Mandarin-English Code-switching Speech Recognition with Self-supervised Speech Representation Models




Liang-Hsuan Tseng   Yu-Kuan Fu   Heng-Jui Chang   Hong-yi Lee

National Taiwan University

AAAI-22

AAAI SAS 2022

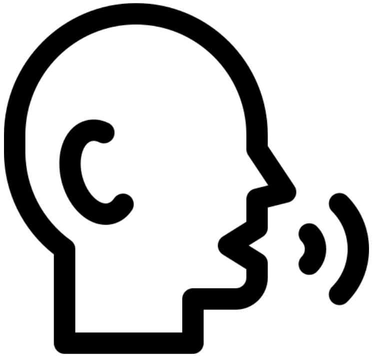
Outline

- **Introduction**
  - What is code-switching(CS)
  - Difficulties in CS ASR
  - The advantage of using SSS representation
- **Methods**
  - Extracting speech representation from SSL model
  - CTC module for CS ASR
  - LID module for CS language identity
  - CTC-LID Jointly framework
- **Experiments**

Introduction

What is Code-switching (CS) ?

Alternating between two or more languages in speech.

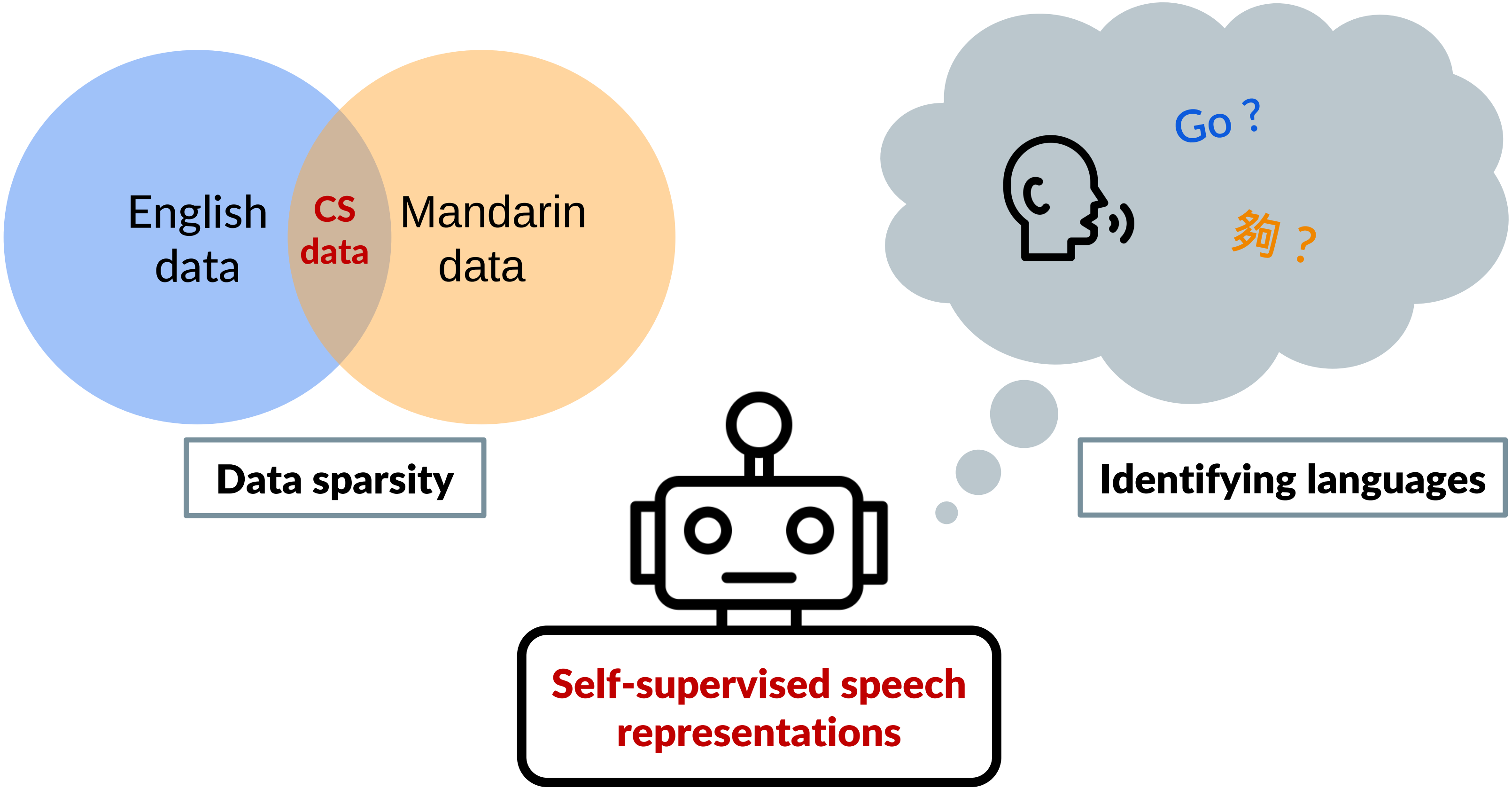


我目前正在做 Self-supervised learning 相關的 research.

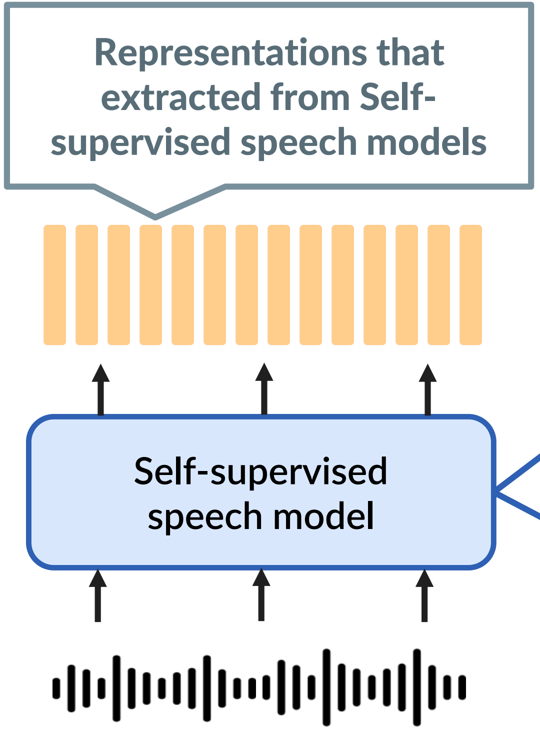
*Orange: Mandarin *Blue: English

We mainly focus on Mandarin-English code-switching, which is quite common in East Asia or South-East Asia

Difficulties of Code-switching(CS) ASR



Self-supervised Speech Representations…?



