A user study to compare two conversational assistants designed for people with hearing impairments

Anja Virkkunen
Aalto University, Finland
anja.virkkunen@aalto.fi

Juri Lukkarila
Aalto University, Finland
juri.lukkarila@aalto.fi

Kalle Palomäki
Aalto University, Finland
kalle.palomaki@aalto.fi

Mikko Kurimo
Aalto University, Finland
mikko.kurimo@aalto.fi

Abstract

Participating in conversations can be difficult for people with hearing loss, especially in acoustically challenging environments. We studied the preferences the hearing impaired have for a personal conversation assistant based on automatic speech recognition (ASR) technology. We created two prototypes which were evaluated by hearing impaired test users. This paper qualitatively compares the two based on the feedback obtained from the tests. The first prototype was a proof-of-concept system running real-time ASR on a laptop. The second prototype was developed for a mobile device with the recognizer running on a separate server. In the mobile device, augmented reality (AR) was used to help the hearing impaired observe gestures and lip movements of the speaker simultaneously with the transcriptions. Several testers found the systems useful enough to use in their daily lives, with majority preferring the mobile AR version. The biggest concern of the testers was the accuracy of the transcriptions and the lack of speaker identification.

1 Introduction

Hearing loss can make the participation in normal conversations an exhausting task, because people with hearing impairments need to focus more on the conversation to be able to keep up (Arlinger, 2003). This can cause the deaf and hard of hearing to withdraw from social interactions, leading to isolation and poorer well-being (Arlinger, 2003). Having access to an automatic speech recognizer (ASR) designed to answer their needs could make participation in everyday conversations considerably easier for them.

People with hearing impairment are a heterogeneous group with significant variability in the degree of hearing loss and its causes. Hearing aids and implants can restore hearing to a degree, but they struggle in noisy environments (Goehring et al., 2016). Many people with hearing impairments also refuse to use aids because they are perceived as uncomfortable or costly (Gates and Mills, 2005). Professional human interpreters can help the deaf and hard of hearing with their near real-time transcription, but they require advance booking and are also costly (Lasecki et al., 2017).

ASR has the potential to both function as a support and a replacement to other solutions. The strengths of ASR include accessibility with little cost, nearly real-time transcription and independence of costly human labour. Furthermore, it can be helpful to anyone irrespective of their degree of hearing loss. The weaknesses of ASR are in robustness and the lack of support for speaker diarization. And even though the accuracy of ASR has improved to a level where it rivals human transcribers (Xiong et al., 2017), noisy environments, accented speakers, and far-field microphones remain a challenge (Yu and Deng, 2015). Additionally, recognizing and conveying paralinguistic features like tone, pitch and gestures is difficult for automatic systems.

The objective of our work is to study the preferences of the deaf and hard of hearing when using ASR-based conversation assistants. We constructed two pilot Finnish language ASR systems for two portable devices with different display options. The first system is a standalone laptop ASR that does not utilize network connection or video camera. The second system is a mobile device wirelessly connected to an ASR server. In the mobile device the ASR transcript is shown in augmented video stream next to the head of the speaker. The purpose of this augmented reality

Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies, pages 1–8
Minneapolis, Minnesota, USA

(AR) view is to reduce visual dispersion, which in this case refers to the need of the user to switch attention between multiple visuals. These two setups were tested by deaf and hard of hearing users, who were then interviewed to find out their preferences. We review the results and compare the feedback the systems received.

1.1 Previous work

Automatic speech recognition research focusing specifically on helping people with hearing impairments has gone on at least since 1996 (Robison and Jensema, 1996). The first assistive ASR system for the Finnish deaf and hard of hearing was devised in 1997 (Karjalainen et al., 1997). Since then, helping the deaf and hard of hearing in their school environment has been a concern in many assistive ASR systems. A lot of this research focuses in providing real-time ASR-generated transcriptions of lectures (Wald, 2006; Kheir and Way, 2007; Ranchal et al., 2013). Major effort is also dedicated to improving the ASR aided learning experience in other ways, such as minimizing visual dispersion (Cavender et al., 2009; Kushalnagar and Kushalnagar, 2014; Kushalnagar et al., 2010), comparing captioning and transcribing of online video lectures (Kushalnagar et al., 2013), and using human editors to correct ASR output (Wald, 2006). Our work focuses more generally on helping people with hearing impairments in conversational situations, not just the school setting.

