Gesture Design
for a Real-time Gesture-to-Speech Conversion System
Based on Space Mapping
Between a Gesture Space and an Acoustic Space

37-097093 Aki Kunikoshi
Hirose & Minematsu Lab
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
Overview

- **Introduction**
  - GMM-based media conversion
    - Fundamentals of Speech Processing
    - Estimation of a mapping function between two spaces
    - Hand-to-Speech system for the five Japanese vowels

- **Nasal Generation**
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation

- **Real-time Hand-to-Speech system**
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations

- **Wrap-up**
# Speech synthesis

<table>
<thead>
<tr>
<th>An example</th>
<th>Symbol input (Text-to-Speech: TTS)</th>
<th>No symbol input</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT&amp;T Natural Voices®</td>
<td>Vocal Tract Lab (‘10 Birkholz)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Articulate speech synthesis</td>
<td>- Articulatory synthesis</td>
</tr>
<tr>
<td>- Simple input method</td>
<td>- Formant synthesis *</td>
</tr>
</tbody>
</table>

* Formant synthesis can be used for TTS as well
Speech synthesis from body motion

Articulatory synthesis

Direct control

Electrical

Formant synthesis

Parametric control

Mechanical
Articulatory synthesis

A synthesis method inspired by the human speech production process

The resonance cavities tune pitches for different vowels
(1780 Donders)
Articulatory synthesis may be used as:

- a tool in basic speech synthesis
- an educational device in the acoustic phonetics
- a way to demonstrate the speech production mechanism
- etc.

Most of them are based on data from one person

They are not easy to adapt to other people
Speech synthesis from body motion

Articulatory synthesis

Direct control

Electrical

Parametric control

Mechanical

Formant synthesis
Formant synthesis

A synthesis method utilizing resonant frequencies of the vocal tract

/i/
/o/
/u/

F1, F2, F3

First Formant [kHz]

Second Formant [kHz]

Male average

Female average

/i/ /e/ /a/ /u/ /o/
Formant synthesis

Speech synthesizer using a pen tablet
(’07 Yabu et al.)

Glove Talk II
(’98 Fels et al.)

These applications are designed to make optimal use of the features specific to the input device.

Their methodologies are difficult to apply to another input device.
Possibility of media selection

- Artists explore various media to pursue the potentials of expression
- For assistive devices, the appropriate input media should be chosen according to the user’s handicap and his or her remaining capabilities

Hand motion

Theremin (1919 Theremin)

Pen stroke

HandSketch (’98 d’Alessandro et al.)

Electric wheelchairs
- EEG (’09 Choi et al.)
- Arm motion (’06 Sakaue et al.)
- Speech (’06 Kojima et al.)
Synthesis based on space mapping

Speech generation from body motion

Mapping from a body motion space to a speech space

Voice Conversion

The speech space of the source speaker

\[ F \]

The speech space of the target speaker

a speech sequence

a speech sequence
Our objective
To establish a media independent methodology for a speech generation system

Speech synthesis based on space mapping to desired input media, for example, hand motion.

A gesture space

\[ \mathcal{F} \]

An acoustic space

How to correspond spaces when input media does not have explicit relationship with an acoustic space?
Speech synthesis from body motion

Articulatory synthesis

Direct control

Electrical

Formant synthesis

Parametric control

Mechanical

Space mapping

Articulatory synthesis

Formant synthesis

Space mapping
Overview

- Introduction

- **GMM-based media conversion**
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels

- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation

- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations

- Wrap-up
The mechanism of Speech production

\[ S(\omega) = H(\omega) \times G(\omega) \]

\[ \log S(\omega) = \log H(\omega) + \log G(\omega) \]
Logarithmic Power Spectrum: \( \log S(\omega) \)

\[
\log S(\omega) = \log H(\omega) + \log G(\omega)
\]

- the vocal tract
- the vocal chord

Lower dimension of cepstrum

Log-spectrum

Spectral envelope

Log-spectrum

IDFT

Cepstrum

DFT

0-17 deg
Feature vectors of speech

Speech signal

Cepstrum vectors

Cepstrum time sequence
CyberGlove

(a) 

