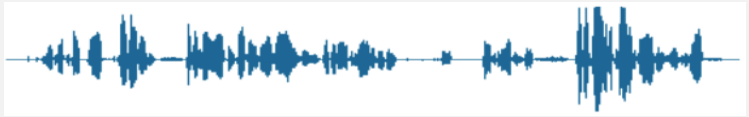
音声認識仮説の曖昧性を考慮する Multi-task End-to-End 音声翻訳

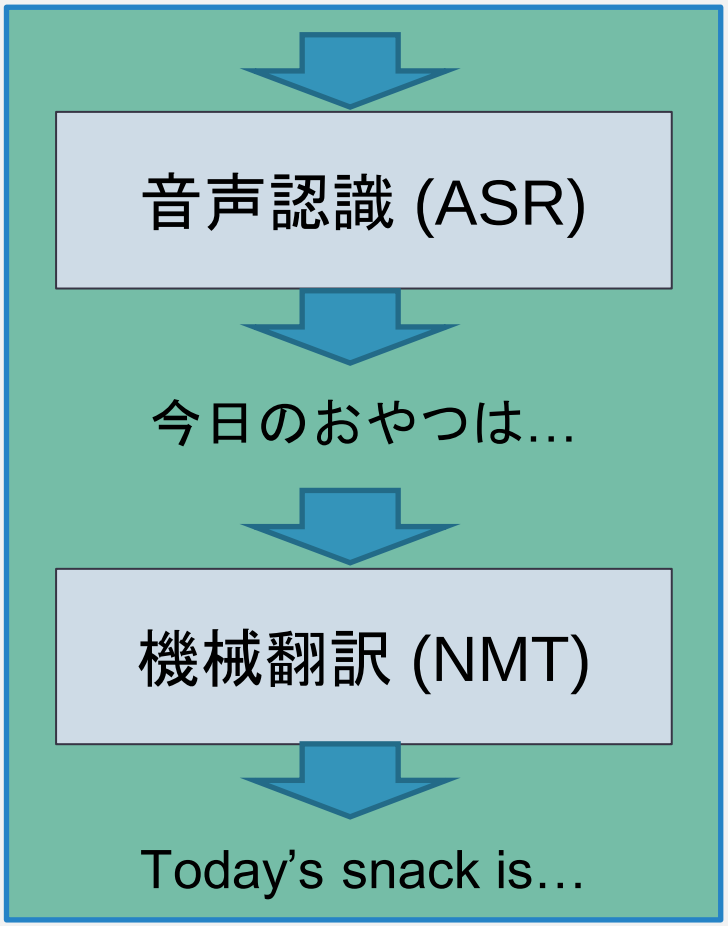
胡 尤佳　　須藤 克仁　　Sakriani Sakti　　中村 哲
奈良先端科学技術大学院大学
理化学研究所 革新知能統合研究センター AIP

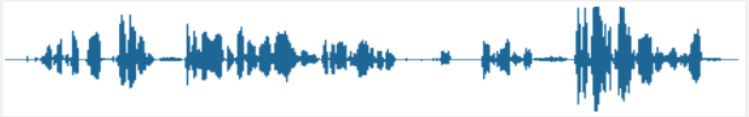
2021/03/15–03/19 @言語処理学会第27回年次大会

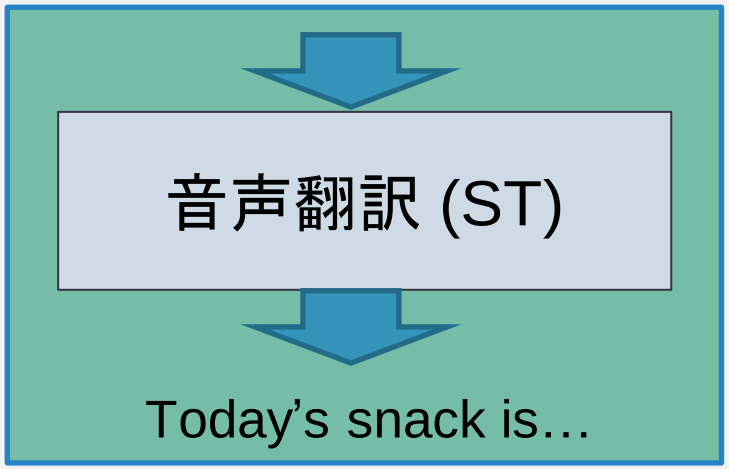
はじめに

- 音声翻訳 Speech Translation (ST)
  - Cascade ST
  - End-to-End ST









- Cascade
  - 音声認識誤りの伝播
- End-to-End
  - Single task だと学習が困難
  - **Pre-train + Multi-task が必要**

