

Semi-supervised Learning by Machine Speech Chain for Multilingual Speech Processing, and Recent Progress on Automatic Speech Interpretation

Satoshi Nakamura,
Sakriani Sakti, and Katsuhito Sudoh

Graduate School of Science and Technology,
Nara Institute of Science and Technology, Japan



NAIST

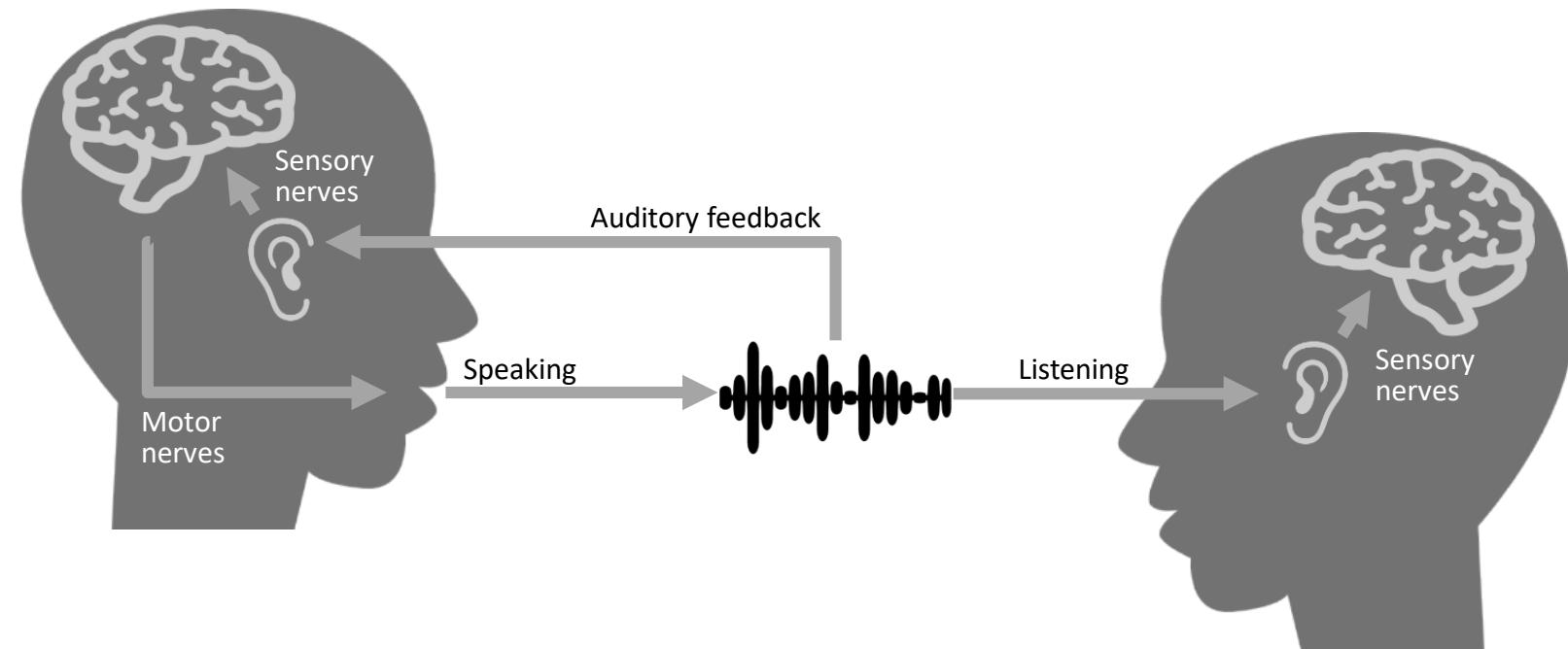
Topics

- ▶ Recent advances in speech processing
 - ASR and TTS research
 - Machine Speech Chain unifies ASR and TTS
 - Application to code switching speech

