The GlottHMM Entry for Blizzard Challenge 2011:

Utilizing Source Unit Selection in HMM-Based Speech Synthesis for Improved Excitation Generation

Antti Suni, Tuomo Raitio, Martti Vainio, and Paavo Alku

antti.suni@helsinki.fi, tuomo.raitio@aalto.fi

2.9.2011

I. Introduction

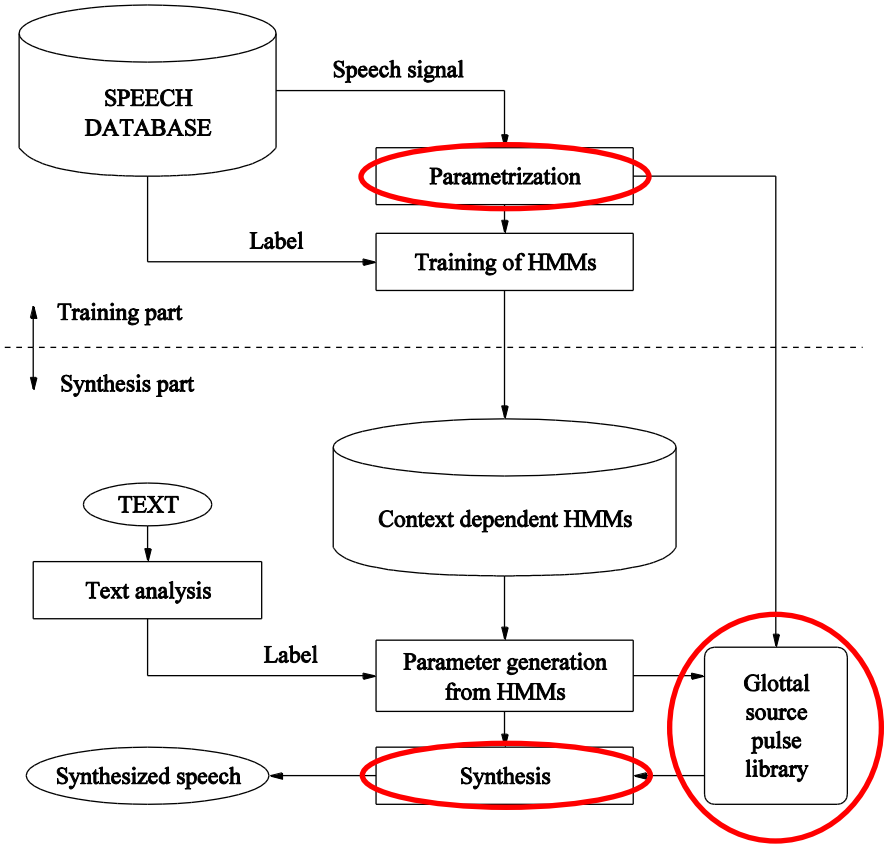
- GlottHMM is a statistical parametric TTS developed in collaboration with Aalto University and University of Helsinki, Finland
- Based on HTS but special vocoder architecture:
  - Glottal inverse filtering based vocoding
  - Detailed parametrization of the voice source

I. Introduction

- This is the second time we participate in the Blizzard Challenge motivated by
  - Improving female voice quality
  - Testing new methods such as:
    - Unit selection scheme for improved excitation generation
    - Stabilized weighted linear prediction (SWLP) for more robust spectral modeling
    - MGE inspired trajectory sharpening based on extrapolation
- Used only 3000 sentences, 16 kHz
- Contextual features extracted from lessemes

II. GlottHMM speech synthesis system

GlottHMM architecture:

