Towards Developing Neural Machine Speech Interpreter that Listens, Speaks, and Listens while Speaking

Sakriani Sakti

Joint work with:
Andros Tjandra, Sashi Novitasari, Tomoya Yanagita,
Sahoko Nakayama, Johanes Effendi, and Satoshi Nakamura

[1]Nara Institute of Science and Technology (NAIST)
[2]RIKEN Advanced Intelligence Project (RIKEN AIP)

What is Speech Interpreter?

Professional Speech Interpreter

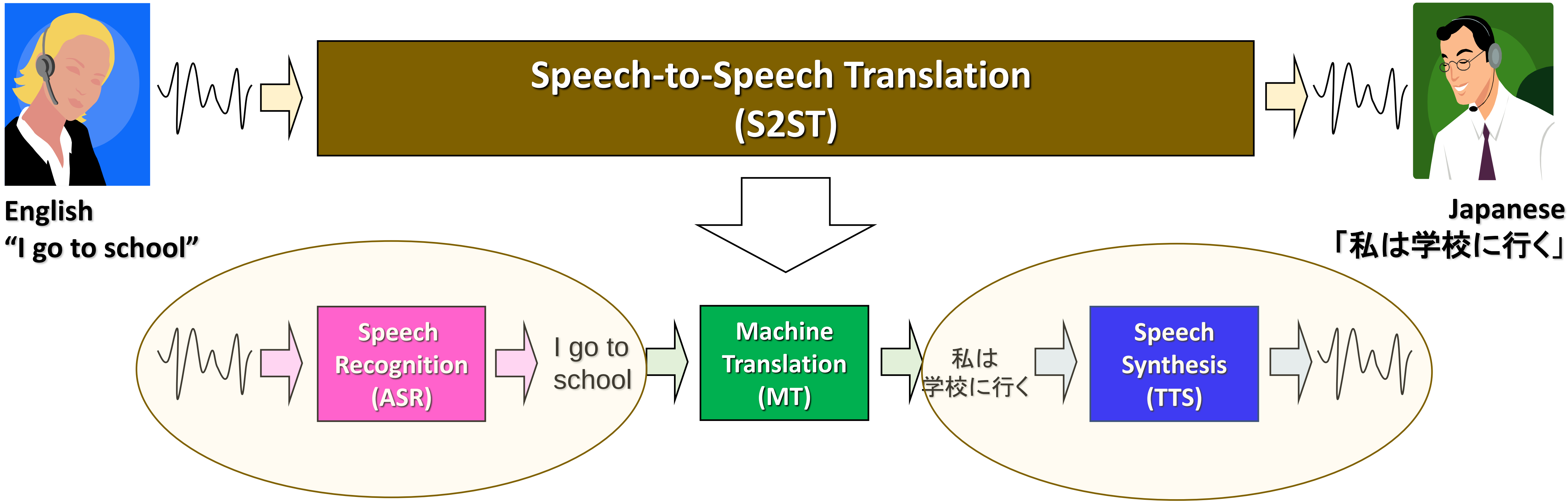
Speaker

Nice to meet you. I am interested in doing business together.

Interpreter

はじめまして。一緒にビジネスをすることに興味があります。

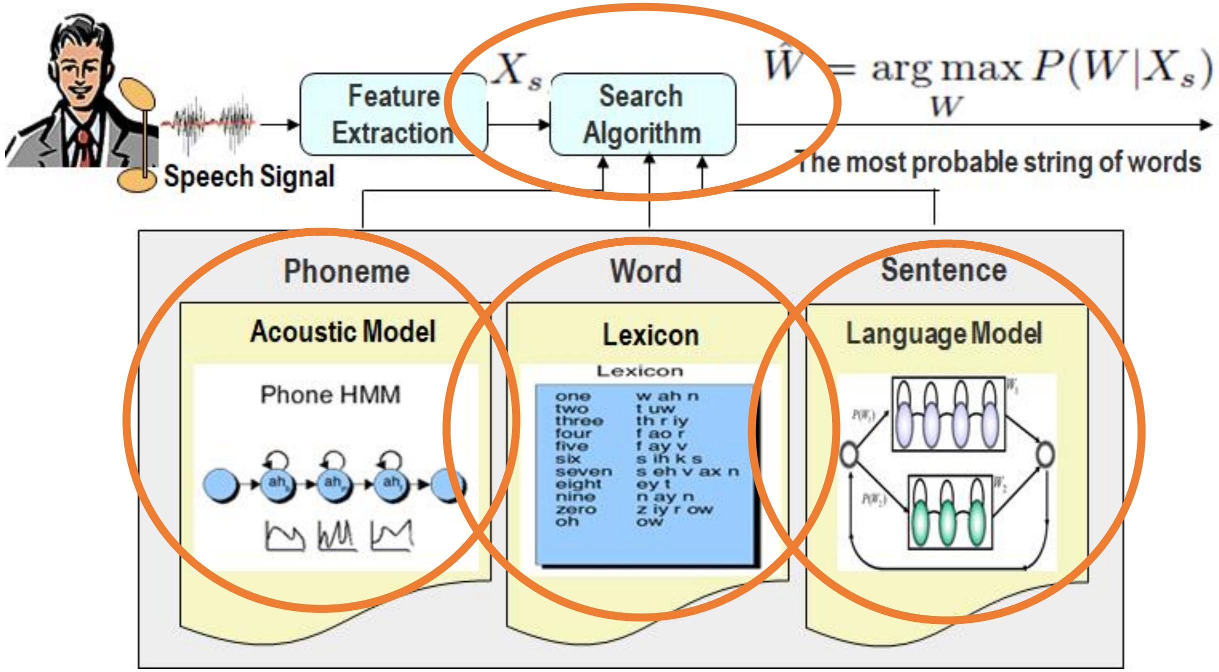
Speech-to-Speech Translation

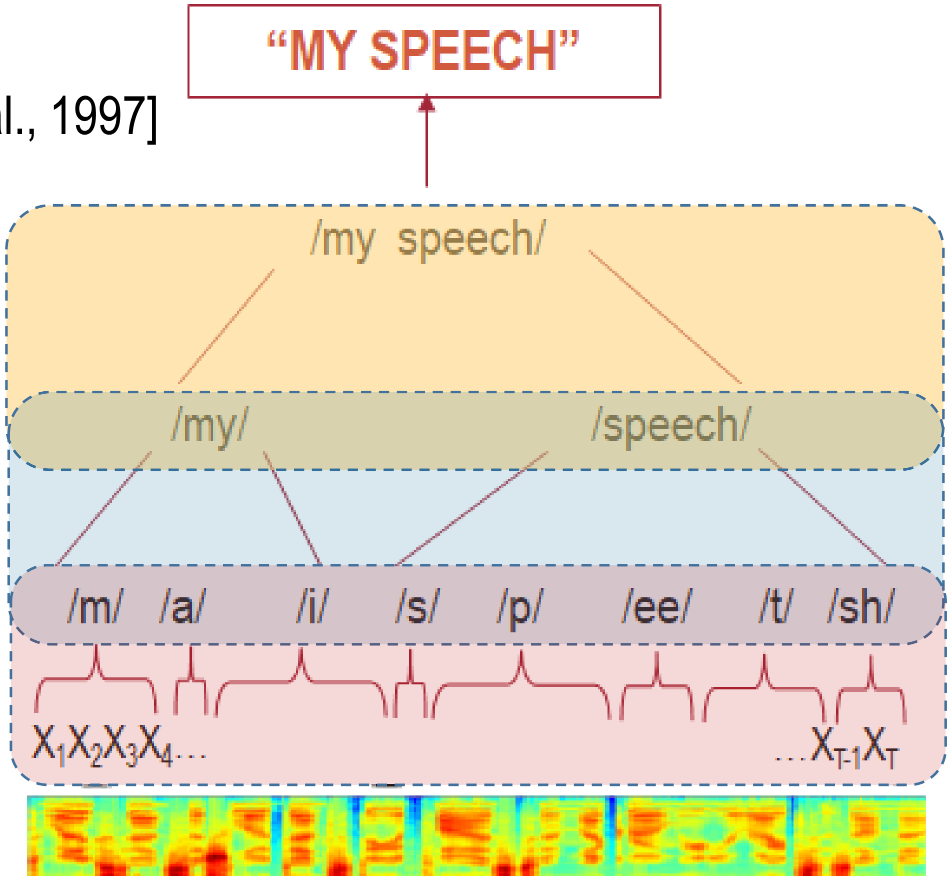


Focus on speech language technologies for S2ST

Automatic Speech Recognition (ASR)

- **Early Technologies for ASR**
  - → Template Matching, Dynamic Programing [Sakoe et al., 1971]
  - → Hidden Markov Modeling [Baum et al., 1966]
  - → Neural Network, TDNN [Waibel et al. 1989], LSTM [Hochreiter et al., 1997]
  - → Weighted Finite State Transducer [Mohri et al., 2002]

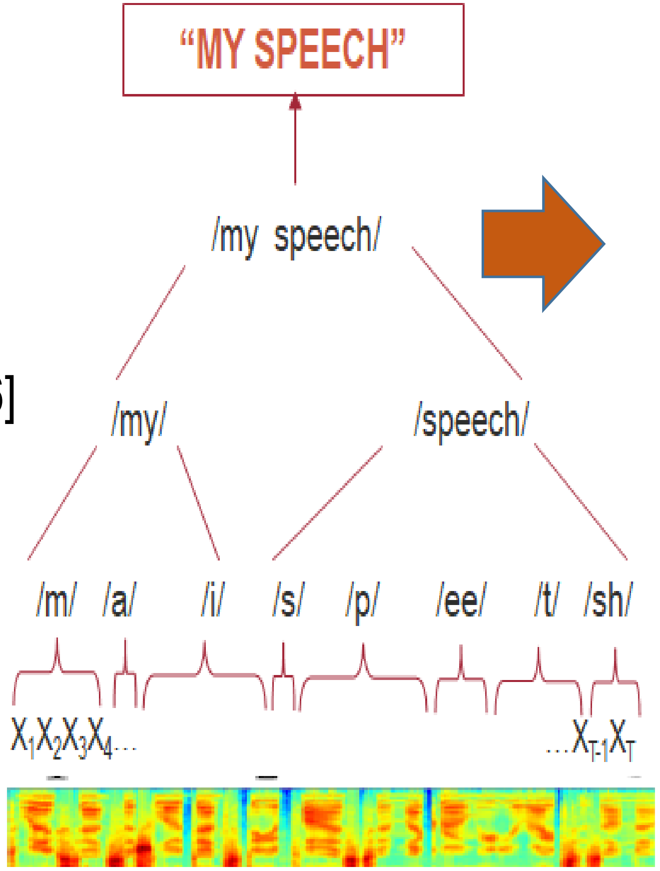


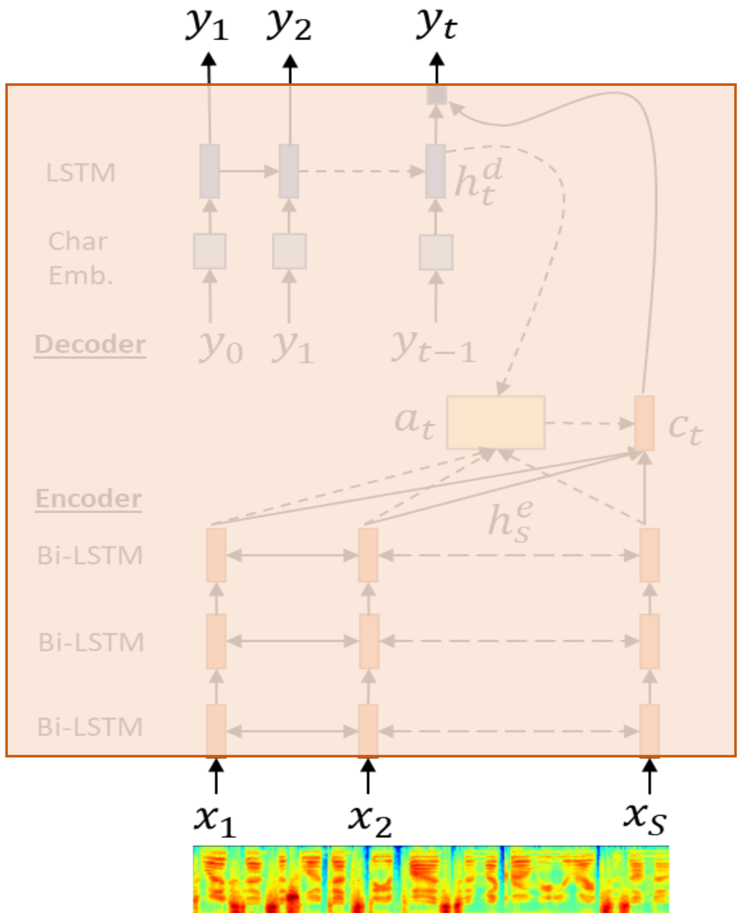


Automatic Speech Recognition (ASR)

- **Recent Technologies for ASR**
  → Hybrid HMM-DNN [Borlard et al., 1993]
  Estimate State Posterior Probability by DNN
  → Connectionist Temporal Classification [Graves et al., 2006]
  Predict Phoneme Label every frame
  → Listen, Attend, and Spell [Chan et al., 2016]
  Sequence-to-sequence modeling

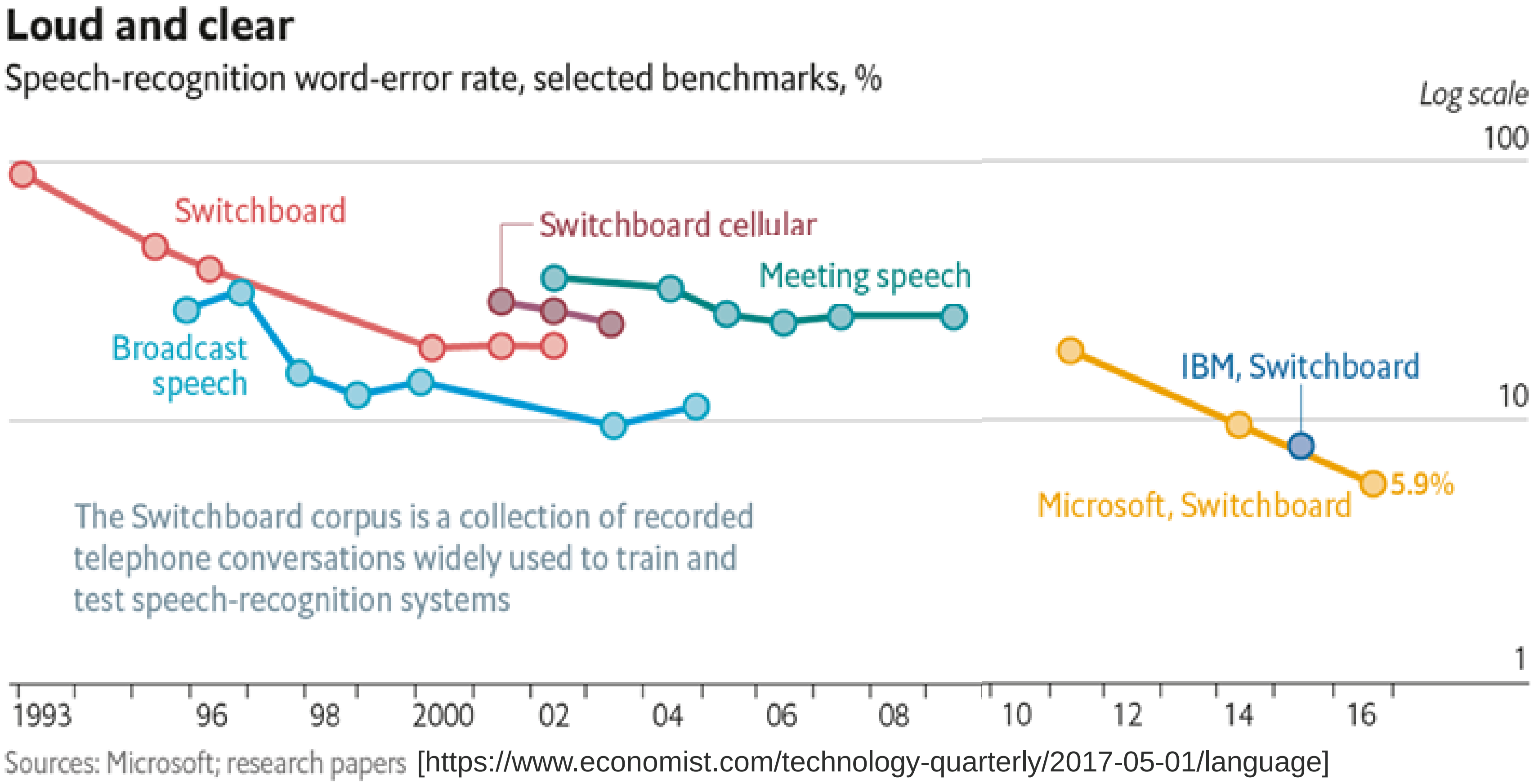
- **Important Factors of Deep Learning**
  → Simplify many complicated hand-engineered models
  → Let the networks find the way that map from speech to text





[LAS, Chan et al. 2015; Figure courtesy of A. Tjandra]

ASR Progress



[https://www.economist.com/technology-quarterly/2017-05-01/language]

Model	N-gram LM		Neural net LM	
	CH	SWB	CH	SWB
Povey et al. [54] LSTM	15.3	8.5	-	-
Saon et al. [51] LSTM	15.1	9.0	-	-
Saon et al. [51] system	13.7	7.6	12.2	6.6
2016 Microsoft system	13.3	7.4	11.0	5.8
Human transcription			11.3	5.9

[Xiaong et al., 2017]

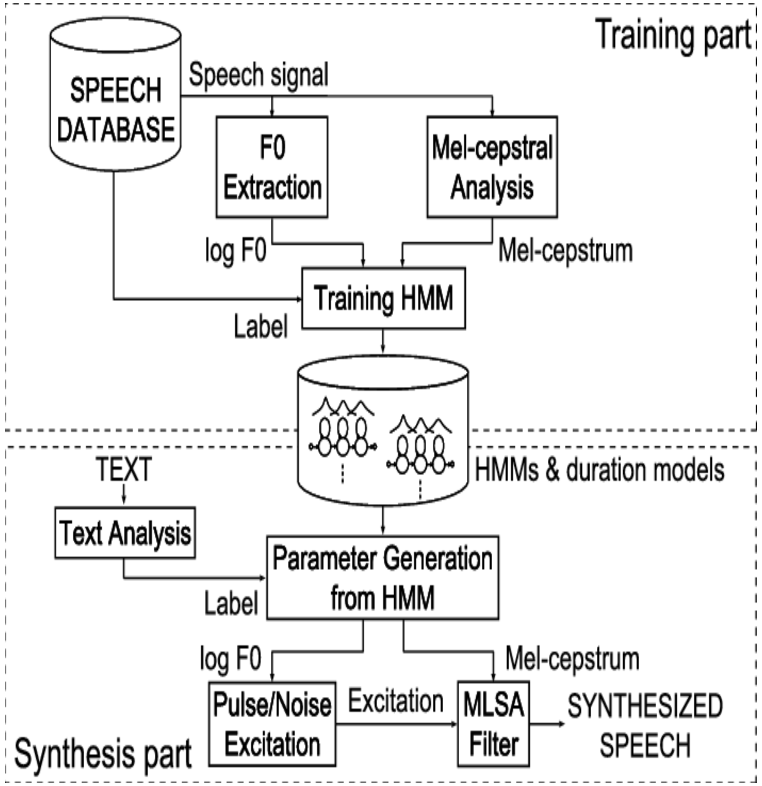
	WER [%]	
	SWB	CH
n-gram	6.7	12.1
n-gram + model-M	6.1	11.2
n-gram + model-M + Word-LSTM	5.6	10.4
n-gram + model-M + Char-LSTM	5.7	10.6
n-gram + model-M + Word-LSTM-MTL	5.6	10.3
n-gram + model-M + Char-LSTM-MTL	5.6	10.4
n-gram + model-M + Word-DCC	5.8	10.8
n-gram + model-M + 4 LSTMs + DCC	**5.5**	**10.3**

[Saon et al., 2017]

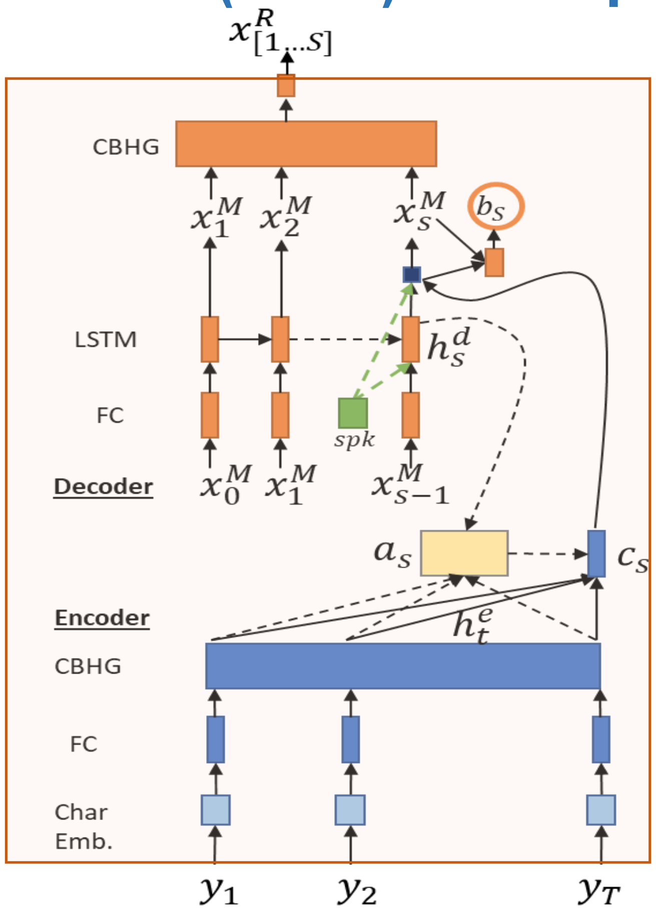
IBM vs. Microsoft: “Human parity” speech recognition record
Makes the same / fewer errors than professional transcriptionists

Text-to-Speech Synthesis (TTS)

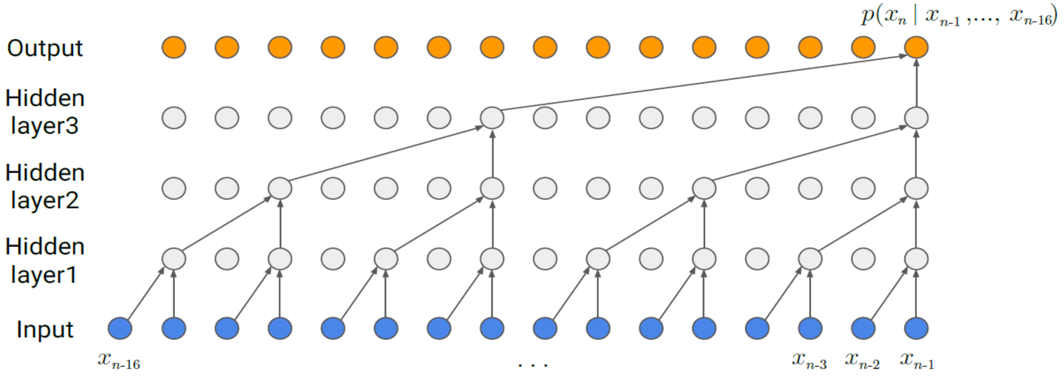
- **From hidden Markov Model (HMM) to Deep Learning**



[Zen et al., 2009]



[Tacotron; Wang et al., 2017;
Figure courtesy of A. Tjandra]



- Conditional WaveNet – TTS

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

[Wavenet; Oord et al., 2016]

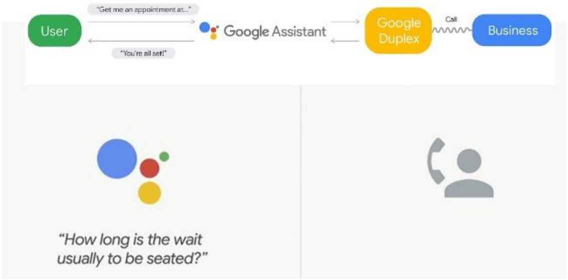
TTS Performance

- **From robot voice to human-like voice**

  [Source: https://www.economist.com/technology-quarterly/2017-05-01/language]

- **Google Duplex**

Google AI



Duplex scheduling a hair salon appointment:

Duplex calling a restaurant:

[Source: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html]

Have we solved all problems?

