HMM-Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering

Master's Thesis Seminar

Tuomo Raitio

tuomo.raitio@tkk.fi

14.5.2008

Background

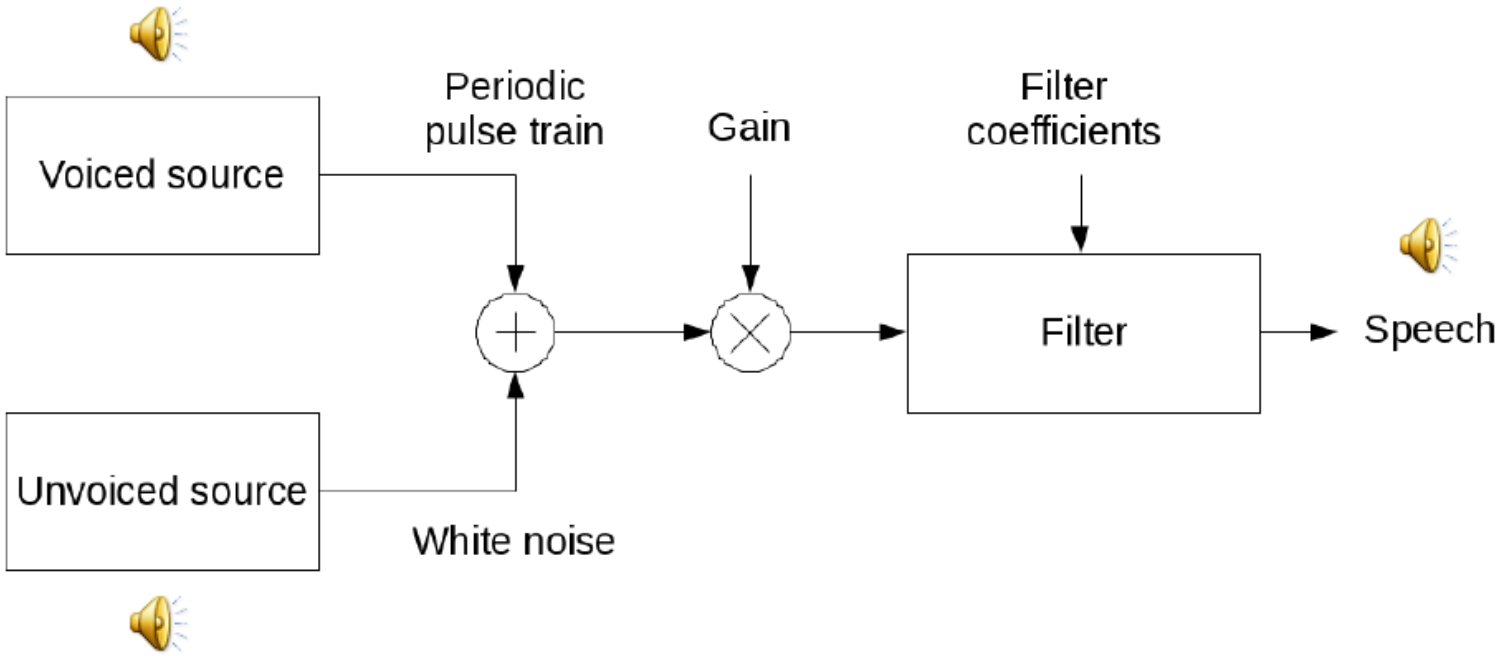
- HMM-based speech synthesis has been developed especially in Japan from the early 90's
- Phonetics and linguistics have been widely studied at the University of Helsinki. Lately, an HMM-based speech synthesizer was adopted to study Finnish speech synthesis
- The human voice production and especially the voice source has been an active research topic at the Helsinki University of Technology

- Collaboration between the Helsinki University of Technology and the University of Helsinki began in 2007 to develop a new HMM-based speech synthesis system

Speech Synthesis

- Speech synthesis is becoming increasingly important in modern information society
- **Text-to-speech (TTS)** systems are the most common and versatile today
- Text-to-speech system generates synthetic speech from arbitrary text

Speech Synthesis



Text-to-Speech (TTS)

Goals of TTS today:

- Create **natural sounding** synthetic speech with
  - Different **speaking styles**
  - Different **speaker characteristics**
  - Expression of **emotions**

- **Flexible** speech synthesis
  - Easy **adaptation** to these properties

Text-to-Speech (TTS)

Currently two major synthesis techniques

1. **Unit selection** based approach
   - Based on selection and concatenation of prerecorded acoustical units
   - Highly natural synthetic speech
   - Poor adaptability to speaking styles, speaker characteristics and emotions
   - Large memory requirement for storing the acoustical units

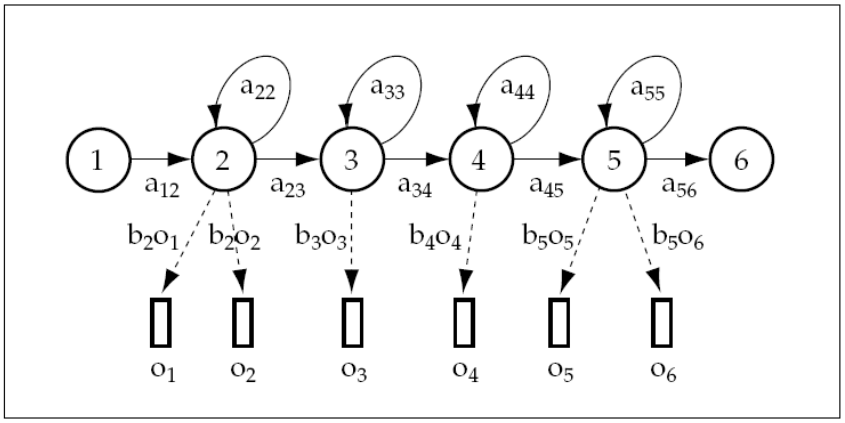
Text-to-Speech (TTS)

2. **HMM-based** approach

- Based on modeling of speech parameters with Hidden Markov Models (HMMs)
- Better adaptability to speaking styles, speaker characteristics and emotions → Flexible speech synthesis
- Small memory requirement

Hidden Markov Models

- Statistical models for various types of sequential data
- A finite state machine which generates a sequence of time observations

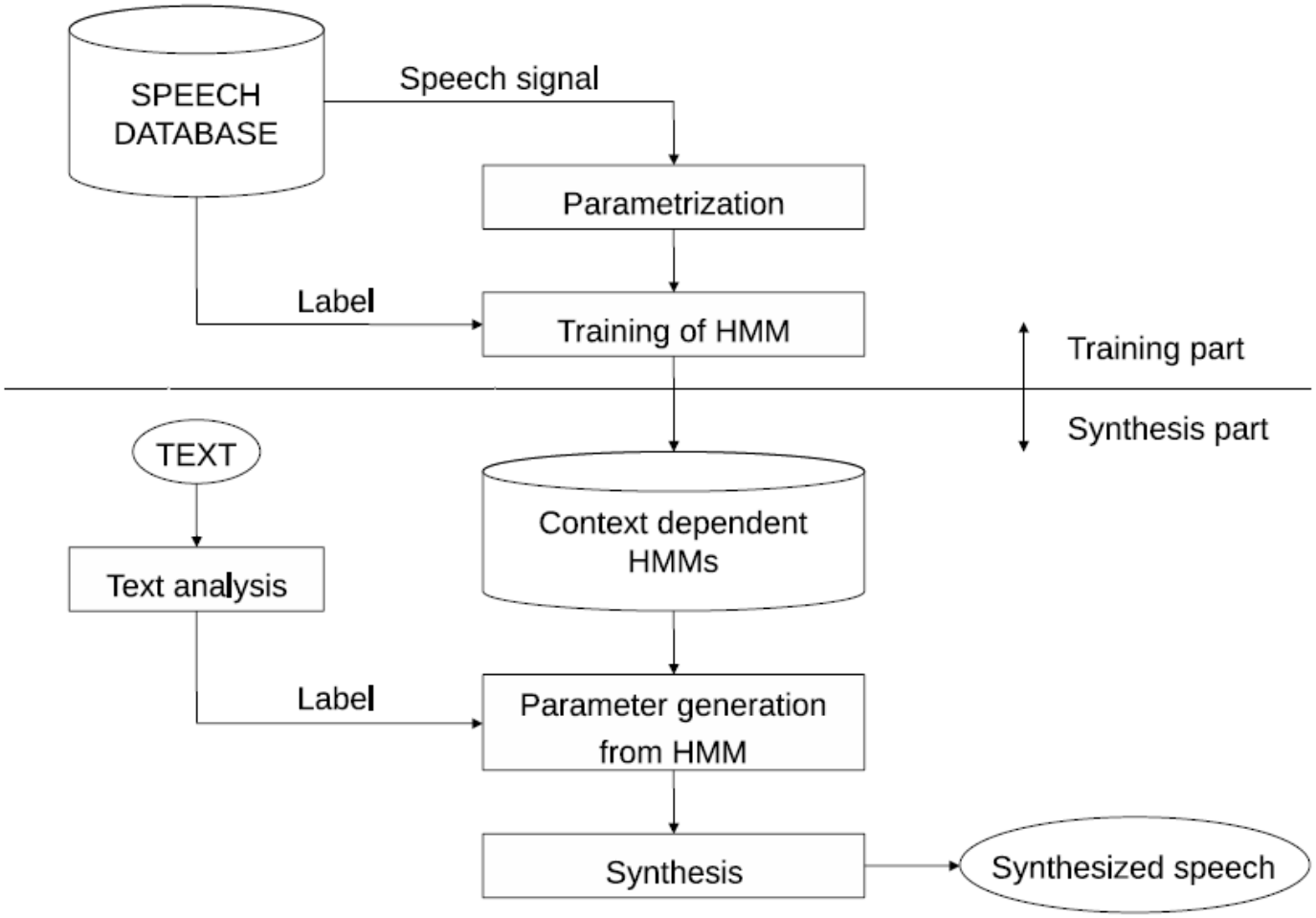


- 6-state left-to-right HMM structure
- a_{ij} - state transition probability from state i to j
- b_i - output probability density
- o_t - observation at time instant t

HMM-based Speech Synthesis

- Two stages
  - **Training**: HMM system is first trained with a speech database
  - **Synthesis**: Speech is synthesized from trained HMM according to text input

HMM-based Speech Synthesis



HMM-based Speech Synthesis

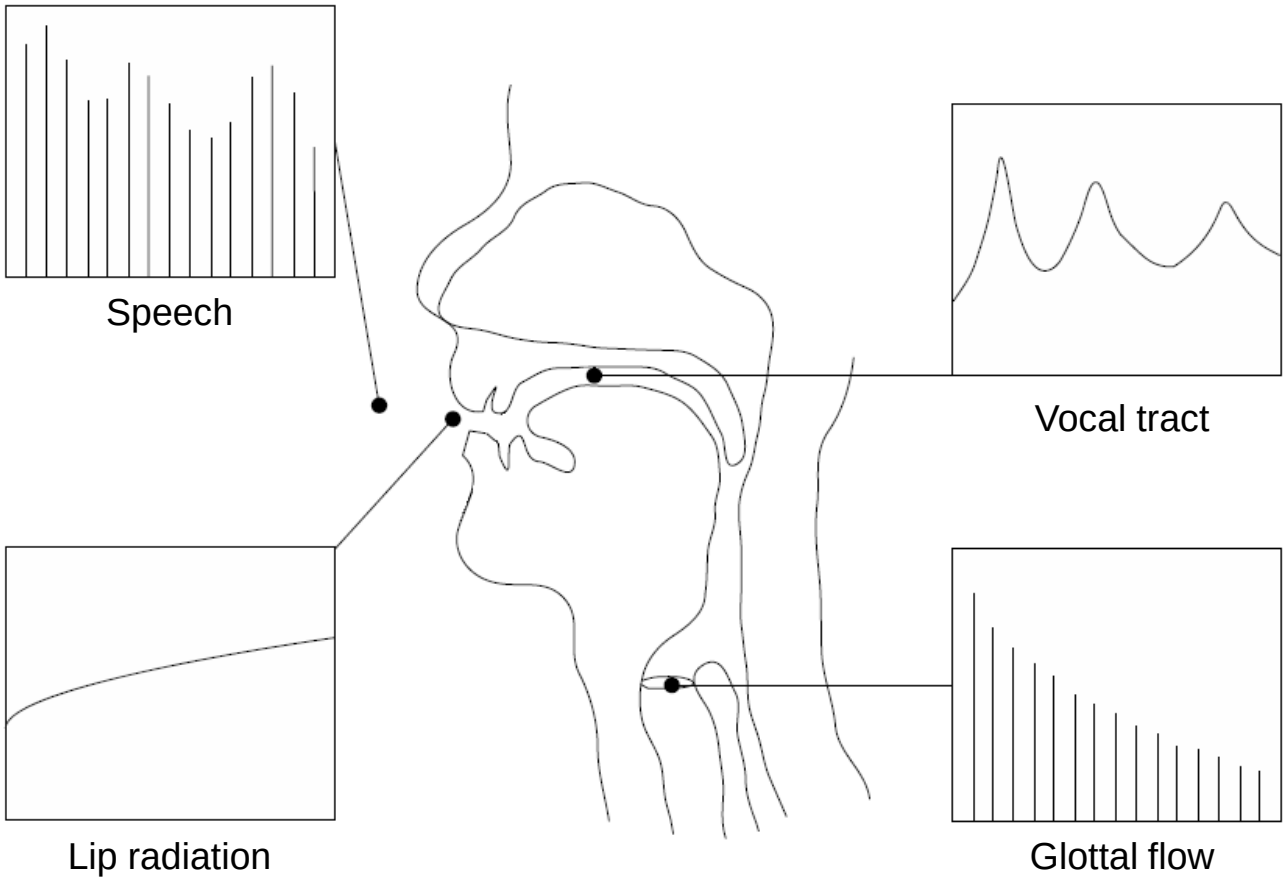
- Problem: HMM-based speech synthesis suffers from degraded naturalness in quality
  - Potential reason is the use of signal generation techniques which are oversimplified to properly mimic natural speech pressure waveforms

