

End-to-end approach to ASR, TTS and Speech Translation

Satoshi Nakamura^{1,2}

with Sakriani Sakti^{1,2}, Andros Tjandra^{1,2}, Takatomo Kano, and Quoc Truong Do

¹Nara Institute of Science & Technology, Japan

²RIKEN, Center for Advanced Intelligence Project AIP, Japan



Outline

- *Machine Speech Chain*

- *Machine Speech Chain: Listening while speaking*

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, “Listening while Speaking: Speech Chain by Deep Learning”, ASRU 2017

- *Speech Chain with One-shot Speaker Adaptation*

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura^{1,2} “Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018

- *End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator*

- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. ICASSP, 2019

- *End-to-end Speech-to-speech Translation*

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

- Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation”, INTERSPEECH2017

Outline

- *Machine Speech Chain*

- *Machine Speech Chain: Listening while speaking*

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, “Listening while Speaking: Speech Chain by Deep Learning”, ASRU 2017

- *Speech Chain with One-shot Speaker Adaptation*

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura^{1,2} “Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018

- *End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator*

- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. ICASSP, 2019

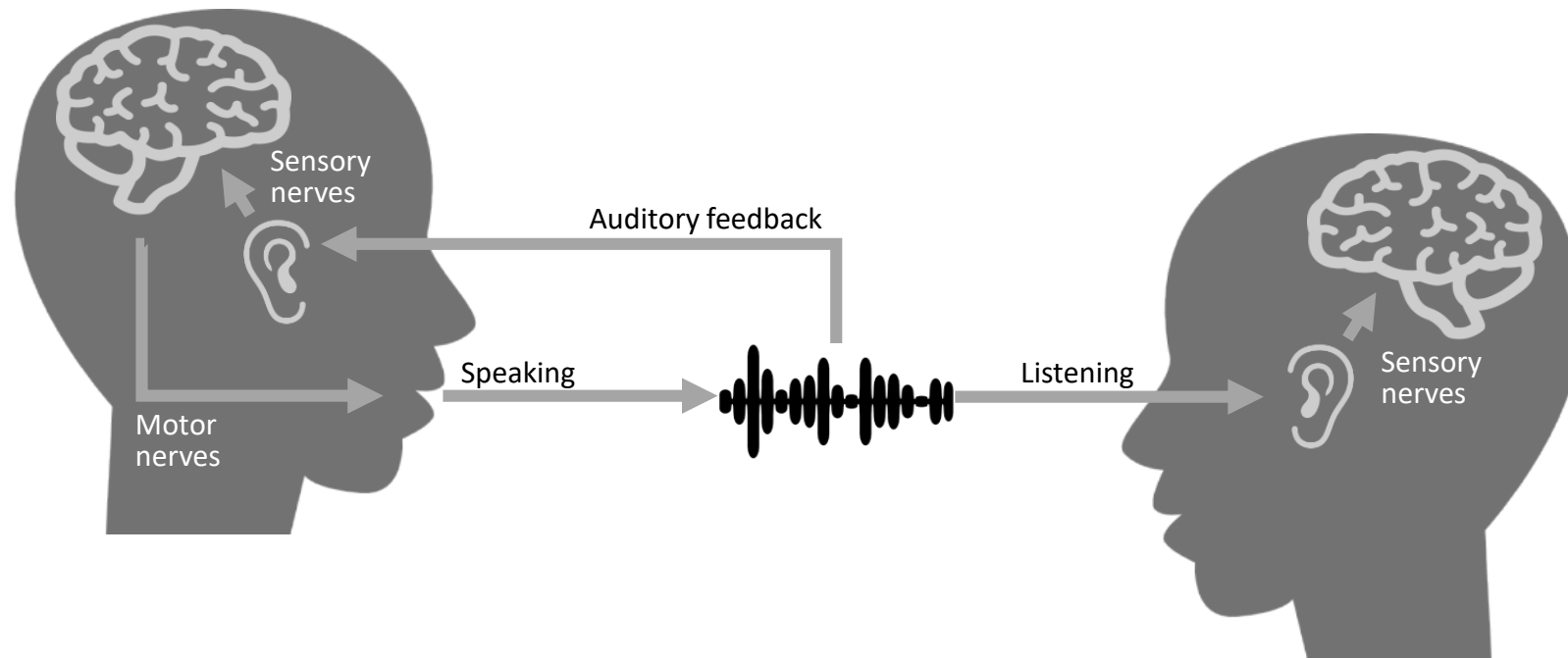
- *End-to-end Speech-to-speech Translation*

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

- Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation”, INTERSPEECH2017

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



Speech Chain: Denes, Pinson 1973

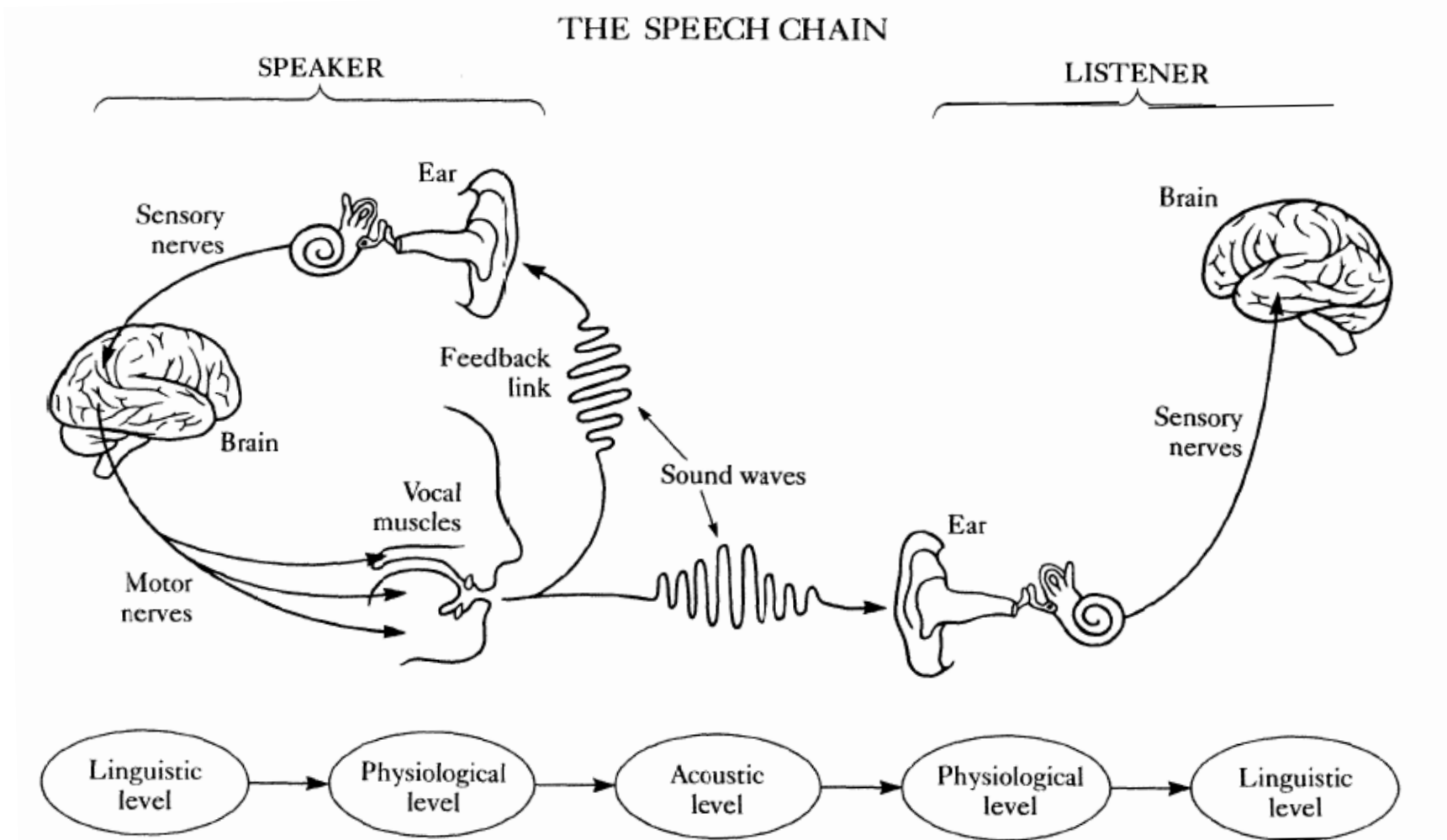


FIGURE 1.1 The speech chain: the different forms of a spoken message in its progress from the brain of the speaker to the brain of the listener.

Delayed Auditory Feedback^{*1,2}

- DAF:
 - It can consist of a device that enables a user to speak into a microphone and then hear his or her voice in headphones a fraction of a second later
- Effects in people who stutter
 - Those who stutter had an abnormal speech–auditory feedback loop that was corrected or bypassed while speaking under DAF.
- Effects in normal speakers
 - DAF in non-stutterers to see what it can prove about the structure of the auditory and verbal pathways in the brain.
 - Indirect effects of delayed auditory feedback in non-stutterers include reduction in rate of speech, increase in intensity, and increase in fundamental frequency in order to overcome the effects of the feedback. Direct effects include repetition of syllables, mispronunciations, omissions, and omitted word endings.