Professional Speech Interpreter

→ The translation process starts before receiving the end of the sentence
→ Has the ability to do simultaneous process

Challenges for machine speech interpreter:

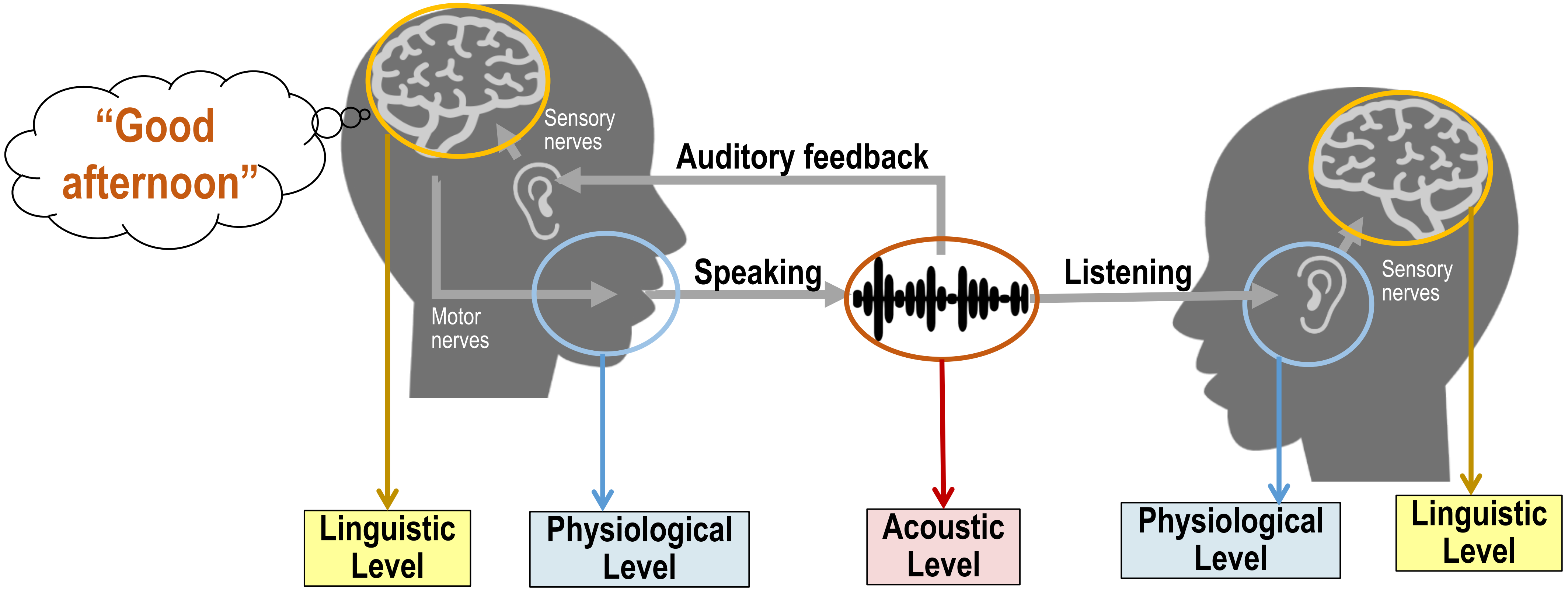
1. Requires the ability to listen while speaking
2. Requires the ability to perform recognition and synthesis speech in real-time

Approach to Problem 1

Listens while Speaking:
Machine Speech Chain by Deep Learning

Human Communication

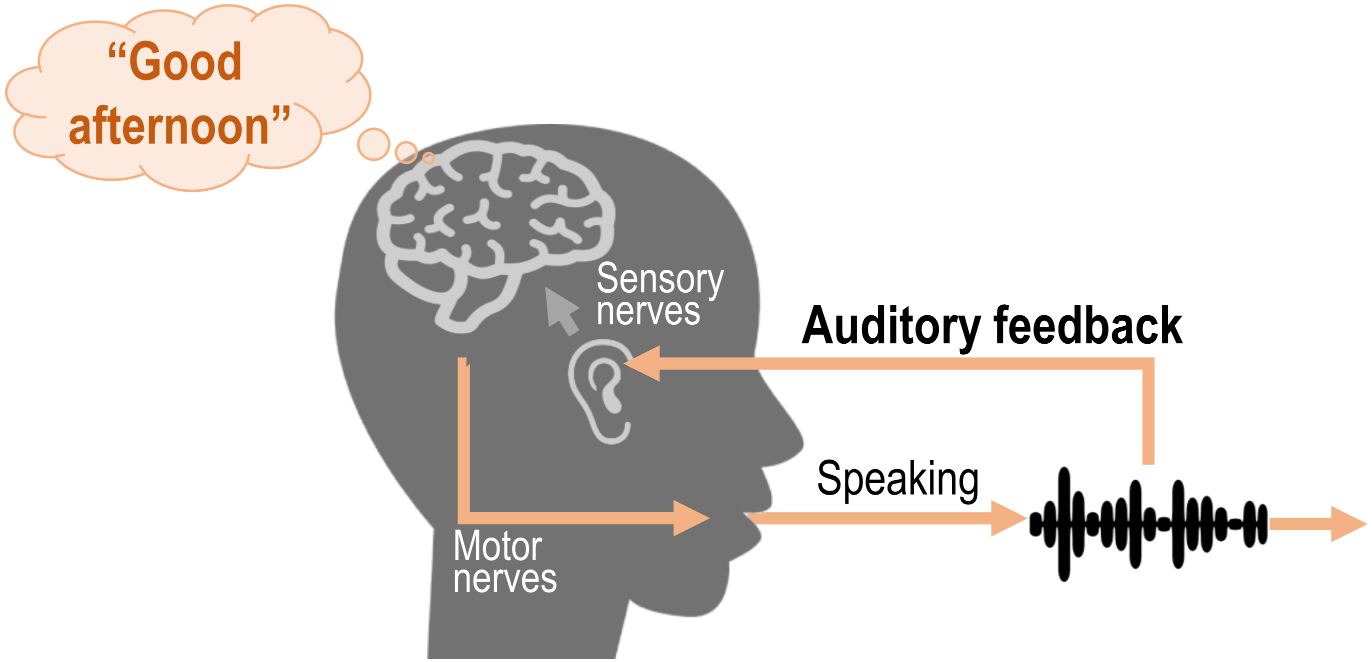
- **Speech Chain** [Denes & Pinson, 1993]



Human Communication

▪ Human: Learning to Listen and Speak

→ Humans learn how to talk by constantly repeating their articulations & listening to sounds produced
→ A closed-loop speech chain mechanism has critical auditory feedback



Children who lose their hearing often have difficulty in producing clear speech

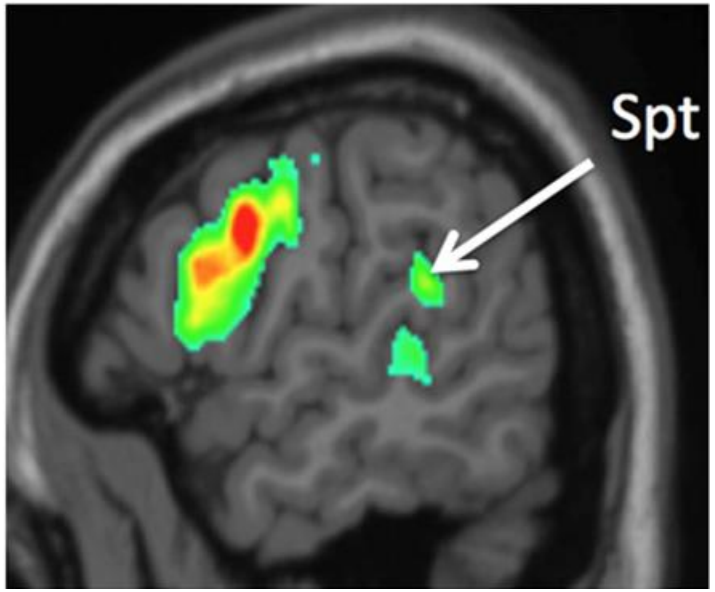
Adults who become deaf after becoming proficient with a language nonetheless suffer speech articulation declines as a result of the lack of auditory feedback

[Waldstein, 1990]

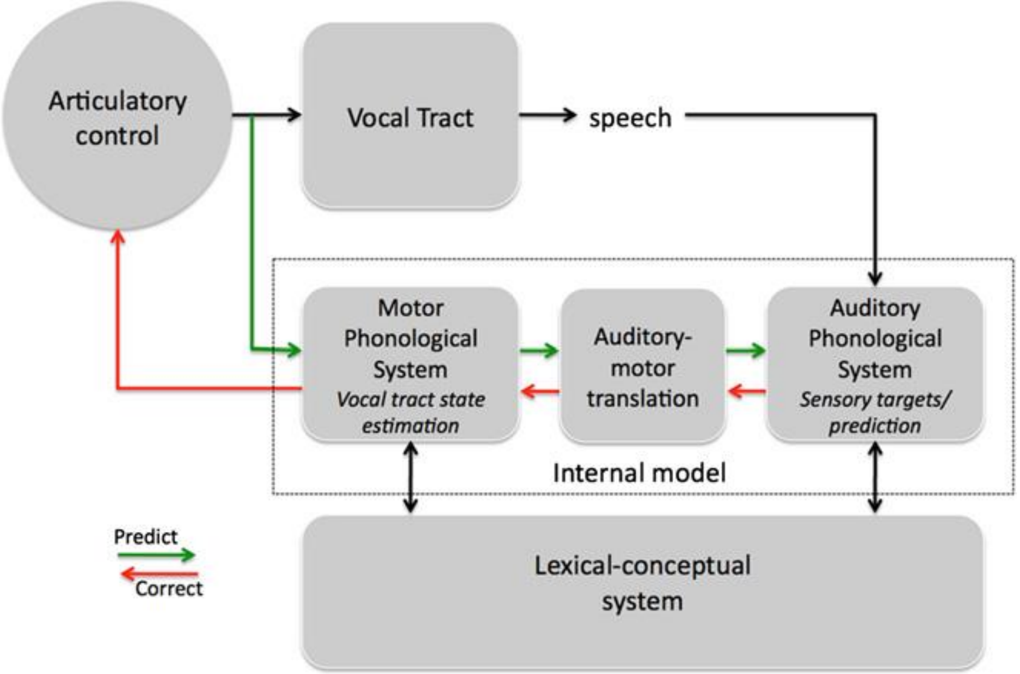
Human Communication

- **Human Brain: Sensorimotor Integration in Speech Processing**

(1) the auditory system is critically involved in the perception of speech
(2) the motor system is critically involved in the production of speech



Spt exhibits sensorimotor response properties, activating both during the passive perception of speech and during covert (subvocal) speech articulation *[Hickok et al, 2003]*



An Integrated State Feedback Control (SFC) Model of Speech Production *[Hickok et al. 2011]*

Human-Machine Communication

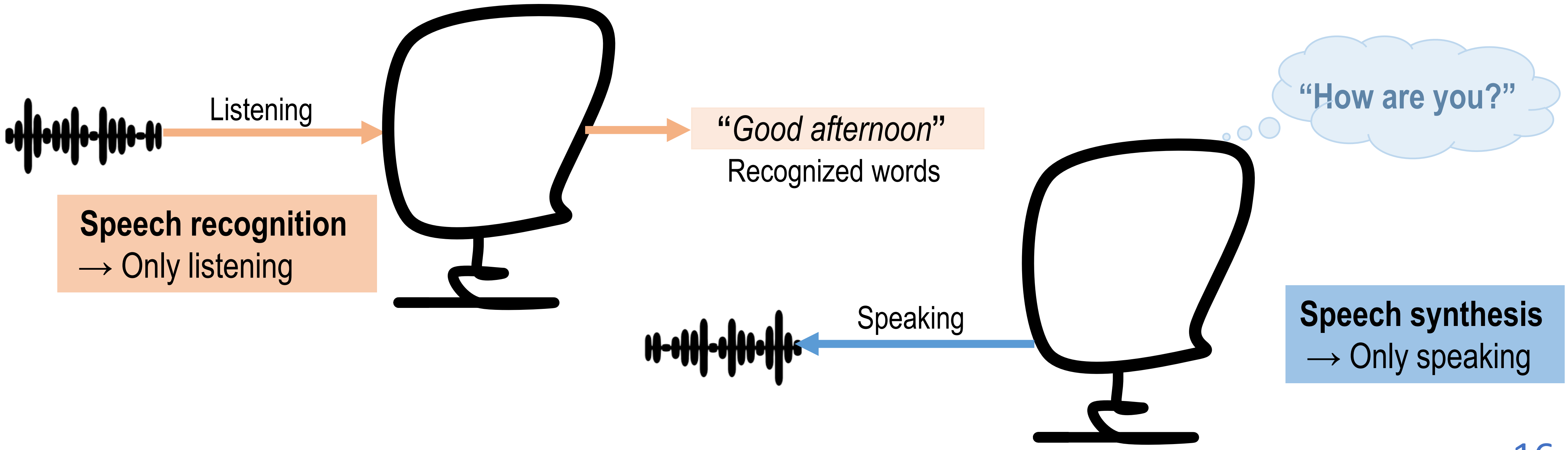
- **Machine: Learning to Listen and Speak**

  → Computers are able to learn how to listen or learn how to speak

  → But, computers cannot hear their own voice

  → The learning to listen and speak is done separately and independently

  → Requires a lot of parallel speech and text to train in a supervised way (more than human need)



Part 1-1
Basic Machine Speech Chain

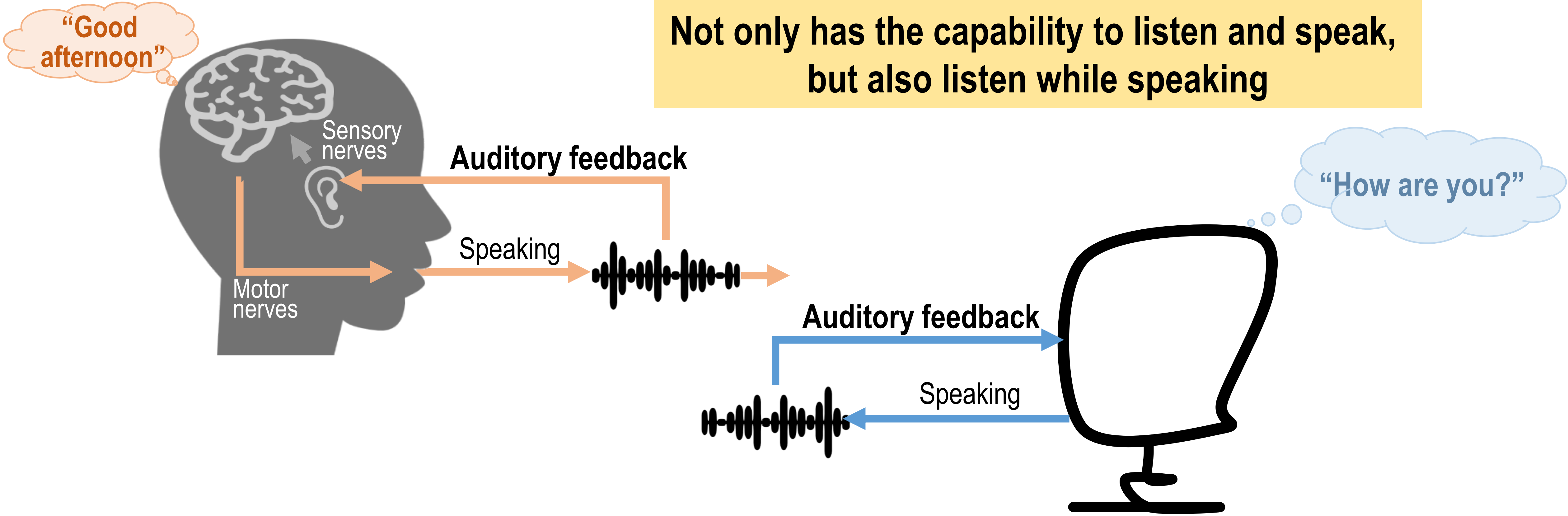
[A. Tjandra, S. Sakti, S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", in Proc. ASRU, 2017]

Machine Speech Chain

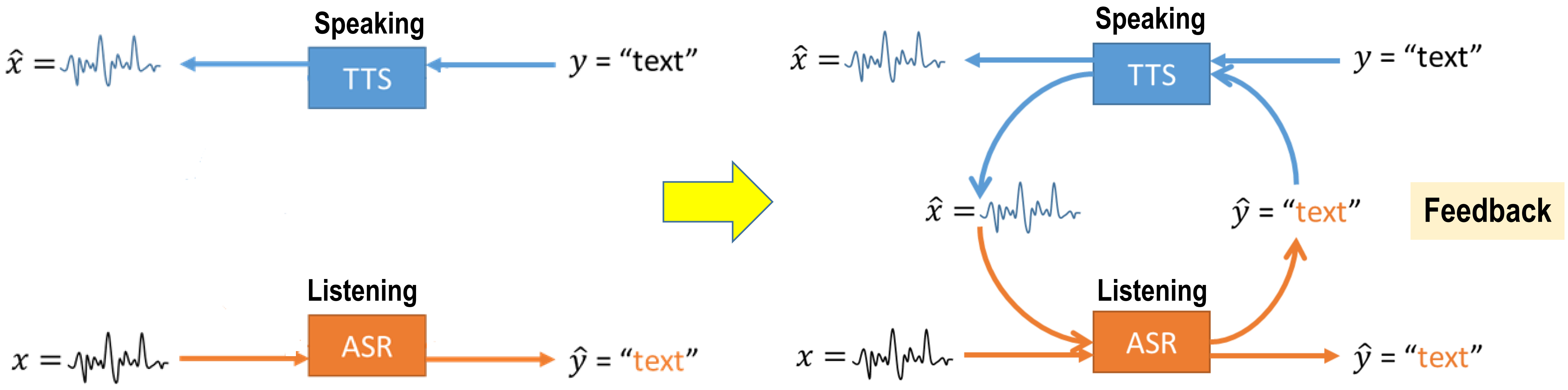
- ## Proposed Method

→ 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



Machine Speech Chain



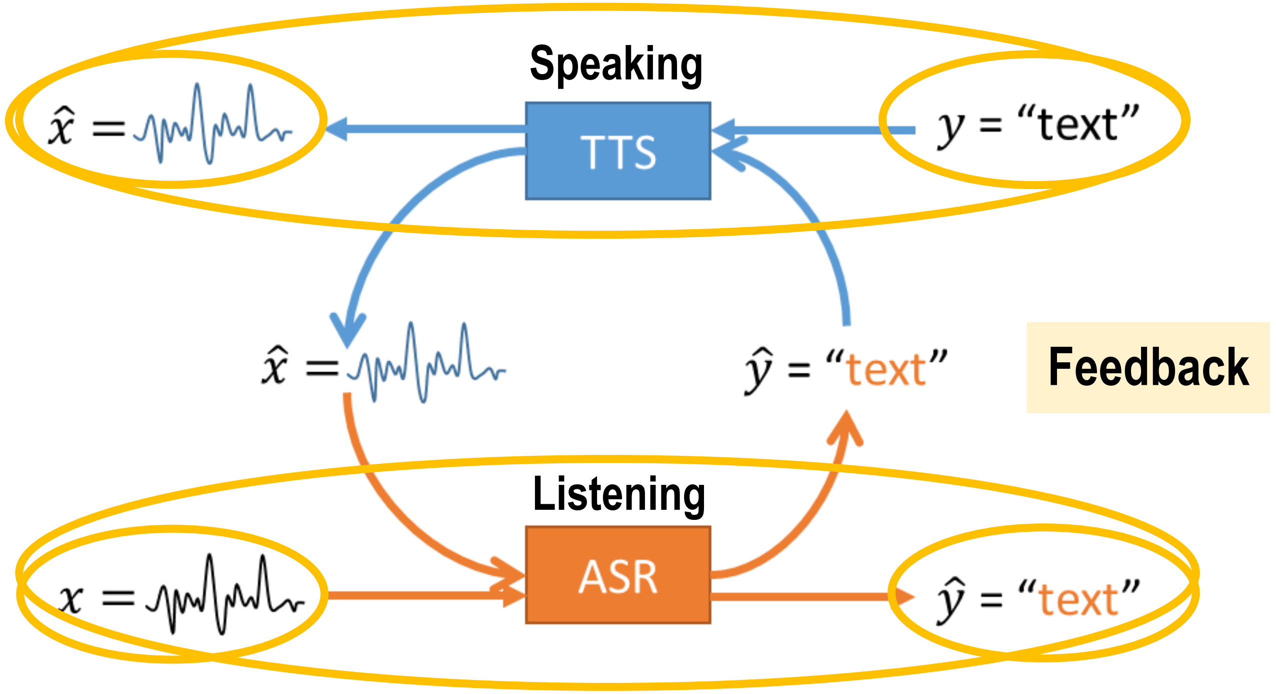
Advantages of closed-loop architecture:

→ **In training stage:**

- Allow to train with labeled and unlabeled data (semi-supervised learning)
- Allow ASR and TTS to teach each other using unlabeled data and generate useful feedback

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

Overall Architecture



Definition:

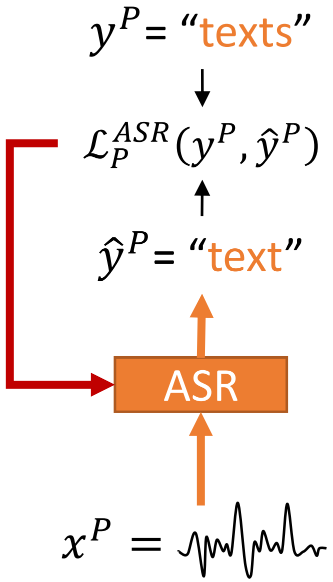
- x = original speech, y = original text
- $\hat{x}$ = predicted speech, $\hat{y}$ = predicted text
- $ASR(x): x \rightarrow \hat{y}$ (seq2seq model transform speech to text)
- $TTS(y): y \rightarrow \hat{x}$ (seq2seq model transform text to speech)

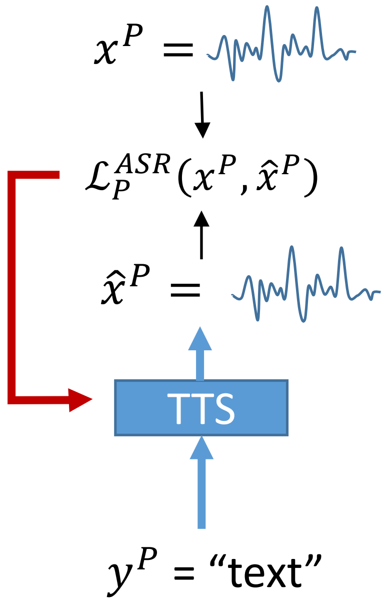
Learning in Machine Speech Chain

Case #1: Supervised Learning with Speech-Text Data

Given a pair speech-text (x^P, y^P)

- Train ASR and TTS in supervised learning
- Directly optimize:
  $\rightarrow ASR$ by minimizing $\mathcal{L}_P^{ASR}(y^P, \hat{y}^P)$
  $\rightarrow TTS$ by minimizing $\mathcal{L}_P^{TTS}(x^P, \hat{x}^P)$
- Update both ASR and TTS independently





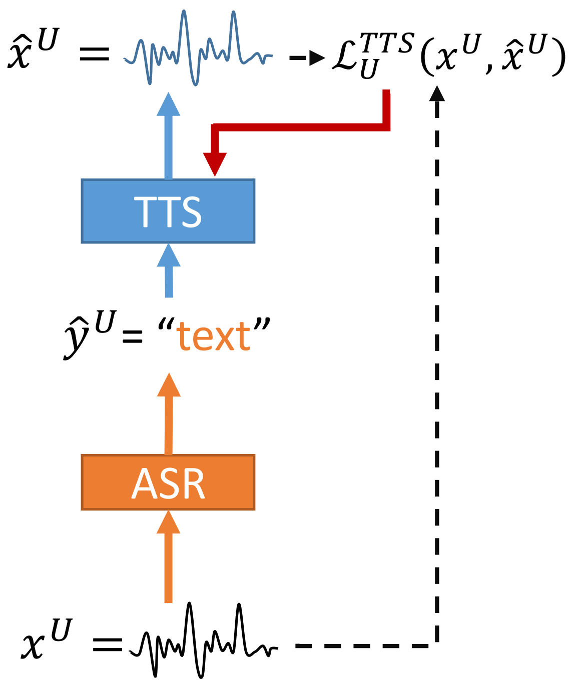
Learning in Machine Speech Chain

Case #2: Unsupervised Learning with Speech Only

Given only speech features x^U

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

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

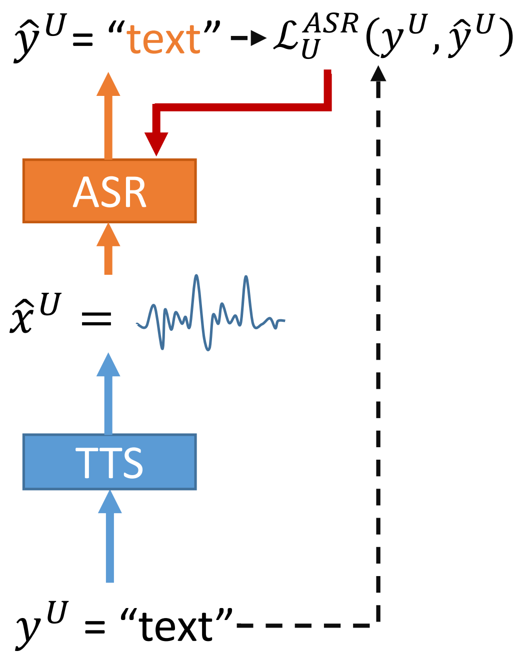


Learning in Machine Speech Chain

Case #3: Unsupervised Learning with Text Only

Given only text features y^U

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



Possible to improve ASR with text only by the support of TTS

Learning in Machine Speech Chain

- **Training Objective**

$$\mathcal{L} = \alpha * \left(\mathcal{L}_P^{ASR} + \mathcal{L}_P^{TTS}\right) + \beta * \left(\mathcal{L}_U^{ASR} + \mathcal{L}_U^{TTS}\right)$$

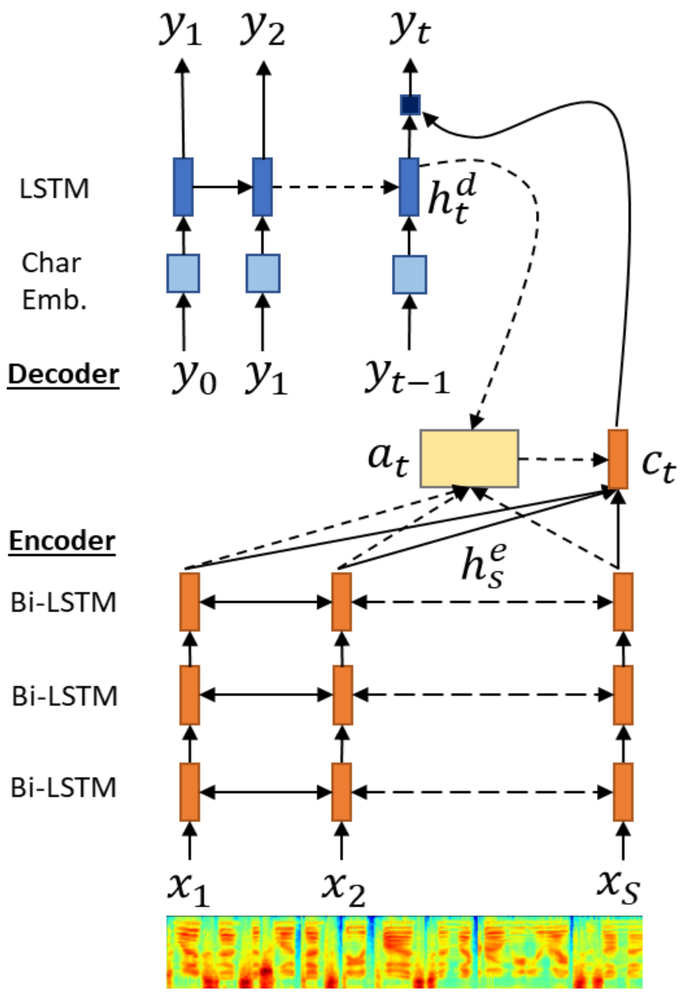
- **Basic Idea**

→ Possible to train the new matters without forgetting the old one

→ $\alpha > 0$: keep using some portions of the loss and the gradient provided by the paired training set

→ $\alpha = 0$: completely learn new matters with only speech or only text

Sequence-to-Sequence ASR



Similar to [LAS, Chan et al. 2015]

Input & output

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

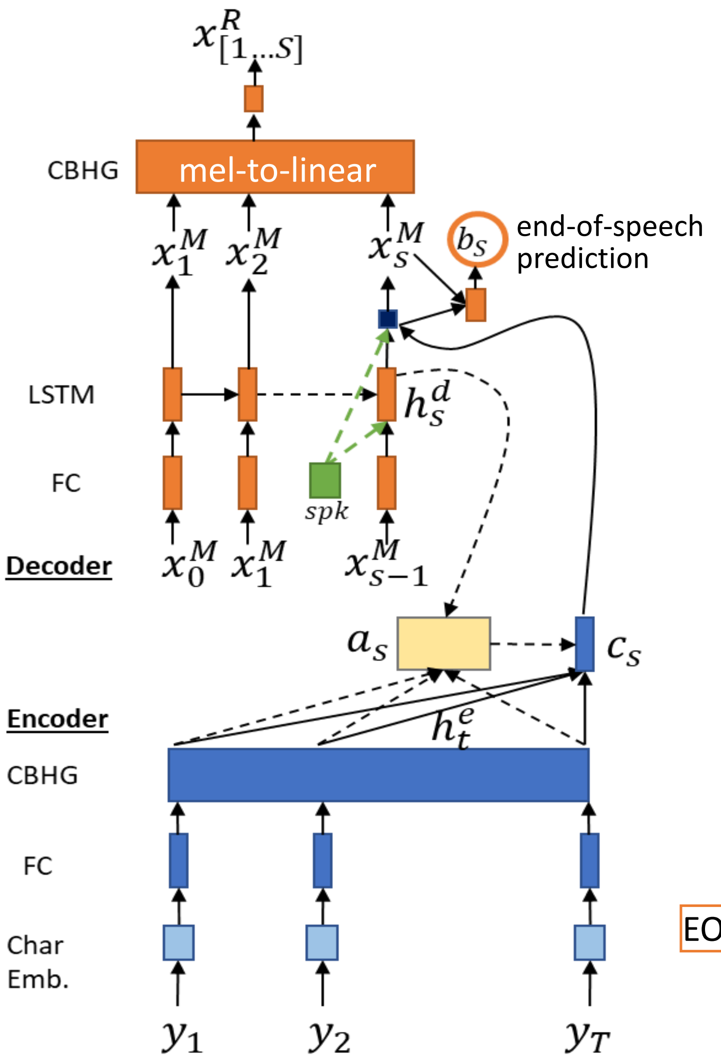
Model states

- $h^e_{[1..S]}$ = encoder states
- h^d_t = decoder state at time t
- a_t = attention probability at time t
  - $a_t(s) = Align\left(h^e_s, h^d_t\right)$
  - $a_t(s) = \dfrac{\exp\left(Score\left(h^e_s, h^d_t\right)\right)}{\sum_{s=1}^{S} \exp\left(Score\left(h^e_s, h^d_t\right)\right)}$
- $c_t = \sum_{s=1}^{S} a_t(s) * h^e_s$ (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



Similar to [Tacotron: Wang et al., 2017]

Input & output

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

Reconst. MSE

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

EOS cross entropy

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

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

Experiments on Single-Speaker Dataset

▪ Features

Speech:

- 80-dim 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

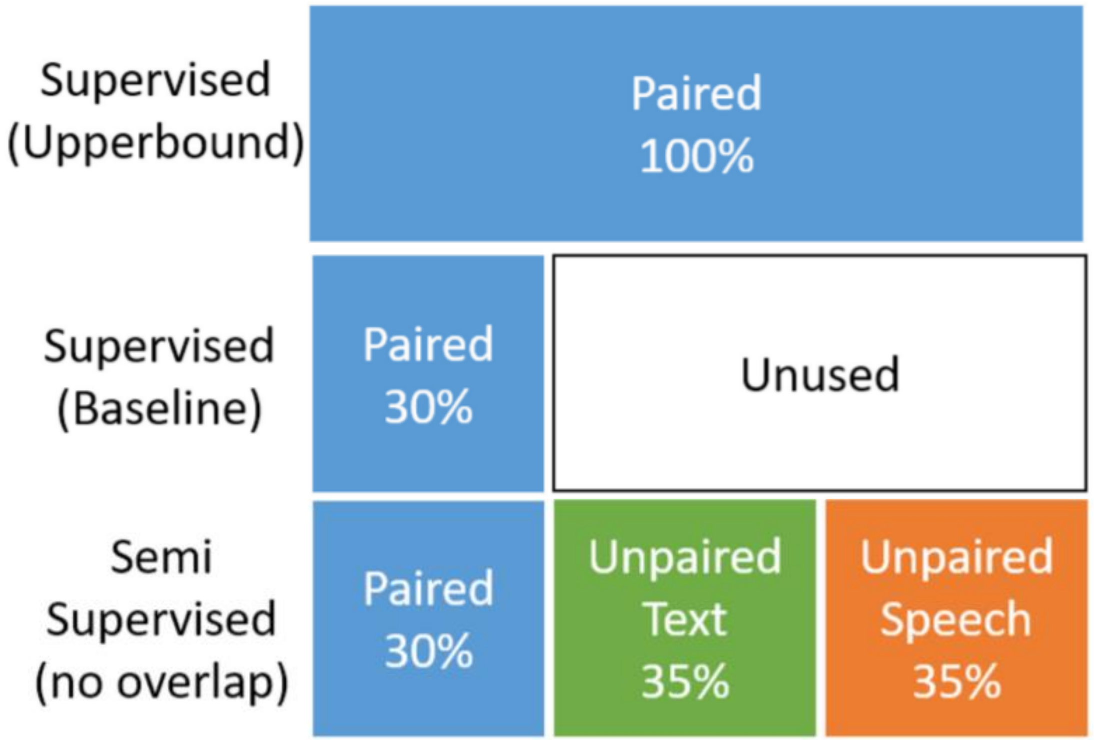
Text:

Character-based prediction

- a-z (26 alphabet)
- 6 punctuation mark (,:'?.-)
- 3 special tags <s> </s> <spc> (start, end, space)

Experiments on Single-Speaker Dataset

- **Data set**
  - Natural speech single-speaker dataset: LJSpeech [Ito et al. , 2017]
  - Training set: 12,314 utts; dev set: 393 utts; test set:393 utts



Experiments on Single-Speaker Dataset

- **ASR results**

Model	Paired	Unpaired Text	Unpaired Speech	CER (%)
Supervised (Baseline)				
Enc-Dec Att	10%	-	-	31.7
Enc-Dec Att	20%	-	-	9.9
Enc-Dec Att	30%	-	-	6.8
Enc-Dec Att	40%	-	-	4.9
Enc-Dec Att	50%	-	-	4.1
Semi-supervised (Speech Chain)				
Enc-Dec Att	10%	45%	45%	12.3
Enc-Dec Att	20%	40%	40%	5.6
Enc-Dec Att	30%	35%	35%	4.7
Enc-Dec Att	40%	30%	30%	3.8
Enc-Dec Att	50%	25%	25%	3.5
Supervised (Upperbound)				
Enc-Dec Att	100%	-	-	3.1

Experiments on Single-Speaker Dataset

- **TTS results**

Model	Paired	Unpaired		L2-norm^2
		Text	Speech	
Supervised (Baseline)				
Enc-Dec Att	10%	-	-	1.05
Enc-Dec Att	20%	-	-	0.91
Enc-Dec Att	30%	-	-	0.71
Enc-Dec Att	40%	-	-	0.69
Enc-Dec Att	50%	-	-	0.66
Semi-supervised (Speech Chain)				
Enc-Dec Att	10%	45%	45%	0.87
Enc-Dec Att	20%	40%	40%	0.73
Enc-Dec Att	30%	35%	35%	0.66
Enc-Dec Att	40%	30%	30%	0.65
Enc-Dec Att	50%	25%	25%	0.64
Supervised (Upperbound)				
Enc-Dec Att	100%	-	-	0.606

Discussion

- **Summary:**
  - **Inspired by the human speech chain, we proposed a machine speech chain to achieve semi-supervised learning**
  - **Enables ASR & TTS to assist each other when they receive unpaired data**
  - **Allows ASR & TTS to infer the missing pair and optimize the models with reconstruction losses**

- **Current Limitations:**
  - **Set of speakers is fixed** → Unable to handle unseen speakers
  - **TTS system only mimics the voices of diff. speakers via speaker's identity by one-hot encoding**
  - **ASR only adapted to a specific set of speaker**

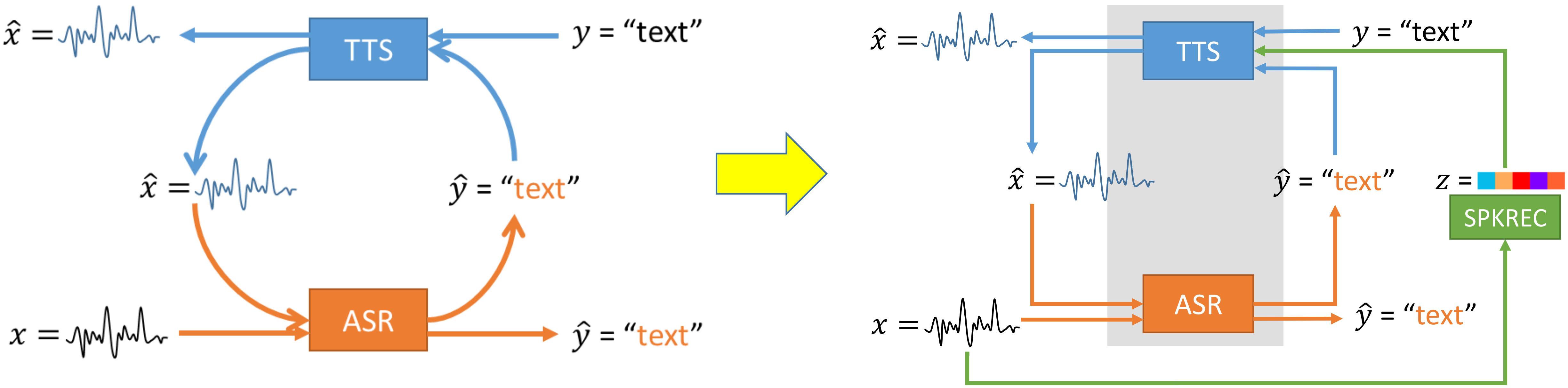
  → Because the TTS unable to produce more voice characteristics from unseen speakers

Part 1-2
Multi-Speaker Machine Speech Chain

[A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", in Proc. INTERSPEECH, 2018]

Multi-Speaker Machine Speech Chain

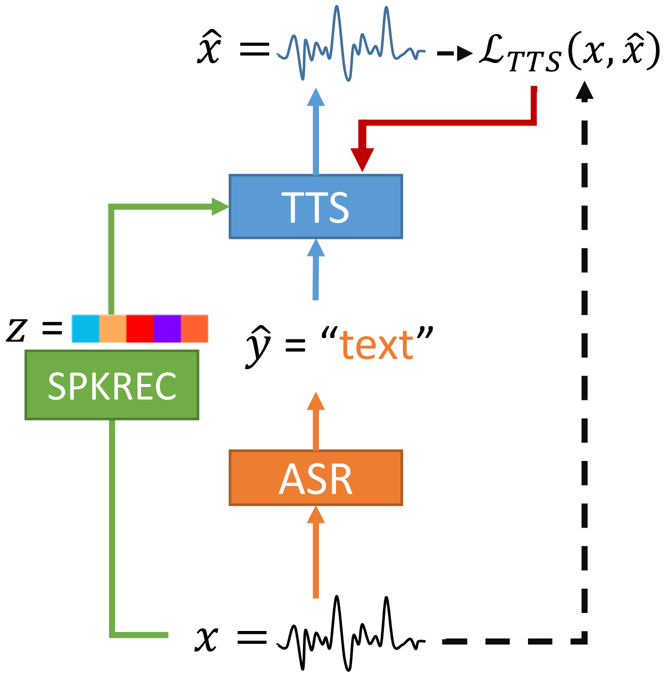
- **Proposed Approach:** **Handle voice characteristics from unknown speakers**
  - → Integrate a speaker recognition system into the speech chain loop
  - → Extend the capability of TTS to handle the unseen speaker using one-shot speaker adaptation
  - → Coupling with ASR, we developed a speech chain framework that is able to adapt new data speech from an unknown speaker



