Towards an articulatory-driven neural vocoder for speech synthesis
Marc-Antoine Georges1,2, Pierre Badin1, Julien Diard3, Laurent Girin1, Jean-Luc Schwartz1, Thomas Hueber1

1Univ. Grenoble Alpes, CNRS, GIPSA-lab, 38000 Grenoble, France.
2Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France.

Introduction

- Classical articulatory synthesizers generate sound with physical models driven by interpretable parameters.
- Machine learning based approaches approximate those physical processes by training a statistical model on parallel articulatory-acoustic recordings, usually using EMA as in (Toda et al, 2008, Zen et al., 2010, Bocquelet et al., 2014).
- We propose a new machine learning approach which incorporates an articulatory model to get interpretable input parameters and a neural vocoder.

Synthesizing method

- Previous approach (Bocquelet et al., 2014)
  - EMA coordinates
  - Articulatory-to-acoustic mapping (DNN)
  - Vocoder (MLSA)
  - Sound
- Proposed approach
  - Articulatory parameters
  - Articulatory-to-acoustic mapping (DNN)
  - Neural vocoder (LPCNet)
  - Sound

- Articulatory parameters reflect elementary articulators activity.
- A Deep Neural Network (DNN), modeling the articulatory-to-acoustic mapping, translates them to filter parameters.
- Together with source parameters, they are sent to a neural vocoder that generates the final waveform.
- Each component is trained with EMA and audio recordings of a reference speaker.

Articulatory model

- The articulatory model follows a Maeda inspired guided Principal Component Analysis (Maeda, 1990), adapted for EMA data (Serrurier et al., 2012).
- Used to translate recorded EMA coordinates to parameters reflecting the activity of the main elementary articulators.

Neural vocoder LPCNet (Valin et al., 2019)

- LPCNet is able to produce high-quality speech sound from a limited set of parameters, describing the activity of the source (f0 & periodicity) and the filter (cepstrum coefficients).
- The explicit dissociation of its inputs between source and filter parameters makes it well suited for the proposed approach.

Experiment

- Using a parallel audio-EMA recordings dataset, we:
  - Fine-tuned a pre-trained LPCNet version to our reference speaker.
  - Trained 2 DNN to predict the filter part of LPCNet parameters (cepstrum coefficients) from EMA for the first network, and articulatory parameters for the second one.
  - Resynthesized test items from the dataset by chaining LPCNet with the DNN articulatory-to-acoustic models.

  Dataset: 1,109 productions (sustained vowels, VCV, words, sentences)

- We compared those resynthesis using PEMO-Q to a LPCNet baseline:
  - The high PEMO-Q value shows the good quality of EMA-based resynthesis.
  - Replacing EMA values by articulatory parameters does not degrade much the resynthesis

Resynthesis examples available at: https://georges.ma/publications/issp2020-abstract/

Summary

We proposed a machine learning based approach to create an articulatory synthesizer from an EMA-audio dataset, that:

- Supported by an articulatory model, provides interpretable input parameters.
- Successfully implements the neural vocoder LPCNet.

Perspectives

- The articulatory synthesizer will be evaluated subjectively with perceptive tests.
- This approach only relies on neural networks and could be adapted to create an end-to-end neural articulatory synthesizer.

References


Acknowledgments

This work has been partially supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).