CLOSED-LOOP TRAINING FOR DIPHONETBASED TTS

Masami Akamine, Senior Fellow
Toshiba Research and Development Center
Personal Background

1985: finished Ph.D.
- State-Space Approach to Synthesis of Minimum Quantization Error Digital Filter without Limit-cycles

Professional experiences
- Toshiba R&D Center for 30 years
  - Speech coding
  - Video coding
  - TTS
  - ASR
  - Spoken dialogue
Research topics on TTS at Toshiba labs

- **Unit concatenation-based speech synthesis**
  - Closed-loop training for diphone-based TTS
  - Multiple unit selection and fusion

- **Statistical speech synthesis**
  - Speaker and language adaptation
  - Cluster Adaptive Training (CAT)
  - Speech synthesis using sub-band basis models (SBM)
  - Band-width expansion based on SBM
  - Dynamic sinusoidal models
  - Complex Cepstrum analysis

Today’s lecture
Outline

1. Diphone-based speech synthesis with TD-PSOLA
2. Closed-loop training of diphone units
3. Closed-loop training of prosody generation model
4. Multiple unit selection and fusion
5. Products and services
6. Conclusions
1. Diphone-based speech synthesis with TD-PSOLA

- Structure of TTS
- 3 types of speech synthesis
  - What is the diphone-based speech synthesis?
- Prosodic modification using TD-PSOLA
- Speech quality problems caused by prosodic modification
Structure of TTS

Text analysis

Prosody generation

Speech synthesis

Synthetic speech

Pronunciation
Accent
Linguistic inf.

Pitch
Duration

Text
Methods of speech synthesis

- Formant synthesis
- Unit concatenation-based speech synthesis
  - Single inventory: diphone-based speech synthesis
  - Multiple inventory: unit selection-based speech synthesis
- Statistical parametrical speech synthesis
  - HMM-based speech synthesis
  - Deep learning, e.g. DNN, RNN, LSTM
Formant synthesis

- **Klattalk (Klatt, 1983)**
  - Based on source–filter model (vocoder)
  - Handcrafted rules to map phoneme to model parameters

Pulse train (voiced)

White noise (unvoiced)

u/v    f0    gain

Synthesis filter

Speech

u/v    f0    gain

F₁,₂,₃  Q₁,₂,₃  gain

F₁  F₂  F₃

http://www.cs.indiana.edu/rhythmmsp/ASA/Contents.html
Diphone-based speech synthesis

Phoneme sequence

$tadai$ma$

Unit selection

Prosodic modification

Unit concatenation

Phoneme sequence

$ta\ ad\ da\ ai\ im\ ma$
Unit selection-based speech synthesis

- Multiple inventory with different spectrum, f0, duration
- Unit selection based on target cost and concatenation cost
- High quality but large data

Linguistic features → Unit selection → Unit concatenation

Target cost
\[ C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \]
Differences in features, e.g. phoneme, position, stress, etc.

Concatenation cost
\[ C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) \]

A. Hunt and A. Black, ICASSP96
Statistical parametric speech synthesis

- Statistical modeling of vocoder parameters
- Parameter generation and speech synthesis

K. Tokuda et al., ICASSP2000
Diphone-based TTS

A research project of TTS started in 1994.
Why diphone-based speech synthesis?

- **Business target**
  - Focused on applications to embedded systems
    - The market of GPS navigation was growing
    - VUI was essential
    - Good synergy with semiconductor business in Toshiba

- **Research objectives**
  - A method to synthesize high quality speech
    - No muffled voice, clear voice
    - Natural prosody, lively intonation
  - Small footprint (memory size): less than 1 MB
  - Low computational complexity: less than 15 MIPS
Memory size and speech quality

- **Poor**
  - Muffled voice and flat f0

- **Fair**
  - E-dictionary

- **Good**
  - Diphone-based synthesis
  - Unit selection-based synthesis

Our target

Memory size [MB]
Prosodic modification

- Diphone-based speech synthesis needs prosodic modification of the diphone unit

- Requirements
  - Prosodic modification of speech with minimum distortions
    - Time-scale modification modifies the duration without changing pitch
    - Pitch-scale modification modifies the pitch without changing its duration
  - C.f., speeding up recordings to shorten the duration changes the pitch

![Image of a tape recorder]
TD-PSOLA

- Time domain - pitch synchronous overlap and add
  - Signal processing to perform time-scale and pitch-scale modification of speech
  - Isolate pitch waveforms in the original signal by windowing the signal
  - Perform the modification by allocating isolated waveforms at new pitch marks
  - Resynthesize the final waveform through an overlap-add operation
Pitch modification

