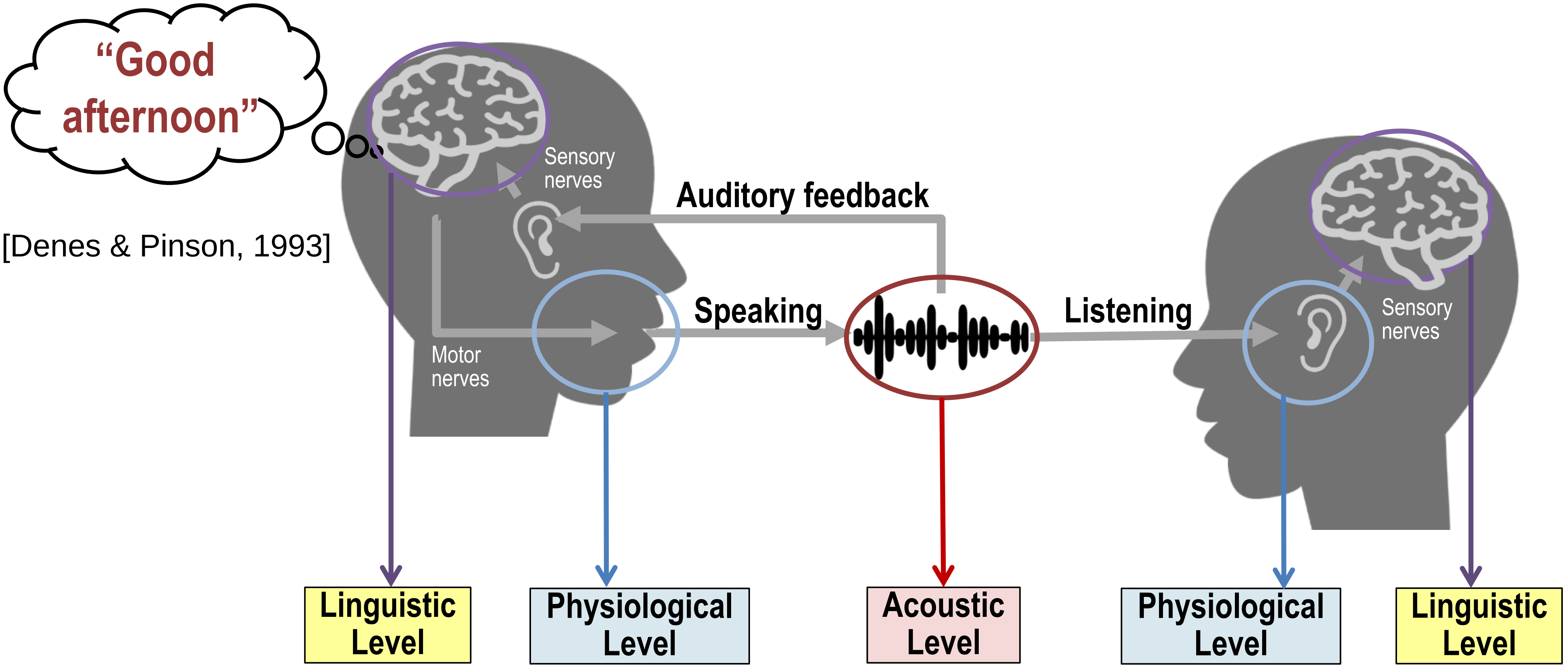
Towards More Human-like Machine Speech Chain

Satoshi Nakamura[1,2], with
Sashi Novitasari[1], and Sakriani Sakti[1,2]
[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Advanced Intelligence Project AIP, Japan

Speech Chain on Human Speech Communication

- In human speech production and hearing
  → Closed-loop speech chain mechanism with auditory feedback



[Denes & Pinson, 1993]

Delayed Auditory Feedback[*1,2]

- DAF for stutter:
  - DAF device that enables a user to speak into a microphone and then hear own voice in headphones a fraction of a second later
  - Stuttering was corrected or bypassed while speaking under DAF.

- Effects in normal speakers
  - DAF is used to see the structure of the auditory and verbal pathways in the brain.
  - Reduction in speaking rate, increase in intensity, and increase in fundamental frequency in order to overcome the effects of the feedback.
  - Repetition of syllables, mispronunciations, omissions, and omitted word endings.

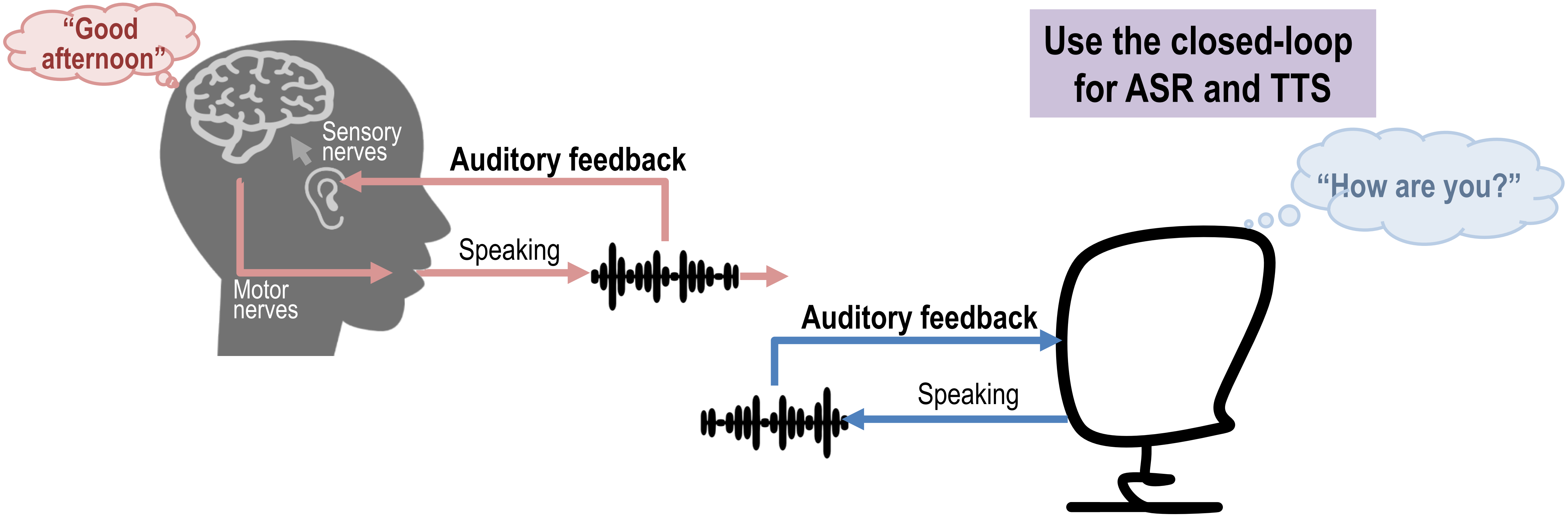
[*1]Bernard S. Lee, “Delayed Speech Feedback”, The Journal of the Acoustical Society of America **22**, 824 (1950);
[*2] Wikipedia “Delayed Auditory Feedback”

Topics

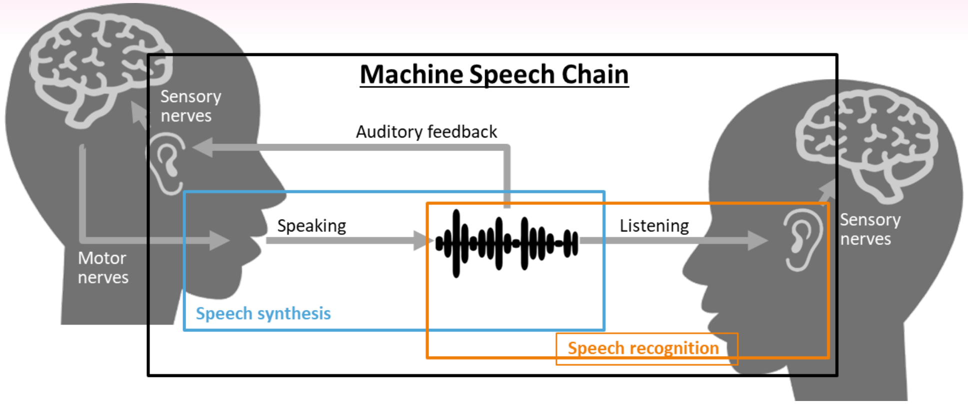
- Machine Speech Chain
  - ASR and TTS research
  - ASR & TTS semi-supervised joint learning
- Neural Incremental ASR and TTS
  - Neural Incremental ASR
  - Neural Incremental TTS
- Incremental Speech Chain
  - Incremental Learning of Speech Chain
- Summary

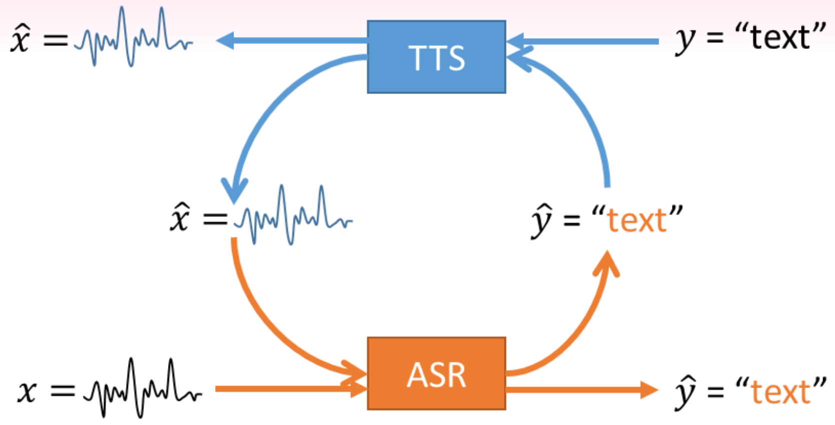
Machine Speech Chain

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



Machine Speech Chain





→ **In training stage:** ASR and TTS teach each other using unpaired data and generate useful feedback

→ **In Inference stage:** Possible to use ASR & TTS module independently, or dependently

→ **Semi-supervised learning:** Allow to train with labeled and unlabeled data

→ **A closed-loop architecture in which the domain of source & target are different**

Two agents: (1) ASR: Speech-to-text vs (2) TTS: Text-to-speech

(Dual learning NMT: both agents text2text, CycleGAN: both agents image2image)

