The GlottHMM Speech Synthesis Entry for Blizzard Challenge 2010

Antti Suni, Tuomo Raitio, Martti Vainio, and Paavo Alku
(antti.suni@helsinki.fi, tuomo.raitio@tkk.fi)

25.9.2010

Aalto University

UNIVERSITY OF HELSINKI

Outline

I. Introduction

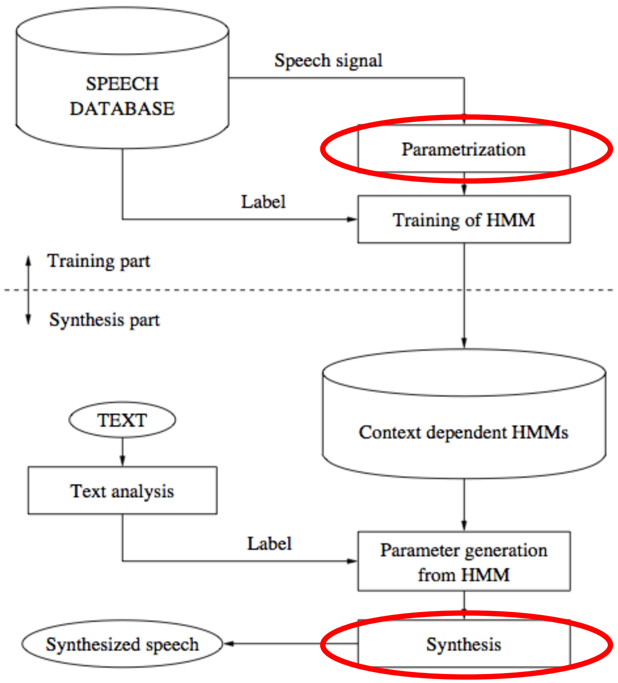
- Finnish speech synthesis has been studied in the University of Helsinki with a special emphasis on speech prosody
- Research on speech processing and acoustics has a long history in Aalto University (formerly known as Helsinki University of Technology, TKK)
- GlottHMM speech synthesis project begun in 2007 as a collaboration between TKK and University of Helsinki
- The aim was to develop a flexible high-quality HMM-based speech synthesis system

I. Introduction

- This is our first entry for Blizzard Challenge motivated by
  - Extensive comparison with other systems
  - Building non-Finnish voices
  - Using our prototype English front-end
  - Testing our prominence based prosody model

