Effects of Speaking Rate on Speech and Silent Speech Recognition

Laxmi Pandey
Inclusive Interaction Lab
University of California, Merced
Merced, California, United States
lpandey@ucmerced.edu

Ahmed Sabbir Arif
Inclusive Interaction Lab
University of California, Merced
Merced, California, United States
asarif@ucmerced.edu

ABSTRACT

Speaking rate or the speed at which a person speaks is a fundamental user characteristic. This work investigates the rate in which users speak when interacting with speech and silent speech-based methods. Results revealed that native users speak about 8% faster than non-native users, but both groups slow down at comparable rates (34–40%) when interacting with these methods, mostly to increase their accuracy rates. A follow-up experiment confirms that slowing down does improve the accuracy of these methods. Both methods yield the best accuracy rates when speaking at 0.75x of the actual speaking rate. A post-hoc error analysis revealed that speech and silent speech methods and native and non-native speakers are susceptible to different types of errors.

CCS CONCEPTS

• **Human-centered computing** → **Text input**; **Empirical studies in interaction design**; **Natural language interfaces**; • **Computing methodologies** → *Speech recognition.*

KEYWORDS

Speech, silent speech, speaking rate, recognition

ACM Reference Format:
Laxmi Pandey and Ahmed Sabbir Arif. 2022. Effects of Speaking Rate on Speech and Silent Speech Recognition. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3491101.3519611

1 INTRODUCTION

Speech and silent speech-based methods have the potential of becoming dominant input modalities, especially when the user's hands are busy performing other tasks, when in public or noisy places, and when touching public devices is to be avoided in times like the current COVID-19 situation. Speech input is a spoken form of communication that enables users to communicate with a computer system using speech commands, whereas silent speech input is an unspoken form of communication that enables communication by visually interpreting the movements of the speaker's lips. Although many identified these as inherently usable and more natural modes of interaction [55, 61], they are not yet reliable [37]. One major issue that adversely affects these methods' performance is inaccurate recognition. To avoid potential speech misrecognition, users often monitor their behaviors to adjust and optimize future task performance according to experienced errors or conflicts [6, 62, 69]. They engage themselves in processes of repairing the errors by either reformulation, simplification, or hyperenunciation [32, 38, 40, 42, 50, 57]. However, peoples' approach to silent speech input to avoid potential misrecognition is unknown.

Speaking rate is a fundamental user characteristics that can influence speech recognition performance due to the variation in acoustic properties of human speech production, such as vowel and consonant duration, the transition between phoneme and stops, and distortions in the temporal and spectral domains [23, 24, 72]. Some studies report that faster speaking rates result in higher error rates [23, 48, 64, 65], whereas some identified slower speaking rates to be more error prone [24, 66]. This disagreement encourages reinvestigation of the effects of speaking rates on speech recognition performance. Besides, no such investigations have been conducted for silent speech recognition. This work explores whether native and non-native speakers interact differently with speech and silent speech-based methods, whether speaking rate affects recognition rates of these methods, the optimal speaking rates for increased accuracy, and whether the effects of speaking rate are different for native and non-native speakers.

2 RELATED WORK

2.1 Speech Interaction

Speech input enables interaction with computer systems via speech recognition. It converts spoken language to text using acoustic and language models. Nowadays, speech interfaces can be found in automobiles [14, 29, 39], smartphones [36], and home assistant devices [47, 58, 63]. These interfaces use automatic speech recognition (ASR) to detect and comply with spoken commands. ASR can potentially improve productivity and user comfort when traditional input methods, like touch and keyboards, are inefficient, difficult, or inconvenient to use [26, 61]. Yet, users of speech input are usually unsatisfied with the quality of interaction due to low recognition accuracy [20]. To avoid potential errors, users tend to modify their speaking styles and patterns [32, 38, 40, 42, 50, 57] by shortening their sentences [34, 57], performing repetition [11, 16], increasing the volume [13, 22], and hyper-articulating [53]. However, studies showed that ASR can fail even when these strategies are applied due to high levels of disfluency, non-canonical pronunciation, accent, speaking rate, and acoustic and prosodic variability [24]. Luce

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3519611

and Pisoni [41] reported that recognition is worse for words that are phonetically similar to other words than for highly distinctive words. Shinozaki and Furui [64] found out that longer words have slightly lower error rates than shorter words. Howes [28] showed that infrequent words are more likely to be misrecognized. A different research found a correlation between large fluctuations in the short-term speaking rate and high recognition errors [2]. Another work reported that male speakers have significantly higher recognition error rates than female speakers due to higher rates of disfluency [1]. Relevantly, misrecognized words were found to have higher pitch and energy than correctly recognized words [27]. Another study revealed that words with more possible pronunciations have higher error rates and longer words have slightly lower error rates [24].

