Self-Supervised Trained (SST) Models for Multilingual ASR

Lucas Ondel, Léa-Marie Lam-Yee-Mui LISN (ex-LIMSI) "Multilingual" -> to bring speech technologies (ASR) for everyone

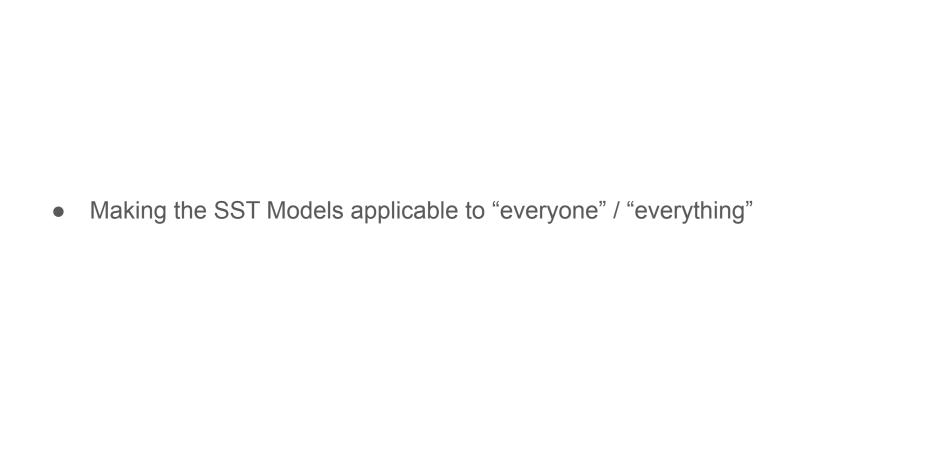
Feature	System	NMI		
		English	Mboshi	Yoruba
MFCC	k-means	31.01	30.20	28.27
	HMM	35.42 ± 0.18	37.14 ± 0.26	36.20 ± 0.31
	SHMM	38.96 ± 0.07	38.95 ± 0.60	38.98 ± 0.15
	H-SHMM	39.75 ± 0.58	42.73 ± 0.97	39.52 ± 0.46
BNF	k-means	36.63	33.79	34.33
	HMM	35.11 ± 0.49	33.39 ± 0.49	35.05 ± 0.73
	SHMM	37.99 ± 0.11	38.91 ± 0.34	40.80 ± 0.14
	H-SHMM	36.49 ± 0.96	41.32 ± 0.44	41.26 ± 0.39
HuBERT	k-means	51.63	39.35	38.93
	HMM	44.82 ± 0.36	37.77 ± 0.66	38.16 ± 0.79
	SHMM	46.97 ± 0.43	51.94 ± 0.16	46.28 ± 0.26
	H-SHMM	45.04 ± 1.23	52.44 ± 0.78	46.63 ± 0.55
XLS-R	k-means	46.76	38.65	41.44
	HMM	44.54 ± 0.62	37.79 ± 1.14	42.59 ± 0.93
	SHMM	49.86 ± 0.84	53.36 ± 0.49	53.26 ± 0.34
	H-SHMM	47.25 ± 0.96	53.97 ± 0.34	52.73 ± 0.24

But SST Models are very heavy machinery:

- costly to adapt
- costly to run

A word about myself:

- speech technologies for "everyone"
- speech decoder for on-device ASR



- SST:
 - use of FNET on top of pre-trained models
 - Can we achieve similar performances ?
 - Computational gain ?
- Multilingual ASR:
 - High-performance decoder for multi-lingual ASR / code-switching scenario
 - Semiring-Linear Algebra-based decoder
- Connecting with compressed SST model:
 - Minimalist Yet High Performing Speech Recognition System (MYHPRS)