唇差分画像と畳み込みニューラルネットワークを用いたリップリーディング

李, 一婷
高島, 悠樹
滝口, 哲也
有木, 康雄

(Citation)
神戸大学都市安全研究センター研究報告,20:73-78

(Issue Date)
2016-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81011508

(URL)
https://hdl.handle.net/20.500.14094/81011508

神戸大学都市安全研究センター
研究報告，第20号，平成28年３月

唇差分画像と畳み込みニューラルネットワークを用いたリップリーディング

Lipreading using a difference image of lips and convolutional neural networks

李　一婷[1)]
Yiting Li
高島　悠樹[2)]
Yuki Takashima
滝口　哲也[3)]
Tetsuya Takiguchi
有木　康雄[4)]
Yasuo Ariki

概要：近年，音声認識はパソコンや携帯電話でのインターフェースとして広まり，キーボードからの入力に替わるハンズフリーな技術として注目されています．しかしながら，雑音が多い状況下で認識性能が著しく低下してしまうことや，話すことが難しい場所では利用できないという問題がある．よって，雑音などの影響を受けない入力方法が求められる．本研究では，このような入力方法の一つのリップリーディングの実現を目標としている．リップリーディングは，唇の形や動きで発話内容を理解する方法である．本稿では，唇連続画像に対してフレーム間の変化を表現できる唇差分画像及び，近年，画像や音声認識に広く使われている畳み込みニューラルネットワーク（Convolutional Neural Network，CNN）を用いる手法を提案し，その有効性を示す．

キーワード：リップリーディング，唇差分画像，畳み込みニューラルネットワーク（CNN）

1.　はじめに

近年，リップリーディングが広く実用化されている．例えば，難聴者とのコミュニケーション、手の指に障碍を持つ人の文字入力や発話することがマナー違反になるような公共の場で利用する新たなインターフェースとして活用することを考えられる．しかし，現在のリップリーディングには，大語彙に対応する状況下で認識性能が著しく低下してしまうという問題点があり，リップリーディングの大きな課題となっている．一方，人間は発話内容を理解する際，様々の情報を統合的に利用している．音声が聞き取り難い場合，発話者の顔，特に唇に注目して発話内容を理解しようとし，逆に，唇の動きと音声が不一致の場合，唇の動きに影響されて発話内容を誤って理解してしまうこともある．これは，マガーク効果と呼ばれ，音韻知覚が音声の聴覚情報のみで決まるのではなく，唇の動きといった視覚情報からも影響を受けることが報告されている[1)]．このように唇の画像情報を利用するリップリーディングは，人間の発話を理解する際には重要なサポートとなる．

リップリーディング技術は，これまでの多くは難聴者を対象としたものであったが，現在では健常者や高齢者などを対象とした場合や，車内や会議室といった様々な実環境下での利用を対象とした場合など多様化して

いる [2)3)4)]．本研究では，健常者を対象としている．従来研究として，唇の連続画像を用いたリップリーディング手法である．しかし，普通の連続画像は発話に対して時間変化に対応できないという問題点がある．そのため，時間的に連続する画像の各フレームに対して前後何フレームの差分を計算し，得られた差分画像を連続画像のかわりに用いる．差分画像は，動きのある物体を検出する場合でよく使われている [5)6)]．さらに，従来手法の被験者発話時微小な位置ズレに弱いという問題点に対して，畳み込みニューラルネットワーク（Convolutional Neural Network，CNN）を用いて特徴抽出を行う手法を用いる．CNN は前段の特徴量を所定の範囲内で畳み込み演算をする畳み込み層と微小な位置ズレを問題にしないようにぼかし効果を加えるプーリング層を交互に配置した多層ニューラツネットワーク（NN）である．CNN では，局所的な平行移動に対して不変な出力がなされるため，唇における発話変動によるスペクトルの微小な変化(位置ズレ，局所的な歪み）や，唇画像における被写体ブレや唇検出(フェイスアライメント）の検出ズレ・トラッキングのズレに対して，頑健な画像特徴量抽出が行えると期待される．

