Convolutive Bottleneck Network 特徴量を用いた構音障害者の音声認識*

☆吉岡利也，中鹿亘，滝口哲也，有木康雄 (神戸大)

1 はじめに

近年，音声認識技術は飛躍的に進歩し，様々な環境や場面での利用が期待されている．例えば，音声による家電製品の操作や，会議音声の議事録化など様々な分野に応用されている．また，これまでの多くは成人を対象としたものであったが，現在では高齢者や子供など成人と発話スタイルの異なる音声であっても高い精度が得られており，多くの人にとって利用する機会が増加している．しかし，構音障害などの言語障害を患う方を対象とした音声認識は非常に少ない．

言語障害の原因の一つとして，脳性麻痺を挙げることができる．脳性麻痺とは，中枢神経系の損傷によって筋肉の制御が不安定になり，痙攣や麻痺，その他の神経障害が起こる症状のことである．脳性麻痺の程度や症状は，人によって様々である．分類としては，1) 痙直型，2) アテトーゼ型，3) 失調型，4) 緊張低下型，5) 固縮型，6) 混合型，の 6 種類に分類することが出来る [1].

本稿では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型の脳性麻痺では，筋肉の随意運動や姿勢の調整を行っている大脳基底核 (大脳皮質，視床や脳幹を結びつけている 神経核の集まり) に損傷を受けたことにより，アテトーゼと呼ばれる筋肉が不随に動き正常に制御できない症状が現れる．この症状は緊張時や意図的な動作を行う場合に多く発生するため，発話時に筋肉の緊張が起こり正しく構音できない場合がある．発話が困難な方でも，手話認識や音声合成システム [2] を使用することでコミュニケーションをとることが可能であるが，脳性麻痺患者の多くは手足が不自由であり，音声に頼るしかない状況が考えられる．そのため，構音障害者のための音声認識には十分なニーズがあり，研究の必要性があるといえる．

構音障害者の発話スタイルは健常者と大きく異なるため，従来の音声認識で用いられている不特定話者モデルでは認識精度が著しく低下する．また，アテトーゼによる筋肉の不随意運動を伴うため，発話内容が同じであっても，発話のばらつきが健常者と比べて大きくなるという課題が考えられる．そこで本稿では，構音障害者の音声のみを用いた特定話者モデルを構築することで，認識精度の改善を行う．また，発話のばらつきが大きくなるという課題に対して，多層ニューラルネットワーク (NN) の一種である，CNN (Convolutional Neural Network) [3, 4, 5] を用いた発話変動にロバストな音声特徴量抽出法を提案する．

音声認識において，最も一般的に用いられている特徴量は MFCC (Mel-Frequency Cepstrum Coefficient) である．MFCC はメル尺度フィルタバンクの短時間対数エネルギー出力系列に対して，離散コサイン変換を適用し，音声のスペクトル包絡成分に対応する低次ケプストラムのみを抽出した特徴量であり，健常者の音声認識において高い精度を示している．しかし，アテトーゼ型の構音障害者の発話は，筋肉の緊張のため不安定になり変動しやすい．そのため，音声の時間フレームごとに独立して計算される MFCC などのケプストラム特徴量では，発話のばらつきを考慮していないため，十分な精度が得られない．そこで本稿では，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，入力データの変動に対する不変性を有する CNN を用いて特徴量抽出を行うアプローチをとる．

CNN は，前段の特徴量を所定の範囲内 (局所受容野) で畳み込み演算をする層 (畳み込み層) と，細かい位置ずれを問題にしないようにぼかし効果を加える層 (プーリング層) を交互に配置した多層 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，構音障害者特有の発話変動によるスペクトルの微小な変化 (位置ずれ，局所的な歪み等) に対して頑健性を持つ．また，構音障害者の音声のみを用いてネットワークを学習することによって，より障害者音声に特化した特徴量抽出が行えると期待される．

CBN (Convolutive Bottleneck Network) [6] とは，通常の CNN と，ボトルネックの構造を持つ 3 層以上の MLP (Multi Layer Perceptron) とが階層的に接続された構成を持つ多層 NN である．ボトルネック層では，極端に少ないユニットで多くの情報を表現しているため，入力データに含まれる本質的な情報が獲得できると考えられる．提案手法では，音声スペクトラムを短時間フレームで分割した 2 次元特徴を入力，その音素ラベルを出力とした CBN を学習し，ボトルネック特徴量の抽出を行う．

