E-019

同時通訳者音声認識に向けた原言語テキストを補助入力とする Transformer 音声認識

Transformer Based Automatic Speech Recognition with Auxiliary Input of Source Language Text Toward Transcribing Simultaneous Interpretation

谷口 秀太† 加藤 恒夫† 田村 晃裕† 安田 圭志‡
Shuta Taniguchi Tsuneo Kato Akihiro Tamura Keiji Yasuda

1. はじめに

同時通訳者の訓練プログラムでは，訓練生はさまざまな通訳テストを受ける．評価者が通訳品質を評価する際には，音声を繰り返し再生する必要がなく，また声に対する主観が評価に影響しないように書き起こしテキストを使用する．そこで音声認識を用いて訓練生の発話を自動的に書き起こすことで評価者の業務を軽減することが期待される．同時通訳者音声には，言い淀み，フィルドポーズ，中断，言い直し等が含まれ，認識が難しい．一方で，同時通訳者訓練用の音声認識は，テスト素材が予め用意されているため，通訳者の翻訳元音声を書き起こした原言語テキストを利用できるという利点がある．本研究では原言語テキストを補助入力として利用することで，その精度向上を目指す．

本論文では，Transformer アーキテクチャ[1]に基づき，音声と原言語テキストを入力とする End-to-end のマルチモーダル音声認識を提案する．本モデルは音声認識の後段に原言語テキストを用いて誤り訂正を行う誤り訂正モデルではなく，処理の途中で 2 つのモダリティを統合し，そのモデルパラメータを End-to-end で最適化する Transformer ベースのエンコーダ・デコーダモデルである．

ただし，現在大規模な同時通訳者コーパスは存在しない．そこで，音声翻訳用の大規模コーパスを用いて原言語テキスト付音声認識のシミュレーションを行った．具体的には音声翻訳の翻訳先言語を原言語テキストとして扱い，原言語テキストがある場合とない場合の音声認識の精度を評価した．

2. 原言語テキストを補助入力とする Transformer 音声認識

提案する原言語テキストを補助入力とする Transformer 音声認識モデルは，音声と原言語テキストを入力するための 2 つの Transformer エンコーダと，認識単語を出力する 1 つの Transformer デコーダで構成される．図 1 にモデルの構造を示す．左から右へ 3 列のブロックは，それぞれテキストエンコーダ，音響エンコーダ，デコーダを表す．ブロックの色分けは Transformer ブロックの種類を表す．灰色のブロックは，マルチヘッド自己注意機構と全結合層に続いて，残差接続，レイヤー正規化，ドロップアウトを持ち，ソースターゲット間の注意機構を持たない．オレンジ色と緑色のブロックはマルチヘッド自己注意機構と全結合層の間に，それぞれ 1 つと 2 つのマルチヘッドソースターゲット注意機構を持つ．

デコーダの Transformer ブロックは，全て音響エンコーダの出力に対してマルチヘッドソースターゲット注意機構を持つ．さらにデコーダの緑色の Transformer ブロックと音響エンコーダのオレンジ色の Transformer ブロックはテキストエンコーダの出力に対するマルチヘッドソースターゲット注意機構を持つ．

