Recent Advances in Speech Processing and Machine Translation Research at NAIST

Dr. Satoshi Nakamura

Director, Data Science Center,
Professor, Graduate School of Science and Technology,
Nara Institute of Science and Technology
Team Leader, Tourism Information Analytics Team, AIP Center, Riken, Japan
Where is NAIST?

30 mins from Osaka

90 mins from Kansai Airport

30 mins from Kyoto

500 km from Tokyo
Campus

Land area: 139,967 m²
Constructed area: 27,392 m²
Extended area: 99,109 m²
Chronology

Established as a National university
Oct 1991

Graduate School of Biological Sciences
Apr 1993

Graduate School of Materials Science
Apr 1994

Oct 1998

Apr 2011

Apr 2017

Apr 2018

©Satoshi Nakamura, AHC Lab, NAIST, Japan. AIST 2019 @IGDTUW India
New Data Science Program for Master Students

Data science will be the BASIS of all programs
The Organization of DSC

Director
S. Nakamura

Director of Research
K. Funatsu (NAIST & U. Tokyo)

Data Science
S. Nakamura, Augmented Human Communication
S. Kanaya, Systems Biology
K. Ikeda, Mathematical Informatics
Y. Matsumoto, Computational Linguistics
N. Ono, Data Science, Systems Biology
K. Sudo, Augmented Human Communication
Y. Suzuki, Data Analytics, Data Engineering
K. Yasuda, Natural Language Processing
H. Tanaka, Augmented Human Communication
R. Eguchi, Material Informatics
K. Funatsu, Chemo-infomatics
Y. Uraoka, Information Device Science
Y. Ishikawa, Information Device Science
M. Hatanaka, Materials Informatics
T. Miyao, Chemo-infomatics
H. Mori (Systems Microbiology)
A. Muto (Systems Microbiology)
Y. Sakumura (Computational Biology)
K. Kunida (Computational Biology)

Materials Informatics

Bioinformatics

Open Innovation

International Collaboration

H. Nojima (Strategy and Planning)

Underline: main affiliation
Research Topics at AHC-lab

- Speech Translation
- Neural Machine Translation
- Multi-language ASR, TTS
- Machine Speech Chain

Simultaneous Speech Translation Project 2017-2021
Research Topics at AHC-lab

Speech Translation
Neural Machine Translation
Multi-language ASR, TTS
Machine Speech Chain

Augmented Communication

Why don’t you join our lab!

I’m looking for a lab.

Spoken Dialog
Multi-modal Dialog

Goal-oriented Dialog
Non goal-oriented Dialog

Deep Neural Network

Natural Language Processing

Nov. 14 2019 ©Satoshi Nakamura, AHC Lab, NAIST, Japan. AIST 2019 @IGDTUW India
Research Topics at AHC-lab

Brain Analysis
- Incongruity measurement
- Cognitive Load
- EEG Hyper Scanning

Speech Translation
- ANR-CREST: TAPAS
- SST&CBT ECA 2019-2024

Affective Computing
- SST, CBT, Early Dementia

Why don’t you join our lab!

Good Morning!

Spoken Dialog
- Multi-modal Dialog

Goal-oriented Dialog
- Non goal-oriented Dialog

Deep Neural Network

Why don’t you join our lab!

I’m looking for a lab.

Natural Language Processing

Incongruity measurement
- Cognitive Load
- EEG Hyper Scanning

Spoken Dialog
- Multi-modal Dialog

Goal-oriented Dialog
- Non goal-oriented Dialog

Deep Neural Network

I’m looking for a lab.

Speech Translation
- ANR-CREST: TAPAS
- SST&CBT ECA 2019-2024

Affective Computing
- SST, CBT, Early Dementia

Why don’t you join our lab!

Good Morning!
Research Topics at AHC-lab

Speech Translation
Neural Machine Translation
Multi-language ASR, TTS
Machine Speech Chain

Affective Computing
SST, CBT,
Early Dementia

Why don’t you join our lab!
I’m looking for a lab.