Windowing to decompose into pitch waveforms

Overlap and add at new pitch marks

[Hamon et.al, ICASSP89]
Pitch and duration modification

- Decompose into pitch waveforms
- Duplicate or delete pitch waveforms
- Resynthesize speech waveforms at new pitch marks
Distortion caused by prosodic modification

- Prosodic modification: resample the spectrum envelope at different harmonic frequencies from original

- Large distortion for high pitch voice and large modification of F0
Handcrafting of diphone units

- Clip small segments from recorded speech and select proper units by hand
  - Trial and error process
  - Labor and time consuming

Diagram:
- Recorded speech → Clipping → Evaluation by listening → Diphone units

Trial and error process
Problems at the time

- Poor speech quality and unnatural prosody
  - Muffled voice
  - Flat intonation
  - Robotic voice

- Time consuming to build up synthesis units
  - 2 or 3 years to develop a prototype of a voice
  - Trial and error
  - Hand tuning
2. Closed-loop training of diphone units

- Formulate the distortion in speech synthesis as error between original and synthesized speech
- Generate diphone units that minimize the distortion

- Two methods of closed-loop training
  - Method 1: Unit selection by closed-loop training
  - Method 2: Analytic generation by closed-loop training
Formulation of distortion in speech synthesis

Introduce recorded voice as reference

Text input and waveform output, no way to calculate the distortion
Closed-loop training of diphone units

- **Method 1: Unit selection by closed-loop training**
  - Set candidates
  - Modify prosody of candidates and resynthesize waveforms
  - Calculate distortion between synthesized and original waveforms in training data
  - Select diphone units that minimize the total sum of distortion

- **Method 2: Analytic generation by closed-loop training**
  - Formulate synthesized waveforms as a function of diphone unit
  - Formulate distortion between synthesized and original waveforms
  - Analytically generate the optimum unit that minimizes the total sum of distortion

  [Kagoshima et al., ICASSP97], [Akamine et al., ICSLP98]
Closed-loop training

Method 1

Prosody Analysis

Prosodic Modification

Candidates

Distortion Evaluation

Units Selection

Training Data

Synthesis Units

Select diphone units that minimize the total sum of distortion
Prosodic modification

Candidate for synthesis unit $u(n)$

Training data $r(n)$

Synthetic speech $\hat{u}(n)$

1. decompose
2. align
3. overlap-add
Distortion evaluation

Training data $r(n)$

Synthetic speech $y(n)$

Distortion

$$e = \sum_n \left( \frac{r(n)}{\bar{r}} - \frac{y(n)}{\bar{y}} \right)^2$$

Squared error with normalized power
Synthesis units selection

Training data

<table>
<thead>
<tr>
<th></th>
<th>( r_1 )</th>
<th>( r_2 )</th>
<th>( r_3 )</th>
<th>( r_4 )</th>
<th>( r_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>( e_{11} )</td>
<td>( e_{12} )</td>
<td>( e_{13} )</td>
<td>( e_{14} )</td>
<td>( e_{15} )</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>( e_{21} )</td>
<td>( e_{22} )</td>
<td>( e_{23} )</td>
<td>( e_{24} )</td>
<td>( e_{25} )</td>
</tr>
<tr>
<td>( u_3 )</td>
<td>( e_{31} )</td>
<td>( e_{32} )</td>
<td>( e_{33} )</td>
<td>( e_{34} )</td>
<td>( e_{35} )</td>
</tr>
<tr>
<td>( u_4 )</td>
<td>( e_{41} )</td>
<td>( e_{42} )</td>
<td>( e_{43} )</td>
<td>( e_{44} )</td>
<td>( e_{45} )</td>
</tr>
</tbody>
</table>

Error matrix

Cost function

\[
C(i_1, \cdots, i_n) = \frac{1}{M} \sum_{j=1}^{M} \min(e_{i_1,j}, \cdots e_{i_n,j}) \quad (n > 1)
\]

\[
C(i) = \frac{1}{M} \sum_{j=1}^{M} e_{ij} \quad (n = 1)
\]

Synthesis units minimizing \( C() \) are selected.
Example (n=2)

<table>
<thead>
<tr>
<th>Training data</th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
<th>$r_4$</th>
<th>$r_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$u_2$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>$u_3$</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$u_4$</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Cost function