Other notable applications include the system of Matthews et al. (2006), where mobile phones were used for delivering human-made transcriptions via text messages. They showed transcriptions could help people with hearing impairments, but lacked the ASR component. The transcription table design from Van Gelder et al. (2005) provided all meeting participants with partial text support. The aim there was to minimize the stigma on the deaf or hard of hearing participant.

The idea to use AR has also been introduced before in the work of Mirzaei et al. (2012, 2014) and Suemitsu et al. (2015). The system of Mirzaei et al. is similar to our mobile AR system, but it is developed for ultra mobile personal computers and has a text-to-speech component. In the work of Suemitsu et al. the focus is on reducing effect of noise with directional microphones and beamforming. As a consequence, their system works well only if the speaker is directly in front of the user. Moreover, both of these works lack the user perspective because the focus is more on the system design.

2 Conversation Assistant

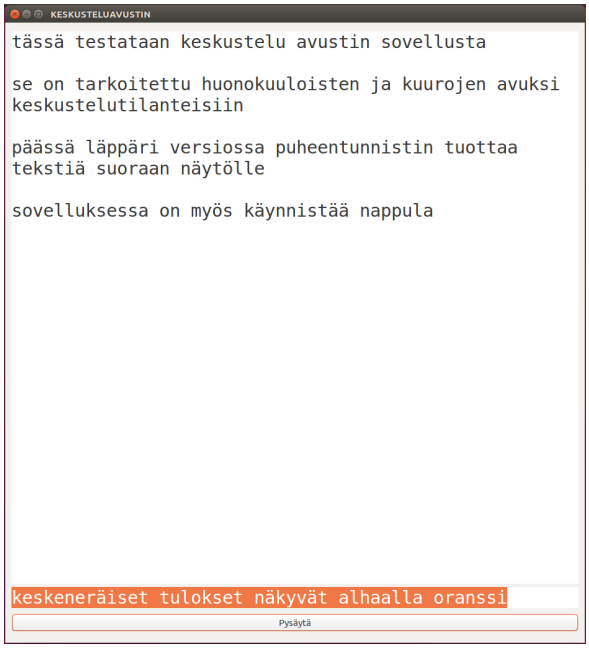
Our aim in building the two Conversation Assistant systems was to provide automatic transcriptions to people with hearing impairments in a useful format. A useful conversation assistant system can (1) recognize large-vocabulary continuous speech in real-time, (2) manage varying acoustic environments with noise, and (3) present the transcriptions in a clear manner. Achieving the first two requirements is possible with modern speech recognition systems, however, their computing power and memory consumption pose limitations on the system design. The minimal solution to the third requirement would be to just display the recognition results on a screen, but looking at the screen would cause the user to miss nonverbal communication, like gestures.

We built two prototypes of the Conversation Assistant system: one running on a laptop and one on a mobile device with augmented reality (AR) capabilities. In both systems, Kaldi Speech Recognition Toolkit (Povey et al., 2011) was used to build the speech recognition models. In addition, we used Gst-Kaldi, a GStreamer plugin, for handling the incoming audio. The source codes for both the laptop version[1] and the mobile AR version[2] are published on Github.

2.1 Laptop

The laptop version was built for the first round of user tests, to find out the preferences of the deaf and hard of hearing for the Conversation Assistant concept in general. We therefore built a decidedly simple system. It ran a Kaldi speech recognition model locally on a laptop. The acoustic model of the automatic speech recognizer was a feed-forward deep neural network, trained on multi-condition data to increase noise robustness. The robustness to noise is important, because conversations are rarely had in silent environment. The data for training the acoustic model came from the SPEECON corpus (Iskra et al., 2002). The language model was trained using the *Kielipankki*[3]

[1] https://github.com/Esgrove/mastersthesis
[2] https://github.com/aalto-speech/conversation-assistant
[3] The Language Bank of Finland



(a) Laptop version. The speech recognition results are at the top of the window. The speech being recognized currently is displayed at the bottom with orange background.

