Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks

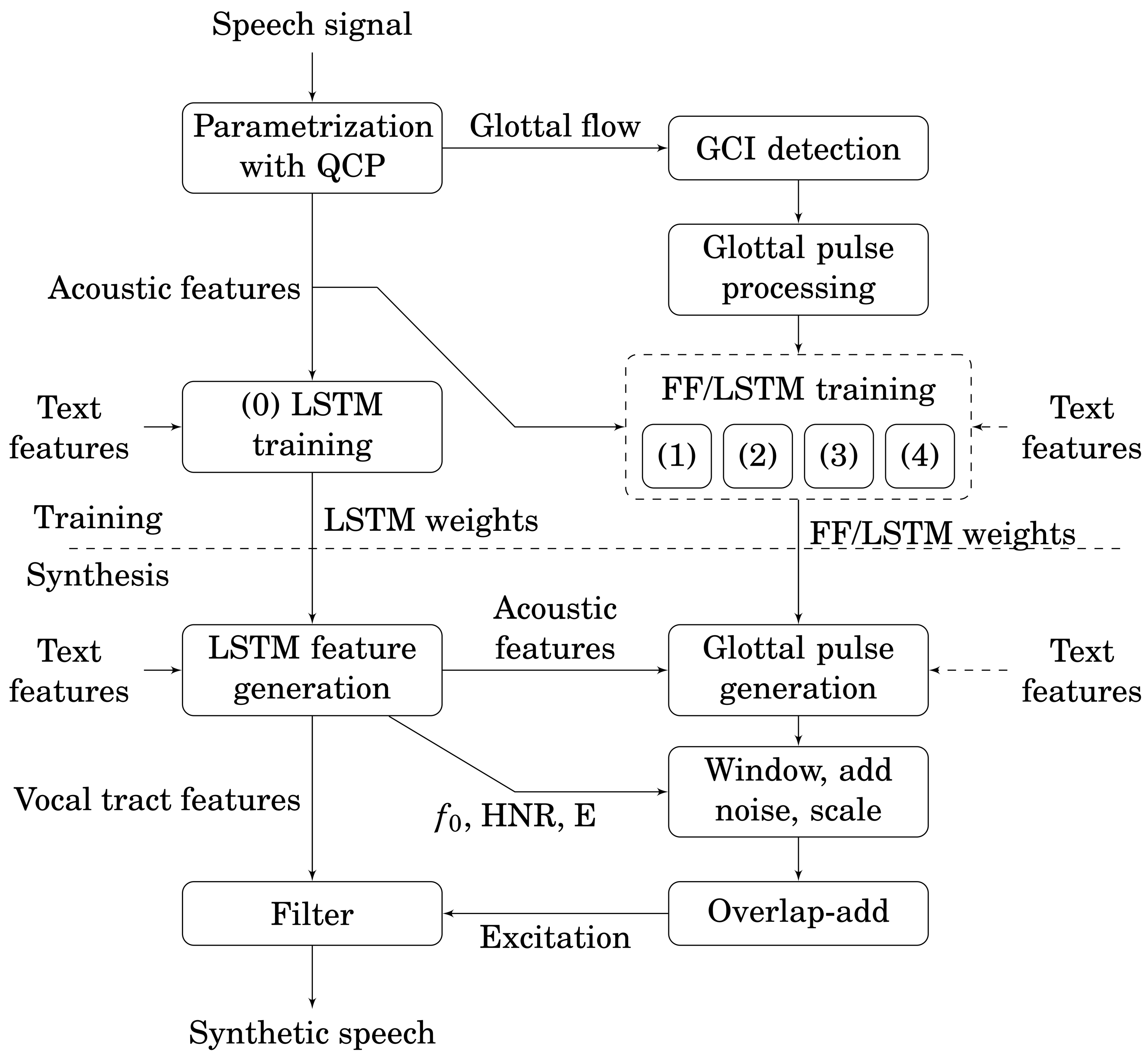
Aalto University
School of Electrical Engineering

Lauri Juvela[1,2], Xin Wang[2,3], Shinji Takaki[2], Manu Airaksinen[1], Junichi Yamagishi[2,3,4], Paavo Alku[1]
[1]Aalto University, [2]National Institute of Informatics, [3]Sokendai University, [4]University of Edinburgh

1 Introduction

- Feedforward DNN [1] and LSTM RNN [2] have been successfully used for acoustic modelling in SPSS
- DNN-based excitation models for glottal vocoding [3] have been shown to increase speech quality [4]
- **This work**: Does LSTM improve excitation modelling and can glottal waveforms be predicted directly from text?

