Aalto University

UTILIZING GLOTTAL SOURCE PULSE LIBRARY FOR GENERATING IMPROVED EXCITATION SIGNAL FOR HMM-BASED SPEECH SYNTHESIS

Tuomo Raitio[1], Antti Suni[2], Hannu Pulakka[1], Martti Vainio[2], and Paavo Alku[1]

[1]Department of Signal Processing and Acoustics, Aalto University

[2]Department of Speech Sciences, University of Helsinki

ICASSP 2011

Contents

I. Background

- The ultimate goal of text-to-speech (TTS) is to generate natural sounding expression from arbitrary text
- Two major TTS trends:

Unit selection

- ❑ Based on concatenating prerecorded acoustical units
- ❑ Yields (almost) natural quality
- ❑ Poor adaptability to speaking styles, speaker characteristics and emotions

Statistical

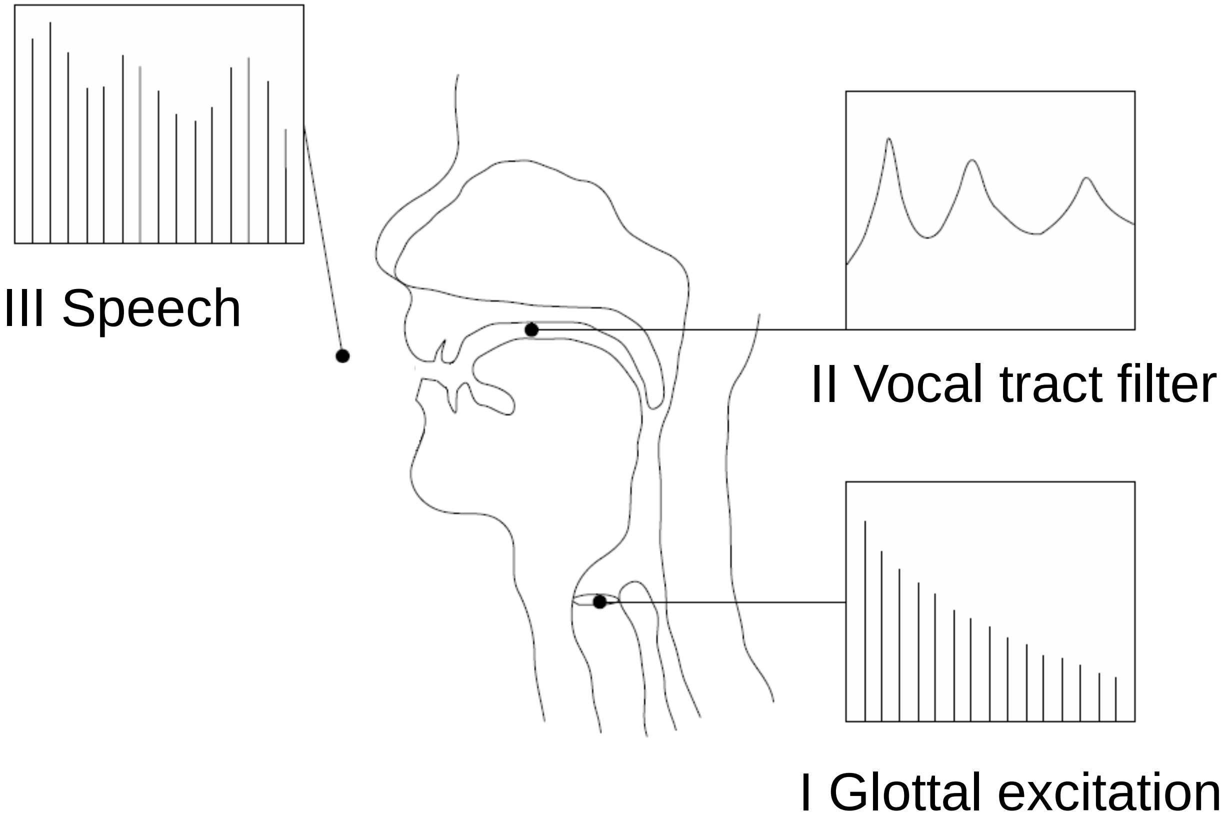
- ❑ Based on modeling speech parameters with Hidden Markov Models (HMMs)
- ❑ Better adaptability to speaking styles, speaker characteristics and emotions

