# Convexity-based Pruning of Speech Representation Models

Teresa Dorszewski, Lenka Tětková, Lars Kai Hansen

*Technical University of Denmark, Section for Cognitive Systems*

IEEE MLSP 2024 London, UK

## Motivation

How and where do speech representation models process information?

Can we prune models based on where they developed the most useful representations?

How do representations change during finetuning?

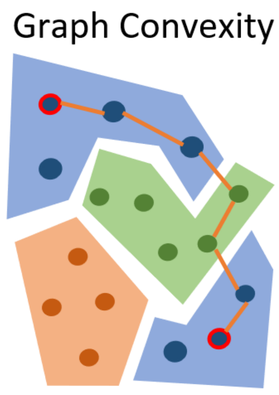
## Methods

**Models:**

Investigation of 4 self-supervised speech representation models
- Wav2vec2 [1], wavLM [2], HuBERT [3], ccc-wav2vec [4]
- All models are trained to perform word classification and speaker identification on speech commands v0.02 [5]

**Convexity:**

Evaluate graph convexity [6] of words, speakers before and after finetuning:
1. Extract latent representations of all input data
2. Build graph with nearest neighbors and distances
3. For all pairs within one class, find shortest path through neighbors
4. Avg. proportion of points in path belonging to same class is the graph convexity score

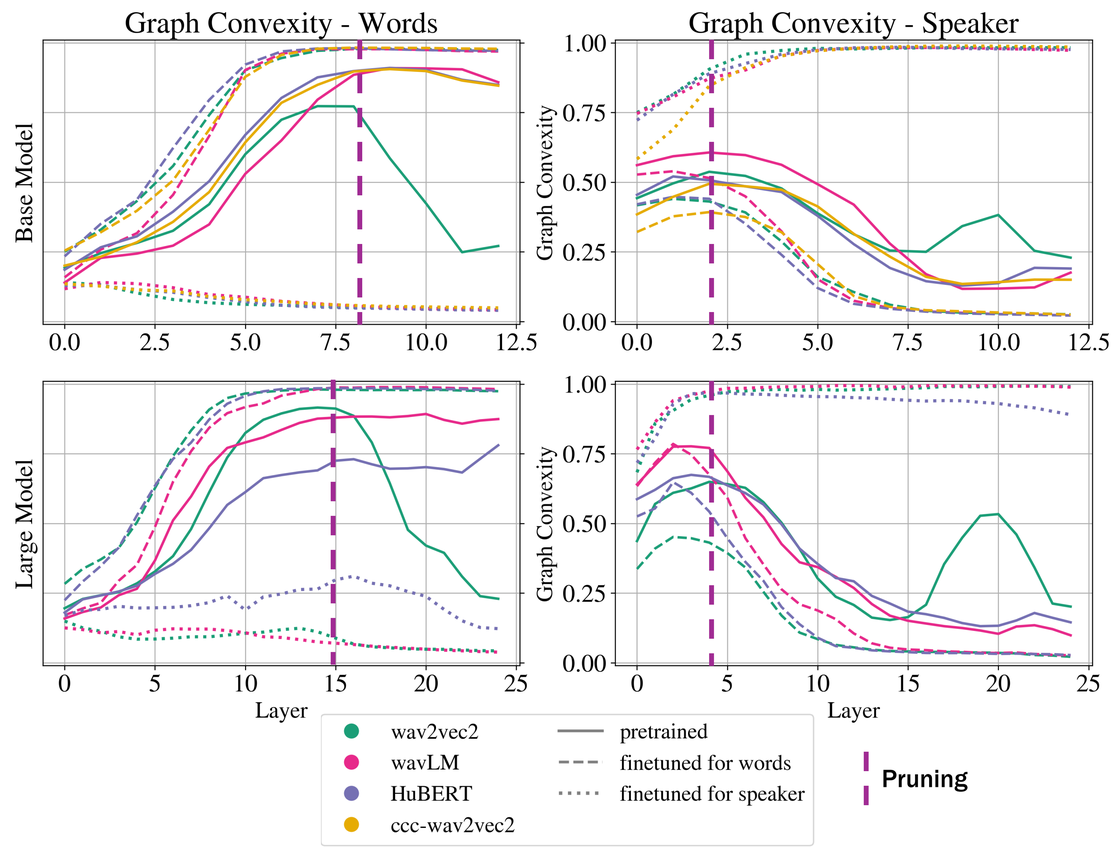



**Pruning:**
- Prune pretrained models, finetune the pruned model
- Delete all layers after the layer with highest convexity score
  - For word classification: Layer 8/15
  - For speaker identification: Layer 2/4

## Results

Models encode information in different layers:
- Word meaning in later layers
- Speaker information in early layers

Finetuning enhances convexity of important information and reduces convexity for less important concepts




Finetuning after convexity-informed pruning leads to comparable accuracy while significantly reducing training and inference time

| | Word classification | Speaker identification |
|---|---|---|
| Change in Accuracy | - 0.25% | + 1.22% |
| Change in Training time | - 23.4% | - 47.8% |
| Change in Inference time | - 25.9% | - 58.7% |

## Conclusion

Models process speaker information early and word content in the middle

Informed pruning can reduce computational resources while maintaining performance

References:
[1] Alexei Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations” Advances in neural information processing systems, vol. 33, pp. 12449–12460, 2020.
[2] Sanyuan Chen et al., “Wavlm: Large-scale self-supervised pretraining for full stack speech processing” IEEE Journal of Selected Topics in Signal Processing, 2022.
[3] Wei-Ning Hsu et al., “Hubert: Self-supervised speech rep-resentation learning by masked prediction of hidden units” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
[4] Vasista Lodagala et al.,“Ccc-wav2vec 2.0: Clustering aided cross contrastive self-supervised learning of speech representations” in 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2022.
[5] Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209,* 2018.
[6] Lenka Tetkova et al., “On convex conceptual regions in deep network representations” arXiv preprint arXiv:2305.17154, 2023.

Check out the paper

**Teresa Dorszewski**
DTU Compute, Section for Cognitive Systems

Teresa Dorszewski

tksc@dtu.ck