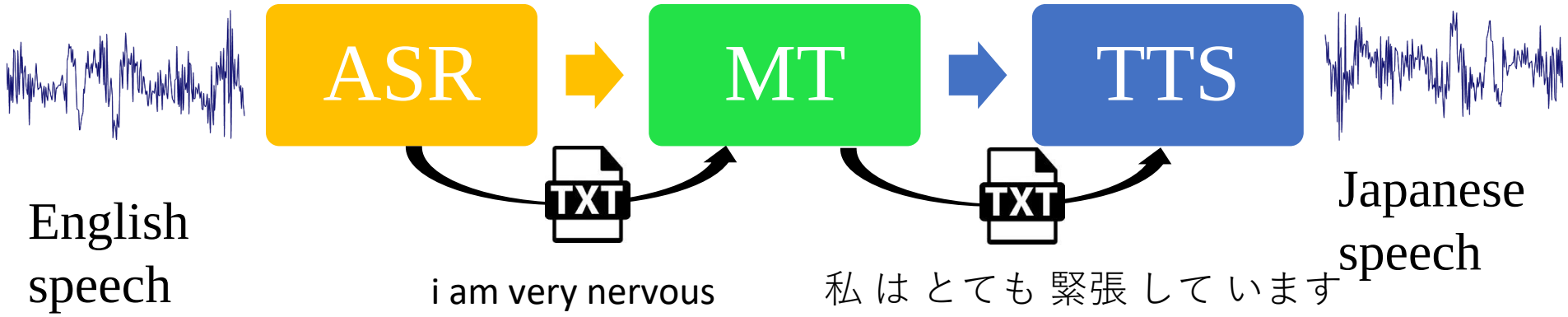
カリキュラムラーニングを用いた音声翻訳の学習戦略の提案

◯叶　高朋, Sakriani Sakti, 中村　哲
奈良先端科学技術大学院大学

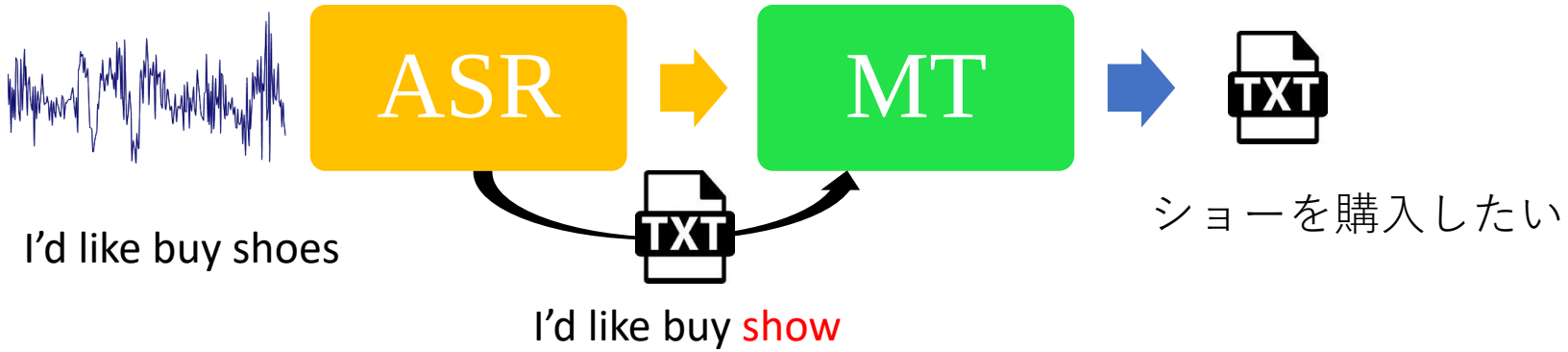
Traditional Speech Translation



Traditional approach in speech-to-speech translation systems

- ✓ construct automatic speech recognition (ASR), machine translation (MT) and text to speech synthesis (TTS)
- ✓ all of which are independently trained and tuned

Traditional Speech Translation



ASR error affect Translation performance.

✓ all of which are independently trained and tuned

✓ NMT module difficult handle input with error word.

Limitations in Traditional Approach

1. Basic unit for information sharing is only words at the text level
   - ✓ Many languages do not have written form

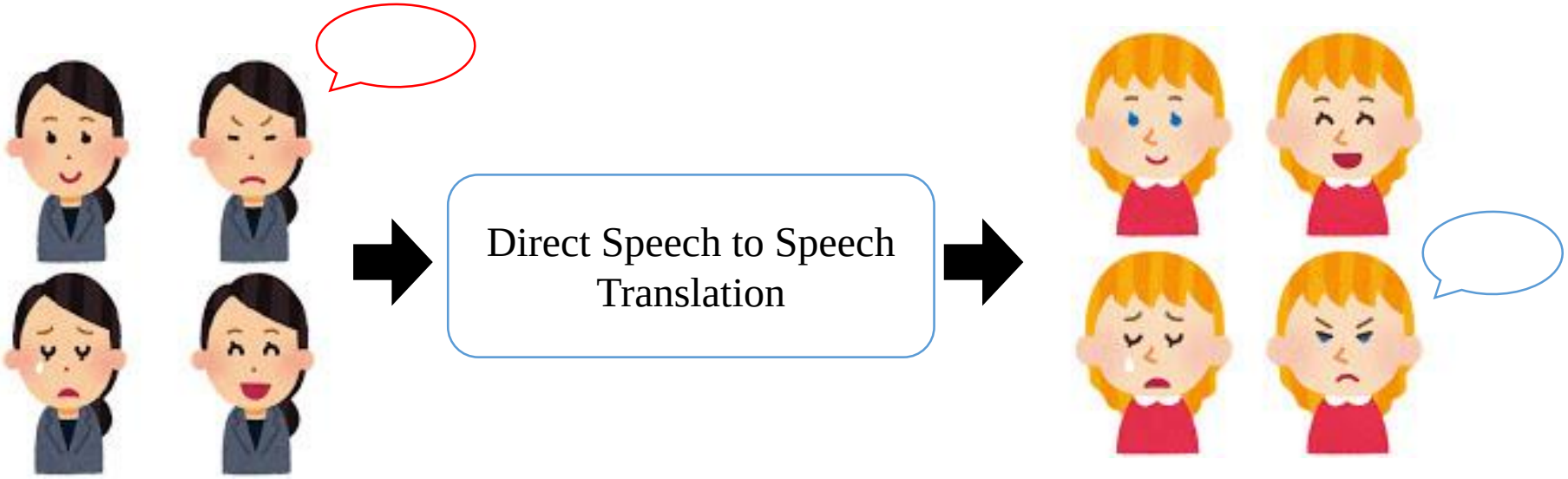
2. Speech acoustics might involve both linguistic and paralinguistic information