Problem: Current HMM-based synthesizers suffer from degraded naturalness in speech quality

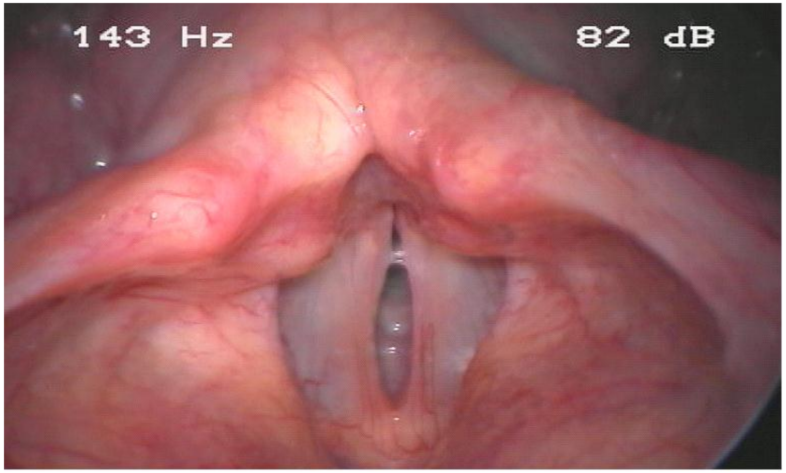
Our approach:

1. Speech is decomposed into the glottal source signal and the vocal tract transfer function
2. Glottal source is further decomposed into several parameters and a glottal pulse library
3. Parameters are modeled in HMMs
4. In synthesis, source signal is reconstructed from the selected glottal pulses and the filtered with the vocal tract filter to create speech

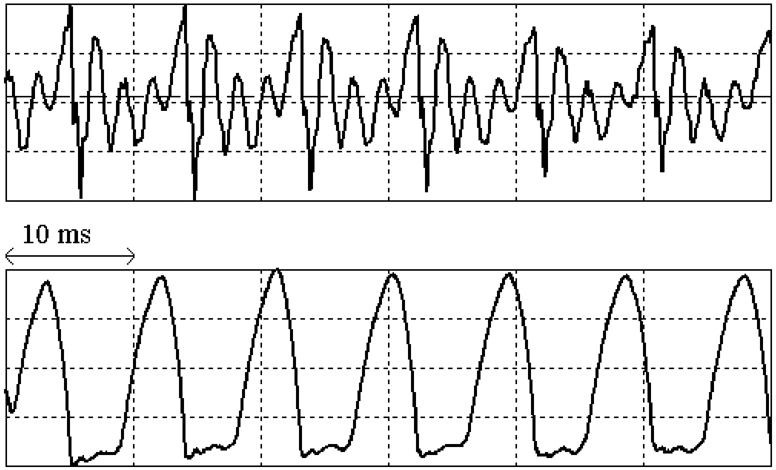
II. Speech Production Mechanism



Glottal Source



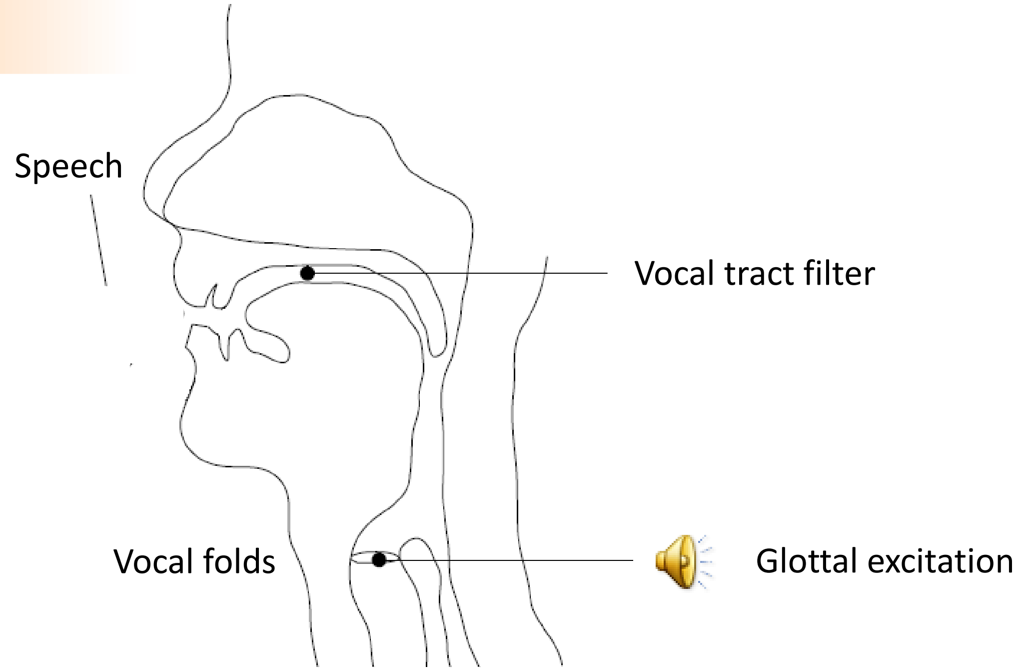
Vibrating vocal folds.



