Acceptability of Speech and Silent Speech Input Methods in Private and Public

Laxmi Pandey
lpandey@ucmerced.edu
Human-Computer Interaction Group
University of California, Merced
Merced, California, United States

Khalad Hasan
khalad.hasan@ubc.ca
University of British Columbia, Okanagan
Okanagan, BC, Canada

Ahmed Sabbir Arif
asarif@ucmerced.edu
Human-Computer Interaction Group
University of California, Merced
Merced, California, United States

ABSTRACT

Silent speech input converts non-acoustic features like tongue and lip movements into text. It has been demonstrated as a promising input method on mobile devices and has been explored for a variety of audiences and contexts where the acoustic signal is unavailable (e.g., people with speech disorders) or unreliable (e.g., noisy environment). Though the method shows promise, very little is known about peoples' perceptions regarding using it. In this work, first, we conduct two user studies to explore users' attitudes towards the method with a particular focus on social acceptance and error tolerance. Results show that people perceive silent speech as more socially acceptable than speech input and are willing to tolerate more errors with it to uphold privacy and security. We then conduct a third study to identify a suitable method for providing real-time feedback on silent speech input. Results show users find an abstract feedback method effective and significantly more private and secure than a commonly used video feedback method.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *Text input*.

KEYWORDS

silent speech, speech, social acceptance, input and interaction, voice assistant, contactless interaction

ACM Reference Format:
Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445430

1 INTRODUCTION

Speech input on mobile devices continues to evolve at a rapid pace as the speech recognition technologies get better at understanding users' voice commands. This method offers the opportunity for faster and seamless hands-free information access, especially when users' hands are busy performing other tasks or when touching public devices is to be avoided in times like the current COVID-19 situation. Prior research showed that speech input is a viable solution for accessing information on small-screen devices where it allows users to access information faster than traditional on-screen input methods [102]. A major challenge with this method, however, is users' reluctance to use speech in pubic places due to privacy and security concerns [36, 37, 78, 92]. Additionally, voice recognition accuracy is heavily affected by ambient noise [69] and the method is not well supported for people with speech disabilities[1].

Silent speech input, which interprets users' lip and tongue motions into text, has been shown as a promising alternative to speech input [32, 38, 43, 54, 55, 104, 110]. Researchers explored different video-based [2, 13, 27, 28] and advanced sensor-based [83, 84, 95, 113] recognition methods where they showed high accuracy in speech recognition with silent speech input. A recent work [110] explored silent speech input on mobile devices, where users expressed a higher level of satisfaction with this input method over the tradition speech input. In spite of promising results, very little is known on factors such as social acceptance and error tolerance that could influence users' willingness to use this input method. Consequently, the extent to which this input method is viable on mobile devices is an open question.

In this paper, we explore users' attitudes towards speech and silent speech input methods with a focus on social acceptability, and user tolerance of recognition errors in these methods. We first conduct a crowdsourced study examining social acceptance of these methods considering different factors, including users' and viewers' perspectives towards using these in different locations and in front of different audiences. Results show that, in general, people prefer using silent speech input over traditional speech input. Since prior research suggests that silent speech input can be error-prone [33, 82], we conducted another study to explore users' attitude towards recognition errors associated with the two methods. Results reveal that users are willing to tolerate more errors with silent speech input than speech input as it offers a higher degree of privacy and security. Inspired by the findings, we further investigate suitable feedback method for silent speech input. Results show that users find both a commonly used video and an abstract (a blinking dot) feedback effective but the latter significantly more private, more secure, and less intrusive than the video feedback.

To summarize, in this work we: i) explore the social acceptance of speech and silent speech input in different social contexts; ii) investigate user tolerance of recognition errors in the two methods;

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445430

[1] SheSpeaks, https://www.shespeaks.com/why-the-speech-impaired-feel-left-out-of-the-voice-assistant-revolution

iii) identify suitable feedback mechanism for silent speech input; and iv) propose a set of recommendations for using silent speech input on mobile devices.

2 RELATED WORK

This work intersects with four areas of interest: speech input, silent speech input, social acceptance of technology, and users' and viewers' perspectives.

2.1 Speech Input

Speech input enabled devices, such as personal voice assistants, allow users to communicate with computer systems using speech commands. Personal voice assistants like Siri, Google Assistant, Alexa, and Cortana can interpret human speech and handle a wide variety of tasks [53, 72]. Research on speech input mainly focused on the recognition of speech [75, 106, 119], language models [7, 17] and voice controlled systems [121]. Clark et al. [29] provides a comprehensive review of the literature on speech-based input and interaction methods. With the recent advances in speech recognition technology [1, 44, 87, 91, 108], today's voice-based commercial products [23, 48, 62, 116–118, 120] can perform streaming, high-accuracy, low-latency speech recognition [15, 68] to revolutionize human-computer interaction [29]. Recently, He et al. [49] presented an end-to-end speech recognizer for on-device speech recognition using a recurrent neural network, which has been deployed in the default Google keyboard on the flagship Pixel phones. Despite its popularity, studies show privacy and security concerns for the use of personal voice assistants and voice search commands in public places [36, 37, 78, 92]. A survey[2] revealed that 39% smartphone users use the built-in voice assistants at home but only 6-14% use these in public [85]. To uphold the privacy and security of users, researchers explored whisper input, which is a variant of speech input with a significantly lower energy than normal speech. These works detected whispered speech using a stethoscopic microphone that contacts the skin behind the ear [79], a throat microphone [59], and a non-contact microphone by placing it very close to the front of the narrowly opened mouth [40]. Recently, Amazon included a whisper mode to their personal voice assistant Alexa[3]. When users whisper to Alexa, it whispers back to them. Some have also incorporated state-of-the-art machine learning techniques to improve the performance of whisper speech recognition [41, 42, 45]. However, whispers with a much lower acoustic power and relatively flat spectrum than regular speech are inherently noise-like, thus are highly susceptible to acoustic interference [76]. Moreover, long-term use of whisper voice might have negative effects on our vocal cords [103].

2.2 Silent Speech Input

Silent speech input enables users to communicate with a computer system using speech commands without the need for producing any audible sound. Unlike speech input, this method allows users to communicate efficiently with computer systems without hurting privacy and security or disrupting the environment. There have been several previous attempts at achieving silent speech communication. Many have explored silent speech enabled input and interaction methods that use different sensors (e.g., electromagnetic articulography (EMA) [38, 43, 50], electroencephalogram (EEG) [88], electromyography (EMG) [56–58, 74, 105, 115], ultrasound imaging [31, 32, 39, 43, 50, 54, 55, 61], vibrational sensors of glottal activity [83, 84, 95, 113], speech motor cortex implants [18], and non-audible murmur (NAM) microphone [51, 52, 80]) to recover the speech content produced without vibration of the vocal folds, by detecting tongue, facial, and throat movements. Some have developed intracortical microelectrode Brain-Computer Interfaces (BCI) to predict user's intended speech information directly from the brain activities involved in the speech production mechanism [24, 30, 89, 111, 112]. Some have also used multimodal imaging systems for speech recognition, focusing mainly on tongue visualization [55]. A recent work developed a wearable interface that places five EMG sensors above the face to capture the neuromuscular signals for silent speech recognition [60]. Most of these works, however, use invasive, impractical, non-portable setup, impeding their scalability in real-world scenarios.

More recently, attempts have been made to enable silent speech communication using video-based recognition, referred to as lip reading [2, 8, 13, 19, 25, 26, 26–28, 86, 109]. For example, a work provided smartphone users access to their phone functionalities through silent speech commands [110]. It used the front camera of a smartphone to capture the motion of the mouth, then recognized the silently spoken commands using deep-learning-based image sequence recognition technology. These works suggest that video-based silent speech input method could be more user friendly and appropriate in private and public settings since it can be used without any wearable devices. It has the potential to facilitate input and interaction on private devices when the hands are not available, as well as on public devices when direct contact is not recommended in times like the current COVID-19 situation. It can also help people with speech disorder, muteness, and blindness to input and interact with computer systems, increasing their access to technologies.

2.3 Social Acceptance of Technology

Previous research has explored social acceptability for body-based and device-based gestures [97–99, 101], around device input [3], head-mounted display (HMD) input [4], and companion drones for blind people [14] in lab or public settings. In a recent work, Baier and Burmester [16] explored the social acceptability of speech input, which revealed that location influences users' willingness to use the method in public spaces. However, no prior study has explored user attitudes and acceptance of using silent speech input. In a different research, Alallah et al. [5] investigated whether social acceptability studies can be conducted on crowdsourced platforms. They showed that crowdsourced platforms could be an alternative to conducting laboratory-style studies for examining social acceptability. Inspired by this work, we conducted our social acceptability study (Study 1) via crowdsourcing.