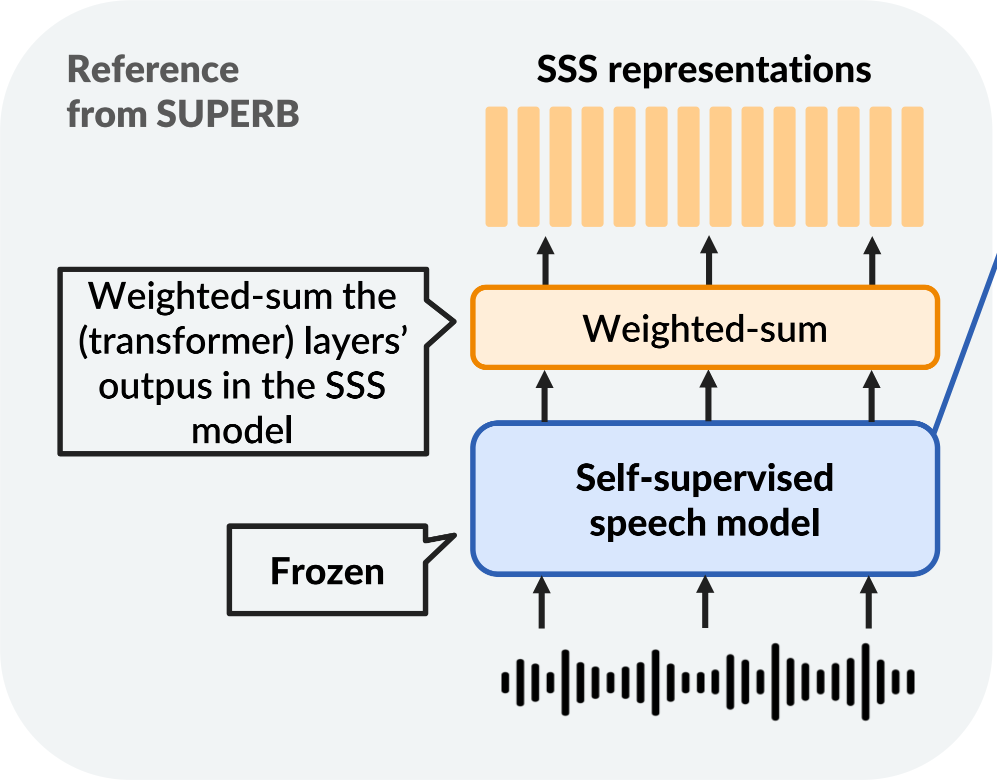
What are the advantages of self-supervised speech representations?

Pre-trained on huge amount of data.
⇒ Bring benefits to CS ASR, which has data sparsity issue.

Contains transformer architecture.
⇒ More contextual info in its speech representations.
⇒ Easier to identify languages.

Methods

Extract Self-supervised Speech(SSS) Representations

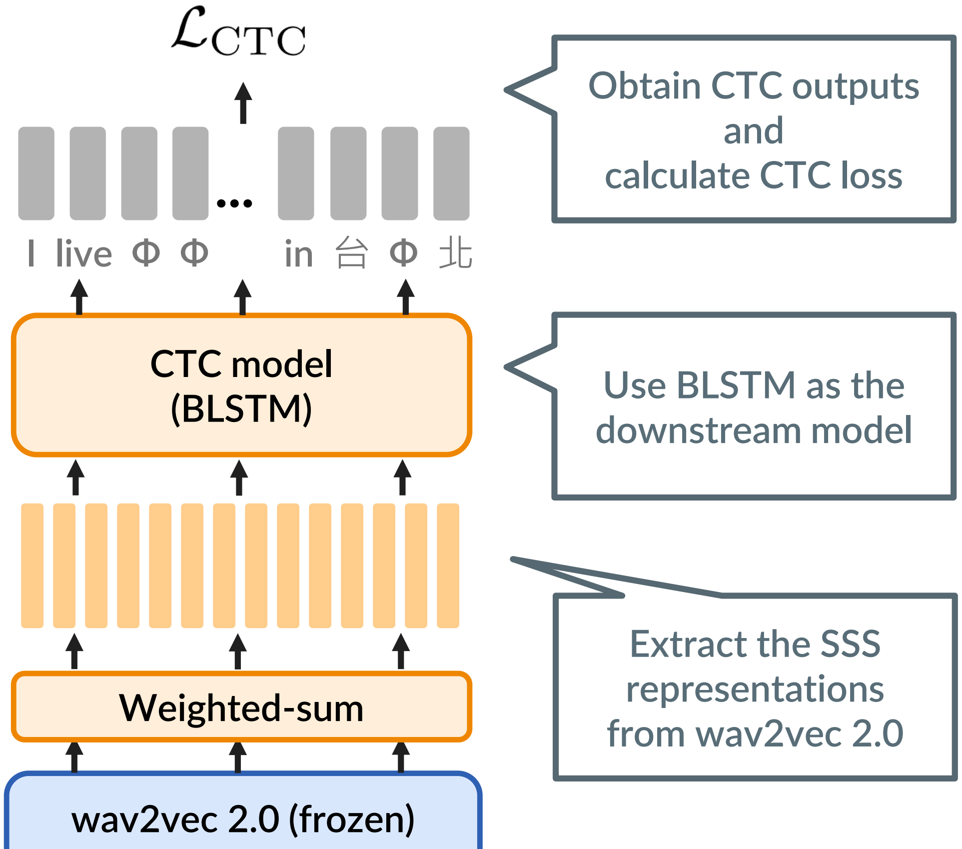


We use **wav2vec 2.0**
Not only because it is powerful…

Model	Layers	Data (hr)	Languages
Base	12	960	1 (EN)
Large	24	60k	1 (EN)
XLSR	24	56k	53

But also because it provides Multi-lingual / Mono-lingual models pre-trained with similar amount of data

CTC module for Code-switching(CS) ASR



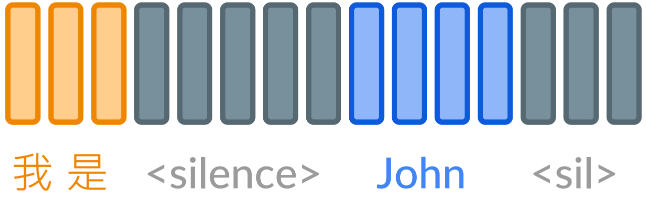
This is a simple CTC framework but with SSS representations as input.