$C(1,2) = (1 + 3 + 2 + 2 + 1) / 5 = 1.8$

$C(1,3) = (1 + 3 + 2 + 2 + 3) / 5 = 2.2$

$C(3,4) = (2 + 1 + 2 + 4 + 3) / 5 = 2.4$

$u_1$ and $u_2$ are selected as synthesis units.
Experiment

**Synthesizer**
- Japanese
- waveform concatenation based
- CV/VC units (diphone)
- one synthesis unit per diphone unit (262 units)
- 11.025 kHz sampling

**Training data**
- 50 min/speaker
- one male and one female speaker
Comparison between CLT and OLT

Open Loop Training

step 1. Calculate LSP parameters of all the candidates.

step 2. Take an average of all the LSP parameters.

step 3. Select the synthesis unit whose LSP parameters are the closest to the average.
Experimental Result

Test sentences  10
Subjects        7 males
Prosody        Natural speech

Female Speech
Male Speech

Preference Score (%)
Method 2: Analytic Generation
Closed-loop training to minimize the distortion

Prosodic modification

Prosody analysis

Speech DB
Training data

Error

Synthesizes speech vector

Minimization

Unknown unit $u$

Automatic generation of optimum units

Optimum unit

Synthesis units
Analytic method

1. Prepare speech segments as training data
2. Set initial synthesis unit vectors
3. Partition training vectors into cluster sets based on the nearest neighbor condition
4. Generate the optimal unit vector that minimizes distortion in each cluster
5. Update the synthesis unit vectors
6. Repeat Step 3 to 5

A generalized method to create a multiple entry for a diphone unit
Partition of training vectors

\[ G_i = \{ r_j : d(r_j, y_{j,i}) < d(r_j, y_{j,k}); \text{ all } k \neq i \} , \]

Given an initial vector \( u_i \), partition \( r_j \) based on the distance

\( r_j \) : training vector
\( y_{j,i} \) : synthesized vector from unit vector \( u_i \)
\( d(*,*) \) : distance measure
Generate the optimal unit vector that minimizes the distortion in each cluster.
Prosodic modification

$y_{j,i}$ is produced by modifying its pitch period and duration so that they are identical with those of the training vector.

(a) Unit vector, (b) Pitch waveform vector, (c) Training vector, (d) Modified vector
Formulation of prosodic modification

\[ y = Au \]

Resynthesized waveform

Original unit

\[ A: \text{overlap and add operation} \]
Distortion

Synthesis unit

\[ y = Au \]

Synthesized vector

\[ y \]

Error signal

\[ e = r - gy \]

Training vector

\[ r \]

Distortion

\[ e = (r - gAu)^T(r - gAu) \]
Automatic generation of the optimum unit

Calculate the distortion over all the training vectors

\[ E = \sum_j (r_j - g_j A_j u)^T (r_j - g_j A_j u) \]

\[ \frac{\partial E}{\partial u} = 0 \]

Solve linear equations

Obtain the optimum unit \( u \)

\[
\left( \sum_{r_j \in G_i} g_{j,i}^2 A_{j,i}^T A_{j,i} \right) u_i = \sum_{r_j \in G_i} g_{j,i} A_{j,i}^T r_j.
\]
Experiments

- Training corpus: 40 minute long, Japanese male and female voices
- Synthesis units: 302 CV-VC units
- Prosody was produced by a method based on closed-loop training
Objective evaluation

Distortion and the number of synthesis units for Method 1 and Method 2
### Subjective evaluation

<table>
<thead>
<tr>
<th></th>
<th>Method 2</th>
<th>Method 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Analytic method (80 %)</td>
<td>Selective method</td>
</tr>
<tr>
<td>Female</td>
<td>Analytic method (76 %)</td>
<td>Selective method</td>
</tr>
</tbody>
</table>

Preference scores for analytic method and selective method
Closed-loop training

- Minimize distortion in synthetic speech

Extract the most likely spectrum over different f0

Optimum unit

Exploit increase of sampling points of spectrum

Same envelope, different f0
3. Closed-loop training of prosody generation model

- Generation model
- Closed-loop training of F0 contour codebook
- Representative vector selection
Prosody generation model

- Grammatical attributes
  - Vector Selection
  - Expansion or Contraction
- Duration
  - Offset Level Prediction
  - Offset adjustment

F0 Contour Codebook

Graphs showing F0 contour selection and adjustment.
Prosody generation algorithm

- Generate F0 contour codebook using closed-loop training method
- Select contour vector based on grammatical attributes
- Predict offset based on grammatical attributes
- Quantification Theory Type I is used for contour vector selection model and offset prediction model
Training of an F0 contour codebook