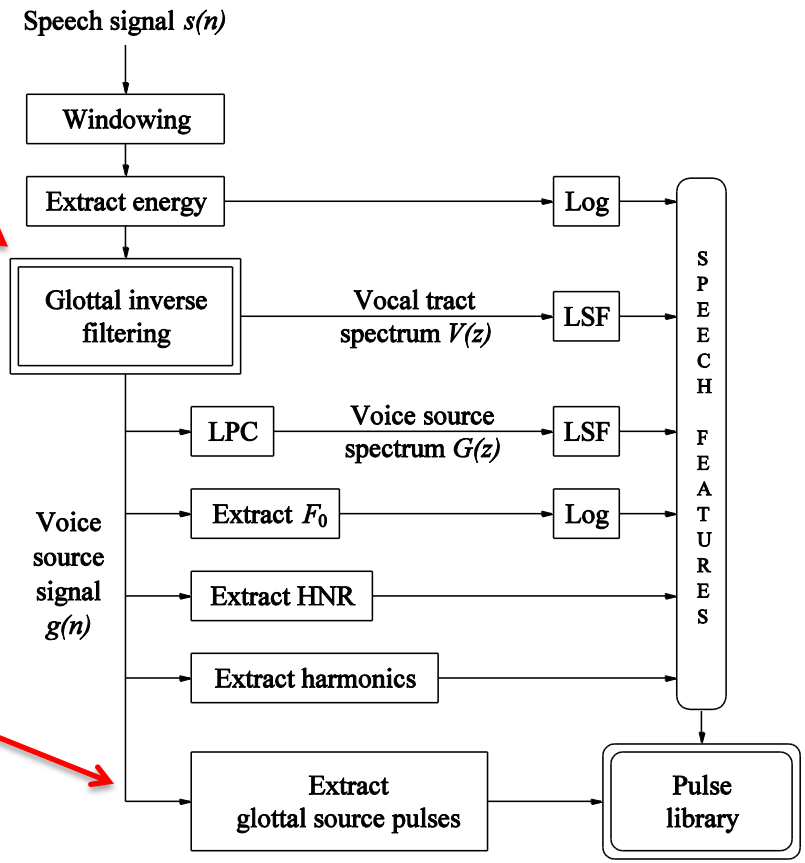


II. Analysis

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

Vocal tract is parameterized with SWLP to line spectral frequencies (LSFs)

Voice is source parametrized with LPC spectral tilt, F0, Harmonic-to-noise ratio (HNR), 10 harmonics, and a **glottal pulse library**

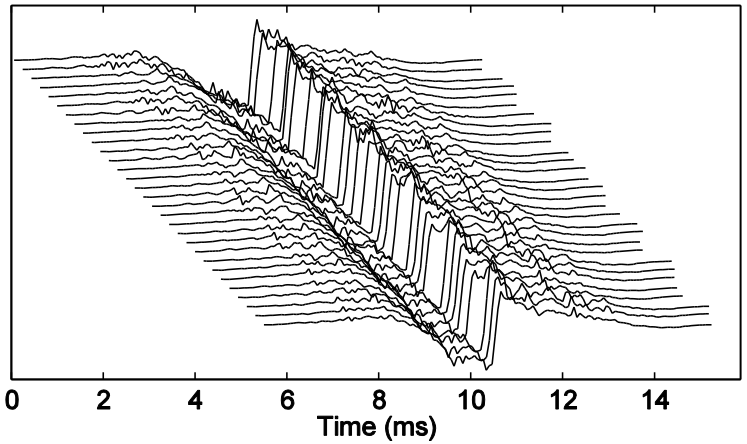


II. Pulse library

Pulse library is constructed by

1. Glottal closure instants (GCIs) are determined from the differentiated glottal flow signal
2. Each complete two-period glottal source segment is extracted and windowed with the Hann window
3. Pulses are linked with the corresponding voice source parameters
4. In addition, a down-sampled (10 ms) version of the pulse waveform is stored for evaluating concatenation cost in synthesis stage

15000 pulses from 20 selected utterances with rich F0 movement and phonetic context



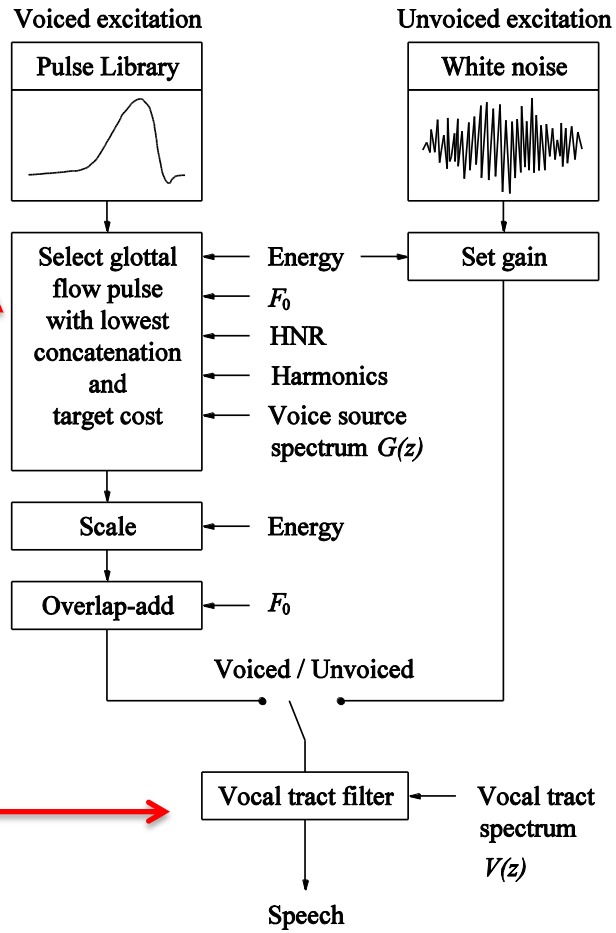
II. Synthesis

In synthesis stage, excitation signal is generated by **selecting the best matching pulses** from the library according to the voice 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



