Convolutional Neural Networks を用いた構音障害者のための音声認識*

☆吉岡利也，中鹿亘，滝口哲也，有木康雄 (神戸大)

1 はじめに

本研究では，アテトーゼ型脳性麻痺による構音障害者を対象とした音声認識の実現を目指している．彼らは意図的な動作時や緊張状態にある場合に筋肉の制御が困難となり，アテトーゼと呼ばれる不随意運動を伴う．この影響により，構音障害者の発話は非常に不安定になるため，従来の音声認識システムでは認識精度が著しく低下する．

音声認識に用いる特徴量として，MFCC (Mel-Frequency Cepstrum Coefficient) が一般的であり，健常者の音声認識において高い性能を示している．しかし，構音障害者の音声認識においては十分な精度が得られていない．我々はこれまでに，PCA (Principal Component Analysis) やランダムプロジェクションを用いた構音障害者の音声認識に適した特徴量抽出法を提案してきた [1, 2]．本稿では，さらなる認識率の改善を目指し，数フレーム程度の部分信号から時間-周波数の 2 次元特徴を抽出し，画像処理の技術を用いて特徴量抽出を行うアプローチをとる．

提案手法では，音声のスペクトログラムから得られた 2 次元特徴を入力層，入力層の音素情報を要素として持つベクトルを出力層とする Convolutional Neural Networks (CNN) [3, 4, 5] を構築し，特徴量抽出に用いる．3 章では，健常者，および構音障害者を対象とした単語認識実験を行い，提案手法の効果について検討する．

2 Convolutional Neural Networks

CNN は LeCun ら [3] によって提案された多層型ニューラルネットワークの一種であり，特に画像処理やパターン認識の分野で高い効果を示している．CNN はフィルタの畳み込み演算を行う畳み込み層 $C_{m\to n}^{p\times q}$ と，平滑化を行うプーリング層 $S_m^{p\times q}$ を交互に積み重ねることで，効果的に 2 次元的な情報を集約させるといった特徴を持つ．

本稿では，Fig. 1 のような構造を持つ CNN を考える．入力層としては，隣接する数フレーム分のメル周波数スペクトラムから抽出した 2 次元特徴 (メルマップ) を用いる (Fig. 2)．出力層には，入力層の各メルマップに対応した音素ベクトル (例えば，入力層が音素/i/のメルマップであれば，出力層は/i/に対応するユニットが値 1 で他のユニットが値 0 のベクトルとなる) を置き，逆誤差伝播 (Back Propagation) 法により層間の結合パラメータを修正する．

未知の音声に対しては，同様にメルマップを抽出し，学習した CNN の入力層とすることで音素情報を含んだ出力層を得る．その出力層に対して PCA を適用し，直交化・次元圧縮を行ったものを特徴量とする．

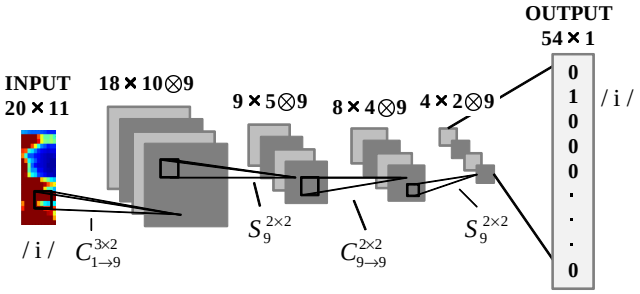

Fig. 1 The proposed CNN architecture. $C_{m\to n}^{p\times q}$ represent convolutional operations with convolution kernels of size $p\times q$. $S_m^{p\times q}$ are subsampling operations with $p\times q$ kernels. The layers corresponding to $C_{m\to n}^{p\times q}$ or $S_m^{p\times q}$ are fully connected; otherwise connected 1 by 1. $i\times j\otimes k$ above each layer means that the layer has k maps of size $i\times j$.

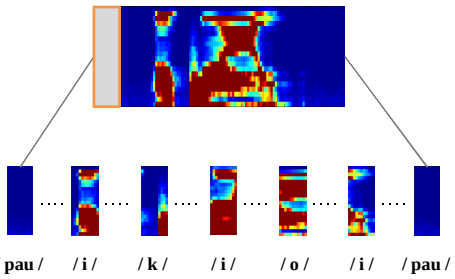

Fig. 2 Mel maps of 20 × 11 pixels are obtained from each speech data.

