Glottal inverse filtering based on articulatory synthesis and deep learning

Ingo Langheinrich, Simon Stone, Xinyu Zhang, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany

Introduction

- The glottis is inaccessible without invasive and obstructive equipment. The glottal flow is therefore very difficult to measure directly.
- However, the glottal flow carries versatile information and might be helpful e.g., for the detection of voice disorders, diseases, speaker identification and verification.
- We propose a new method to estimate the glottal vocal tract excitation from speech signals based on deep learning.
- Since natural reference data is unobtainable at the required scale for deep learning, we used the articulatory speech synthesizer VocalTractLab to generate a large dataset.

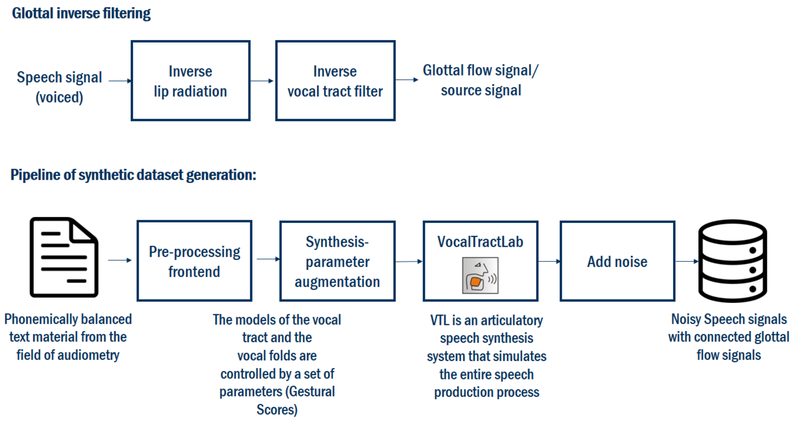

Figure 0: Synthesis Pipeline

Datasets

Since articulatory synthesis is entirely parametric, we created a number of variants of each synthesized sentence by changing:

- the phonation type,
- the mean fundamental frequency,
- the speech rate,
- and by adding noise.

We used two different datasets for evaluation:

- OPENGLOT dataset for objective evaluation,
- the BITS Unit Selection corpus for a qualitative sample and plausibility check of the performance on real human speech.

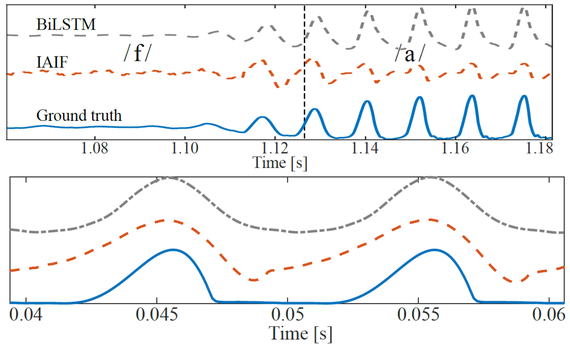

Figure 1: Top: Example glottal flow segments from the VTL synthetic speech. Bottom: Glottal flow segments of isolated vowel sounds of the OPENGLOT dataset (repository III, [a], f0 = 100 Hz).

Model

- The neuronal model was trained entirely on synthetic signals and maps the time-domain representations of speech signals to the corresponding time domain of the glottal flow derivative:
  - For the transformation, we used a bidirectional recurrent neuronal network with long short-term memory units (BiLSTM)
- The model performance was evaluated on the OPENGLOT dataset regarding open-quotient and cross correlation between the ground truth and predicted glottal flow signals. The best model was compared to the reference algorithm IAIF.

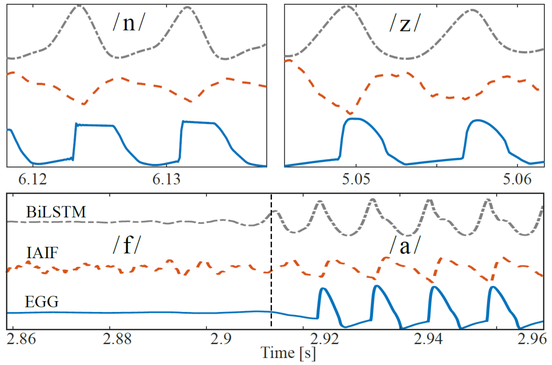

Figure 2: Example glottal flow and EGG segments for the phonemes /n, z/ and an unvoiced/voiced transition from the BITS corpus of natural speech.

Results and Conclusion

- The trained BiLSTM model produced much more plausible glottal flow signals on continuous utterances than IAIF without any manual intervention.
- IAIF achieved slightly higher marks on the computationally synthesized isolated vowel sounds contained in OPENGLOT (Repository I-II).
- Our proposed model produced a more accurate estimation using OPENGLOT's physically synthesized signals (Repository III).
- While IAIF requires the manual specification of some parameters based on the speech signal content, the BiLSTM model has no free parameters