ASR have lost all of their paralinguistic information

Paralinguistic information:
- ✓ is not a factor in written communication
- ✓ cannot even be expressed in words

Objective of this Research

“Direct speech-to-speech translation”

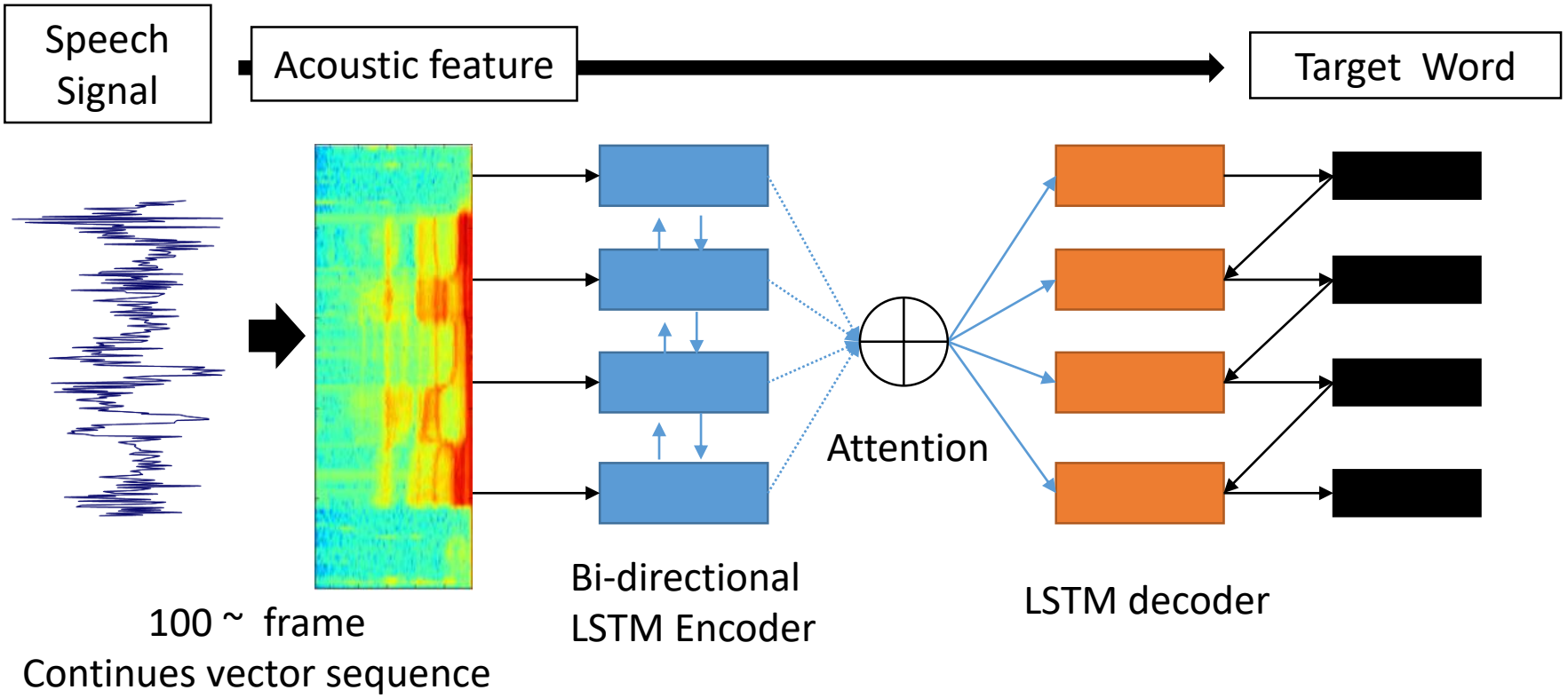


Related Works

- Long 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
- Yoshua Bengio et al. ICML 2009[3]
  - Title: Curriculum Learning
  - A Leaning strategy, learn from easy data to difficult data.

