Flexible speech synthesis based on hidden Markov models

Keiichi Tokuda
Nagoya Institute of Technology

APSIPA ASC 2013, Kaohsiung
November 1, 2013
Towards human-like talking machines

For realizing natural human-computer interaction, speech synthesis systems are required to have an ability to generate speech with:

- arbitrary speaker’s voice
- speaking styles
  (e.g., reading, conversational)
- emotional expressions
  (e.g., happy, angry, sad)
- emphasis
- and so on
History

Rule-based, *formant synthesis* (~’90s)
– Hand-crafting each phonetic units by rules

Corpus-based, *concatenative synthesis* (~’90s~)
– Concatenate speech units (waveform) from a database
  • Single inventory: diphone synthesis
  • Multiple inventory: unit selection synthesis

Corpus-based, *statistical parametric synthesis*
– Source-filter model + statistical acoustic model
  • Hidden Markov model: HMM-based synthesis
General unit-selection synthesis scheme

Whole speech unit database

Selected speech units

---

Target cost

Concatenation cost
Overview of this talk

1. Introduction and background
2. Basic techniques in the system
3. Examples demonstrating its flexibility
4. Discussion and conclusion
HMM-based speech synthesis system

Training part

1. Text analysis
2. Speech signal
3. Excitation parameter extraction
   - Excitation parameters
4. Spectral parameter extraction
   - Spectral parameters
5. Training HMMs

Context-dependent HMMs & state duration models

Synthesis part

1. Text analysis
2. Parameter generation from HMMs
   - Excitation parameters
3. Synthesis Filter
   - Spectral parameters
4. SYNTHESIZED SPEECH
HMM-based speech synthesis system

**Training part**

- **Speech signal**
  - **Text analysis**
    - **SPEECH DATABASE**
    - **Excitation Parameter extraction**
      - **Excitation parameters**
    - **Spectral Parameter Extraction**
      - **Spectral parameters**
  - **Training HMMs**
  - **Labels**
  - **Parameter generation from HMMs**
  - **Excitation parameters**
  - **Spectral parameters**

**Synthesis part**

- **TEXT**
  - **Text analysis**
  - **Labels**
  - **Parameter generation from HMMs**
  - **Excitation generation**
  - **Synthesis Filter**
  - **SYNTHESIZED SPEECH**

**Context-dependent HMMs & state duration models**
Human speech production

Modulation of carrier wave by speech information

- Frequency transfer characteristics
- Magnitude start--end
- Fundamental frequency

Speech

Sound source
- Voiced: pulse
- Unvoiced: noise

air flow
Source-filter model

Source excitation part
- Pulse train
- White noise

Excitation
- $e(n)$

Vocal tract resonance part
- Linear time-invariant system $H(z)$
- Speech

$x(n) = h(n) * e(n)$
ML estimation of spectral parameter

Mel-cepstral representation of speech spectra

\[ H(z) = \exp \sum_{m=0}^{M} c(m) z^{-m} \]

\[ \tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} = e^{-j\tilde{\omega}} \]

ML-estimation of mel-cepstrum

\[ c = \arg \max_{c} p(x \mid c) \]

\( x \) : short segment of speech waveform (Gaussian process)

\( c \) : mel-cepstrum
Synthesis Filter

\[ H(z) = \exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m} \]

\[ F(z) = \sum_{m=0}^{M} c(m) \tilde{z}^{-m} \]

\[ b(m) = c(m) - \alpha b(m + 1) \]
Structure of MLSA filter

\[ H(z) = \exp F(z) \cong \frac{1 + \sum_{l=1}^{L} A_{L,l} \{F(z)\}'}{1 + \sum_{l=1}^{L} A_{L,l} \{-F(z)\}'} \]
Waveform reconstruction

These speech parameters have to be estimated from the text to be synthesized → use of HMM
HMM-based speech synthesis system

SPEECH DATABASE

Text analysis

Speech signal

Excitation Parameter extraction

Excitation parameters

Spectral Parameter Extraction

Spectral parameters

Training HMMs

Labels

Text analysis

Parameter generation from HMMs

Labels

Excitation parameters

Spectral parameters

SYNTHESIZED SPEECH

Excitation generation

Synthesis Filter

CONTEXT-DEPENDENT HMMS & STATE DURATION MODELS
Hidden Markov model (HMM)

\[ a_{ij} : \text{state transition probability} \]

\[ b_q(o_t) : \text{output probability} \]

\[ o \]

\[ q \]

Observation sequence

State sequence
Structure of state output (observation) vector

- **Spectrum part**
  - $c_t$
  - $\Delta c_t$
  - $\Delta^2 c_t$
  - Spectral parameters (e.g., mel-cepstrum, LSPs)

