Is the Same Performance Really the Same?: Understanding How Listeners Perceive ASR Results Differently According to the Speaker's Accent

SEOYOUNG KIM, School of Computing, KAIST, Republic of Korea
YEON SU PARK*, School of Computing, KAIST, Republic of Korea
DAKYEOM AHN*, College of Education, SNU, Republic of Korea
JIN MYUNG KWAK, School of Computing, KAIST, Republic of Korea
JUHO KIM, School of Computing, KAIST, Republic of Korea

Research suggests that automatic speech recognition (ASR) systems, which automatically convert speech to text, show different performances according to various input classes (e.g., accent, age), requiring attention to building fairer AI systems that would perform similarly across various input classes. However, would an AI system with the same performance regardless of input classes really be perceived as fair enough? To this end, we investigate how listeners perceive the ASR system of the same result differently according to whether the speaker is a native speaker (NS) or a non-native speaker (NNS), which may lead to unfair situations. We conducted a study (n = 420), where participants were given one of the ten speech recordings with various accents of the same script along with the same captions. We found that even with the same ASR output, listeners perceive the ASR results differently. They found captions to be more useful for NNS's speech and blamed NNS more for the errors than NS. Based on the findings, we present design implications suggesting that we should take a step further than just achieving the same performance across various input classes to build a fair ASR system.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: perceived fairness of AI, automatic speech recognition (ASR), human-subject experiments

ACM Reference Format:
Seoyoung Kim, Yeon Su Park*, Dakyeom Ahn*, Jin Myung Kwak, and Juho Kim. 2024. Is the Same Performance Really the Same?: Understanding How Listeners Perceive ASR Results Differently According to the Speaker's Accent. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 169 (April 2024), 22 pages. https://doi.org/10.1145/3641008

1 INTRODUCTION

Automatic Speech Recognition (ASR), which automatically converts human speech to text, has allowed various technologies to foster human-to-human communication. For instance, audio conferencing tools (e.g., Zoom, Otter.ai) or video platforms (e.g., YouTube) integrate ASR technology to

*Both authors contributed equally to this research.

Authors' addresses: Seoyoung Kim, School of Computing, KAIST, Daejeon, Republic of Korea, youthskim@kaist.ac.kr; Yeon Su Park*, School of Computing, KAIST, Daejeon, Republic of Korea, yeonsupark@kaist.ac.kr; Dakyeom Ahn*, College of Education, SNU, Daejeon, Republic of Korea, adklys@snu.ac.kr; Jin Myung Kwak, School of Computing, KAIST, Daejeon, Republic of Korea, kwak.jinmyung@kaist.ac.kr; Juho Kim, School of Computing, KAIST, Daejeon, Republic of Korea, juhokim@kaist.ac.kr.

ACM 2573-0142/2024/4-ART169
https://doi.org/10.1145/3641008

help users better understand each other in conversations. In the era of globalization, the ASR system is especially showing its potential in supporting communications in multilingual environments where people speaking different languages interact together with lingua franca. Despite its wide usage and practical values, ASR systems are also known for showing performance disparity regarding the speaker's various characteristics, such as accent, gender, age, and speech habit [28, 47]. For ASR technologies to be inclusive, previous research has raised the importance of reducing the performance gap.

However, would an ASR system with the same performance across various speakers' characteristics actually be fair? Previous work suggests that users' perception or reaction to the outcomes of an AI system depends not only on its performance but also on many different factors such as types of errors [37], explainability [29, 62], or user interface of the system [5]. Even if the performance is the same, an ASR system cannot be considered fair if it brings disparate impact across various subgroups [6]. Thus, to comprehend its impact, it is important to understand how listeners perceive the ASR result, considering that listeners are the primary users of the ASR systems. Although there may exist various speakers' characteristics (e.g., one's own accent, tone of the speech) which may affect how listeners perceive the ASR results, we specifically focus on native speakers (NS) and non-native speakers (NNS). This is because previous communications research suggests that speakers' characteristics, such as accents, influence how listeners perceive their speech; NNS may be perceived more negatively than NS even with the same speech content [30]. Would using ASR systems aggravate this unfavorable situation for NNS? Furthermore, would using an ASR system with the same performance for NS and NNS result differently?

To this end, we aim to understand how listeners perceive the ASR result differently when the speaker is NS and NNS with two conditions: (1) when given the same performance and (2) when there exists a performance disparity similar to the current status of ASR models. We conducted a study (n=420) where we showed a video with one of 10 speech recordings (5 NS, 5 NNS) of reading the same script along with the same caption. The participants were told that the caption was automatically generated by AI and were asked to complete a survey that asks how they (1) perceive the AI system and its output, (2) perceive the speaker and their speech, and (3) attribute the errors of the captions. We found that although the caption's performance was the same, there exist differences in how listeners perceive them. We also found that there exists an even bigger gap in listeners' perception given the performance gap similar to the current status of ASR models. Based on the findings, we present design implications for both ASR model developers and ASR system developers to build a fair and more inclusive ASR system.

The contributions of the paper are as follows:

- Findings from our study on how listeners (1) perceive AI system and its output, (2) perceive the speaker and their speech, and (3) blame the errors in the caption depending on the performance of ASR and whether the speaker is NS or NNS
- Design implications for building fairer ASR systems

2 RELATED WORK

We review previous work on (1) bias in AI systems, (2) users' perception of ASR systems, and (3) listeners' perception of NNS's speech.

2.1 Bias in AI Systems

Previous research demonstrates that AI systems and their algorithms can maintain societal prejudices due to their skewed or imbalanced training data [35]. Since AI systems are dependent on the observable characteristics of the data (e.g., gender, age, skin color, regional accents), they will

learn and reflect certain biases in their outputs. Buolamwini et al. reported that facial recognition algorithms most frequently misclassify darker-skinned women whereas lighter-skinned men show the lowest error rate [10]. Moreover, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software used by the US courts was found to be negatively biased against African-Americans regarding the risk of re-offending, resulting in an unfair decision-making process like harsher sentences [44].

Similarly, ASR systems are found to have certain biases as well. ASR systems are typically trained with English speech data usually spoken by people with certain characteristics, underrepresenting diverse styles of speech [31]. Feng et al. [20] and Koenecke et al. [28] found that ASR systems struggle with speech with large variations—such as gender, age, speech impairment, race, and accents—due to the training set with limited speech diversity. Tatman et al. also found that ASR systems show different performance—word error rate (WER)—between dialects and races; white speakers using Mainstream US English showed the lowest average error rate [49]. Current ASR systems are struggling to achieve high accuracy across diverse speaker groups.

Despite these known biases in ASR systems, little prior work has examined how the users of the ASR systems, i.e., the listeners, would perceive the ASR results even when the ASR systems achieve similar performance between NS and NNS. If the listeners perceive differently despite the same performance level, this may also cause biased and unfair situations. Therefore, our work investigates how listeners perceive the ASR results differently according to whether the speaker is NS or NNS.

2.2 Users' Perceptions of ASR System

Previous research has found that the perceived accuracy may significantly differ from the calculated accuracy of AI systems. Papenmeier et al. found that even with classifiers with equal accuracy, the prediction mistakes made on impossible-to-classify sentences were perceived to be significantly more accurate than easy or difficult-to-classify sentences [42]. Yin et al. also found that people's trust in the model can be significantly affected by the perceived accuracy regardless of its calculated accuracy as well [60].

This can be applied to the ASR systems as well. The performance of the ASR systems is generally calculated with word error rate (WER). However, Mishra et al. [34] found that while WER considers all words as equally important, in practice, users' perceptions of errors are not the same. They noticed that the users' satisfaction regarding the ASR result is dependent on the severity of the errors (e.g., incorrect recognition of names or phone numbers compared to deletion/substitution of function words like *a* or *the*) than just on the overall performance. This is because listeners judge and perceive the accuracy based on whether the transcription of the ASR systems captures the meaning of the speech.

However, little research has focused on how the users' perceived accuracy of the ASR can be different even with the same ASR result depending on whether the speaker is NS or NNS. Hence, our study focuses on how the perceived accuracy of the same caption with the same performance differs depending on NS or NNS.

2.3 Listeners' Perceptions of NNS's Speech

Previous research has investigated the negative influence of NNS's accent on listeners. Lev-Ari et al. found that the NNS's accent serves as a signal that the speaker is an out-group member as well as a factor that makes the speech more difficult to understand [30]. They showed that accented speech reduces "processing fluency". Moreover, the listeners perceived accented speech as less truthful and perceived NNS as less credible [30]. Similarly, it was found that the listeners make lower ratings in terms of attractiveness, benevolence, and trustworthiness to speakers with accents that are

considered foreign or spoken by minorities) [21]. Consequently, listeners' different perceptions according to one's accent may bring severe social consequences such as discrimination [7].

