音声認識出力の曖昧性を考慮した Multi-task End-to-end 音声翻訳と曖昧性の高い音声入力に対する頑健性の分析

胡 尤佳 [1]　須藤 克仁 [1,2]　中村 哲 [1,3]　Sakriani Sakti[1]
[1] 奈良先端科学技術大学院大学　[2] 奈良女子大学
[3]The Chinese University of Hong Kong, Shenzhen
{ko.yuka.kp2,sudoh,s-nakamura,ssakti}@is.naist.jp

概要

音声翻訳は原言語の音声を目的言語の音声やテキストへ変換する技術であり，近年では原言語の音声を直接目的言語のテキストへ翻訳する End-to-end 音声翻訳の研究が進んでいる．本研究では，より音声認識出力の曖昧性を考慮した損失関数を用いた Multi-task 音声翻訳モデルの学習方法を提案し，Hybrid CTC/Attention loss の仕組みへの適用を試みた．実験と分析により，従来モデルと比較し，提案モデルにおける翻訳性能の向上が見られ，曖昧性をある程度多く含む音声入力に対する頑健性の向上が見られることを確認した．

1　はじめに

音声翻訳（Speech Translation; ST）は，原言語の音声を入力とし，目的言語のテキストを出力する技術である．これまで，音声認識モデル（Automatic Speech Recognition; ASR）と機械翻訳モデル（Machine Translation; MT）を組み合わせた Cascade モデルによって実現されてきた．一方，近年ではニューラルネットワークを活用した系列変換技術を用い，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 モデルの研究が進展している．Cascade モデルには，音声認識結果に誤りが含まれる場合に機械翻訳の精度が大きく低下するという課題があり，音声認識の誤りに対して頑健な翻訳モデルの開発が求められる一方，End-to-end モデルでは，事前学習済みの ASR や MT モデルを利用して Encoder や Decoder を初期化したり，ASR や MT を Sub-task として Main-task と同時に学習する Multi-task Learning が精度向上のために必要とされる．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損失が同じように計算されるため，

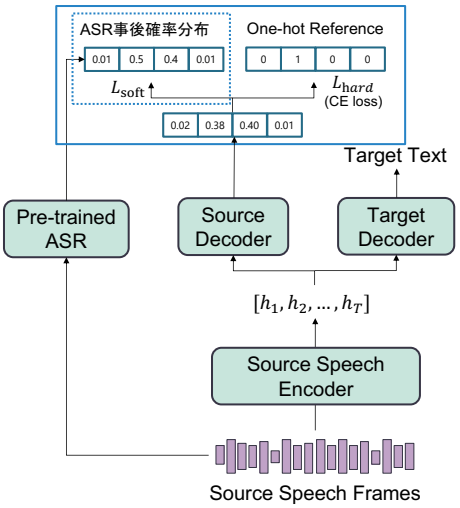


図 1　音声認識出力の曖昧性を考慮した Multi-task End-to-end 音声翻訳．

正解とは異なるが発音が類似している結果と，発音が全く異なる結果が同じ損失として扱われる問題がある．そのため，End-to-end モデルにおいても Cascade モデルと同様に，音声認識結果の曖昧性を考慮した学習手法が必要となる．このような課題を踏まえ，従来の提案手法 [1] では，End-to-end ST において，音声認識結果の曖昧性を考慮した損失関数を提案し，Cross-entropy (CE) loss を用いた Multi-task End-to-end ST モデルへ適用し，効果を確認した．近年では，損失計算において CE loss のみでなく，CTC loss [2] も同時に用いる Hybrid CTC/Attention loss [3] により性能が向上する傾向が見られ，ASR，ST で広く採用される損失の一つとなっている．そこで本研究では，提案手法である音声認識結果の曖昧性を考慮した損失関数を Hybrid CTC/Attention loss [3] へ適用し，ST での有効性を検証する．実験により，提案手法が Hybrid CTC/Attention loss をベースにした Multi-task End-to-end ST においても，音声翻訳の性能向上に寄与することを確認した．また分析によ