- **Excitation part**
  - $p_t$
  - $\Delta p_t$
  - $\Delta^2 p_t$
  - log F0 with V/UV

- $o_t$
Unable to model by continuous or discrete distributions $\Rightarrow$ Multi-space distribution HMM (MSD-HMM)
Multi-space probability distribution HMM
(MSD-HMM)
MSD-HMM for F0 modeling

HMM for F0

\[ \Omega_1 = R^1 \]

Voiced / Unvoiced weights

Voiced

Unvoiced

\[ \Omega_2 = R^0 \]
Contextual factors

Phoneme
- \{preceding, succeeding\} two phonemes
- current phoneme

Syllable
- # of phonemes in \{preceding, current, succeeding\} syllable
- \{accent, stress\} of \{preceding, current, succeeding\} syllable
- Position of current syllable in current word
- # of \{preceding, succeeding\} \{accented, stressed\} syllable in current phrase
- # of syllables \{from previous, to next\} \{accented, stressed\} syllable
- Vowel within current syllable

Word
- Part of speech of \{preceding, current, succeeding\} word
- # of syllables in \{preceding, current, succeeding\} word
- Position of current word in current phrase
- # of \{preceding, succeeding\} content words in current phrase
- # of words \{from previous, to next\} content word

Phrase
- # of syllables in \{preceding, current, succeeding\} phrase

.....

Huge # of combinations ⇒ Difficult to have all possible models
Decision tree-based state clustering [Odell; ’95]

Sharing the parameter of HMMs in the same leaf node
Stream-dependent tree-based clustering (2)

State duration model

Three dimensional Gaussian

HMM

Decision trees for mel-cepstrum

Decision tree for state dur. models

Decision trees for F0
HMM-based speech synthesis system

**Training part**

- **Speech signal**
  - **Speech database**
    - **Text analysis**
      - **Speech signal**
        - **Excitation parameter extraction**
          - **Excitation parameters**
        - **Spectral parameter extraction**
          - **Spectral parameters**
  - **Labels**
  - **Training HMMs**
    - **Parameter generation from HMMs**
      - **Excitation parameters**
      - **Spectral parameters**

**Synthesis part**

- **Text analysis**
  - **Labels**
  - **Excitation generation**
    - **Excitation**
    - **Synthesis filter**
      - **Synthesized speech**
Speech parameter generation algorithm [Tokuda; ’00]

For given sentence HMM, determine a speech parameter vector sequence \( \hat{o} = \left[ \hat{o}_1^\top, \hat{o}_2^\top, \ldots, \hat{o}_T^\top \right]^\top \) which maximizes

\[
P(o | \hat{l}, \hat{\lambda}) = \sum_q P(o | q, \hat{\lambda})P(q | \hat{l}, \hat{\lambda}) 
\approx \max_q P(o | q, \hat{\lambda})P(q | \hat{l}, \hat{\lambda})
\]

\[
\hat{q} = \arg \max_q P(q | \hat{l}, \hat{\lambda})
\]

\[
\hat{o} = \arg \max_o P(o | \hat{q}, \hat{\lambda})
\]
Determination of state sequence (1/3)

- Observation sequence
- State sequence
- State duration

\[ a_{ij} : \text{State transition probability} \]
\[ b_q(o_t) : \text{Output probability} \]

Determine state sequence via determining state durations
Determination of state sequence

\[ P(q \mid \hat{l}, \hat{\lambda}) = \prod_{i=1}^{K} p_i(d_i) \]

- \( p_i(\cdot) \): state-duration distribution of \( i \)-th state
- \( d_i \): state duration of \( i \)-th state
- \( K \): \# of states in a sentence HMM for \( \hat{l} \)

Gaussian

\[ p_i(d_i) = N(d_i \mid m_i, \sigma_i^2) \Rightarrow \hat{d}_i = m_i \]
Speech parameter generation algorithm

For given HMM $\lambda$, determine a speech parameter vector sequence $o = [o_1^T, o_2^T, \ldots, o_T^T]^T$ which maximizes

$$P(o | \hat{l}, \hat{\lambda}) = \sum_q P(o | q, \hat{\lambda})P(q | \hat{l}, \hat{\lambda})$$

$$\approx \max_q P(o | q, \hat{\lambda})P(q | \hat{l}, \hat{\lambda})$$

$$\hat{q} = \arg \max_q P(q | \hat{l}, \hat{\lambda})$$

$$\hat{o} = \arg \max_o P(o | \hat{q}, \hat{\lambda})$$
Without dynamic feature

becomes a sequence of mean vectors
⇒ discontinuous outputs between states
Dynamic features

\[ \Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1}) \]

\[ \Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1} \]
Integration of dynamic features