→ New **glottal inverse filtering** based parametrization and synthesis method that models the **natural behavior of the voice source**!

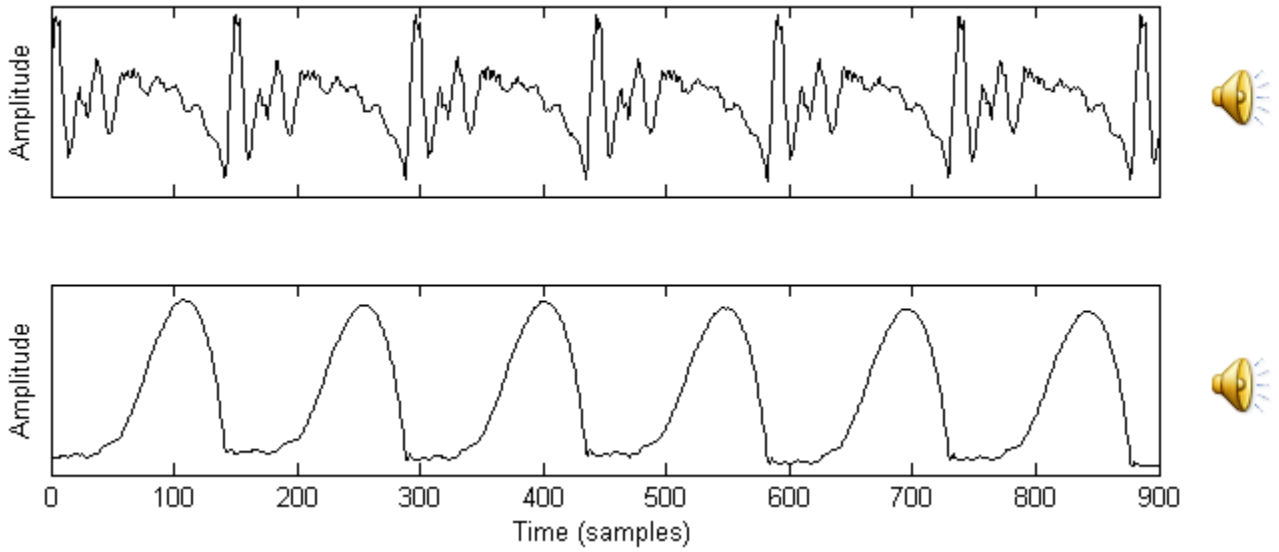
Glottal Inverse Filtering

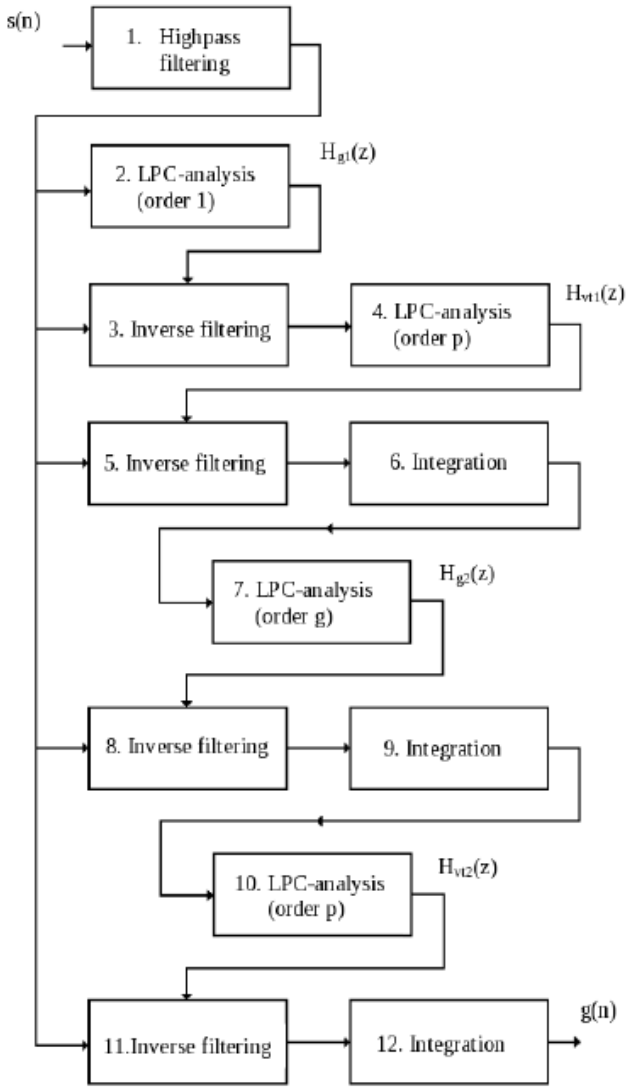
- Glottal inverse filtering estimates the **glottal volume velocity waveform (glottal flow)** by canceling the effects of
  - Vocal tract and
  - Lip radiation

Glottal Inverse Filtering



Glottal Inverse Filtering



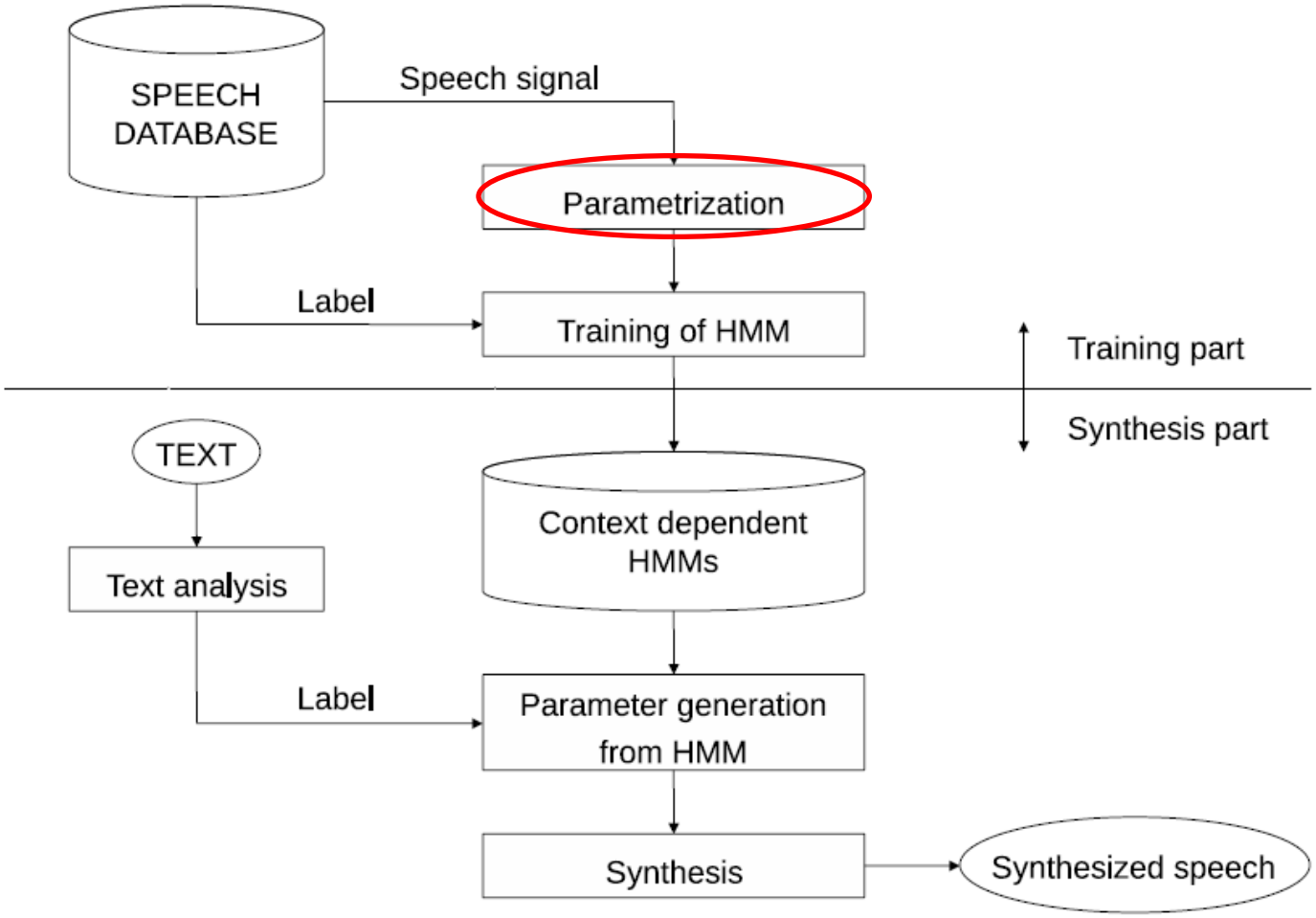


- Iterative Adaptive Inverse Filtering (IAIF)
- Automatically estimates the glottal flow by canceling the effects of the vocal tract and lip radiation
- Based on linear prediction (LP)