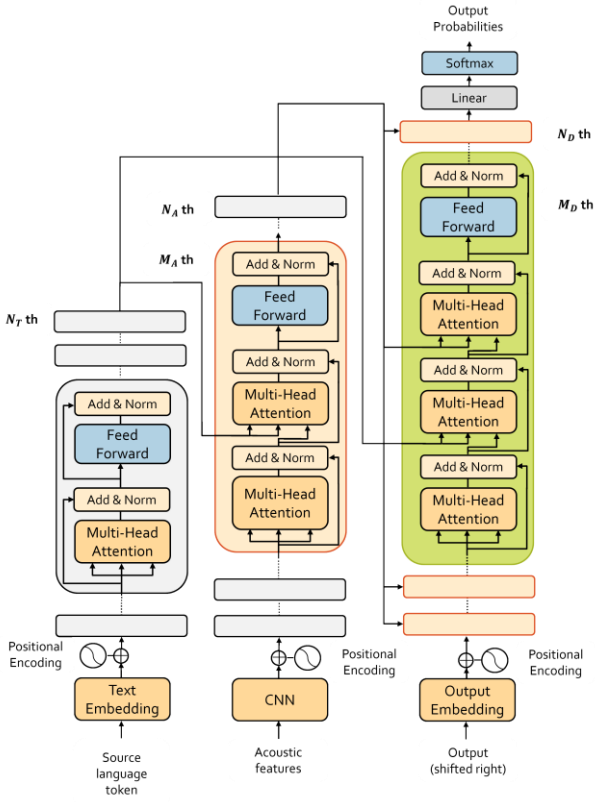


図 1 原言語テキストを補助入力とする Transfornmer 音声認識

2.1 テキストエンコーダ

図 1 の左列のテキストエンコーダはトークン化された原言語テキスト$\boldsymbol{S} = \{s_1, s_2, \ldots s_L\}$を受け取り，$d_M$次元ベクトル$\overline{\boldsymbol{S}}$に埋め込み，位置エンコーディングを付加する．この処理は以下のとおり表される．

$$\overline{\boldsymbol{S}} = \mathrm{TextEmbed}(\boldsymbol{S})$$
$$\boldsymbol{X_S^0} = \overline{\boldsymbol{S}} + \mathrm{PositionEnc}$$

ここで，$\boldsymbol{X_S^0}$は L×d_Mの行列で，最初のテキストエンコーダブロックへの入力を示す．N_T層あるうちの i 番目（$1 \leq i \leq N_T$）の Transformer エンコーダブロック$\mathrm{TfEncBlk}(\cdot)$は，前段のブロックの出力$\boldsymbol{X_S^{i-1}}$を受け取り，同じサイズの行列である$\boldsymbol{X_S^i}$を出力する．

$$\boldsymbol{X_S^i} = \mathrm{TfEncBlk}^i(\boldsymbol{X_S^{i-1}})$$

図 1 に示すように，Transformer エンコーダブロック$\mathrm{TfEncBlk}^i(\cdot)$は，マルチヘッド自己注意機構と全結合層から構成され，それぞれに残差接続，レイヤー正規化が後続する．

† 同志社大学 Doshisha University
‡ マインドワード株式会社 MINDWORD, Co.Ltd,

表 1　デコーダのテキストエンコーダに対する注意機構の位置(M_D)を変化させた時の MuST-C NI 開発セットに対する単語誤り率(WER, %)

	M_A	M_D	WER[%]
1)	11	1	12.7
2)	11	2	12.2
3)	11	3	12.4
4)	11	4	**11.4**
5)	11	5	**11.4**
6)	11	6	11.9

表 2　音響エンコーダのテキストエンコーダに対する注意機構の位置(M_A)を変化させた時の MuST-C NI 開発セットに対する単語誤り率(WER, %)

	M_A	M_D	WER[%]
1)	1	4	12.4
2)	4	4	12.7
3)	7	4	11.7
4)	9	4	11.5
5)	11	4	**11.4**
6)	12	4	12.1

表 3　デコーダの音響エンコーダ，テキストエンコーダに対する注意機構の個数を変化させた時の MuST-C NI 開発セットに対する単語誤り率(WER, %)

	M_A	M_D	WER[%]
1)	11	-	12.3
2)	-	4	11.8
3)	11	4	**11.4**
4)	1:12	4	12.2
5)	11	1:6	**11.4**
6)	1:12	1:6	11.7

表 4　MuST-C En-NI, En-De, Ted-EN-JP, CoVoST2 En-Ja における文書量と時間

	Train	Dev	Test
MuST-C En-Nl	248.3K 434h12m	1.4K 2h33m	2.6K 4h09m
MuST-C En-De	229.7K 400h2m	1.4K 2h33m	2.6K 4h09m
Ted-EN-JP	133.6K 208h37m	1K 1h30m	2K 3h05m
CoVoST En-Ja	289.4K 364h	15.5K 26h	15.5K 25h