Nonetheless, little prior work has investigated how listeners perceive the speaker differently according to whether the speaker is NS or NNS when they also get exposed to the ASR output of their speech. Since ASR technologies are frequently used in computer-mediated communications to support multilingual communication settings, it is important to investigate how ASR technology affects the listener's perception. Hence, our study attempts to investigate how different qualities of captions change the listener's perception of NNS's speech, as well as how this prior perception may result in blaming the speaker for the errors in the ASR-generated captions.

3 METHOD

We investigate how the listeners perceive the ASR system differently between speakers being NS or NNS.

Our research questions are:

- **RQ1**. **[Perception towards the ASR system and its output]** How do listeners perceive the ASR system and its output differently between NS and NNS?
- **RQ2**. **[Perception towards the speaker and their speech]** How do listeners perceive the speaker and their speech differently between NS and NNS when given ASR output?
- **RQ3**. **[Error blaming]** How do listeners of the ASR system attribute the errors differently between NS and NNS?

We investigate these questions in two conditions: (1) when the ASR system shows the same performance regardless of whether the speaker is NS/NNS and (2) when performance disparity exists similar to the current status of ASR systems — NS getting captions of high performance, while NNS getting those of low performance — as shown in Section 4.1.

For RQ1, we investigate (1) how listeners view the quality of the ASR system and (2) how listeners perceive the usefulness of the generated captions.

For RQ2, we investigate how listeners perceive (1) the speaker's accent negatively affecting their understanding of the video content, (2) the speaker's expertise in the subject matter, (3) the reliability of the speaker's explanation, and (4) the speaker's delivery skill.

For RQ3, we investigate how listeners attribute the ASR errors to (1) the ASR system and (2) the speaker.

We conducted a study with 420 participants with 10 speech recordings of five NS and five NNS to investigate how listeners perceive the ASR result differently between the two groups. We explain the details of how we prepared the study and the study procedure below.

3.1 Study Preparation

Here we explain how we set up our study, in terms of how we created the 10 recordings and videos used in the study and how we created the experimental conditions.

3.1.1 Speaker recruitment. We recruited five native speakers and five non-native speakers of English to record the same script. Participant recruitment was conducted both offline and online. For online, we posted recruitment calls in the communities that are likely to have members from diverse countries (e.g., a Facebook group called 'foreign friends work in Korea', international communities at universities). For offline recruitment, we posted flyers on campus. We also asked applicants or international acquaintances to promote the recruitment of participants. For the application, they were asked to input their age, gender, country of origin, country they have lived in the longest, and their use of English. They were also given a short paragraph to read aloud and were asked to record it. We recruited 10 participants by mainly considering (1) their

demographics to balance the age and gender between the two groups, and (2) nationality and their region so that the participants are from different parts of the world. Since one's own accent may differ significantly even within a country [52] for non-native speakers, we tried to select those with strong representative English accents of their own country. For this, we selected NNS who answered the country they lived in the longest corresponded to their nationality. Moreover, since some applicants had weak English accents, two or more authors went over the candidates and selected those whom they both agreed on having strong English accents. Table 1 shows the information of the speakers. We also asked if they had other factors that could potentially affect their speaking ability. Although NNS1 said ADHD, they said it does not affect their pronunciation, but just influences the duration of getting the recording task done. Moreover, NNS4 reported wearing braces, but since wearing braces can be common, we proceeded with the recording.

3.1.2 Audio recording. For each of the 10 selected participants, we proceeded with a remote recording session, which lasted approximately two hours. To avoid external noise during the recording as much as possible, participants were asked to use a microphone and computer in a quiet place during the session. Since the listener's perception towards the speaker and the ASR system may differ greatly according to various factors (e.g., level of the topic, word selection in the speech, grammatical errors), all speakers were given the same script. The script was based on the lectures from a Coursera course 'Understanding the Brain: The Neurobiology of Everyday Life' [1] and was modified so that it contains no grammar errors, contains no domain-specific jargon, and the flow is smooth. The script was about introducing the four functions of the brain: voluntary movement, perception, homeostasis, and higher cognitive abilities. The script consisted of 843 words (10 paragraphs). The speakers recorded under the presence of at least one author and were asked to record at least three times per paragraph using their own pronunciation and voice. If they made mistakes during the recording, such as mispronouncing a word, correcting themselves, or unnatural hesitation, they were asked to say "one, two, three" and then re-start recording the sentence from the beginning. To ensure the quality of the recordings, at the recording session, at least one author went over the recordings and asked for re-recording if there still remained any problem. After they finished the recording, we chose the version with the fewest external noise and unintended pauses between the words compared to the other versions recorded by the same speakers.

3.1.3 Audio post-processing. We post-processed the voice recordings, such as adjusting the average volume of the audio to ensure consistency across all speakers' recordings and experimental conditions. We also trimmed out the mistakes and instances of "one, two, three". Furthermore, we concatenated the paragraph-by-paragraph recordings into one recording, setting the time interval to about two seconds between them so that the listeners could tell that the paragraphs were broken up. We removed any non-voice noise such as background noises at the beginning or end of each paragraph recording.

3.1.4 Experimental Conditions. We had three caption conditions to mimic the three levels of performance of ASR: Word Error Rate (WER) 5%, WER 15%, and WER 30%. We decided on these three WERs considering the current ASR technology performance from Section 4.1 as well as the literature on it. Research suggests that humans can perceive differences in WER that are greater than 5-10% [53] and that 20% is the critical point for the transcription to be useful and acceptable [11, 41]. Therefore, we set our caption conditions as WER 5%, 15%, and 30%.

[1] https://www.coursera.org/learn/neurobiology

Table 1. Demographics and information of speakers (5 NS, 5 NNS) who participated in audio recording

Participant		English as Mother Tongue	Nationality	City and Country Lived the Longest	Age	Gender	Fluent in English*	Close with Accented English Speaker*	Use Accented English*	Use English Frequently*	Speak English Frequently*	Other Factors Affecting Speaking Ability
NS	1	Yes	USA	California, USA	22	Female	7	4	2	7	7	None
	2	Yes	UK	London, UK	23	Female	7	7	1	6	6	None
	3	Yes	New Zealand	Auckland, New Zealand	41	Male	7	7	1	7	7	None
	4	Yes	Australia	Sydney, Australia	24	Female	7	7	3	7	7	None
	5	Yes	UK	Liverpool, UK	25	Male	7	7	7	7	7	None
NNS	1	No	Greece	Athens, Greece	30	Male	6	1	6	7	7	ADHD
	2	No	Ecuador	Quito, Ecuador	25	Female	6	6	6	6	6	None
	3	No	Tajikistan	Panjakent, Tajikistan	25	Female	3	6	7	3	3	None
	4	No	Vietnam	Binh Thuan Province, Vietnam	25	Female	6	6	6	6	6	Wearing Braces
	5	No	China	Anshan, China	23	Male	3	5	6	5	3	None

*The answer was given as 7-point Likert scale (1 = very strongly disagree, 7 = very strongly agree)

Even with the same WER, the perceptions of how listeners perceive the caption could be different depending on the error type in the caption [61]. Therefore, we made the same caption with each of the WER to be used across all of the recordings.

To generate captions with a target WER, we utilized the errors in generated captions from various ASR technologies 2. If the errors appear in the captions of multiple speakers and multiple ASR technologies, it is likely that the error is plausible (e.g., for the error that was originally 'while' but got translated to 'well' by ASR, was shown in six out of ten different speakers, seven out of seven different ASR models). Therefore, we selected frequent errors until two-thirds of the target WER was reached. For the remaining one-third of the target WER, we selected errors by considering the distribution of already selected errors. This is because if the errors are concentrated in a certain part of the recording, the participants may perceive the quality of the transcription differently. Thus, to make sure that the errors are evenly distributed throughout the recording, we partitioned the transcript into micro level (three words per partition; total 281 parts) and macro level (approximately 84 words per partition; total 10 parts) and set the entropy of the sum of substitutions (S), insertions (I), and deletions (D) to be higher when selecting errors to include in the caption. We also kept the capitalization, punctuation marks, and abbreviations (e.g., it's, they're) unchanged from the original script unless they are interpreted differently by ASR technologies.

In this way, the total number of words from the original captions may change. To make sure that the changed caption also appears as the speaker speaks, one author first created ground truth captions per recording by syncing the original script with the timestamps of the speaker's utterance as a subscription text (SRT) file, and then we synced the changed caption to the ground truth caption. We synced the changed caption so that the same number of words can newly appear in each caption. If the number of words does not divide evenly, we made the remaining words appear in later parts, so that the captions do not appear faster than the voice. Finally, to give the impression that the caption is auto-generated by an ASR system, we delayed the caption timestamps by 0.2 seconds and explicitly told the participants that the caption was auto-generated by an AI.