2 Speech synthesis system



- Four different networks were trained for modelling glottal waveforms, the acoustic model (text to acoustic) was shared between the systems

ID System	Input	Output	Network
(0) TXT-LSTM-AC	TXT	AC	LSTM
(1) AC-FF-GL	AC	GL	FF
(2) AC-LSTM-GL	AC	GL	LSTM
(3) TXT-LSTM-GL	TXT	GL	LSTM
(4) TXT+AC-LSTM-GL	TXT + AC	GL	LSTM

2.1 Acoustic features

- Acoustic features derived with QCP inverse filtering
- Analysis and synthesis with a modified GlottHMM [5] vocoder
- Dynamic features were used in the neural networks

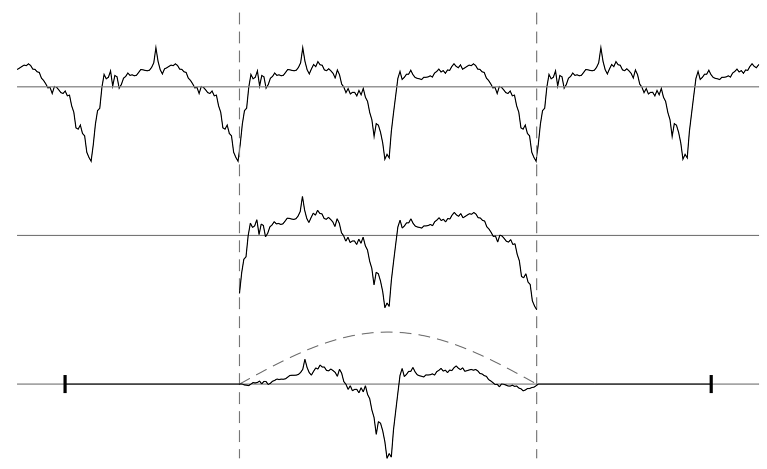
Feature	dim.	Δ dim.
f_0	1	3
VUV	–	1
Energy	1	3
LSF Vocal tract	30	90
LSF Glottal source	10	30
HNR	5	15
total	47	142

2.2 Text features

- Combilex-based linguistic features contain phoneme, syllable, word, phrase, and sentence level information
- 396-dimensional input text features are created with forced alignment

