HUBERTOPIC: ENHANCING SEMANTIC REPRESENTATION OF HUBERT THROUGH SELF-SUPERVISION UTILIZING TOPIC MODEL

Takashi Maekaku[1], Jiatong Shi[2], Xuankai Chang[2], Yuya Fujita[1], Shinji Watanabe[2],
[1]LY Corporation, [2]Carnegie Mellon University

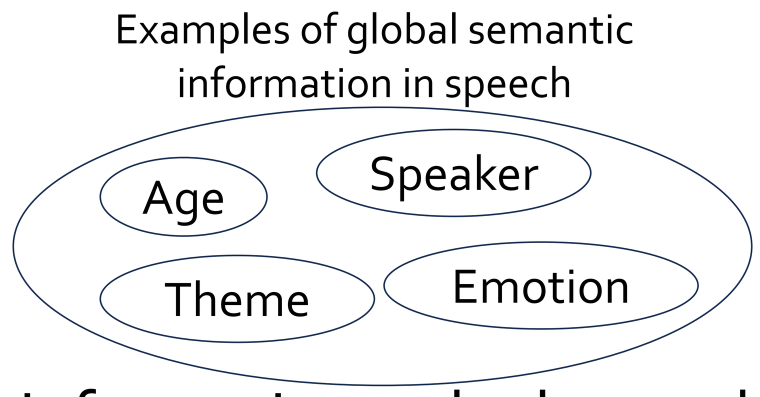
Introduction

Implicit Challenge:

- HuBERT's masked prediction task may not effectively utilize global semantic information

Proposed Solution:

- Enhance HuBERT's representation by utilizing topic labels generated by LDA
- Incorporate a topic classification task into HuBERT, which allows additional global semantic information to be learned



System Description

HuBERT

- Employ the following masked prediction loss

$$\mathcal{L}_{\text{MP}} = -\sum_{d}\sum_{t\in\mathcal{M}^d} \log p_f(z_t^d|\tilde{X}^d)$$

d: Utterance index
t: Timestep of acoustic feature $\tilde{X}$
z: Pseudo label

Proposed Method (HuBERTopic)

- Apply LDA to pseudo-labels to obtain per-utterance topic distributions
- For each d, assign the topic with the highest contribution in its distribution
- Add a topic label classification task to HuBERT

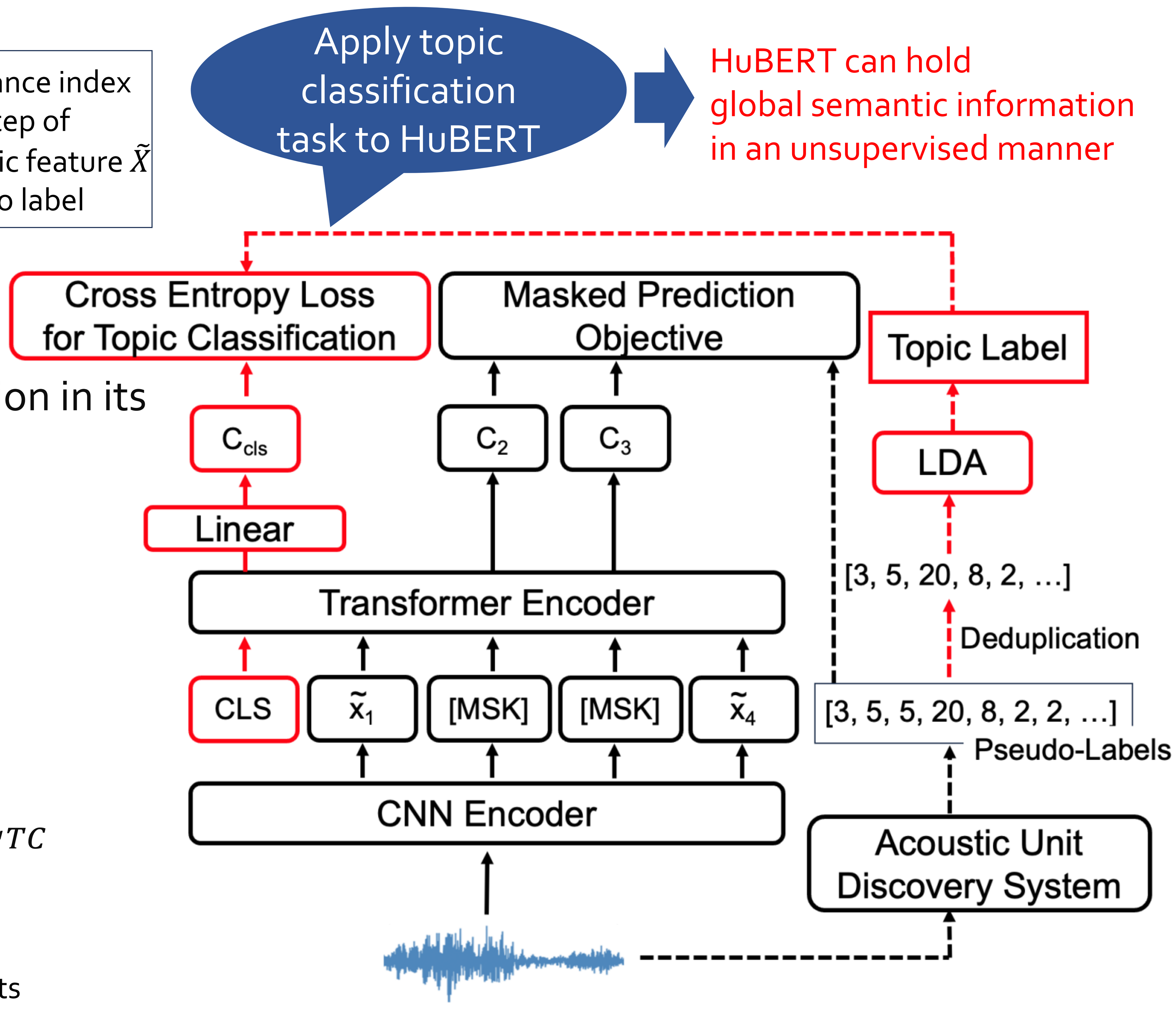
$$\mathcal{L}_{\text{TC}} = -\sum_{d}\sum_{k} \tau_k^d \log(\text{softmax}(c_{\text{cls}}{}^d)_k)$$