(b) Mobile AR version. The speech recognition results are placed in speech bubbles next to the face of the speaker.

Figure 1: User interface screenshots from the software.

corpus and the lexicon was based on morphs (Virpioja et al., 2013) instead of words, because of the morphological complexity of Finnish. The speech recognition model had a word error rate (WER) of 29.7 % when tested on (low-noise) broadcast news data from the Finnish public broadcaster YLE (Lukkarila, 2017). The user interface (UI), shown in Figure 1a, was kept minimal with only a record button and space for the transcription results. A more detailed description of the system is given in (Lukkarila, 2017).

2.2 Mobile AR

The main objective of the mobile AR Conversation Assistant was to move the laptop-based system to a mobile device platform and use AR to bring the transcriptions and a visual of the speaker close to each other. With the latter we aimed to reduce the amount of non-verbal communication the user loses when switching between the speaker and the transcriptions. The reduced computational capabilities of mobile devices necessitated splitting the system into a separate mobile AR application and a speech recognition server.

The mobile application was developed for the iOS platform using a 10.5" iPad Pro tablet from Apple as a test device. In the application, AR camera is used to provide a view of the speaker. The face of the speaker is located locally on the device with a face recognition algorithm provided as part of iOS toolkits. When speech is recorded by the device, the application sends the audio to the server for recognition. The text transcriptions returned from the server are then placed in speech bubbles next to the detected face as shown in Figure 1b. The bubbles follow the face of the speaker, if he or she moves in the screen.

The server side uses an open-source Kaldi Server implementation (Alumäe, 2014) based on Gst-Kaldi. The server is split to a controlling master server unit and workers responsible for the recognition process. Each mobile device connecting to the server needs one worker to do the transcribing and connections between the two are handled by the master server. The master server and the workers can be run on different machines and all the communication happens over the internet. This requires stable connections from all parties, including the mobile client, but on the upside, the system can be scaled up as much as needed.

In the speech recognizer, the acoustic model was a combination of time-delay and long short-term memory neural network trained with Finnish Parliament Speech Corpus (Mansikkaniemi et al., 2017). This training data is significantly larger than the one used for the laptop system, but in-

cludes little background noise. The language model was trained with the same data as in the laptop system, but utilizing a more recent subword modeling optimized for Kaldi's weighted finite state transducer architecture (Smit et al., 2017). The speech recognition model scored WER of 18.56 % on the YLE data which is more than a 10 % absolute improvement over the model in the laptop system (Mansikkaniemi et al., 2017). Even though these evaluations were not performed for noisy tasks, we expect that the improvement is substantial enough to cover the lack of noise-robustness for our user tests. For a more detailed description of the system, see the work in (Virkkunen, 2018).

3 User tests

To evaluate the prototypes, we organized user tests for both versions with deaf and hard of hearing participants. Our aim was to simulate noisy conversational situations to see how much the Conversation Assistant could help the tester in following the conversation. In the user tests of the laptop version, the participant had a conversation with and without the support of the Conversation Assistant. In the mobile AR tests, the comparison was made between a text-only view similar to the laptop version and the AR view seen in figure 1b. The contents of this section are further detailed in (Lukkarila, 2017) and (Virkkunen, 2018).

The test users were recruited with the help of *Kuuloliitto*, the association of the deaf and hard of hearing in Finland. The number of participants for the laptop and mobile AR versions were nine and twelve, respectively. The sample size was limited by the number of volunteers we were able to find. Each participant was also asked to give a written consent and permission to record the test session. Six people took part in both tests so in total there were 15 unique participants. The age of the testers, excluding two who refused to disclose their age, ranged from 15 to 84, with the median age being 55. All except two participants were women. Two of the participants were deaf, but they could communicate verbally. The rest had different degrees of hearing loss and used either hearing aids, cochlear implants or both in their daily lives. Four participants also had used ASR applications before, for example personal voice assistants, automatic video captioning, the Google Translate service and note takers.