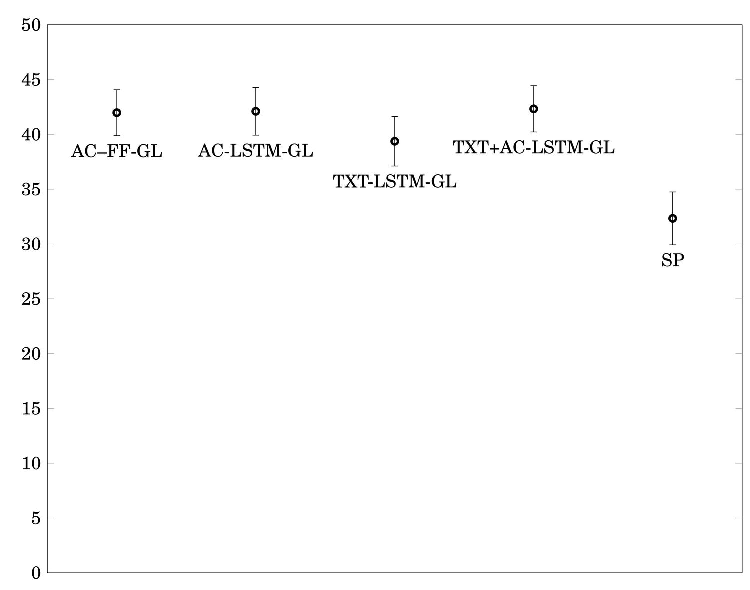
2.3 Glottal waveforms

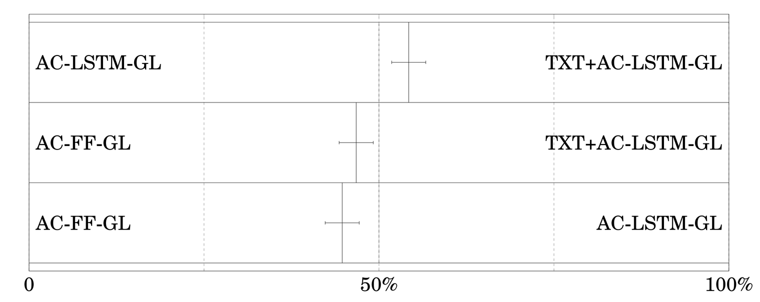
- Obtain glottal flow derivative waveforms with QCP
- Model two-pitch-period segments delimited by GCI
- Window and zero-pad to get fixed length for DNN



3 Listening experiments

- MUSHRA-like test with natural speech with same content as reference
- 30 native English listeners participated
- DNN-based methods outperform single pulse excitation (SP), text-only input is not significantly worse
- Pairwise preference test for the top three methods
- LSTM-based systems outperform the FF-DNN system





4 Summary

- Glottal vocoding was integrated to a framework where text and acoustic features can be used to predict glottal excitation waveforms
- Glottal waveforms can be predicted reasonably well by using only text information
- Use of LSTM slightly improves the excitation modelling performance

References

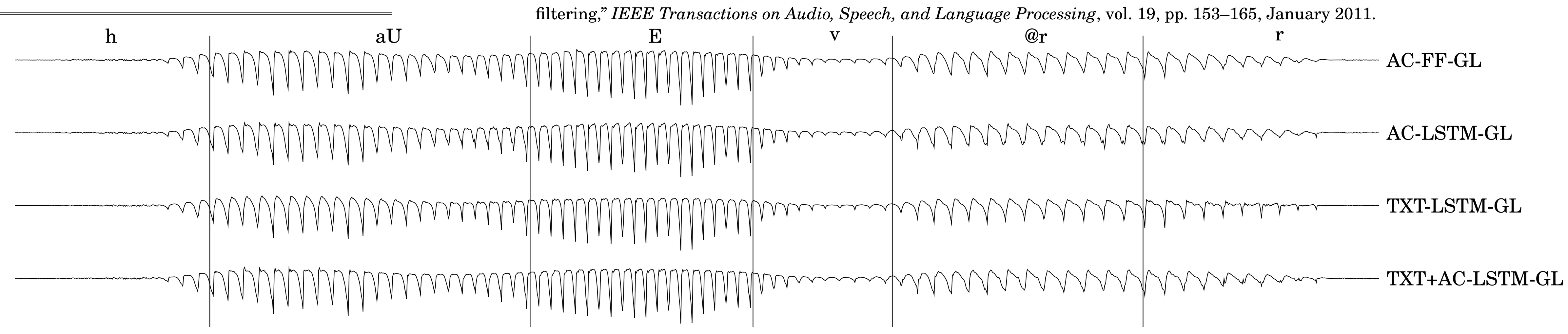
[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. of ICASSP*, pp. 7962–7966, May 2013.

[2] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks.," in *Interspeech*, pp. 1964–1968, 2014.

[3] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *22nd European Signal Processing Conference (EUSIPCO)*, (Lisbon, Portugal), September 2014.

[4] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *"Proc. of ICASSP"*, pp. 5120–5124, Mar. 2016.

[5] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 153–165, January 2011.



Examples of generated glottal excitation waveforms after overlap-add with the word "however"

Department of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, Finland

This work was supported by the Academy of Finland (proj. no. 256961 and 284671), the European Union TEAM-MUNDUS scholarship (TEAM1400081), and the EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF).
E-mail: lauri.juvela@aalto.fi