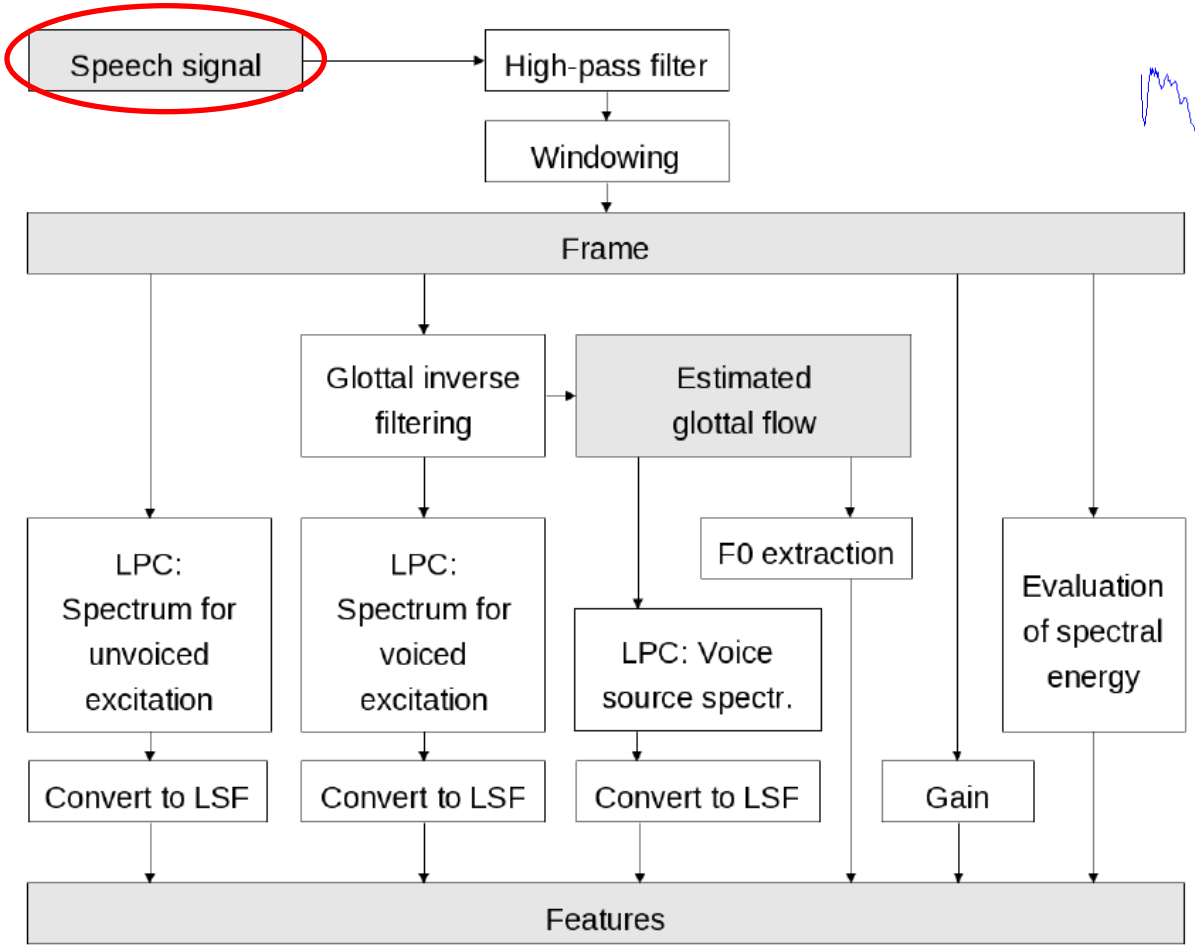
New Text-to-Speech System

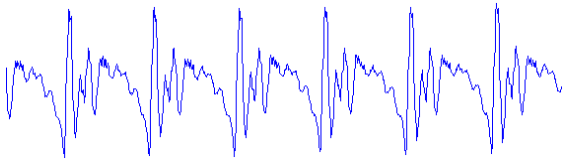
- Improvements to HMM-based speech synthesis:
  - Utilization of **glottal inverse filtering** in order to extract and model the characteristics of the voice source
  - **Individual modeling** of the voice source characteristics in the HMM system
  - Utilization of **natural glottal flow pulses** for creating the voice source