2.2 Effects of Speaking Rate

Different speaking rates can significantly affect speech recognition performance due to a distorted spectrum caused by variations in speaking rate [23, 24, 72]. Natural speaking rate depends on user characteristics like gender, age, accents, and psychological state. Yuan et al. [71] showed that older people speak slowly compared to young adults, and women talk slower than men. Rao and Koolagudi [60] reported that people usually speak fast when in a hurry or angry, and slow when they are tired, sad, or sick. Studies also showed that non-native speakers talk much slower [25] and exhibit more variation in speaking rate than native speakers [8]. However, suprasegmental characteristics between native and non-native speakers in spontaneous speech suggest that non-native speakers are less variable than native speakers [49], which can affect recognition rate [56], particularly for non-native speakers [7, 19, 70]. However, the research community is divided on how speaking rate affects recognition accuracy. Some associated faster speech with higher error rates [23, 48, 64, 65], while others found slow speech to be more error-prone [24, 66].

2.3 Silent Speech Interaction

Video-based silent speech input enables interaction by analyzing lip movements. It captures lip movements with a camera, then recognizes the silently spoken words using image processing and language models. Silent speech input can be effective when speaking audibly could disturb others or disclose confidential information, to understand elderly and children speech, and to provide people with speech and motor impairments access to computer systems [55]. Research found silent speech error-prone due to its dependence on extraneous factors like lighting, skin complexion, posture, head rotation, and facial expression [12, 33, 51, 54, 67]. In recent investigations, users reported a higher level of satisfaction using this method than speech input in some scenarios [55, 68]. However, very little is known about how speaking rate affects silent speech recognition and whether these effects are different than those of speech recognition.

3 EXPERIMENT 1: SPEAKING RATE

This experiment investigates whether native and non-native speakers speak at different rates when interacting with speech and silent speech-based methods.

3.1 Apparatus

We developed a custom app with Android Studio 3.1.4 (Fig 1). Participants used it on their own Android smartphones. Its *landing page* included a drop-down menu to select a recording condition (speech, silent speech) and a Start button to start data collection. The *data collection page* displayed the front camera in real-time, random phrases from a set [44] for participants to speak or silently speak, and a Record and Stop toggle button to start and stop recording, respectively. The app stored all videos locally and automatically logged the duration of each spoken phrase.

3.2 Participants

Twelve volunteers took part in the experiment (Fig. 1). Table 1 presents the demographics of the participants divided into native and non-native groups. Originally, we wanted to recruit equal number of native and non-native speakers, but were unable to do so due to the spread of COVID-19.

3.3 Design and Metrics

The experiment had one within-subjects independent variable: *medium*, with three levels: *baseline*, *speech*, and *silent speech*; and one between-subjects independent variable: *speaker*, with two levels: *native* and *non-native*. The baseline condition recorded participants' speaking rates in human-human communication, while the speech and silent speech conditions recorded their speaking rates with a speech and silent speech recognizer, respectively, through the mobile app. We used a Wizard-of-Oz setup, that is, the app did not include actual recognizers but pretended to accurately recognize all spoken and silently spoken phrases as long as the participant's face was visible to the app. For the baseline condition, we extracted one minute of speech from the conversations we had with the participants during the app installation and demonstration process. In the speech and silent speech conditions, participants spoke and silently spoke 30 phrases from a set [44], respectively (720 phrases, in total). The dependent variables were:

- **Time per phoneme** (TPP) is the average time participants took to utter a phoneme (in milliseconds), calculated using the following equation: $TPP = \frac{\text{time per phrase}}{\text{total phoneme in phrase}}$. Total phoneme in a recognized phrase was counted with the Pronouncing API[1] that uses the Carnegie Mellon University (CMU) Pronouncing Dictionary[2] to identify phonemes.
- **Actual words per minute** (A-WPM) is the most commonly used metrics for calculating speaking rate [3, 15]. It measures the average number of actual words spoken in a minute. This metric is different from the traditional WPM metric that considers five characters as one word regardless of the actual number of words in a phrase [5]. A-WPM is calculated using the following equation: $WPM = \frac{\text{total words}}{\text{number of minutes}}$.

3.4 Procedure

The experiment was conducted remotely via Zoom due to COVID-19. We scheduled individual video calls with each participant. They

[1] Pronouncing API: https://pronouncing.readthedocs.io/en/latest/pronouncing.html#pronouncing.phones_for_word
[2] CMU Pronouncing Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

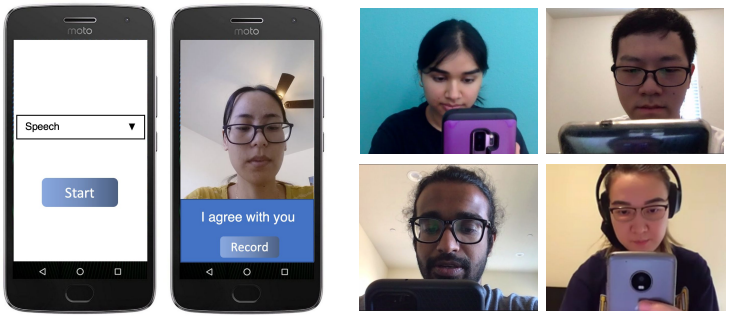

Figure 1: (a) Screenshots of the custom app used in the experiment: the landing page (left) and the data collection page (right). (b) Four volunteers participating in the first experiment through a teleconferencing system.