3.1.5 Video processing. We created videos that experiment participants would watch, with the audio of the processed audio and captions of a given condition. Since the perception towards the speaker may be largely influenced by the appearance of the speaker, we created videos of black

screen. But to make sure that the listener is looking at the video and to filter out those who have not paid attention to the video, we also inserted four pictures of animals at approximately 1/5, 2/5, 3/5, and 4/5 points of the video for 10 seconds each. The images were selected so that each could leave a strong impression so that the participants could pass the attention check question without much effort.

3.2 Study

Here we explain how we conducted our study.

3.2.1 Study Participants. We recruited a total of 420 participants through Prolific [2] to watch a video with the captions of a certain experimental condition. Participants were paid 3.75 GBP (approximately 5 USD) for the task which took about 25 minutes. For each condition (total 30 conditions = 10 videos (5 NS, 5 NNS) × 3 levels of caption conditions (WER 5, 15, 30)), we recruited 14 listeners (7 NS, 7 NNS).

Since the videos were in English, we only recruited participants who could understand spoken English — who answered more than 4 for the questions on how much they can understand spoken English in daily conversation as well as in academic purposes with a 7-point Likert scale (1 = cannot understand at all, 7 = can fully understand).

3.2.2 Study Procedure. We created an interface to conduct the study on Prolific. The main purpose of this interface was to assign a condition for each participant, allow them to watch the video in a controlled environment, and collect their answers for the post-survey regarding their experiences and perceptions.

Video Assignment. The listener's familiarity with the accent may influence how the listener perceives the ASR system and the speaker [2]. Since people are relatively more familiar with NS's accents while less familiar with NNS's accents in the real world, (Section 4.2), we assigned NS's video if the listeners are familiar with the NS's accents or NNS's video if the listeners are not familiar with the NNS's accent.

For this, participants were asked to provide their familiarity with the 10 accents of the speakers in the videos. The participants had to listen to each sound clip, which was around 10 seconds long, and provide their familiarity with the accent on a 7-point Likert scale (1 = not familiar at all, 7 = extremely familiar). These sound clips were chosen based on internal consensus; three of the authors picked multiple clips from each video that represented the characteristics of the accents of the speakers. Among the clips that two or more authors agreed on, the final clips were decided so that all speakers did not share a common line of the script. This was done since listening to the same script multiple times could influence how well the participant hears the sound clip, potentially affecting participants' familiarity with the accent. To prevent any ordering effects, the sound clips were provided in random order.

Based on their familiarity with the speakers' accents, we randomly assigned a NS's video that they had rated 5 or higher, or a NNS's video that they had rated 3 or lower. Hence, if a participant rated their familiarity higher than 4 for all NNS and less than 4 for all NS or if all of the videos that can be assigned are fully assigned already (i.e., 14 participants per each condition who fully completed the task and is not filtered out (Section 3.3)), they completed the task at this point and were paid 0.9 GBP (approximately 1 USD).

Instructions. The interface provided a set of instructions prior to watching the video. This was given so that the participant could prepare the appropriate environment to watch the video.

[2]https://www.prolific.co/

First, the participants were asked to play a 5-second audio clip with a beep sound to unmute or adjust their audio to be able to clearly hear the sound of the assigned video. Second, the main task was explained. The participants were told that there would be a video with captions auto-generated by the AI system. They were warned that the video interactions (i.e., pause, skip forward, skip backward, re-watching) would be disabled once the video started playing. This was done so that the listeners only get exposure to certain parts of the caption once. Furthermore, they were warned that they must not refresh the page as this could also allow them to re-watch the video. Third, they were asked to concentrate on the video, as they would be asked to recall some content of the video later for attention-checking purposes. Similarly, they were also told to concentrate on the screen as animal images would randomly appear in the video, which would also be asked later. This was to make sure that the participants actually listened to the video and watched the screen. However, we did not ask them to always look at the caption so that it could mimic the real-world situation, where they get to switch back and forth between the screen and the captions in their free will.

Video Watching. After the instructions, the participants were shown a video of the assigned condition. Although the speakers all read the same script, the length of the video slightly varied depending on their speaking speed. The average length of the video was 363.5 seconds (SD = 77.4, min = 270.0, max = 494.0). As instructed, all video controls (pause, skip forward, skip backward, re-watching) were disabled. Any action of refreshing the page was recorded if it occurred. A separate progress bar was inserted at the bottom of the video to show the progress so that the participants can track where they are at and how much is left. This was done to help the participants maintain their concentration. After the video ended, they could move on to the post-survey.

Post-Survey. We asked the following main questions to answer our RQs:

- RQ1. [Perception towards ASR system and its output]
  - How would you rate the quality of the AI technology that generated the captions you watched in the video? (1 = very poor, 7 = excellent)
  - Captions were useful for recognizing the speaker's pronunciation. (1 = strongly disagree, 7 = strongly agree)
- RQ2. [Perception towards the speaker and their speech]
  - How much did the speaker's accent negatively affect your understanding of the video content? (1 = never, 7 = every time)
  - The speaker is highly knowledgeable about the subject matter (1 = strongly disagree, 7 = strongly agree)
  - I found the speaker's explanation to be reliable and trustworthy. (1 = strongly disagree, 7 = strongly agree)
  - The speaker is highly skilled in delivering the content. (1 = strongly disagree, 7 = strongly agree)
- RQ3. [Error blaming]
  - How much of the errors in the caption do you think were caused due to the speaker? (e.g., speaker's accent, way of talking, or speaking habits) (1 = never, 7 = every time)
  - How much of the errors in the caption do you think were caused due to the AI? (e.g., AI's low performance of recognizing speech) (1 = never, 7 = every time)

Since questions on error blaming could not be answered if the participant did not notice any errors in the caption, we asked a question whether they noticed any errors, and if they answered "no", error-blaming questions were skipped. Furthermore, through pilot studies, we found that some could not answer questions on the expertise of the speaker and the reliability of the speaker's explanation, we provided an option to skip the question.

For quality control, participants were asked to select an animal image that did not appear in the video. Moreover, they were asked to briefly write the topic of the video.

3.3 Participant Filtering

Before analyzing the data, we first filtered out the participants who failed to pass our quality control questions or who were not suitable for our analysis. The criteria to exclude were:

- The participant refreshed the page once or more.
- The participant was incorrect on the quality-control question regarding animal images from the video.
- The participant was completely wrong in describing the topic of the video.
- The participant did not pay attention to the caption (answered below 3 in question on 'How much attention did you pay to the captions while watching the video? (1 = never, 7 = every time)').
- The participant did not notice any errors in the caption.
- The participant did not show an opposite trend for the same question but when asked in an opposite manner.

We ran the study until each of the 30 conditions gathered responses from 14 participants (7 NS, 7 NNS) after filtering out using the above criteria, resulting in a total of 420 participants. Out of 580 participants who completed the task, 160 were excluded.

4 RESULT

In this section, we first present our analysis of the performance of eight different ASR models. Then, we present the distribution of familiarity towards NS and NNS's accents. Finally, we present the results to understand how listeners perceive the ASR system (RQ1) and the speaker (RQ2) and how they attribute the errors (RQ3) differently when the speaker is NS and when the speaker is NNS. We present these results under two circumstances: (1) given the same ASR performance regardless of the speaker being NS/NNS and (2) given the disparity gap which reflects the current status of ASR systems as in Section 4.1.

4.1 Performance of ASR Model

We used the 10 recordings (5 NS, 5 NNS) that we collected and processed (Section 3.1.2 and 3.1.3) to understand and compare the performance of various ASR models available: (1) three commercialized ASR technology integrated into video platforms or meeting support systems, namely YouTube automatic captioning function [3], Zoom automated captions function [4], and Otter.ai [5]), (2) four ASR APIs, namely Rev AI [6], AssemblyAI [7], Amazon Transcribe [8], and IBM Watson [9]), and (3) one of the state-of-the-art ASR models, namely OpenAI's Whisper [45] in two versions (English-only model and multilingual model, both in base size).

Results (Table 2) show that the performance of ASR models measured by WER varies significantly across different speakers and models: 0.5% (AssemblyAI, Speaker 1 (NS)) to 100% (YouTube, Speaker 10 (NNS), failed to output ASR result). The performance of a single ASR model also varied a lot (std WER of YouTube: 30.3%).

