GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach

Antti Suni
Martti Vainio

Tuomo Raitio
Paavo Alku

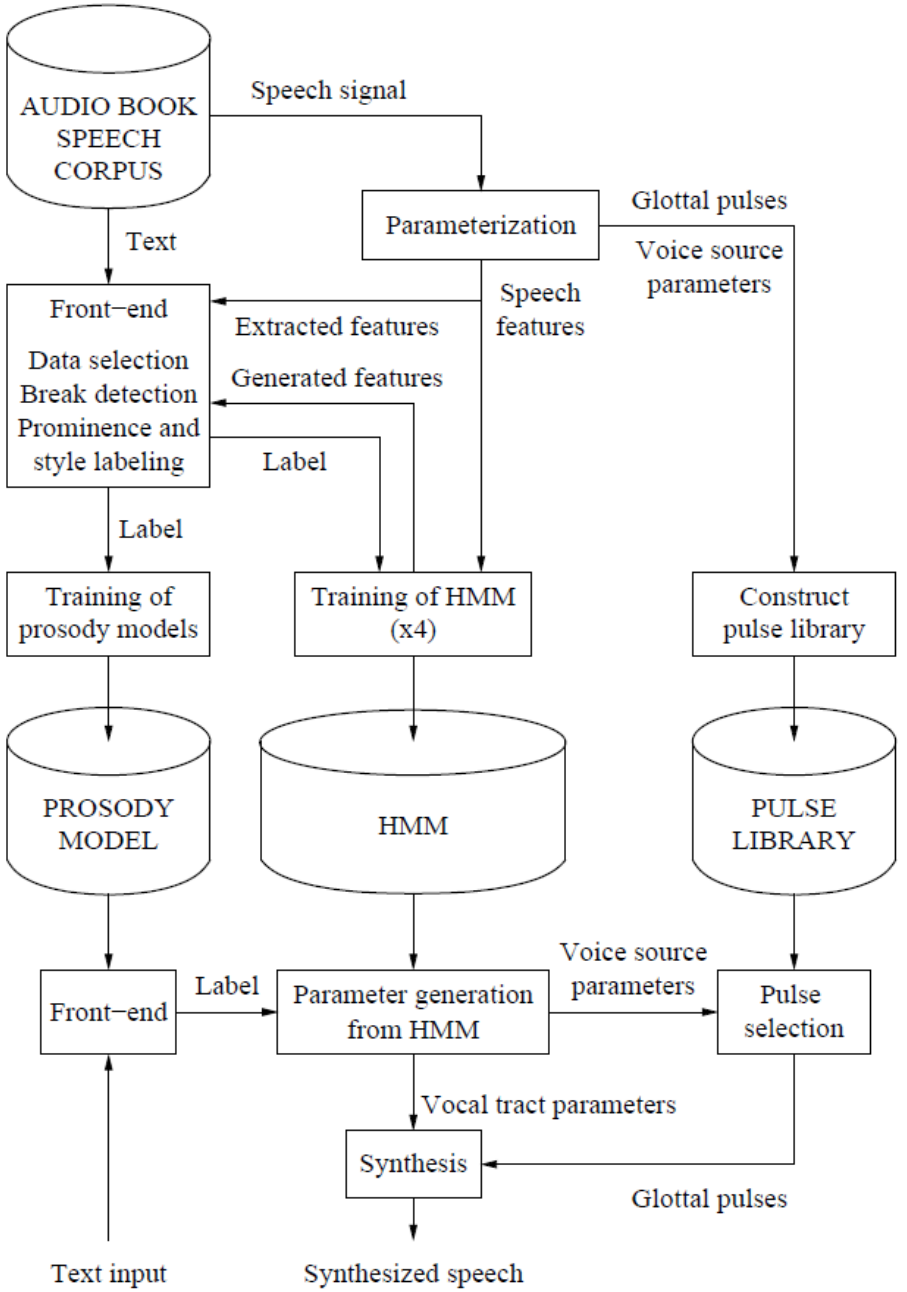
Introduction

- GlottHMM is a statistical parametric TTS system developed in collaboration with University of Helsinki and Aalto University, Finland
- Based on HTS but:
  - Detailed modeling of prosody
  - Special vocoder architecture

Motivation

- This is the third time we participate in the Blizzard Challenge, motivated by:
  - Testing new methods such as glottal flow pulse library based excitation generation and differential LSF based spectral representation
  - Testing our system with challenging speech material:
    - Continuous speech
    - Suboptimal recordings
    - Mixed speaking styles