Speech pressure waveform (upper panel) and estimated glottal excitation (lower panel).

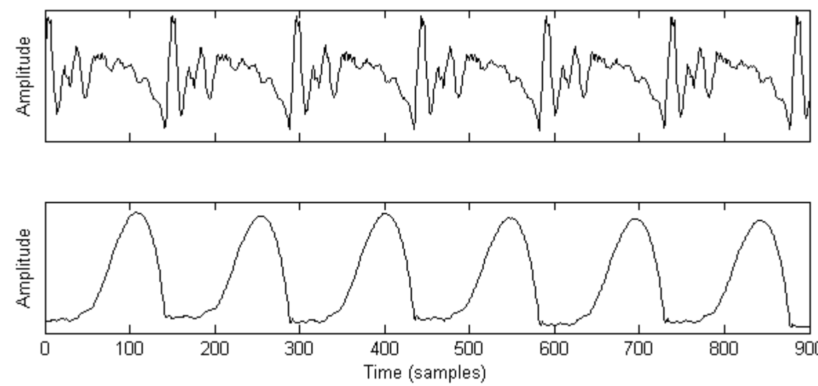
Glottal Source

Glottal inverse filtering estimates the glottal flow and the vocal tract filter from a speech signal

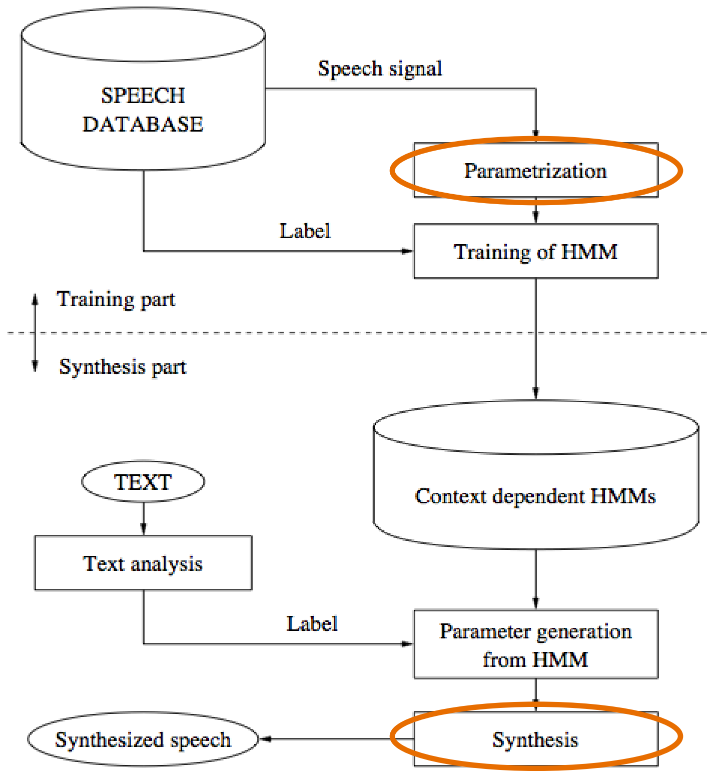


Speech signal

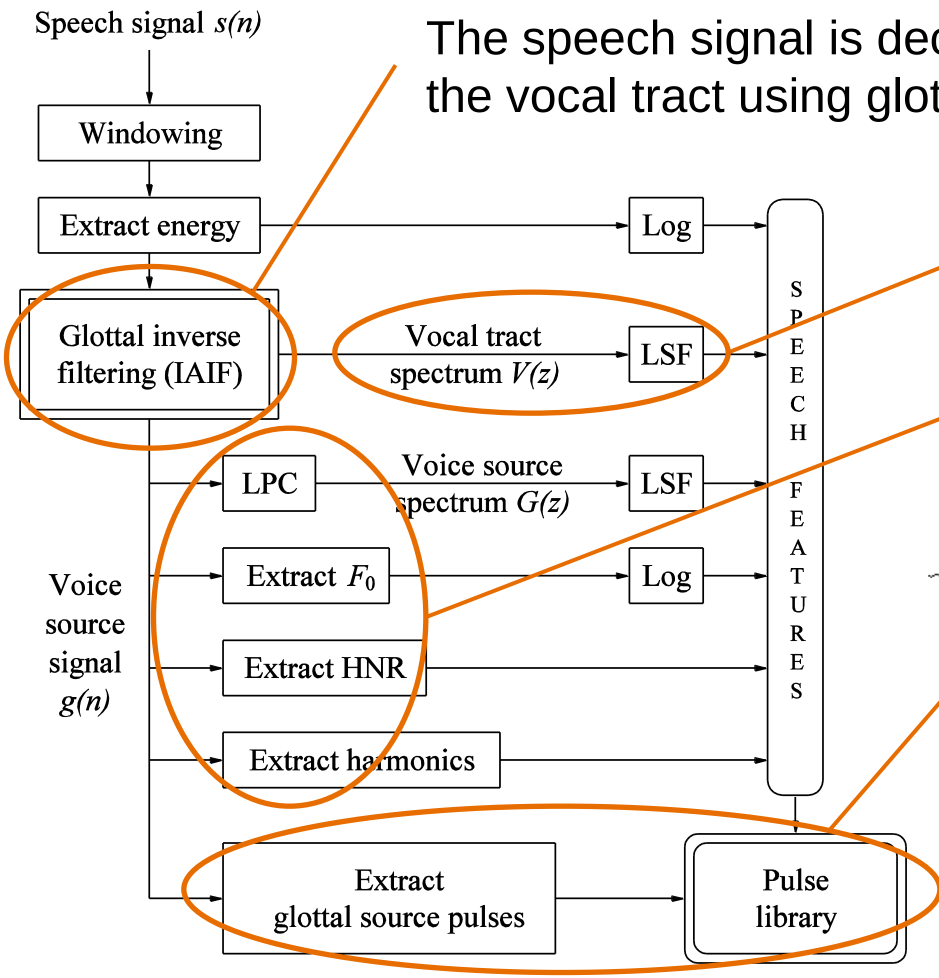
Estimated glottal flow signal



III. Speech Synthesis System