第 2 章で CNN の基本要素について述べる．第 3 章では，提案する CBN の構造，および音声特徴量抽出の流れを示す．第 4 章では，ネットワーク特徴量の認識性能を評価するため，構音障害者の孤立単語認識実験を行う．最後に，第 5 章で本研究の成果をまとめるとともに，今後の研究課題について述べる．

2 Convolutional Neural Networks

ここでは，CNN を構成する基本要素として，畳み込み層，およびプーリング層の処理について述べる．

2.1 畳み込み層

$n_x \times n_x$ の 2 次元特徴 $\mathbf{x}$ に対する，$n_w \times n_w$ のフィルタ $\mathbf{w}$ の畳み込みを考える．この畳み込みを $\mathbf{h} = \mathbf{w} * \mathbf{x}$ と書くと，一般に出力 $\mathbf{h}$ は $n_h \times n_h\,(n_h \equiv n_x - n_w + 1)$ の 2 次元特徴となる．CNN は通常，このようなフィルタを複数個 $\{\mathbf{w}_1, \cdots, \mathbf{w}_L\}$ 持ち，フィルタ毎に異なる出力のセット $\{\mathbf{h}_1, \cdots, \mathbf{h}_L\}$ を持つ．畳み込み演算の結果として得られる同じサイズをもつ 2 次元特徴のセットを特徴マップ (feature

*Speech Recognition for Articulation Disorders using Convolutive Bottleneck Network Features. by Toshiya YOSHIOKA, Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

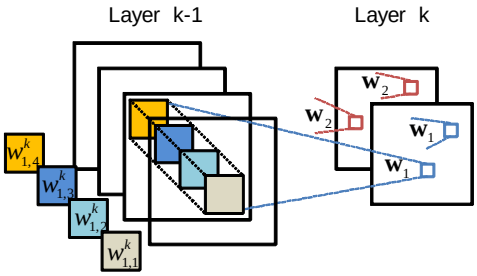

Fig. 1　Example of a convolutional layer.

map) と呼ぶ.

第 $k-1$ 層から第 k 層への計算例を以下に示す. 第 k 層の j 番目の特徴マップ $\mathbf{h}_j^k$ $(\in \mathbb{R}^{n_h \times n_h})$ は，第 $k-1$ 層の全特徴マップ $\{\mathbf{h}_1^{k+1}, \cdots, \mathbf{h}_i^{k+1}, \cdots, \mathbf{h}_I^{k+1}\}$ を元に,

$$\mathbf{h}_j^k = f\left(\sum_{i \in I} \mathbf{w}_{j,i}^k * \mathbf{h}_i^{k-1} + \mathbf{b}_j^k\right) \qquad (1)$$

のように計算される．$\mathbf{w}_{j,i}^k$ は，第 $k-1$ 層 i 番目のマップから第 k 層 j 番目のマップへの結合重み (フィルタ) である．関数 f には何らかの非線形性を持つものが用いられるが，本稿では以下のようなシグモイド関数を選んでいる．

$$f(x) = \frac{1}{1+\exp(-x)} \quad f(x) \in [0:1] \qquad (2)$$

畳み込み層では，特徴マップの画素と 1 対 1 でユニットが対応する前層に対して，その $n_w \times n_w$ の部分集合と上位層のユニットが対応付けられ，両者が結合される (局所受容野)．このように CNN は，層間の結合が位置ごとに局所的になされ，その結合重みが (上位層の全ユニット間で) 共有されている (フィルタの係数 $\mathbf{w}_l$ がマップごとに同じ) ことを特徴とする (Fig. 1).

2.2　プーリング層

これに加えて CNN では，その次の層でプーリングと呼ばれる処理が行われる．畳み込み演算によって得られた特徴マップ $\mathbf{h}_j^k$ の小領域 $\mathbf{P}$ $(m \times m)$ 内の応答をまとめ，次の層の 1 ユニットに結び付ける．この $m \times m$ ブロックはフィルタ出力層で重複させないため，プーリング層のサイズはフィルタ出力層のサイズの m 分の 1，すなわち $n_h/m \times n_h/m$ となる.

