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

Master’s Thesis Seminar
Tuomo Raitio
tuomo.raitio@tkk.fi
14.5.2008
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

Voiced source

Unvoiced source

Periodic pulse train

Gain

Filter coefficients

Filter

Speech
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
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

- 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

Speech

Vocal tract

Lip radiation

Glottal flow
Glottal Inverse Filtering

Speech $\rightarrow$ Glottal Inverse Filtering $\rightarrow$ Glottal flow

Amplitude

Time (samples)
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

1. SPEECH DATABASE
   - Speech signal
   - Label

2. Parametrization
   - Training of HMM

3. TEXT
   - Text analysis
   - Label

4. Context dependent HMMs
   - Parameter generation from HMM

5. Synthesis
   - Synthesized speech

Training part

Synthesis part
Parametrization of Speech

**Speech signal** → **High-pass filter** → **Windowing**

**Frame**

- **Glottal inverse filtering**
- **Estimated glottal flow**
- **F0 extraction**
- **LPC: Spectrum for unvoiced excitation** → **Convert to LSF**
- **LPC: Spectrum for voiced excitation** → **Convert to LSF**
- **LPC: Voice source spectr.** → **Convert to LSF**
- **Gain**

**Evaluation of spectral energy** → **Features**
Parametrisation of Speech

25-ms rectangular window
Parametrization of Speech

1. High-pass filtering
2. Windowing
3. Glottal inverse filtering
4. LPC: Spectrum for unvoiced excitation
   - Convert to LSF
5. LPC: Spectrum for voiced excitation
   - Convert to LSF
6. Estimated glottal flow
7. F0 extraction
8. LPC: Voice source spectrum
   - Convert to LSF
9. Evaluation of spectral energy
   - Gain
10. Features
Parametrization of Speech

1. Speech signal
2. High-pass filter
3. Windowing
4. Frame
   - Estimated glottal flow
   - Glottal inverse filtering
   - LPC: Spectrum for unvoiced excitation
     - Convert to LSF
   - LPC: Spectrum for voiced excitation
     - Convert to LSF
   - F0 extraction
   - LPC: Voice source spectr.
     - Convert to LSF
   - Evaluation of spectral energy
     - Gain
5. Features
Parametrization of Speech

1. Speech signal
2. High-pass filter
3. Windowing
4. Frame
5. Glottal inverse filtering
6. Estimated glottal flow
7. LPC: Voice source spectrum
8. LPC: Spectrum for voiced excitation
9. Convert to LSF
10. LPC: Spectrum for unvoiced excitation
11. Convert to LSF
12. F0 extraction
13. Gain
14. Evaluation of spectral energy
15. Features
Parametrization of Speech

- Speech signal → High-pass filter → Windowing → Frame
  - Glottal inverse filtering → Estimated glottal flow
    - LPC: Spectrum for voiced excitation
      - Convert to LSF
    - LPC: Spectrum for unvoiced excitation
      - Convert to LSF
  - F0 extraction → LPC: Voice source spectr.
    - Convert to LSF
  - Evaluation of spectral energy
    - Gain

Features
# Extracted Speech Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Parameters per frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental frequency</td>
<td>1</td>
</tr>
<tr>
<td>Energy</td>
<td>1</td>
</tr>
<tr>
<td>Spectral energy</td>
<td>5</td>
</tr>
<tr>
<td>Voice source spectrum</td>
<td>10</td>
</tr>
<tr>
<td>Voiced spectrum</td>
<td>20</td>
</tr>
<tr>
<td>Unvoiced spectrum</td>
<td>20</td>
</tr>
</tbody>
</table>
HMM Framework

1. SPEECH DATABASE
   - Speech signal
   - Label
2. Parametrization
3. Training of HMM
4. Context dependent HMMs
   - Parameter generation from HMM
   - Synthesis
5. TEXT
   - Text analysis
   - Label
6. Synthesized speech

Training part
Synthesis part
Synthesis from Parameters

Synthesis

Training part

Synthesis part

Speech signal

Label

Parameter generation from HMM

Parameter generation from HMM

Context dependent HMMs

Training of HMM

Parametrization

Speech signal

Label

TEXT

Text analysis

SPEECH DATABASE

SPEECH DATABASE

Synthesized speech
Synthesis from Parameters

- Random noise
- F0
- Spectral energy
  - Scale
    - Voiced excitation
      - Voiced LSFs
        - Vocal tract filter
          - Lip radiation
            - Gain
              - Match gain
                - Synthesized speech signal
  - Unvoiced excitation
    - Unvoiced LSFs
      - Unvoiced spectrum filter

- Library pulse
- F0
- Gain
  - LPC analysis
    - Pulse train
      - Interpolate and scale
        - Pulse train spectrum
          - Spectral match filter
            - Source spectrum
Synthesis from Parameters

- Random noise
- F0
- Spectral energy

- Library pulse
- F0
- Gain

- LPC analysis

- Pulse train

- Spectral match filter
- Source spectrum

- Scale

- Voiced excitation
  - Voiced LSFs
  - Vocal tract filter
  - Lip radiation
  - Gain

- Unvoiced excitation
  - Unvoiced LSFs
  - Unvoiced spectrum filter

- Match gain

- Synthesized speech signal
Synthesis from Parameters

Random noise
F0
Spectral energy

Library pulse
F0
Gain

LPC analysis
Pulse train
Pulse train spectr.
Source spectrum

Interpolate and scale

Pulse train

Spectral match filter

Scale

Voiced excitation
Voiced LSFs
Vocal tract filter
Lip radiation
Gain
Match gain

Unvoiced excitation
Unvoiced LSFs
Unvoiced spectrum filter

Gain

Synthesized speech signal
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

- Random noise
- F0
- Spectral energy

- Library pulse
- F0
- Gain

- LPC analysis
- Pulse train spectr.
- Source spectrum

- Interpolate and scale
- Pulse train
- Spectral match filter

- Voiced excitation
- Voiced LSFs
- Vocal tract filter
- Lip radiation
- Gain

- Unvoiced excitation
- Unvoiced LSFs
- Unvoiced spectrum filter
- Match gain

- Synthesized speech signal
Synthesis from Parameters

- Random noise
- F0
- Spectral energy

- Library pulse
- F0
- Gain

- LPC analysis
- Pulse train spectrum
- Source spectrum

Scale

Voiced excitation
- Voiced LSFs
- Vocal tract filter
- Lip radiation
- Gain
- Match gain

Unvoiced excitation
- Unvoiced LSFs
- Unvoiced spectrum filter

Speech signal
Speech from Parameters

Voiced source

Unvoiced source

Periodic pulse train

Gain

Filter coefficients

Filter

Speech

White noise
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)

- User interface

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Much Better</td>
</tr>
<tr>
<td>2</td>
<td>Better</td>
</tr>
<tr>
<td>1</td>
<td>Slightly Better</td>
</tr>
<tr>
<td>0</td>
<td>About the Same</td>
</tr>
<tr>
<td>-1</td>
<td>Slightly Worse</td>
</tr>
<tr>
<td>-2</td>
<td>Worse</td>
</tr>
<tr>
<td>-3</td>
<td>Much Worse</td>
</tr>
</tbody>
</table>
Results

Order of Preference

Score

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

![Bar chart showing results for 'proposed', 'no pref.', and 'baseline' categories. The 'proposed' category has a significantly higher percentage than the other two categories.]

- **Proposed**: High percentage
- **No pref.**: Lower percentage
- **Baseline**: Lower percentage compared to proposed, but higher than no pref.
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


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