ZERO-SHOT CODE-SWITCHING ASR AND TTS WITH MULTILINGUAL MACHINE SPEECH CHAIN

Sahoko Nakayama[1,2], Andros Tjandra[1], 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. Introduction

Code-switching(CS):
switching languages within a conversation

CS challenges for ASR & TTS:
need to handle the multilingual input

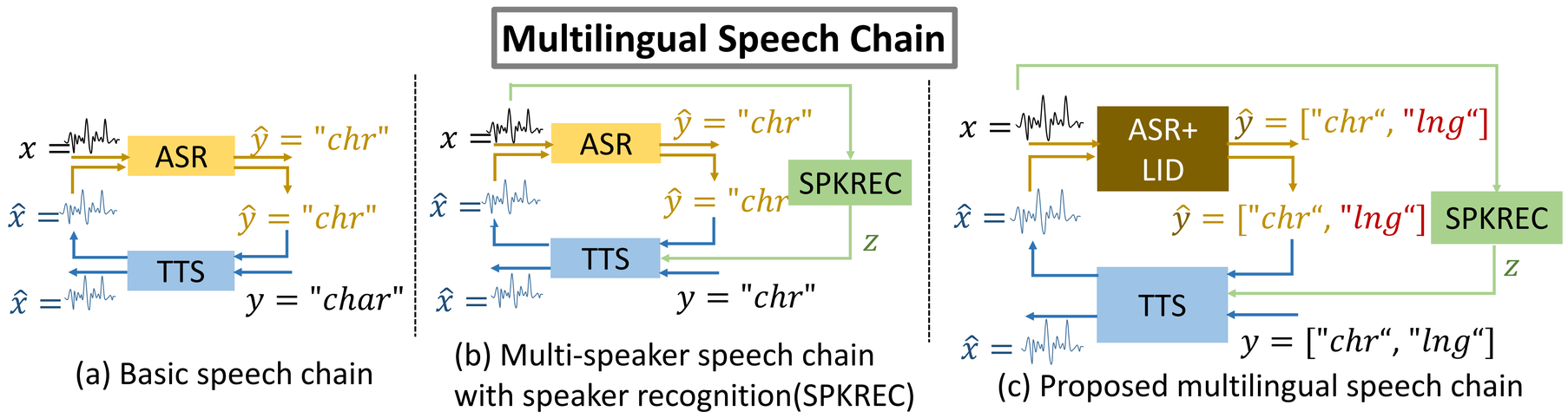
Existing approaches
- Just developed on ASR or TTS
- Only focused on a single language pair
- Trained in a supervised fashion