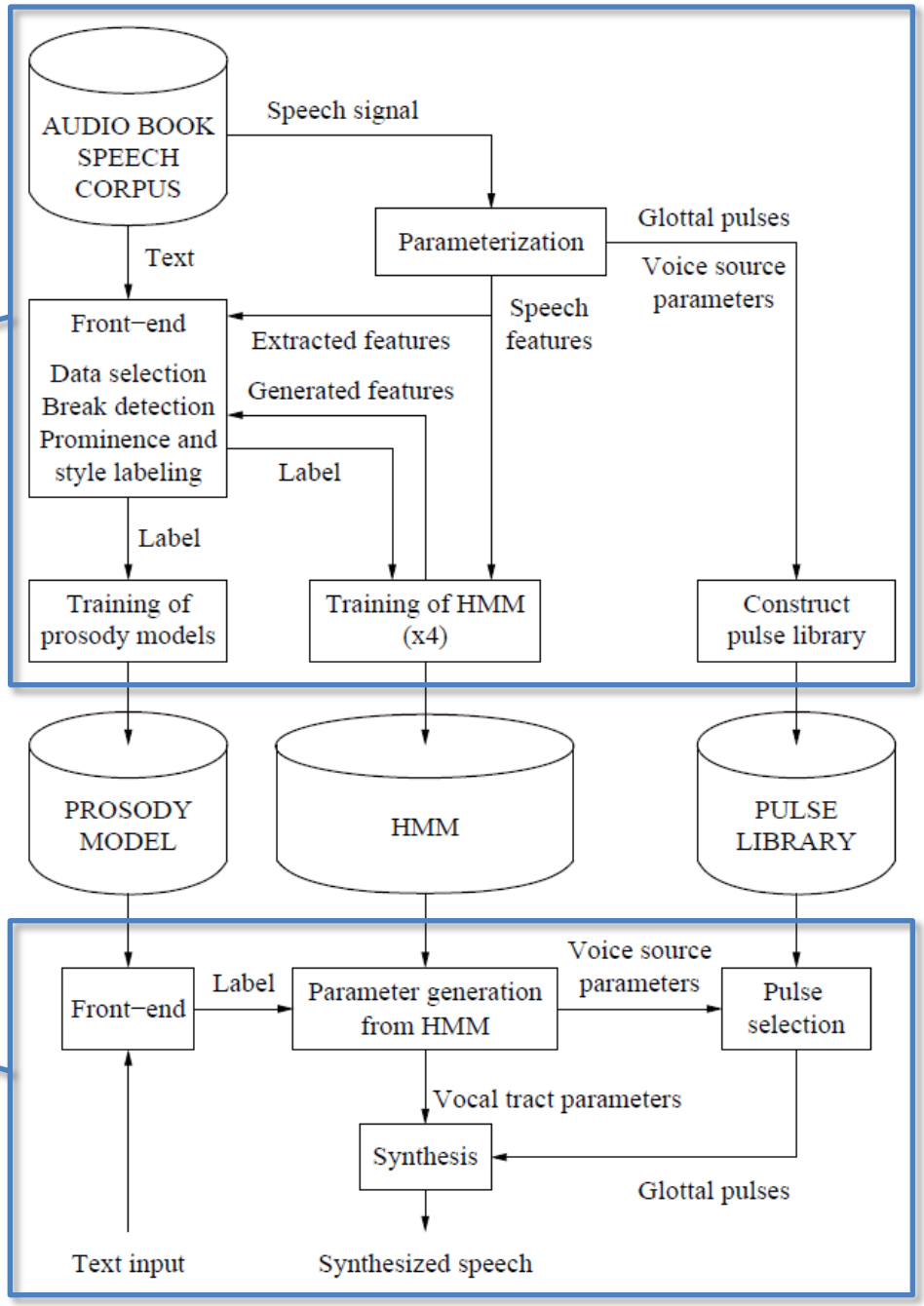
GlottHMM System



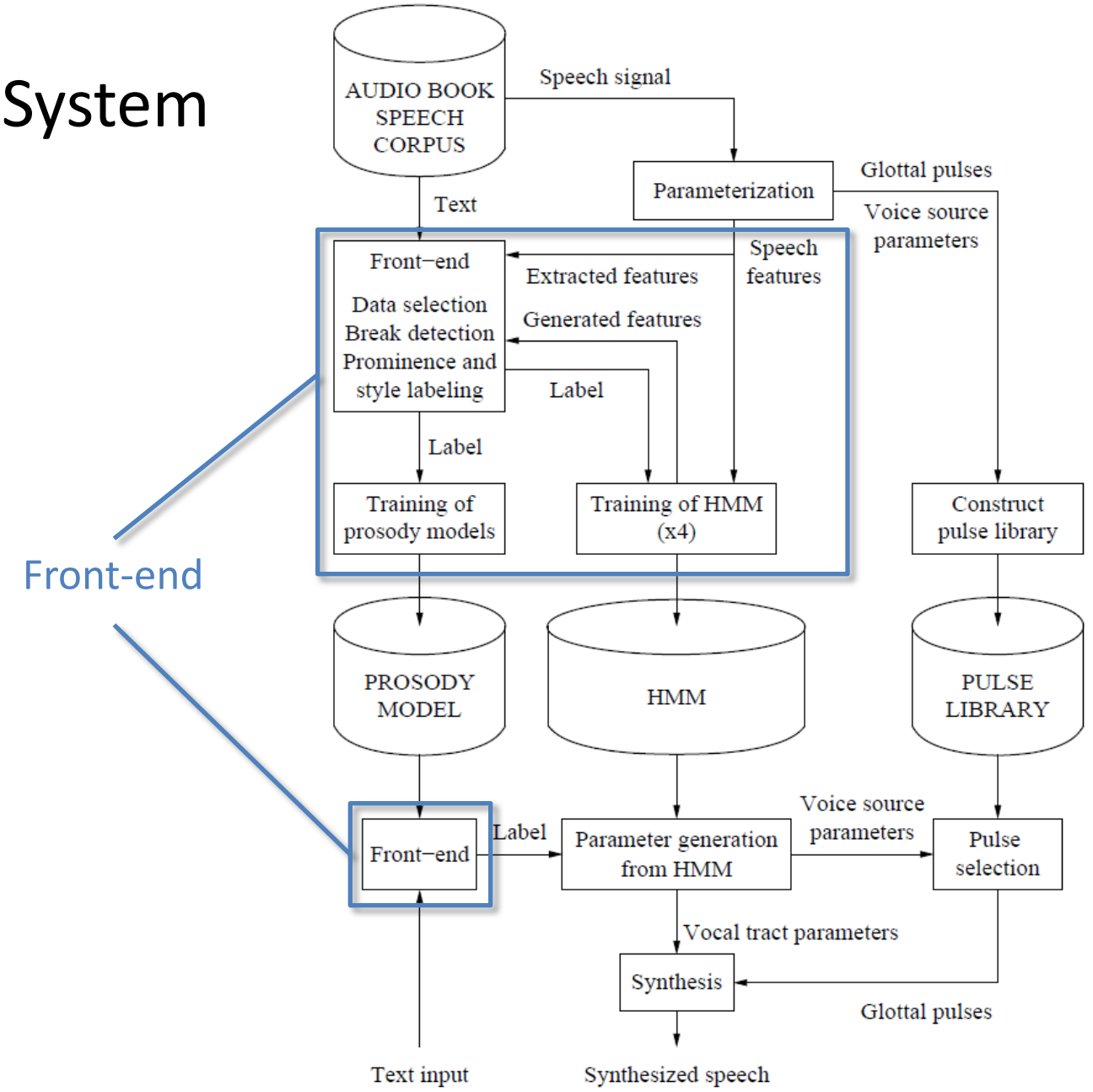
GlottHMM System

Training part

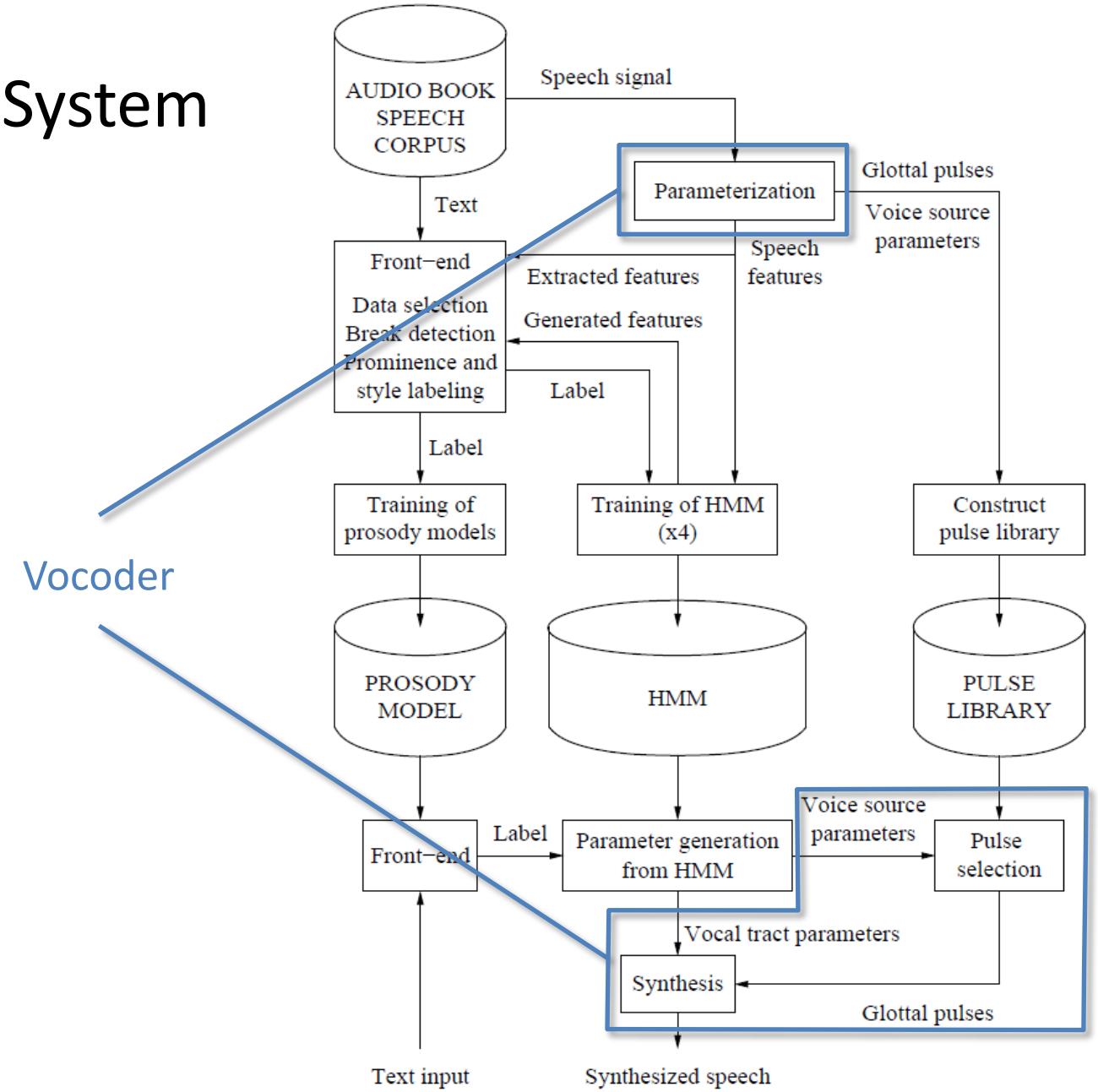
Synthesis part



GlottHMM System



GlottHMM System



Data selection

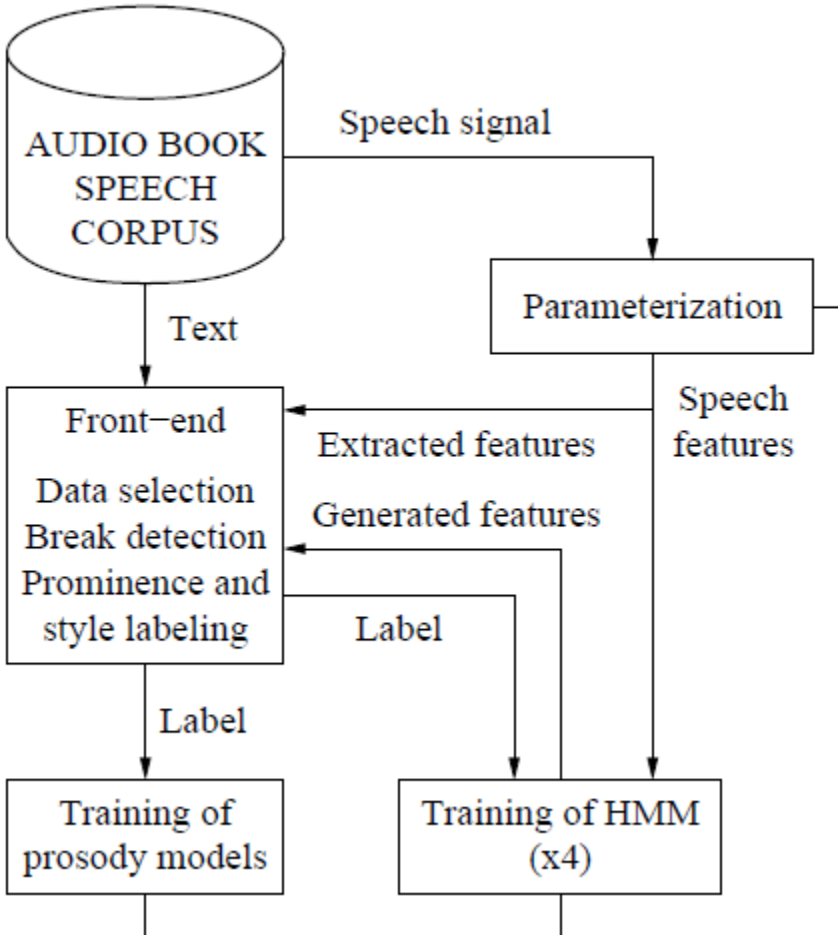
- Only one of the four books were chosen as training material for consistency: “Adventures of Tom Sawyer”
  - It contained least OOV words
  - Unfortunately, it contained more noise than the other books
- Channel differences between the books:

Used book:  Other book: 

- Only sentences that gave 100% confidence scores were selected; a total of 3740 sentences of training material

Front-end

- Compared to e.g. Festival, there is a closer relationship with label generation and the training of speech in our process
- Iterative refinement of labels using acoustic features in
  - Break detection
  - Prominence labeling
  - Style labeling



