Improving Khmer Automatic Speech Recognition by Joint Training with Speaker Recognition and Translation Task

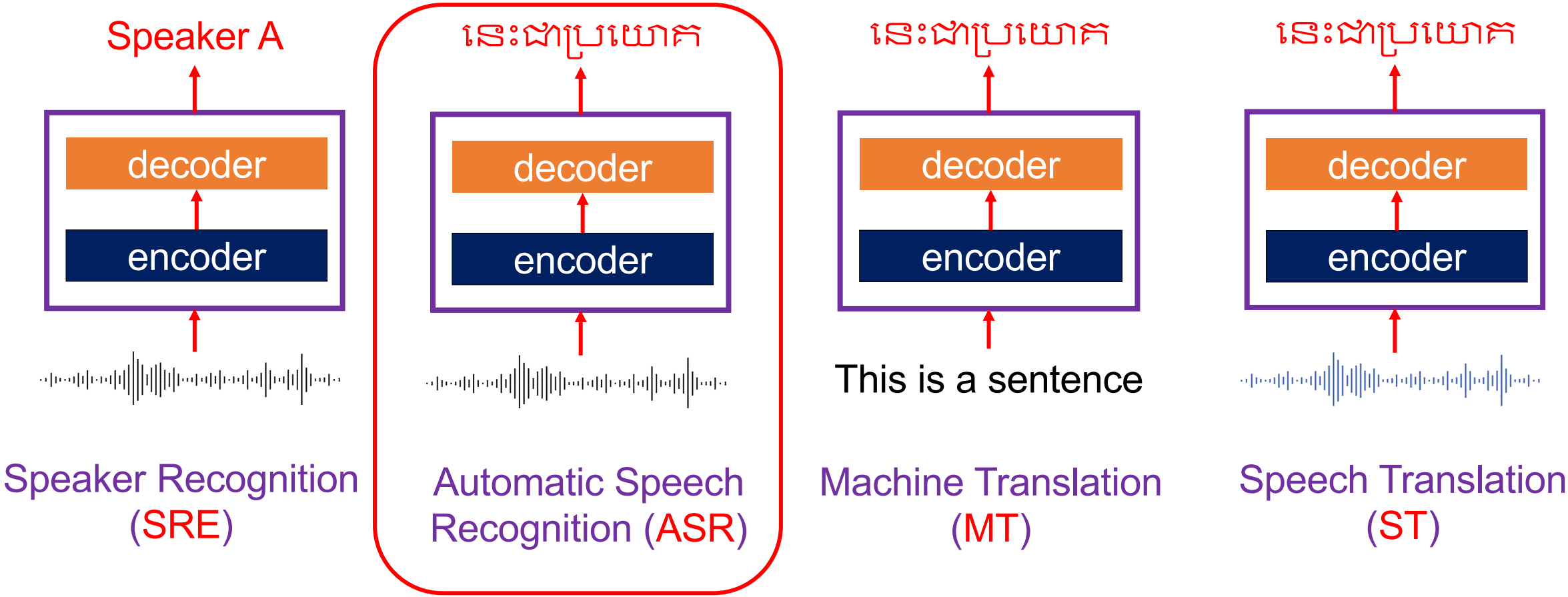
Kak Soky, D3,
Speech and Audio Processing Laboratory,
Kyoto University
@ Online seminar, 29-January-2022

Contents

Introduction

- Automatic speech recognition (ASR) is well performed in resource-rich languages, especially with the end-to-end model architecture.
- However, the ASR performance of Khmer is not yet satisfying:
  - Lack of resources (speech and text).
  - Lack of language processing tools.
  - Lack of a good benchmark ASR system.
- Therefore, corpus creating is conducted, and then propose the methods to improve ASR for Khmer in this work.

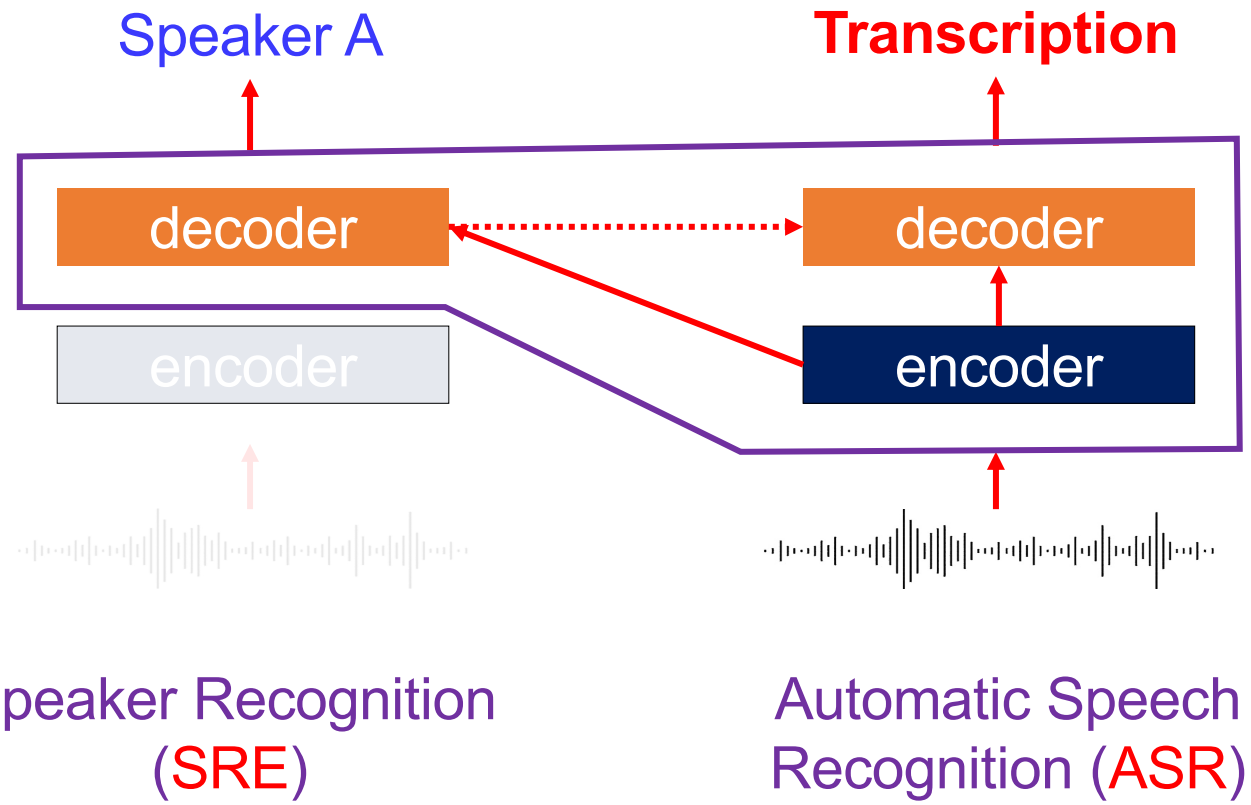
Overview of end-to-end systems



Poor performance in low-resource language

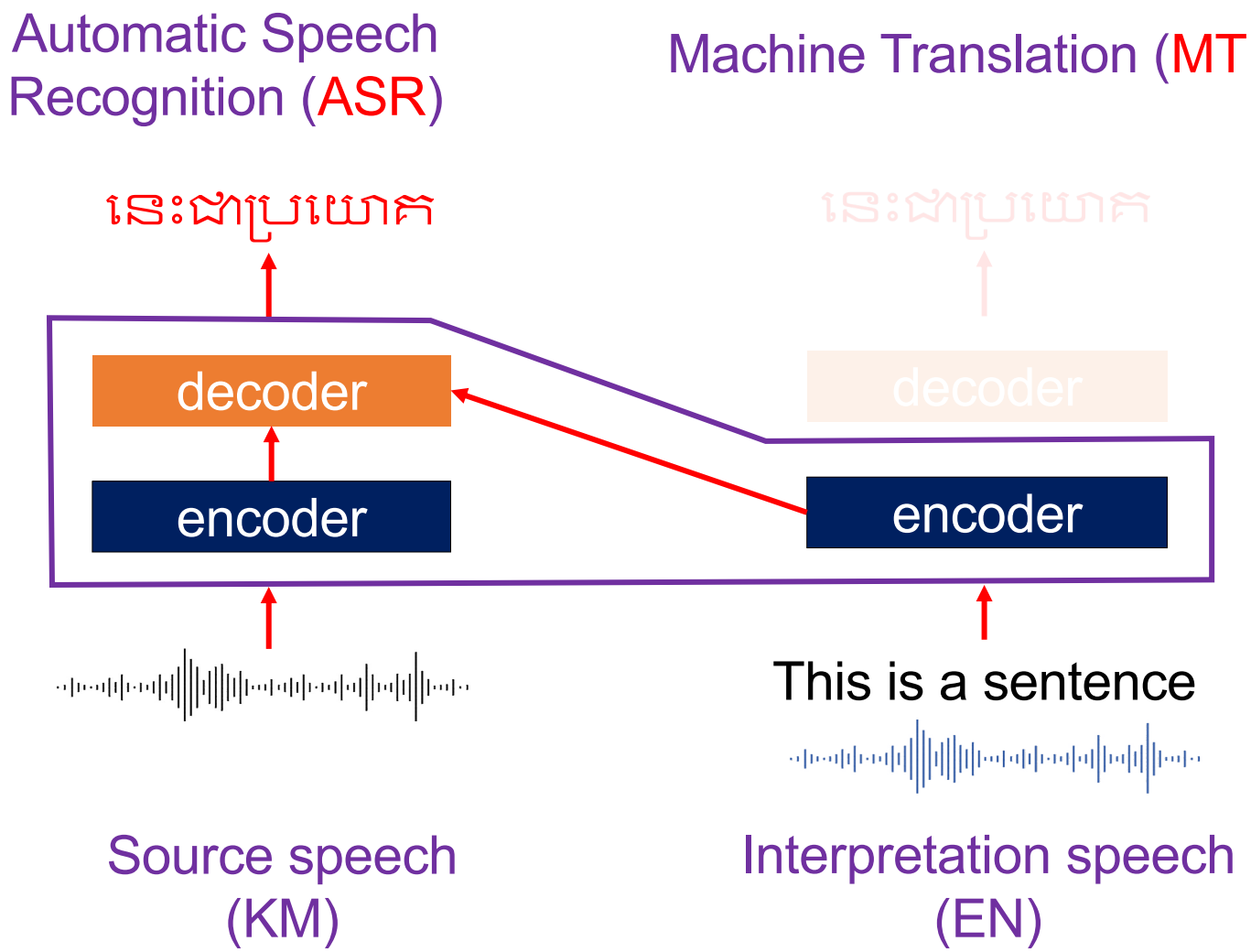
3.1 Proposed joint ASR and SRE

- It is easy to recognize the speech if we can identify the speaker.
- Thus, we propose **a joint ASR and SRE:** using speaker information to leverage ASR task.