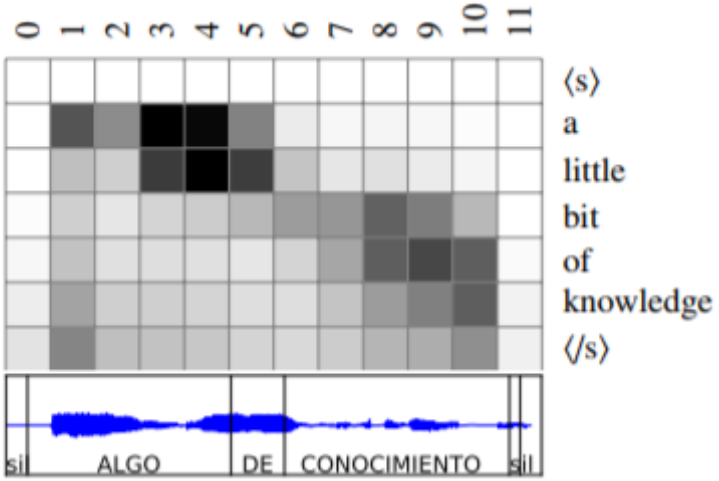
Related Works

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

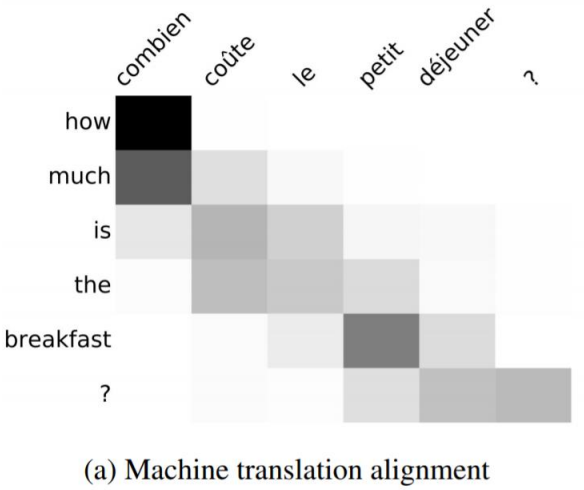


Problems

- 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]

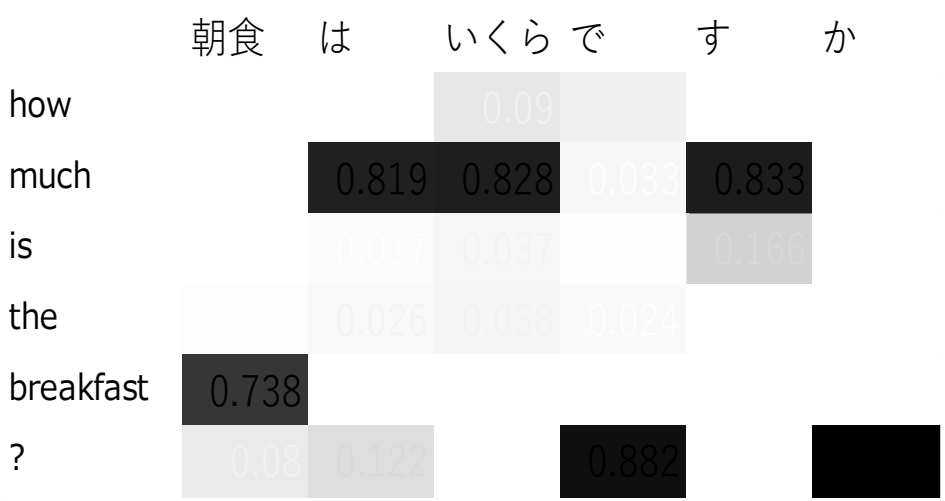


(a) Machine translation alignment

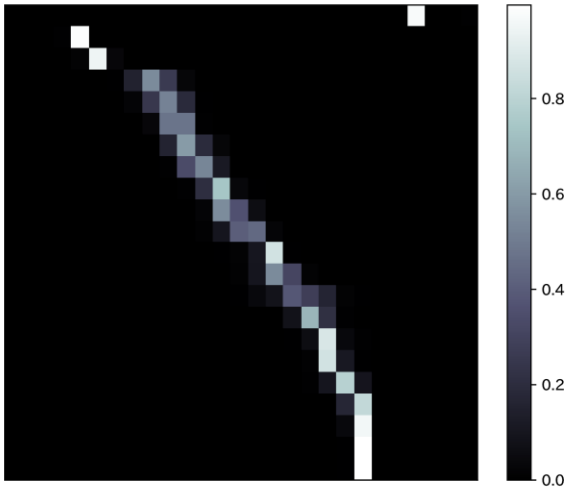
French to English translation attention matrix [2]

Problems

- Syntactically distant language pairs (SVO versus SOV) suffers from long-distance reordering phenomena.



English to Japanese NMT attention matrix



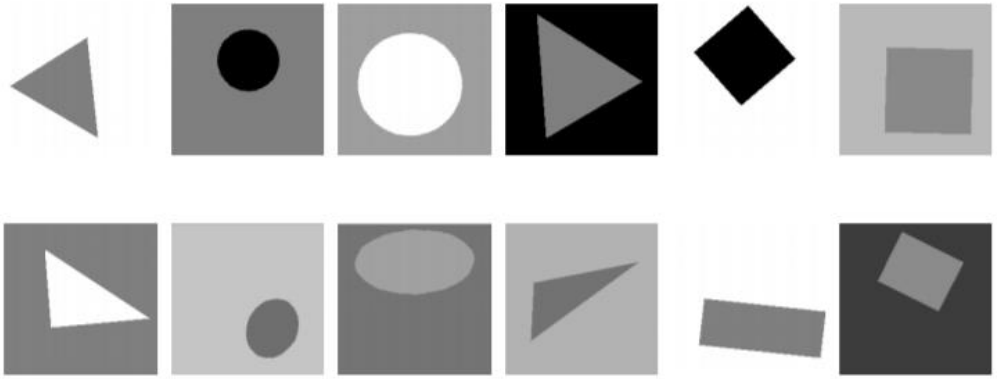
Japanese ASR attention matrix

Proposed method

- A first step we focus Speech to Text direct translation system(ST) on syntactically distant language pairs
- Train attentional model on English-Japanese language pairs with SVO versus SOV word order.
- **To guide the encoder-decoder attentional model to learn this difficult problem**, we proposed a structured-based curriculum learning strategy.

Curriculum Learning

- Curriculum learning [3]
  - One learning paradigm, is inspired by the learning processes of humans and animals that learn from easier aspects and gradually increase to more difficult ones.



