Accessible Human-Error Interactions in AI Applications for the Blind

JONGGI HONG, University of Maryland, USA

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile devices**; **Accessibility technologies**;

Additional Key Words and Phrases: Machine learning, accessibility, speech input, automatic speech recognizer, personalized object recognizer

ACM Reference Format:
Jonggi Hong. 2018. Accessible Human-Error Interactions in AI Applications for the Blind. In *Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2018 International Symposium on Wearable Computers (UbiComp/ISWC'18 Adjunct), October 8–12, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3267305.3267321

1 PROBLEM DOMAIN

People who are blind experience challenges when performing everyday tasks that are heavily dependent on vision such as identifying objects, clothing, and packages of food as well as entering text on a smartphone using touchscreen keyboards. To overcome these challenges, they would typically use other sensory channels such as touch, taste, smell, and hearing. For example, they use a screen reading application such as VoiceOver to identify keys on a touchscreen keyboard or Braille to recognize everyday objects (*e.g.,* by attaching adhesive Braille labels to them). Machine learning (ML) applications such as automatic speech recognition (ASR) and computer vision can make it easier for this population to carry out such tasks by allowing access to the visual world and interactions through preferred modalities. Prior work [2, 34] has shown that indeed speech input is the preferred method for text input on a mobile device for blind people. Also, many computer vision application have been proposed to enable these users navigate in unfamiliar indoor environments [3, 30], access printed text [5, 25], and identify objects of interest [1, 20, 27] using the built-in cameras on their mobile devices.

Although advances in ML promise improved accuracies, ML applications are inherently error prone. The accuracy of ASR systems is reaching 5.1% word error rate for English [33]. In computer vision, the top-5 performance of image classifiers has only 3.7% error rate [26]. Given that these numbers are averages reported on benchmarking datasets, users may face additional challenges when these models are deployed in real-world application especially in conditions deviating from those that the models were trained on. For example, ASR errors can occur due to speaker variation, disfluency, background noise, ambiguity of words, and mistakes from users [14, 18]. Also, object recognition errors, even though models are typically trained on limited numbers of objects, could occur due to lack of discriminative characteristics, lighting conditions, and reflective surfaces, but more so for blind users due to their challenges in photo taking. These could result in different background clutter, scale, viewpoints, occlusion, and image quality than in photos taken by sighted users used in training [20, 35].

Author's address: Jonggi Hong, University of Maryland, 2117 Hornbake Bldg, South Wing, College Park, MD, 20742, USA, jhong12@umd.edu.

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

XXXX-XXXX/2018/8-ART $15.00
https://doi.org/10.1145/3267305.3267321

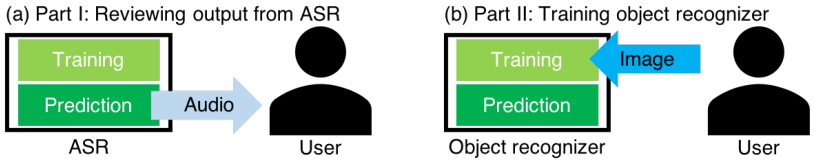

Fig. 1. Two parts of my research on human-error interactions in AI application for the blind.

Even when enabling the end-user to provide the training examples in order to personalize the recognition task (*e.g.,* teachable object recognizers for the blind [20, 23]), errors could result due to the challenges that these users face in providing high-quality or well-framed images as training examples.

While both ASR and object recognition systems can benefit blind people, the **need for visual inspection of their errors** has hold back their usability for this population. For example, reviewing and editing text to correct ASR errors has been known as a bottleneck in text entry through speech not only for sighted users [15, 21, 28] but more so for blind users [2]. Typically, they have to perform these steps non-visually, though voice-only interactions. User interfaces and multimodal interactions that address these challenges (*i.e.,* identifying, avoiding, and handling errors) would be of value for accessibility researchers designing novel assistive technologies that employ ML and AI models.

Driven by the above limitations in the field, the focus of my research is to improve human-error interactions for blind people when using ML applications. My work explores user interfaces across two core AI areas: speech and vision. First, it investigates the challenge of identifying and reviewing ASR errors through audio-only interactions as shown in Figure 1a. Second, it examines methods to help blind users provide good training examples for teachable object recognizers as shown in Figure 1b.