Verifying language identity(LID) in CS speech

Besides of CS ASR, verifying language Identity(**LID**) is also considered as a common technique to help tackling CS problems.

⇒ We want to investigate the effectiveness of SSS representations on LID task.

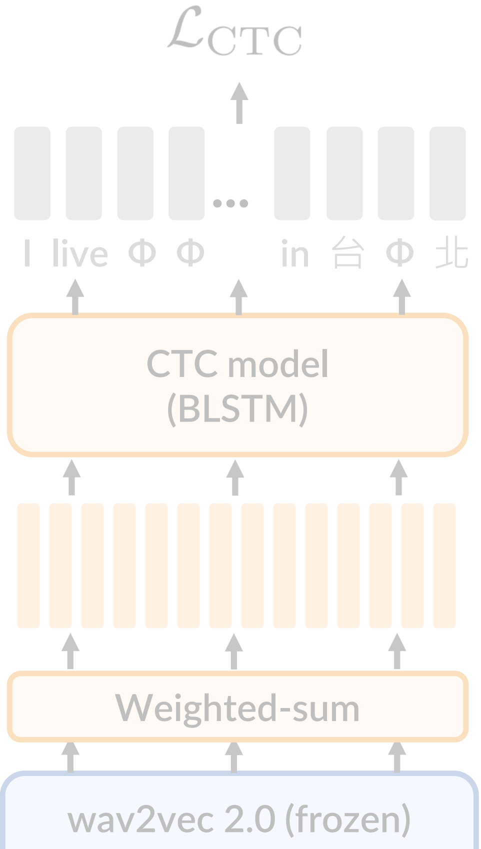
***Frame-level LID!**



Three classes:

Mandarin **English** **Silence**

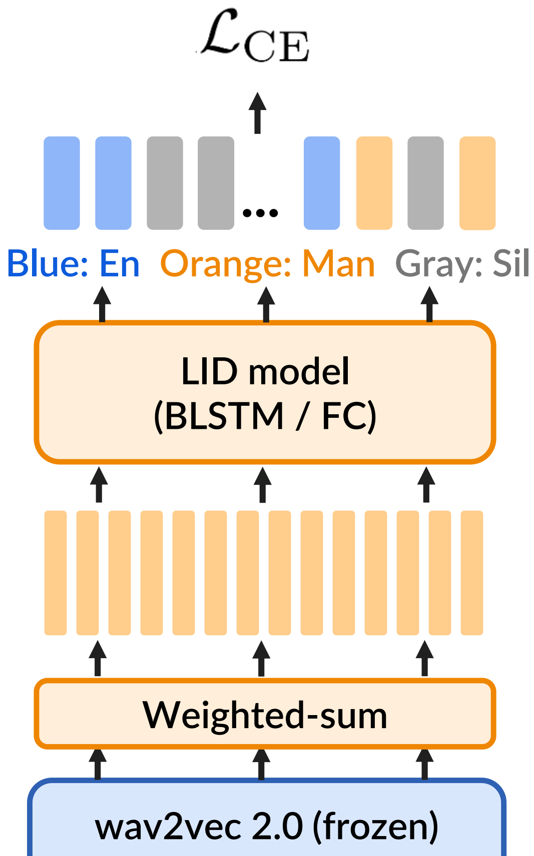
LID module for Code-switching(CS) Language Identity



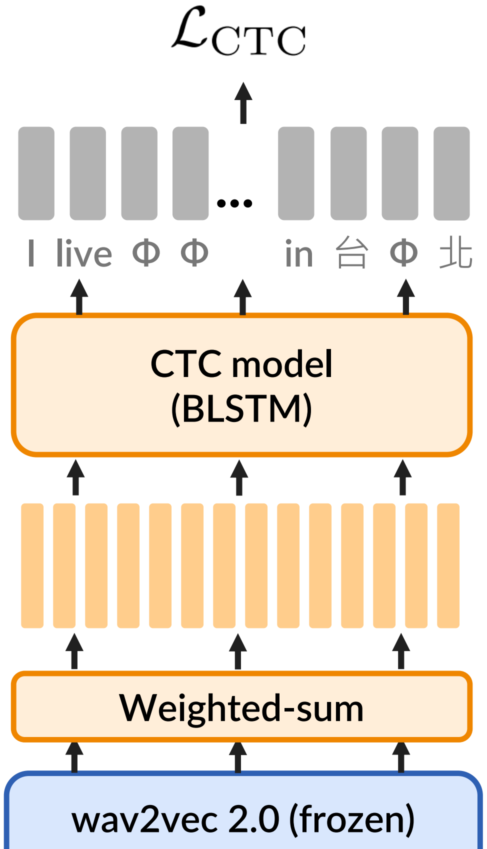
Calculate Cross Entropy loss on frame-level LID predictions

Use Fully Connected network(FC) to see how easy LID task is with the SSS representations

