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[1]　須藤 克仁[1,2]　Sakriani Sakti[1,2]　中村 哲[1,2]
[1]奈良先端科学技術大学院大学
[2]理化学研究所 革新知能統合研究センター AIP
{ko.yuka.kp2, sudoh, ssakti, s-nakamura}@is.naist.jp

1 はじめに

音声翻訳 (Speech Translation; ST) は，原言語の音声を入力として目的言語のテキストを出力する技術である．従来は音声認識モデル (Automatic Speech Recognition; ASR) と機械翻訳モデル (Machine Translation; MT) をつなぎ合わせた Cascade モデルにより実現されていた．それに対し，近年ではニューラルネットワークを用いた系列変換技術により，原言語の音声を直接目的言語のテキストに翻訳する End-to-End モデルの研究が進んでいる．

Cascade モデルにおいては，音声認識の結果に誤りが含まれている場合，機械翻訳の精度も大きく低下するという問題があり，音声認識の誤りに頑健な機械翻訳が必要となる．End-to-End モデルでは，事前学習された ASR，MT モデルによる Encoder と Decoder の初期化や，ASR，MT を Sub-task として，Main-task と同時に学習をする Multi-task Learning が精度向上のために不可欠である．しかし，一般的な Multi-task Learning では，正解と一致しない予測結果に対して等しく損失が計算される．そのため，正解系列と異なるが発音の似た予測結果と，発音の似ていない予測結果の損失が同じになる場合があり問題となる．よって，End-to-End モデルにおいても，Cascade モデルと同様に，音声認識出力の曖昧性を考慮した学習方法が必要となる．

本研究では，End-to-End 音声翻訳モデルにおいて，音声認識出力の曖昧性を考慮した Multi-task による学習方法を提案する．提案手法による実験と分析を通じて，本手法が音声翻訳における音声認識出力の曖昧性に対する頑健性の向上と，翻訳精度の向上に寄与することを確認した．

2 関連研究

End-to-End 音声翻訳における Multi-task Learning [1, 2] では，Main-task である ST-task を学習するだけでなく，Encoder とその隠れベクトルを共有した上で Sub-task である ASR-task の学習も同様にする．

Kano ら [3] は，Multi-task の仕組みを用いて，語順の差異の大きな英日 End-to-End 音声翻訳において，音声認識，機械翻訳といった比較的簡単なタスクから学習し，構造を組み替えながら最終的に，入力音声から出力系列を直接翻訳するモデルをカリキュラム学習する手法を提案し，翻訳精度を向上させた．

Chuang ら [4] は，Multi-task End-to-End 音声翻訳で，Sub-task である ASR-task の学習において，予測単語と正解単語の埋め込みベクトルのコサイン類似度を損失関数の計算に利用し，音声翻訳精度を向上させた．これは，ASR-task において，単語の意味的類似性を考慮することで，音声翻訳の頑健性を向上させる試みと言える．

Osamura ら [5] は，Cascade モデルの音声翻訳において，One-hot ベクトルの代わりに，音声認識の事後確率分布を表すベクトルを機械翻訳モデルへ渡し学習をすることで，音声認識誤りに対する頑健性を向上させた．

本研究では，Kano ら，Chuang らと同様に Multi-task の手法と，Osamura らの手法をもとに，End-to-End においても音声認識出力の曖昧性に頑健な音声翻訳の実現を試みる．

3 関連技術

3.1 End-to-End 音声翻訳

End-to-End 音声翻訳モデルは Encoder-Decoder モデルにより実現される．$\mathbf{X} = (x_1, ..., x_T)$ を原言語の入力音声に対する音響特徴量の系列とし，$\mathbf{Y} = (y_1, ..., y_N)$ を目的言語記号列とする．ここで，$y_i \in V$ であり，V は目的言語の語彙集合，T は音響特徴量のフレーム長，N は目的言語記号列のトークン数を表す．

v を語彙集合 V の元とすると，i 番目の目的言語

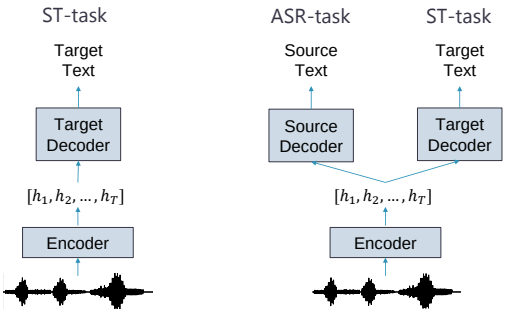

図 1　Single-task ST (左) と Multi-task ST (右)