[3]https://support.google.com/youtube/answer/6373554?hl=en
[4]https://support.zoom.us/hc/en-us/articles/8158289360141-Enabling-automated-captions
[5]https://otter.ai
[6]https://www.rev.ai/
[7]https://www.assemblyai.com/
[8]https://aws.amazon.com/transcribe/
[9]https://www.ibm.com/cloud/watson-speech-to-text

Despite the performance differences across speakers, we could still find a clear tendency for the ASR performance of NNS to be lower than that of NS overall. The average WER of NS varied from 1.3% to 29.8%, while the average WER of NNS varied from 8.0% to 65.6%. The ASR performance was also more stable across NS than NNS (std for NS: 1.2% to 13.8%, NNS: 4.5% to 40.4%). For one NNS, YouTube even failed to output the ASR result, resulting in 100% WER. This result of the disparity gap between NS and NNS also aligned with previous work [8].

Interestingly, all three commercialized online platforms that integrate ASR technology (YouTube, Zoom, Otter.ai) and IBM Watson showed huge performance disparity (approx. 15%) between NS and NNS compared to other models. In addition, considering 20% WER to be the critical point for ASR models to be useful [11, 41], only two of the commercialized ASR technology (YouTube, Zoom) showed the performance of being useful for NS, while none of them were useful for NNS. This could mean that users in the wild would be perceiving a larger performance disparity between NS and NNS compared to what is being reported in previous research or API documents.

Table 2. Average and Standard Deviation of Accuracy (WER) of Five NS and Five NNS for each ASR model

		YouTube	ZOOM STT	Otter.ai	Rev AI	AssemblyAI	Amazon	IBM Watson	Whisper Base.en	Whisper Base
NS	AVR	2.5	8.0	4.3	2.9	1.3	2.1	29.8	2.5	3.4
	STD	2.4	7.1	3.9	2.9	1.2	1.9	13.8	1.6	3.0
NNS	AVR	28.64	24.6	19.1	14.8	8.0	9.0	65.5	11.2	17.0
	STD	40.4	12.1	12.1	7.2	5.4	4.5	19.1	6.2	13.9

4.2 Familiarity towards NS and NNS's accents

Since the listener's familiarity with the speaker's accent may affect the listener's perceptions, we analyzed people's accent familiarity to take this factor into account when assigning conditions in the study. We analyzed 875 participants' (405 NS, 470 NNS) responses to accent familiarity for each of the 10 recordings (5 NS, 5 NNS) on a scale of 1 (not familiar at all) to 7 (extremely familiar). These participants included the main study participants (420 total), as well as participants who (1) dropped out after completing accent familiarity ratings, (2) could not proceed to the main task as the videos that could be assigned were already fully assigned to 14 participants, and (3) got excluded later in the main study (Section 3.3). There were no participants who could not proceed to the main task due to rating their familiarity higher than 4 for all NNS and lower than 4 for all NS.

Figure 1 shows the distribution of accent familiarity towards speakers differed depending on whether the speaker was NS or NNS. Participants were relatively **more familiar with NS's accents** and **less familiar with NNS's accents**: on average, 48.9% of participants were familiar (responded with 4-7) and 5.09% of participants were unfamiliar (responded with 1-3) with NS's accents. While 34.22% of participants were familiar, 88.3% of participants were unfamiliar (responded with 1-3) with NNS's accents.

We also conducted Pearson's Chi-squared test and found that there is a significant relationship between accent familiarity and whether the speaker is NS or NNS ($\chi^2 = 3491.5$, $df = 6$, $p < 0.001$). Cramér's V, which shows how strongly two categorical variables are associated, was 0.63, meaning that the two variables have a strong relationship. Including variables with a strong relationship in the statistical model may result in multicollinearity, which results in high errors in the analysis result. Thus, we decided to exclude accent familiarity as a factor in the following statistical analysis as suggested by Dormann et al. [14].

At the same time, accent familiarity differed even within the NS and NNS groups. For instance, participants were relatively less familiar with NS5's accent compared to NS1's accent, shown in Figure 1.

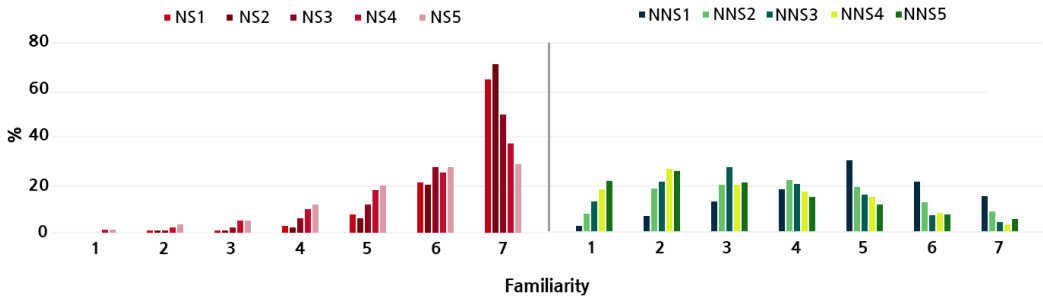

Fig. 1. Listeners' familiarity with accents of NS and NNS

4.3 Perception difference when the performance is the same

4.3.1 *RQ1: Perception towards ASR system and its output.*

Quality of ASR system. We performed Aligned Rank Transform [56] to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners perceive the quality of the ASR system (Figure 2-left). We found that the performance of ASR has a statistically significant effect ($F_{2,417} = 86.07$, $p < .001$, $\eta_p^2 = 0.29$), while **whether a speaker is NS/NNS did not have a significant effect on how listeners perceive the quality of the ASR system** ($F_{1,418} = 3.49$, $p > .05$). We also did not observe a significant interaction between these two factors ($F_{2,414} = 1.77$, $p > .05$).

For the post-hoc pairwise comparison, we used ART-C contrasts with Tukey adjustment [17]. Results showed that listeners **perceived the quality of the ASR system to be significantly better as the WER gets lower** in our experimental conditions: WER 5 ($M = 5.26$, $SD = 1.03$) > WER 15 ($M = 4.60$, $SD = 1.21$, $p < .001$), WER 5 > WER 30 ($M = 3.37$, $SD = 1.44$, $p < .001$), WER 5 > WER 15 ($p < .001$).

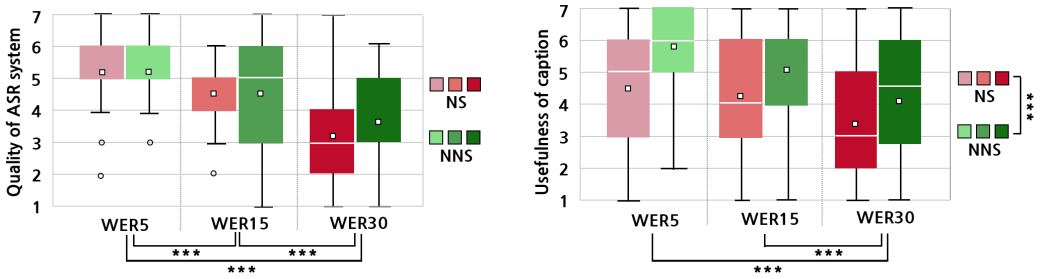

Fig. 2. Listeners' perception towards the ASR system and its output (*** $p < .001$)

Usefulness of captions. We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners perceive the usefulness of the caption (Figure 2-right). We found that whether a speaker is NS/NNS ($F_{1,418} = 39.24$, $p < .001$, $\eta_p^2 = 0.09$) and the performance of ASR ($F_{2,417} = 22.42$, $p < .001$, $\eta_p^2 = 0.10$) have a statistically significant effect on how listeners perceive the usefulness of the caption. However, we did not observe a significant interaction between these two factors ($F_{2,414} = 0.14$, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [17]) results showed that listeners' **perceived usefulness of captions was significantly higher when the speaker is NNS** (M = 5.05, SD = 1.69) compared to when the speaker is NS (M = 4.10, SD = 1.91, $p < .001$). Post-hoc tests also showed that listeners **perceived significantly lower usefulness for WER 30** (M = 3.77, SD = 1.85) **compared to WER 15** (M = 4.74, SD = 1.77, $p < .001$) **or WER 5** (M = 5.21, SD = 1.67, $p < .001$).