Relationship between speech parameter vectors & static feature vectors

\[ \mathbf{o}_t = \begin{bmatrix} c_t^T, \Delta c_t^T, \Delta^2 c_t^T \end{bmatrix}^T \]
Solution for the problem (1/2)

By setting

$$\frac{\partial \log P(O | \hat{q}, \lambda)}{\partial c} = O,$$

we obtain

$$W^T \Sigma_\hat{q}^{-1} Wc = W^T \Sigma_\hat{q}^{-1} \mu_\hat{q},$$

where

$$c = [c_1^T, c_2^T, \ldots, c_T^T]^T$$

$$\mu_\hat{q} = [\mu_{\hat{q}_1}^T, \mu_{\hat{q}_2}^T, \ldots, \mu_{\hat{q}_T}^T]^T$$

$$\Sigma_\hat{q} = [\Sigma_{\hat{q}_1}^T, \Sigma_{\hat{q}_2}^T, \ldots, \Sigma_{\hat{q}_T}^T]^T$$
Solution for the problem (2/2)
Generated speech parameter trajectory

/sil/ /a/ /i/ /sil/
Generated spectra

Spectra changing smoothly at state boundaries
Generated F0

- **natural speech**
- **without dynamic features**
- **with dynamic features**

Frequency vs. time (s) graph showing different patterns for natural speech and speech with and without dynamic features.
Effect of dynamic features

<table>
<thead>
<tr>
<th>log F0</th>
<th>Mel-cepstrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>static+</td>
<td>static+</td>
</tr>
<tr>
<td>$\Delta + \Delta^2$</td>
<td>$\Delta + \Delta^2$</td>
</tr>
<tr>
<td>static</td>
<td>static</td>
</tr>
</tbody>
</table>
HMM-based speech synthesis system

Training part

SPEECH DATABASE

Text analysis

Speech signal

Excitation Parameter extraction

Speech signal

Spectral Parameter Extraction

Training HMMs

Labels

Excitation parameters

Spectral parameters

Synthesis part

TEXT

Text analysis

Labels

Excitation parameters

Spectral parameters

Parameter generation from HMMs

Context-dependent HMMs & state duration models

Excitation generation

Synthesis Filter

SYNTHESIZED SPEECH
Overview of this talk

1. Introduction and background
2. Basic techniques in the system
3. Examples demonstrating its flexibility
4. Discussion and conclusion
<table>
<thead>
<tr>
<th>text</th>
<th>neutral</th>
<th>angry</th>
</tr>
</thead>
<tbody>
<tr>
<td>「授業中に携帯いじってんじゃねえよ！電源切っとけ！」</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Don’t touch your cell phone during a class! Turn off it!”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>「ミーティングには毎週参加しなさい！」</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“You must attend the weekly meeting!”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

trained with 200 utterances
Speaker adaptation (mimicking voices)

MLLR-based adaptation

- Speaker-independent system
- Adaptation data speaker A
- Adapted model

- w/o adaptation (initial model)
- Adapted with 4 utterances
- Adapted with 50 utterances
- Speaker A’s speaker-dependent system
Speaker interpolation (mixing voices)

Linear combination of two speaker-dependent models

Model A

Interpolated model

Model B

A: 1.00 0.75 0.50 0.25 0.00

B: 0.00 0.25 0.50 0.75 1.00
Voice morphing

Two voices:

A ⇔ ⇔ ⇔ ⇔ ⇔ ⇔ ⇔ ⇔ B

A ⇔ ⇔ ⇔ ⇔ ⇔ ⇔ ⇔ ⇔ B

Four voices:

Male1 Female1

Female2 Male2
Interpolation of speaking styles

Base model A

Interpolation

Base model B

extrapolation

Neutral

High Tension
Eigenvoice (creating voices) [Shichiri; ’02]

Mean Calculation

Mean vector $\bar{\mu}$

Speaker 1
Supervector 1

Speaker 2
Supervector 2

Speaker S
Supervector S

PCA

$e(1)$ $\cdots$ $e(k)$ $\cdots$ $e(K)$

Click here for a demo
Multilingual speech synthesis

- Japanese
- American English
- Chinese (Mandarin) (by ATR)
- Brazilian Portuguese (by Nitech, and UFRJ)
- European Portuguese (by Nitech, Univ of Porto, and UFRJ)
- Slovenian (by Bostjan Vesnicker, University of Ljubljana, Slovenia)
- Swedish (by Anders Lundgren, KTH, Sweden)
- German (by University of Bonn, and Nitech)
- Korean (by Sang-Jin Kim, ETRI, Korea)
- Finish (by TKK, Finland)
- Baby English (by Univ of Edinburgh, UK)
- Polish, Slovak, Arabic, Farsi, Croatian, Polyglot, etc.
HMM-based singing synthesis

