Enhancing Speech For Parkinson's Disease Patient

Quan Nguyen, Hanqing Guo, Yuanda Wan, Dr. Qiben Yan

MOTIVATIONS

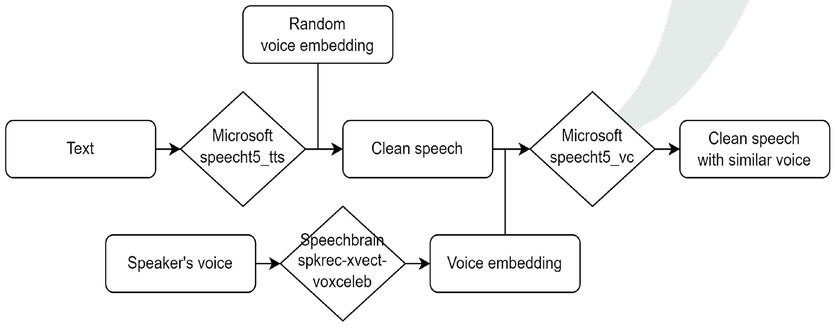
- Parkinson's disease (PD) leads to speech difficulties, characterized by slurred and low intelligibility speech, which impact communication for those affected
- Previous studies have explored enhancing automatic speech recognition (ASR) specifically for Parkinson's patients by developing customized models, techniques to address slurred speech characteristics, but not specifically focusing on improving audio intelligibility
- Our research project explores the use of Generative Adversarial Network (GAN) to enhance speech intelligibility, aiming to generate speech that retains the speaker's characteristics while improving clarity.

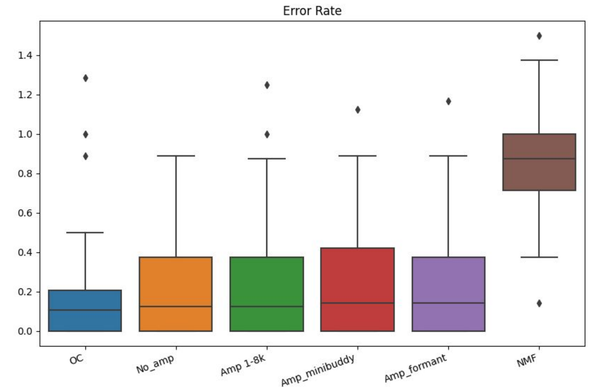
OBJECTIVES

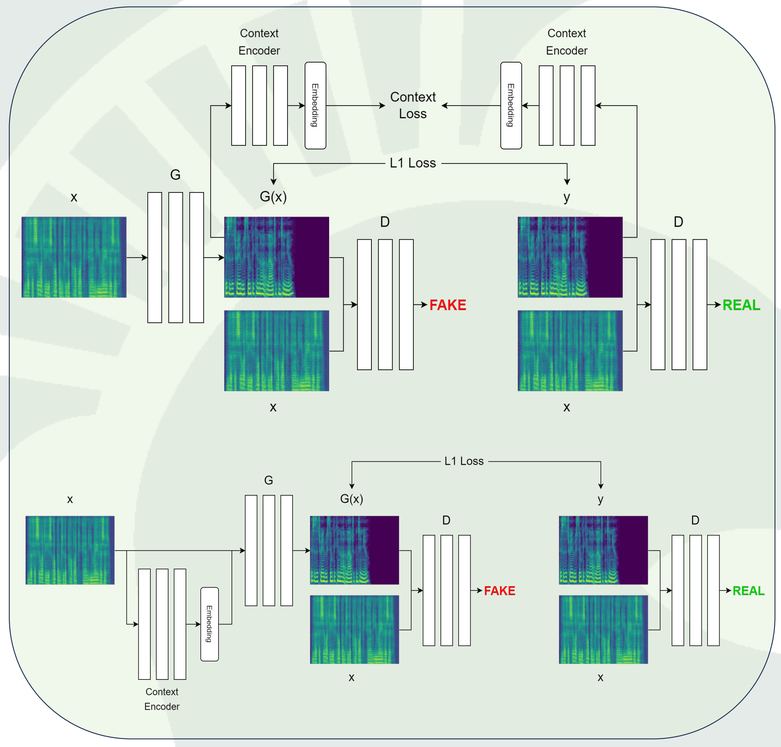
- Address the communication challenges arising from speech impairments in Parkinson's disease and explore innovative approaches to mitigate limitations.
- Retain the unique speaker characteristics in the synthesized speech while significantly improving speech clarity and intelligibility.
- Contribute to the field of speech enhancement for PD patients, aiming to improve their communication experience and overall quality of life.

DATASET

- PD speech: we collected speech from 106 PD patients. Each patient was asked to reading aloud Harvard Sentences (phonetically balanced sentence lists) in 48 kHz. Later we added another 342 speech files
- Clean speech: we synthesized using pretrained text-to-speech (TTS) model resulting in speech with 16 kHz







METHODOLOGY

For the spectral enhancement, we use:

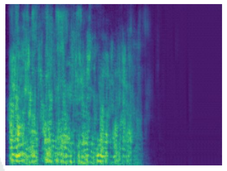
1. Deep Neural Network (DNN), based on Pix2Pix model with further modification (Isola et al., 2017):
   - Use context loss to draw generated and clean speech closer to each other
   - Use context embedding to ensure the context is included during training
2. Other enhancement methods such as Non-negative Matrix Factorization (NMF), Formant Enhancement, etc.

METHODOLOGY (cont.)

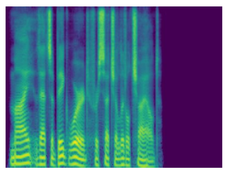
- For the DNN approach, we adapted Pix2Pix model, a cGAN with an encoder-decoder generator and a PatchGAN discriminator suggested by Isola et al. We trained the generator using either ResNet or Unet with skip connections.
- During the training, we standardized all audios to 16kHz, trimming or adding white noise to ensure a consistent length. We then transformed the audio into mel-spectrograms, an image-based representation of the waveform. The model's output, the mel-spectrogram, was used to reconstruct audio using the Griffin-Lim algorithm (Griffin et al., 1984).

DISCUSSION

- The current models exhibit noisy speech, partially attributed to the use of the Griffin-Lim. However, they still fall short in generating mel-spectrograms with sufficient detail for producing intelligible speech



Generated Mel-spectrogram



Ground truth

CONCLUSIONS

- Regarding the results, our models demonstrate limitations in recognizing characteristics of PD speech.
- In our future steps, we intend to explore a range of different architectures and integrate additional techniques from ASR, that designed to address speech challenges in PD patients, to enhance quality of synthesized speech. To strengthen the validity of our findings, we aim to expand the dataset, making our research more robust and applicable, ultimately improving communication for individuals with PD.

REFERENCES

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 32(2), 236-243.
Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).