Propose a joint ASR and MT/ST end-to-end systems

- In multi-lingual events such as international meetings and court proceedings, simultaneous translation/interpretation is often available.
- Thus, we propose **a joint ASR and MT/ST:** using translation knowledge to leverage ASR task.



Corpus Creation of the Extraordinary Chambers in the Courts of Cambodia (ECCC)

Soky et al., Khmer Speech Translation Corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC), In Proc. O-COCOSDA 2021.

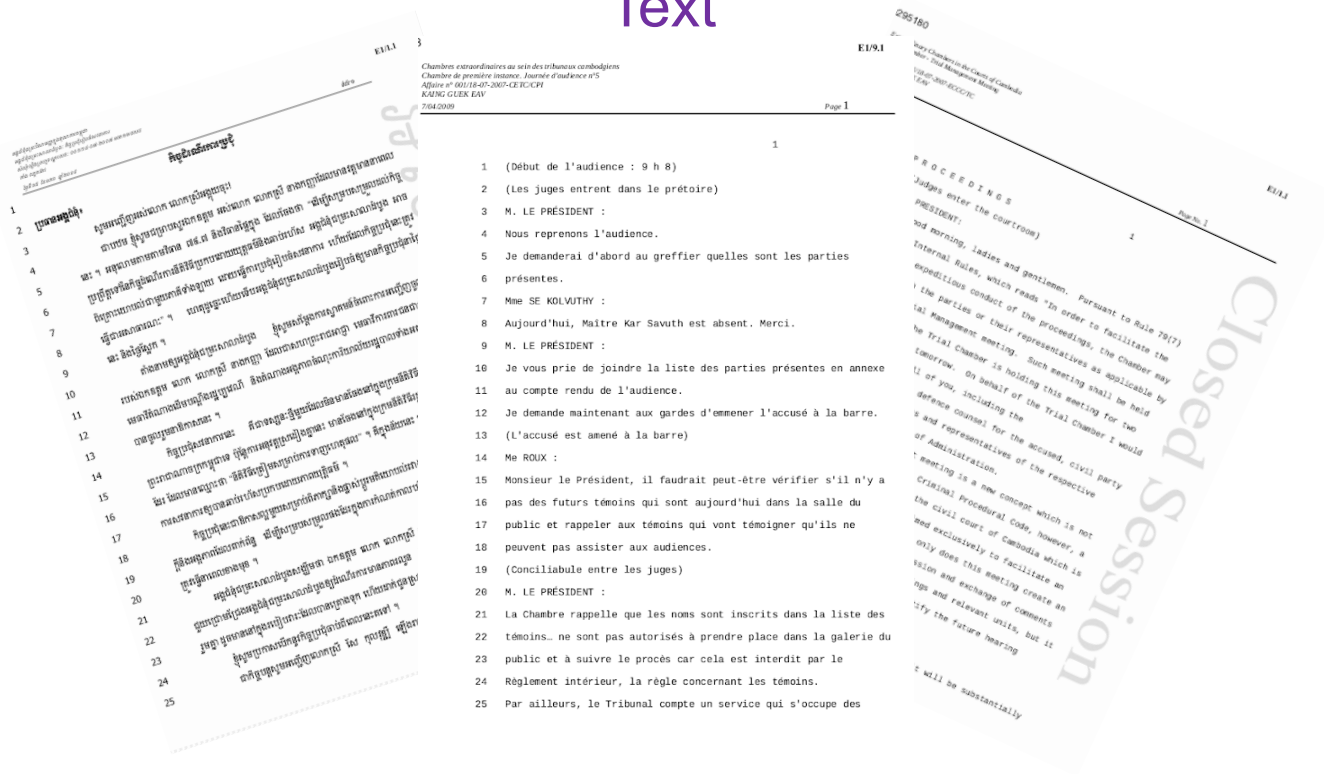
ECCC dataset

- The Extraordinary Chambers in the Courts of Cambodia (ECCC) is a court established to prosecute the senior leaders who committed crimes during the Khmer Rouge regime in Cambodia from 1975 - 1979.
- We collected 222 sessions of audios and text of the first trial which conducted from February 17 to November 27, 2009.
- The trial simultaneously conducted in Khmer, English and French.

ECCC dataset

Video

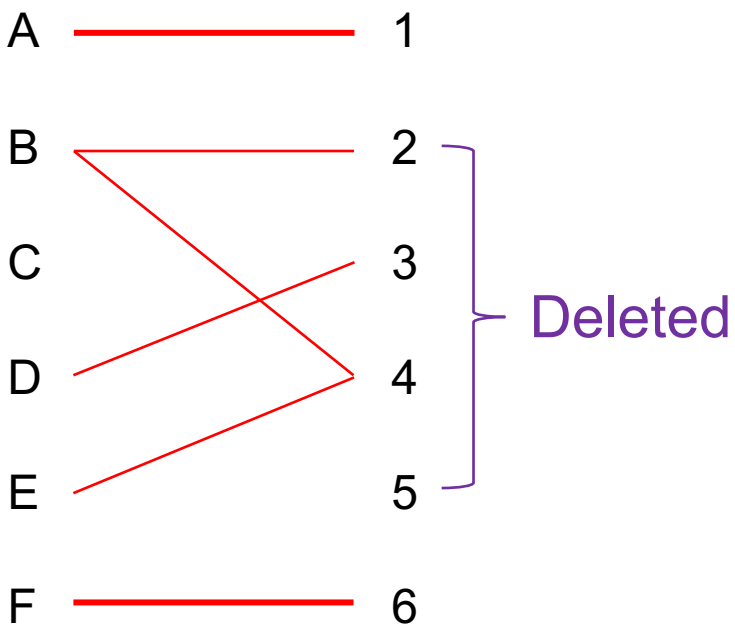
Text

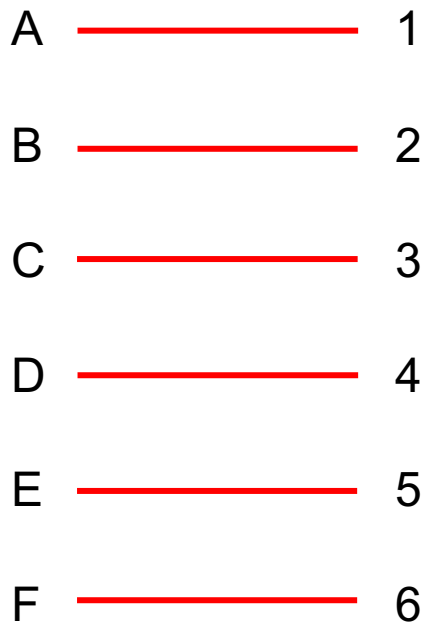


(Début de l'audience : 9 h 8)
(Les juges entrent dans le prétoire)
M. LE PRÉSIDENT :
Nous reprenons l'audience.
Je demanderai d'abord au greffier quelles sont les parties présentes.
Mme SE KOLVUTHY :
Aujourd'hui, Maître Kar Savuth est absent. Merci.
M. LE PRÉSIDENT :
Je vous prie de joindre la liste des parties présentes en annexe au compte rendu de l'audience.
Je demande maintenant aux gardes d'emmener l'accusé à la barre.
(L'accusé est amené à la barre)
Me ROUX :
Monsieur le Président, il faudrait peut-être vérifier s'il n'y a pas des futurs témoins qui sont aujourd'hui dans la salle du public et rappeler aux témoins qui vont témoigner qu'ils ne peuvent pas assister aux audiences.
(Conciliabule entre les juges)
M. LE PRÉSIDENT :
La Chambre rappelle que les noms sont inscrits dans la liste des témoins… ne sont pas autorisés à prendre place dans la galerie du public et à suivre le procès car cela est interdit par le Règlement intérieur, la règle concernant les témoins.
Par ailleurs, le Tribunal compte un service qui s'occupe des

- However, two main problems in this dataset:
  1. Written without word or sentence boudary in Khmer text.
  2. No time aligment between speech and text.

2. Corpus creation: alignment





Traditional method [Braune et. al, 2010; Dyer et. al, 2013; Thompson et. al, 2019]

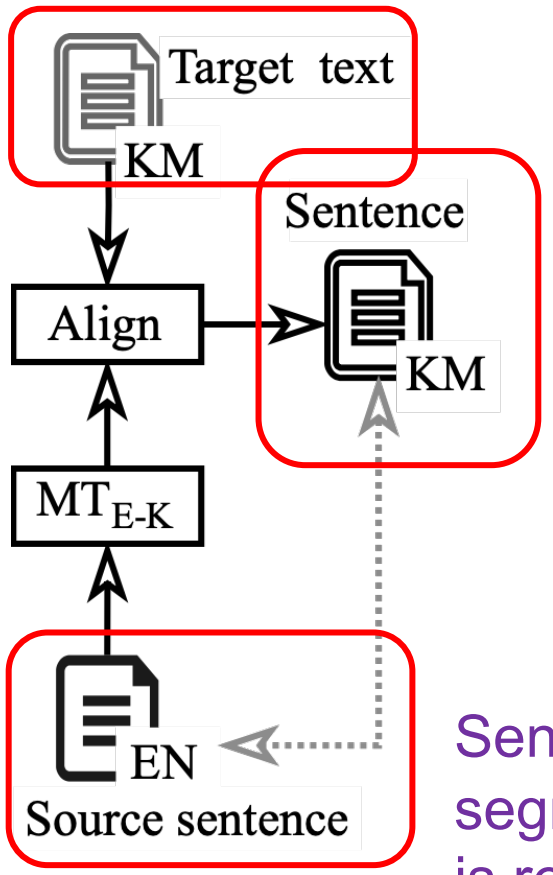
- Source and target languages are in sentence-based.

Proposed a monotonic alignment:

- Suitable for simultaneous translation.
- Only source language requires the sentence-based.

2. Corpus creation

1. Generating monotonic sentences alignment (English to Khmer)



Word tokenization is required.

Sentence boundaries are added.

