The HMM-based Speech Synthesis System
Version 2.0

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako,
Takashi Masuko, Alan W. Black, Keiichi Tokuda

1 Nitech, 2 Titech, 3 Edinburgh Univ., 4 Tokyo Univ., 5 CMU
Overview

- HMM-based speech synthesis [Yoshimura et al.;'00]
  - Statistical parametric speech synthesis
  - Spectrum, excitation, & durations are modeled by HMMs
  - Generates speech waveforms from the HMMs themselves

- HTS [Zen et al.;'02]
  - Open-source software for HMM-based speech synthesis
  - Provides a research & development platform
  - New version (ver. 2.0) was released in Dec. '06

Its new features & future release plans are shown
Outline

- Overview
- HMM-based speech synthesis
  - Corpus-based speech synthesis
  - Overview of HMM-based speech synthesis
- HTS
  - HTS vers. 1.0 ~ 1.1
  - HTS ver. 2.0
- Other applications
- Future release plans
- Conclusions
Corpus-based speech synthesis

- **Unit selection synthesis**
  - Selects appropriate units from a speech database
  - High quality (but sometimes discontinuous)
  - Difficult to change its voice characteristics

- **HMM-based synthesis**
  - Generates speech parameters from statistical models
  - Vocoded (but smooth & stable)
  - Easy to change its voice characteristics

In last years, HMM-based approach is getting popular
Overview of HMM-based speech synthesis

**Training part**

- Speech signal
  - Excitation parameter extraction
  - Spectral parameter extraction
  - Training of HMMs
    - Context-dependent HMMs & state duration models
    - Transformed acoustic models

**Synthesis part**

- Text
  - Text analysis
    - Labels
      - Excitation parameters
        - Synthesis filter
          - SYNTHESIZED SPEECH
    - Spectral parameters
      - Transformed synthesized speech
- SPEECH DATABASE
  - Speech signal
    - Excitation parameters
      - Excitation generation
        - Transform
          - Spectral parameters
            - Parameter generation from HMM
Outline

- Overview
- HMM-based speech synthesis
  - Corpus-based speech synthesis
  - Overview of HMM-based speech synthesis
- HTS
  - HTS vers. 1.0 ~ 1.1
  - HTS ver. 2.0
- Other applications
- Future release plans
- Conclusions
HMM-based speech synthesis system (HTS)

HTS: A toolkit for HMM-based speech synthesis

- HTS website: http://hts.sp.nitech.ac.jp/
- Provides a research platform for HMM-based synthesis
- Released as a patch code for HTK
- Open source, BSD-style license
- Over 4,000 downloads
- Used in various organizations
  e.g.) MSRA, iFlyTek, ATR, Cambridge, Edinburgh, CMU, KTH, IDIAP, DFKI, Bonn, & others...
Features in the past releases

- Dec. '02: ver. 1.0
  * F0 pattern modeling by MSD-HMMs
  * Tree-based clustering based on the MDL criterion
  * Stream-dependent clustering
  * State-duration modeling by Gaussian distributions
  * Speech parameter generation algorithm

- May '03: ver. 1.1
  * Small run-time synthesis engine (hts_engine)
  * HTS voices for Festival

⇒ Speaker-dependent HTS voices can be constructed
HTS ver. 2.0

New features & modifications

- Based on HTK ver. 3.4
- Supports speaker adaptation
  * Adaptation of multi-stream MSD-HMMs
  * Adaptive training
- EM-based speech parameter generation algorithm
- Supports various model structures (e.g., ergodic, full cov.)
- Minor improvements
- Bug fixes
- Adaptation & adaptive training demo
Adaptation (1)

- Adjust general acoustic models to a specific class
  - Widely used in speech recognition
  - Standard approaches: MAP, MLLR & eigenvoice

- Adaptation in HMM-based speech synthesis
  - Good example to show the flexibility of HMM-based syn.
  - Enables us to synthesize various speech using small data
    * Speaker adaptation [Masuko;'97, Tamura;'01, Yamagishi;'06]
    * Speaking style adaptation [Tachibana;'06]
  - Not released in the previous versions
Adaptation (2)

Adaptations in HTS ver. 2.0

- Supports multi-stream MSD-HMMs (MAP & MLLR)
  ⇒ Adapt spectrum & excitation simultaneously [Tamura;01]

- Can use decision trees to define regression classes
  ⇒ Supra-segmental features can be adapted [Yamagishi;'03]

- Adaptation of durations has not been supported
EM-based speech parameter generation algorithm

Speech parameter generation algorithms [Tokuda;'00]

\[ \hat{o} = \arg \max_{o} P(o | \lambda)|_{o=Wc} \quad \text{case 1} \]
\[ \approx \arg \max_{o} \max_{s,m} P(o, s, m | \lambda)|_{o=Wc} \quad \text{case 2} \]
\[ \approx \arg \max_{o} P(o | \hat{s}, \hat{m}, \lambda)|_{o=Wc} \quad \text{case 3} \]

