ASR Posterior-based Loss for Multi-task End-to-end Speech Translation

Yuka Ko[1], Katsuhito Sudoh[1,2], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]

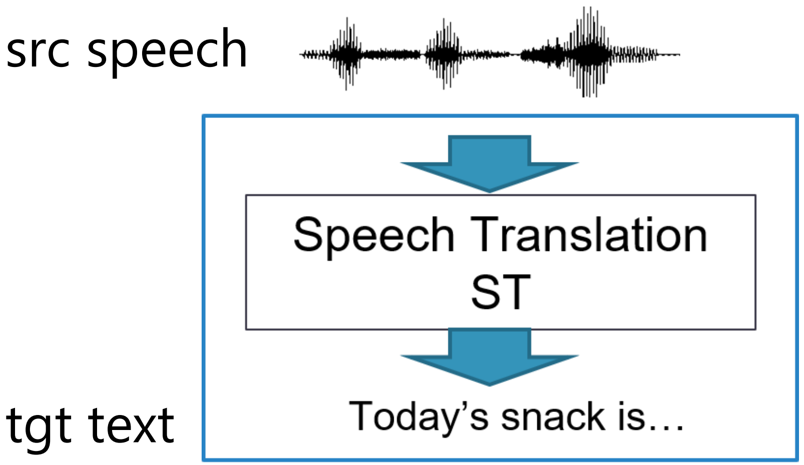
[1]Nara Institute of Science and Technology (NAIST), Japan
[2]RIKEN Center for Advanced Intelligence Project (AIP), Japan

Motivation

- End-to-End Speech Translation (ST)
  - ➢Data size is small : small pairs of src speech - tgt text

    → **Multi-task** is main approach
    - ➢**Cross entropy (CE) loss only using correct answer**

      → Multi-task doesn't consider ASR hypotheses



End-to-End ST

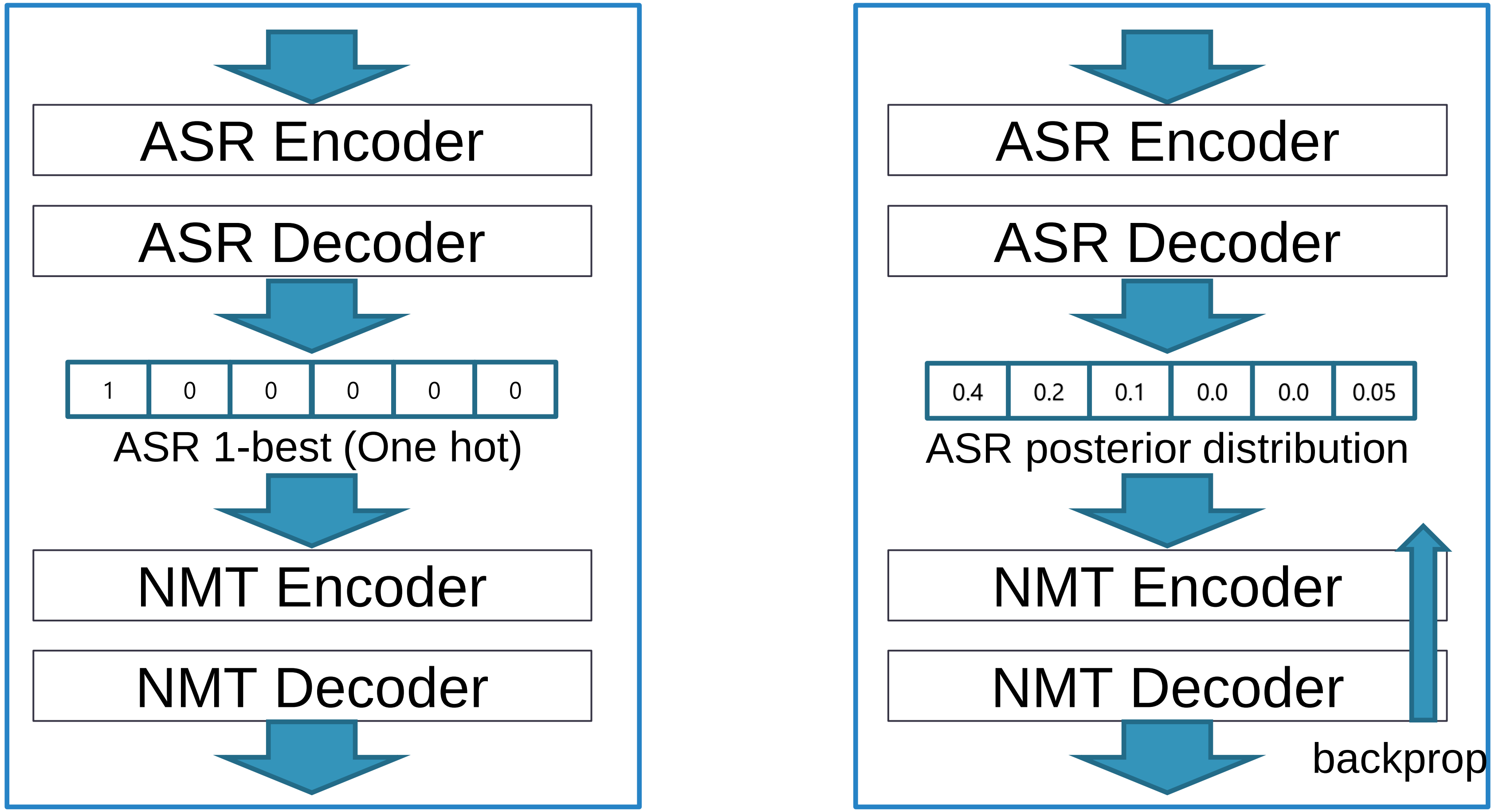
- Proposed
  - ➢Multi-task ST training with ASR posterior distribution