While my research on accessible human-error interaction is focused on speech and image motivated by a specific user group, the methods and techniques in this work should be applicable to other domains within and beyond accessibility where errors in intelligent systems are either introduced by the user through teachable interfaces (*e.g.*, teachable machines for accessibility [19]) or can be visually inspected only (*e.g.*, indoor localization, obstacle detection, and facial expression recognition). Also, to allow for more generalizable findings, my research questions are explored with sighted participants as well and compared to those obtained from blind participants. Other contributions of this work include: (1) an experimental framework for measuring the ability of identifying ASR errors through audio-only interactions, (2) evaluating the effect of different manipulations of synthesized speech on ASR error identification, (3) collection of empirical data on how users with little knowledge on ML interact with teachable machines such as personal object recognizer, and (4) developing interactive guidance to help blind users take good training photos.

2 RELATED WORK

With a focus on error-handling methods, this section discusses related work on ML applications where blind users interact with two different types of data: speech and images.

2.1 Reviewing and Editing Text from Speech Input

Reviewing and editing has been known as a bottleneck in the text entry process using speech input for both sighted and blind users. The correction of ASR errors were over twice as long as the entry time when sighted users entered text with speech input on a desktop computer [21]. Experienced users were able to achieve faster text entry speed by correcting ASR errors faster than novice users when speech input was used by blind users on a desktop [15]. While novice users re-dictated to correct errors and wasted time due to the repeated misrecognition

of redictation, experienced users achieved faster text entry rate by switching modality of entering text from speech to keyboard and mouse. Blind users also spent 80% of text entry time editing text from speech input [2]. Azenot's study showed that identifying ASR errors is a challenging task for blind users because ASR errors sound similar to users' original speech input. However, the ability of identifying ASR errors has not been evaluated quantitatively.

Correcting ASR errors requires first recognizing errors in inputted text, then editing. While my research explores the former, prior studies on editing text from speech input mostly focused on correcting ASR errors with the assumption that users can review the text visually. Many studies combined speech input with other modalities such as pen gestures, touchscreen, or spelling out individual letters [11–13]. Predicting alternative word candidates for correcting ASR errors has been used with keyboard and touchscreen input to allow users to edit words easy and fast [17, 22, 24, 32]. However, these studies depend on the visual feedback which is not accessible to blind users.

2.2 Teachable Object Recognizers

Prior work in object recognizers for blind users usually spans across computer vision systems, crowdsourced platforms, or a hybrid combination. For example, VizWiz [8], BeMyEyes [6], and BeSpecular [7] are crowd-powered object recognition systems where users ask questions with image or real-time videos depicting unknown objects to the user and crowd-workers provide descriptions. Though such crowd-powered applications are a feasible approach, they raise privacy issues since blind individuals have to share images and videos often from their personal environment with unknown crowds, which may not be available 24/7. Generic object recognizer, on the other hand, can enable these users to recognize objects without sighted help [1, 27] thus ensuring privacy. However, object recognizers trained on large available datasets tend to be limited on the number of classes that they can support. Therefore, they have limited application especially for distinguishing fine-grained instances for everyday objects of interest across all blind people.

Teachable object recognizers [20], can allow blind users to personalize the capabilities of an object recognizer by providing a small number of examples for each object of interest in their environment. An early example of this is LookTel [23]; though it was limited in the sense that sighted help was required to get a single high quality training example. As a follow up, Kacorri *et al.* explored whether it was feasible for the blind users do the training by themselves instead. One of the major challenges identified in this prior work was the effect that bad object framing may have in performance degradation since blind participants could not visually inspect their training examples. My research focuses on improving such teachable interfaces through accessible interactions for obtaining better training examples that can reduce errors in object recognizers.