(b) 
18 sensors

18 sensors

○ ○ ○ ⋅ ⋅ ○

DP
PIP
MCP
Thumb

Middle
Index
Ring
Pinky

IP
MCP

PIP
MCP
ABD
IP
The framework of hand motion to speech conversion

\[ h = (h_1, h_2, h_3, \ldots, h_{18}) \]

\[ s = (s_1, s_2, s_3, \ldots, s_{18}) \]
Overview

- **Introduction**
- **GMM-based media conversion**
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- **Nasal Generation**
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- **Real-time Hand-to-Speech system**
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- **Wrap-up**
GMM-based feature mapping

(’98 Kain et al.)

Source feature vector sequence

\[ x \]
\[ x_1 \quad x_2 \quad \cdots \quad X_T \]

Target feature vector sequence

\[ y \]
\[ y_1 \quad y_2 \quad \cdots \quad y_T \]

Input source vector sequence

\[ x' \]
\[ x'_1 \quad x'_2 \quad \cdots \quad x'_T \]

Joint feature vector sequence

\[ z \]
\[ z_1 \quad z_2 \quad \cdots \quad z_T \]

output target vector sequence

\[ y' \]
\[ y'_1 \quad y'_2 \quad \cdots \quad y'_T \]

\[ \dot{y} = \arg\max_y P(y_t | x_t) \]

How to design the correspondence between hand gestures and sounds???
Challenge

How to design the correspondence between hand gestures and sounds

Japanese finger alphabet cannot be used

We derive the correspondence between gestures and sounds
Overview

- Introduction

- **GMM-based media conversion**
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels

- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation

- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations

- Wrap-up
The gesture design for the five Japanese vowels

By considering the topological equivalence between the structure of hand gestures in the gesture space and that of vowel sounds in the vowel space, a quasi-optimal correspondence can be obtained.

’08 Kunikoshi, Qiao et al.

The gesture space
Data Glove data space

The vowel sounds space
F1-F2 space
The Basic 28 Hand Gestures

(’99 Wu et.al.,)
Gesture design for the five Japanese vowels

(’09 Kunikoshi et al.)

The location of 28 gestures in PCA space

Equivalence between two spaces
The preliminary experiment

1

/g/' /i/' /u/' /e/' /o/

2

gestures  cepstrums  augment vectors

3

GMM
## Experimental conditions

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gesture</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature</td>
<td>CyberGlove sensor data (18 deg)</td>
<td>Cepstrum (0-17 deg)</td>
</tr>
<tr>
<td>Sampling period / Shift length</td>
<td>10～20 ms (variable)</td>
<td>1 ms</td>
</tr>
<tr>
<td>Window</td>
<td>------</td>
<td>Hamming</td>
</tr>
<tr>
<td>Training sets</td>
<td>5 vowels and transitions between each vowel pair (5 + 5P_2 = 25)</td>
<td></td>
</tr>
<tr>
<td>Number of training sets</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>GMM</td>
<td>1 mixture</td>
<td></td>
</tr>
</tbody>
</table>
Two styles of Hand-to-Speech generation

articulate

inarticulate
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
Classification of Japanese consonants

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Alveolar</th>
<th>Alveolo-palatal</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
<th>Synthesis methods in our system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fricatives</td>
<td>φ</td>
<td>s</td>
<td>z</td>
<td>ɕ</td>
<td>ʑ</td>
<td>ɕ</td>
<td>waveform concatenation</td>
</tr>
<tr>
<td>Affricates</td>
<td>tʃ</td>
<td>dz</td>
<td>tɕ</td>
<td>dʒ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plosives</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>k</td>
<td>g</td>
<td>Expansion of S2H framework used for vowel sound generation</td>
</tr>
<tr>
<td>Nasals</td>
<td>m</td>
<td>n</td>
<td>ŋ</td>
<td>ŋ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap and Flips</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Vowel’s transition</td>
</tr>
<tr>
<td>Semivowels</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
<td>w</td>
<td></td>
</tr>
</tbody>
</table>

Semivowels
People perceive several Japanese words which include semivowels only with the formant transitions of vowels (’09 Yabu et al.)