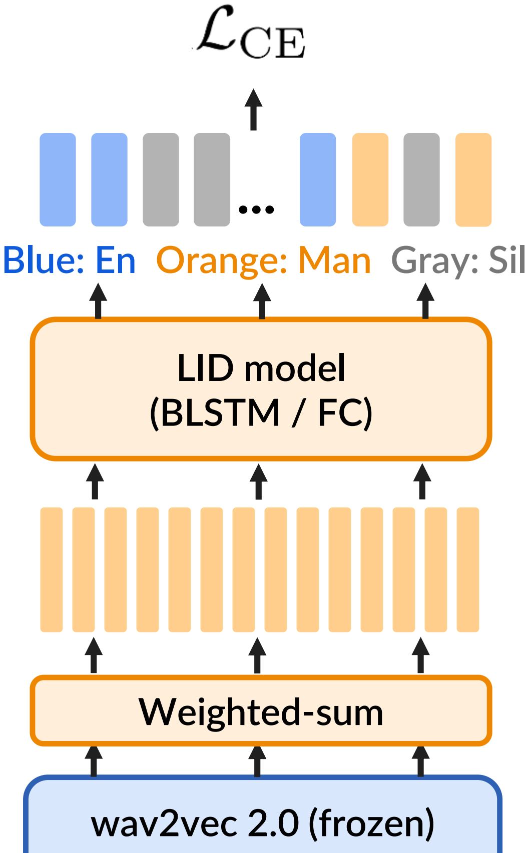
The framework is very Similar with the CTC module