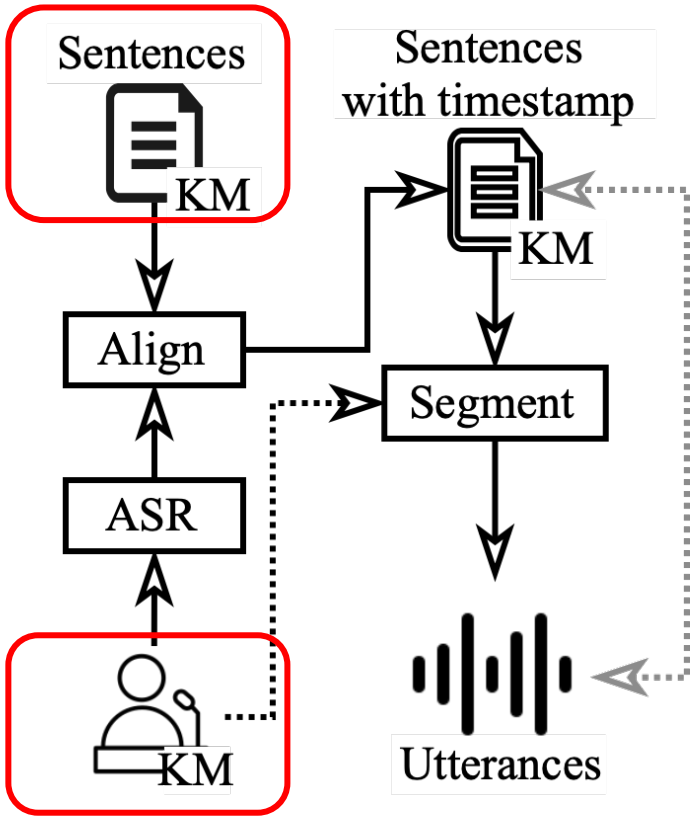
Sentence segmentation is required.

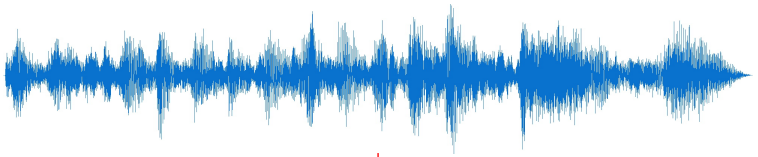
បន្ទាប់ ពី ការ សំរេច របស់ អង្គ ជំនុំជម្រះ តុលាការ កំពូល នៅ ថ្ងៃ ទី ៣ ខែ កុម្ភៈ ឆ្នាំ ២០១២ ដែល បាន បញ្ជាក់ និង ធ្វើ វិសោធន កម្ម ផ្នែក ខ្លះ នៃ សាល ដីកា របស់ អង្គ ជំនុំជម្រះ សាលា ដំបូង និង បាន លុប ចោល សេចក្តី សម្រេច លើ ការ កាត់ ទោស កាំង ហ្គេកអ៊ាវ ត្រូវ បាន រក ឃើញ ថា មាន ទោស យោង តាម មាត្រា ៥ ទី ៦ និង ២៩ នៃច្បាប់ អ.វ.ត.ក. នៃ ឧក្រិដ្ឋកម្ម ដូច ត ទៅ នេះ ដែល ប្រព្រឹត្ត នៅ រាជធានី ភ្នំពេញ និង នៅ លើ ទឹក ដី នៃ ប្រទេស កម្ពុជា ចន្លោះ ថ្ងៃ ទី ១៧ ខែ មេសា ឆ្នាំ ១៩៧៥ ដល់ ថ្ងៃ ទី ៦ ខែ មករា ឆ្នាំ ១៩៧៩។

1. Following the decision of the Supreme Court Chamber on 3 February 2012,
2. which partially confirmed and amended the Trial Chamber Judgement as well as overturning the decision on sentencing,
3. Kaing Guek Eav has been found guilty pursuant to Articles 5, 6 and 29 of the ECCC Law of the following crimes committed in Phnom Penh and within the territory of Cambodia between 17 April 1975 and 6 January 1979.

2. Corpus creation

2. Text-to-speech alignment





```
1560 1680 they
1710 1920 aren't
1920 2100 think
2100 2490 archer
3570 3930 the
4110 4770 public
5010 5340 please
5340 5880 stand
```

1. Please stand up
2. Please be seated
3. Today in the name of Cambodian people and the united nations.

Align

E1_3R_S1.txt

```
0 13320 Let the public please stand
13320 16440 Please invite the public to be seated
16440 26520 Today In the name of the Cambodian people and the United Nations
26520 35130 and pursuant to the Law on the Establishment of Extraordinary Chambers in the Courts of Cambodia
35130 47250 for the Prosecution of Crimes Committed during the Period of Democratic Kampuchea the Trial Chamber of the Extraordinary
47250 47790 promulgated by Royal Kram one thousand four slash zero zero six
47790 56250 dated twenty seven October two thousand fourth today declares open the Initial Hearing
56250 63210 on Case File zero zero one relating to the accused man Kaing Guek Eav alias Duch
63210 72810 aged sixty six who has been charged with crimes against humanity grave breaches of the Geneva Conventions
72810 79800 of twelve August one thousand nine hundred forty ninth and violation of the one thousand nine hundred fifty six
79800 84720 Cambodian Penal Code The bench is composed of judges I myself Nil Nonn
```

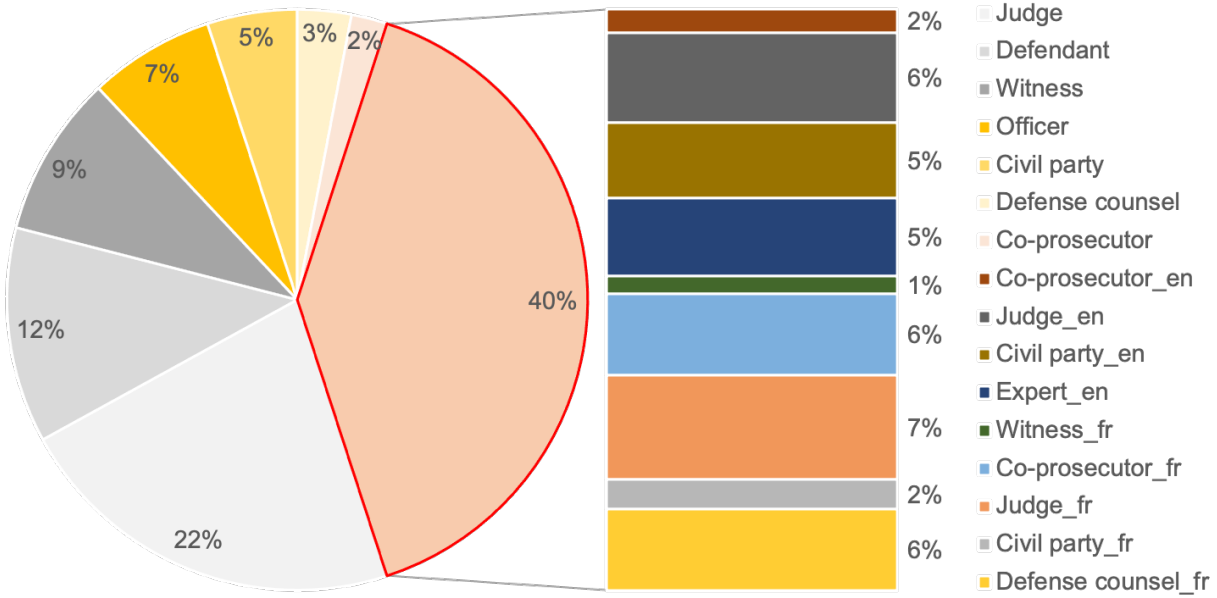
Please stand up

Corpus size and speaker distribution

- Overall statistics of ECCC corpus

Source	Vocabs.	Avg. src words	Avg. length (s)	Target	Utterances	Hours	Target words	Avg. tgt words	Tgt. Vocabs.
KM	9K	25	8.5	EN	65.4K	155	1.24M	19	15K
				FR	64.2K	152	1.33M	21	21K

- 60% of speech is the original speech of Khmer speakers.
- Another 40% is translated from English and French of the judges, lawyers, experts, and officers.



Experiments setup for ASR system

- ECCC:
  - Encoder: 6, Decoder: 6, FFN units: 2048, attention head: 4, attention-dim: 256
  - 45 epochs
  - 5000 BPE tokens per language
  - Train in a single GPU using 64-batch-size

Experiment results of ASR on ECCC corpus

Model	WER (%)		
	Khmer	English	French
w/o augmentation	23.6	6.9	14.5
w/ speed perturbation (SP)	22.2	6.6	14.0
w/ SpecAugment (SA)	21.8	6.4	13.8
w/ SP + SA	21.4	6.0	12.6

- Khmer ASR performance is far bahind English or French.
- The Khmer ASR performance will be address in this talk.

Experiments setup for Translation Models

- Architecture: Transformer
- Encoder: 6, Decoder: 6, FFN units: 2048,
- Attention head: 4, attention-dim: 256
- 100 epochs
- 5000 byte pair encoding (BPE) tokens per language
- Train in a single GPU using 96-batch-size

Baseline results of translation models in ECCC corpus

Source	Target	BLEU		
		MT	Cascade-ST	End-to-end ST
Khmer (KM)	English (EN)	16.63	15.14	13.81
EN	KM	14.44	14.15	14.14
FR	KM	10.54	9.82	10.26
KM	FR	11.53	10.66	9.39

- Machine translation (MT): Text-to-Text translation.
- Speech translation (ST): Speech-to-Text translation.

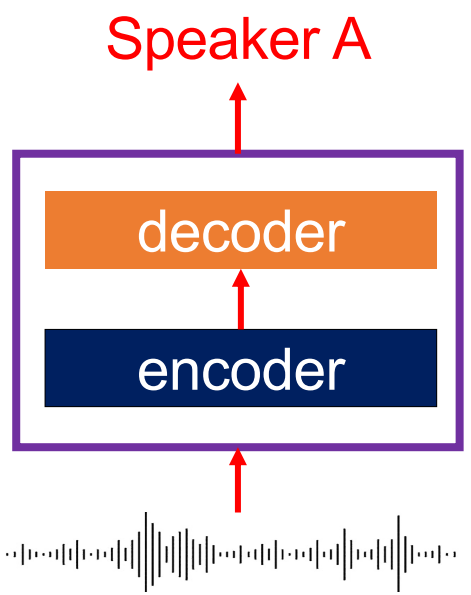
- The translation quality of EN->KM and KM->EN were better because EN to KM was directly aligned in bilingual sentence alignment.
- Whereas KM-FR was indirectly aligned.