目的と動機

- Cascade ST
  - 音声認識誤りに対して頑健な機械翻訳が必要

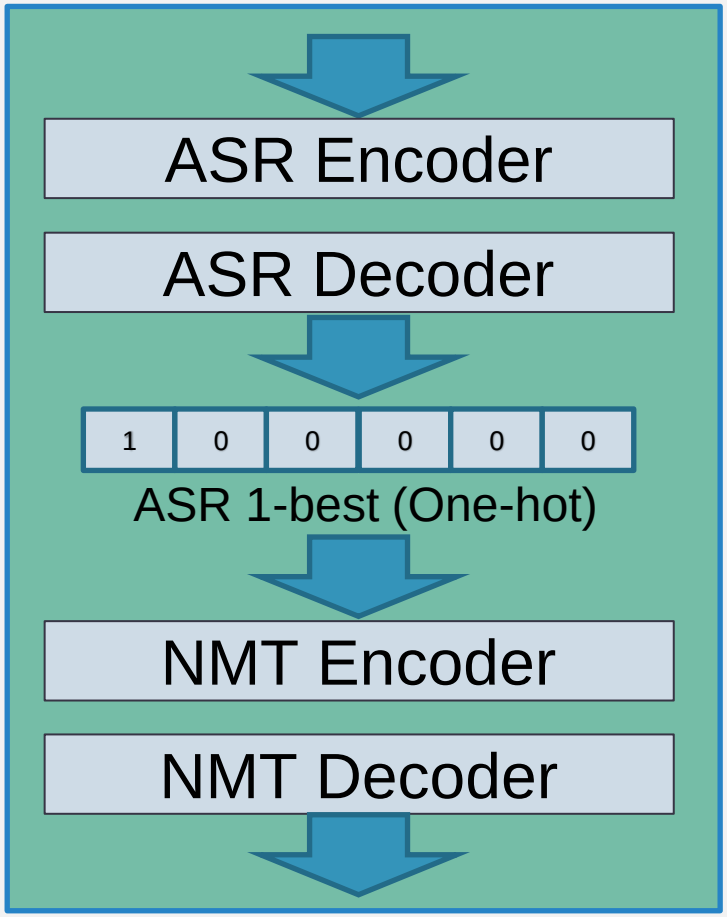
- End-to-End ST
  - 音声認識誤りの問題は本当になくなった？
    - Multi-task での学習中に存在すると仮定
    - 音声認識誤り，曖昧性を考慮する必要

- 提案
  - End-to-End 音声翻訳においても音声認識誤りに頑健なモデルの学習方法

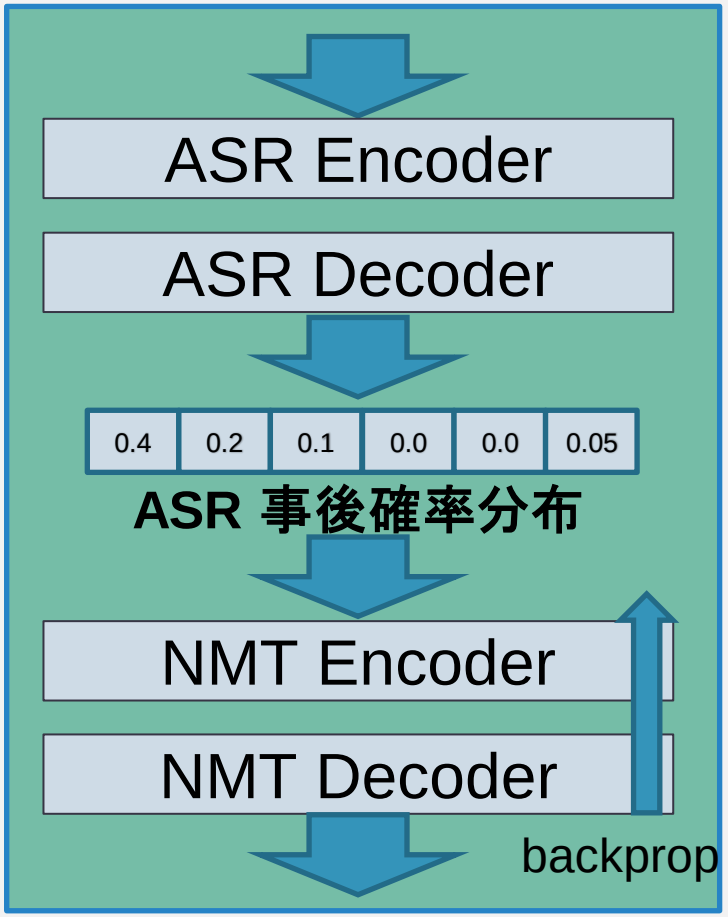
- 目的
  - 音声翻訳の精度改善の期待

関連研究 1：音声認識誤りに頑健な機械翻訳

- [Osamura+, 2018] (cascade)
  - 音声認識出力：1-best → ASR 事後確率分布
  - 音声認識誤りに対して頑健な機械翻訳モデルの提案



