MAUU5900
MASTER THESIS
In
Universal Design of ICT

May 2019

CAPTIONING FOR THE DHH

S310283 Siv Tunold

Department of Computer Science

Faculty of Technology, Art and Design

OSLOMET

CONTENTS

LIST OF FIGURES

1 PREFACE.

I met this man a few years ago - he impressed me! He was both deaf and blind. A well-educated, happy man and he was even playing some kind of football. How has he been able to cope with his life? How is it possible? It is quite impressive isn't?

As being a visually impaired myself, I have been very focused on accessibility and universal design of ICT. What can we do to help the disabled people? How can we make their life easier?

So universal design is important. The society want us to be independent and it's hardly any local bank offices left. The society expect you to pay your bills through internet-banks and buy your train-tickets by mobile. So, all our IT solutions have to be available for all. In Norway we have a law who says that all ICT solutions shall meet the requirements of universal design by 1 January 2021.

I have made some changes to make this document easier to read with a screen-reader. All headings are made very visible by using a background-colours or by a line.. All images are marked as decorative and have no alternative text, but they all have captions. No tables have been used to format text, the list of definitions doesn't look so nice, but it is easier to hear. The document has also passed Words integrated accessibility check.

Some appendices in this document is not screen reader friendly.

1.1 Acknowledgement.

All my thanks to Terje Gjøsæter, my coach and supervisor from OSLOMET.

1.2 Structure of the report

In the first part of the report you will find 4. Problem description, before the Research questions are defined in chapter 5. The methodologies used is described in chapter6. A literature review is done too, 7 .Literature review - Scientific reports and 8. Literature review

- Background and state of the art . The report will have the Result and Findings. The suggested Result and Findings.Future work is at the end.

Among the appendices you will find screen-prints of the prototypes, the interview guides and a description of automatic speech recognition. It also includes a list of definitions, acronyms and the project plan.

2 ABSTRACT.

The main rule in Norway is that all ICT solutions have to be universally designed, and this is covered in the gender equality and anti-discrimination law. One of the most common faults in web-sites today are the lack of adequate subtitling of videos. Audio content have to be presented in different ways, without losing meaning and the captions must be readable by assistive technologies.

When you are unable to hear, you also have problems to communicate with people who can hear, most of them are of are unfamiliar with the use sign-language. This increases the isolation of deaf people from society. Children and adolescents who are DHH gets poorer results at school compared to their peers with normal hearing. They are lonelier and struggling to join the community. Transcribing by ASR might make communication easier.

During interviews, the response to Google's beta version of a transcription app as well as the author's prototype was met with overwhelmingly positive response. One respondent claimed that 'It really would change the way she solves her communication problem; "I don't need as so much imagination as in an *ordinary* conversation". A cheap an easily available smartphone app was considered as appropriate. The fact that many would know the concepts of transcribing and were familiar to dictating would make everyday life easier for the DHHs.

Audio content have to be presented in different ways, without losing meaning. The captions must be readable by assistive technologies. A lot have to be fixed with the beta app before it gets accessible for all – if you are being depend of a braille it is nearly impossible to read the transcribed text.

The speech recognition done by ASR is not perfect. The errors sometimes make the transcribing difficult to trust and may not be usable in more formal situations. Maybe a mutually approval of the transcription of both speakers are necessary.

3 INTRODUCTION.

"Speech you don't understand translated to something you do understand, done by using captioning for the deaf and hard of hearing (DHH)."

The World Health Organization (WHO) estimates that 360 million people have a disabling hearing loss, that is around 5% of the world population. A large (and increasing) number of DHH people lose their hearing later in life, which includes one third of the population over 65. In Norway hearing-related disorders are one of the fastest growing public health challenges. According to (BUFDIR, 2015) 14.5 % of the population in Norway or nearly 700,000 Norwegians have a significant hearing impairment, of whom approx. 3,500 - 4,000 are deaf. Of these, over 3000 children and adolescents are under the age of 20, where education is a very important part of their lives. Research indicates that 1 million Norwegians may have a hearing impairment by 2020 (BUFDIR, 2015). Elderly people often have a hearing impairment. This is due to age related changes in the inner ear, the hearing nerve and in the central hearing lanes. In America they are talking about more than 30 million with some type of hearing loss. Hørselshemmedes landsforbund (HLF) has 55,000 members and is the largest organization for people with disabilities in Norway.

The DHHs struggle to understand the audio input and benefit a lot from visual input. They combine a personal and different mix of lip-reading, sign language interpreters or real-time caption typists to cope. Sign language is as a different language with their own vocabulary, grammar, and syntax. People who recently have become DHH usually do not understand it at all.

Some people do have more than one impairment; they might be both deaf and blind. They need the captions as braille or tactile sign language given by an interpreter.

Captions is a way to make information given by audio available for people who cannot hear it. Captioning is the process of converting the audio content of a television broadcast,

webcast, film, video, live event, or other production into text and displaying the text on a screen or on some kind of monitor as a braille. Captions provide the part of the content available via the audio track. Captions are not only including the dialogue, but also identify who is speaking and include non-speech information conveyed through sound, including meaningful sound effects.

Captions enable people who are deaf or hard of hearing (DHH) to watch synchronized media presentations and videos. It is critical for students who are DHH, and it also aids the reading and literacy skills development of many others, as foreigners and illiterates.

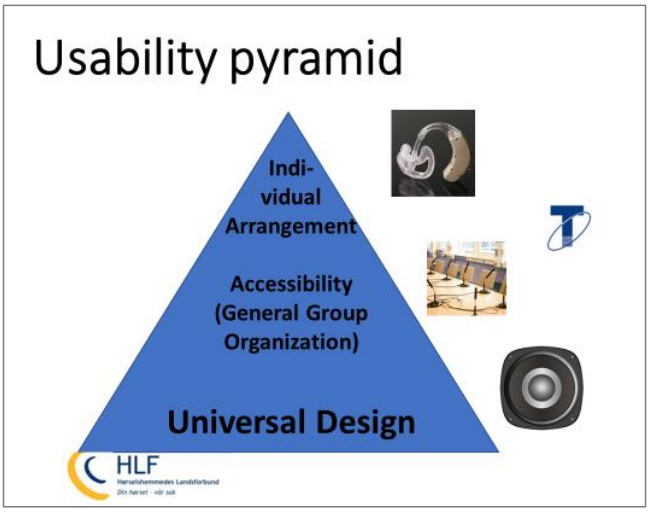


Figure 1 The usability pyramid from HLF, translated from Norwegian

The bottom layer is universal design available for all. Important for the hearing impaired are factors as good sound, acoustic and the background noise. The middle layer is accessibility for a group of people as microphones and Induction loops. The top layer are individual arrangements as a hearing-device or a bilingual interpreter.

But who needs captions? Only the DHH? In fact, there are situations where everyone would think that subtitles would have been nice –as when you are without a headset, when you hardly hear anything because of the noisy environment or when you do not want to disturb anyone – then you are at the bottom layer of the pyramid.

The middle layer is for all the DHH people, they need help to hear the audio. They need captions to “hear” a video, caption for both the dialogs and the sound effects.

It is important to distinguish between persons who are deaf from childhood and has sign language as their native language or people who have become deaf later in life and having a spoken language as mother tongue. Many childhoods deaf have a great need for sign language information.

Within certain parts of the deaf movement, they consider themselves as less disabled and more like a linguistic minority with their own culture. Sign languages are not simply codes for an aural language, but rather an entirely different language with their own vocabulary, grammar, and syntax. You could actually look at it as two persons who talks with each other in a totally different language than your own. Therefore, this audience needs videos in sign language to independently access information online. The deaf-blinds use both voice captioning and a Braille to read a video.

3.1 Accessibility requirements

The accessibility requirements in Norway are regulated by law. ~~Formal requirements - WCAG~~The main rule is that all ICT solutions should be universally designed. These legislations are based on the international guidelines WCAG (W3C 2008) and some more work done in EU. WCAG 2.1 consists of technology-neutral principles, guidelines, and success criteria, these guidelines and success criteria can be satisfied in a variety of ways.

The WCAG guidelines are organized under 4 principles: perceivable, operable, understandable, and robust. In this report the main focus is on perceivable for the DHH – all audio must be presentable to users in ways they can perceive it.

- Make it easier for users to see and hear content.
- Provide captions and other alternatives for multimedia.
- Audio content that can be presented in different ways without losing meaning.
- The captions must be readable by assistive technologies.

For more about the formal requirement see chapter 9. Formal requirements.

3.2 Ethical considerations

The goal for everybody is to be a part of the society, also when you are not able to hear what is said around you. If you need an interpreter when you go to the doctor or at a meeting with your lawyer and always are dependent of having a third person present. How much privacy are you having …?

What about the quality on the interpreting and captioning? Do the DHH understand what have been said? Some misunderstandings might even be quite dangerous. Errors makes the learning process even worse for the students who are trying to understand and learn what the teacher say.

Going for a step by step development might lead to some ethical considerations and problems. Some might find the discrimination hard to accept If they have to wait two years tor the same functionality as others.

An ASR system who knows the speaker are giving better and better quality on the transcriptions, it is keeping learning of what have been said by saving the data and continuously building new statistic of the user's speech. Then you have a privacy issue – your talk and data are saved in the cloud somewhere, just think of the GDPR directives.

4 PROBLEM DESCRIPTION

DHHs struggle to understand audio input and benefit a lot from visual input. They combine a different mix of lip-reading, sign language interpreters or real-time caption typists to cope., some even need braille. People who newly have become DHH usually do not understand sign language.

Norway has through the UN's resolution on human rights for people with disabilities committed themselves to ensuring full participation for all in political and public life. (UN, 2016) - This statutory right is often broken and in many areas. The Norwegian association for the hard of hearing (HLF) calls for more pressure on universal design. A lot have been done already, but there is still much left to do. Hearing-impaired rights are still sometimes being ignored and do need more pressure to make things better.
Many organizations are too kind and patient in their demands for equality. The parliament, which even adopts laws, violates the law of universal design. Subscriptions of parliamentary debates remain in the future.

But who needs captions? Only people who are deaf or hard of hearing (DHH)? According to BBC[1] 80% of closed captions users don't have any hearing loss. In fact, there are situations where everyone could think that captions would have been nice – as when you are without a headset or when you hardly hear anything because of the noisy environment. Or when you do not want to disturb anyone around you. When BBC asked their viewers why they use subtitles, they said that "subtitles were regarded as generally very effective in making programmes understood". Captions enable the DHHs to watch synchronized media

[1] https://www.3playmedia.com/2015/08/28/who-uses-closed-captions-not-just-the-deaf-or-hard-of-hearing/

presentations and videos. It is critical for DHH students and it also aids the reading and literacy skills development of many others, as foreigners and illiterates.

This report has a general view and are not looking at individual solutions. Many in the DHH society are not looking for the perfect solution, but are prepared to make incremental steps to get a better existence. Increasingly many people are getting older, thereby more people are also getting hard of hearing. The possibility of more social integration is very important. To understand and to be understood is vital, loneliness can be a huge problem for many DHHs. Depending of the situation the tolerance of errors can be quite high. It is to be able to get simultaneous captioning of the main keywords that's matter in many unformal social settings.

Privacy is another issue. How private is it when you need a sign interpreter when visiting the doctor, in the bank or when you are talking with your boss at work?

4.1 The requirements of the regulations.

The web is Information. The information can be presented visually or audibly. It can be presented in graphics, video, audio, animation, or in text. The web is increasingly consisting of video and multimedia content. The description does not need to be an exact reproduction of the content, but there should be enough information to understand the most important message or purpose of the audio clip.

The Norwegian are regulated by the laws (Lovdata, 2018); The gender equality and anti-discrimination law and the regulation on universal design of information and communication . New regulation form in EU (EU, 2019) will also affect Norwegian law beyond 2019.

The basic requirements for this law are the Web Content Accessibility Guidelines (WCAG) developed by World Wide Web Consortium (W3C) which provides a technical standard for web content accessibility (W3C 2008). WCAG 2.1 covers a wide range of recommendations for making Web content more accessible. Following these guidelines will make content more

accessible to a wider range of people with disabilities, including accommodations for blindness and low vision, deafness and hearing loss, limited movement, speech disabilities, photosensitivity, and combinations of these, and some accommodation for learning disabilities and cognitive limitations.

The techniques for providing accessibility for users with disabilities are very straightforward - provide captions and transcripts for multimedia content (meaning video content that also has audio) and provide transcripts for audio-only content.