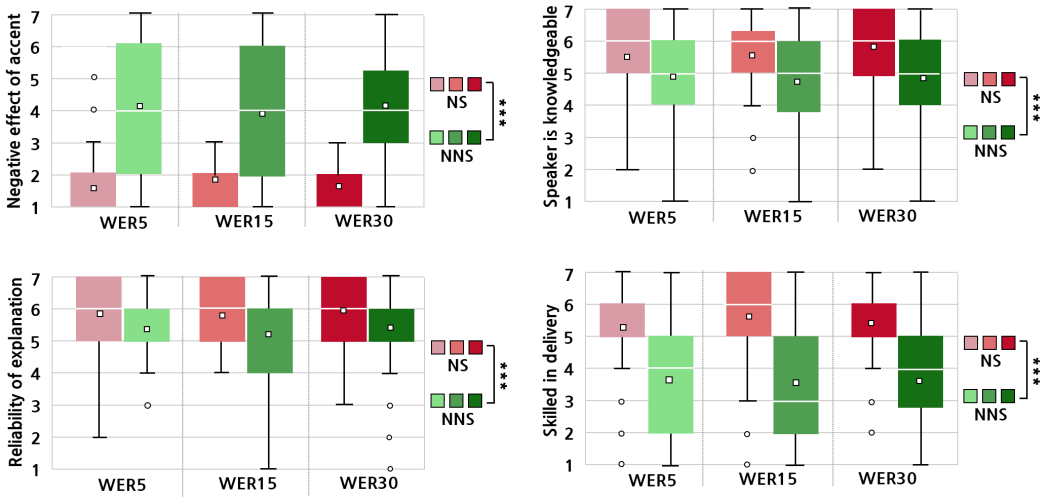

Fig. 3. Listeners' perception towards the speaker and their speech (*** $p < .001$)

4.3.2 RQ2: Perception towards the speaker and their speech. We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on the perception towards the speaker and their speech (Figure 3). We found that **whether a speaker is NS/NNS has a statistically significant effect on the perceptions towards the speaker and their speech**: perceived negative effect of the speaker's accent on their understanding of the video content ($F_{1,418}$ = 312.34, $p < .001$, η_p^2 = 0.43), perceived expertise of the speaker ($F_{1,418}$ = 25.67, $p < .001$, η_p^2 = 0.06), reliability of speaker's explanation ($F_{1,418}$ = 36.10, $p < .001$, η_p^2 = 0.09), and perceived delivery skill of the speaker ($F_{1,418}$ = 195.64, $p < .001$, η_p^2 = 0.32). On the other hand, **performance of ASR did not have a significant effect on the perceptions towards the speaker and their speech**: perceived negative effect of the speaker's accent on their understanding of the video content ($F_{1,417}$ = 0.05, not significant), perceived expertise of the speaker ($F_{1,417}$ = 0.29, not significant), reliability of speaker's explanation ($F_{1,417}$ = 0.14, not significant), and perceived delivery skill of the speaker ($F_{1,417}$ = 0.10, not significant). We also did not observe a significant interaction between these two factors: perceived negative effect of the speaker's accent on their understanding of the video content ($F_{1,414}$ = 1.02, not significant), perceived expertise of the speaker ($F_{1,414}$ = 0.10, not significant), reliability of speaker's explanation ($F_{1,414}$ = 0.09, not significant), and perceived delivery skill of the speaker ($F_{1,414}$ = 0.48, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [17]) results showed that **listeners perceived the speaker and their speech significantly negatively when the speaker is NNS** compared to when the speaker is NS: perceived negative effect of the speaker's accent on their understanding of the video content ($p < .001$, NNS: M = 4.08, SD = 1.77, NS: M = 4.08, SD = 1.77),

perceived expertise of the speaker ($p < .001$, NNS: $M = 5.66$, $SD = 1.72$, NS: $M = 4.87$, $SD = 1.30$), reliability of speaker's explanation ($p < .001$, NNS: $M = 5.35$, $SD = 1.28$, NS: $M = 5.90$, $SD = 0.97$), and perceived delivery skill of the speaker ($p < .001$, NNS: $M = 3.66$, $SD = 1.65$, NS: $M = 5.49$, $SD = 1.32$). This result aligns with previous works that also found that NNS receive unfavorable impressions compared to NS [18].

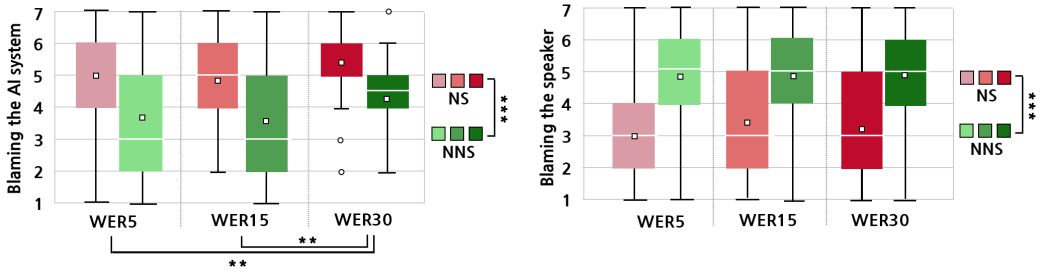

Fig. 4. How much listeners blame errors in the captions on the AI system (left) and the speaker (right) (** $p < .01$, *** $p < .001$)

4.3.3 *RQ3: Error blaming.*

Blaming the ASR system. We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners blame the ASR system for the errors in the caption (Figure 4-left). We found that whether a speaker is NS/NNS ($F_{1,418} = 74.30$, $p < .001$, $\eta_p^2 = 0.15$) and the performance of ASR ($F_{2,417} = 9.75$, $p < .001$, $\eta_p^2 = 0.05$) have a statistically significant effect on how much listeners blame the ASR system for the errors in the caption. However, we did not observe a significant interaction between these two factors ($F_{2,414} = 0.19$, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [17]) results showed that listeners **blamed the ASR system significantly more when the speaker is NS** ($M = 5.10$, $SD = 1.53$) compared to when the speaker is NNS ($M = 3.86$, $SD = 1.69$, $p < .001$). Post-hoc test results showed that listeners **blamed the ASR system significantly more in WER 30** ($M = 4.89$, $SD = 1.53$) than WER 15 ($M = 4.26$, $SD = 1.70$, $p < .01$) or WER 5 ($M = 4.28$, $SD = 1.86$, $p < .01$).

Blaming the speaker. We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners blame the speakers for the errors in the caption (Figure 4-right). We found that whether a speaker is NS/NNS has a statistically significant effect ($F_{1,418} = 112.07$, $p < .001$, $\eta_p^2 = 0.21$), while **performance of ASR did not have a significant effect on how much listeners blame the speaker** for the errors in the caption ($F_{2,417} = 1.32$, not significant). We also did not observe a significant interaction $_{2,414} = 0.65$, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [17]) results showed that listeners **blamed the speaker significantly more when the speaker is NNS** ($M = 4.90$, $SD = 1.52$) compared to when the speaker is NS ($M = 3.28$, $SD = 1.56$, $p < .001$).

4.4 Perception difference when performance disparity exists

When the ASR's performance was given higher when the speaker is NS than when the speaker is NNS (NS: WER = 5, NNS: WER = 30), we found that listeners perceived the ASR system to be of lower quality but found it more useful, while perceiving the speaker negatively when the speaker is NNS. Moreover, listeners blamed the speaker more and the ASR system less when the speaker was NNS. We report detailed results below.

4.4.1 *RQ1: Perception towards ASR system and its output.*

Quality of ASR system. We performed Mann-Whitney U test and found that **listeners perceived the quality of the ASR system significantly higher when the speaker was NS (med = 6) compared to NNS (med = 3), although the performance was lower in NNS's condition** (Figure 5-left) (U = 871, n_{NS} = n_{NNS} = 70, p < .001, r = 0.64).

Usefulness of captions. We performed Mann-Whitney U test and found **no significant difference exists on how listeners perceive the usefulness of captions between when the speaker is NS and when the speaker is NNS, although the performance was lower in NNS's condition** (Figure 5-right) (U = 2117, n_{NS} = n_{NNS} = 70, p > .05).

