Toward Automatic Speech Interpretation

Nara Institute of Science and Technology
Data Science Center, and Graduate School of Science and Technology

Satoshi Nakamura
with
Katsuhito Sudo, Graham Neubig
Sakriani Sakti, Hiroki Tanaka,
Katsuki Chosa, Do Quoc Truong
Speech-to-Speech Translation System

Japanese

「私は学校に行く: Watashi wa Gakko ni iku」

Watashi wa Gakko ni iku

I go to school

Multilingual Speech Recognition

Spoken Language Translation

Multilingual Speech Synthesis

English

I go to school
Speech Translation and Text Translation

Speech Translation
- Translation of spoken languages
- Speech recognition errors
- Translation from source language speech to target language speech (text)
- Short latency for real-time human communication

Translation of Spoken Language
- Object is real-time communication and understanding
- Para-linguistic/non-linguistic information necessary
- Context dependent utterances, non syntactical utterances
- No punctuation
- No upper/lower case
Technical Background around 2000

▶ Corpus-based Approach
  – Statistical modeling and large size training data

▶ Machine Translation
  – Rule based:
    Linguists created translation rules
  – Corpus based:
    • Example-Based
      Automatic extraction of translation rules [M. Nagao 1984 etc.]
    • Statistical MT (Statistical Machine Translation)
      Extract rules statistically based on Noisy Channel Model
      [P. F. Brown, et.al., 1993]
Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Current Project and Data Collection
4. Summary and Future Works
Speech Translation Projects

★ Japan
  – ATR Speech-to-speech Translation (1986-2008)

★ EU
  – EU-Bridge(2012-2014)

★ US
  – DARPA TransTac, Communicator (2006-2010)
  – DARPA GALE(2006-2010)
  – DARPA BOLT(2011-2015)

★ International
  – IWSLT (2004-)
  – U-Star Consortium (2009-)
History of Speech Translation Research in Japan

1986
- Read Speech
  - Syntactically correct
  - Clear utterance
  - Limited domain
    - Ex. “Conference Registration”

1992
- Daily Conversation
  - Standard expression
  - Unclear utterance
  - Limited domain
    - Ex. “Hotel Reservation”

1999
- Wider and Real Domain
  - Wider and real domain
    - “International Travel”
  - Realistic expressions
  - Noisy speech
  - J-E, J-C speech translation

2006
- + More Languages for Translation
  - 21 multilateral text translation

2008
- A-STAR
  - Multilateral translation for 8 Asian languages
  - Network-based S2ST

2010
- C-STAR
  - Multilateral translation for 7 world languages

2011
- IWSLT
  - Evaluation Campaign of S2S technologies

Fundamentals
- Rule-based Technology
  - Hand-made
- Corpus-based Technology
  - Large scale corpus
  - + Machine learning
Mechanism of Speech Translation System

Japanese

「私は学校に行く: Watashi wa Gakko he iku」

Convert Japanese Phoneme sequence
"a", "i", "u", ...

Convert to word sequence
By lexicon and grammar

Convert Japanese word sequence into English word sequence using dictionary

「私は: watashi ha」⇒ “I”
「学校に: Gakko ni」⇒ “to school”
「行く: iku」⇒ “go”

Re-order word sequence
According to English grammar

“I”
“to school”
“go”

Select appropriate waveform to English text from the corpus

Large Scale Japanese Speech Corpora
Large Scale Japanese Text Corpora

Large Scale Parallel Corpora between Japanese and English

Large Scale English Text Corpora
Large Scale English Speech Corpora

I go to school

I go to school
Today I will give a lecture on machine translation.