- ▶ Speech Translation
 - Recent Progress on Automatic Speech Interpretation

Motivation Background

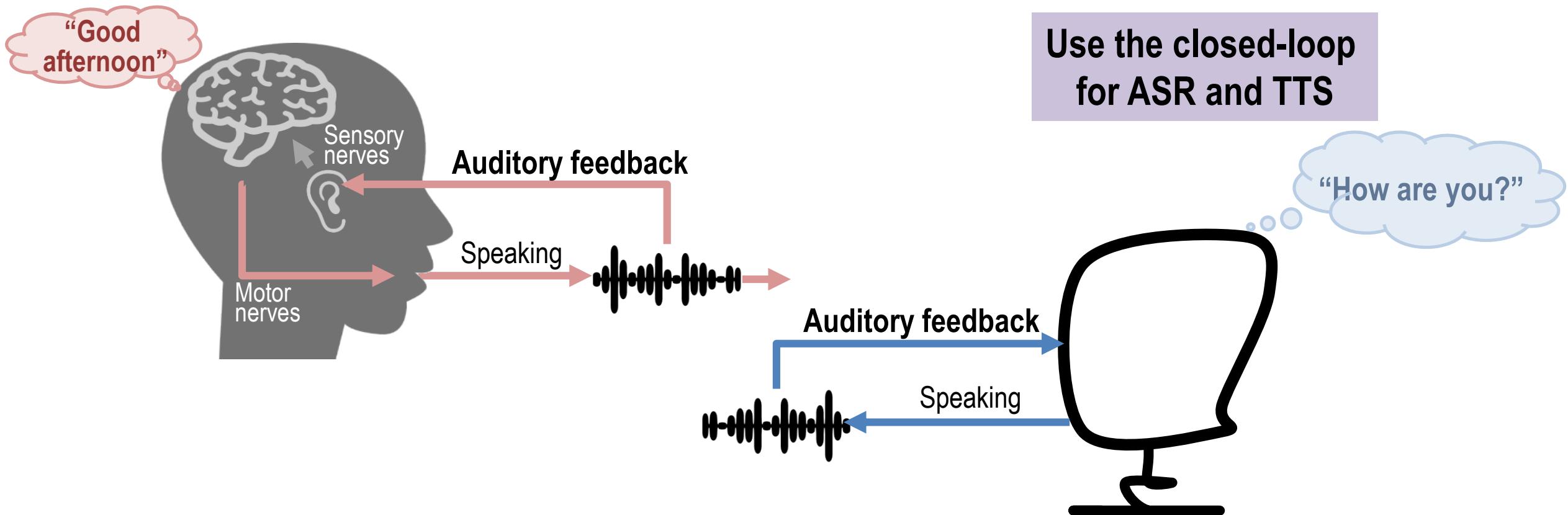
- ▶ In human communication
 - A closed-loop speech chain mechanism has a critical auditory feedback mechanism (“Speech Chain”, Denes, Pison 1973)