Table 1: Demographics of the participants.

	Native (*N* = 4)	**Non-native (*N* = 8)**
Age	22–54 years (M = 32.2, SD = 14.7)	19–33 years (M = 25.8, SD = 4.7)
Gender	1 female, 3 male	4 female, 4 male
Experience with speech	1–8 years (M = 3.5, SD = 3.3)	1–4 years (M = 1.2, SD = 0.8)
Experience with silent speech	None	None

were instructed to join the call from a quiet room to avoid any interruptions during the experiment. In the call, we first explained how speech and video-based silent speech recognition systems work, then demonstrated the custom app and collected their informed consents and demographics using electronic forms. We then shared the app installation file (APK) with them and guided them through the installation process on their smartphones. The data collection session started after that, where the app displayed one phrase at a time. Participants were instructed to press the Record button, speak or silently speak the presented phrase, then pressed the Stop button. They were told that the system will process the spoken or silently spoken phrase when they press the Stop button. If the phrase is correctly recognized, it will display the next phrase, otherwise will ask them to re-speak the same phrase. However, in reality, the app did not include a recognizer, instead pretended to correctly recognize all spoken and silently spoken phrases. The Zoom sessions were recorded to extract one minute of speech for the baseline condition (Section 3.3). Participants were not informed of this during the experiment to avoid a potential Hawthorne effect [43]. Upon completion, participants shared all locally stored video clips and log files with us by uploading those to a cloud storage. They then took part in an interview about their experience with the app. Finally, we debriefed them about the Wizard-of-Oz setup and informed them that clips from the demo and installation Zoom session will be used to measure their natural speaking rates.

4 RESULTS

A complete experiment took 45–60 minutes. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a one-way repeated-measures ANOVA to study the effects of *medium*, a one-way between-subjects ANOVA for the effects of *speaker*, and a mixed-design ANOVA for the *medium* × *speaker* interaction effects [4].

4.1 Time per Phoneme (TPP)

An ANOVA identified a significant effect of medium ($F_{2,11} = 697.59$, $p < .0001$) on TPP. On average, participants took 107.5 ms (SD = 4.6), 161.6 ms. (SD = 5.5), and 178.7 ms (SD = 8.8) to utter a phoneme in the baseline, speech, and silent speech conditions, respectively. An ANOVA also identified a significant effect of speaker ($F_{1,10} = 1212.35, p < .0001$) on TPP. On average, native participants took 161.26 ms (SD = 10.78) to utter a phoneme, while non-native participants took 173.14 ms (SD = 13.36). There was also a medium × speaker interaction effect ($F_{1,20} = 66.02, p < .0001$). Fig. 2 (a) presents average TPP for native and non-native speakers with the three mediums.

4.2 Actual Words per Minute (A-WPM)

An ANOVA identified a significant effect of medium ($F_{2,11} = 1783.18$, $p < .0001$) on A-WPM. On average, participants yielded 109.7 (SD = 4.6), 89.5 (SD = 5.2), and 74.8 (SD = 3.6) A-WPM in the baseline, speech, and silent speech conditions, respectively. An ANOVA also identified a significant effect of speaker ($F_{1,10} = 1467.96, p = .0001$) on A-WPM. On average, native participants yielded 87.49 A-WPM (SD = 9.04), while non-native participants yielded 80.26 A-WPM (SD = 8.42). There was also a medium × speaker interaction effect ($F_{1,20} = 17.18, p < .0001$). Fig. 2 (b) presents average A-WPM for native and non-native speakers with the three mediums.

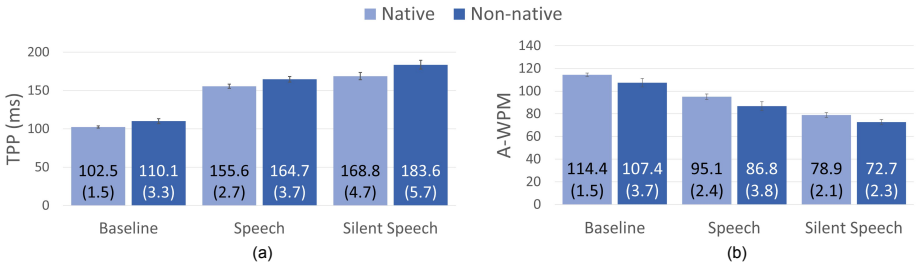

Figure 2: (a) Average time per phoneme (TPP) and (b) average actual words per minute (A-WPM) for native and non-native speakers with the three investigated mediums. The values inside the brackets are standard deviations (SD). The error bars represent ±1 SD.

