P6-13-SLP

Self-Supervised Syllable Discovery Based on Speaker-Disentangled HuBERT

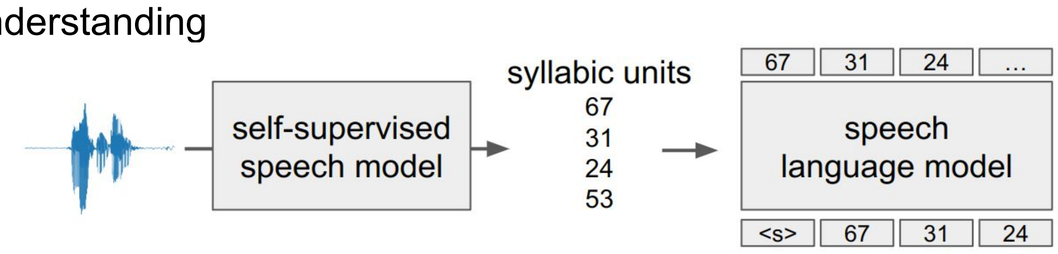
Ryota Komatsu
Independent Researcher

Takahiro Shinozaki
Tokyo Institute of Technology

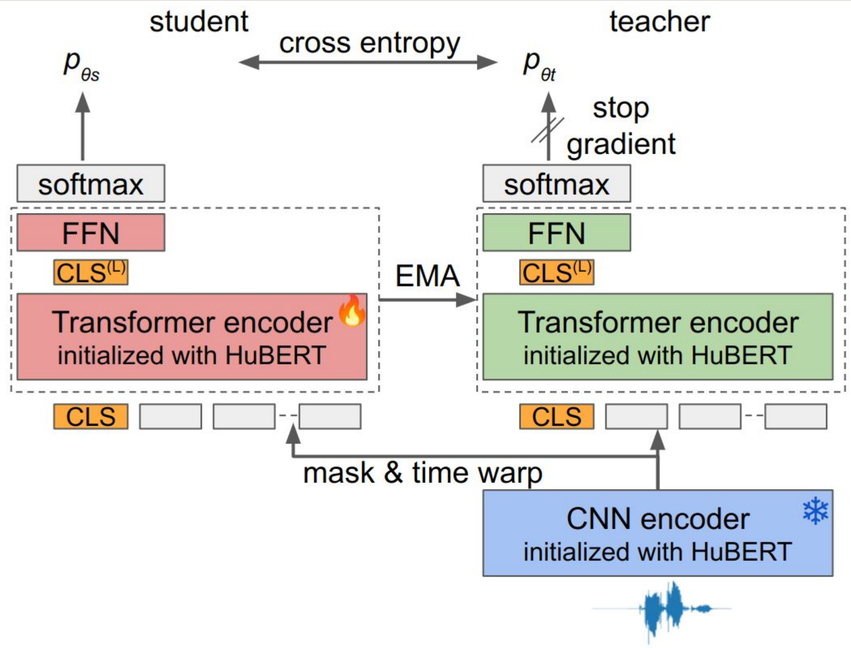
codes & models

1. Motivation

- Self-supervised learning (SSL) of speech representations has become essential for extracting meaningful features from raw audio
- Hidden units obtained by discretizing learned representations highly correlate with linguistic units, e.g., phones, syllables, and words
- By utilizing them as pseudo-transcripts for raw audio, we can develop textless models, including speech language models
- Compared to phonetic units, coarse-grained syllabic units have an advantage in token frequency and potentially enhance semantic understanding



