7a.026-TUT - Improving Speech Recognition Robustness

Report Date	Project Start	Project End	Project Budget	Amount Spent To Date
14 Jan 2019	1 August 2018	31 July 2019		

PROJECT SUMMARY

Speech recognition for general tasks is widely available in many languages by providers such as Google and Microsoft. However, current level technology has not been able to come up with a good quality general speech recognizer. Therefore, a good quality system requires training on case specific datasets for the task at hand. This is not possible with e.g. Google Speech API. Another aspect of this customization is the ease with which it is possible to take additional aspects of the audio into account on subsequent models. Another concern with these cloud services is confidentiality. In some use cases the data cannot be allowed to leave the organization in question. The costs of continued use of the cloud services can also be considerable. The aim of this project is to help Silo.ai to create an in-house solution for speech recognition. A general deep learning -based speech recognizer is trained on open data and other available sources. The general model is used in the creation of better case specific models, which are trained on client data. The resulting models are portable and can be setup in either cloud environments or local servers. The model can easily be combined with or serve as an input for additional ML models, such as sentence classification.

Robustness to noise and other interference in the environment is a crucial feature of a speech recognition system. The system should also be robust to different speakers, especially in public environments, instead of being adapted specifically to each user. The purpose of this research project is to investigate different approaches to make speech recognition systems robust to noisy environments and different speakers. This project will study speech enhancement techniques with next-generation, data-driven approaches.

PROJECT TEAM

Team Member Name	Team Role *(PI, Co-PI, Student, Researcher)*	Academic Site
Moncef Gabbouj	PI	Tampere University
Okko Räsänen	Co-PI	Tampere University
Ali Senhaji	Researcher	Tampere University
Mohammad Al-Sad	Researcher	Tampere University

IAB PROJECT MENTOR(S)

IAB Project Mentor Name	IAB Organization
Jaakko Vainio, Filip Ginter and Peter Sarlin	Silo.ai

PROJECT FUNDED BY

IAB Organization(s)
Silo.ai
Business Finland

CVDI Year 7 Mid-Year Report

08/01/18 – 12/31/18

OVERALL PROGRESS/ACHIEVEMENTS TO DATE

We identified new resources to build a new dataset for the ASR task in Finnish language. We are currently working on making it ready for use by applying alignment techniques. We studied the new data-driven approaches for ASR, as well as the current state-of-art architectures by outlining the engine's capabilities and hyper-parameters. Now we are setting up the training environment to run experiments. As a next step, we are going to augment speech data using the identified techniques and conduct a robustness assessment.

PROJECT DELIVERABLES

Deliverable	Achievements	Remaining To Do
A new Database based on public domain resources for Automatic Speech Recognition task in Finnish language.	- Identified the resources to be used for the new Dataset. - Benchmarked the different methods for transcript alignment methods. *In progress:* Testing the different transcript alignment methods. **60% Complete**	- Conduct performance measures on the aligned transcripts. - Formally describe the audiobooks database. - Perform a preliminary experiment using this database.
A baseline for the task and determine appropriate metrics for robustness.	- Setup the DeepSpeech engine on a local machine. - Outline its input and output restrictions, e.g. data format and size. - Outline the engine's capabilities and hyper-parameters. **50% Complete**	- Fix the dataset that will be used for this project, by merging a portion of the parliament dataset and the audiobook data, and evaluate the baseline. - Designing metrics for robustness.
Augment the merged database to expand the engine training domain and to test its ability in generalizing to unseen environments, speakers and samples.	Benchmarked data augmentation techniques: • Change the speaker acoustics using vocal tract length perturbation (VTLP), spectral shifts, and speech speed distortions. • Modify the acoustic environment, e.g. reverberation and (channel variability). • Add interferences such as city traffic sounds or cafe background noises with increased signal-to-interference ratios. **10% Complete**	- Augment speech data using the identified techniques. - Conduct statistical analysis on the augmented data to ensure effective deviation from its original form. - Conduct a robustness assessment to compare the performance of the model using different data augmentation techniques. • Train a set of models using augmented data by the various methods applied to the original datasets. • Evaluate the performance of the different methods. • Perform statistical analysis to access robustness

Propose a novel data augmentation technique to enhance the deep learning approach for Automatic Speech Recognition Systems in general and for Finnish specifically.	Further work will be reported as we progress. **0% Complete**	