Speech Synthesis as A Statistical Machine Learning Problem

Keiichi Tokuda
Nagoya Institute of Technology

ASRU2011, Hawaii
December 14, 2011
Introduction

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

How we can formulate and understand the whole corpus-based speech synthesis process in a unified statistical framework?
Problem of speech synthesis

We have a speech database, i.e., a set of texts and corresponding speech waveforms. Given a text to be synthesized, what is the speech waveform corresponding to the text?

\[ W \] : texts
\[ X \] : speech waveforms
\[ w \] : text to be synthesized
\[ x \] : speech waveform
Statistical formulation of speech synthesis (1/8)

Bayesian framework for prediction

\[ \text{Draw } \tilde{x} \text{ from } p(x \mid w, X, W) \]

\( W \) : set of texts \quad \{ \text{database} \} \quad \{ \text{Given} \}

\( X \) : speech waveforms

\( w \) : text to be synthesized

\( x \) : speech waveform \quad \{ \text{unknown} \}

1. Estimate predictive distribution given variables
2. Draw sample from the distribution
1. Estimating predictive distribution is hard 😞
   → Introduce acoustic model parameters

\[
p(x \mid w, X, W) \downarrow \text{introduce acoustic model } \lambda
\]

\[
= \int p(x, \lambda \mid w, W, X) d\lambda = \int p(x \mid w, \lambda)p(\lambda \mid W, X) d\lambda
\]

\[
\lambda : \text{acoustic model (e.g. HMM )}
\]
2. Using speech waveform directly is difficult 😞

→ Introduce parametric its representation

\[
p(x \mid w, X, W) = \int p(x \mid w, \lambda)p(\lambda \mid X, W)d\lambda
\]

↓ introduce parametric representation of speech \( o \)

\[
= \int \int p(x \mid o)p(o \mid w, \lambda)p(\lambda \mid X, W)d\lambda do
\]

\( o \) : parametric representation of speech waveform \( x \)
(e.g., cepstrum, LPC, LSP, F0, aperiodicity)
3. Same texts can have multiple pronunciations, POS, etc. 😞
→ Introduce labels

\[
p(x \mid w, X, W) = \int \int p(x \mid o)p(o \mid w, \lambda)p(\lambda \mid X, W)d\lambda do
\]

↓ introduce labels derived from texts, \( l \) & \( L \)

\[
= \int \int \sum_{\forall l} p(x \mid o)p(o \mid l, \lambda)P(l \mid w)p(\lambda \mid X, W)d\lambda do
\]

\( l \) : labels derived from text \( w \)
(e.g. prons, POS, lexical stress, grammar, pause)
4. Difficult to perform integral & sum over auxiliary variables 🙁

→ Approximated by joint max

\[
p(x \mid w, X, W) = \int \int \sum_{\forall l} p(x \mid o)p(o \mid l, \lambda)P(l \mid w)p(\lambda \mid X, W) d\lambda d\sigma \\
\downarrow \text{approximate integral & sum by joint max}
\]

\[
\approx p(x \mid \hat{o})p(\hat{o} \mid \hat{l}, \hat{\lambda})P(\hat{l} \mid w)p(\hat{\lambda} \mid X, W)
\]

where

\[
\{\hat{o}, \hat{l}, \hat{\lambda}\} = \arg\max_{o, l, \lambda} p(x \mid o)p(o \mid l, \lambda)P(l \mid w)p(\lambda \mid X, W)
\]
5. Joint maximization is hard 😞
→ Approximated by step-by-step maximizations

\[
\left\{ \hat{o}, \hat{l}, \hat{\lambda} \right\} = \underset{o,l,\lambda}{\text{arg max}} \ p(x \mid o)p(o \mid l, \lambda)P(l \mid w)p(\lambda \mid X, W)
\]

↓ approx joint max by step-by-step max

\[
\hat{\lambda} = \underset{\lambda}{\text{arg max}} \ p(\lambda \mid X, W) \quad \leftrightarrow \text{training}
\]

\[
\hat{l} = \underset{l}{\text{arg max}} \ P(l \mid w) \quad \leftrightarrow \text{text analysis}
\]

\[
\hat{o} = \underset{o}{\text{arg max}} \ p(o \mid \hat{l}, \hat{\lambda}) \quad \leftrightarrow \text{speech parameter generation}
\]
6. Training also requires parametric form of wav & labels 😞
→ Introduce them & approx by step-by-step maximizations

\[
\hat{\lambda} = \arg \max_{\lambda} p(\lambda \mid X, W)
\]

\[
\downarrow
\]

\[
\hat{\mathcal{L}} = \arg \max_{L} P(L \mid W) \quad \leftarrow \text{labeling}
\]

\[
\hat{O} = \arg \max_{O} p(O \mid X) \quad \leftarrow \text{feature extraction}
\]

\[
\hat{\lambda} = \arg \max_{\lambda} p(\hat{O} \mid \hat{\mathcal{L}}, \lambda) p(\lambda) \quad \leftarrow \text{acoustic model training}
\]