II. Unit selection scheme for the voice source

- The best pulse for each time index is selected by minimizing the **joint cost** composed of target and concatenation costs:
  - **Target cost**: RMS error between the voice source parameters generated by the HMM and the ones stored for each pulse
  - **Concatenation cost**: RMS error between the down-sampled versions of the pulse candidates
- Viterbi search is used for finding the best pulses for each voiced segment

→ Very natural voice source at its best

II. SWLP based spectral modeling

Stabilized weighted linear prediction (SWLP) is used to estimate the vocal tract spectrum

In SWLP analysis, the autocorrelation is weighted by the short time energy window of the signal, thus emphasizing high energy parts

→ Spectrum is less distracted by the harmonics of the excitation signal since the high energy parts are located in the glottal closed phase instants
→ Inverse filtering is more accurate since the excitation is given less weight when determining the vocal tract spectrum

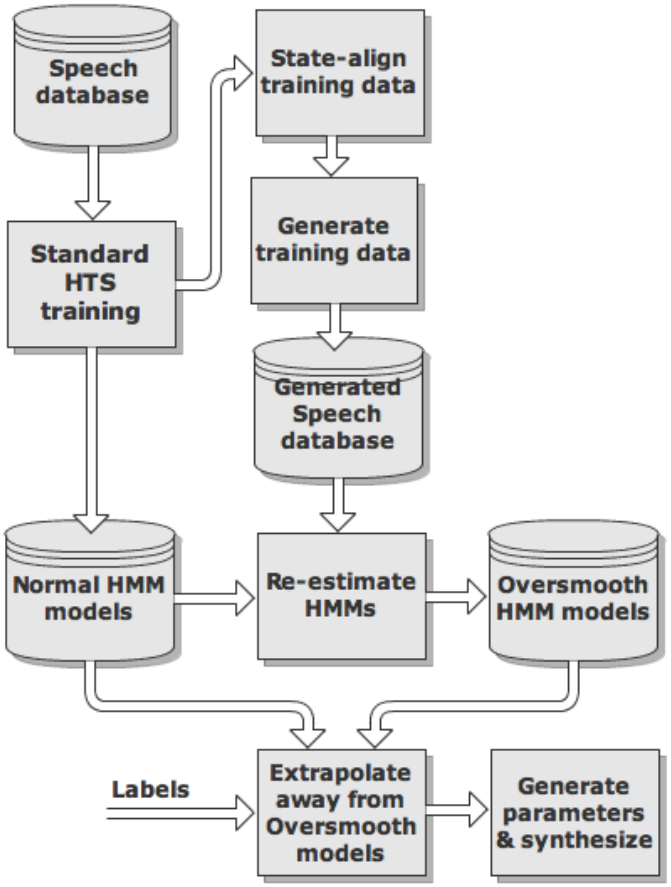
Especially suitable for high-pitched female voices!

II. Reduction of oversmoothing

A simple MGE inspired method to reduce oversmoothing: **Utilize the difference between the original and the over-smoothed model sets**

1. Generate training data
2. Re-estimate HMMs with the generated data to acquire oversmoothed models
3. Extrapolate away from the oversmoothed models

