Integrating Self-Supervised Speech Model with Pseudo Word-Level Targets from Visually-Grounded Speech Model

Hung-Chieh Fang [1*] Nai-Xuan Ye [1*] Yi-Jen Shih [1,2] Puyuan Peng [2]
Hsuan-Fu Wang [1] Layne Berry [2] Hung-yi Lee [1] David Harwath [2]

[1]National Taiwan University, Taiwan [2]The University of Texas at Austin, USA

Motivation & Background

Motivation

- HuBERT [1] is trained with frame-level units as targets, limiting their performance in general spoken language understanding (SLU) tasks.
- Previous approaches use ASR + NLU (traditional) or end-to-end NLU [2] frameworks to alleviate the issue, but these require expensive paired transcripts.

Background: VG-HuBERT [3]

- It is a visually-grounded speech (VGS) model that reaches state-of-the-art performance in word segmentation tasks.
- It is trained with image-speech pairs, eliminating the dependence on paired transcripts.

Can we enrich the semantic information of self-supervised speech models by incorporating word-level units without the need for paired transcripts?

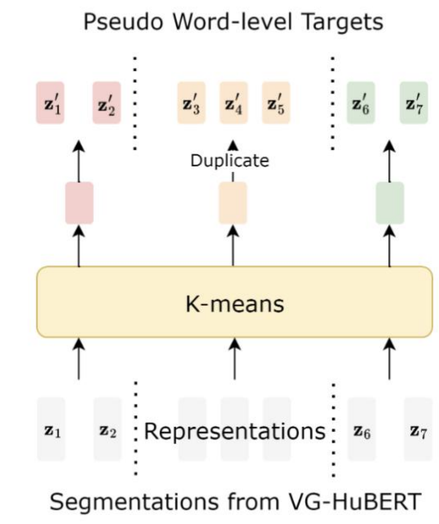
Methodology

Key Idea

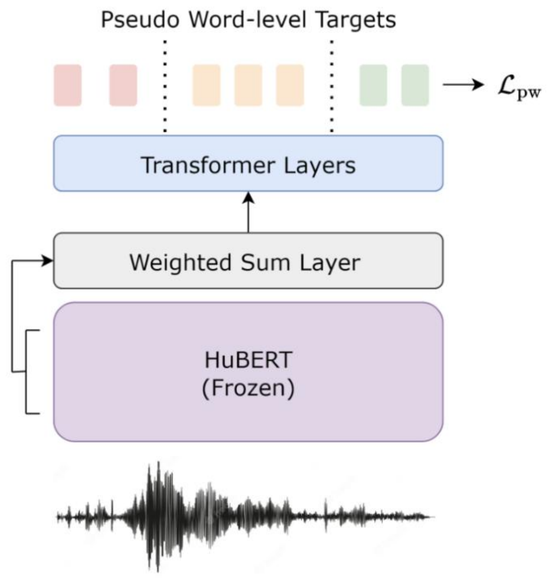
- Utilize word boundaries from VG-HuBERT [3] to provide word-level supervision without paired speech-text data.
- Incorporate frame-level and word-level targets to further enhance semantic information.

Method

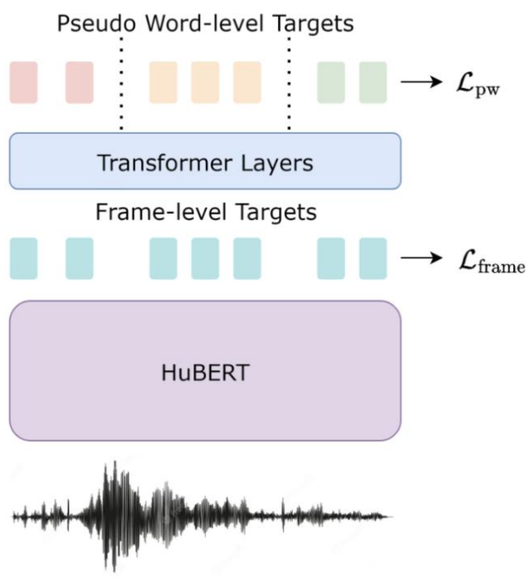
- Aggregate representations within word boundaries from VG-HuBERT [3] to generate pseudo word-level targets.
- Single PW-HuBERT fine-tunes HuBERT with pseudo word-level targets.
- Hierarchical PW-HuBERT employs frame-level targets in shallower layers and pseudo word-level targets in deeper layers, based on the insight from [4] that higher-level information resides in deeper layers.



(a) Pseudo word-level target generation



(b) Single PW-HuBERT



(c) Hierarchical PW-HuBERT

Results

Table 1. HuBERT$_{14}$ is a 14-layer version of HuBERT with the same number of parameters as PW-HuBERT.

Dataset	SLUE		SLUE Phase-2	SNIPS		ZeroSpeech	
	SA	NER	NEL	SF	IC	Lib.	Syn.
Metric	F1 ↑	F1 / Label F1 ↑	Frame F1 / Word F1 ↑	F1 ↑		Similarity ↑	
HuBERT	45.27	51.6 / 64.8	57.54 / 61.14	88.16	98.57	5.71	6.79
HuBERT$_{14}$	44.54	51 / 66.8	58.43 / 61.84	88.18	<u>98.71</u>	5.11	6.63
VG-HuBERT	45.1	41.4 / 52.3	47.11 / 51.52	84.98	98.42	**8.42**†	**9.97**†
Single PW-HuBERT	<u>48.7</u>	<u>52.5</u> / <u>67.3</u>	<u>59.44</u> / <u>63.51</u>	**88.32**	98.44	5.16	6.88
Hierarchical PW-HuBERT	**49.06**	**55.3 / 68.6**	**61.28 / 65.55**	<u>88.25</u>	**98.85**	<u>6.55</u>	<u>9.02</u>

Observation

- Both Single and Hierarchical PW-HuBERT outperform baselines on general SLU tasks, showing the benefits of word-level targets.
- VG-HuBERT shows improvement in ZeroSpeech due to its similar training setup†, but it adversely affects general SLU tasks.

Analysis

Architecture	HuBERT Targets	SLUE SA	SLUE NER	SLUE 2 NEL	SNIPS SF
Hierarchical	✗	44.94	53.6/67.9	59.4	**88.4**
	✓	**49.06**	**55.3/68.6**	**61.3**	88.3

The Effect of Frame-Level Targets

- Hierarchical PW-HuBERT with frame-level targets consistently outperforms its counterpart without.
- Exploiting the synergy between frame-level and word-level targets further enhances training guidance.

Conclusion

- Propose a framework that incorporates pseudo word-level targets from a VGS model into training without the necessity of paired speech-text data.
- Demonstrate the advantage of jointly training with frame-level and word-level targets.

References

[1] Wei-Ning Hsu et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460.

[2] Siddhant Arora et al. "Espnet-slu: Advancing spoken language understanding through espnet". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7167–7171.

[3] Puyuan Peng and David Harwath. "Word Discovery in Visually Grounded, Self-Supervised Speech Models". In: *Interspeech*. 2022.

[4] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. "Layer-wise analysis of a self-supervised speech representation model". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2021, pp. 914–921.