表 5　MuST-C En-NI, En-De, Ted-EN-JP, CoVoST2 En-Ja の各データセットにおけるベースラインと提案モデルの単語誤り率(WER, %)．提案モデルはデコーダのテキストエンコーダに対する注意機構の位置を(M_D=4)とし音響エンコーダのテキストエンコーダに対する注意機構の位置を(M_A=11)とした．

	MuST-C En-Nl	MuST-C En-De	Ted-EN-JP En-Ja	CoVoST2 En-Ja
ベースライン	14.4	14.6	12.1	28.7
提案モデル	**9.5**	**10.6**	**10.9**	**21.2**

2.2 音響エンコーダ

図 1 の中央の音響エンコーダは，音響特徴量の時系列 $\boldsymbol{A} = \{a_1, a_2, \dots a_M\}$を受け取り，2次元畳み込みニューラルネットワーク（2D-CNN）で音響特徴量の局所的な時間パターンを抽出し，位置エンコーディングを付加する．その処理は以下のように示される．

$$\overline{\boldsymbol{A}} = \mathrm{Conv2D}(\boldsymbol{A})$$
$$\boldsymbol{X_A^0} = \overline{\boldsymbol{A}} + \mathrm{PositionEnc}$$

ここで，M×d_Mの行列$\boldsymbol{X_A^0}$は，最初の音響エンコーダブロックの入力である．N_A層あるうちの j 番目（1≦j≦N_A）のTransformerエンコーダブロックに対してテキストエンコーダと同じ方法で$\boldsymbol{X_A^j}$を出力する．ただしM_A番目のブロックだけテキストエンコーダの出力に対するソースターゲット注意機構を持つ．

$$\boldsymbol{X_A^j} = \begin{cases} \mathrm{TfEncBlk}^j\left(\boldsymbol{X_A^{j-1}}\right) & j \neq M_A \\ \mathrm{TfEncAttnBlk}^j\left(\boldsymbol{X_A^{j-1}}, \boldsymbol{X_S^{N_T}}\right) & j = M_A \end{cases}$$

M_A番目のエンコーダブロックは，マルチヘッド自己注意機構と全結合層の間にマルチヘッドソースターゲット注意機構を持ち，それぞれに残差接続，レイヤー正規化が続く．

2.3 デコーダ

図 1 の右側のデコーダは，認識したトークンを自己回帰的に出力する．デコーダは，入力用のトークン$\boldsymbol{W_{0:t-1}} = \{w_0, w_1, \dots, w_{t-1}\}$，$w_0$ =<s>を受け取り，d_M次元ベクトルの時系列$\boldsymbol{W_{0:t-1}}$に埋め込み，以下のように位置エンコーディングを付加する．

$$\overline{\boldsymbol{W}}_{\boldsymbol{0:t-1}} = \mathrm{TextEmbed}(\boldsymbol{W_{0:t-1}})$$
$$\boldsymbol{Z_t^0} = \overline{\boldsymbol{W}}_{\boldsymbol{0:t-1}} + \mathrm{PositionEnc}$$

ここで，$\boldsymbol{Z_t^0}$はt×d_M 行列であり，最初のデコーダブロックに対する t 番目の入力を示す．N_D層あるうちの k 番目のTransformerデコーダブロック（1≦k≦N_D）は，音響エンコーダの出力と前段のTransformerデコーダブロックの出力を受け取り，$\boldsymbol{Z_t^k}$を出力する．ただし，M_D番目のデコーダブロックはテキストエンコーダの出力に対して追加のマルチヘッドソースターゲット注意機構を有する．

$$\boldsymbol{Z_t^k} = \begin{cases} \mathrm{TfDecAttnBlk}^k\left(\boldsymbol{Z_t^{k-1}}, \boldsymbol{X_A^{N_A}}\right) & k \neq M_D \\ \mathrm{TfDec2AttnBlk}^k\left(\boldsymbol{Z_t^{k-1}}, \boldsymbol{X_S^{N_T}}, \boldsymbol{X_A^{N_A}}\right) & k = M_D \end{cases}$$