\(O\): parametric representation of speech waveforms \(X\)

\(L\): labels derived from texts \(W\)
Statistical formulation of speech synthesis (8/8)

Draw $\tilde{x}$ from $p(x \mid w, X, W)$

<table>
<thead>
<tr>
<th>Expression</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{O} = \arg \max_O p(X \mid O)$</td>
<td>$\leftrightarrow$ feature extraction</td>
</tr>
<tr>
<td>$\hat{L} = \arg \max_L P(L \mid W)$</td>
<td>$\leftrightarrow$ labeling</td>
</tr>
<tr>
<td>$\hat{\lambda} = \arg \max_\lambda p(\hat{O} \mid \hat{L}, \lambda)p(\lambda)$</td>
<td>$\leftrightarrow$ acoustic model training</td>
</tr>
<tr>
<td>$\hat{l} = \arg \max_l P(l \mid w)$</td>
<td>$\leftrightarrow$ text analysis</td>
</tr>
<tr>
<td>$\hat{o} = \arg \max_o p(o \mid \hat{l}, \hat{\lambda})$</td>
<td>$\leftrightarrow$ speech parameter generation</td>
</tr>
<tr>
<td>$\tilde{x}$ from $p(x \mid \hat{o})$</td>
<td>$\leftrightarrow$ waveform reconstruction</td>
</tr>
</tbody>
</table>
Overview of this talk

1. Mathematical formulation
2. Implementation of individual components
3. Examples demonstrating its flexibility
4. Discussion and conclusion
HMM-based speech synthesis system

Training part

1. SPEECH DATABASE
2. Text analysis
3. Speech signal
4. Excitation Parameter extraction
5. Spectral Parameter Extraction
6. Excitation parameters
7. Spectral parameters
8. Training HMMs
9. Context-dependent HMMs & state duration models
10. Parameter generation from HMMs
11. Excitation generation
12. Synthesis Filter
13. SYNTHESIZED SPEECH
14. TEXT
15. Excitation parameters
16. Spectral parameters
17. Labels

Synthesis part
HMM-based speech synthesis system

Training part

\[ \hat{L} = \arg \max_L P(L | W, \Lambda) \]

\[ \hat{O} = \arg \max_O p(X | O) \]

Excitation parameters

Spectral parameters

Synthesis part

\[ \hat{l} = \arg \max_l P(l | w, \Lambda) \]

\[ \hat{o} = \arg \max_o P(o | \hat{L}, \hat{\lambda}) \]

\[ \tilde{x} \sim p(x | \hat{o}) \]
HMM-based speech synthesis system

Training part

SPEECH DATABASE

Text analysis

Speech signal

\[ \hat{O} = \arg \max_O p(X \mid O) \]

Excitation parameters

Spectral parameters

Training HMMs

Context-dependent HMMs

& state duration models

Parameter generation

from HMMs & state duration models

Labels

Excitation parameters

Spectral parameters

TEXT

Text analysis

Labels

SYNTHESIZED SPEECH
Human speech production

Modulation of carrier wave by speech information

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

Sound source
- Voiced: pulse
- Unvoiced: noise

Speech

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

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

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

ML-estimation of mel-cepstrum

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

- \( x \): speech waveform (Gaussian process)
- \( c \): mel-cepstrum
Waveform reconstruction

Original speech

F0
Unvoiced / voiced
Mel-cepstrum

Synthesis filter

\[ H(z) \]

Excitation

\[ e(n) \]

reconstructed speech

\[ x(n) \]

Pulse train

White noise

These speech parameters should be modeled by HMM
HMM-based speech synthesis system

Text analysis

Speech signal

Training part

Labels

Excitation parameters

Spectral parameters

Context-dependent HMMs & state duration models

TEXT

Parameter generation from HMMs

Labels

Excitation parameters

Excitation generation

Excitation

Synthesis Filter

SYNTHESIZED SPEECH

SPEECH DATABASE

Text analysis

Labels

Synthesis part

\[ \hat{\lambda} = \arg \max_{\lambda} P(O \mid \hat{L}, \lambda) \]
Hidden Markov model (HMM)