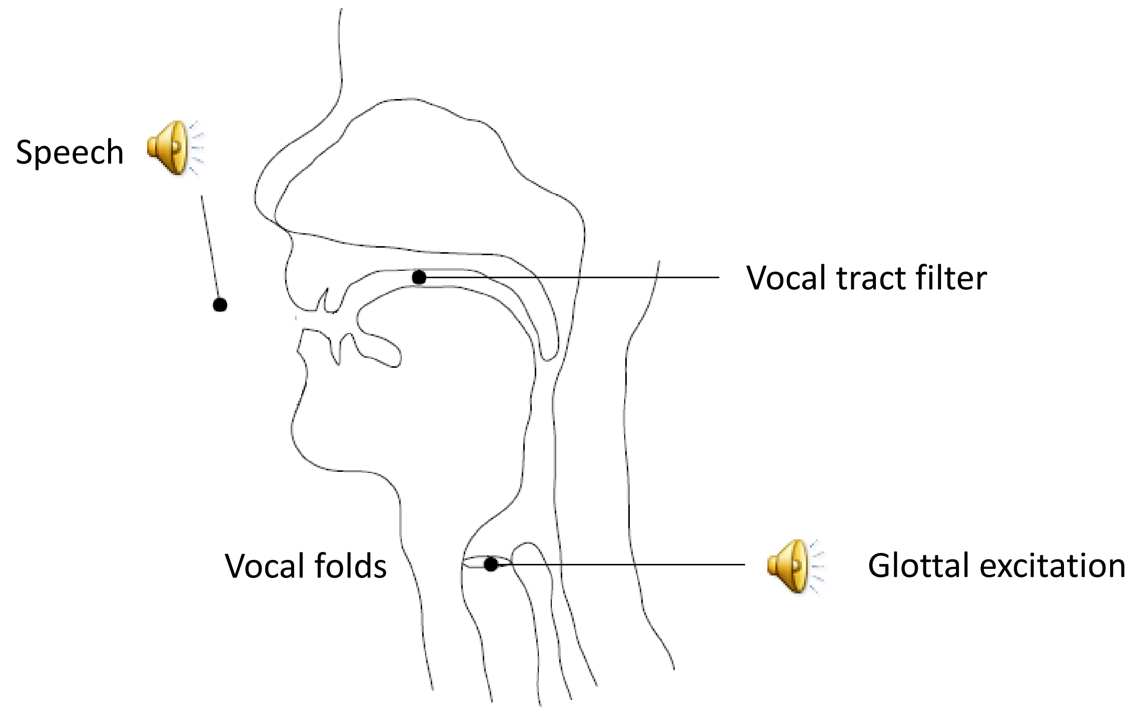
II. GlottHMM speech synthesis system

- GlottHMM is an HMM based speech synthesis system that uses a novel vocoding approach (Raitio *et al.*, 2010, in press)



II. GlottHMM speech synthesis system

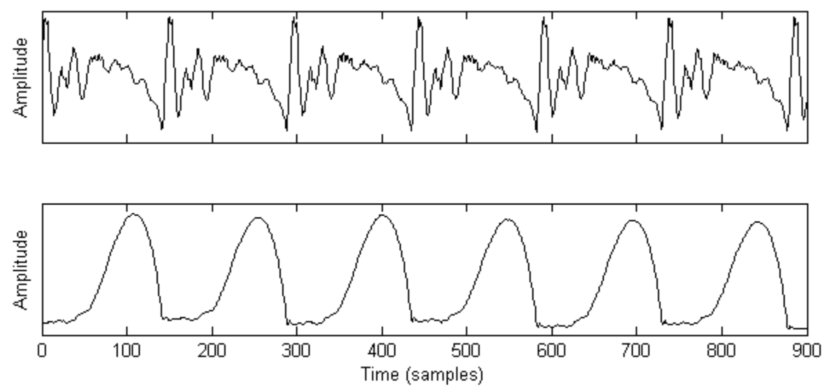
- In speech **analysis**, speech signal is decomposed into the glottal excitation and the model of the vocal tract filter by using **glottal inverse filtering** (IAIF, Alku 1992)



