1R-06

会話音声から句読点付きテキストの End-to-End 認識

野崎 樹文[1]　河原 達也[1]　石塚 賢吉[2]　橋本 泰一[2]

[1]京都大学 大学院情報学研究科　[2]株式会社 RevComm

1. はじめに

一般に，音声認識システムの出力するテキストには句読点が含まれていない。これは音声認識結果の可読性を下げる要因の一つである。また，句読点が含まれていないテキストは，機械翻訳や自動要約などの後続の自然言語処理タスクへの入力として望ましくない。この問題に対し，音声認識結果に対して BERT[1] などの事前学習済みモデルを用いて句読点を挿入する研究が近年盛んに行われている [2, 3]。しかし，これらの研究は句読点予測の際に音声の情報を活用できておらず，また，音声認識誤りの影響を直接受ける。さらに，句読点予測モデルに音声認識結果のテキストを入力する際に，テキストのトークン分割を行う必要があるが，句読点が含まれていないテキストに対するトークン分割は不正確になりやすい。

これに対して本研究では，音声を入力として句読点の付いたテキストを直接出力する End-to-End モデルを提案する。これにより音響情報を使用しながら，音声認識誤りやトークン分割誤りに対して頑健に句読点を予測することを目指す。また，モデルの最終層の出力と句読点付きテキストから計算される誤差に加え，中間層の出力と句読点のついていないテキストから計算される誤差を用いてモデルを学習する方法を提案する。評価実験は日本語と英語の二つのデータセットを用いて行い，BERT を用いて音声認識結果のテキスト情報のみから句読点を予測する従来のシステムと，提案モデルの性能を比較する。

2. 句読点付き音声認識

2.1 問題設定

本研究では，句読点付き音声認識を，音響特徴量系列 $\mathbf{X}$ を入力とし，句読点を含むトークン系列 $\mathbf{y}_{\text{pnct}}$ を出力する問題と定義する。句読点としては，例えば日本語のデータを用いる場合，読点（、），句点（。），クエスチョン（?）の 3 種類を考える。

2.2 カスケード接続型モデル

句読点付き音声認識では，句読点を出力しない音声認識モデルと，句読点予測モデルをそれぞれ別々に学習し，推論時はそれらをカスケードに接続して用いる手法が主流である。句読点予測モデルとしては，近年，BERT などの事前学習済みモデルを用いる研究が盛んに行われている。具体的には，まず，音声認識モデルが出力した認識結果のテキストを BERT の語彙に従ってトークンに分割する。その後，それぞれのトークンに対し，その直後に挿入される句読点の種類に従ってクラス分類を行うタスクで BERT をファインチューニングする。トークンの直後に挿入される句読点がない場合は “句読点なし” というクラスを正解とする。例えば日本語のデータを用いる場合，各トークンを，句読点なし，読点（、），句点（。），クエスチョン（?）の 4 クラスに分類するタスクを学習する。

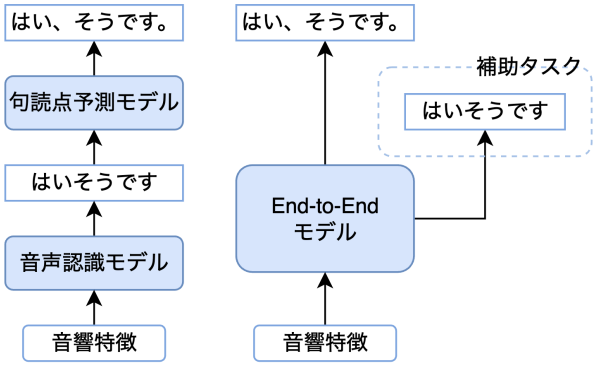

図 1: カスケード接続型モデル（左）と提案モデル（右）の概要図

3. 提案手法

本研究では，音響特徴量系列 $\mathbf{X}$ から句読点を含むトークン系列 $\mathbf{y}_{\text{pnct}}$ を直接 End-to-End に学習・推論するモデルを提案する。提案モデルと，前章で述べたカスケード接続型モデルの概要図を図 1 に示す。提案モデルには，l 層の Transformer Encoder 層を用い，最終層の出力 $\mathbf{X}_l$ に対して CTC 誤差関数で計算される以下の負の対数尤度を最小化するように学習を行う。

$$\mathcal{L}_{\text{CTC}} = -\log P(\mathbf{y}_{\text{pnct}} \mid \mathbf{X}_l) \tag{1}$$