Sample inputs from Basic Shape(top)
and Genome Shape(bottom)

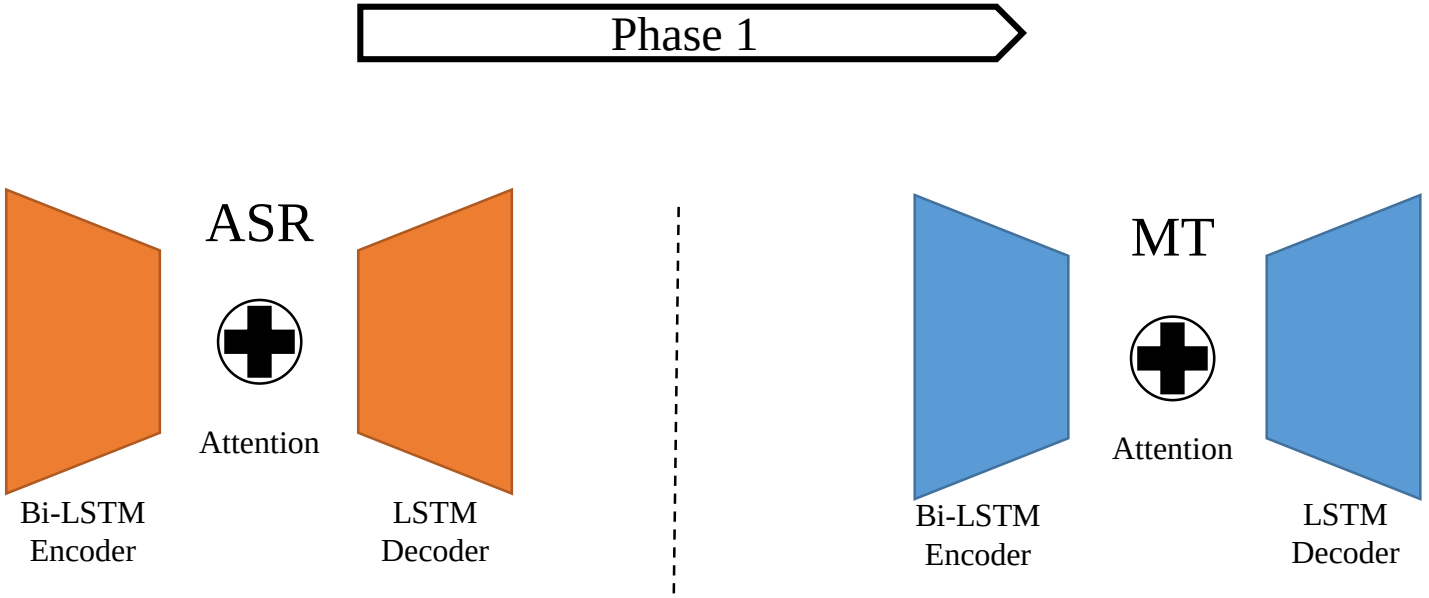
Curriculum Learning

- Original "Curriculum learning"
  - The attentional encoder-decoder architecture trained directly for speech translation tasks using similar but more and more difficult speech translation data

How to correct easy data in translation task? Shorter one? Common word? ...

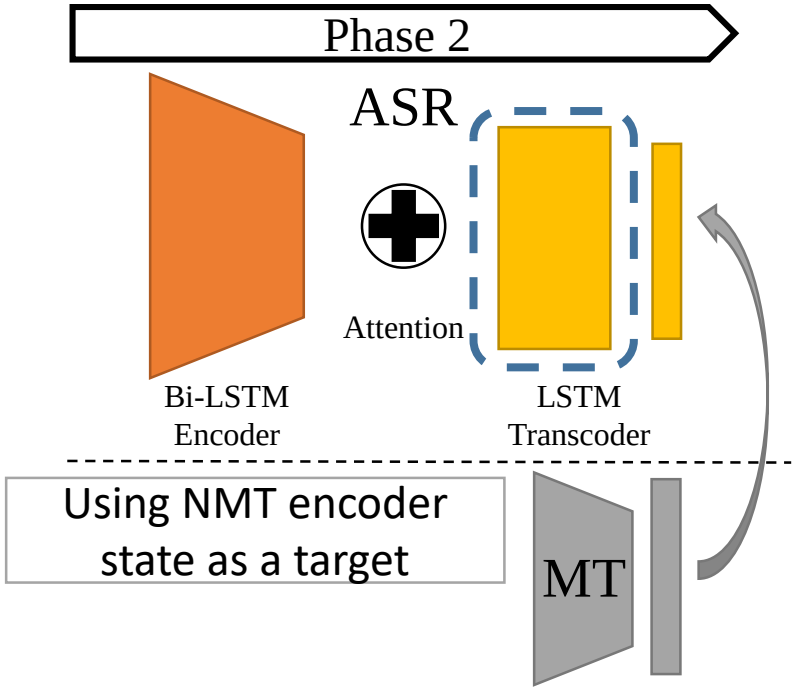
- **Our proposed "Structure Based Curriculum learning"**
  - We train the attentional encoder-decoder architecture by starting **from a simpler task, switch a certain part of the structure in each training phase, and set it to a more difficult target task**.