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

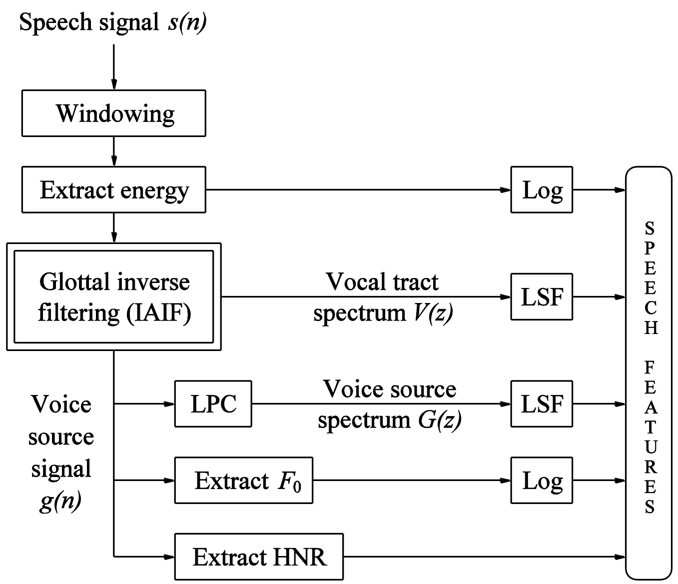
Speech signal

Estimated glottal flow signal



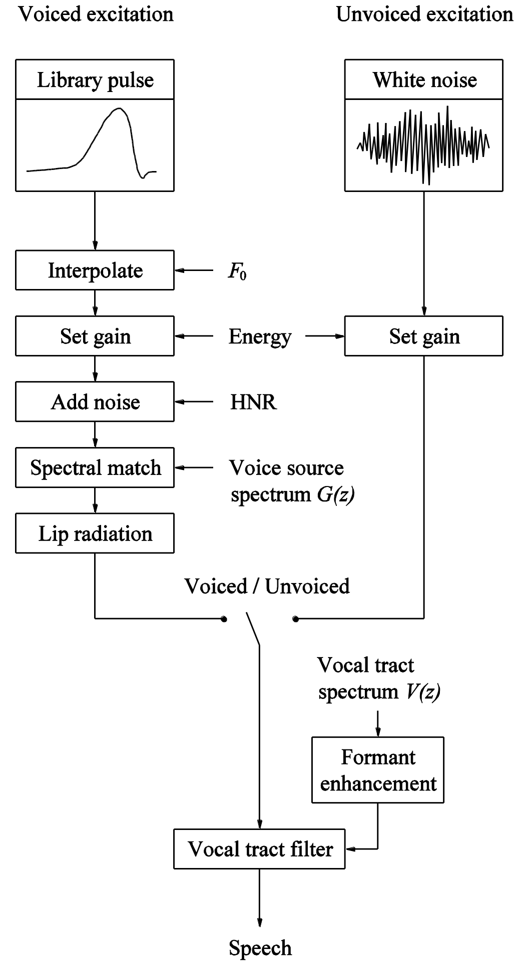
II. GlottHMM speech synthesis system

- Vocal tract is parameterized with line spectral frequencies (LSFs)
- Glottal flow signal is parameterized with
  - F0
  - Harmonic-to-noise ratio (HNR)
  - Source spectrum (LSF)
  - Gain



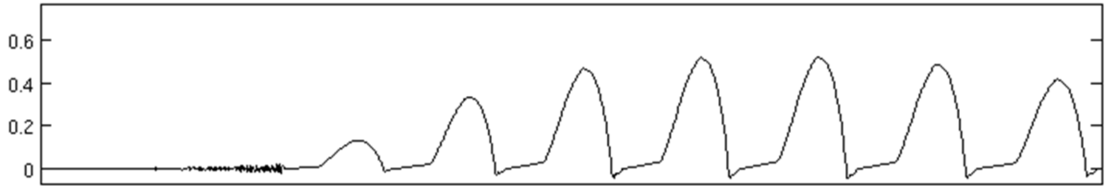
II. GlottHMM speech synthesis system

- In **synthesis** stage, excitation signal is generated by interpolating in time (**F0**) and scaling in magnitude (**Gain**) a natural **glottal flow pulse**
- Pulses are modified to match the **source spectrum** and **harmonic-to-noise ratio** by filtering and adding noise, respectively
- White noise is used as a unvoiced sound source

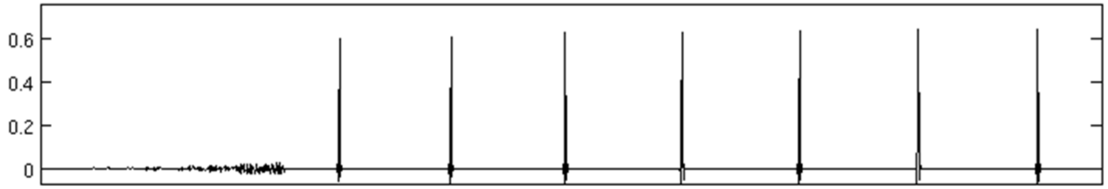


II. GlottHMM speech synthesis system

- The detailed model of the excitation should potentially allow for better control and production of prosody, speaker characteristics and speaking style

