ISS-A-061

3DCNN 発話分類モデルにおける 日本語単語読唇への発話形態の影響

北村 亮太† 寺澤 卓也†

†東京工科大学大学院 バイオ・情報メディア研究科 メディアサイエンス専攻

1. はじめに

読唇認識は、発話時の口の動きを基に単語を認識する。そのため、周囲の音に左右されずに使用できる他、話者が無声であっても使用可能である。

近年の読唇認識分野では、深層学習を用いた研究が行われている。しかし、こうした研究で用いられるデータセット等は、有声のものが多い。また無声と有声では、同じ単語を発話した際でも口の動きが異なる。したがって、無声話者の利用が想定される読唇認識では、声の有無といった発話形態が、深層学習モデルの作成時に与えるデータと評価データで異なると、精度に影響を及ぼす可能性がある。

予備研究[2]では、単語発話時の映像から、深層学習による日本語単語読唇分類モデルを作成し、発話形態の違いがモデルの精度に影響を及ぼすか調査を行った。しかし、話者数の少なさ等から、影響の有無を断定できなかった。そのため、話者数と語彙を増やし、追加実験と検討を行った。

2. データ収集と単語読唇分類モデルの作成

はじめに、単語読唇分類モデル作成のために、発話データの収集と、収集データの前処理を施した。

発話内容は[3]を参考に、「おはよう」「こんにちは」「もしもし」「どういたしまして」「はい」の 5 単語とした。各単語は有声、囁き声、無声の発話形態毎に 1 単語各 50 回発話してもらい、その際の話者の顔正面の映像を撮影した。話者は大学生の男女 10 名で、1 名の話者から 750 回の単語を発話した際の映像を収集した。

データの前処理として、発話時の映像から口の関心領域をサイズ 40×40、RGB の 3 チャンネルで抽出したものをフレームごとに区切り、画像とした。1 つの映像につき、単語の発話シーンが含まれた 89 フレームを対象に関心領域を抽出した。

前処理を施した各データは、モデル作成と作成したモデルを評価するために、話者で分割を行った。被験者 6 名分をモデルの学習データとして扱い、2 名分をモデルのパラメータ調整の為の検証データとした。残り 2 名分を評価データとし、作成された学習モデルの分類精度を算出する。また、学習データは拡張を行った。そのため、発話形態 1 つの最終的なデータ数は、学習データ 7500 個、検証データ 500 個、評価データ 500 個となっている。

次に、作成するモデルは、[3]を参考に 3DCNN による構造のものとした。作成には、発話形態毎の学習データを用いる。そのため、有声モデル、囁き声モデル、無声モデルが作成される。しかし、各モデルは作成の都度、学習の出来にバラツキが生じる事から、1 発話形態につき 10 個モデルを作成した。

3. 評価方法

はじめに、モデルの分類精度を算出する。算出には、発話形態毎に作成したモデルに対し、発話形態毎の評価データを与える。1 発話形態につき各 10 個作成されたモデルの精度の平均を、そのパターンでの分類精度とし、その中での発話形態毎の精度差から分類傾向を判断する。

次に、算出されたパターン毎のモデルの分類精度を比較する。これによって、読唇モデルの構築と利用に、発話形態がどういった影響を与えているかを考察する。

4. 結果

各モデルに評価データを与えた際の分類精度を表 1 に示す。分類精度は、モデルと評価データの発話形態が一致している場合に、最も高くなると予想していた。しかし、囁き声モデルにおいても、有声データを与えた際の精度が最も高いという結果になった。また、無声モデルでの有声データを与えた際と無声データを与えた際の精度差は、非常に小さかった。

表 1 各モデルに評価データを与えた際の分類精度

評価データ→ モデル↓	有声	囁き声	無声
有声	0.548	0.481	0.439
囁き声	0.58	0.517	0.47
無声	0.548	0.441	0.55

5. おわりに

モデルと評価データで発話形態が一致しているときの精度が最も高いという予想に対して、本検証は異なる結果となった。その要因はいくつか考えられるが、今回の実験では確証が得られなかった。

今後は発話データのさらなる収集や分割の見直し、他の深層学習モデルの導入等を含めた検証を行う。

参考文献

[1] 斎藤 歩, 他, 情報処理学会, 第75回全国大会講演論文集, Vol. 1, pp. 467-468, 2013

[2] 北村 亮太, 他, 第 82 回全国大会講演論文集, Vol. 1, pp. 391-392, 2020

[3] 斎藤 剛史, 他, 電子情報通信学会技術研究報告, vol.117, no.513, pp.163-168, 2018.