プーリングにはいくつかのバリエーションがあるが，ここでは以下のような平均プーリング (average pooling) を用いる.

$$\mathbf{h}_{j,\acute{u},\acute{v}}^{k+1} = \frac{1}{m^2} \sum_{(u,v) \in \mathbf{P}} \mathbf{h}_{j,u,v}^k \qquad (3)$$

ここで，j は特徴マップのインデックス，u, v はそれぞれ画像内の位置を示すインデックスである．平均プーリングでは，第 k 層での小領域 P での応答の平均値が第 $k+1$ 層の 1 ユニットとして計算される．この処理によって，フィルタの出力層の隣接ブロックのどれがアクティブとなっても，そのブロックを集積したプーリング層のユニットがアクティブとなるため，微小な並進移動に対する不変性が実現される.

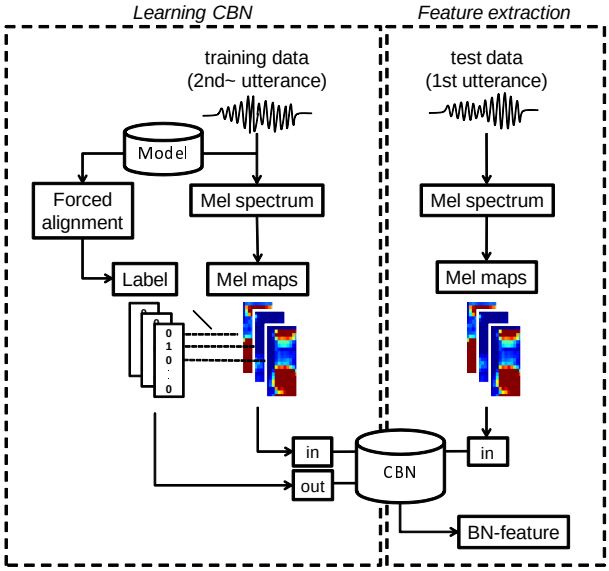

Fig. 3　Flow of the bottle-neck feature extraction using CBN.

3　提案手法

3.1　Convolutive Bottleneck Network

提案手法では，Fig. 2 に示すように，ボトルネックの構造を持つ CNN (以下 CBN) を考える．入力層からの数層は，フィルタの畳み込みとプーリングをこの順で何度か繰り返す構造をとる．つまりフィルタ出力層，プーリング層の 2 層を，プーリング層を次の層の入力層とする形で積み重ねる．出力層は識別対象のクラス数と同じサイズを持つ一次元ベクトルであり，そこに至る何層かは畳み込み・プーリングを挟まない全結合の NN (MLP) とする．提案手法では，MLP を 3 層に設計し，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の重要な情報が集約されていると考えられる．そのため，LDA や PCA と同じような次元圧縮処理の意味合いも合わせ持つ．ボトルネック特徴量は，American Broad News コーパスなどの標準的な評価セットでの改善が報告されており [7]，提案手法においてもこのボトルネック特徴量を音声認識に用いる.

3.2　ボトルネック特徴量の抽出

Fig. 3 は，提案手法による音声特徴量抽出の流れを示している．まず，構音障害者が発話した音声データを用いて，Fig. 2 のようなネットワークの学習を行う．ネットワークの入力層 (in) には，学習データのメル周波数スペクトラムを，オーバーラップを許しながら数フレームごとに分割して得られた 2 次元画像 (以下メルマップ) を用いる．出力層 (out) の各ユニットには，入力層のメルマップに対する音素ラベル (例えば，音素/i/のメルマップであれば，/i/に対応するユニットだけが値 1，他のユニットが値 0 にな

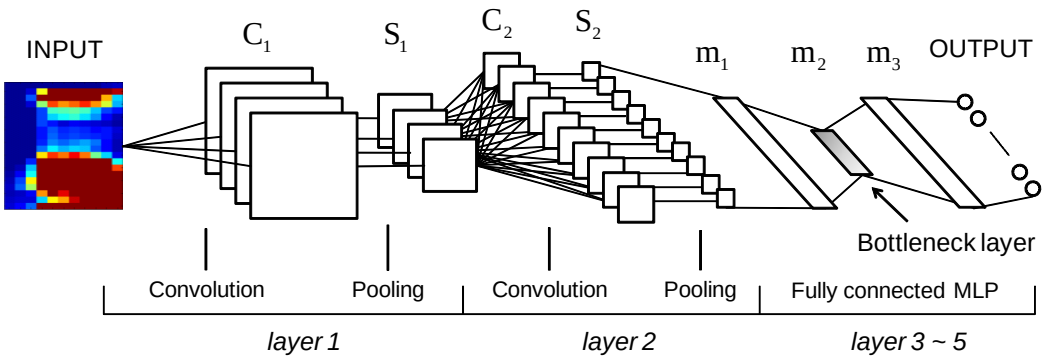

Fig. 2　Convolutive Bottleneck Network (CBN).

