

# 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 $\pm$ 0.18	37.14 $\pm$ 0.26	36.20 $\pm$ 0.31
	SHMM	38.96 $\pm$ 0.07	38.95 $\pm$ 0.60	38.98 $\pm$ 0.15
	H-SHMM	<b>39.75</b> $\pm$ 0.58	<b>42.73</b> $\pm$ 0.97	<b>39.52</b> $\pm$ 0.46
BNF	k-means	36.63	33.79	34.33
	HMM	35.11 $\pm$ 0.49	33.39 $\pm$ 0.49	35.05 $\pm$ 0.73
	SHMM	<b>37.99</b> $\pm$ 0.11	38.91 $\pm$ 0.34	40.80 $\pm$ 0.14
	H-SHMM	36.49 $\pm$ 0.96	<b>41.32</b> $\pm$ 0.44	<b>41.26</b> $\pm$ 0.39
HuBERT	k-means	<b>51.63</b>	39.35	38.93
	HMM	44.82 $\pm$ 0.36	37.77 $\pm$ 0.66	38.16 $\pm$ 0.79
	SHMM	46.97 $\pm$ 0.43	51.94 $\pm$ 0.16	46.28 $\pm$ 0.26
	H-SHMM	45.04 $\pm$ 1.23	<b>52.44</b> $\pm$ 0.78	<b>46.63</b> $\pm$ 0.55
XLS-R	k-means	46.76	38.65	41.44
	HMM	44.54 $\pm$ 0.62	37.79 $\pm$ 1.14	42.59 $\pm$ 0.93
	SHMM	<b>49.86</b> $\pm$ 0.84	53.36 $\pm$ 0.49	<b>53.26</b> $\pm$ 0.34
	H-SHMM	47.25 $\pm$ 0.96	<b>53.97</b> $\pm$ 0.34	52.73 $\pm$ 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

- Making the SST Models applicable to “everyone” / “everything”

- 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)