Fricatives, affricates and plosives
- They are not influenced by the speaking rate of following vowels. (’84 Sagisaka et al.)
- VOT (Voice Onset Time) influences to the perception of those consonants
  →preset waveforms and control VOT with body gestures

Nasals and “Tap and Flap”
The characteristics of nasals are, like vowels, described with resonance and anti-resonance characteristics. →Expansion of S2H framework for vowels
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
Gestures for consonants

When the gestures for vowels are given,
- Which gestures correspond to consonants?
- Which paths are to be used for transitions?

Inappropriate gesture design for consonants causes a lack in the transition parts in synthesized speech (’10 Kunikoshi et al.,)
Speech-to-Hand conversion system

To derive gestures for consonants, an S2H system is made using parallel data for vowels.

**Probabilistic Integration Model** (’10 Saito et al.,)

\[
y^* = \underset{y_t}{\arg \max} P(y_t|x_t)
\]

\[
x : \text{speech} \quad y : \text{gesture}
\]

from the Joint Density Model using parallel data of **vowels and consonants**

\[
= \underset{y_t}{\arg \max} P(x_t|y_t)
\]

\[
P(y_t)
\]

Joint Density Model using parallel data for **vowels**

from the Gesture Model

We endeavoured to come up with gestures for **consonants** using the **Speech to Hand** system, constructed with the **vowel data**
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
The procedure of experiments

1. Choose the gesture design for vowels

2. Make S2H and H2S systems

3. Estimate gestures for consonants

4. Compare those gesture designs
The procedure of experiments

(1) Choose the gesture design for vowels

/a/  /i/  /u/  /e/  /o/

Gesture model  Conversion model  Sound model

Speech  S2H system  Hand  H2S system  Speech
Candidate sets for five hand gestures of vowels

16 easy-to-form gestures are considered as candidate for gestures

The location of 28 gestures in PCA space

Accounting for the equivalence with the F1-F2 chart, 8190 combinations are chosen from $^{15}P_4$
The procedure of experiments

(2) Make S2H and H2S systems
The experimental conditions of the gesture / sound model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gesture</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature</strong></td>
<td>CyberGlove Sensor data (18 dim) # PCA is performed</td>
<td>STRAIGHT-based cepstrum (1-18 deg)</td>
</tr>
<tr>
<td><strong>Sampling period</strong></td>
<td>10～20 ms (variable)</td>
<td>1 ms</td>
</tr>
<tr>
<td><strong>Shift length</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Window</strong></td>
<td>------</td>
<td>Hamming</td>
</tr>
<tr>
<td><strong>training sets</strong></td>
<td>16 gestures and transitions between each pair (16 + 16P2 = 256)</td>
<td>A set (50 sentences) from the ATR phoneme balanced set</td>
</tr>
<tr>
<td><strong># of training sets</strong></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong># of mixtures of GMM</strong></td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>
### The experimental conditions of the joint density model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gesture</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CyberGlove Sensor data</td>
<td>STRAIGHT-based cepstrum</td>
</tr>
<tr>
<td></td>
<td>( 18 deg )</td>
<td>( 1-18 deg )</td>
</tr>
<tr>
<td></td>
<td># PCA is performed</td>
<td></td>
</tr>
<tr>
<td>Sampling period / Shift length</td>
<td>10〜20 ms ( variable )</td>
<td>1 ms</td>
</tr>
<tr>
<td>Window</td>
<td>-----</td>
<td>Hamming</td>
</tr>
<tr>
<td>training sets</td>
<td>The 5 isolated vowels and the 5P2 transitions for each permutation of two vowels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 5 + 5P2 = 25 in total )</td>
<td></td>
</tr>
<tr>
<td># of training sets</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td># of mixtures of GMM</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
The procedure of experiments