る) を割り当てる．音素ラベルを用意するために必要な学習データの音素境界ラベルは，学習データを用いて構築された音響モデルと，その読み上げテキスト (意図された音素列) を用いた強制切り出し (forced alignment) によって求める．CBN はランダムな初期値から学習を開始し，確率的勾配降下法 (Stochastic Gradient Descent, SGD) を用いた誤差逆伝搬により，結合パラメータを修正する．

次に，学習したネットワークを用いて特徴量抽出を行う．学習データと同様にテストデータのメルマップを計算し，学習した CBN への入力とする．その後，畳み込みフィルタとプーリングによって入力データの局所的特徴を捉えて，後部の MLP 層によって音素ラベルへと非線形に変換する．入力データの情報はボトルネック層上に集約されているため，提案手法では，このボトルネック特徴量を用いて音声認識を行う．

4　評価実験

ここでは，構音障害者の音声データのみを用いた孤立単語認識実験を行う．Fig. 2 に示す CBN でボトルネック特徴量を抽出し，認識性能を評価する．

4.1　実験条件

評価対象として，構音障害を患う男性 1 名が発話する ATR 音素バランス単語 (216 単語) を用いた．各単語は連続で 5 回発話されており (Fig. 4)，実験では各発話を切り出した合計 1,080 単語を使用する．本稿では，各単語の第 1 発話 (216 単語) を評価データ，残りの第 2 ～ 5 発話 (864 単語) を CBN，および音響モデルの学習データとする．

音声の標本化周波数は 16kHz，語長 16bit であり，音響分析には Hamming 窓を用いている．STFT におけるフレーム幅，シフト幅はそれぞれ 25ms, 10ms である．本稿で用いる音響モデルは，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．

ボトルネック層のユニット数が 28，30，32 のネットワークを用意し，各ネットワークで得られたボトルネック特徴量 (28，30，32 次元) を音声特徴量として用いる．ケプストラム特徴量である MFCC+ΔMFCC (28，30，32 次元) をベースラインとし，提案手法との比較を行う．

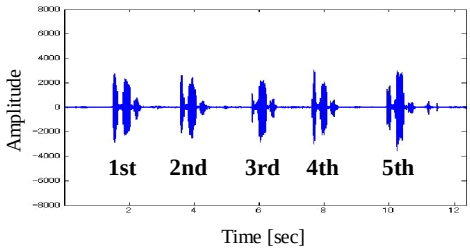

Fig. 4　Example of recorded speech data.

4.2　ネットワークのサイズ

本稿では，Fig. 2 に示すように，2 層の CNN (ここでは，畳み込み層とプーリング層をまとめて 1 層としている) と，ボトルネック層を含む 3 層の MLP とが階層的に接続された 5 層構造のネットワークを考える．入力層には，39 次元のメル周波数スペクトラムをフレーム幅 13，シフト幅 1 で分割したメルマップを用いる．CNN の各層における特徴マップのサイズには Table 1 の値を用いた．畳み込みフィルタの数とサイズ，およびプーリングサイズは，これらの値から一意に決定される．なお，MLP の各層 (ボトルネック層は除く) のユニット数は 108，出力層のユニット数は 54 としている．

Table 1　Size of each feature map. $(k, i \times j)$ indicates that the layer has k maps of size $i \times j$.

C1	S1	C2	S2
13，36×12	13，12×4	27，9×3	27，3×1

4.3　ネットワークの学習方法

ネットワークの学習データには各単語の第 2 ～ 5 発話を用いる．各学習データについて，メル周波数スペクトラムを短時間フレームで分割したメルマップと，その音素ラベルのペアを用意する (Fig. 3)．以降，これらのペアを訓練セットと呼ぶ．本稿で用いるネットワークは，この訓練セットで 100 回の繰り返し学習を行う．

畳み込み層のフィルタ係数 $\mathbf{W}$ は，下式で表される normalized initialization [8] で初期化した．

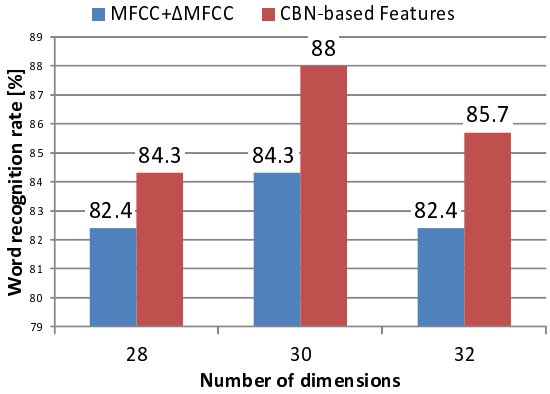

Fig. 5 Word recognition accuracy (%) for the 1st utterances of an articulation disorder using BN-features.