5 DISCUSSION

Both native and non-native speakers spoke at much slower rates compared to their usual speaking rates while using the speech and the silent speech recognizers. On average, participants took 33.4% and 39.8% extra time to utter a phoneme with speech and silent speech, respectively. Likewise, A-WPM dropped by 22.5% and 46.6%, respectively. Consequently, a post-hoc Tukey-Kramer multiple-comparison test identified two distinct groups: {baseline} and {speech, silent speech}. The post-experiment interview revealed that participants spoke slowly while using these methods thinking that it would increase their recognition rates. However, there was no actual effects on phrase recognition as the Wizard-of-Oz approach pretended to correctly recognize all spoken or silently spoken phrases. Since all participants were experienced users of various voice assistant systems, it is likely that the unreliability of these systems encouraged them to reduce the rate of their speech. Relevantly, a participant (female, 27 years, non-native) said, *"It [speaking slowly] is mostly due to lack of proficiency and different accent. I always try to speak slowly and try to match accent to make the speech assistant understand me which is sometimes awkward and irritating"*. Surprisingly, they spoke at a much slower rate when using a silent speech recognizer compared to when using a speech recognizer. This could be either because participants never used a silent speech-based method before or the fact that video-based silent speech recognizers detect speech based on lip movements rather than the sound produced by the speakers (Section 3.4), giving them the impression that the method requires extra finesse for an acceptable accuracy rate. Post-experiment interview revealed that participants overemphasized their lip movements during silent speech to "aid" the recognition process.

Results revealed that non-native speaker spoke at a slower rate than native speakers (about 7% slower TPP). This is not surprising since many studies found out that average speaking rate for non-native speakers is slower than for native speakers as *"a general lack of proficiency and experience can result in slower speaking rates"* [9, 17, 18, 21, 25]. However, both native and non-native speakers slowed down at comparable rates when interacting with speech (~34% slower TPP) and silent speech (~40% slower TPP) recognizers. This finding is interesting as it suggests that these slower speaking rates were not caused by the lack of proficiency or experience but due to the speakers' skepticism about the reliability of the state-of-the-art speech and silent speech recognizers. Based on these findings, we recommend evaluating new speech and silent speech recognizers with both native and non-native speakers of the target language, and report the results of the two groups separately due to their significantly different speaking rates. The fact that users slow down when interacting with speech and silent speech recognizers can also be exploited for improved performance.

We were unable to study any potential effects of recognition error on speaking rate since the Wizard-of-Oz setup collected data without any errors. However, users are likely to adjust their interaction behavior when interacting with an error-prone system, like observed in other recognition systems [6]. Another limitation of the study is using different scenarios in the baseline and the speech conditions. Speaking rate for the baseline was calculated in continuous computer-mediated communication, while the same for the speech and the silent speech were calculated from manually segmented phrases. It is unknown whether the additional latency introduced by the manual segmentation affected the speaking rate in any way. It is also unclear if the speaking rates are different for computer-mediated and face-to-face communications, although prior works reported other behavioral changes [35].

6 EXPERIMENT 2: EFFECTS OF SPEAKING RATE

This experiment studied whether speaking rate affects recognition rates of state-of-the-art speech and silent speech recognizers.

6.1 Participants and Design

We invited the participants of the previous experiment (Section 3.2) to take part in this experiment. The experiment had two within-subjects independent variable: *medium* and *speaking rate*. The former had two levels: *speech* and *silent speech*, and the latter had seven levels: 0.25x, 0.5x, 0.75x, 1x, 1.25x, 1.5x, and 1.75x of the actual speaking rates of the participants. These rates were selected based on YouTube's playback speed scale, ranging from quarter speed (0.25x) to double speed (2x). Among these, we selected the actual rate (1x), the top three slower rates (0.25x, 0.5x, 0.75x) and the top three faster rates (1.25x, 1.5x, 1.75x), resulting in seven rates in total. The experiment had one between-subjects independent variable: *speaker*, with two levels: *native* and *non-native*. Participants spoke 30 phrases from the respective model's training dataset [54, see A.1 & A.5], which were post-processed to achieve the seven

speaking rates, resulting in 30 phrases × 2 mediums × 7 speaking rates = 420 phrases per participant. The dependent variable was the following performance metric:

- **Word accuracy** (WA) measures the total number of words accurately recognized from the total number of spoken words. It is calculated using the following equation, where S is the number of substitutions, D is the number of deletions, I is the number of insertions, N is the number of words in the ground truth: $WA = 1 - \frac{(S+D+I)}{N}$.

6.2 Apparatus and Procedure

We modified the custom app used the previous experiment to replace the phrases [44] with phrases from the examined speech and silent speech recognition models' training datasets. We also included a new condition in the app, where participants are instructed to read the presented phrases. Recorded video clips were time-expanded for the slower rates and time-compressed for the faster rates using the FFmpeg [3] platform. All clips were then processed using two state-of-the-art pre-trained recognition models for speech and silent speech: Kaldi (Api.ai) [59] and LipType [54], respectively.