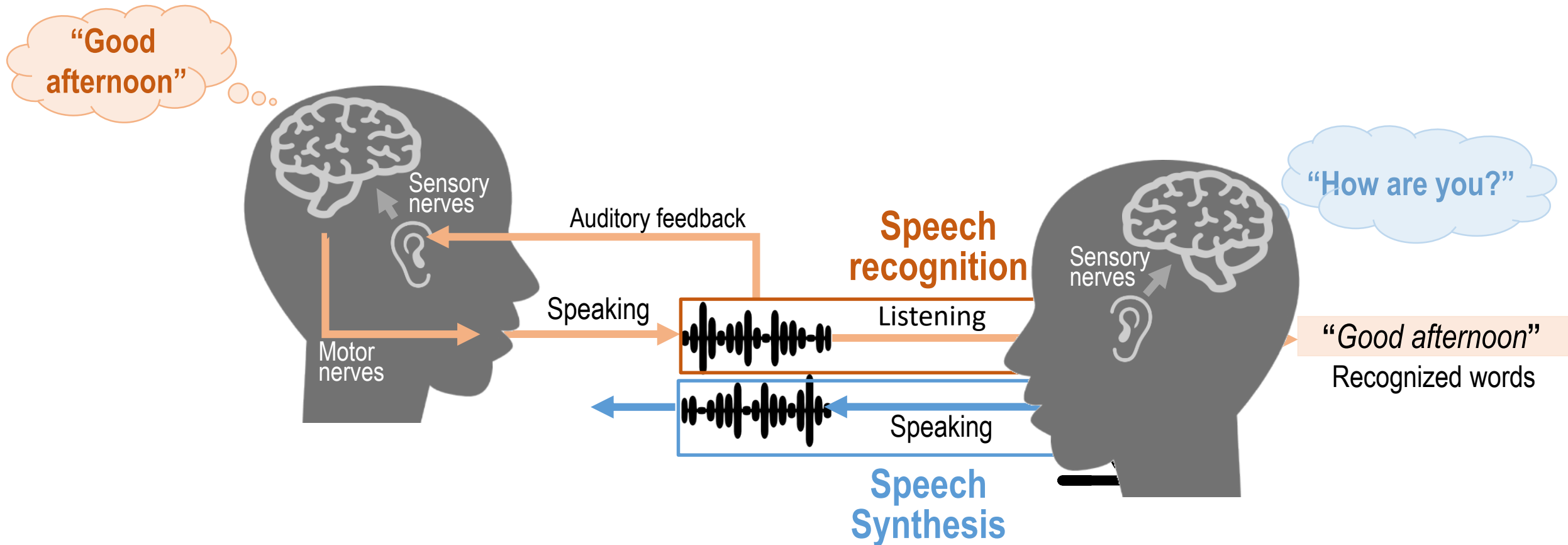
^{*1}Bernard S. Lee, “Delayed Speech Feedback”, The Journal of the Acoustical Society of America **22**, 824 (1950);

^{*2}Wikipedia “Delayed Auditory Feedback”

Human-Machine Interaction

■ Modality in Human-Machine Interaction

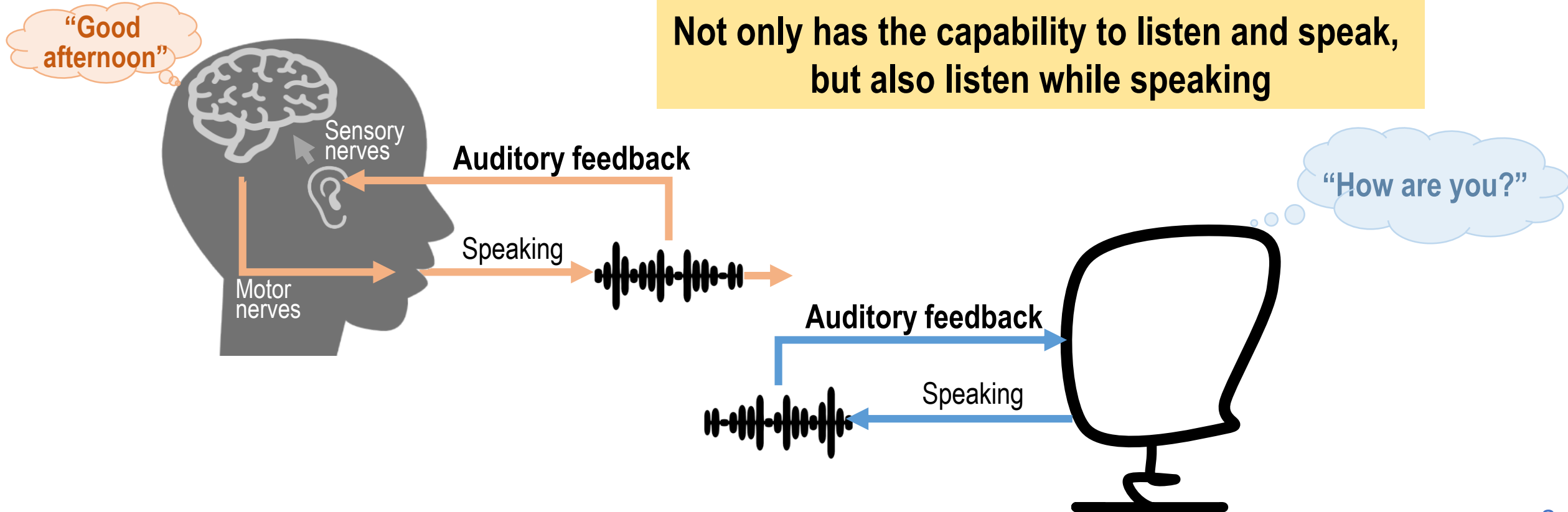
→ Providing a technology with ability to **listen** and **speak**



Machine Speech Chain

■ Proposed Method

- Develop a closed-loop speech chain model based on deep learning
- The first deep learning model that integrates human speech perception & production behaviors



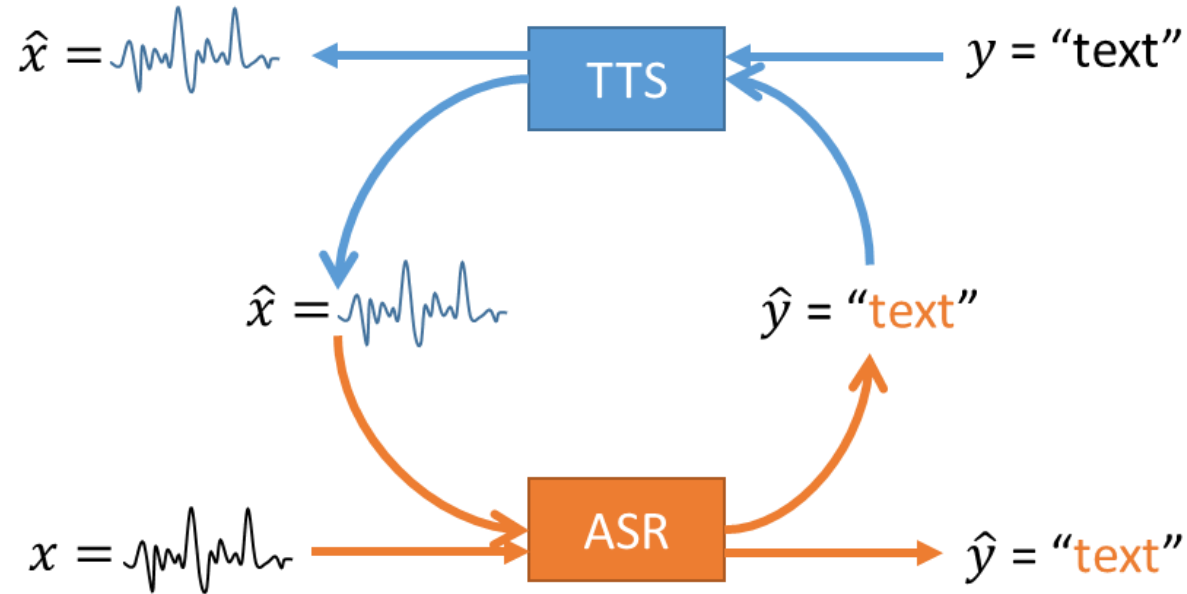
Not only has the capability to listen and speak, but also listen while speaking

Motivation Background

- Despite the close relationship between speech perception & production → ASR and TTS researches have progressed independently

Property	ASR	TTS
Speech features	MFCC Mel-fbank	MGC log F0, Voice/Unvoice, BAP
Text features	Phoneme Character	Phoneme + POS + LEX + ... (Full context label)
Model	GMM-HMM Hybrid DNN/HMM End-to-end ASR	GMM-HSMM DNN-HSMM End-to-end TTS