Break and pronunciation detection

- Phrase breaks were initially placed on punctuation, then recognized by force alignment with optional silence after each word

  cat = (C A T | C A T SIL)

- Word pronunciations were also recognized by force alignment using variants listed in the unilex lexicon

  the = (DH @ | DH EE)

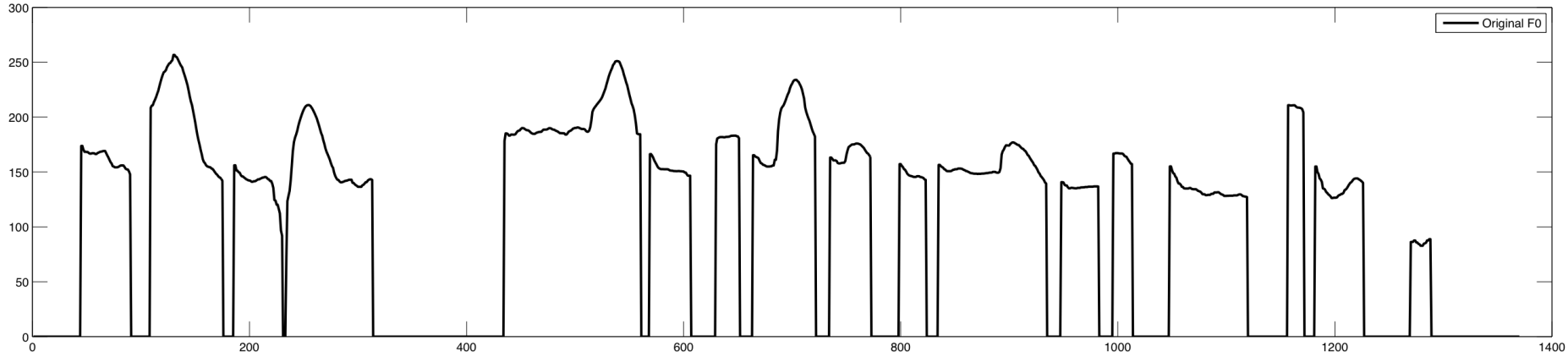
Front-end – Prominence labeling

- Instead of text-based accent model, we use word prominence scale:
  - 0 – no accent
  - 1 – weak accent
  - 2 – strong accent
  - 3 – emphasis

- Annotation is based on the difference between original and generated speech features

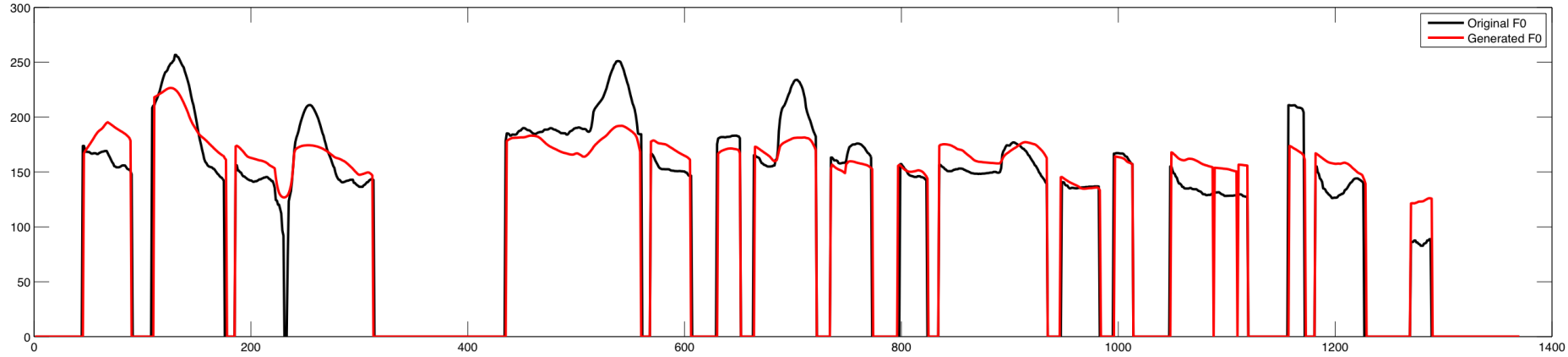
Prominence Labeling – Example

- Extract GlottHMM acoustic features from speech database (F0 shown below)



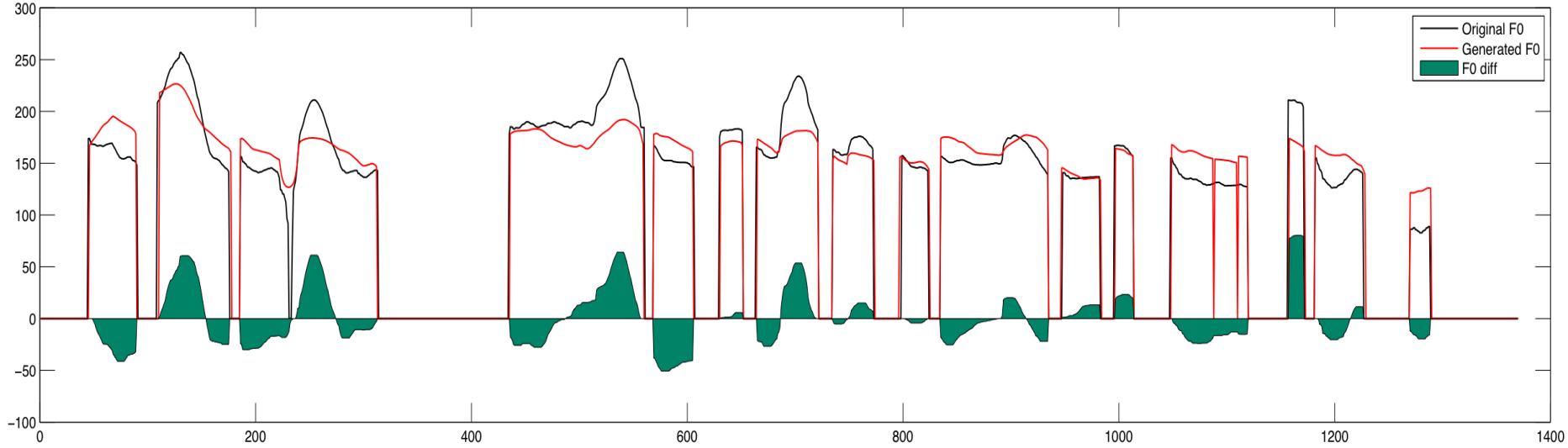
Prominence Labeling – Example

- Train voice without word context features
- Generate aligned features of the training data



