クロスチャネル言語識別における wav2vec を用いた教師なし特徴量学習*

☆吉本拓真 (神戸大/NICT), 沈 鵬, Xugang Lu (NICT),
高島遼一, 滝口哲也 (神戸大), 河井 恒 (NICT)

1 はじめに

言語識別 (Language Identification; LID) システムは、入力された音声がどの言語のものであるかを判別するシステムであり、長年にわたり様々な研究がされている。主には特徴抽出と類似度計算の 2 段階で構成されるが、本稿ではその中でも特徴抽出に着目する。主な特徴抽出手法として、i-vector[1] や x-vector[2] が挙げられる。これらの手法は優れた性能を示すが、音声には言語に関する情報 (音素列情報など) の他にも、ノイズやチャネルといった言語識別には必要のない情報も含まれるため、例えば学習データと異なるチャネルの音声データが入力されると性能が劣化してしまう。先行研究では x-vector のネットワークに加えて自動音声認識 (ASR) モデルを同時に学習し、そこから得られる音素ベクトルを x-vector ネットワークの補助知識とする手法が提案されている [3]。しかしこの手法では、大量の音素ラベル付きデータが必要となり、対応できる言語が限られてしまう。そこで本研究では、自己教師あり学習を用いることで、言語に関する情報を多く含んだ x-vector 抽出モデルの入力として用いる特徴量を作成することを考える。

2 自己教師あり学習を用いた特徴表現学習

2.1 x-vector とロジスティック回帰

本研究のベースとなる LID 手法は、言語情報を含んだ特徴を抽出するための x-vector 抽出モデルと、抽出された x-vector を入力として言語を識別するロジスティック回帰 (LR) モデルからなる。x-vector 抽出モデルでは、N 種類の言語ラベルが付与されている学習データを正しく識別できるように DNN を学習し、テスト時には中間層の出力を特徴表現ベクトル x-vector として抽出する。 x-vector は、言語識別に有用な情報を含んでおり、LR 分析への入力として用いられる。LR の出力は、テストデータに含まれる言語が M 種類あるとしたとき、各言語 $i \in M$ の確率 (スコア) を表す。ここで、学習データとテストデータには $M \subset N$ の関係がある。

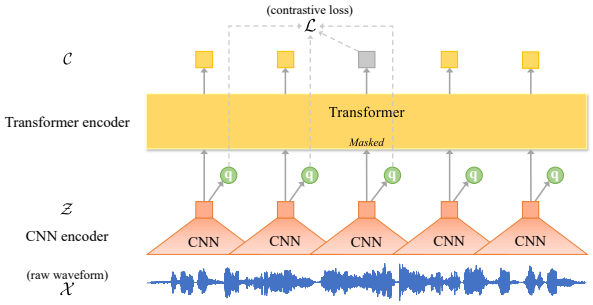

Fig. 1 wav2vec architecture.

2.2 wav2vec 2.0

本研究では、自己教師あり学習による特徴表現の獲得に wav2vec 2.0[4] のモデルを使用する (以降では wav2vec 2.0 を wav2vec と表記する)。wav2vec の構造を Fig. 1 に示す。wav2vec のシステムは音声の時間波形 $\mathcal{X}$ を入力し発話に関する潜在表現 $\mathcal{Z}$ を得る畳み込みニューラルネットワーク (CNN) エンコーダ部と、その潜在表現 $\mathcal{Z}$ から発話全体の情報が考慮された文脈表現 $\mathcal{C}$ を得る Transformer エンコーダ部からなる。また自己教師あり学習をする際は、目的関数を計算するため、量子化された潜在表現 $\mathbf{q}_t$ を生成する量子化モジュールも使用されている。

学習時には BERT[5] における masked language modeling と同様に $\mathcal{Z}$ の一部をマスクし、そのマスクされた各時間ステップごとに、別の時刻の量子化潜在表現とマスクされた時刻の真の量子化潜在表現が正しく識別されるように学習する。文献 [2] では大規模なデータセットで事前学習したのちに特定タスクの少量の学習データで再学習しているが、我々は少量データのみの学習による有効性を確認するため、事前学習は行っていない。

2.3 モデルの概観

今回我々が提案する LID モデル全体の流れを Fig. 2 に示す。wav2vec システムによって得た表現 $\mathcal{C}$ を x-vector 抽出モデルの入力とし、そこから x-vector を抽出しロジスティック回帰 (LR) 分析に通してスコアを得る。なお、x-vector 抽出モデルには extended TDNN を用いている [9]。

*Unsupervised Feature Learning based on wav2vec for Cross-channel Spoken Language Identification. by YOSHIMOTO, Takuma (Kobe Univ./NICT), SHEN, Peng, LU, Xugang (NICT), TAKASHIMA, Ryoichi, TAKIGUCHI, Tetsuya (Kobe Univ.), KAWAI, Hisashi (NICT)

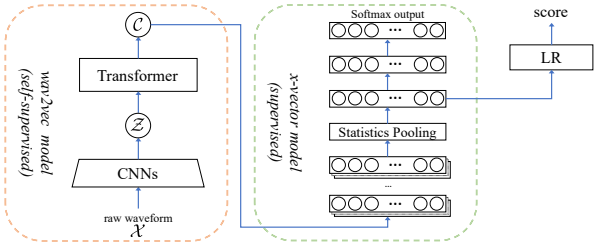

Fig. 2　Overview of the proposed method.