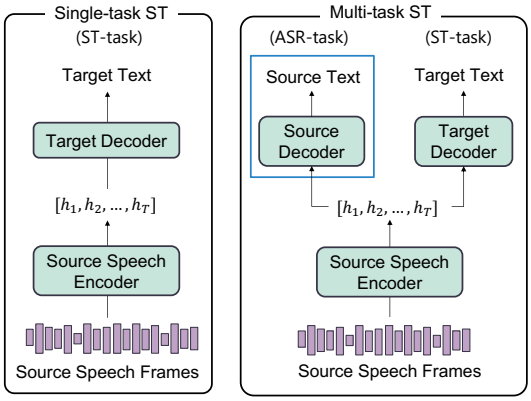


図 2 Single-task ST (左) と Multi-task ST (右).

り，従来モデルと比較し提案モデルでは，曖昧性をある程度含む音声入力に対して翻訳性能の向上がより見られることを確認した．

2 関連研究

近年の End-to-end 音声翻訳においては，性能向上のためのアプローチとして，Multi-task Learning [4, 5] が採用されている．Chuang ら [6] は，Multi-task learning を取り入れた End-to-end 音声翻訳において，ASR-task の学習時に予測単語と正解単語の埋め込みベクトル間のコサイン類似度を損失関数に取り入れる手法を提案した．この手法は，ASR-task における単語の意味的類似性を考慮した学習を行うことで，音声翻訳の頑健性の向上を試みている．また，Osamura ら [7] は，Cascade モデルを用いた音声翻訳において，音声認識結果を One-hot ベクトルではなく，事後確率分布を表すベクトルとして機械翻訳モデルに入力として与え学習する手法を提案し，音声認識誤りに対する頑健性を向上させた．提案手法では，Chuang らの Multi-task Learning の手法と，Osamura らの Cascade ST で ASR 出力を用いる学習手法をもとに，End-to-end ST においても音声認識結果の曖昧性に対する頑健性を高める音声翻訳の実現を目指している．

3 関連技術

3.1 Single-task 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，$\mathbf{X} = (x_1, ..., x_T)$ を原言語の入力音声の音響特徴量の系列，$\mathbf{Y} = (y_1, ..., y_N)$ を目的言語のトークン系列とする．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 は目的言語系列のトークン数を表す．v を語彙集合 V の元とすると，i 番目の目的言語記号の事後確率 $P_{\mathrm{ST}}(y_i = v)$ は以下の式で表される．

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

また，ST の学習時の損失関数 $\mathcal{L}_{\mathrm{ST}}$ は，CE loss を用いると以下の式で表される．

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

ここで $q(y_i, v)$ は，$y_i = v$ の場合は 1，そうでなければ 0 となることを表す[1)]．

3.2 Multi-task End-to-end 音声翻訳

ASR-task を取り入れた Multi-task End-to-end 音声翻訳では，Single-task End-to-end ST と同様，まず入力となる音響特徴量が Encoder により隠れベクトルに変換される．その後，Main-task である ST-task の Decoder と Sub-task である ASR-task の Decoder の両方が用いてモデルが学習される．式 (1) (2) をもとに，P_{ASR} を P_{ST} と同様に定義した上で，ASR 学習時の損失関数 $\mathcal{L}_{\mathrm{ASR}}$ は以下の式で表される．

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

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 λ_{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．

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

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 2 に示す．ここでは，式 (3) における $\mathcal{L}_{\mathrm{ASR}}$ が，hard label (One-hot reference) による CE loss として $\mathcal{L}_{\mathrm{hard}}$ で表されている．

近年では，Attention ベースの Encoder-Decoder モデルと CTC モデルの両方を利用するハイブリッドアプローチが提案され，ASR，ST の性能向上のために広く用いられている．CTC loss [2] では，音声とテキスト間のアライメントを Forward-backward アルゴリズムを用いて学習する．近年の ASR の学習に広く用いられているものの，CTC ベースの ASR のフレームワークには言語モデルが含まれていないことで性能が低下してしまう課題があった．そこ

