SPEECH CHAIN FOR SEMI-SUPERVISED LEARNING OF JAPANESE-ENGLISH CODE-SWITCHING ASR AND TTS

Sahoko Nakayama[1], Andros Tjandra[1,2], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]
[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan
{nakayama.sahoko.nq1, andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

1. Background

Code-switching(CS): Bilingual speakers switch languages within a conversation

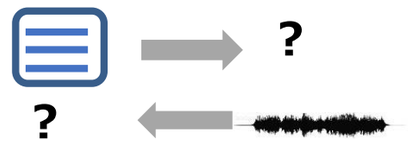
CS challenges: Handle the input in a multilingual setting

Existing approaches:

- ✓ Train either only ASR or only TTS
- ✓ By supervised learning with CS data
  ⇒It requires large amount of parallel CS data

Problems:

Parallel speech & text CS are generally unavailable



2. Proposed Method

Human conversation:

- ✓ Never learn CS at school
- ✓ Learn several monolingual languages, then listen & speak CS in multilingual environments

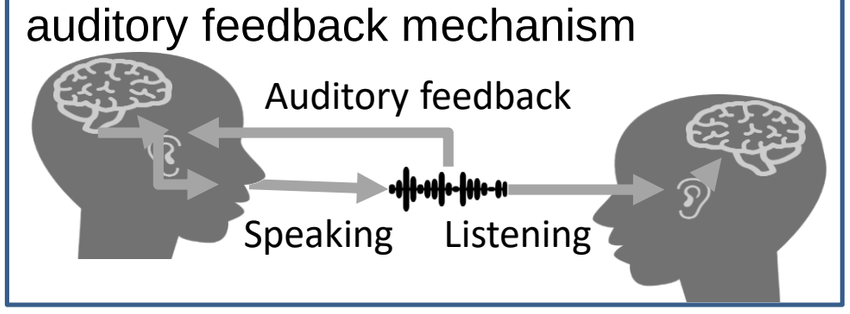
Based on how humans learn CS:

- ✓ **Listening while speaking CS using a speech chain framework**
- ✓ **Enable to perform semi-supervised learning (No need parallel speech & text CS data)**
- ✓ **Aim to improve both ASR and TTS at the same time**

Speech chain [Andros et al., 2017]

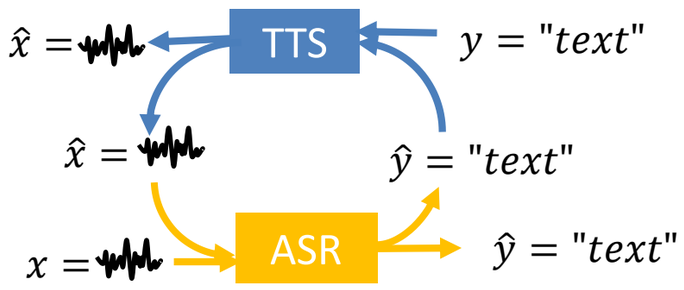
Human speech chain

A closed-loop mechanism has a critical auditory feedback mechanism



Machine speech chain

A closed-loop architecture which allows ASR and TTS to teach each other



Code-switching with Speech Chain

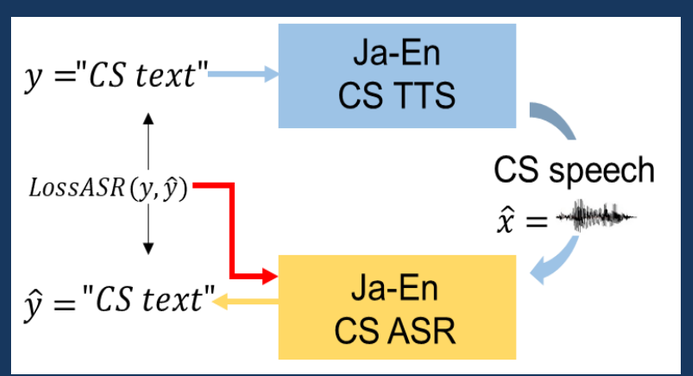
Step1. Supervised learning

Separately train ASR & TTS with parallel speech-text monolingual data

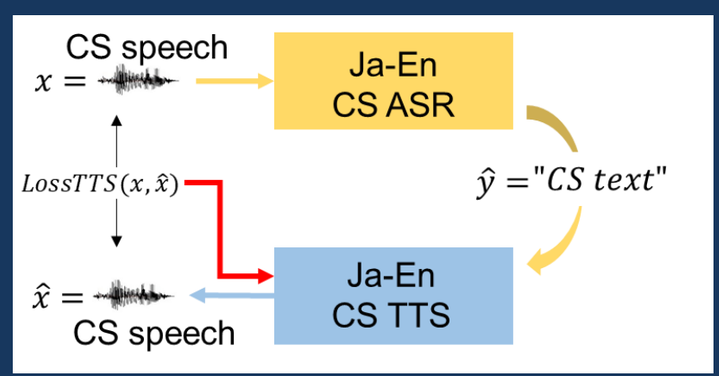
Step2. Unsupervised learning

Perform a speech chain with only CS text or CS speech

➢ **Input CS text into TTS⇒ASR** Calculate the loss $L_{ASR}^{CS}(\hat{y}^{CS}, y^{CS})$



➢ **Input CS speech into ASR⇒TTS** Calculate the loss $L_{TTS}^{CS}(\hat{x}^{CS}, x^{CS})$



➢ **Training objective**:

$$L = \alpha * \left(L_{ASR}^{Mono} + L_{TTS}^{Mono}\right) + \beta * \left(L_{ASR}^{CS} + L_{TTS}^{CS}\right)$$

Possible to train the new matters without forgetting the old one

3. Experimental Results

Set-up:

- ✓ **Model:** Attention-based encoder-decoder ASR, Tacotoron TTS
- ✓ **Data:** En/Ja monolingual BTEC text data
  En-Ja CS text created from BTEC
  } Speech is synthesized by GoogleTTS

ASR & TTS Performances

(in CER & L2-norm squared in log-Mel spectrogram)

Model	Japanese test(JaTTS)		CS test(MixTTS)		English test(EnTTS)	
	ASR	TTS	ASR	TTS	ASR	TTS
Baseline: paired speech-text ⇒ Supervised training						
Ja50k(JaTTS)	2.11%	0.321	33.73%	0.484	81.12%	0.667
En50k(EnTTS)	86.42%	0.373	66.16%	0.469	2.30%	0.417
Ja25k + En25k(MixTTS)	1.71%	0.312	**18.11%**	**0.489**	2.99%	0.437
Speech chain: [paired Ja25k+En25k (MixTTS)] + [unpaired CS (Mix+JaTTS)] ⇒ Semi-supervised training						
+CS20k(Ja+MixTTS)	1.82%	0.305	**5.08%**	**0.372**	4.05%	0.439

Note: MixTTS means using both JaTTS and EnTTS

4. Conclusion

Proposed CS Model based on speech chain:

- ✓ **Allows CS ASR & CS TTS to learn from each other**
- ✓ **Even without any parallel speech & text CS data**

Experimental results reveal:

- ✓ **Maintaining performance in the monolingual setting**
- ✓ **Improved ASR in CS test from CER 18.11% to 5.08%**
- ✓ **Also improved TTS from L2-norm 0.489 to 0.372**

Future Work:

- ✓ **Use natural speech data**
- ✓ **Apply to other languages**