Parametrization of Speech

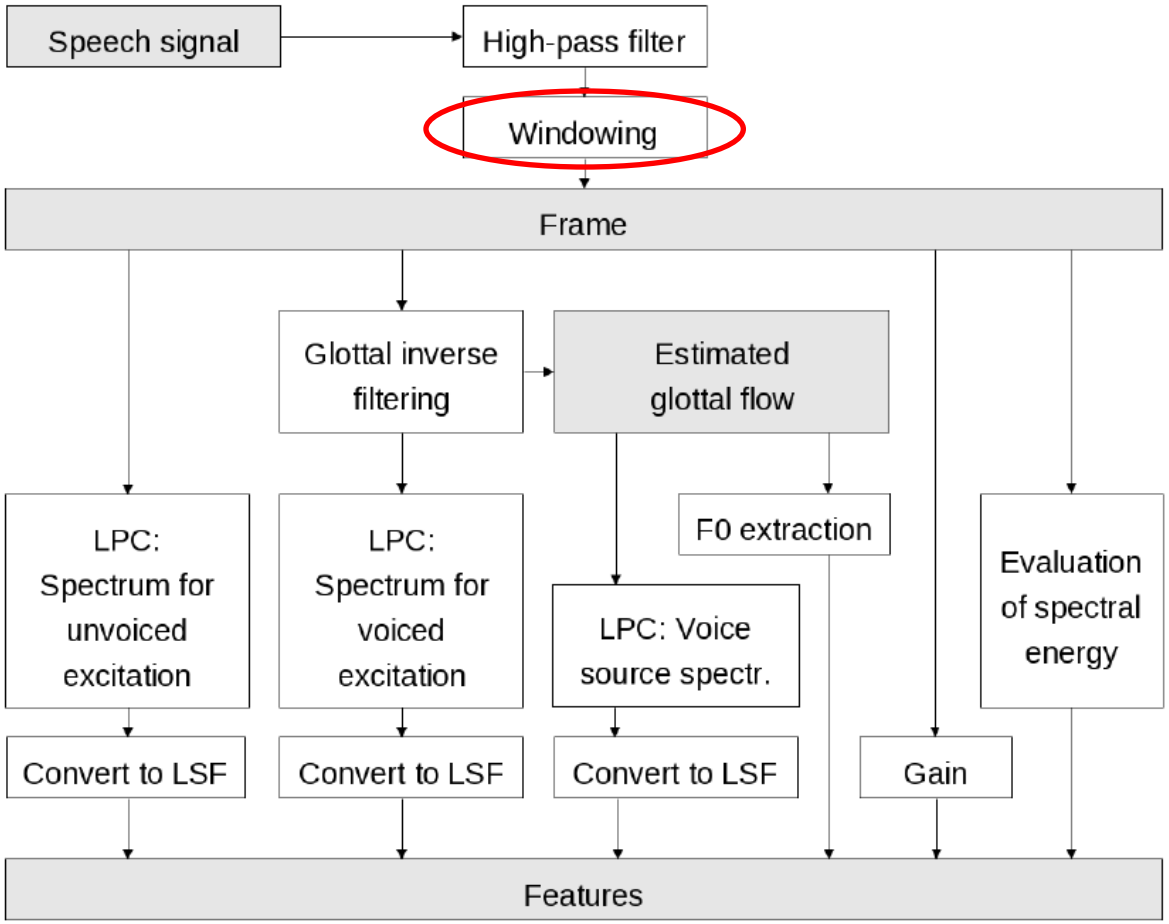


Parametrization of Speech



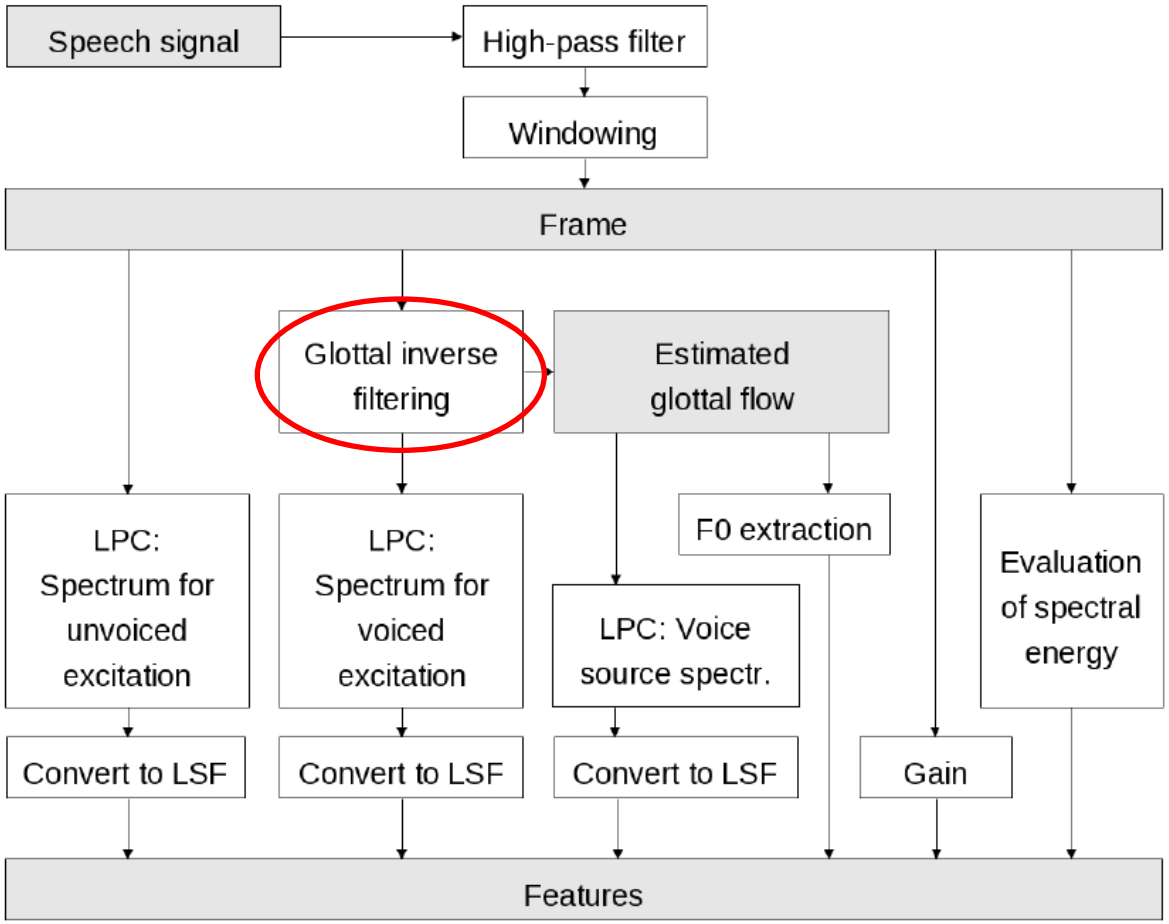


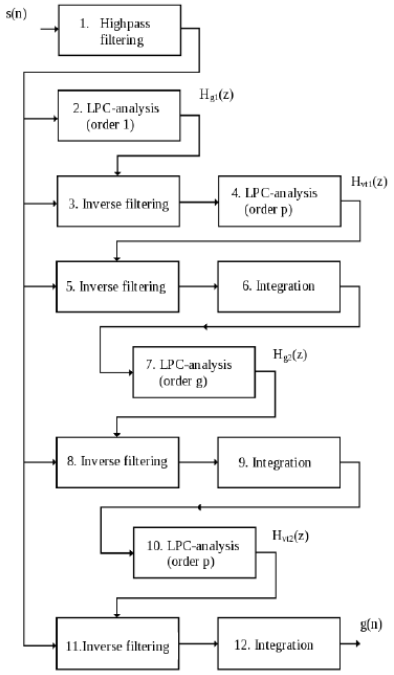
Parametrization of Speech



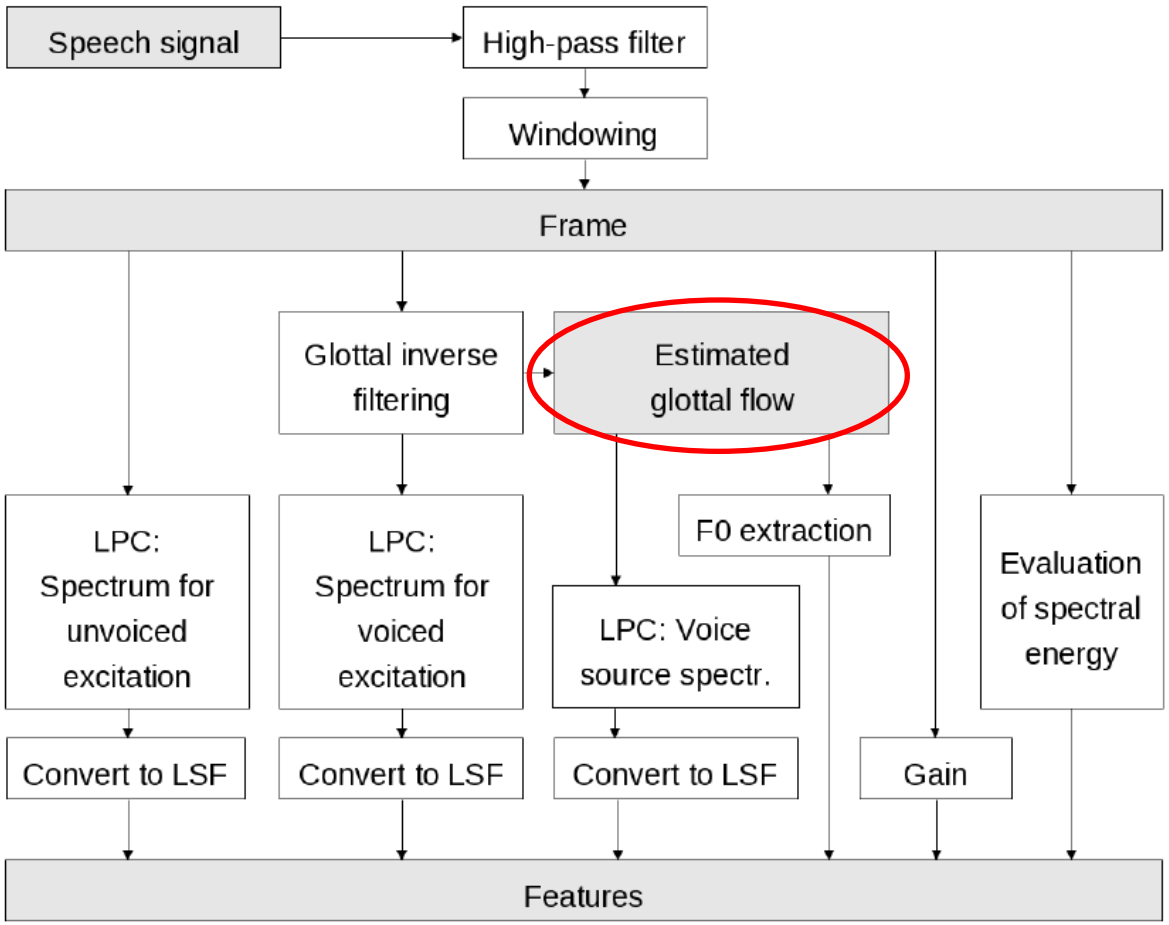
25-ms rectangular window

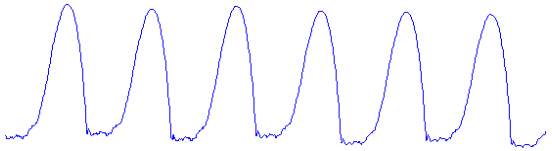
Parametrization of Speech



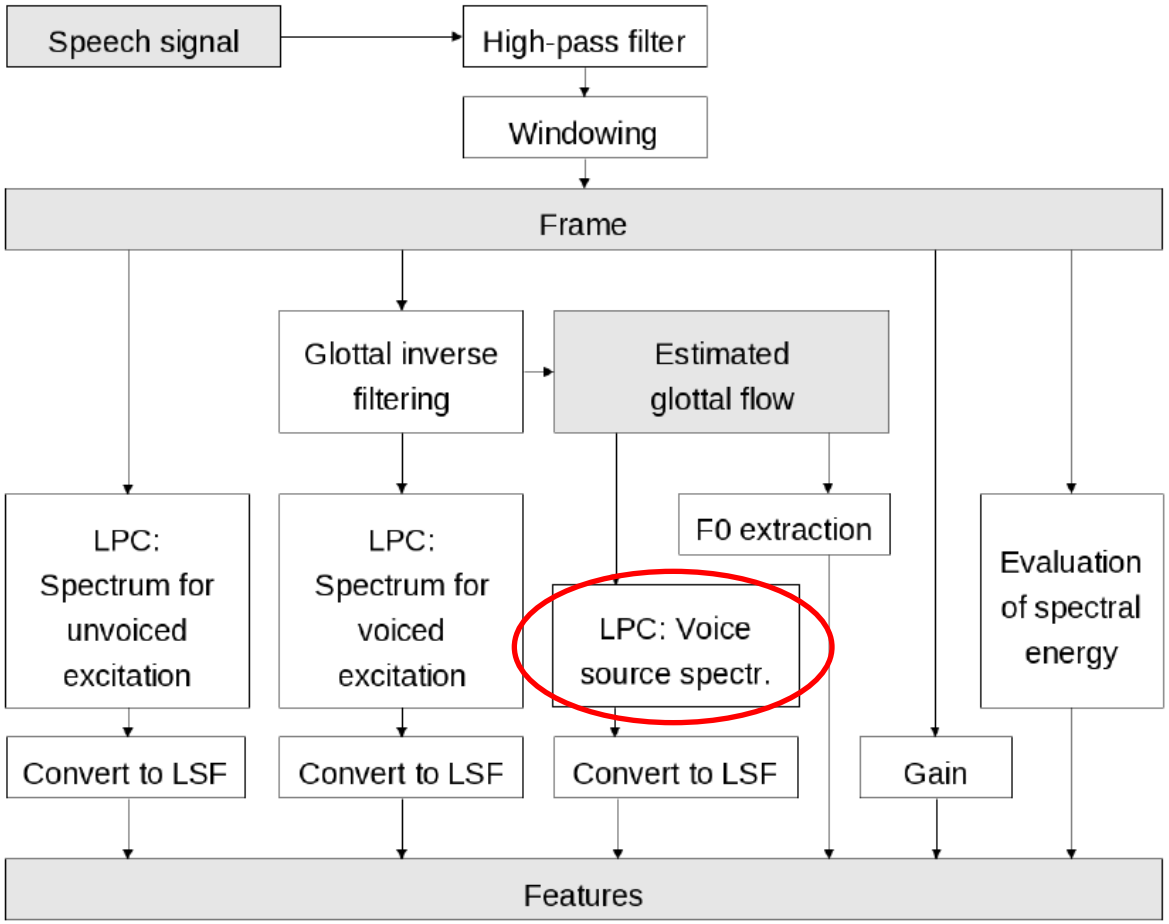


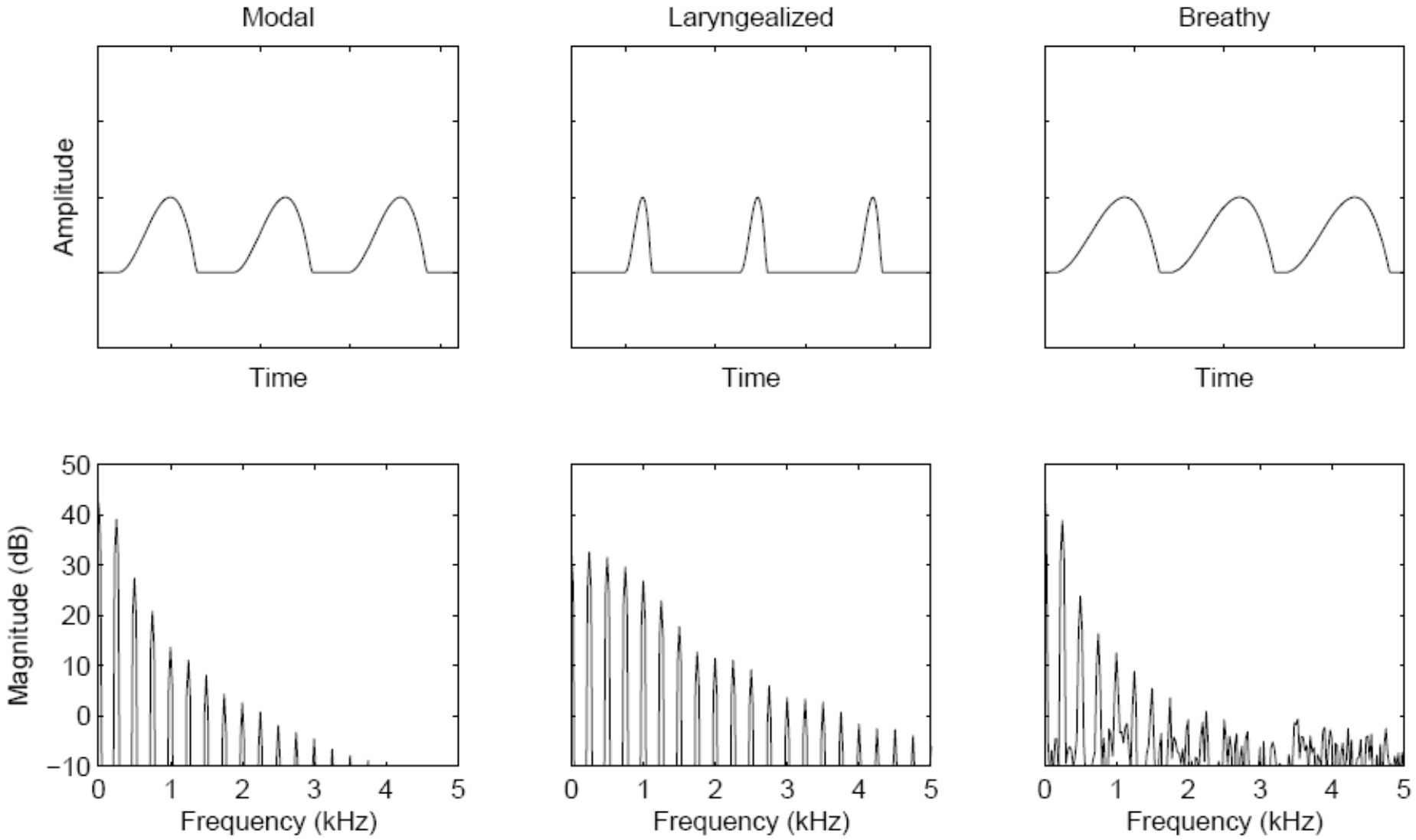
Parametrization of Speech



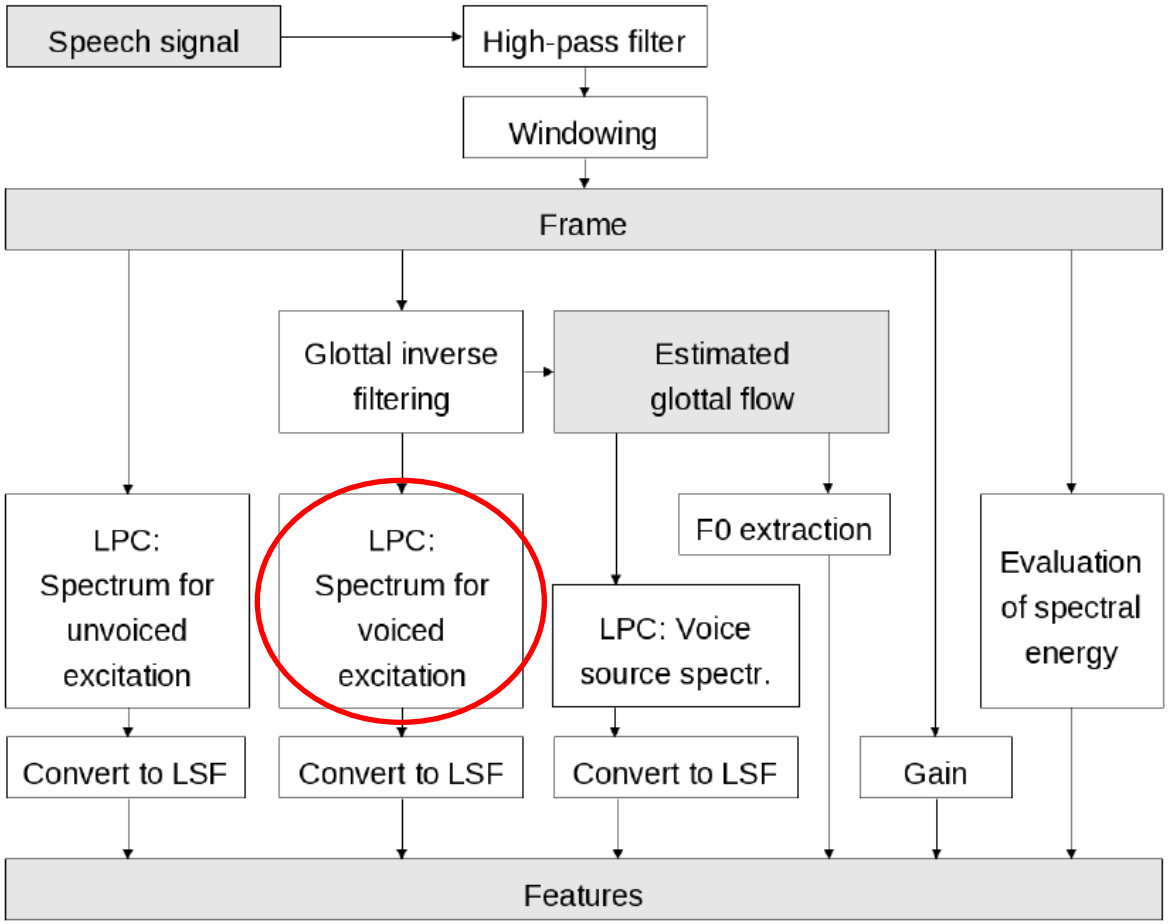


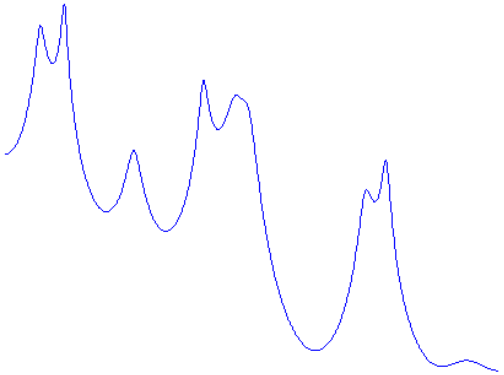
Parametrization of Speech




Modal
Laryngealized
Breathy
Amplitude
Time
Time
Time
Magnitude (dB)
50
40
30
20
10
0
−10
0
1
2
3
4
5
Frequency (kHz)
Frequency (kHz)
Frequency (kHz)