2. 差分画像

各フレームの同じ位置の点の値をデータ列$\{y_i\}(i=1,\cdots,n)$に対して，最小 2 乗法より，直線$y=ai+b$を当てはめる問題を考える．このとき，データ列$\{y_i\}$を直線近似にした際の誤差の平均 2 乗和は以下の式で表される．

$$E=\frac{1}{n}\sum_{i=1}^{n}\{(ai+b)-y_i\}$$

それから，Eを最小化するa，bを求める．直線$\boldsymbol{y}$の傾きaは以下の式より求まる．

$$\mathrm{a}=\frac{\frac{1}{n}\sum_{i=1}^{n}(i\cdot y_i)-\left(\frac{1}{n}\sum_{i=1}^{n}i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}i^2\right)-\left(\frac{1}{n}\sum_{i=1}^{n}i\right)^2}$$

ここで，スペクトル乗法を表すパラメータのnフレームにおけるi番目の値を$c_i(n)$と記す．このとき，時刻nを中心とした区間$[n-\delta,n+\delta]$における$c_i(n)$の値に，直線を当てはめた場合の直線の傾きを$\Delta c_i(n)$で表すと，上の式より，

$$\Delta c_i(n)=\frac{\sum_{k=-\delta}^{\delta}k\cdot c_i(n+k)}{\sum_{k=-\delta}^{\delta}k^2}$$

が成り立つ．kは各フレームに対して前後何フレームの差分を計算するかをあらわす係数，差分フレーム数である．上の式で得られたパラメータを動的特徴と呼ぶ．

唇のオリジナル画像を，Figure 1 に示す．差分画像は，Figure 2 に示すように，各フレームの前後何フレームの差分を計算することで，フレーム間の変化を表現できる．

Figure 1　オリジナル唇画像

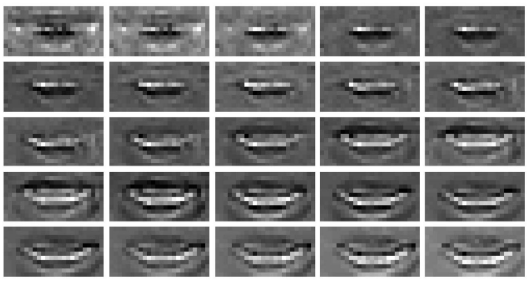

Figure 2　唇差分画像

3. 畳み込みニューラルネットワーク（Convolutional Neural Network，CNN）

CNN では，手書き文字認識の分野で広く使われていることもあって，移動不変性の獲得を目的としている．

移動不変性とは，検知対象が入力データのどこに存在していても正しく検知できることである．

従来のニューラルネットワーク（NN）では，入力の誤差が全てのユニットに伝搬してしまう．そこで CNN では，ある層のユニットへ接続する前層のユニットの数を制限する．この処理を局所接続性，もしくは局所受容野という．受容野を単位に局所性を表現しようとしているため，局所受容野における重みを他の局所受容野と共有する．CNN では入力の２次元特徴に対して，ある局所的なフィルタを畳み込む演算を各層で実行していることになる．そのため，通常の NN ではユニット間の接続に重みが付与されるが，CNN では各フィルタに付与されることになる．局所受容野と共有重みにより，局所的な情報のみをパスするフィルタ（局所的な誤差情報の全体への波及防止）と，入力データにおける認識対象の位置に依存しない接続方法を獲得したことになる．しかし，重みを共有することで，全ての局所受容野が同じ学習をすることになり，全ての局所受容野で類似する特徴しか認識できなくなってしまう可能性がある．そこで CNN では，局所受容野の次元数を増やし各次元それぞれで異なる特徴を獲得する．これに加えて CNN では，プーリングと呼ばれる処理が行われる．その目的は，局所受容野と共有重みで得られた移動不変性をさらに強めることである．そのため，細かい位置ずれを問題にしないように，入力データの低解像度化を行う．

実験で用いた CNN は通常の CNN とボトルネック層を含む３層の MLP(Multi-Layer Perceptron)とが階層的に接続された多層 NN を考える．MLP を３層設計し，中間層のユニット数を少なく抑える（ボトルネック）構成をとっている．ボトルネック特徴量はボトルネック層のニューロンの線形和で構成される空間であり．少ないユニットで多くの情報を表現しているため，入力層と出力層を結び付けるため重要な情報が集約されている．