3　評価実験

3.1　データセット

音声データには、言語識別のコンペティション AP20-OLR Challenge で使用される以下のデータセットを用いた [6, 7, 8]。学習データセットは、AP16-OL7、AP17-OL3、および AP19-OLR-dev/short のデータセットで構成され、標準中国語、広東語、インドネシア語、日本語、ロシア語、韓国語、ベトナム語、チベット語、ウイグル語、カザフ語の全 10 言語 ($N = 10$) について、それぞれ男女 12 名ずつによる 10 時間前後の音声が収録されている。また、テストデータとして用いるのは AP19-OLR-channel であり、このデータには学習データセットとは異なるチャネルで収録された標準中国語、日本語、ロシア語、ベトナム語、チベット語、ウイグル語の全 6 言語 ($M = 6$) の音声が約 1,800 文ずつ収録されている。

3.2　提案手法

まず wav2vec の自己教師あり学習の部分を先に行う。この学習にはテストデータを含む前節で述べたすべてのデータを用いて学習を行い、音声の時間波形 $\mathcal{X}$ から文脈表現 $\mathcal{C}$ を得て、これを x-vector 抽出モデルへの入力とする。ここで、$\mathcal{C}$ の代わりに潜在表現 $\mathcal{Z}$ を x-vector 抽出モデルへの入力とすることも考えられるが、予備実験において $\mathcal{C}$ の方が高い性能を示したため、$\mathcal{C}$ を用いることとした。CNN エンコーダ部および Transformer エンコーダ部の構成は文献 [4] と同様である。ただし、CNN エンコーダ部の最後のブロックに関しては CNN エンコーダの最終的な出力である $\mathcal{Z}$ の次元を変化させるためにチャネル数を 512, 256, 128, 64, 32 と変化させて実験を行った。

次に、wav2vec によって計算された $\mathcal{C}$ を用いて、x-vector 抽出モデルの学習を行う。ここではラベル付きデータ、すなわち学習データのみを用いて教師あり学習を行う。x-vector の次元数は 512 とし、さらに線形判別分析により 100 次元に圧縮したうえで LR に入力している。また、従来手法として、 x-vector 抽出モデルの入力に 30 次元の MFCC を用いた手法と比較した。

Table 1　Experimental results of baseline and the proposed method.

Method	Feature	C_{avg}	EER%
baseline	30-dimensional MFCC	0.32	32.7
proposed	$\mathcal{C}$ (512-dimensional $\mathcal{Z}$)	0.20	20.9
	$\mathcal{C}$ (256-dimensional $\mathcal{Z}$)	0.19	21.1
	$\mathcal{C}$ (128-dimensional $\mathcal{Z}$)	0.16	18.5
	$\mathcal{C}$ (64-dimensional $\mathcal{Z}$)	**0.13**	**13.4**
	$\mathcal{C}$ (32-dimensional $\mathcal{Z}$)	0.21	23.3

3.3　実験結果

従来手法及び提案手法を C_{avg}[6] と EER で評価した結果を Table 1 に示す。ここで C_{avg} は偽陽性率と偽陰性率の平均値を計算したものであり、EER は言語識別の誤り率を示したものである。従来手法と比べ提案手法は C_{avg}、EER ともに良い結果が得られた。また、$\mathcal{Z}$ の次元数がある程度小さいほうが優れた結果となり、次元数が 64 のときにおいて約 60%の改善が見られた。次元数を小さくすることでノイズやチャネルによる影響をさらに減らすことができたためだと考えられる。

4　おわりに

x-vector 抽出モデルへの入力に自己教師あり学習によって得られた特徴表現を用いることで、ラベル付きデータの量が限られている場合でも異なるチャネルの言語識別性能を改善できることを示した。

参考文献

[1] N. Dehak *et al.*, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.
[2] D. Snyder *et al.*, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *ICASSP*, 2018.
[3] Y. Liu *et al.*, "Speaker embedding extraction with phonetic information," *INTERSPEECH*, 2018.
[4] A. Baevski *et al.*, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," arXiv, 2020.
[5] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, 2018.
[6] Z. Li *et al.*, "AP20-OLR Challenge: Three Tasks and Their Baselines," arXiv, 2020.
[7] Z. Tang *et al.*, "AP19-OLR Challenge: Three Tasks and Their Baselines," *APSIPA ASC*, 2019.
[8] KingLine Data Center, AP16-OL7 Multilingual Database, Speechocean Ltd., 2016.
[9] D. Snyder *et al.*, "Speaker Recognition for Multi-speaker Conversations Using X-vectors," *ICASSP*, 2019.