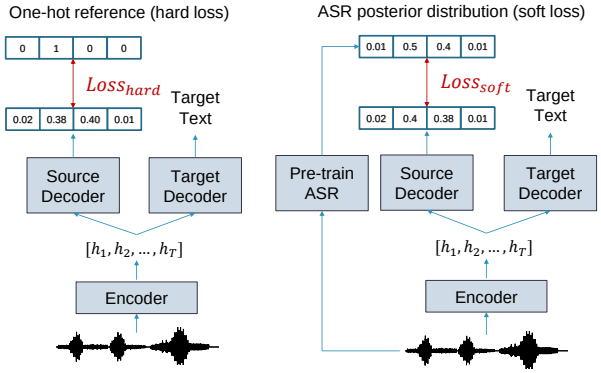

図 2　従来手法 (左) と提案手法 (右)

記号の事後確率は以下の式で表される

$$P_{\mathrm{ST}}(y_i = v) = p(v|\mathbf{X}, y_{<i}). \tag{1}$$

ST の学習時の損失関数 $\mathcal{L}_{\mathrm{ST}}$ は，cross entropy loss を用いて以下の式で表される

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{i=1}^{N}\sum_{v\in V}\delta(v, y_i)\log P_{\mathrm{ST}}(y_i = v). \tag{2}$$

式中の $\delta(v, y_i)$ は，$v = y_i$ のとき 1，そうでなければ 0 とする．

3.2　Multi-task End-to-End 音声翻訳

Multi-task 音声翻訳においては，音響特徴量が Encoder により隠れベクトルに変換され，その後 ST-task (Main-task) の Decoder と ASR-task (Sub-task) の Decoder の両方を用いて学習される．式 (1) (2) より，P_{ASR} を P_{ST} と同様に定義すると，ASR 学習時の損失関数 $\mathcal{L}_{\mathrm{ASR}}$ は以下の式で表される

$$\mathcal{L}_{\mathrm{ASR}} = -\sum_{i=1}^{N}\sum_{v\in V}\delta(v, y_i)\log P_{\mathrm{ASR}}(y_i = v). \tag{3}$$

ST-task の損失関数を $\mathcal{L}_{\mathrm{ST}}$，ASR-task の損失関数を $\mathcal{L}_{\mathrm{ASR}}$，$\mathcal{L}_{\mathrm{ASR}}$ に対する重みを W_{ASR} とすると，学習時全体の損失関数 $\mathcal{L}$ は以下の式で表される

$$\mathcal{L} = (1 - W_{\mathrm{ASR}})\mathcal{L}_{\mathrm{ST}} + W_{\mathrm{ASR}}\mathcal{L}_{\mathrm{ASR}}. \tag{4}$$

3.1 節で述べたモデルを Single-task ST，本節で述べたモデルを Multi-task ST とし，それぞれの概要図を図 1 に示す．本稿では，式 (3) における $\mathcal{L}_{\mathrm{ASR}}$ を，hard label (One-hot reference) による cross entropy loss として $\mathcal{L}_{\mathrm{hard}}$ で表す．

4　提案手法

本手法では，Multi-task End-to-End 音声翻訳モデルの ASR-task の学習の際に，事前学習された ASR の事後確率分布を reference として与え，ASR 出力の曖昧性を考慮するよう ST を学習する手法を提案する．

従来手法と提案手法の概要図を図 2 に示す．3.2 節で述べた従来手法では，正解と一致しない予測結果に対して等しく損失が計算されてしまい，発音の類似した予測結果と類似していない予測結果が同じように損失を計算され，ASR 出力の曖昧性を考慮できないと考えられる．

そこで提案手法では，ASR-task において，One-hot reference の代わりに，ASR 出力の曖昧性を表す ASR 事後確率分布 (ASR posterior distribution) のベクトルを reference として用いる．ASR 事後確率分布は，発音が似ている単語が同じようなスコアを持つことが期待されるため，どの単語が他のどの単語と発音が類似しているか，という情報を保持している．これを reference とすることで，音声認識出力の曖昧性に対して頑健な音声翻訳が学習できると期待できる．Osamura らの方法では，Cascade モデルにおいて，事前学習された ASR により出力される事後確率分布を，One-hot ベクトルの代わりに MT の入力とし，チューニングすることで音声認識出力の曖昧性に対して頑健な翻訳を実現した．それと比較すると，本手法では，ASR 事後確率分布をそのまま入力とするのではなく reference として用い，損失関数の計算方法を変える点で違いがある．

