Individual-basis pronunciation clustering of World Englishes using speech structure analysis

Nobuaki Minematsu
Graduate School of Engineering, The University of Tokyo
Japanese learning and English learning

Learning of a local language
To communicate to native speakers of that language.
Goal of pronunciation training
Intelligible enough pronunciation to native speakers
= Native-sounding pronunciation

Learning of the international language
To communicate among speakers of different native languages.
The total number of English users is 1.5 billions.
Goal of pronunciation training
Intelligible enough pronunciation to UK and US speakers?
= Native-sounding pronunciation?
Many countries adopt English as official language.
A quarter of the 1.5 billions speakers use English as official language.
They speak differently accented English from UK or US accent.
World Englishes

**English as only the common language for everybody**

- Diversity of localized pronunciations of English [Crystal’95]
  - US pron. and UK pron. are just two major examples of World Englishes.
- What is the unit of diversity of World English pronunciation?
  - The minimal unit should be **individual!!** (#speakers = 1.5 billions)
- No English is correct and no English is incorrect.
  - One will be interested in how his pronunciation is compared to the others.

[Expanding circle]

- English as foreign language (750 millions)
- English as second language or official language (400 M)
- English as native language (350 M)

[Kachru 1992]
It is something like “face”.

- Europe, Africa, Asia, Oceania, America, ...
- White, yellow, black, ...

No face is correct and no face is incorrect.

The face functions as identifier of that person.
Yet another approach to CAPT

Computer Aided Pronunciation Training

- Application of ASR (Automatic Speech Recognition) technology
  - To detect pronunciation errors in a given utterance.
  - To calculate pronunciation proficiency score for a speaker.
- Comparison bet. teachers’ model and a learner’s utterance
  - Teachers’ model = native speakers’ model
- Which English should be the target model?
  - US or UK?
  - Meaningful question for learning English?

Yet another approach to CAPT system development

- Location of a learner’s pronunciation in the diversity of World Englishes pronunciations
- Comparison of a learner’s pronunciation to all the others.
  - 1 speaker ↔ 1.5 billions - 1 speakers
How is one’s pron. compared to others’?
How is one’s pron. compared to others’?
How is one’s pron. compared to others’?
How is one’s pron. compared to others’?
Outline of presentation

Actual condition of English (World Englishes)
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

A technical challenge and it’s solution
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

Individual-basis pronunciation clustering
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

Possible applications
- Close and distant speakers finder / World Englishes browser

Conclusions
Outline of presentation

Actual condition of English (World Englishes)
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

A technical challenge and it's solution
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

Individual-basis pronunciation clustering
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

Possible applications
- Close and distant speakers finder / World Englishes browser

Conclusions
Outline of presentation

Actual condition of English (World Englishes)
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

A technical challenge and it’s solution
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

Individual-basis pronunciation clustering
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

Possible applications
- Close and distant speakers finder / World Englishes browser

Conclusions
What is needed for clustering?
What is needed for clustering?

Dendrogram
What is needed for clustering?

Dendrogram

Multi-dimensional Scaling
What is needed for clustering?

- Multi-dimensional Scaling
- Dendrogram
### Clustering English pronunciations

The dataset can be represented as:

$$
\begin{bmatrix}
1 & 2 & \cdots & N \\
1 & d_{11} & d_{12} & \cdots & d_{1N} \\
2 & d_{21} & d_{22} & \cdots & d_{2N} \\
3 & d_{31} & & \ddots & \vdots \\
N & d_{N1} & d_{N2} & \cdots & d_{NN}
\end{bmatrix}
$$

The total number of pairwise distances can be calculated as:

$$| = (500 \times 500 - 500) / 2 = 124,750$$
Clustering English pronunciations

\[ | = \frac{(500 \times 500 - 500)}{2} = 124,750 \]
Clustering English pronunciations

1  2  ...  N
\[ d_{11} \quad d_{12} \quad \ldots \quad d_{1N} \]

\[ d_{21} \quad d_{22} \quad \ldots \quad d_{2N} \]

\[ d_{31} \quad : \quad \vdots \]

\[ d_{N1} \quad d_{N2} \quad \ldots \quad d_{NN} \]

How to calculate pron. distance automatically?

\[ | = \frac{(500 \times 500 - 500)}{2} = 124,750 \]
A big and technical challenge

Pronunciation distance and acoustic distance

To which Minematsu’s natural pronunciation closer?

“Those answers will be straightforward if you think them through carefully first.”