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

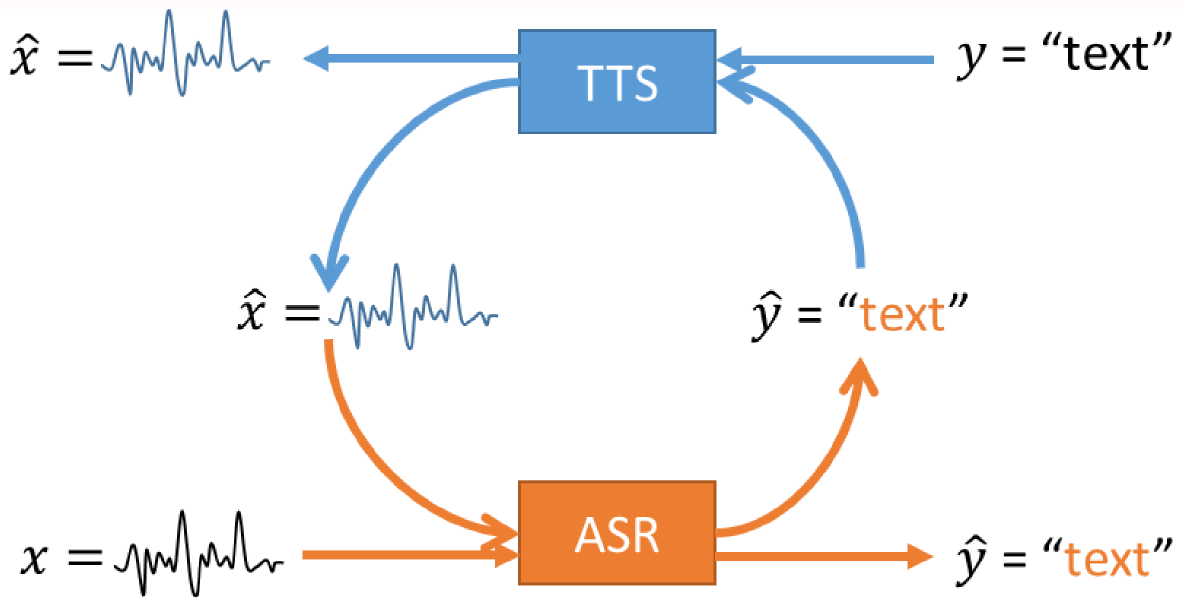
Machine Speech Chain

Andros Tjandra, Sakriani Sakti, Satoshi Nakamura,

"Listening while Speaking: Speech Chain by Deep Learning",

IEEE ASRU 2017

Machine Speech Chain

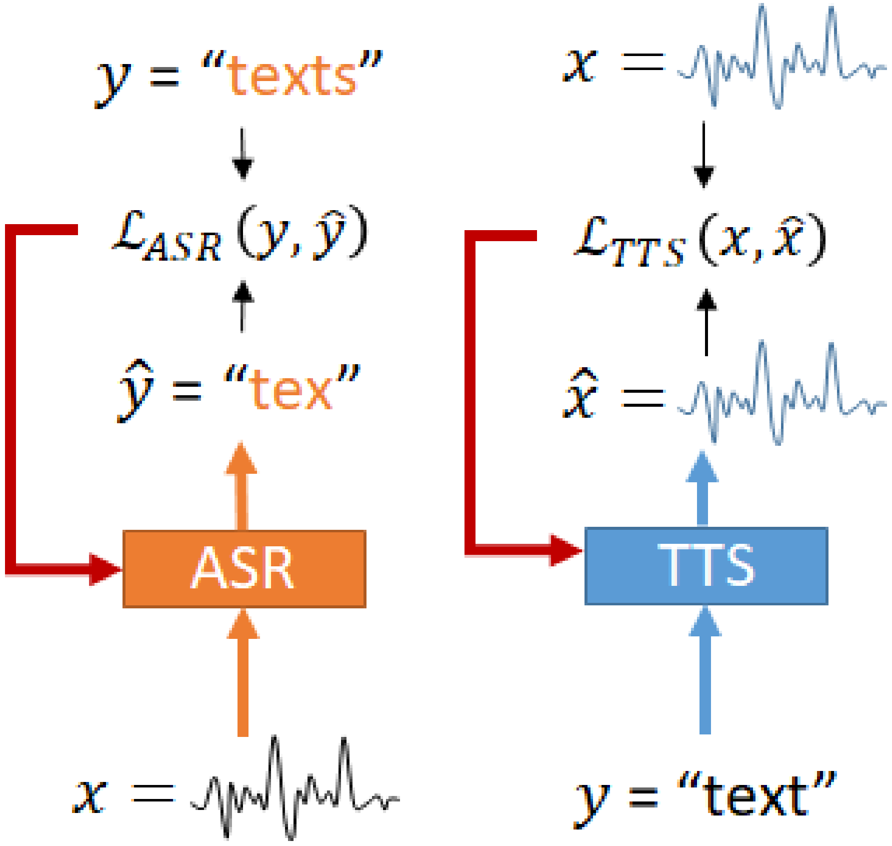


- Definition:
  - x = original speech, y = original text
  - $\hat{x}$ = predicted speech, $\hat{y}$ = predicted text
  - $ASR(x): x \rightarrow \hat{y}$ (seq2seq model transforms speech to text)
  - $TTS(y): y \rightarrow \hat{x}$ (seq2seq model transforms text to speech)

Machine Speech Chain

Case #1: Supervised Learning with Speech-Text Data

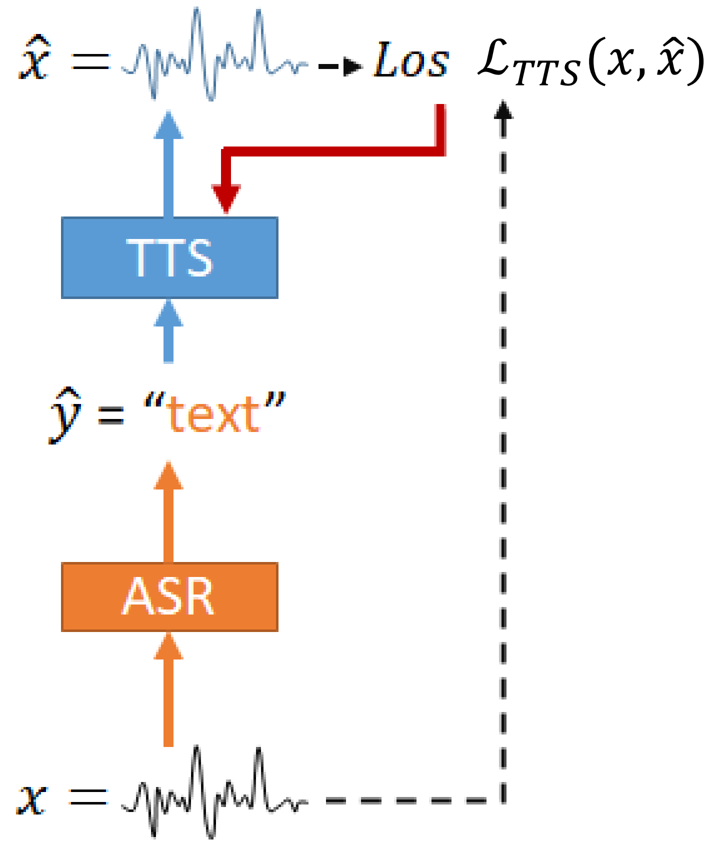
- **Given a pair speech-text $(\boldsymbol{x}, \boldsymbol{y})$**
  - Train ASR and TTS in supervised learning
  - Directly optimized:
    → ASR by minimize $\mathcal{L}_{ASR}(y, \hat{y})$
    → TTS by minimizing loss between $\mathcal{L}_{TTS}(x, \hat{x})$
  - Update both ASR and TTS independently



Machine Speech Chain

Case #2: Semi-supervised 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

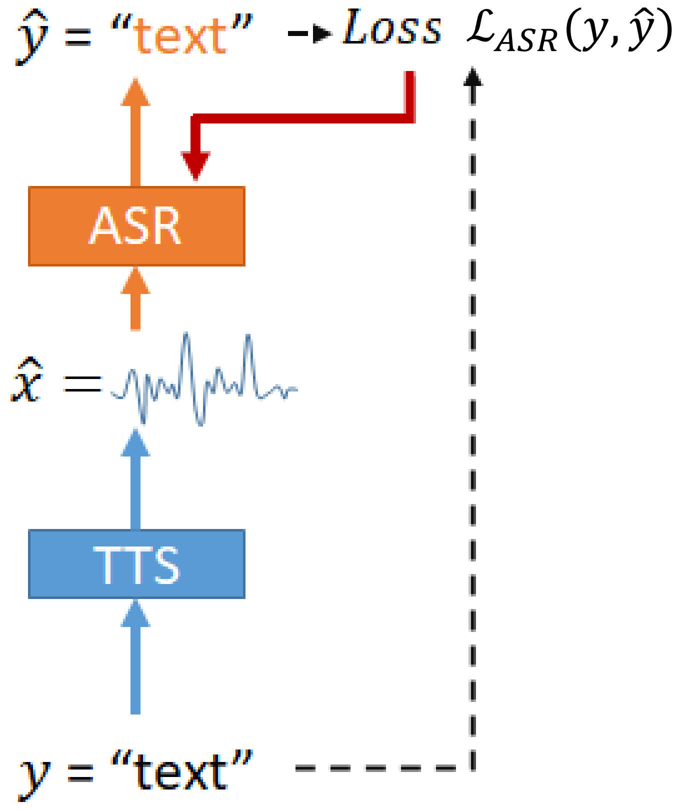
Machine Speech Chain

Case #3: Semi-supervised Learning with Text Only

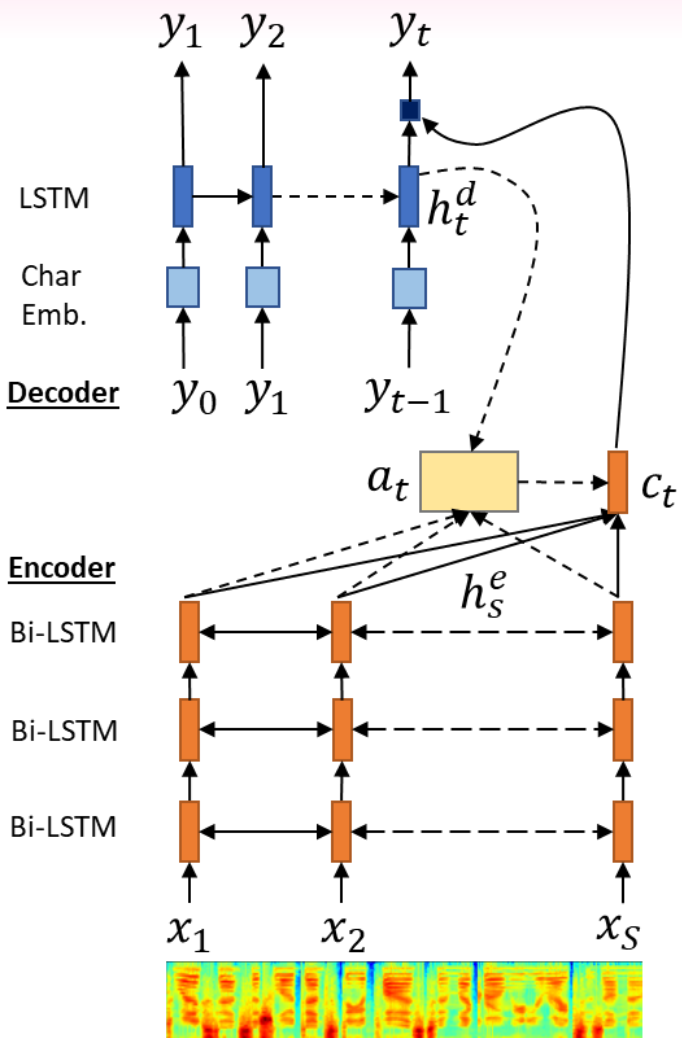
– **Given the unlabeled text features y**

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



Sequence-to-Sequence ASR



Input & output

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

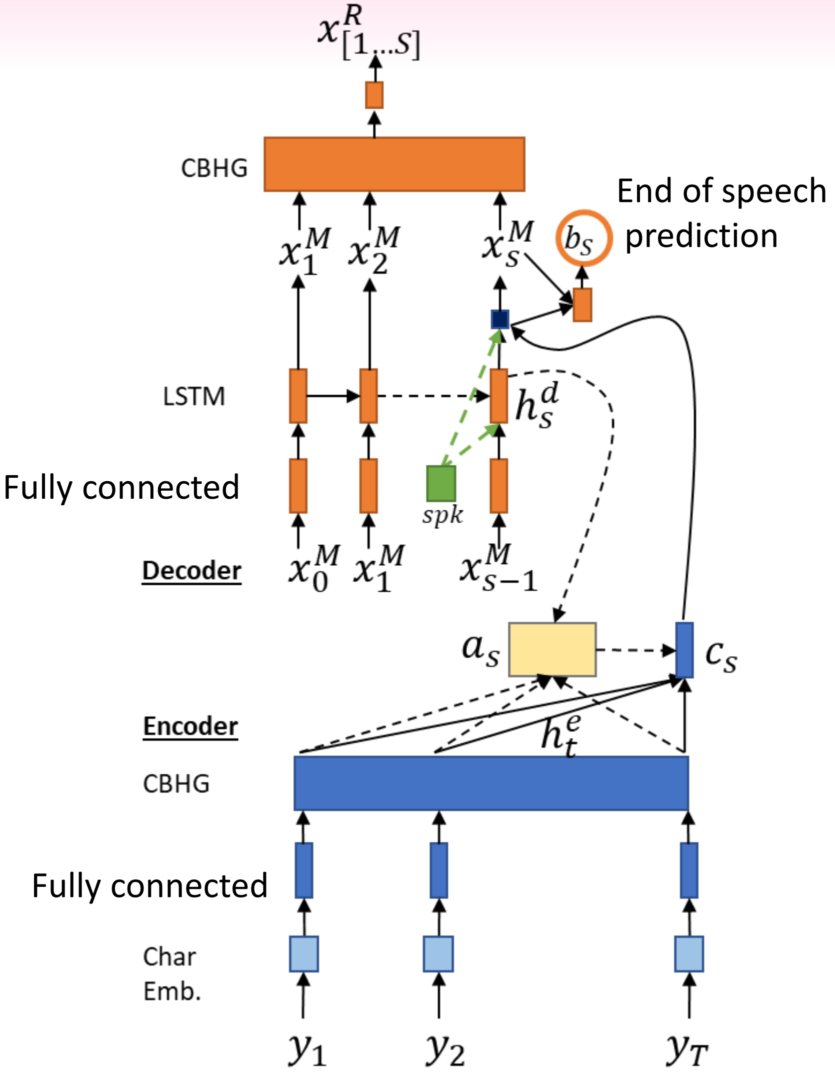
Model states

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

Loss function

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

Sequence-to-Sequence TTS



CBHG: Convolution Bank + Highway + bi-GRU

Input & output

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

Model states

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

Loss function

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

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

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

Model Optimization in Speech Chain

- Combined loss:

$$\ell_{ALL} = \underbrace{\alpha\left(\ell^{P}_{TTS} + \ell^{P}_{ASR}\right)}_{\text{Loss from paired data}} + \underbrace{\beta\left(\ell^{U}_{TTS} + \ell^{U}_{ASR}\right)}_{\text{Loss from unpaired data}}$$

α and β are hyper-parameters for scaling the gradient from paired and unpaired data

