3-1-25-SUP

SLT 2022

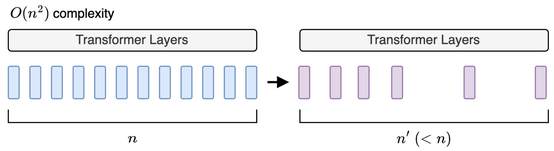
On Compressing Sequences for Self-Supervised Speech Models

Yen Meng[1*], Hsuan-Jui Chen[1*], Jiatong Shi[2], Shinji Watanabe[2], Paola Garcia[3], Hung-yi Lee[1], Hao Tang[4]

[1]National Taiwan University, [2]Carnegie Mellon University, [3]Johns Hopkins University, [4]The University of Edinburgh

Introduction

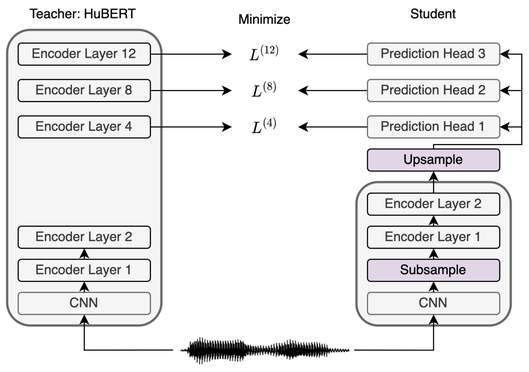
1. Reduce the computation cost in self-supervised speech models via compressing sequences.
2. We propose to use variable-length subsampling for self-supervised speech models.



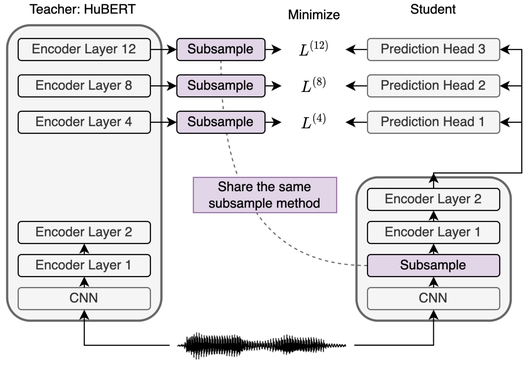
Experiment Framework

We add a subsampling layer in the DistilHuBERT [Chang et al., 2022] with the two following settings:

With Upsample



Subsample Targets



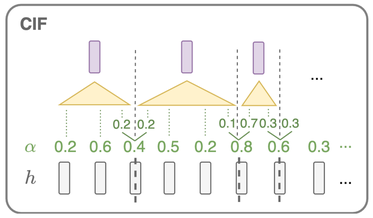
Compressing Sequences with Subsampling

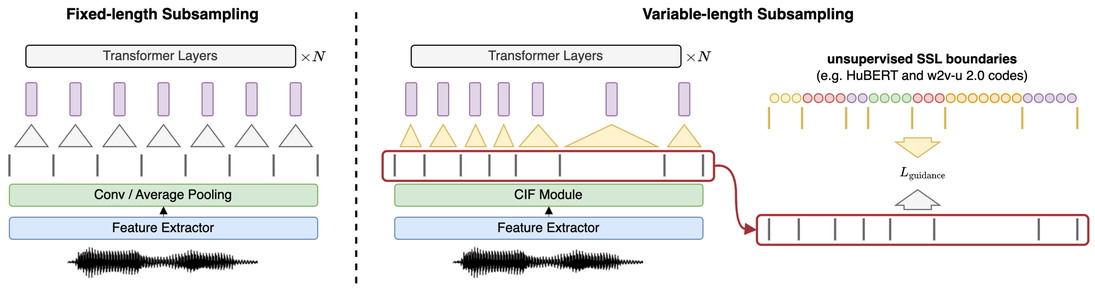
Fixed-length Subsampling

- Naive approach, using convolution or average pooling.

