IMPROVING IMPAIRED SPEECH USING NON-PARALLEL VOICE CONVERSION

Inge Salomons

INTRODUCTION

Research objective: determining to what extent the intelligibility of esophageal speech can be improved using a non-parallel voice conversion model.

- **Esophageal speech** is produced by means of vibrations in the esophagus after the larynx has been removed and differs from natural speech in a wide range of acoustic features, such as fundamental frequency and relative intensity (in dB), resulting in less natural and intelligible speech.
- A **non-parallel voice conversion model** aims to convert source speech to target speech while maintaining the linguistic information, and is trained on a data set of two types of speech that are unrelated regarding the linguistic content.

METHODOLOGY

STEP 1 Adopting a state-of-the-art model by Chen et al. (2018), created to convert dysarthric speech in Mandarin, and learning how to run it. Model is a Generative Adversarial Network, consisting of three interacting models: Generator, Discriminator and Controller. Generator tries to fool the Discriminator, which has to discriminate unimpaired from generated speech. Controller minimizes the loss in impaired speech.

STEP 2 Adapting the data processing procedure to fit the specific training data:

- Enlarging the amplitude clipping range;
- Lowering the silence threshold for trimming.

STEP 3 Training on 85 fragments of esophageal speech vs. 11.189 fragments of unimpaired speech.

STEP 4 Testing on 15 unseen fragments of esophageal speech and evaluating the results.

RESULTS

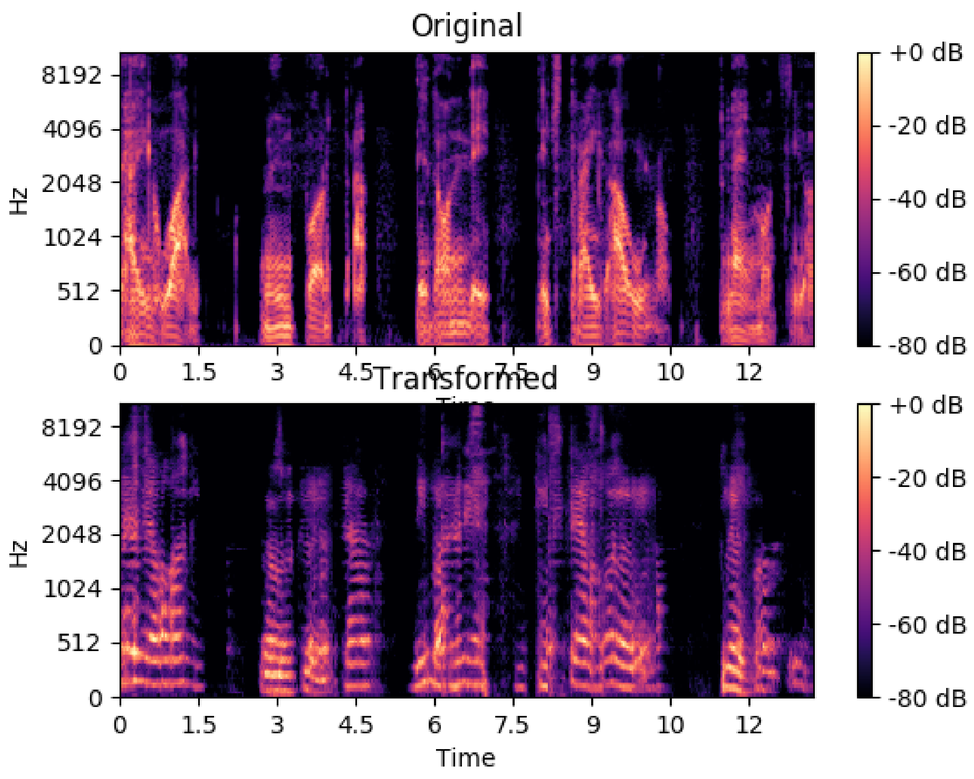

Figure 1. Spectrograms of original and generated speech

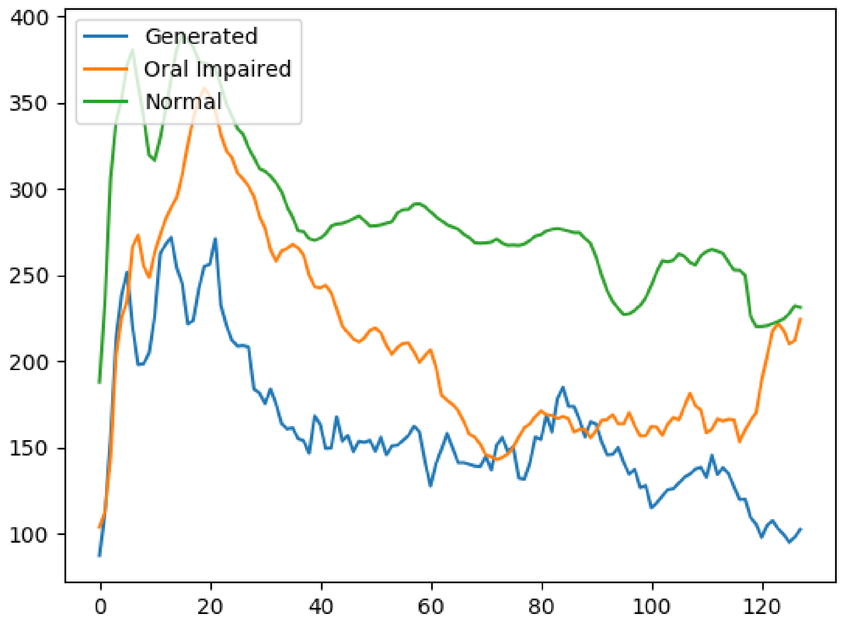

Figure 2. Global variances of original, unimpaired, and generated speech

- Objective intelligibility score (STOI index): 0.42 > 0.34.

DISCUSSION & CONCLUSION

The results show that there is:

- No improvement in intelligibility;
- Improvement in naturalness;
  - Reduced noise;
  - Restored fundamental frequency;
  - Increased relative intensity.

This suggests that even though the naturalness of the speech is improved, non-parallel voice conversion is not suitable for esophageal speech. Since esophageal speech differs from other types of impaired speech in speech production, referring to the absence of the vocal cords, a more specific model that is trained on a larger data set of esophageal speech is more preferable.

REFERENCES

- Chen, Li-Wei, Hung-Yi Lee, and Yu Tsao (2018).Generative Adversarial Networks for Unpaired Voice Transformation on Impaired Speech.
- Mohammadi, Seyed Hamidreza and Alexander Kain (2017). “An overview of voice conversion systems”. In: Speech Communication 88, pp. 65 –82.