$$\mathbf{W} \sim \mathbf{U}\left(-\sqrt{\frac{6}{n_j + n_{j+1}}}, \sqrt{\frac{6}{n_j + n_{j+1}}}\right) \quad (4)$$

ここで U は一様分布の乱数，n_j および n_{j+1} は特徴抽出器の入出力特徴マップの画像数である．識別層の重み，およびバイアスは値 0 で初期化した．これらの値はネットワークの出力と教師データとの二乗誤差を最小とするように誤差逆伝搬法で学習し，訓練セットを 50 個ごとに区切ったミニバッチごとに誤差の平均値で更新した．学習率には 0.1 を用いる．

4.4 評価結果

構音障害者の第 1 発話に対する評価結果を Fig. 5 に示す．Fig. 5 より，特定話者モデルを用いた音声認識において，約 4% (84.3% → 88%) の認識精度の改善がみられた．従来のケプストラム特徴量では考慮していない平行移動不変性によって，構音障害者特有の発話変動によるスペクトルの微小な変化に対応することが可能になったと考えられる．

次に，畳み込みフィルタおよびプーリングの効能を評価するため，ボトルネック構造は同じで，CNN の層数 (畳み込み層，プーリング層) のみが異なる 2 つのネットワークを用意し，提案手法での評価を行った．一つ目は，ボトルネック層を含む 5 層の MLP で構成されるネットワーク (Deep Bottleneck Network, DBN) であり，中間層のユニット数は全て 108 である．二つ目は，CNN を 1 層とし，残りの 4 層を MLP とした 5 層構造のネットワークである．MLP 層への入力特徴マップのサイズは表 1 と同様の値を用いた．畳み込みフィルタの数とサイズ，およびプーリングサイズは，この値から一意に決定される．各ネットワークによる評価結果を Table 2 に示す．

Table 2 より，2 層の CNN と，ボトルネック層を含む 3 層の MLP で構成させる CBN が最も良い性能を示した．DBN では，従来のケプストラム特徴量を下回る結果となっており，障害者音声における CNN 構造の必要性が確認できる．

Table 2 Word recognition accuracy as a function of # of convolutional layers.

# of Convolutional, vs. Fully Connected Layers	Acc.(%)
No conv, 5 full (DBN)	83.3
1 conv, 4 full (CBN)	84.7
2 conv, 3 full (CBN)	**88.0**
Baseline (MFCCs)	84.3

5 結論

本論文では，構音障害者を対象とした音声認識の実現に向けて，障害者音響モデルを用いた認識実験を行った．さらに，筋肉の緊張により発話が変動しやすいという障害者特有の問題に対して，ボトルネックの構成を持つ CNN (CBN) を用いた特徴量抽出法を提案した．構音障害者 1 名を対象とした孤立単語認識実験より，提案手法によって約 4%の認識性能の改善がみられた．

構音障害の症状や引き起こされる発話変動の傾向は，人によって様々であるため，今後は被験者の数を増やして提案手法の有効性を確認する予定である．

参考文献

[1] S.T. Canale, and W.C. Campbell, “Campbell’s Operative Orthopaedics,” Mosby-Year Book, 2002.

[2] C. Veaux *et al.*, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders,” in Interspeech, 2012.

[3] Y. Lecun, and Y. Bengio, “Convolutional networks for images, speech, and time-series,” in The Handbook of Brain Theory and Neural Networks, 3361, 1995.

[4] Y. Lecun *et al.*, “Gradient-based learning applied to document recognition,” in Proceeding of the IEEE, pp. 2278-2324, 1998.

[5] H. Lee *et al.*, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in Advances in Neural Information Processing Systems 22, pp. 1096-1104, 2004.

[6] K. Veselý *et al.*, “Convolutive bottleneck network features for LVCSR,” in ASRU, pp. 42-47, 2011.

[7] C. Plahl *et al.*, “Hierarchical bottle neck features for LVCSR,” in Interspeech, pp. 1197-1200, 2010.

[8] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in International Conference on Artificial Intelligence and Statistics, pp. 249-256, 2010.