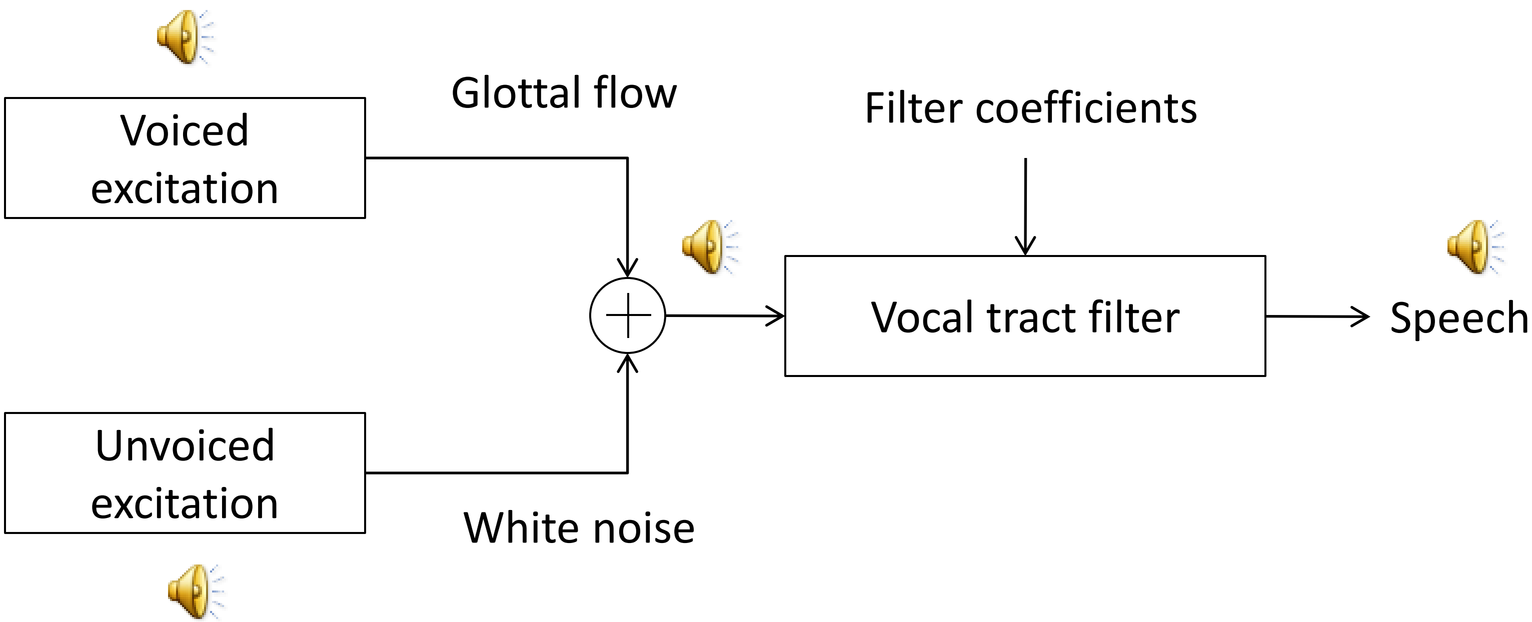


Glottal inverse filtering based excitation



Conventional impulse excitation

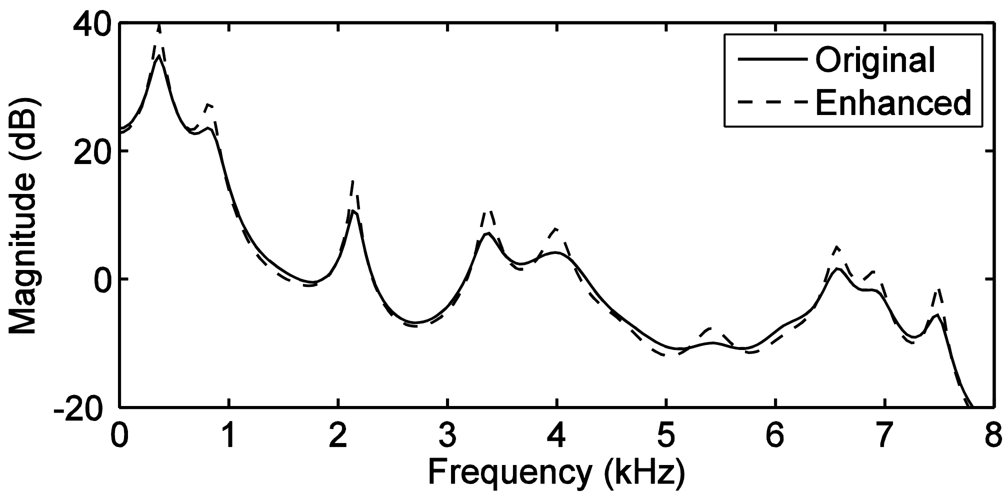
II. GlottHMM speech synthesis system



Aalto University
UNIVERSITY OF HELSINKI

II. GlottHMM speech synthesis system

- Other special solutions:
  - Enhancement of vocal tract information with a new method (Raitio *et al.*, SSW7)