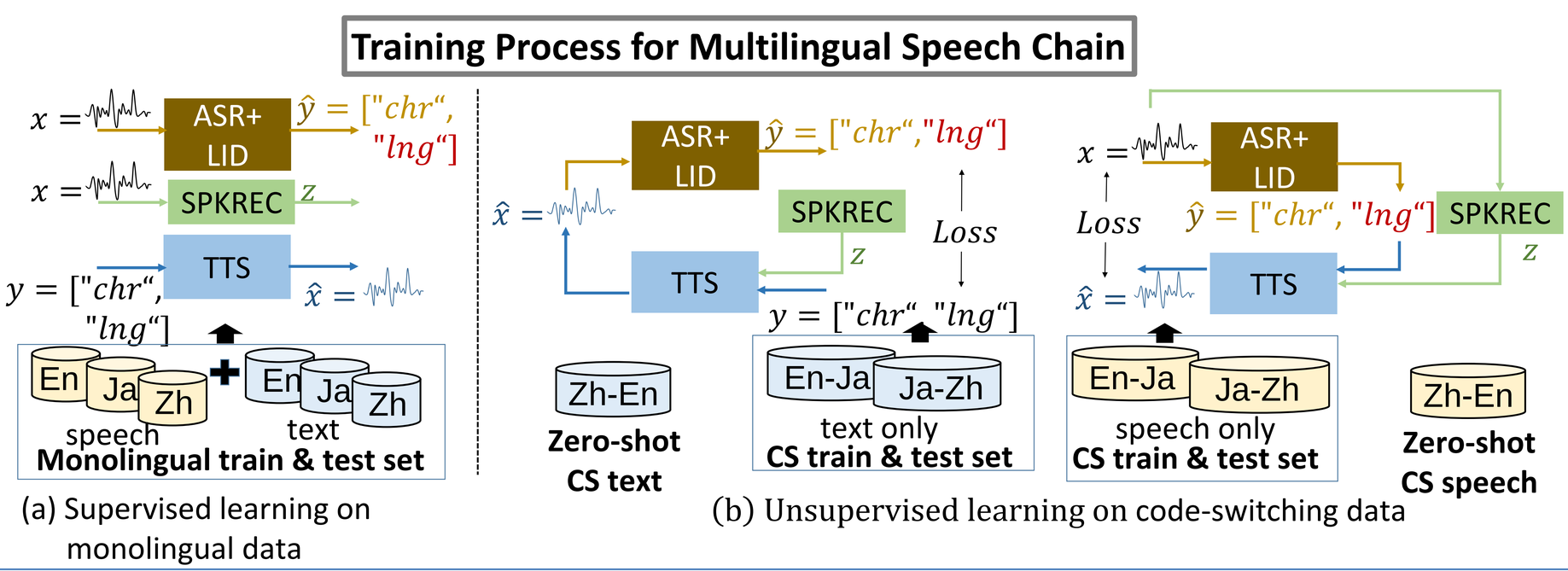
Goal
- CS tasks of ASR and TTS on multiple language pairs
- Trained semi-supervised learning
- **Zero-shot CS**: performing the unknown CS w/o directly learning it

2. Proposed Method

Proposed Machine Speech Chain [Tjandra et al., 2017,2018,2019]:
a closed-loop architecture that enables ASR & TTS to assist each other

Method: embedding **Language Identification Discrimination (LID)** to **machine speech chain**





3. ASR Evaluation

I. Proposed model on zero-shot CS (Known language)

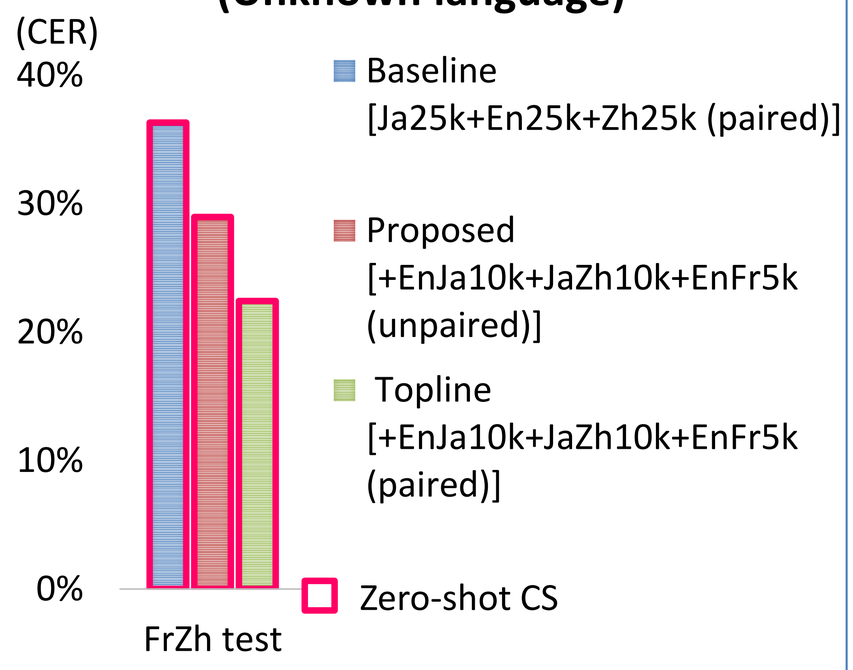
Model	Monolingual (CER)			Code-switching (CER)		
	Ja	En	Zh	EnJaCS	JaZhCS	ZhEnCS
Baseline: Supervised training on monolingual data only						
Ja25k+En25k+Zh25k (paired)	8.85%	8.48%	5.11%	14.06%	16.91%	16.04%
Proposed Machine Speech chain: Semi-supervised training on two CS data						
+ EnJaCS10k+JaZhCS10k (unpaired)	9.18%	12.71%	5.93%	11.56%	8.31%	10.52%
+EnJaCS10k+ZhEnCS10k (unpaired)	8.93%	12.34%	5.67%	11.18%	9.21%	9.71%
+ZhEnCS10k+JaZhCS10k (unpaired)	8.91%	14.45%	6.08%	11.85%	10.40%	11.29%
Topline: Supervised training on CS data						
+EnJaCS10k+JaZhCS10k(paired)	10.18%	12.32%	7.93%	8.94%	6.70%	8.09%
+EnJaCS10k+ZhEnCS10k(paired)	11.04%	10.91%	7.48%	10.81%	7.26%	8.07%
+ZhEnCS10k+JaZhCS10k (paired)	10.98%	11.57%	7.22%	10.34%	7.72%	7.98%
+EnJaCS10k+JaZhCS10k+ZhEnCS10k	10.48%	10.43%	6.88%	8.68%	6.98%	8.05%