Speech Parameterization



The speech signal is decomposed into the glottal source and the vocal tract using glottal inverse filtering (IAIF) [1]

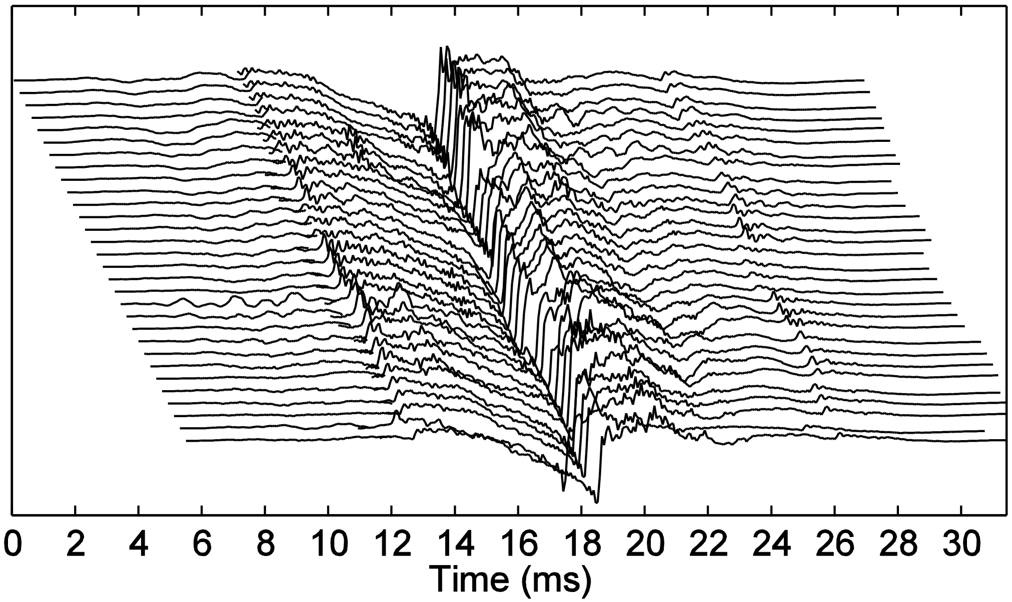
Vocal tract filter is parameterized with LSFs

The glottal source signal is further parameterized into F0, HNR, spectral tilt (LPC), and harmonic magnitudes

Finally, glottal pulses of the source signal are extracted and stored to a library.