Removal of speaker identity related to age, gender, etc.

Pronunciation “skeleton” is needed for comparison.

Only the “accent” aspect of pronunciation is needed for comp.

Speech structure analysis [Minematsu’06]

Speaker-invariant representation of speech or pronunciation
A big and technical challenge

Pronunciation distance and acoustic distance

To which Minematsu’s natural pronunciation closer?

A          X          B

"Those answers will be straightforward if you think them through carefully first."

Removal of speaker identity related to age, gender, etc.

Pronunciation “skeleton” is needed for comparison.

Only the “accent” aspect of pronunciation is needed for comp.

Speech structure analysis [Minematsu’06]

Speaker-invariant representation of speech or pronunciation
A big and technical challenge

Pronunciation distance and acoustic distance
To which Minematsu’s natural pronunciation closer?

“Those answers will be straightforward if you think them through carefully first.”

Removal of speaker identity related to age, gender, etc.
- Pronunciation “skeleton” is needed for comparison.
- Only the “accent” aspect of pronunciation is needed for comp.

Speech structure analysis [Minematsu’06]
- Speaker-invariant representation of speech or pronunciation
A big and technical challenge

Pronunciation distance and acoustic distance
To which Minematsu’s natural pronunciation closer?

Removal of speaker identity related to age, gender, etc.
Pronunciation “skeleton” is needed for comparison.
Only the “accent” aspect of pronunciation is needed for comp.

Speech structure analysis [Minematsu’06]
Speaker-invariant representation of speech or pronunciation

“Those answers will be straightforward if you think them through carefully first.”
Speaker variability = acoustic space deformation

Speaker-invariance = deformation (transform)-invariance

Are there any features that are invariant against any deformation?

Complete transform-invariance measure: f-divergence [Qiao’10]

Every event has to be represented as distribution, not as point.

\[ f_{\text{div}}(p_1, p_2) = \int p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right) dx \quad g : (0, \infty) \to \mathbb{R} \quad \text{and} \quad g(1) = 0 \]

\[ f_{\text{div}}(p_1, p_2) \equiv f_{\text{div}}(P_1, P_2) \quad \rightarrow \text{sufficiency} \]

If \[ \int M(p_1(x), p_2(x)) dx \] is invariant, \[ M = p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right) \]

Speech contrasts (edges) based on f-div. are invariant.
**Invariance in variability**

**Speaker variability = acoustic space deformation**

**Speaker-invariance = deformation (transform)-invariance**

Are there any features that are invariant against any deformation?

**Complete transform-invariance measure: f-divergence [Qiao’10]**

Every event has to be represented as distribution, not as point.

\[ f_{\text{div}}(p_1, p_2) = \int p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right) dx \quad g : (0, \infty) \rightarrow \mathbb{R} \text{ and } g(1) = 0 \]

\[ f_{\text{div}}(p_1, p_2) \equiv f_{\text{div}}(P_1, P_2) \rightarrow \text{sufficiency} \]

If \( \int M(p_1(x), p_2(x))dx \) is invariant, \( M = p_2(x)g \left( \frac{p_1(x)}{p_2(x)} \right) \rightarrow \text{necessity} \)

Speech contrasts (edges) based on \( f \)-div. are invariant.
Invariance in variability

Speaker variability = acoustic space deformation

Speaker-invariance = deformation (transform)-invariance

- Are there any features that are invariant against any deformation?

Complete transform-invariance measure: $f$-divergence [Qiao’10]

- Every event has to be represented as a distribution, not as a point.
- $f_{div}(p_1, p_2) = \int_{\mathcal{X}} p(x) g \left( \frac{p_1(x)}{p_2(x)} \right) dx$ \quad $g : (0, \infty) \rightarrow \mathbb{R}$ and $g(1) = 0$
- $f_{div}(p_1, p_2) \equiv f_{div}(P_1, P_2)$ \quad $\rightarrow$ sufficiency
- If $\int_{\mathcal{X}} M(p_1(x), p_2(x)) dx$ is invariant, $M = p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right)$ \quad $\rightarrow$ necessity
- Speech contrasts (edges) based on $f$-div. are invariant.
Speaker variability = acoustic space deformation
Speaker-invariance = deformation (transform)-invariance
Are there any features that are invariant against any deformation?