3 METHODOLOGICAL APPROACH

As summarized in Figure 1, the goal of my research is to improve the experience of blind users with AI applications by exploring effective human-error interactions in two contexts: reviewing ASR errors in text entry through speech (Part I) and providing good training example in teachable object recognizers (Part II). My work employs a mixed-methods approach. In Part I, the experience of identifying ASR errors was investigated through semi-structured interviews and custom experimental testbeds designed to quantify error recognition accuracies and strategies through both a crowdsourcing platform and a lab setup. In Part II, interaction patterns of non-experts' perception of machine teaching will be explored through a web-application that support remote training of teachable object recognizers. Both training data and user feedback obtained from crowdsourced participants will be employed in the design of novel sound interactions guiding users in obtaining better training examples. Our interactions will be evaluated in the context of a mobile app, that allows blind users train their personal object recognizer in a real-world setting.

Specifically, Part I explores challenges in identifying ASR errors through audio-only interactions through a series of five studies. Since experiences and listening rates for synthesized speech differ across sighted and blind participants due to experience with a screen reader [9], we decouple the ability to identify ASR errors in this context based in experience with a screen reader. Therefore, the first four of these studies involve sighted participants recruited through Amazon mTurk [31] not using a screen reader [16]. Findings and insights from these initial studies are used to guide the experimental design for a follow-up in depth study[1] that investigates experiences and responses across sighted and blind participants in a lab setup. In both cases, error-identification accuracies are measured with carefully engineered stimuli under a number of controlled conditions.

Part II, which is work in progress, focuses on investigating effective interactions for detecting and reducing errors in the context of teachable object recognizers. In this case, users are not simply the consumers of model predictions but can have a more active role by tweaking the behavior of the model with their training examples. Thus new errors can be introduced that are related to their training examples. We decouple these errors based on a) inexperience in machine teaching and b) challenges in photo-taking. We explore the first case with a large-scale study [2] with sighted participants recruited through Amazon mTurk, where we measure accuracies of the user-trained models and qualitatively identify training strategies and their relationship with model performance. The findings and insights from this study will inform the design of interactions for a teachable object recognizer mobile app that will be later evaluated across sighted and blind participants.

4 COMPLETED WORK: REVIEWING ASR ERRORS IN TEXT ENTRY THROUGH SPEECH (PART I)

Insights from user studies with sighted participants in the lab and a crowdsourced platform inform the design of a lab-based user study exploring responses across sighted and blind participants.

4.1 Characterizing Challenges in ASR Error Identification

Though the challenge of identifying ASR errors through audio-only interactions was observed in Azenkot's study [2], the accuracy of ASR error identification was not evaluated in quantitative way. The accuracy measured in this work can serve as a baseline performance for future research on improving the interface for presenting ASR result to users. Different methods of manipulating synthesized speech of ASR result were shown as possible methods to help users identify ASR errors more accurately.

As a first step of Part I, a series of four studies is conducted to characterize this challenge by measuring the accuracy of identifying ASR errors when sighted participants review text through audio-only interactions [16]. The first controlled-lab study is conducted to identify the type of ASR errors that users were likely to miss and measure the baseline accuracy of identifying ASR errors. Participants were asked to enter reference phrases displayed on a screen of a tablet device using speech input and identify ASR errors with only audio of the ASR result. ASR errors are found to be missed most frequently when multiple sequential words sounded like another word or words. The results also show that participants are able to identify less than 50% of ASR errors with synthesized speech rate 200 words per minute (WPM). The next three studies are conducted with crowdsourced participants to evaluate the effect of three different manipulations of synthesized speech (*i.e.*, adjusting speech rate, inserting pauses between words, and repeating the synthesized speech twice) on the accuracy of error identification. In these three studies, we isolated the speech input and revision process by using the ASR result from the first controlled-lab study. Participants were asked to identify ASR errors, given the original phrases on a visual display and the ASR result through audio. The findings showed that slower speech rate (100-200 WPM) and longer pauses (150ms) between words resulted in higher accuracy of identifying ASR errors, but repeating synthesized speech twice does not improve the accuracy. Moreover, slower speech rates provided more time

[1] In preparation for submission at IMWUT.
[2] This is an ongoing study.

for users to identify ASR errors. Pauses made words in audio sound more clear because they removed elision, dropping syllables in words during speech (*e.g.* "I don't know" sounds like "I duno").