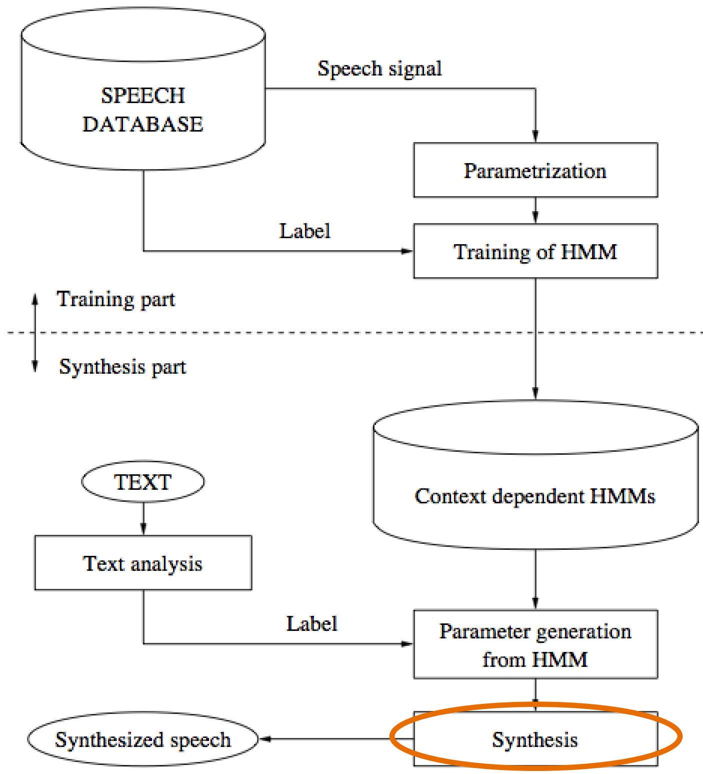
Pulse Library

Consists of hundreds or thousands of glottal flow pulses (and the corresponding voice source parameters)



Windowed glottal volume velocity pulse derivatives from the pulse library of a male speaker

Synthesis

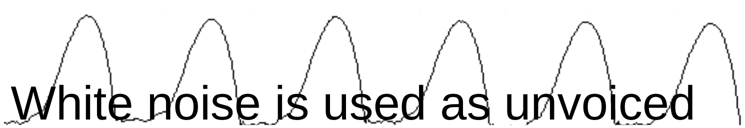


SPEECH DATABASE
Speech signal
Parametrization
Label
Training of HMM
Training part
Synthesis part
Context dependent HMMs
TEXT
Text analysis
Label
Parameter generation from HMM
Synthesized speech
Synthesis

