Improved Masking Strategies for Self-Supervised Speech Representation Learning

Anuj Diwan, David Harwath, Eunsol Choi

Department of Computer Science,
University of Texas at Austin

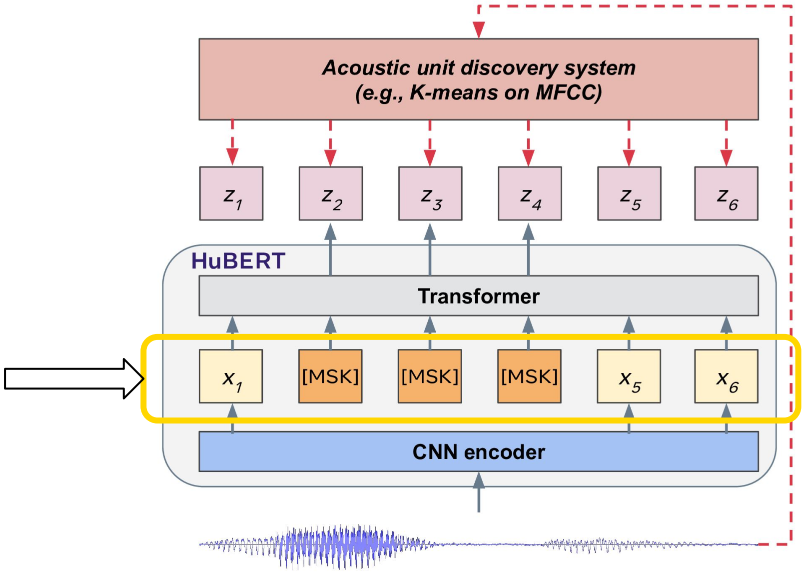
Background

- **Self-supervised** speech representation learning involves training a model with lots of unlabeled speech data to generate powerful latent representations
- Recent approaches like **wav2vec2.0** [1] and **HuBERT** [2] have achieved state-of-the-art performance on many downstream tasks like Librispeech [3] ASR, Keyword Spotting, Phoneme Recognition, and more tasks from the SUPERB [4] benchmark
- In this work, we focus on HuBERT and Masked Language Modelling approaches for speech.

HuBERT

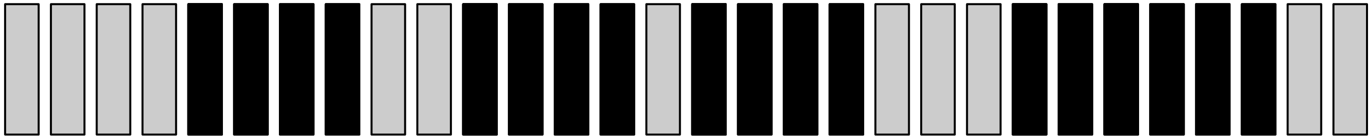
- Involves **masking** a random subset of the input (speech frames) sequence and tasking the model with **predicting** a *discretized* version of the masked input
- Similar approach to Masked Language Models like BERT

Our focus



RandomFrameSpan Masking

- We can call the masking strategy used by HuBERT as *RandomFrameSpan Masking*
- This masking strategy is fairly simple:
  - randomly sample a **proportion *p*** of all input timesteps to be **start indices** of masking spans
  - mask a **fixed number M** of timesteps starting from each start index
  - In HuBERT, p = 0.08, M = 10



n=30
p=1/6
M=4

Better Masking Strategies in textual NLP

- **Random-Token Masking**: Original BERT; pick subword tokens randomly
- **Random-Span Masking**: Sample a span length and span start index randomly, mask entire span (consisting of multiple words)
- **Knowledge Masking** and **Salient Span Masking**: Use parsers to identify meaningful entities/phrases and mask these.

Can we propose improved masking strategies for speech as well?