The experiment used the same procedure as the first experiment except for the demonstration and the post-experiment debrief and interview. The custom app displayed one phrase at a time, and participants were instructed to read it at a rate in which they would usually speak with another person. Note that, despite the different speaking rates, each participant spoke exactly the same number of words in each condition.

7 RESULTS

A complete experiment took about 30 minutes. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a two-way repeated-measures ANOVA to study the effects of *medium* and *speaking rate*, a one-way between-subjects ANOVA to to study the effects of *speaker*, and a mixed-design ANOVA to study the interaction effects [4].

7.1 Word Accuracy (WA)

An ANOVA identified a significant effect of medium ($F_{1,11} = 64769.13, p < .0001$) and speaking rate ($F_{6,66} = 697.21, p < .0001$) on WA. The medium × speaking rate interaction effect was also statistically significant ($F_{6,66} = 33009.02, p < .0001$). Fig. 3 illustrates the average WA of the speech and the silent speech recognition methods with the seven speaking rates. An ANOVA also identified a significant effect of speaker ($F_{1,10} = 805.74, p < .0001$). The speaker × medium ($F_{1,10} = 64.54, p < .0001$) and the speaker × speaking rate × medium ($F_{6,60} = 543.30, p < .0001$) interaction effects were also statistically significant. Fig. 4 illustrates the average WA of the speech and the silent speech recognition methods for native and non-native speakers with the seven examined speaking rates.

[3] A Complete, Cross-Platform Solution to Record, Convert and Stream Audio and Video: https://www.ffmpeg.org

Table 2: Distribution of insertion, deletion, and substitution errors in the phrases recognized by the speech and the silent speech recognizers.

	Speech			Silent Speech		
	All	Native	Non-Native	All	Native	Non-Native
Insertion	38%	12%	29%	2%	4%	22%
Deletion	21%	41%	37%	27%	30%	21%
Substitution	41%	47%	34%	71%	66%	57%

7.2 Error Analysis

We conducted a post-hoc analysis of the recognized phrases at the usual speaking rate (1x) to find out the distribution of insertion errors (extra words are incorrectly inserted), deletion errors (correct words are incorrectly omitted), and substitution errors (words are substituted with incorrect words) [10]. Table 2 presents the results.

8 DISCUSSION

Both speech and silent speech methods performed well with regular (1x) speaking rate. On average, speech and silent speech methods yielded 82% (SD = 4.6) and 80% (SD = 3.5) WA, respectively, with regular speaking rate. The effects of speaking rate was different for native and non-native speakers. At regular rate, speech and silent speech methods were 9.9% and 7.5% more accurate, respectively, for native speakers than non-native speakers. However, for both native and non-native speakers, the performance of the speech recognition method dropped substantially with speaking rates lower than 0.5x and higher than 1.25x (Fig. 4). A post-hoc Tukey-Kramer multiple-comparison test revealed that 0.75x speaking rate was significantly more accurate than the other speaking rates. Likewise, the performance of the silent speech recognition method dropped substantially with rates lower than 0.75x and higher than 1.25x for both native and non-native speakers (Fig. 4). Like speech, a post-hoc Tukey-Kramer multiple-comparison test identified 0.75x as significantly more accurate than the other examined speaking rates. These findings suggest that speaking slightly slower than usual can indeed increase the reliability of speech and silent speech recognizers, regardless of the speaker's proficiency and experience in English. Results also suggest that 0.5–1.25x is the optimal range for speech and 0.75–1.25x is the optimal range for silent speech for higher accuracy rates. We speculate, this is due to the fact that much faster speaking rates can cause frequent and stronger pronunciation changes while much slower speaking rates tend to add unnecessary pauses between phonemes [45]. The average natural speaking rate was slower in this experiment than the first experiment since, here, participants read the phrases, which is slower than speaking [30, 31, 46, 52].

Error analysis revealed that silent speech had 94.7% lower insertion errors than speech. We speculate, this is because ambient noise in the audio affected the recognition of the speech method. Silent speech, in contrast, uses visual information for recognition, thus was not affected by background noise. Interestingly, speech committed 11% higher deletion errors and 38% higher substitution errors for native speakers than non-native speakers. This could be because faster rates resulted in overlaps between the words, making it difficult to segment them. Silent speech also resulted in 81%

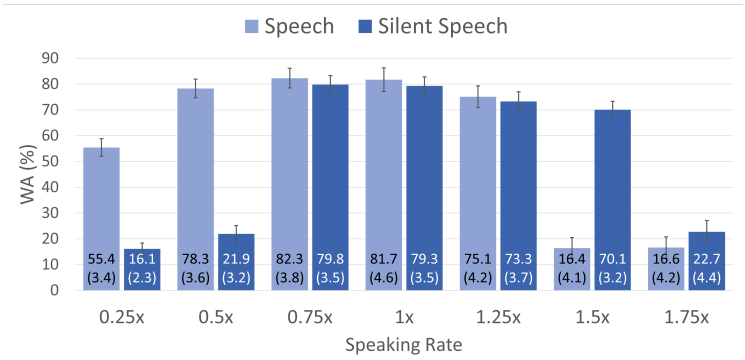

Figure 3: Average word accuracy rates (%) of the speech and the silent speech recognition methods with the seven examined speaking rates. The values inside the brackets are standard deviations (SD). The error bars represent ±1 SD.