\( o \): Observation vector sequence (including delta & delta-delta)
\( c \): Speech parameter sequence (static feature only)
\( s, m \): State & mixture sequences (\( \hat{s}, \hat{m} \): Best ones)

- 1: EM algorithm  2: recursive search  3: linear equations
- Only case 3 was released in the previous versions
- HTS ver. 2.0 additionally provides case 1 algorithm
Outline

- Overview
- HMM-based speech synthesis
  - Corpus-based speech synthesis
  - Overview of HMM-based speech synthesis
- HTS
  - HTS vers. 1.0 ~ 1.1
  - HTS ver. 2.0
- Other applications
- Future release plans
- Conclusions
Integration of HTS in other speech synthesis systems

- **Licence issues**
  - Once patch is applied, users must obey the HTK license
  - hts_engine: small synthesis engine independent of HTK
    ⇒ Can be used in other systems

- **Other synthesis systems using hts_engine**
  - ATR XIMERA [Kawai;'04]
    ⇒ Uses hts_engine as a target prediction module
  - Festival [Black;'06]
    ⇒ Includes hts_engine as a waveform synthesis module
  - DFKI OpenMARY [Schröder;'06]
    ⇒ Porting hts_engine into Java & integrating to OpenMARY
Other applications

Applications of HTS

- Synthesis-related topics
  * Human motion [Mori;'05, Niwase;'05, Hofer;'07]
  * Audio-visual (lip & speech) [Tamura;'99, Sako;'00]
  * Face animation [Govokhinai;'06]
  * Very low-bitrate speech coder [Hoshiya;'03]
  * Acoustic-articulatory inversion mapping [Richmond;'06]

- Recognition-related topics
  * Online hand-writing [Ma;'07]
  * Automatic evaluation of ASR systems [Terashima;'03]
  * Prosodic events (accent, tone) [Emoto;'03, Wang;'06]
Other applications (1)

Human motion synthesis
- Record human movements by a motion capture system
- HMMs are estimated in the same way used in speech syn.
- Synthesize motions

Finger alphabets  Walking motion  Head motion
Other applications (2)

Audio-visual synthesis

- Record both lip movies & speech
- Extract lip & speech features
- Estimate HMMs & synthesize

Model-based (coordinates)  Image-based (PCA)
Other applications (3)

Online handwriting recognition

- Observation: continuos (F0 value, pen down) or discrete (unvoiced, pen up)

  ⇒ Can be modeled properly by MSD-HMMs [Tokuda;'02]

- Properly handle imaginary strokes

  ⇒ Better discrimination between similar characters [Ma;'07]
Outline

• Overview

• HMM-based speech synthesis
  - Corpus-based speech synthesis
  - Overview of HMM-based speech synthesis

• HTS
  - HTS vers. 1.0 ～1.1
  - HTS ver. 2.0

• Other applications

• Future release plans

• Conclusions
Internal versions

Additional features provided in the internal versions

- Hidden semi-Markov models (HSMMs) [Zen;'07]
- Parameter generation considering global variance [Toda;'07]
- Variational Bayes [Nankaku;'03]
- Trajectory HMMs [Zen;'07]
- Interpolation [Yoshimura;'03]
- Shared tree construction [Yamagishi;'01]
- Advanced adaptation (CSMAPLR) [Nakano;'06]
- Adaptation & adaptive training of state duration models
- Structured precision matrix models (e.g., EMLLT, SPAM)
- Multiple-regression HMMs [Nose;'06], eigenvoice [Shichiri;'03]
Future release plans

- **Aug. '07: ver. 2.0.1**
  - Bug fixes & minor changes
  - C/C++ API version of hts_engine
  - Interpolation
  - Supports band structure in MLLR transforms
  - Demo using MGC-LSPs in addition to mel-cepstrum

- **Mar. '08: ver. 2.1**
  - HSMM training & adaptation
  - Parameter generation considering global variance
  - Advanced adaptation (CSMAPLR)
Outline

• Overview
• HMM-based speech synthesis
  - Corpus-based speech synthesis
  - Overview of HMM-based speech synthesis
• HTS
  - HTS vers. 1.0 ~ 1.1
  - HTS ver. 2.0
• Other applications
• Future release plans
• Conclusions
Conclusions

- HMM-based speech synthesis system (HTS) ver. 2.0
  - Based on HTK ver. 3.4
  - Supports adaptation & adaptive training
  - EM-based speech parameter generation algorithm
  - Bug fixes, minor improvements

- Future release plans
  - Aug. '07: ver. 2.0.1 ⇒ Bug fixes, API ver. of hts_engine
  - Mar. '08: ver. 2.1 ⇒ HSMM, GV, & CSMA PLR

Please use it & give us your feedback!!