Prior research also showed that social acceptability has a significant implication for technological acceptance as they are directly connected to peoples' preferences on using new technologies [63, 114]. To examine the social acceptance of new technologies,

[2]Creative Strategies, https://creativestrategies.com/voice-assistant-anyone-yes-please-but-not-in-public

[3]Digital Trends, https://www.digitaltrends.com/home/how-to-enable-whisper-mode-on-alexa

researchers conducted studies from users' perspective and/or viewers' perspective [3, 4, 97–99]. To investigate users' perspective, researchers either provided participants with a first-hand experience using a new technology or showed them video clips on how the technology could potentially be used [3]. Later, participants were asked to consider themselves as users of the technology and express their opinion on using it in different contexts. While there are many social acceptability studies conducted from the users' perspective, less attention has been paid to examine social acceptance from viewers' standpoints. A few studies investigated social acceptance from the viewers' perspective where researchers elicited opinions from people watching others using a new technology in different contexts. Montero et al. [77] showed that considering viewers viewpoint is important, especially when using the technology in public places, as users' interactions with the technology might draw bystanders' (or the viewers') unwanted attention. Consequently, viewers' perspective are explored for wearable e-textile interface [93, 94], Augmented Reality (AR) in public space [34], and public interfaces (e.g., public performance act) [96]. Additionally, some studies considered both the users' and the viewers' perspectives while evaluating the social acceptance of new technologies, such as gestural interaction on mobile devices [77], head-worn devices [4, 5, 65, 70], data glass [64], and around device input methods [3]. These studies were commonly conducted by examining observers' impression on watching other people interacting with a technology — either in a real-world setting or in a video. In this paper, we examined the social acceptability of speech and silent speech input from both users' and viewers' perspectives.

3 STUDY 1: SOCIAL ACCEPTABILITY

3.1 Input Modalities

Researchers have explored a number of voice and non-voice input modalities to interact with mobile devices. For instance, they investigated using speech and silent speech input methods that range from noticeable to inconspicuous [24, 38, 56, 61, 84, 110, 121]. Speech or voice input, which is commercially available on smartphones, requires users to make voice commands to send instructions to mobile devices. This input modality is explicit and commonly draws co-located observers' attention due to the nature of its input visibility – thus can make users feel awkward or uncomfortable with the presence of nearby users. On the other hand, silent speech input, which recognizes speech without requiring users to make acoustic signals, interprets users' commands on smartphones by tracking tongue and lip movements. This input method is more subtle than the speech input, and used when acoustics is not an option (e.g., speech-impaired people) or it is undesired (e.g., during a confidential conversation or communication in public places). On one hand, using explicit input modalities can convey clear instructions to the devices; however, this form of input might be less socially acceptable due to the visibility to co-located people. On the other hand, subtle inputs are less explicit; however, co-located observers might not readily interpret these commands, making the interaction more acceptable. Therefore, we first conduct a study to explore the social acceptability of these two input modalities.

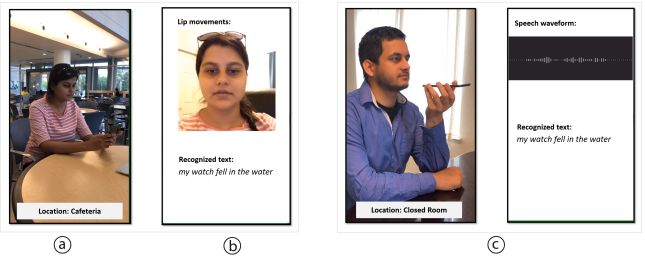

Figure 1: Two example videos used in the survey: (a) a user is interacting with a mobile device with silent speech input in a public place, (b) a video clip showing users lip movements and the recognized text, and (c) another video showing a user using speech input on a mobile device in a private room.

3.2 Crowdsourced Study

As discussed in the related work, researchers explored social acceptability for a wide range of input modalities, such as smartphone gestures [97, 98, 101], around-device interaction [3], and hand-to-face input methods [67, 107]. They used two common approaches: (i) allowing participants to use the technology in a particular context (e.g., public places) and (ii) showing participants videos of how the technique can be used. To collected feedback, participants are commonly asked to imagine using it in other contexts (e.g., workplace) and provide their feedback on a 5-point Likert scale. Due to the spread of COVID-19, we were unable to recruit participants to run a study in a public place. Thus, we used the second approach for our study.

Crowdsourcing platforms have now become increasingly popular to conduct HCI user studies [4, 5]. They provide researchers with an easy access to large and diverse groups of participants. Additionally, these platforms have been considered as cost-efficient solutions to run user studies remotely. Though there has been concern about the data quality from crowdsourced studies, researchers have taken certain measures to remove outliers, which have been almost as effective as laboratory or field studies [4, 5, 20, 46]. Consequently, we decided to use crowdsourcing platforms to run our first study.

3.3 Online Survey

We created an online survey with Qualtrics to collect responses from participants. Figure 2 shows a sample of questions from the survey. We divided the survey questions into four sections: (i) *Demographics*: 14 questions to collect demographic information (e.g., age, gender) and prior experience (e.g., experience with smartphones and voice input) from participants; (ii) *Users' perspectives*: 6 questions asking users to share their experience of using speech and silent speech input methods by considering themselves as users of the modalities; (iii) *Observers' perspectives*: 6 questions were used to explore observers' perspective, i.e., seeing other people using the input modalities and (iv) *Overall preference*: 6 questions asking participants to provide their overall preference of using the input modalities on mobile devices. These questions were designed using both open-ended questions, single/multiple-choice questions, and 5-point Likert scale questions. The open-ended questions were used

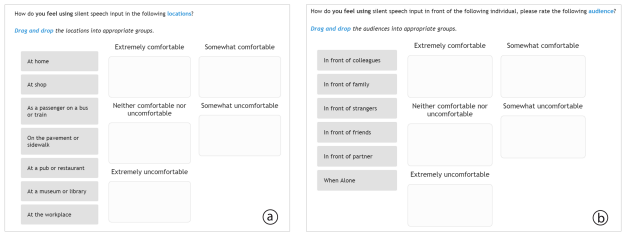

Figure 2: Example of survey questions to collect users feedback on using silent speech input (a) in seven locations; and (b) in front of six audiences.

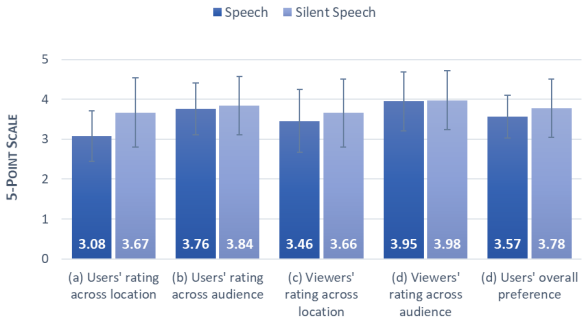

Figure 3: Medians of social acceptability for two input methods from users' perspective across (a) location, (b) audiences and from viewers' perspective across (c) location and (d) audiences, and (e) users' overall preference for two input methods. The error bars represent ±1 standard deviation (SD).

to collect descriptive responses (e.g., justifying their response to a question), while the other types of question were used to collect their preference/perception of using the input modalities and demographic information. When designing the questionnaire, we used similar questions and location-audience contexts used in previous work on social acceptance [3–5, 97, 99]. We also followed many steps listed by Boateng et al. [21], including item generation, context validity, pre-testing with a pilot study, item reductions and others.

Researchers explored a number of ways to measure social acceptabilities of the methods under investigation. One of the commonly used methods is to elicit participants' responses to social acceptability questions through the 'audience-and-location' axes [3–5, 97], where participants are asked to provide their social comfortness of using a method in front of different audiences and locations. Participants commonly respond by indicating how comfortable they were using the method on a 5-point Likert scale – Extremely comfortable, Somewhat comfortable, Neither comfortable nor uncomfortable, Somewhat uncomfortable, and Extremely uncomfortable. Therefore, we used six audiences (i.e., alone, partner, family, friends, colleagues, and strangers) and seven locations (i.e., home, shop, bus or train, pavement or sidewalk, pub or restaurant, museum or library, and workplace) to explore participants' impression of using the two input methods (i.e., speech and silent speech). As participants might not be familiar with a input method, we used a set of video clips showing users using the two methods to interact with a mobile device in two different contexts – in a busy café surrounded by strangers and at home when alone.

3.4 Participants and Study Procedure

To recruit participants, we posted the survey as a task in Amazon Mechanical Turk (AMT), a popular Crowdsourcing platform. All AMT users (i.e., workers) could see the task, however, only the workers who owned a smartphone and had a minimum of 70% approval rate on their previously completed tasks could participate. Workers were compensated with USD $1.50 for their time. We collected data from 109 crowdsourced participants. 62 of them were from the U.S., 6 were from India, 2 were from Brazil, and 1 was from Germany. 8 of them were in the age range of 18–24 years, 28 were in 25–34 years, 18 were in 35–44 years, 10 were in 45–54 years, 5 were in 55–64 years, and 2 were 65 years or older.