Parametrization of Speech

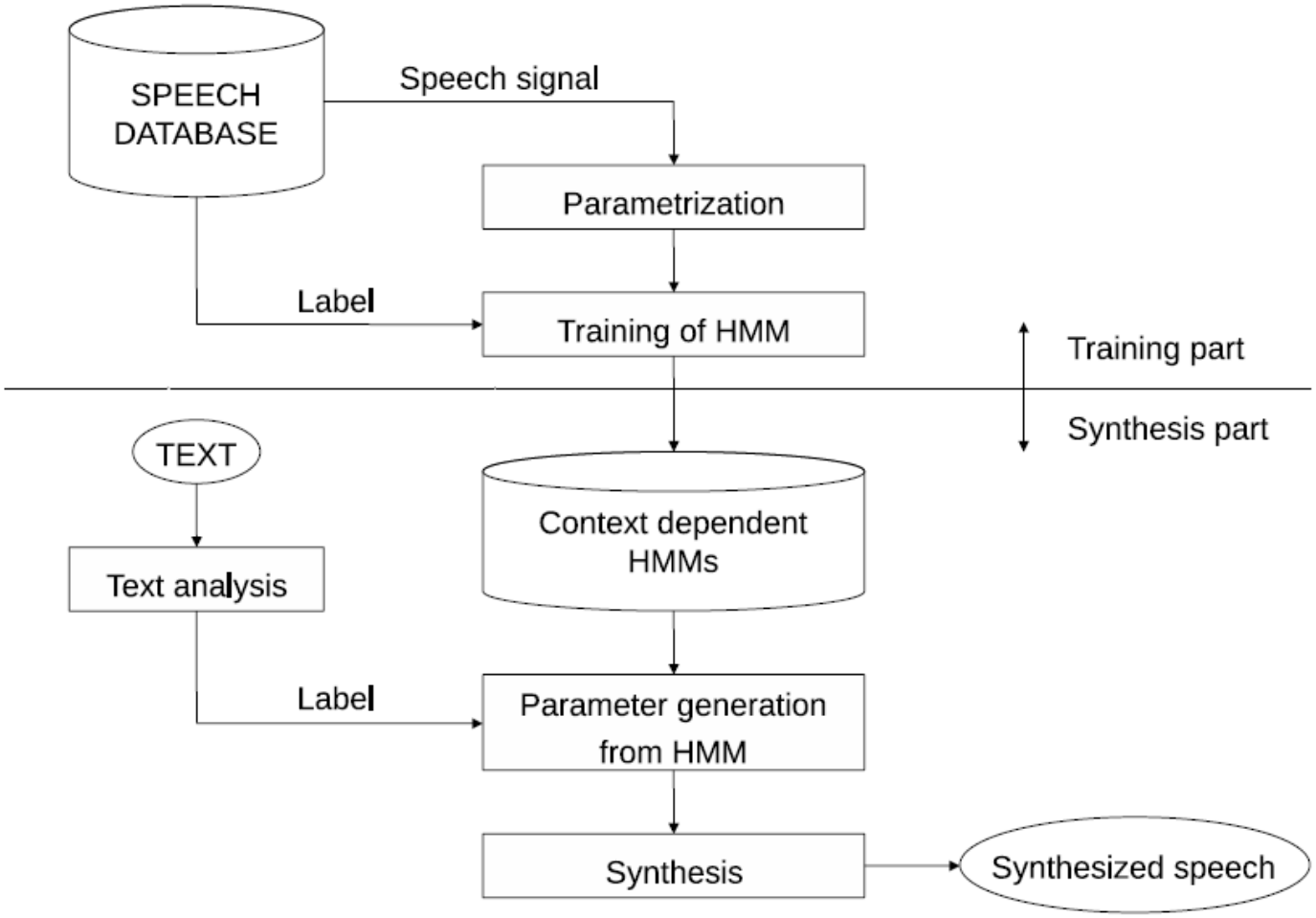


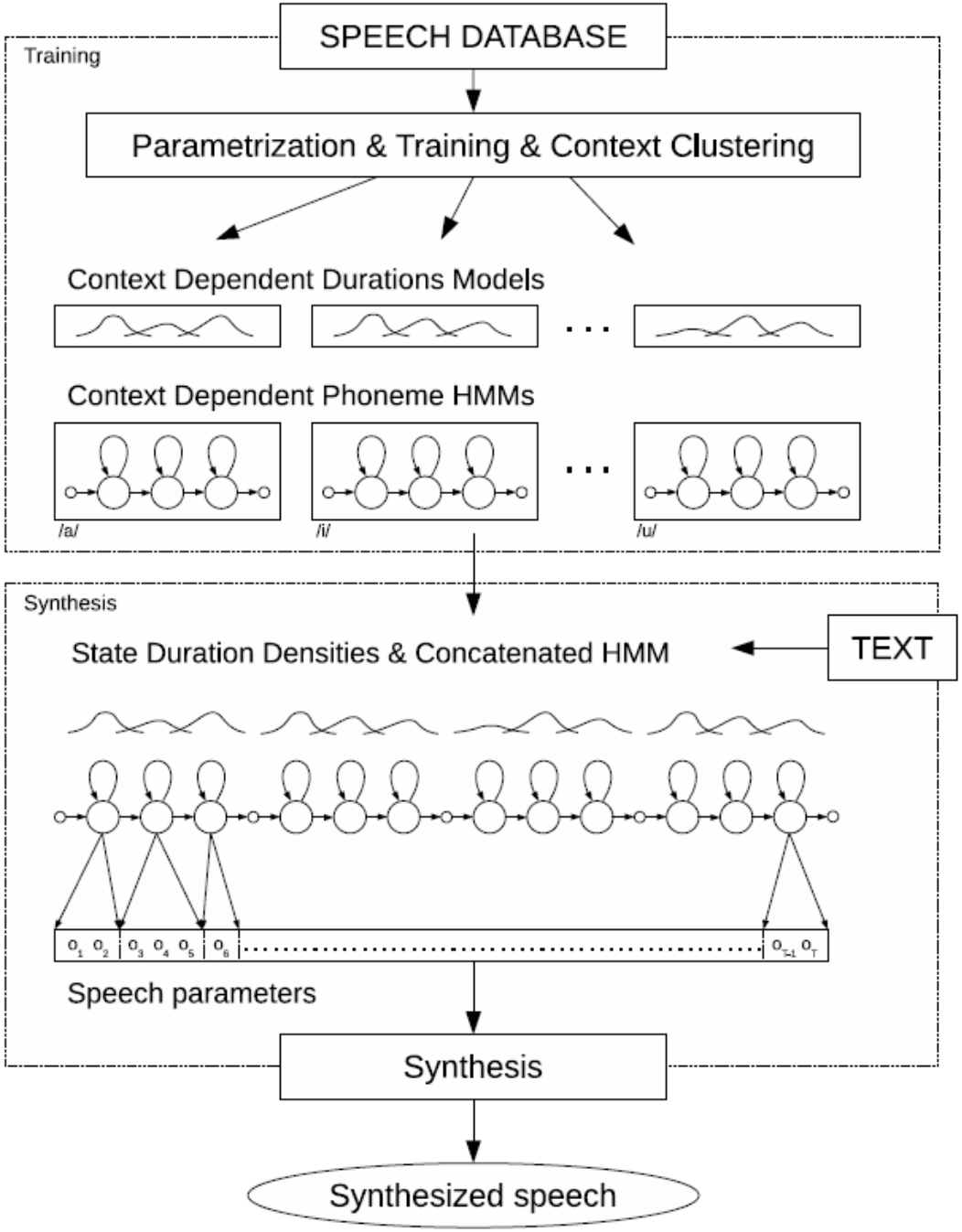


Extracted Speech Features

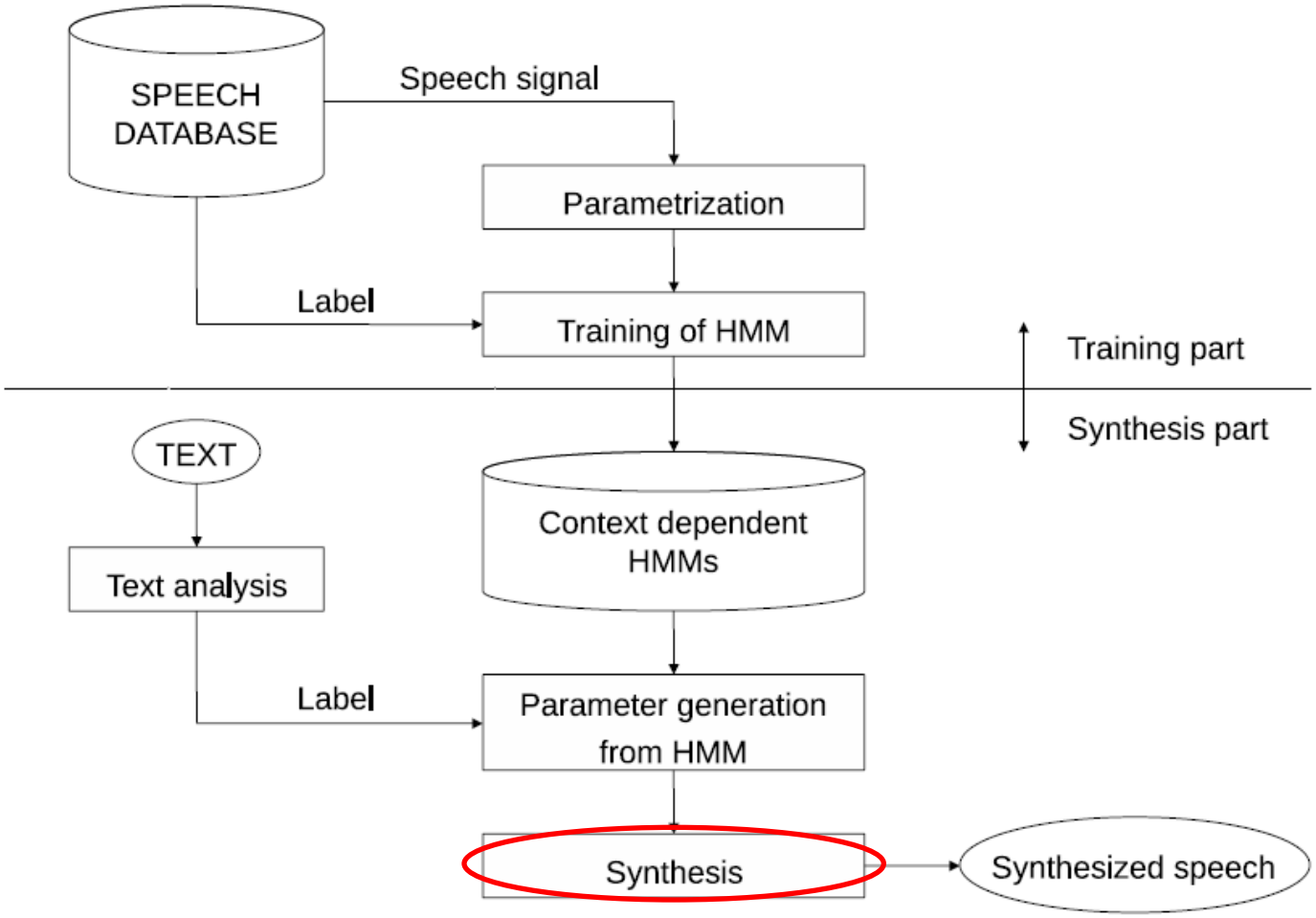
Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Spectral energy	5
Voice source spectrum	10
Voiced spectrum	20
Unvoiced spectrum	20