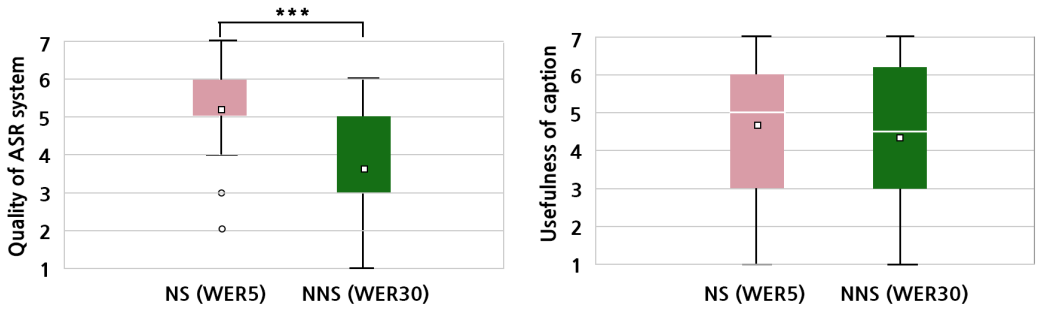

Fig. 5. Listeners' perception towards the ASR system and its output given performance disparity between NS and NNS (*** p < .001)

4.4.2 *RQ2: Perception towards the speaker and their speech.*

We performed Mann-Whitney U test and found that listeners perceived significantly negatively when the speaker is NNS compared to when the speaker is NS when performance disparity exists (Figure 6). Listeners perceived the **negative effect of the speaker's accent on their understanding of the video content to be higher** (U = 385, med_{NS} = 1, med_{NNS} = 4, n_{NS} = n_{NNS} = 70, p < .001, r = 0.84), **expertise of the speaker to be lower** (U = 1620, med_{NS} = 6, med_{NNS} = 5, n_{NS} = 66, n_{NNS} = 63, p < .05, r = 0.22), **reliability of the speaker's explanation to be lower** (U = 1409, med_{NS} = 6, med_{NNS} = 6, n_{NS} = 63, n_{NNS} = 59, p < .05, r = 0.24), and **delivery skill of the speaker to be lower** (U = 995, med_{NS} = 6, med_{NNS} = 4, n_{NS} = n_{NNS} = 70, p < .001, r = 0.59) **when the speaker is NNS** compared to when the speaker is NS.

4.4.3 *RQ3: Error blaming.*

We performed Mann-Whitney U test and found that listeners blamed the ASR system and the speaker significantly differently according to whether the speaker is NS/NNS even when performance disparity exists (Figure 7). Listeners blamed **the ASR system significantly less** (U = 1791, med_{NS} = 6, med_{NNS} = 4.5, n_{NS} = n_{NNS} = 70, p < .001, r = 0.27) and **the speaker significantly more** (U = 913, med_{NS} = 3, med_{NNS} = 5, n_{NS} = n_{NNS} = 70, p < .001, r = 0.63) **when the speaker is NNS** compared to when the speaker is NS.

5 DISCUSSION

We discuss our interpretation of the results, design implications for building a more inclusive ASR model and its applications, and the generalizability of the results.

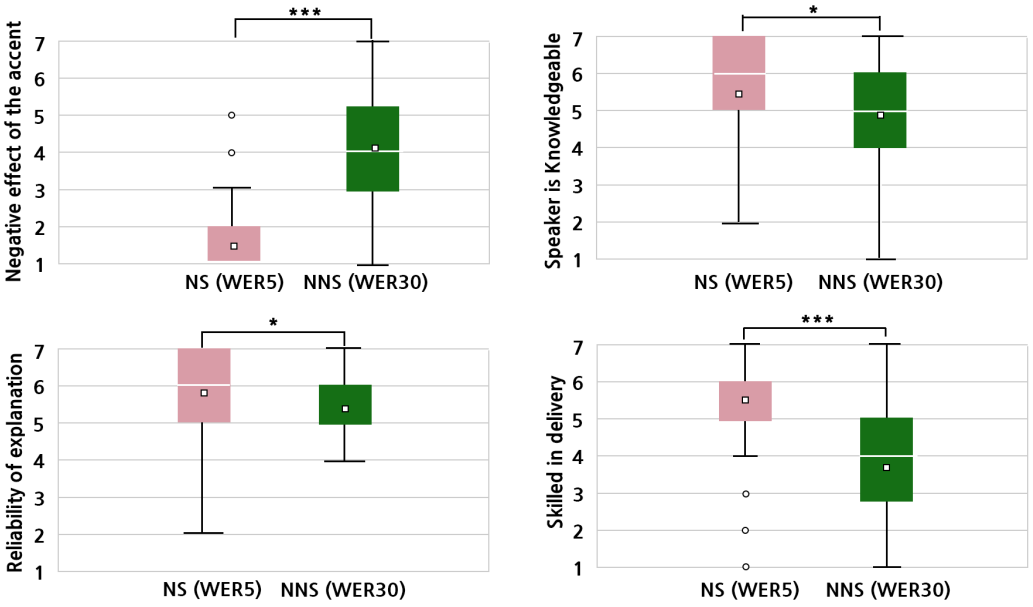

Fig. 6. Listeners' perception towards the speaker and their speech given performance disparity between NS and NNS (* $p < .05$, *** $p < .001$)

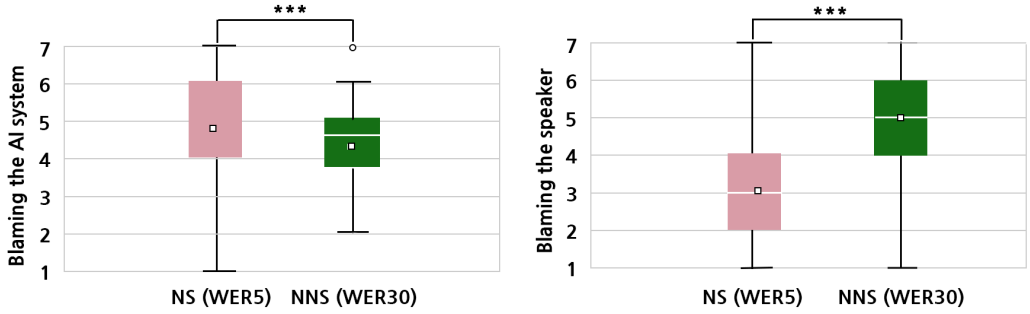

Fig. 7. How much listeners blame errors in the captions on the AI system (left) and the speaker (right) given performance disparity between NS and NNS (*** $p < .001$)

5.1 Interpretation of Results

Here we discuss our interpretation of the results and their implications.

5.1.1 Performance Disparity of ASR System and Change of Its Usage. Our study has found that even with the same performance, ASR is more useful for understanding NNS's speech than NS's speech (Figure 2). In contrast, when the usage of ASR was first introduced in computer-mediated communications, many studies focused on how it can support NNS in understanding NS's speech [11, 16, 24, 39–41, 46, 59]. ASR systems were relatively not helpful for supporting the understanding of NNS's speech and even suggested removing NNS's captions as NNS was bothered by the low-quality captions compared to NS [22]. This could be because previously, ASR models were based on hidden Markov models (HMM) to convert audio to phoneme [55], so the performance of an ASR system

for NNS could have been much lower. However, with the adoption of deep learning in ASR, there was a huge advance in the performance for various accents during the past decade [55], which may lead to a change in ASR's usage towards understanding NNS's speech. Moreover, since the room for increasing NNS's performance still remains significant (Section 4.1), we expect if this improvement is made in the future, this could also contribute to the change in ASR's usage.

To keep in line with this change in ASR's usage and make ASR technologies more useful, much research needs to be conducted. Previous studies have pointed out that WER 20% is a critical point for captions to be useful [11, 41]. However, they were investigated in the shoes of NNS understanding NS's speech. Our results suggest that the critical point of the captions' usefulness could be even different according to who the speaker is; the critical point of captions' usefulness for NNS's speech could be lower than that of NS's since our results suggest that the usefulness of captions of same performance differs according to whether the speaker is NS or NNS. Thus, investigating critical points of the captions' usefulness across various speakers is needed.

Furthermore, extensive research is needed to improve the UI/UX of ASR systems in alignment with the change in ASR's usage. Previous work focused more on ways to increase the usefulness of the ASR results of NS's speech, such as allowing NS to highlight important parts of their speech in the transcript [39] or allowing NS to edit their transcript [22]. However, the same UI/UX may not be applicable to NNS [22] as they face a higher cognitive load when communicating with others as they are relatively less familiar with the language [25]. Moreover, while NNS face more speaking anxiety than NS [3], their anxiety levels may further increase if they encounter errors in the captions of their own speech than NS. Thus, improving the UI/UX of ASR systems for NNS's speech should be further investigated to fully support the usefulness of ASR systems for NNS's speech.

5.1.2 Aggravating Unfavorable Situations for NNS. Even without ASR, previous work has suggested that NNS are more likely to be situated in an unfavorable situation or receive a more negative impression, such as being perceived as less credible compared to NS [30]. Our result also shows that NNS are perceived negatively (Figure 3). Now, with the use of ASR, our results suggest that in addition to receiving a more negative impression than NS, they also get more blame for errors in the captions than NS by using ASR technology for their speech (Section 4.4). Although ASR has the potential to support an understanding of NNS's speech, this suggests that the use of ASR technology can bring a more unfavorable environment for NNS by deteriorating the situation.