The survey was self-paced and the workers were asked to first watch the video clips for an input method, then respond to the questions related to that method. We also clearly instructed them not to relate comfort with physical comfort (e.g., tiredness), rather focus on social and mental aspects of it when providing their responses. Similar strategies were applied in previous studies exploring the social acceptance of new input modalities [3].

As mentioned earlier, data collected from crowdsourcing platforms sometimes raises concerns due to the lack of direct supervision of the workers. Thus, we used the following criteria to remove outliers from our data. (i) Duplicate IP address: we removed any data with the same IP address. This outlier removal technique was also used in prior studies [4, 5]. (ii) Time threshold: as participants were required to watch a set of videos before responding to the questions, they had to spend a minimum time to watch the videos and read and understand the questions before answering them. Consequently, any responses that were submitted within 3 minutes of start were excluded from our analysis. (iii) Incorrect answers: there were a few open-ended questions asking participants to provide justifications for their responses. Any data with incorrect, incomplete, or random answers were rejected. This process excluded in total 38 participants. Hence, we analyzed the data from 71 participants.

3.5 Results

We used non-parametric analyses on the data and, thus, median values are reported. We also report the effect size (r) for the Wilcoxon signed-rank test. Since r for the Friedman test is calculated for pairwise comparison and there is not an agreed method for calculating the confidence interval [100], Kendall's W is most commonly used to assess agreement among the raters. Hence, we report W for the Friedman test. Both r and W use the Cohen's interpretation where 0.1 constitutes a small, 0.3 constitutes a medium, and > 0.5 constitutes a large effect. We aggregated users ratings for each input across all the locations and audiences.

Figure 3 (a) and (b) show the median of social acceptability for each input across locations and audiences, respectively, from users' perspective. A Wilcoxon signed-rank test revealed significant differences between the speech and silent speech input methods across locations ($z = -4.59, p < .05, r = 0.54$). However, we found no significant difference between aggregated values for two input

methods across audiences ($z = -1.36, p = .17, r = 0.16$). Figure 3 (c) and (d) show the median of social acceptability ratings for each input across locations and audiences, respectively, from viewers' perspective. A Wilcoxon signed-rank test showed that silent speech input was significantly different from speech input ($z = -2.5, p < 0.05, r = 0.30$) across locations. However, we did not find any significant difference between two input methods across audiences ($z = -1.14, p = .26, r = 0.14$). We also asked participants to provide their preference for using the two input methods to interact with mobile devices across locations and audiences. Figure 3 (e) shows the results. A Wilcoxon signed-rank test revealed significant differences between speech and silent speech input methods ($z = -3.27, p < .05, r = 0.39$). We recommend caution in interpreting the "not significant" results since they yielded a small effect size ($r < 0.3$).

3.6 Discussion

The results suggest that social acceptability for the two input modalities from users' and viewers' perspectives were different across locations as users considered the less noticeable input method (e.g., silent speech) as their preferred method to interact with mobile devices. Similar findings were revealed in a prior work [5], where they suggested that less noticeable input methods (e.g., ring and touchpad) are more socially acceptable than noticeable ones (e.g., hand gestures) to interact with an HMD. The results also show that participants preferred to use silent speech input over speech input. In subjective feedback, participants expressed their interest in using silent speech input as it is more subtle and provide a high degree of privacy and security than the other method. One participant (male, 35–44 years) commented, *"I would still feel that I have a high level of privacy when using silent input"*. Another participant (female, 35–44 years) wrote, *"I prefer whisper or silent because it doesn't bother others and can be used in quiet places like libraries"*.

Though the results showed users' interest in using silent speech input, there are several key questions remain unknown that could influence their attitude towards using the method. For instance, researchers showed that silent speech input could be prone to high error rates [33, 69, 82, 90]. Consequently, silent speech recognition accuracy could be a key factor in adopting the method. However, little is know of users' error tolerance level for silent speech input. Additionally, silent speech input recognition on mobile devices depends primarily on capturing users tongue and lip movements via the front camera. Thus, providing appropriate real-time feedback on input recognition is critical for the acceptance of the method. Therefore, in the next two studies, we explore error tolerance and suitable feedback mechanism for silent speech input.

4 STUDY 2: ERROR TOLERANCE

Since the survey results revealed that users put much emphasis on privacy and security, we conducted a Wizard-of-Oz study to investigate whether they are willing to compromise the accuracy of an input method for increased privacy and security.

4.1 Apparatus

We developed a custom client/server web application with HTML5 and JavaScript for the Wizard-of-Oz study. The client and server communicated with each other using WebRTC[4]. The client interface looked and felt like the interface depicted in Fig. 1. It was launched on a Google Chrome mobile web browser (v71.0.3578.98) on a Motorola Moto G[5] Plus smartphone (150.2x74x7.7 mm, 155 g) at 1080x1920 pixels. The server was hosted on a HP Pavilion 15 laptop computer running on Linux 16.04 at 1920×1080 pixels. The server interface was launched on a Google Chrome web browser (v74.0.3729.157), which included dedicated buttons for each condition for the researcher (wizard) to display the spoken and silently spoken phrases on the client side. Both devices were connected to a fast and reliable Wi-Fi network. There were no network dropouts during the study.

4.2 Participants

Twelve volunteers from the local university community participated in the user study. Their age ranged from 22 to 25 years (M = 24.25, SD = 1.48). Four of them identified as women and eight as men. They were all experienced smartphone (at least 5 years of experience, M = 7.25, SD = 1.48) and voice assistant (at least one year of experience, M = 2.5 years, SD = 0.65) users. Most of them used multiple voice assistants, including Alexa, Cortana, Google Assistant, and Siri. Two participants used these voice assistants almost every day, eight of them used these occasionally, and the remaining two rarely used these.

4.3 Design

The study used a within-subjects design. The independent variables were *method* and *injected error rate* and the dependent variables were the qualitative metrics. In summary, the design was:

> 12 participants ×
> 2 methods (speech and silent speech, counterbalanced) ×
> 5 injected error rates (0%, 5%, 10%, 15%, and 20%, randomized) ×
> 12 phrases from the MacKenzie and Soukoreff [73] set = 1,440 phrases, in total.

4.4 Error Injection

Injected errors are commonly used in text entry research to study the effect of errors on performance and preference [6, 9, 11, 66]. In the study, we injected 0%, 5%, 10%, 15%, and 20% misrecognition errors. A misrecognition error occurs when the recognizer incorrectly recognizes a word [12], for example, "take a coffee break" ("coffee" was replaced with "toffee"). The total number of misrecognition errors in a condition was calculated using the following equation: $(w \times e)/100$, where w is the total number of words in *all* presented phrases in the condition and e is the target error rate. We injected errors at word level since both speech and silent speech methods work at either word or phrase level. To inject errors, we randomly replaced a word consisting more than three letters with a similar sounding word, excluding the first word. To assure that all participants encountered the same errors, we randomly pre-selected a subset of phrases from the MacKenzie and Soukoreff [73] set, then used those with the methods in a counterbalanced order. The error

[4]Real-time communication for the web, https://webrtc.org

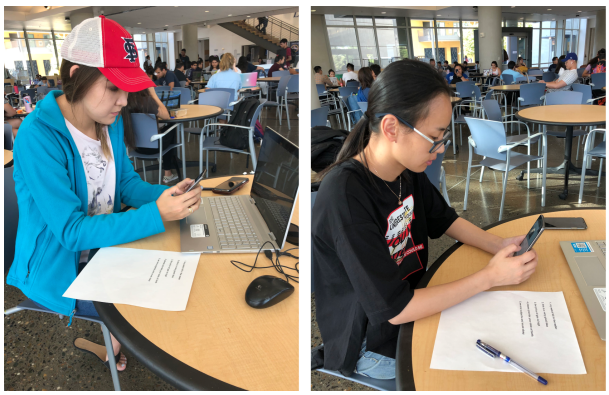

Figure 4: Two participants taking part in the second study at a cafeteria.

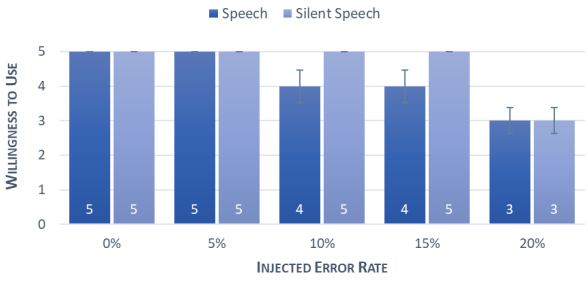

Figure 5: Median willingness to use ratings for speech and silent speech with the five injected error rates on a 5-point Likert scale, where where 1 to 5 represented Very unlikely to Very likely. The error bars represent ±1 standard deviation (SD).

injection rates were selected based on the findings of a prior investigation the reported that user performance tend to drop significantly when error rate of an input method reaches 20% [12].