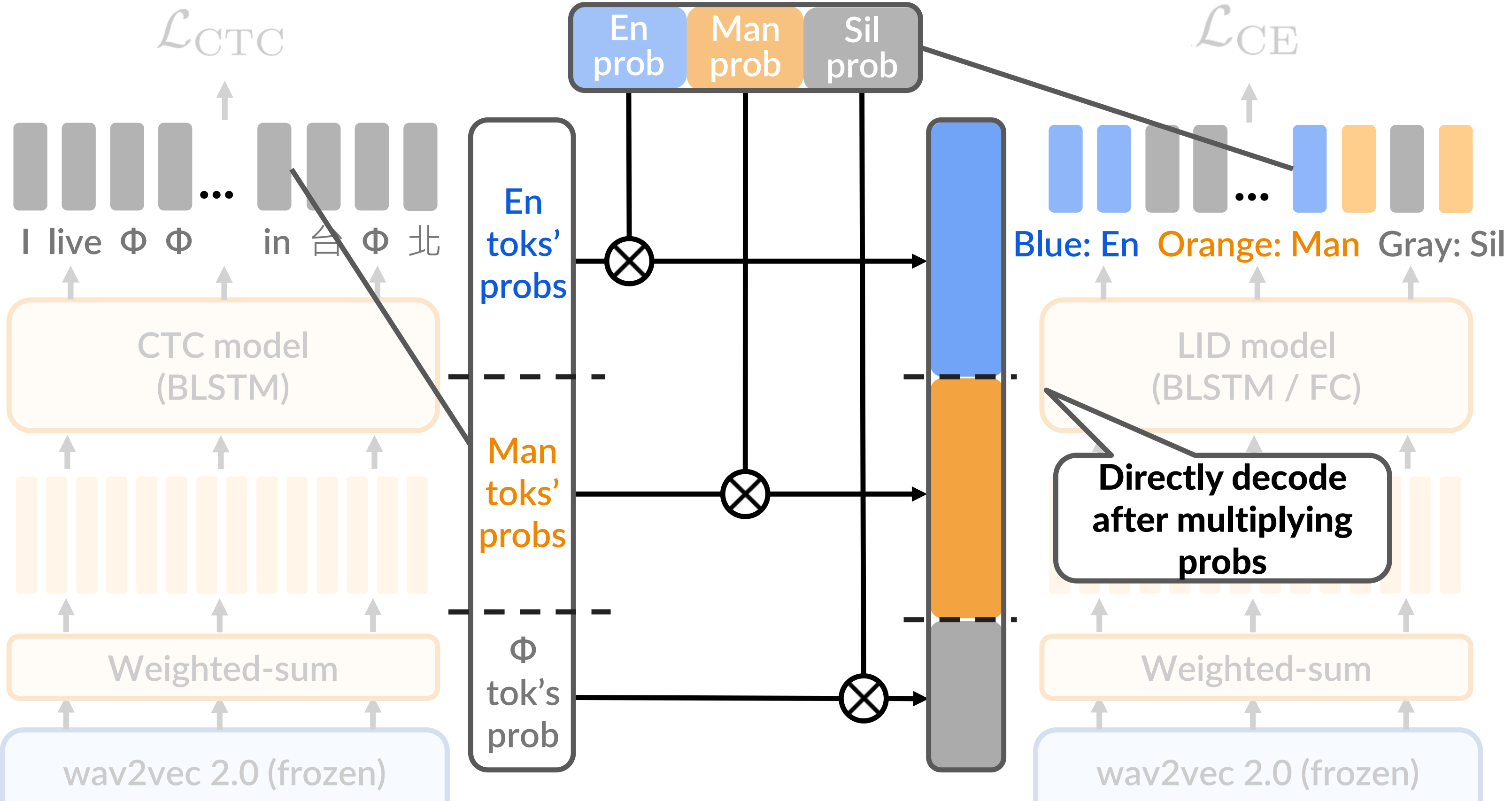
The CTC module and the LID module



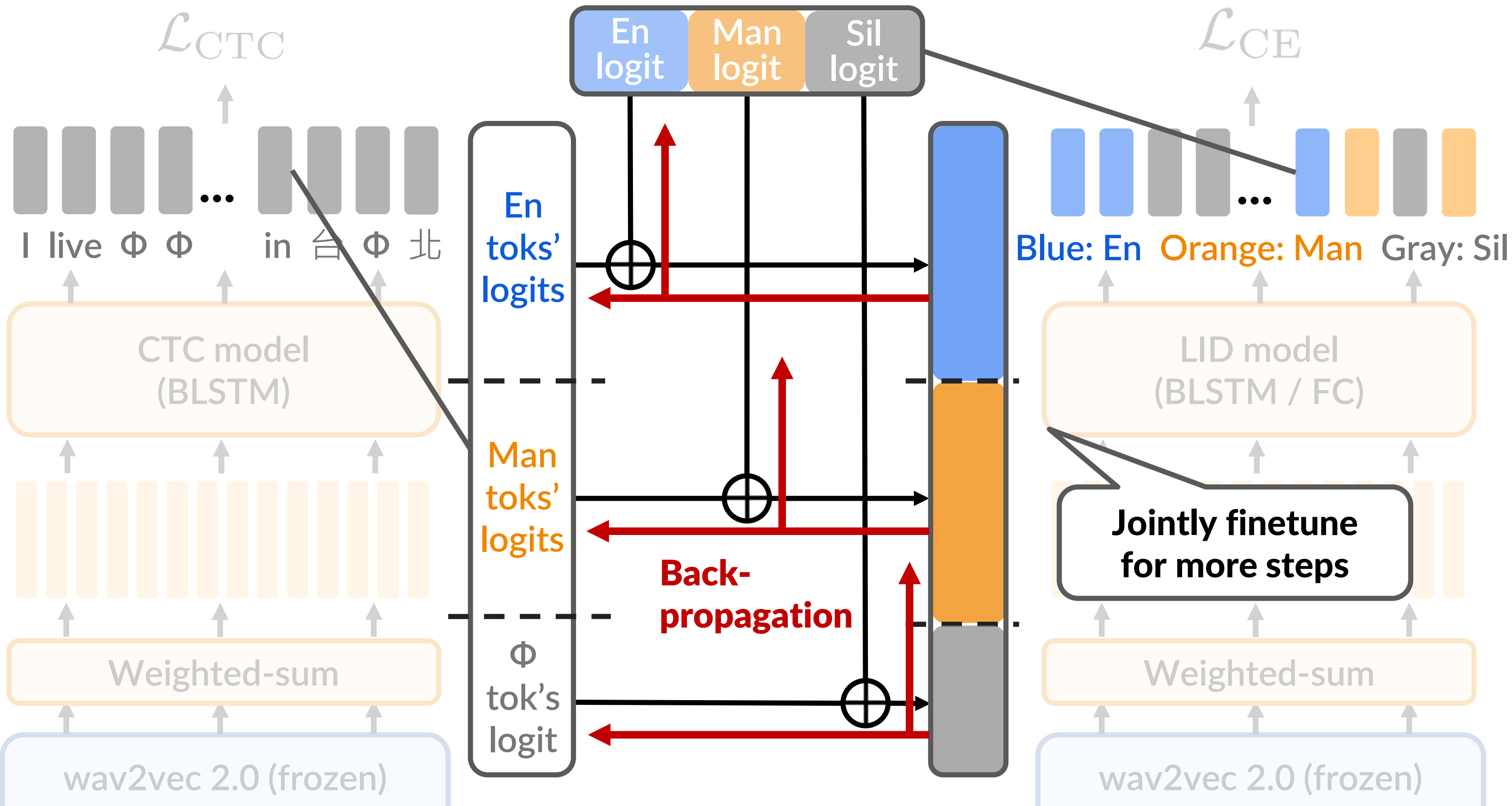
How can the LID module benefit the CTC module...?



Directly multiply LID probs to CTC probs (Li et al. 2019)



Add LID logits to CTC logits and jointly train (proposed)



Experiments

Data

- SEAME corpus : A Mandarin-English Code-switching Dataset
- Language distribution:

Testing sets

	train	val	dev-man	dev-sge
Duration (hours)	93.0	4.7	7.5	4.0
Mandarin	23.4%	23.7%	23.5%	9.8%
English	21.4%	21.1%	11.2%	48.7%
Code-switching	55.1%	55.2%	65.3%	41.5%

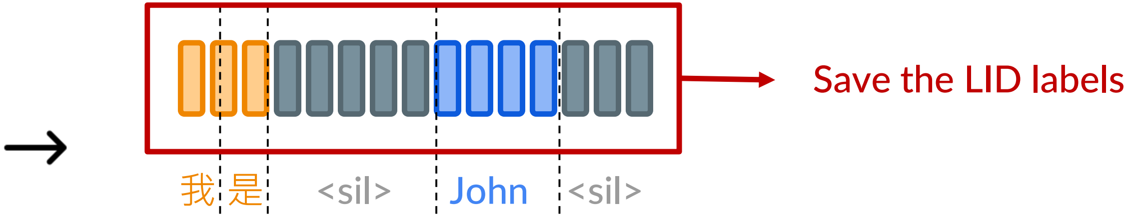
Mandarin Dominant

English Dominant

Data – How to Form LID labels

Input paired speech and text data.

Use the Montreal-Force Aligner to define word boundaries.



Use word segments to label LID in frame-level, then save the LID labels.

CS ASR with only CTC module but different inputs

Token Error Rate(TER):
Similar to WER but compute on Tokens(Mandarin characters and English words, in our scenario).

TER(%) on SEAME testing sets

	Method	dev-man	dev-sge
Traditional	(a) fbank	29.4	40.8
SSS representation from different wav2vec 2.0 models	(c) Base	20.8	30.8
	(e) Large	19.8	29.7
	(h) XLSR	**19.4**	**28.8**

Outperformed fbank by over 30% relative TER.

Multi-lingual pre-trained model gives the best performance.

CS LID with different models and inputs

Accuracy (%) of LID on SEAME testing sets

Method	dev-man	dev-sge
(a) fbank + FC	59.4	37.7
(b) fbank + BLSTM	84.7	82.3
(c) Base + FC	75.0	68.5
(d) Base + BLSTM	91.9	89.5
(e) Large + FC	77.9	71.8
(f) Large + BLSTM	92.3	89.9
(g) XLSR + FC	76.4	69.7
(h) XLSR + BLSTM	**92.7**	**90.0**

*FC:
Fully Connected

*BLSTM:
Bidirectional LSTM

Even the simple FC prediction head can score high accuracy

Highest accuracy when using the representation from the multi-lingual pre-trained model