Utilizing [Deep speaker; Li et al., 2017]

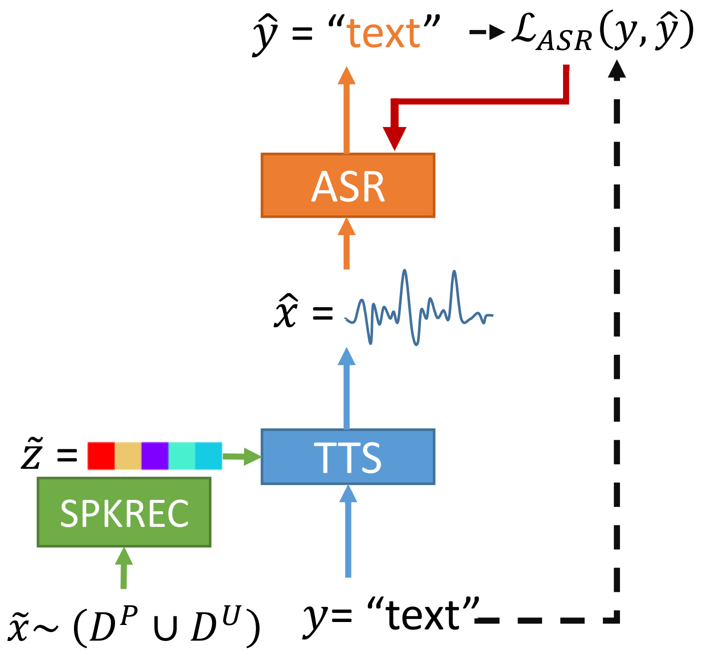
Learning in Multi-Speaker Speech Chain

- **Train with Speech only:**
  ASR→TTS



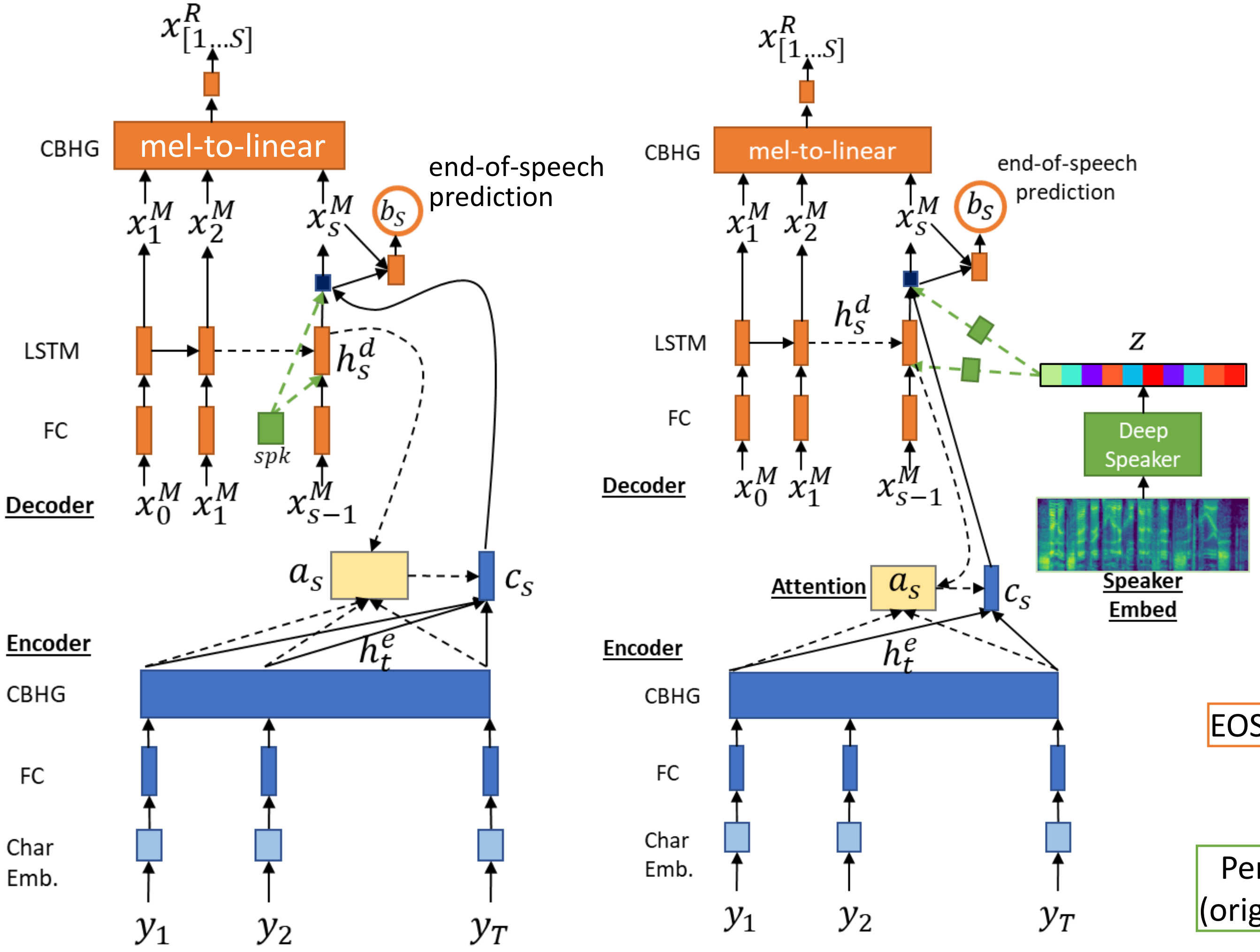
→ ASR predicts most possible transcription $\hat{y}$
→ SPKREC provides a speaker embedding z
→ Based on $[\hat{y}, z]$, TTS tries to reconstruct speech $\hat{x}$

- **Train with Text only:**
  TTS→ASR



→ Sample a speaker vector $\tilde{z}$ from available speech
→ TTS generates speech features $\hat{x}$ based on $[y, \tilde{z}]$
→ Given $\hat{x}$, ASR tries to reconstruct text $\hat{y}$

Sequence-to-Sequence TTS



Input & output

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

Reconst. 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 (original vs gen sp) $\mathcal{L}_{TTS3}(z, \hat{z}) = 1 - \frac{\langle z, \hat{z} \rangle}{\|z\|_2 + \|\hat{z}\|_2}$

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

Experiments on Multi-Speakers

- **Data set: Wall Street Journal (WSJ) [Paul et al. , 1992]**
  - **Training set: Supervised (paired text & speech)**
    - WSJ SI-84 dataset (baseline)
      (7138 utterances, ~16 h, 84 speakers)
    - WSJ SI-284 dataset (upperbound)
      (37318 utterances, ~81 h, 284 speakers)
  - **Training set: Unsupervised (unpaired text & speech)**
    - WSJ SI-200 dataset
      (30180 utterances, ~66 hours, 200 speakers)
    - Notes: SI-200 doesn't overlap with SI-84
  - **Development set:** dev93
  - **Evaluation set:** eval92

ASR Results

Model	CER (%)
Supervised training: WSJ train_si84 (16hrs speech, paired) -> Baseline	
Att Enc-Dec	17.35
Supervised training: WSJ train_si284 (81 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
Proposed speech chain	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 (S 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			

Discussion

- **Summary:**
  - **Improved machine speech chain to handle voice characteristics from unknown speakers**
    → TTS can generate speech with similar voice characteristic only with one-shot speaker examples
    → ASR also get new data from the combination between a text sentence and an arbitrary voice characteristic
  - **By combining both models, we could train with auxiliary feedback loss**

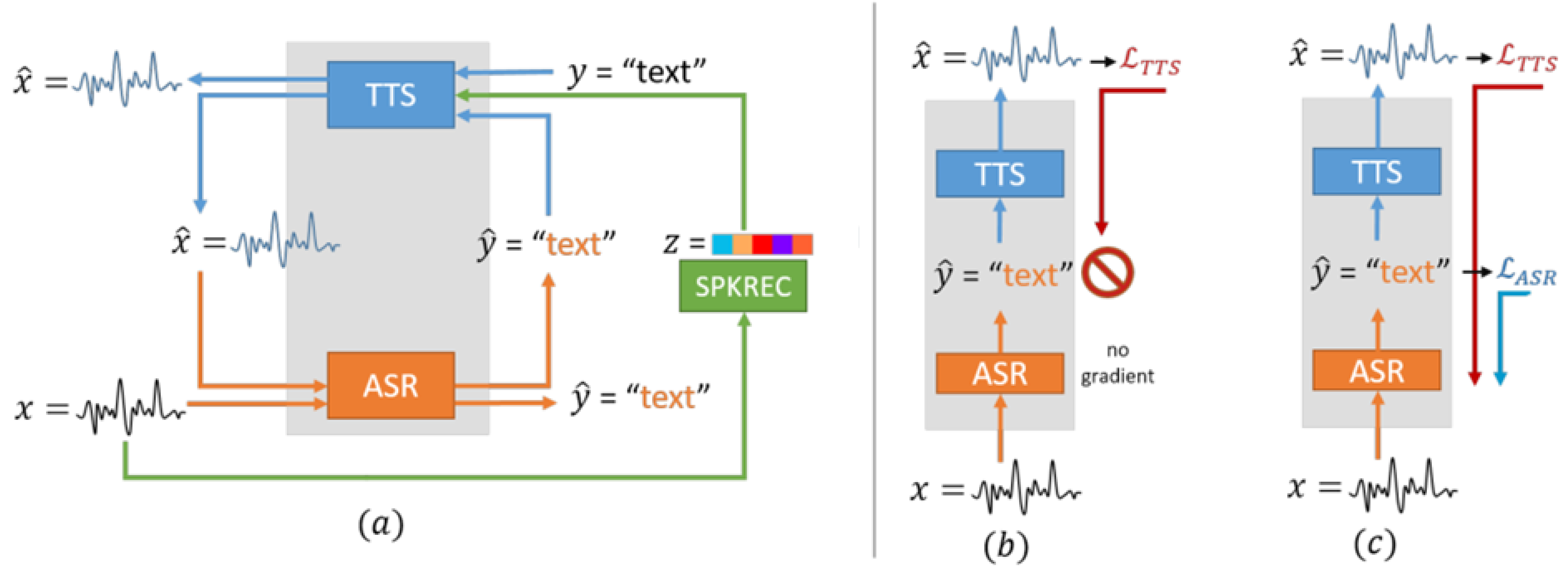
- **Current Limitations:**
  - **If we only have the text:** Perform TTS →ASR, and update only ASR with feedback loss
  - **If we only have speech:** Perform ASR →TTS, and update only TTS with feedback loss
  - **Backpropagation the error from the reconstruction loss through ASR is challenging due to the output of the ASR discrete tokens**

Part 1-3
Machine Speech Chain
with End-to-end Feedback Loss

[A. Tjandra, S. Sakti, S. Nakamura, "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 $x = TTS(\hat{y}, z)$

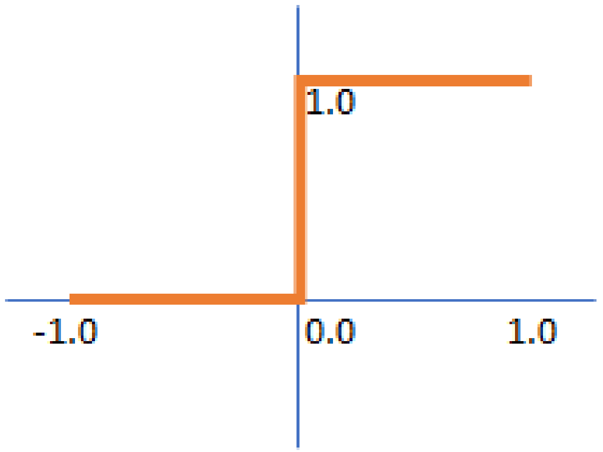
a) Speech chain loop with speaker embedding module z

b) Original: feedback $\mathcal{L}_{TTS}$ can't be backpropagated through variable $\hat{y}$

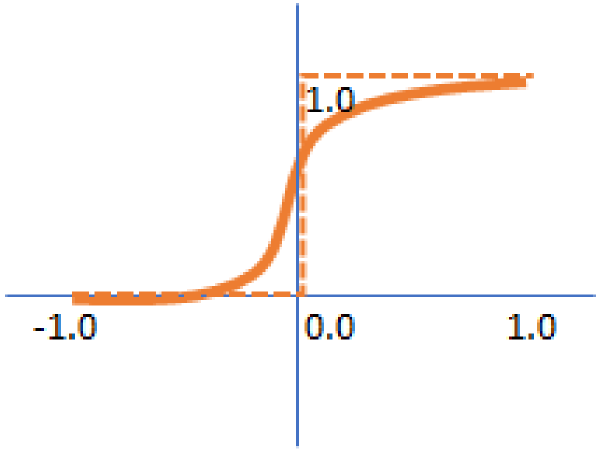
c) **Proposal**: Estimate gradient through variable $\hat{y}$ with straight-through estimator

Basic Idea

- Based on Two ICLR 2017 Papers [Jang et al. 2017, Maddison et al. 2017]
- Example: Discrete Node with Step Function



Almost everywhere, a small change in input results in no change in output
→ Gradient is zero



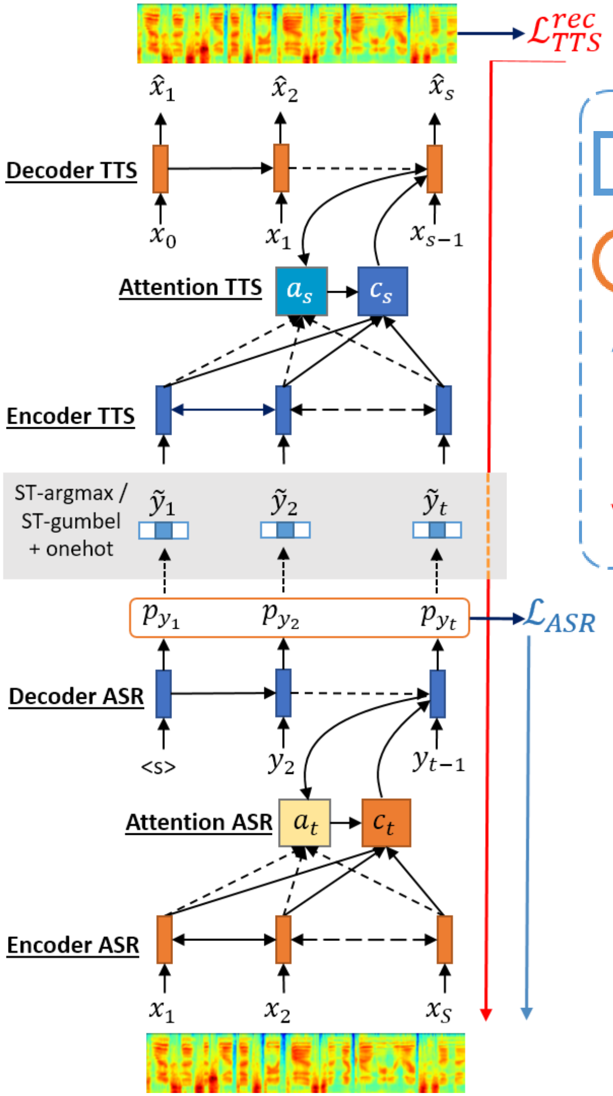
A common trick is to use a continuous approximation
→ But, they fail to produce discrete outputs

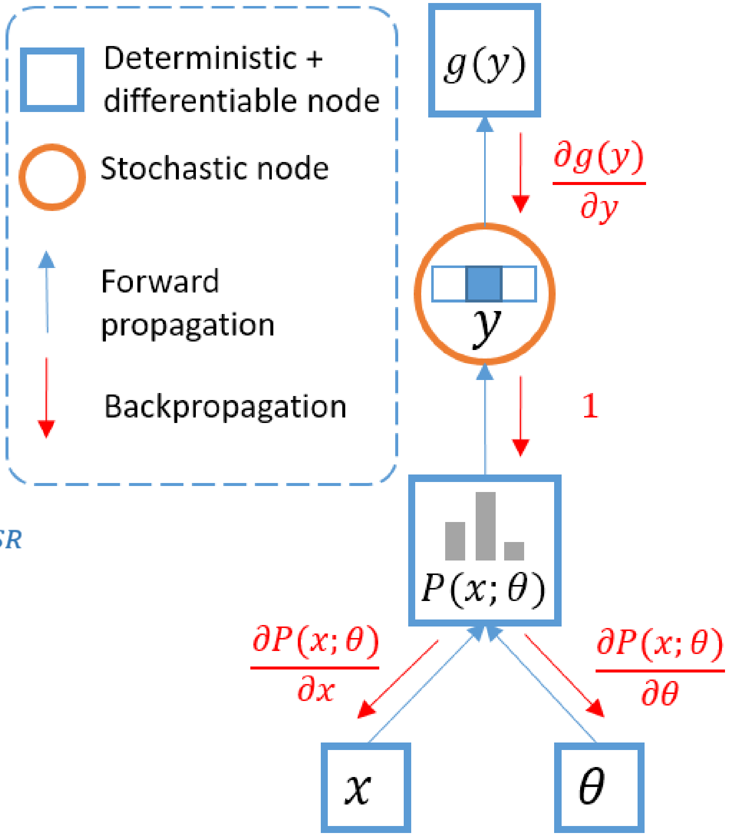
Gumbel-Softmax Distribution

Provides a simple method for draw samples from a categorical distribution with class Probabilities
→ Use softmax function as approximate of argmax function

[Source: https://uoguelph-mlrg.github.io/spaceNet_overview2/]

Straight-Through Estimator





Replace gradient from computation node $\frac{\partial \tilde{y}_t}{\partial p_{y_t}}$ with identity

a) **ST-argmax**

Deterministic choosing token by highest probability

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

$$\tilde{y}_t = argmax_c\, p_{y_t}[c]$$

b) **ST-Gumbel softmax**

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

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

$$\tilde{y}_t \sim Categorical\left(p_{y_t}[1], .., p_{y_t}[C]\right)$$

τ = temperature
h_t^d=logit ASR

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

$$\frac{\partial \mathcal{L}_{TTS}^{rec}}{\partial \theta_{ASR}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_{TTS}^{rec}}{\partial \tilde{y}_t} \cdot \frac{\partial \tilde{y}_t}{\partial p_{y_t}} \cdot \frac{\partial p_{y_t}}{\partial \theta_{ASR}}$$

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

Experiments on Multi-Speakers WSJ Task

Data set

- **Training set: Supervised (paired text & speech)**
  WSJ SI-284 dataset [Paul et al. , 1992]
  (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

Discussion

- **Summary:**
  - **Improved machine speech chain mechanism**
    $\rightarrow$ Allow backpropagation through discrete output with a straight-through estimator
  - **Future work:**
    It is necessary to validate the effectiveness of the approach in various languages

Part 1-5
From Speech Chain to Multimodal Chain

[Johanes Effendi, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, "Augmenting Images for ASR and TTS through Single-loop and Dual-loop Multimodal Chain Framework," Proc. of INTERSPEECH, Oct 2020]

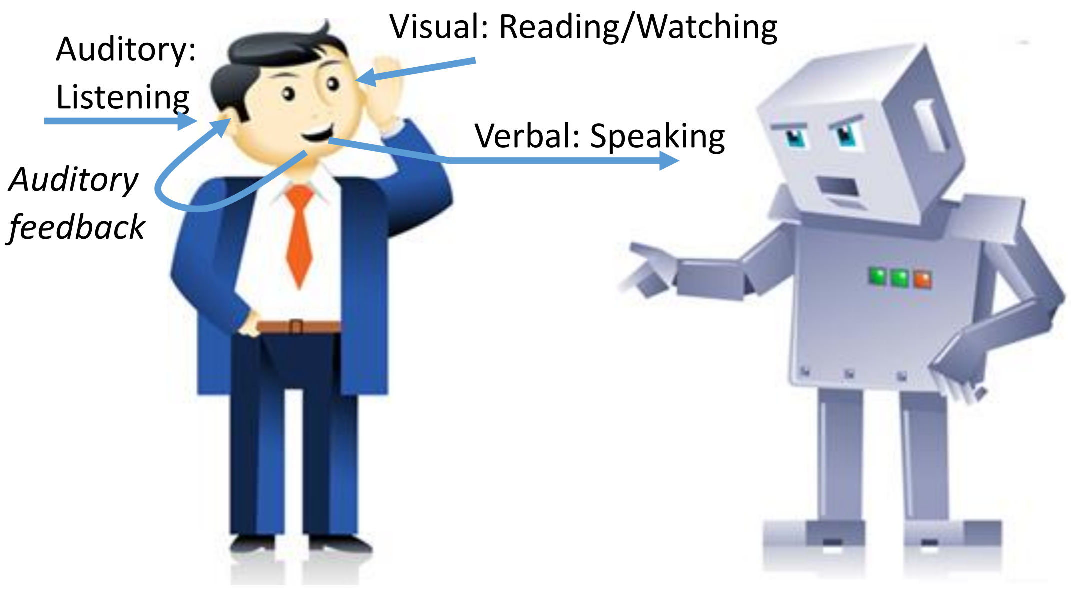
Human Communication

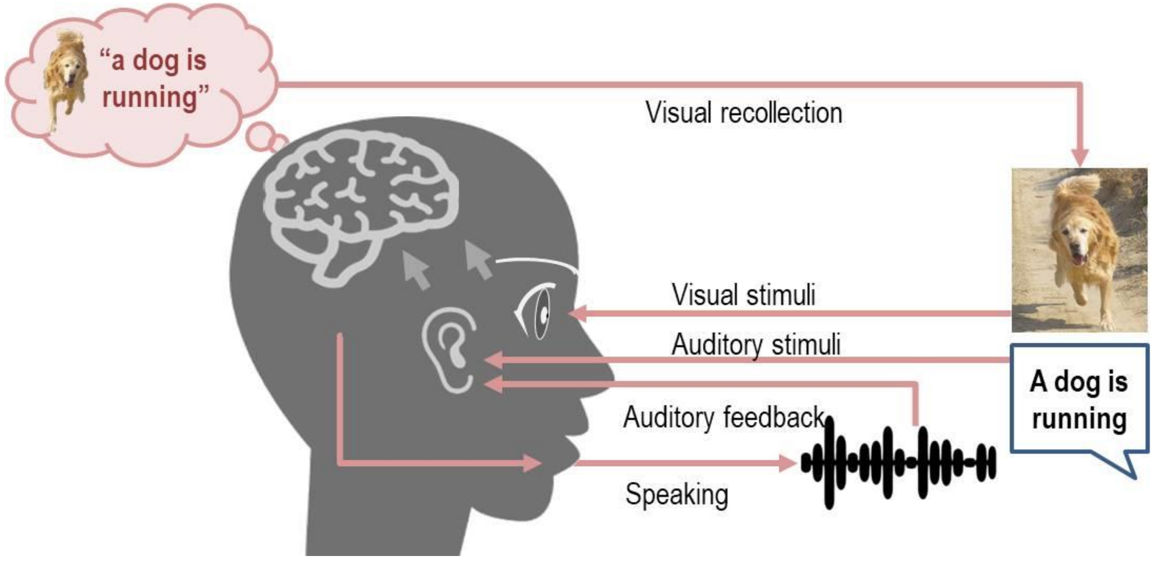
- **Human Communication:**
  - The most common way for humans to communicate is by speech
  - But, a language system cannot know what it is communicating without a connection to the real world by image perception