Complete transform-invariance measure: $f$-divergence [Qiao’10]

- Every event has to be represented as distribution, not as point.
- $f_{div}(p_1, p_2) = \int p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right) dx$ \quad $g : (0, \infty) \rightarrow R$ and $g(1) = 0$
- $f_{div}(p_1, p_2) \equiv f_{div}(P_1, P_2)$ \quad $\rightarrow$ sufficiency
- If $\int M(p_1(x), p_2(x)) dx$ is invariant, $M = p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right)$ \quad $\rightarrow$ necessity
- Speech contrasts (edges) based on $f$-div. are invariant.
**Definition of $f$-divergence**

$$f_{\text{div}}(p_1, p_2) = \int p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right) dx$$

**Examples of $f$-divergence**

- $g(t) = t \log(t) \quad \rightarrow \quad f_{\text{div}} = \text{KL divergence}$.
- $g(t) = (\sqrt{t} - 1)^2 \quad \rightarrow \quad f_{\text{div}} = \text{Hellinger distance}$.
- $g(t) = \sqrt{t} \quad \rightarrow \quad -\log(f_{\text{div}}) = \text{Bhattacharyya distance}$. 
Several examples of $f$-div.

**Definition of $f$-divergence**

$$ f_{\text{div}}(p_1, p_2) = \int p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right) dx $$

**Examples of $f$-divergence**

- $g(t) = t \log(t)$  \quad $\rightarrow$  \quad $f_{\text{div}} = \text{KL} - \text{divergence}$
- $g(t) = (\sqrt{t} - 1)^2$  \quad $\rightarrow$  \quad $f_{\text{div}} = \text{Hellinger distance}$
- $g(t) = \sqrt{t}$  \quad $\rightarrow$  \quad $- \log(f_{\text{div}}) = \text{Bhattacharyya distance}$
Several examples of $f$-div.

**Definition of $f$-divergence**

$$f_{\text{div}}(p_1, p_2) = \int p_2(x) g \left( \frac{p_1(x)}{p_2(x)} \right) dx$$

**Examples of $f$-divergence**

- $g(t) = t \log(t) \quad \rightarrow \quad f_{\text{div}} = \text{KL divergence.}$
- $g(t) = (\sqrt{t} - 1)^2 \quad \rightarrow \quad f_{\text{div}} = \text{Hellinger distance}$
- $g(t) = \sqrt{t} \quad \rightarrow \quad -\log(f_{\text{div}}) = \text{Bhattacharyya distance.}$
Invariance against deformation

Topological invariance

Each event changes but the system or organization does not change.
Invariance against deformation

Topological invariance

Each event changes but the system or organization does not change.
Invariance against deformation

**Topological invariance**

- Each event changes but the system or organization does not change.
Classical theory in phonology

Theory of relational (topological) invariance

Proposed by R. Jakobson

Distinctive features are used to explain the invariance.

We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.

Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. with their relations to the other sounds.
Theory of relational (topological) invariance
Proposed by R. Jakobson
Distinctive features are used to explain the invariance.

We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.

Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. with their relations to the other sounds.
Classical theory in phonology

Theory of relational (topological) invariance

- Proposed by R. Jakobson
- Distinctive features are used to explain the invariance.

We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.

Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. with their relations to the other sounds.
Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack ........
Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack ..........

A common paragraph to pron. structure

Pron. distance calculation using structure
Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack .........
Clustering English pronunciations

How to calculate pron. distance automatically?

\[ | = \frac{(500 \times 500 - 500)}{2} = 124,750 \]
Outline of presentation

**Actual condition of English (World Englishes)**
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

**A technical challenge and it’s solution**
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

**Individual-basis pronunciation clustering**
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

**Possible applications**
- Close and distant speakers finder / World Englishes browser

**Conclusions**
Outline of presentation

Actual condition of English (World Englishes)
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

A technical challenge and its solution
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

Individual-basis pronunciation clustering
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

Possible applications
- Close and distant speakers finder / World Englishes browser

Conclusions
Speech Accent Archive

Real WE data with IPA transcription [Weinberger’13]

http://accent.gmu.edu
Everybody's read speech of the common paragraph

The paragraph is designed on AE phoneme coverage.

Everybody can join the data collection project of SAA.

#speakers is about 1.8 thousands.

IPA narrow transcriptions are provided (#symbols = 153).

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.
Everybody’s read speech of the common paragraph

The paragraph is designed on AE phoneme coverage.

Everybody can join the data collection project of SAA.

#speakers is about 1.8 thousands.

IPA narrow transcriptions are provided (#symbols = 153).