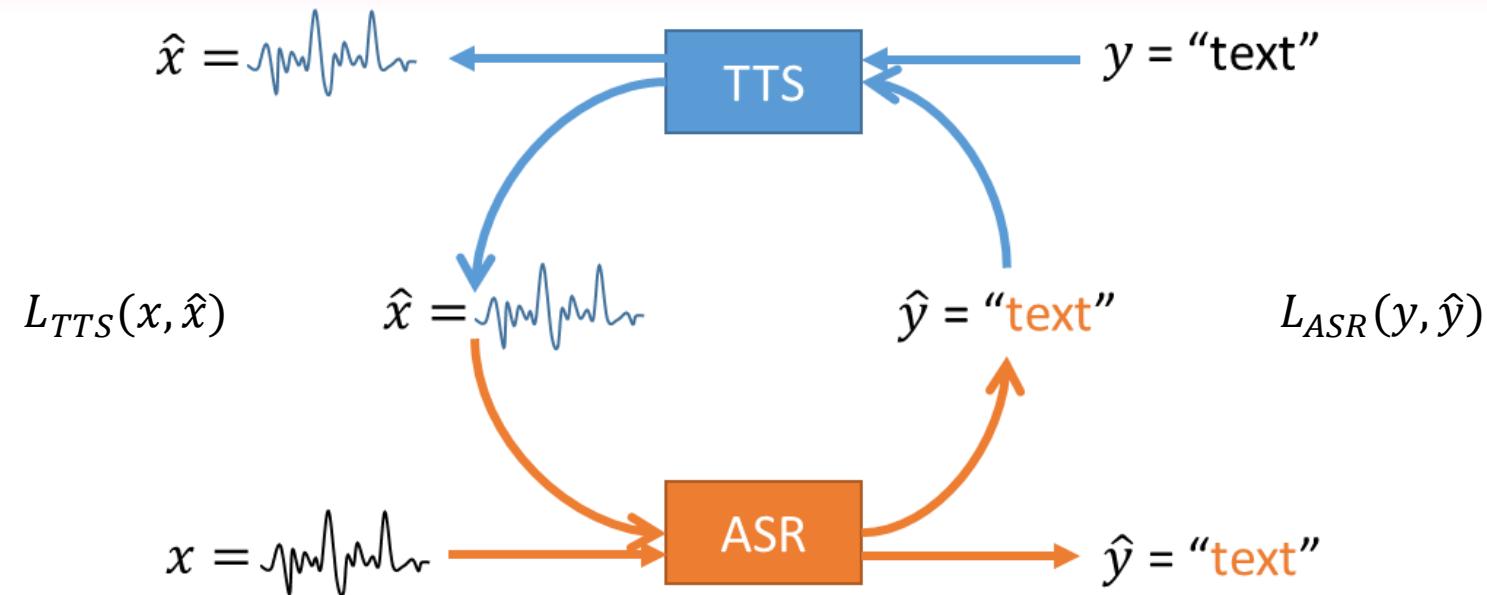
Machine Speech Chain

■ Proposed Method

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



Machine Speech Chain



▶ Definition:

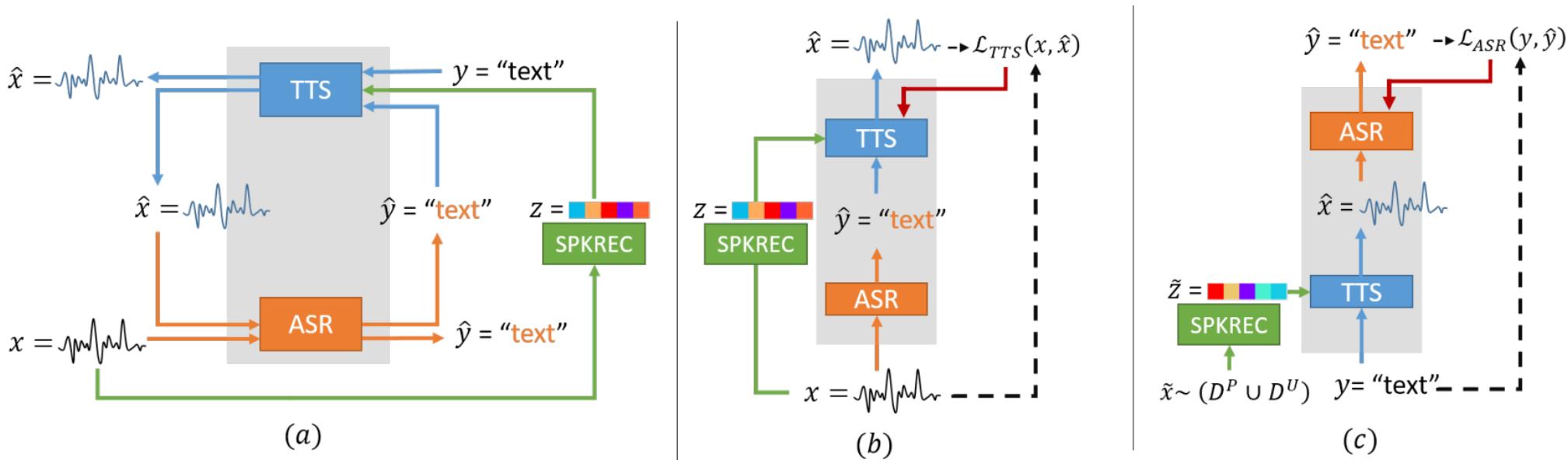
- x = original speech, y = original text
- \hat{x} = predicted speech, \hat{y} = predicted text
- $ASR(x): x \rightarrow \hat{y}$ (seq2seq model transforms speech to text)
- $TTS(y): y \rightarrow \hat{x}$ (seq2seq model transforms text to speech)

Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", Proc. IEEE ASRU 2017

Speech Chain with One-shot Speaker Adaptation

► Proposed model

- Train ASR and TTS models **using unpaired data** and small amount of paired data.
- Speaker individuality is generated by SPKREC embedding.



Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", Proc. INTERSPEECH 2018

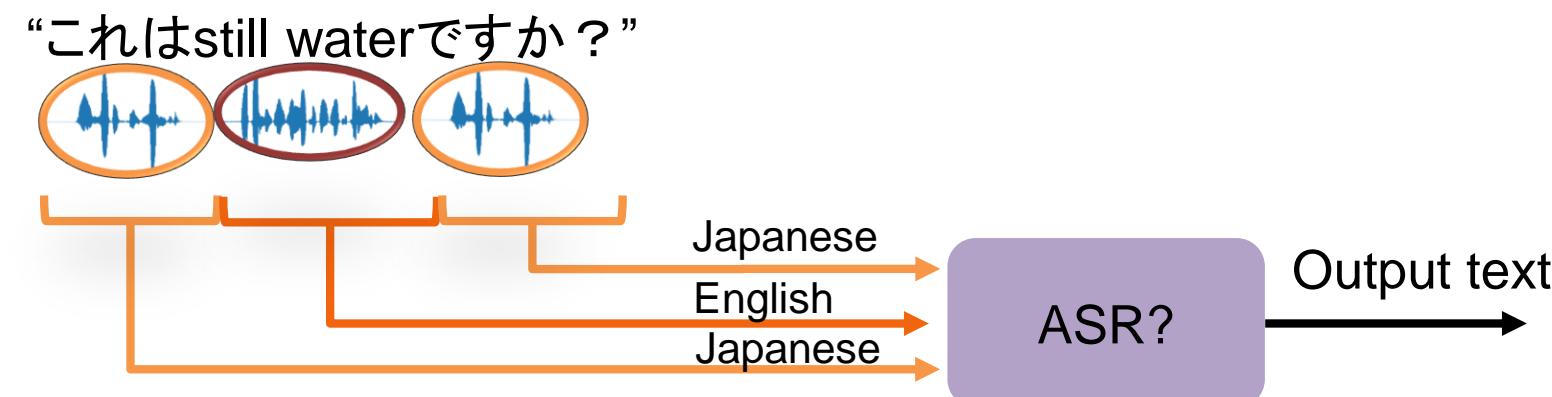
Code-switching Challenges For ASR

- Typical case where paired speech and transcription are difficult to collect.

- Standard ASR is monolingual



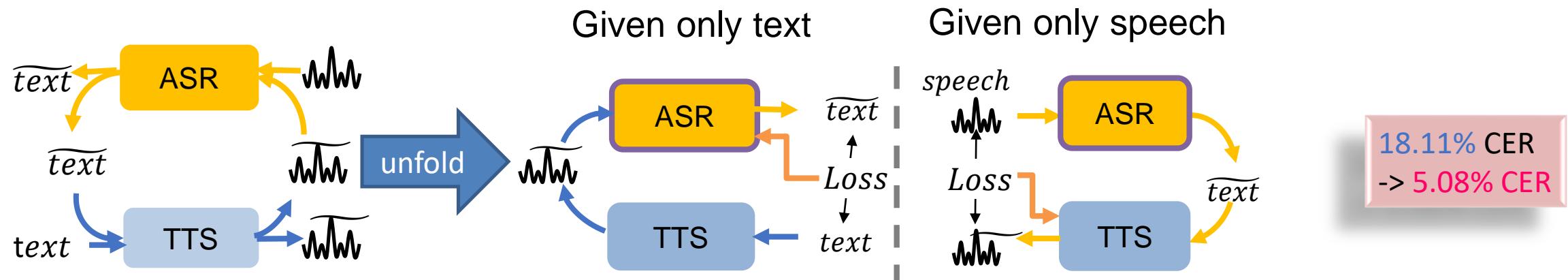
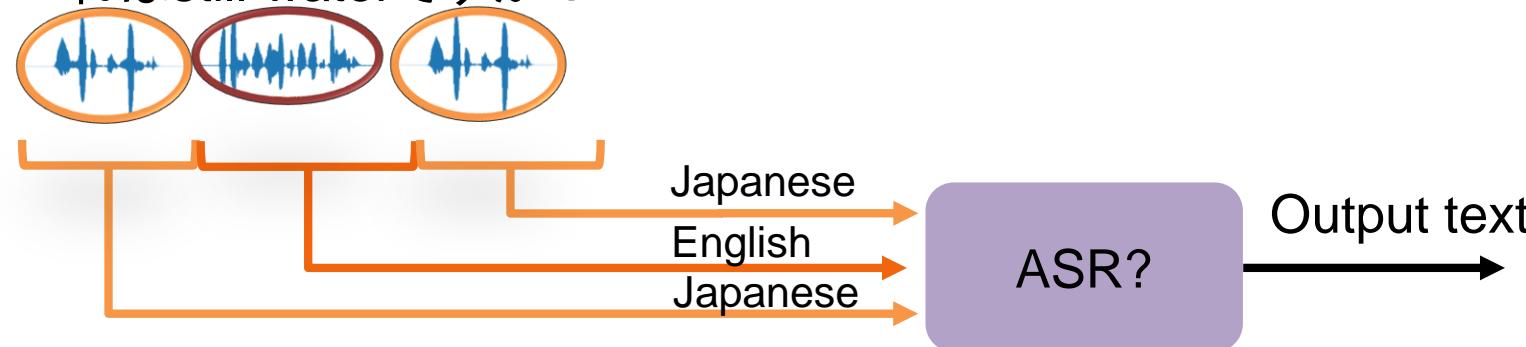
- Challenge with Code-switching : Mixed multilingual input