(3) Estimate gestures for consonants
<table>
<thead>
<tr>
<th></th>
<th>/a/</th>
<th>/i/</th>
<th>/u/</th>
<th>/e/</th>
<th>/o/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td><img src="image1" alt="Sample 1" /></td>
<td><img src="image2" alt="Sample 1" /></td>
<td><img src="image3" alt="Sample 1" /></td>
<td><img src="image4" alt="Sample 1" /></td>
<td><img src="image5" alt="Sample 1" /></td>
</tr>
<tr>
<td>Sample 2</td>
<td><img src="image1" alt="Sample 2" /></td>
<td><img src="image2" alt="Sample 2" /></td>
<td><img src="image3" alt="Sample 2" /></td>
<td><img src="image4" alt="Sample 2" /></td>
<td><img src="image5" alt="Sample 2" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>/m/</th>
<th>/n/</th>
<th>/r/</th>
<th>/b/</th>
<th>/p/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td><img src="image1" alt="Sample 1" /></td>
<td><img src="image2" alt="Sample 1" /></td>
<td><img src="image3" alt="Sample 1" /></td>
<td><img src="image4" alt="Sample 1" /></td>
<td><img src="image5" alt="Sample 1" /></td>
</tr>
<tr>
<td>Sample 2</td>
<td><img src="image1" alt="Sample 2" /></td>
<td><img src="image2" alt="Sample 2" /></td>
<td><img src="image3" alt="Sample 2" /></td>
<td><img src="image4" alt="Sample 2" /></td>
<td><img src="image5" alt="Sample 2" /></td>
</tr>
</tbody>
</table>
The procedure of experiments

(4) Compare those gesture designs
The cepstral RMSE (Root Mean Square Error) between the input speech and the output speech

<table>
<thead>
<tr>
<th>Letter</th>
<th>Average of the 8190 gesture designs</th>
<th>One of the appropriate designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>i</td>
<td>0.29</td>
<td>0.28</td>
</tr>
<tr>
<td>u</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>e</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>o</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>na</td>
<td>0.30</td>
<td>0.31</td>
</tr>
<tr>
<td>ni</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>nu</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>ne</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>no</td>
<td>0.27</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
Subjective evaluation - AB preference test

- A and B were the output speech of the H2S systems which were trained by the conventional and the proposed design, respectively.
- A sample has /na/, /ni/, /nu/, /ne/, /no/, 5 data synthesized by both designs, 10 data in total.
- 15 Japanese native speakers select the preferred one from a pair of synthesized speech in terms of naturalness.
- Options are A, B and No preference.
Gesture design comparison

<table>
<thead>
<tr>
<th>Gesture</th>
<th>Male Average</th>
<th>Female Average</th>
<th>Previous</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/u/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/a/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/o/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/e/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/n/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{Inul} \leq \min(\text{Inal, Inil, Inel, Inol}) \]
# The experimental condition

Feature, sampling period / shift length, window are the same as the previous one

<table>
<thead>
<tr>
<th>training sets</th>
<th>Vowels</th>
<th>Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The 5 isolated vowels and the 5P2 transitions for each permutation of the two vowels: 1 set</td>
<td>&lt; Previous method &gt;&lt;br&gt;Consonant and transition parts of /na/, /ni/, /nu/, /ne/, /no/, which were recorded from the same speaker: 1 sets&lt;br&gt;5 x 2 = 10 (7,359 and 9,533 frames for each) in total</td>
</tr>
<tr>
<td></td>
<td>5 + 5P2 = 25 (125,147 frames) in total</td>
<td>&lt; Proposed method &gt;&lt;br&gt;Consonant and transition parts of /na/, /ni/, /nu/, /ne/, /no/, which are synthesized by S2H system: 10 sets&lt;br&gt;5 x 2 x 10 = 100 (10,370 and 5,044 frames for each) in total</td>
</tr>
<tr>
<td># of mixtures of GMM</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>

"The experimental condition"
Naturalness

- Proposed: 48%
- No preference: 27%
- Previous design: 24%
Synthesized speech for /na/