Synthesis

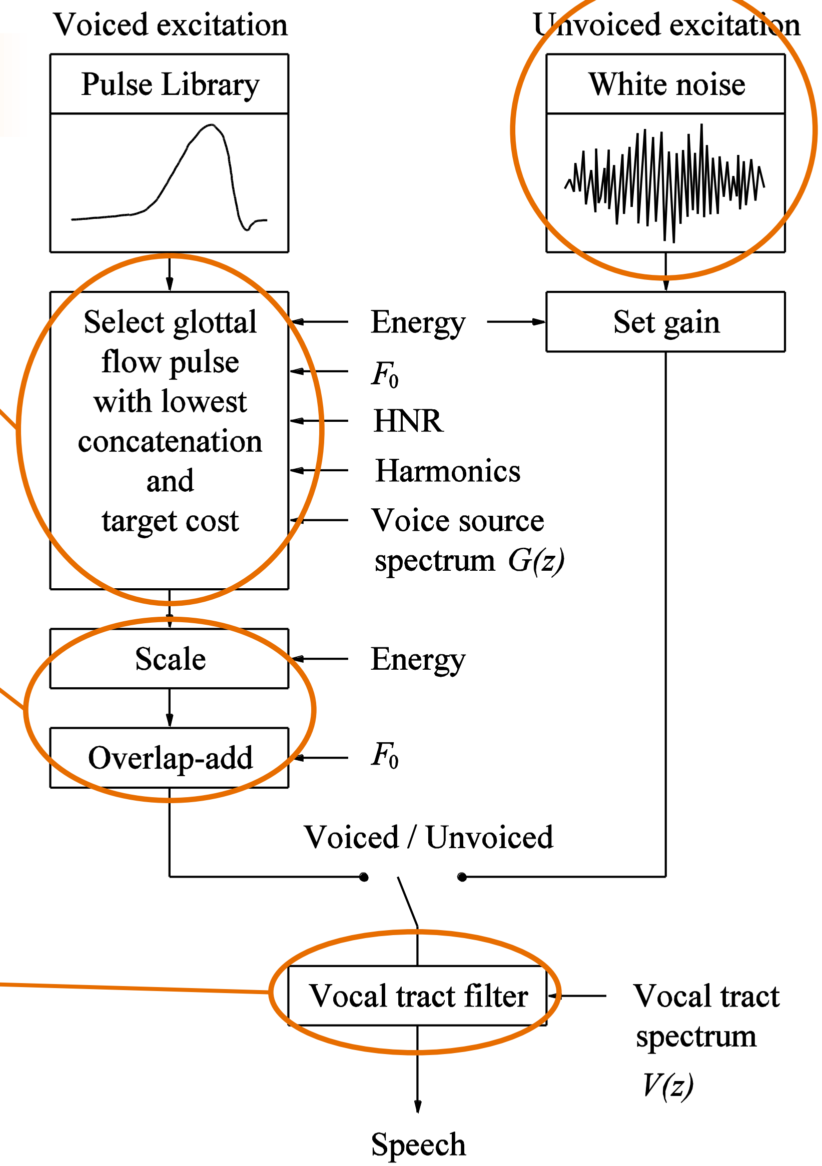
In synthesis stage, excitation signal is generated by selecting the best matching pulse from the library according to the source features