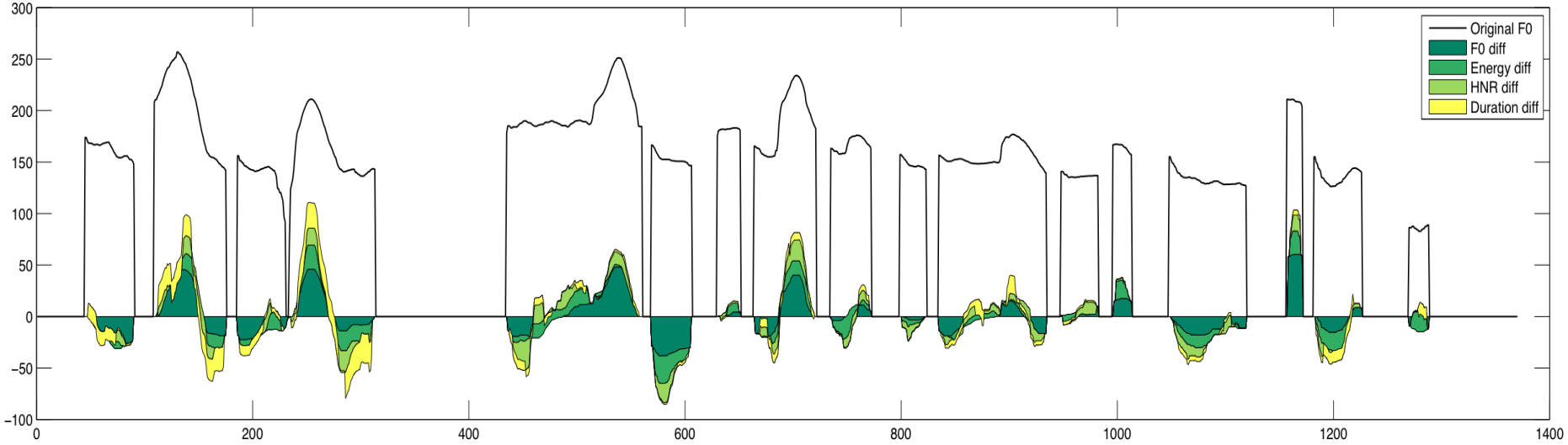
Prominence Labeling – Example

- Evaluate difference between original and generated features



Prominence Labeling – Example

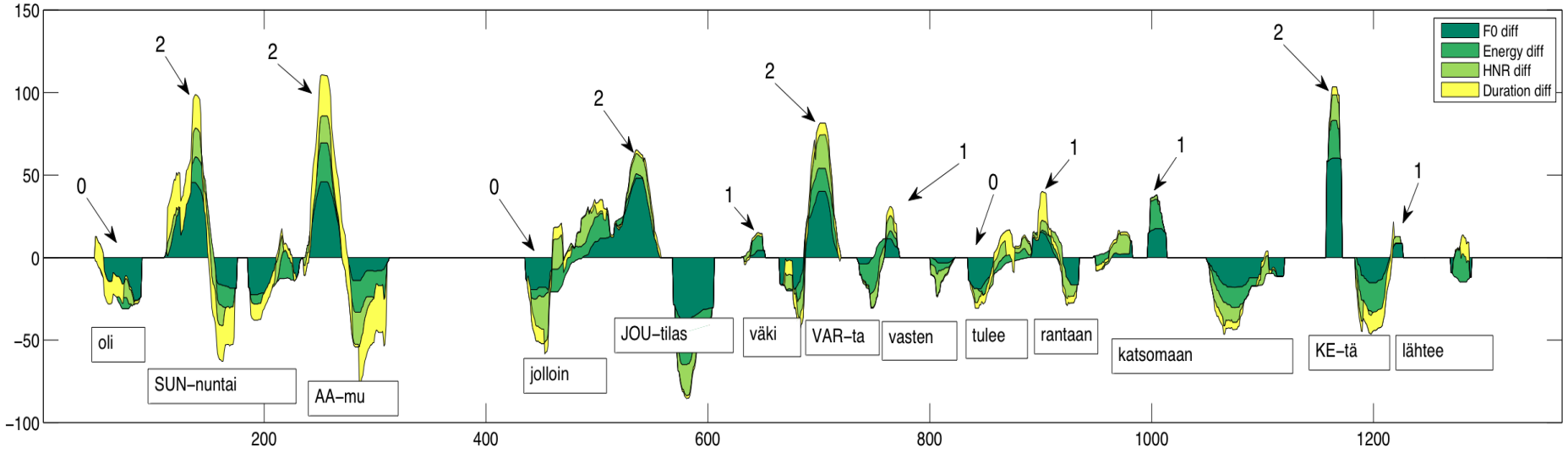
- Calculate weighted sum of differences of features:
  - F0, Energy, HNR, duration



Prominence Labeling – Example

- Discretize differences into four prominence levels
- Add prominence contextual features into labels

  → Train final voice



Front-end – Prominence prediction

- In synthesis, word prominences were predicted with CART using
  - average prominence of the word in training data
  - part-of-speech
  - unigram frequency
  - positional features

- Additionally, some rules related to paragraph context were applied, for example, decreasing the prominence of previously mentioned words in pre-modified nominal phrases

Front-end – Phrase style labeling

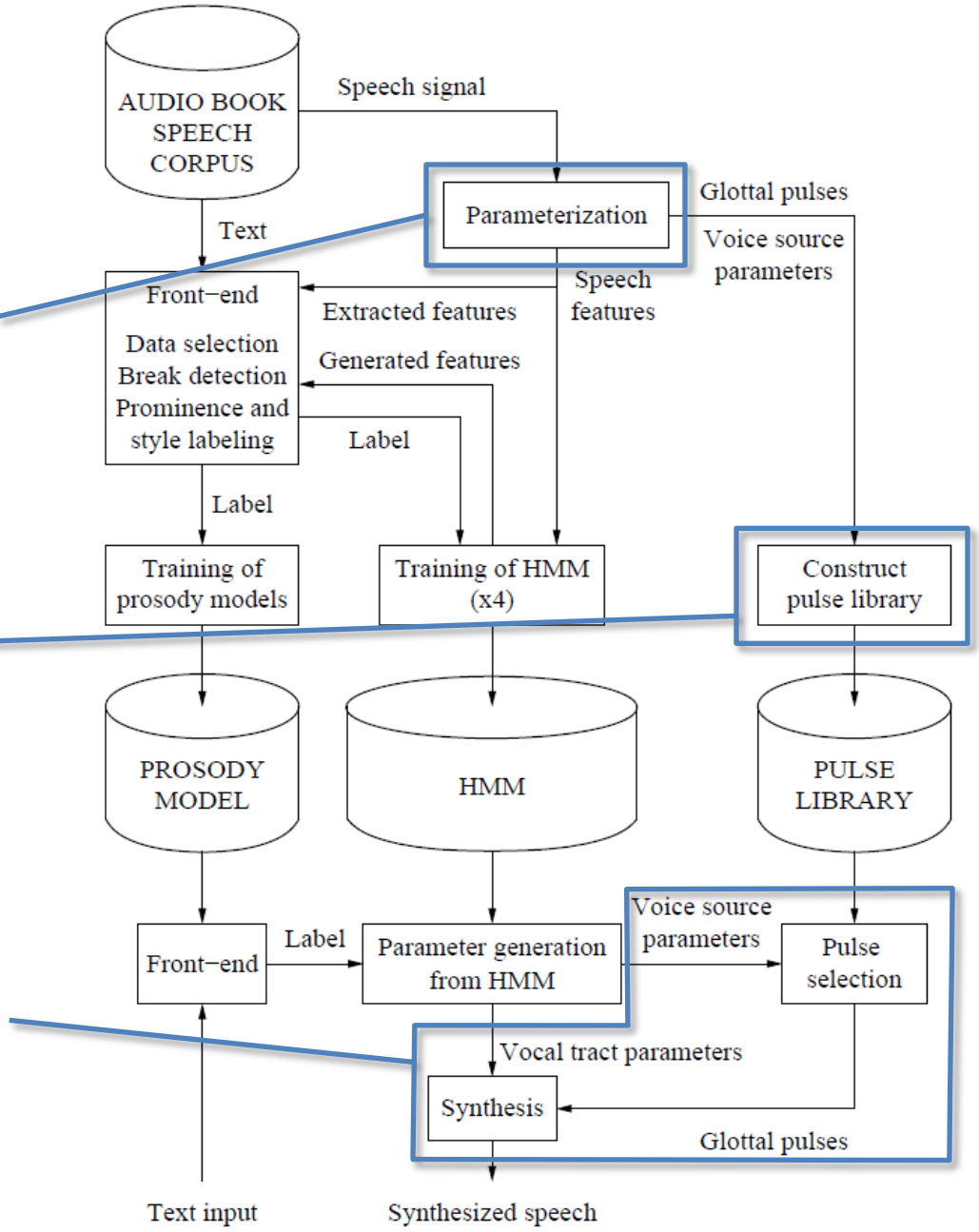
- Speaking style was annotated using the prominence method for phrase level
- Styles were discretized into two levels of involvement, roughly corresponding to normal narrative style and quotations
- This stabilized the normal style synthesis, but the quotation style could not be used due to erratic pitch and artefacts