Machine Speech Chain



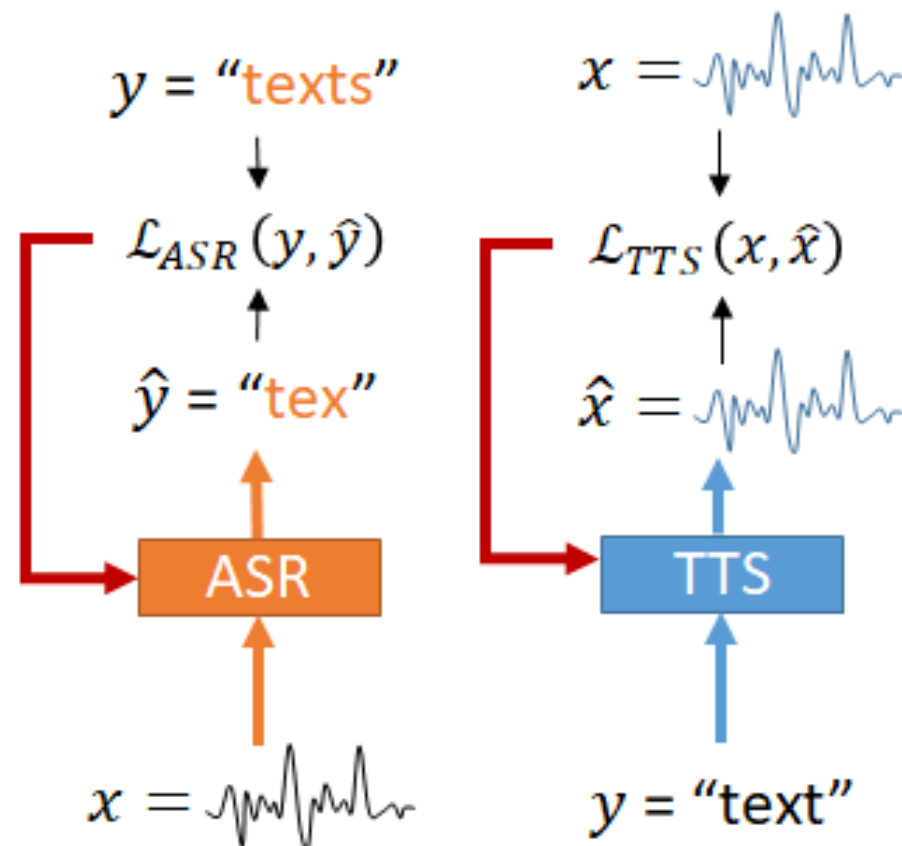
- Definition:

- x = original speech, y = original text
- \hat{x} = predicted speech, \hat{y} = predicted text
- $ASR(x): x \rightarrow \hat{y}$ (seq2seq model transforms speech to text)
- $TTS(y): y \rightarrow \hat{x}$ (seq2seq model transforms text to speech)

Machine Speech Chain

Case #1: Supervised Learning with Speech-Text Data

- **Given a pair speech-text (x, y)**
 - Train ASR and TTS in supervised learning
 - Directly optimized:
 - ASR by minimize $\mathcal{L}_{ASR}(y, \hat{y})$
 - TTS by minimizing loss between $\mathcal{L}_{TTS}(x, \hat{x})$
 - Update both ASR and TTS independently

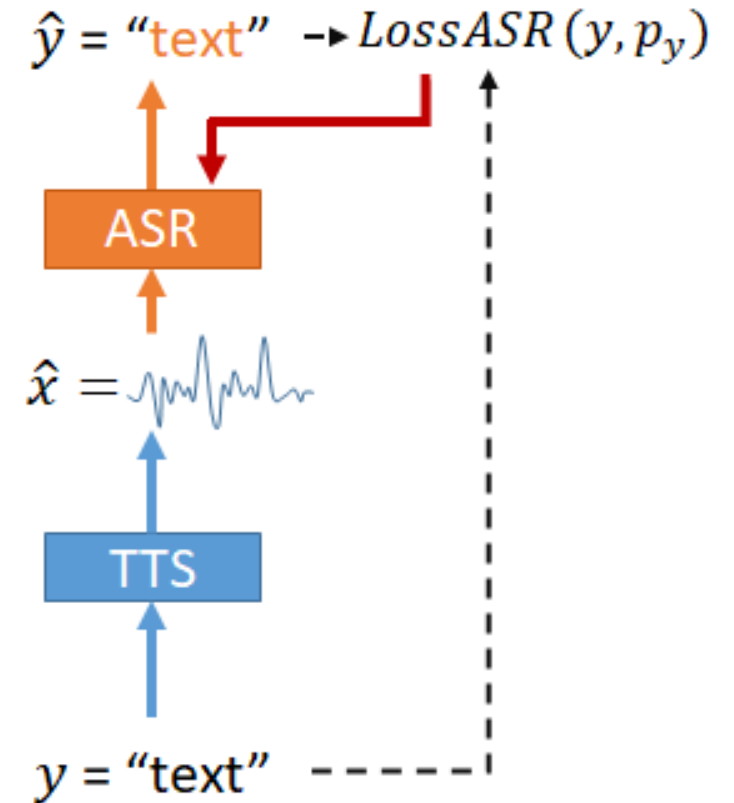


Machine Speech Chain

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**



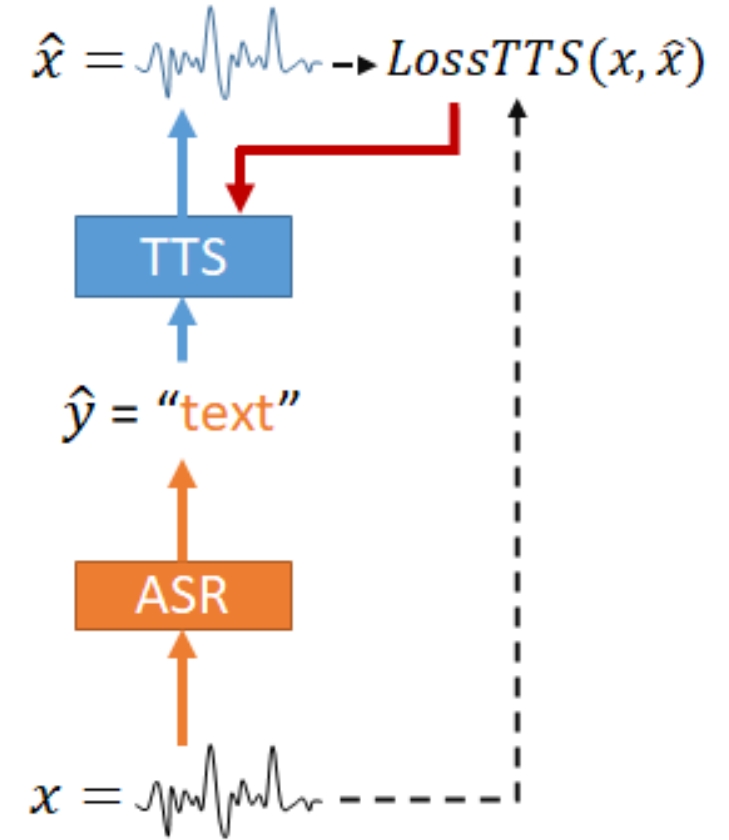
Machine Speech Chain

Case #3: Unsupervised Learning with Speech Only

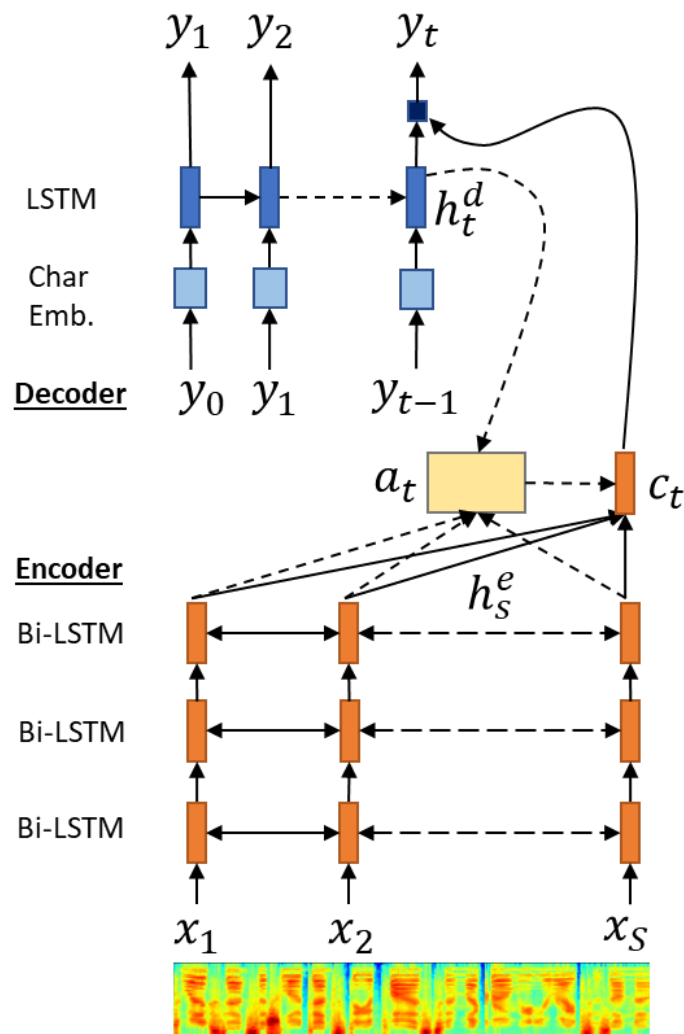
- **Given the unlabeled speech features x**