ASR 事後確率分布は，事前学習された ASR を用いて得られた，各トークンに対するスコアを持ったベクトルの softmax を取り，soft label とする．soft label において i 番目のトークン v のスコアを $P_{\mathrm{soft}}(i, v)$ とすると，提案手法よる損失 $\mathcal{L}_{\mathrm{soft}}$ は以下の式で表される

$$\mathcal{L}_{\mathrm{soft}} = -\sum_{i=1}^{N}\sum_{v\in V}P_{\mathrm{soft}}(i, v)\log P_{\mathrm{ASR}}(y_i = v). \tag{5}$$

表 1 実験に用いた Fisher CallHome Spanish コーパス[1)]

	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

本実験では，$\mathcal{L}_{\text{ASR}}$ を以下の式として定義し，$\mathcal{L}_{\text{hard}}$ と $\mathcal{L}_{\text{soft}}$ の割合を，重み W_{soft} で調整できるようにした

$$\mathcal{L}_{\text{ASR}} = (1 - W_{\text{soft}})\mathcal{L}_{\text{hard}} + W_{\text{soft}}\mathcal{L}_{\text{soft}}. \quad (6)$$

5 実験

本実験では，データセットとして，Fisher Spanish CallHome Spanish コーパス [6] を用い，Spanish-English (Es-En) の音声翻訳モデルを作成した．本データセットは，170 時間のスペイン語による電話での日常会話音声と書き起こし，それらに該当する英語テキストにより構成されている．音響特徴量は，Kaldi [7] により抽出した，3 次元の pitch が付加された 83 次元の Fbank+pitch を用いた．テキストは句読点，記号を取り除き小文字化し，音響特徴量はフレーム長 3000，テキストは文字数が 400 より大きいものを取り除いた．実験に用いたデータセットの内訳を表 1 に示す．ASR，ST モデルの学習，評価はともに表 1 のデータを用いた．Tokenizer は SentencePiece [8] により，最大語彙数 1000 として，train データからスペイン語と英語のトークンを共有した辞書を作成し，train, dev, test データに適用した．

ASR，ST モデルは ESPnet [9] を用い，Transformer [10] により作成した．モデルの設定を表 2 に示す．soft label の作成に必要な事前学習 ASR モデルは，学習後，dev データの WER が最も低いモデルを用いた．ST モデルは，学習後，dev データの BLEU スコア [11] が高いモデルを上から 5 つ取り出し，model averaging をし，最終的なモデルとして test データで評価した．

本実験では，式 (4) (6) における W_{ASR} を 0.3 とし，$W_{\text{soft}} = \{0.0, 0.3, 0.5, 0.7, 1.0\}$ の 5 つの場合に分けて実験をし，BLEU スコアにより評価した．ベースラインである cross entropy loss を用いた条件は $W_{\text{soft}} = \{0.0\}$ に対応し，提案手法である soft loss を用いた条件は $W_{\text{soft}} = \{0.3, 0.5, 0.7, 1.0\}$ に対応する．

1) train データは，音声のスピードを 0.9 倍と 1.1 倍に調整したものを加えるデータ拡張をした．

表 2 ASR，ST モデルの設定

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

6 実験結果と分析

Fisher CallHome Spanish コーパスの test データによる BLEU スコアの結果を表 3 に示す．softW_{soft}-hard$(1 - W_{\text{soft}})$ が，各 W_{soft} のパラメータでの実験を表す．例として，soft0.0-hard1.0 が，$W_{\text{soft}} = \{0.0\}$ での実験を表す．また，BLEU 4-ref は 4 つの reference を用いた BLEU スコア，BLEU 1-ref は 1 つの reference を用いた BLEU スコア，↘はベースラインと比較し BLEU スコアが低下した結果，**太字**は BLEU スコアが最も向上した結果を表す．