4.5 Procedure

We conducted a Wizard-of-Oz study to control the error rate in each condition. Before the study, participants were told that the purpose of the study was to compare the performance of multiple speech and silent speech recognition methods that may vary in accuracy rate. The study took place at a campus cafeteria. We picked a public place for the study since its purpose was to investigate whether users were willing to tolerate more errors for the sake of increased privacy and security. Note that the survey results suggested that users are likely to be more conscious about their privacy and security when in public. Upon arrival, we demonstrated the speech and silent speech methods on the smartphone and explained the study procedure to each participant. We then collected their consents. The study started after that, where participants were instructed to enter short English phrases from the MacKenzie and Soukoreff [73] set using either speech or silent speech at varying injected error rates. The methods were counterbalanced and the error rates were randomly injected to mitigate any potential learning effects. The interface displayed one phrase at a time. Participants were instructed to tap on the screen when they were done speaking or silently speaking the phrase. They all sat at a table in the cafeteria (Fig. 4). A researcher (the wizard) sat at a nearby table with the server interface launched on a laptop computer. Upon completion of each phrase, she pressed a key to display the recognized phrase and the next phrase on the smartphone. Participants were asked to speak or silently speak a phrase again when the phrase contained a misrecognized word. Upon completion of each condition (method × injected error rate), participants completed a short questionnaire that asked them to rate their willingness to use the examined methods on a 5-point Likert scale. Upon completion of the complete study, they completed the NASA-TLX questionnaire [81] to rate the methods' perceived workload. We then held a debrief session to explain the study's actual purpose. A complete study session took about 60 minutes.

4.6 Results

We used non-parametric analyses on the data, thus report median values. We also report the effect size r and Kendall's W for the Wilcoxon signed-rank and Friedman tests, respectively (see Section 3.5).

4.6.1 Willingness to Use. A Friedman test identified a significant effect of condition on willingness to use ($\chi^2(9) = 94.04, p < .0001, r = 0.87$). There was a significant effect of injected error rate on willingness to use for both the speech ($\chi^2(4) = 38.06, p < .0001$) and silent speech ($\chi^2(4) = 48.00, p < .0001$) methods. A Dunn's multiple comparisons test identified a significant difference in willingness to use between the methods with both 10% ($z = 2.75, p < .05$) and 15% ($z = 2.83, p < .05$) error rates. Fig. 5 illustrates median willingness to use for both methods with the five injected error rates.

4.6.2 Perceived Workload. A Wilcoxon Signed-Rank test identified a significant effect of method on temporal demand ($z = -1.1, p < .05, r = 0.61$) and overall performance ($z = -2.24, p < .05, r = 0.65$). However, no significant effect was identified on mental demand ($z = -1.93, p = .05, r = 0.55$), physical demand ($z = -0.93, p = .35, r = 0.27$), effort ($z = -1.45, p = .15, r = 0.42$), or the level of frustration ($z = -0.99, p = .32$). Fig. 6 illustrates median Raw TLX (RTLX) scores for both methods. We analyzed the subscales individually, which is a common modification made to NASA-TLX [47]. Note that the evidence is inconclusive about whether RTLX is more sensitive, less sensitive, or equally sensitive compared to the original version, thus Hart [47] left it to the researchers' discretion.

4.7 Discussion

Results revealed that 0% and 5% error rates yielded the highest and 20% error rate yielded the lowest willingness to use ratings for both methods. This is not surprising since prior investigations reported that user performance with an input method is the best between 0% and 5% error rates, slightly drops between 5% and 10% error rates, and the worst at 20% error rate [10, 12]. Interestingly, for 10% and 15% error rates, the willingness to use ratings for speech dropped at a higher rate that silent speech (Fig. 5). A post hoc analysis failed to identify a significant difference between 0–5% and

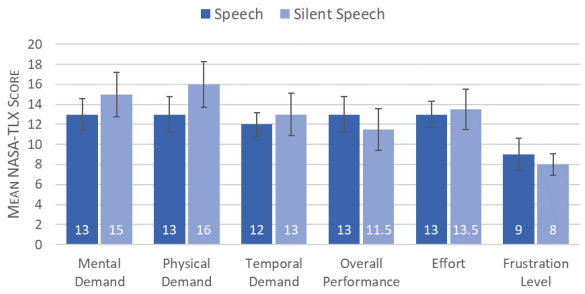

Figure 6: Median RTLX scores of the workload related to speech and silent speech methods. The error bars represent ±1 standard deviation (SD).

10–15% error rates for silent speech, while these two groups were significantly different for speech. This suggests that users were willing to tolerate more errors in silent speech. When asked about this during the debrief session, all participants (100%) responded that it was mostly due to concerns about their privacy and security. They feared that speech will violate their privacy and security in public places, especially when they are surrounded by unknown people. One participant (female, 22 years) commented, *"Sometimes, I feel very hesitant to type with my voice publicly because I always feel that someone else is listening to me"*. In contrast, participants felt that silent speech is more private and more secure, thus were willing to compromise accuracy to some extent. One participant (male, 23 years) commented, *"[Silent speech] is very useful for sharing important information in public"*.

There was a significant difference in temporal demand and overall performance for the two methods. Most participants felt that silent speech required more time to use than speech (Fig. 6). The debrief session revealed that it was because participants silently spoke the phrases at a much slower rate than speech assuming that it will increase the method's accuracy (although in reality it had no effect since we used a Wizard-of-Oz setup). This also significantly affected their overall rating of the method. There was no significant difference in mental demand, physical demand, effort, and frustration. However, we recommend caution in interpreting these results since in the study participants used the methods while seated at a table. Although we did not instruct them on how to hold the device, they all held the device with both hands for clear view of the interface (Fig. 1) and rested their elbow on the table for comfort (Fig. 4). Hence, the results may differ when the methods are evaluated in a standing position or while walking.

5 STUDY 3: VISUAL FEEDBACK

Providing appropriate feedback on the system status is the key usability principle while designing any system. Efficient visual feedback helps users to interpret the system status correctly, enabling them to access information rapidly and accurately [71]. However, designing effective visual feedback for mobile devices is challenging due to their limited display space. Besides, some participants of Study 2 commented that the video feedback method occupies much of the smartphone real estate, leaving a little or no space for additional input and interaction tasks (Fig. 1). We, therefore, conducted a user study to find out whether it is feasible to replace the commonly used video feedback with a more compact, abstract feedback method.

5.1 Apparatus

We used the same client/server architecture as the last study, but with an updated user interface (Fig. 7). Further, we hosted the app on GitHub[5] to enable people outside the campus network access the client. Six participants used Apple iOS-based smartphones, while the remaining six used Android-based smartphones. Ten of them used a Google Chrome mobile web browser (> v84), while the remaining two used a Safari browser (> v85) to access the client app. The wizard used a Microsoft Surface Book 3 (34.3 cm display, i7 CPU at 1.90GHz, 16GB RAM) to launch the server interface on a Google Chrome web browser (v85.0.4183.102). We did not record any network dropouts during the study.

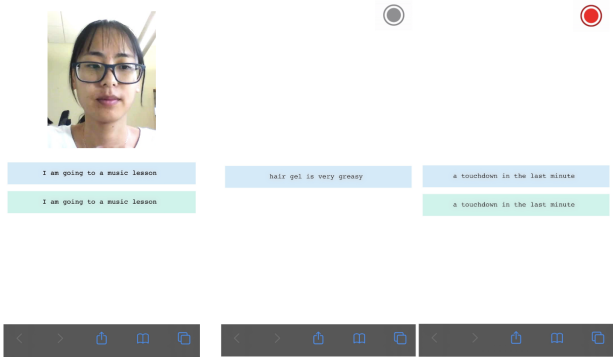

Figure 7: The two visual feedback methods used in the study: (1) video feedback that always displays the video captured by the device's front-facing camera on the screen (left) and (2) abstract feedback that displays a grey or a blinking red dot at the top right corner of the device based on whether the camera can see the lips or not, respectively (right).

5.2 Feedback Methods

We implement the following two types of visual feedback:

- **Abstract feedback.** The abstract feedback method is designed to provide minimal feedback on silent speech input. For this, we used a grey dot at the top right corner of the device that turns red and starts blinking when the system tracks the lips (similar to the video recording button on most mobile device). The dot turns grey and stops blinking when the device is unable to see the lips. We use this feedback as it offers a higher level of privacy (does not show users' face or lips) and use minimum screen space on the device.
- **Video feedback.** The video feedback method provides detailed information about users' lip by showing the video captured by the device's front-facing camera. We place the video on the screen as constant feedback to users about the

[5] GitHub Pages, https://pages.github.com

systems status. Though this form of feedback provides precise information on whether the camera can see users' lips, it consumes a considerable portion of the screen real-estate.