Divide the sentence into small phrases and translate:

**Today I will give a lecture on machine translation.**

- Score translations with translation model (TM), reordering model (RM), and language model (LM)
Translation Model Creation

- Perform automatic alignment of parallel text
- Extract phrases from the aligned text for translation

| ホテルの受付 (hoteru no) → hotel |
| ホテルの受付 (hoteru no) → the hotel |
| 受付 (uketsuke) → front desk |
| ホテルの受付 → hotel front desk |
| ホテルの受付 → the hotel front desk |
Statistical MT

- Translation Model, Reordering Model, Language Model

Source and target language parallel text corpus
Parameter estimation
Translation model
Phrase substitution
Input text (Source Language)

Reordering model

Target language text corpus
Parameter estimation
Language model
Grammatical correctness

Machine Translation
Decoding

Translation text (Target Language)
Parallel Corpus

Japanese:
“mado wo aketemo iidesuka”

English:
1. may i open the window
2. ok if i open the window
3. can i open the window
4. could we crack the window
5. is it okay if i open the window
6. would you mind if i opened the window
7. is it okay to open the window
8. do you mind if i open the window
9. would it be all right to open the window
10. i’d like to open the window
## ATR BTEC Corpus

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic</strong></td>
<td>12.2%</td>
<td>(7)</td>
</tr>
</tbody>
</table>
| • greet someone  
• ask a question  
• state one’s purpose  
• …            |
| **Trouble** | 12.1%      | (20)  |
| • luggage  
• emergency  
• medicine  
• assistance  
• …            |
| **Shopping** | 10.0%     | (13)  |
| • buy something  
• gather information  
• price  
• wrapping  
• …            |
| **Move**    | 8.4%       | (8)   |
| • transportation  
• buy a ticket  
• rental car  
• trouble  
• …            |
| **Stay**    | 8.2%       | (11)  |
| • make/change a reservation  
• check-in  
• trouble  
• …            |

### Subcategories

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sightseeing</strong></td>
<td>7.7%</td>
<td>(11)</td>
</tr>
<tr>
<td><strong>Restaurant</strong></td>
<td>7.3%</td>
<td>(11)</td>
</tr>
<tr>
<td><strong>Communication</strong></td>
<td>6.4%</td>
<td>(6)</td>
</tr>
<tr>
<td><strong>Airport</strong></td>
<td>5.5%</td>
<td>(14)</td>
</tr>
<tr>
<td><strong>Business</strong></td>
<td>5.3%</td>
<td>(26)</td>
</tr>
<tr>
<td><strong>Contact</strong></td>
<td>4.0%</td>
<td>(6)</td>
</tr>
<tr>
<td><strong>Airplane</strong></td>
<td>3.6%</td>
<td>(11)</td>
</tr>
<tr>
<td><strong>Homestay</strong></td>
<td>2.3%</td>
<td>(11)</td>
</tr>
<tr>
<td><strong>Study Overseas</strong></td>
<td>1.6%</td>
<td>(14)</td>
</tr>
<tr>
<td><strong>Drink</strong></td>
<td>1.3%</td>
<td>(4)</td>
</tr>
<tr>
<td><strong>Exchange</strong></td>
<td>1.2%</td>
<td>(5)</td>
</tr>
<tr>
<td><strong>Snack</strong></td>
<td>1.2%</td>
<td>(4)</td>
</tr>
<tr>
<td><strong>Beauty</strong></td>
<td>0.8%</td>
<td>(5)</td>
</tr>
<tr>
<td><strong>Go Home</strong></td>
<td>0.6%</td>
<td>(4)</td>
</tr>
<tr>
<td><strong>Research</strong></td>
<td>0.1%</td>
<td>(12)</td>
</tr>
</tbody>
</table>
Mechanism of Speech Translation System

Convert Japanese Phoneme sequence "a", "i", "u", ...

Convert to word sequence by lexicon and grammar

Convert Japanese word sequence into English word sequence using dictionary

Re-order word sequence according to English grammar

Select appropriate waveform to English text from the corpus

Large Scale Japanese Speech Corpora

Large Scale Japanese Text Corpora

Large Scale Parallel Corpora between Japanese and English

Large Scale English Text Corpora

Large Scale English Speech Corpora

Digital revolution for under resourced languages in Asia 2019

http://www.naist.jp/
Speech and Language Corpus for ASR

<table>
<thead>
<tr>
<th>Language</th>
<th>Acoustic model</th>
<th>Language model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>4,200 speakers (271 hrs)</td>
<td>852k sentences</td>
</tr>
<tr>
<td>English</td>
<td>532 speakers (202 hrs)</td>
<td>710k sentences</td>
</tr>
<tr>
<td></td>
<td>US, BRT, AUS</td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>536 speakers (249 hrs)</td>
<td>510k sentences</td>
</tr>
<tr>
<td></td>
<td>Beijing, Shanghai, Canton, Taiwan</td>
<td></td>
</tr>
</tbody>
</table>
Speech to Speech Translation