また，学習を安定させるために，中間層の出力に対しても正解系列との誤差を計算する手法 [4] を用いる。[4] では最終層と中間層の出力に対して同じ系列を用いて誤差を計算するが，本研究では，最終層の出力に対しては $\mathbf{y}_{\text{pnct}}$，中間層の出力に対しては句読点を含まないトークン系列 $\mathbf{y}_{\text{unpnct}}$ を用いて誤差を計算する。中間層である $\lfloor l/2 \rfloor$ 層目の出力 $\mathbf{X}_{\lfloor l/2 \rfloor}$ に対する誤差は以下のようになる。

$$\mathcal{L}_{\text{inter}} = -\log P(\mathbf{y}_{\text{unpnct}} \mid \mathbf{X}_{\lfloor l/2 \rfloor}) \tag{2}$$

最終的な誤差関数は式 (1), (2) の線形和で表される。

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}} \tag{3}$$

ここで $\lambda_{\text{CTC}}, \lambda_{\text{inter}}$ はそれぞれの誤差の項にかかる重みづけの係数である。

推論時は中間層における予測は用いず，最終層における予測のみを用いる。

An end-to-end model for conversational-speech-to-punctuated-text recognition: *Jumon Nozaki, Tatsuya Kawahara (Kyoto Univ.), Kenkichi Ishizuka, and Taiichi Hashimoto (Revcomm, Inc.).*

表 1: カスケード接続型モデルと End-to-End モデル（提案モデル）の比較

モデル	JCALL CER (%)↓	句読点予測精度 (F1)↑ 、	。	?	平均	パラメータ数 (M)↓	MuST-C WER (%)↓	句読点予測精度 (F1)↑ ,	.	?	平均	パラメータ数 (M)↓
カスケード接続型	15.2	49.9	63.9	66.3	60.0	126	22.4	61.3	89.4	63.6	71.4	128
+ 中間層学習	14.6	50.2	64.5	62.7	59.2	126	20.1	**62.5**	89.6	67.4	73.2	128
End-to-End	15.0	**61.5**	70.1	73.9	68.5	**18**	29.2	55.8	91.4	61.3	69.5	**18**
+ 中間層学習	**14.2**	61.0	**70.4**	**74.3**	**68.6**	**18**	**19.7**	62.1	**92.6**	**70.6**	**75.1**	**18**

4. 評価実験

4.1 データセット

日本語データとして，インサイドセールスにおける営業担当者と顧客，コールセンターにおけるオペレータと顧客の会話音声を録音した独自のデータセット（JCALL）を用いる。読点（、），句点（。），クエスチョン（?）の 3 つを句読点とみなす。

英語データとして MuST-C コーパス [5] を用いる。これは主に音声翻訳の研究に用いられるデータセットであるが，本研究では，英独音声翻訳のデータセットのうち，英語音声と句読点付き英語スクリプトをペアデータとして用いる。コンマ（,），ピリオド（.），クエスチョン（?）の 3 つを句読点とみなす。前処理として，スクリプトは全て小文字に変換する。