5.3 Participants

Twelve participants (6 female, 6 male) aged 23 to 34 years (M = 28.75, SD = 2.89) participated in this study. All the participants reported being right-handed, using smartphones for the last 8.58 years (SD = 2.29), and using at least one voice assistant system for 2.26 years (SD = 2.24). None of the participants had prior experience using silent speech input. Note that none of the participants participated in the previous studies.

5.4 Error Injection

We injected errors in this study for two reasons. First, to increase the validity of the study since none of the current recognition systems are 100% accurate. Besides, a fully accurate system would have altered some participants about the Wizard-of-Oz setup. Second, to investigate whether users perceive the frequency in which errors occur differently with different feedback methods. For error injection, we used the same approach as the previous study. However, here we maintained a constant 5% error rate over all sessions and injected tracking error rather than misrecognition error. The 5% error rate was chosen as it was found to be an acceptable error rate in various text entry system [6, 9, 11]. A tracking error occurs when the system fails to track the lips because they are out of sight or range, or due to technical issues, resulting in missing words in the final text, for example, "take it to the recycling depot" ("recycling" is removed). We injected tracking error since the purpose of visual feedback on a recognition system is usually to inform users that it is receiving the tracking signals. Hence, tracking error is more appropriate to evaluate the efficiency of visual feedback than misrecognition error.

5.5 Design

The study used a within-subjects design. The independent variables was *feedback* and the dependent variables were the qualitative metrics. In summary, the design was:

12 participants ×
2 feedback methods (video and abstract, counterbalanced) ×
30 phrases from MacKenzie & Soukoreff set [73] with 5% injected error = 720 phrases, in total.

5.6 Procedure

The study was conducted remotely due to the spread of COVID-19. We scheduled a video call with each participant ahead of time. They were told that the purpose of the study was to evaluate two different types of visual feedback on a working silent speech recognizer. They were instructed to join the call from a quiet room to avoid any interference during the study. A researcher (the wizard) demonstrated the system and the feedback methods, explained tracking error (that the inability to track the lips results in missing words in the recognized phrase), collected their consents and demographics, and provided all instructions via the video call. The researcher provided the participants with a link to the client app, which they accessed on their smartphone using their preferred web browser. They were instructed to activate the airplane mode but keep the Wi-Fi enabled to avoid any interruptions due to incoming calls. The system displayed one phrase at a time. Participants were asked to silently speak the phrase then tap on the screen to see the recognition and the next phrase. The researcher displayed the recognized phrase and updated the presented phrase using the server interface. We did not instruct the participants on how to hold the device but informed them that the blinking red dot will turn grey when the system cannot track the lips during the graphical feedback condition. The researcher observed all interactions with the smartphone to manually turn the blinking red dot to grey when the front-facing camera is unlikely to capture the lips due to the holding posture or angle. Error correction was not required in this study. Upon completion of the study, participants completed a short questionnaire that asked them to rate various aspect of the two feedback methods on a 5-point Likert scale. We then held a debrief session to inform the participants about the actual nature of the study. The complete study session was recorded using a screen recorder.

5.7 Results

We used non-parametric analyses on the data, thus report median values. We also report the effect size r for the Wilcoxon signed-rank test.

A Wilcoxon signed-rank test identified a significant effect of feedback on whether the method provides enough details about lip detection ($z = -2.06, p < .05, r = 0.6$), occludes, interrupts, and interferes with the task at hand ($z = -2.84, p < .01, , r = 0.82$), and compromise privacy and security ($z = -2.41, p < .05, r = 0.7$). However, there was no significant effect on effectiveness ($z = -0.30, p = .76, r = 0.09$), perceived speed ($z = -1.34, p = .18, r = 0.39$), perceived accuracy ($z = -0.71, p = .48, r = 0.2$), or the overall preference ($z = -1.56, p = .12, r = 0.45$). Fig. 8 illustrates median ratings of all aspects of the two feedback methods.

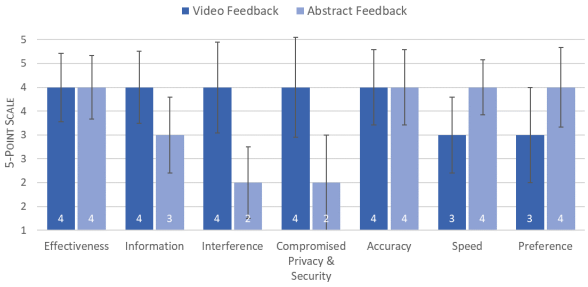

Figure 8: Median ratings of various aspects of the two feedback methods on a 5-point Likert scale, where where 1 to 5 represented Strongly disagree to Strongly agree. The error bars represent ±1 standard deviation (SD).

5.8 Discussion

Participants found both feedback methods equally effective. They found the video feedback significantly more informative than abstract feedback. This is not surprising since video feedback

displayed a real-time video captured by the device's front-facing camera. Interestingly, participants found the abstract feedback to be the least intrusive (does not occlude, interrupt, or interfere with the task at hand) and most private and secure (does not compromise the user's privacy and security). Once participant (female, 31 years) commented, *"I have privacy concerns with video feedback, I don't want to see my phone camera on when using apps all the time"*. Another participant (male, 27 years) wrote, *"In my opinion, the video feedback mode will always gonna be a concern for my privacy and security"*. In terms of willingness to use, participants were slightly leaning towards the abstract feedback, but this difference was not statistically significant (medium effect size). This is not necessarily a bad thing since it can be interpreted as, users are impartial about the methods, thus using an abstract feedback method is an acceptable design choice. Participants found both methods to be equally reliable (did not compromise accuracy), but interestingly they felt the system with video feedback was slower (statistically not significant) although both used the same Wizard-of-Oz setup. We speculate this is because participants were looking at the video while speaking, which increased the mental demand due to information processing, giving them the impression that it was slower. One limitation of these findings is the lack of generalizability in terms of personality, culture, and ethnic background. Although, the study questionnaire used questions from the SUS questionnaire [22] and custom questions prepared following the Dix et al. [35] guideline, they were not formally validated for the effects of personality, culture, and ethnic background.

6 FINAL REFLECTION

Our general intuition may provide initial guidance regarding speech and silent speech input that the latter is likely to be more acceptable than the former due to the nature of the method (it is subtle and less visible). However, without empirical data, it is difficult to come to a conclusion as users' perception towards using the method might be influenced by various factors, such as where they are using the method, in front of whom they are using it, and their acceptance towards the errors committed by the methods. The study results confirm that silent speech input is more socially acceptable as it is subtle, more secure, and less attention-seeking than speech input. Moreover, our results affirm that users are willing to accept more recognition errors with silent speech input than speech input. This is primarily due to the fact that the method is more private, secure, and does not trigger feelings of discomfort. Consequently, users expressed their intention to use the method even with a higher rate of errors than speech input. However, they also showed their preference in limiting the error rate within a reasonable threshold (e.g., 5–10%) for both input methods. We also observed that there is a possible linkage between perceived privacy and security and feedback design for silent speech input. Though video feedback provides users with detailed information (e.g., whether lip movements are captured by the camera), participants expressed their concerns about using this feedback method as it may operate in an always-on manner, continually tracking and analyzing lip movements from the camera. These results further confirm users' strong intention to ensure a high level of privacy and security while inputting on mobile devices.

7 LIMITATIONS AND FUTURE WORK

In this paper, we took a step toward understanding users' perception about using silent speech input method from social acceptance, error tolerance, and feedback design perspectives. While an in-the-wild study would have provided further insights into these issues in more realistic usage contexts, due to the COVID-19 pandemic, it was not an option available to us. Our results encourage a further exploring on these issues in an in-the-wild study. For Study 2 and 3, we recruited participants from a western country which limits the generalizability of the data across different culture and ethnic background. We acknowledge that a larger and more diverse sample would have further affirmed the findings. Additionally, in Study 3, we investigated only one type of abstract feedback (e.g., blinking dot) for silent speech input, leaving out other possible abstract feedback (e.g., sinusoid icons) that could also influence users' impression towards silent speech input. Further investigation is needed to identify any differences or similarities between a wider range of feedback methods. Last but not least, our studies were conducted with Wizard-of-Oz mimicking a mobile silent speech input method. Hence, we were unable to study other technical factors (e.g., silent speech processing delay) that could have affected users' willingness to use the method. It would be interesting to develop an app to enable silent speech input on mobile devices to perform a longitudinal study examining users' perception towards the input modality.