3.1 Test setup

The test setup simulated a conversation of two people in a noisy environment. The test administrator and the participant would sit face-to-face at a table surrounded by loudspeakers playing a looped noise recording from a busy cafe. The participant had the laptop/mobile device in front of them and they could freely alternate between following the application and their conversation partner. In the case of mobile AR, the participant could choose to hold the device in their hands or place it on a tripod. The test was designed to last for one hour and the feedback was collected using a questionnaire. The overall structure and content of the questionnaires is the same between the two user tests, but small changes were made to the mobile AR version to reflect the changes in the system.

The test and the questionnaire had four sections: introduction, word explaining, conversation, and debriefing. In the introduction the test participant was familiarized with the test plan and the Conversation Assistant. The participant was also asked to fill in their background information in the questionnaire. In the debriefing section, the participant was asked to give overall feedback on the system.

The first task, word explaining, consisted of the test administrator explaining words from a list to the participant, who tried to guess the word in question. Halfway through the task, the participant was asked to switch between the compared methods (without versus with Conversation Assistant or text view versus AR view). After finishing the word list or running out of time, the participant gave feedback in the questionnaire. In the second task, the conversation, the test administrator conversed with the test participant on a range of common topics from hobbies to food and travel for 10-15 minutes. Switch between the compared methods was made at the midpoint again. And after a time limit set for the task was reached, the participant was asked to give feedback on the questionnaire.

3.2 Results

The questionnaire had questions with both written and numeric format. Question with written feedback were concerned with the potential use cases of the system and its strengths and weaknesses. The numeric questions assessed the perceived quality of the system and the preferences of the testers. The numeric questions further break

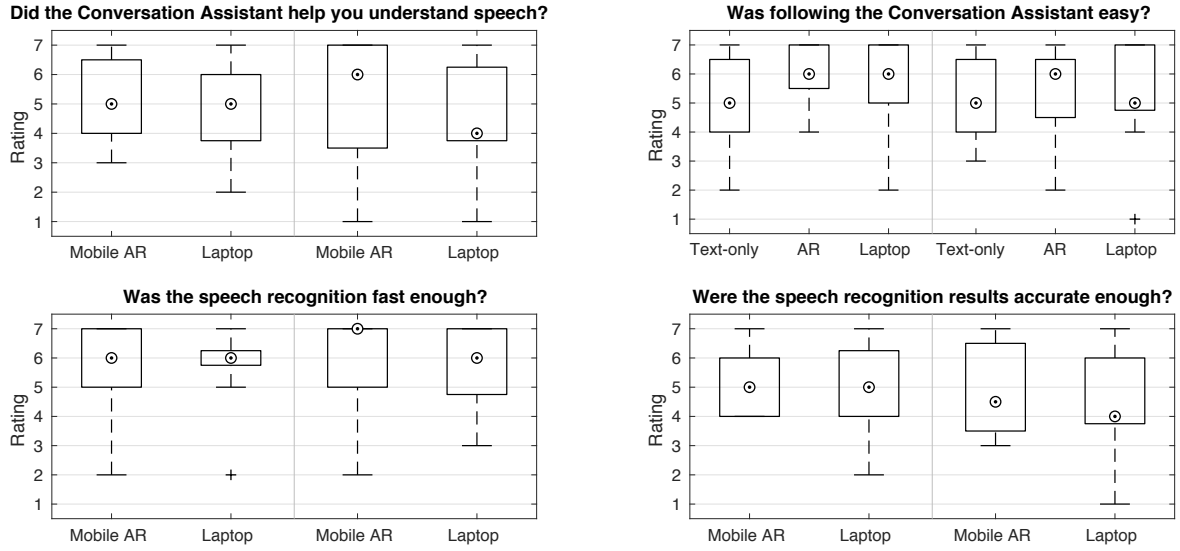


Figure 2: Rating results from task 1, word explanation (left side of each plot) and task 2, conversation (right side of each plot). The range from one to seven is the same as in Figure 3.

down into binary choices, one multiple-choice question and rating questions. The rating questions had a range from one (negative) to seven (positive). Questions with numeric format also had text fields where the testers could elaborate their choices if they wanted.