Variable-length Subsampling

- Incorporates the idea of Continuous Integrate-and-Fire (CIF) [Dong et al., 2020].
- Additional segmentation guidance with pre-extracted boundaries.





Main Results

Evaluation on downstream tasks with the **Subsample Targets** setting. We experiment with segmentation guidance using smoothed HuBERT codes and unsupervised ASR boundaries.

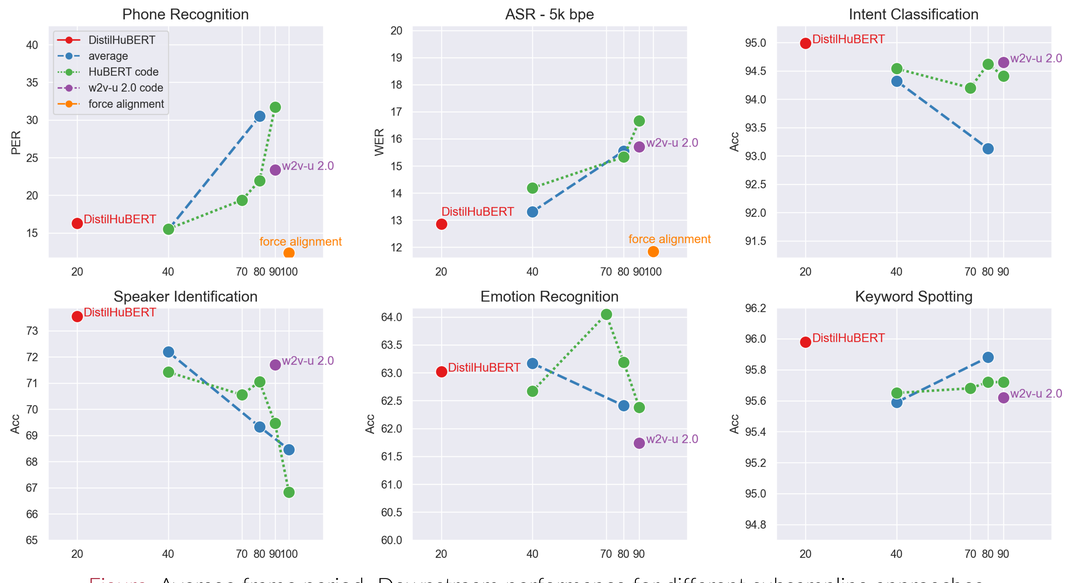


Figure: Average frame period. Downstream performance for different subsampling approaches.

Discussions

Downstream Performance

1. The variable-length subsampling recovers the performance for PR and ASR at around the phone duration (80-90ms).
2. The utterance-level tasks are less affected by subsampling.
3. Using the unsupervised ASR as guidance gives a better performance-efficiency trade-off.

Runtime Efficiency

We report the average multiply-accumulate operations (MACs). The reduction in MACs is consistent with the sequence length.

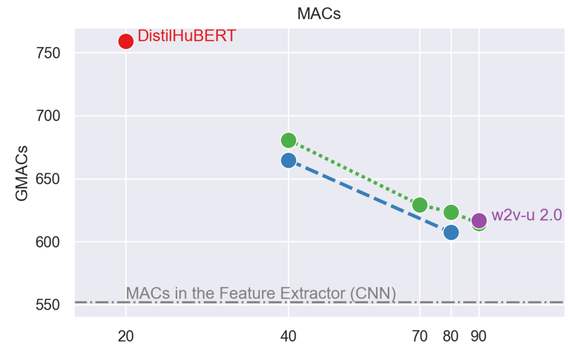


Figure: Average frame period vs. MACs.

Conclusion

1. Different tasks have different preferred frame rates.
2. Our proposed variable-length subsampling works particularly well under low frame rates.

Acknowledgments

This work was supported by JSALT 2022 at JHU, with gift-funds from Amazon, Microsoft, and Google. We thank the Taiwan Web Service and the National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computing and storage resources.