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

Machine Speech Chain

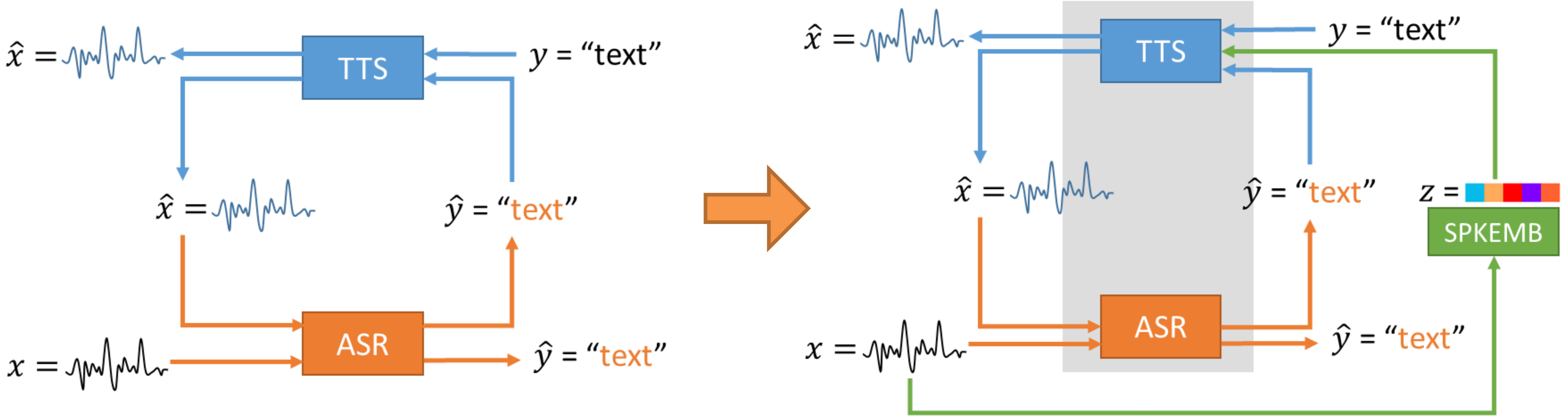
Andros Tjandra, Sakriani Sakti, Satoshi Nakamura,

"Machine Speech Chain with One-shot Speaker Adaptation",

INTERSPEECH 2018

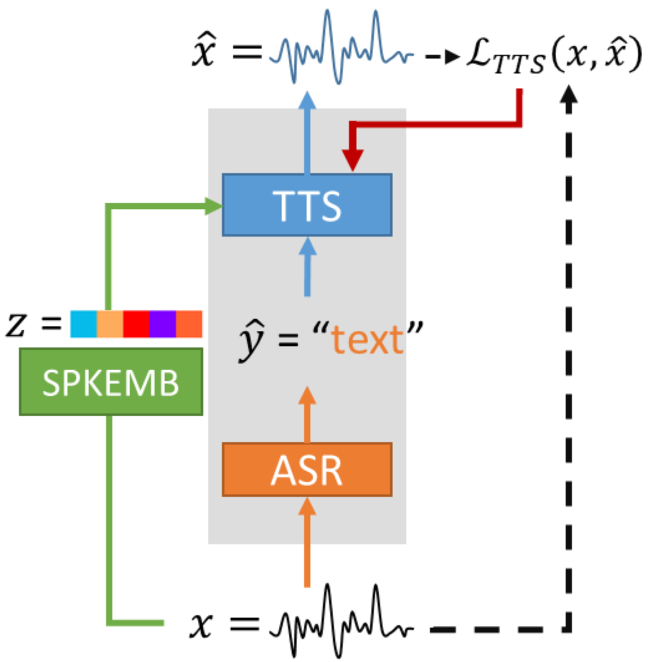
Multi-Speaker Speech Chain

Adding a speaker embedding as conditional input for TTS



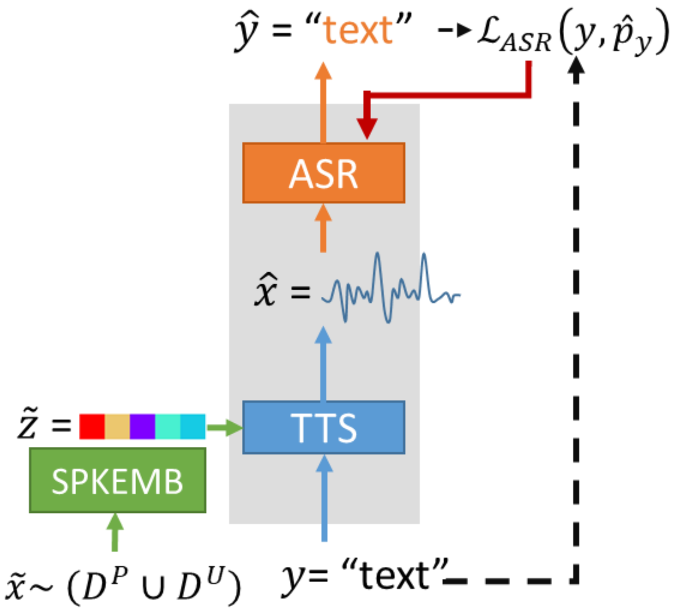
Training Unpaired Data Scenario

Train with unpaired speech



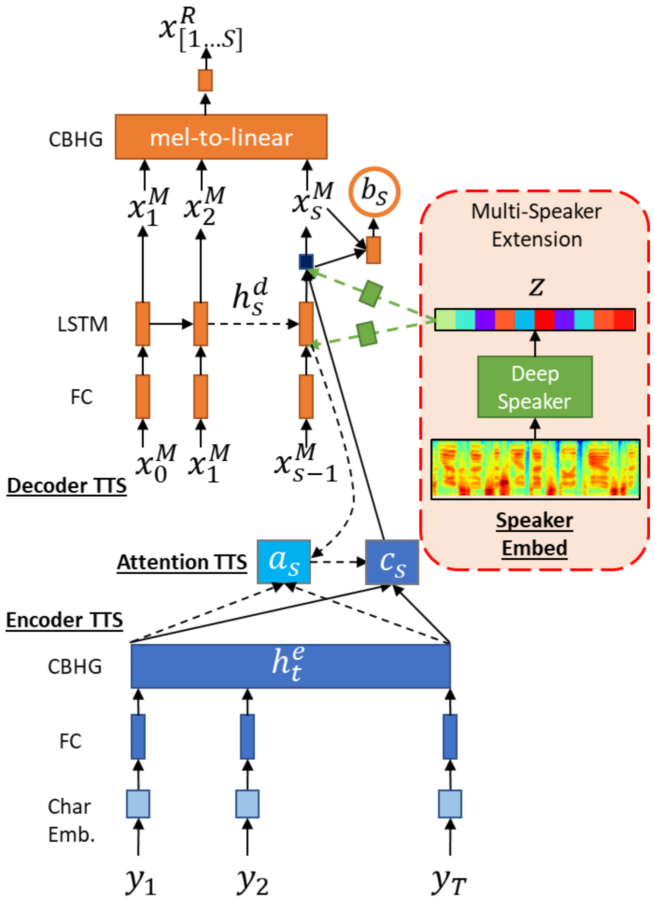
1. ASR predict best transcription $\hat{y}$ given x
2. SPKEMB generate speaker embedding z
3. TTS reconstructs $\hat{x}$ given $[\hat{y}, z]$

Train with unpaired text



1. Sample a speaker embedding $\tilde{z}$ from any speech
2. TTS generates speech feature $\hat{x}$ given $[y, \tilde{z}]$
3. ASR reconstruct text $\hat{y}$ given $\hat{x}$

Tacotron + Multi-speaker Adaptation



Input & output

- $\boldsymbol{x^R} = [x_1, \dots, x_S], \boldsymbol{x^M} = [x_1, \dots, x_S]$ (mel & linear spectrogram)
- $\boldsymbol{y} = [y_1, \dots, y_T]$ (text)
- $\boldsymbol{z}$ (speaker embedding vector)

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

Reconstruction MSE $\mathcal{L}_{TTS1}(x, \hat{x}) = \frac{1}{S}\sum_{s=1}^{S}(x^M_s - \hat{x}^M_s)^2 + (x^R_s - \hat{x}^R_s)^2$

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

Perceptual loss between original & generated speech $\mathcal{L}_{TTS3}(z, \hat{z}) = 1 - \frac{\langle z, \hat{z}\rangle}{\|z\|_2 + \|\hat{z}\|_2}$

$$\mathcal{L}_{TTS}(x, \hat{x}, b, \hat{b}) = \mathcal{L}_{TTS1}(x, \hat{x}) + \mathcal{L}_{TTS2}(b, \hat{b}) + \mathcal{L}_{TTS3}(z, \hat{z})$$

Experiment Results on WSJ

Table 1: *Character error rate (CER (%)) comparison between results of supervised learning and those of a semi-supervised learning method, evaluated on test_eval92 set*

Model	**CER (%)**
Supervised training: WSJ *train_si84* (paired) → Baseline	
Att Enc-Dec [19]	17.01
Att Enc-Dec [20]	17.68
Att Enc-Dec (ours)	17.35
Supervised training: WSJ *train_si284* (paired) → Upperbound	
Att Enc-Dec [19]	8.17
Att Enc-Dec [20]	7.69
Att Enc-Dec (ours)	7.12
Semi-supervised training: WSJ *train_si84* (paired) + *train_si200* (unpaired)	
Label propagation (greedy)	17.52
Label propagation (beam=5)	14.58
Proposed speech chain (Sec. 2)	**9.86**

Table 2: *L2-norm squared on log-Mel spectrogram to compare the supervised learning and those of a semi-supervised learning method, evaluated on test_eval92 set. Note: We did not include standard Tacotron (without SPKREC) into the table since it could not output various target speaker.*

Model	**L2-norm^2**
Supervised training: WSJ *train_si84* (paired) → Baseline	
Proposed Tacotron (Sec. 4) (ours)	1.036
Supervised training: WSJ *train_si284* (paired) → Upperbound	
Proposed Tacotron (Sec. 4) (ours)	0.836
Semi-supervised training: WSJ *train_si84* (paired) + *train_si200* (unpaired)	
Proposed speech chain (Sec. 2 + Sec. 4)	**0.886**

Topics

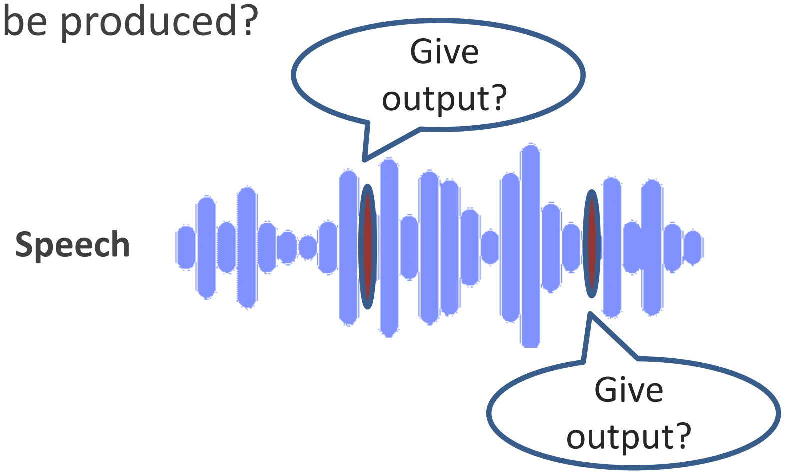
- Machine Speech Chain
  - ASR and TTS research
  - ASR & TTS semi-supervised joint learning
- Neural Incremental ASR and TTS
  - Neural Incremental ASR
  - Neural Incremental TTS
- Incremental Speech Chain
  - Incremental Learning of Speech Chain
- Summary

Incremental Speech Recognition

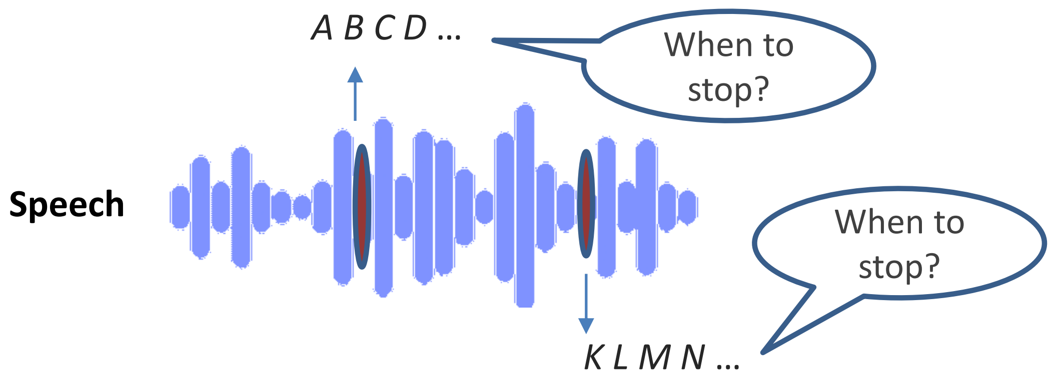
- **ISR** begins the speech recognition without waiting the speech to finish (low delay)
  - Recognize the speech part-by-part in several incremental steps
  - Input: a short part of the speech

- **Challenge:** How to do an incremental step?

1) **Input boundary decision**
   When the transcription of a short speech part can be produced?



2) **Output boundary decision**
   When to stop the output prediction of the current speech part and and move to the next?