1. ASR predicts the most possible transcription \hat{y}
2. Based on \hat{y} , TTS tries to reconstruct speech features \hat{x}
3. Calculate $\mathcal{L}_{TTS}(x, \hat{x})$ 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

- $x = [x_1, \dots, x_S]$ (speech feature)
- $y = [y_1, \dots, y_T]$ (text)

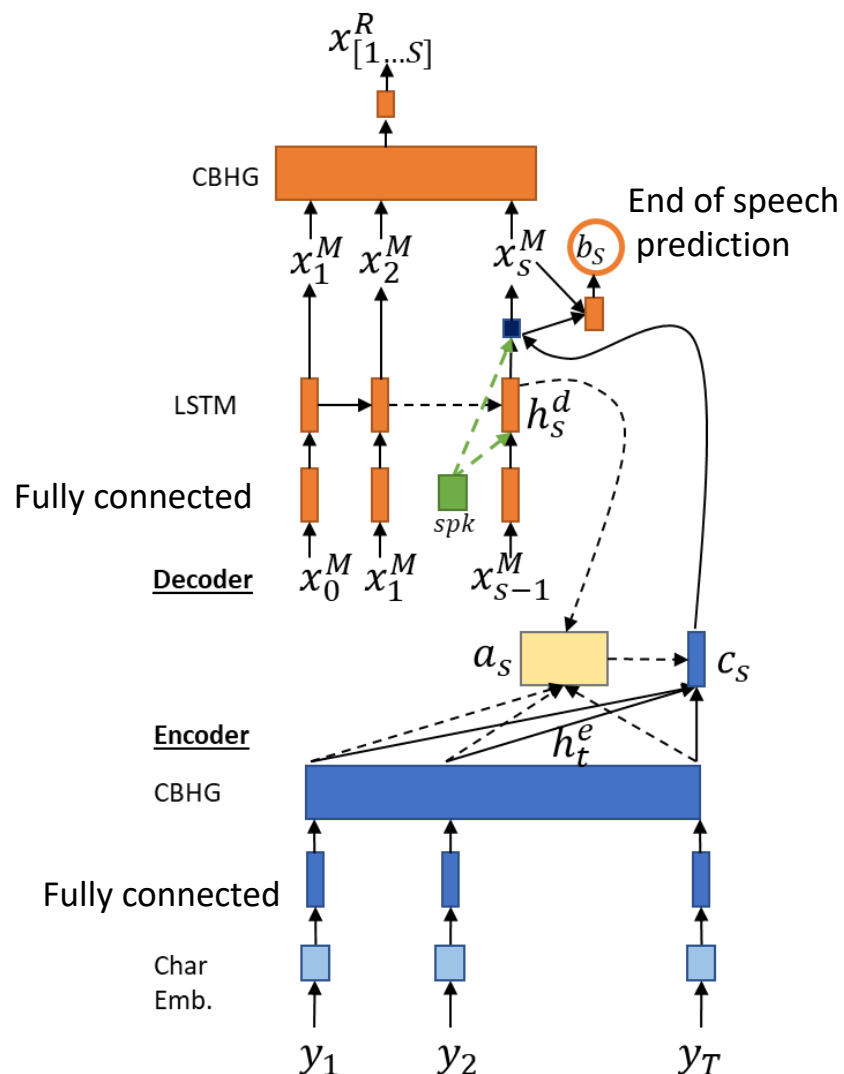
Model states

- $h_{[1..S]}^e$ = 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

$$\mathcal{L}_{ASR}(y, p_y) = -\frac{1}{T} \sum_{t=1}^T \sum_{c \in [1..C]} 1(y_t = c) * \log p_{y_t}[c]$$

Sequence-to-Sequence TTS



Input & output

- $x^R = [x_1, \dots, x_S]$ (linear spectrogram feature)
- $x^M = [x_1, \dots, x_S]$ (mel spectrogram feature)
- $y = [y_1, \dots, y_T]$ (text)

Model states

- $h_{[1...S]}^e$ = encoder states
- h_s^d = decoder state at time t
- a_s = attention probability at time t
- $c_s = \sum_{t=1}^S a_s(t) * h_t^e$ (expected context)

Loss function

$$\mathcal{L}_{TTS1}(x, \hat{x}) = \frac{1}{S} \sum_{s=1}^S (x_s^M - \hat{x}_s^M)^2 + (x_s^R - \hat{x}_s^R)^2$$

$$\mathcal{L}_{TTS2}(b, \hat{b}) = -\frac{1}{S} \sum_{s=1}^S (b_s \log(\hat{b}_s) + (1 - b_s) \log(1 - \hat{b}_s))$$

$$\mathcal{L}_{TTS}(x, \hat{x}, b, \hat{b}) = \mathcal{L}_{TTS1}(x, \hat{x}) + \mathcal{L}_{TTS2}(b, \hat{b})$$

19: *# Loss combination:*

20: Combine all weighted loss into a single loss variable

$$L = \alpha * (L_P^{TTS} + L_P^{ASR}) + \beta * (L_U^{TTS} + L_U^{ASR}) \quad (5)$$

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

$$G_{ASR} = \nabla_{\theta_{ASR}} L \quad (6)$$

$$G_{TTS} = \nabla_{\theta_{TTS}} L \quad (7)$$

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

$$\theta_{ASR} \leftarrow \text{Optim}(\theta_{ASR}, G_{ASR}) \quad (8)$$

$$\theta_{TTS} \leftarrow \text{Optim}(\theta_{TTS}, G_{TTS}) \quad (9)$$

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:

Data	Hyperparameter			ASR	TTS		
	α	β	gen. mode	CER (%)	Mel	Raw	Acc (%)
Paired (10k)	-	-	-	10.06	7.07	9.38	97.7
+Unpaired (40k)	0.25	1	greedy	5.83	6.21	8.49	98.4
	0.5	1	greedy	5.75	6.25	8.42	98.4
	0.25	1	beam 5	5.44	6.24	8.44	98.3
	0.5	1	beam 5	5.77	6.20	8.44	98.3

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

Data	Hyperparameter			ASR	TTS		
	α	β	gen. mode	CER (%)	Mel	Raw	Acc (%)
Paired (80 utt/spk)	-	-	-	26.47	10.21	13.18	98.6
+Unpaired (remaining)	0.25	1	greedy	23.03	9.14	12.86	98.7
	0.5	1	greedy	20.91	9.31	12.88	98.6
	0.25	1	beam 5	22.55	9.36	12.77	98.6
	0.5	1	beam 5	19.99	9.20	12.84	98.6

Acc: End of speech prediction accuracy

Outline

- ***Machine Speech Chain***

- ***Machine Speech Chain: Listening while speaking***

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, “Listening while Speaking: Speech Chain by Deep Learning”, ASRU 2017

- ***Speech Chain with One-shot Speaker Adaptation***

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura^{1,2} “Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018

- ***End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator***

- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. ICASSP, 2019

- ***End-to-end Speech-to-speech Translation***

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

- Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation”, INTERSPEECH2017



while



Speech Chain with One-shot Speaker Adaptation

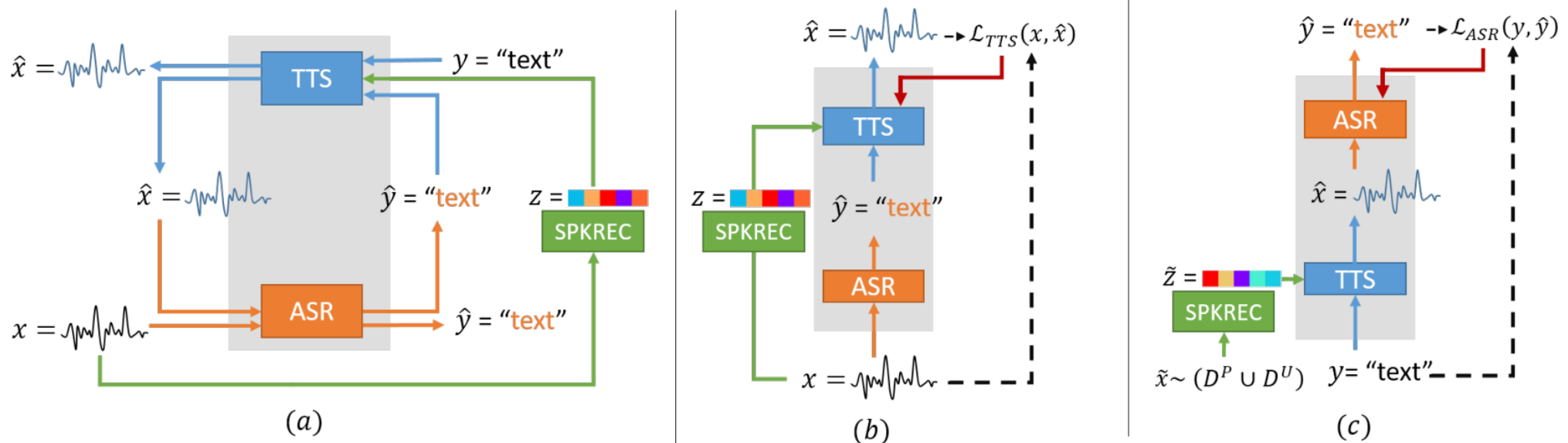
Andros Tjandra^{1,2}, Sakriani Sakti^{1,2}, Satoshi Nakamura^{1,2}

