TaFiDiAI: Taligenkänning för Finlandssvenska Dialekter genom Artificiell Intelligens - Automatic Speech Recognition of Finnish Swedish Dialects through Artificial Intelligence.

Espinosa Leal, Leonardo; Forss, Thomas; Hägglund, Susanne; Kuvaja Adolfsson, Kristoffer ; Rautanen, Kimmo; Shcherbakov, Andrey; Tigerstedt, Christa

DOI:
10.5281/zenodo.7495136

Published: 30/12/2022

Document Version
Final published version

Document License
CC BY-NC-SA

TaFiDiAI: Taligenkänning för Finlandssvenska Dialekter genom Artificiell Intelligens

Automatic Speech Recognition of Finnish Swedish Dialects through Artificial Intelligence

Leonardo Espinosa-Leal[1], Thomas Forss[2], Susanne Hägglund[3], Kristoffer Kuvaja-Adolfsson[1], Kimmo Rautanen[3], Andrey Shcherbakov[1] and Christa Tigerstedt[1]

[1]Graduate School and Research, Arcada University of Applied Sciences, Helsinki, Finland.
[2]StageZero Technologies, Helsinki, Finland.
[3]Åbo Akademi University, Experience Lab, Faculty of Education and Welfare Studies, Vasa, Finland.