HMM Framework

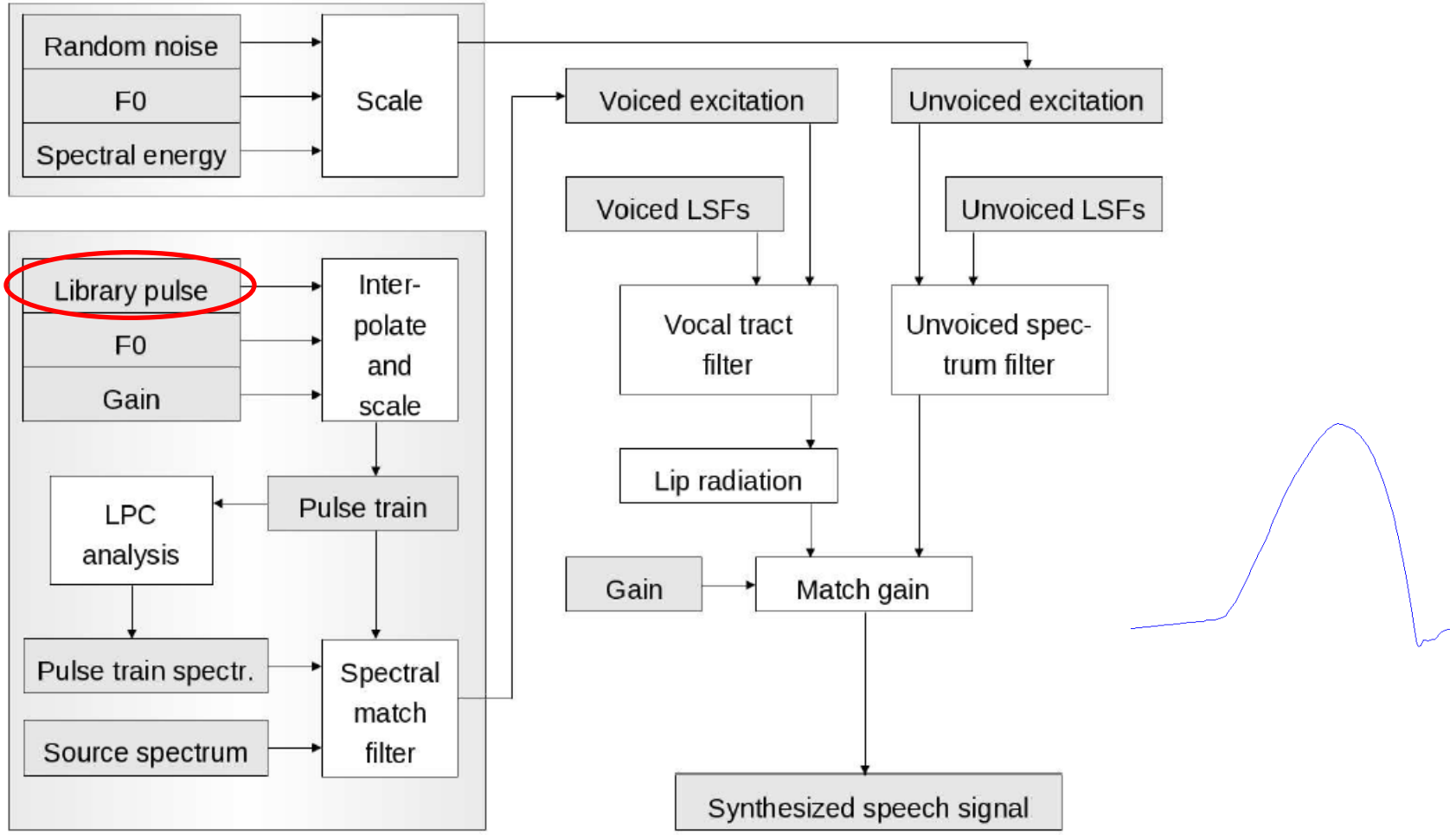



SPEECH DATABASE
Training
Parametrization & Training & Context Clustering
Context Dependent Durations Models
Context Dependent Phoneme HMMs
/a/
/i/
/u/
Synthesis
State Duration Densities & Concatenated HMM
TEXT
Speech parameters
Synthesis
Synthesized speech