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

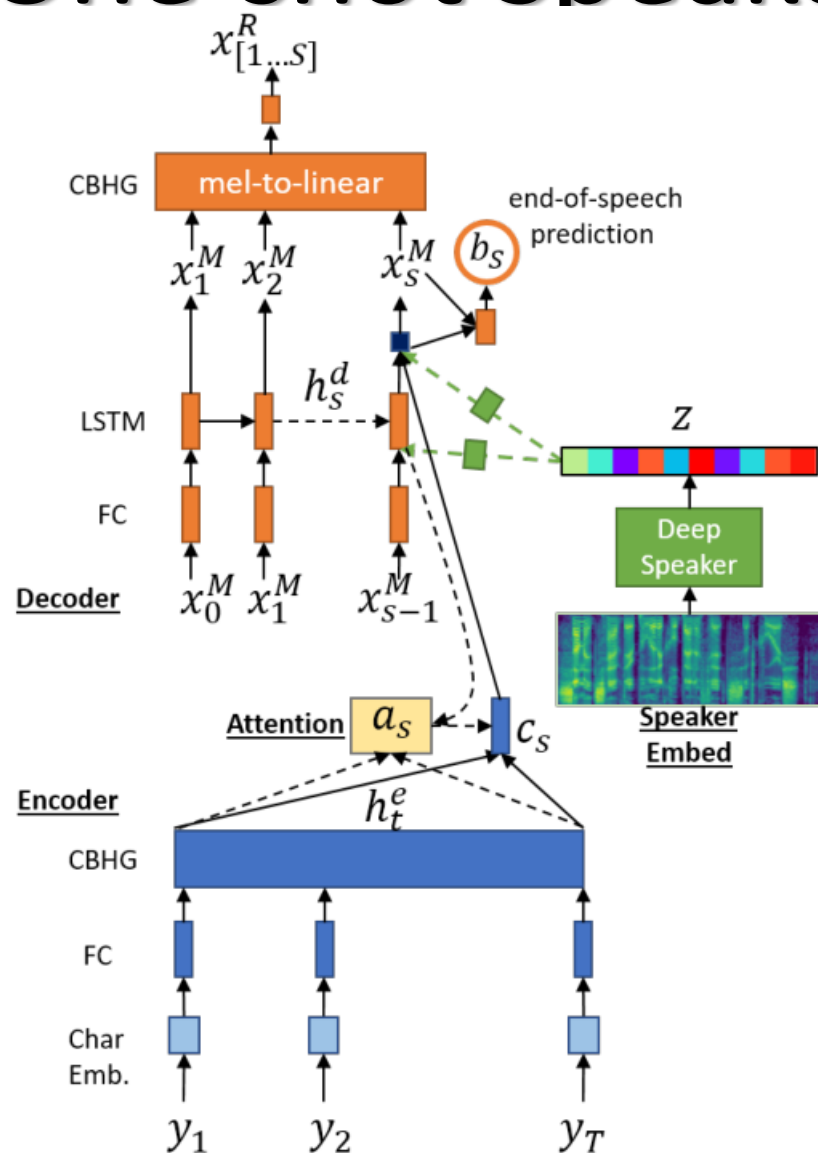


Sequel: Speech Chain with One-shot Speaker Adaptation

- 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



- 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).




ASR Results

Model	CER (%)
Supervised training: WSJ train_si84 (16hrs speech, paired) -> Baseline	
Att Enc-Dec	17.35
Supervised training: WSJ train_si284 (66 hrs speech, paired) -> Upperbound	
Att Enc-Dec	7.12
Semi-supervised training: WSJ train_si84 (paired) + train_si200 (unpaired)	
Label propagation (greedy)	17.52
Label propagation (beam=5)	14.58
<i>Proposed speech chain</i>	9.86







TTS Results

- **Text:** “the busses aren’t the problem, they actually provide a solution”

- Single Speaker (LJSpeech) (p = paired, u = unpaired)

Baseline (P 30%)	Sp-Chain (P 30% + U 70%)	Full (P 100%)
		

- Multispeaker (WSJ)

Speaker	Baseline (P si84)	Sp-Chain (P si84 + U si200)	Full (P si284)
Female A			
Male B			

Outline

- ***Machine Speech Chain***

- ***Machine Speech Chain: Listening while speaking***

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, “Listening while Speaking: Speech Chain by Deep Learning”, ASRU 2017

- ***Speech Chain with One-shot Speaker Adaptation***

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura^{1,2} “Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018

- ***End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator***

- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. ICASSP, 2019

- ***End-to-end Speech-to-speech Translation***

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

- Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation”, INTERSPEECH2017



while



End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator

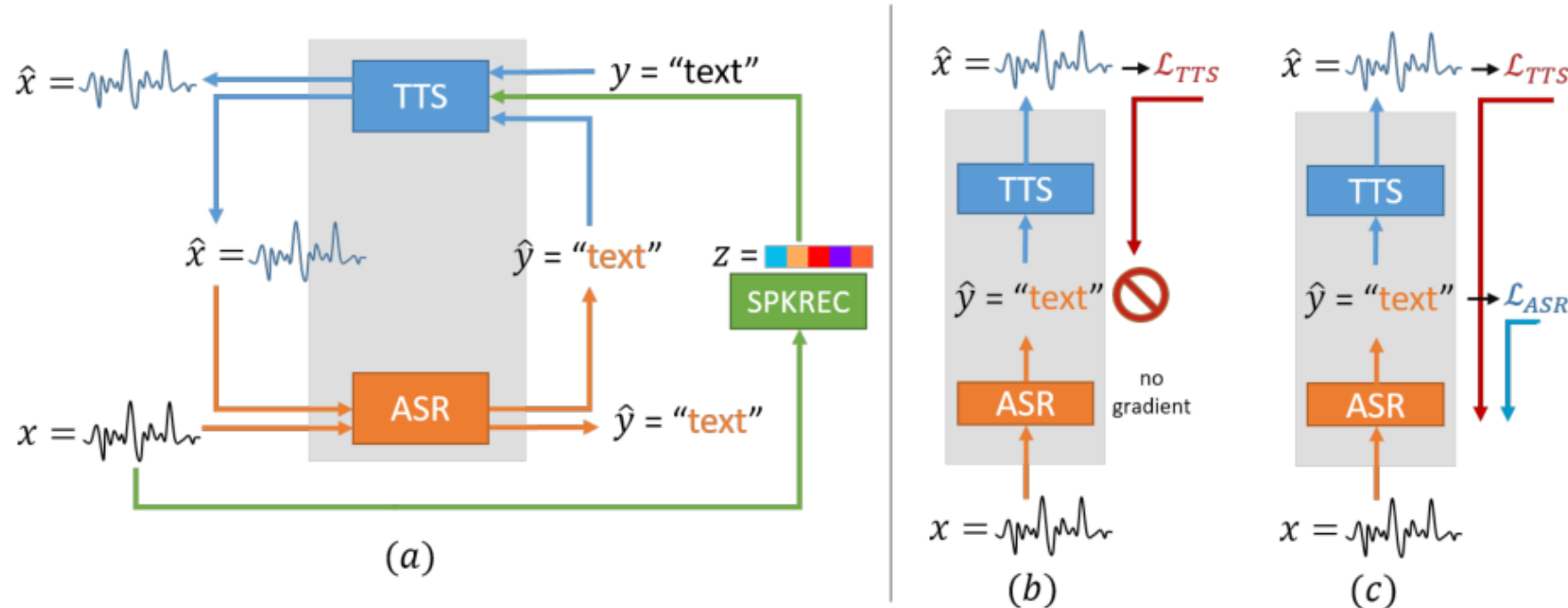
Andros Tjandra^{1,2}, Sakriani Sakti^{1,2}, Satoshi Nakamura^{1,2}

"End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. ICASSP, 2019



Straight-Through Estimator for Speech Chain

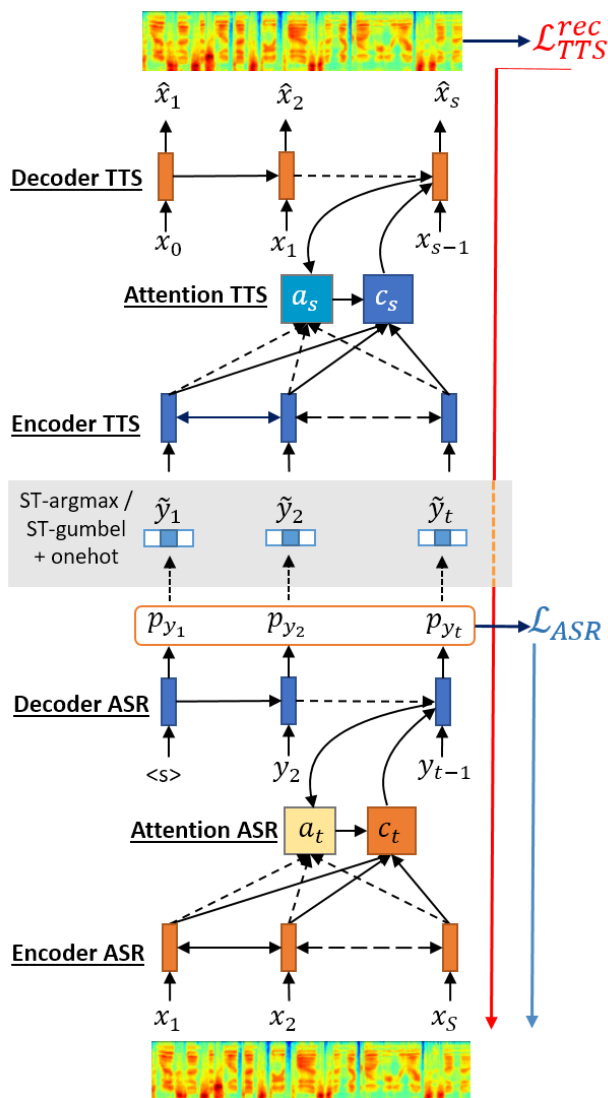
- **Proposed Approach:** Handle backpropagation through discrete nodes