Language used as pair | Language used as unpair | Zero-shot CS

Model:
Encoder-decoder attention ASR
Data: Monolingual BTEC
CS created from BTEC

- Improved ASR in the multilingual CS test
- Performed well on an unknown CS test

II. Proposed model on zero-shot CS (Unknown language)



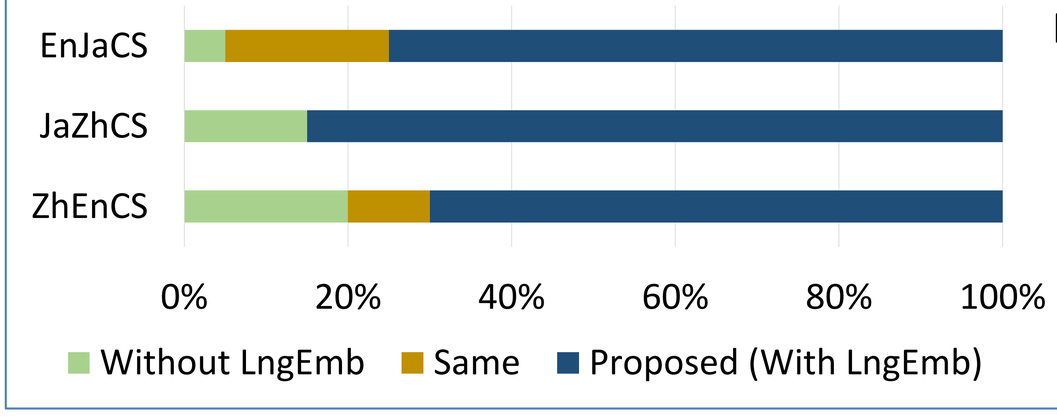
- Also improved on zero-shot CS with unknown Fr language

III. Proposed model on zero-shot CS (CS natural speech)

- Improved even in the case of using CS natural speech (see paper for more details)

4. TTS Evaluation

AB preference subjective evaluation



Setup
Model: Tacotron TTS,
DeepSpeaker SPKREC

- Maintained TTS quality better
- Especially on the switch position between two languages

5. Conclusion

- Introduced a zero-shot CS ASR &TTS
- Proposed multilingual machine speech chain with LID
- Improved the performance of the multilingual CS
- Also performed well on the unknown CS without directly learning it