A Joint Training Automatic Speech and Speaker Recognition

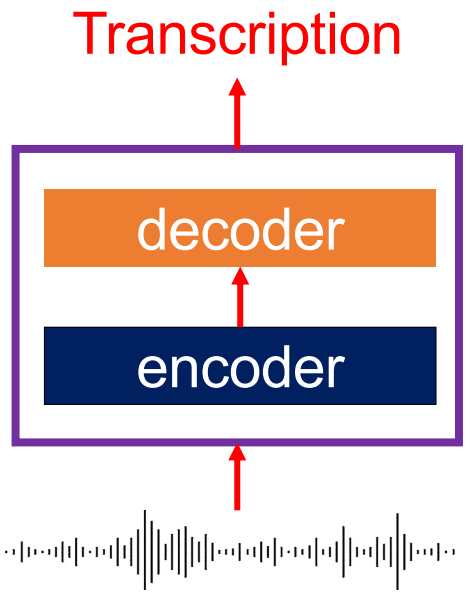
Soky et al., On the Use of Speaker Information for Automatic Speech Recognition in Speaker-imbalanced Corpora, In Proc. APSIPA ASC 2021.

1. Introduction

- In speech, we can:
  - generate meta information (Speaker information) using SRE,
  - And decipher speech content (transcription) using ASR.



Speaker Recognition (SRE)

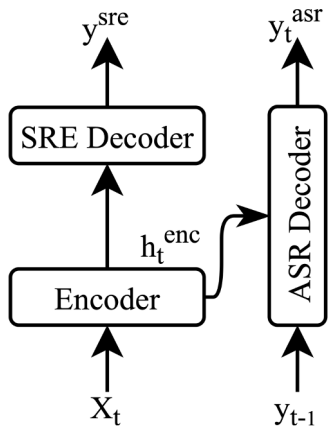


Automatic Speech Recognition (ASR)

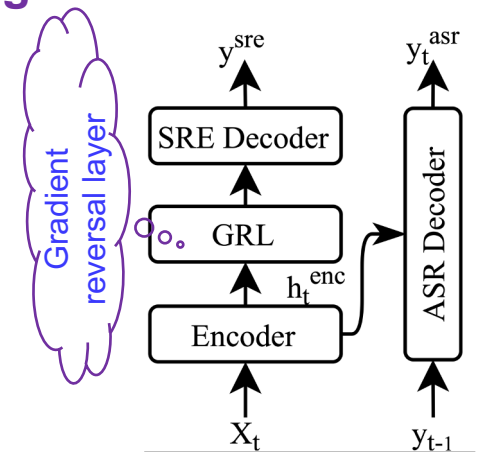
They can perform together and simultaneously

2. Conventional joint ASR and SRE (1/2)

- ASR and SRE presented in:
  - MTL [Tang et al., 2016]: a unification of transcribing the speech and identifying the speakers simultaneously by sharing the same speech feature extraction layers.
  - AL [Ganin et al., 2015]: a similar to MTL, but learns a speaker-invariant model which reduces the effects of speaker variability.
  - However, they do not use speaker information explicitly for ASR.



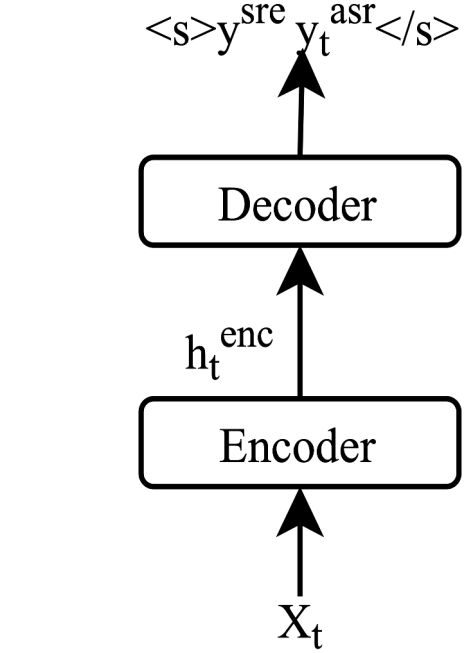
Multitask learning (MTL)



Adversarial Learning (AL)

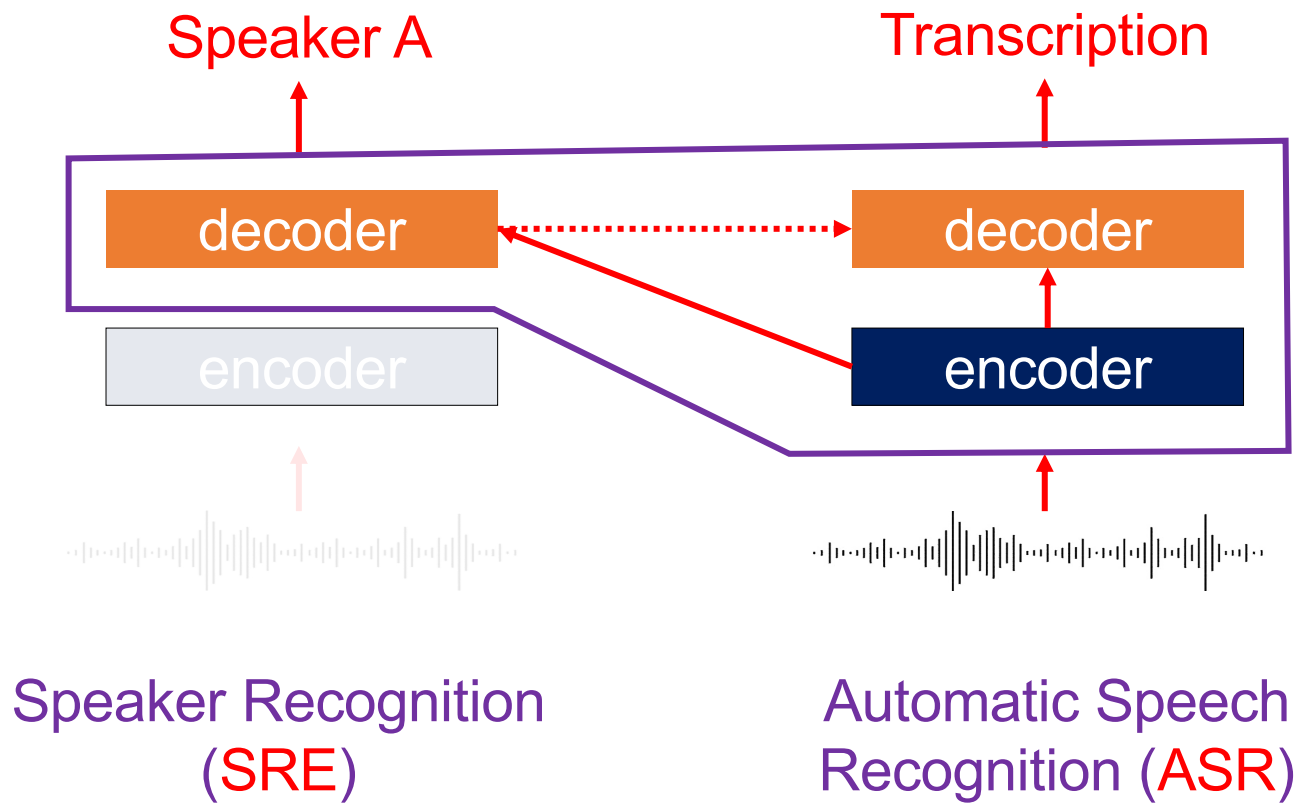
2. Conventional joint ASR and SRE (2/2)

- SAug [Li et al., 2019]: a fully E2E system integrating SRE and ASR, but uses only a single decoder to perform both ASR and SRE.
- Speaker embedding [Delcroix et al., 2018]: a use of speaker information to improve ASR, but it doesn't conduct the SRE.



Speech Attribute Augmentation (SAug)

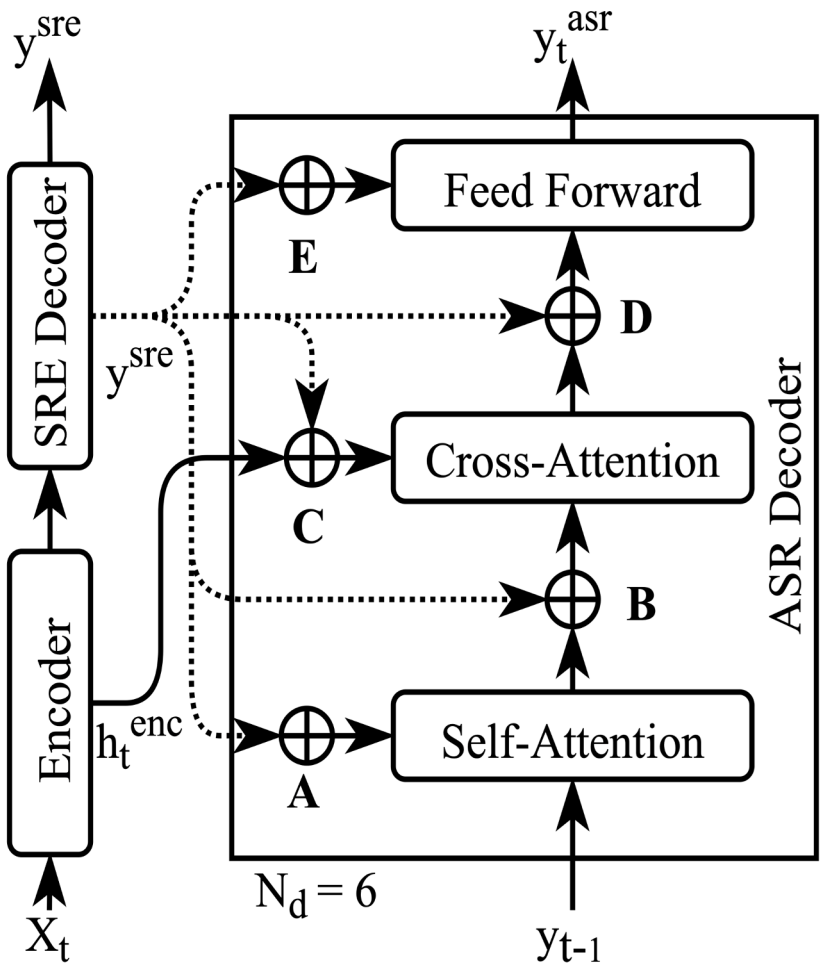
3.1 Proposed joint ASR and SRE



A joint ASR and SRE: using speaker information to leverage ASR task.