Re-synthesized speech

Proposed design

Previous design
Overview

❖ Introduction
❖ GMM-based media conversion
  ◆ Fundamentals of Speech Processing
  ◆ Estimation of a mapping function between two spaces
  ◆ Hand-to-Speech system for the five Japanese vowels
❖ Nasal Generation
  ◆ Japanese consonant classification
  ◆ Gesture design based on Speech-to-Hand system
  ◆ Evaluation of proposed designs
  ◆ Subjective evaluation
❖ Real-time Hand-to-Speech system
  ◆ The problem of real-time H2S system
  ◆ The effect of alpha & delta
  ◆ The improved real-time H2S system
  ◆ Subjective user evaluations
❖ Wrap-up
How to make the H2S system

Model-Integration VC (MIVC)

Gesture Model

Joint Model

parallel data for vowels

S2H system

Gesture

parallel data for consonants

Speech

Joint Model

H2S system

conventional VC
The problem of real-time H2S system

/n/ part of /nu/, which is synthesized with recorded data (offline)

- Why does it appear?
- Why did not it appear in the offline experiments?

The difference between real-time and offline systems is gesture data, both in training and in conversion phase.
The output of S2H system

Estimated gesture for /nu/, which is derived with an S2H system

Impossible to change gestures so fast!

Solution 1: Considering the weight of the models in MIVC
Solution 2: Considering the delta features in the S2H system
Overview

- **Introduction**

- **GMM-based media conversion**
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels

- **Nasal Generation**
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation

- **Real-time Hand-to-Speech system**
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations

- **Wrap-up**
The weight of the models

\[ \hat{h}_t = \arg\max_h P(h|s) = \arg\max_h P(s|h)P(h) \]

To obtain the optimal \( \hat{s} \)
the following likelihood function is maximized:

\[ \mathcal{L}(h_t; s_t, \lambda^{(z)}, \lambda^{(s)}) \equiv P(s_t|h_t, \lambda^{(z)})P(h_t|\lambda^{(s)})^\alpha \]

Joint Density Model
using parallel data for vowels

from the Gesture Model

Smaller \( \alpha \) Larger

difficult-to-form Gestures easy-to-form
articulate Trajectory flat
Conversion considering delta features

Gesture

Ignoring the correlation of the feature vectors causes the discontinuity of estimated sequences

Conversion considering delta features

(’07 Toda et al.)

MIVC considering delta features

(’11 Saito et al.)

Apply this to our S2H systems

\[ \hat{h} = \arg\max_h P(S|H, \lambda^{(Z)}) P(H|\lambda^{(G)}) \]

\[ S_t = [s_t^\top, \Delta s_t^\top]^\top \]

here,

\[ H_t = [h_t^\top, \Delta h_t^\top]^\top \]
The experimental conditions of the gesture model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>CyberGlove Sensor data (16 dim) # PCA is performed</td>
</tr>
<tr>
<td>Sampling period / Shift length</td>
<td>10～20 ms (variable)</td>
</tr>
<tr>
<td>training sets</td>
<td>16 gestures and transitions between each pair (16 + 16P2 = 256)</td>
</tr>
<tr>
<td># of training sets</td>
<td>1</td>
</tr>
<tr>
<td># of mixtures of GMM</td>
<td>static: 64, static+delta: 256</td>
</tr>
</tbody>
</table>
The experimental conditions of the joint density model

<table>
<thead>
<tr>
<th></th>
<th>Gesture</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature</strong></td>
<td>CyberGlove Sensor data (16 deg)</td>
<td>STRAIGHT-based cepstrum (0-15 deg)</td>
</tr>
<tr>
<td></td>
<td># PCA is performed</td>
<td></td>
</tr>
<tr>
<td><strong>Sampling period</strong></td>
<td>10~20 ms (variable)</td>
<td>8 ms</td>
</tr>
<tr>
<td><strong>Shift length</strong></td>
<td></td>
<td># 1 of every 8 data is chosen</td>
</tr>
<tr>
<td><strong>Window</strong></td>
<td>------</td>
<td>Hamming</td>
</tr>
<tr>
<td><strong>training sets</strong></td>
<td>The 5 isolated vowels and the ( 5P_2 ) transitions for each permutation of two vowels (( 5 + 5P_2 = 25 ) in total)</td>
<td></td>
</tr>
<tr>
<td><strong># of training sets</strong></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong># of mixtures of GMM</strong></td>
<td>static: 8, static+delta: 16</td>
<td></td>
</tr>
</tbody>
</table>
The effect of alpha