8 CONCLUSION

In this paper, we investigated users' impression towards using silent speech input method on mobile devices from social acceptance, error tolerance, and feedback design perspectives. In a crowdsourced survey, we found out that in general people preferred using silent speech input over the traditional speech input. We also observed that users were more comfortable using silent speech input in different public and private locations but expressed their concerns about input recognition, privacy, and security issues. Consequently, we conducted a study examining users' error tolerance with both input methods, where results revealed their willingness to tolerate more errors for the sake of privacy and security. In the final study exploring suitable feedback for silent speech input, we observed that users found both a video and an abstract feedback methods effective. Yet, they found the latter to be significantly more private and secure than the commonly used video feedback. We learned that designing solutions for silent speech input requires careful consideration of various factors and privacy concerns as well as people's tolerance towards using it on mobile devices.

REFERENCES

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 10 (Oct. 2014), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Deep Lip Reading: A Comparison of Models and an Online Application. *arXiv:1806.06053 [cs]* (June 2018). http://arxiv.org/abs/1806.06053 arXiv: 1806.06053.

[3] David Ahlström, Khalad Hasan, and Pourang Irani. 2014. Are You Comfortable Doing That? Acceptance Studies of around-Device Gestures in and for Public Settings. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/2628363.2628381

[4] Fouad Alallah, Ali Neshati, Yumiko Sakamoto, Khalad Hasan, Edward Lank, Andrea Bunt, and Pourang Irani. 2018. Performer vs. Observer: Whose Comfort Level Should We Consider When Examining the Social Acceptability of Input Modalities for Head-Worn Display?. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (Tokyo, Japan) *(VRST '18)*. Association for Computing Machinery, New York, NY, USA, Article 10, 9 pages. https://doi.org/10.1145/3281505.3281541

[5] Fouad Alallah, Ali Neshati, Nima Sheibani, Yumiko Sakamoto, Andrea Bunt, Pourang Irani, and Khalad Hasan. 2018. Crowdsourcing vs Laboratory-Style Social Acceptability Studies? Examining the Social Acceptability of Spatial User Interactions for Head-Worn Displays. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173884

[6] Ohoud Alharbi, Ahmed Sabbir Arif, Wolfgang Stuerzlinger, Mark D. Dunlop, and Andreas Komninos. 2019. WiseType: A Tablet Keyboard with Color-Coded Visualization and Various Editing Options for Error Correction. In *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019* (Kingston, Canada) *(GI'19)*. Canadian Human-Computer Communications Society, Waterloo, CAN, Article 4, 10 pages. https://doi.org/10.20380/GI2019.04

[7] Cyril Allauzen and Michael Riley. 2011. Bayesian Language Model Interpolation for Mobile Speech Input. In *Interspeech 2011*. 1429–1432.

[8] Ibrahim Almajai, Stephen Cox, Richard Harvey, and Yuxuan Lan. 2016. Improved Speaker Independent Lip Reading Using Speaker Adaptive Training and Deep Neural Networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2722–2726. https://doi.org/10.1109/ICASSP.2016.7472172 ISSN: 2379-190X.

[9] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2010. Predicting the Cost of Error Correction in Character-Based Text Entry Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/1753326.1753329

[10] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2010. Predicting the Cost of Error Correction in Character-Based Text Entry Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/1753326.1753329

[11] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2014. User Adaptation to a Faulty Unistroke-Based Text Entry Technique by Switching to an Alternative Gesture Set. In *Proceedings of Graphics Interface 2014* (Montreal, Quebec, Canada) *(GI '14)*. Canadian Information Processing Society, CAN, 183–192.

[12] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2014. User Adaptation to a Faulty Unistroke-Based Text Entry Technique by Switching to an Alternative Gesture Set. In *Proceedings of Graphics Interface 2014 (GI '14)*. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 183–192. http://dl.acm.org/citation.cfm?id=2619648.2619679

[13] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. LipNet: End-to-End Sentence-level Lipreading. *arXiv:1611.01599 [cs]* (Dec. 2016). http://arxiv.org/abs/1611.01599 arXiv: 1611.01599.

[14] Mauro Avila Soto and Markus Funk. [n.d.]. Look, a guidance drone! Assessing the Social Acceptability of Companion Drones for Blind Travelers in Public Spaces. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2018-10-08) *(ASSETS '18)*. Association for Computing Machinery, 417–419. https://doi.org/10.1145/3234695.3241019

[15] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-End Attention-Based Large Vocabulary Speech Recognition. *arXiv:1508.04395 [cs]* (March 2016). http://arxiv.org/abs/1508.04395 arXiv: 1508.04395.

[16] Monique Faye Baier and Michael Burmester. 2019. Not Just About the User: Acceptance of Speech Interaction in Public Spaces. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 349–359. https://doi.org/10.1145/3340764.3340801

[17] Brandon Ballinger, Cyril Allauzen, Alexander Gruenstein, and Johan Schalkwyk. 2010. On-Demand Language Model Interpolation for Mobile Speech Input. In *Interspeech*. 1812–1815.

[18] Jess Bartels, D. Andreasen, P. Ehirim, Hui Mao, and P. Kennedy. 2008. Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex. *Journal of Neuroscience Methods* (2008). https://doi.org/10.1016/j.jneumeth.2008.06.030

[19] Helen L. Bear and Richard Harvey. 2019. Alternative Visual Units for an Optimized Phoneme-Based Lipreading System. 18 (2019), 3870. https://doi.org/10.3390/app9183870

[20] Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The Viability of Crowdsourcing for Survey Research. *Behavior Research Methods* 43, 3 (March 2011), 800. https://doi.org/10.3758/s13428-011-0081-0

[21] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in public health* 6 (2018), 149.

[22] John Brooke. 1996. SUS: A Quick and Dirty Usability Scale. *Usability evaluation in industry* (1996), 189.

[23] Christopher Ralph Brown. 2008. Automatic Pruning of Grammars in a Multi-Application Speech Recognition Interface. https://patents.google.com/patent/US20080059195/en

[24] Jonathan S. Brumberg, Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. 2010. Brain-Computer Interfaces for Speech Communication. *Speech Communication* 52, 4 (April 2010), 367–379. https://doi.org/10.1016/j.specom.2010.01.001

[25] Joon Son Chung and Andrew Zisserman. 2016. Out of Time: Automated Lip Sync in the Wild. In *ACCV Workshops*. https://doi.org/10.1007/978-3-319-54427-4_19

[26] Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in Profile. In *BMVC*. https://doi.org/10.5244/C.31.155

[27] Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in the Wild. In *Computer Vision – ACCV 2016 (Lecture Notes in Computer Science)*, Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.). Springer International Publishing, Cham, 87–103. https://doi.org/10.1007/978-3-319-54184-6_6

[28] Joon Son Chung and Andrew Zisserman. 2018. Learning to Lip Read Words by Watching Videos. *Computer Vision and Image Understanding* 173 (Aug. 2018), 76–85. https://doi.org/10.1016/j.cviu.2018.02.001

[29] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and et al. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (Jun 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[30] Charles S. DaSalla, Hiroyuki Kambara, Yasuharu Koike, and Makoto Sato. 2009. Spatial Filtering and Single-Trial Classification of Eeg During Vowel Speech Imagery. In *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology (i-CREATe '09)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/1592700.1592731

[31] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone. 2006. Prospects for a Silent Speech Interface Using Ultrasound Imaging. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2006.1660033 ISSN: 2379-190X.

[32] B. Denby and M. Stone. 2004. Speech Synthesis from Real Time Ultrasound Images of the Tongue. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–685. https://doi.org/10.1109/ICASSP.2004.1326078 ISSN: 1520-6149.

[33] Li Deng and Xuedong Huang. 2004. Challenges in Adopting Speech Recognition. *Commun. ACM* 47, 1 (Jan. 2004), 69–75. https://doi.org/10.1145/962081.962108

[34] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. [n.d.]. *In situ with bystanders of augmented reality glasses: perspectives on recording and privacy-mediating technologies*. Association for Computing Machinery, 2377–2386. https://doi.org/10.1145/2556288.2557352

[35] Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale. 2003. *Human-Computer Interaction (3rd Edition)*. Prentice-Hall, Inc., USA.

[36] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. 2014. Voice Activated Personal Assistant: Acceptability of Use in the Public Space. In *Human Interface and the Management of Information. Information and Knowledge in Applications and Services (Lecture Notes in Computer Science)*, Sakae Yamamoto (Ed.). Springer International Publishing, Cham, 324–334. https://doi.org/10.1007/978-3-319-07863-2_32

[37] Christos Efthymiou and M. Halvey. 2016. Evaluating the Social Acceptability of Voice Based Smartwatch Search. In *AIRS*. https://doi.org/10.1007/978-3-319-48051-0_20

[38] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman. 2008. Development of a (silent) Speech Recognition System for Patients Following Laryngectomy. *Medical Engineering & Physics* 30, 4 (May 2008), 419–425. https://doi.org/10.1016/j.medengphy.2007.05.003

[39] Victoria M. Florescu, L. Crevier-Buchman, B. Denby, T. Hueber, Antonia Colazo-Simon, Claire Pillot-Loiseau, P. Roussel-Ragot, C. Gendrot, and S. Quattrocchi. 2010. Silent Vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface. In *INTERSPEECH*.