- Purpose
  - ➢**Robust multi-task End-to-End ST for ASR hypotheses**

Related Works

Related1 : Robust cascade ST for ASR hypotheses

- [Osamura+, 2018] (cascade ST)
  - ASR output : 1-best → **ASR posterior distribution**
  - Proposed : Robust NMT for ASR hypotheses

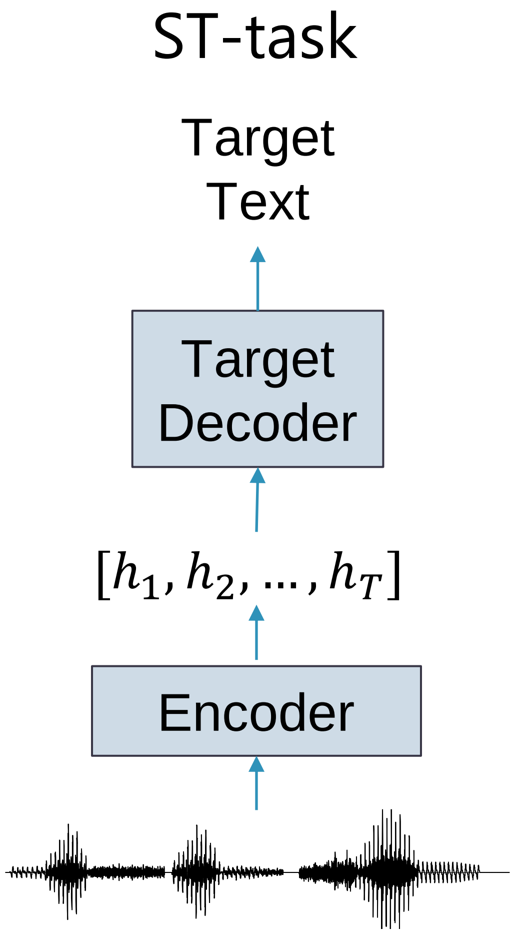


Previous

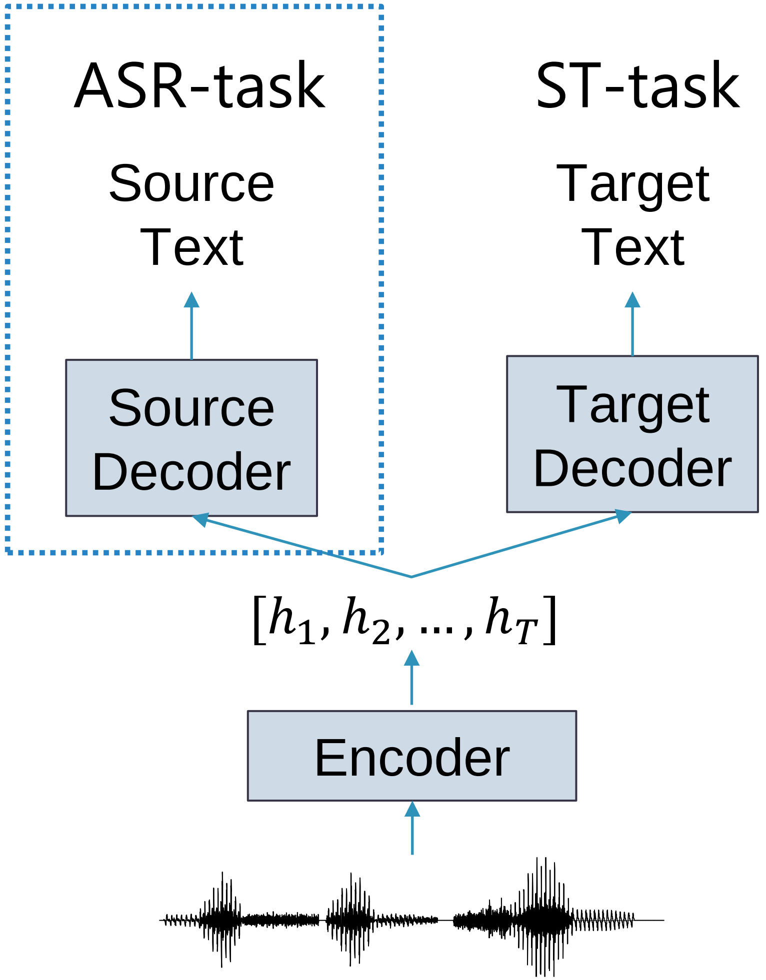
[Osamura+, 2018]

Related 2 : Single-task and Multi-task ST

- Single-task End-to-End ST



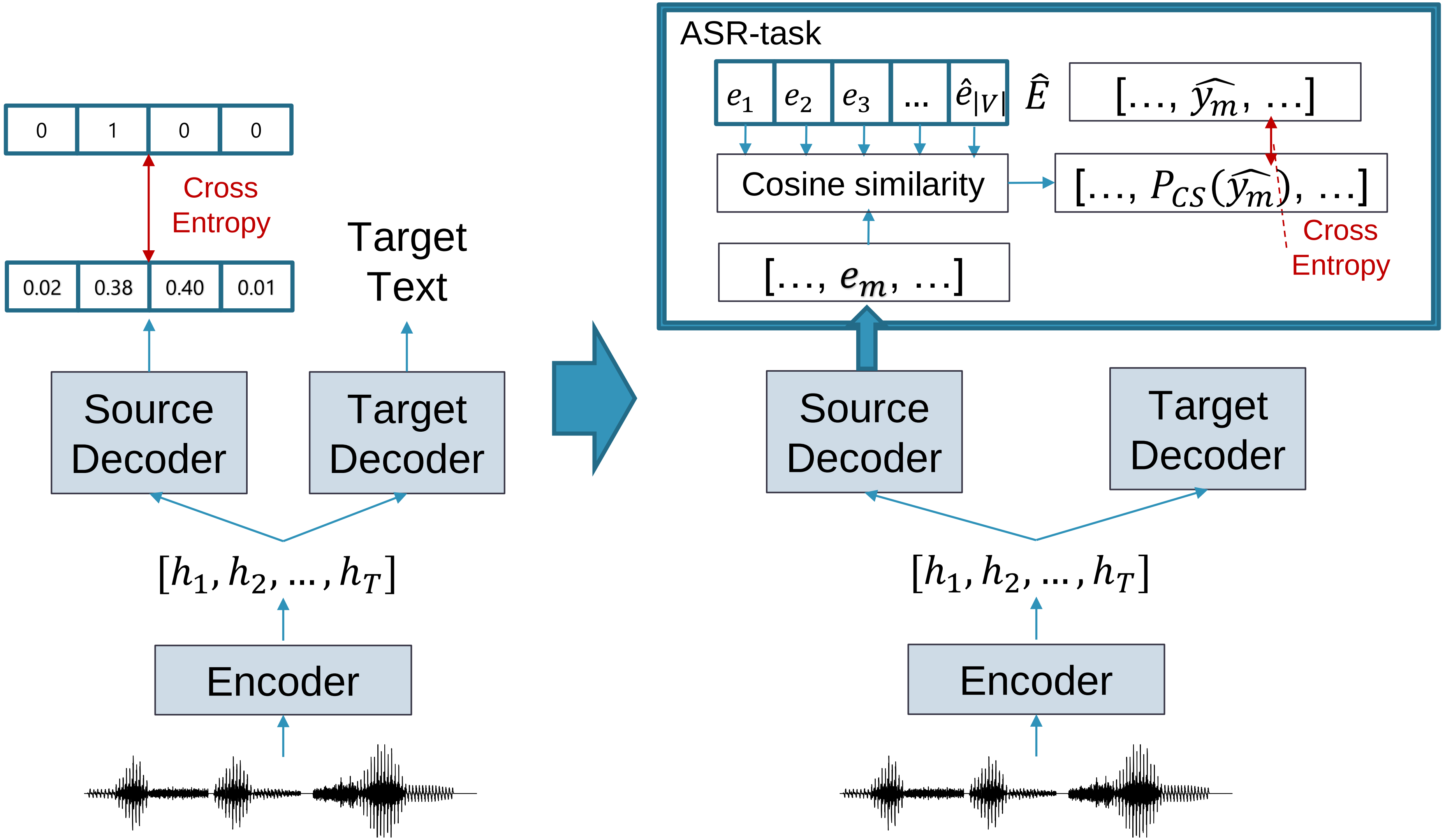
- Multi-task End-to-End ST [Weiss+ 2017]