2. Baseline method: Self-Distillation HuBERT [3]



Main points

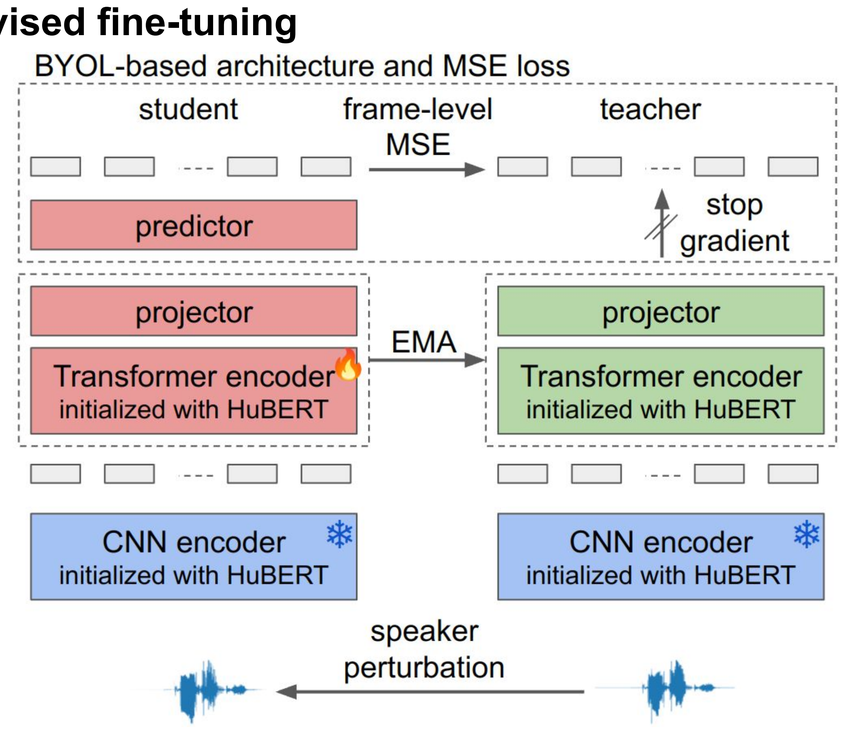
- Syllabic organization naturally emerges within outputs from the student's intermediate (9-th) Transformer layer through sentence-level self-distillation (DINO) fine-tuning of the pretrained HuBERT
- Sentence-level representation is aggregated through self-attention layers using a special CLS token concatenated with the input speech feature sequence

Problem

- Given the speaker ID X and the student's final softmax category Y, we observed that 61% of the entropy (uncertainty) of X was reduced after observing Y on the Librispeech test set, indicating that the model learns to predict **speaker identity** rather than **linguistic content**

3. Proposed method

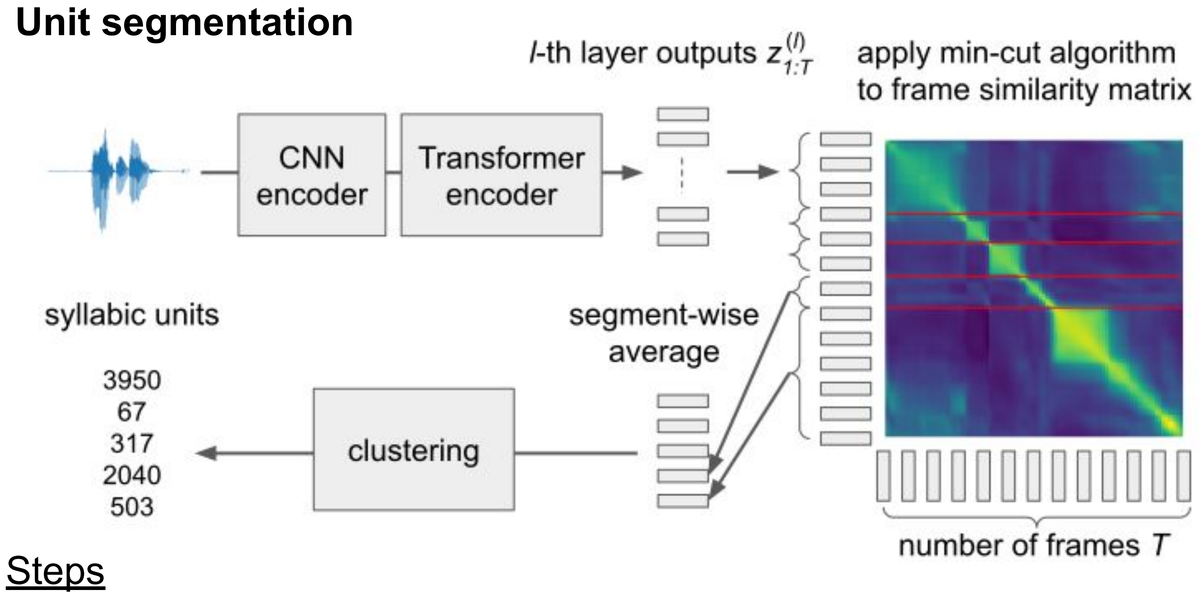
Self-supervised fine-tuning



Key idea

- Following **BYOL**, a SSL framework, we adopt the MSE loss and incorporate a predictor in the student. This approach has shown superior performance compared to DINO in image segmentation
- To prevent the model from learning speaker identity with the CLS token, we remove it and compute the loss at the **frame-level**
- We constrain the model to extract consistent features between the original speech and its **speaker-perturbed** version

