# Project proposal v2

## Determining number of different speakers during the day

### Introduction and motivations

Determining the number of face to face conversations people have in everyday life would be an interesting measure for numerous purposes. Some example uses could be; analysis of social networks, workspace dynamics, diagnosis of depression and stress. In this project we will develop a mobile app which detects starts of conversations and use principles of machine learning to determine how many different people the user communicates with in everyday scenarios.

### Proposed methods

#### Android application

Our application will need to collect microphone data and identify how many different people the owner speaks to during the day. To achieve that the app will need to be running in the background and detect when a conversation is taking part. How to effectively achieve that without draining too much battery during the day is still up for discussion. For a proof of concept.

#### Sound signal preprocessing

Recent work has shown the usefulness of convolutional neural networks (CNN) to extract features from speech spectrograms. Thus first preprocessing step of raw audio will be converting the signal from the time domain to a frequency domain. For that we will first usea pre-emphasis filter to amplify higher frequencies, then we will split the signal in short time frames (cca. 30ms). On each frame we then apply a short-time Fourier transform to get a frequency power spectrum. We could then do further processing with extracting mel-scaled filter banks and mel-frequency cepstral coefficients. But according to empirical evidence these steps might not be necessary for neural network based models, due to the fact that those transformations are linear and should easily be bypassed by a neural network. Sound signal processed this way will serve as data representation for our model

#### Speaker counting

To efficiently discern different speakers we need to somehow map their speech data to a lower dimensional embedding space, where proximity of different signals will imply a similar or same speaker. For this we propose to implement a convolutional siamese network. Siamese networks have been extensively used in facial recognition and image clustering. The core idea is that they optimize embedding generation in such way that distances between embeddings of same class instances are minimised and distances between embeddings of different class instances are maximised. For training we will use VoxCeleb or similar dataset. We will train a siamese network separately on a GPU enabled workstation in the PyTorch framework. Then we will convert a trained PyTorch model to a TensorFlow model using ONNX and from there convert it to a mobile acceptable format using the TensorFlow Lite Converter.

When the model is successfully ported to an Android device, it will be run each time the application detects a conversation happening. We propose to split continuous speech data in to smaller chunks and produce embeddings for each of them. When the conversation ends we would use one of classical clustering methods like k-means to make clusters of speech chunk embeddings. The number of clusters will represent the number of speakers taking part in the conversation.

### Timeline

* Week 1,2:
  + Prepare the application to collect microphone data
* Week 3,4,5:
  + Train a siamese network
* Week 6,7,8:
  + Transfer trained network to the mobile device
  + implement sound data preprocessing on mobile device
  + implement sound data framing
* Week 9:
  + Implement clustering
  + testing
* Week 10,11:
  + Report writng

### Related work

* [Signal preprocessing approaches](https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html)
* [No Fuss Distance Metric Learning using Proxies](https://arxiv.org/abs/1703.07464)
* [SPEAKER DIARIZATION USING DEEP NEURAL NETWORK EMBEDDINGS](https://danielpovey.com/files/2017_icassp_diarization_embeddings.pdf)
* [LEARNING EMBEDDINGS FOR SPEAKER CLUSTERING BASED ON VOICE EQUALITY](https://core.ac.uk/download/pdf/159488805.pdf)
* [Siamese Neural Networks for One-shot Image Recognition](https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf)
* [Speaker Verification using Convolutional NeuralNetworks](https://arxiv.org/pdf/1803.05427.pdf)
* [DEEPCLUSTERING WITH ADYNAMICAUTOENCODER: FROMRECONSTRUCTION TOWARDSCENTROIDSCONSTRUCTION](https://arxiv.org/pdf/1901.07752.pdf)

### Resources

* Potential datasets:
  + [VoxCeleb](http://www.robots.ox.ac.uk/~vgg/data/voxceleb/)