- Duration corpus
- F0 Contour corpus
- Vector modification
- Error calculation and clustering
- Codebook generation
- Optimal offset level
- Approximation error for Ci

Training data for estimating approx. error and offset
Generate the optimal vector that minimizes the distortion in each cluster
Codebook generation

F0 contour control model

segmental contour

duration matrix

representative vector

unit vector

offset level

Total approximation error

\[ E_i = \sum_{\mathbf{r}_j \in G_i} (\mathbf{r}_j - (D_j \mathbf{c}_i + b_{ij} \mathbf{i}))^T (\mathbf{r}_j - (D_j \mathbf{c}_i + b_{ij} \mathbf{i})) \]

Codebook generation

\[
\frac{\partial E}{\partial \mathbf{c}_i} = 0 \\
\left( \sum D_j^T D_j \right) \mathbf{c}_i = \sum D_j^T (\mathbf{r}_j - b_{ij} \mathbf{i})
\]
Representative vector selection

- Approximation error prediction model
  - Quantification Theory Type I

- Training data
  - Approximation error corpus obtained through codebook training
## Quantification theory Type I

<table>
<thead>
<tr>
<th>attribute</th>
<th>category</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>first</td>
<td>$a_1$</td>
</tr>
<tr>
<td>(a)</td>
<td>middle</td>
<td>$a_2$</td>
</tr>
<tr>
<td></td>
<td>last</td>
<td>$a_3$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$b_1$</td>
</tr>
<tr>
<td>(b)</td>
<td>2</td>
<td>$b_2$</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>$b_{11}$</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>$b_{12}$</td>
</tr>
<tr>
<td>(c)</td>
<td>next</td>
<td>$c_1$</td>
</tr>
<tr>
<td></td>
<td>next but one</td>
<td>$c_2$</td>
</tr>
<tr>
<td></td>
<td>next but over one</td>
<td>$c_3$</td>
</tr>
<tr>
<td>(d)</td>
<td>next</td>
<td>$d_1$</td>
</tr>
<tr>
<td></td>
<td>next but one</td>
<td>$d_2$</td>
</tr>
<tr>
<td></td>
<td>next but over one</td>
<td>$d_3$</td>
</tr>
</tbody>
</table>

(a) position of current phrase in sentence  
(b) current phrase length

Predicted value:

$$a_2 + b_2 + c_3 + d_1 = x$$
Offset level prediction

- Grammatical attributes → Offset level prediction → Offset level

- **Offset level prediction model**
  - Quantification Theory Type I

- **Training data**
  - Optimal offset level corpus obtained through codebook training
Experiment

- **Data sets**
  - Japanese
  - Male voice
  - 866 sentences (6398 phrases)
  - Automatic phoneme labeling
  - Automatic F0 extraction + manual correction

- **F0 contour codebook**
  - 48 vectors (6 accent types $\times$ 8 vectors)

- **Grammatical attributes**
  - Accent type, phrase length, stress, part of speech, etc.
Listening tests

- **Speech samples**
  - (a) synthetic speech by proposed method
  - (b) synthetic speech by conventional method (Shiga et al., ICSLP94)
  - (c) synthetic speech with original prosody

- **Naturalness test**: “Which is more natural, (a) or (b) ?”

- **Individuality test**: “Which resembles (c), (a) or (b) ?”

<table>
<thead>
<tr>
<th>preference score</th>
</tr>
</thead>
</table>
| naturalness test          | 84.3%  
| individuality test        | 92.9%  

Examples of F0 contour

- Original
- Proposed method
Memory size and speech quality

Quality

good

fair

poor

Memory size [MB]

Closed-loop training

Unit selection-based synthesis

E-dictionary

Diphone-based synthesis

Quality

poor

fair

good

1

10

100

1000

Memory size [MB]
Demonstration

- Japanese for car navigation systems
  - Footprint: 500 kB (Japanese) / 800 kB (English)

- UK English for car navigation systems
4. Multiple unit selection and fusion

- Motivation
- Unit selection
- Units fusion
Motivation

- Scalable systems with different corpus sizes and stable voice quality
- Change the number of units to be fused depending on the corpus size
- Multiple unit selection and fusion, when using a small corpus → stable sound