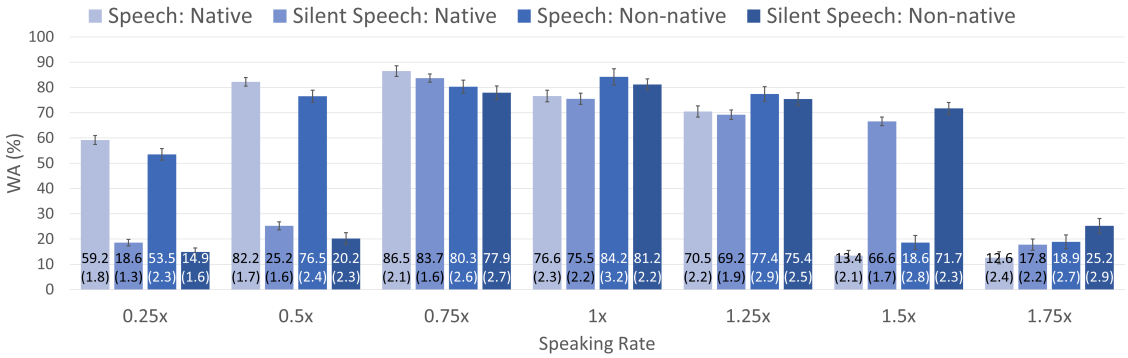

Figure 4: Average word accuracy rates (%) of the speech and silent speech recognition methods for native and non-native speakers with the seven examined speaking rates. The values inside the brackets are standard deviations (SD). The error bars represent ±1 SD.

lower insertion errors, 42% higher deletion errors, and 16% higher substitution errors than non-native speakers, presumably for the same reasons. Silent speech had 73.1% higher substitution errors than speech, which could be due to the difficulty in distinguishing between different homophones with visual information as multiple characters can produce the same lip movement sequence, such as for the letters 'p' and 'b'.

9 CONCLUSION

The findings of this work highlight the importance of considering speaking rate in speech and silent speech-based interfaces. While designing interfaces for these methods, the recognition algorithms must be optimized for varying speaking rates and the characteristics of native and non-native speakers. Error analysis presented in this work could be used to identify areas that require extra effort to increase the respective method's accuracy rates. The findings could also provide guidance to users on improving speech and silent speech input performance.

REFERENCES

[1] Martine Adda-Decker and Lori Lamel. 2005. Do Speech Recognizers Prefer Female Speakers?. In *Ninth European Conference on Speech Communication and Technology*.

[2] Anastasios Anastasakos, Richard Schwartz, and Han Shu. 1995. Duration modeling in large vocabulary speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, 628–631.

[3] Aries Arditi and Jianna Cho. 2005. Serifs and Font Legibility. *Vision research* 45, 23 (2005), 2926–2933.

[4] Ahmed Sabbir Arif. 2021. Statistical Grounding. In *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice* (1 ed.). Association for Computing Machinery, New York, NY, USA, 59–99. https://doi.org/10.1145/3447404.3447410

[5] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2009. Analysis of Text Entry Performance Metrics. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*. 100–105. https://doi.org/10.1109/TIC-STH.2009.5444533

[6] 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 event-place: Montreal, Quebec, Canada.

[7] Tim Ashwell and Jesse R Elam. 2017. How Accurately Can the Google Web Speech API Recognize and Transcribe Japanese L2 English Learners' Oral Production?. *Jalt Call Journal* 13, 1 (2017), 59–76.

[8] Melissa M Baese-Berk and Tuuli H Morrill. 2015. Speaking Rate Consistency in Native and Non-native Speakers of English. *The Journal of the Acoustical Society of America* 138, 3 (2015), EL223–EL228.

[9] Melissa M. Baese-Berk and Tuuli H. Morrill. 2015. Speaking Rate Consistency in Native and Non-Native Speakers of English. *The Journal of the Acoustical Society of America* 138, 3 (Sept. 2015), EL223–EL228. https://doi.org/10.1121/1.4929622 Publisher: Acoustical Society of America.

[10] L. Bahl and F. Jelinek. 1975. Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition. *IEEE Transactions on Information Theory* 21, 4 (July 1975), 404–411. https://doi.org/10.1109/TIT.1975.1055419 Conference Name: IEEE Transactions on Information Theory.

[11] Linda Bell and Joakim Gustafson. 1999. Interaction with an Animated Agent in a Spoken Dialogue System. In *Sixth European Conference on Speech Communication and Technology*.