Launched in November 2007
The first network-based STS translation service
“Shabette Hon’yaku”
「しゃべって翻訳」

- Japanese–English
- NTTDoCoMo

VoiceTra

- “VoiceTra” Network-based Speech Translation released on Jul. 2010
- 21 language pair for Text I/O
- 6 language pair for Speech I/O
800k download and 4M access worldwide as of 2011.3.

Japanese, English, Mandarin, Taiwanese Mandarin, German, French, Dutch, Danish, Italian, Spanish, Portuguese, Brazilian Portuguese, Russian, Arabic, Hindi, Indonesian, Malay, Thai, Tagalog, Vietnamese, Korean

※ Language in red can be input/output in voices.
※ There is no text input support for Hindi or Vietnamese.
Performance Improvements

Subjective Evaluation % of ABC

| A | Good |
| B | Fair |
| C | Acceptable |
| D | Nonsense |
| NIL | No Output |

Word Error Rate %

# utterances used for adaptation
Basic Travel Expression Corpus: Parallel Sentences

Japanese

English

Chinese

Korean

New lang.

Parallel sentences

BTEC
Standardization Image

Server A (ex. Japan)  S2S  Server B (ex. Thailand)

- User interface
- User interface standardization
- Parallel corpus, format, lexicon
- Parallel corpus, Speech data, lexcon
- Data transfer (ASR results, MT results etc)
- XML format
- HTTP protocol
- Processing modules

Parallel corpus, Speech data, lexcon
Data transfer (ASR results, MT results etc)
Activity start for standardization of Network-based S2ST at ITU-T SG16

- Session period: October, 2009 to March, 2010
- NICT is the editor for S2ST standardization at ITU-T SG16, WP2 Q21/22

<table>
<thead>
<tr>
<th>Document</th>
<th>Title</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.745</td>
<td>Functional Requirements for Network-based S2ST</td>
<td>- Definition of Network-based S2ST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Functions and service requirements of network-based S2ST</td>
</tr>
<tr>
<td>H.625</td>
<td>Architectural Requirements for Network-based S2ST</td>
<td>- Requirements of S2ST architecture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Definition of interface for Network-based S2ST</td>
</tr>
</tbody>
</table>

- Not only language conversion but also potentially added module like sign language are taken into account:
  S2ST -> Modality conversion
Research Topics at NAIST

Integrating fundamental technologies into the augmented human-communication systems
Recent Progress of ASR after 2000

Traditional Technologies
- Template Matching, Dynamic Programing [Sakoe 71]
- Hidden Markov Modeling, N-Gram Model [Mercer 83, etc]
- Neural Network, TDNN [Waibel 89], LSTM [Hochreiter 97]
- Weighted Finite State Transducer [Mohri 2006]
- Big Training Data, Data Collection through Trial Service

Deep Learning (Hinton visited MSR)
- DNN-HMM [Hinton 2012]
  - Estimate State Posterior Probability by DNN
- Connectionist Temporal Classification [Graves 2013]
  - Predict Phoneme Label every frame
- Listen, Attend, and Spell [Chan 2016]
  - CTC + Attention: End-to-end modeling
Recent Speech Synthesis

▶ Traditional Technologies
  – Formant-based Synthesis, Waveform Concatenation
  – Statistical Speech Synthesis: HTS
    • Speech Synthesis by HMM

▶ Deep Learning
  – WaveNet
    • Waveform Convolution
  – Tacotron
    • End-to-end speech synthesis with character input. Waveform generation by Griffin-Lim
  – Tacotron2:
    • Tacotron + WaveNet
Recent MT progress

▶ Traditional Technologies
  - Rule-based MT：
    Linguists generate translation rules
  - Corpus-based MT:
    • Example-Based: Automatic rule extraction from corpus [M. Nagao84, Sato et.al.,89, Sumita et. al., 91 ]
    • Statistical MT: Statistical Modeling of MT. Extraction of model parameters from corpus and MT based on Noisy Channel Model [P. F. Brown, et.al. 93]
    • Phrase-base SMT
    • Tree-to-string
      – Statistical MT based on Tree Structure

