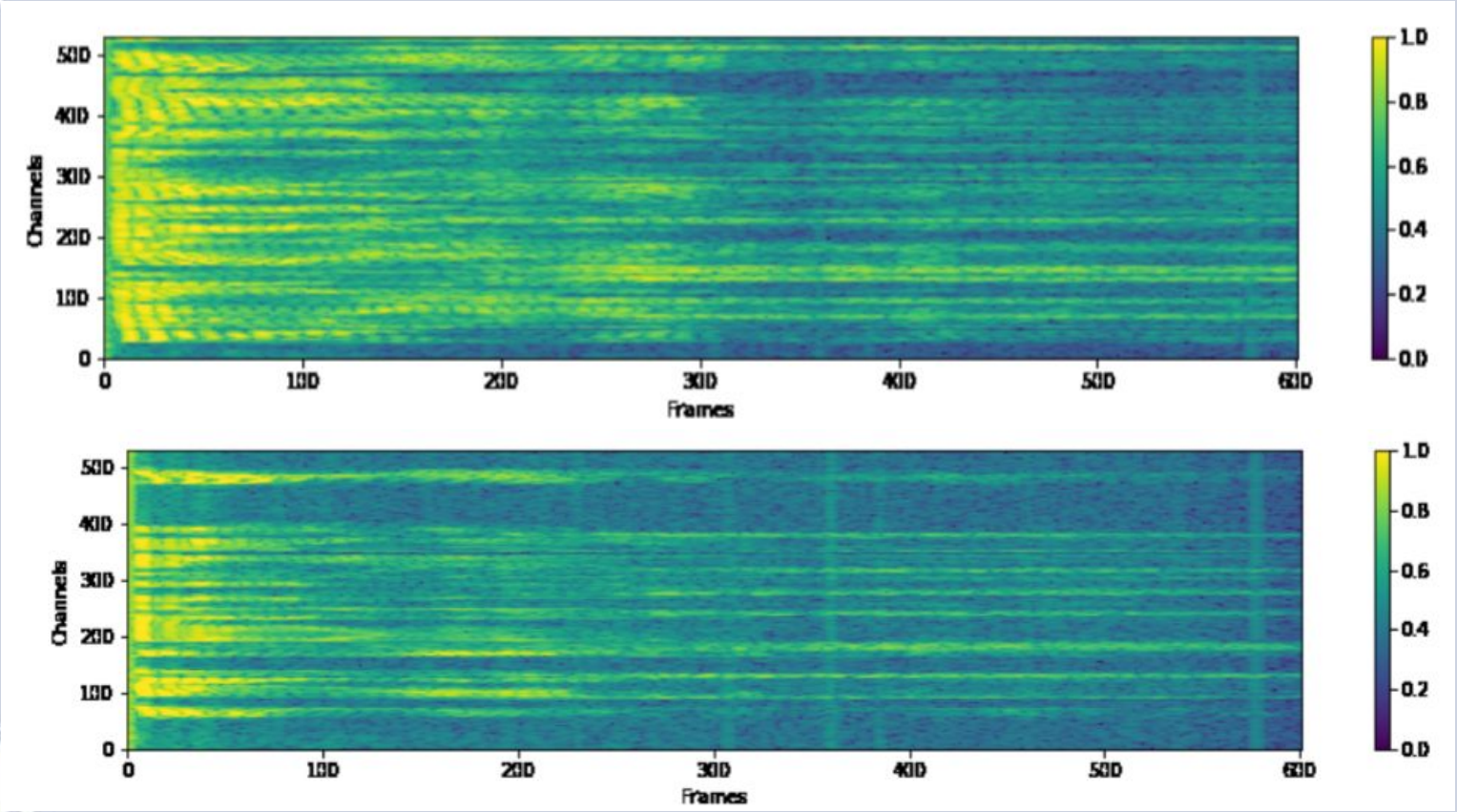
VoiceSplit: Targeted Voice Separation by Speaker-Conditioned Spectrogram

Course: SCC5830 - Image Processing - 2020
Name: Edresson Casanova, Pedro Regattieri Rocha
NUSP: 11572715 , 8531702
Email: edresson@usp.br, pedro.regattieri.rocha@usp.br
Project Github: https://github.com/Edresson/VoiceSplit

Introduction

- This Project goal is the development of a system that, given an audio input, is able to separate overlapping voices through the use of Spectrograms, based on the characteristics of each speaker's speech patterns.

Introduction



Introduction

- For the development of this work we used the Google VoiceFilter system.
- This work's contributions are as follows:
  - Proposes improvements for the architecture of the VoiceFilter model;
  - It is the first work that compares the performance of different loss functions in this scenario;
  - All of our experiments are open source and can be used freely by the community.