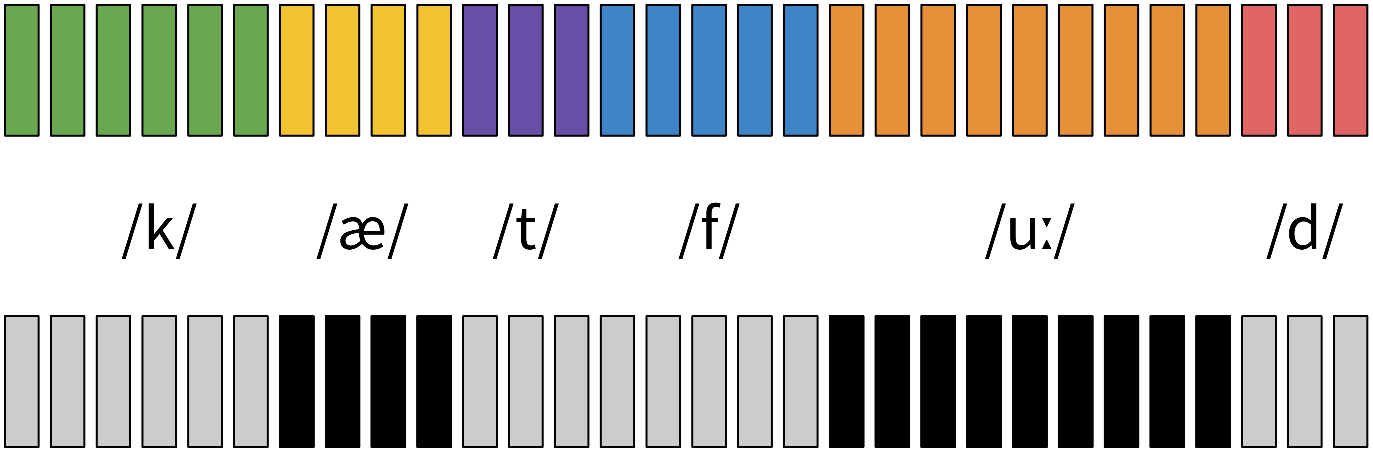
RandomPhoneme Masking

- We propose a **new masking strategy** where we mask **spans of entire phonemes**
- Given a list of phoneme boundaries for each utterance i.e. the start and end frame indices for each phoneme in the utterance.
    - fix a **proportion *q*** of total frames we intend to mask and a phoneme **span length *m***
    - randomly sample a phoneme index i and mask out all phonemes from i to i+m-1. Repeatedly do this until the total frames masked hit the proportion *q*
    - *q*=0.56 is chosen such that the number of frames masked is approximately the same as RandomFrameSpan masking
- This is a more linguistically-driven masking strategy that intuitively should result in a harder pretraining loss

How to obtain a list of phoneme boundaries for unlabelled data?

Phoneme Segmentation

- The phoneme segmentation task involves segmenting a given input speech utterance into its constituent phonemes i.e. outputting a sequence of phoneme boundaries



Fully Unsupervised Phoneme Segmentation

- Since we are developing a masking strategy for SSL, we cannot assume access to ground truth text data
- We use an unsupervised phoneme segmentation strategy proposed by Kreuk et. al. [6]. This approach neither requires labels for training nor for inference
- The algorithm uses contrastive loss to learn latent frame representations that distinguish adjacent frame pairs from non-adjacent frame pairs.
- During inference, a similarity score is computed for each adjacent frame pair and the lowest similarity score pairs are identified as phoneme boundaries

Supervised Phoneme Segmentation

- If one has access to the *ground truth text* of the speech utterance as well as a phonetic dictionary that maps words to phonemes (like in the traditional ASR pipeline), one can easily extract phoneme boundaries by:
  - Training a traditional HMM-based HCLG ASR system on labelled speech data
  - For a given speech utterance, run forced Viterbi alignment using the trained model on the speech and the corresponding ground truth phoneme sequence. This will time-align the ground truth phoneme sequence, giving phoneme boundaries as desired
- This requires access to ground truth text data for both the training set used to train the ASR system and the test set whose boundaries need to be found
- Kaldi [5] has off-the-shelf scripts to run forced Viterbi alignment
- This is an 'oracle' experiment; uses near-perfect phoneme boundaries

Evaluation

- The SUPERB [4] benchmark consists of a set of downstream speech tasks that can be used to evaluate pretrained speech models
- We focus on ASR (Automatic Speech Recognition) , PR (Phoneme Recognition), KS (Keyword Spotting). The metrics are WER (lower is better), PER (lower is better), Accuracy (higher is better) respectively.
- For each task, the model parameters are frozen. Then,
  - All layers of the model are summed (with learnable weights for each layer) to generate the final representation
  - A task-specific head is placed on top of this representation for generating the task output.

Experimental Setup - Pretraining and Eval

- **Datasets**
  - We use the 960-hr Librispeech data for pretraining the model
- **Model**
  - We use the HuBERT Base model for all our experiments. We train our k-means clusters using the 6th layer of the Facebook-provided pretrained checkpoint.
  - We modify the HuBERT dataloader and training code from fairseq [8] to support phoneme-based masking.
  - We initialize our pretraining expts using the Facebook-provided pretrained checkpoint rather than pretraining from scratch. We train for 40k additional steps.
- **Evaluation**
  - We use the provided SUPERB scripts in the s3prl toolkit.

Experimental Setup - Phoneme Segm.

- Unsupervised Approach
  - Off-the-shelf phoneme segmentation model released by Kreuk. et. al. trained on Buckeye [7] corpus and train-other-500 set of Librispeech. We run the entire Librispeech corpus through the model to generate phoneme segmentations
- Supervised Approach
  - We train a TDNN-HMM ASR model using the standard Kaldi Librispeech ASR recipe on 960-hr Librispeech train set
  - We run forced Viterbi alignment using the `align_fmllr.sh` Kaldi script on the Librispeech corpus

Experimental Results

Masking Strategy	Span Length	Boundary Type	ASR (WER)	PR (PER)	KS (Acc)
None (original ckpt)	-	-	7.09	6.10	96.55
RandFrameSpan	-	-	7.18	5.61	96.62
RandPhoneme	1	Supervised	7.13	5.57	96.72
	2	Supervised	7.11	5.53	**96.88**
		Unsupervised	**7.05**	**5.51**	96.52

Future Work

- Training for more steps to (hopefully) demonstrate larger gains
- Data-driven analysis (like PMI Masking) to find spans that are potentially even more meaningful than phonemes/resemble phonemes
- Reducing dependency on external tools like phoneme segmentation using the above
- Using phoneme-based ideas to change the discretization strategy itself (force all frames within a phoneme to have the same discrete index, for example)

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, & Michael Auli. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, & Abdelrahman Mohamed. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.

[3] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5206-5210).

[4] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, & Hung-yi Lee. (2021). SUPERB: Speech processing Universal PERformance Benchmark.

[5] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.

References

[6] Felix Kreuk, Joseph Keshet, & Yossi Adi. (2020). Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation.

[7] Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication, 45(1):89–95, 2005. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2004.09.001.

[8] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, & Michael Auli. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling.