1 Projekt beskrivelse

1.1 Motivation

We see a tendency in voice conversion being more normalised in our everyday. Both on social media, such as Snapchat where filters give you a wacky new voice, but also in more business oriented cases like call centers trying to make their voice sound different so they may seem more reliable from a customers point of view. We also see voice conversion spread into the news as Deep-Fakes, manipulating and fooling people. In a world where most people still see their voice as unique as their fingerprint, we find it interesting to explore the world of converting voices in a believable way. This raises the question of who can we trust and how easy is it to fool these people?

1.2 Scope

The scope of this paper is to investigate State of the art of Voice Conversion models and improve these to reduce the time and data needed for voice conversion. We would like to focus on zero-shot conversion as this seems to widen the use and gives the oppotunity for multiple applications without having to train the model all over again.
We want to use already existing methods, and try to find the minimum of ressources/data/features etc. to create a convincing VC. The State of the art tells us that Autoencoder from WaveNet would be a main goal, but to validate this model we could also look at GAN (starGAN / ada-GAN). Our success criteria would be to implement the models in such a way that the final product would be able to succeed a Turing Test.

1.3 Two possible issues of interest

1) Accent Conversion
A particular interesting case is to work with accent conversion in which the source speakers voice is maintained after the conversion but the converted accent matches that of the target speaker.
Concerns

- **Definitions:** How do we define accent and voice?
- **Data:** How do we split audio data into voice features and accent features? In some papers they focus on acoustic features (often Mel-frequency cepstral coefficients) and articulatory features (often electromagnetic articulography) - but we aim to use audio data only. Some papers also address the use phonetic posteriograms(?).
- **Availability:** It is unfortunately difficult to find available code for Accent conversion.
- **Goals:** How well will the possible outcome be? Is it possible to achieve anything meaning full within the limited time frame of the project considering the above limitations?

2) Optimise VC to reduce amount of time and data
A different case is to improve current VC models to better perform zero-shot conversion with less data of the target speaker. The goal would be to implement real time solutions for VC.
Concerns

- **Use case:** Whereas accent conversion can improve understanding and help in business the use of regular VC seems to be mainly for entertainment or crimes.
- **Data ethics:** Using someones voice data to mimic this person requires their acceptance.