Attention-based ST with Curriculum Learning



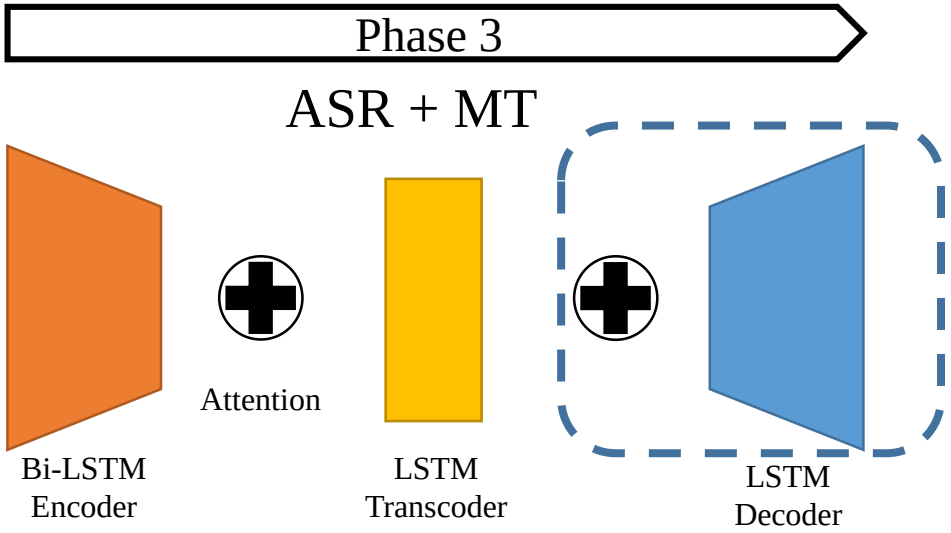
As before, we train the attentional-based encoder-decoder neural network for a standard ASR and MT task

Attention-based ST with Curriculum Learning



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

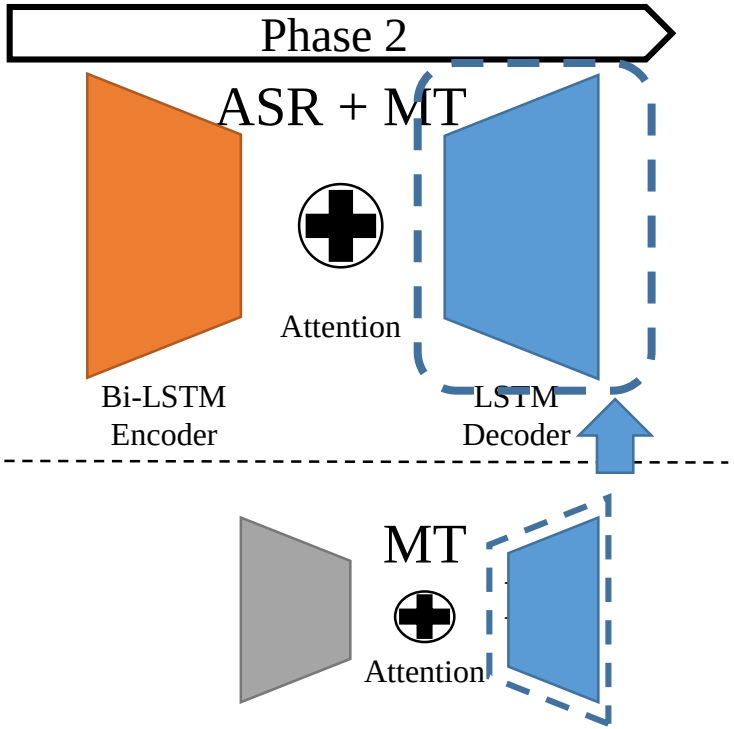
Attention-based ST with Curriculum Learning



Slow track

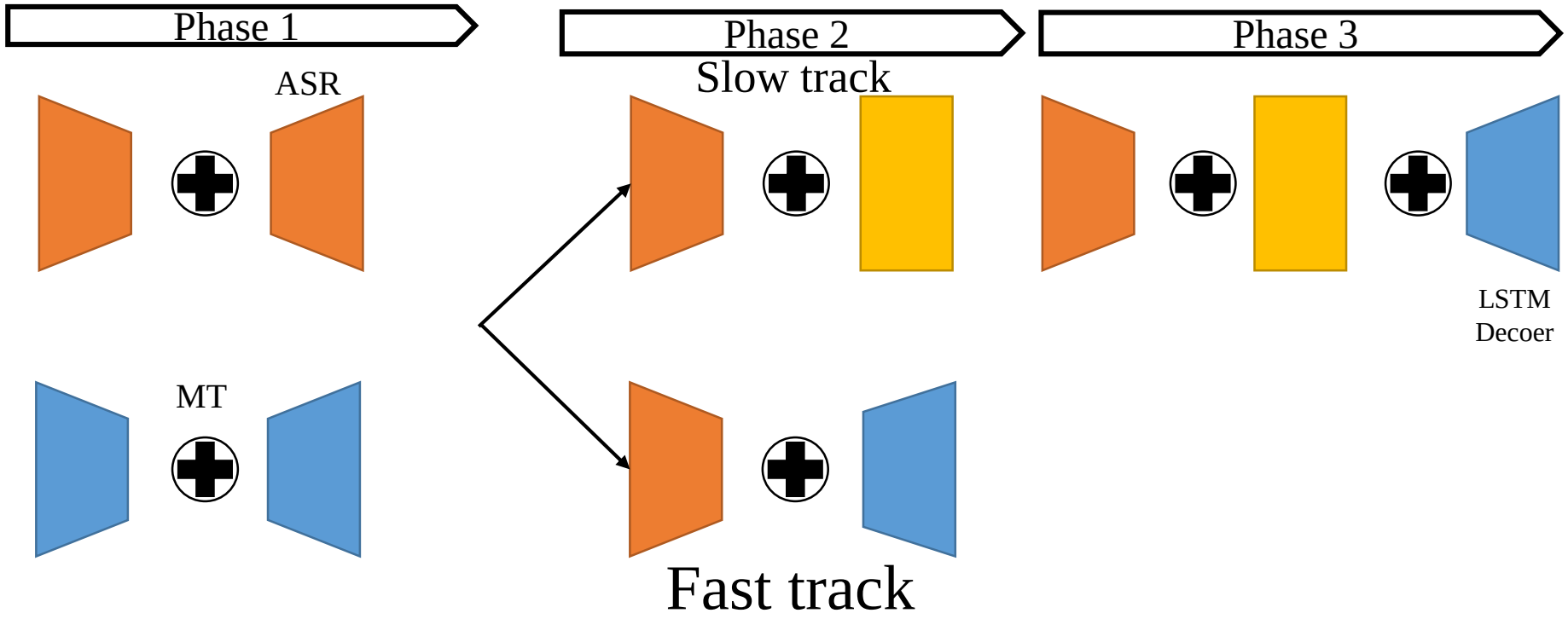
We combine the MT attention and decoder modules to perform the speech translation task from the source speech sequence to the target word sequence