▶ Deep Learning
  - Neural Machine Translation [2014]
    • Combination of Encoder and Decoder by LSTM
  - Attention NMT [2015]
    • Add Attention to encoder and decoder
  - Self Attention NMT [2017]
  - Self attention by multiple heads. Transformer.
Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Speech Translation with Para-linguistic Information
4. Current Project and Data Collection
5. Summary and Future Works
Communication with Translation

- Input: Text, Speech, Video, Gesture
- Output: Text, Speech, Video, Gesture

End-to-end Process

1. Simultaneity, Incremental, Latency,
2. Para/non linguistic information

Source Language

Target Language

Speech

"to o kyo e i ku"

MT results
/I/go/to/Tokyo/

TTS results
"ai go tu tokyo/

Linguistic Information

Paralinguistic Information

Discourse Context

Domain knowledge, Ontology

Realtime Incremental MT Conversion

Dialog Control

Paralinguistic Information

Personality, Prosody

Personality, Prosody

Text

Text

Image PR

Image

PR

Text

Text

Text

Text

Image Syns.

End-to-end Process

Communication

Input:

- Text
- Speech
- Video
- Gesture

Output:

- Text
- Speech
- Video
- Gesture

Linguistic Information

Paralinguistic Information,

Emotion, Style, Personality, Prosody, Gesture

Speech

⇒ Text

ASR

Image

⇒ text

PR

Text

⇒ Speech

TTS

Text

⇒ Image

Image Syns.

Text

⇒ Text

Image

⇒ Text

Text

⇒ Image

Image Syns.
E-J Interpretation Example

(1) The relief workers (2) say (3) they don’t have (4) enough food, water, shelter, and medical supplies (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

(1) 救援担当者は (9) 生きるための (8) 食料を求め (7) 村を荒らし回っている (6) 大量の難民達の (5) 世話をするための (4) 十分な食料や水, 宿泊施設, 医療品が (3) 無いと (2) 言っています.
Problem: Delay (Ear-Voice Span)

こんにちは、駅はどこですか?

Hello, where is the station?
Simultaneous Incremental Speech Interpretation

Hello, the station where is it?

But, this is not easy!
Can We Do the Same in Automatic Speech Interpretation?

Four problems:

- **Segmentation:** When do we start interpretation?
- **Prediction:** Can we predict things that haven't been said?
- **Rewording:** Can we reword sentences to be conducive to simultaneous interpretation?
- **Evaluation:** How do we decide which results are better?
Re-ordering

- Crucial for translation accuracy:

**Normal phrase-based translation:**

こんにちは 駅は どこですか
Hello, where is the station

**Translation with early timing:**

こんにちは 駅は どこですか
Hello, the station where is it
Lexicalized Reordering Model

- Probabilistically models reordering for increased accuracy of translation
- Given current phrase and next phrase:

**Monotone:**

背の高い男  
the tall man

**Swap:**

太郎を訪問した  
visited Taro

**Discontinuous Right:**

私は太郎を訪問した  
I visited Taro

**Discontinuous Left:**

背の高い男を訪問した  
visited the tall man

- “monotone” + “discontinuous right” = “right probability”
First, temporarily choose strings according to method one

Next, if that phrase's **right probability** exceeds a threshold, actually translate the words in the cache

---

**Example (threshold = 0.8):**

```
hello  where      is      the      station
```

- "hello" phrase exists
- "hello where" phrase missing
- wait
- choose "hello"
- right probability is 0.9 > 0.8
- translate "hello"
- "where is" phrase missing
- wait
- choose "where is"
- right probability is 0.6 < 0.8
- do not translate yet
- "the station" utterance ends
- translate "where is the station"

---

Threshold 1.0 = traditional, 0.0 = method one

Fujita, et. al., 2013
Comparison Across Settings

- Delay decreases in all settings
- Better delay/accuracy tradeoff for long sentences, similar languages
Experiments (IWSLT2013)

- Contents: TED Talk (English⇒Japanese)
  - Translation (Caption)
    vs. Interpretation

- Human Interpreter
  Three professionals with different skills

<table>
<thead>
<tr>
<th>Skill Rank</th>
<th># Years of Interpreter Experiences</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>15 years</td>
</tr>
<tr>
<td>A</td>
<td>4 years</td>
</tr>
<tr>
<td>B</td>
<td>1 year</td>
</tr>
</tbody>
</table>
SS2S vs. Human Interpreter Results on TED Talks

- **A Rank (4 Year)**: 4 yr. exp
- **B Rank (1 Year)**: 1 yr. exp.