Code-switching Challenges

- Typical case where paired speech and transcription are difficult to collect.
 - Challenge with Code-switching : Mixed multilingual input

“これはstill waterですか？”

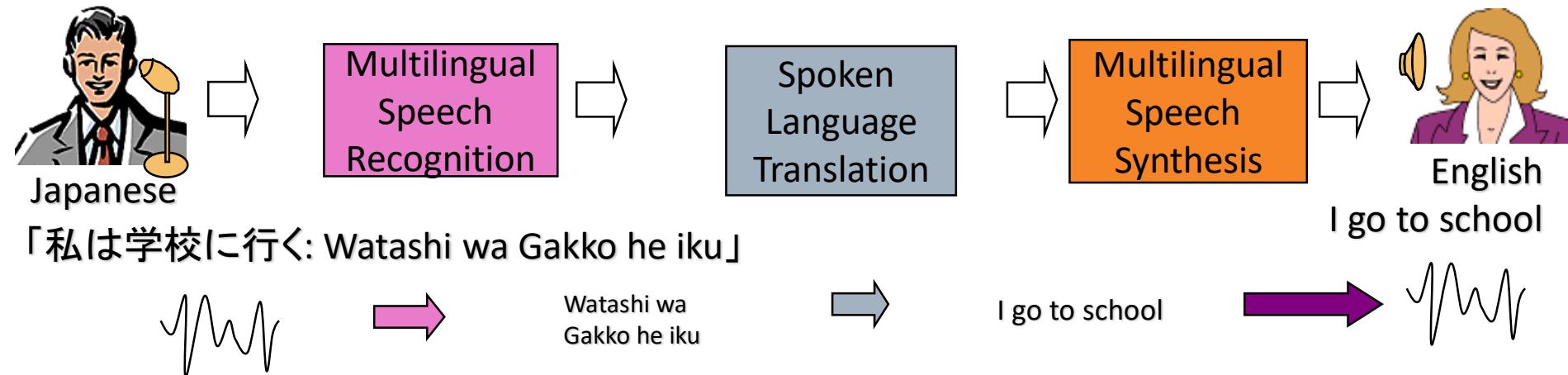


Topics

- ▶ Recent advances in speech processing
 - ASR and TTS research
 - Machine Speech Chain unifies ASR and TTS
 - Application to code switching speech