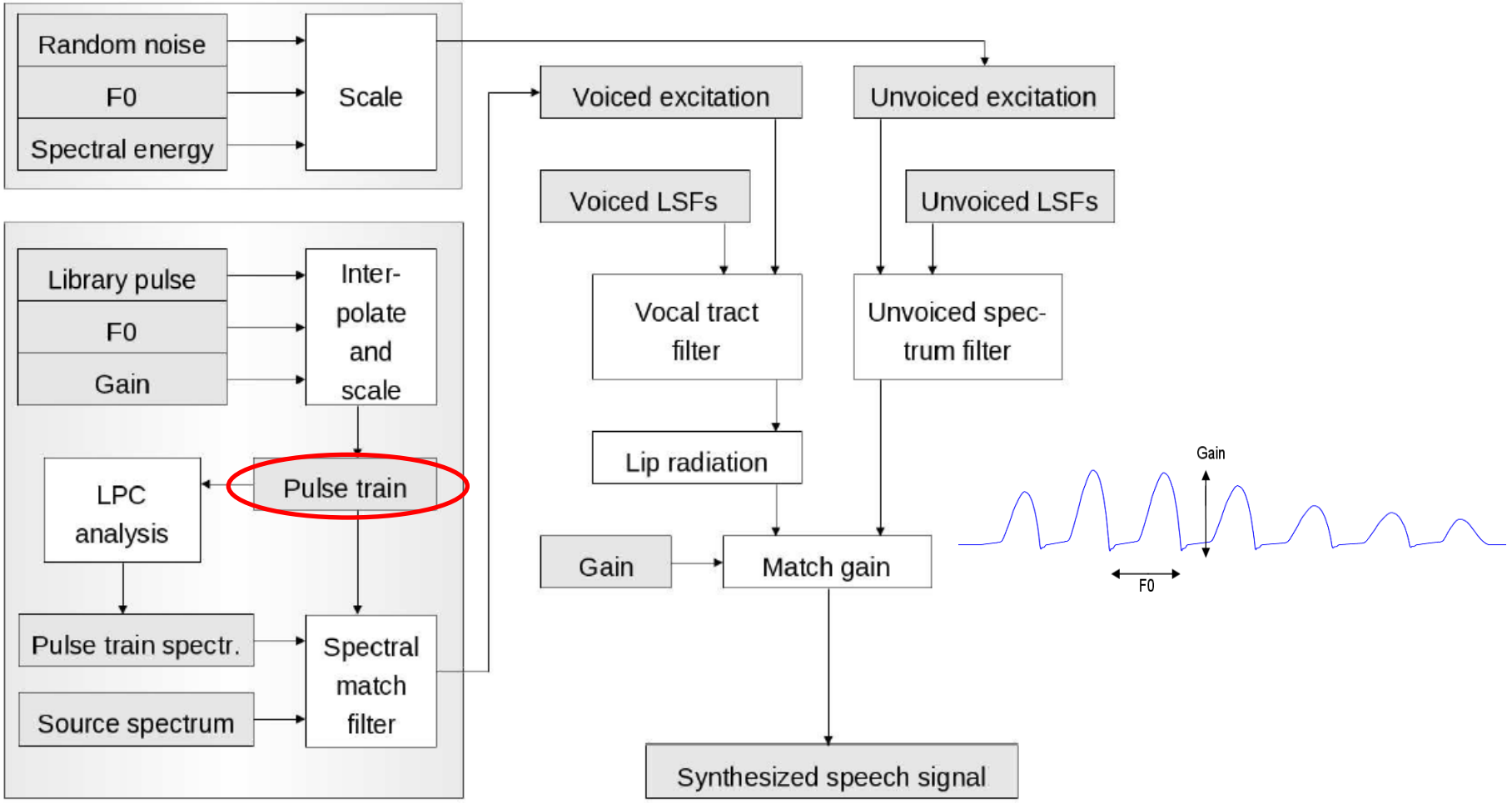
Synthesis from Parameters



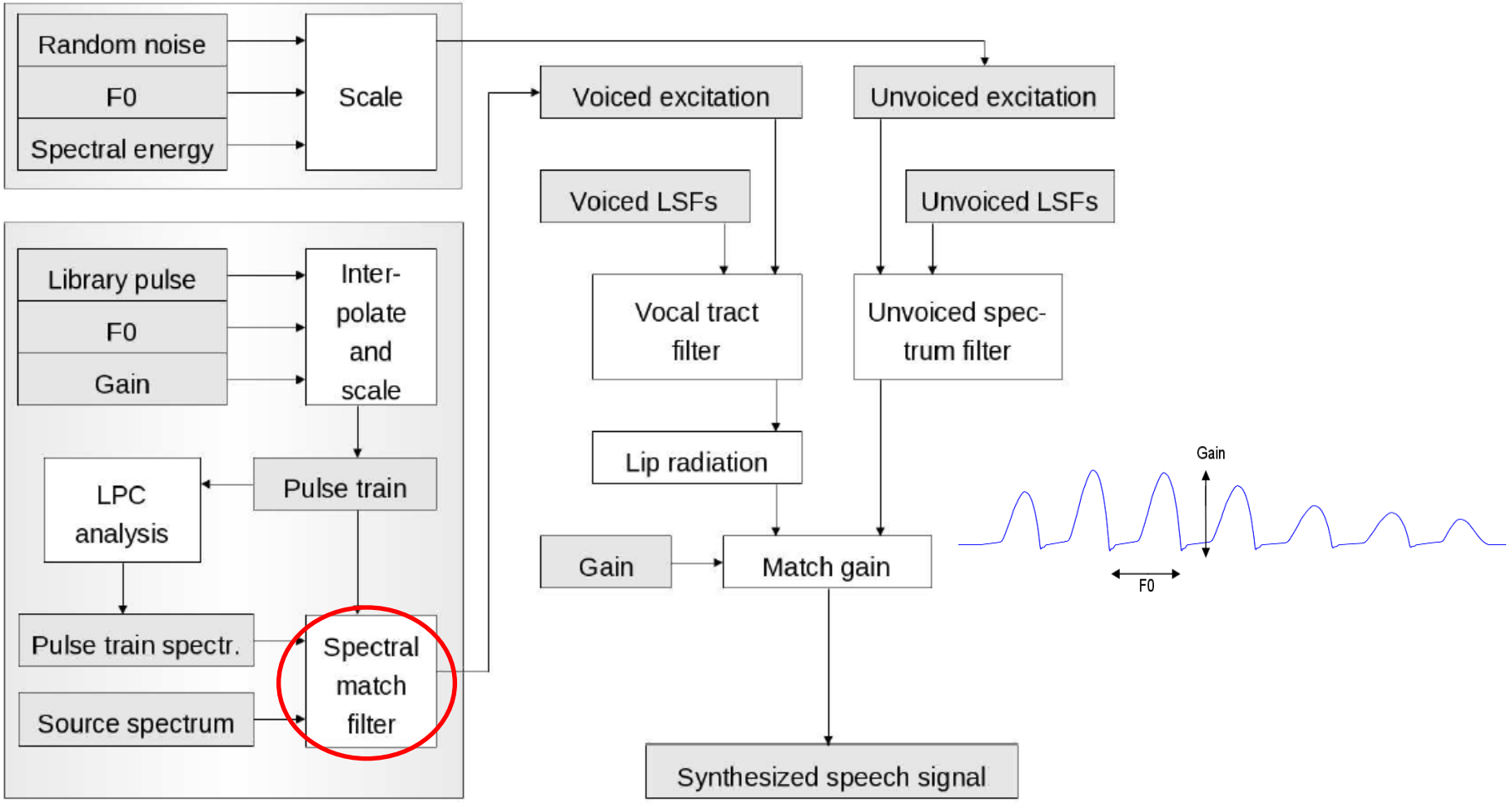
Synthesis from Parameters



Synthesis from Parameters

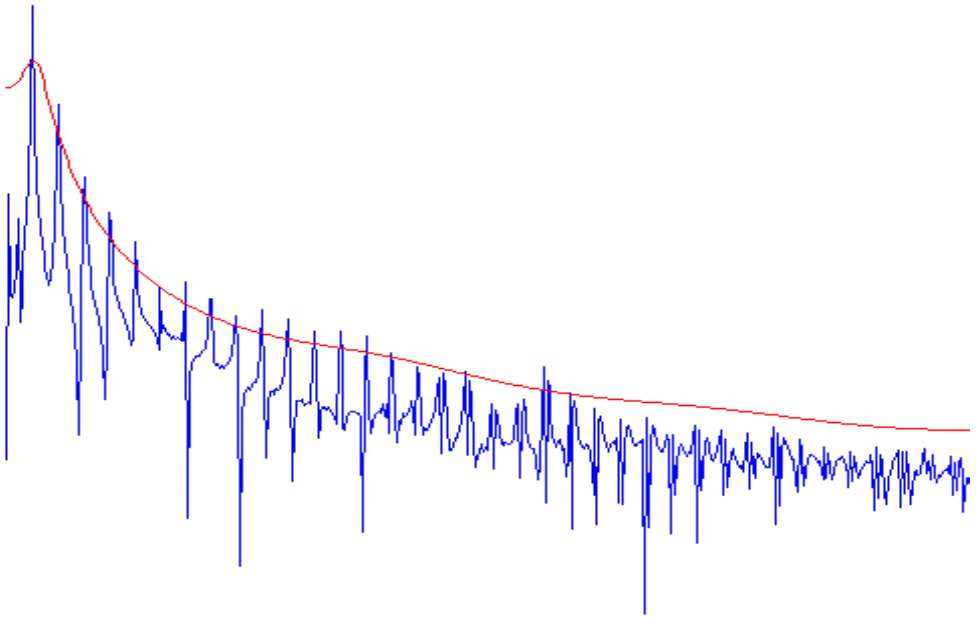


Synthesis from Parameters

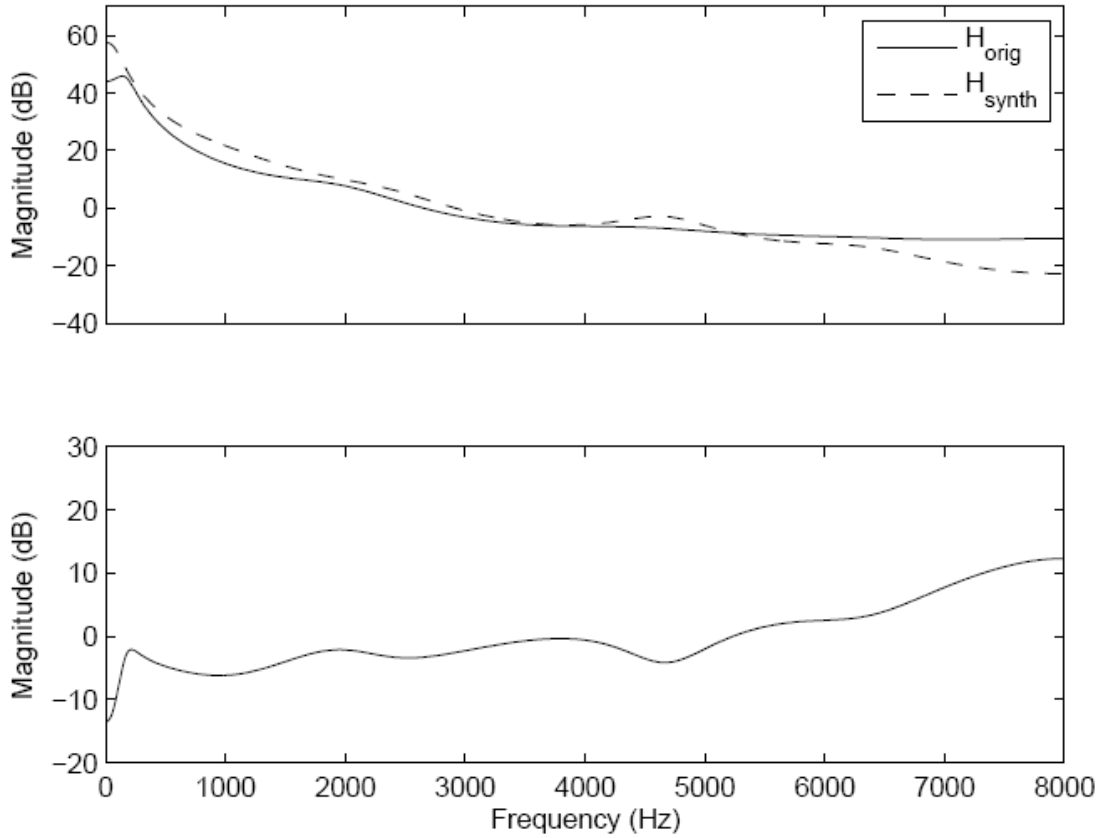


Spectral Matching of the Voice Source

- The spectrum of the pulse train is further modified with an adaptive IIR filter to imitate the natural variation in the voice source