Please call Stella. Ask her to bring these things with her from the store:
Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.
Everybody’s read speech of the common paragraph

The paragraph is designed on AE phoneme coverage.

Everybody can join the data collection project of SAA.

#speakers is about 1.8 thousands.

IPA narrow transcriptions are provided (#symbols = 153).

Please call Stella. Ask her to bring these things with her from the store:
Six spoons of fresh snow peas, five thick slabs of blue cheese, and
maybe a snack for her brother Bob. We also need a small plastic snake
and a big toy frog for the kids. She can scoop these things into three red
bags, and we will go meet her Wednesday at the train station.
Everybody’s read speech of the common paragraph

The paragraph is designed on AE phoneme coverage.

Everybody can join the data collection project of SAA.

#speakers is about 1.8 thousands.

IPA narrow transcriptions are provided (#symbols = 153).

Please call Stella. Ask her to bring these things with her from the store:
Six spoons of fresh snow peas, five thick slabs of blue cheese, and
maybe a snack for her brother Bob. We also need a small plastic snake
and a big toy frog for the kids. She can scoop these things into three red
bags, and we will go meet her Wednesday at the train station.
Supervised learning of a distance predictor

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(d_{11})</td>
<td>(d_{12})</td>
<td>...</td>
<td>(d_{1N})</td>
</tr>
<tr>
<td>2</td>
<td>(d_{21})</td>
<td>(d_{22})</td>
<td>...</td>
<td>(d_{2N})</td>
</tr>
<tr>
<td>3</td>
<td>(d_{31})</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>(d_{N1})</td>
<td>(d_{N2})</td>
<td>...</td>
<td>(d_{NN})</td>
</tr>
</tbody>
</table>
Supervised learning of a distance predictor

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$d_{11}$</td>
<td>$d_{12}$</td>
<td>...</td>
<td>$d_{1N}$</td>
</tr>
<tr>
<td>2</td>
<td>$d_{21}$</td>
<td>$d_{22}$</td>
<td>...</td>
<td>$d_{2N}$</td>
</tr>
<tr>
<td>3</td>
<td>$d_{31}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>$d_{N1}$</td>
<td>$d_{N2}$</td>
<td>...</td>
<td>$d_{NN}$</td>
</tr>
</tbody>
</table>
## Supervised Learning of a Distance Predictor

### N Speakers

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{N1}$</td>
<td>$d_{11}$</td>
<td>$d_{21}$</td>
<td>$d_{31}$</td>
<td>$d_{41}$</td>
</tr>
<tr>
<td>$d_{N2}$</td>
<td>$d_{12}$</td>
<td>$d_{22}$</td>
<td>$d_{32}$</td>
<td>$d_{42}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$d_{NN}$</td>
<td>$d_{1N}$</td>
<td>$d_{2N}$</td>
<td>$d_{3N}$</td>
<td>$d_{4N}$</td>
</tr>
</tbody>
</table>
Supervised learning of a distance predictor

### N speakers

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$d_{11}$</td>
<td>$d_{12}$</td>
<td>$d_{1N}$</td>
</tr>
<tr>
<td>2</td>
<td>$d_{21}$</td>
<td>$d_{22}$</td>
<td>$d_{2N}$</td>
</tr>
<tr>
<td>3</td>
<td>$d_{31}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>$d_{N1}$</td>
<td>$d_{N2}$</td>
<td>$d_{NN}$</td>
</tr>
</tbody>
</table>
Supervised learning of a distance predictor

$N$ speakers

\[
\begin{bmatrix}
1 & d_{11} & d_{12} & \cdots & d_{1N} \\
2 & d_{21} & d_{22} & \cdots & d_{2N} \\
3 & d_{31} & \vdots & \ddots & \vdots \\
N & d_{N1} & d_{N2} & \cdots & d_{NN}
\end{bmatrix}
\]
Supervised learning of a distance predictor

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d1</td>
<td>d2</td>
<td>d3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d11</td>
<td>d12</td>
<td>d13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d21</td>
<td>d22</td>
<td>d23</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d31</td>
<td>d32</td>
<td>d33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N speakers
Supervised learning of a distance predictor

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(d_{11})</td>
<td>(d_{12})</td>
<td>...</td>
<td>(d_{1N})</td>
</tr>
<tr>
<td>2</td>
<td>(d_{21})</td>
<td>(d_{22})</td>
<td>...</td>
<td>(d_{2N})</td>
</tr>
<tr>
<td>3</td>
<td>(d_{31})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>(d_{N1})</td>
<td>(d_{N2})</td>
<td>...</td>
<td>(d_{NN})</td>
</tr>
</tbody>
</table>