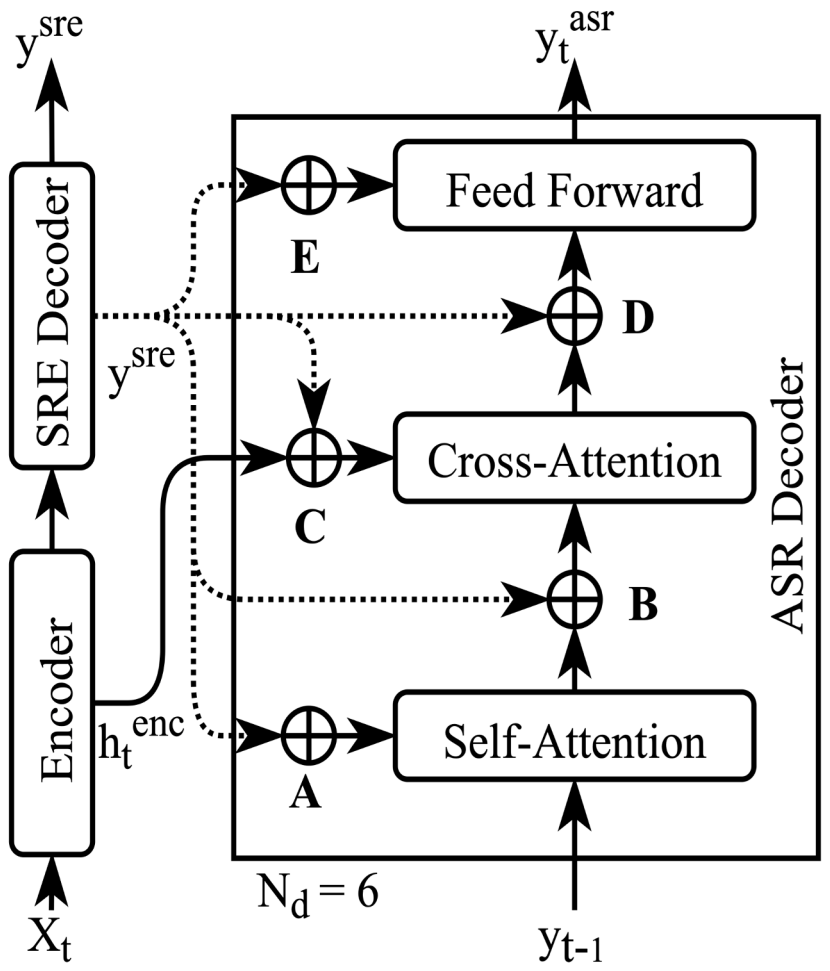
3.2 Model architecture (1/2)

- Given an acoustic features $X_t = \{x_1,\ldots,X_n\}$, the model generates:
  - A speaker ID, y^{sre}
  - And vocabulary tokens $y_t^{sre} = \{y_1,\ldots,y_m\}$
- Unlike the previous speaker embedding:
  - The proposed method is a direct use of the speaker IDs as a speaker embedding in the ASR decoder of end-to-end network.



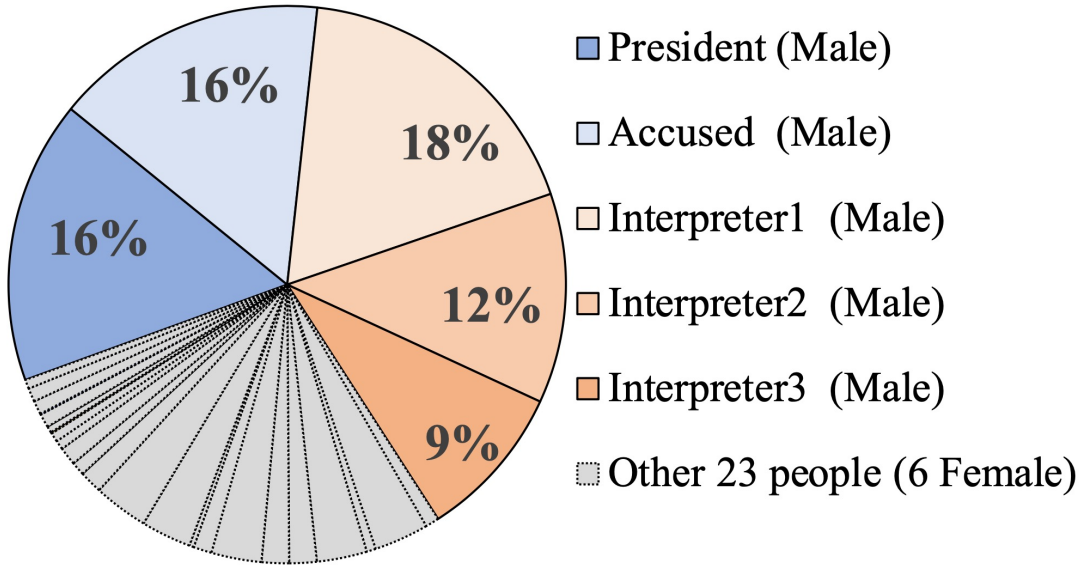
3.2 Model architecture (2/2)

- The speaker output (y^{sre}) injected into:
  - A: self-attention
  - B: after self-attention
  - C: cross-attention
  - D: after cross-attention
  - E: after feed-forward
  - AC: combination of A and C
  - BD: combination of B and D
  - And layer-wise of the ASR decoder



4.2 Corpus size and data setup

- There are 28-speaker (Gr28) in total:
  - 70% of speech is from the five major speakers.
  - Another 30% is from other 23 people.
- Thus, we propose to combine the other 23 speakers into one speaker, therefore only 6-speaker (Gr6).



Dataset	# utterances	# hours	# characters
Training	75,170	176	6.02M
Test	3,733	10	294K
Total	78,903	186	6.30M

4.3 Result of ASR and SRE systems

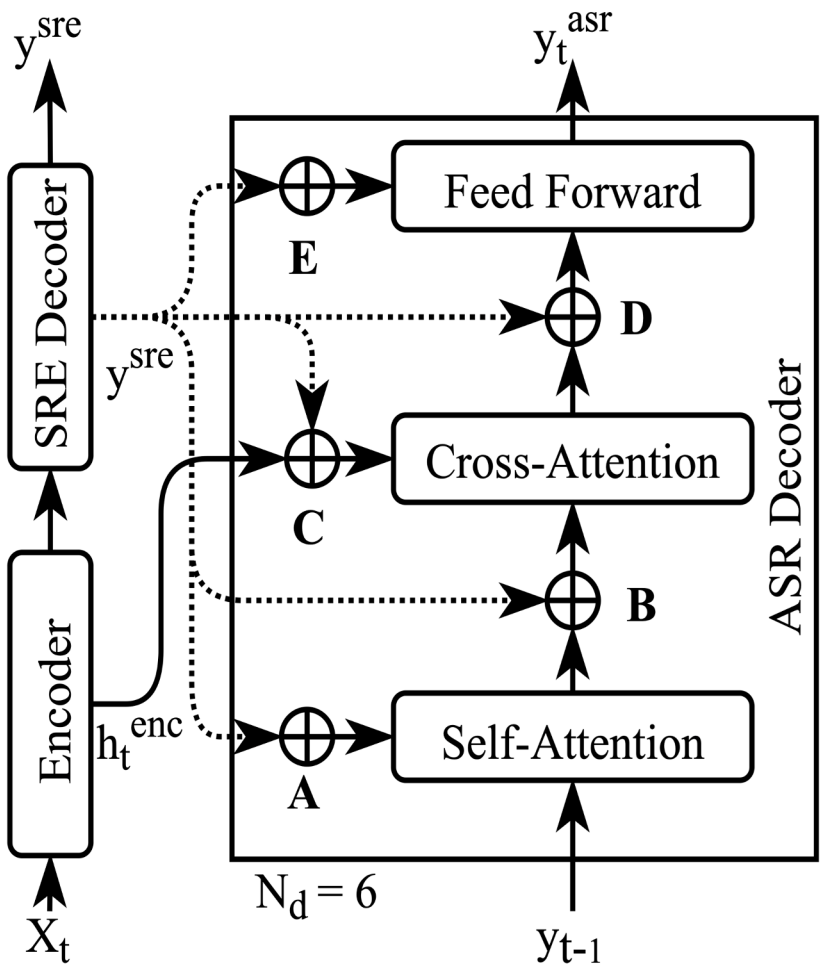
System	SRE (%incorrect)	ASR (%CER)
Baseline (Gr6):		
- X-vector	9.72	-
- Transformer	-	7.46
Joint SRE and ASR (Gr6):		
- Multitask learning (MTL)	9.09	7.30
- Adversarial learning (AL)	75.16	7.30
- Speech attribute augmentation (SAug)	**8.81**	7.37
Proposed method		
- Speaker embedding at A & C, all layers (Gr6)	8.97	**7.21**
- Speaker embedding at A & C, all layers (Gr28)	11.27	7.26

- SAug has a better result for only SRE, but it performs worse in terms of ASR.
- The proposed method improves not only ASR but also SRE.
- Gr6 gives better performance than Gr28, showing that the combination of minor speakers is critical to solve the speaker-imbalanced problem.