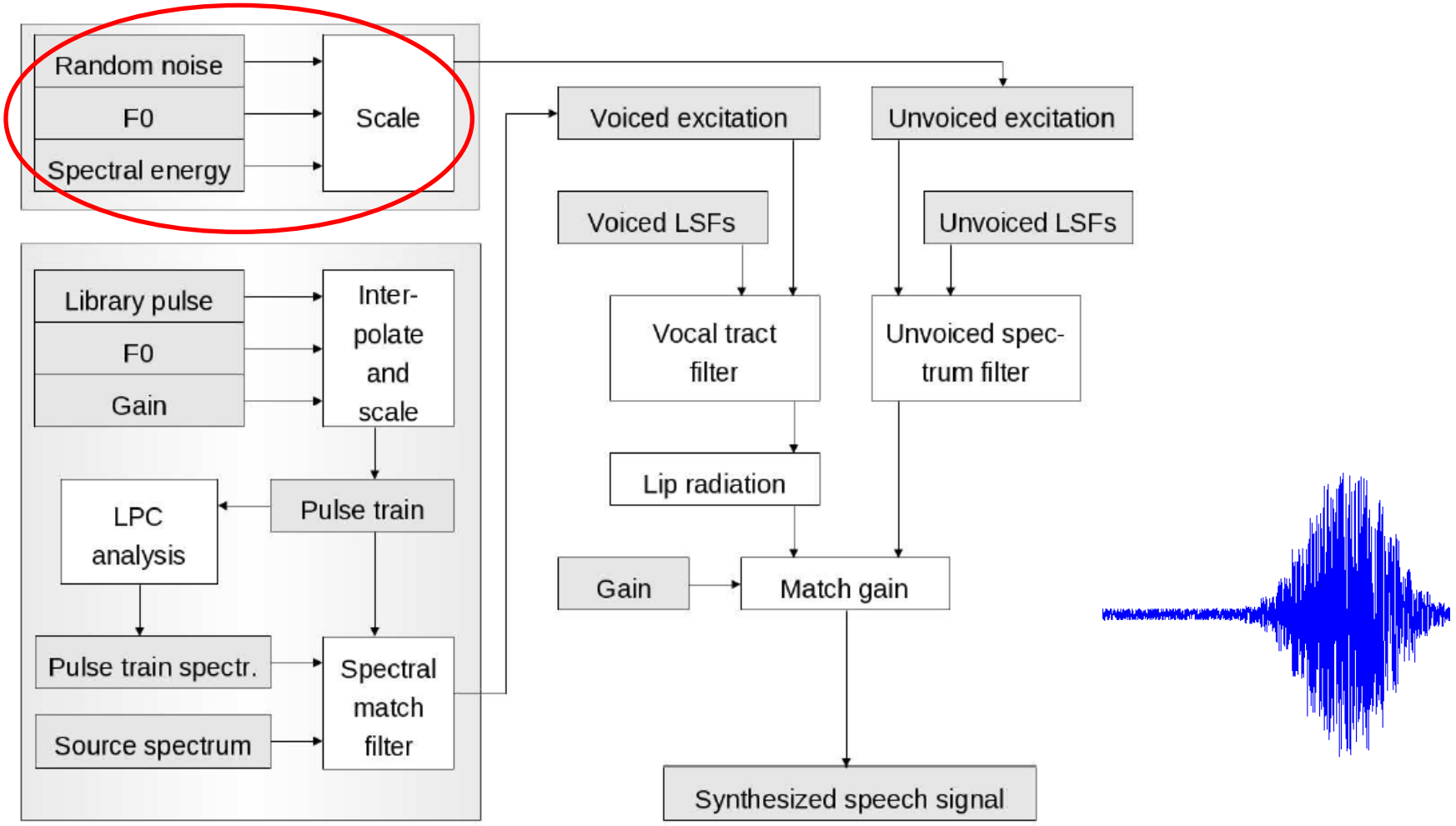


Spectral Matching of the Voice Source

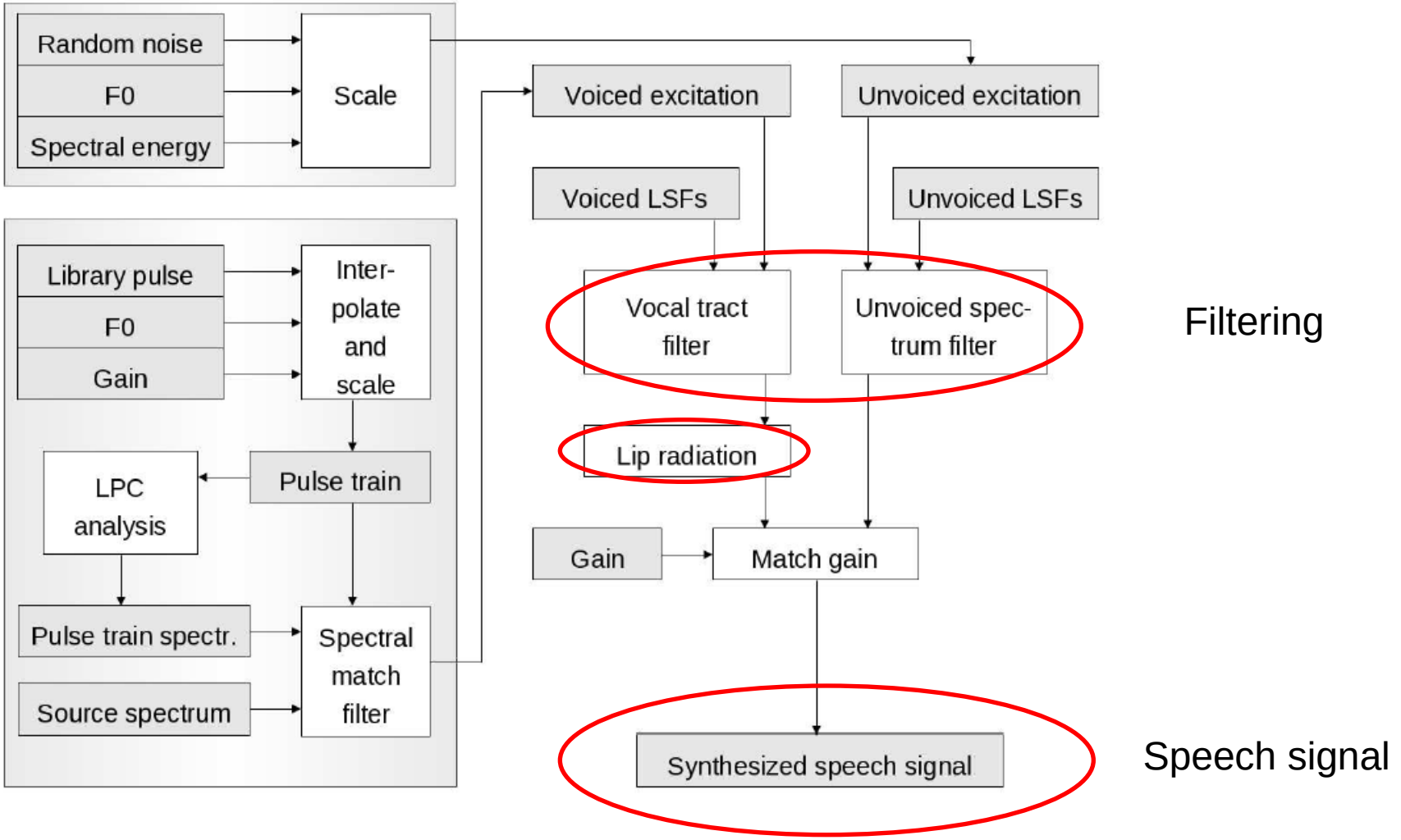


$$H_{\text{match}}(z) = \frac{H_{\text{orig}}(z)}{H_{\text{synth}}(z)}$$

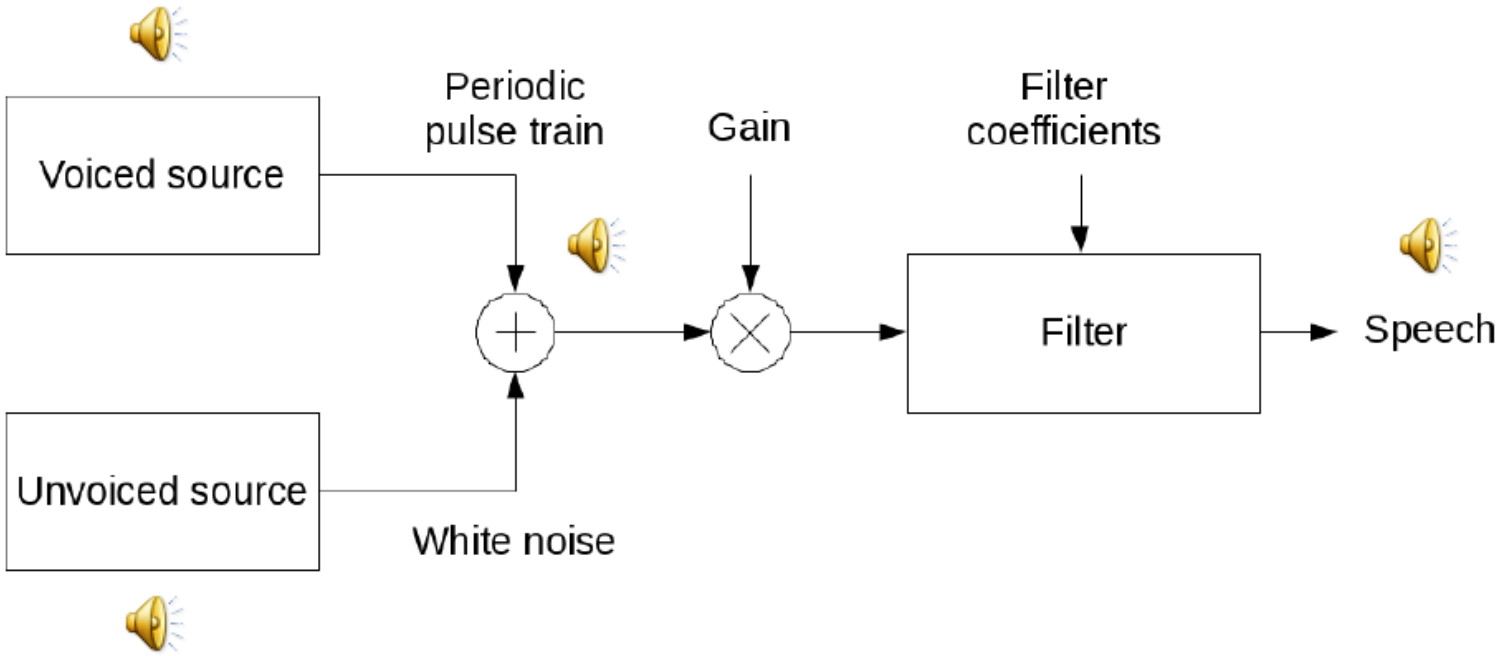
Synthesis from Parameters



Synthesis from Parameters



Speech from Parameters



Listening Tests

Two listening tests:

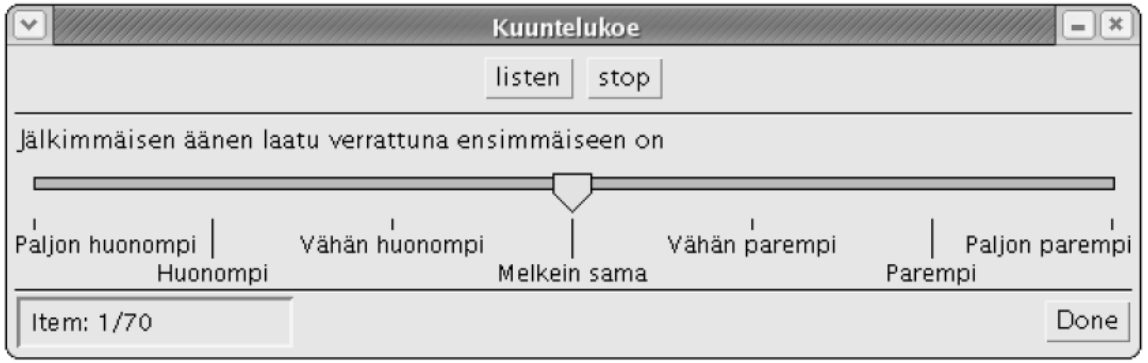
- **Category Comparison Rating (CCR) test**
  - New system was compared to natural speech and traditional HMM-based speech synthesizer
- **Pair Comparison test**
  - New system was compared to traditional HMM-based speech synthesizer

CCR Test

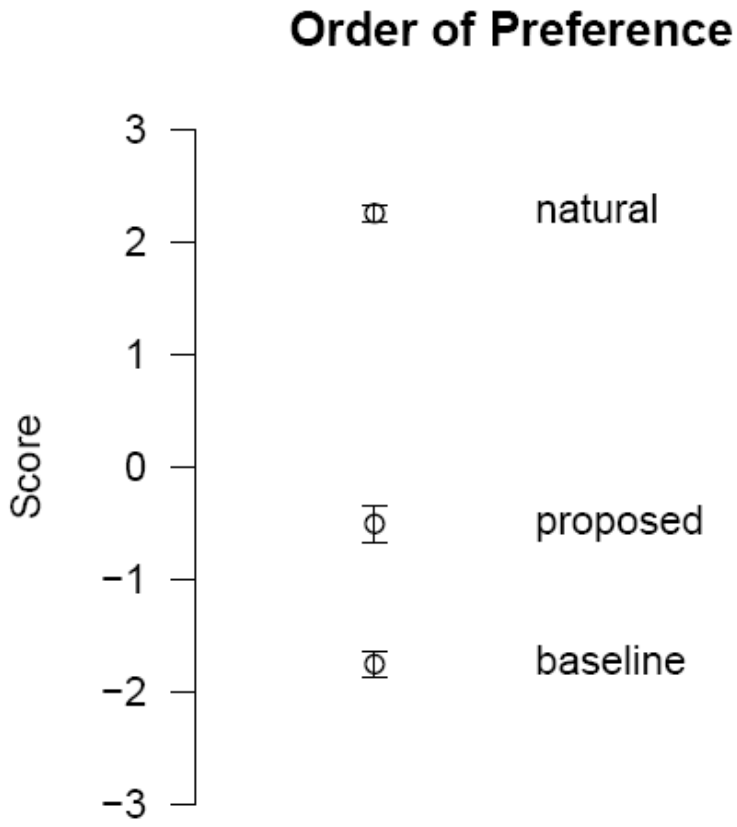
- Listeners assessed the quality of the sample **A** compared to the quality of sample **B** on the 7-point Comparison Mean Opinion Score (CMOS)

3	Much Better
2	Better
1	Slightly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

- User interface



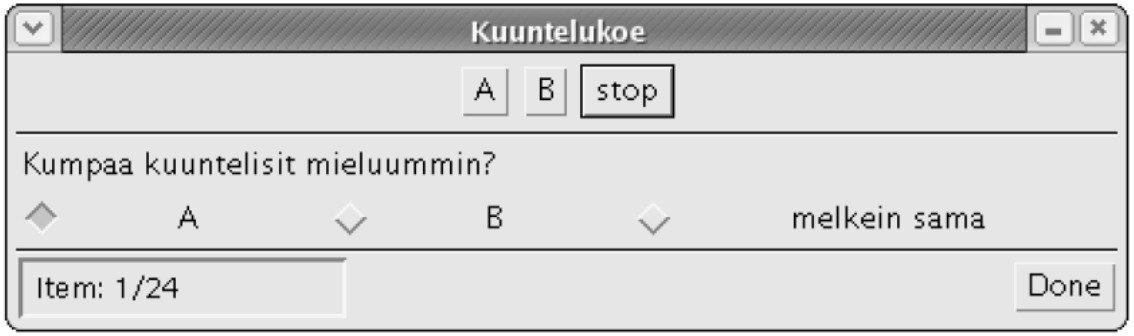
Results


Order of Preference
Score
3
2
1
0
−1
−2
−3
natural
proposed
baseline

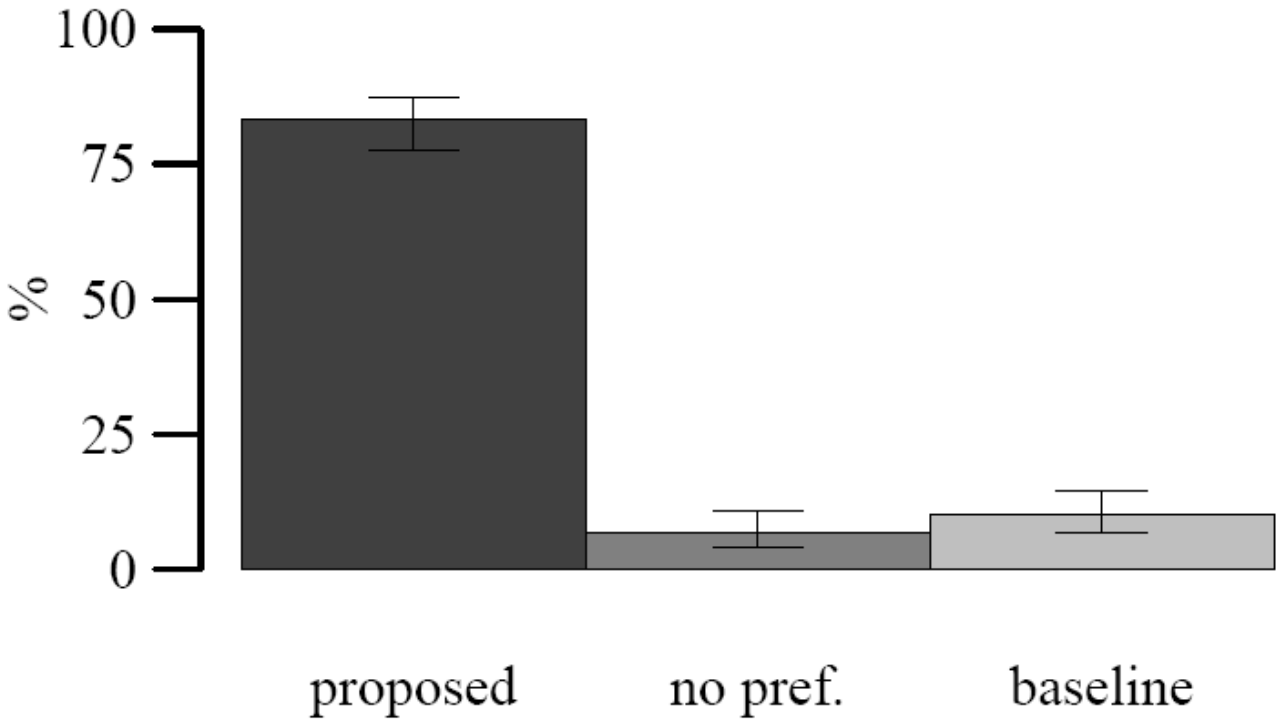
Pair Comparison test

- Subjects listened to samples **A** and **B**, and **selected** the one they would rather listen to

- User interface



Results



Listening Tests

- The listening test show that
  - The new TTS system is able to generate **highly natural synthetic speech** with **specific speaker characteristics**
  - The quality of the new TTS system is considerably better compared to a traditional HMM-based TTS system

Samples

- Sample 1 
- Sample 2 
- Sample 3 

Further Development

- Development of the new TTS system continues to fully utilize the new techniques introduced in this work

References

- Tokuda, K., Zen, H. & Black, A. W. An HMM-based speech synthesis system applied to English, *Proceedings of 2002 IEEE Workshop on Speech Synthesis* pp. 227–230, 2002

- Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, *Speech Communication* 11(2-3): 109–118, 1992

- HMM-based speech synthesis system. http://hts.sp.nitech.ac.jp