1) CE loss では，過学習を避けるために label smoothing[8, 9] を適用することが一般的であり，本研究における CE loss では重み 0.1 の label smoothing を導入している．

で，Hybrid CTC/Attention [3] のようなハイブリッドアプローチを採用することで，CTC ベースの ASR に言語モデル制約を含めることが可能となる．本研究では、Multi-task ST における ASR-task loss $\mathcal{L}_{\text{ASR}}$ に Hybrid CTC/Attention を採用した．この場合 $\mathcal{L}_{\text{ASR}}$ は，CTC loss と Attention ベースの loss（ここでは CE loss）である $\mathcal{L}_{\text{Att}}$ の割合を調整する重み λ_{CTC} を用いて以下の式で表される．

$$\mathcal{L}_{\text{ASR}} = (1 - \lambda_{\text{CTC}})\mathcal{L}_{\text{Att}} + \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}}. \tag{5}$$

4 提案手法

Multi-task ST での 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 ST を学習する手法を提案する．提案手法の概要図を図 1 に示す．3.2 節における従来手法では，正解と一致しない予測結果に対して損失が等しく計算されることで，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る場合があり，ASR 出力の曖昧性を考慮したモデルの学習が難しいと考えられる．そこで提案手法では，ASR-task において，正解の One-hot reference のみでなく，ASR 出力の曖昧性を表す ASR 事後確率分布のベクトルも reference として用いる．ASR の事後確率分布は，ある単語が他の単語とどれほど発音が類似しているかという情報をスコアで表し，ASR の事後確率分布を用いてモデルを学習することで，音声認識出力の曖昧性に対して頑健な音声翻訳が期待できる．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 $P_{\text{soft}}(i, v)$ とすると，提案手法で追加される損失 $\mathcal{L}_{\text{soft}}$ は以下の式で表される

$$\mathcal{L}_{\text{soft}} = -\sum_{i=1}^{N}\sum_{v\in V} P_{\text{soft}}(i, v) \ln P_{\text{ASR}}(y_i = v). \tag{6}$$

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．P_{ASR} は，ST モデル内の ASR Decoder から取得された確率分布であり，事前学習済み ASR モデルから得られる P_{soft} とは異なる．$\mathcal{L}_{\text{soft}}$ を元の CE loss である $\mathcal{L}_{\text{hard}}$ と，重み λ_{soft} により加重混合することで，式 (5) における $\mathcal{L}_{\text{Att}}$ は次のように表され，本研究では ASR Posterior-based Loss (ASR-PBL) と呼ぶ．

$$\mathcal{L}_{\text{Att}} = (1 - \lambda_{\text{soft}})\mathcal{L}_{\text{hard}} + \lambda_{\text{soft}}\mathcal{L}_{\text{soft}}. \tag{7}$$

5 実験

実験では，Fisher Spanish Corpus [10] を用い，スペイン語音声から英語テキストへの音声翻訳モデルを作成した．ASR，ST モデルは ESPnet [11] を用い，Transformer [12] ベースのモデルを作成した．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．本実験では，式 (5) における λ_{CTC} は 0.5 に固定し，式 (4) (7) における λ_{ASR} を $\{0.3, 0.4, 0.5\}$，$\lambda_{\text{soft}} = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ の場合に分けて実験をした．Fisher dev を検証データ，Fisher dev2, test を評価データとして用い，検証データを用いた際に，それぞれの手法において最も良い BLEU スコアとなったモデルを採用し，評価データにより評価した．その他の実験設定は付録 A.1 に記載する．

6 実験結果と分析

Fisher dev, dev2, test データにおける BLEU スコアの結果を表 1 に示す．実験結果から，提案手法がいずれの評価データにおいても従来手法と比較し BLEU が向上し，提案手法が Hybrid CTC/Attention loss ベースの Multi-task ST においても，性能向上に寄与することが分かった．