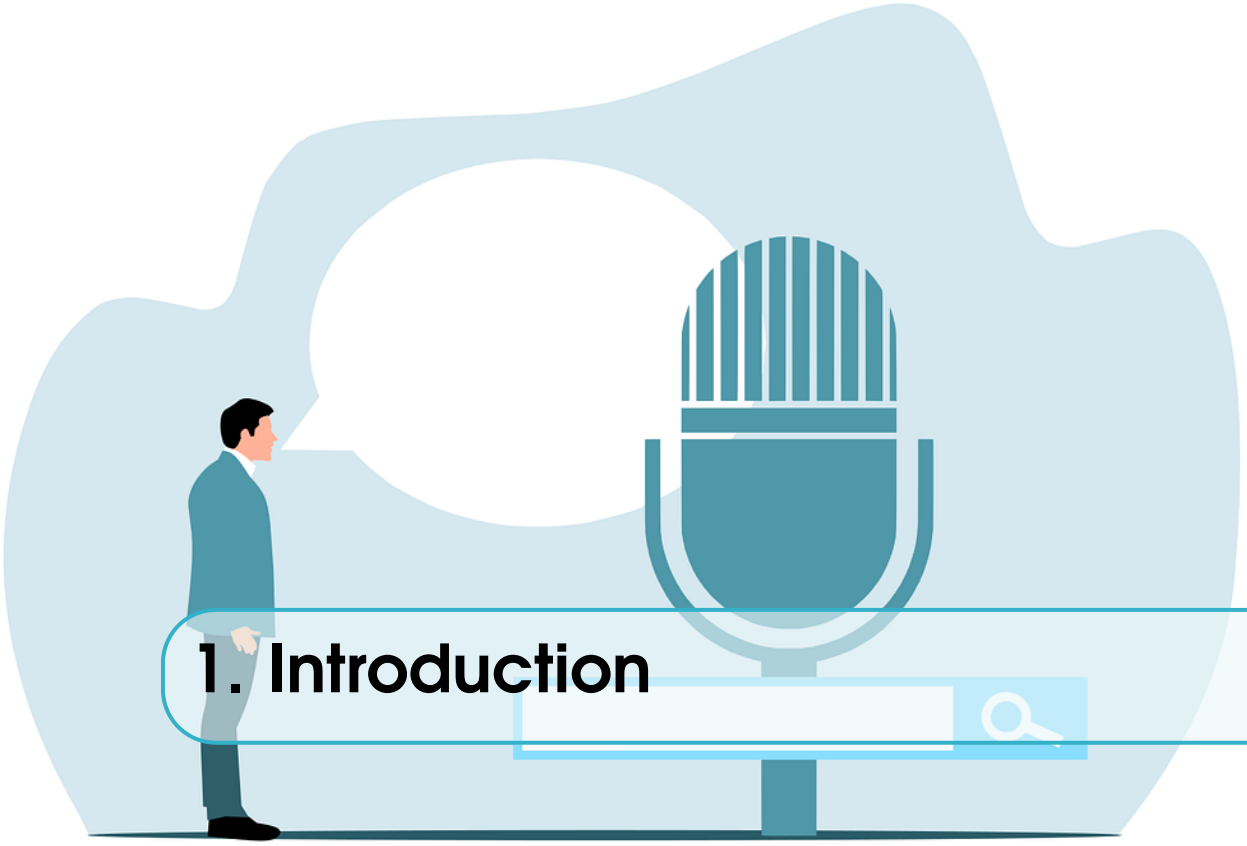
PROJECT POWERED BY

This research was done with the economic support of Svenka Kulturfonden.

Corresponding author: Leonardo Espinosa-Leal, leonardo.espinosaleal@arcada.fi

First release, December 2022

Contents



1. Introduction

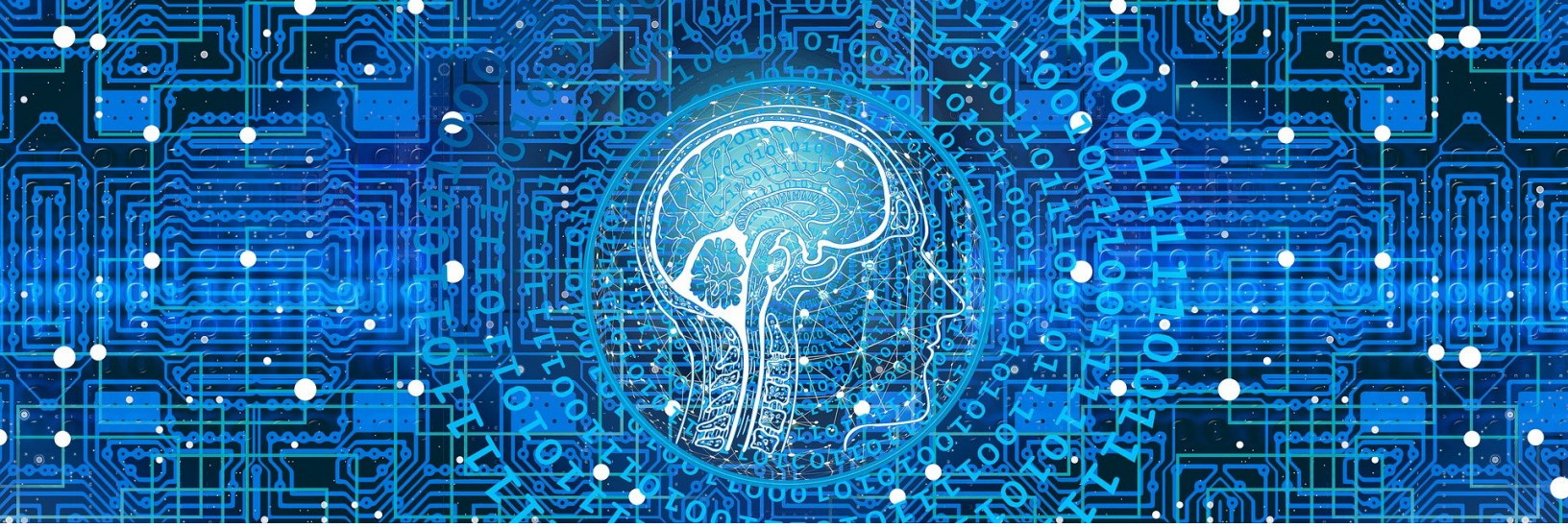
1.1 Motivation

In recent years, the use of Artificial Intelligence (AI) technologies in the area of services are gaining strong acceptance in the society, mainly due to the advances in the development of algorithms able to perform specific tasks as accurately as a human level. Despite the many successes in fields such as banking, healthcare, telecommunications, cybersecurity, education and many others, many challenges still need to be addressed (Espinosa-Leal et al., 2020). For instance, the use of virtual or embodied AI interacting with customers or users is still an area under development. For instance, the lack of clear communication with an assistant (embodied or virtual) AI robot can have a substantial effect on the level of adoption (and trust) of these technologies (Espinosa-Leal et al., 2020; Majd et al., 2021).

