音声特徴量に基づく感情推定用 2D-CNN の検討

1230332　坂口　白磨　【 コンピュータ構成学研究室 】

1　はじめに

近年，AI を用いて人の話し声から話し手の感情を推定する研究が進められている. 話し声には，イントネーションや大きさなどの特徴があり，これらの特徴から感情を推定する. 先行研究では，音声データから音声特徴量である MFCC(Mel-Frequency Cepstrum Coefficients) を生成し，2D-CNN を用いた LIGHT-SERNET で感情を推定する方法が提案されている [1]. 本研究では，LIGHT-SERNET の構成において，感情推定の精度向上に大きく寄与する機能を分析した. そして分析した結果から LIGHT-SERNET の改良モデルを提案する.

2　LIGHT-SERNET[1]

LIGHT-SERNET は，低レベル特徴の並列抽出部と高次特徴の段階的抽出部から構成される. 前者は，3 種類の異なるカーネルサイズを持つ 2D-CNN Block(Conv+BN+ReLU+AvgPool) が並列に配置された構成で，後者は，5 つの 2D-CNN Block が縦列接続された構成になっている.

2.1　低レベル特徴の並列抽出部

まず 3 種類の 2D-CNN Block のうち，1 種類の 2D-CNN Block のみを用いた場合の推定精度を比較した. その結果，2D-CNN Block(kernel size=9x1) を使用した場合の推定精度が高かった. 次に 3 種類の 2D-CNN Block のうちの 2 種類の 2D-CNN Block を用いた場合の推定精度を比較した. その結果，2D-CNN Block(kernel size=9x1) を含む場合の推定精度が高かった. さらに，2D-CNN Block(kernel size=9x1) のみを使用した上で，Conv2d のフィルタ数を 64 に増やし，AvgPool2d を削除すると，推定精度が向上した.

2.2　高次特徴の段階的抽出部

まず，4 層目の 2D-CNN Block の AvgPool2d を削除した結果，削除する前より推定精度が向上した. 一方で 5 層目の 2D-CNN Block の GAP(GlobalAvgPool2d) を全結合層に置き換えた結果，置き換える前より推定精度が低下した.

3　提案モデル

図 1 は，提案モデルの構成である. 分析より，MFCC のスペクトル特徴を抽出する 2D-CNN Block(kernel size =9x1) が推定精度の向上に大きく寄与していると考えられたため，提案モデルは，2D-CNN Block(kernel size=9x1) を 4 層に増やすことで，MFCC のスペクトル特徴量をより強調して抽出する構成とした. また 1 層目の Conv2d のフィルタ数を 64 に増やし，AvgPool2d を削除した.

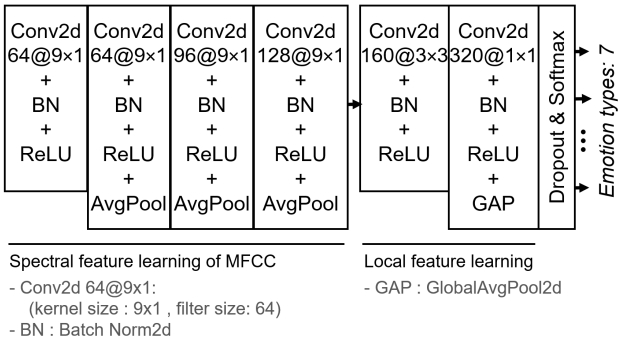

図 1　提案モデルの構成

4　評価・まとめ

モデルの評価には，5-fold 交差検証を使用した. 5-fold 交差検証を 5 セット行い，それぞれの平均値を推定精度とし，提案モデルと LIGHT-SERNET を比較した. データの分割方法は，535 個の音声データを 6 分割し，1 つをテストデータ，そして残りの 5 つのデータについては，1 つを検証データ，4 つを学習データとした. 以下の表 1 は，5-fold 交差検証による提案モデルと LIGHT-SERNET の推定精度である.

表 1　5-fold 交差検証によるモデルの推定精度

	1	2	3	4	5	Avg
文献 [1]	85.4	85.8	88.3	86.1	85.8	86.3
提案モデル	90.1	88.1	88.9	91.2	87.6	89.2

5-fold 交差検証で学習した結果，提案モデルは LIGHT-SERNET の推定精度より約 3%高かった. また，モデルのパラメータ数に関して，LIGHT-SERNET の 425,607 個よりわずか 4%の増加で精度を向上できた.

本研究で LIGHT-SERNET のモデル構成を分析した結果，MFCC のスペクトル特徴量の抽出が推定精度の向上に寄与していることが分かった. 今回は 1 種類のデータセットのみを使用したため，今後は提案モデルを他のデータセットへも適用し，汎用性について評価してみたい.

参考文献

[1] A. Aftab, et al., “LIGHT-SERNET: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition,”ICASSP 2022, pp. 6912–6916, April. 2022.