[Tamura et al., ICASSP2005]
Unit selection

F0, duration, phoneme
↓
unit pre-selection by target cost
↓
search for the best path based on target cost and concatenation cost
↓
select N units based on the best path, target cost and concatenation cost
↓
selected units

p1, p2, p3, p4
↓

u1, u2, u3, u4, u5
Waveform generation

- Voiced waveform
  - Plural unit fusion
  - Overlap-add synthesis
  - Unit concatenation

- Unvoiced waveform
  - Unit concatenation

TD-PSOLA
Unit fusion

1. Selected units
2. Unit fusion
3. Windowing
4. Adjust duration
5. Adjust phase
6. Fused unit
7. Average
Demonstration: a boy and a girl

Quarrel about cakes

あ～、お姉ちゃんが僕の分食べたー

私の方が1つ上なんだから、いいでしょ

お母さんに言いつけるぞー

すぐ言いつける、この言いつけ虫

お姉ちゃんだって、お菓子虫じゃないかい！
Products and services
5. Products and services

1. Human machine interface
   • In-car navigation systems
   • Home appliances, TVs, air conditioners, fridges
   • E-book reader
2. Contents creation service
   • E-learning
   • Multi-media texts for education
3. Smart-phone applications
Businesses in Toshiba

Middleware licensing

Value addition to Toshiba products

Micro processor

Semiconductor devices
Products and services using TTS

- Chip
- Software
- Contents

**Chip**
- FAX
- Telephone

**Software**
- GPS car navigation
- Video game software
- PC software
- Service for mobile phone
- Smart phone Tablet PC

**Contents**
- Elevator
- NEWS
- Internet service

Toshiba Leading Innovation
ASR & TTS on Bluetooth™ chipset for in-car use

You can make hands-free call while driving.

Call Bob.

- **ASR:**
  - Small vocabulary, e.g. Commands, digit
  - American/European languages, 9 in total
  - Focus on noise robustness especially for car

---

**Micro processor**

- Grapheme-to-phoneme translator

**Bluetooth™**

**Mobile-phone**

**ASR**

**TTS**

**phoneme sequence**
User’s TTS voice creation

Open TTS voice creation to public (http://tospeak.ivc.toshiba.co.jp/grcd)

- Provide a pipeline system to create new voices
- A user uploads recorded voice of 100 sentences
- The process completes in 1 hour
- Created more than 800 voices
Variation

Speaker variation

Speaking style variation

- (.wav)
- (.wav)
- (.wav)
- (.wav)
- (.wav)
- (.wav)
- (.wav)
- (.wav)
- (.wav)

reading
formal conversation
reading reminder alert
Scalability

MOS

Footprint

2.5MB

30MB

60MB

100MB

[MB]
RECAIUS™: Cloud service of media processing

- Contents creation using TTS
- Transcription service using ASR
- Speech to speech translation applications
- Spoken dialogue

https://developer.recaius.io/jp/top.html

https://www.toshiba.co.jp/cl/pro/recaius/index_j.htm
CAT-based statistical models

- Cluster adaptive Training
- The means of the distributions are defined as a linear combination of P clusters (factors):

- CAT weight vector \( \lambda_i \) represents continuous voice space

\[
\mu_m^{(s)} = \sum_{i=1}^{P} \lambda_i^{(s)} \mu_{c(m,i)}
\]

- \( \mu_{c(m,i)} \): the mean vector of the \( i \) cluster
- \( \lambda_i \): the weight for Cluster \( i \)
- The variance is shared across all clusters
CAT: Model structure

- Unlike eigenvoice models, CAT doesn’t need to force the clusters to be orthogonal.
- Decision trees can be independent, i.e., tree intersection

=> more context can be represented efficiently

Single tree
8 leaves => 8 context

CAT Tree intersection
7 leaves => 12 context
Control of voice character using CAT weights

- Create a wide variety of TTS voices by controlling elements of voice characteristics such as the age, gender, and brightness of the voice.

\[
\lambda = \lambda^{(0)} + \sum_{e=1}^{E-1} w^{(e)} \cdot (\lambda^{(e)} - \lambda^{(0)});
\]

Average voice model \(\lambda^{(0)}\)
6. Conclusions

- **Diphone-based TTS**
  - Prosodic modification using TD-PSOLA
  - Distortion in speech quality when large modification, high pitch voice

- **Closed-loop training of diphone units**
  - Method 1: Unit selection by closed-loop training
  - Method 2: Analytic generation by closed-loop training

Minimize distortion between synthesized and original waveforms

*cf. MGE training for HMM-TTS*  [Wu et al., ICASSP2006]

- **Closed-loop training of pitch contour codebook**
- **Applications to embedded systems**