Wav2vec-Switch: Contrastive Learning from Original-noisy Speech Pairs for Robust Speech Recogn

Yiming Wang Jinyu Li Heming Wang Yao Qian Chengyi Wang Yu Wu

Microsoft Corporation

Highlights

- We propose a method to encode noise robustness into contextualized representations of speech via contrastive learning.
- The quantized representations of the original-noisy speech pair are switched as additional prediction targets of each other. By doing this, it enforces the network to have consistent predictions for the original and noisy speech, thus allows to learn contextualized representation with noise robustness.
- 2.9–4.9% relative WER reduction on the synthesized noisy LibriSpeech data, compared to a data augmentation baseline.
- No deterioration on the original clean data.
- 5.7% relative WER reduction on CHiME-4 real 1-channel noisy data. Matches or even surpasses those with well-designed speech enhancement components.

Motivation

- Self-supervised Learning (SSL) has been shown promising for low-resource ASR.
- Noise robustness is another challenge for ASR. No much work for improving noise robustness for SSL in ASR.
- Most existing noise robustness approaches add a dedicated enhancement/denoising frond-end, increasing model complexity.
- In wav2vec 2.0, contextualized representation is learned by predicting quantized targets from masked input. If we want the representation robust to noise, the representation of an original speech should also be able to predict the target of its noisy version and vice versa.
- Therefore, we enforce the prediction consistency constraint in the contrastive loss without adding any complexity to networks.

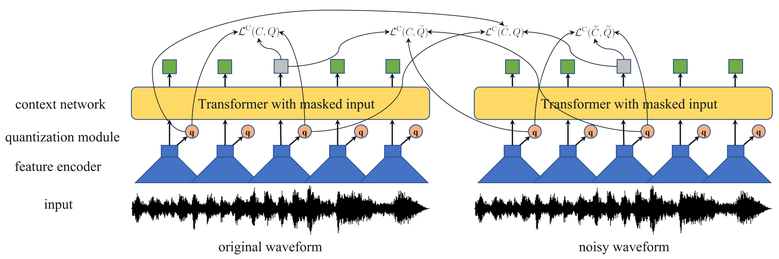
wav2vec 2.0

$$\begin{aligned}&\text{feature encoder } f: \mathcal{X} \mapsto \mathcal{Z}, \quad Z = [\mathbf{z}_1, \dots, \mathbf{z}_{T'}] \in \mathcal{Z} \\ &\text{context network } g: \mathcal{Z} \mapsto \mathcal{C}, \quad C = [\mathbf{c}_1, \dots, \mathbf{c}_{T'}] \in \mathcal{C} \\ &\text{quantization module } h: \mathcal{Z} \mapsto \mathcal{Q}, \quad Q = [\mathbf{q}_1, \dots, \mathbf{q}_{T'}] \in \mathcal{Q}\end{aligned} \tag{1}$$

$$\mathcal{L}(C, Q) = \sum_{t=1}^{N} \mathcal{L}_t(C, Q)/N \tag{2}$$

$$\mathcal{L}_t(C, Q) = -\log \frac{\exp(\text{sim}(\mathbf{c}_t, \mathbf{q}_t))}{\sum_{\mathbf{q}^- \in Q_t^-} \exp(\text{sim}(\mathbf{c}_t, \mathbf{q}^-))}, \quad Q_t^- = \{\mathbf{q}_1^-, \dots, \mathbf{q}_K^-\} \tag{3}$$

wav2vec-Switch



$$\begin{aligned}&\text{original-noisy speech pair } X, \widetilde{X} \in \mathbb{R}^{B \times T} \\ &C = g(f(X)), \qquad Q = h(f(X)) \\ &\widetilde{C} = g(f(\widetilde{X})), \qquad \widetilde{Q} = h(f(\widetilde{X}))\end{aligned} \tag{4}$$

$$\mathcal{L}_{\text{switch}}(C, Q, \widetilde{C}, \widetilde{Q}) = \mathcal{L}(C, Q) + \mathcal{L}(\widetilde{C}, \widetilde{Q}) + \lambda \left(\mathcal{L}(C, \widetilde{Q}) + \mathcal{L}(\widetilde{C}, Q) \right) \tag{5}$$

Experiments

Dataset: LibriSpeech + MUSAN noise corpus.
Results: WERs on synthesized noisy data are improved relatively by 7.1–11.0% without LM, or 2.9–4.9% with LM, and remain the same on the original test sets.

Table 1: Results on the original and synthesized noisy LibriSpeech test sets under the matched condition.

	LM	**Original**		**Noisy** (5–10 dB)	
		test-clean	test-other	test-clean	test-other
wav2vec 2.0	N	5.9	**13.4**	15.6	33.1
	Y	**2.6**	**6.6**	8.0	21.3
+ MUSAN *music+noise* (5–10 dB) (Baseline)	N	6.1	14.1	8.2	19.8
	Y	**2.6**	6.7	3.4	10.2
wav2vec-Switch	N	**5.8**	13.6	**7.3**	**18.4**
	Y	2.7	6.7	**3.3**	**9.7**

Experiments (cont.)

Table 2: Results on the synthesized noisy sets under different mismatched noisy conditions (without LM).

	music+noise (0–5 dB)		*speech* (5–10 dB)		*speech* (0–5 dB)	
	test-clean	test-other	test-clean	test-other	test-clean	test-other
wav2vec 2.0 + MUSAN *music+noise* (5–10 dB)	11.0	26.1	25.7	52.9	54.7	82.4
wav2vec-Switch	**9.7**	**24.5**	**23.8**	**50.4**	**52.7**	**80.7**
Gain (%)	11.8	6.1	7.4	4.7	3.7	2.1

Dataset: CHiME-4 + MUSAN noise corpus. The real 1-channel track is for testing.
Results: The relative improvement from the corresponding baseline is 7.8% without LM (16.5 vs. 17.9), or 5.7% with LM (6.6 vs. 7.0). It is also worth noting that, without any speech enhancement, our self-supervised approach followed by a simple CTC fine-tuning achieves better results than those using carefully designed enhancement algorithms. The additional data we were using was just the unlabeled 960-hour LibriSpeech audio and the MUSAN corpus.

Table 3: Results on the CHiME-4 real 1-channel dev/eval sets.

	continual pre-training	**LM**	dev	eval
Chen et al. (Kaldi Baseline) [1] (2018)		Y	5.6	11.4
Du et al. [2] (2016)		Y	4.6	9.2
Wang et al. [3] (2020)		Y	**3.5**	6.8
wav2vec 2.0 + MUSAN *music+noise* (5–10 dB) (Baseline)	N	N	10.6	17.6
		Y	3.7	7.2
	Y	N	10.7	17.9
		Y	4.6	7.0
wav2vec-Switch	N	N	10.2	16.8
		Y	3.6	7.1
	Y	N	**10.0**	**16.5**
		Y	**3.5**	**6.6**

References

[1] Szu-Jui Chen et al. "Building State-of-the-art Distant Speech Recognition Using the CHiME-4 Challenge with a Setup of Speech Enhancement Baseline". In: *Proc. Interspeech.* 2018.

[2] Jun Du et al. "The USTC-iFlytek system for CHiME-4 challenge". In: *Proc. CHiME-4 Challenge.* 2016.

[3] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. "Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR". In: *IEEE ACM Trans. Audio Speech Lang. Process.* 28 (2020), pp. 1778–1787.