**Accurate**

**RIBES**

**Fast**

**Dealy (Sec)**

**By Phrase**

**By Sentence**

=A rank: 4 yr. exp
=B rank: 1 yr. exp.

\[\approx\]

B rank human interpreter with 1 year experience
Translation Timing Control by Syntactic Prediction, 2015

- **Syntactic Prediction**
  - Incremental bottom up parsing
  - Feature extraction and syntactic prediction

- **Wait MT output when specific labels appear.**
  - Control MT output timing according reordering

---

<table>
<thead>
<tr>
<th>Incremental parsing and syntactic prediction</th>
<th>in the next 18 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>i'm going to take [NP] (waiting)</td>
<td>i'm going to take you on a journey</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MT results</th>
<th>18 分である</th>
</tr>
</thead>
<tbody>
<tr>
<td>[NP]を行っています</td>
<td>皆さんを旅にお連れします</td>
</tr>
</tbody>
</table>

Sample 1 ,2015

Conventional Automatic Speech Interpretation with Delay to Wait for Speech End (HirofumiSeo-trad.mp4)
Sample 2, 2015

Actual Interpreter
(HirofumiSeo-interpreter.mp4)
Sample 3, 2015

- Proposed Automatic Speech Interpretation (HirofumiSeo-simul.mp4: )
Statistical Translation Frameworks

Symbolic Models

**Phrase-based MT [Koehn+ 03]**

- *he* has a *cold*
- 吾 は 風邪 を 引いている

**Tree-to-String MT [Liu+ 06]**

- S
  - VP
    - PRP, he
    - VBZ, has
    - DET, a
    - NN, cold

Continuous-space (Neural) Models

**Encoder-Decoder [Sutskever+ 14]**

- *he* has a *cold* <s>
- 彼 は 風邪 を 引いて いる

**Attentional [Bahdanau+ 15]**

- P(e_i|F,e_1,...,e_i-1)
Encoder-decoder Model

- Memorize input sentence by LSTM recurrent neural network
- Generate output sentence by LSTM recurrent neural network

This is a machine translation.
Better Memorization of Sentence and Looking-back Mechanism

- Weighted-sum by the attention

This is a machine translation of the sentence: "これは機械翻訳です" (This is a machine translation).
Results (Neubig, et.al, WAT2015)

Confirm what we know: Neural reranking helps automatic evaluation.

Show what we didn't know: Also help manual evaluation.
Wait-k Algorithm


Wait-k Algorithm

Original:ブッシュ 大統領 は プーチン と 会談 する

Transliteration:President Bush meets with Putin

Satoshi Nakamura, NAIST

Wait-k Algorithm

従来法 delay President Bush delay meets with Putin

提案法 delay President Bush meets with Putin

Controllable! Prediction!
Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Current Project and Data Collection
4. Summary and Future Works
Objectives

- Incremental Automatic Speech Interpretation Algorithm
- Corpus Collection
- Evaluation Measure

Duration: 2017-2021, 5 years

Member:

- Leader: Satoshi Nakamura (NAIST) Leader
- Acoustic Signal Processing: Hiroshi Saruwatari (U. Tokyo)
- Speech Recognition: Sakriani Sakti (NAIST), Tatsuya Kawahara (Kyoto U)
- Machine Translation: Katsuhito Sudo, Yuji Matsumoto (NAIST)
- Speech Synthesis: Tomoki Toda (Nagoya U), Shinnosuke Takamichi (U.Tokyo), Sakriani Sakti (NAIST)
- Audio-visual Translation: Shigeo Morishima (Waseda U)
- Cognitive Load Measurement: Hiroki Tanaka (NAIST)
- Corpus Collection: Katsuhito Sudo, Manami Matsuda (NAIST)
Project Overview