4.2 Going Deeper into ASR Error Identification across Sighted and Blind Users

My research further investigates the experience of using speech input and the accuracy of identifying ASR errors with blind participants. A user study with semi-structured interview and tasks of entering text using speech input is conducted with sighted and blind participants. The results of semi-structured interview show both similarities and differences in experience with speech input between these two user groups. For example, blind participants used speech input more frequently and were more concerned about ASR errors than sighted participants. Though participants in both groups reported during the interview that identifying ASR errors is not a challenging task, results show that they only identified about 40% of the errors with the default speech rate of 175 WPM. No significant difference was found across the two groups, indicating that the ability of identifying ASR errors in audio-only interactions might not improve significantly with experience with speech synthesis. Given that identifying ASR errors is still challenging for blind users who had more experience with synthesized speech than sighted users, further research on method of presenting ASR results to users needs to be conducted. A potential solution could be predicting the words that are likely to be missed by users using machine learning so that such words are highlighted to allow users to pay more attention to them.

5 PROPOSED WORK: ERROR-INTERACTION IN TEACHABLE OBJECT RECOGNIZERS (PART II)

Part II addresses challenges in reducing recognition errors when training object recognizer through two studies.

5.1 Characterizing Challenges in Machine Teaching by Non-Experts

Object recognizers trained on images of objects taken by users who have little or no knowledge of ML may have performance degradation due to the misperception of what it means to provide training examples. In this first real world study, crowdsourced sighted participants are asked to take images of objects for training remotely a personalized object recognizer. Crowdsourcing is used to collect a dataset of object images that are used to train and test in real time a convolutional neural network. Images will be analyzed qualitatively and quantitatively to uncover user strategies for machine teaching with little examples. For example, qualitative measures will include the image quality, variation in viewpoints and discriminative visual characteristics included in the multiple training examples, as well as changes in training strategy after testing. On the other hand quantitative measures will be based on recognition accuracies across objects, varying number of examples, and users. The data, insights, and findings from this study will be used to guide the design and implementation of novel interactions to support better machine teaching strategies that will be evaluated in the next study. Currently, I'm working on implementing a web application that will allow the crowd to train and test a personal object recognizer in their real world environment. The models for the personalized object recognizers will be trained remotely on our servers given users training images by employing transfer learning on Google's Inception-v3 model [29].

5.2 Exploring Interactions for Machine Teaching across Sighted and Blind Users

Prior work in teachable object recognizers [20] showed that the absence of object in an image was one of the main reasons of performance degradation. It indicates that beyond potential misperceptions in machine teaching, performance degradation could be caused from the difficulty of framing the object in the camera scope during training. In the next study, different approaches in automatically inferring and communicating potential errors to guide the users during photo-taking will be evaluated. The teachable object recognizer will be implemented as a mobile app with the training and testing running on a server with GPUs to support real time feedback to the user. Guidance on good versus bad training examples will be communicated through novel interactions informed by

our previous study. Both the application and the interactions will be evaluated through a user study conducted with two user groups: blind and sighted participants serving as control. In a between subject design we will explore the effect of automatic guidance in taking good training photos beyond just the built-in instructions on how to train the object recognizer. Interactive feedback will be evaluated subjectively through an interview after task performance and quantitatively based on the performance of object recognizers. While prior work explored teachable object recognizers in a Wizard of Oz study [20], this study requires a working application. I have developed a working prototype of the teachable object recognizer as an iOS app, where the training happens in a remote server using transfer learning [4]. Specifically, the last layers of Google's Inception v3 [29], pre-trained on ImageNet [10], are fine-tuned with classes and training examples provided by the user.

6 CONTRIBUTION

My research focus is accessible human-error interactions in AI applications for the blind; specifically in the context of text input through ASR and object recognition through computer vision. While both ASR and object recognition systems can benefit blind people, the need for visual inspection of their errors has hold back their usability for this population. In this context, my work focuses on a) quantifying the challenge of identifying and reviewing ASR errors through audio-only interactions for both sighted and blind users, and b) exploring accessible interactions that help non-experts and blind users provide good training examples for teachable object recognizers. I believe that the methods and techniques employed in this work are applicable to other domains within and beyond accessibility where errors from intelligent systems can be impacted by the user interaction with teachable interfaces or they be visually inspected only. Thus, to allow for more generalizable findings, my research questions are explored both with sighted and blind participants.