Universal design does not only make your pages more accessible for disabled people but makes a better user interface for everybody. It does not have to be more expensive either, but remember to think about the universal design already in the design phase. Captions gives the pages better Search Engines Optimizing (SEO), pure audio is not easy searchable.

4.2 Make the WEB accessible.

A very relevant issue within universal design relates to people that are hard of hearing and concerns the lack of captioning in videos and movie clips published on the web (W3C, 2018b). Ranging from short video clips on social media to live streaming of conferences, press conferences, and live web. The ultimate goal is:

Make the web accessible for everyone, regardless of their different handicaps if any.

This research report is looking for resource efficient ways to do captioning of speech. This could be achieved through using technology for automatic speech recognition (ASR), alternatively by an increased use of existing manuscripts published as text files. The report also includes some examples of how this can be achieved and could help us move accessibility for the hard of hearing people one step further. How could captioning be done?

"Captions" commonly refer to on-screen text specifically designed for hearing impaired viewers, while "subtitles" are straight transcriptions or translations to another language of

the dialogue. Captions are usually positioned below the person who is speaking, and they include descriptions of sounds such as gunshots, closing doors or music. This means captions done by ASR actually are subtitles unless somebody has processed them and included sound effects too.

4.3 Socializing

Human communication is an important part of everyday life. Whether you are cutting your hair at a hairdresser, ordering food at a restaurant, asking for directions, receiving a phone call from a family member or talking to a colleague. Communication is a basic human need in life and is very powerful in its ability to connecting individuals, the exchange of information, and building of relationships.

Those who are deaf have limited communication accessibility compared to the one's hearing, and by default, they will obtain less public information and face more obstacles during social interactions. The DHH still possess the same need to communicate, especially in one-on-one situations in everyday interactions. Often, when communication is attempted between hearing and deaf individuals, there are awkward exchanges, confusion, misunderstanding and as a result, cultural misconceptions are developed. Sometimes the DHHs withdraw, and loneliness could be a huge problem for some.

Loneliness and being able to manage yourself are two important issues, to enable users to interact with the not disabled community in daily activities. What about warnings and alarms coming from the intercom, by a blinking light at the railway station - but if you can't see and hear it?

With today's existing technology and software, a solution needs to be generated to maximize accessibility and improve the quality of communication interactions between deaf and hearing individuals.

Voice based communications are the most common form of daily one-on-one interactions. Using a text- and braille-based strategy approach makes this possible for the deaf-blinds as well. It could also play an underlying role in improving literacy for the deaf.

ASR and the wide use of smart phones and their apps have allowed huge inroads when preparing DHH students to be effective and productive in the hearing workplace (Easton, 2017). This paper presents both a hearing instructor's experiences and a deaf researcher's observations when preparing deaf and hard of hearing students as computer technicians for the hearing work place. A hard of hearing person finds conversations in groups difficult, especially with background noise are challenging. It can be difficult to remember all the conversation and to use the information in follow up work after the group conversation. They want to explore use of automatic speech recognition to facilitate communication between DHH and hearing persons in meetings when an interpreter is not available and cost-effective methods to facilitate DHH individual's communication in the workplace are not available.

4.4 Error tolerance.

The correctness and the quality of the captioning achieved is an issue. The different kinds of errors between sources of captions shows that not all errors are equally important.

ASR tends to replace words with others that are phonetically similar, but differ sometimes in meaning, while humans tend to replace words with ones that have similar meanings, just some misplaced letters, spelling errors which are easier for readers to interpret. Even speaker-dependent ASR has difficulty recognizing domain-specific jargon. The captions need a low error rate so that the reader is not derailed during the reading process.

Marking potential errors in the captions feels confusing for the readers (Raja S. Kushalnagar, Lasecki, & Bigham, 2014). It is even worse to read the marking of potential errors with colours and fonts when using a braille (writers' own comment).

4.5 Fast speaking – is challenging.

Some people talk fast, sometimes shortening of content is required to make the caption readable. Especially considering the recipient group - depending on age and education, the ability to read text fast is varying too.

In the survey of (Raja S. Kushalnagar et al., 2014) they showed that the average speaking rate was around 170 words per minute (WPM). In order to transcribe an average lecture, a transcriber has to maintain a rate of at least 170 WPM. Even proficient touch typists are only able to enter around 75 WPM on QWERTY keyboards.

Parallelize the number of typists is a solution which is potentially much more effective than stenography of QWERTY keyboards (Lasecki et al., 2017)

4.6 Norwegian is a small language.

One of the main problems in a small country as Norway, it is not many people who have Norwegian as a native language. We even have two written languages, Bokmål and Nynorsk, and there are of course different dialects too. The easily available software is usually using the world languages as English, Chinese and Spanish, and free good solutions will probably not soon be available for us.

Pronouncing problems - as an example like foreigners do have or deaf people usually have, make it hardly possible at all to transcribe their speech automatically.

5 RESEARCH QUESTIONS

RQ1: What features would be important for users being both DHH and visually impaired when they want to use ASR captioning?

RQ2: Are DHH users willing to try a non-optimal solution as a start, to facilitate the development of an incrementally improving versions of an ASR system, or would they rather wait for the «perfect» system?

6 METHODOLOGY.

The methodology work with this report consists of a literature review, some interviews and discussions around an external Google-based prototype.

The focus in this report has been on quality not quantity numbers. There have been few respondents in the interview group, but the main goal has been to find tendencies in the community, not a statically significance.

6.1 Literature review

More details in the literature review are described in the chapters 7. Literature review - Scientific reports and 8. Literature review - Background and state of the art.

6.2 Prototype

This report has been under progress for the last three years and it seems like my original idea for a prototype was quite good after all as both Google and Microsoft have done something similar. Google has released their first release of an ASR app and MS Office 365 has integrated both ASR and translation in their laptop office tools.

The work started by creating a simple prototype in PowerPoint. This was made to have a starting point for the discuss in the interviews. The simple “paper”-prototype is to be found in appendix 14.4. But in February 2019 Google released a beta version of their entirely new transcription app - Google Live Transcribe, for the newest Android mobiles. This beta app is quite good and gives a more realistic impression with ASR integrated. I decided to use this one as a prototype instead. The beta app is to be found in the appendix 14.5.

6.3 Interviews

Some DHH representatives were interviewed about their thoughts and opinions about captions as an aid. The interviews do not give the solution, but just a tendency. There were two groups of candidates; DHH and non-DHH and only grown-ups in the age from 30 to 80 years old.

Some of the topics which were discussed during the interviews;

- Captioning, how? Interpreter, sign language, humans captioning in real time, captioning, of films, simultaneous captioning by ASR.
- Do you know what ASR is?
- The quality of the captioned text is affected by the (un)known speaker, dictation ability, and cost. What do you think?
- When is there a need for an interpreter, in which situations? At work, at school, or just for socialization. What about having a third person present? Correctness of the interpreting?
- Quality of the captioned speech by ASR is affected by the (un)known speaker, dictation ability, and cost. Incremental steps, rather than a big one. Incremental steps provide faster results and corrections along the way would be possible.
- One more device? The world is full of gadgets and apps. Maybe an app for your mobile.
- An app, a useful app also for hearings too. Something in it for all, easily accessible, easy to learn how to use. More people will get familiar with the concept of automated captioning, more comfortable and get better at dictating.
- Show and discuss the different prototypes.

The interview-objects were shown the beta release of Google Live Transcribe. The app was used to give them a feeling of how it could be, and the not always correct transcription done by ASR. See Appendix 14.5

7 LITERATURE REVIEW - SCIENTIFIC REPORTS

7.1 Confidence

The ASR technologies have seen major progress in their accuracy and speed the last years. Due to their cheap and scalable ability (compared to other captioning alternatives) to generate real-time text from live audio or recordings, ASR systems have a potential for the task of captioning.

The researchers have been looking at the relationship between ASR errors and the impact it has on the understandability of a text for DHH users and tried to describe a prediction model that represents the impact of ASR errors present in the text on its comprehension.

The errors in the ASR output however is still a present challenge in the use of a fully automatic system (Kafle & Huenerfauth, 2016). Their research of are looking more closely into the impact of different inaccurate transcriptions from the ASR system and how they affect the understandability of captions for DHH individuals.

One fundamental problem that makes real-time captioning difficult; sequential keyboard typing is much slower than speaking and research findings suggest that captions need to be provided within 5 seconds so that the student can participate.

(Lasecki et al., 2017) ASR is inexpensive and available on-demand, but its accuracy is not good enough, sometimes it drops below 50% when the ASR solution is not trained on the speaker, captioning multiple speakers, or when not using a high-quality microphone. Software used to assist real-time captionists may often make errors that can change the meaning of the original speech. As DHH people use context to compensate for errors, they often have trouble following the speaker.

A study ((Raja S. Kushalnagar et al., 2014)on using ASR software in captioning lectures showed 75% accuracy on untrained ASR software and could reach 90% under ideal single

speaker conditions. This is still too low for use by deaf students. The challenges for the ASR are modern classroom lectures that have extensive technical vocabulary, poor acoustic quality, multiple information sources, or speaker accents.

ASR tends to replace words with others that are phonetically similar, but differ sometimes in meaning, while humans tend to replace words with ones that have similar meanings, just some misplaced letters. Additionally, the errors ASR makes can often change the meaning of the text. Human captioners on the other hand, will typically only omit words they do not understand, or make accidental spelling errors which are easier for readers to interpret. Even speaker-dependent ASR has difficulty recognizing domain-specific jargon. The captions need a low error rate so that the reader is not derailed during the reading process. The participants' perception of the different kinds of errors between sources of captions, not all errors are equally important.

Since the speech rate is faster than the average caption reading rate, captioners tend not to include all spoken information so that readers can keep up with the transcript. Instead, they abbreviate sentences or drop words. This makes the flow smoother.

(Kafle & Huenerfauth, 2016) said recent advancements in the accuracy of ASR technologies are potential candidate for the task of captioning. Over the past few decades, ASR technologies have seen major progress in their accuracy and speed. Due to their cheap and scalable ability (compared to other human captioning alternatives) to generate real-time text from live audio or recordings, ASR systems have a potential for the task of captioning. The errors in the ASR output however is still a present challenge in the use of a fully automatic system. Their research is looking more closely into the impact of different inaccurate transcriptions from the ASR system and how they affect the understandability of captions for DHH.

Researchers (Kafle & Huenerfauth, 2016) have begun to investigate the suitability of ASR to automate or semi-automate the process of captioning with the use of ASR in various application settings. This is an example from this report of a not correct ASR caption:

An example - Correct text:

"The meeting today has been cancelled and is scheduled for next Thursday."

The same example, but with errors made by ASR:

1. The meet in today has been cancelled an is scheduled for next Thursday.
2. The meeting today has been capital and is skidoo for next Thursday.

The average understanding of the last sentence was 40%. For some more examples with the prototype see Some transcription tests were run.

7.2 Accuracy of captions.

As (Parton, 2016) writes in her report – Nevertheless, even for a first draft, accuracy might be a problem, but the reaction to the severity of the inaccuracy is mixed, ranging from 'devastating', 'a barrier to communication' and 'humorous' to 'a fairly good job'. An often-heard recommendation is to start with auto-captions and then edit and correct it to reduce the number of errors and fix any timing issues.

Mark-up

Recent advances in ASR (Berke & Caulfield, 2017)have made this technology a potential solution for transcribing audio input in real-time for people who are DHH. However, ASR is still imperfect; users must cope with errors in the output. They have conducted two studies comparing various methods of visually presenting the ASR output with certainty values, how ASR captioning could be used with confidence display mark-up.

Users preferred captioning styles which they were already most familiar with (that did not display confidence information), and they were concerned about the accuracy of ASR

systems. While the participants expressed interest in systems that display word confidence during captions, they were concerned that text appearance changes may be distracting.

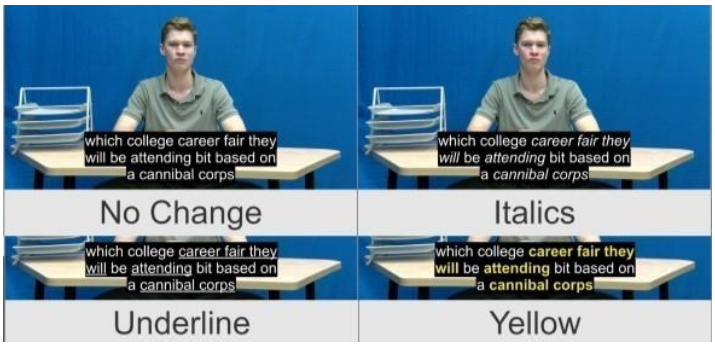


Figure 2 The prototype tool examined in this work, with different styles of confidence display mark-up

Feedback given by some of the participants:

Among those that shared negative experiences, a common theme that arose was that they were dissatisfied with ASR accuracy and they felt that it was too frustrating to follow the captioning. As some said - “It becomes really annoying!”