4.2 実験設定

カスケード接続型モデルの音声認識には複数の Transformer Encoder 層を用い，CTC 誤差関数で学習する。学習には音声と句読点を除いたテキストをペアデータとして用いる。Transformer の層数は 12，隠れ層の次元は 256，ヘッド数は 4 とする。入力特徴量としては 80 次元の対数メルスペクトルを使用し，学習時は SpecAugment を用いたデータ拡張を行う。出力の語彙としては JCALL は文字単位の 1,923 種，MuST-C は SentencePiece で作成した 2000 種を用いる。句読点予測モデルには，JCALL に対しては東北大学，MuST-C に対しては Google 社が公開している Base サイズの事前学習済み BERT モデルを用いる。これらの BERT モデルの最終層に句読点のクラス分類のための線形層を追加し，クロスエントロピー誤差でファインチューニングを行う。ファインチューニングのデータには JCALL，MuST-C それぞれの学習データを用いる。

提案モデルは、正解系列に句読点付きのテキストを使用する以外は、カスケード接続型モデルの音声認識と同様のアーキテクチャと学習方法で訓練する。

カスケード接続型モデルの音声認識と提案モデルの双方とも，式 (2) で表される，中間層の出力に対して句読点なしテキストを用いて計算される誤差を用いて学習する場合と用いない場合の実験を行う。用いる場合は式 (3) の $\lambda_{\text{CTC}}, \lambda_{\text{inter}}$ はそれぞれ 0.5，0.5 とする。

4.3 評価方法

音声認識の精度を測る指標として JCALL は文字誤り率（CER），MuST-C は単語誤り率（WER）を用いる。計算の際には出力されたテキストの句読点を全て除いてから計算する。句読点予測の精度を測る指標としては，句読点の種類ごとに予測の F1 スコアを計算する。モデルが出力する予測系列は音声認識誤りを含むため，単純に正解系列と比較して F1 スコアを計算することはできない。そこで，音声認識の認識誤り率の計算と同様の手法で予測系列と正解系列のアラインメントをとってから計算を行う。これに加えて，ベースラインモデルと提案モデルの総パラメータ数を比較する。

4.4 実験結果

実験結果を表 1 に示す。JCALL を用いた実験では，提案モデルはカスケード接続型と同等の音声認識精度を保ちながら，高い句読点予測精度が得られた。中間層での学習を行った場合，音声認識精度，句読点予測精度ともに改善の傾向が見られた。MuST-C を用いた実験では，中間層での学習を用いない場合，提案モデルは WER が高くなったが，中間層での学習を用いると，カスケード接続型と同等の WER に改善した。句読点予測精度に関しては，提案モデルはカスケード接続型モデルと同等かそれ以上の精度であった。JCALL を用いた実験においてはカスケード接続型モデルの句読点予測精度が低いのは，日本語は単語区切りに空白を用いないため，句読点予測モデルに入力する前のトークン分割が不正確になりやすいことが原因として考えられる。また、どちらのコーパスを用いた実験でも，提案モデルはカスケード接続型のモデルと比較して，大幅にパラメータ数が少ない。

5. おわりに

本研究では，音声を入力として句読点の付いたテキストを直接出力する End-to-End モデルを提案した。評価実験により，提案モデルは従来のカスケード接続型モデルと比較して少ないパラメータで高い精度の認識ができることを示した。また，中間層の出力に対して句読点なしテキストを正解として学習することの有効性を示した。今後は，異なるアーキテクチャを用いたモデルの改善や，リアルタイム音声認識への応用を検討していく。

参考文献

[1] J. Devlin *et al.*: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL*, 4171–4186, 2019.

[2] M. Karan *et al.*: “ “ Transfer learning for punctuation prediction,” *APSIPA*, 268–273, 2019.

[3] J. Yi *et al.*: “Focal Loss for Punctuation Prediction,” *INTERSPEECH*, 721–725, 2020.

[4] J. Lee *et al.*: “Intermediate loss regularization for ctc-based speech recognition,” *ICASSP*, 6224–6228, 2021.

[5] MA Di. Gangi *et al.*: “MuST-C: a Multilingual Speech Translation Corpus” *NAACL-HLT*, 2019.