最後に線形変換に続くソフトマックス関数から，最も確率の高い単語$\hat{w}_t$を自己回帰的に出力する．

$$p(V) = \mathrm{Softmax}(\mathrm{Linear}(\boldsymbol{Z_t^{N_D}}))$$
$$\hat{w}_t = \underset{v \in V}{\mathrm{argmax}}\, p(v)$$

ここで，V と$p(V)$は語彙と事後確率を表す．

2.4 教師あり学習

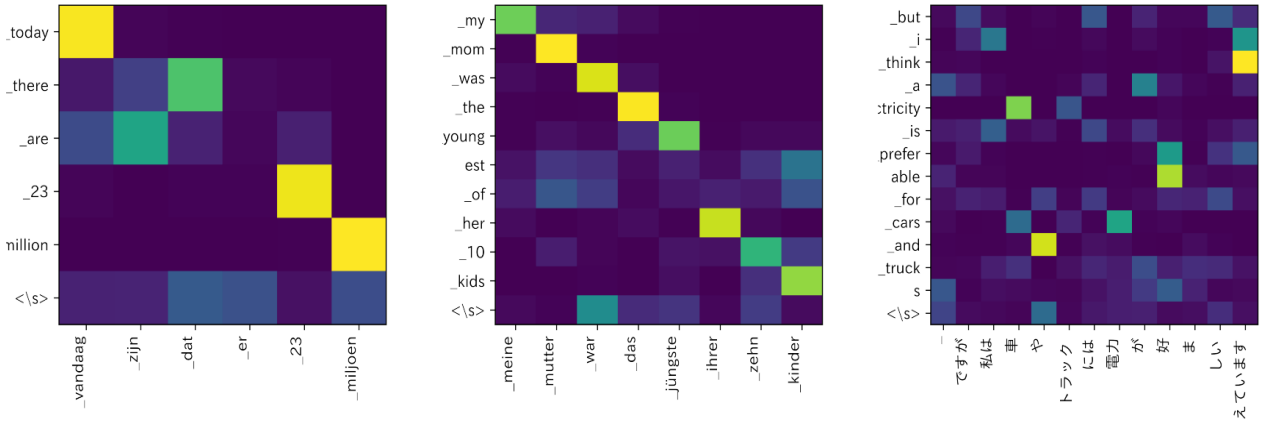
提案モデルは、音声と原言語テキストを入力とし，書き起こしテキストを教師信号として End-to-end で学習される．次式のクロスエントロピー誤差を最小にする．

$$L = -\sum_{t=1}^{T} \log p(w_t | w_{0:t-1}, \boldsymbol{S}, \boldsymbol{A})$$

ここで Tはトークン列の長さである

表 6　MuST-C En-Nl，MuST-C En-De，TED-EN-JP の各データセットにおいて，提案モデルによって改善された認識文の例．提案モデルで修正された単語と原文の単語には下線を引いた．

	A)MuST-C En-Nl	B) MuST-C En-De	C) TED-EN-JP
書き起こし	today there are 23 million	my mom was the youngest of her 10 kids	but i think electricity is preferable for cars and trucks
原言語テキスト	Vandaag zijn dat er 23 miljoen	meine mutter war das jungste ihrer zehn kinder	ですが私は車や トラックには電力が 好ましいと考えています
ベースライン認識結果	there are 23 billion	my mom was the young as the pretend kid	but i think a electricity is profitable for cars and drugs
提案モデルの認識結果	today there are 23 million	my mom was the youngest of her 10 kids	but i think a electricity is preferable for cars and trucks



A) MuST-C En-Nl　　*B) MuST-C En-De*　　*C) TED-EN-JP*

図 2　表 6 に示した３つの例に対する提案モデルのデコーダからテキストエンコーダへの注意機構．横軸は原言語トークン，縦軸は認識されたトークンを示す．黄色いセルは 1 に近い重み，暗いセルは 0 に近い重みを表し，提案モデルで修正されたトークンには対応する原言語トークンまたはそれに近いトークンが注目されている．

