深層学習を用いた

リップリーディングのための映像データベースの構築と評価

Building and Evaluation of Movie Database for Lip Reading with Deep Learning

5115E016-2　橋本 直矢　　指導教員　尾形 哲也　教授

HASHIMOTO Naoya　　Prof. OGATA Tetsuya

概要： 本研究では，唇の形・動きといった視覚情報を手掛かりとして情報を読み取る方法(リップリーディング)を深層学習に用いるための映像データベースの構築を行った．リップリーディングは近年深層学習により他手法と比べ高い精度で実現されている．従来リップリーディングの学習に用いた音声発話の映像データベースは収集に高度な環境や設備を必要とするため，一般的に深層学習で必要とされる大規模データの収集が難しいといった課題があった．本研究ではタブレット端末といった比較的簡易なツールを用い 40 人分のデータ収集を行った．データベースの読み上げ文章は著作権フリーな文章から構成することでデータベースの公開を可能とし，同様の方法でさらに大規模なデータを収集できることを可能にした．

キーワード： リップリーディング，大規模映像データベース，深層学習

Keywords: Lip Reading, Large Movie Database, Deep Learning

１.はじめに

実環境下における雑音にロバストな音声認識の構築を目指した際，唇画像という視覚情報をリップリーディングが考えられる．視覚情報の利用により，雑音が大きい環境下でもその影響を受けにくい音声認識を実現できる．

リップリーディングに深層学習を適用した研究が従来手法より高い性能を持つことが示されているが，学習のためのデータ収集には高度な環境が必要なため，一般に深層学習に必要とされるような大規模なデータの収集が難しかった．本研究では，比較的容易な環境でデータ収集を行い，実用的なリップリーディングのための大規模データセット作成を目指す．

２．従来手法と課題

近年画像認識では Convolutional Neural Network (CNN)をはじめとした深い階層を持つニューラルネットワークが高い認識性能を示している[1]．音響特徴と画像特徴両者を用いた音声認識である視聴覚音声認識において，野田らはリップリーディングの学習に CNN を用いることで従来手法より高い性能を持つことを示した[2]．また，5000 時間の大量のニュース映像(英語)からなるデータベースを深層学習することでも高い認識性能を発揮することが示されている[3]．このことより大規模なデータを CNN で学習させることが認識精度向上につながることがわかる．

前者が使用したデータベースは約 1 時間と小規模であり，さらに高速度カメラを使用するなど大規模データ収集には一般的には困難な条件であった．後者は英語のデータベースであった．これらを踏まえて日本語のリップリーディングのデータベース構築に必要な条件を定義する．

① 様々なデータに対応できるよう収録環境・人数が多様である
② 学習に必要な音素を不可分なく含む
③ データの追加を容易にする為，特殊ではない環境で収録されている

簡易な環境でバランス良く音素を含むデータを収集することが，日本語における実用的なリップリーディング用映像データベース作成への一助となると考えられる．

３．新しいデータベースの構築

タブレット端末上で文章表示と動画撮影を用

い 40 人分の音声発話の映像の収集・作成を行った. 読み上げ文章には松永らが提唱する著作権フリーな文章から, 言語における音の最小単位である音素のバランスを考慮して収集した文章データベースを使用した[4]. 実際のリップリーディングは様々な環境下で行われることが想定されるが, 学習データ作成の為の音声認識を可能にする為, 雑音が比較的少なく, 逆光により顔が映らない状態にならないという条件のみ揃えた上で収集を行なった.

4. 評価実験

収集したデータベースを以下の方法に基づき学習・評価を行う.

リップリーディングにおける CNN の学習に時系列を考慮した CNN である Time Delay CNN(図 1)を導入し, 色チャネル方向に発話の時間方向に近接している複数フレームの画像を入力することで認識精度が向上することが確かめられている[5].

同様の実験を新しく収集したデータベースでも行い, 色チャネル方向に N(=1,3,5)枚入力したデータで学習を行った.

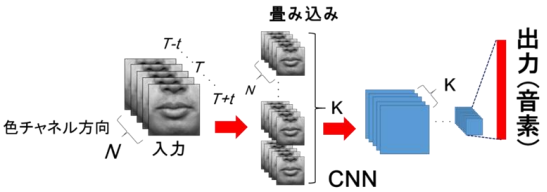

図 1 Time Delay CNN の概略図 [5]より引用

5. 結果

本研究で作成した映像データベースの概要を表 1 に示す. この映像データベースを用いた Time Delay CNN による学習結果を表 2 と図 2 に示す.

従来手法と比べ簡易な環境でも, 従来同様近接フレーム枚数を増加させると認識精度が向上する結果が得られている. ここで N は時間方向に入力した近接フレームの枚数, 認識率は全て音素認識率で算出している.

6. 結論

今回は従来と比べ簡易な環境, 著作権フリーなデータベースを用い, 従来のデータと同傾向の結果を出すことで, 将来的にデータを収集できる礎石を築いた.

今後は音響特徴との統合学習や, 単語やその並び方の集積である言語モデルを考慮した学習が考えられる.

表 1 構築した映像データベース

話者数	男性 31 人, 女性 9 人
読み上げに用いた文章	音素バランス文章データベース 700 文
映像時間	約 35 時間
フレームレート	30 fps
画像	BMP 形式, 約 380 万枚

表 2 構築した映像データベースの認識結果

N	1	3	5
音素認識率[%]	27.65	40.42	49.50

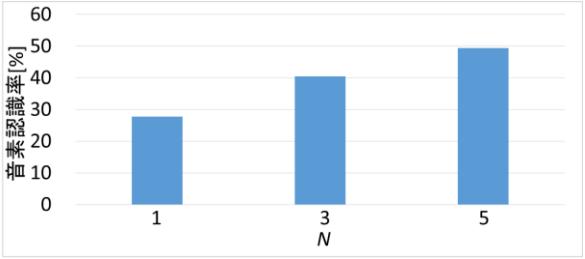

図 2 構築した映像データベースの認識結果

注:

[1] K. He, X. Zhang, S. Ren, J. Sun, ” Residual Learning for Image Recognition”, arxiv1512.0338, 2015

[2] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning", Applied Intelligence, Vol.42, Issue 4, pp. 722-737, 2015

[3] J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, ”Lip Reading Sentences in the Wild”, arXiv:1611.05358v1, 2016

[4] 松永寛之, 橋本直矢, 佐々木 一磨, 中臺 一博, 尾形 哲也:, “音素バランスを考慮した読み上げ用フリー文章データベースの構築手法”, 人工知能学会, 2016

[5] 橋本直矢, 佐々木一磨, 中臺一博, 尾形哲也, ” 時系列を考慮した Convolutional Neural Network による視覚音声認識のための音素識別”, 日本ロボット学会, 2016