Feedback loss: $\mathcal{L}_{TTS}(x, \hat{x})$ where $x = TTS(\hat{y}, z)$

- Speech chain loop with speaker embedding module z
- Original: feedback \mathcal{L}_{TTS} can't be backpropagated through variable \hat{y}
- Proposal:** Estimate gradient through variable \hat{y} with straight-through estimator

Straight-Through Estimator



a) ST-argmax

Deterministic choosing token by highest probability

$$p_{y_t}[c] = \frac{\exp(h_t^d[c])}{\sum_{i=1}^C \exp(h_t^d[c])}$$

$$\tilde{y}_t = \operatorname{argmax}_c p_{y_t}[c]$$

b) ST-Gumbel softmax

Sampling a token from $p_{y_t}[c]$:

$$p_{y_t}[c] = \frac{\exp((h_t^d[c] + g_c)/\tau)}{\sum_{i=1}^C \exp((h_t^d[c] + g_c)/\tau)}$$

$$\tilde{y}_t \sim \operatorname{Categorical}(p_{y_t}[1], \dots, p_{y_t}[C])$$

τ = temperature

New gradient \mathcal{L}_{TTS} w.r.t. θ_{ASR}

$$\begin{aligned} \frac{\partial \mathcal{L}_{TTS}^{rec}}{\partial \theta_{ASR}} &= \sum_{t=1}^T \frac{\partial \mathcal{L}_{TTS}^{rec}}{\partial \tilde{y}_t} \cdot \frac{\partial \tilde{y}_t}{\partial p_{y_t}} \cdot \frac{\partial p_{y_t}}{\partial \theta_{ASR}} \\ &\approx \sum_{t=1}^T \frac{\partial \mathcal{L}_{TTS}^{rec}}{\partial \tilde{y}_t} \cdot \mathbb{1} \cdot \frac{\partial p_{y_t}}{\partial \theta_{ASR}}. \end{aligned}$$

Experiments on Multi-Speakers WSJ Task

■ Data set

- **Training set: Supervised (paired text & speech)**
WSJ SI-284 dataset (upperbound)
(37318 utterances, ~81 h, 284 speakers)
- **Development set:** dev93
- **Evaluation set:** eval92

Model	CER (%)
Baseline	
Enc-Dec Att-MLP [Kim et al., 2017]	11.08
Enc-Dec Att-MLP-Loc [Kim et al., 2017]	8.17
Enc-Dec Att-MLP [Tjandra et al., 2017]	7.12
Enc-Dec Att-MLP-MA (ours) [Tjandra et al., 2018]	6.43
Proposed Method	
Enc-Dec Att-MLP-MA SP-Chain ST argmax	5.75
Enc-Dec Att-MLP-MA SP-Chain ST gumbel	5.70

Outline

- ***Machine Speech Chain***

- ***Machine Speech Chain: Listening while speaking***

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, “Listening while Speaking: Speech Chain by Deep Learning”, ASRU 2017

- ***Speech Chain with One-shot Speaker Adaptation***

- Andros Tjandra, Sakriani Sakti, Satoshi Nakamura^{1,2} “Machine Speech Chain with One-shot Speaker Adaptation”, Proceedings of INTERSPEECH 2018

- ***End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator***

- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. ICASSP, 2019

- ***End-to-end Speech-to-speech Translation***

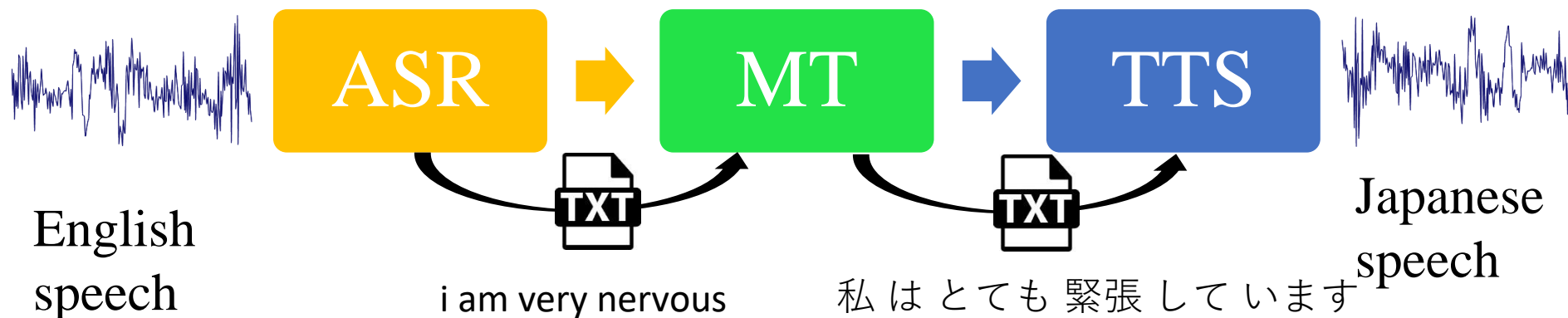
- ***Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation***

- Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation”, INTERSPEECH2017

Structure Based Curriculum Learning for End-to-end Direct English-Japanese Speech Translation

Takatomo Kano, Sakriani Sakti, Satoshi Nakamura, “Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation”, INTERSPEECH2017

Traditional Speech Translation



Traditional approach in speech-to-speech translation systems

✓ construct

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

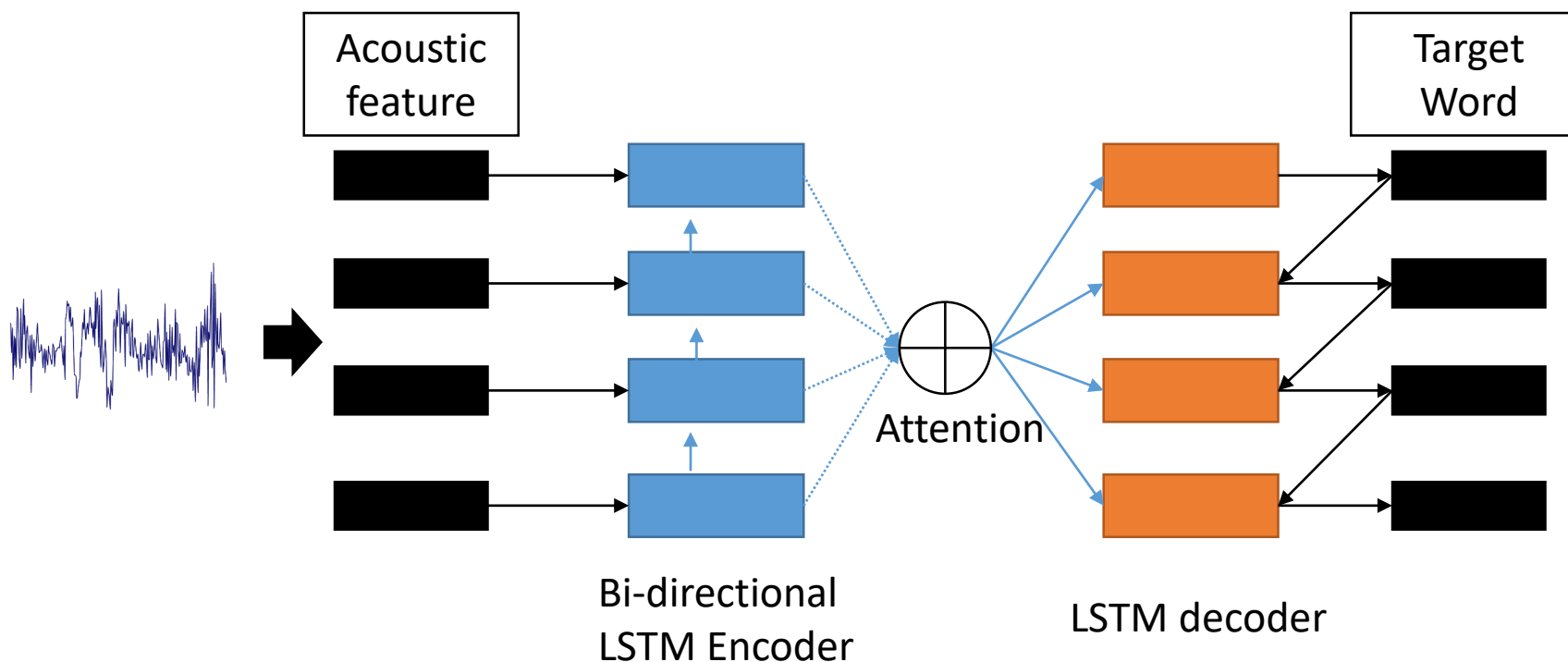
✓ all of which are independently trained and tuned