[12] Andre-Pierre Benguerel and Margaret Kathleen Pichora-Fuller. 1982. Coarticulation Effects in Lipreading. *Journal of Speech, Language, and Hearing Research* 25, 4 (1982), 600–607.

[13] Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic Alignment Between People and Computers. *Journal of pragmatics* 42, 9 (2010), 2355–2368.

[14] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300270

[15] Ronald P Carver. 1976. Word Length, Prose Difficulty, and Reading rate. *Journal of Reading Behavior* 8, 2 (1976), 193–203.

[16] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 337–348. https://doi.org/10.1145/3202185.3202749

[17] Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative Assessment of Second Language Learners' Fluency by Means of Automatic Speech Recognition Technology. *The Journal of the Acoustical Society of America* 107, 2 (2000), 989–999.

[18] Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative Assessment of Second Language Learners' Fluency: Comparisons Between Read and Spontaneous Speech. *the Journal of the Acoustical Society of America* 111, 6 (2002), 2862–2873.

[19] Ronald Cumbal, Birger Moell, José Lopes, and Olov Engwall. 2021. "You don't understand me!": Comparing ASR Results for L1 and L2 Speakers of Swedish. In *Proc. Interspeech 2021*. 4463–4467. https://doi.org/10.21437/Interspeech.2021-2140

[20] 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

[21] T Derwing and MJ Munro. 2001. What Speaking Rates Do Non-Native Listeners Prefer? *Applied Linguistics* 22, 3 (Sept. 2001), 324–337. https://doi.org/10.1093/applin/22.3.324

[22] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is It OK If I Eat You?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[23] Eric Fosler-Lussier and Nelson Morgan. 1999. Effects of Speaking Rate and Word Frequency on Pronunciations in Convertional Speech. *Speech Commun.* 29, 2 (Nov. 1999), 137–158.

[24] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which Words are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase Speech Recognition Error Rates. *Speech Communication* 52, 3 (2010), 181–200.

[25] Susan G. Guion, James E. Flege, Serena H. Liu, and Grace H. Yeni-Komshian. 2000. Age of Learning Effects on the Duration of Sentences Produced in a Second Language. *Applied Psycholinguistics* 21, 2 (June 2000), 205–228. https://doi.org/10.1017/S0142716400002034 Publisher: Cambridge University Press.

[26] Alexander G. Hauptmann and Alexander I. Rudnicky. 1990. A Comparison of Speech and Typed Input. In *Proceedings of the Workshop on Speech and Natural Language* (Hidden Valley, Pennsylvania) *(HLT '90)*. Association for Computational Linguistics, USA, 219–224. https://doi.org/10.3115/116580.116652

[27] Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and Other Cues to Speech Recognition Failures. *Speech Communication* 43, 1-2 (2004), 155–175.

[28] Davis Howes. 1954. On the Interpretation of Word Frequency as a Variable Affecting Speed of Recognition. *Journal of Experimental Psychology* 48, 2 (1954), 106.

[29] Zhang Hua and Wei Lieh Ng. 2010. Speech Recognition Interface Design for In-Vehicle System. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Pittsburgh, Pennsylvania) *(AutomotiveUI '10)*. Association for Computing Machinery, New York, NY, USA, 29–33. https://doi.org/10.1145/1969773.1969780

[30] E. Jacewicz, R. Fox, and L. Wei. 2010. Between-speaker and within-speaker Variation in Speech Tempo of American English. *The Journal of the Acoustical Society of America* 128 2 (2010), 839–50.

[31] Ewa Jacewicz, Robert A Fox, Caitlin O'Neill, and Joseph Salmons. 2009. Articulation Rate Across Dialect, Age, and Gender. *Language variation and change* 21, 2 (2009), 233.

[32] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/2484028.2484092

[33] Timothy R Jordan and Sharon M Thomas. 2011. When Half a Face is as Good as a Whole: Effects of Simple Substantial Occlusion on Visual and Audiovisual Speech Perception. *Attention, Perception, & Psychophysics* 73, 7 (2011), 2270.

[34] Alan Kennedy, Alan Wilkes, Leona Elder, and Wayne S Murray. 1988. Dialogue with Machines. *Cognition* 30, 1 (1988), 37–72.

[35] Sara Kiesler, Jane Siegel, and Timothy W. McGuire. 1984. Social Psychological Aspects of Computer-Mediated Communication. *American Psychologist* 39, 10 (1984), 1123–1134. https://doi.org/10.1037/0003-066X.39.10.1123 Place: US Publisher: American Psychological Association.

[36] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA) *(CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/2854946.2854961

[37] Heidi Horstmann Koester. [n.d.]. *Abandonment of Speech Recognition by New Users*. https://www.resna.org/sites/default/files/legacy/conference/proceedings/2003/Papers/ComputerAccess/Koester_CA_Abandonment.htm

[38] Dounia Lahoual and Myriam Frejus. 2019. When Users Assist the Voice Assistants: From Supervision to Failure Resolution. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299053

[39] Victor Ei-Wen Lo and Paul A Green. 2013. Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and the Related Literature. *International Journal of Vehicular Technology* 2013 (2013).