Spoken Dialog
Multi-modal Dialog

Goal-oriented Dialog
Non goal-oriented Dialog

Deep Neural Network

Augmented Communication

BrainAnalysis

RIKEN AIP
Tourism Information Analytics Analytics PJ
2017-2026

Data Analytics
Caption Generation

Natural Language Processing

Data Science Center (2017–)

Nov. 14 2019
©Satoshi Nakamura, AHC Lab, NAIST, Japan. AIST 2019 @IGDTUW India
Topics

- Recent advances in speech processing
  - ASR and TTS research
  - Machine Speech Chain unifies ASR and TTS
  - Single speaker to multi-speaker

- Speech Translation
  - Direct speech to text translation by DNN
Recent Progress of ASR

Traditional Technologies
- Template Matching, Dynamic Programming [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
Saon, et al. “English Conversational Telephone Speech Recognition by Humans and Machines”, INTERSPEECH 2017

Table 1: Word error rates on SWB and CH for human transcribers before and after quality checking contrasted with the human WER reported in [1].

<table>
<thead>
<tr>
<th>Transcriber Type</th>
<th>WER SWB</th>
<th>WER CH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriber 1 raw</td>
<td>6.1</td>
<td>8.7</td>
</tr>
<tr>
<td>Transcriber 1 QC</td>
<td>5.6</td>
<td>7.8</td>
</tr>
<tr>
<td>Transcriber 2 raw</td>
<td>5.3</td>
<td>6.9</td>
</tr>
<tr>
<td>Transcriber 2 QC</td>
<td>5.1</td>
<td>6.8</td>
</tr>
<tr>
<td>Transcriber 3 raw</td>
<td>5.7</td>
<td>8.0</td>
</tr>
<tr>
<td>Transcriber 3 QC</td>
<td>5.2</td>
<td>7.6</td>
</tr>
<tr>
<td>Human WER from [1]</td>
<td>5.9</td>
<td>11.3</td>
</tr>
</tbody>
</table>


Table 3: Word error rates for LSTMs, ResNet and frame-level score fusion results across all testsets (36 n-gram LM).

<table>
<thead>
<tr>
<th>Model</th>
<th>SWB</th>
<th>CH</th>
<th>RT’02</th>
<th>RT’03</th>
<th>RT’04</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM (baseline)</td>
<td>7.7</td>
<td>14.0</td>
<td>11.8</td>
<td>11.4</td>
<td>10.8</td>
</tr>
<tr>
<td>LSTM1 (SA-MTL)</td>
<td>7.6</td>
<td>13.6</td>
<td>11.5</td>
<td>10.7</td>
<td></td>
</tr>
<tr>
<td>LSTM2 (Feat. fusion)</td>
<td>7.2</td>
<td>12.7</td>
<td>10.7</td>
<td>10.2</td>
<td>10.1</td>
</tr>
<tr>
<td>ResNet</td>
<td>7.6</td>
<td>14.5</td>
<td>12.2</td>
<td>12.2</td>
<td>11.5</td>
</tr>
<tr>
<td>ResNet+LSTM2</td>
<td>6.8</td>
<td>12.2</td>
<td>10.2</td>
<td>10.0</td>
<td>9.7</td>
</tr>
<tr>
<td>ResNet+LSTM1+LSTM2</td>
<td>6.7</td>
<td>12.1</td>
<td>10.1</td>
<td>10.0</td>
<td>9.7</td>
</tr>
</tbody>
</table>

Table 4: WER on SWB and CH with various LM configurations.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWB</td>
</tr>
<tr>
<td>n-gram</td>
<td>6.7</td>
</tr>
<tr>
<td>n-gram + model-M</td>
<td>6.1</td>
</tr>
<tr>
<td>n-gram + model-M + Word-LSTM</td>
<td>5.6</td>
</tr>
<tr>
<td>n-gram + model-M + Char-LSTM</td>
<td>5.7</td>
</tr>
<tr>
<td>n-gram + model-M + Word-LSTM-MTL</td>
<td>5.6</td>
</tr>
<tr>
<td>n-gram + model-M + Char-LSTM-MTL</td>
<td>5.6</td>
</tr>
<tr>
<td>n-gram + model-M + Word-DCC</td>
<td>5.8</td>
</tr>
<tr>
<td>n-gram + model-M + 4 LSTMs + DCC</td>
<td><strong>5.5</strong></td>
</tr>
</tbody>
</table>
Recent Speech Synthesis

- Formant-based Synthesis, Waveform Concatenation

- Statistical Speech Synthesis: HTS
  - Speech Synthesis by HMM

- WaveNet
  - Waveform Convolution

- Tacotron
  - End-to-end speech synthesis with character input. Waveform generation by Griffin-Lim

- Tacotron2:
  - Tacotron + WaveNet
Outline

- **Machine Speech Chain**
  - *Machine Speech Chain: Listening while speaking*
  - *Speech Chain with One-shot Speaker Adaptation*
    - Andros Tjandra, Sakriani Sakti, Satoshi Nakamura1,2 “Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018

- **End-to-end Speech Translation**
  - *Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation*
Motivation Background

- In human communication
  - A closed-loop speech chain mechanism has a critical auditory feedback mechanism
  - Children who lose their hearing often have difficulty to produce clear speech
FIGURE 1.1 The speech chain: the different forms of a spoken message in its progress from the brain of the speaker to the brain of the listener.
Delayed Auditory Feedback*1,2

▶ DAF:
  – Device that enables a user to speak into a microphone and then hear in headphones a fraction of a second later

▶ Effects by DAF to people who stutter

▶ Effects in normal speakers
  – DAF effects prove about the structure of the auditory and verbal pathways in the brain.
  – Indirect effects include reduction in rate of speech, increase in intensity, and increase in fundamental frequency
  – Direct effects include repetition of syllables, mispronunciations, omissions, and omitted word endings.

*2 Wikipedia “Delayed Auditory Feedback”
**Proposed Method**

→ Develop a closed-loop speech chain model based on deep learning

Not only has the capability to listen and speak, but also listen while speaking
Machine Speech Chain

Definition:
- $x = \text{original speech}$, $y = \text{original text}$
- $\hat{x} = \text{predicted speech}$, $\hat{y} = \text{predicted text}$
- $\text{ASR}(x): x \rightarrow \hat{y}$ (seq2seq model transforms speech to text)
- $\text{TTS}(y): y \rightarrow \hat{x}$ (seq2seq model transforms text to speech)
Given a pair speech-text \((x, y)\)
- Train ASR and TTS in supervised learning
- Directly optimized:
  - \(\text{ASR}\) by minimize \(L_{\text{ASR}}(y, \hat{y})\)
  - \(\text{TTS}\) by minimizing loss between \(L_{\text{TTS}}(x, \hat{x})\)
- Update both ASR and TTS independently
Machine Speech Chain

Case #2: Unsupervised Learning with Text Only

– Given the unlabeled text features $y$

1. TTS generates speech features $\hat{x}$
2. Based on $\hat{x}$, ASR tries to reconstruct text features $\hat{y}$
3. Calculate $\mathcal{L}_{ASR}(y, \hat{y})$ between original text features $y$ and the predicted $\hat{y}$

Possible to improve ASR with text only by the support of TTS
Case #3: Unsupervised Learning with Speech Only

Given the unlabeled speech features $x$

1. ASR predicts the most possible transcription $\hat{y}$
2. Based on $\hat{y}$, TTS tries to reconstruct speech features $\hat{x}$
3. Calculate $L_{TTS}(x, \hat{x})$ between original speech features $x$ and the predicted $\hat{x}$

Possible to improve TTS with speech only by the support of ASR
Sequence-to-Sequence ASR