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

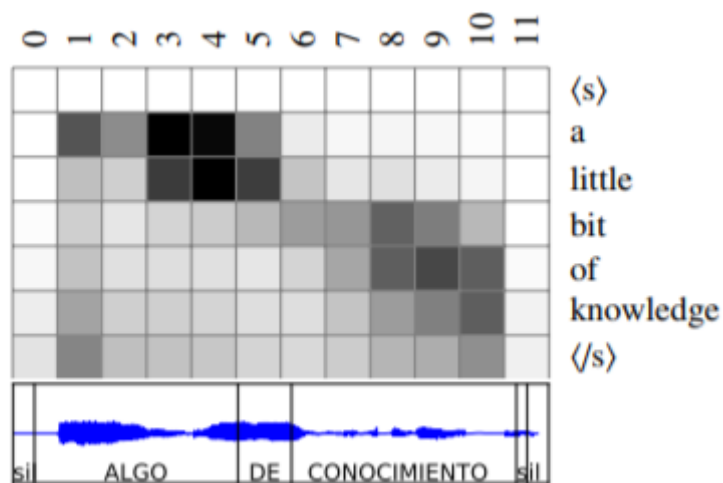
Related Works^[2]

- 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]



(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.

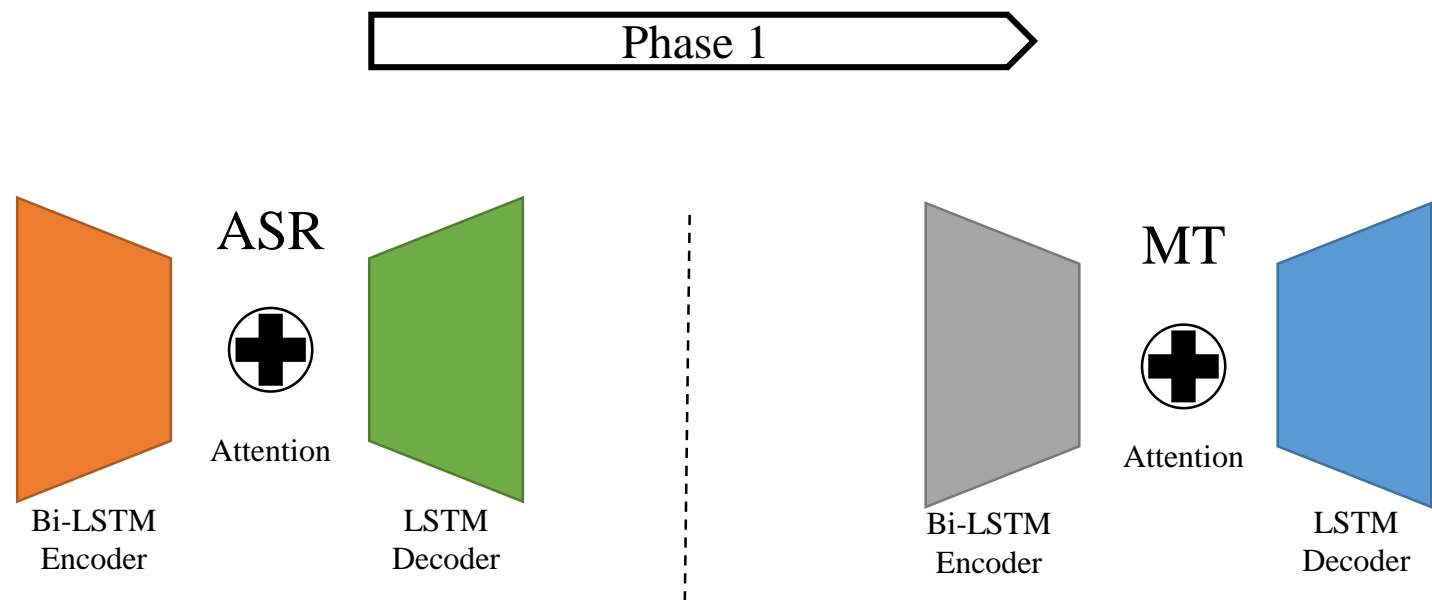


English to Japanese translation attention matrix

Proposed method

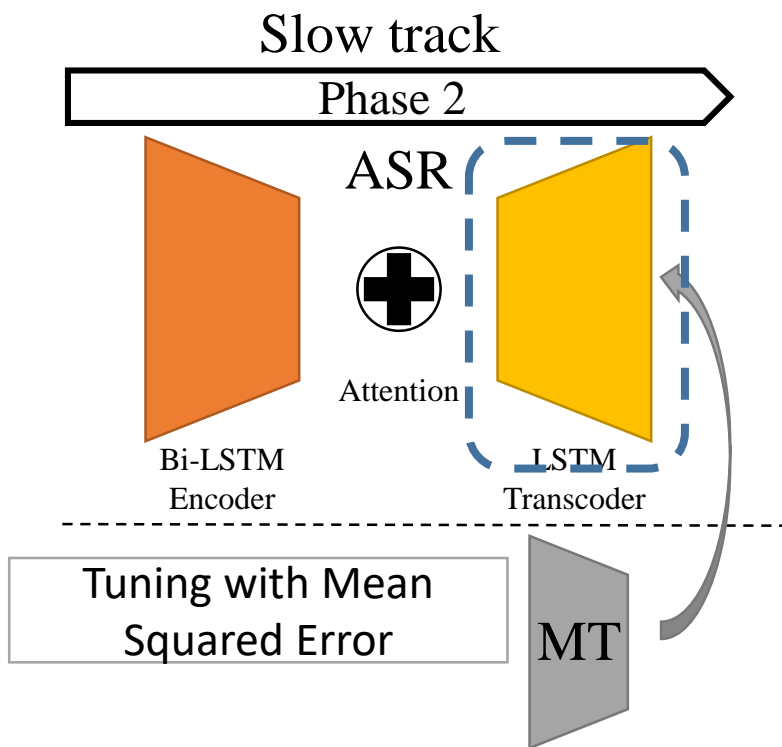
- A first attempt to build direct speech-to-text direct translation system (ST) on syntactically distant language pairs
- To guide the encoder-decoder attentional model to learn this difficult problem, we proposed a **structured-based curriculum learning** strategy.

Attention-based ST with Curriculum Learning

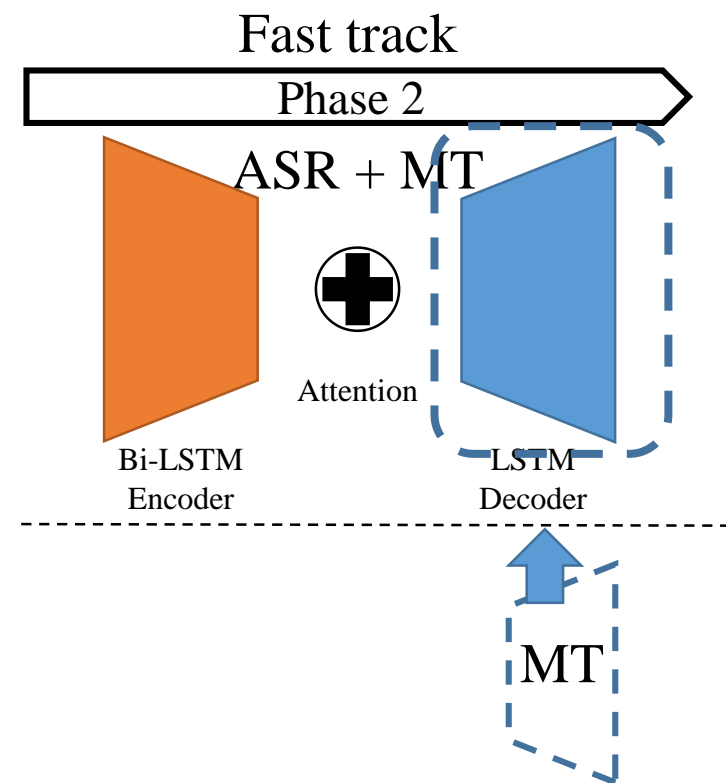


Train the attentional-based encoder-decoder neural network for a standard ASR and MT task