Jointly Training - Results

Only CTC module

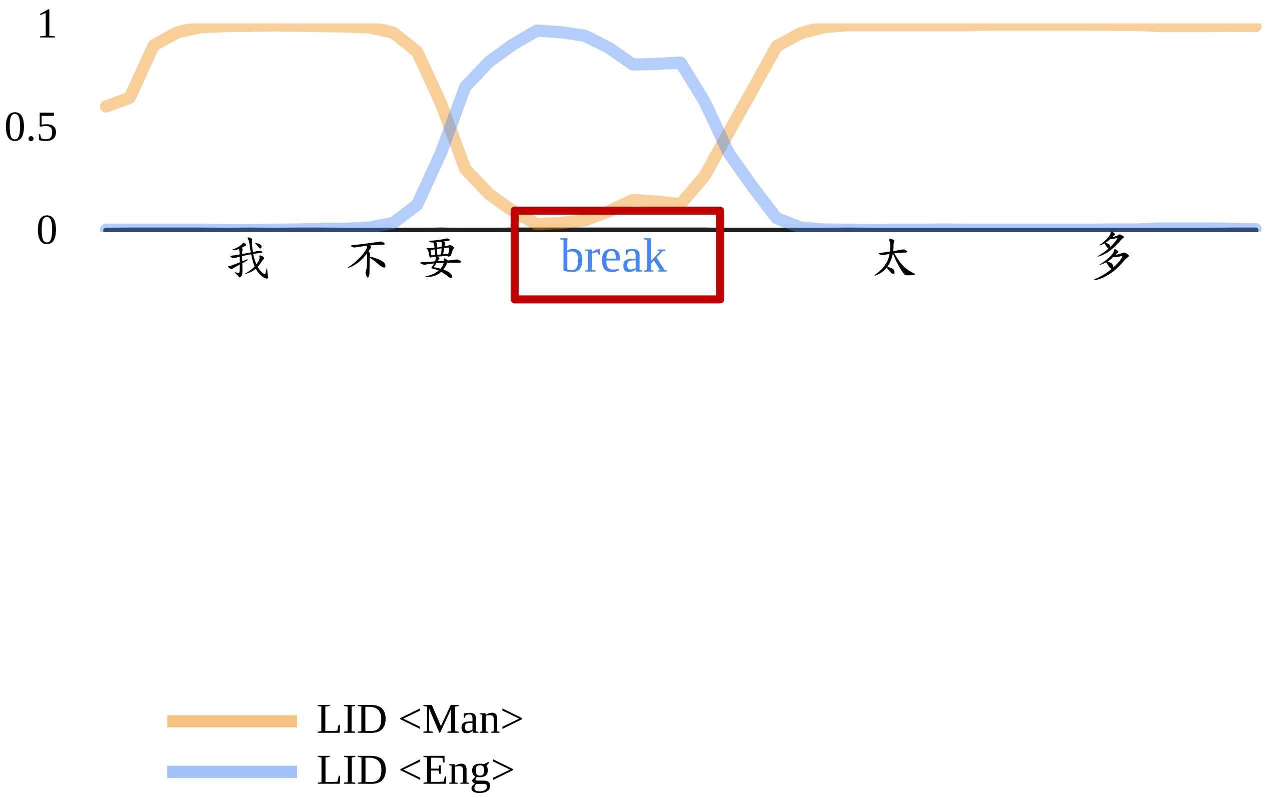
Directly decode by multiplying LID probs to CTC probs (Li et al. 2019)

TER(%) on SEAME testing sets

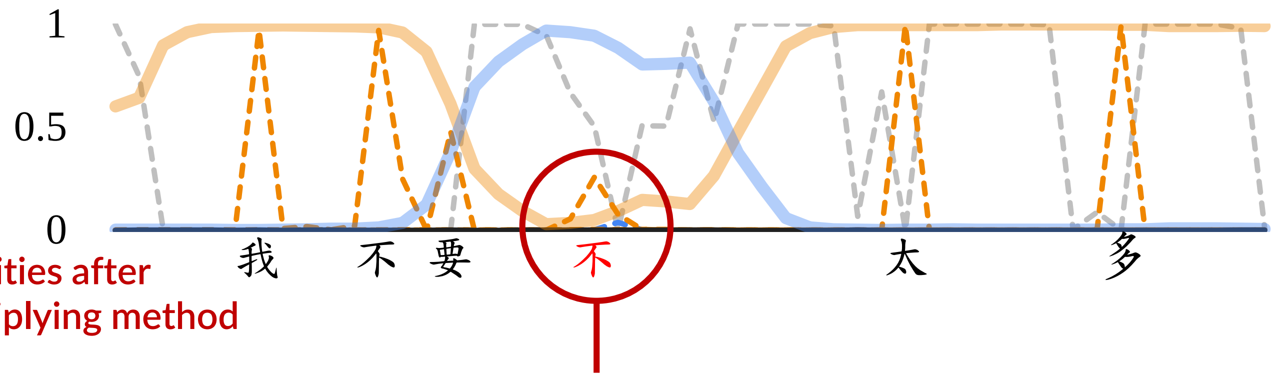
Method	dev-man	dev-sge
(I) CTC only		
(a) Large	19.8	29.7
(b) XLSR	19.4	28.8
(III) CTC-LID (directly decode)		
(c) Large	20.9	31.3
(d) XLSR	20.2	29.9
(III) CTC-LID (jointly finetune)		
(e) Large	19.9	29.8
(f) XLSR	**18.8**	**28.5**

Add LID logits to CTC logits then jointly finetune (proposed)

Analysis on our jointly training method

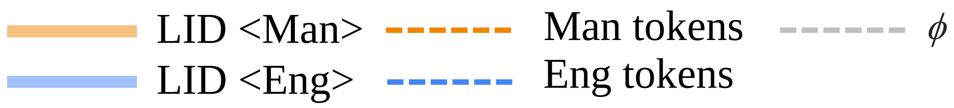


Analysis on our jointly training method

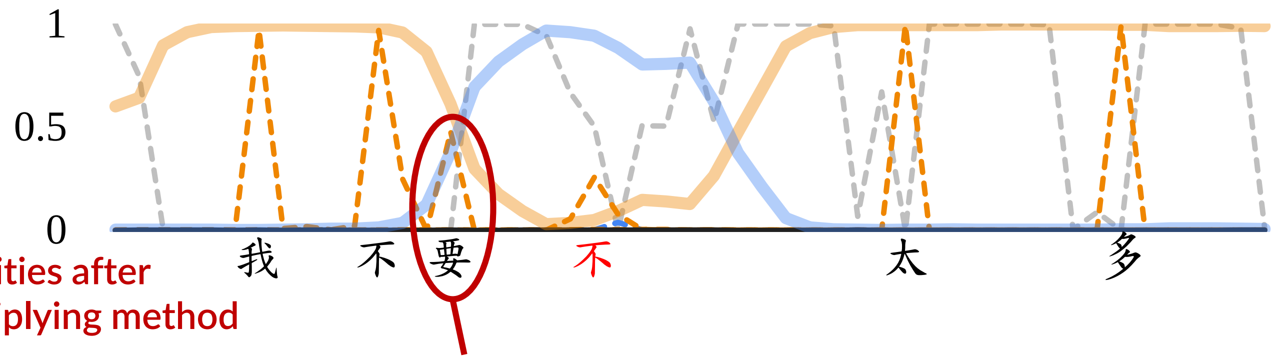


CTC probabilities after
Directly multiplying method

The Mandarin word “不” is wrong but it still cannot be brought down by directly multiplying LID probabilities to it.