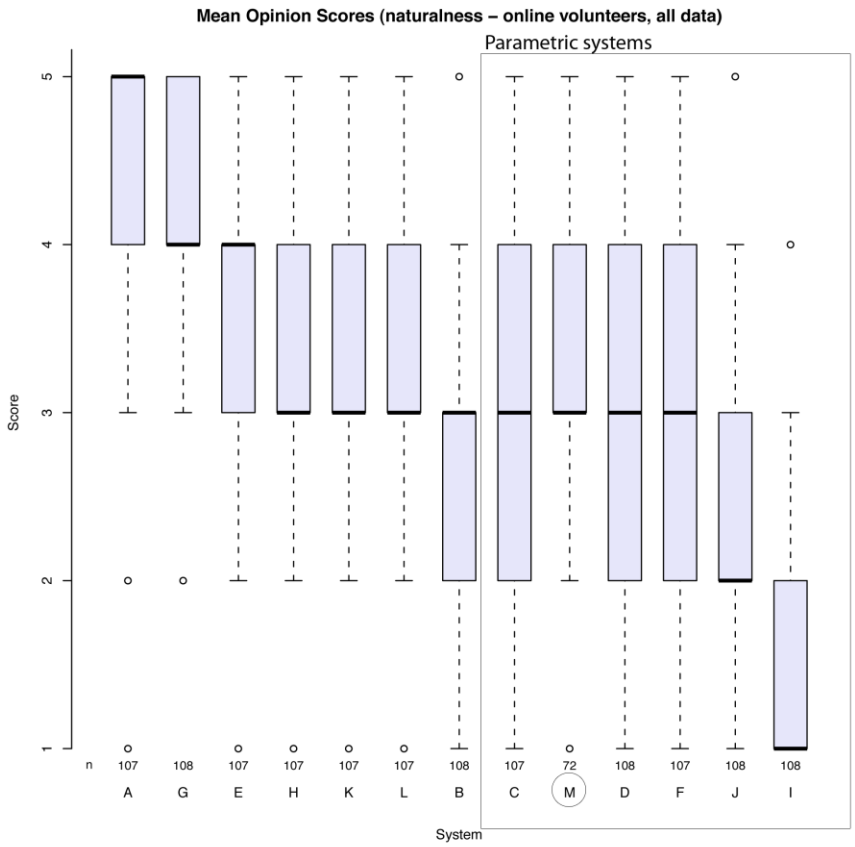
Weights tuned manually as in GV, subjectively better performance compared to GV

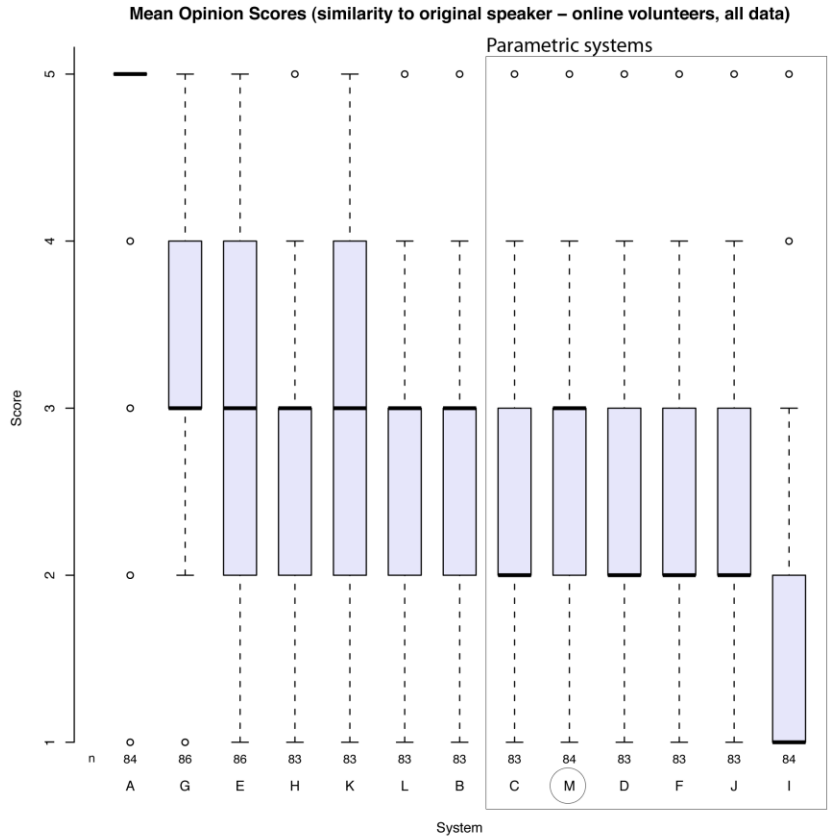


III. Results

- Overall MOS not significantly better than the HMM-based benchmark system
- Better quality compared to our last year's female voice
- Clear improvement on speaker similarity, likely due to the new pulse library method as well as SWLP parameterization

III. Results


Mean Opinion Scores (naturalness – online volunteers, all data)
Parametric systems
Score
1
2
3
4
5
n
107 108 107 107 107 107 108 107 72 108 107 108 108
A G E H K L B C M D F J I
System


Mean Opinion Scores (similarity to original speaker – online volunteers, all data)
Parametric systems
Score
1
2
3
4
5
n
84 86 86 83 83 83 83 83 84 83 83 83 84
A G E H K L B C M D F J I
System

III. Results




Thank You! Questions?