4.4 Comparison of embedding options in all layers

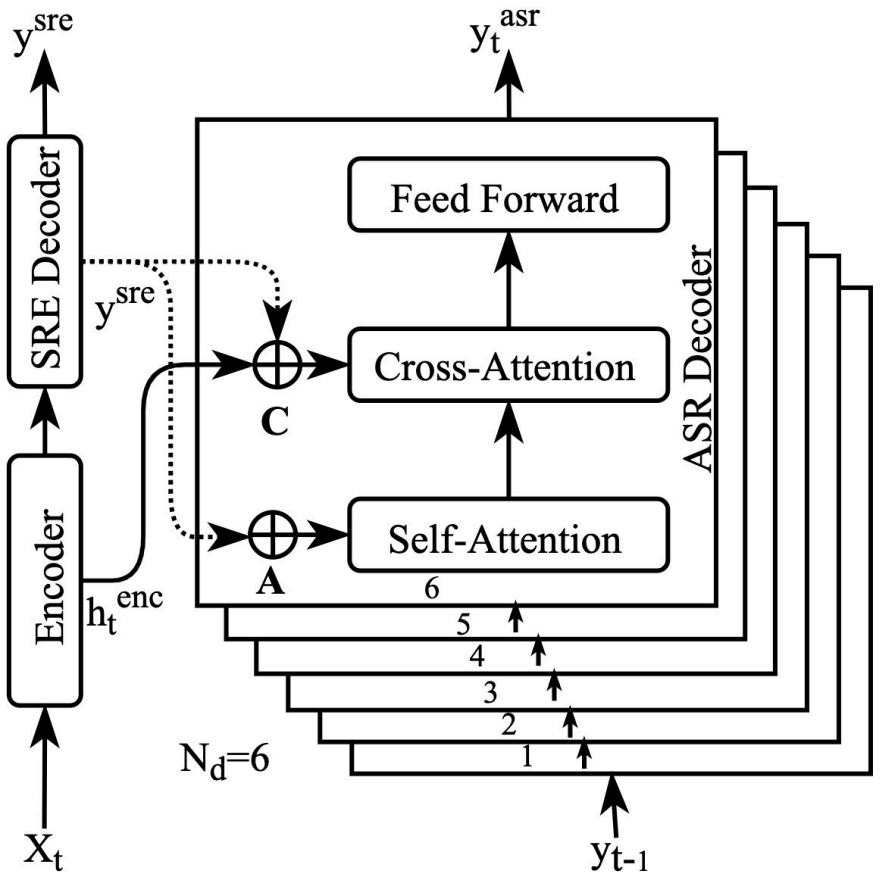
Embedded option in all layer (Gr6)	SRE (%incorrect)	ASR (%CER)
Option A	9.08	7.30
Option B	9.10	7.33
Option C	9.21	7.26
Option D	9.02	7.33
Option E	9.18	7.26
Option AC	8.97	**7.21**
Option BD	8.92	7.40

- The combined options AC is the most effective in both tasks because:
  - Option A is effective in term of SRE
  - Option C is effective in term of ASR



4.5 Comparison of layer-wise application of AC

Embedded layer of AC option (Gr6)	SRE (%incorrect)	ASR (%CER)
Layer 1	9.18	7.20
Layer 2	9.10	7.26
Layer 3	9.35	7.33
Layer 4	9.26	7.30
Layer 5	9.24	7.35
Layer 6	9.91	7.26
Layer 1, 2	9.02	7.28
Layer 1, 2, 3	9.10	7.24
Layer 1, 2, 3, 4	9.08	7.27
Layer 5, 6	9.48	7.29
Layer 4, 5, 6	9.61	7.28
Layer 3, 4, 5, 6	9.30	7.27



- Embedding into a single layer is effective in the ASR but degraded the SRE performance.
- Embedding into lower layers shows a better improvement of ASR and SRE together. This is reasonable as the speaker information is reduced in the ASR decoder.

A Joint Training of Automatic Speech Recognition and Machine Translation

Rejected from ICASSP2022

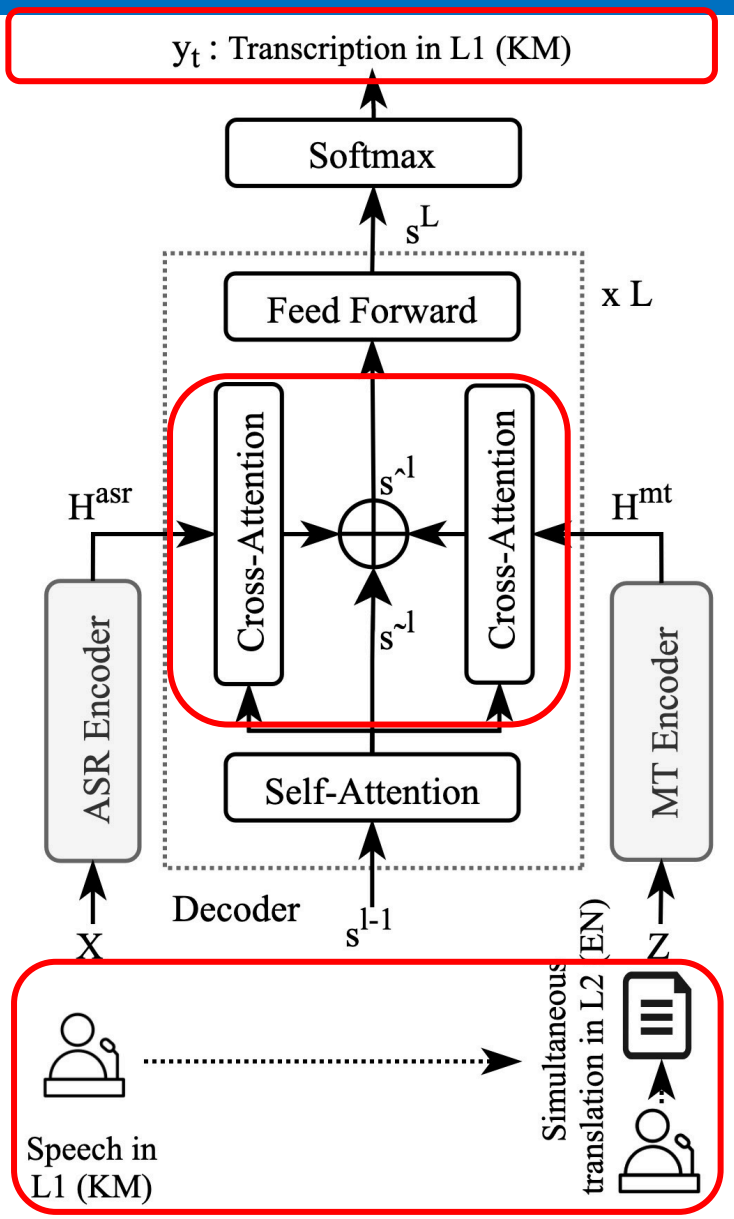
Resubmitted to Interspeech2022

Motivation

- In the multilingual meetings or court proceedings, the speech and translation are usually available.
- [Paulik et al., Khadivi et al., 2005] improved ASR by integrating statistical machine translation, however, those systems were independently trained.
- We propose a joint training end-to-end ASR-MT to improve ASR of a low-resource language using knowledge of simultaneous translation from a resource-rich language.
- For example, Japanese people recognize an American movie that has a simultaneous Japanese subtitle.

Proposed architecture

- Input:
  - speech utterances in L1, {X_1 , X_2 , ..., X_e },
  - their translations in L2, {Z_1 , Z_2 , ..., Z_e },
- Process:
  - sum cross-attention of ASR and MT
- Output:
  - text transcription in L1, {Y_1 , Y_2 , ..., Y_e },

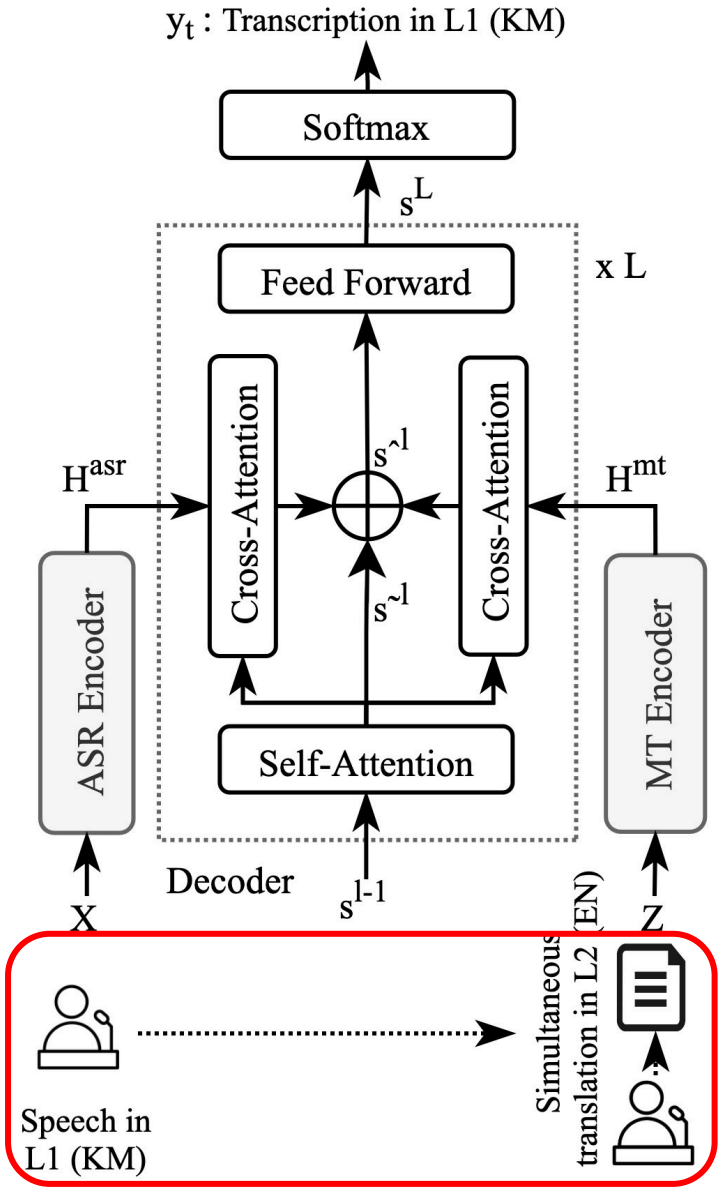


Example: Joint ASR and MT (1/4)

The input is the simultaneous of Speech (Khmer) and translation text (English).