また，提案モデルの曖昧性の高い音声入力に対する頑健性が高まっているかを検証するため，音声入力の曖昧性の大きさを ASR の WER を測り，異なる WER の範囲における BLEU の平均を従来モデルと提案モデルで比較した．Fisher test における各 ASR WER の範囲における BLEU を図 3 に示し，各 ASR WER の範囲における音声入力のサンプル数を図 4 に示した[2)]．図 3 より，WER が 5%から 40%の範囲でほとんどの場合，提案手法が従来手法よりも良いスコアを得ていることがわかった．それに対し，0% および 0%から 5%の範囲では提案手法による向上が見られなかった．これらのサンプルは，ASR による誤りがほとんど含まれておらず，図 4 をみると，0% から 5%の範囲ではサンプル数が非常に少なく，全体の翻訳性能の変動には大きく影響しないと見ら

2) 横軸の WER は soft label を得るために用いた事前学習済み ASR に基づいて計算されている．各範囲での BLEU は，SacreBLEU [14] を使用し Corpus BLEU を計算した．また，WER が 50%を超えるサンプルの出現数は少なかったため，除外している．

表 1 Fisher dev, dev2, test における BLEU の結果．実験における λ_{CTC} は 0.5 に固定している．

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	Transformer ASR-MTL [13]	-	-	46.64	47.64	46.45
	CE (従来手法)	0.5	-	47.18	47.43	46.59
	ASR-PBL (提案手法)	0.3	0.7	47.20	**48.36**	**46.82**

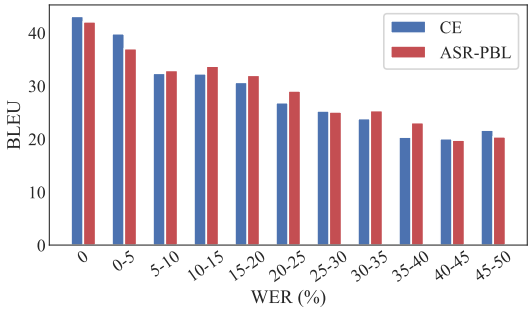


図 3 Fisher test における各 WER での従来手法と提案手法の BLEU の比較.

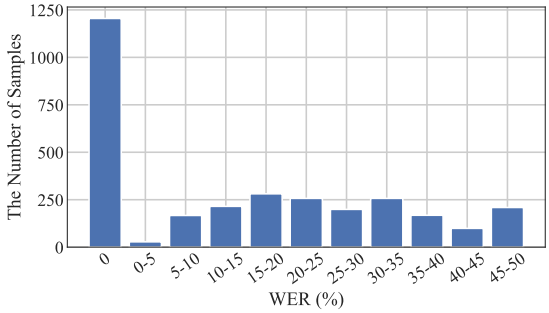
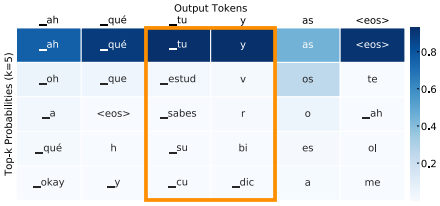


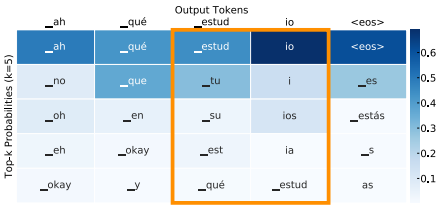
図 4 Fisher test における各 WER でのサンプル数の違い.