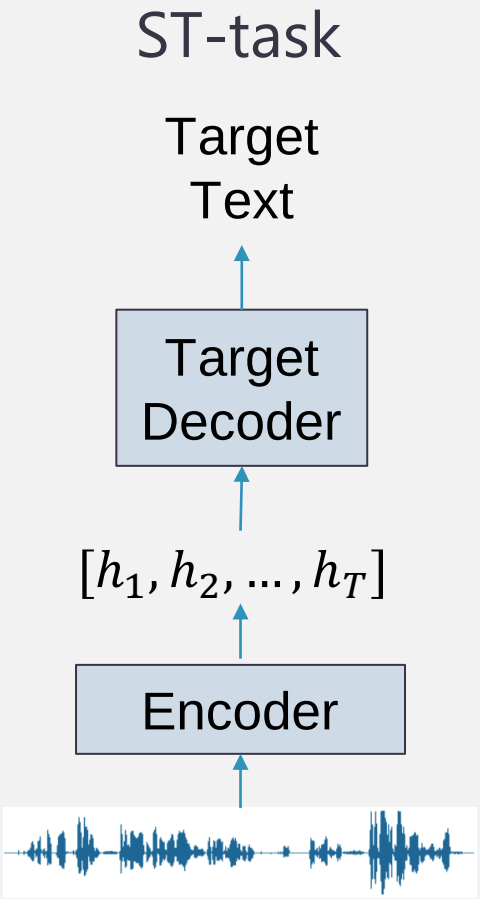
従来手法

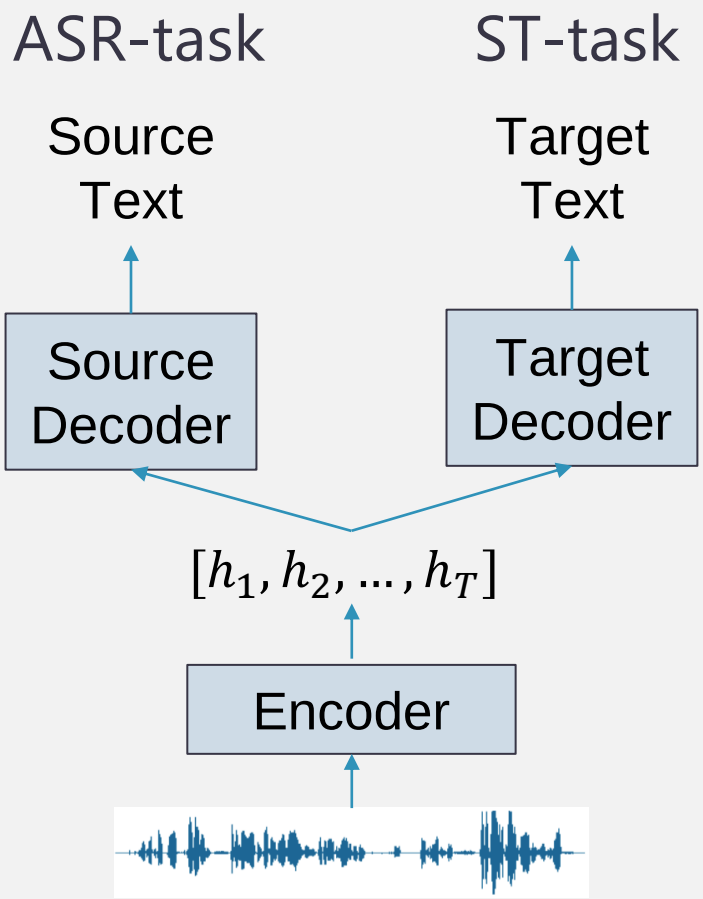


Osamuraらの手法

関連研究 2 : Multi-task End-to-End 音声翻訳

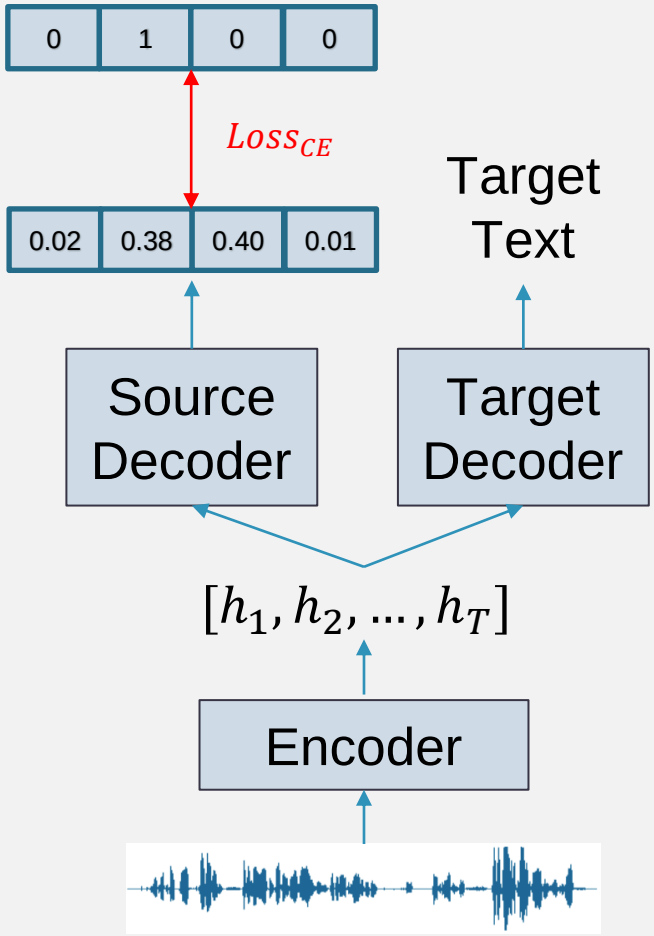
- Single-task End-to-End ST
- Multi-task End-to-End ST

