1R-04

wav2vec2.0 の事前学習モデルを用いた咽喉マイク音声認識

増田　光汰†　緒方　淳‡　西田　昌史†　綱川　隆司†　西村　雅史†
静岡大学†　産業技術総合研究所‡

1 はじめに

咽喉マイクは咽喉付近の皮膚振動を直接受け取るマイクであり、一般的なマイクよりも外部雑音の影響を抑制した音声の収録が可能である。そのため高雑音環境下での音声認識への活用に向けた研究が行われている[1-3]。しかし咽喉マイクで収録された音声は高域が欠如するなど、接話マイクで収録した音声と音響特徴量が著しく異なるため、接話マイク音声で学習されている一般的な音声認識システムでは咽喉マイク音声の認識精度は低下する。そのため咽喉マイク音声を認識するモデルを構築する必要があるが、咽喉マイクを用いて収録された大規模な音声データベースは存在しない。

今回我々は Baevski ら[4]によって提案された音声特徴表現学習のフレームワークである wav2vec 2.0 に着目した。wav2vec 2.0 では、正解ラベル無しの大量の音声データから汎用的な表現を事前学習(自己教師あり学習)し、目的のタスクにおける少量の文字起こしされた音声データでファインチューニングを行うことで高い認識性能を達成できることが示されている。

本研究では多言語音声からなる巨大なデータベースに基づいて構築された wav2vec2.0 の事前学習モデルをベースとして、少量の咽喉マイク音声によるファインチューニングを実施した。また認識精度の向上のために日本語大規模音声データである CSJ コーパスもファインチューニングに用いることを検討した。

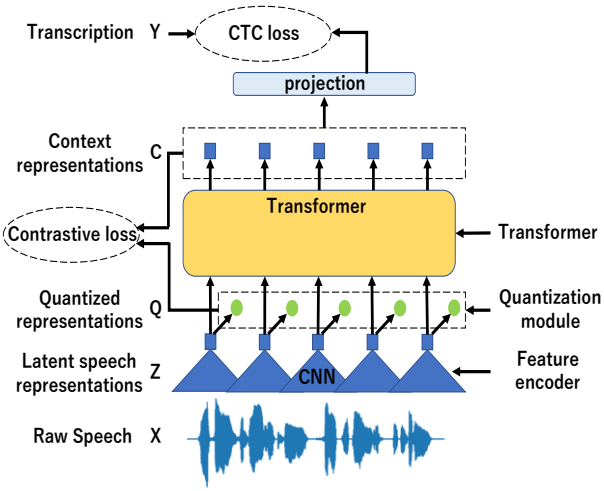

図 1：wav2vec 2.0 モデル構造

2 wav2vec 2.0

wav2vec 2.0 のモデル構造を図 1 に示す。Feature encoder 部は CNN を用いた temporal convolution を含むいくつかのブロックで構成されており、入力される生の音声データはゼロ平均と単位分散に正規化され、7 層の CNN を通じて音声波形 X をベクトル系列 Z へと変換する。Quantization module 部ではベクトル系列 Z に対してベクトル量子化を行うことによって量子化した音声潜在表現へと離散させる。ベクトル量子化とは、いくつかの代表点から構成されているコードブックの中から最も近い表現(Code vector)を選択する操作を指す。Transformer 部では文脈化された表現を得ることを目的としており、ベクトル系列 Z の一部をマスクし、文脈表現 C を得る。マスクされた箇所に対応する出力部分と量子化した音声潜在表現に着目して対照学習(Contrastive loss)によって音声表現の学習を行う。ファインチューニングでは Transformer 部の最終出力 C に対して線形層を加え、正解ラベル Y との CTC loss を最小化することで学習を行っている。

Throat microphone speech recognition using wav2vec2.0 pre-trained model
Kohta Masuda†,Jun Ogata‡, Masafumi Nishida†,
Takashi Tsunakawa†, Masafumi Nishimura†
†Graduate School of Integrated Science and Technology, Shizuoka University
‡National Institute of Advanced Industrial Science and Technology