[40] Masaaki Fukumoto. 2018. Silentvoice: Unnoticeable Voice Input by Ingressive Speech. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/3242587.3242603

[41] Shabnam Ghaffarzadegan, Hynek Bořil, and John H. Hansen. 2017. Deep Neural Network Training for Whispered Speech Recognition Using Small Databases and Generative Model Sampling. *International Journal of Speech Technology* 20, 4 (Dec. 2017), 1063–1075. https://doi.org/10.1007/s10772-017-9461-x

[42] Shabnam Ghaffarzadegan, Hynek Bořil, and John H. L. Hansen. 2016. Generative Modeling of Pseudo-Whisper for Robust Whispered Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 10 (Oct. 2016), 1705–1720. https://doi.org/10.1109/TASLP.2016.2580944 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[43] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green. 2010. Isolated Word Recognition of Silent Speech Using Magnetic Implants and Sensors. *Medical Engineering & Physics* 32, 10 (Dec. 2010), 1189–1197. https://doi.org/10.1016/j.medengphy.2010.08.011

[44] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947 ISSN: 2379-190X.

[45] Đorđe T. Grozdić and Slobodan T. Jovičić. 2017. Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (Dec. 2017), 2313–2322. https://doi.org/10.1109/TASLP.2017.2738559 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[46] Anja S. Göritz, Kathrin Borchert, and Matthias Hirth. 2019. Using Attention Testing to Select Crowdsourced Workers and Research Participants. *Social Science Computer Review* (June 2019), 0894439319848726. https://doi.org/10.1177/0894439319848726 Publisher: SAGE Publications Inc.

[47] Sandra G Hart. 2006. NASA-task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[48] Hideki Hashimoto, Yoshifumi Nagata, Shigenobu Seto, Yoichi Takebayashi, Hideaki Shinchi, and Koji Yamaguchi. 1997. Speech Recognition Interface System Suitable for Window Systems and Speech Mail Systems. https://patents.google.com/patent/US5632002/en

[49] Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein. 2019. Streaming End-to-End Speech Recognition for Mobile Devices. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6381–6385. https://doi.org/10.1109/ICASSP.2019.8682336 ISSN: 2379-190X.

[50] Panikos Heracleous and Norihiro Hagita. 2011. Automatic Recognition of Speech Without Any Audio Information. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2392–2395. https://doi.org/10.1109/ICASSP.2011.5946965 ISSN: 2379-190X.

[51] Panikos Heracleous, Tomomi Kaino, H. Saruwatari, and K. Shikano. 2007. Unvoiced Speech Recognition Using Tissue-Conductive Acoustic Sensor. *EURASIP J. Adv. Signal Process.* (2007). https://doi.org/10.1155/2007/94068

[52] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-Speech Enhancement Using Body-Conducted Vocal-Tract Resonance Signals. *Speech Communication* 52, 4 (April 2010), 301–313. https://doi.org/10.1016/j.specom.2009.12.001

[53] Matthew B. Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly* 37, 1 (Jan. 2018), 81–88. https://doi.org/10.1080/02763869.2018.1404391 Publisher: Routledge _eprint: https://doi.org/10.1080/02763869.2018.1404391.

[54] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. 2007. Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 1. I–1245–I–1248. https://doi.org/10.1109/ICASSP.2007.366140 ISSN: 2379-190X.

[55] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips. *Speech Communication* 52, 4 (April 2010), 288–300. https://doi.org/10.1016/j.specom.2009.11.004

[56] Charles Jorgensen and Sorin Dusan. 2010. Speech Interfaces Based Upon Surface Electromyography. *Speech Communication* 52, 4 (April 2010), 354–366. https://doi.org/10.1016/j.specom.2009.11.003

[57] C. Jorgensen, D.D. Lee, and S. Agabont. 2003. Sub Auditory Speech Recognition Based on Emg Signals. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, Vol. 4. 3128–3133 vol.4. https://doi.org/10.1109/IJCNN.2003.1224072 ISSN: 1098-7576.

[58] S. Jou, Tanja Schultz, Matthias Walliczek, F. Kraft, and Alexander H. Waibel. 2006. Towards Continuous Speech Recognition Using Surface Electromyography. In *INTERSPEECH*.

[59] Szu-Chen Jou and et al. 2004. *Adaptation for Soft Whisper Recognition Using a Throat Microphone.*

[60] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[61] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[62] Peter F. King. 2003. Server Based Speech Recognition User Interface for Wireless Devices. https://patents.google.com/patent/US6532446B1/en

[63] Marion Koelle, Swamy Ananthanarayan, and Susanne Boll. 2020. Social Acceptability in HCI: A Survey of Methods, Measures, and Design Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3313831.3376162

[64] Marion Koelle, Abdallah El Ali, Vanessa Cobus, Wilko Heuten, and Susanne CJ Boll. 2017. All about Acceptability? Identifying Factors for the Adoption of Data Glasses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 295–300. https://doi.org/10.1145/3025453.3025749

[65] Marion Koelle, Matthias Kranz, and Andreas Möller. 2015. Don't Look at Me That Way! Understanding User Attitudes Towards Data Glasses Usage. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Copenhagen, Denmark) *(MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 362–372. https://doi.org/10.1145/2785830.2785842

[66] Andreas Komninos, Emma Nicol, and Mark Dunlop. 2020. Investigating Error Injection to Enhance the Effectiveness of Mobile Text Entry Studies of Error Behaviour. arXiv:2003.06318 [cs.HC]

[67] DoYoung Lee, Youryang Lee, Yonghwan Shin, and Ian Oakley. 2018. Designing Socially Acceptable Hand-to-Face Input. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 711–723. https://doi.org/10.1145/3242587.3242642

[68] Xin Lei, A. Senior, A. Gruenstein, and Jeffrey Scott Sorensen. 2013. Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices. In *INTERSPEECH*.

[69] Ning Li, Tuoyang Zhou, Yingwei Zhou, Chen Guo, Deqiang Fu, Xiaoqing Li, and Zijing Guo. 2019. Research on Human-Computer Interaction Mode of Speech Recognition Based on Environment Elements of Command and Control System. In *2019 5th International Conference on Big Data and Information Analytics (BigDIA)*. 170–175. https://doi.org/10.1109/BigDIA.2019.8802812

[70] Andrés Lucero and Akos Vetek. [n.d.]. NotifEye: using interactive glasses to deal with notifications while walking in public. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology* (New York, NY, USA, 2014-11-11) *(ACE '14)*. Association for Computing Machinery, 1–10. https://doi.org/10.1145/2663806.2663824

[71] Patrick J. Lynch. 1994. Visual Design for the User Interface, Part 1: Design Fundamentals. *Journal of Biocommunication* 21 (1994), 22–22. http://trantor.sheridanc.on.ca/sys32a1/manual/appendix/gui1.html

[72] Gustavo López, Luis Quesada, and Luis A. Guerrero. 2018. Alexa Vs. Siri Vs. Cortana Vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In *Advances in Human Factors and Systems Interaction (Advances in Intelligent Systems and Computing)*, Isabel L. Nunes (Ed.). Springer International Publishing, Cham, 241–250. https://doi.org/10.1007/978-3-319-60366-7_23

[73] I. Scott MacKenzie and R. William Soukoreff. [n.d.]. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2003-04-05) *(CHI EA '03)*. Association for Computing Machinery, 754–755. https://doi.org/10.1145/765891.765971

[74] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. 2005. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* 331–336. https://doi.org/10.1109/ASRU.2005.1566521

[75] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Haşim Sak, Alexander Gruenstein, Françoise Beaufays, and Carolina Parada. 2016. Personalized Speech Recognition on Mobile Devices. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5955–5959. https://doi.org/10.1109/ICASSP.2016.7472820 ISSN: 2379-190X.

[76] I. Mcloughlin, J. Li, and Yan Song. 2013. Reconstruction of Continuous Voiced Speech from Whispers. In *INTERSPEECH*.

[77] Calkin S. Montero, Jason Alexander, Mark T. Marshall, and Sriram Subramanian. [n.d.]. Would you do that? understanding social acceptance of gestural interfaces. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services* (New York, NY, USA, 2010-09-07) *(MobileHCI '10)*. Association for Computing Machinery, 275–278. https://doi.org/10.1145/1851600.1851647

[78] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. 2015. Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *International Journal of Human–Computer Interaction* 31, 4 (April 2015), 307–335. https://doi.org/10.1080/10447318.2014.986642 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2014.986642.

[79] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 5. V–708. https://doi.org/10.1109/ICASSP.2003.1200069 ISSN: 1520-6149.