Substantial advances in the field of Automatic Speech recognition (ASR) have surged in recent years, mainly due to the development in software, datasets and theory, in particular of deep learning techniques (Nassif et al., 2019). Big tech companies have invested many resources in developing algorithms able to surpass human accuracy in tasks such as speech recognition, sentiment analysis, among others. These advances have improved the quality in areas such as services and healthcare. However, one of the critical issues is that most of the AI developments have targeted populations' languages such as English, Spanish, French, Mandarin, or Arabic. Languages with a high representative number of speakers and economic interest. Unfortunately, the situation with languages spoken by populations with a small number of speakers has been a subject of interest mainly for academicians and enthusiasts. The case of dialects within minority languages is even worse. There is an elevated risk of the exclusion of those groups that do not have access to state-of-the-art technologies (Espinosa-Leal, 2022).

1.2 Objective

In this project, we propose bonding academic and industrial partners to address the issue of the lack of developments in the area of automatic speech recognition of the spoken dialects of Swedish in Finnish territory. Our goal is to gather open-access labelled speech dialect data for Swedish speaker population from across Finland to develop a set of ASR technologies and then test them in the field. This project has been built upon the discussion and ideas surged from the project MäRI *(människa -robotinterakt)* Funded by Svenska Kulturfonden for the period 2020-2021 (Hägglund et al., 2022). In that project, Its goal is to design applications for human-robot interaction, explore the experiences of trust in them by stakeholders in healthcare, and to share these insights with educators in caring science and nursing on the 2 and 3 levels. Thus, its scope is rather general and broad. The current proposal focuses exclusively on one aspect of human-robot interaction, namely on language. Gathering, labelling and testing Swedish dialects aims at improving the inclusion of the multidialectal population.



2. Speech Data

For the data collection side of this project, we wanted data from native Swedish speakers across different regions in Finland to provide speech data. We started by creating a website for collection of speech data, where people were asked to donate data by using their microphone on mobile phones or laptops while speaking freely about different topics. We first tried to get the general public to provide speech by marketing the project in Facebook groups, using a Swedish-speaking YouTuber to spread content about it, and speaking in different media such as radio. While we were able to gather some data this way, it was not cost-efficient enough nor fast enough to be able to meet our budget in the project.

We then switched approach and instead arranged a competition between different student organizations across Universities in Finland. We offered a price to the top 3 organizations that provide speech for the project. This approach had a better success in the end.

2.1 Data gathering

Stage Zero was responsible for the initial data collection for TaFiDiAI which started the 25.8.2021. In theory, the collection was still going on in April 2022, but since the end of January 2022 there had hardly been any visitors on the website, and no one was recording speech in posterior dates. The server that hosted the service was switched off at the end of the summer of 2022.

The material was collected via the website, https://stagezero.fi/snacka/ on which the user could record their speech. The user was prompted to choose the dialect they spoke or the region and dialect that had influenced their speech the most ("Vilken dialekt eller stadsmål snackar du?" and "Om du inte pratar dialekt, välj då den regionen och dialekten eller stadsmålet som ditt tal mest har influerats av)".

The data collection was advertised in the media and through Facebook campaigns. Despite this, only a small number of people who visited the site recorded any speech.

Student organizations were then approached and prompted to take part in the data collection in the form of a contest. The student organizations who participated got a link that they could spread to their members. When the members used the link, the users were registered for the organization and the organization's ID was stored in the filename. The organizations were paid a compensation of 20 euros per hour of recorded speech and a bonus of 300/150/100 euros for first, second, and third placement in the contest.

The data was split into the following categories and people were asked to discuss or give opinions on the different topics: Corona, the archipelago, climate change, Swedish in Finland, "dettaomdetta" discussing some questions, healthcare and doctors visits, sports and moving, and a topic to speak freely about anything. Before starting collection, people were asked to self-report some demographics' data, this data was: region of their speech dialect down to a city or municipality level. The majority of the speakers were University students, however, we did not collect age data for the speakers. Figure 2.1 shows the distribution of dialects by regions of Finland.

The material has not been transcribed and has therefore not been prepared in a digital format. The material has been described regarding regions, number of recordings and length. No information about the speakers' age or gender had been collected.