7 ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under award number 1816380. I am grateful for support and resources provided by Professor Leah Findlater at University of Washington and by Professor Hernisa Kacorri at University of Maryland as well as for the collaborations with Christine Vaing, Kyungjun Lee, June Xu, and Jaina Gandhi.

REFERENCES

[1] Aipoly. 2016. Vision through artificial intelligence. http://aipoly.com/

[2] Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 11.

[3] Jinqiang Bai, Dijun Liu, Guobin Su, and Zhongliang Fu. 2017. A Cloud and Vision-based Navigation System Used for Blind People. In *Proceedings of the 2017 International Conference on Artificial Intelligence, Automation and Control Technologies (AIACT '17)*. ACM, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3080845.3080867

[4] Jonathan Baxter. 1998. Theoretical models of learning to learn. In *Learning to learn*. Springer, 71–94.

[5] Pooja Belunkhi and Ravindra Patil. 2018. A Portable Camera-Based Assistive Text Reader for Blind Persons. *Framework* 5, 04 (2018).

[6] BeMyEyes. 2016. Lend you eyes to the blind. http://www.bemyeyes.org/

[7] BeMyEyes. 2016. Let blind people see through your eyes. https://www.bespecular.com/

[8] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.

[9] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A Large Inclusive Study of Human Listening Rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 444.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.

[11] W Feng. 1994. Using handwriting and gesture recognition to correct speech recognition errors. *Urbana* 51 (1994), 61801.

[12] Arnout RH Fischer, Kathleen J Price, and Andrew Sears. 2005. Speech-based text entry for mobile handheld devices: an analysis of efficacy and error correction techniques for server-based solutions. *International Journal of Human-Computer Interaction* 19, 3 (2005),

279–304.
[13] Kazuki Fujiwara. 2016. Error correction of speech recognition by custom phonetic alphabet input for ultra-small devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 104–109.
[14] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2008. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates. *Proceedings of ACL-08: HLT* (2008), 380–388.
[15] Christine A Halverson, Daniel B Horn, Clare-Marie Karat, and John Karat. 1999. The Beauty of Errors: Patterns of Error Correction in Desktop Speech Systems.. In *INTERACT*. 133–140.
[16] Jonggi Hong and Leah Findlater. 2018. Identifying Speech Input Errors Through Audio-Only Interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 567.
[17] David Huggins-Daines and Alexander I Rudnicky. 2008. Interactive asr error correction for touchscreen devices. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*. Association for Computational Linguistics, 17–19.
[18] Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech communication* 45, 4 (2005), 455–470.
[19] Hernisa Kacorri. 2017. Teachable machines for accessibility. *ACM SIGACCESS Accessibility and Computing* 119 (2017), 10–18.
[20] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. 2017. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5839–5849.
[21] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 568–575.
[22] Yuan Liang, Koji Iwano, and Koichi Shinoda. 2014. Simple gesture-based error correction interface for smartphone speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*.
[23] LookTel. 2016. Instantly recognize everyday objects. http://www.looktel.com/recognizer
[24] Jun Ogata and Masataka Goto. 2005. Speech repair: quick error correction just by using selection operation for speech input interfaces. In *Ninth European Conference on Speech Communication and Technology*.
[25] K-NFB Reader. 2014. Easy access to print and files, anytime, anywhere. https://knfbreader.com/
[26] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. 2018. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548* (2018).
[27] SeeingAI. 2017. An app for visually impaired people that narrates the world around you. https://www.microsoft.com/en-us/seeing-ai
[28] Ben Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65.
[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
[30] YingLi Tian, Xiaodong Yang, Chucai Yi, and Aries Arditi. 2013. Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments. *Machine vision and applications* 24, 3 (2013), 521–535.
[31] Amazon Mechanical Turk. 2018. Human intelligence through an API. https://www.mturk.com/
[32] Lijuan Wang, Tao Hu, Peng Liu, and Frank K Soong. 2008. Efficient handwriting correction of speech recognition errors with template constrained posterior (TCP). In *Ninth Annual Conference of the International Speech Communication Association*.
[33] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. The Microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 5255–5259.
[34] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014. Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3123–3132.
[35] Yu Zhong, Pierre J Garrigues, and Jeffrey P Bigham. 2013. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 20.