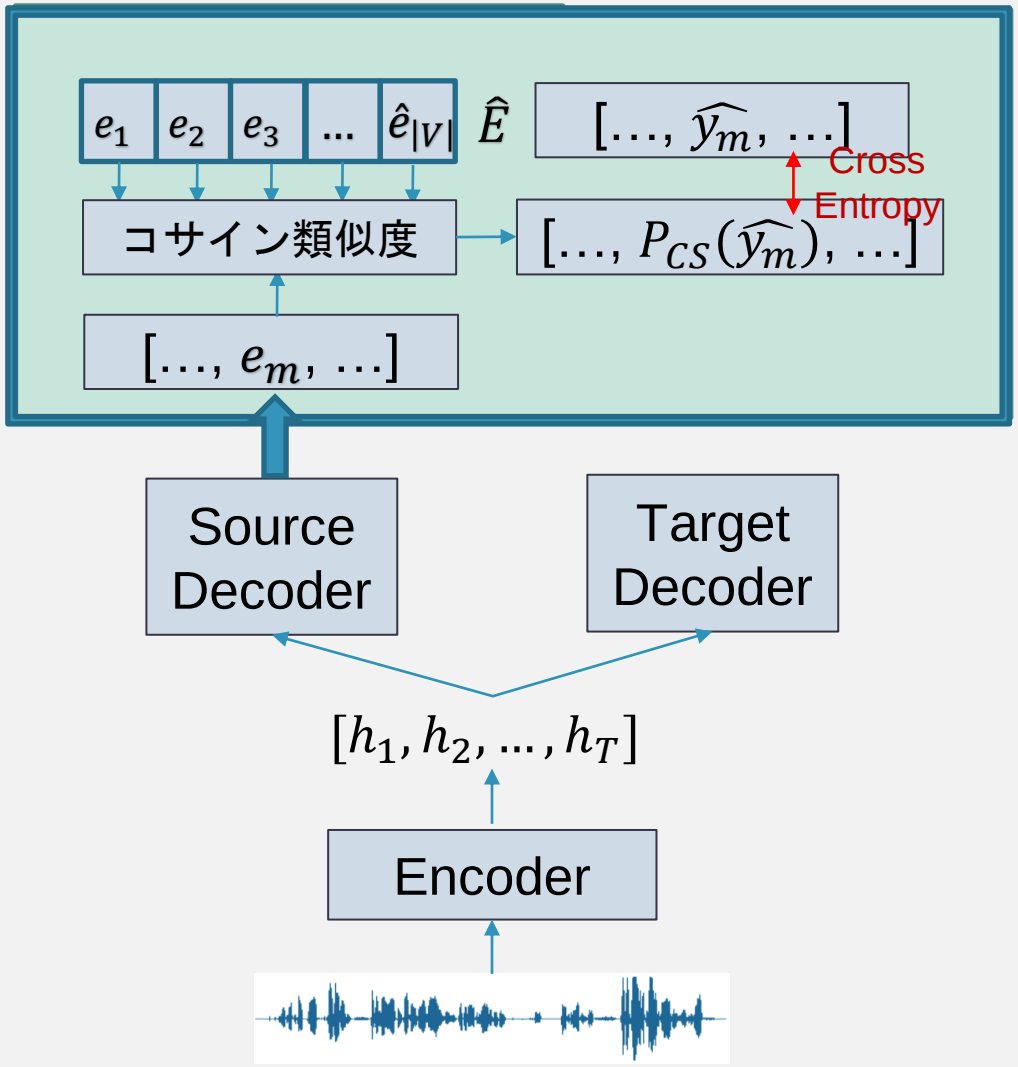




関連研究 3：単語分散表現の類似度を用いたST

- 一般的な Multi-task ST
  - 参照：One-hot ベクトル
- [Chuang+, 2020] 参照，予測間の意味的類似度を用いた誤差

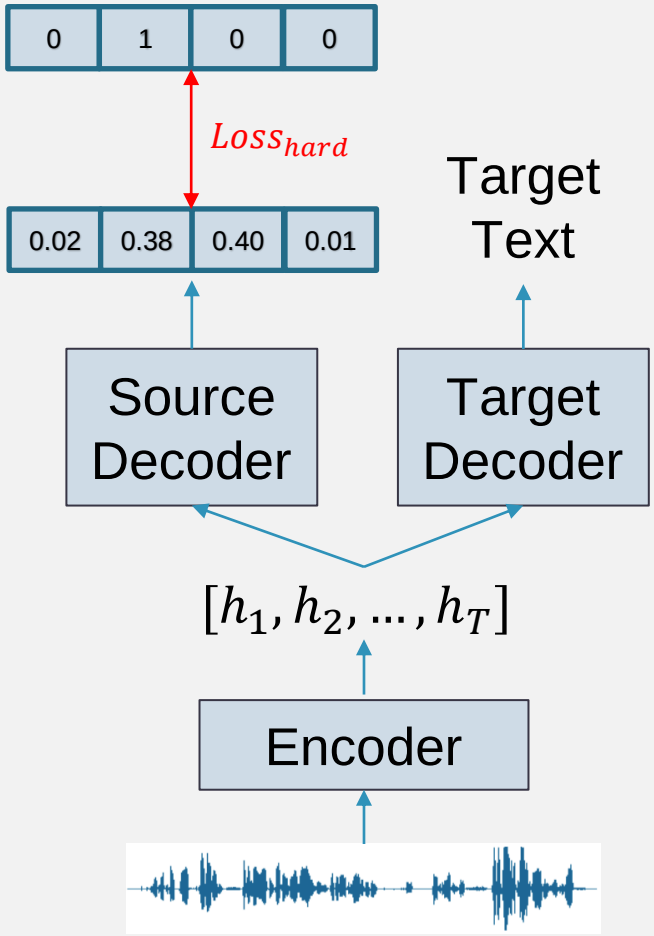




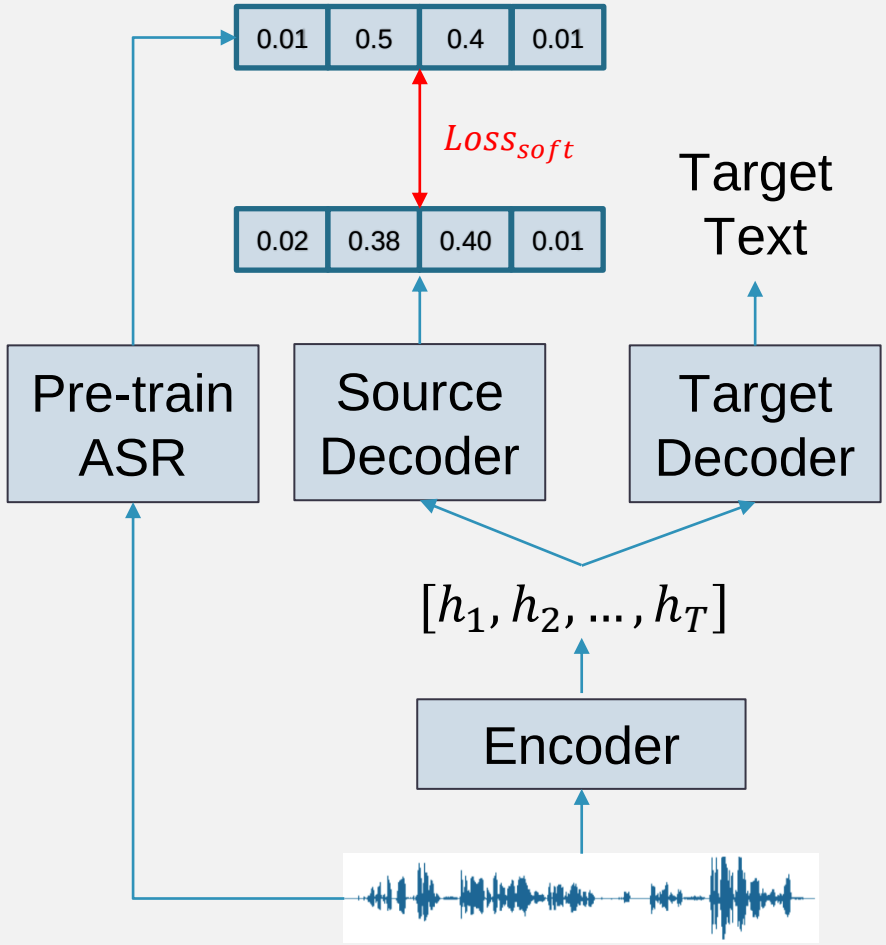
提案手法

従来手法と提案手法

- 従来手法
  - 参照：One-hot ベクトル
    - Hard target loss

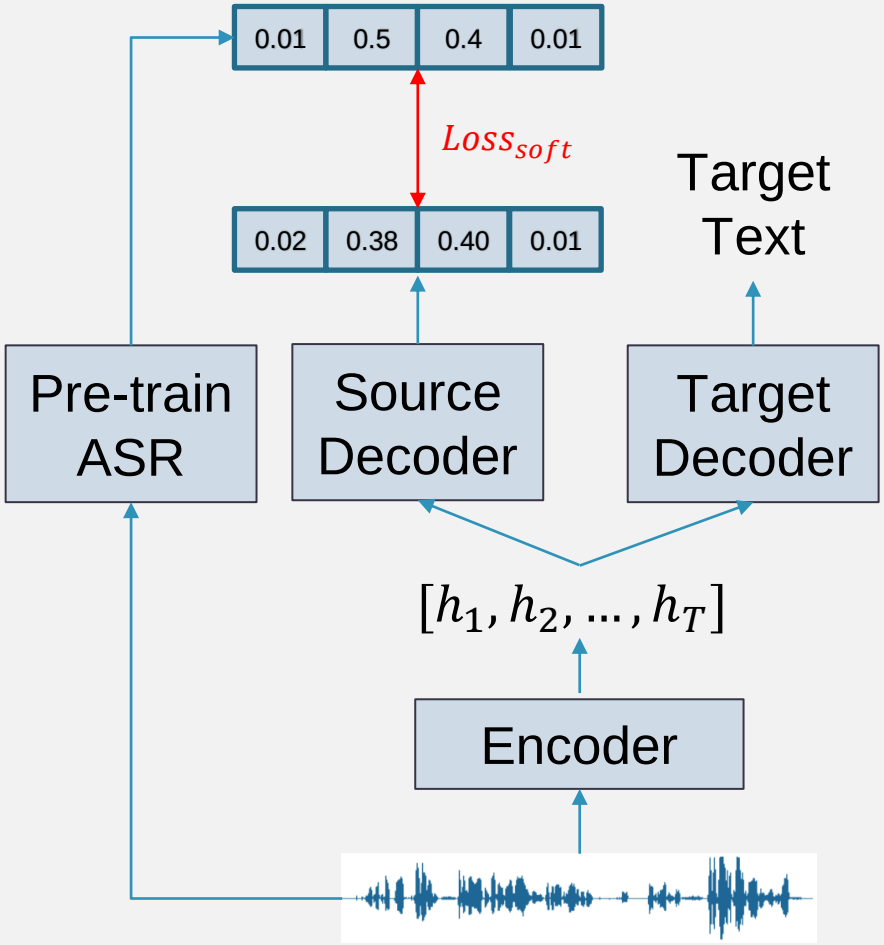


- 提案手法
  - 参照：ASR 事後確率分布
    - Soft target loss



従来手法と提案手法

- 従来手法
  - 参照：One-hot ベクトル
    - Hard target loss
- 提案手法
  - 参照：ASR 事後確率分布
    - Soft target loss



実験

実験：データセット

Data	src-tgt	Speech feature
Fisher Spanish CallHome コーパス	Es-En (Spanish-English)	fbank + pitch (80+3=83次元)

Dictionary
SentencePiece 1000 Es-En Joint

	Dataset	Size
Train	fisher_train	415869 (138623 * 3)
Dev	fisher_dev	3973
Test	fisher_dev2	3957
	fisher_test	3638
	callhome_devtest	3956
	callhome_evltest	1825

実験：学習時モデルパラメータ

- 実装: ESPnet
- Pre-train ASR
  - Dev best (in accuracy)
- ST
  - $\lambda_{soft} = \{0, 0.3, 0.5, 0.7, 1.0\}, \lambda_{asr} = 0.3$
    - Baseline : $\lambda_{soft} = 0.0$
      - 一般的なcross entropy loss / label smoothing loss
