Universal Paralinguistic Representations of Speech Using Self-Supervised Conformers

Problem: Create representation for paralinguistic speech tasks.

Evaluation method: Evaluate 9 "eval dataset" using 5 models using linear probes (3 types) on time-averaged embeddings tasks.

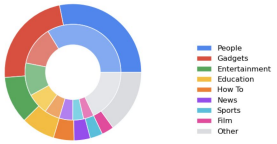
Downstream datasets:

Table 2: Downstream evaluation datasets. *Results in our study used a subset of Voxceleb filtered according to YouTube's privacy guidelines.

Dataset	Target	Classes	Samples	Avg length (s)
VoxCeleb* [32]	Speaker ID	1,251	12,052	8.4
VoxForge [33]	Language ID	6	176,438	5.8
Speech Commands[34]	Command	12	100,503	1.0
Masked Speech [19]	Mask wearing	2	36,554	1.0
ASVSpoof [20]	Synthetic or not	2	121,461	3.2
Euphonia [22]	Dysarthria	5	15,224	6.4
CREMA-D [35]	Emotion	6	7,438	2.5
IEMOCAP [21]	Emotion	4	5,531	4.5
SAVEE [36]	Emotion	7	480	3.8

Models: Trained 5 Conformer models (600M, 1B, 8B params) on the 1M hour YT-U dataset using modified Wav2Vec 2.0 loss without quantization

YT-U dataset:
1M hours of unlabeled speech



Video categories by length (outer) and number (inner)

Main results: Conformer Applied to Paralinguistics layer 12 (CAP12): 600M+ parameter Conformer architecture.

1. CAP12 outperforms previous embeddings
2. CAP12 often outperforms previous SOTA
3. A single CAP representation is near optimal for all tasks

Model	Voxceleb1†	Voxforge	Speech Commands	Masked Speech‡	ASVSpoof 2019**	Euphonia#	CREMA-D	IEMOCAP	SAVEE††
Prev SoTA	-	95.4 [37]	**97.9** [38]	73.0 [39]	5.11 [17]	45.9 [11]	74.0* [40]	67.6+ [17]	84.0* [36]
Baselines									
YAMNet++ [1]	10.9	79.8	78.5	59.7	9.23	43.0	66.4	57.5	69.2
TRILL [1]	12.6	84.5	77.6	65.2	7.46	48.1	65.7	54.3	65.0
FRILL [18]	13.8	78.8	74.4	67.2	7.45	46.6	71.3	57.6	63.3
COLA [2]	11.7	71.0	60.6	65.0	4.58	47.6	69.3	63.9	59.2
ASR Emb [11]	5.2	98.9	96.1	54.4	11.2	**54.5**	71.8	65.4	85.0
Conformers									
Best per-task§ (model, layer #)	**53.5** (XXL-YT, 25)	**99.8** (G-YT, 19)	97.5 (CAP, 16)	**74.2** (XL-LL RA, 5)	**2.5** (CAP, 12)	53.6 (CAP, 13)	87.2 (G, 26)	**79.2** (CAP, 15)	**92.5** (CAP, 15)
Best CAP per task (layer #)	50.3 (11)	99.7 (14)	97.5 (16)	73.4 (10)	**2.5** (12)	53.6 (13)	**88.2**§ (12)	**79.2** (15)	92.5 (15)
Best single layer (CAP12)	51.0§	99.7	97.0	68.9	**2.5**	51.5	**88.2**§	75.0	81.7

Additional results:

1. 3 second context windows are 99% as performant
2. Speech emotion recognition tasks require larger context
3. (CAP12 incorrect) ∩ (other embedding correct) is small
4. CKA analysis shows representation similarity **across different networks**
5. Best internal representations are between 40%-60% of the way through the network, regardless of depth