- $a_{ij}$: state transition probability
- $b_q(o_t)$: output probability

Observation sequence $o = o_1 \ o_2 \ o_3 \ o_4 \ o_5 \ \cdots \ o_T$

State sequence $q = 1 \ 1 \ 1 \ 1 \ 2 \ \cdots \ 2 \ 3 \ \cdots \ 3$
Structure of state output (observation) vector

\[ o_t \]

\[ c_t \]
\[ \Delta c_t \]
\[ \Delta^2 c_t \]
\[ p_t \]
\[ \Delta p_t \]
\[ \Delta^2 p_t \]

Spectral parameters (e.g., mel-cepstrum, LSPs)

\[ \Delta \]
\[ \Delta \Delta \]

Excitation part

log F0 with V/UV

\[ \Delta \]
\[ \Delta \Delta \]
Observation of F0

Unable to model by continuous or discrete distributions
⇒ Multi-space distribution HMM (MSD-HMM)
Multi-space probability distribution HMM (MSD-HMM)

\[ \Omega_1 = R^{n_1} \]
\[ \Omega_2 = R^{n_2} \]
\[ \Omega_M = R^{n_M} \]
MSD-HMM for F0 modeling

HMM for F0

1

2

3

$\Omega_1 = R^1$

Voiced

$w_{1,1}$

$\Omega_2 = R^0$

Unvoiced

$w_{1,2}$

Voiced / Unvoiced weights

$w_{2,1}$

$w_{3,1}$

Unvoiced

$w_{2,2}$

$w_{3,2}$
Structure of state-output distributions

- $o_t$
- $c_t$
- $\Delta c_t$
- $\Delta^2 c_t$
- $p_t$
- $\Delta p_t$
- $\Delta^2 p_t$

Spectrum

Log F0

- Voiced
- Unvoiced

- Voiced
- Unvoiced

- Voiced
- Unvoiced
Contextual factors

Phoneme

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

Syllable

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

Word

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

Phrase

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

.....

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

Sharing the parameter of HMMs in same leaf node
Stream-dependent tree-based clustering (1)

Spectrum & excitation have different context dependency $\rightarrow$ Build decision trees separately

- Decision trees for mel-cepstrum
- Decision trees for F0
State duration modeling

HMM (Hidden Markov Model)
- State duration prob. depends only on transition prob.
- State duration probability exponentially decreases

HSMM (Hidden Semi Markov Model)
- HMM + explicit duration model ⇒ HSMM

\[ P(q | \hat{l}, \hat{\lambda}) = \prod_{i=1}^{K} p_i(d_i) \]
Stream-dependent tree-based clustering (2)

State duration model

Three dimensional Gaussian

HMM

Decision trees for mel-cepstrum

Decision trees for F0

Decision tree for state dur. models
HMM-based speech synthesis system

\[ \hat{o} = \arg \max_{o} P(o | \hat{l}, \hat{\lambda}) \]
Composition of sentence HMM for given text

This sentence HMM gives \( p(o \mid \hat{l}, \hat{\lambda}) \)
Speech parameter generation algorithm [Tokuda; ’00]

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

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

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

\[
\Downarrow
\]

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

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

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

- \( p_i(\cdot) \): 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 $o = [o_1^T, o_2^T, \ldots, o_T^T]^T$ 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

\[ o_t = \left[ c_t^T, \Delta c_t^T, \Delta^2 c_t^T \right]^T \]
By setting
\[ \frac{\partial \log P(Wc | \hat{q}, \lambda)}{\partial c} = O, \]
we obtain
\[ W^T \Sigma^{-1}_{\hat{q}} Wc = W^T \Sigma^{-1}_{\hat{q}} \mu_{\hat{q}}, \]
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)

\[ W^T \Sigma_{\hat{q}}^{-1} W = \Sigma_{\hat{q}}^{-1} \mu_{\hat{q}} \]

\[ \begin{align*}
W^T & = \begin{bmatrix} 1 & 0 & 0 & \cdots & 1 \\
0 & 0 & 1 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\end{bmatrix} \\
\Sigma_{\hat{q}} & = \begin{bmatrix} \Sigma_{q_1} & \cdots & \Sigma_{q_T} \\
\Sigma_{q_1} & \cdots & \Sigma_{q_T} \\
\vdots & \ddots & \vdots \\
\Sigma_{q_1} & \cdots & \Sigma_{q_T} \\
\end{bmatrix} \\
W & = \begin{bmatrix} 1 & 0 & 0 & \cdots & 1 \\
0 & 0 & 1 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\end{bmatrix} \\
c & = \begin{bmatrix} c_1 \\
c_2 \\
\vdots \\
c_T \end{bmatrix} \\
\mu_{\hat{q}} & = \begin{bmatrix} \mu_{q_1} \\
\mu_{q_2} \\
\vdots \\
\mu_{q_T} \end{bmatrix} \]