    - $Loss_{ASR} = \lambda_{soft} Loss_{soft} + \left(1 - \lambda_{soft}\right) Loss_{hard}$
    - $Loss = \lambda_{ASR} Loss_{ASR} + (1 - \lambda_{ASR}) Loss_{ST}$
  - Label smoothing weight (0.0 : cross entropy)
    - 実験1： ASR-task 0.0 / ST-task 0.0 [論文に掲載]
    - 実験2： ASR-task 0.0 / ST-task 0.1 [追加掲載]
    - 実験3： ASR-task 0.1 / ST-task 0.1 [省略/付録]

結果：Pre-train ASR WER

- 実験に用いた Pre-train ASR モデルの WER
  - Dev best model in epoch 30

	WER
fisher_dev	30.152
fisher_dev2	29.124
fisher_test	27.184
callhome_devtest	49.138
callhome_evltest	49.646

実験1：ASR-task 0.0 / ST-task 0.0

- BLEU スコア
- ST-task : cross entropy / ASR-task : cross entropy
- 全体的なBLEUの向上が見られた (↓：低下 / **太字：最大)**

		Baseline	Proposed			
	softλ_{soft}-hard$(1-\lambda_{soft})$	soft0.0-hard1.0	soft0.3-hard0.7	soft0.5-hard0.5	soft0.7-hard0.3	soft1.0-hard0.0
fisher dev	BLEU 4-ref	41.04	40.99 ↓	41.40	41.20	**41.51**
	BLEU 1-ref	23.97	23.88 ↓	24.12	24.00	**24.33**
fisher dev2	BLEU 4-ref	42.14	42.05 ↓	42.28	**42.45**	42.22
	BLEU 1-ref	25.17	25.23	25.22	25.30	**25.32**
fisher test	BLEU 4-ref	41.17	41.38	**41.41**	41.18	41.39
	BLEU 1-ref	24.77	25.02	24.93	24.82	**25.01**
callhome devtest	BLEU 1-ref	14.83	**15.23**	15.00	15.01	14.95
callhome evltest	BLEU 1-ref	14.81	**15.26**	15.10	14.78 ↓	15.09

実験1：Fisher test $\lambda_{soft} = 0.5$

- Label smoothing weight = 0.0 (cross entropy loss)

	Example
Label Ground Truth (Es) Ground Truth (En) Baseline (En) Proposed (En)	20051028_180633_356_fsp-A-016164-016487 sí pero o sea sigue siendo bastante **intensi** yes but it's still pretty **intensive** yes but that keeps getting pretty **unthinkable** (->"inconcebible", "impensable" (Es)) yes but that keeps being pretty **intense**
Label Ground Truth (Es) Ground Truth (En) Baseline (En) Proposed (En)	20051028_180633_356_fsp-A-033453-034134 es es mejor en el sentido que uno okay que hay menos **riesgos** pero ay that is the best in the sense that one okay that there are less **risks** but ay it's it's better in the sense that you don't that there are less **colds** but there are (->resfriados (Es)) it's it's a best in the sense that you don't that there are less **risks** but