Need to learn short input-short output alignments

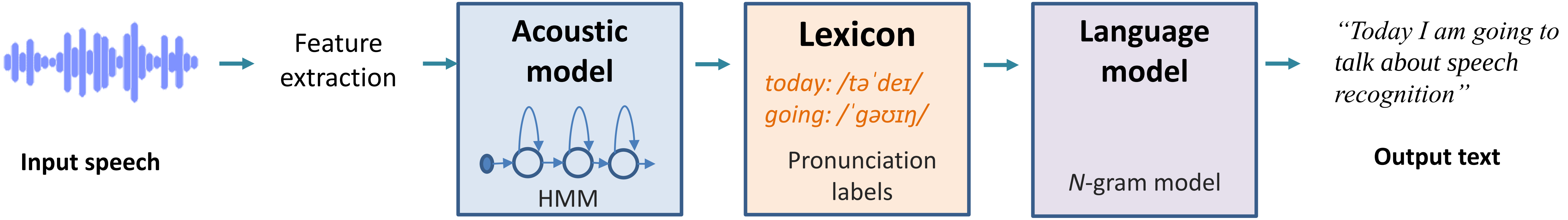
Neural Incremental Speech Recognition

Incremental Speech Recognition

Related Works

A. Statistical approach (Pipeline)

- Hidden Markov model (HMM) ASR [Rabiner, 1989; Juang and Rabiner, 1991]
- 3 parts: Acoustic model, lexicon, language model



- Low delay speech recognition by performing left-to-right input processing (unidirectional)
- Not end-to-end

Neural Incremental Speech Recognition

How to achieve an ISR system that can:

1. reduce delay,
2. keep the system complexity, and
3. maintain a close performance

of the standard neural ASR system?

Proposal

Neural ISR construction by employing sources (architecture, knowledge) from standard neural ASR.

Neural Incremental Speech Recognition

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura,

“Sequence-to-sequence Learning via Attention Transfer for Incremental Speech Recognition”,

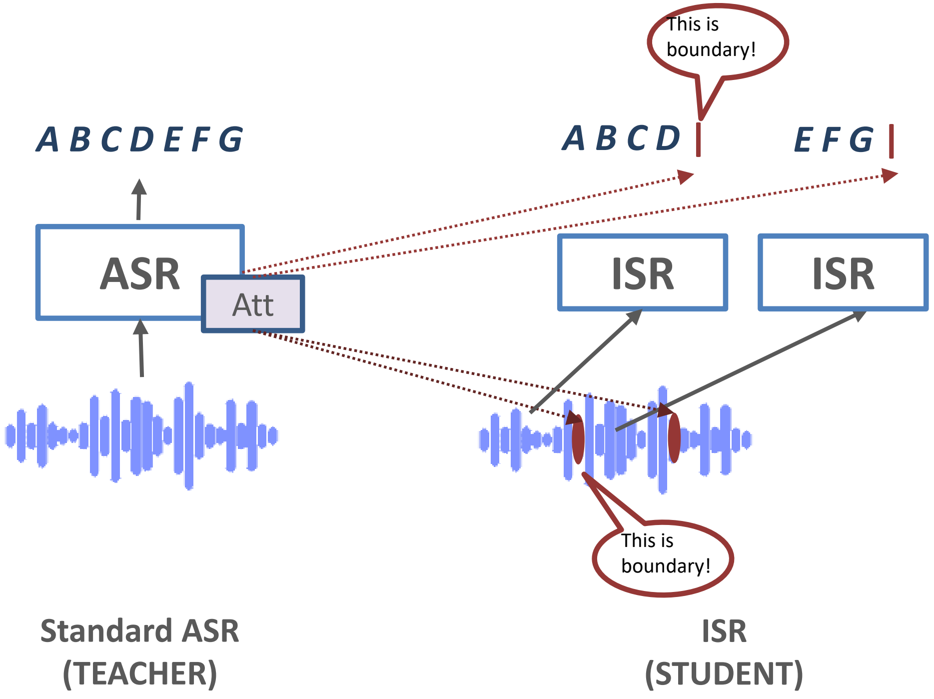
Interspeech 2019

Attention-Transfer ISR

Attention-Transfer Incremental Speech Recognition (AT-ISR)
[Novitasari et al., 2019]

- **Aim**
  ISR (student) learns to mimic the attention-based alignment generated by a standard seq2seq ASR (teacher)
  - ISR architecture : Same as the teacher (seq2seq)
  - Incremental step : Learn through attention transfer from the teacher ASR

AT-ISR Training



- **Attention transfer** : Attention knowledge transfer from teacher to student model
  - Prev. works → image recognition tasks
    - Teach another model [Zaguruyko and Komodakis, 2017]
    - Domain transfer (image to video) [Li et al., 2017]
  - Has not been utilized for ISR construction yet

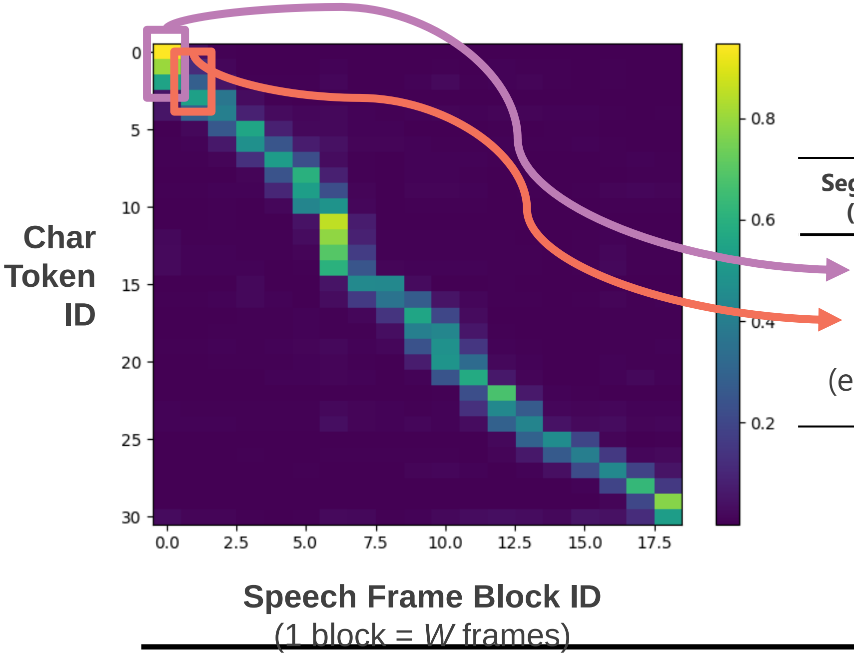
Attention-Transfer ISR

Attention Transfer

Train ISR (student) to learn the attention-based alignment from a standard seq2seqASR (teacher)

1) Extract speech-text alignment from attention matrix generated by the teacher ASR during teacher-forcing text generation (alignment pair = high attention score):

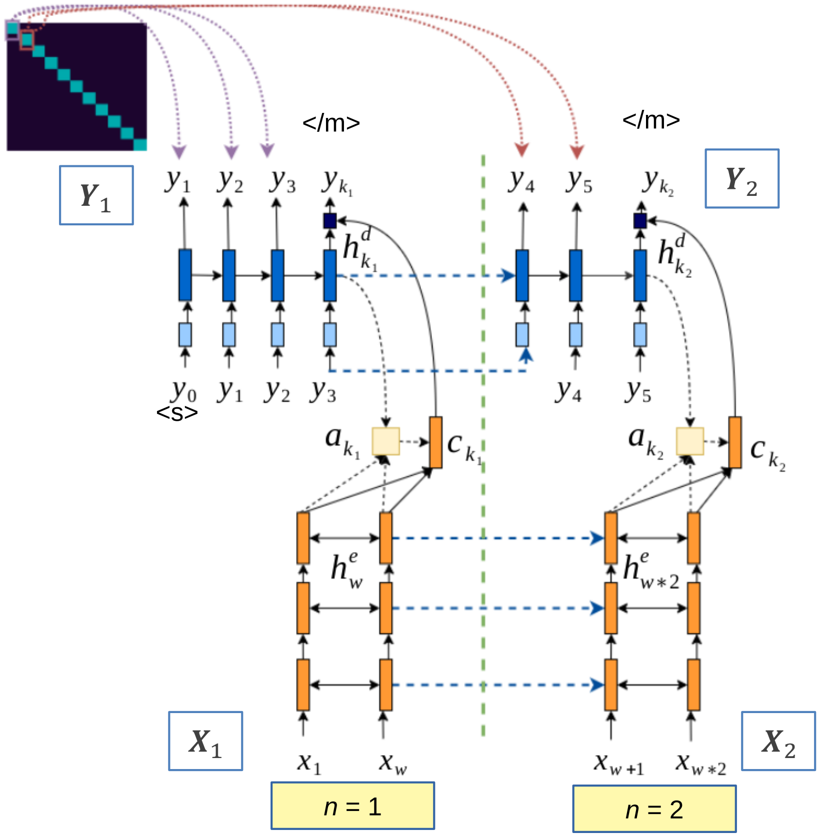
<u>**Teacher ASR attention matrix**</u>



Alignment

Seg. ID (n)	Speech seg. (X_n)	Text seg. (Y_n)
1	$x_1 - x_w$	$y_1 - y_3$
2	$x_{w+1} - x_{w+2}$	$y_4 - y_5$
(etc.)		

2) Train the ISR by using Y_n + </m> as the target of X_n



ISR delay can be managed by changing X_n and Y_n size during training

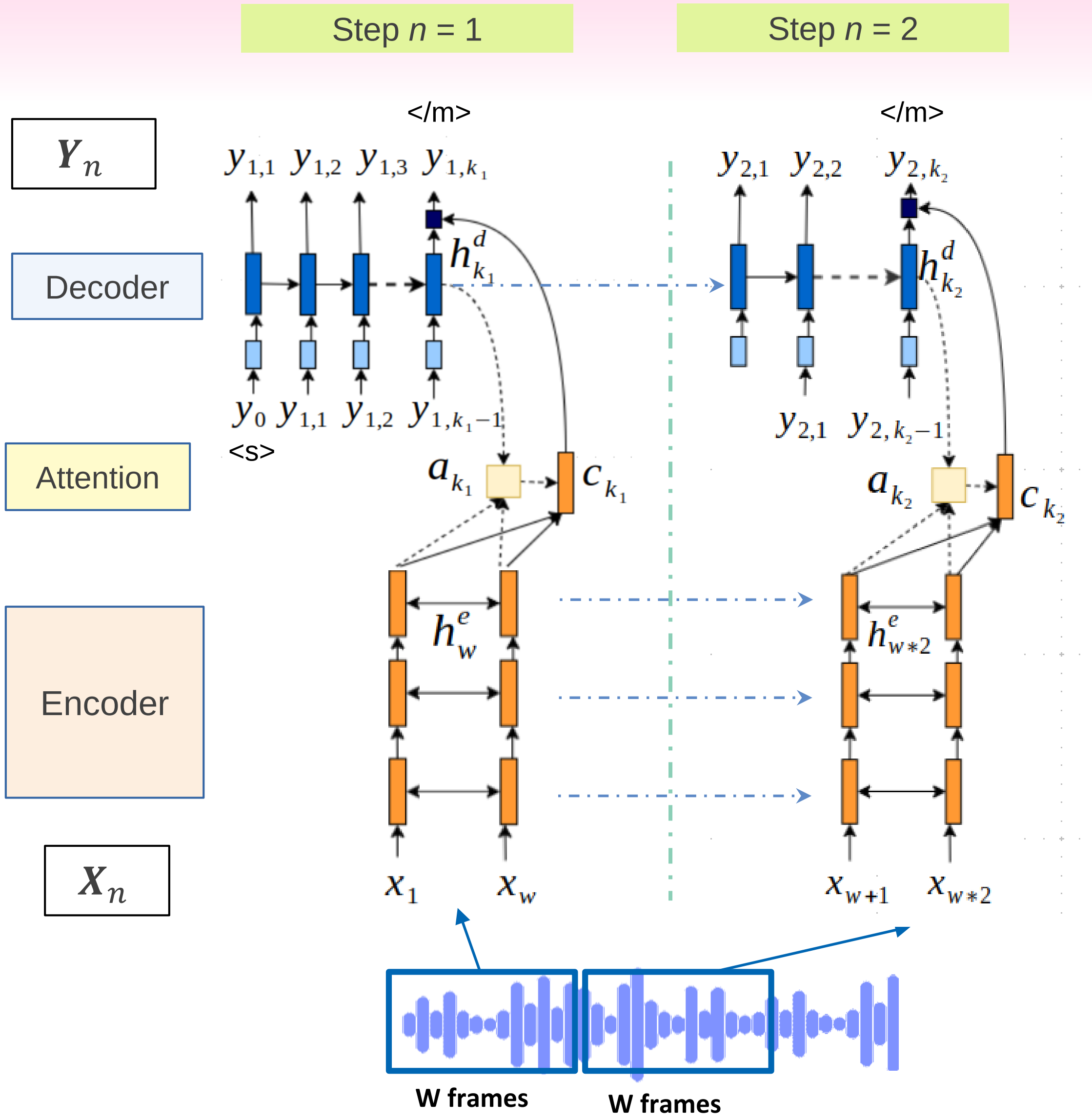
e.g. higher delay : combine several segments into one

Attention-Transfer ISR

AT-ISR Recognition Method

- Given: Full speech ($\boldsymbol{X}$), length S
- Recognize the speech segment-by-segment sequentially based on a fix-sized input window
- For each incremental recognition step n:
  1. **Encode** $\boldsymbol{X}_n$, a W speech frames from $\boldsymbol{X}$ $(W < S)$
  2. **Decode** for $\boldsymbol{Y}_n$ that aligns with $\boldsymbol{X}_n$, until an *end-of-block* *</m>* token is predicted or max. length is reached
     - **Attend** the input $\boldsymbol{X}_n$
  3. **Shift** the input window W frames by keeping the model's state

  (Total step number: $N = \frac{S}{W}$)
- Incremental step:
  - Input boundary : last speech frame in the input window
  - Output boundary : *</m>* token in the output text
- Alignment learning → Attention transfer

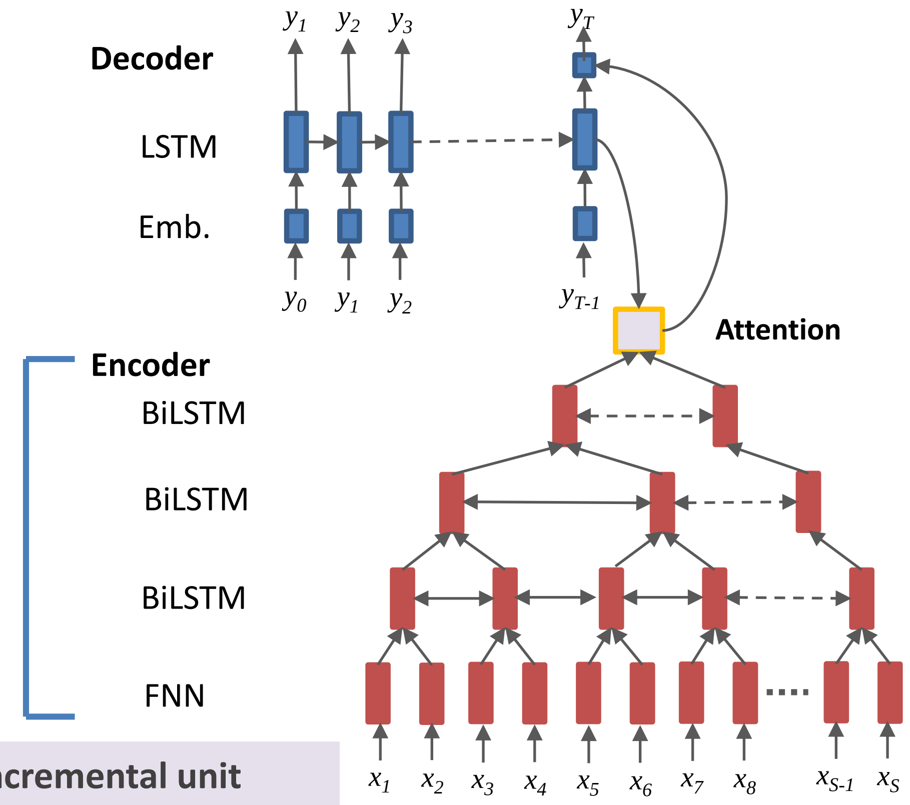


AT-ISR Performance

Model Configuration

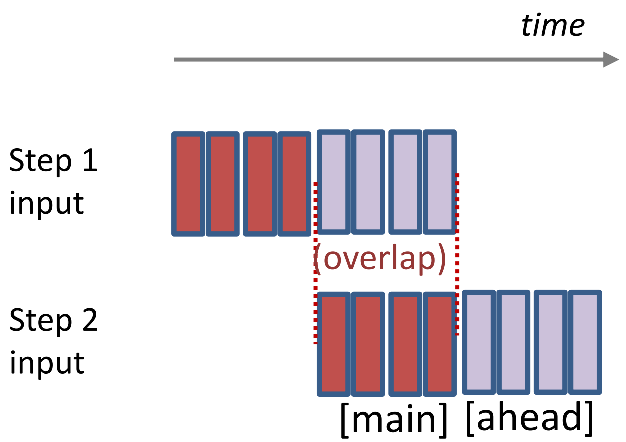
- AT-ISR/Teacher ASR structure: Seq2seq (identical)

Hierarchical sub-sampling
8 speech frames into 1 encoder state



AT-ISR basic incremental unit
8 speech frames = 1 block (0.14 sec)

- AT-ISR with input overlap :
  - Main frames : Aligns with output text seg.
  - Look-ahead frames : Next to the main input (contextual input)



AT-ISR Performance

Evaluation Setting

ISR performance evaluation was made by comparing various model:

- **Non-incremental ASR :** Topline
  - Standard seq2seq ASR (Our Att Enc-Dec; teacher)
  - Other existing neural ASR

- **Incremental ASR:**
  - Baseline neural ISR:
    - Seq2seq ISR without attention transfer
    - Incremental steps were taught by using alignments from forced-alignment by HMM ASR
  - Proposed ISR: AT-ISR (attention transfer; student)
  - Other existing neural ISR: Unidirectional LSTM + CTC [Hwang and Sung, 2016]

Evaluation metric:
- CER, WER
- Delay (speech input size)

AT-ISR Performance

Speech recognition performance of character-level models trained on WSJ dataset

Model	Delay (sec)		CER (%)
	Input	Computation	
Non-incremental ASR (Topline)			
Att Enc-Dec (ours)	7.88 (avg)	0.32 (avg)	6.26
BiLSTM-CTC [1]			8.97
Joint CTC+Att [1]			7.36
Baseline neural ISR			
Input/step: 1 *m* + 1 *la*	0.24	0.02	20.15
Input/step: 1 *m* + 4 *la*	0.54	0.05	11.95
Proposed AT-ISR			
Input/step: 1 *m* + 1 *la*	0.24	0.02	**18.37**
Input/step: 1 *m* + 4 *la*	0.54	0.05	**7.52**
Other existing neural ISR			
LSTM-CTC beam search [2]	-	-	10.96

CER diff.: 1.3%

Result

- Avg. utterance length: 7.88 sec
- Machine: Intel® CoreTM i7-9700K CPU @ 3.60GHz (NVIDIA GeForce RTX 2080Ti GPU)
- ISR performance limitation: short-segment-based recognition (incomplete information)
- Contextual input (*la*) improves performance

AT-ISR performs well with a short delay by learning non-incremental ASR's knowledge

*Note

m = main input block

la = look-ahead block (contextual input)

1 block = 8 frames = 0.14 sec

[1] Suyoun Kim, Takaaki Hori, and ShinjiWatanabe. Joint CTC-attention based end-to-end speech recognition using multitask learning. In Proceedings of ICASSP, pages 4835-4839, New Orleans, USA, 2017.

[2] Kyuyeon Hwang and Wonyong Sung. Character-level incremental speech recognition with recurrent neural networks. In Proceedings of ICASSP, pages 5335 - 5339, Shanghai, China, 2016.

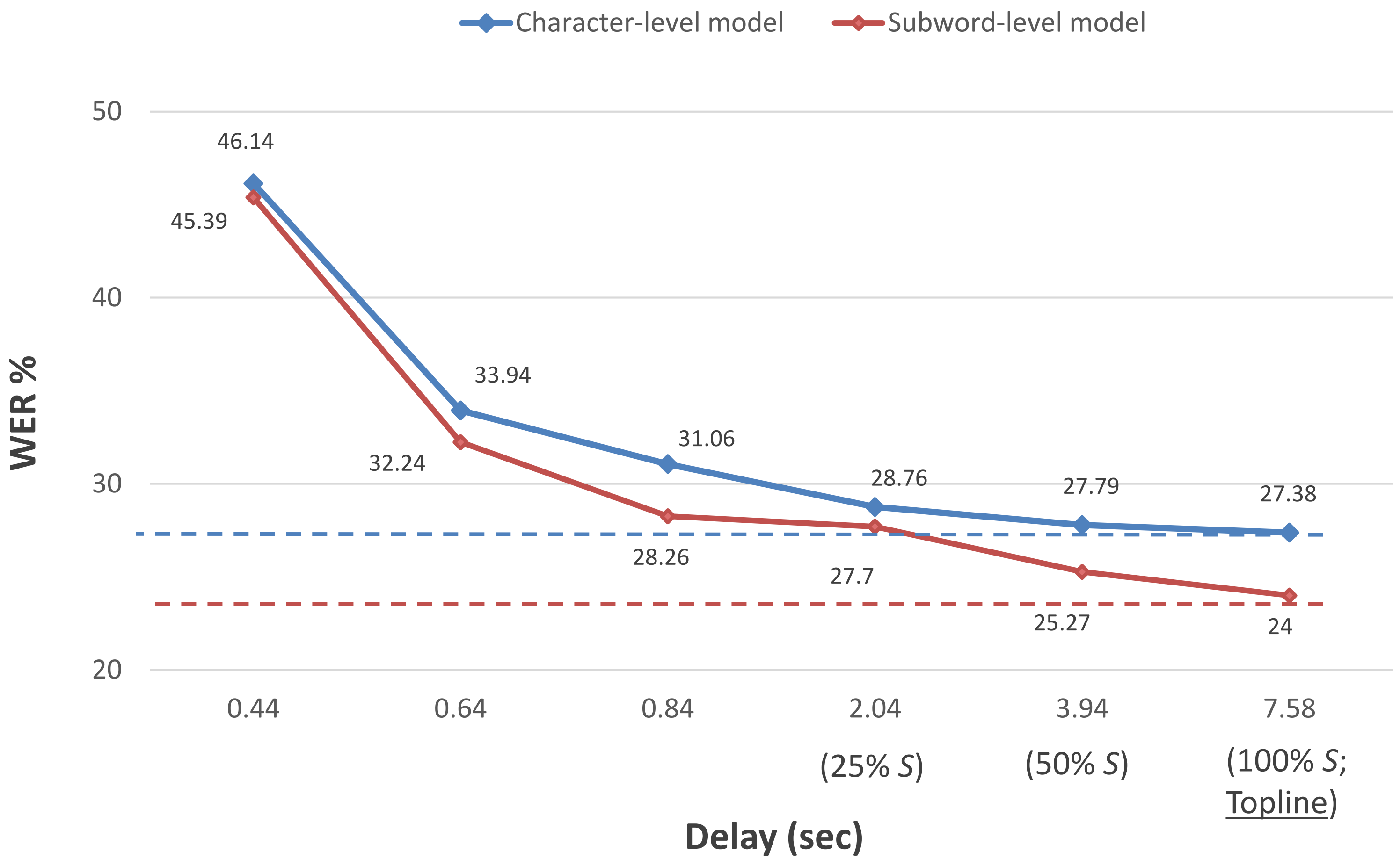
AT-ISR Performance and Delay

ISR Delay

How did the ISR delay affected the ISR performance?

- Trade-off: Higher delay, lower WER
- **Subword-level ISR**
  - Lower WER than character-level ISR
  - Keep word context longer than characters
- **Character-level ISR**
  - Maintains the teacher's performance better than the subword-level ISR
  - ISR with delay 2.04 starts to have a close performance to the teacher ASR

WER (%) of AT-ISR trained on TED-LIUM dataset



*S = average full-utterance length (7.58 sec)

Neural Incremental Speech Recognition

Summary – AT-ISR

Neural ISR system (AT-ISR) with a low recognition delay without increasing the complexity of the standard ASR system

1. AT-ISR with delay < 1 sec. achieved a close performance to standard ASR with delay > 7 sec.
2. AT-ISR as an ISR framework with an efficient development mechanism and reliable performance via attention transfer that applies an identical architecture as the standard ASR

Recent ISR Trend

- Streaming ASR with RNN-Transducer (RNN-T) [Saitnah et al., 2020; Li et al., 2020]
- Streaming transformer ASR [Miao et al., 2020; Moritz et al., 2020; Tsunoo et al., 2020]

Topics

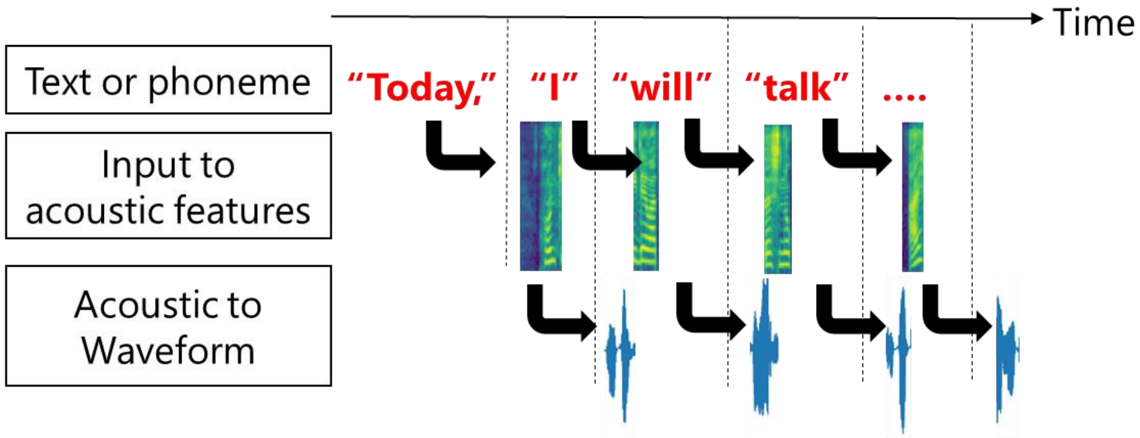
- Machine Speech Chain
  - ASR and TTS research
  - ASR & TTS semi-supervised joint learning
- Neural Incremental ASR and TTS
  - Neural Incremental ASR
  - Neural Incremental TTS
- Incremental Speech Chain
  - Incremental Learning of Speech Chain
- Summary

Incremental Text-to-speech

Incremental Text-To-Speech(iTTS)

Speech is synthesizes a speech in shorter delay (e.g. word).

