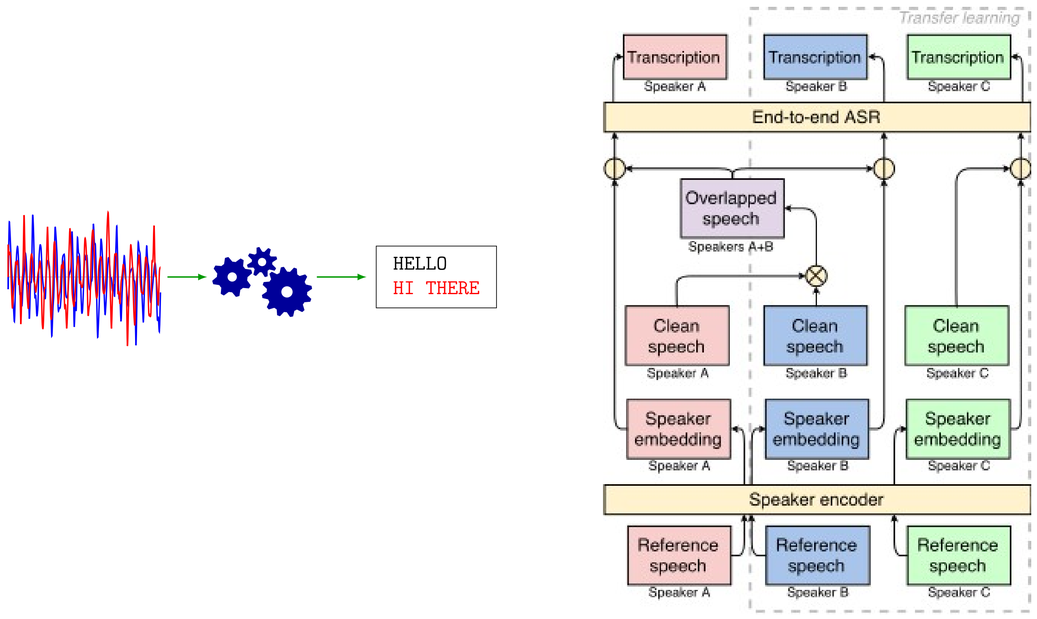
Pavel Denisov
Contact: pavel.denisov@ims.uni-stuttgart.de

TRANSFER LEARNING FOR END-TO-END SPEECH RECOGNITION AND BEYOND

Mutli-Speaker Speech Recognition [1]