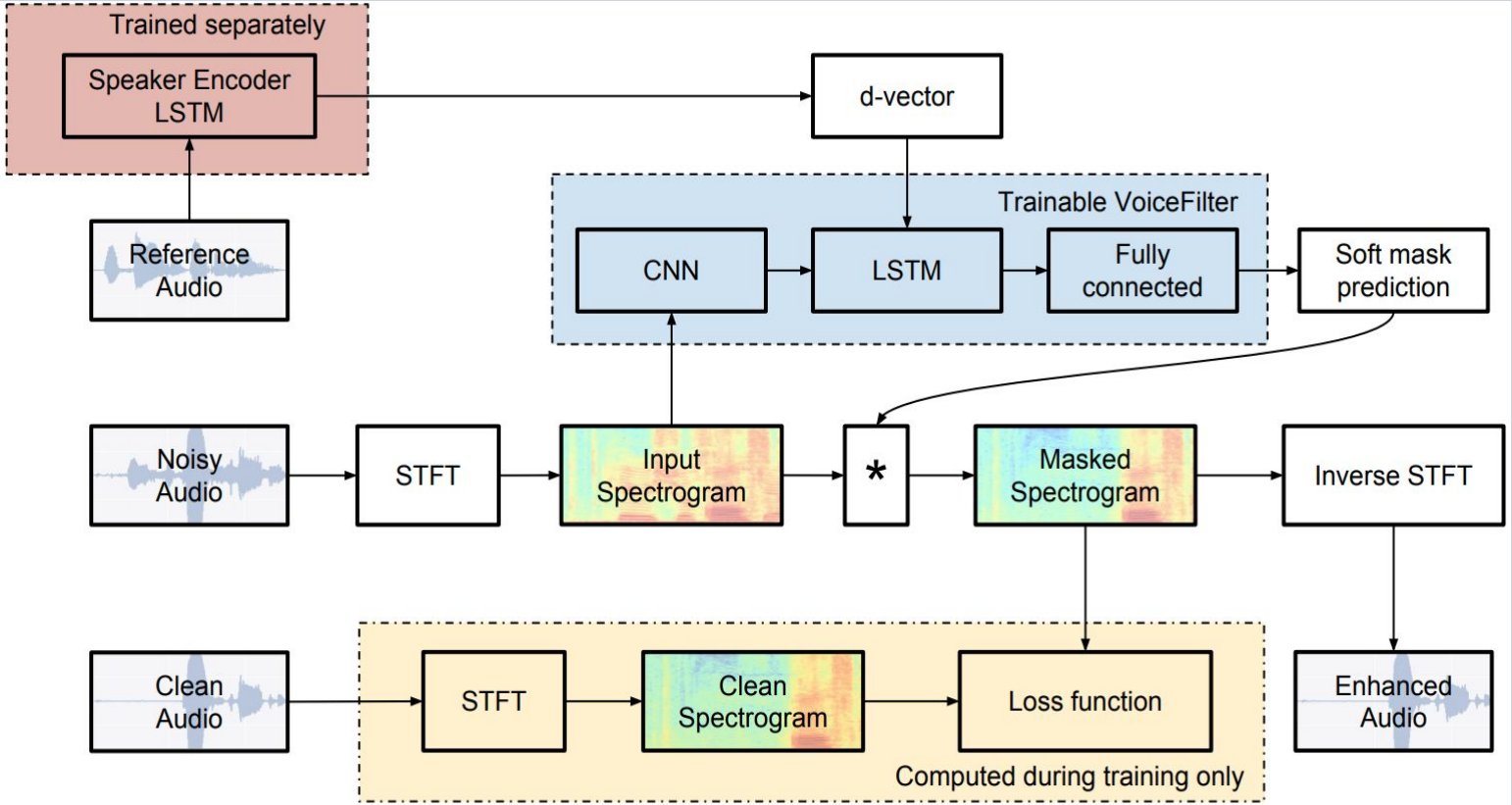
VoiceFilter

- VoiceFilter consists of two parts trained separately:
  - The Speaker Encoder;
  - The VoiceFilter Network, which uses the output of the Speaker Encoder as an additional input.

Speaker Encoder

- The Speaker Encoder is nothing more than a speaker identification/verification system. In the context of VoiceFilter it is used to extract a compressed representation (a vector of size 256) of the speech characteristics of an speaker.
- We used a speaker encoder trained with the GE2E loss that was trained with Mel Spectrograms.
- Two versions: GE2E2k and GE2E3k
- Speaker encoder in the original work was trained with **many many** more hours !

VoiceFilter



Data Pre-processing

- Uses some Image Processing themes we have seen in class
- Mel Spectrograms
- Fourier Transforms

Experiments

- All experiments use the GE2E2k speaker encoder unless stated otherwise.
- Experiment 1 - Reproduces VoiceFilter, uses MSE as loss function.
- Experiment 2 - Reproduces VoiceFilter, and also uses the same loss function, Power-Law Compressed function.
- Experiment 3 - Reproduces VoiceFilter, uses SI-SNR as loss function.

Experiments

- Experiment 4 - Similar to Experiment 3, however additionally the ReLU activation function is replaced with the Mish activation function.
- Experiment 5 - Similar to Experiment 3, however we use the GE2E3k speaker encoder.

Results

Experiment	Avg SI-SNRi
VoiceFilter [Wang et al., 2018]	**10.55729**
1	6.02260
2	5.69875
3	5.66147
4	6.49116
5	**6.55238**

Demos

- Colab notebooks Demo:
  - Exp 1: https://shorturl.at/eBX18
  - Exp 2: https://shorturl.at/oyEJN
  - Exp 3: https://shorturl.at/bInEW
  - Exp 4: https://shorturl.at/qFJN8
  - Exp 5 (best): https://shorturl.at/kvAQ8
- Site demo for the best experiment:

  https://edresson.github.io/VoiceSplit/

Conclusions

- Although we could not reproduce Google's values as we lacked their data, we did find promising results about the use of the MSE loss function and Mish activation function.
- This may lead to the improvement of automatic speech recognition software and separating samples for populating data sets.

References

S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. Neural voice cloning with a few samples. In **Advances in Neural Information Processing Systems**, pages 10019–10029, 2018.

Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In **Advances in neural information processing systems,** pages 4480–4490, 2018.

WANG, Yuxuan et al. Tacotron: Towards end-to-end speech synthesis. **arXiv preprint arXiv:1703.10135**, 2017.

PING, Wei et al. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. **arXiv preprint arXiv:1710.07654**, 2017.