Attention-based ST with Curriculum Learning



The model now predicts the corresponding word sequence in the target language given the input speech

Attention-based ST with Curriculum Learning



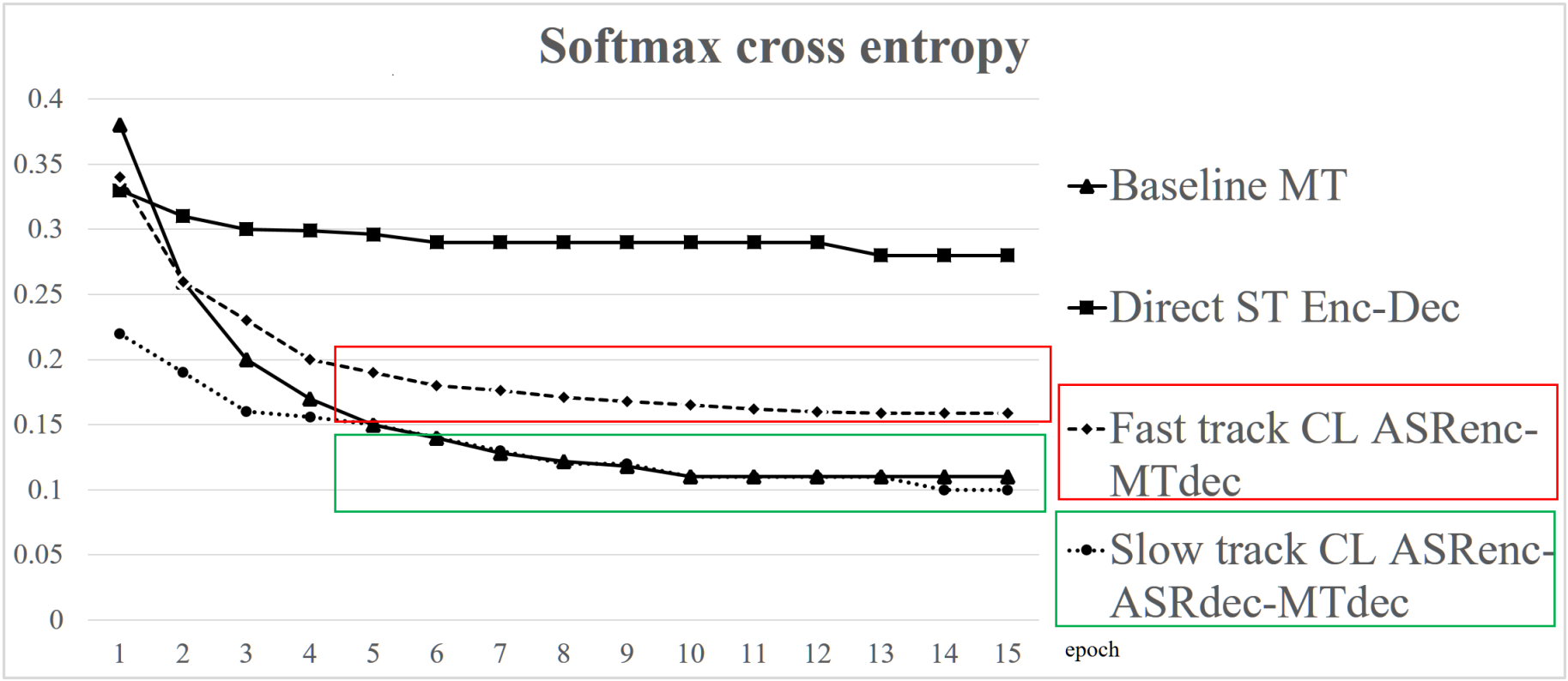
Attentional-based neural trained for ASR and text-based MT tasks and gradually train the network for end-to-end ST tasks.

Experimental Set-up

System settings	
ASR	
Input units	23
Hidden units	512
LSTM layer depth	2,2
MT	
Source Vocabulary	27293
Target Vocabulary & Output size	33155
Input units & Embed size	128
Hidden units	512
LSTM layer Depth	2,2
Optimizer	
Adam	