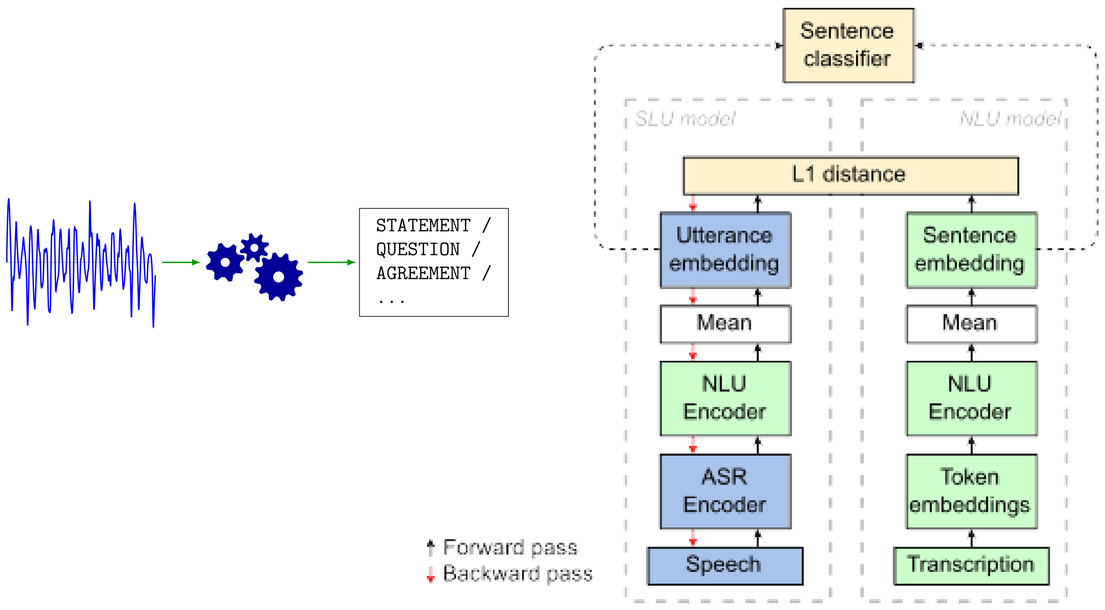


Models
- End-to-end ASR: ESPnet hybrid CTC/attention encoder-decoder
- Speaker embeddings extractor: Kaldi x-vector

Datasets
- Training and evaluation (overlapped speech): wsj0-2mix, wsj0-3mix
- Speaker embedding extraction (clean speech): WSJ0
- Additional training (clean speech): LibriSpeech 100

	WER, % (↓)			
	2 speakers		3 speakers	
	dev	eval	dev	eval
Baseline	79.6	85.7	95.9	96.0
+ speaker embeddings	11.4	22.1	95.6	95.7
+ parameters transfer	8.8	16.9	22.7	45.3
+ multi-condition training	**8.5**	**14.6**	**21.7**	**42.9**

Spoken Language Understanding : Classification [2]

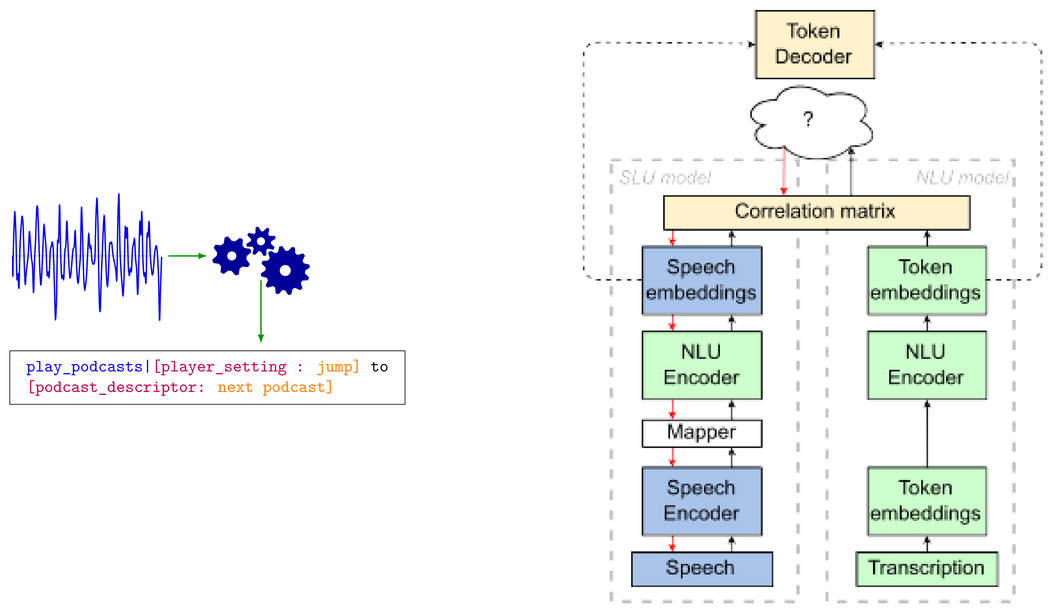


Models
- End-to-end ASR: ESPnet hybrid CTC/attention encoder-decoder
- NLU: Sentence-BERT

Datasets
- NXT-format Switchboard Corpus (SwDA)
- ICSI Meeting Recorder Dialog Act Corpus (MRDA)
- Fluent Speech Commands (FSC)

	Accuracy, % (↑)		
	SwDA	MRDA	FSC
0	58.60	60.18	91.12
4	58.71	60.47	**95.54**
10	**60.22**	**61.32**	95.49
Pipeline	*57.23*	*64.06*	*94.57*