Vocoder

Speech is decomposed into the vocal tract filter and voice source signal using glottal inverse filtering

Glottal flow pulses are extracted and stored with corresponding voice source features

Glottal flow pulses are selected according to voice source features and concatenated. Excitation is filtered with vocal tract parameters to generate speech

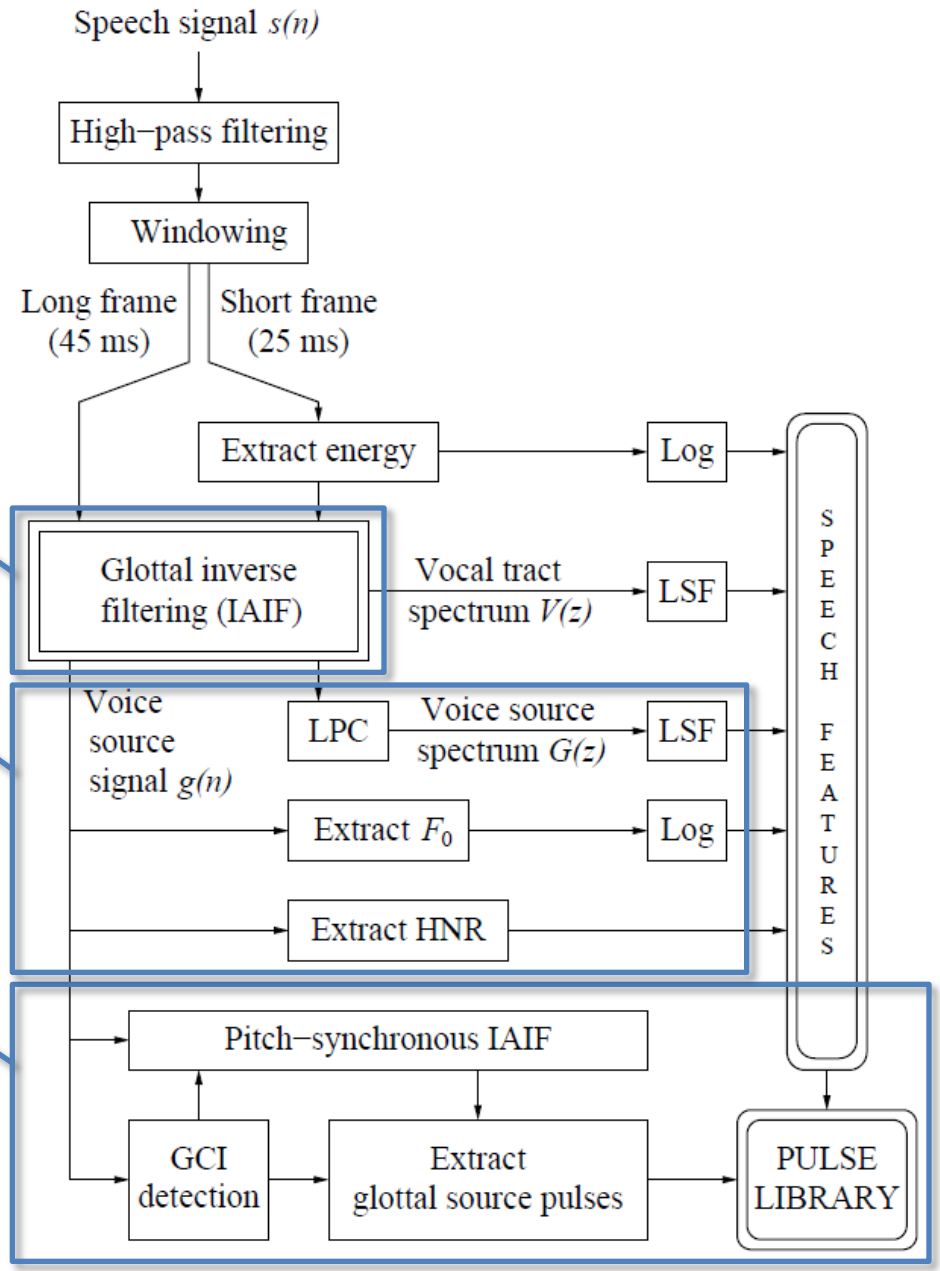


Parameterization

Speech is decomposed into the vocal tract filter and voice source signal using glottal inverse filtering

Detailed voice source analysis

GCIs are detected and glottal flow pulses are extracted and stored with corresponding voice source features



Differential LSFs

- Line Spectral Frequencies (LSFs) are know to correlate with each other, which may cause problems in HMM training
- As a solution, we used differential LSFs instead
- dLSF vector contained 31 coefficients:

1	First LSF
2—30	Difference between the adjacent LSFs
31	Distance of last LSF to π (Pi)

- Square root of the dLSFs were used for training in order to make distributions more Gaussian

In synthesis, the differential LSFs were equalized so that the sum of the 30 first LSF distances matched the 31th, the distance to π

Parameterization

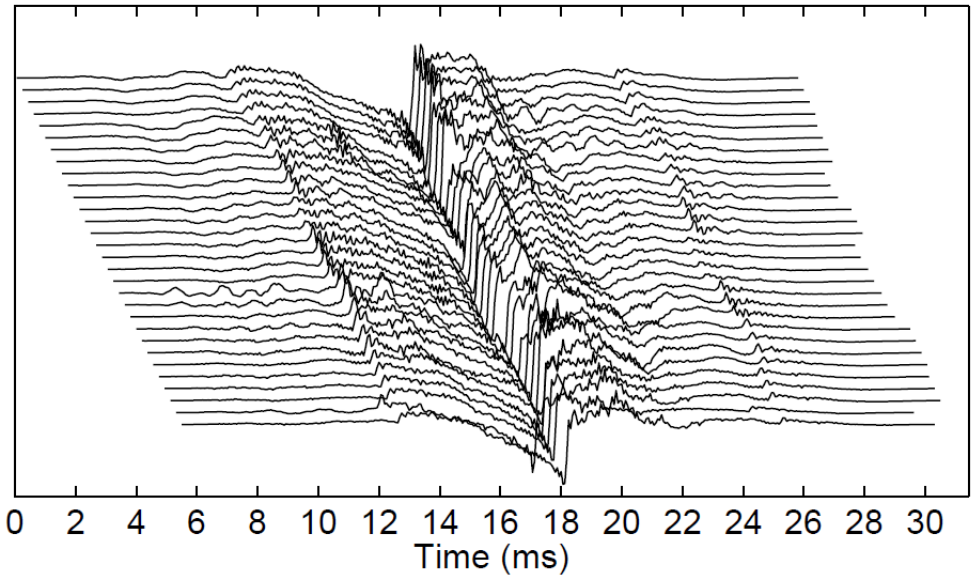
Speech Feature	Parameters per Frame
Log-Fundamental frequency	1
Log-Energy	1
Harmonic-to-noise ratio	5
Voice source spectrum LSFs	10
Vocal tract spectrum dLSFs	31
Pulse library	20 000 pulses