7.3 Wearable

The article (Schipper & Brinkman, 2017) have explored the possibility of improving everyday communication for those who are DHH by a head-worn video see-through augmented reality application. Three different captioning modes who were tested.

1. Locked mode - the captions always stay within the field of view of the user.
2. Delayed mode - the captions are following the user’s direction of view but move at a delay when the user rotates their head.
3. Bubble mode - involves dynamically placing captions on the speaker.

They concluded that the optimal caption placement for head-worn augmented reality applications was different from what is optimal for stationary screens or for hand-held applications. They believe the delayed mode (2) improves readability.

How to control the caption device is an issue to beware of as the researchers (Easton, 2017) also experienced while they tested an ASR app for smartphones. They also suggested an ASR app that does not require the use of a send button, which did not include finger clicking in the caption process.

According to (Easton, 2017) ASR and the wide use of smart phones and their apps have allowed huge inroads when preparing DHH students to be effective and productive in the hearing workplace. In future studies the authors would like to implement a watch instead of a traditional smart phone and an ASR app that does not require the use of a send button.

7.4 Multi-Handicapped

Captions provide DHH users access to the audio component of speech, web videos and television. While hearing consumers can watch and listen simultaneously, the transformation of audio to text requires deaf viewers to watch two simultaneous visual streams: the video and the textual representation of the audio.

Prior research has shown that the cognitive process of reading a transcript or caption that constantly changes is very different from the cognitive process of reading print that does not change during the course of reading[1]

Caption interfaces could help viewers to adapt and follow better the wide variety of speeds and complexity of different content categories (Raja S Kushalnagar, Lasecki, & Bigham, 2013). This can be a problem when the video has a lot of text or the content is dense, or if you are DHH and visually impaired.

Unlike print, captions force readers to read the text at a variable pace; and the readers cannot control or predict that pace. Viewers need time to read text and the captions can be hard to read when you try to watch the continuously changing surroundings.

Humans use all their five senses (sight, hearing, taste, smell and touch) to understand and evaluate their surroundings, but electronic devices communicate with us using predominantly just two of them: sight and hearing. If you miss one of these senses, you have to use the other. However, having two impairments, if you are being both deaf and blind, you will meet a more complicated existence, with only taste, smell and touch are left to use. Of the remaining, the touch sense is most likely to be used.

The tactile sense can be a means of communication to provide some kind of information to sensory disabled individuals. Conclusions: A lack of acceptance emerged from the discussion of capabilities and limitations of haptic assistive technologies.

Tactile – braille

Text, the captions, can be written in Braille making it possible to read the text by using your fingertip. With a Braille display attached to your mobile or laptop you can read all the text on the screen.

Haptic feedback

Human beings have five senses, but electronic devices communicate with us using predominantly just two of them: sight and hearing. Haptic feedback is the use of touch to communicate with users. Most people are familiar with the vibration in a mobile phone or the rumble in a game controller – but haptic feedback is much more.

While the tiny devices that create vibrations as a mobile phone are probably the best-known haptic technology, there are many other ways to simulate touch. It is still under development, but somebody is creating a jacket (Delazio et al., 2018) that allows the

profoundly deaf to “feel” and understand speech. The jacket features dozens of embedded sensors that vibrate in specific patterns to represent words.

To make it possible for a deaf person to be able to understand a speaker, the most commonly known method consists in that this person learns to read lips. There are, however, situations in which this person cannot see the speaker, as a message from the speaker transmitted by an intercom.

Visual and auditory inputs are converted in haptic feedback via different actuation technologies (Sorgini, Caliò, Carrozza, & Oddo, 2018). The information is presented in the form of static or dynamic stimulation of the skin. If you are both visually impaired and DHH you have a great problem in understanding captions, you are not able to see them, only by using a braille connected to your phone or laptop you are able to read them.

As they say in their article (Sorgini et al., 2018) future researches are oriented towards the optimization of the stimulation parameters together with the development of miniaturized, custom-designed and low-cost aids operating in synergy in networks, aiming to increase patients’ acceptability of these technologies.

Haptic technology communication recreates the sense of touch by applying forces, vibrations, or motions to the user. This mechanical stimulation can be used to assist in the creation of virtual view of the speaking surroundings; as the direction of who is speaking. A tactile stimulation device is capable of transforming an audio signal into vibrations which can be sensed by the skin of the DHH telling him the direction to the person speaking.

The article (Findlater et al., 2019) uses three different design scenarios (smartphone, smartwatch, head-mounted display) and two output modalities (visual and haptic), and probes issues related to social context of use. While most participants were highly interested in being aware of sounds, this interest was affected by communication preference as sign or oral communication or both. Almost all participants wanted both visual and haptic feedback and more than two thirds preferred the feedback on separate devices (e.g., haptic

on smartwatch, visual on head-mounted display). The social context and other findings related to sound type, full captions, sound filtering, notification styles, can provide direct guidance for the design of future mobile and wearable sound awareness systems.

Sign language

Unfamiliarity with sign-language increases the isolation of deaf people from society, captioning might make communication easier.

The translation between sign and text is a complete machine translation challenge, because the two languages have different structures and grammars. In the work done by (Luqman & Mahmoud, 2018), they propose a rule-based machine translation system to translate Arabic text into Arabic sign language. The findings in their translation system provides an accurate translation for more than 80% of the translated sentences.

8 LITERATURE REVIEW - BACKGROUND AND STATE OF THE ART

A lot have already been done to bridge this communication gap for the deaf. There are existing stand-alone tools and applications that can assist with communications between the deaf and hearing. Some are video-based interpreting platforms and others are text-based applications such as email, notepad and messenger applications. One could also use voice-to-text dictation software such as Siri or Dragon as well.

Human communication is an important part of everyday life. Those who are deaf have limited communication accessibility compared to the hearing, obtain less public information and face more obstacles during social interactions. Regardless of the variables and literacy comprehension, all individuals are driven by the need to communicate on a daily basis.

8.1 Captions – how to do it.

You must be honest and trustworthy as a captioner. Caption the speech as it has been said (use spoken dialect and language) and do not forget the audio and visual effects. Include as much of the original language as possible. Editing the original transcription may be necessary to provide time for the caption to be possible to read and for it to be in synchronization with the audio. Captions must be timed-synced with the audio. Information could be missed if the action and the words do not match up.

Remember captions should be **ACCRE – Accurate, Consists, Clear, Readable** and **Equal** ((DCMP, 2018). It will not be easily readable if the text is too small or if it does not contrast enough with the background. Spelling errors may cause unnecessary confusion over the content. It is also important to label who is speaking so that the dialog is easy to follow.

Accurate. Errorless captions are the goal for each production. Errors may confuse the reader.

Consists. Uniformity in style and presentation of all captioning features is crucial for viewer understanding. Font size and background colours matter.

Clear. A complete textual representation of the audio, including speaker identification and non-speech information, provides clarity.

Readable. Captions are displayed with enough time to be read completely. They must be in synchronization with the audio and are not obscured by (nor do they obscure) the visual content.

Equal. Equal access requires that the meaning and intention of the material is completely preserved.

8.2 How to easily check a web page?

Some easy checks you can do to check how accessible your own web page is for the DHH (DIFI, 2018).

Prerequisite: Users need to be able to read what they can't hear.

Solution: Present text options for video and audio files.

Single test: Turn off the sound and check if you still do understand the content.

Prerequisite: Users with sign language as native speakers have limited literacy.

Solution: Present the important information in sign language (not a legal requirement, only recommended).

Single test: Think about what your most important information is and see if there is a sign language version of this.

Prerequisite: Make the video more available for everybody.

Solution: Think before you start filming. Maybe the speaker could describe the sounds and surroundings as well too?

Single test: Close your eyes and just listen to the caption text. (You are both deaf and blind) Is it possible to understand what is going on in the video?

8.3 Captioning standards.

Does it matter what the captions look like or how they are displayed? (DCMP, 2018)

These standards are based on best practices and are consistent with the mandates by the Federal Communications Commission (FCC). Elements of Quality Captioning from (DCMP, 2018)

Past research (Raja S. Kushalnagar et al., 2014) has indicated that the majority of deaf high school graduates' literacy level are lower than their hearing students at the same level. Which makes the error problem even bigger for the understanding of the captions. They (Raja S. Kushalnagar et al., 2014) have been looking the relationship between ASR errors and the impact it has on the understandability of a text for DHH users and tried to describe a prediction model. Word Error Rate (WER).

(Schipper & Brinkman, 2017) The researchers evaluated ASR performance by focusing on improving the Word Error Rate (WER) metric. They have proposed a new captioning-focused evaluation metric that better predicts the impact of ASR recognition errors on the usability of automatically generated captions for people who are DHH.

In a side-by-side comparison of pairs of ASR text output (with identical WER), the texts preferred by their new metric were also preferred by DHH participants. The new metric had significantly higher correlation with DHH participants' subjective scores on the usability of a caption. The researchers suggest this new metric could be used to select ASR systems for captioning applications, and to consider when optimizing ASR systems.

8.4 CROWD captioning

While ASR technology has improved, there are still errors in the output, especially in noisy and complex environments. To boost the accuracy of imperfect ASR, some researchers have

created systems in which humans to fix mistakes in ASR output or have crowdsourced the task of transcribing audio. The business model for these services requires some regular payment for the human labour.

The article (Raja S. Kushalnagar et al., 2014) shows that both hearing and DHH participants preferred and followed collaborative captions better than those generated by automatic speech recognition (ASR) or professionals due to the more consistent flow of the resulting captions.

As (Kawas, Karalis, Wen, & Ladner, 2016)) tells many DHH students choose real-time captioning to access mainstream classes with hearing teachers and students. Today, real-time captioning solutions such as CART (communication access real-time translation) and C-Print have human captioners attend students' classes either in person or over a remote Internet call and transcribe course content as the instructor speaks. Captioners must be scheduled in advance, they have limited availability and tend to be costly. This is making them difficult to offer by budget-conscious schools. ASR is a machine technology that converts speech to text in real time, offering a low-cost alternative to human captioners. Poor accuracy and latency have been serious limitations of ASR as a real-time captioning tool, preventing it from catching on as a legitimate option for DHH students. However, as the technology improves ASR becomes increasingly viable as a classroom captioning tool.

ASR is a technology that recognizes and transcribes spoken language into readable text in real time. ASR has a wide range of applications such as supporting digital personal assistants like Siri and providing closed captions for online videos on websites like YouTube and Coursera. However, ASR faces issues with accuracy, latency and context formalization due to the variability and complexity of human speech. Factors like background noise, speaker accent, speech rate, speaking style and spontaneous speech can degrade the quality of ASR transcriptions. Many applications leverage ASR technology, though they tend to be for personal use or one-on-one communication. For example, Dragon NaturallySpeaking

software is known for providing high quality speech-to-text for memo dictation. Skype Translator converts speech to text and provides machine translation in real time over a Skype video call, allowing speakers of different languages to communicate. These kinds of applications were not designed for real-time captioning in a classroom setting.

As (Kawas et al., 2016) tells ASR is a desirable real-time captioning solution because of its low cost compared with human captioning services or sign language interpreters. ASR is not perfect, but combined with human labour it can make a better caption of speech on shorter time. Respeaking another attempt to maintain high quality and low cost by using ASR as a starting point for human typists.

Scribe.

Researchers have thus begun to devise ASR-based solutions that can benefit DHH users. E-Scribe proposed the technology (Lasecki et al., 2017) for a web-based ASR captioning solution for DHH users. The APEINTA (Kawas et al., 2016) system provided automatic speech-to-text and text-to-speech on multiple platforms for DHH students inside and outside the classroom.

In the article (Lasecki et al., 2017) the researchers discuss how the Scribe system can combine human labour, ASR and machine intelligence in real time to reliably convert speech to text in short time, with less than 4 second latency. They discuss how collaborative captioning can reduce the cost and give better quality and availability. Human-powered caption services are expensive, not available on demand, and their captions may not match their consumers' reading speed and abilities. While ASR is cheaper and potentially always available, it generally produces unacceptable error rates, not only in most lecture environments, but also in many unconstrained real-world environments.