It can synthesize a speech before finishing text input.



Challenges

How to improve speech quality?

Speech quality of Incremental TTS

How to estimate target prosody from an incomplete sentence?

target prosody is typically calculated from long-window features. (e.g. co-articulation)

-> predicts next information(e.g. word) at step of input-to-acoustic-features.

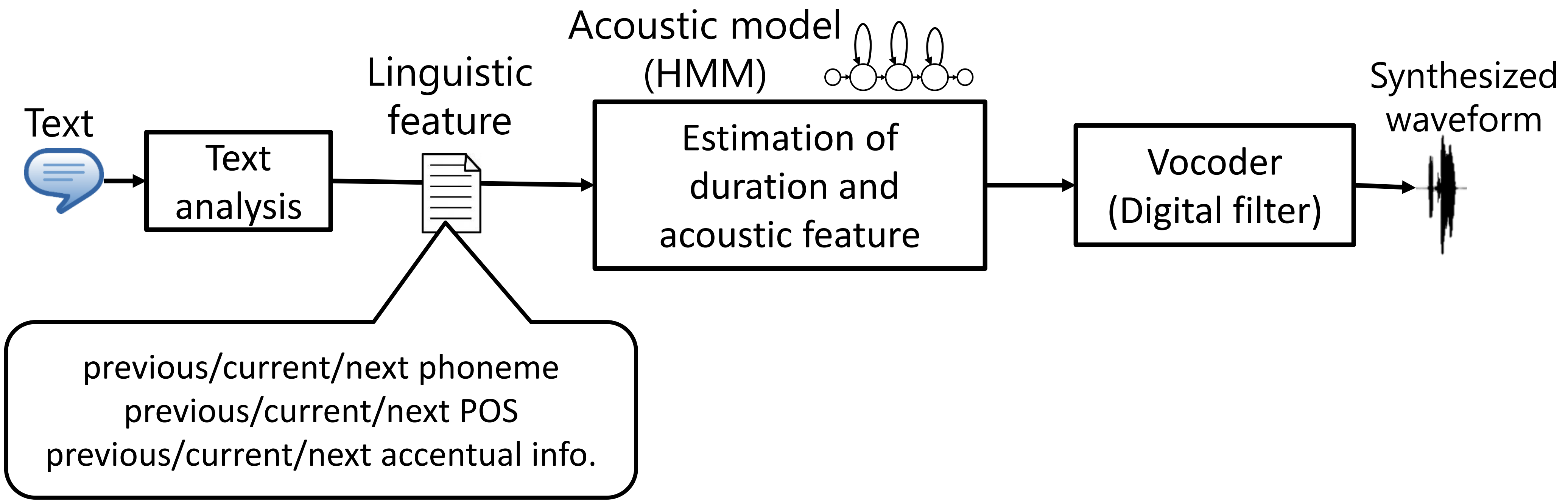
-> wait next word when synthesizing a current word.

Related Works of iTTS(1/2)

Statistical approach (pipeline)

Hidden Markov model TTS[Baumann et al., 2014.],[Pouget et al., 2015],[Yanagita, et al., 2018]

No neural End-to-end iTTS approach



Related Works of iTTS(2/2)

End-to-end TTS [Wang, et al., 2017.], [Sotelo, et al., 2017], [Shen, et al., 2018.]

Encoder-decoder with an attention mechanism

-> Output prediction starts after the input sequence.

The speech is also synthesized Sentence-by-sentence.

-> **It can generate High quality speech close to human.**

Challenge of the neural iTTS system.

More natural synthesized speech

Neural iTTS[Yanagita et al., 2019]

no wait next word for synthesis.

Control output sequence with stop flag

Prefix-to-Prefix Framework [Ma et al., 2020]

wait next word for synthesis.

Control output sequence with attention weight and stop flag

One word look ahead at least for synthesis.

Neural iTTS

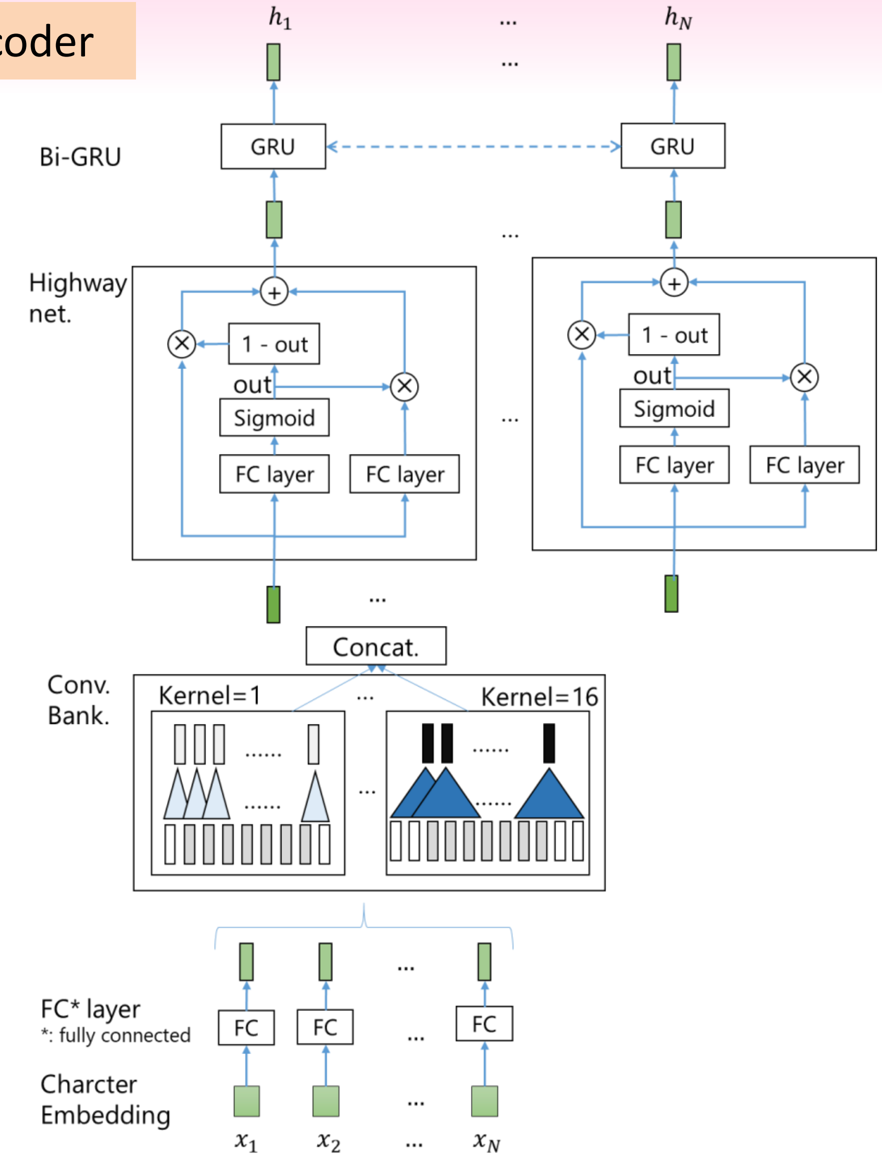
Tomoya Yanagita, Sakriani Sakti and Satoshi Nakamura,

"Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework",

10th Speech Synthesis Workshop (SSW10) , Sep. 2019

End-to-End TTS

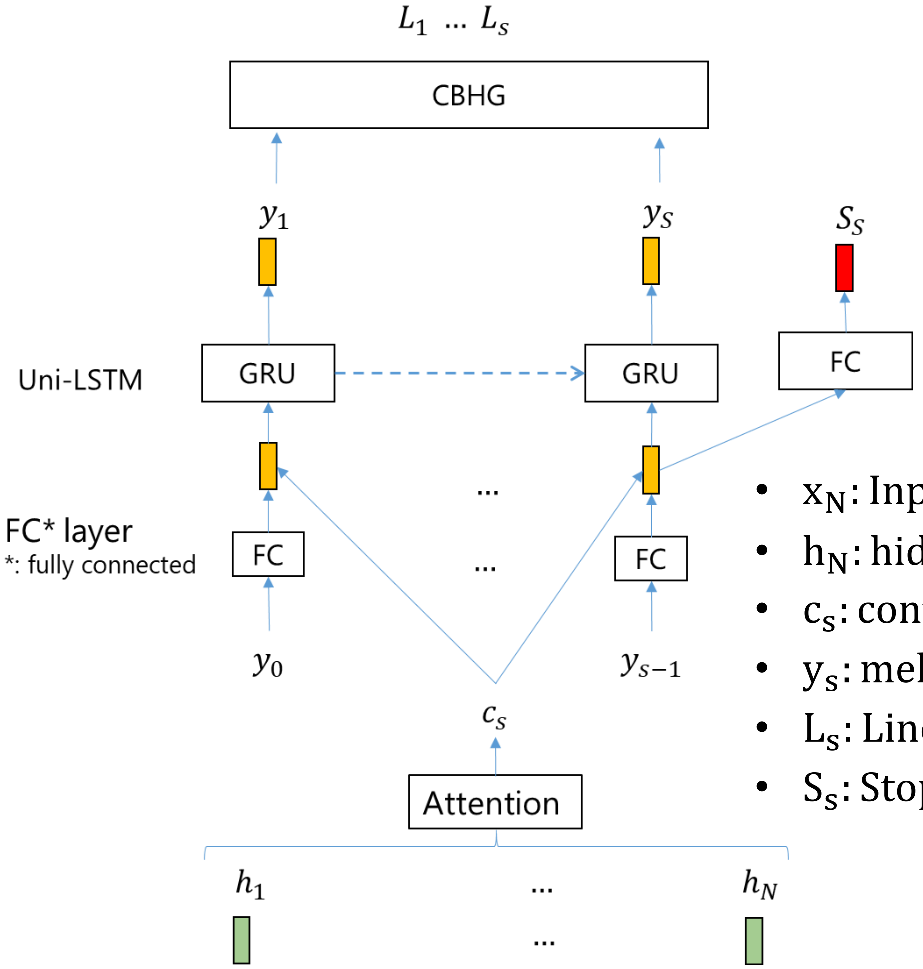
Encoder



Decoder with attention

We use Tacotron[Wang, et al., 2017.].

Stop flag prediction to control output seq. is also used.



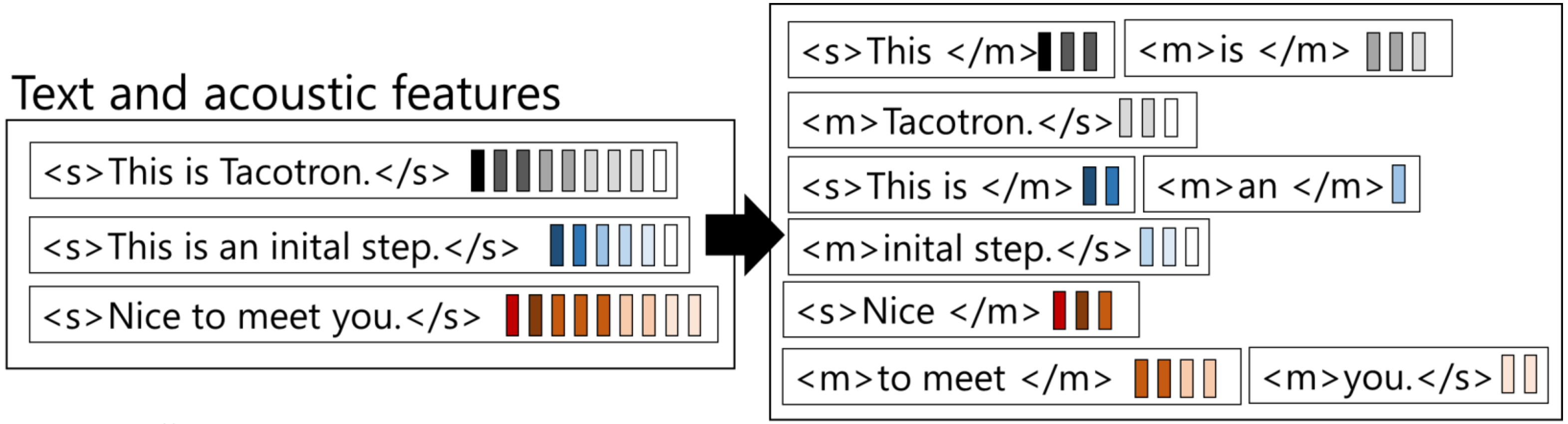
- x_N: Input sequence (N length)
- h_N: hidden representatiaon of encode
- c_S: context vector (S length)
- y_S: mel spectrogram
- L_S: Linear spectrogram
- S_S: Stop flag