Attention-based ST with Curriculum Learning

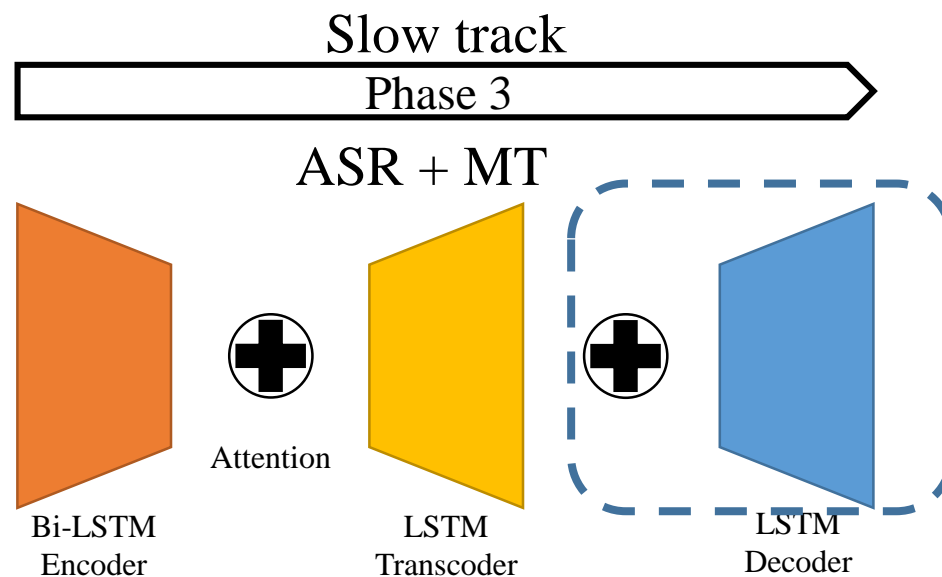


The model's objective now is to predict the word representation (like the MT encoder's output)



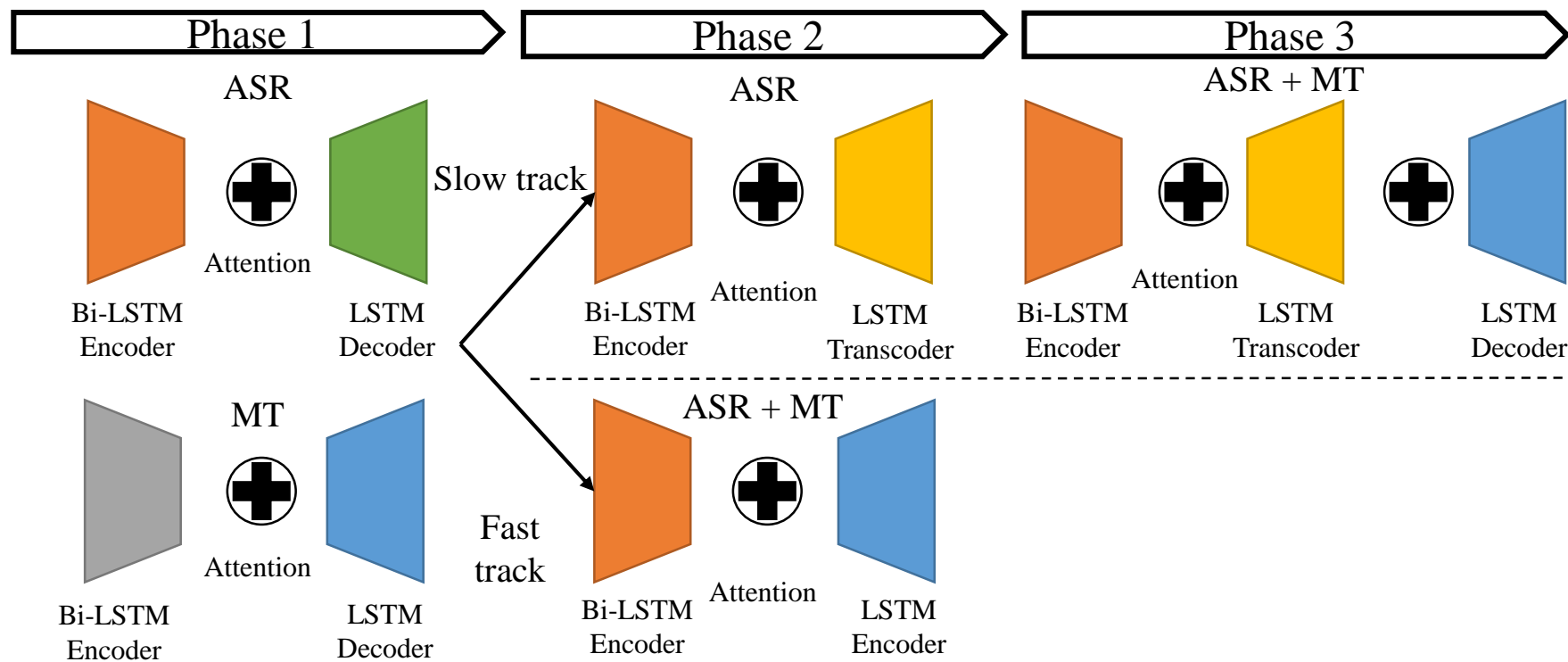
The model now predicts the corresponding word sequence in the target language given the input speech

Attention-based ST with Curriculum Learning



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



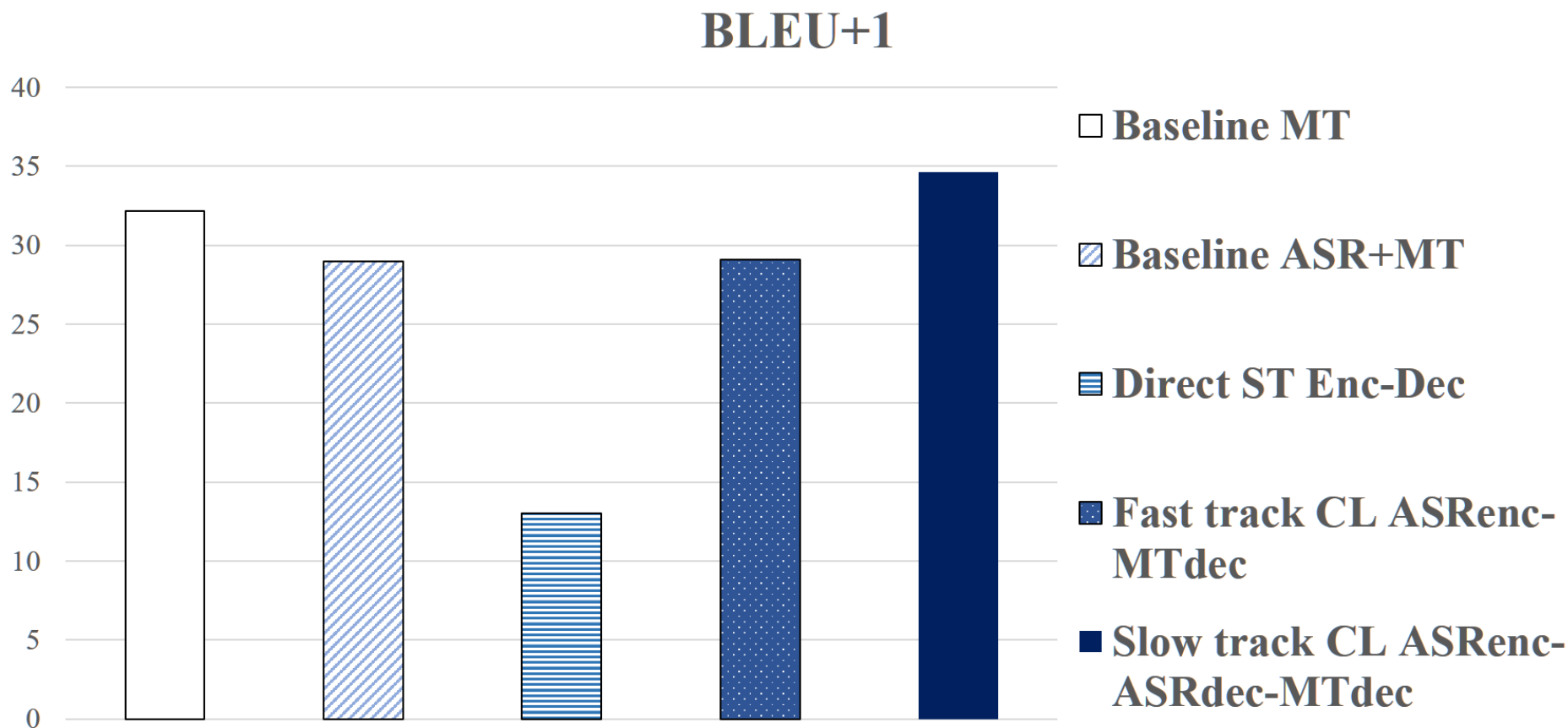
Attentional-based neural trained for ASR and text-based MT tasks and gradually train the network for end-to-end ST tasks.

Experimental Set-up

System settings	
ASR	
Input units	23
Hidden units	512
Output units	27293
LSTM layer depth	2
MT	
Source Vocabulary	27293
Target Vocabulary & Output size	33155
Input units & Embed size	12823
Hidden units	512
LSTM layer Depth	2
Optimizer	
Adam	

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

Translation Accuracy



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

Overall Summary:

- *Machine Speech Chain by ASR-TTS coupling*
 - *Machine Speech Chain: Listening while speaking*
 - *Speech Chain with One-shot Speaker Adaptation*
 - *End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator*
- *End-to-end Speech-to-speech Translation*
 - *Structure Based Curriculum Learning for End-to-end English-Japanese Speech Translation*
- *Future Works*
 - *Multi-modal speech chain*
 - *Advanced ASR-TTS modules*
 - *Advanced MT modules*
 - *Learn human perception and cognitive process*