3 評価実験

提案手法の認識性能を評価するため，健常者，および構音障害者を対象とした単語認識実験を行った．

3.1 実験条件

健常者の音声データとして，話者 5 名 (男性 3 名，女性 2 名) が発声したラベル付き ATR 音素バランス単語 (210 単語) を用い，男性 1 名を評価データ，残りの男女 4 名を学習データとして使用する．音素数は 54，音響モデルは 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．

また，障害者の音声データとして，構音障害を患う男性 1 名の発話を収録した．発話内容は，健常者と同様に ATR 音素バランス単語 (210 単語) とし，各単語は連続で 5 回発話されている (Fig. 3)．各単語の第 1 発話を評価データ，第 2〜5 発話を学習データとして使用する．音素数，および音響モデルの構造は健常者のものと同様である．

*Speech Recognition for Articulation Disorders Based on Convolutional Neural Networks. by Toshiya YOSHIOKA, Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

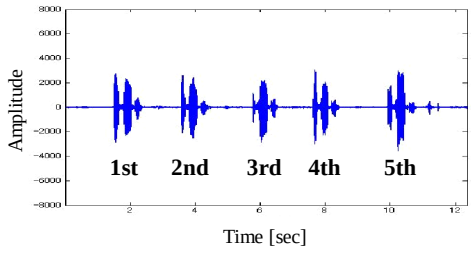

Fig. 3　Example of recorded speech data.

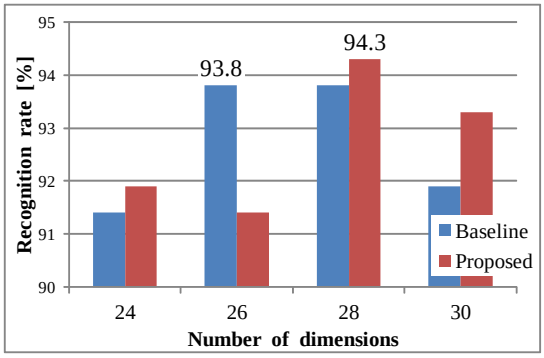

Fig. 4　Word recognition results [%] for a physically unimpaired person.

特徴量は，健常者，構音障害者ともに，CNN で得られた出力層を PCA で 24，26，28，30 次元に圧縮したものを用いる (提案手法)．MFCC+Δ (24，26，28，30 次元) をベースラインとし，提案手法との比較を行う．

3.2　評価結果

まず，健常者音声に対する実験結果を Fig. 4 に示す．Fig. 4 より，提案手法によって認識率が 93.8% から 94.3% に改善された．この結果より，正確な音素ラベルが与えられている健常者音声であれば提案手法の効果が確認できた．

Fig. 5 は，構音障害者の音声に対する単語認識実験の結果である．障害者の場合，ベースラインと比較して認識率が大きくて低下してしまった．その大きな要因として，アライメントの精度が不十分であることが挙げられる．構音障害者の音声データは音素ラベルが与えられていないため，HTK の強制アライメント機能を利用している．しかし，子音の欠如や，構音障害者特有の言いよどみ (吃音) の影響により，データによってアライメントの精度にバラつきが生じている．そのため，CNN の教師あり学習が上手くいっておらず，認識に有効な特徴量が得られなかったと考えられる．

4　おわりに

本稿では，画像処理やパターン認識の分野で効果が確認されている CNN をベースとした特徴量抽出法を検討した．

評価実験では，健常者と構音障害者の両者に対し

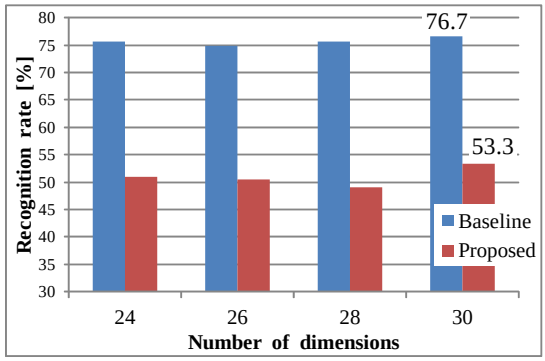

Fig. 5　Word recognition results [%] for a person with an articulation disorder.

て単語認識実験を行った．その結果，健常者については提案手法の有効性が確認できたが，構音障害者の音声では，アライメントの精度が不十分であることから認識率の改善が見られなかった．

今後は，構音障害者の音素体系に注目してより正確な音素ラベルを得ると共に，畳み込み層のフィルタの数やサイズ，プーリング層のプーリングサイズを工夫するなどして [6]，精度の改善に取り組んでいく予定である．

参考文献

[1] H. Matsumasa *et al.*, “PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders,” Interspeech2007, pp. 1150-1153, 2007.

[2] T. Yoshioka *et al.*, “Evaluation of Random-Projection-Based Feature Combination on Dysarthric Speech Recognition,” American Journal of Signal Processing, 3(3), pp. 41-48, 2013.

[3] Y. Lecun *et al.*, “Gradient-based learning applied to document recognition,” Proc. IEEE, pp. 2278-2324, 1998.

[4] Y. Lecun *et al.*, “Convolutional Networks and Applications in Vision,” Proc. International Symposium on Circuits and Systems, IEEE, pp. 253-256, 2010.

[5] T. Nakashika *et al.*, “Local-feature-map Integration Using Convolutional Neural Networks for Music Genre Classification,” Interspeech2012, 2012.

[6] L. Deng *et al.*, “A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion,” ICASSP2013, pp. 6669-6673, 2013.