1) HMM training

Any score with lyric

2) Singing generation

synthesized singing
Flexibility in singing synthesis

• Sing a popular song

• Sing by a famous person’s voice
  (from a TV program NHK “Science zero”)

• Rap singing
MMDAgent

A toolkit for building voice interaction systems

- Fully open-source toolkit with open interfaces
- Tightly integrated speech recognition/synthesis engines
- 3-D scene rendering fully compatible with CG tools
- HMM-based flexible and expressive speech synthesis (neutral, angry, bashful, happy, sad)
Summary

Statistical approach to speech synthesis

• Whole speech synthesis process is described in a unified statistical framework
• It can provide flexibility: various voices, emotional expressions, speaking styles, etc.

Future work
• Still we have many problems should be solved:
  • Direct modeling of speech waveform
  • Importance of the database
Sagisaka;'92 - "ATR nu-TALK speech synthesis system," ICSLP, '92.
Black;'96 - "Automatically clustering similar units...," Euro speech, '97.
Imai;'88 - "Unbiased estimator of log spectrum and its application to speech signal...," EURASIP, '88.
Tokuda;'94 - "Mel-generalized cepstral analysis -- A unified approach to speech spectral...," ICSLP, '94.
Imai;'83 - "Cepstral analysis synthesis on the mel frequency scale," ICASSP, '83.
Itakura;'75 - "Line spectrum representation of linear predictive coefficients of speech...," J. ASA (57), '75.
Shinoda;'00 - "MDL-based context-dependent subword modeling...," Journal of ASJ(E) 21(2), '00.
Yoshimura;'98 - "Duration modeling for HMM-based speech synthesis," ICSLP, '98.
Tokuda;'00 - "Speech parameter generation algorithms for HMM-based speech synthesis," ICASSP, '00.
Hunt;'96 - "Unit selection in a concatenative speech synthesis system using...," ICASSP, '96.
Donovan;'95 - "Improvements in an HMM-based speech synthesiser," Eurospeech, '95.
Kawai;'04 - "XIMERA: A new TTS from ATR based on corpus-based technologies," ISCA SSW5, '04.
Hirai;'04 - "Using 5 ms segments in concatenative speech synthesis," Proc. ISCA SSW5, '04.
References (2/4)

Rouibia;'05 - "Unit selection for speech synthesis based on a new acoustic target cost," Interspeech, '05.
Mizutani;'02 - "Concatenative speech synthesis based on HMM," ASJ autumn meeting, '02.
Ling;'07 - "The USTC and iFlytek speech synthesis systems...," Blizzard Challenge workshop, 07.
Ling;'08 - "Minimum unit selection error training for HMM-based unit selection...," ICASSP, 08.
Plumpe;'98 - "HMM-based smoothing for concatenative speech synthesis," ICSLP, '98.
Wouters;'00 - "Unit fusion for concatenative speech synthesis," ICSLP, '00.
Okubo;'06 - "Hybrid voice conversion of unit selection and generation...," IEICE Trans. E89-D(11), '06.
Aylett;'08 - "Combining statistical parametric speech synthesis and unit selection..." LangTech, '08.
Pollet;'08 - "Synthesis by generation and concatenation of multiform segments," Interspeech, '08.
Tachibana;'05 - "Speech synthesis with various emotional expressions...," IEICE Trans. E88-D(11), '05.
Kuhn;'00 - "Rapid speaker adaptation in eigenvoice space," IEEE Trans. SAP 8(6), '00.
Shichiri;'02 - "Eigenvoices for HMM-based speech synthesis," ICSLP, '02.
Fujinaga;'01 - "Multiple-regression hidden Markov model," ICASSP, '01.
Nose;'07 - "A style control technique for HMM-based expressive speech...," IEICE Trans. E90-D(9), '07.
Yoshimura;'01 - "Mixed excitation for HMM-based speech synthesis," Eurospeech, '01.
Zen;'07 - "Details of the Nitech HMM-based speech synthesis system...", IEICE Trans. E90-D(1), '07.
Abdl-Hamid;'06 - "Improving Arabic HMM-based speech synthesis quality," Interspeech, '06.
References (3/4)