れる．また，0%のサンプルについては，1 文字や 2 文字といった短い単語で構成されるものが多く，認識が容易なものがほとんどであった．これらの結果は，提案手法が WER がとても小さいサンプル (0%および 0%から 5%の範囲) と，とても大きいサンプル (40%より大きい範囲) では有効性が見られないものの，約 5%から 40%といったある程度の WER の範囲で曖昧性を多く含む音声に対しては有効に利用でき，頑健性が向上したことを示している．

従来手法および提案手法における，ASR-task での上位 5 トークンの事後確率スコアを図 5 に示す．従来手法の確率分布では，最上位のトークン（例：**_tu** や **y**）に高い確率が割り当てられており，ST モデルが ASR-task で **estudias** と出力される可能性をほとんど考えていない．それに対して，提案手法



(a) CE ASR-task (従来手法)



(b) ASR-PBL ASR-task (提案手法)

図 5 上位 5 トークンの出力確率分布のスコアの一例．実際の出力結果を付録 A.2 に示す.

の事後確率では，**_estud** だけでなく**_tu** の事後確率も高くなっている．提案手法では **tuyas** の単語の一部の**_tu** に対してもスコアが大きく与えながら，**estudias** の一部である**_estud** が最も可能性が高いと考え，結果を **study** として正しく翻訳することができている．この結果は，提案モデルが **estudio**，**tuyas**，および **estudias** の間の発音の類似性，曖昧性に基づいて出力を予測できたことを示唆している．

7 まとめと今後の展望

本研究では，音声認識出力の曖昧性に対して頑健な音声翻訳のさらなる性能向上を目指し，音声認識出力の事後確率分布を用いた学習手法を，Hybrid CTC/Attention loss ベースの音声翻訳に適用した．実験結果から，提案手法による BLEU スコアの向上が見られ，ある程度曖昧性を含む音声入力に対しての頑健性がより高まっていることを示した．今後の課題として，提案手法の同時音声翻訳への適用や，英日のような他言語への適用が考えられる．

謝辞

本研究は JSPS 科研費 JP21H05054，JP21H03467，JP24KJ1695 の助成を受けたものである．

参考文献

[1] Yuka Ko, Katsuhito Sudoh, Sakriani Sakti, and Satoshi Nakamura. ASR Posterior-based Loss for Multi-task End-to-end Speech Translation. **Proc. Interspeech 2021**, pp. 2272–2276, 2021.

[2] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In William W. Cohen and Andrew W. Moore, editors, **Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006**, Vol. 148 of **ACM International Conference Proceeding Series**, pp. 369–376. ACM, 2006.

[3] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. **IEEE Journal of Selected Topics in Signal Processing**, Vol. 11, No. 8, pp. 1240–1253, 2017.

[4] 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.

[5] 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.

[6] 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.

[7] 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**.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In **2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016**, pp. 2818–2826. IEEE Computer Society, 2016.

[9] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, **Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada**, pp. 4696–4705, 2019.

[10] 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.

[11] 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.

[12] 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.

[13] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. ESPnet-ST: All-in-One Speech Translation Toolkit. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020**, pp. 302–311. Association for Computational Linguistics, 2020.

[14] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[15] 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.

[16] 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.

[17] 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.

A　付録 (Appendix)

A.1　実験設定

実験て用いた Fisher Spanish Corpus 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．実験に使用したデータのサイズ，事前学習した ASR モデルの WER，soft label のサイズと 1-best WER（その際の Decoding beam size）を表 2 に示す．音響特徴量は，Kaldi [15] 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．Tokenizer は SentencePiece [16] 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．ST モデルは，学習後，dev データの BLEU スコア [17] 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．

表 2　実験で用いた Fisher データのサイズ，事前学習した ASR モデルの WER，soft の label 1-best WER.

Model	dev	dev2	test	Soft labels 1-best
Data size	3.9k	3.9k	3.6k	415.8k
Decoding beam size		10		1
Pre-trained ASR WER	30.2	29.1	27.2	9.3

A.2　出力結果の例

従来手法と提案手法の出力結果の例を表 3 に示す．

表 3　従来手法と提案手法の出力結果例.

	ASR output	ST output
Example		
Reference	ah qué *estudias*	oh what do you *study*
CE (従来手法)	ah qué tuyas	ah what are you doing
ASR-PBL (提案手法)	ah qué **estudio**	oh what do you **study**