Several numeric results (Table 1, Figure 3 and top-left of Figure 2) show that the participants found both Conversation Assistant systems useful. Majority of the participants would adopt a Conversation Assistant system in their daily lives. Potential use cases and environments mentioned included daily conversations, meetings, museums, restaurants, live television, lectures and office. Speed and ease of following the output in both systems were also rated favorably with few exceptions in Figure 2. Those who disagreed said that the following suffered mostly from recognition errors. Another point raised was the movement and positioning of the speech bubbles in the mobile AR version which felt distracting to some.

Speech recognition errors were the biggest problem reducing the perceived utility of the sys-

Would you use an application like the Conversation Assistant in your daily life?	
Yes: 83 %	No: 17 %
Which mode would you prefer?	
With AR view: 67 %	Text-only view: 33 %
Which one of the following options is better?	
Text appears faster (<1 sec), but contains more mistakes: 67 %	Text appears slower (>1 sec), but contains less mistakes: 33 %

Table 1: Results to the binary questions asked in the user tests of the mobile AR version.

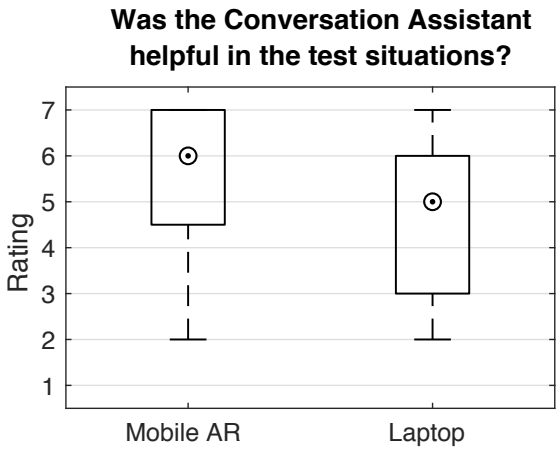


Figure 3: Rating of the overall usefulness of the system. The range is from one (negative) to seven (positive). In the box plots, the circle with the dot marks the median. The bottom and top of the box correspond to 25th and 75th percentiles, respectively. The sample is skewed if the median is not at the middle of the box. The whiskers show the extreme data points that are not considered outliers. The data points outside the whiskers are marked with plus signs.

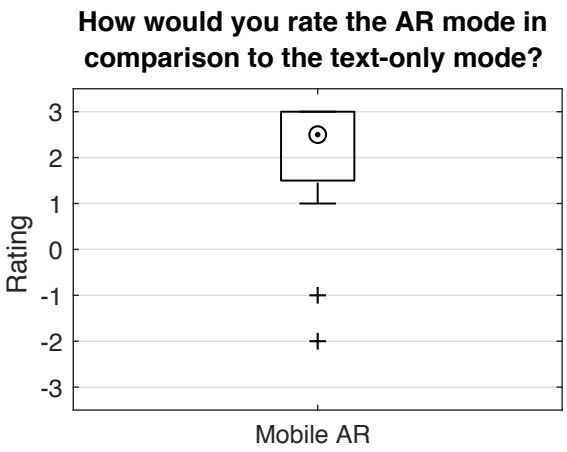


Figure 4: In the user tests of the mobile AR version, the testers were asked to rate whether they preferred having AR view over the text-only view. The range is from minus three (much worse) to three (much better).

tem. The ratings are similar for the two systems, but the AR mobile has less negative ratings overall. In Figure 2 it can be seen that ratings of the accuracy reflect the ratings of the helpfulness. Many felt the transcriptions helped them get an idea of the conversation, but that they could not solely rely on the transcriptions because of the errors. Moreover, a couple of participants noted they used the Conversation Assistant only a little because they could use lip reading instead. They would need transcriptions only in group conversations or in cases where the face of the speaker cannot be seen.