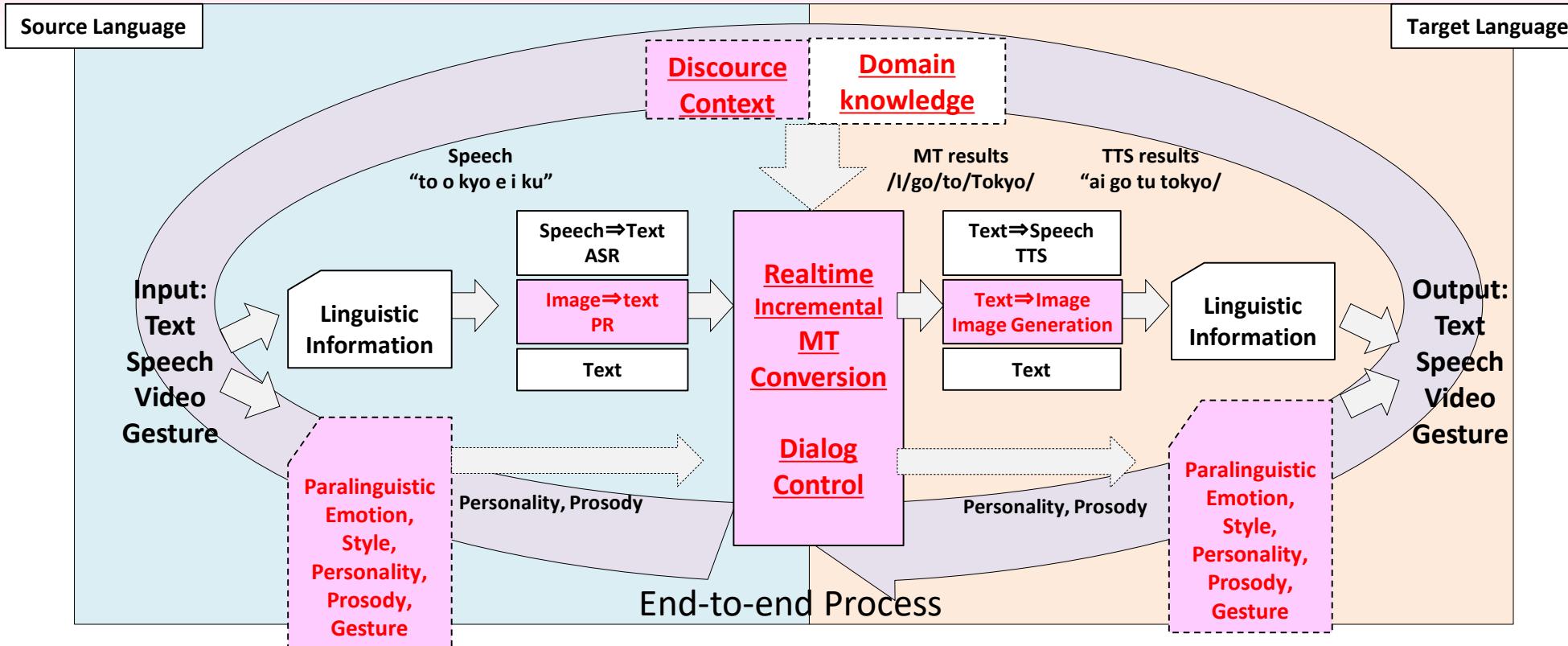
- ▶ Speech Translation
 - Recent Progress on Automatic Speech Translation

Speech Translation



- Cascaded process of speech recognition, machine translation, and speech synthesis.
- Machine translation of ASR transcripts.

Cross-lingual Communication



Communication

- ① Simultaneity, Incremental, Latency,
- ② +Para/non linguistic information

Human Interpreter [A.Mizuno 2016]

E-J Translation Example

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplies (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

(1) 救援担当者は (9) 生くるための (8) 食料を求めて (7) 村を荒らし回っている (6) 大量の難民達の (5) 世話をするための (4) 十分な食料や水, 宿泊施設, 医療品が (3) 無いと (2) 言っています.

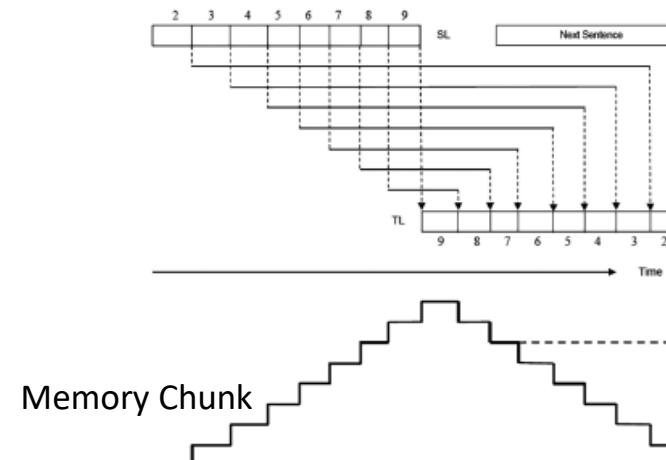
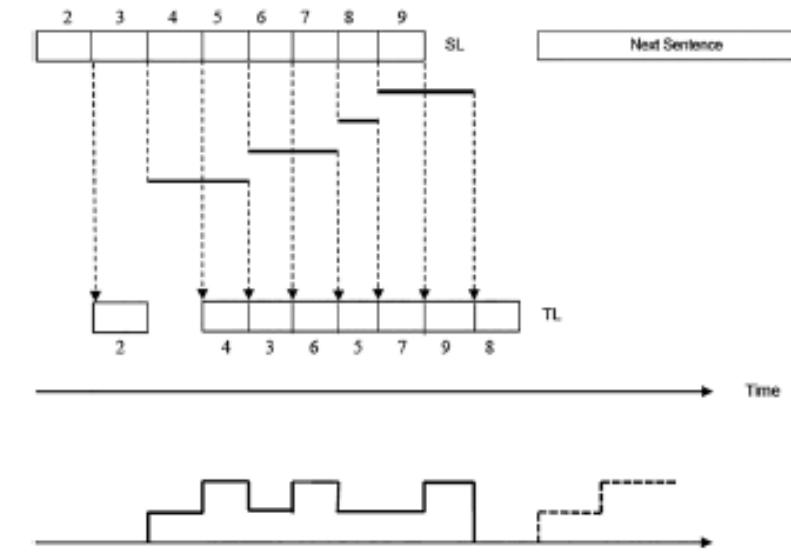