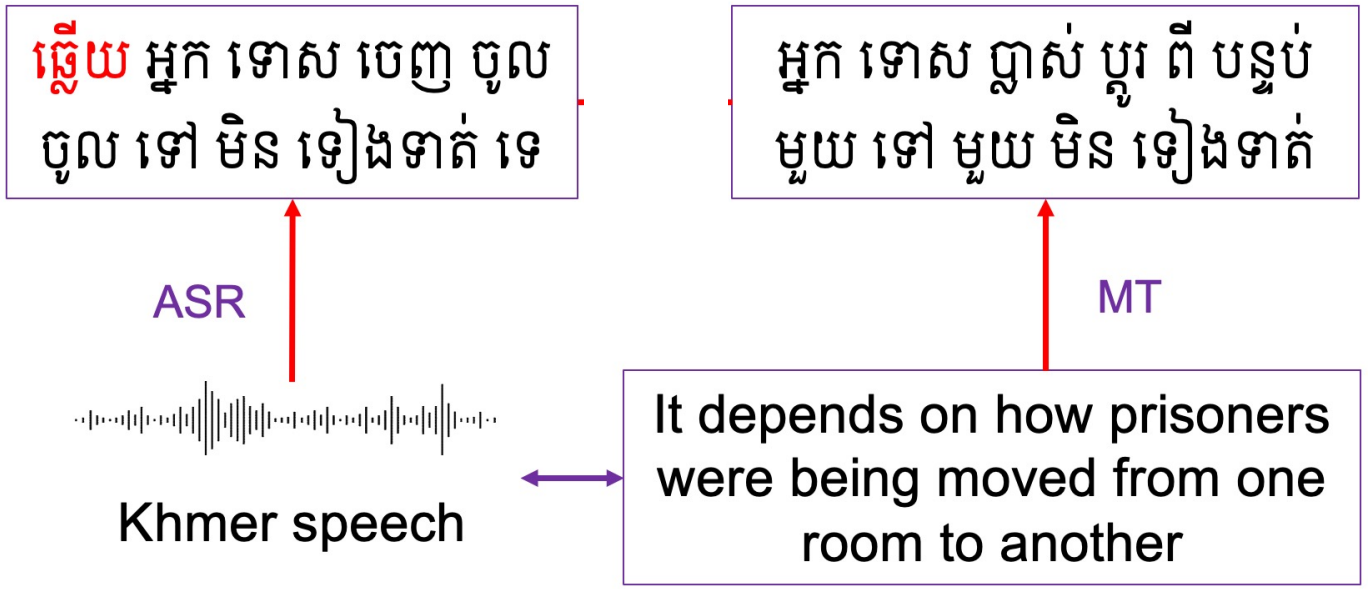
Khmer speech

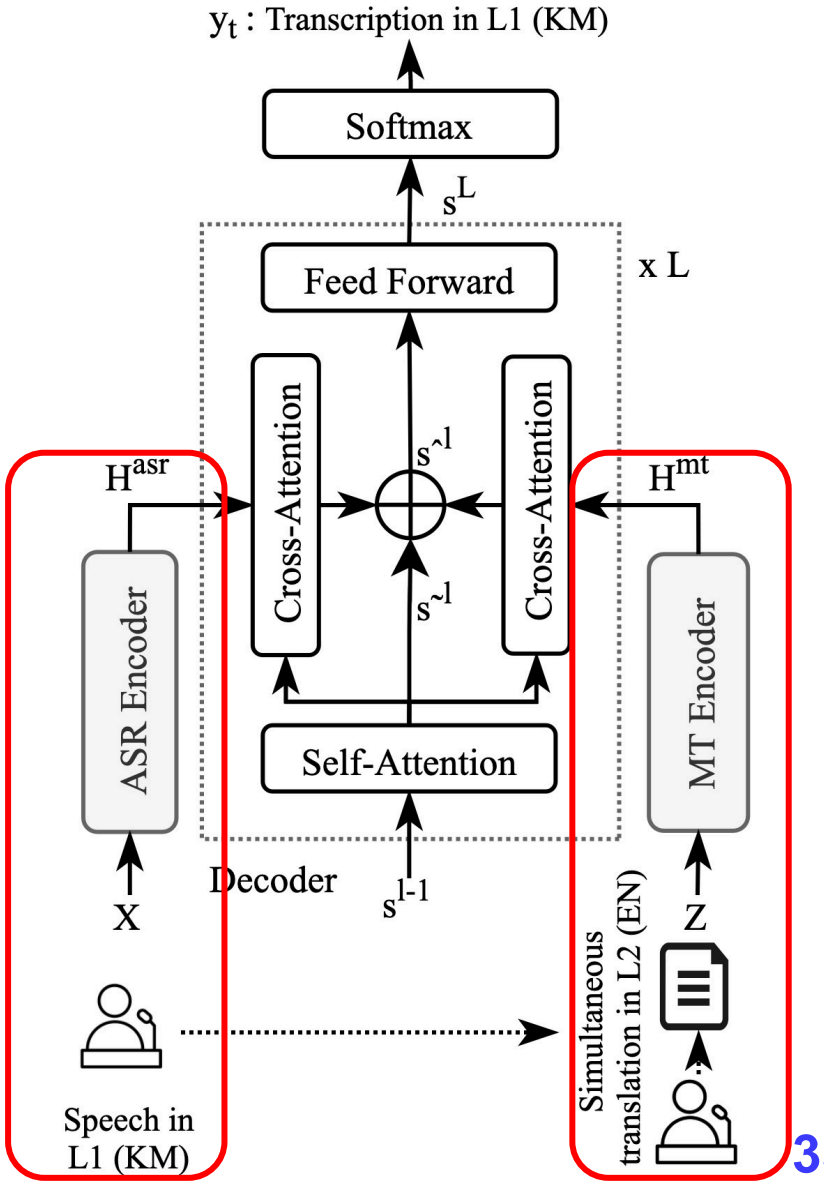
It depends on how prisoners were being moved from one room to another



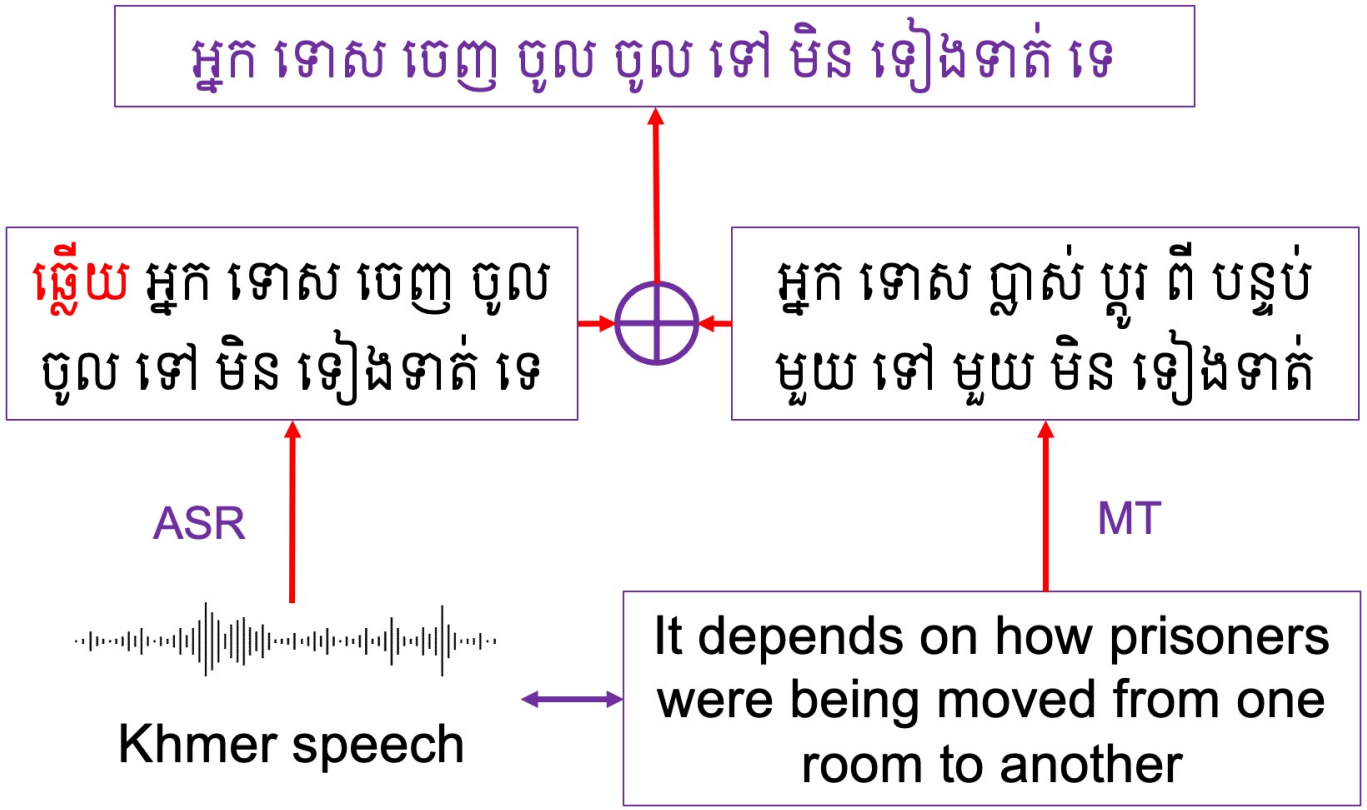
Example: Dual encoders (2/4)

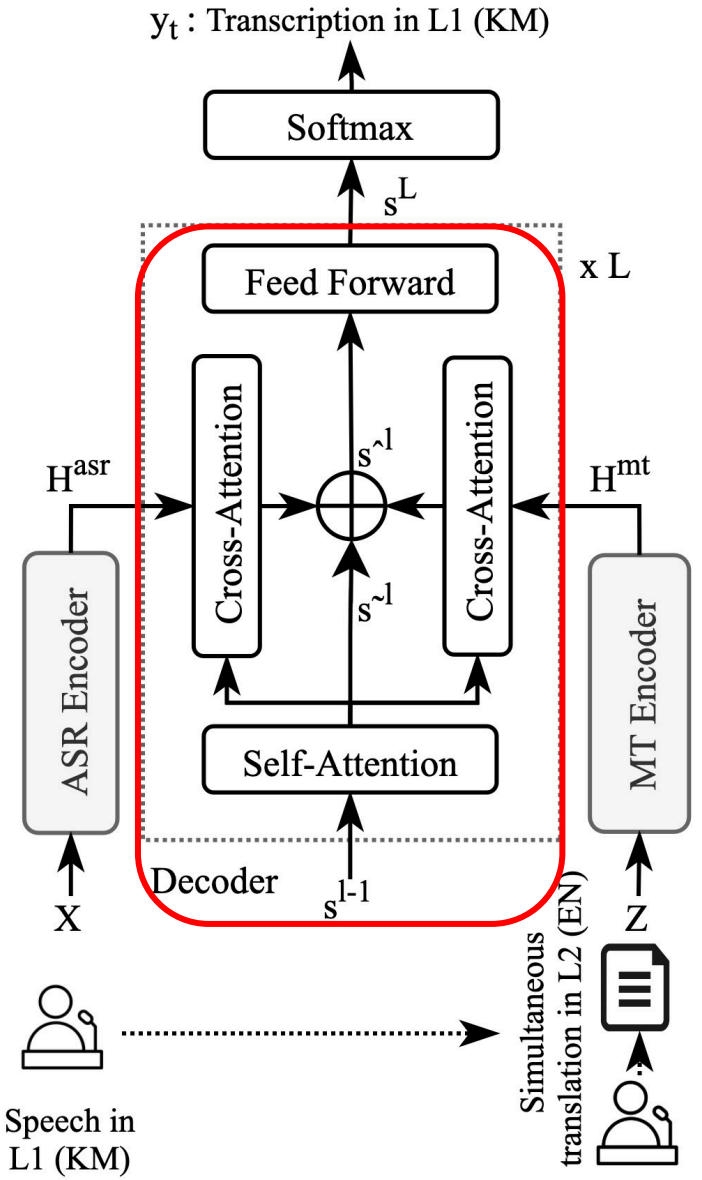
Each encoder produces different representation output.