Task 1: Incremental Speech Interpretation Algorithm
- Incremental ASR
- Incremental MT
- Incremental TTS
- Extraction of Paralinguistics
- Paralinguistic MT
- Paralinguistic TTS
- Caption Generation

Task 2: Paralinguistic Speech Translation
- Face modeling
- Speaking Face Conversion
- Speaking Face MT

Task 3: Video MT
- Task 3: Video MT

Task 4: Real Time Cognitive Load Measurement by Human Sensing
- 2x 32ch EEG, Gaze, Heart rate

Task 5: Corpus Collection and Prototyping
- Collect 400 hours Data of Japanese and English Speech Interpretation
- Building Prototype of the Incremental Speech Interpretation System
2012-2016
- Source speech: MP4 (TED), MP3 (CNN), PCM
- Interpreter speech: 24bit 48kHz PCM
  - Skill: S (10 years+), A(3 years+), B
  - Some data includes speech of multiple interpreters

<table>
<thead>
<tr>
<th>Translation direction</th>
<th>Domain</th>
<th>Source Speech</th>
<th>Interpreter Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>#files</td>
<td>#hours</td>
</tr>
<tr>
<td>E-&gt;J</td>
<td>TED</td>
<td>74</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>13</td>
<td>0.731</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>87</td>
<td>15.9</td>
</tr>
<tr>
<td>J-&gt;E</td>
<td>TED</td>
<td>60</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>CSJ</td>
<td>31</td>
<td>5.51</td>
</tr>
<tr>
<td></td>
<td>NHK</td>
<td>10</td>
<td>0.304</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>101</td>
<td>17.7</td>
</tr>
</tbody>
</table>
NAIST Interpreter Corpus 2018

As of 2018

- Source speech: MP4 (TED, TEDx), PCM (CSJ)
- Interpreter speech: 16bit 16kHz PCM
  - Skill: S (10 years +), A (3 years +), B
  - For training set. Total 100 hours by the rank A interpreters
  - For test set. Total 24 hours by one from all rank interpreters

<table>
<thead>
<tr>
<th>Translation direction</th>
<th>domain</th>
<th>Source speech</th>
<th>Interpreter speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#files</td>
<td>#hours</td>
<td>#files</td>
</tr>
<tr>
<td>E-&gt;J</td>
<td>TED</td>
<td>302</td>
<td>66.8</td>
</tr>
<tr>
<td></td>
<td>TED (test)</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>318</td>
<td>70.8</td>
</tr>
<tr>
<td>J-&gt;E</td>
<td>CSJ</td>
<td>146</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>TEDx (test)</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>165</td>
<td>37</td>
</tr>
</tbody>
</table>
Books (Japanese version)
Contents

1. History of Automatic Speech Translation Research
2. Automatic Speech Interpretation Technologies
3. Current Project and Data Collection
4. Summary and Future Works
Summary

- **Remarkable progress**
  - By Statistical Machine Translation
  - Deep Neural Network
  - Progress in Speech Translation

- **Automatic Speech Interpretation**
  - Data Collection
  - Develop Algorithms both for Automatic Speech Interpretation and Interpreter Support System

- **Further Research**
  - Para-linguistics/ Multi-modal
  - Context/ Situation Dependency
  - Common Sense and Domain Knowledge
  - Semantics, Discourse Analysis
  - Towards Better Communication
Communication with Translation

Communication
① Simultaneity, Incremental, Latency,
② Para/non linguistic information
Research Focus Up to Now

Emphases Speech Translation
→ Translates speech while preserving emphasis information

(1) Emphasis estimation (ES) systems:
Estimate emphasis information given speech & a corresponding word sequence

(2) Emphasis translation (ET) systems:
Translate estimated emphasis information into another language
Speech Translation Samples

English-Japanese Emphases Translation

English | ASR | MT | TTS | Japanese
---|---|---|---|---
natural | | | |

natural | | |

ET(CRF) | | |

ET(CRF)+pause | | |

natural | | | baseline

natural | | |

ET(CRF) | | |

ET(LSTM) | | |