Fig.4 Translation to seek syntactic correspondence and its load
The dotted line of lower right indicates assumed load when next sentence comes in before the completion of translation of previous sentence.

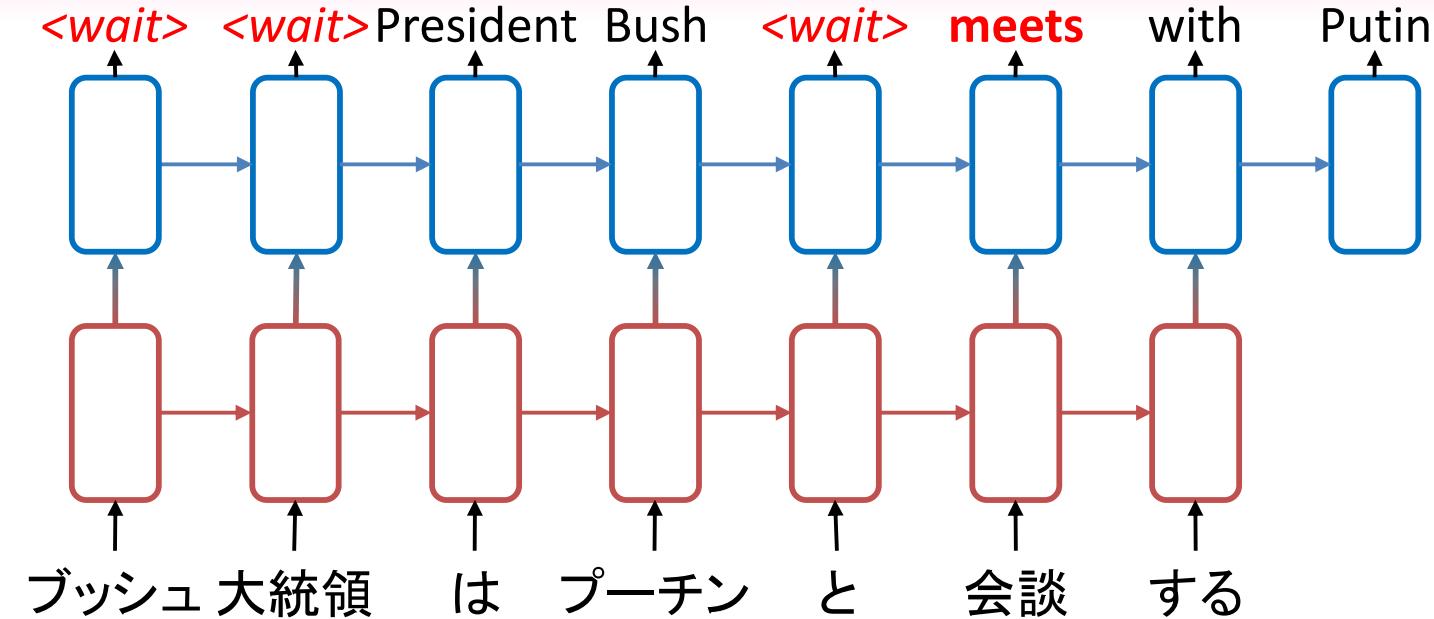
Necessary #Chunk > 3 !

(1) 救援担当者達の (2) 話では (4) 食料, 水, 宿泊施設, 医薬品が, (3) 足りず (6) 大量の難民達の (5) 世話が出来ないとのことです. (7) 難民達は今村々を荒らし回って, (9) 生くるための (8) 食料を求めているのです.