実験2：ASR-task 0.0 / ST-task 0.1

- Label smoothing weight = 0.0
  - ASR-task の Hard loss はcross entropy loss
- 全体的なBLEUの向上が見られた (↓：低下 / **太字：最大**)
  - Soft lossのみを用いると精度が低くなる傾向

		Baseline	Proposed			
	softλ_{soft}-hard$(1-\lambda_{soft})$	soft0.0-hard1.0	soft0.3-hard0.7	soft0.5-hard0.5	soft0.7-hard0.3	soft1.0-hard0.0
fisher dev	BLEU 4-ref	44.08	**44.91**	44.72	44.38	43.99 ↓
	BLEU 1-ref	25.68	25.69	**25.86**	25.58 ↓	25.39 ↓
fisher dev2	BLEU 4-ref	45.07	45.70	45.54	45.69	45.07
	BLEU 1-ref	27.00	27.13	27.01	27.33	26.99 ↓
fisher test	BLEU 4-ref	44.69	45.04	**45.29**	44.81	44.84
	BLEU 1-ref	26.78	26.73 ↓	**27.16**	26.61 ↓	26.55 ↓
callhome Devtest	BLEU 1-ref	15.83	16.09	16.28	15.96	**16.34**
callhome evltest	BLEU 1-ref	15.85	16.01	**16.80**	15.42 ↓	16.22

まとめ

- 音声認識の事後確率分布による End-to-End ST の学習
  - 音声認識の曖昧性に対するロバスト性を期待
  - BLEUの向上が期待できることを示した
    - 全体的な BLEU の向上が見られた
  - Soft loss のみの場合，参照としての信用度が低く精度が下がる可能性

- 今後の展望
  - Pre-train ASR の性能ごとの信頼度の検証
  - テスト時の出力結果の分析
    - Main-task (ST) 出力とSub-task (ASR) 出力の照らし合わせ
    - Decoder 出力分布の確認
  - 発音，読み情報を利用した誤差計算 [Salesky+, 2020]

参考文献

- [Osamura+, 2018] Osamura, Kaho, et al. "Using spoken word posterior features in neural machine translation." *architecture* 21 (2018): 22.
- [Anastasopoulos+, 2018] Anastasopoulos, Antonios, and David Chiang. "Tied multitask learning for neural speech translation." *arXiv preprint arXiv:1802.06655* (2018).
- [Chuang+, 2020] Chuang, Shun-Po, et al. "Worse WER, but Better BLEU? Leveraging Word Embedding as Intermediate in Multitask End-to-End Speech Translation." *arXiv preprint arXiv:2005.10678* (2020).
- [Watanabe+, 2018] "Espnet: End-to-end speech processing toolkit." *arXiv preprint arXiv:1804.00015* (2018).
- [Kikui+, 2003] "Creating corpora for speech-to-speech translation." *Eighth European Conference on Speech Communication and Technology*. 2003.
- [Salesky+, 2020] Elizabeth Salesky and Alan W. Black. Phone features improve speech translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 2388–2397. Association for Computational Linguistics, 2020.

Appendix

実験：学習時モデルパラメータ

	ASR	ST
epoch	30	
encoder layers	12	
encoder units	2048	
decoder layers	6	
decoder units	2048	
attention dimention	256	
attention heads	4	
batch size	64	
accum grad	2	4
gradient clipping	5	
transformer learning late	5	2.5
transformer warmup steps	25000	
decode beam size	1	4
label smoothing weight	0.1	{0, 0.1}
dropout	0.1	
model average	1	5

実験3：ASR-task 0.1 / ST-task 0.1

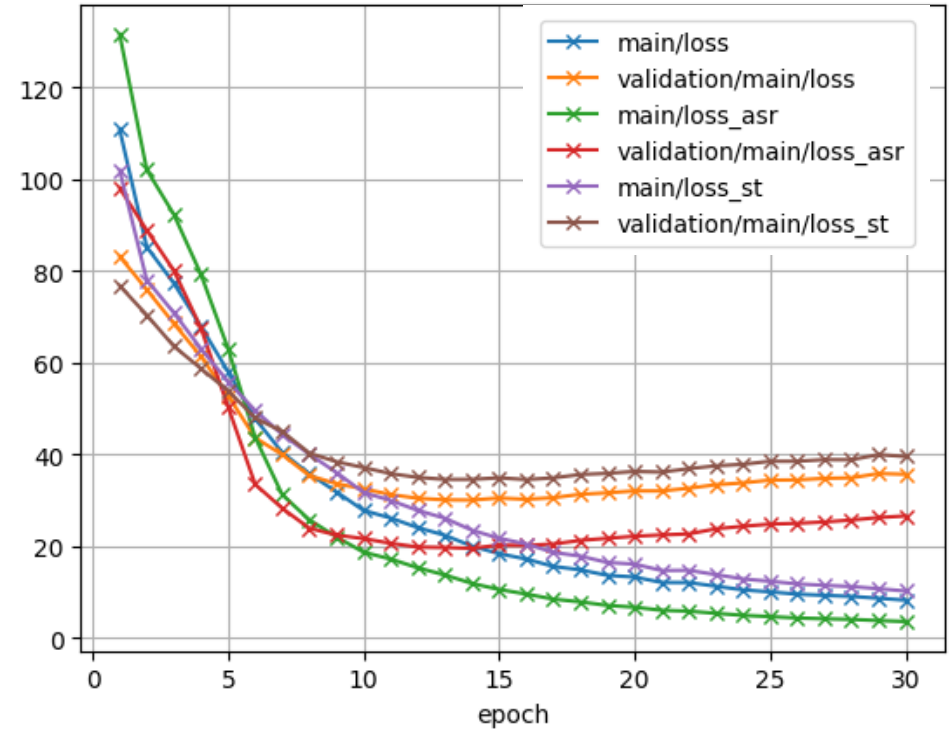
- Label smoothing weight = 0.1 (ASR Hard lossに対して)
- 全体的なBLEUの向上が見られなかった(↓：低下 / **太字：最大**)