[Weiss+ 2017] Sequence-to-Sequence Models Can Directly Translate Foreign Speech

Related3 : Robust Multi-task ST for semantic similarity

- Using cosine similarity in ASR-task loss calculation [Chuang+, 2020]

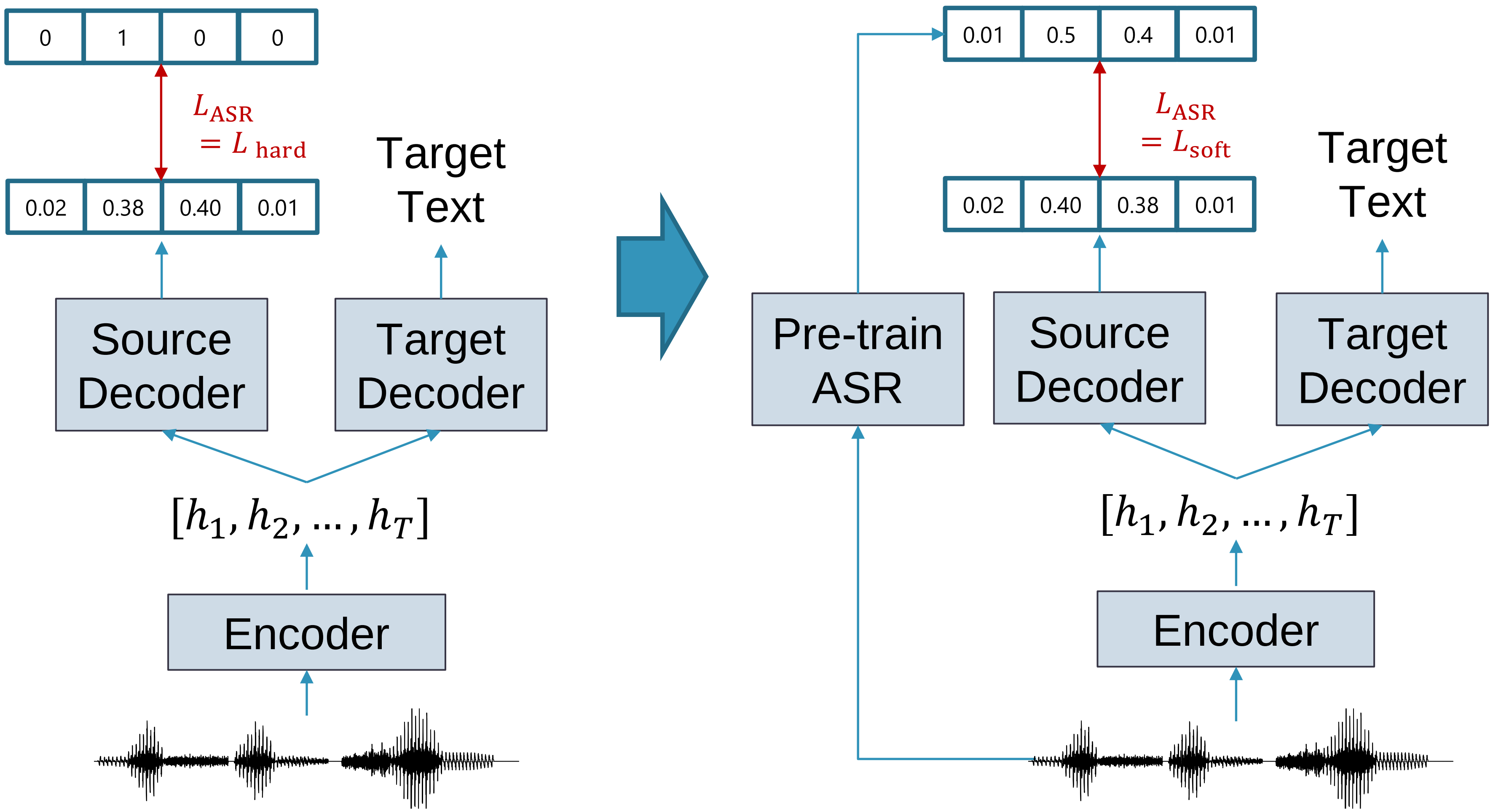


Proposed Method

Previous and Proposed

- Previous : L_{hard}
- Reference : One-hot
  - △ Same score in mistaken word
  - △ Not consider acoustic similarity

- Proposed : L_{soft}
- Reference : **ASR posterior distribution**
  - ○ Consider acoustic similar word & unsimilar word