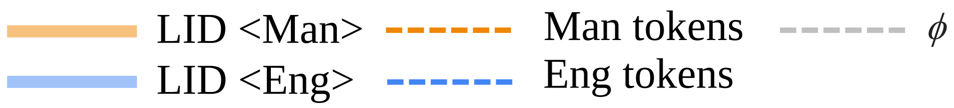


Analysis on our jointly training method

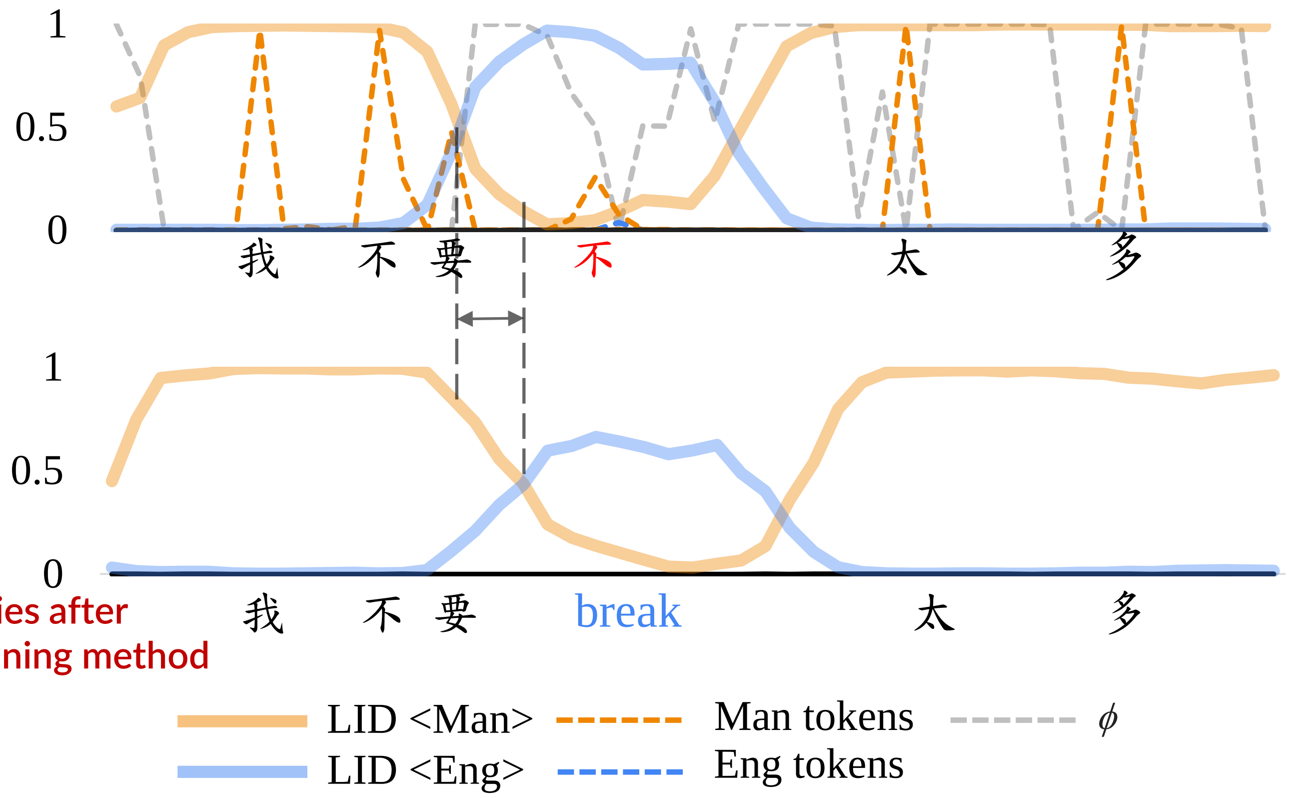


CTC probabilities after
Directly multiplying method

The probability of the Mandarin word “要”
is compressed by the misalignment
between the CTC and the LID outputs.

