Self-Supervised Domain Adaptation for Multilingual ASR

Mingkun Huang
ByteDance Speech & Audio

Outlines

- Introduction
- Multilingual ASR
- Language Identification
- Data Augmentation
- Unsupervised Domain Adaptation
- Conclusion

Introduction

Dataset

Language	Train (#hours)	Test (#hours)	#Duplications
Hindi	96	6	22
Marathi	94	5	31
Odia	94	5	73
Tamil	40	5	1.3
Telugu	40	5	1.3
Gujarati	40	5	1.1

Multilingual ASR

Baseline System

- Data: Uniformly sample from each language.
- Model: Wav2vec 2.0 base.
- Criterion: CTC.
- Output units: Characters (shared vocabulary).
- Decoding: WFST beam search with mixture language model.

Multilingual ASR

Baseline Results

Language	*Baseline Dev	*Baseline Test	Our Dev	Our Test
Hindi	40.41	37.2	23.95	27.63
Marathi	22.44	29.04	11.75	86.64
Odiya	39.06	38.46	28.77	27.45
Tamil	33.35	34.09	17.47	33.8
Telugu	30.62	31.44	18.76	31.78
Gujarati	19.27	26.15	12.78	23.11
Average	**30.73**	32.73	**18.91**	38.4

Notation:
*: provided hybrid system.
Dev: provided test set.
Test: blind test set without label.

Compare with hybrid system:
Worse on **Marathi**.

Language Identification

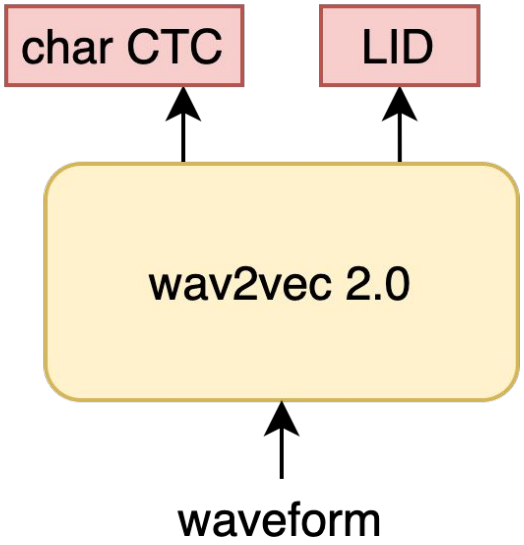
Oracle LID

Language	Mix. LM	Sep. LM
Hindi	23.95	13.42
Marathi	11.75	5.08
Odiya	28.77	25.21
Tamil	17.47	16.77
Telugu	18.76	17.66
Gujarati	12.78	12.26
Average	18.91	**15.07**

Separate language model decoding gives **20% relative** improvements over mixture language model decoding.

Language Identification

Multitask CTC + LID



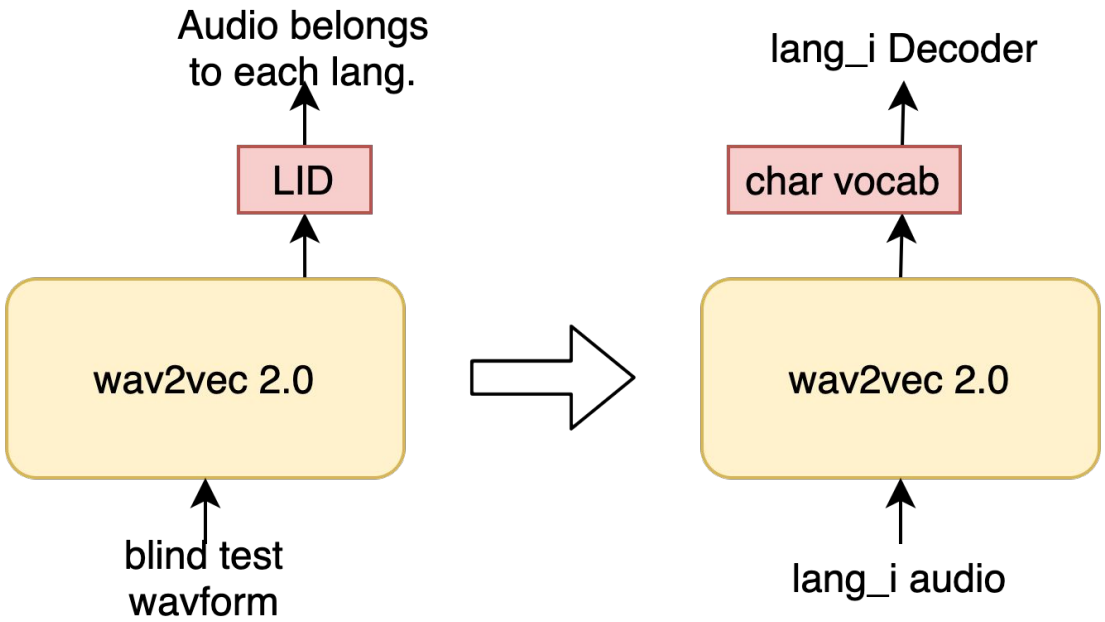
Language	Mix. LM	Sep. LM	CTC+LID
Hindi	23.95	13.42	13.42
Marathi	11.75	5.08	5.09
Odiya	28.77	25.21	23.88
Tamil	17.47	16.77	16.71
Telugu	18.76	17.66	17.12
Gujarati	12.78	12.26	12.12
Average	18.91	**15.07**	**14.72**

CTC+LID even better than oracle LID decoding.

On Dev set!

Language Identification

Blind test decoding process



Data Augmentation

Even though LID works well on dev set (100% Acc),

it can not distinguish **Marathi from Hindi** on test set.

MixAudio

Randomly sample other language audio as background noise.

$$\text{snr} = 10 \log_{10} \frac{\text{speech}^2}{(\lambda \cdot \text{noise})^2}$$

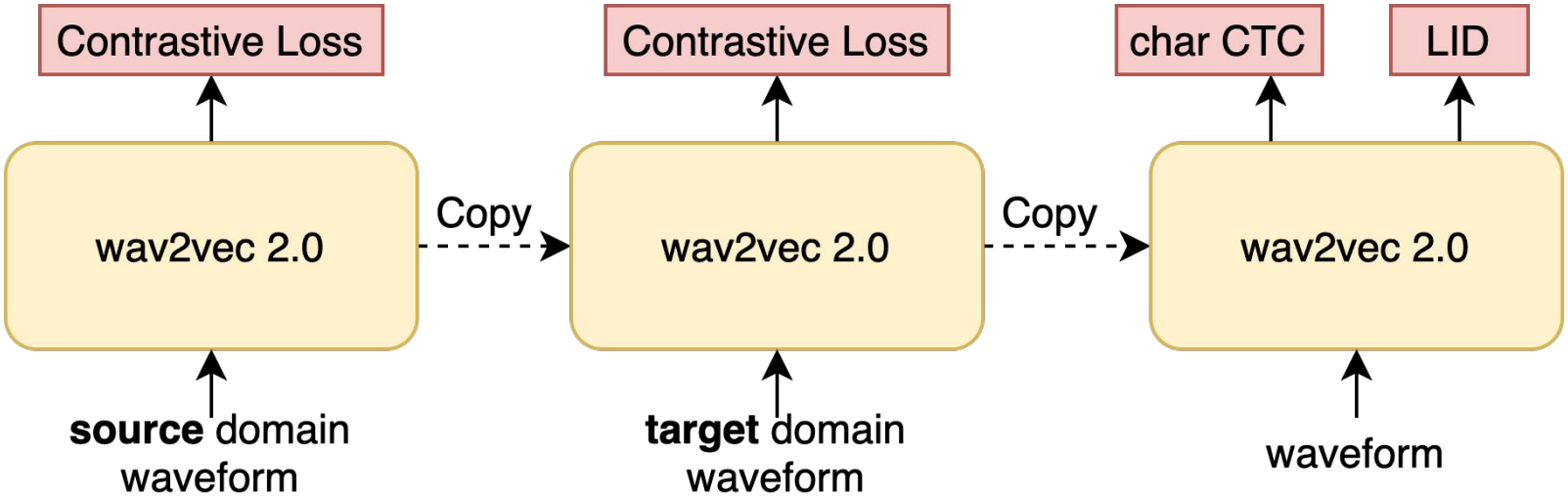
10 < snr < 20

Data Augmentation

Language	CTC+LID		MixAudio	
	Dev	**Test**	**Dev**	**Test**
Hindi	13.42	27.63	13.77	17.87
Marathi	5.09	86.64	4.69	58.78
Odiya	23.88	27.45	22.55	17.74
Tamil	16.71	33.8	17.20	30.69
Telugu	17.12	31.78	17.71	27.67
Gujarati	12.12	23.11	12.44	23.62
Average	**14.72**	38.4	**14.73**	29.39

MixAudio gives about **35% relative** Improvements for the first three languages.

Unsupervised Domain Adaptation



Unsupervised Domain Adaptation

Language	MixAudio		UDA	
	Dev	Test	Dev	Test
Hindi	13.77	17.87	12.33	16.59
Marathi	4.69	58.78	4.46	15.65
Odiya	22.55	17.74	23.07	17.81
Tamil	17.20	30.69	16.73	28.59
Telugu	17.71	27.67	17.03	25.37
Gujarati	12.44	23.62	11.92	21.3
Average	**14.73**	29.39	**14.26**	20.89

Conclusion

- We use **multitask (CTC + LID) training** to make full use of separate language modeling for multilingual ASR.
- We propose a simple data augmentation method, **MixAudio**, to improve the generalization ability of LID.
- We propose a **self-supervised domain adaptation** framework to fully leverage target domain unlabelled data.
- Combining the above techniques gives **absolute 17%** improvements over the baseline multilingual system.

Thanks

Q&A