alpha = 0
conventional VC

When alpha is small, outputs are out of range

As alpha gets larger, transitions become flat

Appropriate value is 0.6

Re-synthesized /na/
The effect of delta

- **Static**
- **Static+Delta**

<table>
<thead>
<tr>
<th>Time [ms]</th>
<th>Sensor output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>400</td>
<td>240</td>
</tr>
<tr>
<td>800</td>
<td>140</td>
</tr>
</tbody>
</table>

Estimated gesture for /na/

- **Static**
- **Static+Delta**

/na/ /au/
How to make the H2S system

Model-Integration VC (MIVC)

Gesture Model

Joint Model

S2H-delta
alpha = 0.6

parallel data for consonants

Speech

Gesture

H2S system

parallel data for vowels

conventional VC
Previous S2H vs Improved S2H

Time [ms]

Sensor output

- static+delta
- static

65
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
The experimental conditions of the joint density model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gesture</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature</strong></td>
<td>CyberGlove Sensor data</td>
<td>STRAIGHT-based cepstrum</td>
</tr>
<tr>
<td></td>
<td>(16 deg)</td>
<td>(0-15 deg)</td>
</tr>
<tr>
<td></td>
<td># PCA is performed</td>
<td></td>
</tr>
<tr>
<td><strong>Sampling period/Shift length</strong></td>
<td>10～20 ms (variable)</td>
<td>8 ms</td>
</tr>
<tr>
<td></td>
<td></td>
<td># 1 of every 8 data is chosen</td>
</tr>
<tr>
<td><strong>Window</strong></td>
<td>------</td>
<td>Hamming</td>
</tr>
<tr>
<td><strong>training sets</strong></td>
<td>The 5 isolated vowels and the 5P_2 transitions for each permutation of two vowels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 sets of /na/, /ni/, /nu/, /ne/, /no/</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5 + 5P_2 + 5 x 10 = 75 in total)</td>
<td></td>
</tr>
<tr>
<td><strong># of mixtures of GMM</strong></td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
As with the Theremin, pitch / volume control parts of the system are designed...
Synthesized speech for /na/
Overview

❖ Introduction
❖ GMM-based media conversion
  ◆ Fundamentals of Speech Processing
  ◆ Estimation of a mapping function between two spaces
  ◆ Hand-to-Speech system for the five Japanese vowels
❖ Nasal Generation
  ◆ Japanese consonant classification
  ◆ Gesture design based on Speech-to-Hand system
  ◆ Evaluation of proposed designs
  ◆ Subjective evaluation
❖ **Real-time Hand-to-Speech system**
  ◆ The problem of real-time H2S system
  ◆ The effect of alpha & delta
  ◆ The improved real-time H2S system
  ◆ **Subjective user evaluations**
❖ Wrap-up
Listening tests

- Speed of samples are about 1.2 morae/sec. Considering the average speed of finger alphabet (2 morae/sec), samples in double speed are prepared as well as those in normal speed, to mitigate the influence of the speaking speed on impression.
- 10 samples below are randomized and shown 6 Japanese native speakers.
- They are asked to answer (1) what did the speaker say (2) what do you think about the emotion of the speaker.
- They are informed that the samples are Japanese words used in daily life.