**Input & output**
- $x = [x_1, ..., x_S]$ (speech feature)
- $y = [y_1, ..., y_T]$ (text)

**Model states**
- $h_{[1..S]}^e$ = encoder states
- $h_t^d$ = decoder state at time $t$
- $a_t$ = attention probability at time $t$
  - $a_t(s) = \text{Align}(h_s^e, h_t^d)$
  - $a_t(s) = \frac{\exp(\text{Score}(h_s^e, h_t^d))}{\sum_{s=1}^{S}\exp(\text{Score}(h_s^e, h_t^d))}$
  - $c_t = \sum_{s=1}^{S} a_t(s) * h_s^e$ (expected context)

**Loss function**
$$L_{ASR}(y, p_y) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{c \in [1..C]} 1(y_t = c) * \log p_{y_t}[c]$$
**Sequence-to-Sequence TTS**

**Input & output**
- $x^R = [x_1, ..., x_S]$ (linear spectrogram feature)
- $x^M = [x_1, ..., x_S]$ (mel spectrogram feature)
- $y = [y_1, ..., y_T]$ (text)

**Model states**
- $h^e_{1..S} = $ encoder states
- $h^d_s = $ decoder state at time $t$
- $a_s = $ attention probability at time $t$
- $c_s = \sum_{s=1}^{S} a_s(t) * h^e_t$ (expected context)

**Loss function**
\[
L_{TTS1}(x, \hat{x}) = \frac{1}{S} \sum_{s=1}^{S} (x^M_s - \hat{x}^M_s)^2 + (x^R_s - \hat{x}^R_s)^2 \\
L_{TTS2}(b, \hat{b}) = -\frac{1}{S} \sum_{s=1}^{S} (b_s \log(\hat{b}_s) + (1 - b_s) \log(1 - \hat{b}_s)) \\
L_{TTS}(x, \hat{x}, b, \hat{b}) = L_{TTS1}(x, \hat{x}) + L_{TTS2}(b, \hat{b})
\]
19: **# Loss combination:**

20: Combine all weighted loss into a single loss variable

\[ L = \alpha \ast (L_P^{TTS} + L_P^{ASR}) + \beta \ast (L_U^{TTS} + L_U^{ASR}) \]  

21: Calculate TTS and ASR parameters gradient with the derivative of \( L \) w.r.t \( \theta_{ASR}, \theta_{TTS} \)

\[ G_{ASR} = \nabla_{\theta_{ASR}} L \]  

\[ G_{TTS} = \nabla_{\theta_{TTS}} L \]  

22: Update TTS and ASR parameters with gradient descent optimization (SGD, Adam, etc)

\[ \theta_{ASR} \leftarrow Optim(\theta_{ASR}, G_{ASR}) \]  

\[ \theta_{TTS} \leftarrow Optim(\theta_{TTS}, G_{TTS}) \]  

23: **until** convergence of parameter \( \theta_{TTS}, \theta_{ASR} \)
Experimental Set-up

Features

- **Speech:**
  - 80 Mel-spectrogram (used by ASR & TTS)
  - 1024-dim linear magnitude spectrogram (SFFT) (used by TTS)
  - TTS reconstruct speech waveform by using Griffin-Lim to predict the phase & inverse STFT

- **Text:**
  - Character-based prediction
    - a-z (26 alphabet)
    - 6 punctuation mark (,,:'?.-)
    - 3 special tags <s> </s> <spc> (start, end, space)
Experiments on Single-speaker

Dataset:
- BTEC corpus (text), speech generated by Google TTS (using gTTS library)
- Supervised training: 10000 utts (text & speech paired)
- Unsupervised training: 40000 utts (text & speech unpaired)

Result:

<table>
<thead>
<tr>
<th>Data</th>
<th>Hyperparameter</th>
<th>ASR</th>
<th>TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>β</td>
<td>gen. mode</td>
</tr>
<tr>
<td>Paired (10k)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+Unpaired (40k)</td>
<td>0.25</td>
<td>1</td>
<td>greedy</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1</td>
<td>beam 5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1</td>
<td>beam 5</td>
</tr>
</tbody>
</table>

