6N-01

自動会話収集システムにおいて話者識別に用いる深層学習モデルの検討

井腰　三四郎†　　相場　亮‡

芝浦工業大学　大学院理工学研究科†‡

1. はじめに

現在、ICT において人間の動向に関するデータの重要性が上昇している。データ活用の規模が広がるほど必要となるデータの数は増大していき、そのために、より効率的なデータ収集方法が求められ、IoT デバイスなどによる個人の行動データの自動収集などが活用されてきた。収集されるデータは人間の行動パターンや心拍数など様々であるが、現在注目されているものの一つとして音声が挙げられる。

音声は、音声の字幕化や翻訳、聞き取り辛い音声や発音の修正、変換などに活用される。近年は声だけではなく会話にも注目が行き、議論における話し方の効果や、個人が成す役割の分析なども行われている。しかし、会話データの収集には収録用の設備やその後のデータ整理方法などの多くの課題が存在する。その中でも話者識別は大きな課題であり、自動化によって全体的な効率化をすることが可能である。

本研究では ICT 活用における効率的な会話データ収集システムにおいて、話者識別に有効な深層学習モデルの検討を行う。これは会話の音声データ収集コストの削減や、議事録などの個人ごとの会話や発言を自動収集するシステムへの活用を想定している。

2. 関連研究

話者識別に関する研究は多く行われており、主に混合ガウス分布(GMM)による識別が基礎となっている。その後 Support Vector Machine (SVM)が識別方法として導入され、GMM を特徴量に落とし込む手法が提案された。さらに、i-vector[1]が提案されてからはそれが主流になっていった。しかし、i-vector は識別に用いる特徴量を音響特徴量ではなく、GMM のモデルパラメータから生成されたものを用いている。そのため、話者の特徴を正常に認識できない可能性がある。

また近年、音声分野における深層学習に関する研究の活発化によって、X-vector[2]という方式が i-vector に代わる新たな特徴抽出器として登場した。これは、特徴抽出部と識別部からなる深いニューラルネットワークを話者識別ができるように訓練を行い、音声から話者識別において有用な情報のみを抽出する特徴抽出器を構築している。音声は可変長の時系列データであり、ニューラルネットに入力するデータも可変だが、X-vector ではプーリング層を特徴抽出部の間に用いることで、一定次元数の特徴量を出力する。

2019 年には法政大学大学院デザイン工学研究科の竹内によって深層学習を用いた話者識別の研究行われている[3]。これは ATR503 文[4]を音読した音声のスペクトログラムに注目し、深層学習によって話者識別を行っている。5 人の被験者に対して、30 分の録音データから最大 84%の話者識別に成功しており、本研究においても有効な手法である。しかし、十分な精度を出す為には長時間の音声データが必要であり、多くの時間が必要となる。

3. 音声データの準備と処理

実験用の音声データには jvc コーパス[5]を使用する。これは 100 名の声優や俳優から得られた様々な音声が含まれているもので、今回は音素バランス文を読み上げている parallel100 から 5 名分の学習用データ作成に使用した。
音素バランス文とは日本語に存在する音素を偏りなく含んだ文章であり、学習用データ中の音素による学習結果の偏りがない。

また、学習においてはより多くのデータが必要になるため、音声を一定時間ごとに分割したものをスペクトログラムに変換し、学習用データとして使用した。

4. 実験手法

本研究では、音声処理によって得られたスペクトログラムから deep neural network(DNN)によって個々人の音声に含まれる特徴を学習し、話者識別を行った。分類器はすべて python によって設計されており、機械学習用のライブラリである keras を用いた。実験には以下に示すよ

On Deep Learning Models for Speaker Identification of an Speech Acquisition System
†Sanshiro Ikoshi
‡Akira Aiba
††‡Graduate School of Science and Engineering, Shibaura Institute of Technology

うな複数の学習モデルを使用し、各モデルの判別精度を比較した。以下、実験結果については代表的なものを述べる。

5. 実験結果

5.1. 8層のDNN学習モデル

図1は8層のDNNによる学習モデルである。損失関数は2乗平均誤差法、最適化アルゴリズムにはAdaGradを用いている。

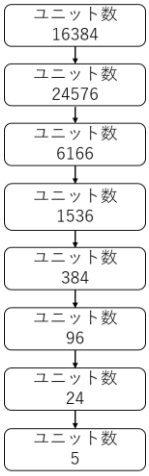

図1　8層のDNN学習モデル

実験の結果、テストデータの正答率は約75%となった。

5.2. VGGNet

VGGNet[6]はAlexNet[7]よりも深い層を持つCNNで、3×3の畳み込み層と全結合3層で構成されている。小さいフィルターの畳み込み層を連続して重ね、プーリング層で特徴マップを半分にすることを繰り返す。本研究では学習済みVGG16(図2)を5クラス分類に対応させて学習を行った。

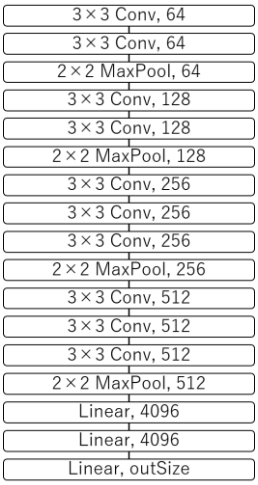

図2　VGG16

実験の結果、テストデータの正答率は約82%となった。

6. 考察

図1の学習モデルの結果から、単純なDNN学習モデルであっても、スペクトログラムによる話者識別に対して一定の効果があることが確認できる。学習用データや深さを増やせばさらに精度を高める事ができると思われるが、多クラス分類における学習効率などを考えると現実的ではないと考えられる。

VGGNetの結果からは、精度に関しては図1の学習モデルと大差はないが、学習用データの増加による精度向上や他クラス分類においては将来性があると考えられる。

7. まとめ

本研究では、話者識別に有効な深層学習モデルの検討を行った。将来的に多数の話者識別を行うことを考えると、画像識別において優秀なモデルを転移学習やファインチューニングすることで、より効率的な話者識別を行うことが可能になるのではないかと考える。

参考文献

[1] 小川　哲司、塩田　さやか、“i-vectorを用いた話者認識“、日本音響学会誌、70巻、6号、p. 332-339(2004)

[2] David Snyder,et al.“X-vectors: Robust DNN Embeddings for Speaker Recognition”、2018 IEEE International Conference on Acoustics, Speech, and Signal Processing(2018)

[3] 竹内　涼平、“深層学習を用いた話者特定システムの検討”、法政大学大学院紀要. デザイン工学研究科編、法政大学大学院デザイン工学研究科、pp.1-4(2019)

[4] 匂坂　芳典、浦谷　則好“ATR音声・言語データベース”音響誌、48巻、12号、pp. 878-882(1992)

[5] Shinnosuke Takamichi,et al.“JVS corpus:free Japanese multi-speaker voice corpus”、arXiv preprint, 1908.06248(2019)

[6] Karen,Simonyan,Andrew,Zisserman.“Very Deep Convolutional Networks for Large-Scale Image Recognition”、arXiv preprint arXiv:1409.1556v6(2015)

[7] Alex Krizhevsky,Ilya Sutskever & Geoffrey E. Hinton“ImageNet Classification with Deep Convolutional Neural Networks”、Advances in Neural Information Processing Systems 25(2012)