4. 提案手法

Figure 3 に提案手法の流れを示す．画像に対しては，Constrained Local Model(CLM)を用いて，画像上の目，口，鼻，眉，輪郭の位置決定(フェイスアライメント）を行い，唇領域を抽出する．次に，抽出した唇画像の各画素の時系列に対して，3 次スプライン補間を適用し，各画素のサンプリング数を増やす．そのままのフレームレートでの特徴量抽出を行うと，認識率の低下を招く恐れがあるためである．

最後に，唇画像列を，事前に学習しておいたボトルネック構造のネットワークに入力し，画像のネットワークからボトルネック特徴量を抽出する．ここで，音素ラベを教師信号として用いる．その後，画像ボトルネック特徴量を Hidden Markov Model（HMM）の入力とすることで，リップリーディングを実現する．

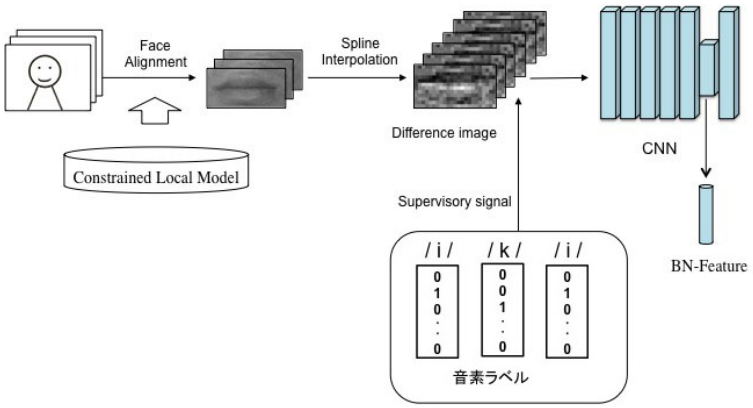

Figure 3　提案手法の概要

5. 評価実験

(1) 実験条件

本稿では，モデルの有効性を確認するため，孤立単語認識の実験を行う．実験用データとして，健常者男性 1 名が発話する ATR 音素バランス単語 B セット(216 単語)を用いる．また，CNN 及び音響モデルの学習データとし

て，同一話者が発話する ATR 音素バランス単語 A セット偶数番（2620 単語）を用いる．これらのデータは，発話時に顔正面に置いたビデオカメラから録画したものであり，音声データと画像データからなる．録画のフレームレートは 60fpsである．またフェイスアライメントによって得られた唇領域画像のサイズ･縦横比は，フレーム毎に異なる．そのため，バイキュービック法を用いて，全ての唇領域画像を縦 12pixel，横 24pixel，縦横比 1：2 にリサイズした．加えて，画像 CNN の入力として輝度画像を用いるため，リサイズ後に RGB 画像から輝度画像への変換を行った．輝度 Y は，R 成分，G 成分，B 成分の重み付き和により計算される．

$$Y = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B$$

実験で用いる読唇のためのモデルは，54 音素の monophone-HMM で，各 HMM の状態数は 5，状態あたりの混合分布数は 8 である．CNN の各層おける特徴マップのサイズには Table 1 の値を用いて実験を行った．C と P はそれぞれ畳み込み層とプーリング層と表す．MLP は M1，M2 と M3 からなる．

Table 1　畳み込みニューラルネットワークのパラメータ数

	Input	C	P	M1	M2	M3	Output
Visual CNN	1, 12×24	13, 8×20	13, 4×10	108	30	108	54

(2) 結果と考察

Figure 1 に，話者一名の評価データ(216 単語）に対する単語認識結果を示す．縦軸は認識精度(corr[%] = H / N × 100，N = 全単語数，H = 正解単語数）を表し，横軸は，唇差分画像の特徴量とそのフレーム関数，静的特徴量と DCT 特徴量を示している．静的特徴はオリジナル唇画像を提案手法と同じ構造の CNN で学習してから得られた特徴である．まず唇差分画像の特徴量用いた場合は，静的特徴と比べて，認識精度の改善が最大 12.82% の認識率の改善が見られた．これは，唇差分画像はフレーム間での時間的変化を表示することが可能になったためと考えられる．また，唇差分特徴のフレーム関数 k は 3 と設定した時，得られた認識精度は一番高いのは 71.76% となって，ベースラインの DCT 特徴量より 19.91% の認識率改善があり，オリジナル画像から求められた静的特徴量より 12.82% の認識率改善が見られた．これは，単語の発音が短いが，フレームームの数が高くなると，フレーム間の動きを明らかになくなる可能性があると考えられる．