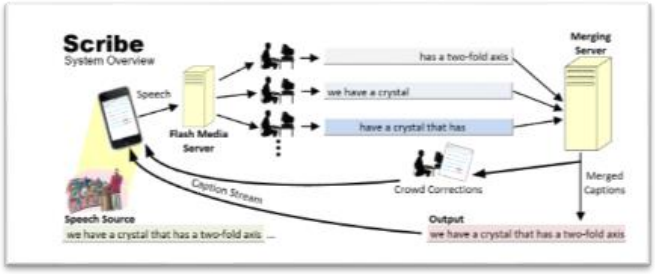


Figure 3 Example of how Scribe do collaborative captioning.

Scribe allows users to caption audio on their mobile device. The audio is sent to multiple amateur captioners who use Scribe's web-based interface to caption as much of the audio as they can in real-time. These partial captions are sent to their server to be merged into a final output stream, which is then forwarded back to the user's mobile device.

Ai-live.

It is challenging for anyone who is deaf or hard of hearing to fully participate in the verbal communications which are taking place every minute of every day. Ai-Live captioning provides access to spoken dialogue displayed on a screen and delivered in real-time. They use real people, not computers to convert speech into text to ensure the text matches what's being said. This all takes place using the internet to stream the words live to laptops, tablets and smartphones (Live, 2017).

Figure 4 Open caption OC by AI-media

8.5 Re-speaking

Although it can be expected that developments in ASR will continue to improve accuracy rates, the use of a human intermediary to improve accuracy through re-speaking the words into a high-end microphone and/or correcting mistakes in real time as they are made by the ASR software could sometimes help compensate for some of ASR's current limitations. Re-speaking could be an attempt to maintain high quality and low cost by using ASR as a starting point for human typists.

ASR is a desirable real-time captioning solution because of its low cost compared with human captioning services or sign language interpreters. ASR is not perfect but combined with human labour can make a better caption of speech on shorter time (Kawas et al., 2016).

Being a personal interpreter is challenging and very demanding. Missions over one hour often require several interpreters so they can get their measured break. This is making interpreter-services expensive. A cheaper alternative might be a crowd-powered speech transcription.

Speaker dependent ASR knows the speaker and is continuously learning and thereby gives better and better quality on the captions. Respeaking gives you a personal interpreter, who have a less demanding job and by use of cloud services, do not have to be at the same location as you.

8.6 Speech in Norwegian translated to another language.

ASR solution which also are able to translate the text to another language. Just think about your grandmother, at her holiday, traveling to Paris for her first time in her life.

In the report (Le, Lecouteux, & Besacier, 2018) tells that automatic quality assessment of spoken language translation, also named confidence estimation ,is an important topic because it allows us to judge whether a system produces user-acceptable outputs or not. By using an interactive speech-to-speech translation, it helps to judge whether a translated term is uncertain, in which case we can ask the speaker to rephrase or repeat the term. For

speech-to-text applications, it may tell us whether output translations are worth correcting, or whether they require retranslation from scratch.

8.7 Lip reading.

Lipreading is the process of understanding and interpreting speech by observing a speaker's lip movements. By visually interpreting the movements of the lips, face and tongue when normal sound is not available. The variation in size and colour of different people's lips will result in different outputs. It relies also on information provided by the context, knowledge of the language, and any residual hearing.

Speech to sign (automatically) there are a lot of ongoing researches. Many reports have already been published this year. Some interesting reports titles might be; A review of hand gesture and sign language recognition techniques (Cheok, Omar, & Jaward, 2019) and Multiple Proposals for Continuous Arabic Sign Language Recognition (Hassan, Assaleh, & Shanableh, 2019).

The aim of report (Salik et al., 2019) is to create a model of how to reconstruct intelligible acoustic speech signals from silent videos from various poses of a person which their system Lipper has never seen before. By silent video feeds from multiple cameras recording a subject to generate intelligent speech of a speaker. Lipper, therefore, is a vocabulary and language agnostic, speaker independent and a near real-time model that deals with a variety of poses of a speaker.

The human language cannot be bound to a fixed set of words or languages, as they have to map lip movements directly to sound. *Some late work (Fenghour, Chen, & Xiao, 2019)* show how an existing deep learning architecture for automatically lip reading individuals can be adapted it so that it can be made speaker independent, and by doing so, improved accuracies can be achieved on a variety of different speakers.

8.8 Students

(Raja S. Kushalnagar et al., 2014) reports fewer DHH students complete their higher educations than hearing students. Equal access to communication is fundamental to the academic success of students. Captioning options are necessary for the DHH students, but are limited due to cost, availability, and quality concerns.

People tend to incorrectly assume that the captioner can capture the speech accurately. They also assume that deaf students can read the verbatim text and understand the full richness of the speech. That is the main issues with captions; speed and flow – the captions must be in sync with the speaker, but not faster than the reader can read. Quality is also important, a low error rate so that the reader is not derailed during the reading process.

Real-time captioning enables deaf and hard of hearing (DHH) people to follow classroom lectures and other aural speech by converting it into visual text with less than a five-second delay. Keeping the delay short allows end-users to follow and participate ((Raja S. Kushalnagar et al., 2014)

The adaptation is bad for hearing impaired children and young people in the Norwegian schools and kindergartens, these institutions fail in including hearing impaired(Kermit, 2018) has currently reviewed the existing research on how it is to be a hearing impaired in Nordic schools or kindergartens. Hearing-impaired children and young people sometimes must spend a lot of energy in trying to understand what is said in the classroom. This group of children and adolescents gets poorer results at school compared to their peers with normal hearing. They are lonelier and struggling to join the community. Some of the research that the report deals with is about universal design - or lack of this.

Microphone systems and individual technical aids are important for hearing impaired, there are a little expertise in using these aids. But it is not enough that the school or kindergarten has the aids. Those who work there must also have expertise in how to use them and understand the limitations of the technology. Many of those in charge of education have low

competence on the needs of hearing impaired for adaptation and facilitation, the report concludes.

Educational videos tend to be 'heavier' and more textual. Typically, the presenter uses slides, text or other structured visual materials in their presentations along with their narration. Presenters also often include a great deal of non-verbal contents, e.g., software demonstrations or experiment manipulations to illustrate a lecture. As a result, it becomes very difficult for viewers to watch both the video and the captions at the same time. With the advent and popularity of online education, especially MOOCs, it becomes even more imperative to offer adaptable and optimized captioning displays to give maximum benefit to the viewers.

8.9 ASR of lectures

Legislation requires that educational materials produced by staff should be accessible to disabled students (Wald 2005). Speech materials therefore require captioning, ASR provides the potential to make teaching accessible to all and to assist learners to manage and search online digital multimedia resources. The automatic provision of accessible synchronised lecture notes enables students to concentrate on learning. The potential of using ASR to provide automatic transcription of speech in higher education classrooms has been demonstrated by the Liberated Learning Initiative.

8.10 Ablecenter - Wirelessly sending images of the board for zooming.

Figure 5 AbleCenter mounted in the ceiling, using wireless communication with smart phones and laptops.

This solution focus on the visually impaired, but the technology is interesting. The main device is mounted on the ceiling and sends images wirelessly to the data, and the user do not need any own specific equipment, his ordinary laptop would do.

8.11 Internet of things (IoT).

In these times of internet, the Internet of things (IoT) are an important issue. There are more and more sensors and wireless connections everywhere. The technology should be there when you need it and you should preferably not need any special equipment. Being dependent of special equipment is stigmatizing and often making the users feeling discriminated. Being DHH is not necessarily your only handicap, you could be blind or sitting in a wheelchair too. If you go to the cinema and the hearing-impaired are placed in front of the audience and the wheelchairs in special place in the back and you have both handicaps - it is not very including.

8.12 voice-to-text recognition software.

Quality costs! Free ASR products vs professionally products.

There are some free out there already. Many products have integrated automatic speech recognition as Apple's Siri, Andriod's woman and Microsoft Office. Most of us have already spoken with Siri. You do not have to pay extra for these.

More costly ASR are product as Dragon (not in Norwegian) and Tuva from Max Manus (only in Norwegian). It is not quite free, but the new version of Office 365 is by using automated speech recognition to create text of your speech, it is even possible to translate your speech in real time thinking of teachers teaching foreign pupils.

Free good ASR solutions - there are some out there already, most of us have already spoken with Siri. Many products have integrated automatic speech recognition as Apple's Siri, Andriod's woman and Microsoft Office. You do not have to pay extra for these.

A lot of work is ongoing. The newest version of Microsoft Office has ASR included, Google has just released a beta version of an app for the DHH called Google transcribe, this app is used as the prototype in this report.

8.13 Available products for live transcribes

The Google accessibility service have just released an app for the DHH. This is just an early release (February 2019), only available for Android phones 5.0 and later. It supports transcription in over 70 languages and dialects. Bilingual support with quick switch between two languages. Transcribe only works with an internet connection, to access the cloud-based transcription service. The immediate response in Norwegian media[2] was positive.

The Live Transcribe offer text representations of spoken conversations as they're happening, while Sound Amplifier has a dynamic audio processing effect to make speech and other

[2] https://www.dinside.no/mobil/tale-til-tekst-pa-null-komma-svisj/70737787

sounds easier to hear. Transcribe performs real-time transcription of speech to text, you can see words appear on your phone as they are soon as they are spoken, and it helps with everyday conversations going on in the world around you.

It can be very challenging and expensive for deaf people to arrange for real-time transcriptions of conversations, both in personal environments and professional settings. These tools give people an easier and cheaper access to this kind of technology. Someone can't afford adding even a small amount to the device they already have. Meanwhile, smartphones are becoming so sophisticated, and are ubiquitous." Google's approach with its new apps "is a pretty smart idea." Standard hearing aids are often relying on button cell batteries and don't have the same kind of processing power as smartphones.

Apple, Google's only real competitor in mobile operating systems, has rolled out its own array of impressive features aimed at the blind and deaf community, and most are built directly into Apple's operating systems.

9 FORMAL REQUIREMENTS

9.1 Formal Requirements - WCAG

Web Content Accessibility Guidelines (WCAG) developed by World Wide Web Consortium (W3C) provides a technical standard for web content accessibility ((W3C, 2018a).

WCAG 2.1 consists of technology-neutral principles, guidelines, and success criteria that reflect properties of Web content, these guidelines and success criteria can be satisfied in a variety of ways. As the Web evolves, the guidelines can assist technology developers and authors in ensuring that Web content becomes more usable to users in general.

The WCAG guidelines are organized under 4 principles: **perceivable**, **operable**, **understandable**, and **robust**. In this report the main focus is on perceivable for the DHH. All audio must be presentable to users in ways they can perceive it, also when using assistive technologies - make it easier for users to see and hear content.

As WCAG (W3C, 2018a) states for each guideline, there are success criteria to be tested, which consist of three levels: A - the lowest, AA, and AAA. In terms of captioning: 1.2.2 Captions (pre-recorded): Captions are provided for all pre-recorded audio content in synchronized media, except when the media is a media alternative for text and is clearly labelled as such. (Level A).

Other requirements:

1.1.1 Non-text Content.

1.2.1 Audio-only and Video-only (Pre-recorded).

1.4.2 Audio Control.

1.4.7 Low or No Background Audio.

2.1.1 Keyboard.

2.4.7 Focus Visible (Level AA).

9.2 The Norwegian requirements

~~Formal requirements - WCAG~~The main rule in Norway is that all ICT solutions should be universally designed.

This applies to websites and automats. Both private and public works, teams and organizations must comply with the regulations. The Norwegian are regulated by the laws; The Gender Equality Act and the Prohibition of Discrimination (Equality and Discrimination Act) and Regulation on universal design of information and communication technology (ICT) solutions. These legislations are based on the international guidelines WCAG (W3C 2008). Some more work in EU will affect our regulation[3] .

New ICT solutions had to be universally designed after 1 July 2014. All ICT solutions, including existing solutions, must follow the requirements for universal design before 1 January 2021. It is a requirement that ICT solution is a head/main release, in addition to being straight against general. If it is not a main discharge, it does not require universal design.

In the prescriptions, the main solution is defined as follows:

> *ICT solutions that are an integral part of the way the company informs and offers its services to the general public and which is linked to the company's general function.*

The Directorate for Administration and ICT (DIFI) supervises that the requirement is followed. Information and guidance will be among the most important tasks for the audit.

[3] ttps://uu.difi.no/nyhet/2018/09/eus-webdirektiv-blir-en-del-av-norsk-regelverk

The Ministry of Local Government and Modernization (KMD) is the appeal body for decisions that are taken by the audit.

The most common faults they find in addition to the lack of adequate subtitling of videos and keyboard control challenges are poor contrast between text and background, error messages in forms and the possibility to enlarge content.