Pulses are modified by scaling the magnitude and then overlap-added

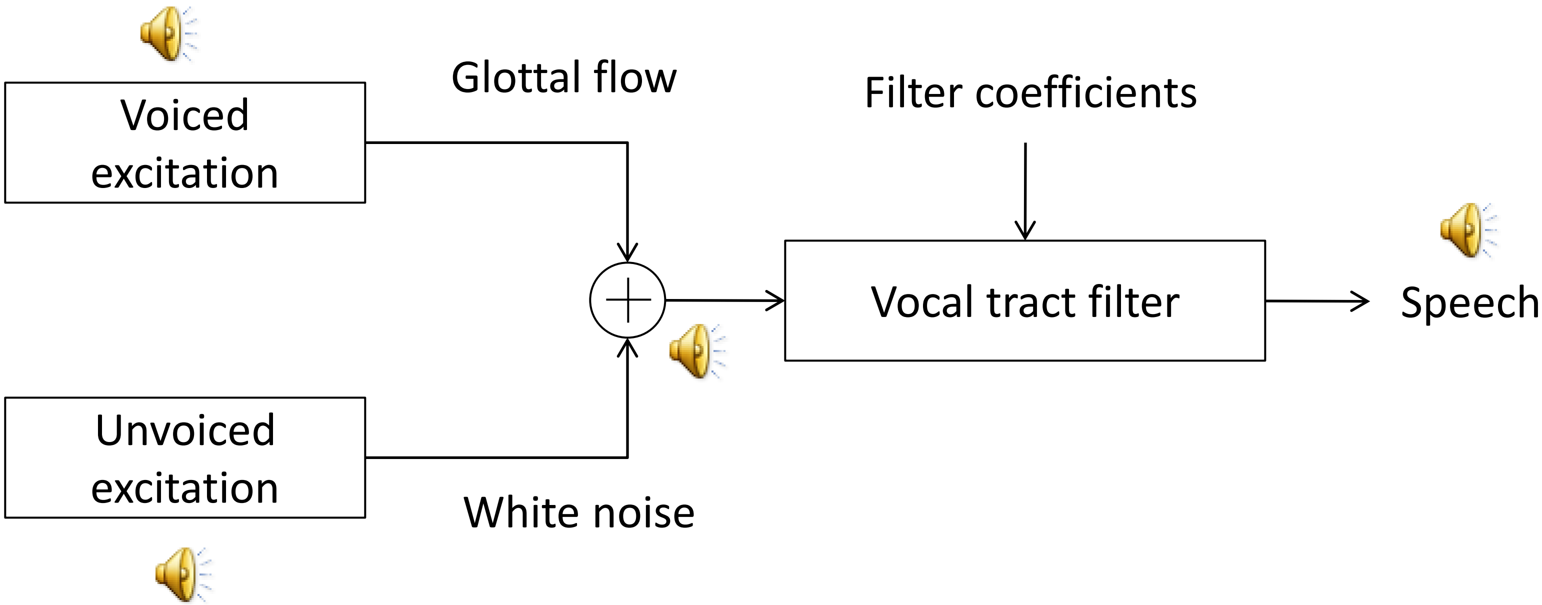


White noise is used as unvoiced excitation

Finally, excitation is filtered with the vocal tract filter to generate speech

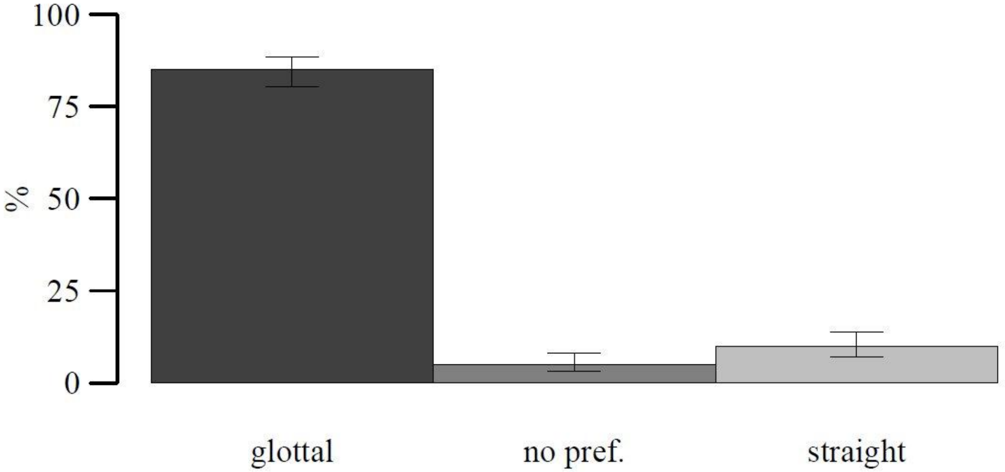


Synthesis


Glottal flow
Filter coefficients
Voiced excitation
Vocal tract filter
Speech
Unvoiced excitation
White noise

IV. Results and Samples

Previously, we have used only **one glottal pulse per utterance.**



Results of the listening test [2] comparing our synthesis method to the most widely used high-quality vocoder STRAIGHT.