Furthermore, ASR models are now expanding their usage beyond supporting multilingual communications and being incorporated into technologies for other purposes, ranging from generating automatic transcripts in video-sharing platforms to intelligent personal assistants (IPA). NNS using those technologies may face difficulties as they are less comfortable with the language than NS in the first place. For instance, previous work found that NNS face more difficulties using IPA compared to NS as they have a hard time syntactically constructing commands in a foreign language [57]. On top of this unfavorable situation for NNS, the performance disparity of ASR may aggravate this situation; NNS may more heavily focus on their pronunciation when using IPA [58] due to the low performance, resulting in unequal user experience compared to NS. Likewise, although ASR itself may not be the main purpose of a certain technology, the performance disparity of ASR may significantly aggravate NNS's technology usage.

Likewise, Kentaro Toyama's Law of Amplification [50, 51] states that technology may widen the existing inequality unless there is an effort to reduce the inequality. Thus, to better accommodate ASR technology in various scenarios, it is crucial to be aware of the additional inequalities it could bring and design systems in a way that will decrease the unfair situation between the user groups.

5.1.3 Reason for Difference in Listeners' Perceptions Despite Identical ASR Performances. Our results suggest that although the performance of ASR is the same, listeners perceive the ASR result differently (Section 4.4). Previous work has suggested perceived accuracy may be different from the model accuracy [43]. However, in our study, the reason why the listeners' perception was different was not because the perceived accuracy was different; we found no significant difference in how they perceived the quality of the ASR system between NS and NNS (Section 4.3.1).

Instead, as our result (Section 4.2) suggests, this could be because people are more familiar with NS's accent, allowing them to hear better. Since NNS's speech was harder to hear, this could have led to higher usefulness of ASR output for NNS's speech even with the same ASR performance (Figure 2). Another potential reason behind blaming NNS more for the errors could be because people may have a subconscious bias that the NS's accent is 'better' than that of the NNS. Despite the effort of acknowledging diverse English accents as 'World Englishes' [26] in academia, bias on accents still persists in the world; previous studies have found that NNS aspired to match US or British accents and considered these accents to be the 'correct' accent [2, 9, 33].

This familiarity and bias towards certain accents may be shaped by education and media exposure. When learning English, students get frequent exposure to US or British accents in English language teaching materials [9]. English listening comprehension tests, such as TOEFL iBT, also only encompass accents of NS from certain countries (e.g., North America, the U.K., New Zealand, or Australia) [19, 32]. Furthermore, nationality is often treated as a proxy to qualify as a 'good' English teacher in the recruitment process [1]. Moreover, media may also play a significant role in one's familiarity towards certain accents and further shaping a subconscious bias. Research shows that in the media, 'non-standard' English speakers rarely appear compared to 'standard' English speakers, and even if they do so, are depicted as less favorable regarding their social or economic status and physical appearance [15]. On the other hand, news reporters are trained to speak in a certain accent [36]. Thus, to further accommodate English varieties and reduce a subconscious bias, societal measures should be taken for a gradual change in how people think and perceive different accents.

5.2 Design Implications

We present design implications for building inclusive ASR systems based on our study results.

5.2.1 Developing ASR Models. Despite various attempts to reduce the disparity gap in research [12, 54], similar to previous studies [13], our results also suggest that a huge performance disparity exists between NS and NNS (Section 4.1), while this disparity may form unfavorable situations for certain user groups. Therefore, researchers developing ASR models should put a much higher priority on reducing the disparity gap. In addition, our results show that the commercialized online platforms integrating ASR technology (e.g., YouTube, Zoom, Otter.ai) showed huge performance disparity compared to other models. This shows the needs for a quicker integration of state-of-the-art ASR models into commercialized services. Moreover, it is also important to explicitly address the performance disparity across diverse speaker groups in the reports or API documentations so that the platform builders could clearly be aware of the issue and select an appropriate model according to their purpose.

Furthermore, our results suggest that there still exist unfavorable situations for NNS regarding listeners' perceptions although model performance is the same (Section 4.3). Thus, for situations where listeners' perceptions are important, such as when ASR is used in job interviews, ASR models can be trained to incorporate listeners' perceptions, inspired by Reinforcement Learning from Human Feedback (RLHF) [38, 48, 63]. We first construct a speech and ASR result dataset with listeners' perceptions (e.g., level of blaming speakers) annotated. Then, we can train a model

to predict listeners' perceptions. Based on the prediction model, we integrate the difference in listeners' perceptions between NNS and NS as the reward signal in the ASR model, reducing the difference. This is different from RLHF, as our suggested reward allows the ASR model to learn to have similar listeners' perceptions towards different speaker groups. Although this may result in performance trade-offs, a fairer ASR model may be more desirable than one with just a higher average performance depending on one's purpose.

5.2.2 Systems Utilizing ASR Models. With the fast-evolving AI technologies and their rapid development, there exist various ASR models with different performances. Hence, when building systems that utilize ASR models, selecting an appropriate ASR model while considering its purpose would be important. Previously, this selection of the ASR model could have been solely dependent on model performance, but our results suggest that listeners' perception should be another important factor to consider. The choice of the ASR model can even depend more on differences in listeners' perception in circumstances where it is important to prevent yielding any discriminating or unfair decisions. For instance, when building an ASR system to support job interviews, preventing speakers from receiving an unfair impression while using the system should be a more important factor to consider when selecting which ASR model to incorporate compared to casual conversation settings. Selecting the ASR model focusing only on its high performance neglecting listeners' different perceptions can cause group inequity between NS and NNS. Thus, system developers should take a step back and consider the societal impact the system they are building would bring as their system would be used by various users and may shape unconscious bias.

Furthermore, speaker-adaptive ASR system is needed where different UI/UX could be provided based on who the speaker is. Previously, listener-adaptive subtitling systems have been proposed, where the rendering of subtitles is personalized to the listener and their device environment [4]. We suggest that UI/UX personalization also needs to be considered in the perspective of speakers to mitigate the unfair environment for NNS when designing an ASR system. For instance, considering our results that users tend to blame the NNS more, when NNS is speaking, the ASR system could explicitly indicate that the model is not performing well so that the listeners could be aware of AI's fault instead of focusing on NNS's pronunciation.

5.3 Generalization of Results

Our study focused on investigating how listeners perceive ASR results differently according to the speaker's accent and how it can result in unfavorable situations for NNS. This result may generalize to other AI systems where users may make value judgments on the input itself or the person who generates the input. For instance, for the output of a grammatical error correction model showing similar accuracy and errors for the same lines of writing of NS and NNS, readers may attribute the errors differently as they may have a prejudice that NS is better at using the language.

This biased perception difference not only occurs between NS and NNS. For example, when a handwriting recognition model fails to recognize certain handwriting, users may perceive differently based on the value judgment of how much the handwriting is aesthetically appealing. This may result in more severe consequences if it is related to diverse accessibility issues. For instance, if the gesture recognition model fails to recognize someone with motor impairments, people may blame the person for the errors, which may induce unjust outcomes. Thus, our results and design implications could be applied to other AI systems as well.

6 LIMITATION & FUTURE WORK

We acknowledge the limitations of our study and present possible future work.

First, although we divided speakers into NS and NNS, the way of speaking can vary greatly within each group. People from the same country can have different accents depending on their socioeconomic and sociolinguistic factors [52]. In addition, listeners may perceive ASR results differently for NS from countries where they use multiple languages as their official language (e.g., India, Kenya, the Philippines). Moreover, there exist other factors of speech that may influence the results that we did not take into consideration: the speaker's age, gender, voice tone, and speech fluency and speed could affect the listener's overall experience and perception of the speaker [23, 27]. Although focusing on NS and NNS's accents could be a meaningful start in understanding differences in listeners' perceptions, future studies could include more speakers considering various factors and investigating how these factors play a role in how listeners perceive.

Second, the listener's familiarity and personal preference towards how the speaker speaks may also affect how the listener perceives the ASR result. Since we found that the people are more likely to be familiar with NS's accent, while not being familiar with NNS's accents (Section 4.2), we assigned listeners who are non-familiar with speaker's accents for NNS, while assigning listeners who are familiar with speaker's accents for NS in our study. However, further study is needed on how the listener's perception differs in other combinations, such as being familiar with NNS's accents.

Lastly, our experiments mainly focused on a single scenario (i.e., a listener watching a video of a speaker asynchronously with auto-generated captions), which may differ from other scenarios using ASR systems. For instance, listeners may perceive differently if they watch a speaker synchronously. In addition, the listener's perception could be influenced by the listeners listening together; the composition of other listeners (e.g., other listeners being the same ethnicity as the speaker) or reactions of other listeners (e.g., applause) could impact their perceptions. Thus, future work could investigate other scenarios for a more generalizable result.