Human Communication

▪ Human Communication:

- Human communication is multisensory and involves several communication channels (auditory and visual channels)
- Human perceives these multiple sources of information together to build a general concept

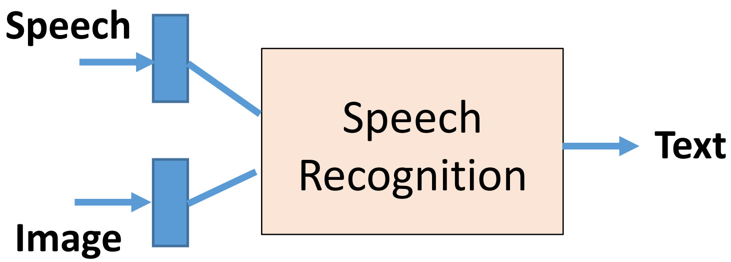


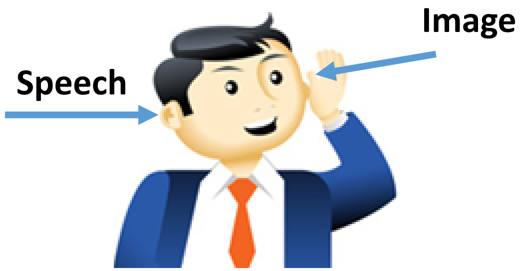


Each modalities shares complementary behaviour

Multimodal System

- ## Multimodal System:
  - The idea of incorporating visual information for speech processing is not new
    → Audio-visual ASR [Petridis et al., 2017, Chung et al., 2017, Afouras et al., 2018]
  - But most approaches are usually made by simply concatenating the information
    → Inefficient to effectively fuse information
  - These methods require all information from different modalities available altogether
    → Parallel data is often unavailable





Machine Speech Chain

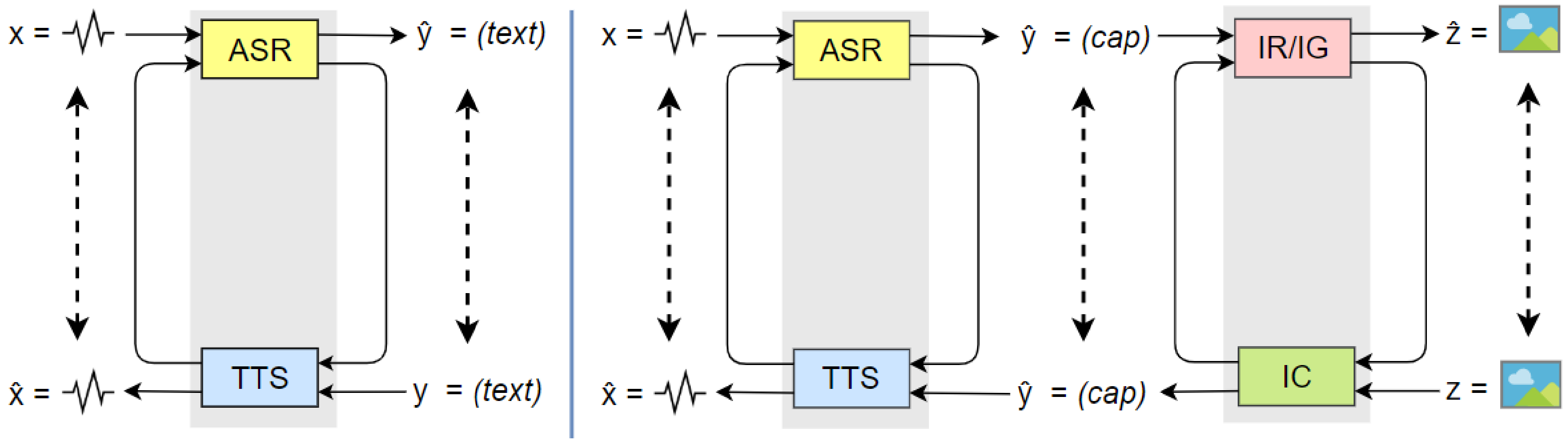
- **Machine Speech Chain:**
  - The approach let us free from the need for a large size of parallel speech-text data
  - It provides possibilities to improve ASR & TTS performance in semi-supervised learning by allowing ASR and TTS to teach each other, given only text or only speech data.

Unfortunately,

- **Although it reduces the requirement of having a full amount of paired data, it is still required to have a large size of unpaired data**
- **This study is limited only to speech and textual modalities**
- **Natural communication is multimodal that involves not only the auditory system but also visual sensory**

Multimodal Machine Speech Chain (MMC)

- **Multimodal Machine Speech Chain (MMC):**
  - Expanding the speech chain (a) into a multimodal chain (b)
  - Design a closely-knit chain architecture that connects ASR, TTS, IC, and IR/IG
  - Can be trained in semi-supervised fashion by assisting each other given incomplete data
  - Leveraging cross-modal data augmentation within the chain



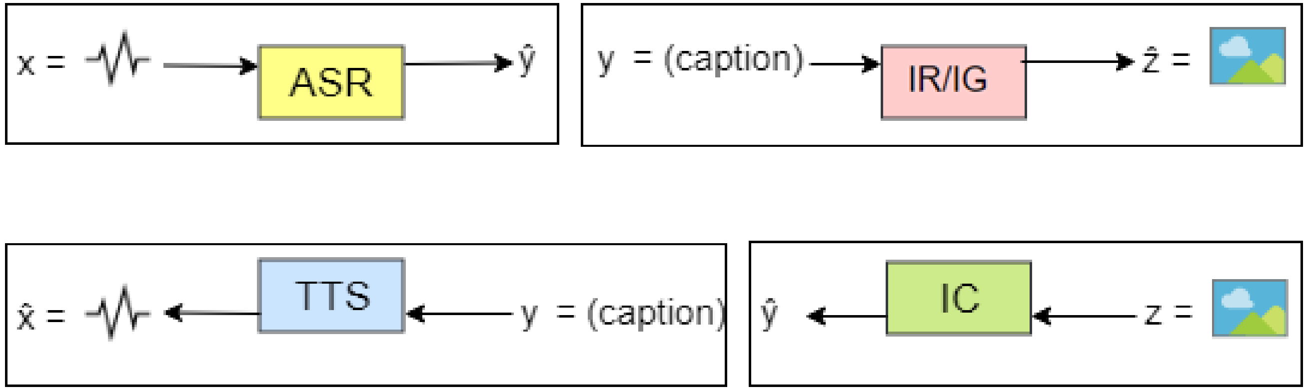
- **Research Question:** **Can we still improve ASR even no speech/text data available?**

Learning in MMC Framework

Case #1: Supervised Learning with Speech-Text-Image Data

- **Paired speech-text-image data exist:**

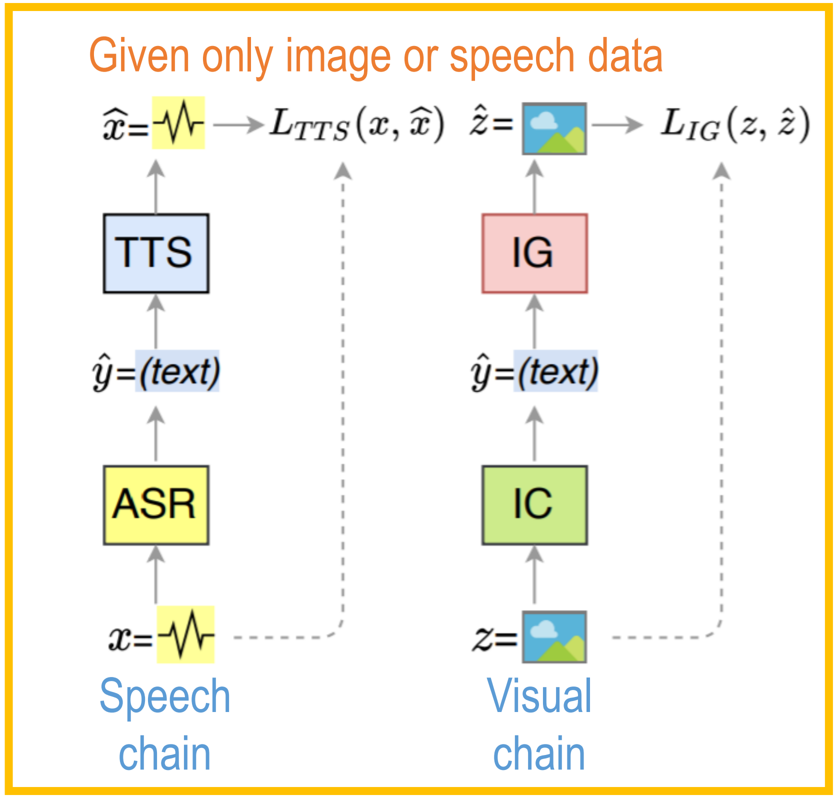
  → Separately train ASR, TTS, IC and IR/IG (supervised learning)

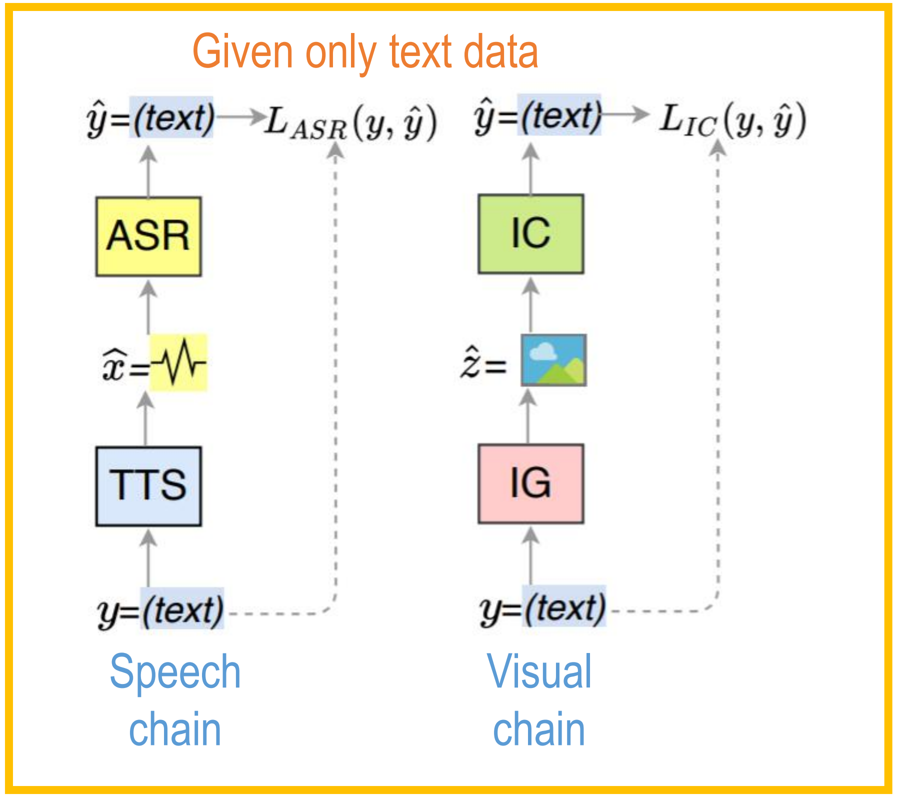


Learning in MMC Framework

Case #2: Unsupervised Learning with Unpaired Data

- **Unpaired speech-text-image data exist:**
  → Perform speech chain (ASR-TTS) with unpaired speech-text data
  → Perform image chain (IC-IG) with unpaired image-text data





Learning in MMC Framework

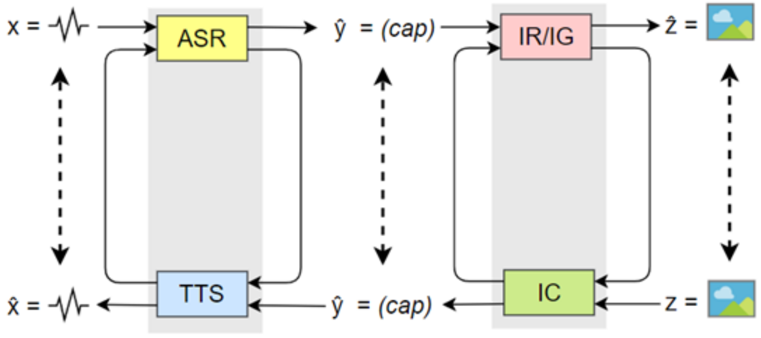
Case #3: Unsupervised Learning with Speech or Image Only Data

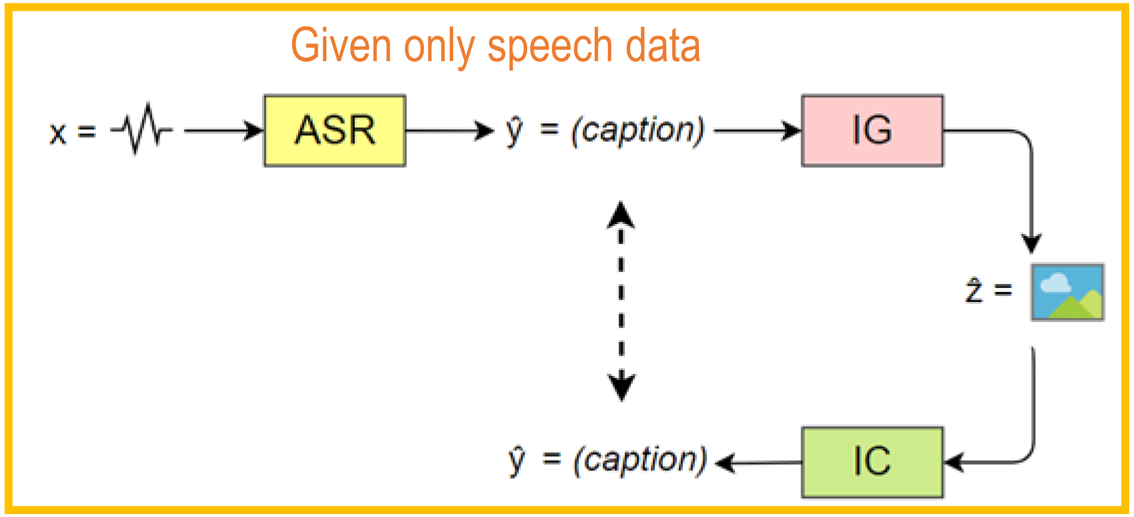
- **Single data exist (speech/text/image only):**

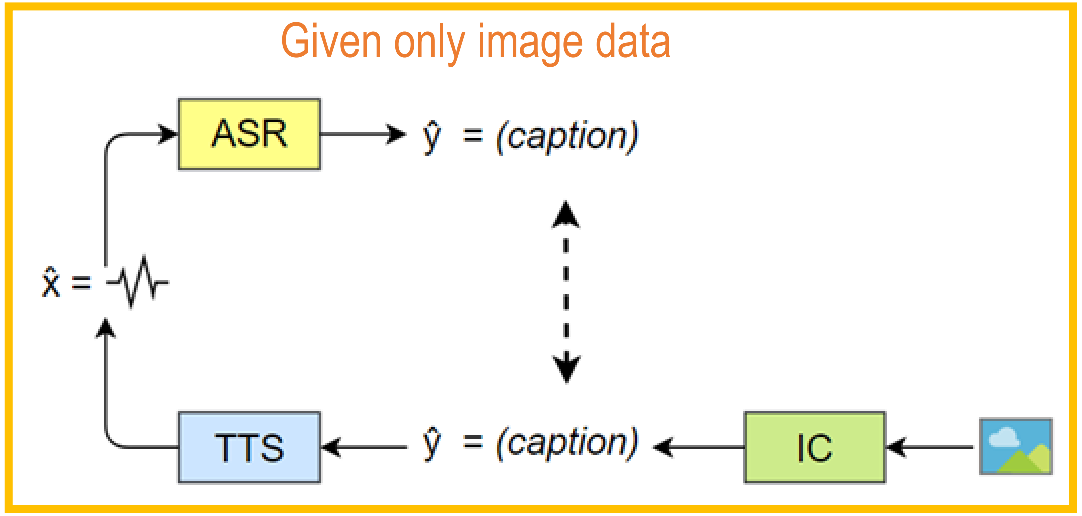
→ **Only text data:** Perform unrolled process
(1) TTS→ASR to update TTS&ASR and generate speech data
(2) IG→IC to update IG&IC and generate image data

→ **Only speech data:**
Unrolled process ASR→TTS to update TTS&ASR and generate text
then use it on IG→IC to update IG&IC and generate image data

→ **Only image data:**
Unrolled process IC→IG to update IG&IC and generate text,
then use it on TTS→ASR to update TTS&ASR and generate speech





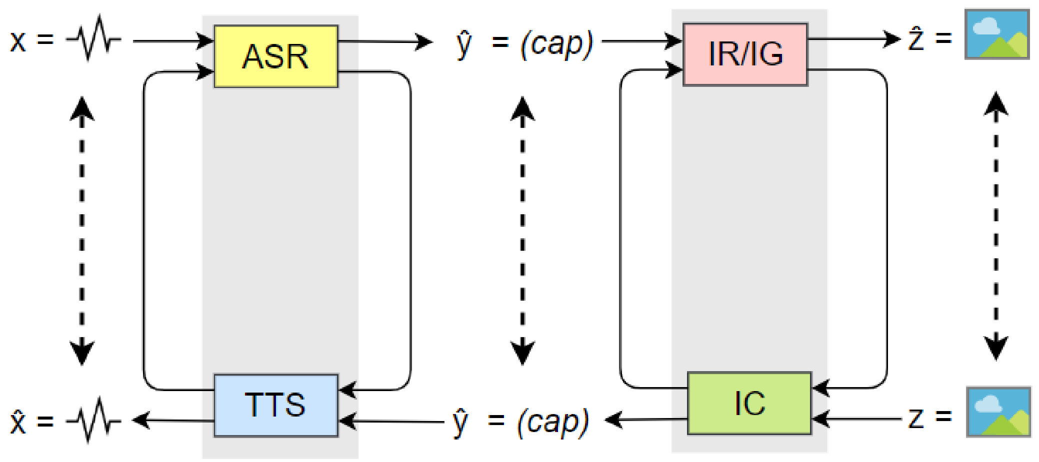


Multimodal Machine Speech Chain

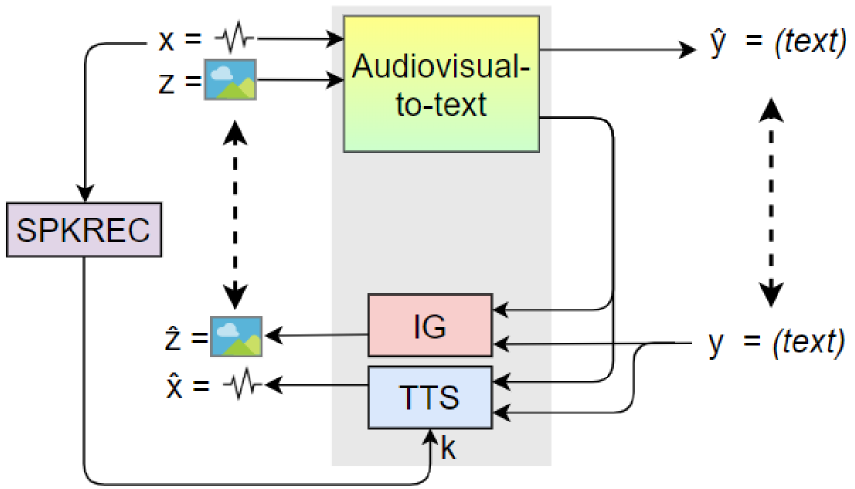
- **Multimodal Machine Speech Chain:**

→ Alternative framework: Simplified single-loop multimodal chain

→ Human brain process visual and auditory components of speech in a unified manner [Calvert, 2001]



MMC1-IR/IG - Dual loop multimodal chain (Proposed)

