6S-07

CNNを用いた日本語単語読唇における異なる発話形態の影響

北村 亮太† 寺澤 卓也†
東京工科大学メディア学部†

1 はじめに

昨今検索などに用いられる音声認識技術は、周囲の音が大きい環境では認識がされにくく、静穏な環境では心理的要因から使用を躊躇う事が考えられる。

これに対し読唇はサイレント音声認識と呼ばれる分野の内の画像によるアプローチ方法に分類され[1]、発話時の口の動きを基に単語を認識する。その為、周囲の音に左右されずに使用できる他、話者は無声による使用も可能である。

近年の読唇認識分野では深層学習を用いた研究が見られる。中でも、英語フレーズ読唇での学習モデルの構築とそのテストにおいて同一の発話形態を用いた際と異なる発話形態を用いた際では、前者の方が分類精度が高くなるという結果が提示されている[2]。

一方で、日本語単語読唇において同様の検証が為されていない事から、日本語においても検証を行う必要があると考えた。そこで、本研究では有声、囁き声、無声による日本語単語発話時の動画を基に発話形態毎の3DCNNモデルを構築し、モデルの分類精度から発話形態の違いが日本語単語読唇分類精度に影響を及ぼすのか検証した。

2 単語読唇分類モデルの作成

単語読唇分類モデルの作成を行うにあたり、発話データの収集、収集データの前処理を予め行う必要がある。

2.1 発話データの収集

データ収集における発話内容は「おはよう」、「こんにちは」、「もしもし」の3単語とした。これらの単語を有声・囁き声・無声の発話形態毎に1単語各50回ずつ発話をしてもらい、その際の話者の顔正面の動画撮影を行った。

撮影にはSONYのHDR-CX590を使用し、解像度1920×1080、フレーム率29.97フレーム/秒の設定で行った。動画は、いずれかの発話形態で1単語の発話を録画開始から3秒以内に行い、4秒経過した時点で停止したものを1データとした。

収集には男性大学生被験者5名に協力してもらい、各被験者から450本の発話データを収集した。

2.2 データの前処理

発話時の映像から、口の関心領域(Region of Interest: ROI)部分を抽出したものをフレーム毎に区切ることで画像とした。前処理にはOpenCVとdlibを用いた。

3DCNNモデルは作成時に各画像の高さ、幅、チャンネル、フレーム数を統一する必要がある。各発話データにおける撮影時の条件より、発話区間は動画の先頭から90フレーム以内に存在する。この事から、1つの動画の撮影開始から89フレーム目までの89枚を1データとすることでフレーム数の統一を行った。また、データ内のROI画像はいずれも40×40サイズ、RGBによる3チャンネルとした。

2.3 モデルの作成

発話形態に応じたデータを学習に用いることで、発話形態毎のモデルを作成した。モデル作成の為の学習に用いるデータは被験者5名の内の4名のものとし、残り1名のデータはテストデータとして扱う。1種類のモデルに用いるデータは3単語600データとなっており、この内の120データを検証用データとし、残りを学習用データとして扱う。各モデルの学習はバッチサイズ5、エポック数30とし、最適化関数にAdamを用い、学習率を0.0001とした。

3 評価方法

評価は各発話形態でのモデル、テストデータによる組み合わせ毎に表1の混同行列と式(1)を用いることで分類精度を算出する。また、式(2)、式(3)、式(4)を用いて、各モデルの単語毎の評価指標として適合率、再現率、F値を算出する。

これによってモデル同士を比較し、深層学習読唇モデルの構築において発話形態を考慮すべきかの考察を行う。更に、単語と発話形態の組み合わせにおける分類傾向などを探る。

Effects of different utterances on Japanese lip reading using CNN
† Ryota Kitamura, Takuya Terasawa
(School of Media Science, Tokyo University of Technology.)