10 RESULT AND FINDINGS.

10.1 Interview Findings

Do you know what ASR is?

The interviewed group had hardly ever heard about ASR, but when I used Siri as an example most of them nodded recognizable. The quality is affected by the (un)known speaker, dictation ability, and cost. The quality of the ASR captioned text had to be of acceptable quality, that's important one said. The possibility to read the captions and correct the text if needed was also an issue.

Both button and speech commands are needed to control the app. As some of them said; I can't hear the speech nor see the captions or any button! You need the possibility to read it by using a braille.

How to make audio available to the DHH?

By use of a personal interpreter who interprets into sign language, by captioning of films and television or simultaneous captioning the speech in real time done by humans. None in the interviewed group was borne deaf and didn't have sign language as their native language, so for them captioning was the only alternative.

They had never heard about simultaneous captioning done by ASR before.

DHH

In which situations?

One of the elderlies in the group told that she never had thought about the need of captions or an interpreter. If not born deaf, captioning is probably better than using sign. But as one of the elderlies in the interview group said: I don't see well, nor do I hear well, but I will never be deaf-blind – captioning is not for me.

Privacy?

At the doctor, a third person is present? The group's responses were divided - one in my family goes with me. Another one said - I need the possibility for privacy. I can use a publicly authorized interpreter and they have confidentiality.

In more formal meetings. Correctness?

Correctness matter! I want good quality on the captioned text.

At work. Requires prior knowledge of interpreter.

Talk to me, I am the one who has the knowledge and not the interpreter, he only does the speaking.

At school, lectures. Being a student.

None in the interview group have recently been students.

Socialization

This is an important aspect. A lower entrance threshold and something useful for most people. Everybody has heard about it. Low-threshold offers might even be better in some situations than a personal interpreter.

Maybe an app which is useful for hearings too - you can dictate shopping lists, to-do lists and make small notes. Generally known app among most people makes talking to a DHH less intimidating.

The world must also accept the technical development. This could be a social problem. Technology may exclude people.

Incremental steps?

Incremental steps, rather than a big one. Incremental steps provide a faster results and corrections along the way would be possible. Nice said some, getting started with it, socializing matters. But the captions have to be correct was a demand from others.

The world is full of gadgets and apps.

It would be been nice to use the mobile phone as much as possible; I even have an induction loop on my smartphone.

The elderly was more sceptic – I don't have a smartphone; they are too difficult to use; I am old and visually impaired.

Costs. What does it mean to you?

When we talk about me as DHH NAV is involved they will pay. Free means an app available for everybody.

NON-DHH

Do you know how to communicate with the DHHs?

None of the candidates had thought much of the DHH at all. There were no DHHs in their surroundings at all, just some elderly who needed louder audio and voices. The answers were depending on whether they had a communication issue with some DHH close related to them or not. As one lady said – "*Give a small touch to give the DHH a sign. Stand in front of him, with your faces against each other. Talk loud and clear, and do not talk too fast."*

To understand the DHH might be more problematic. If lip-reading is not possible, then the best alternative must be communicating by pen and paper, or maybe use iPhone's Siri as one suggested.

Costs. What does it mean to you? Buy a non-costly app?

We discussed public vs private costs. Use it to dictate, for your own need? To be prepared? A useful app also for the hearings too? The interviewed group had not even thought about it.

If it is were very expensive – NAV had to be involved, but if I need it, I would happily pay for a not too expensive app. A low-threshold offers could be important, it would be easy access for everybody. But as also was said - if a cost, even if it is symbolic price, I have to really have a need for it. If anything, trigger the need - I would buy it.

Quality is affected by the (un)known speaker, dictation ability, and cost.

I would use it for a simple text. I want the ability to correct the text too. A dictating tool must be easy to use. Both buttons and speech commands are necessary. A beginner's tutorial is needed. I do believe in such an app.

The interviewed group prefer good quality. If training makes my results of the dictation better, I would do the training.

If I could trust the results from the app, if I can check the text result; I would feel more comfortable. I am not able to check what the interpreter are signing…

Training

Quality is affected by the (un)known speaker and dictation ability. One of the men said he happily would do all the training he had to if he had to regularly communicate with DHHs, but not do the training just for being prepared f to be able to communicate with them some time.

This looks like a complex tool to use; some training is needed – I would prefer a tutorial.

Socialization is an important aspect.

Something for most people, I think it would be nice if everybody had heard about it. Something for all, easily to learn how to use, People will get better at dictating. Commonly

known among people makes life easier. The hearing people would just know have to do it, have to communicate with DHHs.

Another gadget?

Easily accessible is important. Something for all of us, easy to learn how to use, people get better at dictating. I prefer an app.

10.2 Prototype findings - LIVE TRANSCRIBE

Google was the first one to release a prototype of this concept, speech to text using an ordinary smartphone - and it is free to use. The main response was overwhelming - "Wow - it is so exciting!", "Spread the good news!" But some were also quite unsure about the quality of the recognized speech.

The newspaper Dagbladet[4] wrote after their first test that they were very impressed by how quickly it converted the speech to text. They tested by playing a radio broadcast, and the text on the screen was not many seconds after what is said. It is not 100 percent accurate, but still good enough for you to get the essence of what is said.

If you want, you can also make the phone vibrate a little when something is said after a silent break.

One of the reports interview objects had her own business. She is completely deaf and runs her own workshop, where she meets customers daily. All communication with customers is done by using pen and paper. She became very positive when I first talked about the transcribing app. This could make her customer services much easier.

Some statements found on Google's own material on the web:

[4] https://www.dinside.no/mobil/tale-til-tekst-pa-null-komma-svisj/70737787

- “We can now do things that weren't even remotely possible a few years ago, like jump into conversations at the lunch table or casually join in when the opportunity arises.”
- It really would change the way I solve my communication problem.

HLF had a presentation on Facebook of Goole Live Transcribe[5] and immediately got many happy responses from their members.

> *“Spread the good news!”*
>
> *“THIS one we do have to test more!”*
>
> *“This one looks actually very good! I didn't need as so much imagination as in an "ordinary" conversation, it is easier to understand what have been said. “*
>
> *“Does it translate into Norwegian from English?”*
>
> *“Maybe if you talk one at a time. Tried it in the lunch break at work today, it didn't work.”*
>
> *“It goes smoothly when there is not much background noise. I tested it in lunch break at work today. “*
>
> *“It is so exciting!”*
>
> *“You might send message by talking it in, but not if you speak a dialect (iPhone).”*

[5] https://www.facebook.com/hlfhorselshemmedeslandsforbund/videos/22336242268936

Some transcription tests were run.

The quality of the transcription was not impressive at all. It was tried in a simple situation too, and might work well with some goodwill.

Figure 6 A not perfect example of transcription at the hairdresser - booking a new hour.

In more complex settings it was hardly possible to understand the transcription at all.

Figure 7 Transcription of the TV2s news at the television.

Figure 8 A small conversation between to persons in their living rom.

The author's own comments.

My first impression was – this is quite impressive, but a very early release… Being a visually impaired person myself, I have been looking at this app thinking of people who are both DHH and visually impaired – their life is quite challenging. What about the quality and correctness of the text? The recognized text is changing. The app is constantly correcting the already recognized speech (what have most likely been said - based on statistics). I may perceive that something is happening in the screen, but not what happens. If, on the other hand, the text is read with a screen reader or braille display, it will be almost impossible to see these constant changes.

The screen size on a mobile is quite small especially if you are dependent of zooming the text, a pad might be better. To have some easy techniques as speech commands, to set punctuation and make a new paragraph would have been nice – it makes a long text easier to read. When you are talking to more than one person and you can't either see or hear who is speaking, it would have been quite nice if the app had marked the recognized speech by the direction of it or an alias. In the report (Berke & Caulfield, 2017) they used colour to

mark text, which was not an exact recognition of the speech, colour coding would have been awful or me to read. The possibility to easily correct the recognized words might have been a nice functionality to have too.

11 DISCUSSION

Most people I talked to during the work with this report had not heard of ASR technology at all before, and only a few of them had ever heard about simultaneous captioning for the DHH done by ASR. However, when I used Siri as an example most of them nodded in recognition.

The report is based on a qualitative survey by doing some interviews of available DHHs. How does error in the automated captioned text affect the user experience of the caption? This is an important question, to do a good evaluation of the quality and experience of the captioned text. In this master thesis, the task was considered as a too big task to include to.

What are captions? What is the difference between captions and subtitles? **"**Captions" commonly refer to on-screen text specifically designed for hearing impaired viewers, while "subtitles" are straight transcriptions or translations to another language of the dialogue. Usually captions are positioned below the person who is speaking, and they have to include descriptions of sounds such as gunshots, closing doors or music. Editing the original transcription may be necessary to provide time for the caption to be possible to read and for it to be in synchronization with the audio. Creating good captioning of a film is more than just recognising the speech as ASR will do for you. Some extra work has to be done, as captions made by ASR are only just subtitles unless they are evaluated and edited by some humans afterwards. Probably it is more correct to say transcribing by ASR and not captioning by ASR!

Remember there are some caption standards, captions should be **ACCRE – Accurate, Consists, Clear, Readable** and **Equal** ((DCMP, 2018). You must be honest and trustworthy as a captioner. Caption the speech as it has been said (use spoken dialect and language) and do not forget the audio and visual effects. Include as much of the original language as possible.

Today it is even more important to be independent and to be able to manage things by yourself; it is what the society expect of you. Automats for buying groceries in the shop is actually easier as a DHH than using a shopkeeper who talks to you. To be able to manage minor daily activities yourself without having an interpreter around by using captioning - as taking the bus, visit the café, do some shopping or just talk to your neighbour about the weather.

People often have more than one diversity, being DHH might not be your only challenge. As our world's population are getting elder, more and more people will have both a low vision, a bad hearing and cognitive challenges. If you cannot see or hear the captions, you need a braille to be able to read them.

The captioned text is often continuously changing as long as you continue to speak, because the ASR systems are based on statistics, and continuously trying to get a better caption as more as you speak. The changing text is impossible to read if you are visually impaired or struggling with cognitive challenges it is also quite problematic. Marking words in the text to show which of the words who has an unsecure recognition might give a better understanding of the quality of the captioned text, but are difficult to catch for the same reason as mention above.

By having an app for smartphones, the transcribing solution is easily accessible for everybody. Being free or having a low cost makes it even easier. If it were some functionality in the app, which could be useful for the non-DHH as well, people might feel more comfortable using transcribing of speech to communicate with the DHH. Maybe some DHH would feel less lonely then and hopefully not that often feel that people were avoiding communicating with them.

How can easy and unplanned integration without an interpreter be more possible, to be present and available in your social surroundings? We are talking about an average of maybe 2x10 min of pure socializing during a day outside your closest relationships, at work or at

school. Many of those who had tested Live Transcribe tried it in the lunch break, which should be an indication of where they would like to use such an app.

As one of the elderlies I interviewed said “I manage pretty well, just hear a little badly”. It might be difficult for some to accept their need of some extra assistance to hear, some do not even know it could be possible to get an assistive aid.

Being DHH at an hospital or when you are getting older and need to be at a care for elderly, and you are not able to hear the nurses’ speech. Getting a personal interpreter might be difficult, but access to a solution of easy Captioning might help. If the nurses had the solution on their own device or smartphone, the elderly did not have to be afraid of the technology either.

Stigmatization of the DHH is also an important issue to be aware of. As one DHH said: “Talk to me, I am the one who has the knowledge, not the interpreter, he only does the talking”. To have a low vision and having to use a white stick is quite difficult for some to accept, they do not want their diversity to be visible to others. I guess many DHH feel the same way as well, if they have to use fancy glasses, a helmet, special gloves or some other device you have to put in ad off. (Schipper & Brinkman, 2017) (Salik et al., 2019).

User experience designers and developer should incorporate accessibility to make their designs work better for more people in more situations. Addressing accessibility, usability, and inclusion together more effectively lead to a more accessible, usable, and inclusive web for everyone[6]. Remember to think about the universal design already in the design phase.

[6] https://www.w3.org/WAI/fundamentals/accessibility-usability-inclusion/

This is something the marketing people like to talk about – better Search Engines Optimizing (SEO)! Captions gives the pages better SEO; pure audio is not easy searchable.

A personal interpreter will give the best quality of the interpretation, but the cost of and availability of interpreters are an important issue. A lot have been done already as the possibility to have sign interpreting by is just a video call away. For simultaneous captions, maybe a mix of human and ASR would give a cheaper solution of acceptable quality.