III. On Modeling of prosody

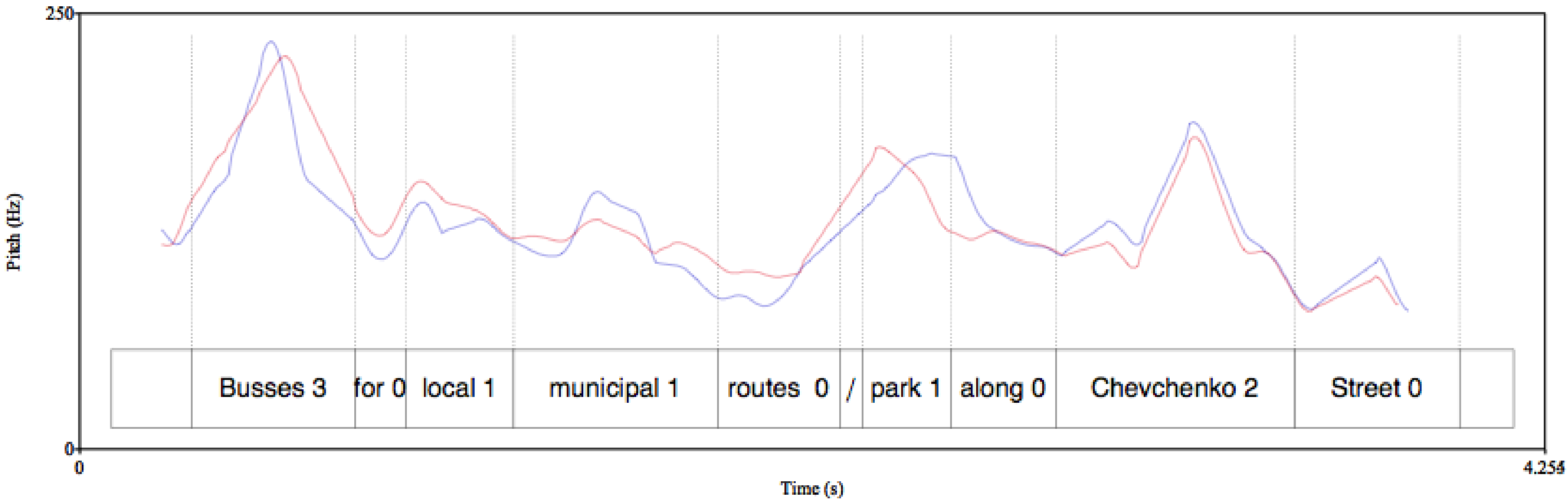
- Due to statistical averaging effect, accurate prosodic labels are crucial for expressive parametric synthesis
- In order to model sentence-level prosody, we used **perceptual prominence** (on 0-3 scale)
  - automatic annotation using acoustic features: F0, energy, harmonics, HNR, duration
  - prominence prediction with CART using shallow features + syntactic phrases

III. On Modeling of Prosody, Example

Synthesizer can reproduce the intended prominences fairly well, allowing controlled production of emphasis, nuclear and pre-nuclear accents

original (blue)

synthesis with hand-labelled prominences (red)



IV. Speech in noise

- Special voices (ES2 and MS2) were built by utilizing several aspects observed in the **Lombard effect**. Modifications were made on several levels of synthesis:
    - **Phonological**: Prominence of stressed syllables of content words was increased and intra-utterance silences were removed
    - **Parameter generation**: Rate of speech was lowered, pitch was raised and pitch range compressed
    - **Vocoder**: More post-filtering was used to produce clearer formant information