表 1　モデル毎の混同行列

		Predicted label		
		こんにちは(A)	もしもし (B)	おはよう (C)
True label	こんにちは(A)	TA	FB(A)	FC(A)
	もしもし (B)	FA(B)	TB	FC(B)
	おはよう (C)	FA(C)	FB(C)	TC

$$\text{正解率} = \frac{TA+TB+TC}{TA+FA+TB+FB+TC+FC} \quad \cdots\cdots\cdots\cdots \quad (1)$$

$$\text{単語}x\text{の適合率}Px = \frac{Tx}{Tx+Fx(y)+Fx(z)} \quad \cdots\cdots\cdots \quad (2)$$

$$\text{単語}x\text{の再現率}Rx = \frac{Tx}{Tx+Fy(x)+Fz(x)} \quad \cdots\cdots\cdots \quad (3)$$

$$\text{単語}x\text{の F 値} = \frac{2\times Px\times Rx}{Px+Rx} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (4)$$

4　結果

作成した 3 種類のモデルに対し、テストデータを与える事で組み合わせ毎に分類精度(正解率)を算出した(表 2)。

[2]の研究と類似した結果が得られると見込んでいたが、同一の発話形態によるモデルとテストデータの組み合わせで高精度を出したのは、共に無声の場合のみであった。囁き声モデルにおける結果は他の 2 種類のモデルの各分類精度と比べ横並びになっている為、一概に無声データの分類に秀でているとは断定できない。しかし、有声モデルにおいては囁き声データを与えた際と他のデータを与えた際では分類精度が大きく乖離している。そこで有声モデルに関して、モデルの作成と同じ発話形態である有声データと分類精度の高かった囁き声データの分類傾向を確認した(図 1，図 2)。

有声データを与えた際、「もしもし」の F 値が 0.79 であったのに対し、「こんにちは」の F 値は 0.40、「おはよう」の F 値が 0.28 と低く、この 2 単語が有声モデルの分類精度の低下要因となっていた。一方で、囁き声データでは最も低かった F 値が「もしもし」の 0.78 であった。この事から、有声データと比べ全体の F 値が高く、各単語の分類も比較的正確に行われていた。

表 2　各組み合わせにおけるモデルの分類精度

テストデータ → 分類モデル　↓	有声	囁き声	無声
有声	**0.54**	**0.82**	**0.48**
囁き声	**0.49**	**0.55**	**0.64**
無声	**0.47**	**0.54**	**0.89**

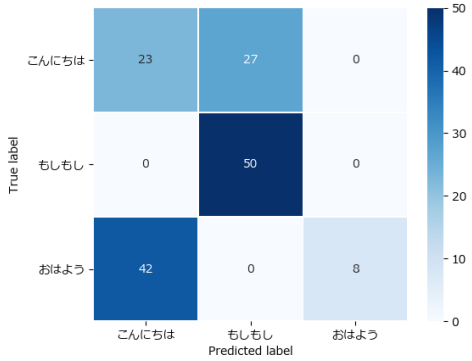

図 1　有声モデルに有声データを与えた際の混同行列

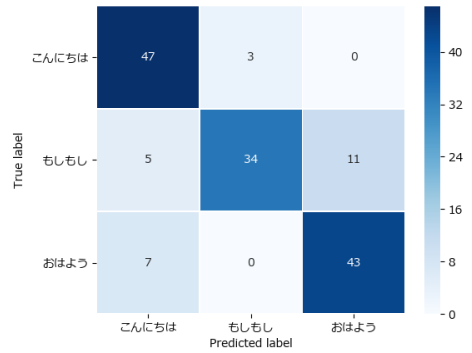

図 2　有声モデルに囁き声データを与えた際の混同行列

5　おわりに

本研究では、発話形態毎のモデルとデータの組み合わせでの分類精度と分類傾向を把握した。一方で、こうした結果となった要因の特定、発話形態の影響の有無の断定には至らなかった。

しかし、今回はデータ数が少ない事から被験者 1 名あたりのデータが学習に大きく影響していると考えられ、中でも発話時の口の動きの大きさが発話形態に左右されにくい人のデータがこれにあたる可能性が高い。

その為、上記の傾向が見られる被験者のデータを用いずにモデルの作成を行う等、引き続き要因の特定を行い、発話形態の違いによる日本語単語読唇分類精度の影響の有無を明確にする。

参考文献

[1] 斎藤 剛史, “サイレント音声認識の研究動向 読唇技術を中心として”, 電子情報通信学会技術研究報告, Vol. 117, No. 251, pp. 77-81, 2017

[2] S. Petridis, J. Shen, et al., “Visual-Only Recognition of Normal, Whispered and Silent Speech”, in IEEE ICASSP, 2018, pp. 6219-6233.