The researches (Le, Lecouteux, & Besacier, 2018) want us to do a confidence estimation , because it allows us to know whether a system produces user-acceptable outputs or not. For speech-to-text applications, it may tell us whether output translations are worth correcting, or whether they require retranslation from scratch.

An easier first solution might be to let the two parts in the transcribing process, both the speaker and receiver, to confirm the correctness and understanding of the transcribed speech.

The DHH need a haptic feedback to be aware of audio changes in their surroundings. A lot of research is ongoing about how to give feedback as an aid– a watch, a glove, a jacket. A lack of acceptance emerged from the discussion of capabilities and limitations of haptic assistive technologies. (Sorgini, Caliò, Carrozza, & Oddo, 2018). Some of the suggested equipment's are quite bulky, but they might be a good start.

(Soviak et al., 2016) tells that specialized haptic devices are very limited and/or are exuberantly expensive and bulky. In this paper, they describe a low-cost haptic-glove system, which can potentially enable usable tactile interaction with GUIs. The vision is to enable blind userst to connect it to any computer or smartphone, and it might be something for the deaf as well.

Haptic feedback means to get the information presented in the form of static or dynamic stimulation of the skin. Sensory substitution systems for visual and hearing disabilities are enabling users to interact with the not disabled community in daily activities.

The reviewed literature in the report (Sorgini et al., 2018) provides evidences that sensory substitution aids are able to mitigate in part of the deficits in language learning, communication and navigation for deaf, blind and deaf–blind individuals. The tactile sense can be a means of communication to provide some kind of information to disabled individuals and they have analysed haptic sensory substitution technologies. Evidences shows that sensory substitution aids are able to mitigate some of the deficits in language learning, communication and navigation for deaf, blind and deaf–blind individuals. The researchers say future researches shall go towards miniaturized, custom-designed and low-cost haptic interfaces and integration with personal devices such as smartphones for a major diffusion of sensory aids among disabled. The development of miniaturized, custom-designed and low-cost aids operating in synergy in networks, aiming to increase patients' and the public acceptability of these technologies.

The search for reports at ORIA about "haptic feedback" and "smartphone" gives quite many hits in the last to year indicating a lot of have been done on the topic.

A cheap and easily available ASR solution is important to make people more comfortable with dictating, which gives a better captioned result, with less errors. The demand for quality on the captions is depending on the situation, how formal it is, whether it is planned or not. In an informal situation, easy access is often more important. "The quality might even be better than the guessing I usually do." Communication between two persons might be the easiest to caption. In more formal situation, quality matter! In a meeting, many to many are talking is challenging – who is speaking, you need the haptic signing as well. In more formal official meeting sign or live captioning are common.

How is it possible for a DHH to talk with a non-DHH? Depending on how long you have been DHH you might be able to read lips, otherwise pen and paper will do, or you might be happy to have your personal interpreter nearby. It is also possible to get interpreting to sign by taking a video call to a pool of interpreters. As a DHH craftsman said – "access to easy captioning would be absolutely brilliant".

11.1 The app

This report has been under progress for the last three years and it seems like my original idea for a prototype was quite good, after all as both Google and Microsoft have done something similar. Google has released their first version of an ASR app and MS Office 365 has integrated both ASR and translation in their laptop office tools. Google's app was very exciting and many DHH were fascinated and found it very interesting – they thought it actually could help them.

Some simple tests of the app were run, for some examples see 10.2 Some transcription tests were run.

11.2 One more gadget?

Do you really want an extra gadget to carry around, to put on and off, to look after or just another app on your smartphone? Some users might even think an app on a smartphone is too difficult to use. Maybe the best solution is to have both alternatives.

By having an app for smartphone, the caption solution is easily accessible for everybody. Being free or having a low cost makes it even easier. If it were some functionality in the app, which could be useful for the non-DHH as well, might make people feel more comfortable using captioning to speak with the DHH. Maybe some DHH would feel less lonely then and hopefully not often feel that people were avoiding communicating with them.

A personal interpreter will give the best quality of the interpreting, but the cost of and availability of interpreters are an important issue. A lot have been done already, as the possibility to have sign interpreting is just a video call away.

For simultaneous captions, maybe a mix of human and ASR would give a cheaper solution of acceptable quality.

Maybe some more advanced settings in the app would give better accessibility too.

The possibility to use the app by voice commands is important too, you might have a motoric impairment, your fingers are busy doing something else or you just have wet or dirty fingers.

DHHs have a bad pronunciation or might even be dumb. They communicate – maybe by writing; thereby the app would need the possibility to write text. Maybe the possibility to pre code some handy sentences or questions would have been nice too?

Maybe it would have been nice to have some settings to turn on and off, a more complex functionality in the app, as mark-ups and word-correction facility.

The possibility to set:

- A bigger screen size would be nice – maybe to choose between smartphone or pad.
- Maybe have possibility to set a delay on the captioned text to prevent the continuously changing text.
- Mark-up on/off or use of different mark-up stiles.
- Correction the transcribe words done by selecting from a list or by writing if wanted.

11.3 Translation

If the starting text before translation is a result of captioning done by ASR it would probably give an unacceptable quality of the result which probably not is good enough to use in even informal situations in your daily life at all. Software in captioning lectures showed 75% accuracy on untrained ASR software and could reach 90% under ideal single speaker conditions 80% accuracy of an ASR transcription and 75% accuracy of the translation of it, gives you less than 60% accuracy on the result – which is not good quality at all. Maybe it could be used on simple sentences using simple formulations with a better result.

11.4 Captured ASR text with confidence mark up

ASR has recent advances, and two studies conducted by (Berke & Caulfield, 2017) have tried by comparing various methods to visually presenting the ASR output with certainty values - how ASR captured text could be shown with confidence display mark-up.

Users preferred captioning styles with which they were already most familiar with, which did not display confidence information at all. While the participants expressed interest in systems that display word confidence during captions, they were concerned that text appearance changes may be distracting. The users were concerned about the accuracy of ASR systems. Among those that shared negative experiences, a common theme that arose was that they were dissatisfied with ASR accuracy and they felt that it was too frustrating to follow the captioning.

Among those that shared negative experiences, a common theme that arose was that they were dissatisfied with ASR accuracy and they felt that it was too frustrating to follow the captioning. As some said - “It becomes really annoying!”, if you are having some kind of cognitive challenges, the situation will be even worse – not usable at all for some.

If needing a braille and having a conjunctive impairment as well – it is getting even more difficult. Being visually impaired and the use of visually mark-up does not sound good at all. However, how visually are the text, when you are reading it with a braille or a screenreader?

11.5 Limitations in this reports research.

The planned questionnaire in this report was not run. I did not have access to enough mail addresses of potential respondents among the DHHs, or support from HLF for the dissemination of the questionnaire.

12 CONCLUSION

Hørselshemmedes landsforbund (HLF) keep calling for more pressure on universal design. The hearing-impaired want to participate in society as everyone else.

A lot have already been done, but we are by no means finished yet. At the beginning of March 2018, HLF had to cancel its Central Board meeting because NAV was unable to write with interpreters. This is an example of hearing-impaired rights are being ignored and emphasizes why we need more pressure in this field. The right to a representative is enshrined in the National Insurance Act and Norway has through the UN's resolution on human rights for people with disabilities committed themselves to ensuring full participation for all in political and public life.

The possibility for full participation in social, public life is depending of a cooperation with the non-DHH, so the knowledge and ability to use ASR among the non DHH can facilitate better transcriptions. A cheap and easily accessible app which also have some useable functionality for everybody would be nice to have.

Maybe the best way is to take many small, but quick incremental steps instead of one big slow step towards a “perfect” solution.

The quality gap between transcription done by a human interpreter an ASR of a meeting, might be quite huge. Even the Norwegian parliament violates its own law of universal design, subscriptions of parliamentary debates are still not accessible.

The current ASR quality is not perfect, and may not be for many years either. Waiting for the perfect result might be a never-ending story, as one of the interviewed persons said; the possibility for participation in social life is more important than the perfect caption. A middle way to go might be to have some kind of approval of the transcription of the parts in the conversation – the transcribed text is correct and understood. In an informal setting it might

be enough that the parts know about it and nod to show the approval. In a formal setting – saving the mutually approved conversation might be correct to do.

Being both DHH and visual impaired is another big challenge, or maybe having a cognitive impairment too. To be able to read the text with a reading list must of course be possible. But if both parts are familiar to the transcription process using a transcription app it gets easier. Shorter and simpler sentences might help many. Avoiding the continuously changing text might help lot too, for example by pre-setting some seconds delay for the text.

Transcription of lectures might help all the students, by helping them to keep focus on the teacher and what he is saying instead of using all their energy on writing notes.

13 FUTURE WORK

Universal design has become a very important aspect of developing new solutions, and the companies must follow the regulations. The attention is high, the CEO of Microsoft has even called accessible technology a “human right”.

The mobile-technology allows a quicker and cheaper access to communication for the DHH, even less stigmatizing as opposed to being dependent on the availability of an interpreter.

The speech-recognition technology is becoming more and more commonly used in everyday interactions. Make the ASR technology more known among people by making it more commonly used. Then hearing people would know what it is and how to use it and more comfortable about using it.

With accents, talking speeds, and other vocal variables voice-recognition technology is not always perfect and the ability to write text is necessary too.

As said in (Sorgini et al., 2018) future researches shall go towards miniaturized, custom-designed and low-cost haptic interfaces and integration with personal devices such as smartphones for a major diffusion of sensory aids among disabled.

The quality of ASR caption is not always perfect. Does it have to be nearly perfect or would it help with the possibility to correct the captions?

In the (Berke & Caulfield, 2017)they tried to mark words in the captioned text of the speech which was not exactly recognized. The use colour coding might be awfully for some to read. The concept of marking insecurity of words with colours or font does not sound good at all for me as visually impaired. Some people with cognitive challenges as many elderlies get would probably have the same feeling too. Maybe the possibility to turn the mark-up on and off would help.

The possibility to easily correct the recognized captioned text might have been a nice functionality to have too. Dragon and Tuva does allow for voice to text “corrections, but their current format seems a bit too complicated for a mobile interface. Functions similar to the autocorrect feature as found on smartphones applications could be used to give a better accessibility. Instead of using ordinary keyboard typing, it could provide an alternate replacement words based on sound or simple sweep over the screen with a finger or two.

It is easy to imagine in a not-so-distant future smartphone apps will be increasingly aware of a person’s needs and becoming self-adjusting. The smartphones already understand where you are and if you set some noise cancelling parameters - the next time you go there, the app would know. This is not exactly rocket science, but it still needs to be executed well. Just imagine you are at the same place, but last time you were sitting in the bus, this time you are standing on the street.

The researchers and developers are trying to use the newest state of the art technology that we already have. Sometimes they just try to make a new system, but some users can’t even afford a small cost to upgrade the device they already have to a newer version or even less to buy a new device. Meanwhile, smartphones are becoming very sophisticated and are ubiquitous.

Ordinary hearing aids often rely in button cell batteries and don’t have the same kind of processing power as a smart phone, they have less processing power and there is a continuously big focus on development of mobile technology.

14 APPENDICES

14.1 APPENDIX A - DEFINITIONS

ASR: Automatic Speech Recognition.

ASR captions: Captions of speech done by automatic speech recognition.

Audio Description: While most people are familiar with closed captioning, audio description is another important accessibility requirement for video content. Audio description narrates the relevant visual information in a video to make your content accessible to blind and low vision users.

Automatic Speech Recognition (ASR): Automatic converting speech to text. This methodology is based on an acoustic model, a dictionary and a huge statistic of what people usually say. Most recently, the field has benefited from advances in deep learning and big data. Works well in ideal situations with high-quality audio equipment, but degrades quickly in real-world settings. ASR is speaker dependent, has difficulty recognizing domain-specific jargon, and adapts poorly to changes, such as when the speaker has a cold.

Braille: It is a tactile writing system used by people who are visually impaired. It is traditionally written with embossed paper. Each character has a rectangular block, called cells, which have tiny bumps called raised dots. The number and arrangement of these dots distinguish one character from another.

Braille embosser: You can type braille with a braille writer, such as a portable braille note taker or computer that prints with a braille embosser.

Captions: Aim to describe to the deaf and hard of hearing all significant audio content - spoken dialogue and non-speech information such as the identity of speakers and, occasionally, their manner of speaking - along with any significant music or sound effects using words or symbols.

CART: Computer-Aided Real-time Transcription.

Closed caption: The term "closed" (versus "open") indicates that the captions are not visible until activated by the viewer, usually via the remote control or menu option.

collaborative captions: Done by humans. Humans tend to replace words with ones that have similar meanings.