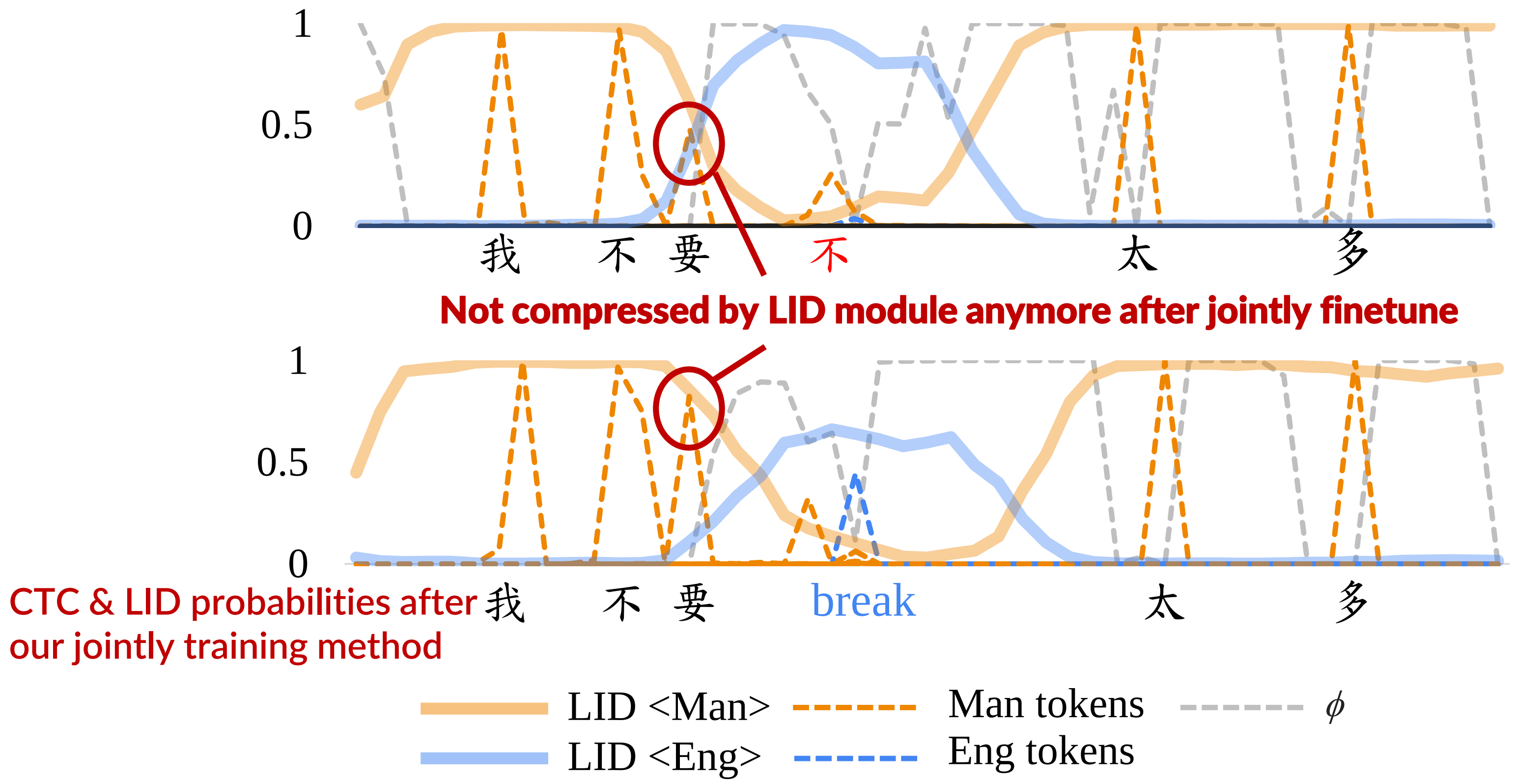


Analysis on our jointly training method

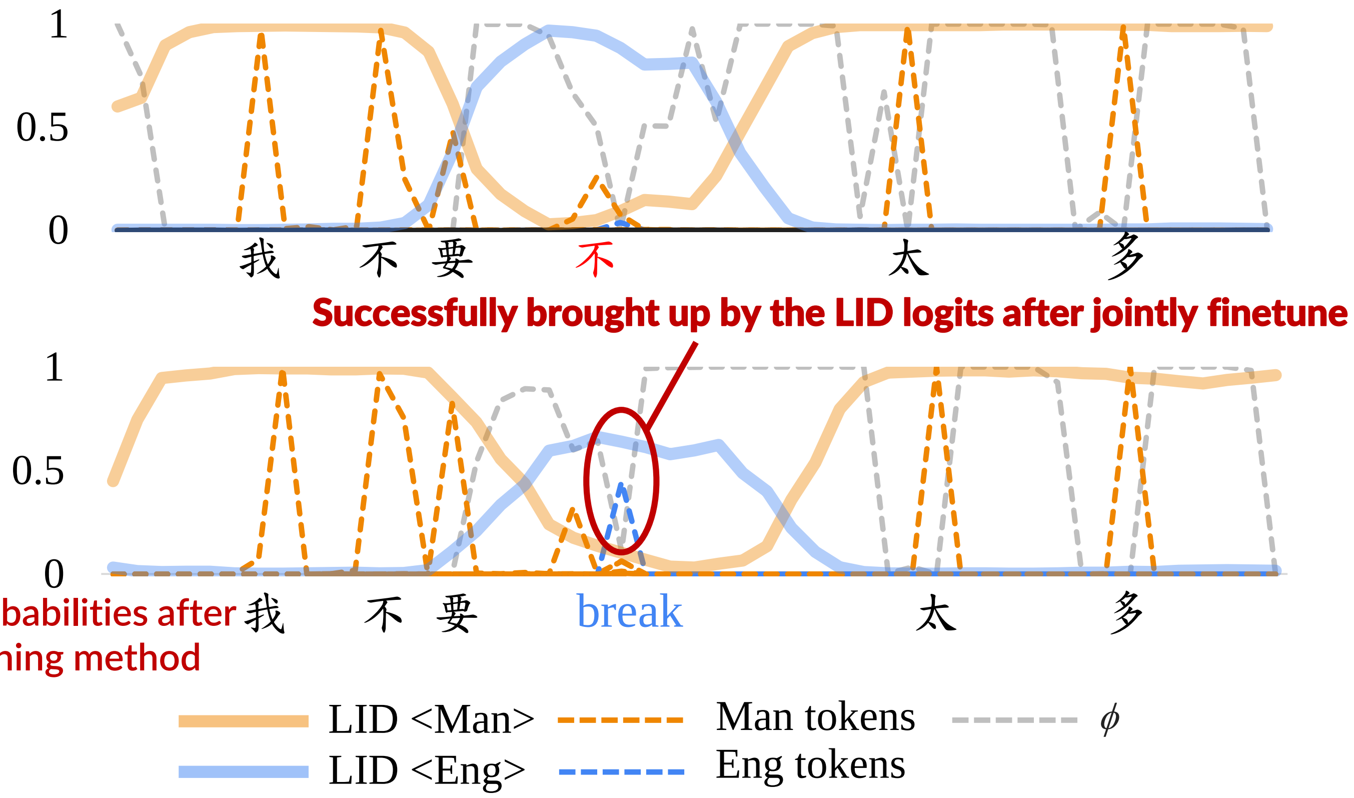


LID probabilities after our jointly training method

Analysis on our jointly training method



Analysis on our jointly training method



CTC & LID probabilities after our jointly training method

Conclusion

- Self-supervised Speech Representations performs well on code-switching tasks
- Multi-lingual pre-trained model gives better performance on our code-switching tasks
- Proposed a method to boost CS-ASR performance by jointly training with LID module

More details can be found in our paper
Thanks for Listening !

National
Taiwan
University

AAAI
SAS
2022

Paper:

Contact: