Improving Self-supervised Pre-training using Accent-Specific Codebooks

Darshan Prabhu[†★], Abhishek Gupta[†★], Omkar Nitsure[★], Preethi Jyothi[★] and Sriram Ganapathy[✚]
[★] IIT Bombay and [✚] IISc Bangalore
[† Authors contributed equally to this work]

1 What is the central idea?

What?

Handling accents in ASR

New architecture to improve ASR adaptation to **varying accents**

Why?

High variance due to accents

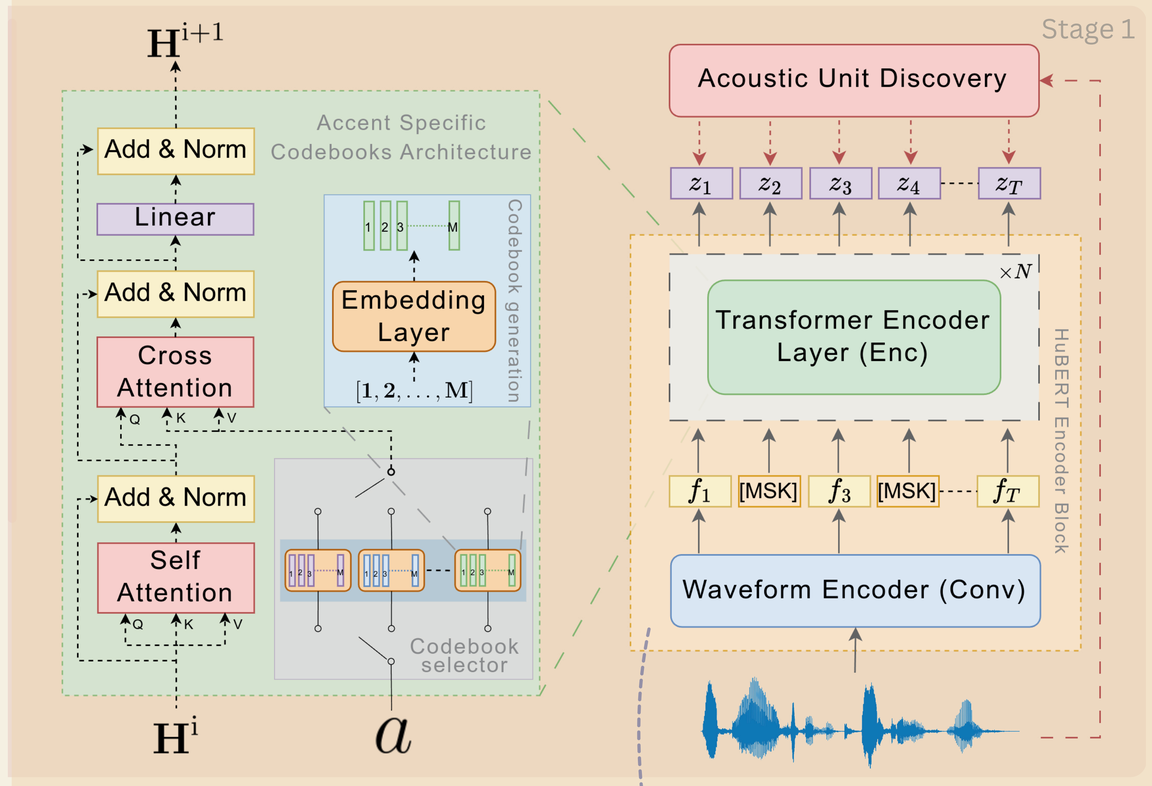
India alone has more than **19,500** languages or dialects.

How?

Codebooks and Cross Attention

Introduce **accent specific information** using codebooks and cross attention.

2 What are we proposing?



Can ASR be made Accent Aware?

3 About the Dataset

Common Voice moz://a

- Australia, Canada, Scotland, England, USA
- HongKong, India, Ireland, Africa, South Wales, Newzealand, Malaysia, Singapore, Phillippines

4 Key Results

METHOD	SIZE	OVERALL	SEEN	UNSEEN
HUBERT	104M	13.1	9.1	17.1
+ LS CKPT	104M	9.7	6.3	13.1
+ FROZEN	74M	9.3	6.0	12.5
MTL	74M	9.4	6.0	12.8
DAT	74M	9.3	6.0	12.5
OURS	76M	**8.9**	**5.9**	**11.9**

METHOD	OVERALL	ACCENTS					
		ARABIC	HINDI	KOREAN	MANDARIN	SPANISH	VIATNAMESE
HUBERT	22.6	20.2	17.8	17.3	25.8	20.4	33.7
MTL	23.0	21.0	18.1	17.6	26.4	20.9	34.1
DAT	22.9	20.7	18.2	17.4	26.2	20.9	34.1
OURS	**21.7**	**19.9**	**16.5**	**16.4**	**24.8**	**19.8**	**32.7**

Word Error Rate (WER %) comparison on L2-Arctic dataset

SOME INTERESTING FINDINGS

- **CODEBOOKS** ARE MORE **USEFUL** AT **LOWER** ENCODER LAYERS
- HAVING **TOO FEW** OR **TOO MANY** CODEBOOK ENTRIES IS **BAD**
- **FROZEN** AND **RANDOM** CODEBOOKS PERFORM **EQUALLY** WELL
- CODEBOOKS **BENEFIT** ACCENTED ASR ON BOTH **SMALL** AND **LARGER DATASETS**

5 Future Work

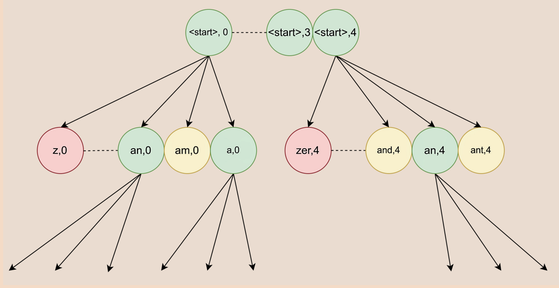
1. **CODEBOOK SIZE** — Codebook size is a **hyperparameter** that needs to be **finetuned** for each task.
2. **CODEBOOK EXPLOSION** — Employing **one codebook** per **accent** is **expensive**.
3. **MULTI-ACCENTED SEARCH** — Joint beam search allows for each utterance at test time to **commit** to a **single** seen **accent**.

Zero Shot Transfer · **Information in Codebooks** · **Ablation Study** · **Joint Beam Search** · **Overall Results**

ACCENT REMOVED	SEEN ACCENTS					UNSEEN ACCENTS								
	AUS	CAN	UK	SCT	US	AFR	HKG	IND	IRL	MAL	NWZ	PHL	SGP	WLS
OURS	3.7	7.5	5.5	4.7	5.9	10.8	14.7	11.6	11.4	16.7	9.2	14.9	18.3	7.3
AUSTRALIA	**4.9**	7.4	5.6	4.9	6.0	10.9	14.9	11.4	11.3	17.2	**9.9**	14.8	18.2	7.6
CANADA	3.7	7.6	5.5	4.8	6.1	10.7	14.8	11.6	**11.5**	16.6	9.1	14.8	18.4	7.0
ENGLAND	4.1	7.4	**5.7**	4.7	6.1	**11.0**	14.6	11.7	11.5	**16.6**	9.3	15.0	18.6	**7.7**
SCOTLAND	3.6	7.4	5.5	**5.2**	5.8	10.7	14.3	11.3	11.5	16.5	9.0	14.7	18.6	7.1
US	3.8	7.7	5.6	4.4	6.1	10.8	14.8	11.9	11.5	17.2	9.2	**14.9**	18.5	7.3
US + CANADA	3.6	**8.0**	5.6	4.8	**6.4**	10.7	**15.4**	**12.3**	11.9	17.3	9.1	**14.9**	**18.9**	7.0

*Numbers reported in this poster are Word Error Rates (WERs).

What is joint-beam search?



Poster by: **Darshan Prabhu**
Email: **darshanp@cse.iitb.ac.in**

INTERSPEECH 2024