



Aalto University
School of Electrical
Engineering

Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system

Lauri Juvela¹, Bajibabu Bollepalli¹, Junichi Yamagishi^{2,3}, Paavo Alku¹

¹ *Aalto University, Department of Signal Processing and Acoustics, Finland*

² *National Institute of Informatics, Japan*

³ *The Centre for Speech Technology Research, University of Edinburgh, United Kingdom*

lauri.juvela@aalto.fi

Outline

Introduction

- Glottal vocoding
- Text-to-speech system
- Excitation processing for neural net

Reducing mismatch

- Training with generated acoustic features
- Inverse filtering with generated vocal tract filter

Experiments

- Objective measures
- Listening test and audio samples

Conclusion

Introduction

- Recent interest in integrating waveform generation and acoustic modeling in TTS ("neural vocoding", end-to-end)
- Why predict speech signal directly? We already can predict parametric envelopes and excitation signals are simpler to model
- Previously we proposed using DNN to generate glottal excitation waveforms [Juvela et al., 2016]
- Problem: **mismatch** resulting from training acoustic and excitation models separately, but using them as a stacked model

Glottal vocoding

- Vcoders typically used in text-to-speech synthesis are based on feeding a voice **source** signal excitation to a vocal tract **filter**
- Conventional approach: use impulse train excitation

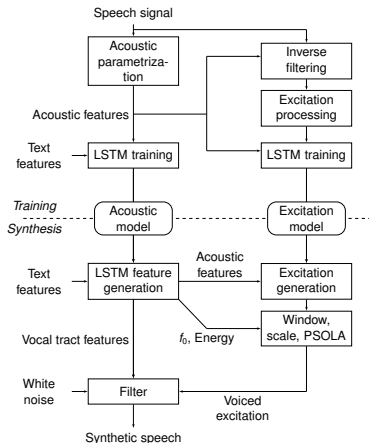


- Modeling **glottal waveforms** (volume velocity through larynx) avoids making strong simplifying assumptions



- Glottal excitation waveforms can be estimated with glottal inverse filtering (we use QCP [Airaksinen et al., 2014])

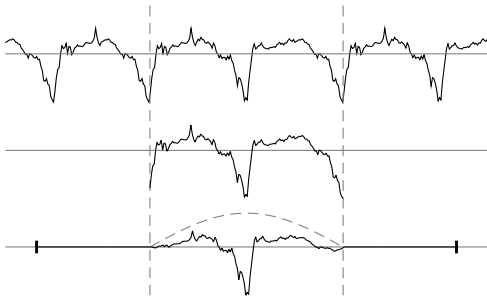
Text-to-speech system – (1) AC-GL



- **Left side:** acoustic model for mapping linguistic specification to acoustic features
- **Right side:** excitation model for mapping acoustic features to glottal excitation waveforms
- Re-structure this to reduce mismatch between training and test time

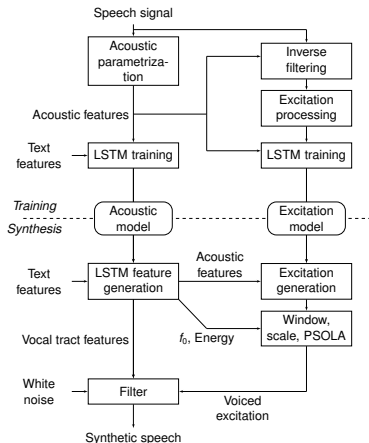
Excitation processing for neural net

- Start with the estimated glottal flow derivative
- Center at a glottal closure instant and take two pitch periods
- Cosine window and zero-pad to fixed length



Training with generated acoustic inputs – (2)

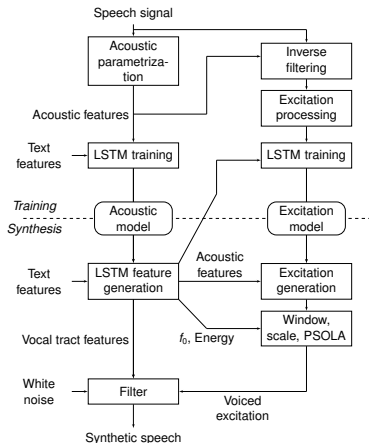
GEN-AC-GL



- Only generated acoustic features are available at synthesis time
- Train excitation model with generated acoustic features

Training with generated acoustic inputs – (2)

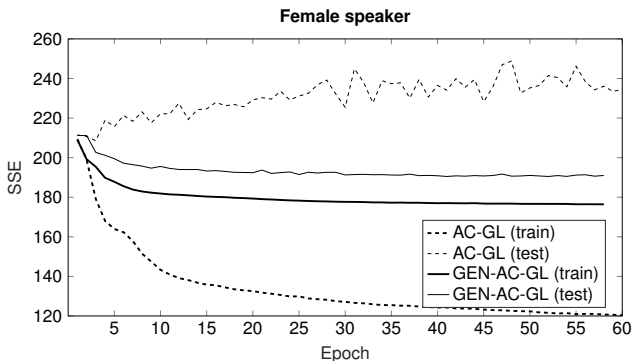
GEN-AC-GL



- Only generated acoustic features are available at synthesis time
- Train excitation model with generated acoustic features

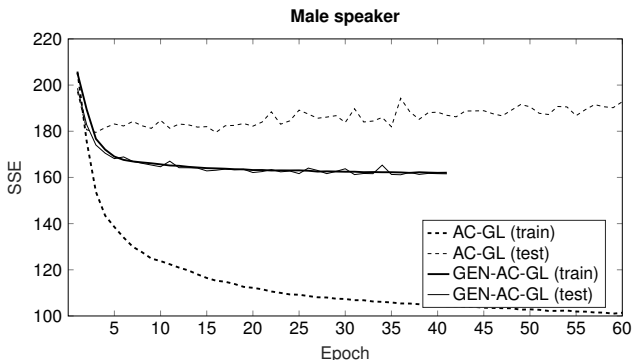
Excitation model training and test errors

- Excitation model overfits to natural acoustic input features and fails to generalize on generated input (dashed lines)
- Training on generated acoustic features works as expected



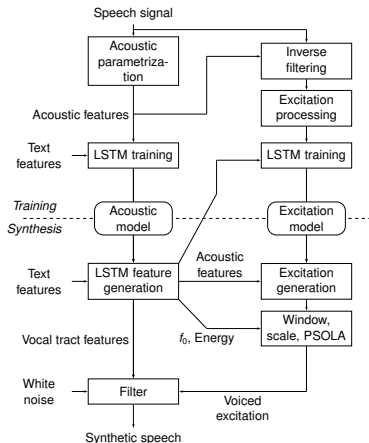
Excitation model training and test errors

- Excitation model overfits to natural acoustic input features and fails to generalize on generated input (dashed lines)
- Training on generated acoustic features works as expected



Re-inverse filtering with generated vocal tract –

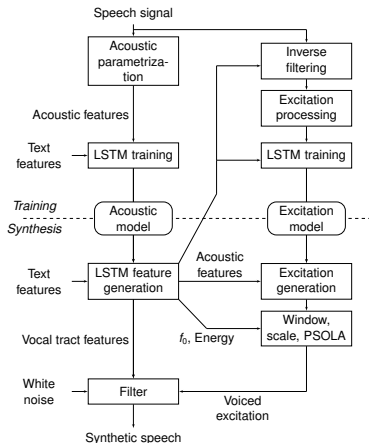
(3) GEN-AC-GL-RE



- Generated vocal tract filter differs from original and chance for perfect reconstruction is lost
- Perform inverse filtering with generated vocal tract filter
- Similar to speech coding, where inverse filtering is done with quantized filter parameters

Re-inverse filtering with generated vocal tract –

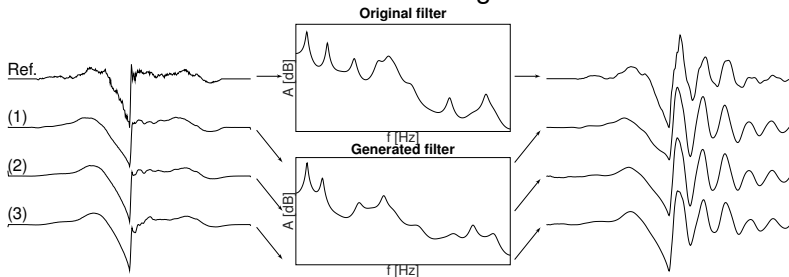
(3) GEN-AC-GL-RE



- Generated vocal tract filter differs from original and chance for perfect reconstruction is lost
- Perform inverse filtering with generated vocal tract filter
- Similar to speech coding, where inverse filtering is done with quantized filter parameters

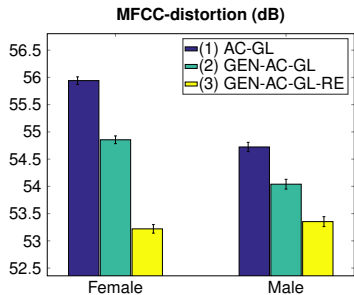
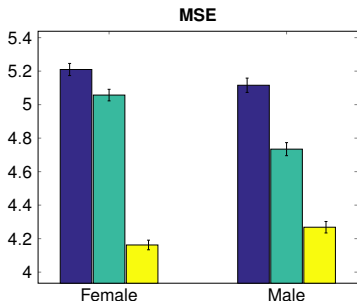
Measuring error in speech waveform domain

- Target waveform is reference pulse filtered with original vocal tract filter
- Generated excitations are filtered with generated vocal tract



Objective measures

- Test set objective errors, mean squared error and MFCC distortion
- Using generated acoustic features for training improves performance, and re-doing inverse filtering helps further



Listening test (A-B preference)

- Listeners were asked to indicate their preference based on overall quality
- 50 listeners participated on CrowdFlower
- Audio samples available online

https://users.aalto.fi/~ljuvela/reducing_mismatch/

speaker	(1)	(2)	(3)	neutral	<i>p</i> -value
Female	7.98	14.79		77.2	0.0008
	9.30		12.98	77.7	0.0464
Male		5.81	12.79	81.4	0.0002
	6.30	12.60		81.1	0.0007
	9.34		11.09	79.6	0.2176
		7.36	12.02	80.6	0.0105

Limitations

- Training models with MSE regression will always regress towards mean
- Result: loss of stochastic variability and high frequencies
- In this paper, we still applied additive harmonic-to-noise ratio (HNR) modification and excitation spectral matching
- In future, use generative models which enable sampling from a distribution (e.g. generative adversarial networks)

Conclusion

- Proposed two methods to reduce mismatch in training glottal excitation models for TTS
- Major mismatch between using original or generated acoustic inputs in training
- Main contribution is studying what kind of data (input acoustic features and target excitation waveforms) should be used
- For full effect, more powerful generative models are needed

References

- [Airaksinen et al., 2014] Airaksinen, M., Raitio, T., Story, B., and Alku, P. (2014). Quasi closed phase glottal inverse filtering analysis with weighted linear prediction.
Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 22(3):596–607.
- [Juvela et al., 2016] Juvela, L., Bollepalli, B., Airaksinen, M., and Alku, P. (2016). High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network.
In *"Proc. of ICASSP"*, pages 5120–5124.