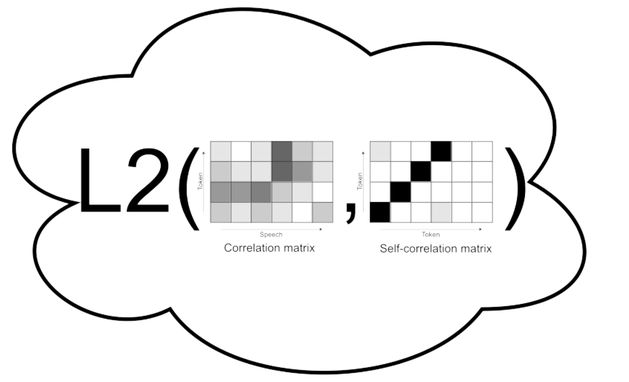
Spoken Language Understanding : Generation



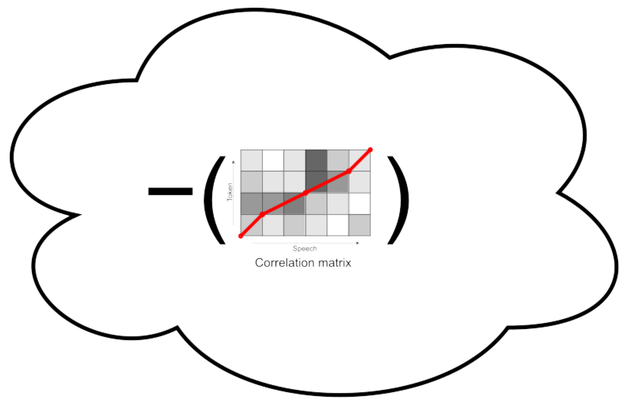
Models
- Speech Encoder: XLS-R
- Mapper: Conformer
- NLU: mBART50 Encoder-Decoder

Datasets
- English: SLURP (intent classification, slot filling), SLUE (NER)
- Mandarin: CATSLU (intent classification, slot filling)
- French: MEDIA (slot filling), PortMedia (slot filling)
- Italian: PortMedia (slot filling)

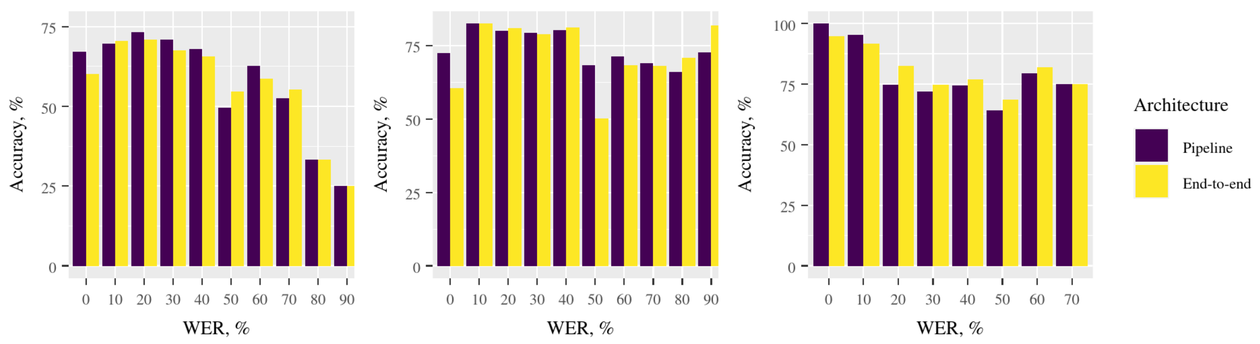
Alignment loss



OR



Spoken Language Understanding : Classification (continued) [2]



(a) SwDA (b) MRDA (c) FSC

Figure 1. Accuracy comparison for the utterances grouped by ASR WER.

References

[1] Pavel Denisov, Thang Vu. 2019. End-to-End Multi-Speaker Speech Recognition using Speaker Embeddings and Transfer Learning. In *Proceedings of Interspeech*.
[2] Pavel Denisov, Thang Vu. 2020. Pretrained Semantic Speech Embeddings for End-to-End Spoken Language Understanding via Cross-Modal Teacher-Student Learning. In *Proceedings of Interspeech*.