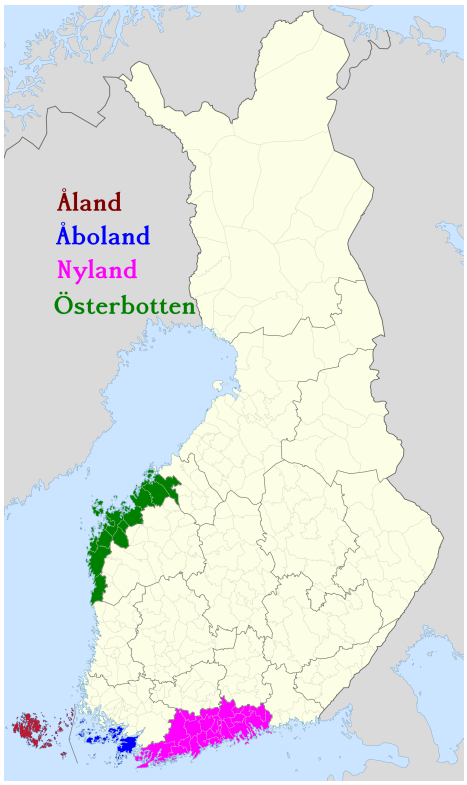


Figure 2.1: Map of Finland with the regions where the dialects were collected.

Region	Number of recordings	Minutes
Åboland / Turunmaa	107	256
Åland / Ahvenanmaa	14	8
Nyland / Uusimaa	146	254
Österbotten / Pohjanmaan maakunta	185	282
Total	452	812

Table 2.1: Distribution of recorded audio datasets by region, number of recordings and total time.

2.1.1 Description

The material consists of 452 recordings (+ 8 files that contained no speech). The total number of hours of collected speech is approximately 13 hours and 30 minutes (812 minutes). Table 2.1 shows a simple summary of the total number of recordings and total minutes of speech per region

3. Automatic Speech Recognition

An investigation for available ASR (Automatic Speech Recognition) tools focused on finding tools that were open-sourced and fully transparent was done by a team of researchers at Arcada. The most promising tools they found were Kaldi-ASR (Povey et al., 2011) and Coqui-AI-TTS[1] however neither tool was usable as it is, and before future development we found the most successful one, the KBLab at National Library of Sweden released their *wav2vec* model (Malmsten et al., 2022) that far exceeded any previous implementation attempted of Kaldi and Coqui systems. The model provided by the authors outperforms any other Swedish speaking one, including the models we used, trained only for Finnish-Swedish language.

3.1 Trained Model

The final model used for implementing a PoC ASR service and for the experiments was an acoustic model Wav2Vec 2.0 large VoxRex Swedish-C which outperforms existing models for Swedish ASR. Being a bidirectional transformer, this model imposed some limitations on the ability to transcribe streaming speech in real time. The quality of transcription by this model is greatly enhanced by providing it with audio data in chunks length over 10 seconds. This defined data transfer requirements from the client node to the server where the 1.2Gb model resides.

3.1.1 Experimental design

The objective of the experiment was to compare the assessment of the quality of ASR by human end-users and by existing statistical indicators used in ML. This would give us an evaluation of the usefulness of the metrics used for training the models. By design,

[1] https://github.com/coqui-ai/STT

volunteers had a possibility to access the perceived quality of transcription on an integer scale 0...3 from unacceptable to good. This assessment will be used to find a correlation between human and statistical evaluation of the ASR quality.

3.1.2 Data gathering strategy

Collected data consisted of audio recordings and automatic transcription of both predefined phrases that volunteers read aloud and their freely formulated utterances on certain given topics. Their own assessment for each transcription was also included into data. For the freely formulated utterances, a manual transcription was made by the researchers and included into data. The feedback is obtained from transforming the input as coloured smiled icons (4) into numeric values (0...3). All data was saved on the server only. The audio recordings and transcriptions along with assessments the participants made are only identifiable by a session ID based on time stamp.