Trajectory HMM

\[ P(o \mid l, \lambda) = P(Wc \mid l, \lambda) \] is not a proper distribution of \( c \)

<table>
<thead>
<tr>
<th></th>
<th>Conventional HMM</th>
<th>Trajectory HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td>[ \arg \max_{\lambda} P(O \mid \hat{L}, \lambda) ]</td>
<td>[ \arg \max_{\lambda} P(C \mid \hat{L}, \lambda) ]</td>
</tr>
<tr>
<td><strong>Synthesis</strong></td>
<td>[ \arg \max_{o} P(o \mid \hat{l}, \hat{\lambda}) \mid_{o=Wc} ]</td>
<td>[ \arg \max_{\hat{c}} P(c \mid \hat{l}, \hat{\lambda}) ] → [ \arg \max_{\hat{l}} P(c \mid \hat{l}, \hat{\lambda}) ]</td>
</tr>
</tbody>
</table>

Solve inconsistency between training & synthesis

⇒ improving the model accuracy
Generated spectra

Spectra changing smoothly between phonemes
Generated F0

natural speech

without dynamic features

with dynamic features
Effect of dynamic features

<table>
<thead>
<tr>
<th>log F0</th>
<th>Mel-cepstrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>static $\Delta + \Delta^2$</td>
<td>static $\Delta + \Delta^2$</td>
</tr>
<tr>
<td>static</td>
<td>static</td>
</tr>
<tr>
<td>$\Delta + \Delta^2$</td>
<td>Smooth!</td>
</tr>
</tbody>
</table>

**Effect of dynamic features**

The table shows the effect of different dynamic features on Mel-cepstrum. The top row represents static features with $\Delta + \Delta^2$, and the row below shows the impact on Mel-cepstrum with $\Delta + \Delta^2$ and static features. The last row indicates the smoothness of the effect, marked as "Smooth!".
Overview of this talk

1. Mathematical formulation
2. Implementation of individual components
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

- 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

Model B

Interpolated model

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:
Interpolation of speaking styles

Base model A

Interpolation

Neutral

High Tension

Base model B

extrapolation
Eigenvoice (producing voices) [Shichiri; ’02]

System Overview

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 Vesnicher, 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.
Singing voice synthesis [Oura; ’10] (1/2)

MusicXML

Musical score

Trained HMMs

Male:
Female:

Singing voice database

Synthesized singing voice

Male:
Female:
Adaptation:

Enable computers to sing any song
Overview of this talk

1. Mathematical formulation
2. Implementation of individual components
3. Examples demonstrating its flexibility
4. Discussion and conclusion
Inclusion of all components

Problem of statistical parametric speech synthesis

\[
\hat{x} \text{ from } P(x \mid w, X, W) = \sum_{l,L} \int \int P(x \mid c) \ P(c \mid l, \lambda) \ P(l \mid w, \Lambda) \times P(\lambda \mid C, L) \times P(L \mid W, \Lambda) P(C \mid X) P(\Lambda) d\lambda d\Lambda dc dC
\]

Waveform generation  Parameter generation  Text processing

Posterior of model parameter

Text processing  Speech analysis  Prior
Relaxing approximations

Marginalizing model parameters

➡ Variational Bayesian acoustic modeling for speech synthesis [Nankaku;’03]

Marginalizing labels

➡ Joint front-end / back-end model training [Oura;’08]

Inclusion of waveform generation part

➡ Waveform-level statistical model [Maia;’10]
Summary

Statistical approach to speech synthesis

- Whole speech synthesis process is described in a statistical framework
- It gives a unified view and reveals what is correct and what is wrong
- Importance of the database

Future work

- Still we have many problems should be solved:
  - Speech waveform modeling
  - Combination with text processing part, etc.
Final message

Is speech synthesis a messy problem?

No!

Let us join speech synthesis research!

Thanks!
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.
Rouibia; '05 - "Unit selection for speech synthesis based on a new acoustic target cost," Interspeech, '05.
Huang; '96 - "Whistler: A trainable text-to-speech system," ICSLP, '96.
Mizutani; '02 - "Concatenative speech synthesis based on HMM," ASJ autumn meeting, '02.
Ling; '07 - "The USTC and iFlytech 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.
References (4/4)

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.
Hofer; '07 - "Speech driven head motion synthesis based on a trajectory model," SIGGRAPH, '07.
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 Yamagichi, Tomoki Toda, Yoshihiko Nankaku, Kei Hahimoto, and Sayaka Shiota for their help.