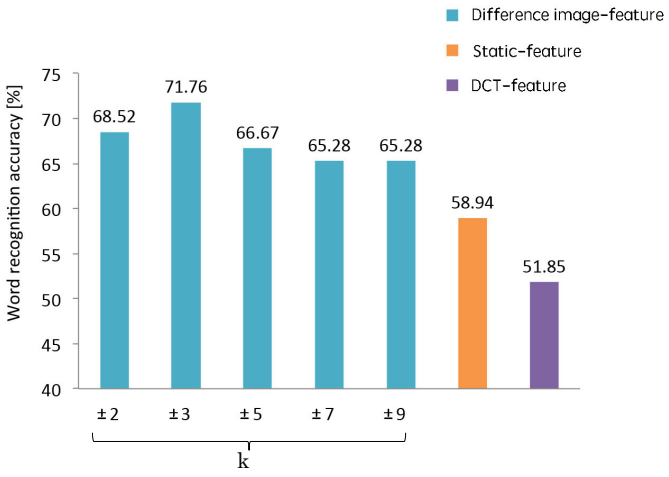

Figure 3　認識結果

6.　おわりに

本稿では，健常者一名を対象とし，唇の動きから発話内容を読み取るリップリーディング(読唇）手法を提案した．孤立単語認識実験を行った結果，唇の差分画像を用いて，認識率の改善を見られ，特に各フレームの前後 3 フレームの差分を計算して，CNN の入力を用いた場合，ベースラインの DCT 特徴に対して 19.91% の認識

率の改善があり，提案手法の有効性が確認できた．今後の課題として，画像のネットワーク構造の変更とメルマップ・唇領域画像の処理の再改良が必要である．また，実環境でのユーザビリティが向上すると考えられるため，被験者数を増やして実験を行う必要がある．

参考文献

1） McGurk, H., & MacDonald, J.: Hearing lips and seeing voices, *Nature*, *264*, pp. 746-748, 1976.

2） 中川聖一：音声認識研究の動向，電子情報通信学会論文誌,Vol. J83-D, No. 2, pp. 433-457, 2000.

3） 馬場朗，ほか: 高齢者音響モデルによる大語彙連続音声認識，電子情報通信学会論文誌，Vol. J85-D2, No. 3, pp. 390-397, 2002.

4） 鮫島充，ほか: 子供音声認識のための音響モデルの構築および適応手法の評価，電子情報通信学会技術研究報告，SP2004, pp. 109-114, 2002.

5） Bruzzone L., Prieto D. F.: Automatic analysis of the difference image for unsupervised change detection, Geoscience and Remote Sensing, IEEE Transactions on, Vol. 38, No. 3, pp. 1171-1182, May 2000.

6） Cheng Y H, Wang J.: A Motion Image Detection Method Based on the Inter-Frame Difference Method, Applied Mechanics & Materials, Vols. 490-491, pp. 1283-1286, 2014.

筆者： 1）李一婷，システム情報学研究科，学生；2）高島悠樹，システム情報学研究科，学生；3）滝口哲也，都市安全研究センター，准教授；4）有木康雄，都市安全研究センター，教授

Lipreading using a difference image of lips and convolutional neural networks

Yiting Li
Yuki Takashima
Tetsuya Takiguchi
Yasuo Ariki

Abstract

In recent years, speech recognition spread throughout to a computer or mobile phone interface, and it has attracted attention as a hands-free technology that replaces the input through a keyboard. However, the recognition performance significantly decreases under a noisy condition, or in some conditions which are hard to speak loudly. In order to deal with the problem, as a noise-robust method, lipreading has been studied.

In this paper, we investigated a difference image of lips, which can be used to express the difference between frames of an image sequence. Then we propose a robust feature extraction method using Convolutional Neural Networks (CNN). CNN has showed success in achieving translation invariance for many image-processing tasks. The success of CNN is largely attributed to the use of local filtering and pooling in the CNN architecture.

Its effectiveness is confirmed by word recognition experiments. The experiment results showed that the CNN-based feature extraction method, which used the difference image of lips, outperformed the conventional method.