Previous and Proposed

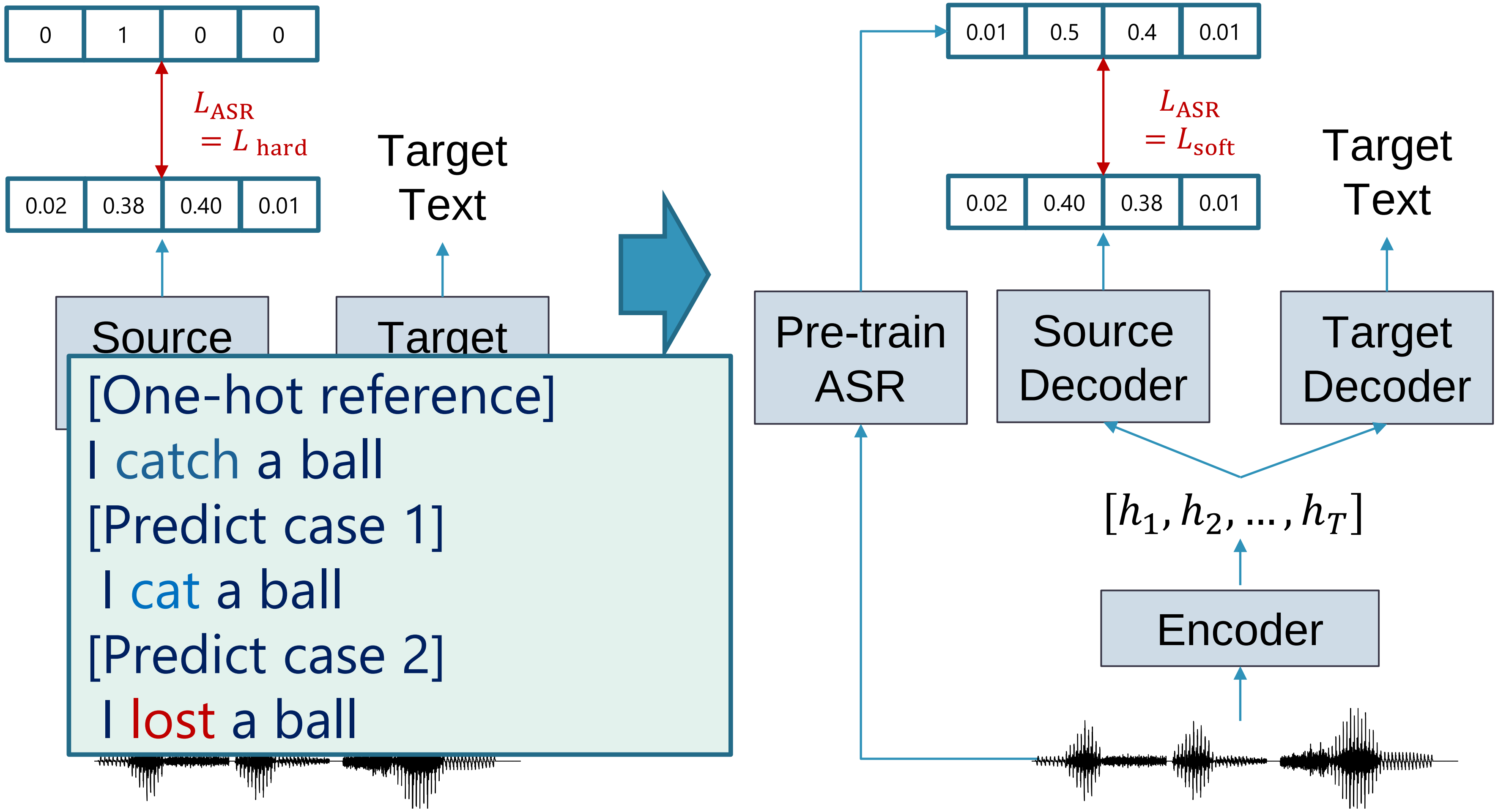
- Previous : L_{hard}
- Reference : One-hot

△ Same score in mistaken word

△ Not consider acoustic similarity

- Proposed : L_{soft}
- Reference : **ASR posterior distribution**

〇 Consider acoustic similar word & unsimilar word



Experiments

Experiment : Dataset

data	src-tgt	speech feature
Fisher Spanish	Es-En	fbank + pitch (80+3=83dim)

BPE model	dict size
SentencePiece	1000 Es-En Joint

	dataset	data size
Train	fisher_train	140K (Before 3 times augmentation)
Dev	fisher_dev	3.9K
Test	fisher_dev2	3.9K
	fisher_test	3.6K

Experiment : Model Parameter

- Implement : ESPnet [Watanabe+, 2018], Transformer
- Pre-trained ASR model (for **soft labels** of ASR posterior distributions)
  - 30 epochs
  - Dev best model (in accuracy)
- ST model
  - 30 epochs
  - Checkpoint averaging : 5 (in BLEU)
  - L_{ST} : CE + Label smoothing
  - ASR-task loss
    - Baseline : L_{ASR} : {CE, CE + Label smoothing}
    - Proposed : **ASR Posterior-based Loss**
      - Dev best model in BLEU

$$\lambda_{\mathrm{ASR}} = \{0.3, 0.4, 0.5\}, \lambda_{\mathrm{soft}} = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$$

$$L_{\mathrm{ASR}} = \lambda_{\mathrm{soft}} L_{\mathrm{soft}} + (1 - \lambda_{\mathrm{soft}}) L_{\mathrm{hard}}$$

  - $L = \lambda_{\mathrm{ASR}} L_{\mathrm{ASR}} + (1 - \lambda_{\mathrm{ASR}}) L_{\mathrm{ST}}$

Result : Pre-trained ASR WER

- WER in dev best accuracy in epoch 30 (beam size 10)
- Our model without language model (LM) and CTC
  - When decoding soft labels, we set beam size 1