Parameter Training

Stream number	Stream type	Features	Order
1	MSD	Fundamental frequency	1
2	normal	Harmonic-to-noise ratio	5
3	normal	Voice source spectrum	10
4	normal	Energy + dLSFs	32

Pulse Library

- Pulse library was built from 10 diverse utterances selected for phonetic and F0 range coverage
- It contained a total of 22 414 windowed two-period glottal flow derivatives linked with the corresponding voice source parameters



Visualization of 30 glottal flow pulses from the library

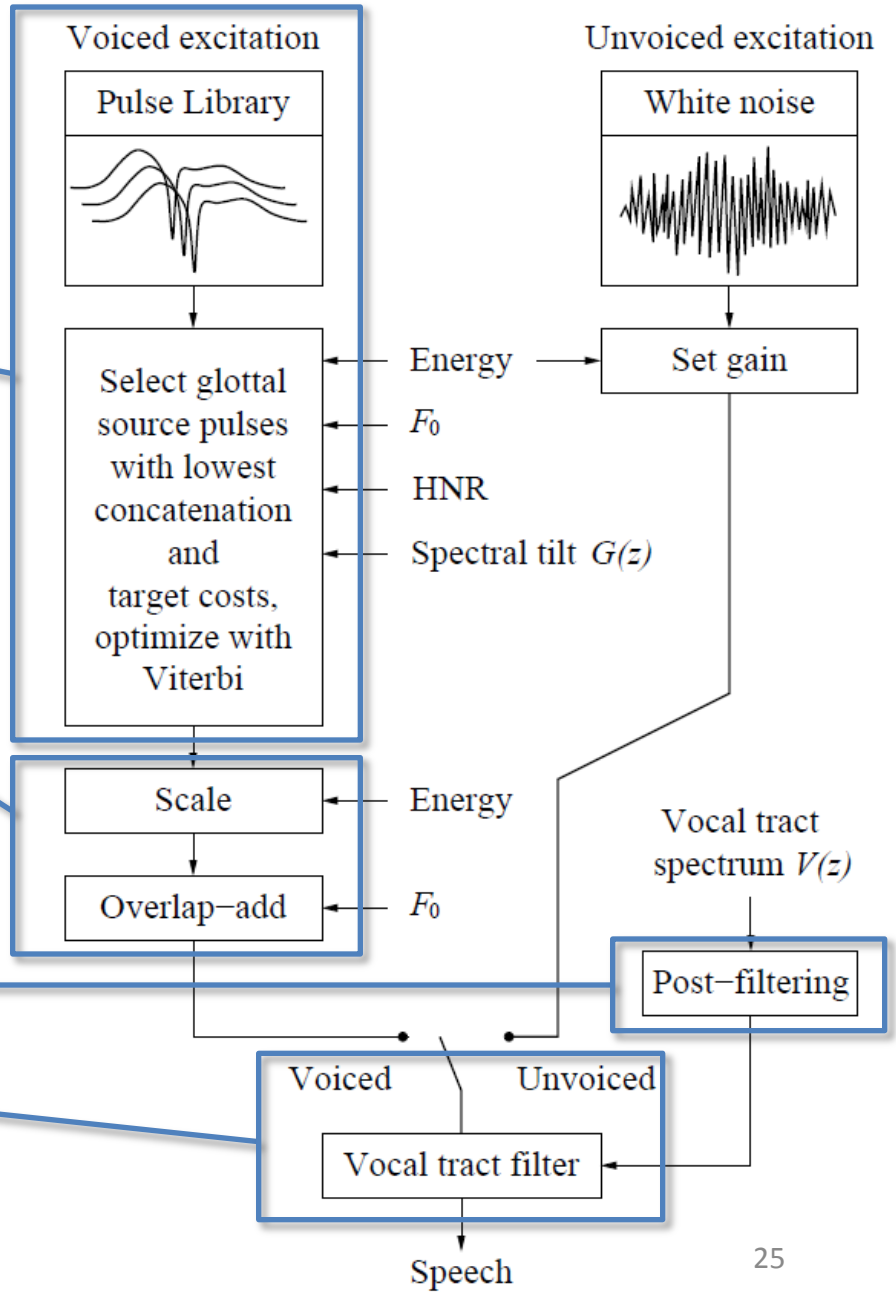
Synthesis

Voiced excitation is generated by selecting the best matching pulses from the library according to the voice source features

Selected pulses are scaled in energy and overlap-added to generate continuous pulse train