N speakers
Supervised learning of a distance predictor

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$d_{11}$</td>
<td>$d_{12}$</td>
<td>...</td>
<td>$d_{1N}$</td>
</tr>
<tr>
<td>2</td>
<td>$d_{21}$</td>
<td>$d_{22}$</td>
<td>...</td>
<td>$d_{2N}$</td>
</tr>
<tr>
<td>3</td>
<td>$d_{31}$</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>$d_{N1}$</td>
<td>$d_{N2}$</td>
<td>...</td>
<td>$d_{NN}$</td>
</tr>
</tbody>
</table>

N speakers
Supervised learning of a distance predictor

\[ \begin{pmatrix} \ p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ p_{31} \\ \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix} \]
Supervised learning of a distance predictor

\[ \{d_{mn}\} \approx \{p_{mn}\} \]
Reference distance bet. two speakers

IPA-based reference distances \( \{d_{mn}\} \)

- Automatic alignment bet. two transcriptions using DTW
  - DTW(=Dynamic Time Warping) minimized the accumulated distortion.

- DTW requires a distance matrix of all the 153 IPA symbols used.
  - 20 productions of each symbol
  - HMM is built for each symbol (SD-HMM)
  - Acoustic distance is obtained from each HMM (phone) pair.
Reference distance between two speakers

IPA-based reference distances \(d_{mn}\)

- Automatic alignment between two transcriptions using DTW
  - DTW (= Dynamic Time Warping) minimized the accumulated distortion.

DTW requires a distance matrix of all the 153 IPA symbols used.

- 20 productions of each symbol
- HMM is built for each symbol (SD-HMM)
- Acoustic distance is obtained from each HMM (phone) pair.
Reference distance bet. two speakers

Selection of speakers from the SAA

- No word-level insertion or deletion in the utterances
- Only 498 speakers are selected out of about 1,800.

\[
\{d_{mn}\} = \text{Fig. 4 The inter-speaker distance information matrix in gray-level image}
\]
IPA-based clustering of AE and GE speakers

Visualization of the distance matrix by its tree diagram

- German speakers are picked up (#speakers = 9)
- The same number of AE speakers are randomly selected.
- GE16 has lived in USA for some years.
Experimental conditions (1/3)

Sample selection from the SAA corpus

The utterances without word-level insertion/deletion were selected.

The utterances with heavy noises were removed.

381 speakers’ data remained (not 498).

#speaker pairs = 381 x 380 / 2 = 72,390

Sorted based on the IPA-based reference inter-speaker distances.

Divided into 2 speaker-pair groups (even / odd numbered groups)

2-fold cross validation

Unit of building a pronunciation structure

1 paragraph → 9 phrases → 9 structures

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.
UBM + spk-adaptation for structure extraction

- 9 phrase HMMs were built using all the data.
- 24MFCC + their Δ, #states/phrase = 3 x #phonemes in that phrase
- UBM-HMM (phrase HMM) is adapted to each speaker.
- MLLLR adaptation (#classes=32)

Structure extraction from the adapted phrase HMM

- BD calculation by a unit of phoneme (3 HMM states)
- The total number of phoneme-to-phoneme contrasts (edges): 2,804
Experimental conditions (3/3)

**Difference matrix** \( \{ D^l \} \)

- Speaker S’s matrices: \( S^1, S^2, ..., S^l, ..., S^9 \)
- Speaker T’s matrices: \( T^1, T^2, ..., T^l, ..., T^9 \)
- Difference matrices:

\[
D_{ij}^l(S_{ij}^l, T_{ij}^l) = \left| \frac{S_{ij}^l - T_{ij}^l}{S_{ij}^l + T_{ij}^l} \right|
\]

\[
D^l = \{ D_{ij}^l \}
\]

**Support Vector Regression**

- \( \epsilon \)-SVR in LIBSVM
- Radial basis kernel function
- Total number of variables: 2,804

\[
D^1, D^2, ..., D^8, D^9 \xrightarrow{\text{Support Vector Regression}} d_{st}
\]
Experimental conditions (3/3)