Unit segmentation

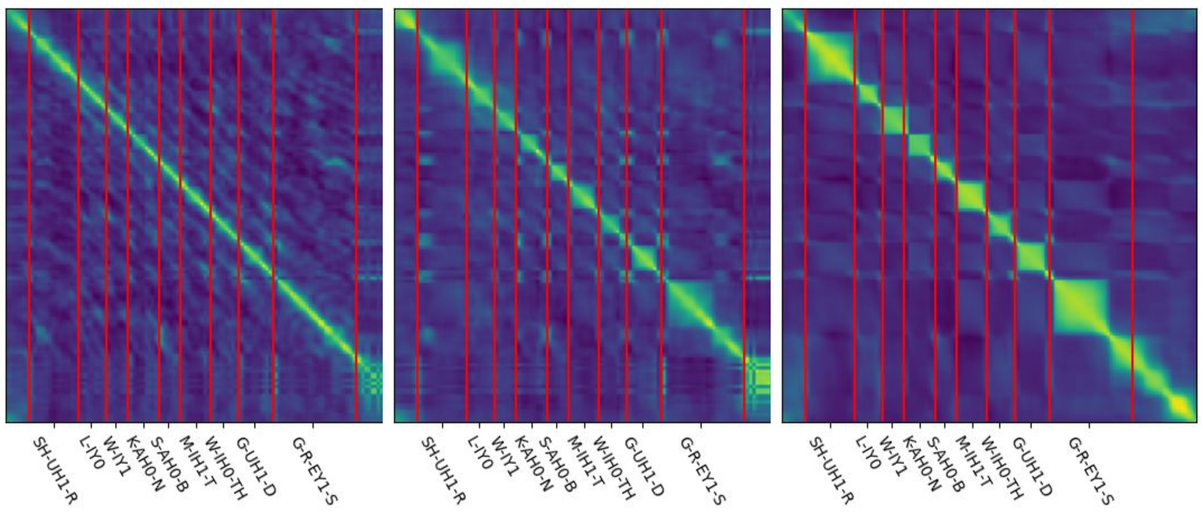


Steps

1. Obtain outputs from the 8-th Transformer layer of the student
2. Create the frame similarity matrix
3. Apply the minimum cut algorithm to the frame similarity matrix
4. Average pooling within segments
5. Two-step clustering on averaged representations

4. Results

Example of frame similarity matrices



HuBERT [1] | Self-Distillation HuBERT [3] | Ours

Red lines indicate reference syllable boundary

Results of syllable discovery and speaker identification

Model	Syllable segmentation				Syllabic unit quality			Speaker Identification
	Precision	Recall	F1	R-value	Syllable purity	Cluster purity	Mutual info.	Accuracy↓
HuBERT[1]	51.4	31.4	39.0	50.1	33.1	28.4	3.54	67.2
VG-HuBERT[2]	65.3	64.3	64.8	70.0	53.4	43.6	4.66	37.4
Self-Distillation HuBERT[3]	64.3	71.0	67.5	70.7	54.1	**46.2**	4.76	47.6
Ours	**73.3**	67.6	70.3	74.6	**59.4**	44.5	**5.08**	**26.6**
ablation study								
Ours -frame-wise BYOL +frame-wise DINO	64.3	65.1	64.7	69.8	59.1	42.9	5.06	32.8
Ours -frame MSE +CLS MSE	70.0	**73.8**	**71.9**	**75.5**	55.7	45.7	4.91	28.9

Findings

- In HuBERT, frame similarity is limited to short spans
- Self-Distillation HuBERT obtains larger structures, but their representations are relatively speaker dependent
- Our block structures are the clearest, and their boundaries roughly match the references
- In our DINO variant, segmentation scores dropped, likely because the classes activated in the softmax were fewer than the number of syllables
- With a sentence-level MSE using the CLS token, the speaker dependence of speech features slightly increased, and the syllabic unit quality degraded
- Overall, our proposed method performs the best, validating the efficacy of speaker disentanglement in syllable discovery

[1] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," TASLP, vol. 29, pp. 3451–3460, 2021.
[2] P. Peng *et al.*, "Syllable discovery and cross-lingual generalization in a visually grounded, self-supervised speech model," in Proc. Interspeech, 2023, pp. 391–395.
[3] C. J. Cho *et al.*, "SD-HuBERT: Sentence-Level Self-Distillation Induces Syllabic Organization in Hubert," in Proc. ICASSP, 2024.