L13 - Spectrogram Player

Michele Perrone, Paolo Sani

Work Overview

- **Datasets**: LJSpeech 1.1 (speech), GTZan (music), Urban8k (urban sounds)
- **Techniques tested**: Griffin-Lim, MelGan, HifiGan, Uniglow, SqueezeNet
- **Metrics**: mean logarithmic spectral distance (LSD), Itakura-Saito distance (ISD), RMS, cosine similarity, Pearson’s correlation coefficient (PCC)

Datasets

- 20 random files chosen from each dataset
- **Urban8k**: only fold 1 considered
- **LJSpeech**: entire dataset considered
- **GTZan**: 2 files for each musical genre
- Motivation: the three datasets cover a wide range of possible audio content

Techniques:

- **Griffin-Lim**
- **MelGan**
- **HifiGan, Uniglow, SqueezeNet**: part of the “NeMo” toolkit in the Text To Speech (TTS) collection, developed by NVIDIA. These models are not designed to accommodate easily the computation of the input mel-spectrogram from outside their pipeline.
- Limitations: all DL models are trained on speech only.
- Other techniques: Deep Griffin-Lim, UnivNet. These are not available as pre-trained.

Metrics

- Used to compare the original audio with the reconstructed one:
  - Mean logarithmic spectral distance (**LSD**)
  - Itakura-Saito distance (**ISD**)
  - Root mean square error (**RMSE**)
  - Cosine similarity
  - Pearson’s correlation coefficient (**PCC**)
- Other metrics: we could use other techniques that evaluate the perceptual audio quality (e.g. **PESQ**), leaving behind the comparison with the original audio recording

References

- GitHub Repository (source code, bibliography): https://github.com/michele-perrone/SpectrogramPlayer
- Google Drive folder (datasets, Google Colab notebooks): https://drive.google.com/drive/folders/1PYkuReth5-53ZeL4B551olfubPTSbo2p?usp=sharing