Figure 4 shows that the user testers preferred the AR view over having text-only view similar to the laptop version. Majority of the answers cited the ability to use lip reading in AR view as the decisive factor. Two participants thought the views would have different use cases, for example the text-only view could be useful in meetings. One found the text-only view cleaner and easier to follow than the AR view. Some testers also worried that pointing the device camera at the conversation partner in the mobile AR version would feel inappropriate.

Figure 5 shows the participants preferences for end-user devices. It is clear from the answers that the device needs to be mobile to be of any use. Most people would like to have the application on their smartphones, as it is a device most people own and carry around everywhere. Tablets and laptops also get many votes, especially from working age people. Smart glasses got votes from three curious participants, though we anticipated more. We hypothesize the image most people have of

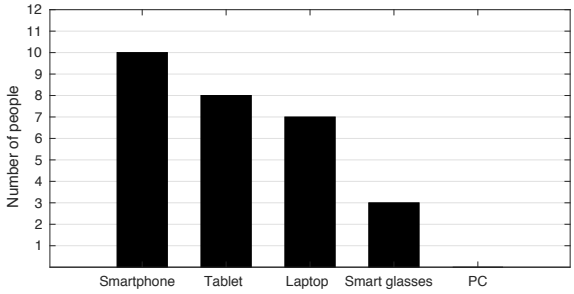


Figure 5: Which device the participants would like to use the Conversation Assistant on.

AR glasses is that they are bulky, impractical and stand out. This could explain the lack of appeal the glasses had among the participants.

4 Conclusions

We evaluated two prototypes of a conversational assistant for the deaf and hard of hearing by user tests. The results show that it is already possible to build an assistive application the deaf and hard of hearing find useful with current technology. In the written feedback many expressed the urgent need for this type of application. Several people noted they would download the mobile application if it were available, despite its flaws. Majority of the test users preferred the mobile AR version because it supported their use of lip reading. A couple of participants saw potential use for both versions depending on the situation.

Accuracy of the transcriptions was the biggest issue in need of improvement according to the participants. Several testers also noted that group conversations in noise are the most difficult to follow. For them, speaker diarization would be the feature that would make the Conversation Assistant truly useful. Both systems also lack direct unmediated eye contact which could be potentially solved with AR glasses. However, to be widely adopted, the glasses would have to be unobtrusive and light weight.

Acknowledgments

This work was funded by the Academy of Finland as part of the project ”Conversation assistant for the hearing impaired” (305503) and Kone Foundation. The writers would also like to thank Katri Leino, Tanel Alumäe and Kuuloliitto for help and resources.

References

Tanel Alumäe. 2014. Full-duplex speech-to-text system for estonian. In *Baltic HLT 2014*, pages 3–10, Kaunas, Lithuania.

Stig Arlinger. 2003. Negative consequences of uncorrected hearing loss-a review. *International journal of audiology*, 42:2S17–2S20.

Anna C Cavender, Jeffrey P Bigham, and Richard E Ladner. 2009. Classinfocus: enabling improved visual attention strategies for deaf and hard of hearing students. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 67–74. ACM.

George A Gates and John H Mills. 2005. Presbycusis. *The Lancet*, 366(9491):1111–1120.

Tobias Goehring, Xin Yang, Jessica JM Monaghan, and Stefan Bleeck. 2016. Speech enhancement for hearing-impaired listeners using deep neural networks with auditory-model based features. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 2300–2304.

Dorota J. Iskra, Beate Grosskopf, Krzysztof Marasek, Henk van den Heuvel, Frank Diehl, and Andreas Kiessling. 2002. Speecon - speech databases for consumer devices: Database specification and validation. In *LREC*.