Simultaneous Speech Translation with Adaptive Delay¹³

- ▶ Define a special token <wait>

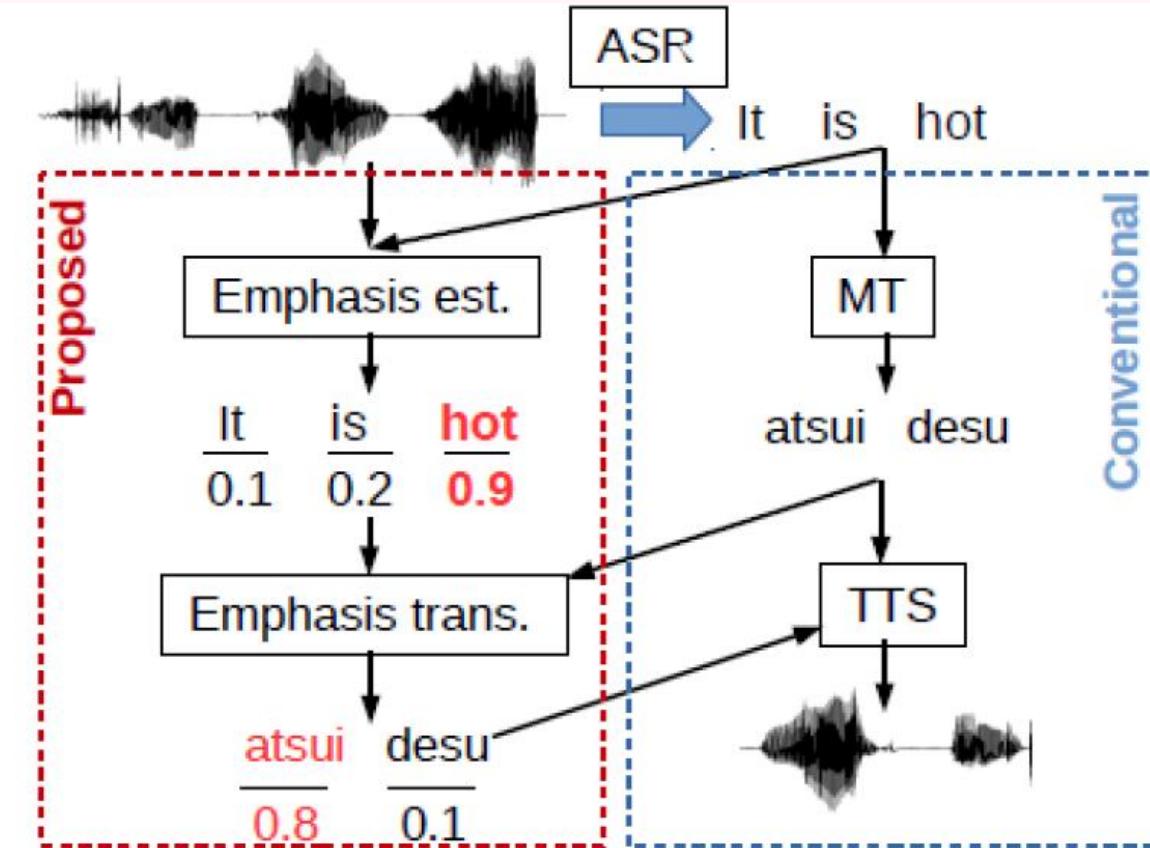


Example (1)

Source	he	did	n	'	t	care	for	swimm	ing	.
Reference	彼	は	水泳	が	得意	で	は	な	かっ	た。
Wait-k (k=3)	w	w	w	彼	は	野球	を	飲	み	ま せ ん で し た。
(gross)	<i>He did not drink baseball.</i>									
Proposed ($\alpha=0.03$)	彼	は	w	w	w	w	w	泳	ぐ	の が 好きではなかった。
(gross)	<i>He did not like swimming.</i>									

Katsuki Chousa, Katsuhiro Sudoh, and Satoshi Nakamura. 2019. Simultaneous Neural Machine Translation using Connectionist Temporal Classification. arXiv preprint , 1911.1193

Paralinguistic Speech Translation



Q. T. Do, S. Sakti, S. Nakamura, "Sequence-to-Sequence Models for Emphasis Speech Translation". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10):1873–1883, 2018

Summary

► Machine Speech Chain

- Semi-supervised training using unpaired data
- Code switching speech, under-resource language, and continuous learning

► Speech Translation

- Simultaneous speech translation
 - Segmentation, anticipation, rewording, evaluation
- Paralinguistic speech translation
 - Emphasis, and emotions

► Future works

- Understanding and interpretation
- Context, situation and multi-modality
- Common sense, knowledge, and cross-cultural knowledge