Hemptinne;'06 - "Integration of the harmonic plus noise model into the...," Master thesis, IDIAP, '06.
Banos;'08 - "Flexible harmonic/stochastic modeling...," V. Jornadas en Tecnologias del Habla, '08.
Cabral;'07 - "Towards an improved modeling of the glottal source in...," ISCA SSW6, '07.
Maia;'07 - "An excitation model for HMM-based speech synthesis based on ...," ISCA SSW6, '07.
Ratio;'08 - "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," Interspeech, '08.
Drugman;'09 - "Using a pitch-synchronous residual codebook for hybrid HMM/frame...," ICASSP, '09.
Dines;'01 - "Trainable speech synthesis with trended hidden Markov models," ICASSP, '01.
Sun;'09 - "Polynomial segment model based statistical parametric speech synthesis...," ICASSP, '09.
Bulyko;'02 - "Robust splicing costs and efficient search with BMM models for...," ICASSP, '02.
Shannon;'09 - "Autoregressive HMMs for speech synthesis," Interspeech, '09.
Zen;'06 - "Reformulating the HMM as a trajectory model...", Computer Speech & Language, 21(1), '06.
Wu;'06 - "Minimum generation error training for HMM-based speech synthesis," ICASSP, '06.
Hashimoto;'09 - "A Bayesian approach to HMM-based speech synthesis," ICASSP, '09.
Wu;'08 - "Minimum generation error training with log spectral distortion for...," Interspeech, '08.
Toda;'08 - "Statistical approach to vocal tract transfer function estimation based on...," ICASSP, '08.
Oura;'08 - "Simultaneous acoustic, prosodic, and phrasing model training for TTS...," ISCSLP, '08.
Ferguson;'80 - "Variable duration models...," Symposium on the application of HMM to text speech, '80.
Levinson;'86 - "Continuously variable duration hidden...," Computer Speech & Language, 1(1), '86.
Beal;'03 - "Variational algorithms for approximate Bayesian inference," PhD thesis, Univ. of London, '03.
Masuko;'03 - "A study on conditional parameter generation from HMM...," Autumn meeting of ASJ, '03.
Yu;'07 - "A novel HMM-based TTS system using both continuous HMMs and discrete...," ICASSP, '07.
Qian;'08 - "Generating natural F0 trajectory with additive trees," Interspeech, '08.
Latorre;'08 - "Multilevel parametric-base F0 model for speech synthesis," Interspeech, '08.
Tiomkin;'08 - "Statistical text-to-speech synthesis with improved dynamics," Interspeech, '08.
Wu;'08 - "Minimum generation error criterion considering global/local variance...," ICASSP, '08.
Toda;'09 - "Trajectory training considering global variance for HMM-based speech...," ICASSP, '09.
Saino;'06 - "An HMM-based singing voice synthesis system," Interspeech, '06.
Tsuzuki;'04 - "Constructing emotional speech synthesizers with limited speech...," Interspeech, '04.
Sako;'00 - "HMM-based text-to-audio-visual speech synthesis," ICSLP, '00.
Tamura;'98 - "Text-to-audio-visual speech synthesis based on parameter generation...," ICASSP, '98.
Haoka;'02 - "HMM-based synthesis of hand-gesture animation," IEICE Technical report,102(517), '02.
Niwase;'05 - "Human walking motion synthesis with desired pace and...," IEICE Trans. E88-D(11), '05.
Ma;'07 - "A MSD-HMM approach to pen trajectory modeling for online handwriting...," ICDAR, '07.
Morioka;'04 - "Miniaturization of HMM-based speech synthesis," Autumn meeting of ASJ, '04.
Kim;'06 - "HMM-Based Korean speech synthesis system for...," IEEE Trans. Consumer Elec., 52(4), '06.
Klatt;'82 - "The Klatt-Talk text-to-speech system," ICASSP, '82.
Keiichi Tokuda would like to thank HTS working group members, including Heiga Zen, Keiichiro Oura, Junichi Yamaguchi, Tomoki Toda, Yoshihiko Nankaku, Kei Hahimoto, and Sayaka Shiota for their help.
Some samples (icon is 🎧) are released under following license.

This voice is free for use for any purpose (commercial or otherwise) subject to the pretty light restrictions detailed below.

# Carnegie Mellon University  #
# Copyright (c) 2003  #
# All Rights Reserved.  #
# Permission to use, copy, modify, and licence this software and its documentation for any purpose, is hereby granted without fee, subject to the following conditions:  #
# 1. The code must retain the above copyright notice, this list of conditions and the following disclaimer.  #
# 2. Any modifications must be clearly marked as such.  #
# 3. Original authors’ names are not deleted.  #
# THE AUTHORS OF THIS WORK DISCLAIM ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.  #
# See http://www.festvox.org/cmu_arctic/ for more details  #