k: Index of topic dimension K
$c_{\text{cls}} \in \mathbb{R}^K$, τ_k^d : One-hot representation of topic label

- Total loss is calculated as a weighted sum of L_{MP} and L_{TC}

$$\mathcal{L} = (1-\rho)\mathcal{L}_{\text{MP}} + \rho\mathcal{L}_{\text{TC}}$$

→ ρ was set to 0.01 in the following experiments



Experiments

Results

- ASR (Fig.1)
  - The performance of HuBERTopic outperformed the baseline
    - 👉 The topic classification task enhanced semantic information useful for the ASR task
  - Improvement was less significant in the 960h scenario
    - 👉 Tuning of K and the relative benefits of the auxiliary task versus more data need further investigation
- SUPERB (Fig.2)
  - HuBERTopic shows overall improvement, with notable gains in PR and SD
    - 👉 Likely due to enhanced phonetic and speaker discrimination

Training Data (SSL)	Model	K(Num of topics) (0itr/1itr)	WER(↓)			
			dev-clean	dev-other	test-clean	test-other
LS-100h	HuBERT	-	17.1	33.5	17.3	35.3
	HuBERTopic	30/200	**16.1**	**32.9**	**16.6**	**34.1**
LS-960h	HuBERT	-	7.4	14.2	7.4	14.2
	HuBERTopic	30/30	**7.2**	**14.1**	7.4	**13.7**

Fig. 1 ASR results

Training Data (SSL)	Model	K (0itr/1itr)	PR(↓)	ER(↑)	IC(↑)	SID(↑)	SD(↓)	SF(↑)	KS(↑)	SE(↑)
LS-100h	HuBERT	-	13.89	60.24	88.72	60.48	8.86	80.62	94.22	2.48
	HuBERTopic	30/30	**12.97**	**60.92**	**90.64**	**61.82**	**8.59**	**81.05**	**94.87**	**2.50**
LS-960h	HuBERT	-	5.04	**64.12**	97.57	**79.34**	7.49	88.61	**96.04**	2.53
	HuBERTopic	30/30	**4.83**	64.10	97.68	78.98	**6.93**	88.76	95.26	2.53
	HuBERTopic	150/1000	4.84	63.61	**98.10**	79.21	7.07	**88.79**	95.81	**2.55**

Fig. 2 SUPERB results

Topic Analysis

- Calculate purity scores with various attributes (Fig.3)

$$\text{Purity}(\Omega,\Lambda) = \frac{1}{N}\sum_{k}\max_{j}|\omega_k \cap \lambda_j|$$

Number of data in ω_k most frequently assigned to λ_j

$\Omega = \{\omega_1, \omega_2, \cdots, \omega_K\}$: a set of attribute labels

$\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_J\}$: a set of topic labels

- HuBERTopic yields higher scores than random cases
  - 👉 Indicate that topic label contains these semantic information

Attribute	K	Purity	
		Proposed	Random
Gender	2	0.978	0.503
Speaker	30	0.075	0.011
Book	30	0.081	0.024
Chapter	30	0.061	0.009

Fig. 3 Purity between the topic label and each attribute label