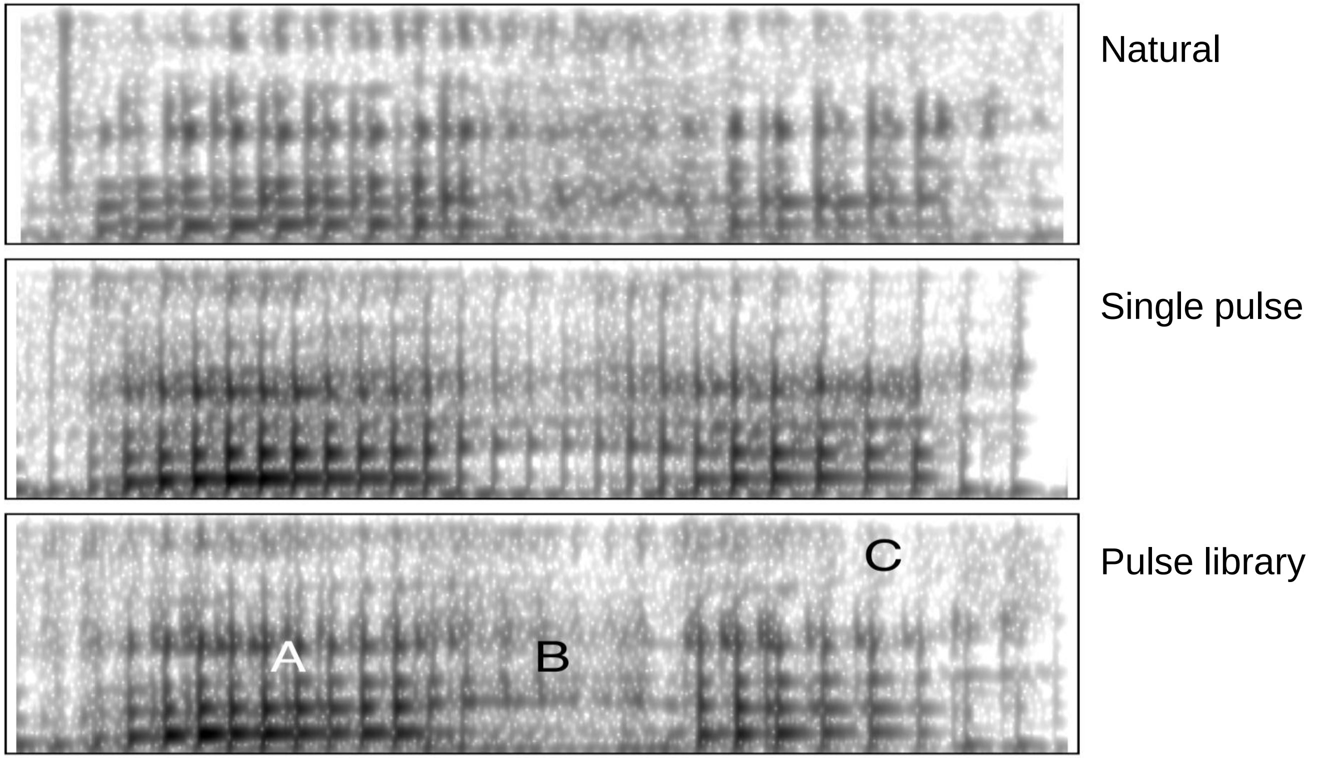
Single pulse technique

Samples:

English	Male	Female

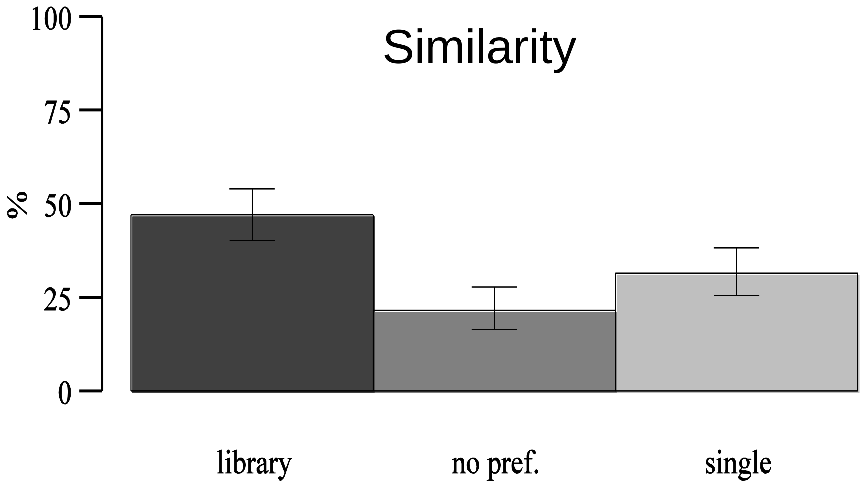
Blizzard Challenge			
English			
Mandarin			

Pulse library technique



Spectrograms (0–8000 Hz) of the word “vähän” (little). Note the improved modeling of A) diplophony B) voiced fricatives C) high frequencies.

Pulse library vs. single pulse technique



Pulse library method is slightly preferred over the single pulse technique and is more similar to the original speaker

Pulse library technique

Pulse library (ICASSP'11)	1pulse	pulselib
Finnish		
Finnish		
English		
English		

Summary

- New physiologically motivated high-quality speech synthesizer
- Allows for better reproduction and control over the speech characteristics
- Pulse library generates more natural excitation and is preferred over single pulse technique

References

[1] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Communication, vol. 11, no.* 2–3, pp. 109–118, 1992.

[2] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, “HMM-based speech synthesis utilizing glottal inverse filtering,” *IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 19, no. 1, pp. 153–165, Jan. 2011.*

Thank you!