Computer-Aided Real-time Transcription (CART): CART is a reliable real-time captioning service, but is also expensive. Trained stenographers' type in shorthand on a "steno" keyboard that maps multiple key presses to phonemes that are expanded to verbatim text. Stenography requires 2-3 years of training to consistently keep up with natural speaking rates that average 141 words per minute (WPM) and can reach 231 WPM.

DHH: Deaf or Hard of Hearing.

DIFI: Direktoratet for forvaltning og IKT. The Norwegian Directorate for Management and ICT.

Haptic feedback: is the use of touch to communicate with users. Most people are familiar with the vibration in a mobile phone or the rumble in a game controller, but if you are both blind and deaf haptic touch can be used to tell you who is talking and if the person is laughing or is angry.

HLF: Hørselshemmedes Landsforbund, The Norwegian Hearing-handicapped country federation

Lip reading: It is a technique of understanding speech by visually interpreting the movements of the lips, face and tongue when normal sound is not available. It relies also on information provided by the context, knowledge of the language, and any residual hearing.

Non-Verbatim Captioning: computer-based macro expansion services like CPrint.

Open caption: "Open" captions are visible on screen to all viewers. (Vs. Closed).

Respeaker: a trained human who repeats the running commentary (with careful enunciation and some simplification and mark-up for input to the automated text generation system.

Re-speaking: In re-speaking, a person listens to the speech and enunciates clearly into a high-quality microphone. Could help the ASR to produce captions with high accuracy. Requires extensive training, since simultaneous speaking and listening is challenging.

Sign language: It is a language which chiefly uses manual communication to convey meaning. This can involve simultaneously combining hand shapes, movement and orientation of the hands, arms or body, and facial expressions to convey a speaker's ideas. Sign languages depend on the respectively spoken language and will differ from country to country.

Speaker dependent ASR: As TUVA. The system analyses the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that use training is called "speaker dependent". The start quality in the text result is good from the speaker dependent system and it will not give the best result the first time, but it gets better and better.

Speaker independent ASR: As Dragon, Google and Iphone's Siri. Systems that do not use training are called "speaker independent" systems. The start quality in the text result is best from the speaker independent system and it will give the same good results each time.

Subtitles: A textual version of a film or television program's dialogue that appears onscreen. "Subtitles" assume the viewer can hear but cannot understand the language or accent, or the speech is not entirely clear, so they transcribe only dialogue and some on-screen text.

Visual description: Description of the visual impression of the film as an angry face or a beautiful woman. Description for a visual impaired person.

14.2 APPENDIX B - ACRONYMS

DHH: Deaf or hard of hearing

DIFI: (Direktoratet for forvaltning og IKT) Directorate for Administration and ICT.

HLF: (Hørselshemmedes landsforening) The Norwegian association for the hard of hearing.

KMD: (Kommunal- og moderniseringsdepartementet) Ministry of Local Government and Modernization.

SEO: Search Engines Optimizing.

W3C: World Wide Web Consortium.

WCAG: Web Content Accessibility Guidelines. Web Content Accessibility Guidelines (WCAG) is developed by World Wide Web Consortium (W3C), and it provides a technical standard for web content accessibility.

WER: Word error rate.

WPM: Words per Minute.

14.3 APPENDIX C - WHAT IS AUTOMATIC SPEECH RECOGNITION (ASR)?

ASR. This methodology is based on an acoustic model, a dictionary and a huge statistic database of what people usually say. The speech industry players include Google, Microsoft, IBM, Baidu, Apple, Amazon, Nuance, SoundHound, IflyTek, CDAC, some systems are free of charge and some quite costly. Most recently, the field has benefited from advances of deep learning and big data.

The system analyses the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy by learning. Systems that do not use training are called "speaker independent" systems. Systems that use training is called "speaker dependent". The start quality in the text result is best from the speaker independent system and it will give the same good results each time, but the speaker dependent systems are giving better and better quality each time.

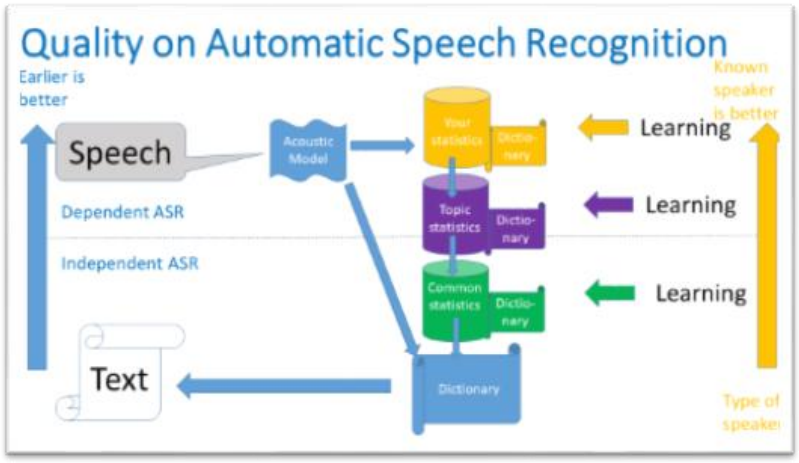


Figure 8 Quality on Automatic speech recognition are multi levelled.

Usually it is talked of a dependent or independent speaker. Actually, there is third variance too – the ASR system do not know the speaker, but the speaker knows how to dictate an

ASR system. The text result from a good speaker would be something in between the dependent or independent speaker's results. You get better quality on the ASR by training, letting the system know the speaker and giving continuously better dictionary and statistics.

The performance of speech recognition systems is usually evaluated in terms of accuracy and speed. The accuracy is usually rated with word error rate and real time used before you got the answer.

ASR Quality are sensible for the environment, the sound surroundings. The speaker's pronation will affect the result. Bad pronunciation because you have some illness or you are a foreigner makes it nearly impossible to use ASR. The microphones quality will also affect the results.

Usefulness of the captioning depends heavily on the error rate.

ASR is inexpensive and available on-demand, but its accuracy is not good enough, sometimes it drops below 50% when it is not trained on the speaker, captioning multiple speakers, or when not using a high-quality microphone. Software used to assist real-time captionists may often make errors that can change the meaning of the original speech. As DHH people use context to compensate for errors, they often have trouble following the speaker.

14.4 APPENDIX D - THIS RESEARCH'S PROTOTYPE – A MOBILE APP

Screenprints

Figure 9 A prototype of an ASR app for your smartphone

Extras functionality

- Save you dictated text and
  - send it as a mail
  - Save it in the cloud.
- Create a group conversation – a meeting, by connecting some mobiles by Bluetooth.

  Then it will be possible to identify the origin of the speech – who said it.

- Even translation of captioned speech to another language would be a possible feature.

14.5 APPENDIX E - GOOGLE TRANSCRIBE – FIRST RELEASE OF A NEW MOBILE ASR APP

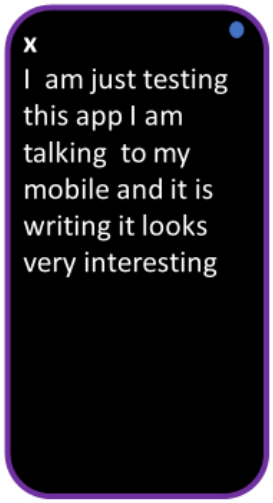


Figure 10 Screen-print of Google Transcribe

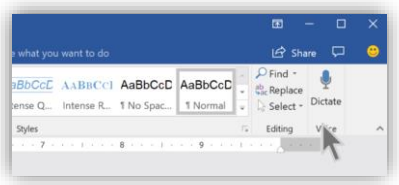
Although it's pretty straightforward, Google Live Transcribe does have some nice features that make it a little easier to use in the moment. A blue circle in the corner pulses slightly to show the ambient noise level so that you can visually see if you need to move the microphone closer to the speaker for it to work. It is also hit to bring up a keyboard to type out replies if you need to. Finally, if somebody starts speaking after a period of silence, Live Transcribe will give you a haptic signal, a vibrate in the phone to let you know to look at it to see what is being said.

Google says that keeping it simple and easy-to-use was a major goal and one of the reasons it chose not to include the option to save transcriptions. Privacy concerns may also have factored in. Google noted that, in addition to not saving transcripts, it's also not storing

audio or transcriptions on its servers, nor is it using any of that data to improve its algorithms. Supports multiple and external microphones, Google has also launched an Android sound engineer, which can then act as a kind of hearing aid by amplifying the sounds around you, where you can also choose to filter out background noise.

14.6 APPENDIX F - OFFICE 365 IN THE 2019 VERSION

For speech recognition within Word, Outlook, and PowerPoint, you must buy an Office 365 subscription, which includes Dictation.



""

Figure 11 Office 365, click to start dictating

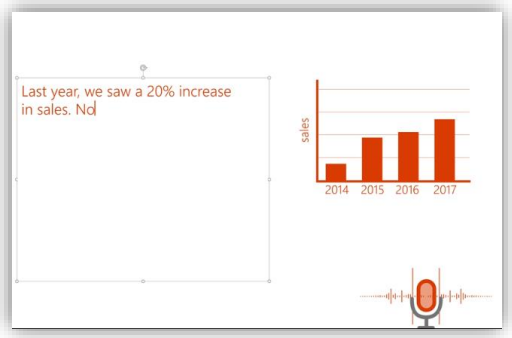


Figure 12 Office 365, dictating in Word, a box for speech to text

14.7 APPENDIX G - QUESTIONNAIRE

This was planned as a quality survey. It would not be enough respondents to give any statistic-based significance, or a hypothesis test, but it could give an impression of the meanings of the DHH society. This questionnaire does mainly focus on persons with some kind of hearing impairment – deaf or hard of hearing.

The focus has been on the quality on the captioning, error tolerance and a simple solution – what were acceptable and no questions about the use of interpreter services and quality and availability on these services.

Example from google form.

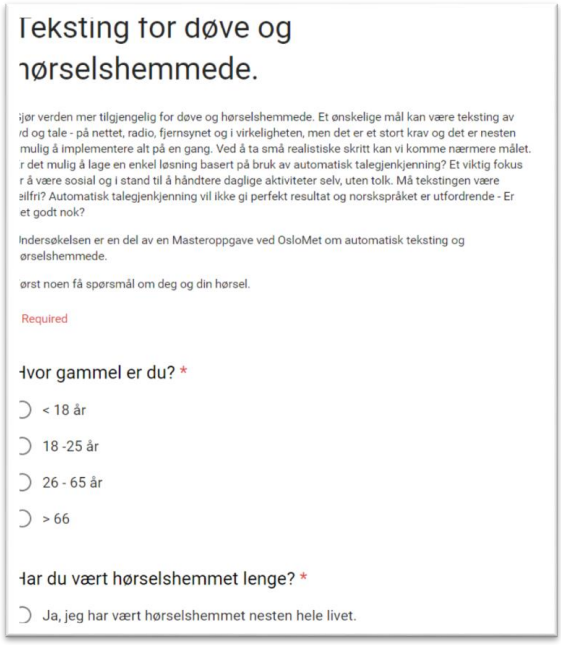


Figure 13 Example of some questions from the Google form in Norwegian.

Technical issues – accessibility

Creating and reading a questionnaire with a screenreader is an issue, and quite challenging. I ended up using Google Forms, but it is not a very screenreader friendly tool. Because of lack of time to find another tool and rewrite the questionnaire in it and because my main focus is not blind or visually impaired respondents, I continued using Google Forms.

English version of the questionnaire - Captioning for deaf and hearing impaired.

The questionnaire is a Google Form and written in Norwegian. This is a Translated version.

What life phase are you in? *

- School / student
- in work
- pensioner

Have you been DHH for a long time? *

- Yes, I have been a DHH almost all of my life.
- No. I became a DHH in adulthood, because of an injury / illness.
- No. My hearing became progressively worse with age.
- No. I'm not hearing impaired.

Who do you usually use as an interpreter? *

- Someone in my close family.
- An acquaintance or omnivore.
- A colleague or fellow student.
- A personal Interpreting Service via Smartphone.
- A public paid personal interpreter.

How often do you use an interpretation service? *

- Daily

- Weekly
- Monthly
- Rarely or never

How often do you feel the need for an interpretation service? *

- Daily
- Weekly
- Monthly
- Rarely or never
- Requirements for quality of the captions.

May automatic speech recognition done by your mobile or laptop be an option for you in some situations? Which requirement for the quality of the captions do you have? If the captions are edited humanly, it gives a better quality of the captions, but it also gives some delay and a third person will be present in your conversation.