Acc: End of speech prediction accuracy
Experiments on Multi-speakers

- **Dataset**
  - BTEC ATR-EDB corpus (text & speech) (25 male, 25 female)
  - Supervised training: 80 utts / spk (text & speech paired)
  - Unsupervised training: 360 utts / spk (text & speech unpaired)

- **Result**

<table>
<thead>
<tr>
<th>Data</th>
<th>Hyperparameter</th>
<th>ASR</th>
<th>TTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>β</td>
<td>gen. mode</td>
</tr>
<tr>
<td>Paired (80 utt/spk)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+Unpaired (remaining)</td>
<td>0.25</td>
<td>1</td>
<td>greedy</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1</td>
<td>greedy</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1</td>
<td>beam 5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1</td>
<td>beam 5</td>
</tr>
</tbody>
</table>

Acc: End of speech prediction accuracy
Speech Chain with One-shot Speaker Adaptation

Andros Tjandra$^{1,2}$, Sakriani Sakti$^{1,2}$, Satoshi Nakamura$^{1,2}$

“Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018
Sequel: Speech Chain with One-shot Speaker Adaptation

Motivation
- Previous model able to improve single-speaker result significantly
- Limitation: couldn’t train on unseen speaker (discrete speaker embedding)

Proposed model
One-shot Speaker Adaptation on TTS

- Instead of using discrete speaker index (one vector for one speaker)
- We generate a vector given a short utterance by using DeepSpeaker (speaker recognition model)
- Take the last layer before softmax as embedding $z$
- Integrate the information with Tacotron’s decoder for generation

Figure 2: Proposed model: sequence-to-sequence TTS (Tacotron) + speaker information via neural speaker embedding (DeepSpeaker).
<table>
<thead>
<tr>
<th>Model</th>
<th>CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised training:</strong></td>
<td></td>
</tr>
<tr>
<td>WSJ train_si84 (16hrs speech, paired) -&gt; Baseline</td>
<td></td>
</tr>
<tr>
<td>Att Enc-Dec</td>
<td>17.35</td>
</tr>
<tr>
<td><strong>Supervised training:</strong></td>
<td></td>
</tr>
<tr>
<td>WSJ train_si284 (66 hrs speech, paired) -&gt; Upperbound</td>
<td></td>
</tr>
<tr>
<td>Att Enc-Dec</td>
<td>7.12</td>
</tr>
<tr>
<td><strong>Semi-supervised training:</strong></td>
<td></td>
</tr>
<tr>
<td>WSJ train_si84 (paired) + train_si200 (unpaired)</td>
<td></td>
</tr>
<tr>
<td>Label propagation (greedy)</td>
<td>17.52</td>
</tr>
<tr>
<td>Label propagation (beam=5)</td>
<td>14.58</td>
</tr>
<tr>
<td><strong>Proposed speech chain</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9.86</td>
</tr>
</tbody>
</table>
Topics

- Recent advances in speech processing
  - ASR and TTS research
  - Machine Speech Chain unifies ASR and TTS
  - Single speaker to multi-speaker

- Speech Translation
  - Direct speech to text translation by DNN
Structure Based Curriculum Learning for End-to-end Direct English-Japanese Speech Translation

Recent MT progress

- 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 [Koen+ 2003]

- Tree-to-string
  - Statistical MT based on Tree Structure

- Neural Machine Translation
  - Combination of Encoder and Decoder by LSTM [Sutskever+ 14]

- Attention NMT [Bahdanau+ 15]
  - Add Attention to encoder and decoder

- Self Attention NMT [Vaswani+ 17]
  - Self attention by multiple heads. Transformer.
Traditional approach in speech-to-speech translation systems