Proposed method

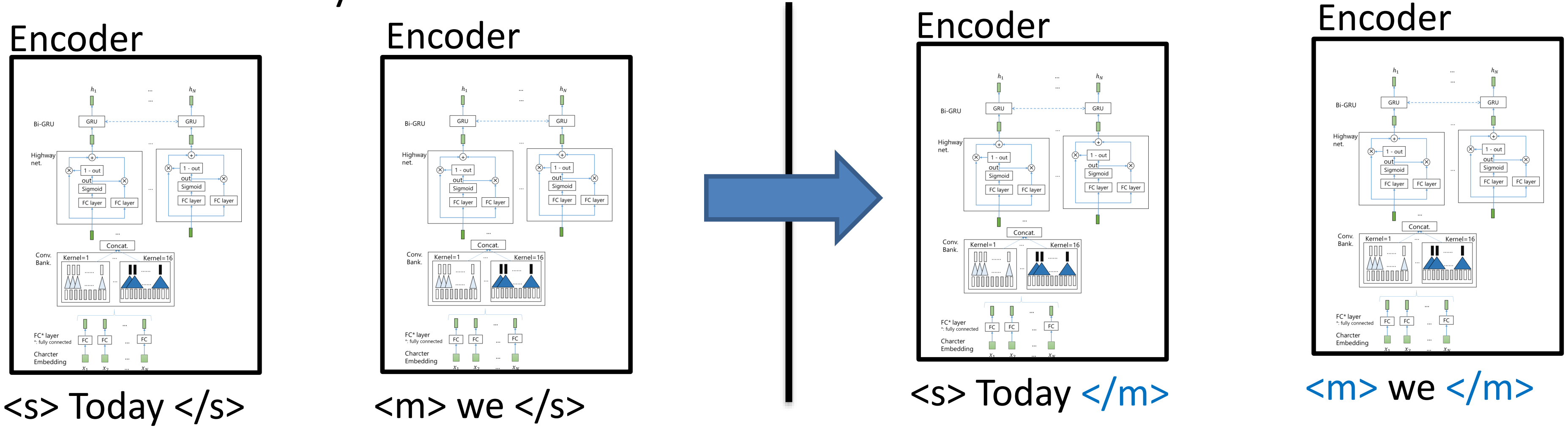
<s>: sentence start
</s>: sentence end
<m>: middle sentence start
</m>: middle sentence end

Dataset sentences are divided into three parts.

- use location symbol to indicate locations
- use all data for training



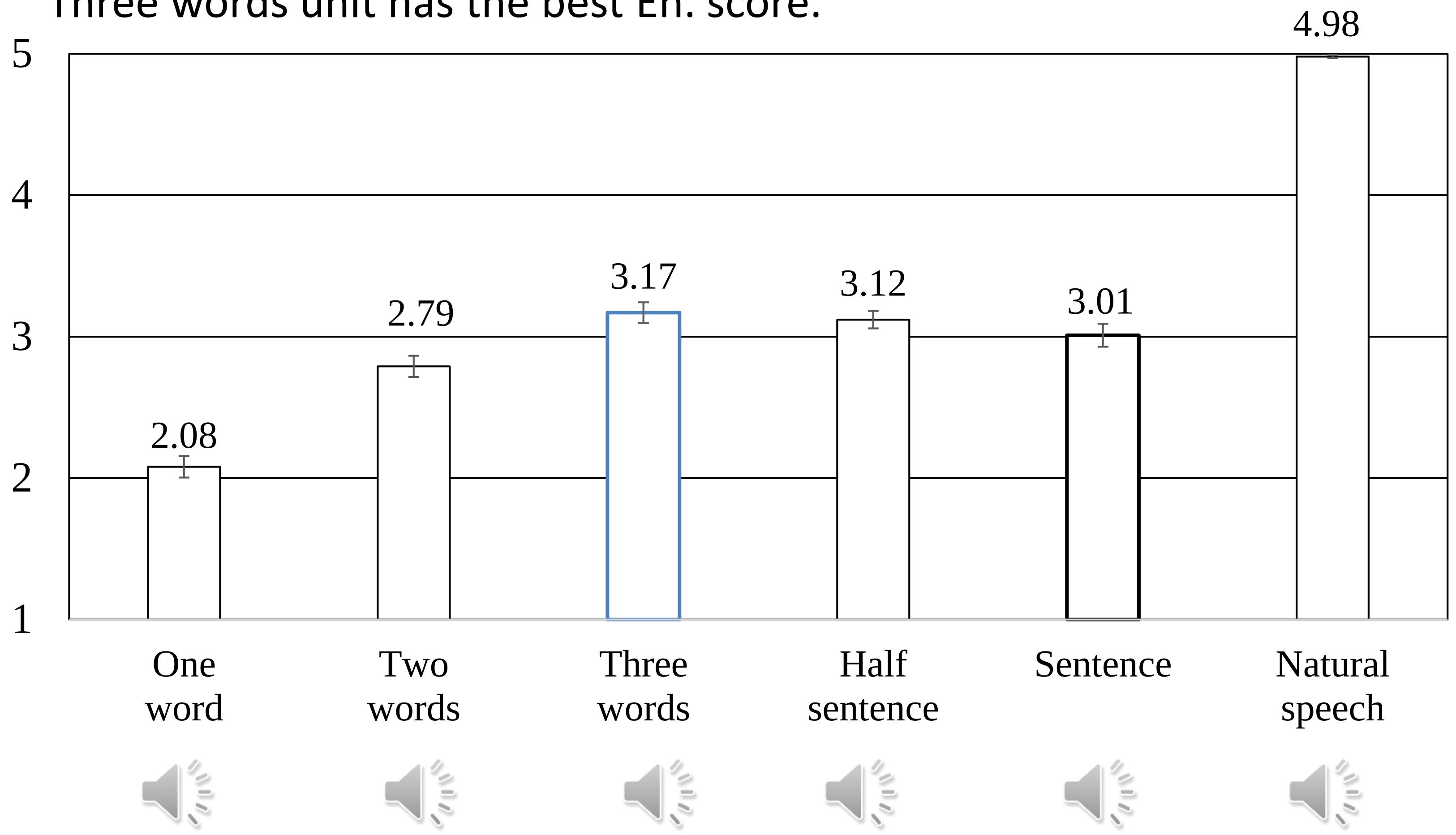
Inference: Ex. “Today we talk about TTS.”



Result of English MOS

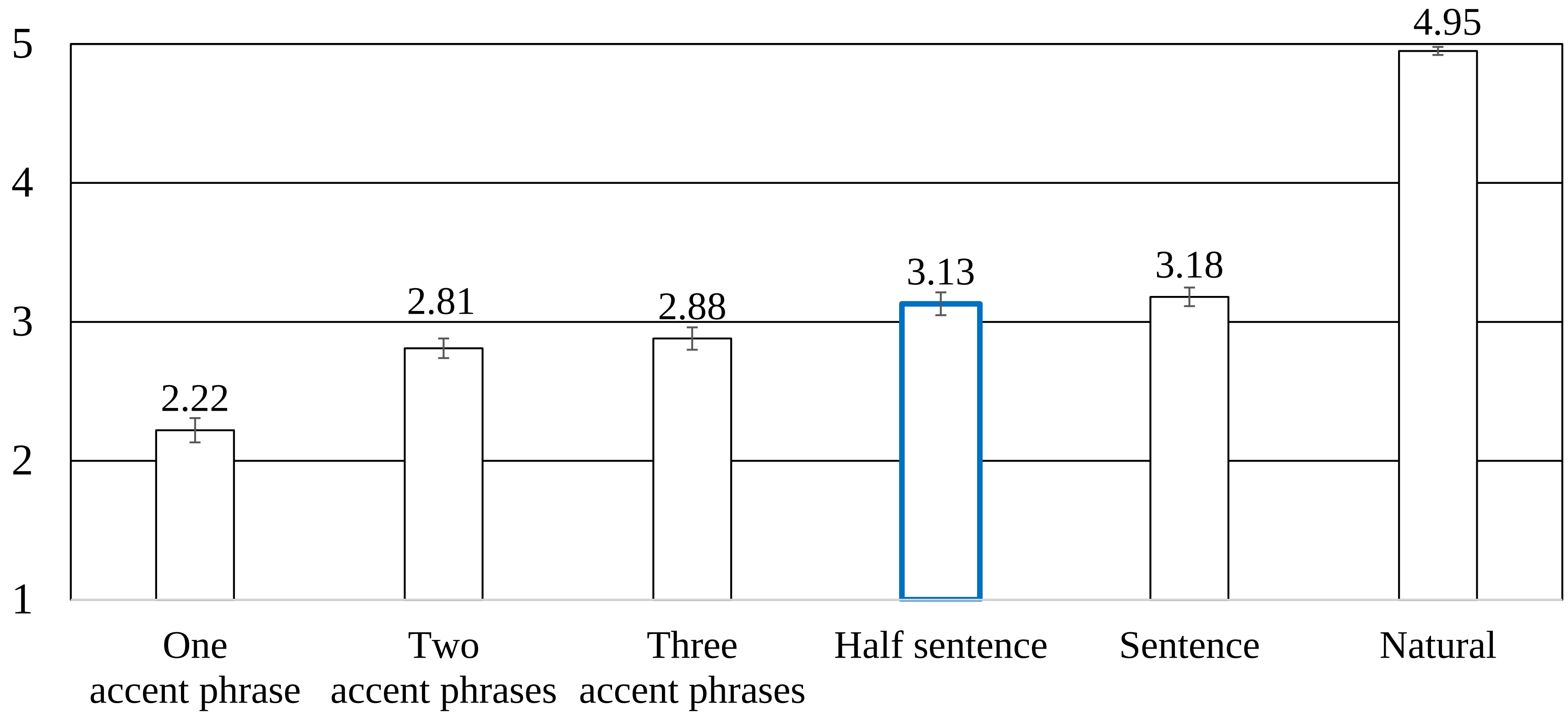
Still big gap between natural speech and synthesized speech.

Three words unit has the best En. score.



Result of Japanese MOS

Still big gap between natural speech and synthesized speech.
Half sentence unit ≒ the full sentence units (Ja.).



Topics

- Machine Speech Chain
  - ASR and TTS research
  - ASR & TTS semi-supervised joint learning
- Neural Incremental ASR and TTS
  - Neural Incremental ASR
  - Neural Incremental TTS
- Incremental Speech Chain
  - Incremental Learning of Speech Chain
- Summary

Incremental Speech Chain

Sashi Novitasari, Andros Tjandra, Tomoya Yanagita,
Sakriani Sakti and Satoshi Nakamura,

"Incremental Machine Speech Chain Towards Enabling Listening while Speaking in Real-time",

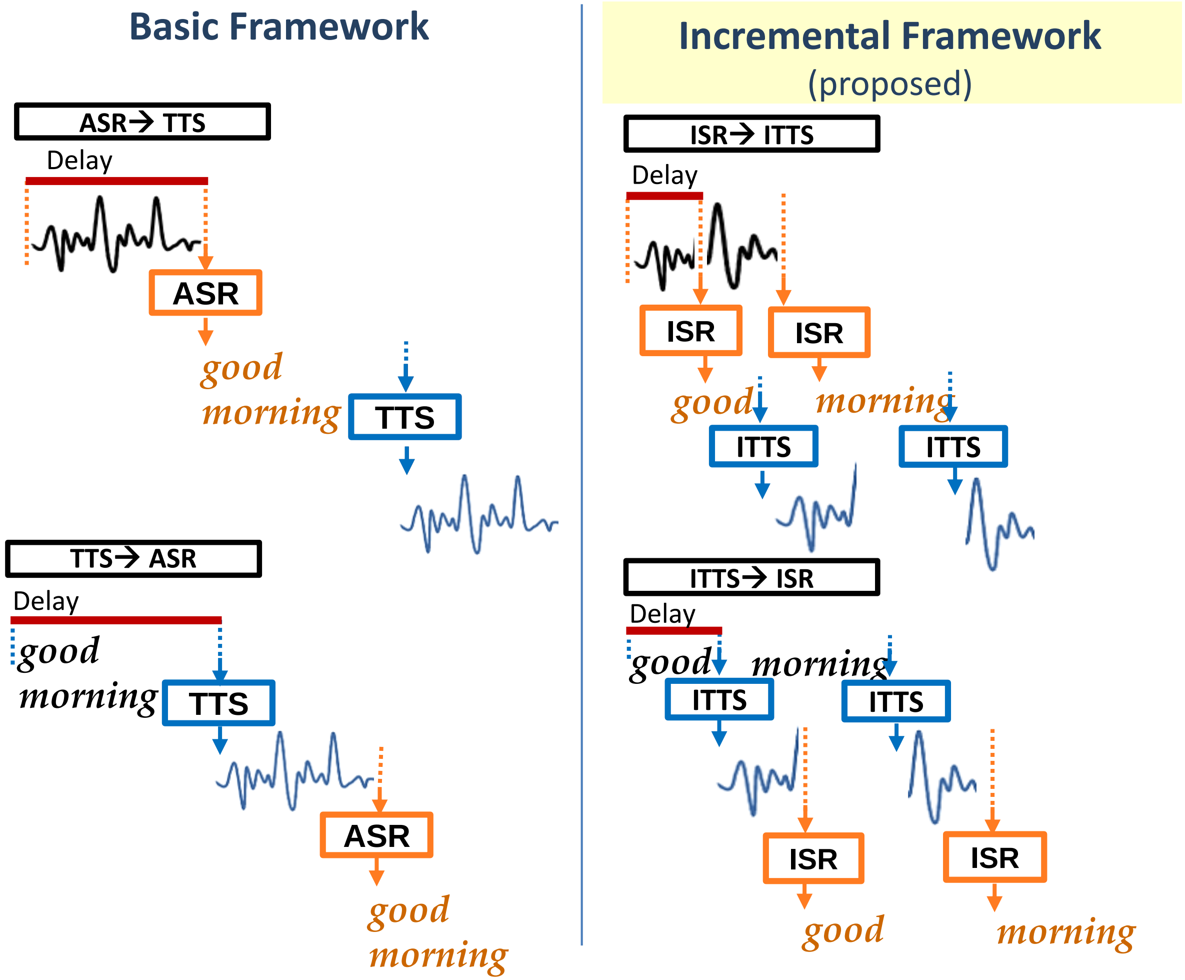
INTERSPEECH 2020

Incremental Machine Speech Chain

Closed short-term feedback loop between incremental ASR (ISR) and incremental TTS (ITTS)