Difference matrix $\{D^l\}$

- Speaker $S$’s matrices: $S^1, S^2, \ldots, S^l, \ldots, S^9$
- Speaker $T$’s matrices: $T^1, T^2, \ldots, T^l, \ldots, T^9$
- Difference matrices:

$$D^l_{ij}(S^l_{ij}, T^l_{ij}) = \begin{vmatrix} S^l_{ij} - T^l_{ij} \cr S^l_{ij} + T^l_{ij} \end{vmatrix}$$

$$D^l = \{D^l_{ij}\}$$

Support Vector Regression

- $\epsilon$-SVR in LIBSVM
- Radial basis kernel function
- Total number of variables: 2,804

$$D^1, D^2, \ldots, D^8, D^9$$
Results

Correlation between \{d_{st}\} and \{p_{st}\}

\{p_{st}\}

\{d_{st}\}
Comparison to other methods

**Naive automation of reference distance preparation**
- Phoneticians’ transcription of utterances ← Automate!!
- DTW between a pair of IPA transcriptions to calculate $d_{mn}$.

**Phoneme error detection system with phoneme HMMs**
- SAA’s IPA transcripts are converted to AE phoneme transcripts.
- Phoneme acoustic models (#phonemes = 42, cf. 153 in IPA)
  - Monophone HMMs trained with all the data of SAA.
- ASR grammar for phoneme error detection
  - Word-based network grammar built from the SAA corpus
- DTW between a pair of obtained phoneme transcriptions $\rightarrow p_{mn}$
  - Use the phoneme distance matrix obtained from the set of HMMs

---

**Fig. 2** An example of word-based grammar
Correlation between \( \{d_{st}\} \) and \( \{p_{st}\} \)

<table>
<thead>
<tr>
<th></th>
<th>error detection performance</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ideal error detector</td>
<td>100%</td>
<td>0.88</td>
</tr>
<tr>
<td>actual error detector</td>
<td>73.4%</td>
<td>0.31</td>
</tr>
<tr>
<td>structures + SVR</td>
<td>---</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Fig. 8 Correlation of the predicted distances and the reference distances.
Outline of presentation

Actual condition of English (World Englishes)
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

A technical challenge and it’s solution
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

Individual-basis pronunciation clustering
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

Possible applications
- Close and distant speakers finder / World Englishes browser

Conclusions
Outline of presentation

**Actual condition of English (World Englishes)**
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

**A technical challenge and it’s solution**
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

**Individual-basis pronunciation clustering**
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

**Possible applications**
- Close and distant speakers finder / World Englishes browser

**Conclusions**
Close & distant speakers finder

Locating your pronunciation in the diversity of WE

Collecting speech samples from high school students

- School A (Japan), School B (Korea), School C (China), School D (Poland)...

- Easy conversation partners and challenging conversation partners
Close & distant speakers finder

Locating your pronunciation in the diversity of WE

Collecting speech samples from high school students

School A (Japan), School B (Korea), School C (China), School D (Poland)...

Easy conversation partners and challenging conversation partners
Close & distant speakers finder

Locating your pronunciation in the diversity of WE

Collecting speech samples from high school students
- School A (Japan), School B (Korea), School C (China), School D (Poland)...
- Easy conversation partners and challenging conversation partners
Close & distant speakers finder

Locating your pronunciation in the diversity of WE

Collecting speech samples from high school students
  School A (Japan), School B (Korea), School C (China), School D (Poland)...
  Easy conversation partners and challenging conversation partners
World Englishes browser

Linking all the English material on the net to the MAP

Asking all the speakers of the material to read the Stella paragraph.
World Englishes browser

Linking all the English material on the net to the MAP

Asking all the speakers of the material to read the Stella paragraph.
World Englishes browser

Linking all the English material on the net to the MAP
Asking all the speakers of the material to read the Stella paragraph.
Linking all the English material on the net to the MAP
Asking all the speakers of the material to read the Stella paragraph.
Outline of presentation

**Actual condition of English (World Englishes)**
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

**A technical challenge and it’s solution**
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

**Individual-basis pronunciation clustering**
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

**Possible applications**
- Close and distant speakers finder / World Englishes browser

**Conclusions**
Outline of presentation

Actual condition of English (World Englishes)
- Learning of a local lang. and that of the international lang.
- A new approach to develop CAPT systems for learning English

A technical challenge and it’s solution
- To which is Minematsu’s pronunciation closer?
- Speech (pronunciation) structure analysis

Individual-basis pronunciation clustering
- Speech Accent Archive
- IPA-based reference pronunciation distance between two speakers
- Prediction of the reference distance by using structure analysis

Possible applications
- Close and distant speakers finder / World Englishes browser

Conclusions
Thank you. Any questions?
Thank you. Any questions?