Model	LM	CTC	dev	dev2	test
Our ASR			30.2	29.1	27.2
ESPnet	✓	✓	24.2	23.6	22.8

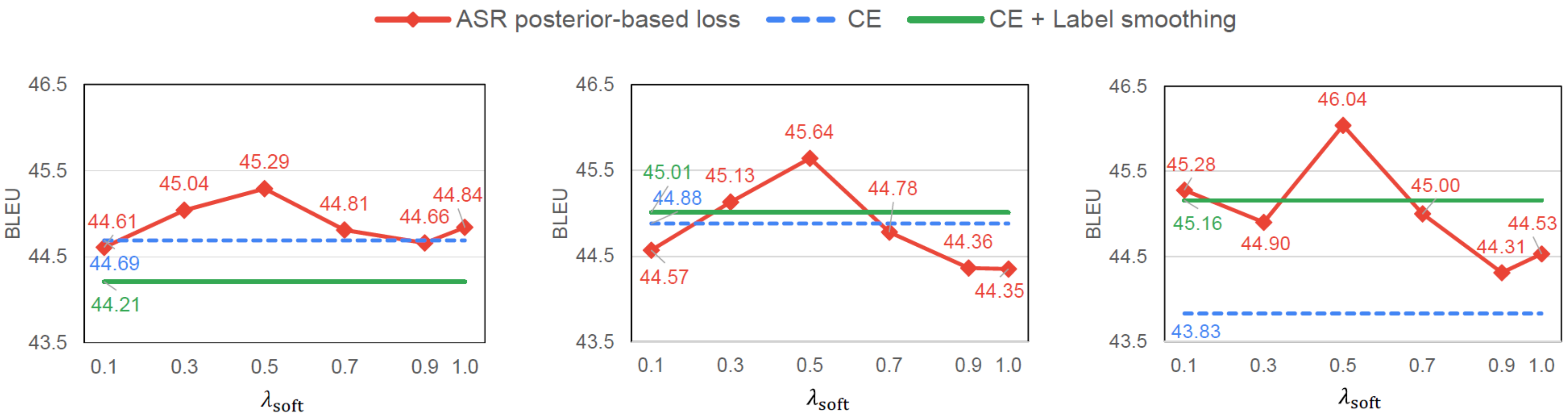
BLEU of test data (dev best on each parameters)

Model				BLEU		
Task	ASR-task loss	λ_{ASR}	λ_{soft}	dev	dev2	test
Single-task ST	-	-	-	41.10	41.61	40.66
Multi-task ST	CE (Baseline)	0.4	-	44.50	46.20	44.88
	CE + Label smoothing (Baseline)	0.5	-	45.29	46.34	45.16
	ASR Posterior-based loss (Proposed)	0.4	0.5	**45.54**	**46.46**	**45.64**

BLEU of test data (dev best on difference λ_{ASR})

Model				BLEU		
Task	ASR-task loss	λ_{ASR}	λ_{soft}	dev	dev2	test
Single-task ST	-	-	-	41.10	41.61	40.66
Multi-task ST	CE	0.3	-	44.08	45.07	**44.69**
	CE + Label smoothing		-	44.23	45.25	44.21
	ASR Posterior-based loss (Proposed)		0.9	**44.99**	**45.73**	44.66
	CE	0.4	-	44.50	46.20	44.88
	CE + Label smoothing		-	44.61	45.30	45.01
	ASR Posterior-based loss (Proposed)		0.5	**45.54**	**46.46**	**45.64**
	CE	0.5	-	43.64	45.28	43.83
	CE + Label smoothing		-	45.29	46.34	45.16
	ASR Posterior-based loss (Proposed)		0.5	**45.46**	**46.37**	**46.04**