[80] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 5. V–708. https://doi.org/10.1109/ICASSP.2003.1200069 ISSN: 1520-6149.

[81] NASA. 1986. *NASA Task Load Index (TLX) V 1.0: Paper and Pencil Package.* Technical Report. Human Performance Research Group, NASA Ames Research Center, Moffett Field, CA, USA.

[82] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, and Herve Glotin. 2000. Audio-Visual Speech Recognition. (2000), 86.

[83] L.C. Ng, G.C. Burnett, J.F. Holzrichter, and T.J. Gable. 2000. Denoising of Human Speech Using Combined Acoustic and Em Sensor Signal Processing. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, Vol. 1. 229–232 vol.1. https://doi.org/10.1109/ICASSP.2000.861925 ISSN: 1520-6149.

[84] Sanjay A. Patil and John H. L. Hansen. 2010. The Physiological Microphone (pmic): A Competitive Alternative for Speaker Assessment in Stress Detection and Speaker Verification. *Speech Communication* 52, 4 (April 2010), 327–340. https://doi.org/10.1016/j.specom.2009.11.006

[85] Dr Marta Perez Garcia, Sarita Saffon Lopez, and Hector Donis. 2018. Everybody is Talking About Virtual Assistants, But How are People Really Using Them?. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*. 1–5.

[86] Stavros Petridis and Maja Pantic. 2016. Deep Complementary Bottleneck Features for Visual Speech Recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2304–2308. https://doi.org/10.1109/ICASSP.2016.7472088 ISSN: 2379-190X.

[87] J.W. Picone. 1993. Signal Modeling Techniques in Speech Recognition. *Proc. IEEE* 81, 9 (Sept. 1993), 1215–1247. https://doi.org/10.1109/5.237532 Conference Name: Proceedings of the IEEE.

[88] Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz. 2009. EEG-based Speech Recognition - Impact of Temporal Effects. In *BIOSIGNALS*. https://doi.org/10.5220/0001554303760381

[89] Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz. 2009. EEG-Based Speech Recognition - Impact of Temporal Effects. In *BIOSIGNALS*. https://doi.org/10.5220/0001554303760381

[90] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. 2003. Recent Advances in the Automatic Recognition of Audiovisual Speech. *Proc. IEEE* 91, 9 (Sept. 2003), 1306–1326. https://doi.org/10.1109/JPROC.2003.817150 Conference Name: Proceedings of the IEEE.

[91] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. https://infoscience.epfl.ch/record/192584 Conference Name: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding Number: CONF Publisher: IEEE Signal Processing Society.

[92] S. Prabhakar, S. Pankanti, and A.K. Jain. 2003. Biometric Recognition: Security and Privacy Concerns. *IEEE Security Privacy* 1, 2 (March 2003), 33–42. https://doi.org/10.1109/MSECP.2003.1193209 Conference Name: IEEE Security Privacy.

[93] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4884–4895. https://doi.org/10.1145/2858036.2858130

[94] Halley P. Profita. [n.d.]. Designing wearable computing technology for acceptability and accessibility. 114 ([n. d.]), 44–48. https://doi.org/10.1145/2904092.2904101

[95] T.F. Quatieri, K. Brady, D. Messing, J.P. Campbell, W.M. Campbell, M.S. Brandstein, C.J. Weinstein, J.D. Tardelli, and P.D. Gatewood. 2006. Exploiting Nonacoustic Sensors for Speech Encoding. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 2 (March 2006), 533–544. https://doi.org/10.1109/TSA.2005.855838 Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[96] Stuart Reeves, Steve Benford, Claire O'Malley, and Mike Fraser. 2005. Designing the Spectator Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 741–750. https://doi.org/10.1145/1054972.1055074

[97] Julie Rico and Stephen Brewster. [n.d.]. Gestures all around us: user differences in social acceptability perceptions of gesture based interfaces. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (New York, NY, USA, 2009-09-15) *(MobileHCI '09)*. Association for Computing Machinery, 1–2. https://doi.org/10.1145/1613858.1613936

[98] Julie Rico and Stephen Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 887–896. https://doi.org/10.1145/1753326.1753458

[99] Julie Rico and Stephen Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 887–896. https://doi.org/10.1145/1753326.1753458

[100] Judy Robertson and Maurits Kaptein. [n.d.]. *Modern Statistical Methods for HCI*. Springer.

[101] Sami Ronkainen, Jonna Häkkilä, Saana Kaleva, Ashley Colley, and Jukka Linjama. 2007. Tap Input as an Embedded Interaction Method for Mobile Devices. In *Proceedings of the 1st International Conference on Tangible and Embedded Interaction* (Baton Rouge, Louisiana) *(TEI '07)*. Association for Computing Machinery, New York, NY, USA, 263–270. https://doi.org/10.1145/1226969.1227023

[102] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 159:1–159:23. https://doi.org/10.1145/3161187

[103] A. Rubin, V. Praneetvatakul, Shirley Gherson, C. Moyer, and R. Sataloff. 2006. Laryngeal Hyperfunction During Whispering: Reality or Myth? *Journal of voice : official journal of the Voice Foundation* (2006). https://doi.org/10.1016/J.JVOICE.2004.10.007

[104] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The Tongue and Ear Interface: A Wearable System for Silent Speech Recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC '14)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/2634317.2634322

[105] Tanja Schultz and Michael Wand. 2010. Modeling Coarticulation in Emg-Based Continuous Speech Recognition. *Speech Communication* 52, 4 (April 2010), 341–353. https://doi.org/10.1016/j.specom.2009.12.002

[106] Mike Schuster. 2010. Speech Recognition for Mobile Devices at Google. In *PRICAI 2010: Trends in Artificial Intelligence (Lecture Notes in Computer Science)*, Byoung-Tak Zhang and Mehmet A. Orgun (Eds.). Springer, Berlin, Heidelberg, 8–10. https://doi.org/10.1007/978-3-642-15246-7_3

[107] Marcos Serrano, Barrett M. Ens, and Pourang P. Irani. 2014. Exploring the Use of Hand-to-Face Input for Interacting with Head-Worn Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3181–3190. https://doi.org/10.1145/2556288.2556984

[108] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. 1995. Speech Recognition with Primarily Temporal Cues. *Science* (1995). https://doi.org/10.1126/science.270.5234.303

[109] Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining Residual Networks with Lstms for Lipreading. *INTERSPEECH* (2017). https://doi.org/10.21437/INTERSPEECH.2017-85

[110] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[111] P. Suppes, B. Han, and Z. Lu. 1998. Brain-Wave Recognition of Sentences. *Proceedings of the National Academy of Sciences of the United States of America* (1998). https://doi.org/10.1073/pnas.95.26.15861

[112] P. Suppes, Z. Lu, and B. Han. 1997. Brain Wave Recognition of Words. *Proceedings of the National Academy of Sciences of the United States of America* (1997). https://doi.org/10.1073/pnas.94.26.14965

[113] Ingo R. Titze, Brad H. Story, Gregory C. Burnett, John F. Holzrichter, Lawrence C. Ng, and Wayne A. Lea. 1999. Comparison Between Electroglottography and Electromagnetic Glottography. *The Journal of the Acoustical Society of America* 107, 1 (Dec. 1999), 581–588. https://doi.org/10.1121/1.428324 Publisher: Acoustical Society of America.

[114] Ying-Chao Tung, Chun-Yen Hsu, Han-Yu Wang, Silvia Chyou, Jhe-Wei Lin, Pei-Jung Wu, Andries Valstar, and Mike Y. Chen. 2015. User-Defined Game Input for Smart Glasses in Public Space. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3327–3336. https://doi.org/10.1145/2702123.2702214

[115] Michael Wand and Tanja Schultz. 2011. Session-Independent Emg-Based Speech Recognition. In *BIOSIGNALS*. https://doi.org/10.5220/0003169702950300
[116] Dean Weber. 2001. Interactive User Interface Using Speech Recognition and Natural Language Processing. https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2001026093
[117] Dean Weber. 2002. Object Interactive User Interface Using Speech Recognition and Natural Language Processing. https://patents.google.com/patent/US6434524B1/en
[118] Dean Weber. 2002. Object Interactive User Interface Using Speech Recognition and Natural Language Processing. https://patents.google.com/patent/US6434524B1/en
[119] D. Zaykovskiy. 2006. Survey of the Speech Recognition Techniques for Mobile Devices.
[120] You Zhang, Jeffery J. Faneuff, William Hidden, James T. Hotary, Steven C. Lee, and Vasu Iyengar. 2010. Automobile Speech-Recognition Interface. https://patents.google.com/patent/US7826945B2/en
[121] Yu Zhong, T. V. Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P. Bigham. 2014. Justspeak: Enabling Universal Voice Control on Android. In *W4A 2014*. http://dl.acm.org/citation.cfm?id=2596720