ACKNOWLEDGMENTS

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-01347, Video Interaction Technologies Using Object-Oriented Video Modeling). We also thank the speakers who participated in the audio recording and the study participants from Prolific.

REFERENCES

[1] So-Yeon Ahn. 2019. Decoding "good language teacher"(GLT) identity of native-English speakers in South Korea. *Journal of Language, Identity & Education* 18, 5 (2019), 297–310.

[2] So-Yeon Ahn and Hyun-Sook Kang. 2017. South Korean university students' perceptions of different English varieties and their contribution to the learning of English as a foreign language. *Journal of Multilingual and Multicultural Development* 38, 8 (2017), 712–725.

[3] Minoo Alemi, Parisa Daftarifard, and Roya Pashmforoosh. 2011. The impact of language anxiety and language proficiency on WTC in EFL context. *Cross-Cultural Communication* 7, 3 (2011), 150–166.

[4] Mike Armstrong, Andy Brown, Michael Crabb, Chris J Hughes, Rhianne Jones, and James Sandford. 2016. Understanding the diverse needs of subtitle users in a rapidly evolving media landscape. *SMPTE Motion Imaging Journal* 125, 9 (2016), 33–41.

[5] Verena Bader and Stephan Kaiser. 2019. Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization* 26, 5 (2019), 655–672.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review* (2016), 671–732.

[7] Rusty Barrett, Jennifer Cramer, and Kevin B McGowan. 2022. *English with an accent: Language, ideology, and discrimination in the United States.* Taylor & Francis.

[8] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).

[9] Louisa Buckingham. 2015. Recognising English accents in the community: Omani students' accent preferences and perceptions of nativeness. *Journal of Multilingual and Multicultural Development* 36, 2 (2015), 182–197.

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[11] Changyan Chi, Qinying Liao, Yingxin Pan, Shiwan Zhao, Tara Matthews, Thomas Moran, Michelle X Zhou, David Millen, Ching-Yung Lin, and Ido Guy. 2011. Smarter social collaboration at IBM research. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 159–166.
[12] Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. *arXiv preprint arXiv:2207.11345* (2022).
[13] Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Performance disparities between accents in automatic speech recognition. *arXiv preprint arXiv:2208.01157* (2022).
[14] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46.
[15] Marko Dragojevic, Dana Mastro, Howard Giles, and Alexander Sink. 2016. Silencing nonstandard speakers: A content analysis of accent portrayals on American primetime television. *Language in Society* 45, 1 (2016), 59–85.
[16] Andy Echenique, Naomi Yamashita, Hideaki Kuzuoka, and Ari Hautasaari. 2014. Effects of video and text support on grounding in multilingual multiparty audio conferencing. In *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology*. 73–81.
[17] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784
[18] Elizabeth Elliott and Amy-May Leach. 2022. False impressions? The effect of language proficiency on cues, perceptions, and lie detection. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* (2022).
[19] ETS. 2023. TOEFL iBT Listening Section. https://www.ets.org/toefl/test-takers/ibt/about/content/listening.html
[20] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122* (2021).
[21] Jairo N Fuertes, William H Gottdiener, Helena Martin, Tracey C Gilbert, and Howard Giles. 2012. A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology* 42, 1 (2012), 120–133.
[22] Ge Gao, Naomi Yamashita, Ari MJ Hautasaari, Andy Echenique, and Susan R Fussell. 2014. Effects of public vs. private automated transcripts on multiparty communication between native and non-native English speakers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 843–852.
[23] James D Harnsberger, Rahul Shrivastav, William S Brown Jr, Howard Rothman, and Harry Hollien. 2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice* 22, 1 (2008), 58–69.
[24] Ari Hautasaari and Naomi Yamashita. 2014. Do automated transcripts help non-native speakers catch up on missed conversation in audio conferences?. In *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology*. 65–72.
[25] Helen Ai He, Naomi Yamashita, Ari Hautasaari, Xun Cao, and Elaine M Huang. 2017. Why did they do that? Exploring attribution mismatches between native and non-native speakers using videoconferencing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 297–309.
[26] Braj B. Kachru. 1992. World Englishes: approaches, issues and resources. *Language Teaching* 25, 1 (1992), 1–14. https://doi.org/10.1017/S0261444800006583
[27] Sei Jin Ko, Charles M Judd, and Diederik A Stapel. 2009. Stereotyping based on voice in the presence of individuating information: Vocal femininity affects perceived competence but not warmth. *Personality and Social Psychology Bulletin* 35, 2 (2009), 198–211.
[28] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.
[29] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.
[30] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology* 46, 6 (2010), 1093–1096.
[31] Karen Livescu. 1999. *Analysis and modeling of non-native speech for automatic speech recognition.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[32] Roy C Major, Susan F Fitzmaurice, Ferenc Bunta, and Chandrika Balasubramanian. 2002. The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL quarterly* 36, 2 (2002), 173–190.

[33] Robert M McKenzie. 2008. The role of variety recognition in Japanese university students' attitudes towards English speech varieties. *Journal of Multilingual and Multicultural Development* 29, 2 (2008), 139–153.

[34] Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. 2011. Predicting human perceived accuracy of ASR systems. In *Twelfth Annual Conference of the International Speech Communication Association*.

[35] FMNSKL a AG NINAREH MEHRABI and Fred Morstatter. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv 1908.09635* (2019).

[36] Poppy Noor. 2021. 'I had to change who I am': 'bison' reporter Deion Broxton on his TV accent struggle. https://www.theguardian.com/us-news/2021/apr/02/deion-broxton-bison-montana-journalist-accent

[37] Atsunori Ogawa, Takaaki Hori, and Atsushi Nakamura. 2016. Estimating Speech Recognition Accuracy Based on Error Type Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 12 (2016), 2400–2413. https://doi.org/10.1109/TASLP.2016.2603599

[38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[39] Mei-Hua Pan, Naomi Yamashita, and Hao-Chuan Wang. 2017. Task rebalancing: Improving multilingual communication with native speakers-generated highlights on automated transcripts. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 310–321.

[40] Yingxin Pan, Danning Jiang, Michael Picheny, and Yong Qin. 2009. Effects of real-time transcription on non-native speaker's comprehension in computer-mediated communications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2353–2356.

[41] Yingxin Pan, Danning Jiang, Lin Yao, Michael Picheny, and Yong Qin. 2010. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1725–1734.

[42] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[43] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How Accurate Does It Feel?–Human Perception of Different Types of Classification Mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] ProPublica. 2016. Machine Bias: Risk Assessments in Criminal Sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[46] Nobuhiro Shimogori, Tomoo Ikeda, and Sougo Tsuboi. 2010. Automatically generated captions: will they help non-native speakers communicate in english?. In *Proceedings of the 3rd international conference on Intercultural collaboration*. 79–86.

[47] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Interspeech 2019*. ISCA. https://doi.org/10.21437/interspeech.2019-1427

[48] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[49] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*. 934–938. https://doi.org/10.21437/Interspeech.2017-1746

[50] Kentaro Toyama. 2011. Technology as Amplifier in International Development. In *Proceedings of the 2011 IConference* (Seattle, Washington, USA) *(iConference '11)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/1940761.1940772

[51] Kentaro Toyama. 2015. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs.

[52] Peter Trudgill. 1997. *The social differentiation of English in Norwich*. Springer.

[53] Ron Van Buskirk and Mary LaLomia. 1995. The just noticeable difference of speech recognition accuracy. In *Conference companion on Human factors in computing systems*. 95.

[54] Irina-Elena Veliche and Pascale Fung. 2023. Improving Fairness and Robustness in End-to-End Speech Recognition Through Unsupervised Clustering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[55] Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry* 11, 8 (2019), 1018.

[56] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[57] Yunhan Wu, Martin Porcheron, Philip Doyle, Justin Edwards, Daniel Rough, Orla Cooney, Anna Bleakley, Leigh Clark, and Benjamin Cowan. 2022. Comparing Command Construction in Native and Non-Native Speaker IPA Interaction through Conversation Analysis. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–12.

[58] Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R Doyle, Leigh Clark, and Benjamin R Cowan. 2020. See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In *22nd international conference on human-computer interaction with mobile devices and services*. 1–9.

[59] Lin Yao, Ying-xin Pan, and Dan-ning Jiang. 2011. Effects of automated transcription delay on non-native speakers' comprehension in real-time computer-mediated communication. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*. Springer, 207–214.

[60] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[61] Atif Zafar, Burke Mamlin, Susan Perkins, Anne M Belsito, J Marc Overhage, and Clement J McDonald. 2004. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics* 73, 9-10 (2004), 719–730.

[62] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. 1–8. https://doi.org/10.1109/CIG.2018.8490433

[63] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

Received July 2023; revised October 2023; accepted November 2023