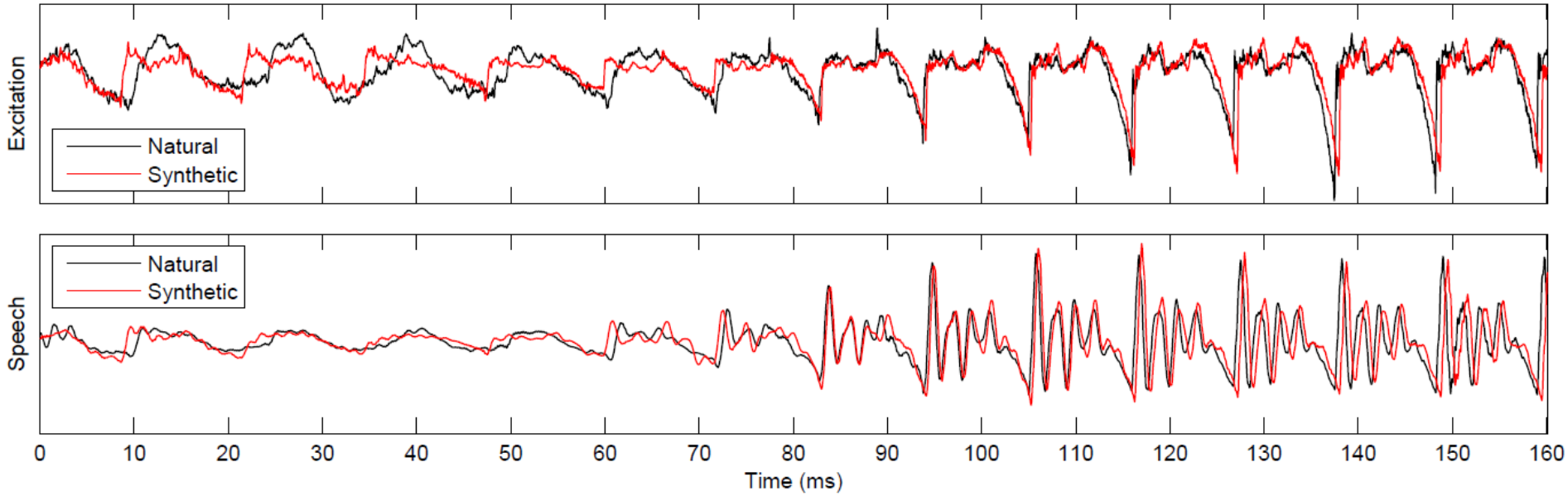
Formant enhancement

Excitations are combined and filtered with vocal tract filter



Estimated glottal flow signal of natural speech segment /ho/

Synthetic glottal flow excitation signal



Natural speech segment /ho/

Synthetic speech segment /ho/

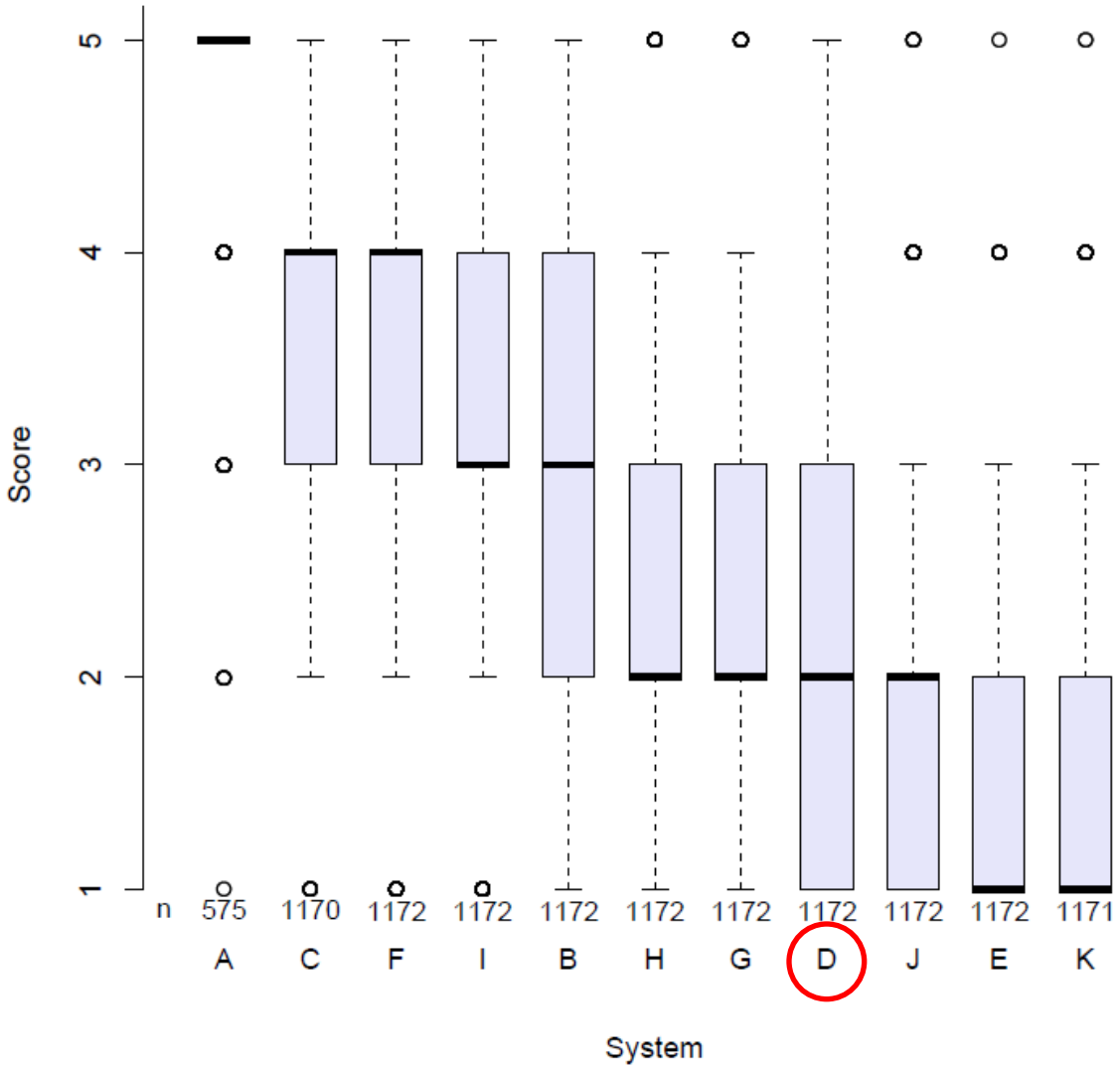
Post-Processing

- Test sentences were synthesized applying both parameter generation considering global variance (GV) and formant-enhancement
  - Successful for compensating for muffled speech but resulted in a harsh quality
- Finally some room reverberation was added to samples but probably this had not much effect

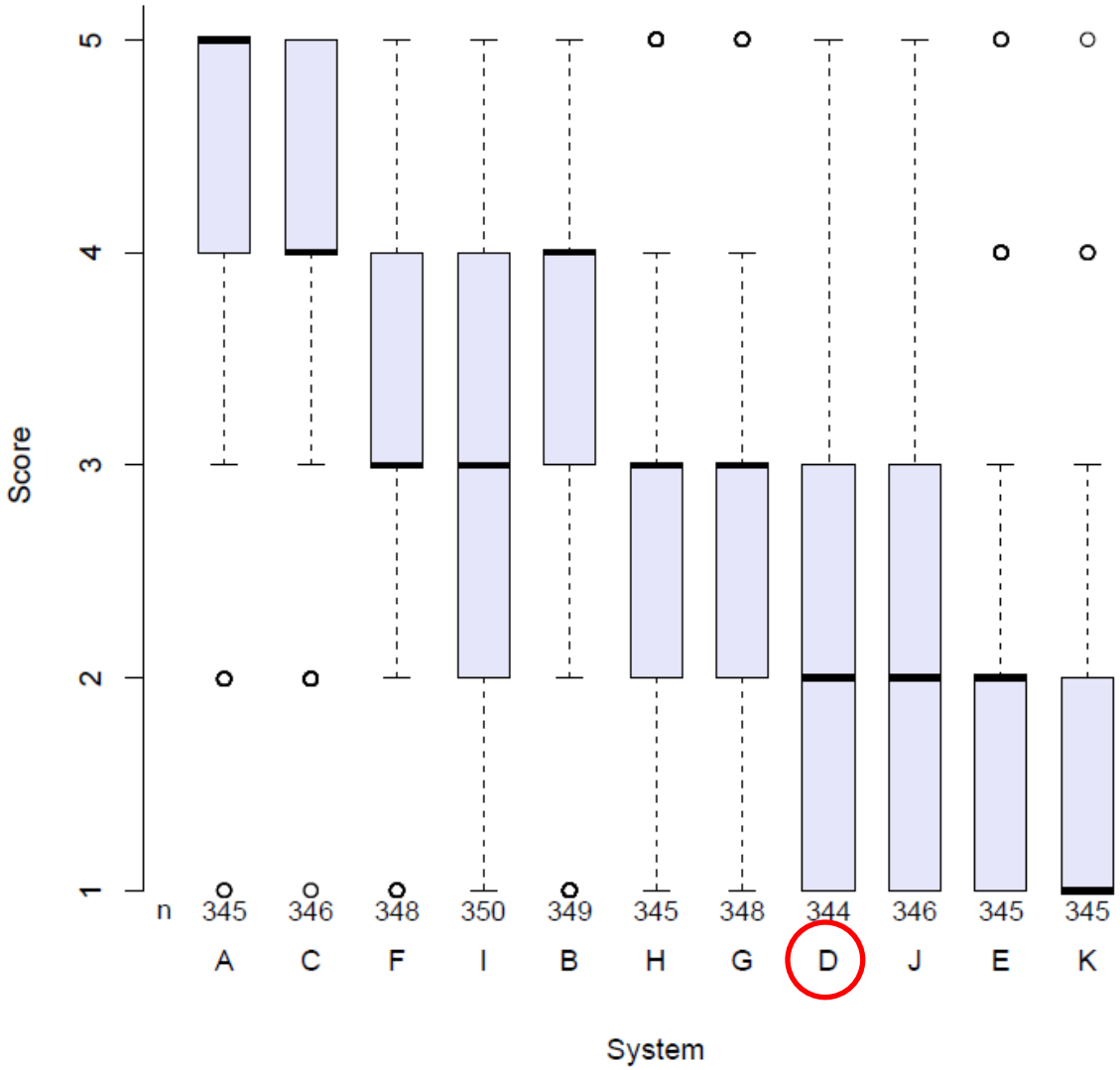
Results

- MOS and SIM were not especially good due to:
  - Strong post-processing, resulting in artificial tone of voice
  - Choice of training material