3. 実験

提案モデルを 2 段階で評価した．まず，開発セットを用いて，デコーダと音響エンコーダのどのブロックがテキストエンコーダに対してマルチヘッドソースターゲット注意機構を持つべきかを検討した．次に，テキストエンコーダを持たず，原言語テキストを用いないベースラインモデルである Transformer と提案モデルの性能を比較した．性能の測定は単語誤り率(Word Error Rate,WER)で行った．

3.1　データセット

評価は，MuST-C[2]のドイツ語訳(En-De)とオランダ語訳(En-Nl)，CoVoST 2[3]の日本語訳(En-Ja)，そして筆者らのオリジナル TED-EN-JP データセットの 4 つの音声翻訳データセットを使用して行った．これらのデータセットは，翻訳元となる英語音声，英語の書き起こし文，その翻訳文から構成される．英語音声の対象翻訳文を我々の実験で扱う原言語テキストとして使用した．

MuST-C の作成方法を参考にして，筆者らオリジナルの TED-EN-JP データセットを作成した．TED-EN-JP データセットの作成方法は以下である．

・データのダウンロード

TED の英語スピーチ 628 時間分，315k 文からなる英語書き起こしテキスト，およびその日本語訳をダウンロードした．

・テキストアライメント

字幕単位の英語書き起こしテキストを文単位に分割・結合し，Vecalign[4]を用いて英語書き起こしテキストと日本語訳文のアライメントを行った．それだけではアライメントの精度が上がらなかったため，英語書き起こしテキストに付与されていた字幕時刻を使って，最も近い字幕時刻を持つ日本語訳を抽出しアライメントを行った． そして Vecalign による文レベルのアライメントが時刻に基づく字幕レベルのアライメントと整合するか検証し，不整合な文は除外した．

・音声-テキストのアライメント

テキストアライメントとは別に，強制アライナ Gentle を用いて英語音声と英語書き起こしテキストとのアライメントを行い，正解単語精度が 96%未満の音声データを除外した．最後に，テキストアライメントと音声-テキストアライメントの積集合を TED-EN-JP データセットとして抽出した．

2 つのアライメント後の TED-EN-JP データセットは，213 時間の英語音声と 136k の英語書き起こし文・日本語訳文から構成された．テキストアライメントでは字幕時刻を使ったフィルタリングを追加したため，抽出データ量は他のオリジナルの MuST-C よりも少ない．また開発セットとして 1000 文，テストセットとして 2000 文を無作為に抽出した．

3.2 実験設定

2 種類の実験を行った．1 つ目は MuST-C En-Nl データセットを用いてデコーダ，音響エンコーダのそれぞれに対してテキストエンコーダに対する注意機構を持つ Transformer ブロックを変化させ，単語誤り率から最適な組み合わせを求める．またデコーダ，音響エンコーダそれぞれにおいてテキストエンコーダに対する注意機構を持つ Transformer ブロックの数を増減させたときに性能がどのように変化するかを調査する．2 つ目に 3.1 章で述べた 4 つのデータセットを用いて，本研究のモデルとベースラインモデルである Transformer を比較した．

提案モデルは 6 層のテキストエンコーダブロック（N_T = 6），12 層の音響エンコーダブロック（N_A = 12），6 層のデコーダブロック（N_D = 6）からなる．ベースラインモデルのエンコーダブロックとデコーダブロックはそれぞれ提案モデルと同じ 12 層と 6 層である．提案モデルとベースラインモデルの埋め込みベクトルの次元数d_Mを 256 として，注意機構におけるヘッド数を 4，全結合層の次元数を 2,048 に設定した．

音響特徴量は 10ms 間隔でサンプリングされた 80 次元のログメルフィルタバンク出力であり，その時間分解能は 2D-Convolutional ネットワークを用いることで 4 分の 1 に低減される．MuST-C と TED-EN-JP データセットの英語書き起こし文と原言語テキストを SentencePiece unigram モデル[5]により 5000 個のサブワードにトークン化した．ただし，日本語は 8000 個のサブワードにトークン化した．CoVoST 2 データセットについては，150 個のサブワードにトークン化した．バッチサイズは 64 であり，70 エポックまで学習させ，5 エポックの early stopping を行った．また，データ拡張には SpecAugment[6]を用いてモデルの学習を行った．ベースコードには Espnet2 toolkit [7]を用いた．