- In a meeting where confidential information is exchanged, like with a doctor, lawyer or similar, you expect the captions to be: *
- Correct. Content like movie captioning.
- Simultaneously. May contain spelling mistakes.
- Verbatim, but may contain misspelled words.

When you are at work / school, you expect the captions to be: *

- Correct. Content like movie captioning.
- Simultaneously. May contain spelling mistakes.
- Verbatim, but may contain misspelled words.

When you're in general social settings like in town, in the store, a break at school, you expect the captions to be: *

- Correct. Content like movie captioning.
- Simultaneously. May contain spelling mistakes.
- Verbatim, but may contain misspelled words.

You watch a random video on the Internet. You expect the text to be: *

- Correct. Content like movie captioning.
- Simultaneously. May contain spelling mistakes.
- Verbatim, but may contain misspelled words.
- Do you have some comments?

Do you have any thoughts about the future, about some aids could make your daily life easier for you as a DHH?

Open question.

14.8 APPENDIX H - INTERVIEW

Some topics to discuss during the interviews.

Captioning, how? Interpreter, sign language, humans captioning in real time, captioning, of films.

Simultaneous captioning. Do you know what ASR is?

- Quality is affected by the (un)known speaker, dictation ability, and cost.
- Show and discuss the different prototype.
- **Exciting news**:
- Live Transcribe
-

Feltkode endret

DHH

- In which situations?
- Privacy. At the doctor. A third person is present?
- In more formal meetings. Correctness?
- At work. Requires prior knowledge of interpreter.
- At school, lectures. Being a student.
- Incremental steps, rather than a big one. Incremental steps provide faster results and corrections along the way would be possible.
- The world is full of gadgets and apps. Will it be too much with one more for you? Had it been nice to use the mobile phone as much as possible?
- Easily accessible. Something for all, easy to learn how to use, people get better at dictating.
- Useful app for the hearings too, you can dictate a shopping list, to-do lists, e-mails. A generally known app among people makes talking to a DHH by captioning less scary.

- Costs. What does it mean to you? Public vs private costs. (Expensive – NAV involved. Cheap - Acceptable for private. Free, for everybody)

NON-DHH

- Do you know how to communicate with the DHHs?
- Prepared to become DHH yourself?
- Is someone in close relation to you a DHH?
- Costs. What does it mean to you? Public vs private costs. (Expensive – NAV involved. Cheap - Acceptable for private. Free, for everybody).
- Socialization is an important aspect. Something for most people, everybody has heard about it. Quality vs low-threshold offers.
- Quality is affected by the (un)known speaker, dictation ability, and cost.
- Easily accessible. For all, easily to learn how to use, People get better at dictating
- Commonly known among people makes the DHH less intimidating.
- Buy a non-costly app? Use it to dictate, for your own need? To be prepared?
- Useful app also for the hearings too, you can dictate a shopping list, to-do lists, e-mails. A generally known app among people makes talking to a DHH by captioning less scary.
- The world is full of gadgets and apps. Just use your mobile phone?

14.9 APPENDIX I - PROJECT PLAN.

Not all tasks were finished within planned deadline, but all them were finished within the end date of report delivering.

The work with the questionnaire and the main Master report was very time consuming, much more than planned!

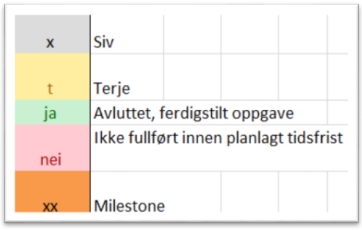


Figure 14 Codes and colours used in the project plan. Marked week is the finish work within week number, not the estimated effort.

		September				Oktober					November				Desember		
Description	Aug (31)	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
Search for people to ask	nei	ja				x											
Tidy up report after Phase I	nei	ja															
Create the questionaire. Try it out.	nei		nei				x										
Evaluate researh questions - still relevant?	nei		ja														
Plan who to ask and how to run the questionaire.				nei					x								
Make an plan for the interviews.					nei					x							
Finish the prototype.					nei					x							
Create first version of Phase II report				nei													
Reportvalidation					t												
revise first version of Phase II report together with supervisor						x											
Create the interview - Questions. Try it out.						x											
check the questionaire						t											
Run the questionaire							x										
Start documenting the results from the questionaire.								x									
Plan how to ask and how to run the interviews									x								
Start running the interviews											x						
Create second version of Phase II report									x								
check the report Phase II										t							
revise second version of Phase II report together with supervisor											x						
New revised project plan											xx						
An individual written report (10000-15000 words)											xx						
Individual oral presentation of results at the mini-conference.																xx	

Figure 15 FASE I - 2017

	August	September				October					November				Desember	
Description	31-35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Search for people to ask	nei	ja														
Tidy up report after Phase I	nei	ja														
Create the questionaire. Try it out.	nei		nei			ja	x									
Evaluate researh questions - still relevant?	nei		ja													
Plan who to ask and how to run the questionaire.				nei		ja										
Create the interview - Questions. Try it out.						Nei	Ja									
Check the questionaire							t									
Run the questionaire										nei		ja				
Start documenting the results from the questionaire.											Nei	ja	x			
Plan how to ask and how to run the interviews									Nei		Ja					
Start running the interviews											x	ja				
Create temporary version of Phase II report									nei			ja				
Make an plan for the interviews.					nei					x	ja					
Finish the prototype.					nei					ja						
check temporary report Phase II										t		x		ja		
Start running the interviews											nei		x		ja	
Run the questionaire												Ja	x			
New revised project plan															xx	
An individual written report (5000 words)															xx	
Individual oral presentation of results at the mini-conference.																xx

Figure 16 FASE II - 2018

2019 uken r	Description	Jan (1)	Feb (6)	Mar (10)	Apr (14)	Mai (19)	Jun (23)
1	start working the the statistical analyze of the questionaire	stoppet					
2	Finish Prototype	ja					
3	Stop the questionaire	stoppet					
6	Finish the interview round	nei					
8	Document the interviews			nei			
9	Document the questionaire			stoppet			
10	Create first version of Phase III report				ja		
12	revise first version of Phase III report together with supervisor				t		
13	Run through the literture review, all still relevant or something new reseaarch?				x		
15	Finalize the documentation of the questionaire.					stoppet	
16	Finalize the documentation of the interviews					x	
18	Abstract					x	
19	conclusion					x	
21	Finalize findings					x	
22	Individual students will be assessed based on the written Master thesis (30000-40000 words in APA style 6th Edition) This part of the examination counts 90% of the final grade.					xx	
24	Individual oral presentation (30 minutes).						xx

Figure 17 FASE III - 2019

HØGSKOLEN I OSLO OG AKERSHUS

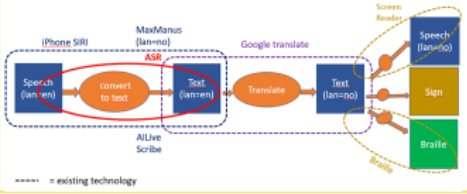
Universal design of ICT – Master

Siv Tunold

Introduction

Hearing-related disorders are one of Norway's fastest growing public health challenges. Make the world more accessible for the DHH.
One goal and solution are automatic captioning of speech, at the web and in real life.

CAPTIONING FOR THE DHH



Speech you don't understand translated to something you do understand.
Could be done by using captioning for the deaf and hard of hearing (DHH).

Method

No special equipment needed.

Just an simple app.

Always available for everybody.

Research questions:

RQ1: What features would be important for users being both DHH and visually impaired when they want to use ASR captioning as an assistive technology?
RQ2: Are DHH users prepared to accept non-optimal, but incrementally improving versions of an ASR system, or would they rather wait for the «perfect» system?

Conclusion and Future work

Upcoming technology

- Lip reading
- Text to sign (automatically converted)
- Haptic feedback
- Cheap and good ASR solutions
- Translate from speech to an other language.
- Respeaking, for making better use of ASR.
- ASR is consciously getting better.

15 References

Berke, L., & Caulfield, C. (2017). Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One

BUFDIR. (2015). Norge universelt utformet 2025 - Invitasjon til deltakelse i konsultasjonsprosess. Retrieved from https://www.bufdir.no/uu/Regjeringens_handlingsplan_for_universell_utforming/Konsultasjon/Innspillsmoter_2013/Horselshemmedes_Landsforbund/.

Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics, 10*(1), 131-153.

DCMP. (2018). A Definition of Captioning. Retrieved from http://www.captioningkey.org/quality_captioning.html#1

Delazio, A., Nakagaki, K., Klatzky, R. L., Hudson, S. E., Lehman, J. F., & Sample, A. P. (2018). *Force jacket: Pneumatically-actuated jacket for embodied haptic experiences.* Paper presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

DIFI. (2018). Hvordan teste universell utforming av ditt nettsted? Retrieved from https://uu.difi.no/krav-og-regelverk/kom-i-gang/hvordan-teste-universell-utforming-av-ditt-nettsted.

Easton, J. M. M. S. L. E. D. (2017). Personal Perspectives on Using Automatic Speech Recognition to Facilitate Communication between Deaf Students and Hearing Customers.

EU. (2019). WAD vUs webdirektiv om universell utforming av offentlige nettsteder og mobilapplikasjoner

Fenghour, S., Chen, D., & Xiao, P. (2019). *Contour mapping for speaker-independent lip reading system.* Paper presented at the

Eleventh International Conference on Machine Vision (ICMV 2018).
Findlater, L., Chinh, B., Jain, D., Froehlich, J., Kushalnagar, R., & Lin, A. C. (2019). *Deaf and Hard-of-hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies.* Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI).
Hassan, M., Assaleh, K., & Shanableh, T. (2019). Multiple Proposals for Continuous Arabic Sign Language Recognition. *Sensing and Imaging, 20*(1), 4.
Kafle, S., & Huenerfauth, M. (2016). *Effect of Speech Recognition Errors on Text. Understandability for People who are Deaf or Hard of Hearing*. Paper presented at the 7th Workshop on Speech and Language Processing for Assistive Technologies, INTERSPEECH.
Kafle, S., & Huenerfauth, M. (2017). Evaluating the Usability of Automatically Generated Captions for People who are Deaf or Hard of Hearing
Kawas, S., Karalis, G., Wen, T., & Ladner, R. E. (2016). *Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students*. Paper presented at the Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility, Reno, Nevada, USA. https://dl.acm.org/citation.cfm?doid=2982142.2982164
Kermit, P. S. (2018). Hørselshemmede barns og unges opplæringsmessige og sosiale vilkår i barnehage og skole: Kunnskapsoversikt over nyere nordisk forskning.
Kushalnagar, R. S., Lasecki, W. S., & Bigham, J. P. (2013). *Captions versus transcripts for online video content.* Paper presented at the Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility.

Kushalnagar, R. S., Lasecki, W. S., & Bigham, J. P. (2014). Accessibility Evaluation of Classroom Captions. *ACM Trans. Access. Comput., 5*(3), 1-24. doi:10.1145/2543578

Lasecki, W. S., Miller, C. D., Naim, I., Kushalnagar, R., Sadilek, A., Gildea, D., & Bigham, J. P. (2017). Scribe: deep integration of human and machine intelligence to caption speech in real time. *Commun. ACM, 60*(9), 93-100. doi:10.1145/3068663

Live, A. (Writer). (2017). Open caption OC by AI lIVE

Forskrift om universell utforming av IKT-løsninger., (2018).

Luqman, H., & Mahmoud, S. A. (2018). Automatic translation of Arabic text-to-Arabic sign language. *Universal Access in the Information Society*, 1-13.

Parton, B. (2016). Video captions for online courses: do YouTube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible, and Distance Learning, 20*(1), 8-18.

Salik, K. M., Kumar, Y., Jain, R., Aggarwal, S., Shah, R. R., & Zimmermann, R. (2019). Lipper: Speaker Independent Speech Synthesis using Multi-View Lipreading.

Schipper, C., & Brinkman, B. (2017). Caption Placement on an Augmented Reality Head Worn Device.

Sorgini, F., Caliò, R., Carrozza, M. C., & Oddo, C. M. (2018). Haptic-assistive technologies for audition and vision sensory disabilities. *Disability and Rehabilitation: Assistive Technology, 13*(4), 394-421.

UN. (2016). Convention on the Rights of Persons with Disabilities (CRPD) (Vol. Article 29 – Participation in political and public life).

W3C. (2018a). Web Content Accessibility Guidelines (WCAG) 2.1. Retrieved from https://www.w3.org/TR/WCAG21/

W3C. (2018b). What's New in WCAG 2.1. Retrieved from https://www.w3.org/WAI/standards-guidelines/wcag/new-in-21/