Mean Opinion Score

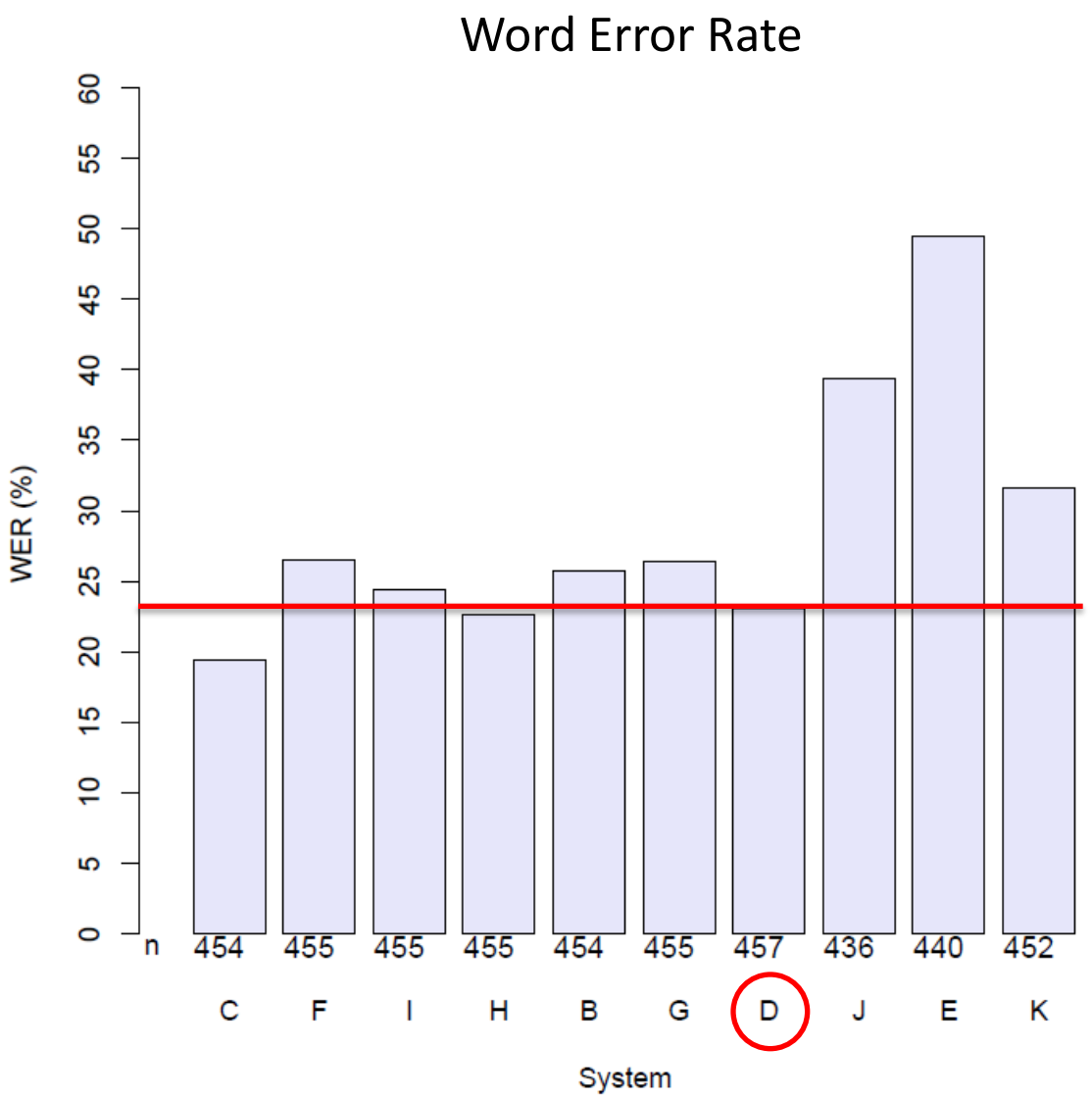


Similarity


Score
1
2
3
4
5
n
345
346
348
350
349
345
348
344
346
345
345
A
C
F
I
B
H
G
D
J
E
K
System

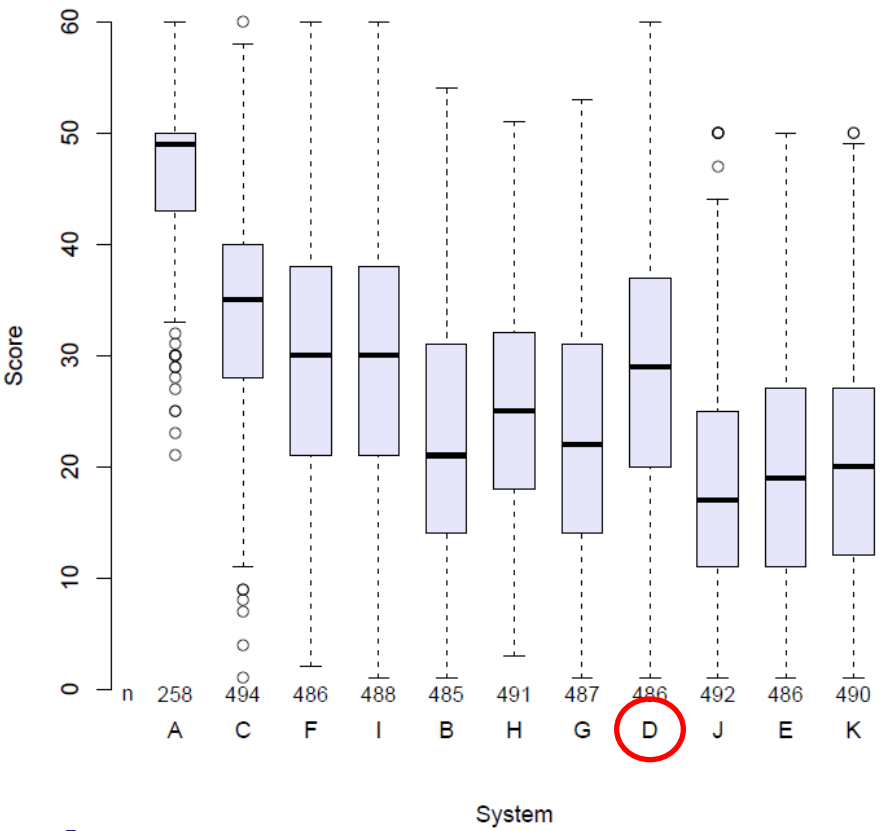
Results

- Intelligibility was the second best after natural speech and system H

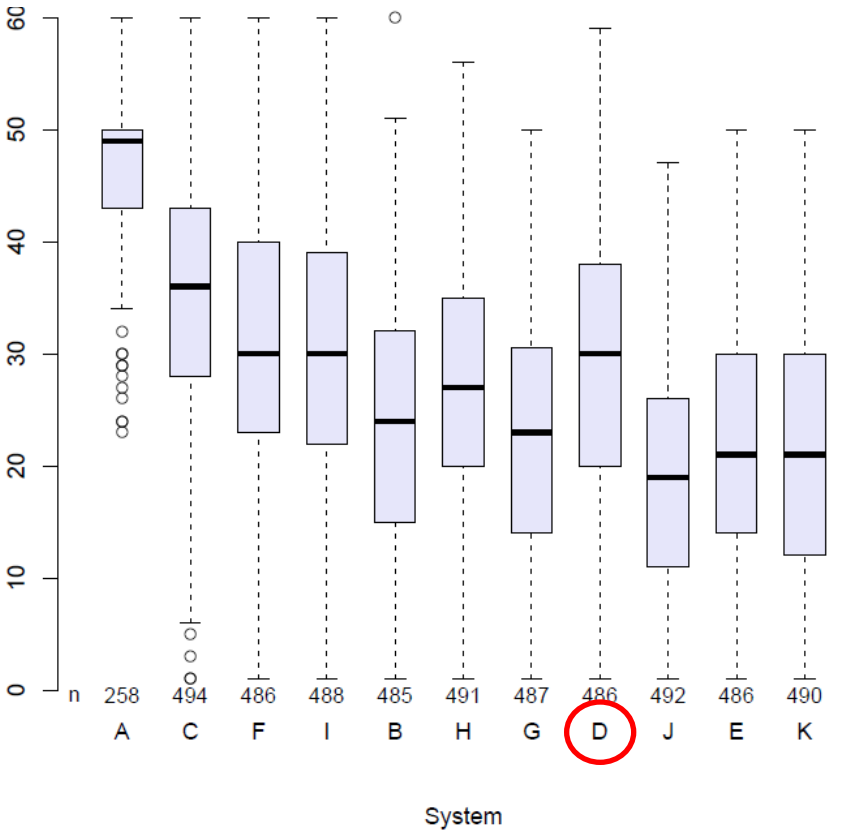


- Prosody characteristics were above average; especially prominence and speech pauses were rated among the top systems

Stress MOS



Pauses MOS



Samples

Synthesized paragraph:

Conclusions

- This year's challenge was difficult due to challenging speech material:
  - Noisy, low-quality recordings
  - Various speaking styles
- Nevertheless, we achieved a clean, intelligible synthetic voice with above average prosody characteristics
- More work required on:
  - Improving the robustness of the vocoder
  - Prosody annotation with unsupervised methods

Thank you for your attention.

Questions?