3.1.3 User interface

Drawing on experience from previous research projects: MäRI and AFORA, in these project several aspects of human-Robot interaction were studies. From the obtained knowledge in these projects, a rough draft was made for the user interface of our system, focusing on simplicity and being self-explanatory. The data gathering system was then implemented as a web application using for the most part JavaScript on the client side and a Python back-end hosted on a CSC cloud server. In Figure 3.1 the first page of the model is presented, here the user start the interaction with the model and stops once the sentence is read. The sentences are chosen randomly from a selected set of sentences. In Figure 3.2, the page where the user introduces the feedback after automatic transcription is shown. Here, the user values the quality of the transcription by using a qualitative scale of colours. Several volunteering students and staff at Arcada tested the application and provided feedback on the usability and design of the user interface implementation. This then led to the final changes of the application before it was deployed for data gathering. The web interface can be found https://stt-dev.techlabs.fi:8080/. The application for testing was coded to run in any browser, but it is limited to one user at the time.

Figure 3.1: First page of the interface to record audio and transcription, used to collect data in the experiment.

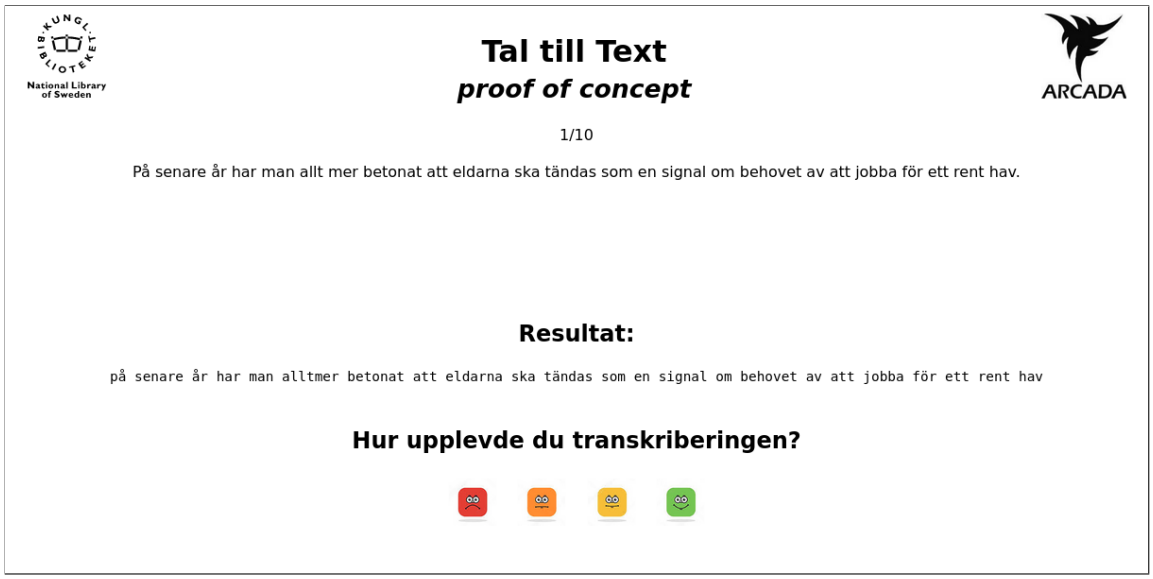


Figure 3.2: Page for feedback from the user after the transcription.

4. User testing

4.1 Approach

Adopting a human centred design perspective, we explored how informants perceive the proof of concept (PoC) model in terms of trustworthiness (reliability, functionality, and helpfulness, see for instance Mcknight et al. (2011)). Increasing the understanding of how people prefer to use automated systems can help steer the direction of developing AI and support sustainable implementation of automatic speech recognition systems (ASR) and, thus, better serve human needs (Shneiderman, 2022).

4.2 How the data was gathered

End user tests were conducted in October, November, and December 2022 throughout parts of Swedish-speaking Finland, mainly in Ostrobothnia and Nyland. 12 informants in Ostrobothnia and Åboland, and 7 in Nyland, participated in the study, invited through purposive sampling. The important criteria was a vivid and strong dialect as a first language.

This stage involved testing, interviewing and survey. The interviews followed upon the above mentioned tests. Interviews lasted about 10-25 min. On top of this participants took part in a survey about speech to text systems and about the test they just did. Interviews are transcribed verbatim and the audio files are deleted when the transcription is ready. Participant were give a consent form before the testing and data collection started.

Studies were carried out either in a university setting or in the field, such as in informants' homes or workplaces. Informants received no reimbursements for devoting their time and effort to participate in the study.

The aim of the study was twofold. On the one hand, we measured perceived usefulness and perceived ease-of-use of the Proof of Concept model. On the other hand, we explored trusting beliefs regarding reliability, functionality, and helpfulness. Together with infor-

mants, we pondered use cases and scenarios, with a special focus on healthcare. Please note that the overall aim was not to provide a diagnostic snapshot of the PoC. Rather, end users' perception of the speech to text service, their individual acceptance, and potential postadoption behaviour.

The results from the experiments performed at Helsinki are not consigned in this report.

Demographics

Dialects from across Ostrobothnia, Larsmo to the north and Närpes to the south, are represented in the sample. 5 from southern, 3 from central, and 3 from northern parts of Ostrobothnia. Several studies were carried out in Åboland, but only one study gave valid data of a Åbo regional dialect speaker. The respondents in Nyland represented regions like Åboland, Ostrobotnia and Helsinki. Data collection is continuing in Nyland.

4.3 Results

All informants were asked to evaluate the ASR proof of concept in terms of usefulness and ease-of-use through the Technology acceptance Model TAM (Lewis, 2019), aiming to quantify likelihood of technology acceptance. TAM has 12 items, where the first six assess perceived usefulness (PU) and the additional six items assess perceived ease of use (PEU). The survey was modified slightly in this study as the workplace context of the tasks was played down. Tasks carried out as a client, consumer, societal member and/or patient/client were stressed instead. The survey's focus on future use remained intact, and ratings concern likelihood of future use instead of experiential ratings of the PoC. Ratings of the transcripts were given in the system itself and are analysed by the Arcada team.

Semi structured interviews were also carried out and a Job to be done-element was included. Jobs-to-be-done (JTBD) is a tool based on the idea that whenever users use a product, they do it for a particular progress or outcome that they need to accomplish. JTBD are at the same time a representative of user needs (Riedmann-Streitz, 2018).

Taken together, users participating in the Vasa team studies struggled somewhat to see how ASR technology would be useful to them when carrying out service tasks in their everyday life. However, the informants slightly agree that voice-based interfaces and speech technology systems would be easy to use. One informant maybe, seven don't, and four do identify personal use cases in healthcare, which is the context this study focuses on. Preliminary analysis of the Nyland data confirms this finding.

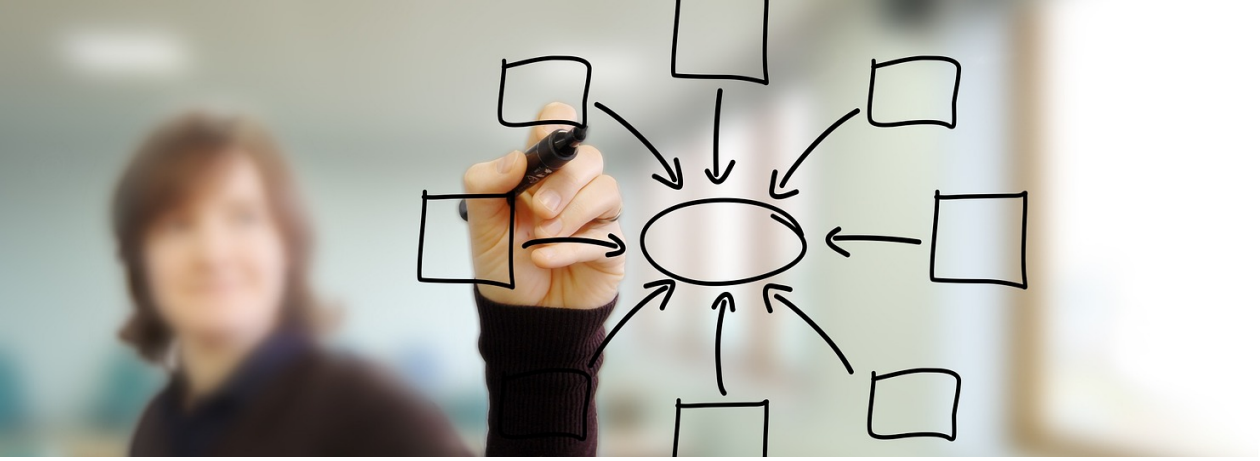
The suggested use cases are 1) to be (better) understood (by another) and 2) to improve one's own competence in and/or understanding of dialects and their speakers. Two informants identify a rather instrumental factor that did not fall under these two outcomes: accomplishment of tasks quicker. The success criteria are thus both subjective, in terms of improving task efficacy and one's skills in dialects, and social, relating to understanding other parties in the communication and to oneself being understood by them.