[40] M. Lohse, K. J. Rohlfing, B. Wrede, and G. Sagerer. 2008. "Try Something Else!" — When Users Change Their Discursive Behavior in Human-robot Interaction. In *2008 IEEE International Conference on Robotics and Automation*. 3481–3486. https://doi.org/10.1109/ROBOT.2008.4543743

[41] Paul A Luce and David B Pisoni. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing* 19, 1 (1998), 1.

[42] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. 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, 5286–5297. https://doi.org/10.1145/2858036.2858288

[43] Ritch Macefield. 2007. Usability Studies and the Hawthorne Effect. *Journal of Usability Studies* 2, 3 (May 2007), 145–154.

[44] 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

[45] Fernando Martinez, Daniel Tapias, Jorge Alvarez, and Paloma Leon. 1997. Characteristics of Slow, Average and Fast speech and Their Effects in Large Vocabulary Continuous Speech Recognition. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*. 469–472.

[46] Sarah C Mason. 2019. Is There a Correlation Between Oral Reading Rate and Social Conversational Speaking Rate? (2019).

[47] Graeme McLean and Kofi Osei-Frimpong. 2019. Hey Alexa... Examine the Variables Influencing the Use of Artificial Intelligent In-home Voice Assistants. *Computers in Human Behavior* 99 (2019), 28–37.

[48] Nikki Mirghafori, Eric Foster, and Nelson Morgan. 1995. Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes. In *Fourth European Conference on Speech Communication and Technology*.

[49] Tuuli Morrill, Melissa Baese-Berk, and Ann Bradlow. 2016. Speaking rate consistency and variability in spontaneous speech by native and non-native speakers of English. In *Proceedings of the International Conference on Speech Prosody*, Vol. 2016. 1119–1123.

[50] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. 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/10.1145/3173574.3173580

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

[52] Meghan Neumer. 2013. The Relationship Between Natural Speech Rate and Oral Reading Fluency Rate and Reading Comprehension Among Third Grade Students. (2013).

[53] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and speech* 41, 3-4 (1998), 419–442.

[54] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Yokohama, Japan, 19 pages. https://doi.org/10.1145/3411764.3445565

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

[56] Seongjin Park and John Culnan. 2019. A Comparison Between Native and Non-native Speech for Automatic Speech Recognition. *The Journal of the Acoustical Society of America* 145, 3 (2019), 1827–1827.

[57] Hannah R.M. Pelikan and Mathias Broth. 2016. *Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk*. Association for Computing Machinery, New York, NY, USA, 4921–4932. https://doi.org/10.1145/2858036.2858478

[58] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. *Voice Interfaces in Everyday Life*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[59] 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.

[60] K Sreenivasa Rao and Shashidhar G Koolagudi. 2013. Robust Emotion Recognition using Speaking Rate Features. In *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer, 85–94.

[61] 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

[62] Marit Ruitenberg, Elger Abrahamse, Elian De Kleine, and Willem B Verwey. 2014. Post-error Slowing in Sequential Action: An Aging Study. *Frontiers in psychology* 5 (2014), 119.

[63] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[64] T. Shinozaki and S. Furui. 2001. Error Analysis Using Decision Trees in Spontaneous Presentation Speech Recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.* 198–201. https://doi.org/10.1109/ASRU.2001.1034621

[65] M.A. Siegler and R.M. Stern. 1995. On The Effects of Speech Rate in Large Vocabulary Speech Recognition Systems. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 612–615 vol.1. https://doi.org/10.1109/ICASSP.1995.479672

[66] Matthew A Siegler and Richard M Stern. 1995. On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems. In *1995 international conference on acoustics, speech, and signal processing*, Vol. 1. IEEE, 612–615.

[67] Brent Spehar, Stacey Goebel, and Nancy Tye-Murray. 2015. Effects of Context Type on Lipreading and Listening Performance and Implications for Sentence Processing. *Journal of speech, language, and hearing research* 58, 3 (2015), 1093–1102.

[68] 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

[69] Lijun Wang, Weigang Pan, Jinfeng Tan, Congcong Liu, and Antao Chen. 2016. Slowing After Observed Error Transfers Across Tasks. *PloS one* 11, 3 (2016), e0149836.

[70] Zhirong Wang, T. Schultz, and A. Waibel. 2003. Comparison of Acoustic Model Adaptation Techniques on Non-native Speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2003.1198837

[71] Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an Integrated Understanding of Speaking Rate in Conversation. In *Ninth International Conference on Spoken Language Processing*.

[72] Xiangyu Zeng, Shi Yin, and Dong Wang. 2015. Learning Speech Rate in Speech Recognition. *arXiv preprint arXiv:1506.00799* (2015).