3.3 実験結果

テキストエンコーダに対する注意機構を持つ音響エンコーダとデコーダそれぞれのブロックを変化させたところ，音響エンコーダ 12 層のうち 11 番目の層(M_A=11)，デコーダ 6 層のうち 4 番目の層(M_D=4)において単語誤り率が最小になった．

表 1 と表 2 にそれぞれ，M_Aを 11 に固定しM_Dを変化させた場合（表 1）と，M_Dを 4 に固定しM_Aを変化させた場合（表 2）の単語誤り率の推移を示す．表 1 より，M_D=4 と M_D=5 で最小の単語誤り率が得られ，入力に近いブロックや最終段ブロックにテキストエンコーダに対する注意機構を有しても最小の単語誤り率が得られないことがわかる．また，表 2 より，M_A=11 のときに最小の単語誤り率が得られた．これらの結果は，音響エンコーダとデコーダにおいてテキストエンコーダからの出力を受け取ったブロックの後段のブロックでテキストエンコーダからの情報を処理していることを示している．

表 3 は，音響エンコーダとデコーダそれぞれにおいてテキストエンコーダに対する注意機構を持つブロックに注目し，ブロック数を増減させた場合の単語誤り率の変化が示される．条件 1)と 2)は，デコーダや音響エンコーダでテキストエンコーダに対する注意機構を用いない場合に単語誤り率が増加したことを示す．条件 4)，5)，6)は，すべてのブロックでテキストエンコーダに対する注意機構を用いても，単語誤り率がそれ以上低下しないことが示される．表 5 は，4 つのデータセットのテストセットに対して，ベースラインモデルと，M_A=11，M_D=4 とし提案モデルの単語誤り率を示したものである．提案モデルはすべてのテストセットで単語誤り率を低減させた．また，英語と同じ語族のオランダ語やドイツ語では，語族の異なる日本語よりも誤りの低減率が高いことがわかった．

表 6 は，MuST-C En-Nl，MuST-C En-De，TED-EN-JP の各データセットにおいて，提案モデルによって改善された認識例である．ベースラインモデルで誤認識した単語を，提案モデルは原言語テキストを用いることで正しく認識できている．図 2 は，表 6 に示した例について，提案モデルにおけるデコーダからテキストエンコーダの注意機構の重みを示した図である．修正された出力トークンには，対応する原言語トークンかそれに近いトークンに対する重みが大きくなっている．

4. 結論

同時通訳者音声の認識に向けて，原言語テキストを補助入力とする Transformer ベースの End-to-end 音声認識を検討した．音声翻訳用のコーパスで英語音声と原言語テキストをシミュレートして，原言語テキストを補助入力とする有効性を検証した．実験の結果，4 つのデータセットすべてにおいて，本モデルは補助入力なしのベースラインモデルと比較して，単語誤り率を有意に減少させた．今後は，原言語テキストを用いた同時通訳コーパスを用いて，提案モデルを評価する予定である．

参考文献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.
[2] R. Cattoni, M. A. D. Gangi, L. Bentivogli, M. Negri, and M. Turchi, “MuST-C: A multilingual corpus for end-to-end speech translation,” Computer Speech & Language, vol. 66, pp. 1–14, 2021.
[3] C. Wang, A. Wu, J. Gu, and J. Pino, “CoVoST 2 and massively multilingual speech translation,” in Proc. Interspeech 2021, 2021, pp. 2247–2251.
[4] B. Thompson and P. Koehn, “Vecalign: Improved sentence alignment in linear time and space,” in Proc. EMNLP 2019, 2019, pp. 1342–1348.
[5] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in Proc. EMNLP 2018, 2018, pp. 66–71.
[6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in Proc. Interspeech 2019, 2019, pp. 2613–2617.
[7] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-End Speech Processing Toolkit,” in Proc. Interspeech 2018, 2018, pp. 2207– 2211.