- Reduce feedback delay within machine speech chain training
- Improve ISR and ITTS learning quality
- Enable immediate feedback generation during inference

Move a step closer for ASR and TTS that can adapt to real-time environment unsupervisedly
→ **Similar to human**



Unrolled processes in machine speech chain loop

Incremental Machine Speech Chain: Training

Two training phases:

1. ISR and ITTS supervised-independent training

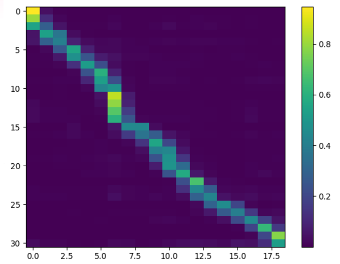
2. ISR and ITTS joint training via short-term feedback loop

ISR and ITTS Independent Training

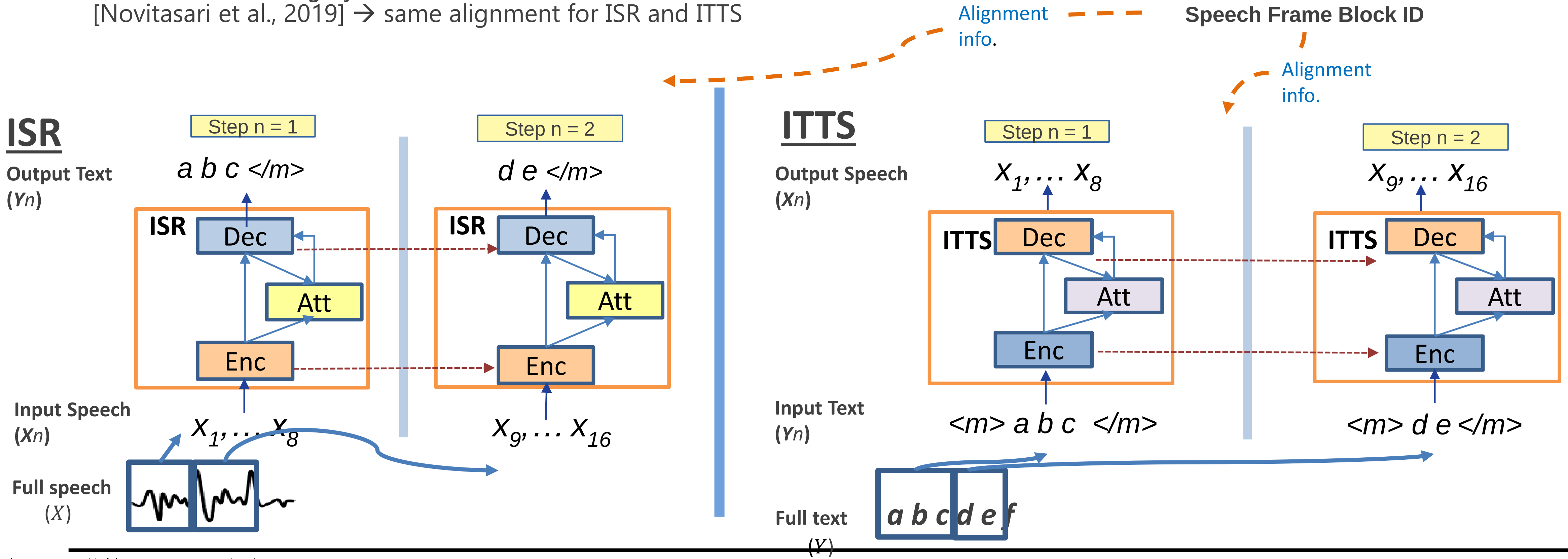
- Incremental : Predict a complete output sequence in *N* steps, for each step *n* :
  1. Encode a segment of input from input window
  2. Decode and predict a segment of output
  3. Shift the input windows
- ISR and ITTS training by attention transfer from non-incremental ASR [Novitasari et al., 2019] → same alignment for ISR and ITTS

Attention alignment from non-incremental ASR

Text Token ID



Speech Frame Block ID



Learning Approach

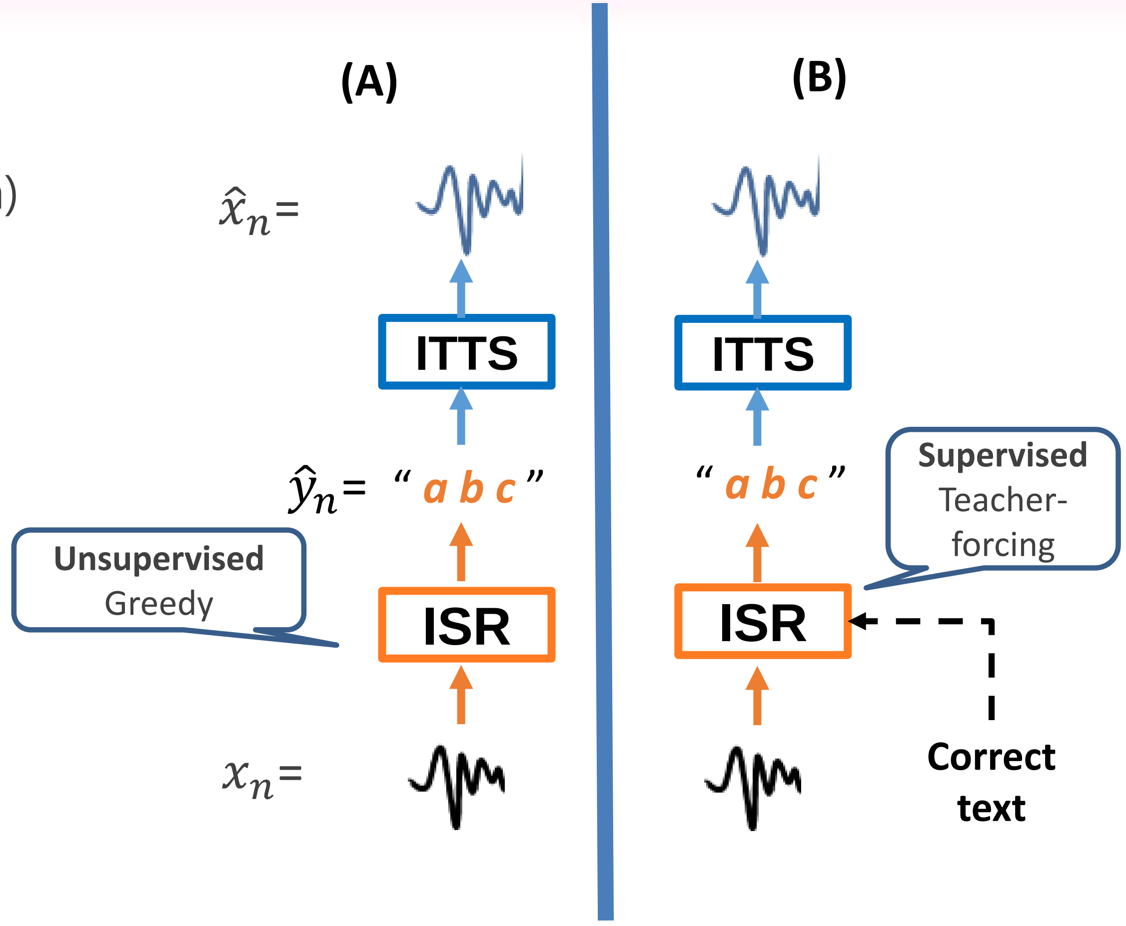
Exploration on 2 learning approaches:

A) Semi-supervised incremental machine speech chain

1) ISR/ITTS independent training : supervised
2) ISR/ITTS joint training: unsupervised (unlabeled data)
→ Same as the original basic machine speech chain

B) Supervised incremental machine speech chain

1) ISR/ITTS independent training : supervised
2) ISR/ITTS joint training : supervised (labeled data)



Unrolled process examples in joint training
(ITTS-to-ISR follows similar mechanism)

Speech Chain Result (SI-84)

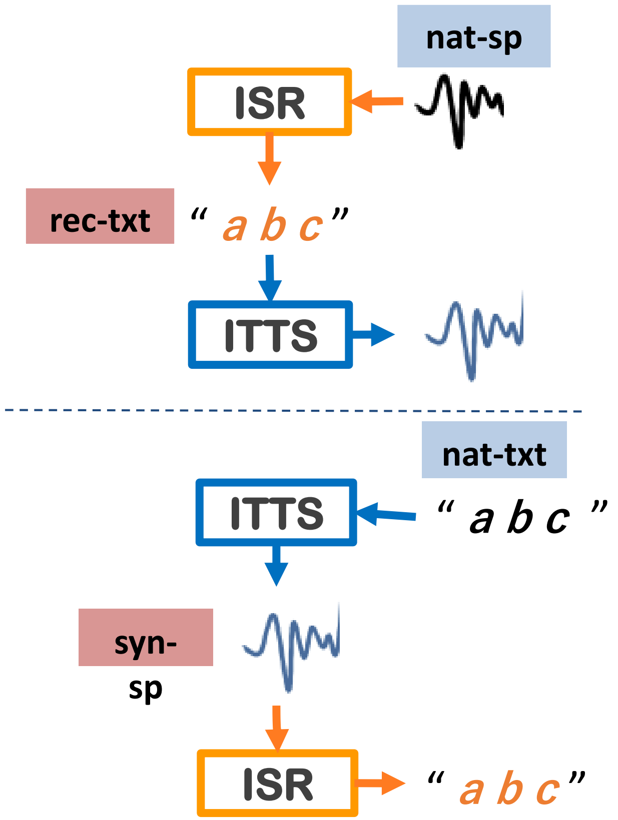
ASR (CER%) and TTS (log Mel-spectrogram L2 loss) performances

Data	ASR (CER%)				TTS (L2-norm)^2			
	Std. (delay: 7.88 sec)		Incr. (delay: 0.84 sec)		Std. (delay: 103 chars)		Incr. (delay: 30 chars)	
	nat-sp	*syn-sp*	*nat-sp*	*syn-sp*	*nat-txt*	*rec-txt*	*nat-txt*	*rec-txt*
Independent Training								
Indep-trn *SI-84*								
Indep-trn *SI-284*								
Machine Speech Chain								
Indep-trn (SI-84) + chain-trn-greedy (SI-200)								
Indep-trn (SI-84) + chain-trn-teachforce (SI-200)								

□ Incremental machine speech chain

- Incremental system able to reduce delay with a close performance to non-incremental system
- Incremental machine speech chain improves ISR and ITTS

- **Baseline**:
  ISR and ITTS *indep-trn SI-84*
- **Topline**:
  Standard systems (std.) *indep-trn SI-284*
- **Input type**:



Incremental Machine Speech Chain

Incremental machine speech chain

Short-term feedback loop for ISR/ITTS development by mimicking human speech chain

- Reduced the delay with a close performance to the basic framework
- Improve ISR and ITTS (natural/synthetic input)
- Synthetic input processing: demonstration of real-time feedback generation

A step to achieve system that can listen while speaking in real-time

Summary

- Machine Speech Chain
  - ASR and TTS research
  - ASR & TTS semi-supervised joint learning
- Neural Incremental ASR and TTS
  - Neural Incremental ASR
  - Neural Incremental TTS
- Incremental Speech Chain
  - Incremental Learning of Speech Chain
- Future work
  - Semi-supervised learning and online incremental learning
  - Much lower delay
  - Application to Incremental Dialogue System and Speech Translation System
  - Multi-modality, Code Switching