IV. Speech in noise

- **Vocoder**: Vocal tract length was shortened slightly to match the raised pitch and raised formant frequencies
- **Vocoder**: The spectral tilt of the glottal source signal was decreased, concentrating more energy in formant frequencies
- **Finally**, the resulting signal waveform was companded in order to make the loudness of the speech as high and uniform as possible

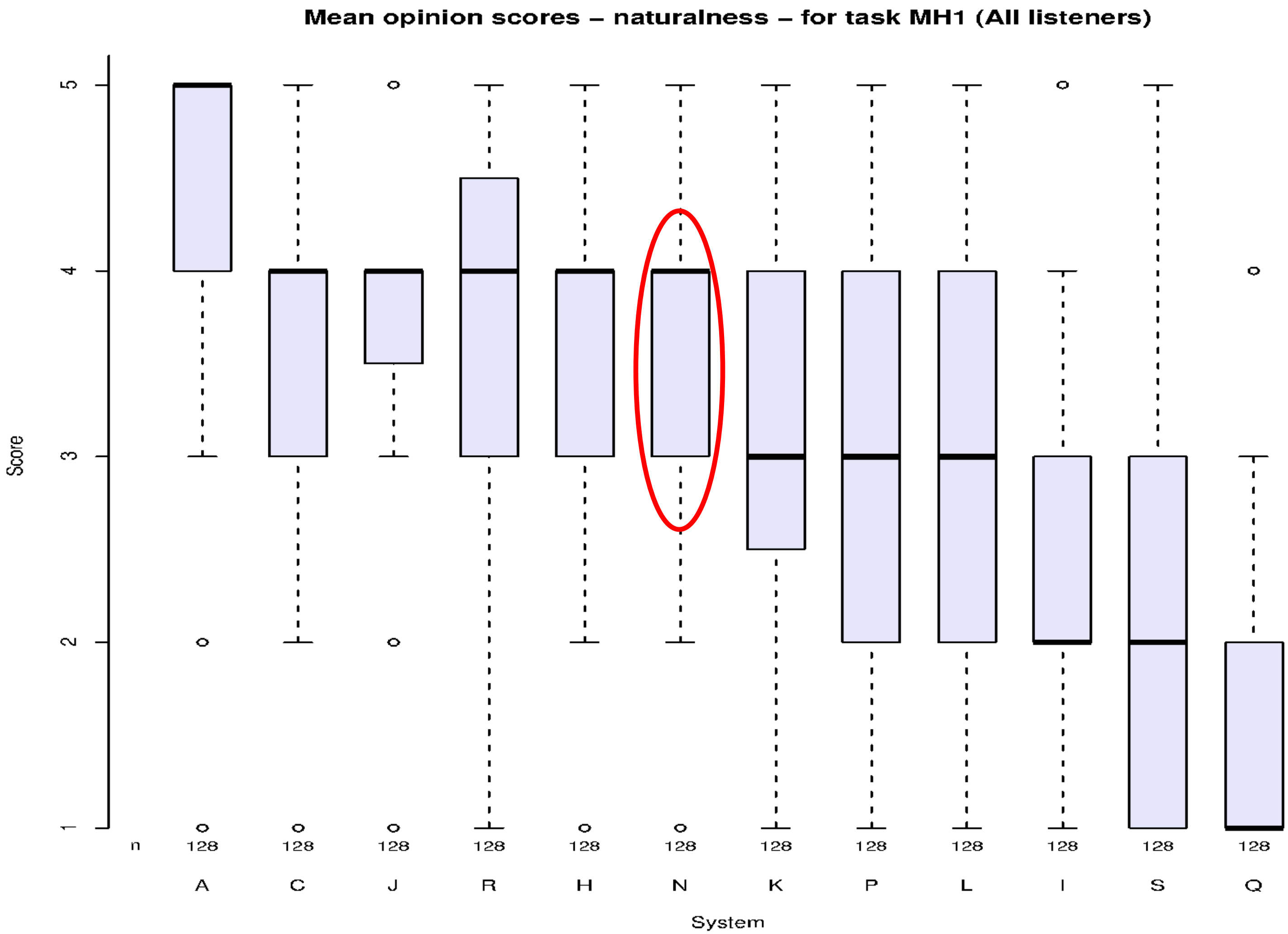
EH1	ES2
(*EH2)	