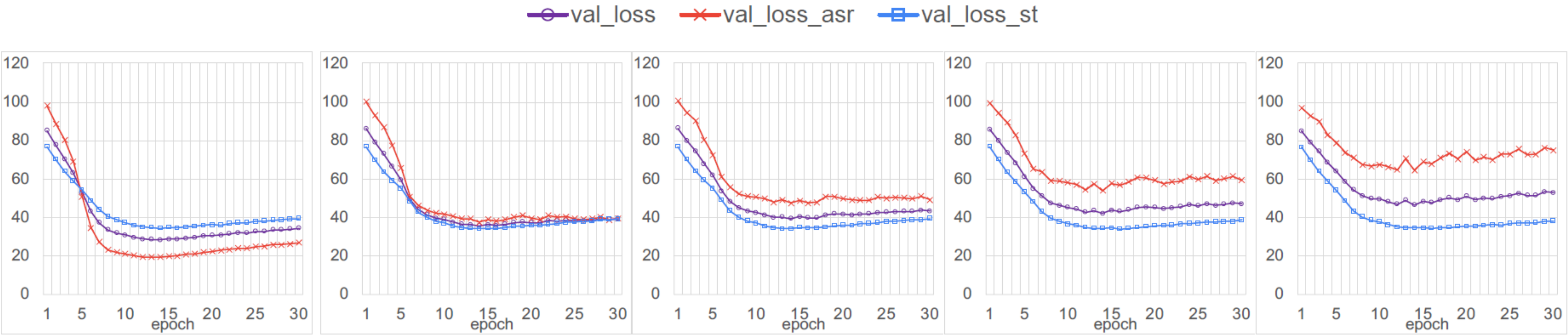
BLEU in Fisher test on different λ_{ASR}



(a) $\lambda_{\text{ASR}} = 0.3$ (b) $\lambda_{\text{ASR}} = 0.4$ (c) $\lambda_{\text{ASR}} = 0.5$

Dev loss in each λ_{soft} ($\lambda_{\text{ASR}} = 0.4$)

- Total loss is higher than CE
  - However (b) 0.3, (c) 0.5 improved
- Maybe proposed soft loss work as regularization
  - It is good in 0.5, in this method



(a) $\lambda_{\text{soft}} = 0.0$ (CE) (b) $\lambda_{\text{soft}} = 0.3$ (c) $\lambda_{\text{soft}} = 0.5$ (d) $\lambda_{\text{soft}} = 0.7$ (e) $\lambda_{\text{soft}} = 1.0$

Examples in Fisher test $\lambda_{\mathrm{ASR}} = 0.4$ $\lambda_{\mathrm{soft}} = 0.5$

	Example
Ground Truth (Es)	para relajacio´n
Ground Truth (En)	for relaxation
CE + Label smoothing (En)	for relationships (↔relaciones (es) ?)
ASR Posterior-based loss (En)	for **relaxing**
Ground Truth (Es)	quie´n no su sobrina
Ground Truth (En)	who no your niece
CE + Label smoothing (En)	who his nephews (↔sobrino (es) ?)
ASR Posterior-based loss (En)	who are your **nieces**

Conclusion

- **Purpose** : Robust Multi-task End-to-End ST for ASR hypotheses
- **Proposed** : ASR posterior-based loss
  - ➢ Using ASR posterior distribution in End-to-End ST training
- **Results** : BLEU improvement in proposed
  - ➢ Baseline : CE, CE + Label smoothing
  - ➢ Mixing hard loss & soft loss was better
- **Future work**
  - ➢ More analysis on ASR-task output
  - ➢ Using pre-trained ASR models of different performances
  - ➢ Using pronunciation information (phone [Salesky+, 2020]) in loss calculation
  - ➢ Fine-tuning (first train with L_{hard}, after training with L_{soft})
- Our code is available at:
  - ➢ https://github.com/ahclab/st-asrpbl

Reference

- [Osamura+, 2018] K. Osamura, T. Kano, S. Sakti, K. Sudoh, and S. Nakamura, "Using Spoken Word Posterior Features in Neural Machine Translation," Proceedings of the 15th International Workshop on Spoken Language Translation, 181-188, Oct. 2018, vol. 21, p. 22, 2018.
- [Weiss+, 2017] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-tosequence models can directly translate foreign speech. Proc. Interspeech 2017, pages 2625–2629.
- [Chuang+, 2020] S. Chuang, T. Sung, A. H. Liu, and H. Lee, "Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 5998–6003.
- [Watanabe+, 2018] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen et al., "ESPnet: End-to-End Speech Processing Toolkit," Proc. Interspeech 2018, pp. 2207–2211, 2018.
- [Salesky+, 2020] E. Salesky and A. W. Black, "Phone Features Improve Speech Translation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 2388– 2397.