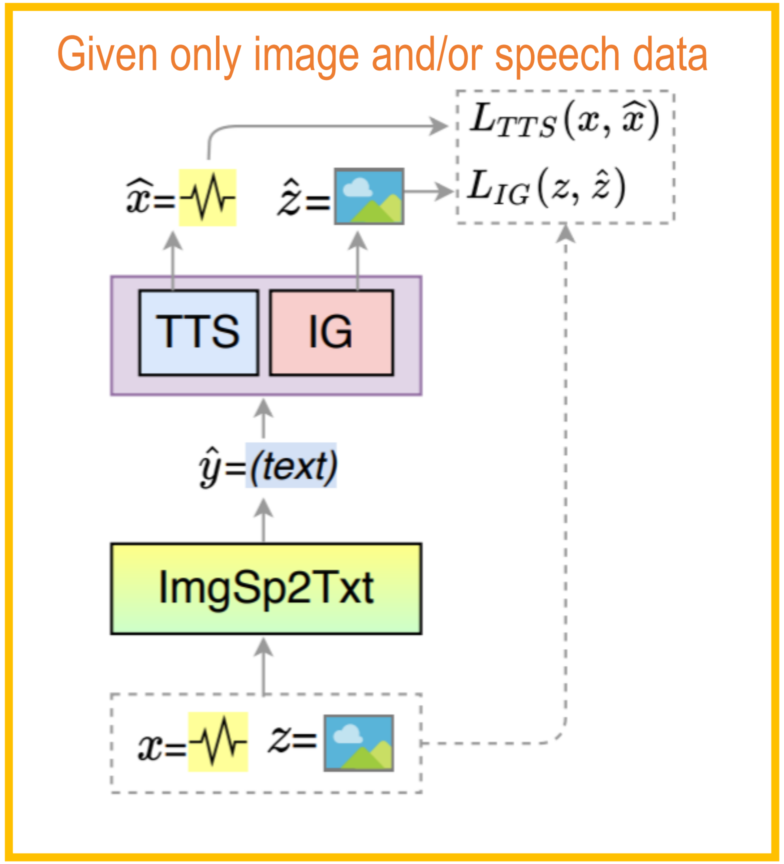


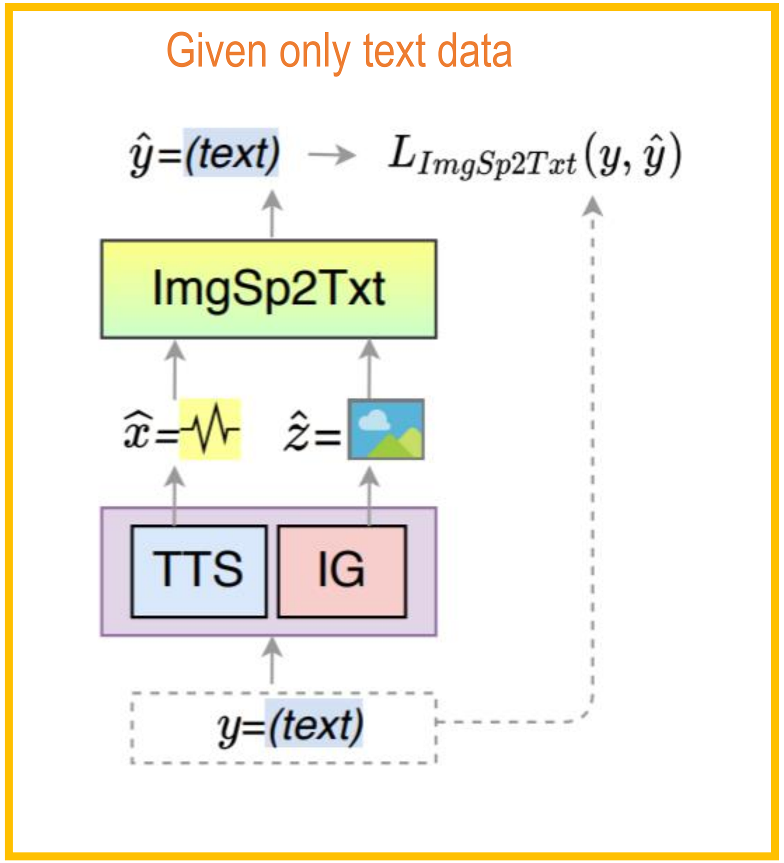
MMC2 – Single loop multimodal chain (Proposed)

Learning in MMC2 Framework

Case #2: Unsupervised Learning with Unpaired Data

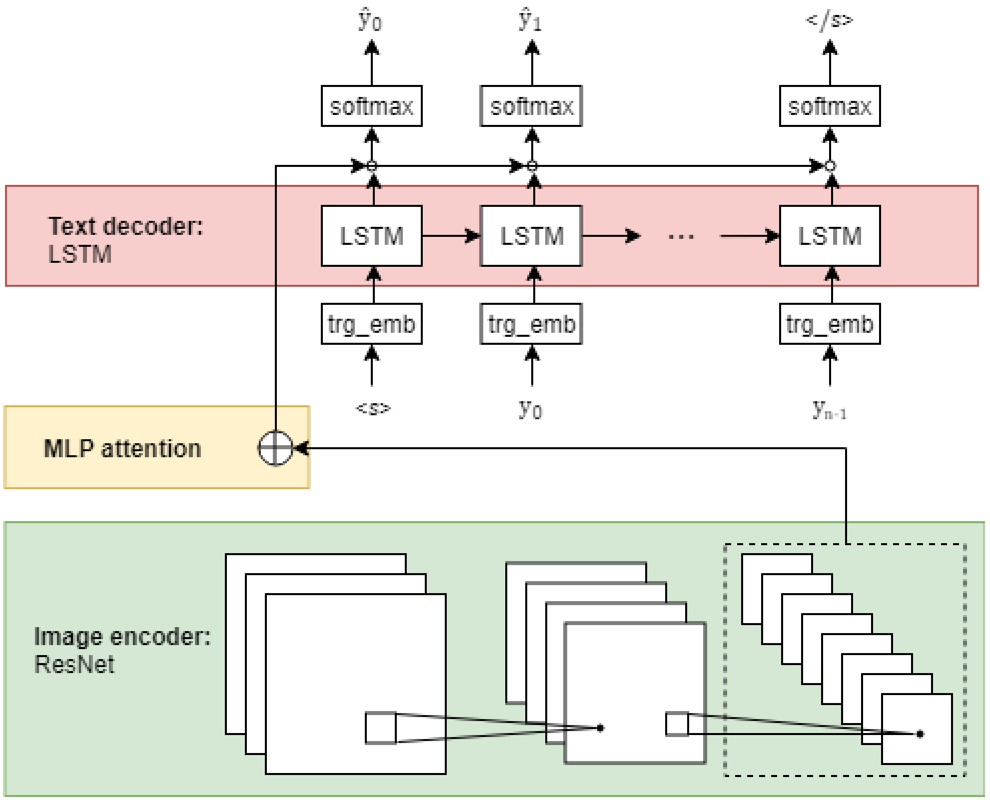
- **Unpaired speech-text-image data exist:**





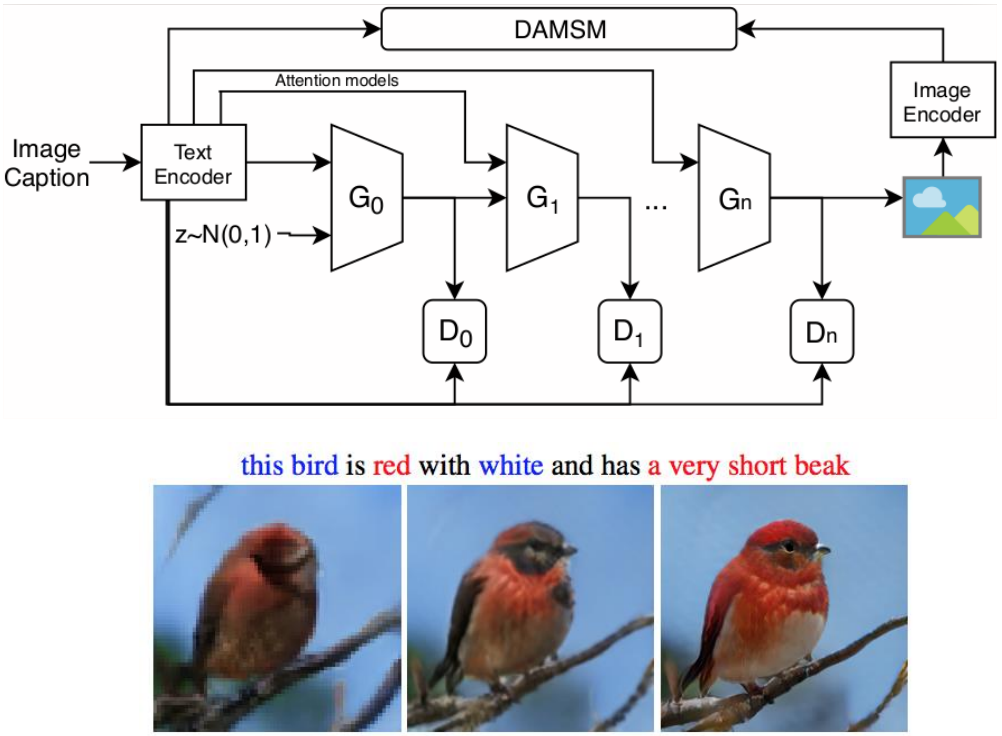
IC and IG Architecture

- **Image Captioning:**



Similar to "Show, attend and tell" [Xu et al. 2015]

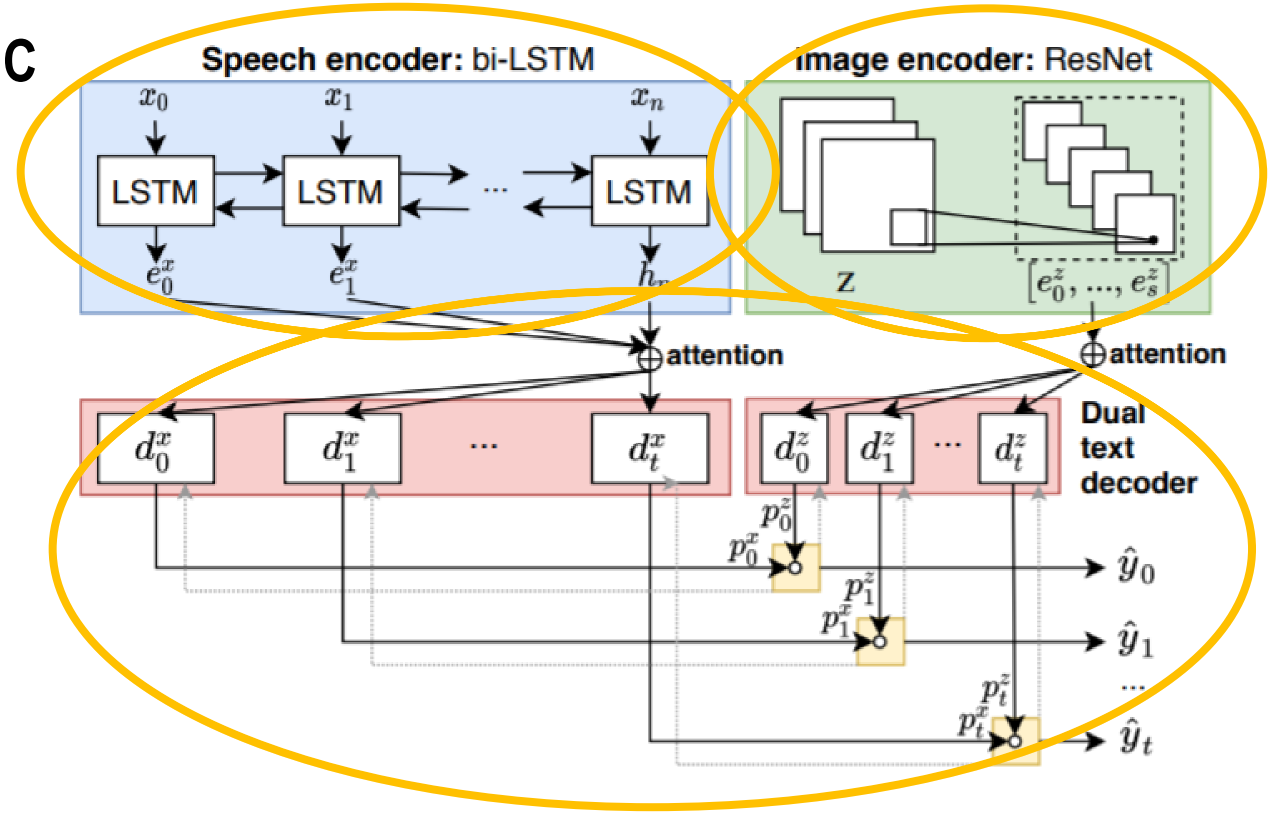
- **Image Generation:**



Similar to "AttnGAN - Multistep image generation using adversarial loss*" [Xu et al., 2017]*

ImgSp2Txt Architecture

- **Image-Speech to Text:**
  - → **Encoders have similar settings with ASR & IC**
  - → **Combine output layer of IC and ASR**
    - use combination for the next step
    - both models trained altogether
  - → **When both image and speech are available:**
    - use average of both output layer
  - → **When only image or speech are available:**
    - use the corresponding modality output layer



Data Set-Up

- **Data**
  → Flickr8k [Ratschian et al., 2010]
  → 8k images, 5 captions/image
  → 65 hours of natural speech
  multi-speaker English speech
  [Harwath and Glass, 2015]

Type	Speech	Text	Image	# Image
Multimodal Paired	○	○	○	800
Multimodal Unpaired	Δ	Δ	Δ	1500
Speech only	Δ	x	x	1850
Image only	x	x	Δ	1850

○ : available paired
Δ : available but unpaired
x : unavailable

ASR-TTS Results

Training	Data Type	#Image	ASR (CER) ↓	IC (BLEU4) ↑	TTS (L2^2 Norm) ↓	IG (Inception) ↑
MMC 1 Dual-loop (Semi-supervised)	Multimodal (P)	800	36.35	12.75	0.77	5.90
	+ Multimodal (U)	1500	15.10	13.22	0.59	8.29
	+ Sp only (U)	1850	12.37	13.28	0.56	9.12
	+ Img only (U)	1850	12.06	13.29	0.56	9.11
Topline MMC1 (Supervised)	Multimodal (P)	6000	5.76	19.91	0.50	9.66
MMC 2 Single-loop (Semi-supervised)	Multimodal (P)	800	26.67	32.23	0.77	5.90
	+ Multimodal (U)	1500	14.88	55.15	0.65	10.12
	+ Sp only (U)	1850	13.81	58.03	0.62	10.65
	+ Img only (U)	1850	12.32	59.66	0.61	9.95
Topline MMC2 (Supervised)	Multimodal (P)	6000	5.16	79.88	0.50	9.66

ASR could still be improved even without speech and text data

Discussion

- **Summary:**
  - Machine speech chain enables semi-supervised learning without parallel data
  - We upgrade the speech chain into the multimodal chain
    by jointly training IC and IG model in a loop connection
  - It allows us to improve ASR even only image data is available

Professional Speech Interpreter

→ The translation process starts before receiving the end of sentence
→ Has the ability to do simultaneous process

Challenges for machine speech interpreter:

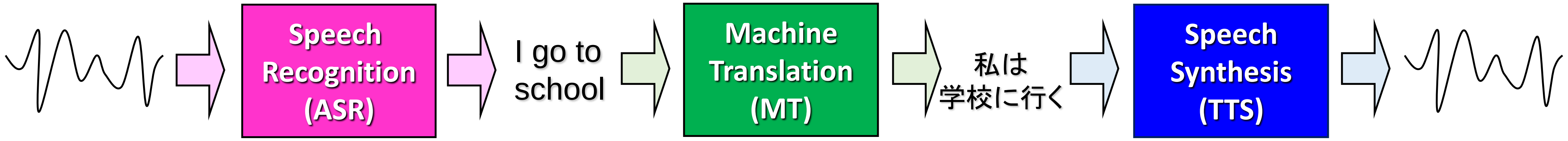
1. Requires the ability to listen while speaking
2. Requires the ability to perform recognition and synthesis speech in real-time

Approach to Problem 2

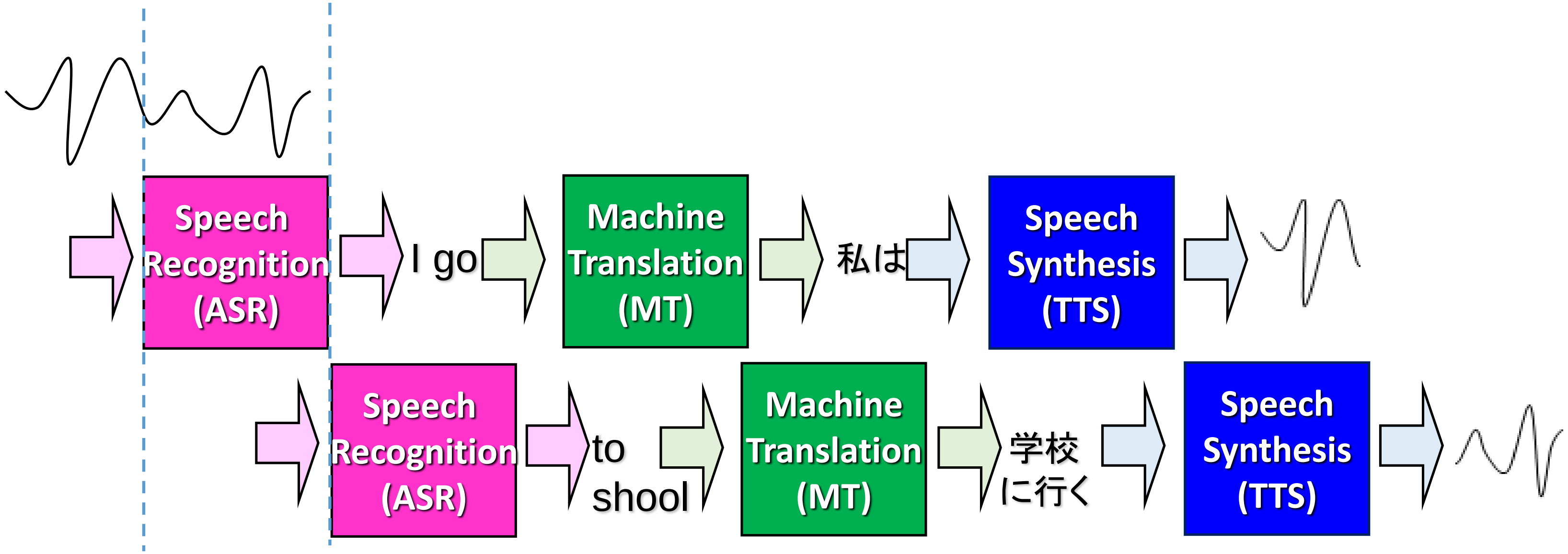
Incremental ASR and TTS for Real-time Machine Speech Interpreter

Real-time Machine Speech Interpreter

- Traditional Speech Translation



- Real-time Machine Speech Interpreter

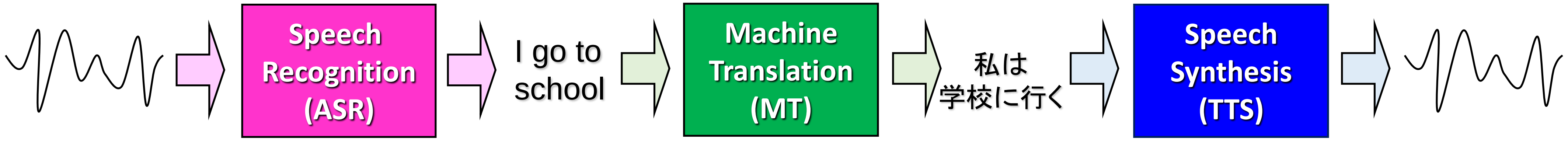


Part 2-1
Neural Incremental Speech Recognition

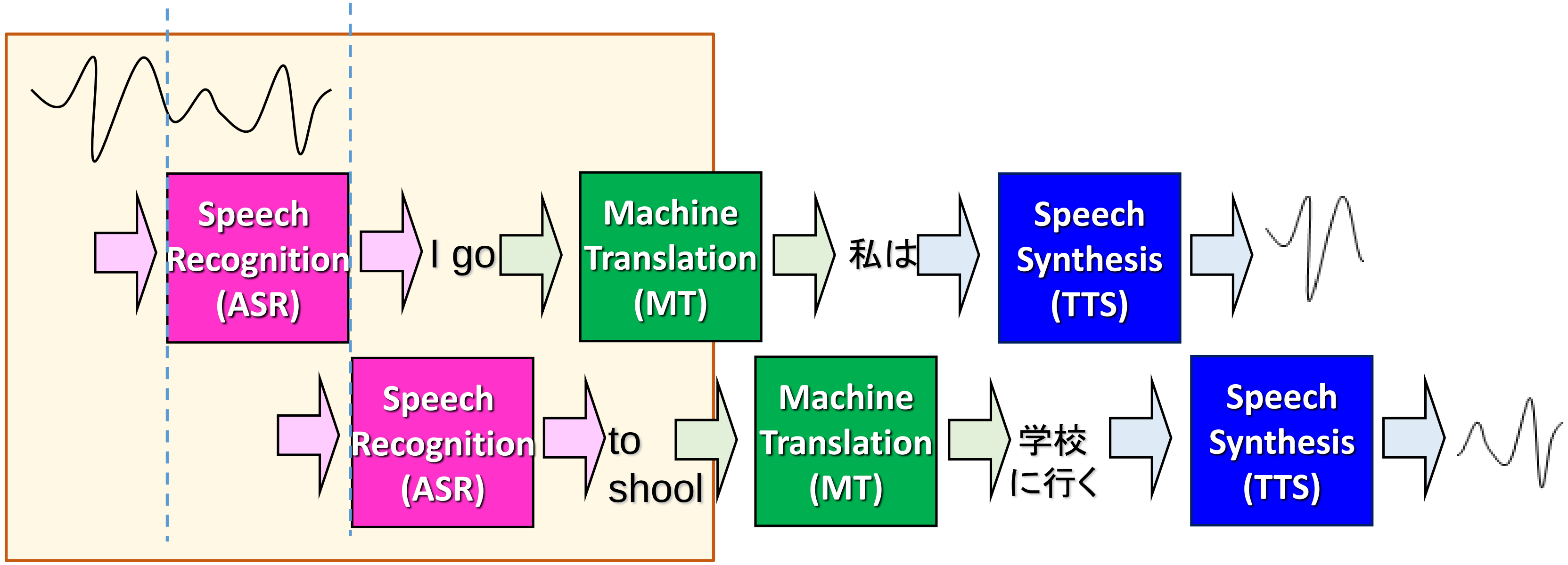
[S. Novitasari, A. Tjandra, S. Sakti, and S. Nakamura, "Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition," in Proc. INTERSPEECH, 2019]