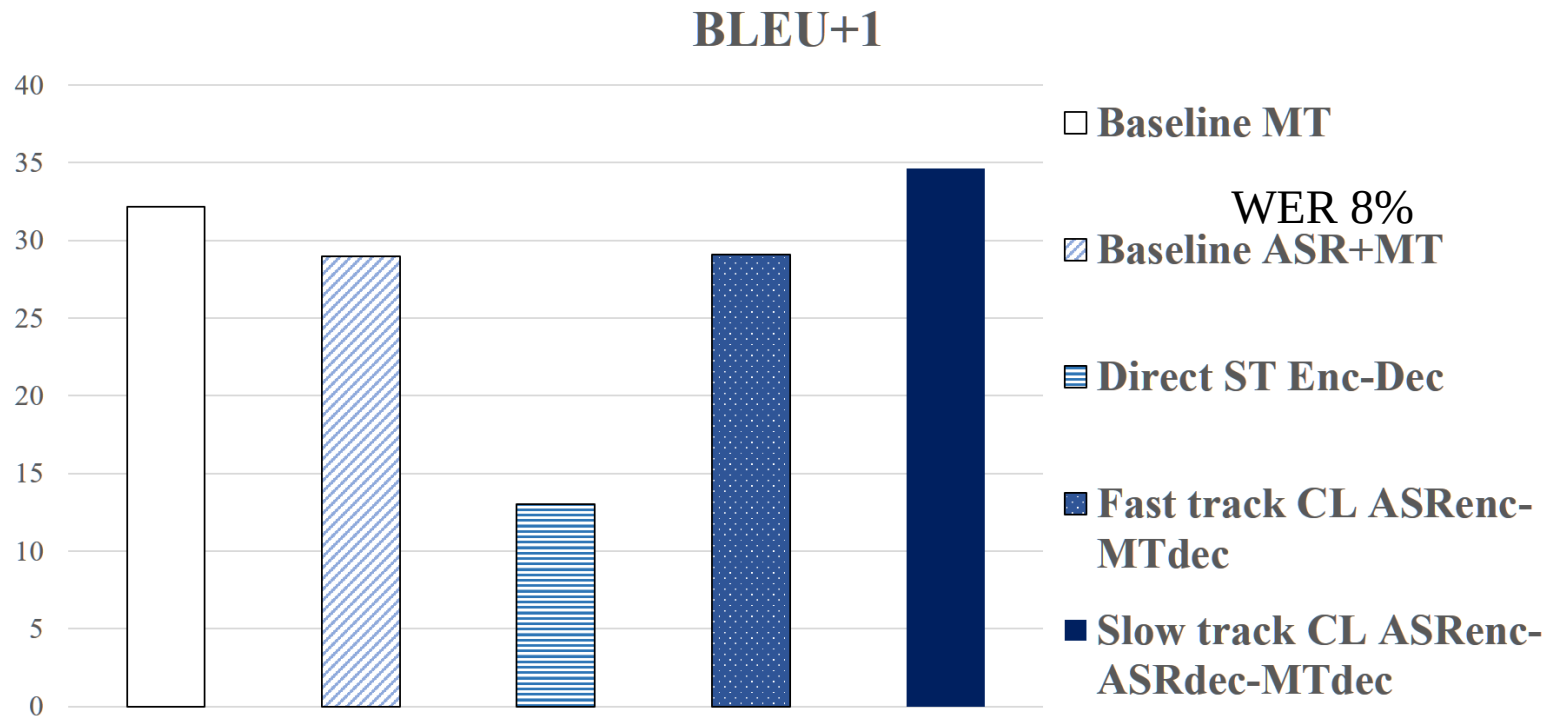
Data settings	
BTEC Para-text	
Train utterance	45,000
Valid utterance	5,000
Test utterance	500
BTEC Speech	
Train utterance	45,000
Test utterance	500
Speech feature	F-bank 23dim
ASR word error rate	512
Other	
We use Google TTS system to generate BTEC speech	

Learning loss for each epoch



- ✓ Using CL-based proposed method, we can further decrease the loss
- ✓ Specifically, the one that trained with CL type 1 – Slow Track successfully outperformed the text-based MT system.

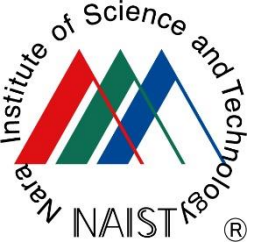
Translation Accuracy



- ✓ Best performance was achieved by proposed Slow Track model
- ✓ Surpassed the text-based MT and cascade ASR+MT systems.

Conclusion

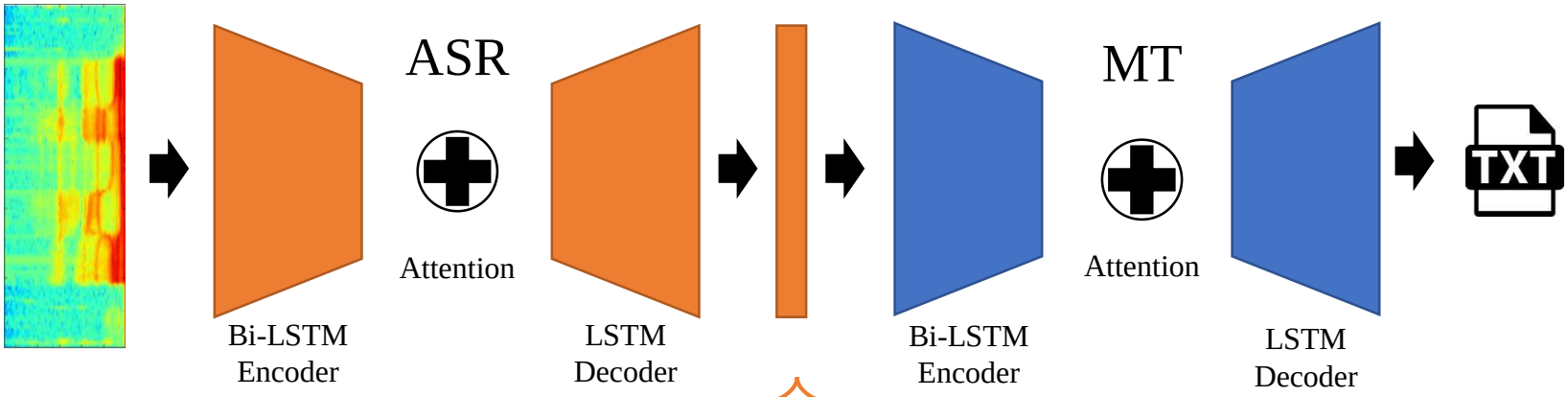
- To using pre-trained model and extend and training deep network are succeed.
- We achieved English-Japanese end-to-end speech to text translation without being affected by ASR error.
- Experimental results demonstrated that the learning model is stable and its translation quality outperformed the standard MT system.
- In future works we intent to consider natural speech, and expand the speech-to-text translation task to a speech-to-speech translation task.