Matti Karjalainen, Peter Boda, Panu Somervuo, and Toomas Altosaar. 1997. Applications for the hearing-impaired: Evaluation of Finnish phoneme recognition methods. In *Fifth European Conference on Speech Communication and Technology*.

Richard Kheir and Thomas Way. 2007. Inclusion of deaf students in computer science classes using real-time speech transcription. In *Proceedings of the 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, ITiCSE '07, pages 261–265. ACM.

Raja Kushalnagar and Poorna Kushalnagar. 2014. Collaborative gaze cues and replay for deaf and hard of hearing students. In *Computers Helping People with Special Needs*, pages 415–422, Cham. Springer International Publishing.

Raja S. Kushalnagar, Anna C. Cavender, and Jehan-François Pâris. 2010. Multiple view perspectives: Improving inclusiveness and video compression in mainstream classroom recordings. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, pages 123–130, New York, NY, USA. ACM.

Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2013. Captions versus transcripts for online video content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 32:1–32:4, New York, NY, USA. ACM.

Walter S. Lasecki, Christopher D. Miller, Iftekhar Naim, Raja Kushalnagar, Adam Sadilek, Daniel Gildea, and Jeffrey P. Bigham. 2017. Scribe: Deep integration of human and machine intelligence to caption speech in real time. *Commun. ACM*, 60(9):93–100.

Juri Lukkarila. 2017. Developing a conversation assistant for the hearing impaired using automatic speech recognition. MSc Thesis, Aalto University.

André Mansikkaniemi, Peter Smit, Mikko Kurimo, et al. 2017. Automatic construction of the Finnish parliament speech corpus. In *INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association*.

Tara Matthews, Scott Carter, Carol Pai, Janette Fong, and Jennifer Mankoff. 2006. Scribe4me: Evaluating a mobile sound transcription tool for the deaf. In *UbiComp 2006: Ubiquitous Computing*, pages 159–176. Springer.

Mohammad Reza Mirzaei, Seyed Ghorshi, and Mohammad Mortazavi. 2012. Combining augmented reality and speech technologies to help deaf and hard of hearing people. In *2012 14th Symposium on Virtual and Augmented Reality*, pages 174–181.

Mohammad Reza Mirzaei, Seyed Ghorshi, and Mohammad Mortazavi. 2014. Audio-visual speech recognition techniques in augmented reality environments. *The Visual Computer*, 30(3):245–257.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Rohit Ranchal, Teresa Taber-Doughty, Yiren Guo, Keith Bain, Heather Martin, J Paul Robinson, and Bradley S Duerstock. 2013. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, 6(4):299–311.

Joseph Robison and Carl Jensema. 1996. Computer speech recognition as an assistive device for deaf and hard of hearing people. In *Biennial Conference on Postsecondary Education for Persons Who Are Deaf or Hard of Hearing (7th, Knoxville, Tennessee*, volume 948, page 154. ERIC.

Peter Smit, Sami Virpioja, and Mikko Kurimo. 2017. Improved subword modeling for WFST-based speech recognition. In *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, pages 2551–2555, Stockholm, Sweden.

Kazuki Suemitsu, Keiichi Zempo, Koichi Mizutani, and Naoto Wakatsuki. 2015. Caption support system for complementary dialogical information using see-through head mounted display. In *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*, pages 368–371.

Joris Van Gelder, Irene Van Peer, and Dzmitry Aliakseyeu. 2005. Transcription table: Text support during meetings. In *IFIP Conference on Human-Computer Interaction*, pages 1002–1005. Springer.

Anja Virkkunen. 2018. Automatic speech recognition for the hearing impaired in an augmented reality application. MSc Thesis, Aalto University.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical report, Aalto University.

Mike Wald. 2006. Captioning for deaf and hard of hearing people by editing automatic speech recognition in real time. In *Computers Helping People with Special Needs*, pages 683–690, Berlin, Heidelberg. Springer.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2410–2423.

Dong Yu and Li Deng. 2015. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, London.