実験結果から，test データ全体として見ると，ベースラインと比較し，提案手法で BLEU スコアが低下したものよりも，向上したものの方が多いことが分かる．また，Fisher test データにおいては，いずれの場合においても，提案手法がベースラインを上回る結果となった．また，いずれの test データにおいても，提案手法である soft1.0-hard0.0，soft0.5-hard0.5 の場合の結果が，ベースラインである soft0.0-hard1.0 の結果を上回った．よって，提案手法が音声翻訳の精度改善に効果的であるということが分かった．

soft0.5-hard0.5 での Fisher test における出力結果を表 4 に示す．例 1 では，ベースラインにおいて，"intensive"と出力されるべきところが，"unthinkable"と誤って出力されている．ここで，"unthinkable"のスペイン語としては"inconcebible", "impensable"が該当する．これらは正解である"intensive"に発音が似ており，ASR-task の出力でこれらの単語を予測して

表 3 test データでの BLEU スコアの結果 (↘ はベースラインと比較し BLEU スコアが低下した結果)

		Baseline	Proposed			
		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

表 4 Fisher test の出力結果の例 (soft0.5-hard0.5)

例 1

Label	20051028_180633_356_fsp-A-016164-016487
Ground Truth (Es)	sí pero o sea sigue siendo bastante intensi
Ground Truth (En)	yes but it's still pretty intensive
Baseline (En)	yes but that keeps getting pretty unthinkable
Proposed (En)	yes but that keeps being pretty intense

例 2

Label	20051028_180633_356_fsp-A-033453-034134
Ground Truth (Es)	es es mejor en el sentido que uno okay que hay menos riesgos pero ay
Ground Truth (En)	that is the best in the sense that one okay that there are less risks but ay
Baseline (En)	it's it's better in the sense that you don't that there are less colds but there are
Proposed (En)	it's it's a best in the sense that you don't that there are less risks but

しまった結果，誤った翻訳結果を出力したと考えられる．これに対し提案手法では，"intensive"に近い"intense"に翻訳できている．例 2 では，ベースラインにおいて，"risks"と出力すべきところを"colds"と誤訳している．ここで，"colds"のスペイン語として，"resfriados"が該当し，正解である"riesgos"に発音が似ている．そのため，ASR-task における予測誤りをもとに，誤った出力をしたと考えられる．これに対し提案手法では，正しく"risks"と翻訳できている．

これらの例においては，ベースラインと比較し，提案手法の方では音声認識出力の曖昧性を考慮して，正しく出力できていると考えられる．

7 まとめと今後の展望

本研究では，音声認識の事後確率分布を用いて，End-to-End 音声翻訳の学習をする方法を提案し，音声認識出力の曖昧性に対して頑健な音声翻訳を期待した．実験結果から，提案手法により BLEU スコアの向上が見られ，本手法の有効性を確認できた．

今後の課題として，ASR-task の出力分布の可視化をした上での定量的な分析，label smoothing loss [12] を用いた実験での有効性の検証が必要だと考えられる．

また，従来研究において，Salesky ら [13] は，音素情報を読み情報として ST に用いることで，BLEU スコアを向上させ，読み情報が ST の学習に効果的であると報告している．それを踏まえた上で，音素，音節情報を損失関数の計算に用いて，字面から組み取れない読みの曖昧性を考慮した ST モデルを作成できると期待しており，今後実験を進める予定である．

謝辞

本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

参考文献

[1] Antonios Anastasopoulos and David Chiang. Tied multitask learning for neural speech translation. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 82–91. Association for Computational Linguistics, 2018.

[2] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pp. 1123–1127. ISCA, 2019.

[3] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. End-to-end speech translation with transcoding by multitask learning for distant language pairs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1342–1355, 2020.

[4] Shun-Po Chuang, Tzu-Wei Sung, Alexander H. Liu, and Hung-yi Lee. Worse wer, but better bleu? leveraging word embedding as intermediate in multitask end-to-end 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. 5998–6003. Association for Computational Linguistics, 2020.

[5] Kaho Osamura, Takatomo Kano, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. Using spoken word posterior features in neural machine translation. *Proceedings of the 15th International Workshop on Spoken Language Translation, 181-188, Oct. 2018*.

[6] Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association, 2004.

[7] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, No. CONF. IEEE Signal Processing Society, 2011.

[8] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018.

[9] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. Espnet: End-to-end speech processing toolkit. *Proc. Interspeech 2018*, pp. 2207–2211, 2018.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[12] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[13] 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.