<table>
<thead>
<tr>
<th>Intention</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>nani1</strong> “What”, when you do not understand and want to ask the speaker</td>
<td>x2</td>
</tr>
<tr>
<td><strong>nani2</strong> “What”, when you are upset with the speaker</td>
<td>O</td>
</tr>
<tr>
<td><strong>iina1</strong> “OK”, as the back-channel feedback</td>
<td>O</td>
</tr>
<tr>
<td><strong>iina2</strong> “Jealous of you”, when you are jealous of the speaker</td>
<td></td>
</tr>
<tr>
<td><strong>iie</strong> “No”, without any emotion</td>
<td></td>
</tr>
</tbody>
</table>
Phoneme-based intelligibility

- **Nearly half of the phonemes** generated by our H2S system were perceived properly (Nasals by Yabu’s device: 65 %)
- Speaking rate does not influence the intelligibility a lot.
- The most common error was to add /n/ at the beginning of the word.
- The second most common one was to drop /i/.

Prolonged sound and duplication of /n/ are ignored, e.g. 「んあ(na)」 「なー(na-)」 「んな(nna)」 are considered as /na/.
Synthesized speech for /aiueo/

Re-synthesized generated by H2S

Estimated gesture for /n/

Because gestures for /i/ and /n/ were very similar, their spectrums were treated very similarly in the system, although energy of /i/ is larger than that of /n/.
Emotions estimated from the generated speech

<table>
<thead>
<tr>
<th>samples</th>
<th>Expected to be</th>
<th>Impression</th>
</tr>
</thead>
<tbody>
<tr>
<td>nani1</td>
<td>asking</td>
<td>asking(2), confirming, <strong>disappointed</strong>, bad temper</td>
</tr>
<tr>
<td>nani1 x 2</td>
<td>asking</td>
<td><strong>asking</strong>(2), blunt, bored, small anger</td>
</tr>
<tr>
<td>nani2</td>
<td>anger</td>
<td><strong>surprised</strong>(3), <strong>sleepy</strong>(3), happy</td>
</tr>
<tr>
<td>nani2 x 2</td>
<td></td>
<td><strong>surprised</strong>(2), <strong>anger</strong>(2), happy</td>
</tr>
<tr>
<td>iina1</td>
<td>back-channel feedback</td>
<td><strong>sleepy</strong>(2), <strong>at ease</strong>, anger, jealous</td>
</tr>
<tr>
<td>iina1 x 2</td>
<td></td>
<td><strong>sleepy</strong>, sad</td>
</tr>
<tr>
<td>iina2</td>
<td>jealous</td>
<td><strong>sleepy</strong>(2), <strong>exhausted</strong>, sad</td>
</tr>
<tr>
<td>iina2 x 2</td>
<td></td>
<td>hurry, anger</td>
</tr>
<tr>
<td>iie</td>
<td>(no emotion)</td>
<td><strong>sleepy</strong>, <strong>at ease</strong>, cheerful, jealous</td>
</tr>
<tr>
<td>iie x 2</td>
<td></td>
<td>sad</td>
</tr>
</tbody>
</table>

- As intelligibility gets higher and as the speaking speed gets closer to that which people hear in daily conversation, the emotion of the speaker is more easily perceived.
Overview

- Introduction
- GMM-based media conversion
  - Fundamentals of Speech Processing
  - Estimation of a mapping function between two spaces
  - Hand-to-Speech system for the five Japanese vowels
- Nasal Generation
  - Japanese consonant classification
  - Gesture design based on Speech-to-Hand system
  - Evaluation of proposed designs
  - Subjective evaluation
- Real-time Hand-to-Speech system
  - The problem of real-time H2S system
  - The effect of alpha & delta
  - The improved real-time H2S system
  - Subjective user evaluations
- Wrap-up
Wrap-up

- We implemented a speech synthesizer from the media which does not have an explicit relationship with acoustic space, hand gestures, based on space mapping.

- We proposed one framework to derive the gestures for consonants when the correspondence for vowels is given.

- Listeners evaluated an H2S system, which exploits gesture data for consonants derived from an S2H system, can generate more natural sounds than those trained with quasi-optimal gesture design chosen from gesture candidates.

- Subjective user evaluations showed that almost a half of the phonemes, which are generated by our H2S system are perceived correctly and this system is effective enough to generate emotional speech.
Thank you very much

Any questions?