The identified use cases in healthcare include:

- When communicating with authorities, and with care providers in general, due to their terminology. It's easier to be able to speak one's own dialect and to be understood

when doing so.
- Calling for making appointments to doctors and other professionals in care, as it would save time.
- ASR systems could support language skills of care professionals in their patient communication, e.g. if they do not understand a particular dialectal word.
- Improving and strengthening patient safety and security as there's a resource available for interpreting and understanding dialects and to be able to speak one's own language (=dialect).

Should ASR technology be implemented in healthcare, values that would be of great importance for users in the Vasa team studies include "to be fully understood" and "the system to function correctly". If not, their sense of trust and sense of security is at stake. Another value is humour and a positive ambiance around the novel technology in patient-care staff encounters. Security as well as presence of/personal contact with a human being are also expressed as important values. Finally, acknowledgement of what has been said, for example through an audio receipt, and reliability of the system are noted as important values.



5. Conclusions

It's noteworthy that although many identify benefits and values of an ASR technology recognising dialects, they don't necessarily include themselves in the target group or adopter group. This is mainly due to the fact that they are already being well understood in care contexts, but also a general mistrust in ASR technology competence and reliability, and a preference for human contact and interaction in care contexts. Among the data gathered by the Vasa team, a required leap of faith is clearly visible. Users want proof that the ASR technology actually works, and they want to see for themselves that it works. If they see convincing evidence of great functionality, then perhaps trust will follow. Many mention that as of now, they are very sceptical as to whether they would trust an ASR system.

The study conducted by the Vasa team points to challenges in relation to the functionality and reliability of the PoC system, to end users' requirement that ASR technologies simply must work well, and to the necessity to identifying contexts where its use would be meaningful to humans in their (inter)action. Many informants were tentatively positive towards the technology and didn't reject using it, once it is developed to reach a certain level of reliability and competence to recognize their accent.

The interview data collected by the Vasa team suggests that future acceptability and use of ASR technology depend on the context (home, leisure time, professional work, or healthcare), end goal of the user (service encounters with great consequences or basic, instrumental tasks) and level of significance (literal recognition of words or general picture; severe illnesses in care or routine checkups), and user preferences, prior experiences, and language skills.

Bibliography

Espinosa-Leal, L. (2022). Inclusion in human-robot interaction. https://zenodo.org/record/6337488.

Espinosa-Leal, L., Chapman, A., and Westerlund, M. (2020). Autonomous industrial management via reinforcement learning. *Journal of intelligent & Fuzzy systems*, 39(6):8427–8439.

Hägglund, S., Tigerstedt, C., Biström, D., Adolfsson, K. K., Penttinen, J., Wingren, M., Andersson, S., Backholm-Nyberg, Y., Fant, H., and Espinosa-Leal, L. (2022). En glad maskin - processrapport från märi-projektet, människa-robotinteraktion i vårdkontext. https://zenodo.org/record/6520248.

Lewis, J. R. (2019). Comparison of four tam item formats: Effect of response option labels and order. *Journal of Usability Studies*, 14(4):224–236.

Majd, A., Biström, D., Tigerstedt, C., and Espinosa-Leal, L. (2021). Social and service robots deployed for social distancing-optimization and placement. In *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–9. IEEE.

Malmsten, M., Haffenden, C., and Börjeson, L. (2022). Hearing voices at the national library – a speech corpus and acoustic model for the swedish language. https://arxiv.org/abs/2205.03026.

Mcknight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, 2(2):1–25.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech*

Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Riedmann-Streitz, C. (2018). Redefining the customer centricity approach in the digital age. In *International Conference of Design, User Experience, and Usability*, pages 203–222. Springer.

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.