Towards an articulatory-driven neural vocoder for speech synthesis

Marc-Antoine Georges[1,2], Pierre Badin[1], Julien Diard[2], Laurent Girin[1], Jean-Luc Schwartz[1], Thomas Hueber[1]

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

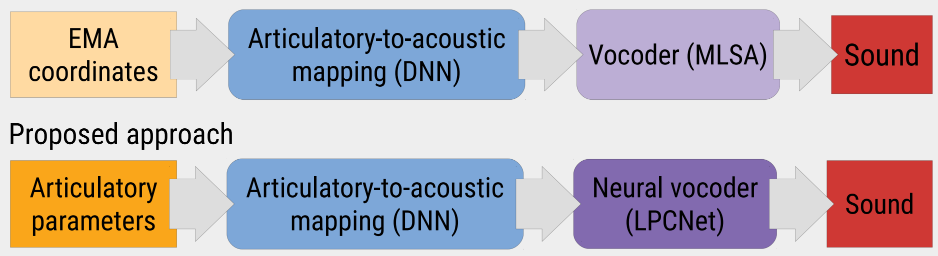
Introduction

- 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)

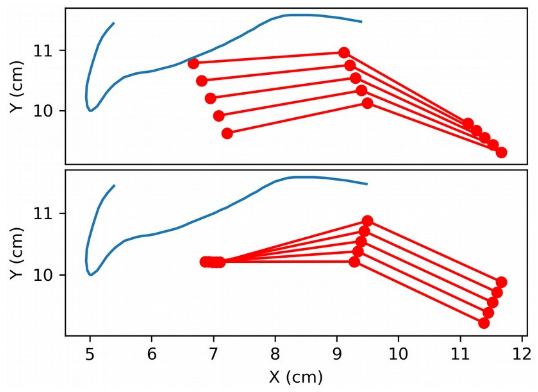
Proposed approach



- 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

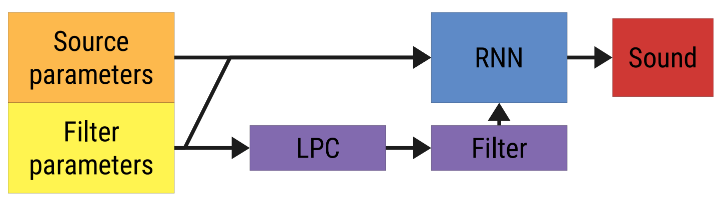
- The articulatory model follows a Maeda inspired guided Principal Component Analysis (Maeda, 1990), adapted for EMA data (Serrurier et al., 2012).
- Used to translate recorded EMA coordinates to parameters reflecting the activity of the main elementary articulators.



Influence of the *JawHeight* (top) and *TongueDorsum* (bottom) parameters on tongue position

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.



LPCNet architecture overview

Experiment

Using a parallel audio-EMA recordings dataset, we:

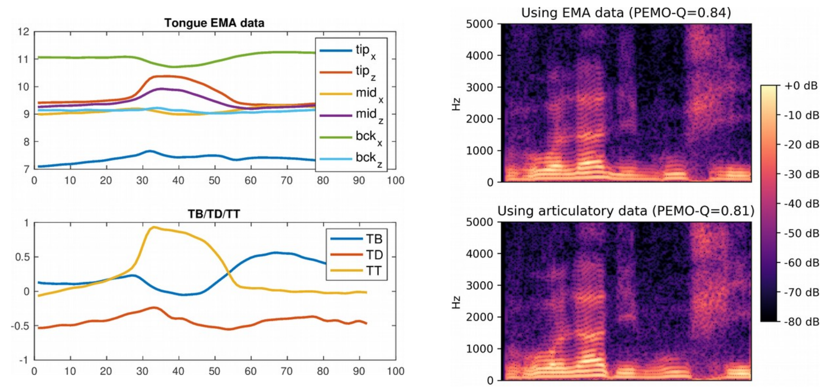
- Fine-tuned a pre-trained LPCNet version to our reference speaker.
- Trained 2 DNN to predict the filter part of LPCNet parameters (cepstrum coefficients) from EMA for the first network, and articulatory parameters for the second one.
- Resynthesized test items from the dataset by chaining LPCNet with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

- Bocquelet, F., Hueber, T., Girin, L., Badin, P., and Yvert, B. (2014). Robust articulatory speech synthesis using deep neural networks for BCI applications. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (September) :2288–2292.
- Maeda, S. (1990). Compensatory Articulation During Speech : Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model. *Speech Production and Speech Modelling*, pages 131–149.
- Serrurier, A., Badin, P., Barney, A., Boë, L. J., and Savariaux, C. (2012). The tongue in speech and feeding : Comparative articulatory modelling. *Journal of Phonetics*, 40(6) :745–763.
- Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical mapping between ar- ticulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3) :215–227.
- Valin, J.-M. M. and Skoglund, J. (2019). LPCNet : Improving neural speech synthesis through linear prediction. (1) :5891–5895.
- Zen, H., Nankaku, Y., and Tokuda, K. (2011). Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech and Language Processing*, 19(2) :417–430.

Acknowledgments

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