3 実験

3.1 実験条件

ファインチューニングに用いる学習データとしては、接話マイク(CM：Close-talk mic)と咽喉マイク(TM：Throat mic)のパラレルデータとして男性話者 20 名から収録した音素バランス文読み上げ音声(約 10 時間)を用いた。評価データとしても、同様に男性話者 10 名から収録した新聞記事読み上げ音声(約 50 分)を使用した。いずれのデータも静かな環境で収録されており、評価データには学習データの話者は含まれていない。

wav2vec2.0 の事前学習モデルとしては、53 言語 56,000 時間の音声データで事前学習されたものを用いた。この事前学習モデルに対して接話マイク音声、咽喉マイク音声でファインチューニングを行ったモデルを Fine-CM、Fine-TM とする。なお少量の学習音声データだけでは言語情報を十分獲得できないことから、CSJ コーパスで学習した 5-gram 言語モデルの併用も検討した。また事前学習モデルに対して、CSJ コーパスによるファインチューニングを行ったモデルを用意し、さらにそのモデルに対してそれぞれの学習データによってファインチューニングしたモデル(Fine-CSJ-CM、Fine-CSJ-TM)も用意した。

それぞれのモデルに対して、接話マイク音声、咽喉マイク音声で評価を行う。評価指標には CER(Character Error Rate:文字誤り率)を用いた。

3.2 実験結果

それぞれのモデルとテストデータで認識性能を評価した結果を表 1 に示す。(1)の Fine-CM に対して CM で評価を行った場合の CER と(4)の Fine-TM に対して TM で評価を行った場合の CER は 25%程度であり、咽喉マイク音声データであっても接話マイクの場合に近い認識性能を得られることを確認した。また、モデルとテストデータのマイクが異なる(2)の場合、CER が大幅に増加したが、(3)の場合の CER は大きく増加することがなかった。接話マイク音声でファインチューニングを行ったモデルに対して、咽喉マイクで評価を行った場合には、高域の情報の欠如により、認識精度が低下することが考えられる。

表 1：認識結果

	model	Test data	CER[%]	+LM CER[%]
(1)	Fine-CM	CM	23.29	20.93
(2)	Fine-CM	TM	38.86	43.79
(3)	Fine-TM	CM	26.74	24.87
(4)	Fine-TM	TM	25.86	24.23
(1')	Fine-CSJ-CM	CM	11.57	ー
(2')	Fine-CSJ-CM	TM	29.3	ー
(3')	Fine-CSJ-TM	CM	13.72	ー
(4')	Fine-CSJ-TM	TM	14.38	ー

また認識性能の改善を目的として Fine-CM や Fine-TM に対して CSJ による言語モデル(LM)を使用した評価を行ったが、文字 LM だけでは大きな認識性能の改善は得られなかった。一方、CSJ によるファインチューニングを行ったモデルによる認識実験(1')~(4')では、(1)~(4)の認識結果と比較して、どの場合でも大幅な認識性能の改善を得られた。このことから、日本語の汎用的な言語情報を事前に学習することが、認識精度の向上にとって特に重要であることを確認できた。

4 おわりに

今回の実験では、咽喉マイク音声のデータ量不足という問題点を wav2vec2.0 の事前学習モデルを活用することで解決することを試みた。

少量の咽喉マイク音声データであっても、接話マイク音声データに近い認識精度を得られることや、CSJ コーパスによるファインチューニングによって認識精度がさらに向上することが確認できた。ただ、鈴木らの研究での提案手法と比較すると認識精度は十分ではなく、改善の余地がある。

謝辞

この成果は，国立研究開発法人新エネルギー・産 業技術総合開発機構(NEDO)の委託業務(JPNP20006) の結果得られたものです。

参考文献

[1] M Graciarena, et al.：Interspeech, Vol.9, No.6, pp.1-4, 2004.
[2] S Dupont, et al.：ISCA Workshop Robustness Issues Conveys Interact., pp.1-4, 2004
[3] 鈴木貴仁ほか:情報処理学会論文誌, Vol.62, No.6, pp1373-1381, 2021
[4] A Baevski, et al. “wav2vec2.0: A Framework for Self-Supervised Learning of Speech Representations”, 2020