Real-time Machine Speech Interpreter

- ## Traditional Speech Translation



- ## Real-time Machine Speech Interpreter



Neural ASR

- **Current Attention-based Seq2Seq Neural ASR**

→ Standard attention has a "global" property :

Attend whole input sequence and calculate the expected context

→ The output is generated after receiving the entire input sequence

→ Requires a significant delay to recognize long utterances

- **Related Works on Incremental ASR (ISR)**

→ Local attention mechanism [Bahdanau et al., 2014; Tjandra et al., 2017]:

Limit the area of attention, but without reducing the latency

→ Character-level Incremental Speech Recognition with Recurrent Neural Networks [Hwang et al., 2016]

- Unidirectional RNN-CTC
- Requires depth-pruning in beam search during output generation

→ Neural transducer [Jaitly et al., 2016]

- Seq2seq model: recognize speech part-by-part
- Requires to infer alignment during training
- Use dynamic programming to approximate best alignment

Neural ASR

- **Problems**

  → Most existing Neural ISR models utilize different frameworks and learning algorithms that are more complicated than standard Neural ASR

  → The models need to decide: (1) Incremental steps
(2) Learn the transcription that aligns with the current short speech segment

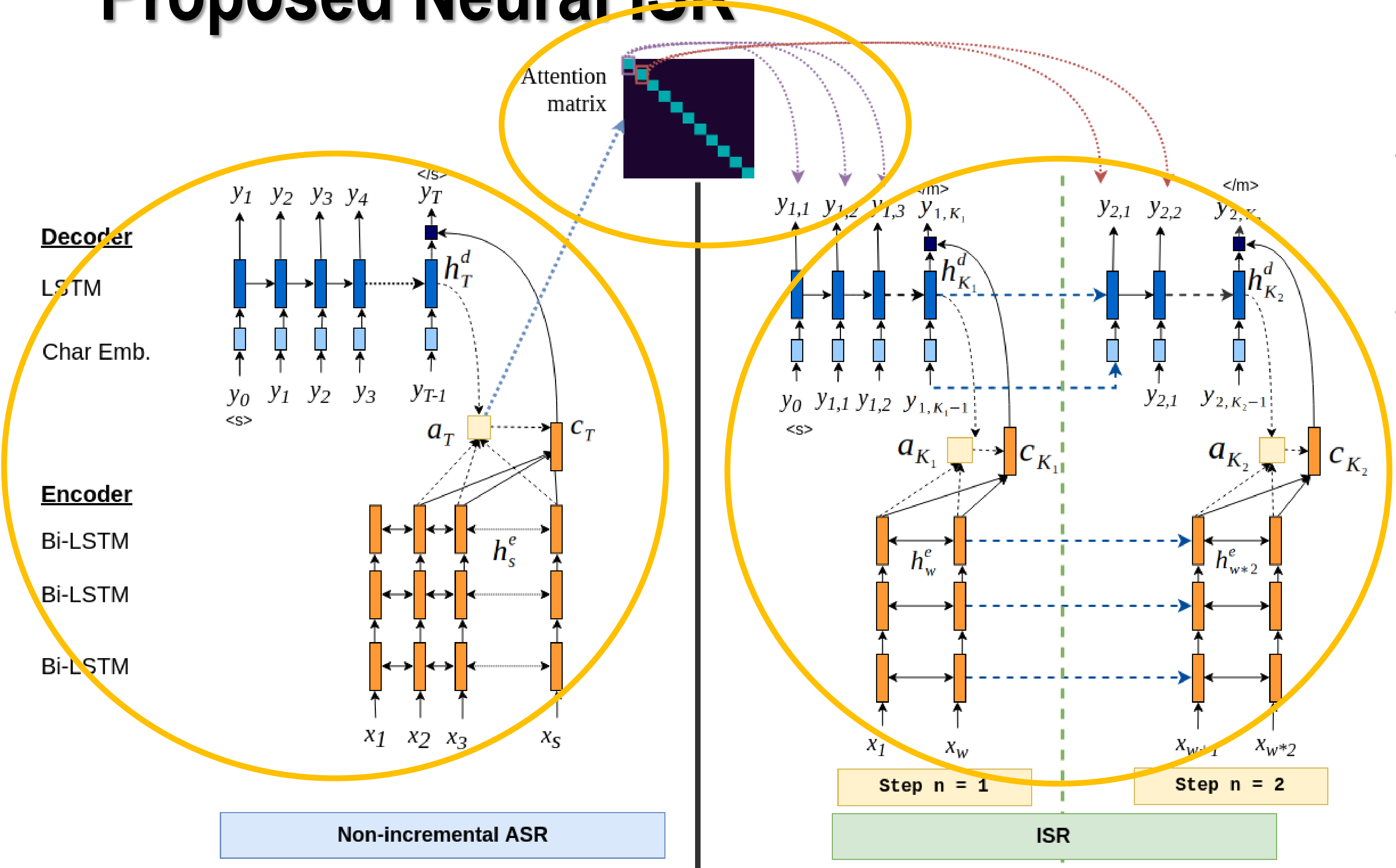
- **Proposed Neural ISR**

  → Employing the original architecture of attention-based ASR with shorter sequences

  → Perform attention transfer: full-utterance ASR as the teacher model and ISR as the student model

  → Mimics the speech-text alignment produced by standard ASR

Proposed Neural ISR



- **In training,** if it reaches the last token aligned to the block, the decoder will learn to output an <\m> symbol
- In actual use, decoding in each step stops when:
  - <\m> predicted, or
  - token length. reaches a threshold (max)

  <\m> → end-of-block
- **Notation:**

x – speech frame
y – token
w – frame block size
h^e – encoder state
h^d – decoder state
a – attention
c – attention context

Experimental Results

▪ Data set

- **Training set: Supervised (paired text & speech)**
  WSJ SI-284 dataset [Paul et al. , 1992]
  (37318 utterances, ~81 h, 284 speakers)
- **Development set:** dev93
- **Evaluation set:** eval92

Main (blocks)	Look-ahead (blocks)	Delay (sec)	CER %	WER %
Non-incremental				
CTC [Kim et al.,2017]		7.88 (avg)	8.87	-
Att Enc-Dec Location [Kim et al.,2017]			8.17	18.60
Joint CTC+Att (MTL) [Kim et al.,2017]			7.36	-
Att Enc-Dec (teacher,greedy)			6.80	17.40
Proposed ISR				
1	1	0.24	19.78	43.54
1	4	0.54	**8.71**	**22.54**

Discussion

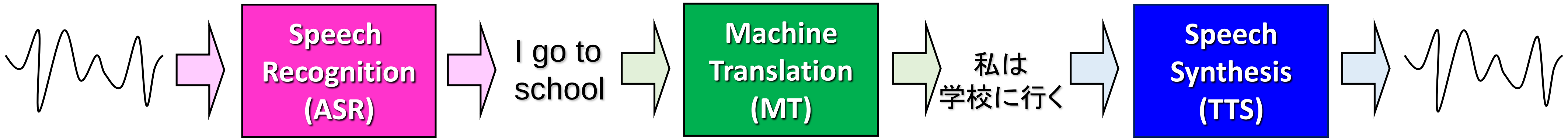
- **Summary:**
  - Develop incremental ASR that employ original architecture of neural ASR
  - Perform transfer learning:
    - Treat standard ASR as a teacher model and ISR as a student model
  - Experimental results:
    - Successfully reduced the delay
    - Achieved comparable performance than standard ASR that wait until the end

Part 2-2
Neural Incremental Speech Synthesis

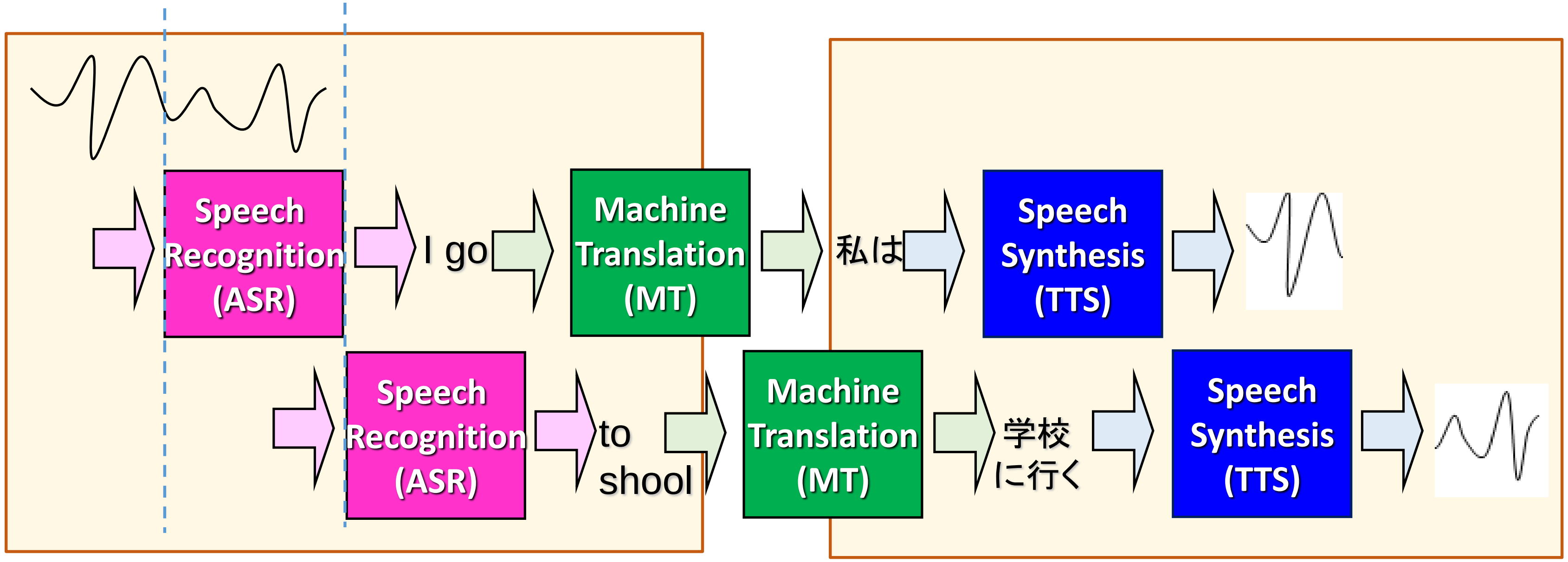
[T. Yanagita, S. Sakti, S. Nakamura, "Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework," in Proc. Speech Synthesis Workshop (SSW), 2019]

Real-time Machine Speech Interpreter

- Traditional Speech Translation

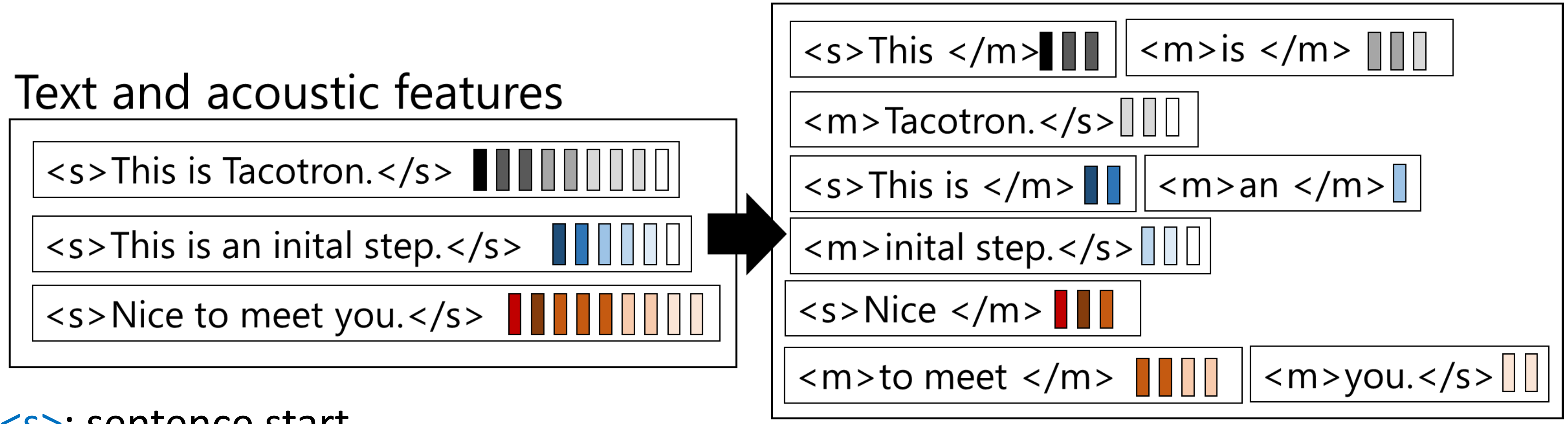


- Real-time Machine Speech Interpreter



Neural ITTS

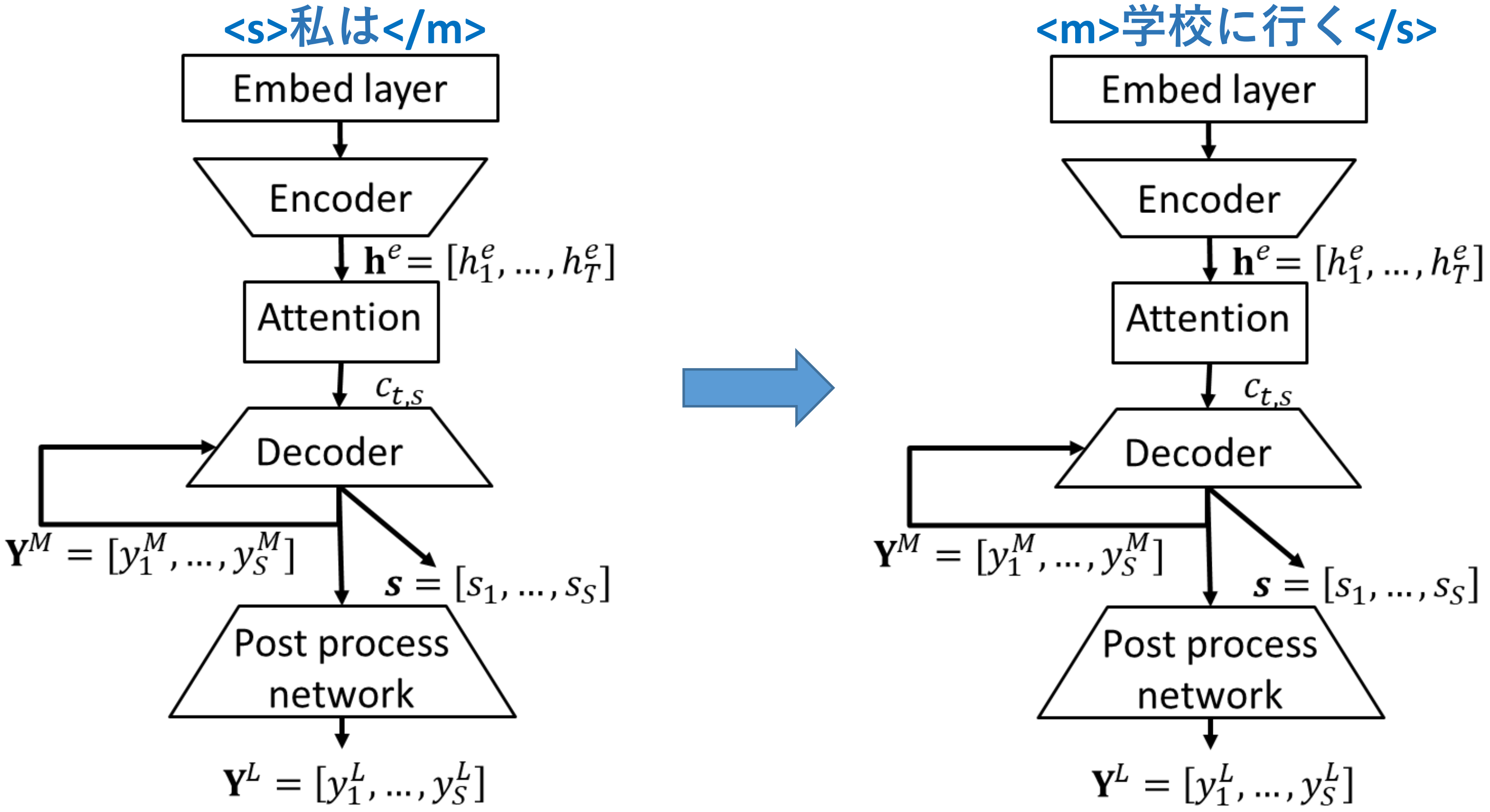
- ## Training
  - Randomly splitting the sentence into shorter parts



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

Neural ITTS

- **Architecture**
  - **Full-sentence:** <s> 私は学校に行く </s>
  - **ITTS perform incremental on shorter units**



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

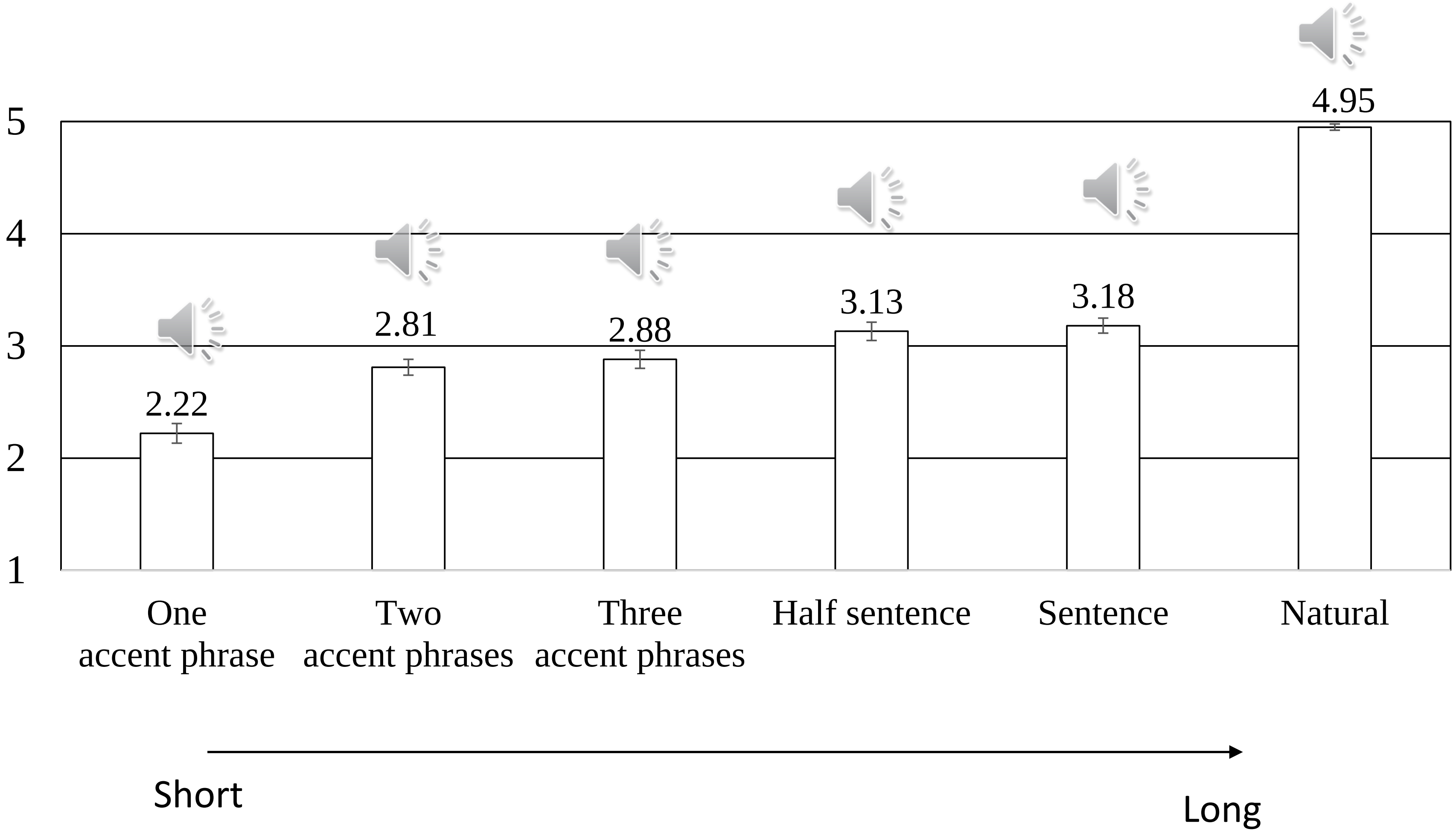
Experiment Set-Up

- **Set-Up:**

Language	Japanese
Dataset	JUST(-10h) [Sonobe et al., 2017]
Sampling rate	22.05kHz
Train/Dev/Test set	5k/100/100
Input dimensions	45 phoneme symbols 20 accent types
Output Acoustic features	80 dim. Mel-spectrogram 1024 linear-spectrogram
Frame shift Frame lenght	12.5ms 5ms
Waveform generation	Griffin-Lim algorithm