- automatic speech recognition (ASR)
- machine translation (MT)
- text to speech synthesis (TTS)

all of which are independently trained and tuned
Related Works

- **L.Duong et al. NAACL 2016 [1]**
  - Title: An Attentional Model for Speech Translation Without Transcription
  - Spanish to English speech-to-text direct translation with attentional encoder decoder networks

- **Alexandre Berard et al. NIPS workshop 2016 [2]**
  - Title: Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation
  - French to English speech-to-text direct translation with attentional encoder decoder networks
Related Works\textsuperscript{[2]}

- End-to-end Speech-to-text translation with attentional model
Problems

- Their works are only applicable for similar syntax and word order (SVO-SVO) [1,2]
- For such languages, only local movements are sufficient for translation.

Spanish to English translation attention matrix [1]

French to English translation attention matrix [2]
• Syntactically distant language pairs (SVO versus SOV) suffers from long-distance reordering phenomena.

<table>
<thead>
<tr>
<th></th>
<th>朝食</th>
<th>は</th>
<th>いくら</th>
<th>で</th>
<th>す</th>
<th>か</th>
</tr>
</thead>
<tbody>
<tr>
<td>how</td>
<td>0.09</td>
<td>0.037</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>much</td>
<td>0.819</td>
<td>0.828</td>
<td>0.033</td>
<td>0.833</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is</td>
<td>0.018</td>
<td>0.037</td>
<td>0.168</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>0.026</td>
<td>0.038</td>
<td>0.024</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breakfast</td>
<td>0.738</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td>0.08</td>
<td>0.24</td>
<td>0.882</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

English to Japanese translation attention matrix
A first attempt to build direct speech-to-text direct translation system (ST) on syntactically distant language pairs.

To guide the encoder-decoder attentional model to learn this difficult problem, we proposed a structured-based curriculum learning strategy.
Attention-based ST with Curriculum Learning

Phase 1

Train the attentional-based encoder-decoder neural network for a standard ASR and MT task
Attention-based ST with Curriculum Learning

The model’s objective now is to predict the word representation (like the MT encoder’s output)

The model now predicts the corresponding word sequence in the target language given the input speech
We combine the MT attention and decoder modules to perform the speech translation task from the source speech sequence to the target word sequence.
Attention-based ST with Curriculum Learning

Attentional-based neural trained for ASR and text-based MT tasks and gradually train the network for end-to-end ST tasks.
## Experimental Set-up

<table>
<thead>
<tr>
<th>System settings</th>
<th><strong>ASR</strong></th>
<th>Data settings</th>
<th>BTEC Para-text</th>
<th>BTEC Speech</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input units</td>
<td>23</td>
<td>Train utterance</td>
<td>45,000</td>
<td>45,000</td>
<td></td>
</tr>
<tr>
<td>Hidden units</td>
<td>512</td>
<td>Test utterance</td>
<td>500</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Output units</td>
<td>27293</td>
<td>Source Vocabulary</td>
<td>27293</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM layer depth</td>
<td>2</td>
<td>Target Vocabulary &amp; Output size</td>
<td>33155</td>
<td>F-bank 23dim</td>
<td>We use Google TTS system to generate BTEC speech</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Input units &amp; Embed size</td>
<td>12823</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hidden units</td>
<td>512</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM layer Depth</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Optimizer
- Adam
Translation Accuracy

BLEU+1

- Baseline MT
- Baseline ASR+MT
- Direct ST Enc-Dec
- Fast track CL ASRenc-MTdec
- Slow track CL ASRenc-ASRdec-MTdec
Overall Summary:

▶ Recent advances in speech processing

▶ ASR and TTS research at NAIST
  ▶ Machine Speech Chain unifies ASR and TTS
  ▶ Single speaker to multi-speaker

▶ Speech Translation research at NAIST
  ▶ Direct speech to text translation by DNN
    - Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation

▶ Future Works
  - Advanced MT modules by Deep Learning
  - Learn human perception and cognitive process