V. Results

- English
  - MOS scores were consistently higher than STRAIGHT-based HTS baseline systems, but we could not compete with the best voices
  - Only average intelligibility and low similarity scores
  - Problems: artefacts and unstable F0 contour, voicing problems in stop consonants, low similarity due to the use of single glottal pulse
  - Prominence modeling was not advantageous with short sentences

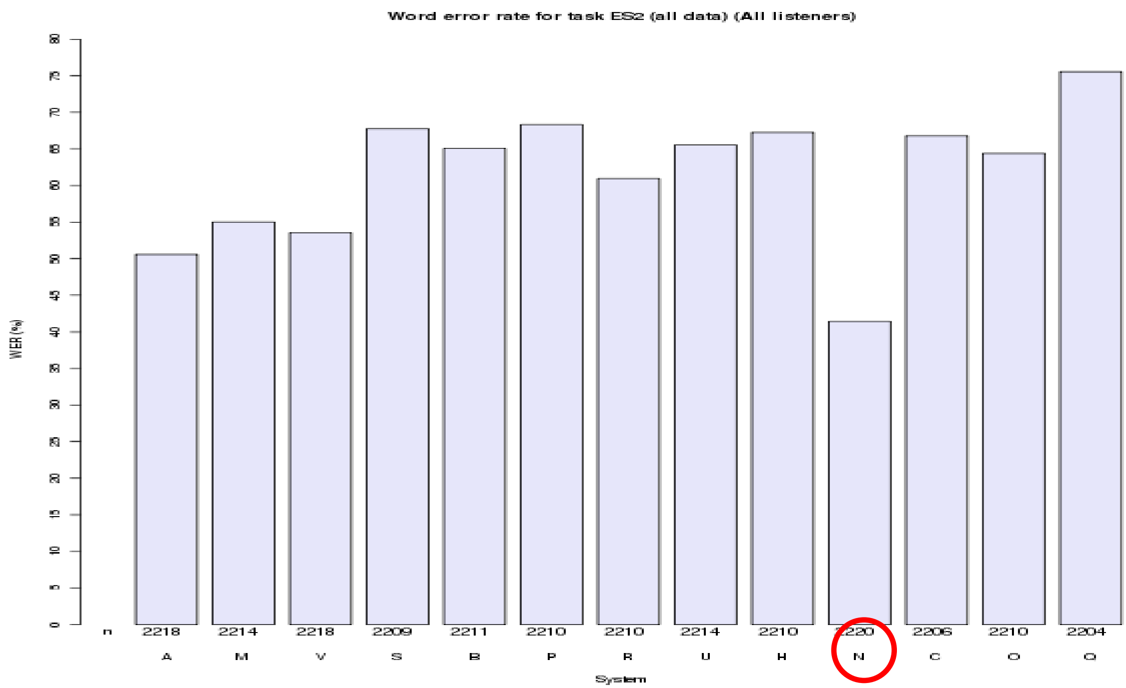
V. Results

- Mandarin:
  - Our system ranked among the best on MOS on task MH1
  - Again, similarity scores were not good
  - On intelligibility test, only the original speaker ranked significantly higher than our system


Mean opinion scores – naturalness – for task MH1 (All listeners)
Score
5
4
3
2
1
n
128
128
128
128
128
128
128
128
128
128
128
128
A
C
J
R
H
N
K
P
L
I
S
Q
System

V. Results

- Speech in noise:
  - Our voices had the lowest word error rates by a clear margin, even compared to natural speech



VI. Conclusions

- Separation of glottal source and vocal tract filter characteristics enabled large modifications in speech in noise task
- On other tasks, MOS scores were generally good
- Similarity scores low; current source modeling with single glottal pulse insufficient $\rightarrow$ ongoing work with speaker specific multiple pulse techniques

Thank you! Questions?

References:

- Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., “HMM-based speech synthesis utilizing glottal inverse filtering”, IEEE Trans. Audio, Speech, and Language Processing, (in press).
- Alku, P., “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering”, Speech Commun., 11(2–3):109–118, Jun. 1992