Experiment Results: Subjective Evaluation

- **MOS Results:**



Discussion

- **Summary:**
  - **Develop neural ITTS based on sequence-to-sequence framework**
  - **Experimental results:**
    - Linguistic feature of accent phrase is critical
      when the next linguistic features are missing
    - The optimum incremental synthesized units was between
      the three accent phrases and the half-sentence units

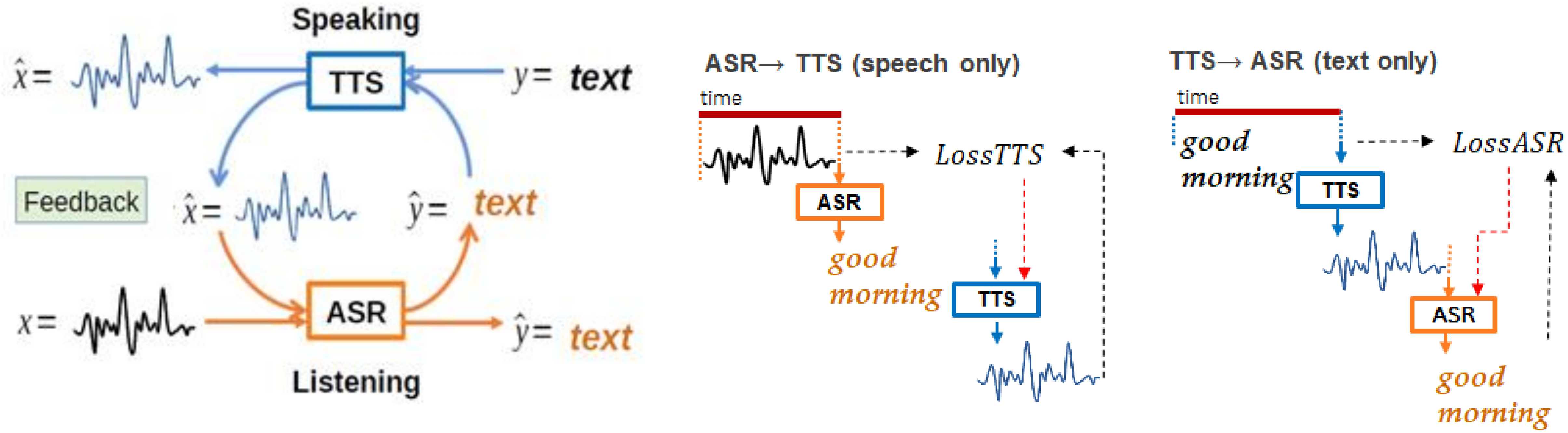
Approach to Problem 1&2

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

[Sashi Novitasari, Andros Tjandra, Tomoya Yanagita, Sakriani Sakti, Satoshi Nakamura, "Incremental Machine Speech Chain Towards Enabling Listening while Speaking in Real-time," Proc. of INTERSPEECH, Oct 2020]

Machine Speech Chain

- **Speech Chain Mechanism**
  - ASR to TTS process using speech data only to improve TTS
  - TTS to ASR process using text data only to improve ASR



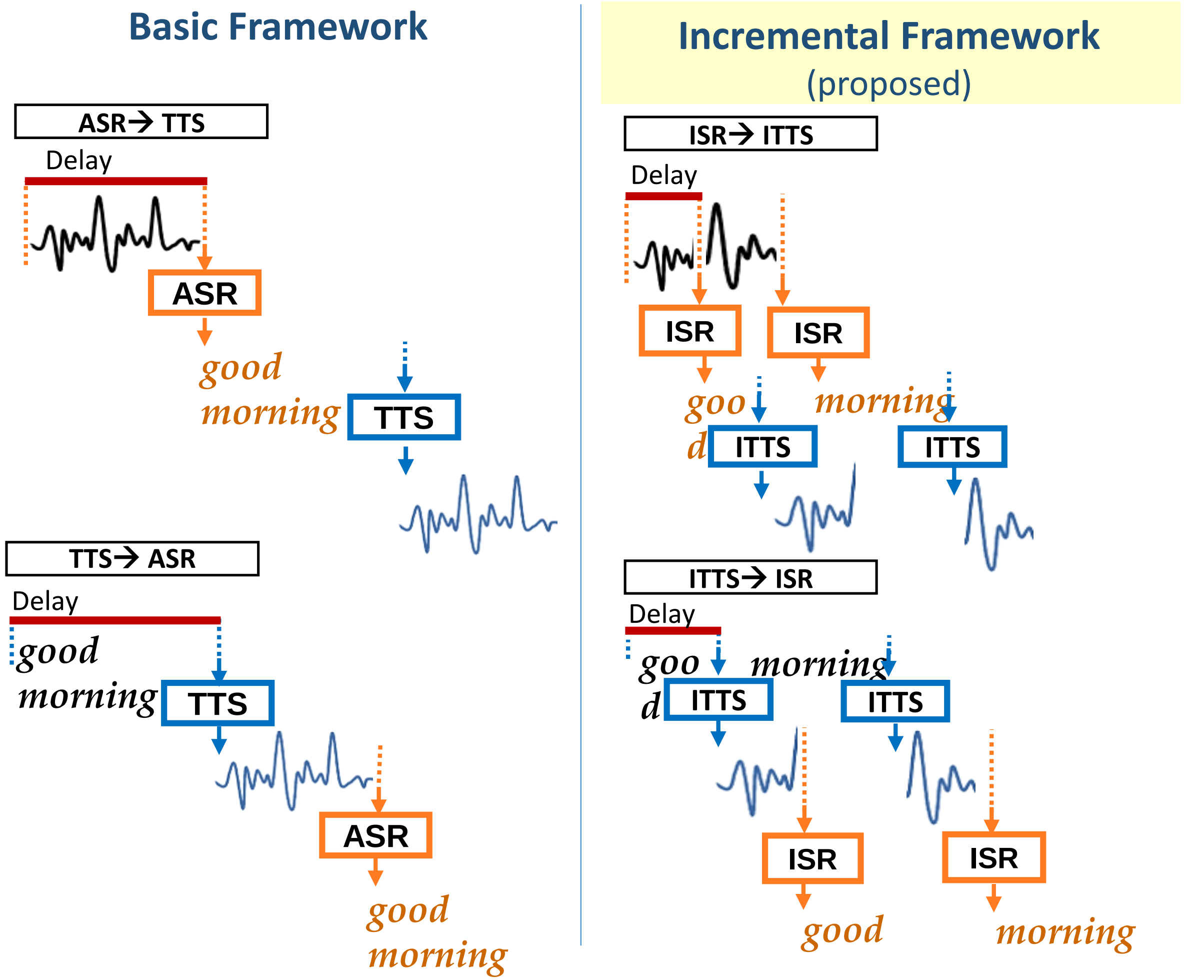
Full-utterance-based ASR and TTS → High delay

Incremental Machine Speech Chain

- **Objective:**

A 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

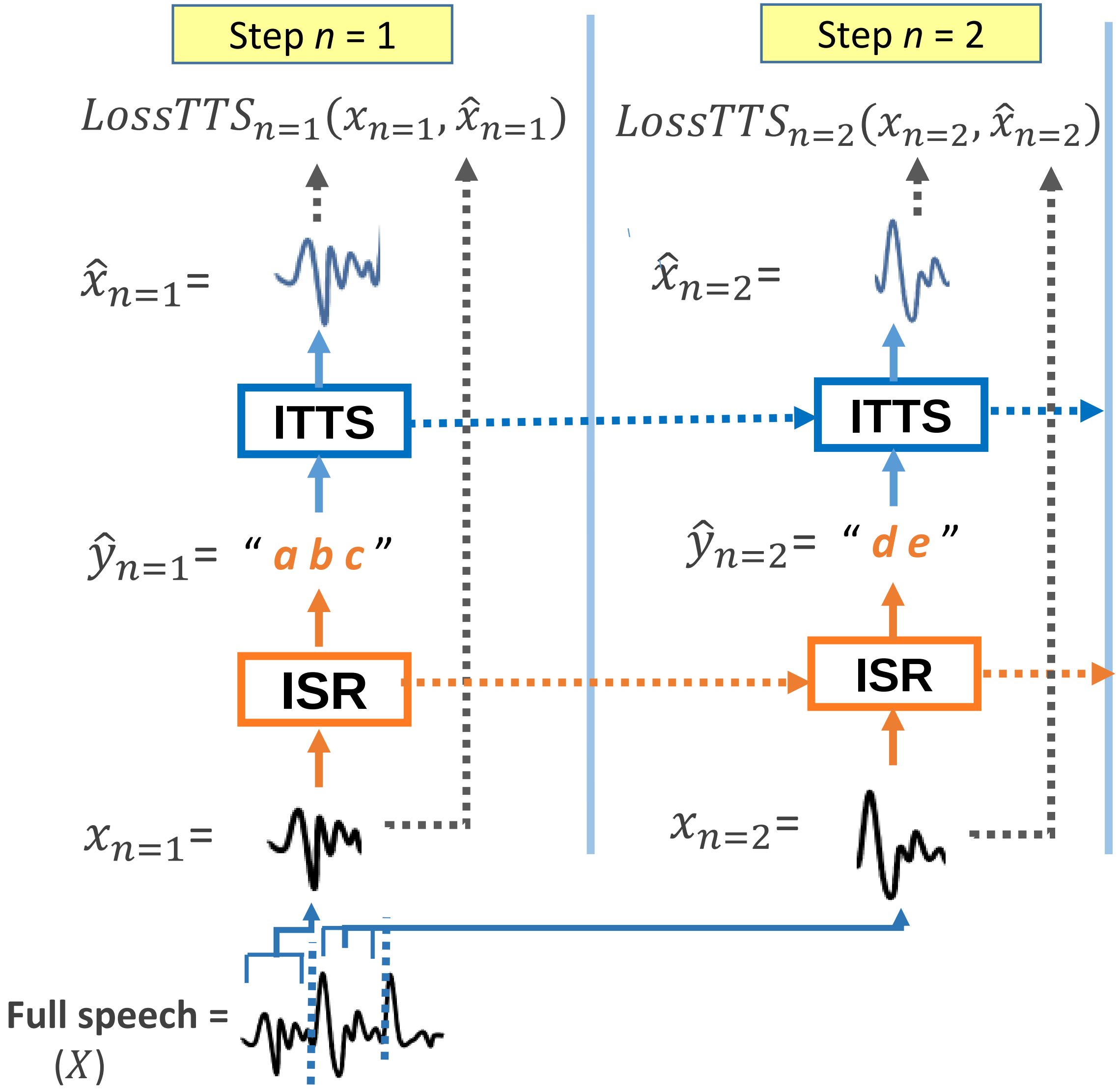


Learning in Incremental Machine Speech Chain

ISR and ITTS Joint Training

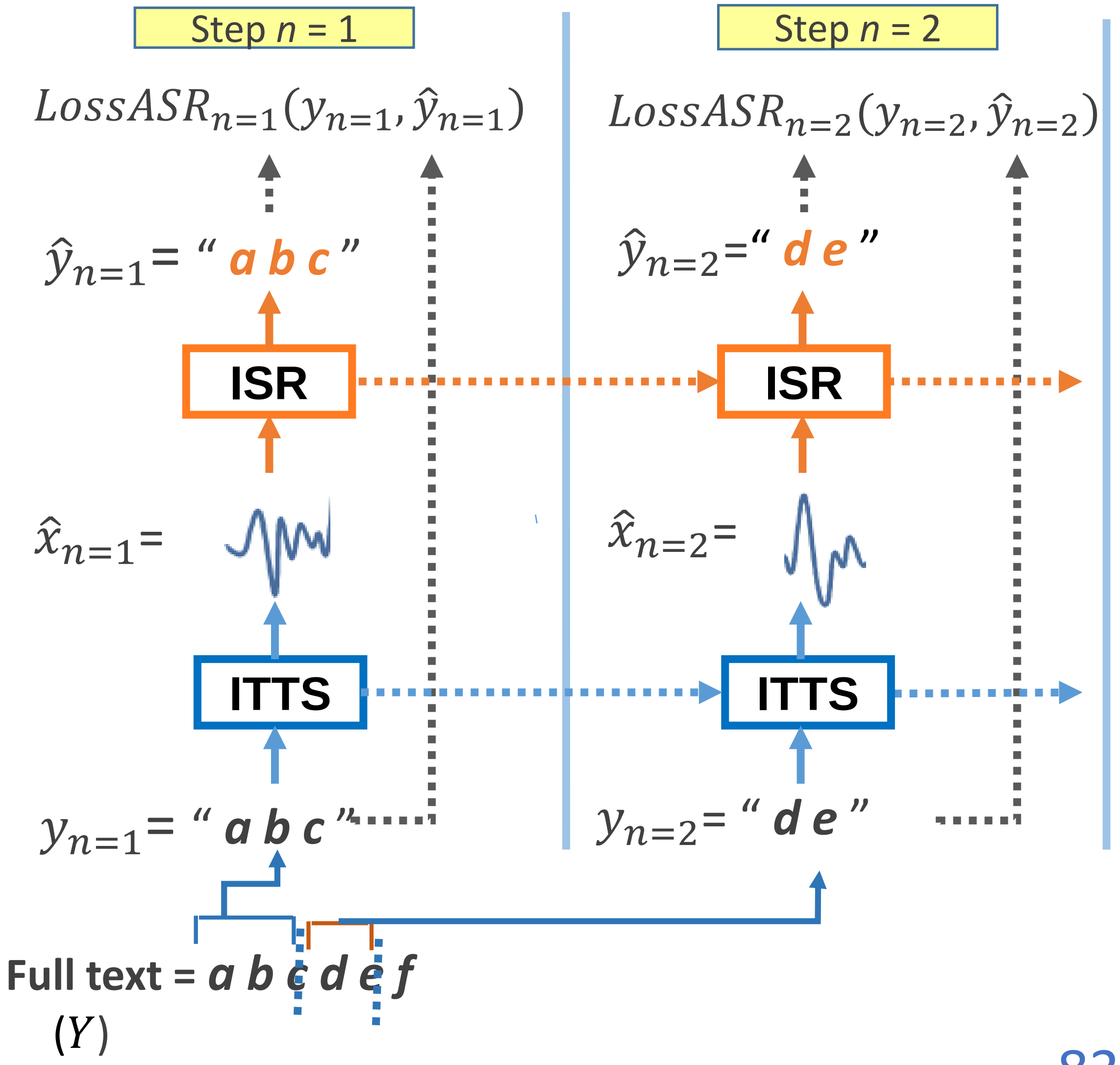
- A short-term feedback loop between the components
- Segment-based output passing
- Unrolled processes

  a. ISR-to-ITTS
  For each step n, ISR predicts $\hat{Y}_n$ from X_n, and then ITTS predicts $\hat{X}_n$ from ISR output $\hat{Y}_n$



Learning in Incremental Machine Speech Chain

- **ISR and ITTS Joint Training**
  - **A short-term feedback loop between the components**
  - **Segment-based output passing**
  - **Unrolled processes**
    a. ISR-to-ITTS
    For each step *n*, ISR predicts $\hat{Y}_n$ from X_n , and then ITTS predicts $\hat{X}_n$ from ISR output $\hat{Y}_n$
    b. ITTS-to-ISR
    For each step *n*, ITTS predicts $\hat{X}_n$ from Y_n, and then ISR predicts $\hat{Y}_n$ from ITTS output $\hat{X}_n$



Incremental Machine Speech Chain

- **ASR and TTS Results:**

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*	17.33	27.03	17.81	44.54	0.99	1.02	1.04	3.62
Indep-trn *SI-284*	7.16	9.60	7.97	19.99	0.75	0.77	0.84	1.31
Machine Speech Chain								
Indep-trn (SI-84) + chain-trn-greedy (SI-200)	11.21	11.52	14.23	32.43	0.80	0.82	0.86	1.35
Indep-trn (SI-84) + chain-trn-teachforce (SI-200)	7.27	6.30	9.43	12.78	0.77	0.80	0.79	1.26

Conclusions and Future Directions

Conclusions and Future Directions

- **Conclusions:**
  - We have constructed a machine speech chain that can listen, speak, and listen while speaking
  - Currently, we mostly utilize it to achieve semi-supervised learning
  - On the other hand, we have also constructed ISR and ITTS
  - Combined ISR and ITTS within incremental machine speech chain framework

- **Future Directions:**

Develop a Real-Time Neural Machine Speech Interpreter that Listen, Translate, Speak, and Listen while Speaking and Translating

Citations

- **[Sakoe et al., 1971]** – H. Sakoe and S. Chiba, “A dynamic programming approach to continuous speech recognition,” in Proc. 7th ICA, 1971
- **[Baum et al., 1966]** – L.E Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," The Annals of Mathematical Statistics, 1966
- **[Waibel et al. 1989]** – A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, “Phoneme recognition using time-delay neural networks,” IEEE Trans. on Acoustic Speech, and Signal Processing, 1989
- **[Hochreiter et al., 1997]** – S. Hochreiter; J. Schmidhuber, "Long short-term memory," Neural Computation, 1997
- **[Mohri et al., 2002]** – M. Mohri, F. Pereira, and M.I Riley, “Weighted finite-state transducers in speech recognition,” Computer Speech and Language, 2002
- **[Bourlard et al., 1993]** – H. Bourlard, N. Morgan, “Connectionist Speech Recognition: A Hybrid Approach,” Kluwer Academic Publishers, 1993
- **[Graves et al., 2006]** – A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, ´”Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with RNN,” in Proc. ICML, 2006
- **[Chan et al., 2016]** – W. Chan, N. Jaitly, Q. Le, and O. Vinyals, et al., “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. ICASSP, 2016
- **[Xiong et al., 2017]** -- W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition“, Microsoft Research Technical Report MSR-TR-2016-71, 2017
- **[Saon et al., 2017]** -- G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L. Lim, B. Roomi, P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines“, ASRU 2017
- **[Zen et al., 2009]** -- H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” Speech Communication, 2009
- **[Wang et al., 2017]** -- Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio et al., “Tacotron: A fully end-to-end text-to-speech synthesis model,” arXiv preprint, arXiv:1703.10135, 2017
- **[Oord et al., 2016]** A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016
- **[Denes & Pinson, 1993]** -- P. Denes and E. Pinson, “The Speech Chain”, ser. Anchor books. Worth Publishers, 1993
- **[Kikui et al. , 2003]** – G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, “Creating Corpora for Speech-to-Speech Translation,” Proc. of European Conf. on Speech Communication and Technology, 2003
- **[Li et al. , 2017]** – C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, “Deep speaker: an end-to-end neural speaker embedding system,” arXiv preprint arXiv:1705.02304, 2017
- **[Paul et al. , 1992]** – D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in Proc. of the workshop on Speech and Natural Language, ACL, 1992
- **[Bahdanau et al. , 2016]** – D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in Proc. ICASSP, 2016
- **[Tjandra et al. , 2017]** – A. Tjandra, S. Sakti, and S. Nakamura, “Local monotonic attention mechanism for end-to-end speech and language processing,” in Proc. IJCNLP, 2017
- **[Hwang et al. , 2016]** – K. Hwang and W. Sung, “Character-level incremental speech recognition with recurrent neural networks,” in Proc. ICASSP, 2016
- **[Jaitly et al. , 2016]** – N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, and S. Bengio, “An online sequence-to-sequence model using partial conditioning,” in Proc. NIPS, 2016
- **[Ito et al. , 2017]** – K. Ito, “The LJ speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017
- **[Kurematsu et al. , 1990]** – A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, ”ATR Ja Speech Database as a Tool of Speech Recog. & Synthesis,” Speech Communication, 1990
- **[Sonobe et al. , 2017]** – 23] R. Sonobe, S. Takamichi, and H. Saruwatari, “Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis,” arXiv preprint arXiv:1711.00354, 2017

Publications

General Machine Speech Chain Framework
- A. Tjandra, S. Sakti, S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", in Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, 2017
- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", in Proc. INTERSPEECH, 2018
- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. IEEE ICASSP, 2019
- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain," IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), Vol. 28, pp. 976-989, 2020

Multilingual Machine Speech Chain
- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Speech Chain for Semi-Supervised Learning of Japanese-English Code-Switching ASR and TTS", in Proc. SLT, 2018
- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Zero-shot Code-switching ASR and TTS with Multilingual Machine Speech Chain,“ in Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, 2019
- S. Novitasari, A. Tjandra, S. Sakti, S. Nakamura, "Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis", in Proc. SLTU, 2020

Multimodal Machine Speech Chain
- J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, “Listening while Speaking and Visualizing: Improving ASR through Multimodal Chain,” in Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, 2019
- J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, "Augmenting Images for ASR and TTS through Single-loop and Dual-loop Multimodal Chain Framework," in Proc. of INTERSPEECH, pp. to appear, 2020

Incremental (Real-time) ASR and TTS
- S. Novitasari, A. Tjandra, S. Sakti, and S. Nakamura, "Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition," in Proc. INTERSPEECH, 2019
- T. Yanagita, S. Sakti, S. Nakamura, "Incremental TTS for Japanese Language," in Proc. INTERSPPECH, 2018
- T. Yanagita, S. Sakti, S. Nakamura, "Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework," in Proc. SSW, 2019

Incremental (Real-time) Machine Speech Chain
- S. Novitasari, A. Tjandra, T. Yanagita, S. Sakti, S. Nakamura, "Incremental Machine Speech Chain for Enabling Listening while Speaking in Real-time," in Proc. of INTERSPEECH, pp. to appear, 2020

Thank you