Example: A joint via cross-attention (3/4)





Example: A joint via cross-attention (4/4)

Reference

អ្នក ទោស ចេញ ចូល ចេញ ចូល ទៅ មិន ទៀងទាត់ ទេ

1 error

អ្នក ទោស ចេញ ចូលចូល ទៅ មិន ទៀងទាត់ ទេ

2 errors

ឆ្លើយ អ្នក ទោស ចេញ ចូល ចូល ទៅ មិន ទៀងទាត់ ទេ

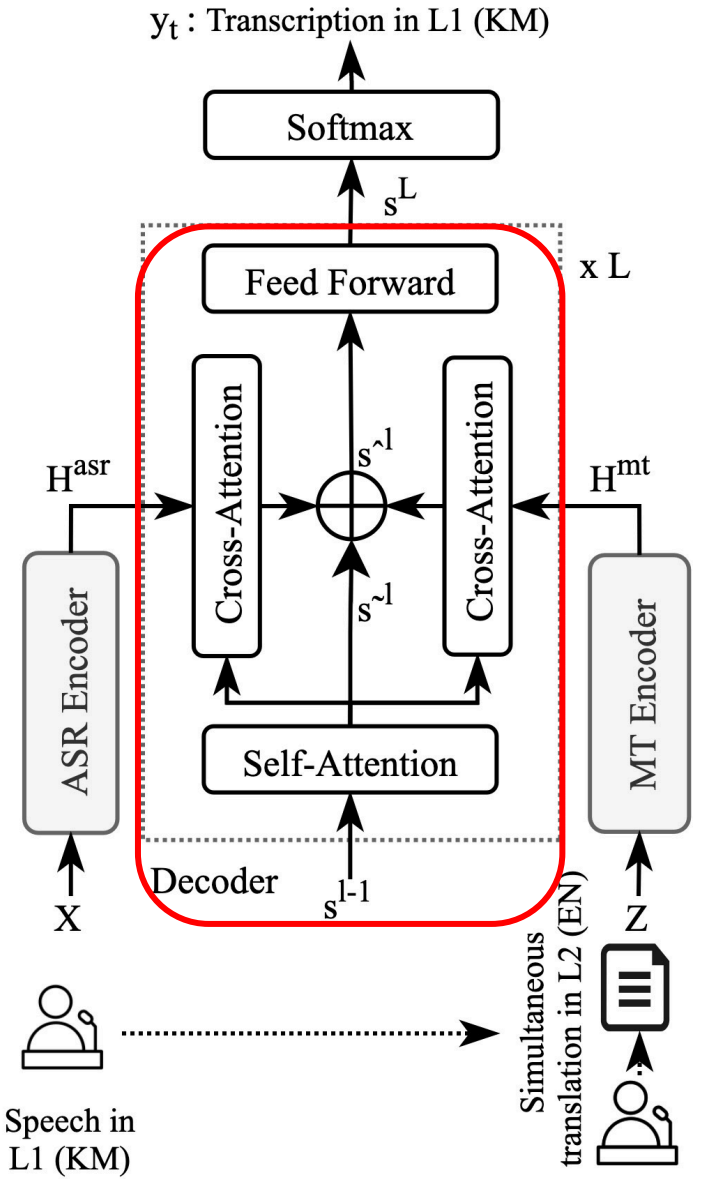
អ្នក ទោស ឆ្លាស់ ប្តូរ ពី បន្ទប់ មួយ ទៅ មួយ មិន ទៀងទាត់

ASR

MT

Khmer speech

It depends on how prisoners were being moved from one room to another



Experiments setup

- ECCC:
  - Encoder: 6, Decoder: 6, FFN units: 2048, attention head: 4, attention-dim: 256
  - 45 epochs
  - 5000 BPE tokens per language
- Fisher-CallHome:
  - Encoder: 12, Decoder: 6, FFN units: 2048, attention head: 4, attention-dim: 256
  - 50 epochs
  - 1000 BPE tokens per language
- Train in a single GPU using 64-batch-size

Experiment results on ECCC corpus

- **Target**: Improving Khmer ASR (KM-ASR)
- **Input**: KM-ASR and EN-KM/FR-KM MT

Model	WER (%)		
	Baseline-ASR	Joint ASR-MT (EN-KM)	Joint ASR-MT (FR-KM)
w/o augmentation	23.6	22.2	**22.0**
w/ speed perturbation (SP)	22.2	**21.1**	21.4
w/ SpecAugment (SA)	21.8	**20.5**	20.6
w/ SP + SA	21.4	**19.5**	20.2

- Joint training model improved over the baseline in both language pairs.
- With EN-KM MT shows the best peformance by reducing 1.9% of WER.

ASR improvement on each group of speakers

Speaker Group	# Hour	Baseline WER(%)	Joint_{en} WER(%)	Relative (%)
Witness	5	23.4	19.7	15.8
Co-prosecutor	2	**19.7**	19.5	1.0
Civil-party	0.7	**15.3**	13.7	10.5
Judge	0.3	**17.0**	17.1	-

- The “Witness” and “Civil-party” reduce a large WER:
  - These speaker groups include are the victims of the Khmer Rouge regime who are elderly and illiterate.
  - They cannot pronounce words correctly and they are disfluency and emotions in their speech during the trial
- However, the performance is comparable in “Judge” and “Co-prosecutor” because they are fluently for their speech.

ASR improvement on baseline WER distribution

Baseline WER (%)	# utterance	Baseline WER(%)	Joint_{en} WER(%)	Relative (%)
0-10	1,137	4.5	5.3	-
10-20	810	14.9	14.2	4.7
20-30	538	25.8	23.6	8.5
30-40	248	37.8	32.5	14.0
40-50	165	49.4	43.3	12.3
50-100	303	88.1	75.3	14.5

- The worse baseline ASR was, the more improvement is achieved.
- The best improvement reduced the WER by 14.5% relative.

ASR improvement on baseline BLEU distribution

MT BLEU (%)	# utterance	Baseline WER(%)	$Joint_{en}$ WER(%)	Relative (%)
0-10	895	23.4	21.7	7.3
10-20	1205	20.2	18.4	8.9
20-30	572	20.5	18.6	9.3
30-40	268	18.7	17.1	8.6
40-50	126	19.3	18.2	5.7
50-100	136	23.5	18.6	20.9

- The better MT performance resulted in better improvement in the transcription of speech.
- The best peformance reduced the WER by 20.9% relative.

Results on Fisher-Callhome Spanish corpus

- **Corpus**: Spanish to English speech translation.
- **Target**: Improving the Spanish speech transcription
- **Input**: Spanish ASR and English to Spanish MT

Model	Fisher						CallHome			
	dev		dev2		test		devtest		evltest	
	ASR	Joint	ASR	Joint	ASR	Joint	ASR	Joint	ASR	Joint
w/ SP	24.2	**24.0**	23.6	**23.1**	21.5	21.7	41.1	**40.5**	41.4	**41.0**
w/ SP+SA	23.1	**22.8**	22.5	**22.3**	20.8	**20.5**	40.2	**39.5**	39.6	**39.4**

- The proposed method shows the improvement by reducing the WER of Spanish by 0.7% (1.7% relative).
- These indicate that the proposed method is generalized to other corpora.

Conclusion

- Alignment sentence in monotonic manner is useful for no sentence boundary languages (source language is in sentence level).
- Integrating speaker information into the ASR decoder is not effective only ASR but also SRE. (can be extended to multilingual system)
- Identifying and clustering speakers was effective to solve the problem of speaker-imbalanced. (Unseen speaker should be investigated)
- The MT knowledge has shown to be useful for enhancing the transcription of speech, especially the low-performance of ASR with the higher translation quality. (Multi-source ASR/ST should be investigated)

Question?

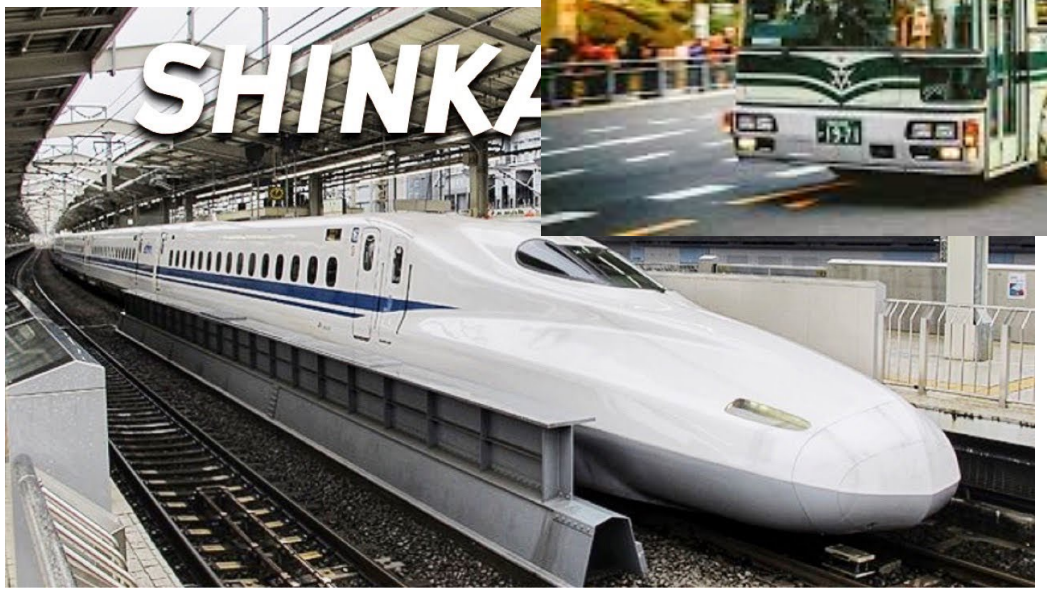
How is the life in Japan?

Beautiful and clean

?

Safety and Healthy

Public transportation is convenience.



Everywhere is Japanese language!

Me, know nothing in Japanese!

Japanese people
are too silent !
No words,
it seems to be
not friendly.