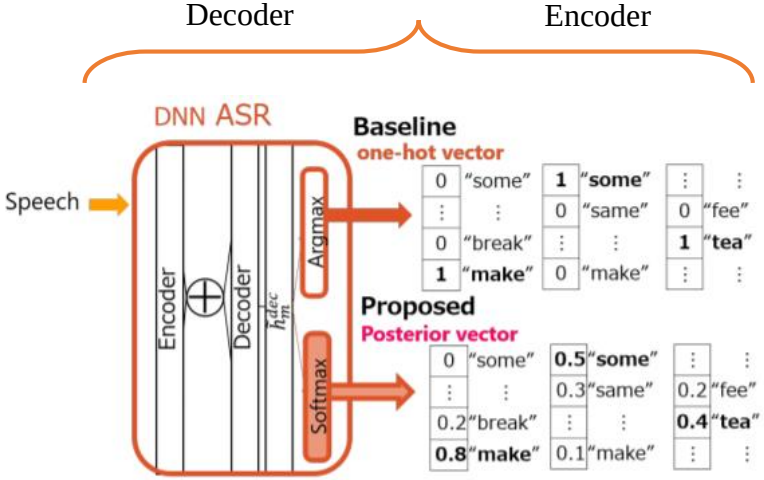


Thank you for your listening.

Additional page

- “Using Spoken Word Posterior Features in Neural Machine Translation” (IWSLT18 S.Sakti et al.)

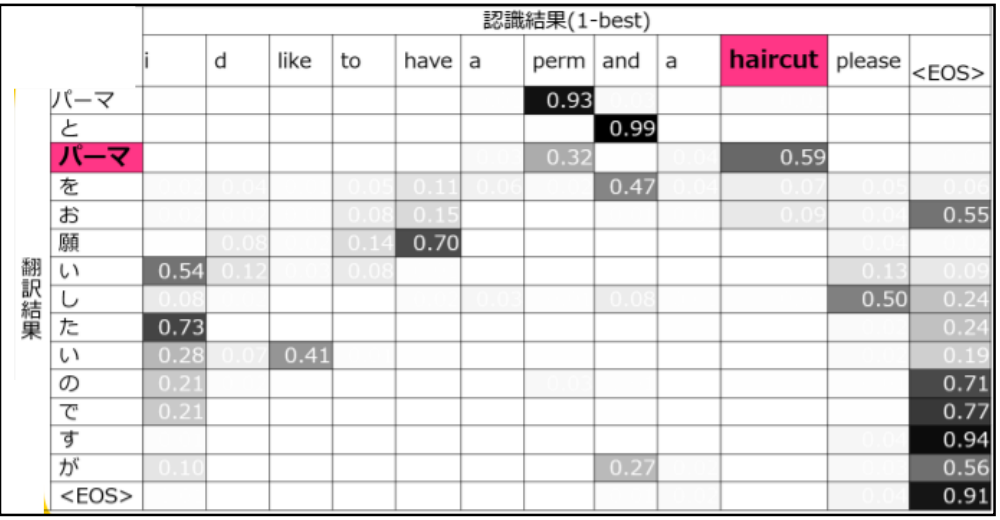




Easy to analysis
Easy to training

Weakly End to End
Huge parameters
ASR and NMT have to use same vocab

Additional page

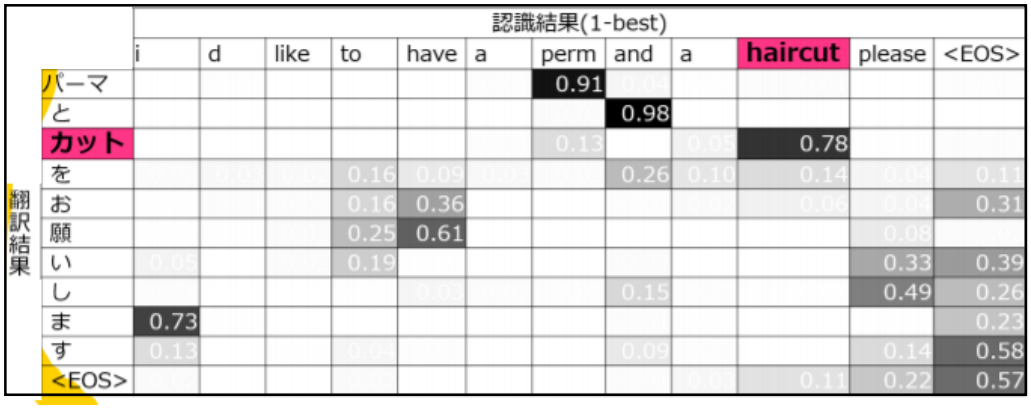


翻訳結果 \ 認識結果(1-best)	i	d	like	to	have	a	perm	and	a	haircut	please	<EOS>
パーマ							0.93					
と								0.99				
パーマ							0.32			0.59		
を								0.47				
お												0.55
願					0.70							
い	0.54											
し											0.50	0.24
た	0.73											0.24
い	0.28		0.41									0.19
の	0.21											0.71
で	0.21											0.77
す												0.94
が								0.27				0.56
<EOS>												0.91

Similar meaning or usage word will embedding similar space.

Some time it makes attention error.

Additional information (pronounce similarity) can fix above problem and get good attention result



翻訳結果 \ 認識結果(1-best)	i	d	like	to	have	a	perm	and	a	haircut	please	<EOS>
パーマ							0.91					
と								0.98				
カット										0.78		
を												
お												0.31
願												
い											0.33	0.39
し											0.49	0.26
ま	0.73											0.23
す												0.58
<EOS>												0.57