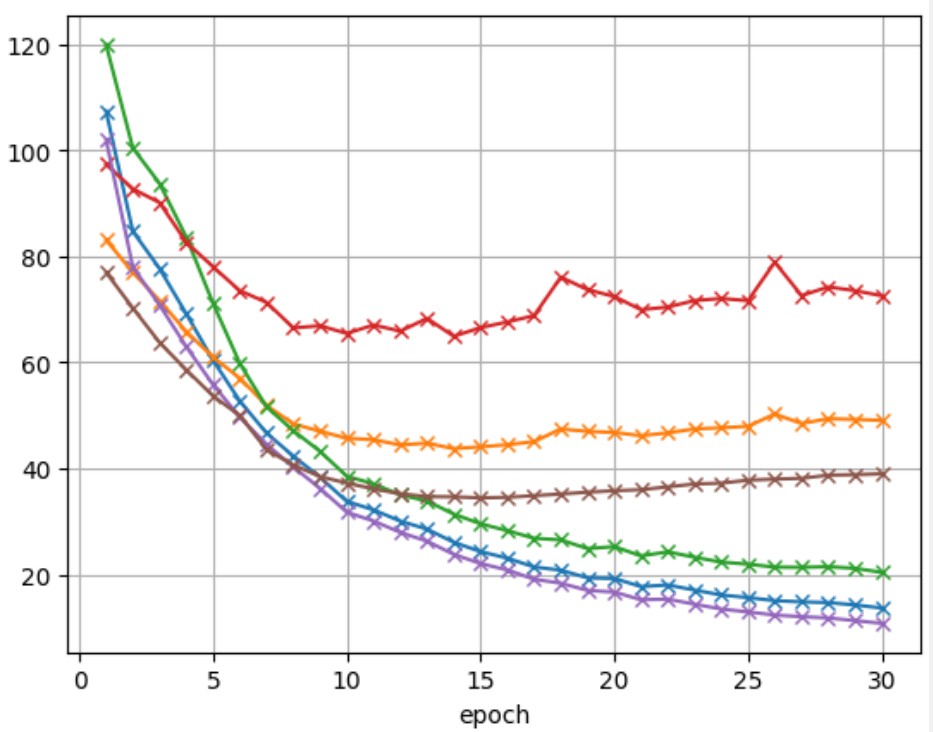
		Baseline	Proposed			
	softλ_{soft}-hard$(1-\lambda_{soft})$	soft0.0-hard1.0	soft0.3-hard0.7	soft0.5-hard0.5	soft0.7-hard0.3	soft1.0-hard0.0
fisher dev	BLEU 4-ref	43.82	44.28	**44.47**	44.70	43.99
	BLEU 1-ref	25.64	**25.78**	25.45 ↓	25.72	25.39 ↓
fisher dev2	BLEU 4-ref	45.33	45.39	45.51	**46.25**	45.07 ↓
	BLEU 1-ref	27.11	27.16	27.22	**27.42**	26.99 ↓
fisher test	BLEU 4-ref	44.15	44.24	44.57	**45.04**	44.84
	BLEU 1-ref	26.72	26.45 ↓	**26.74**	26.72 ↓	26.55 ↓
callhome devtest	BLEU 1-ref	15.93	15.85 ↓	15.85 ↓	16.17	**16.34**
callhome evltest	BLEU 1-ref	**16.22**	15.81 ↓	16.06 ↓	15.52 ↓	16.22

実験 : 実装モデル

- 実装 : ESPnet
- Pre-train ASR model
  - transformer
  - Epoch : 100
  - Dev Best model
- ST model
  - transformer
  - Epoch : 100
  - Checkpoint averaging : 5
  - $W_{soft} = \{0, 0.3, 0.5, 0.7, 1.0\}$, $W_{asr} = 0.3$
    - $loss_{asr} = W_{soft} loss_{soft} + (1 - W_{soft}) loss_{hard}$
    - $loss = W_{asr} loss_{asr} + (1 - W_{asr}) loss_{st}$

loss_asr soft0.0 vs soft1.0





- soft_weight = 0.0
- soft_weight = 1.0