発話に不自由のある聴覚障害者の発話音声認識の検討*

☆柿原康博, 滝口哲也, 有木康雄 (神戸大), 三谷信之, 大森清博 (福祉のまちづくり研究所)

1 はじめに

本研究では，聴覚障害者のためのコミュニケーション支援技術 (発話訓練など) の開発の第一歩として，(発話に不自由のある) 聴覚障害者発話の音声認識の検討を行う．一般に聴覚障害者とは聞こえの不自由な人を指すが，聞こえの程度は聴覚障害の原因や種類によって異なる．発話の程度に関しても，聴覚障害になった時期が音声言語の獲得時期の前であるか後であるかによって異なり，発話訓練の有無にも左右される [1]．発話訓練を受けていても，先天聾である場合や音声言語の習得前に高度・重度難聴となった場合，発話のスタイルは独特であり，健常者とのコミュニケーションが難しい場合がある．本研究では，聴覚障害者の発話のみを用いた特定話者モデルによる認識を行った．これは，聴覚障害者の発話スタイルは健常者と大きく異なり，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 (79.1%→3.8%) するためである．実験として，音声認識において最も一般的に用いられている MFCC(Mel-Frequency Cepstrum Coefficient) を用いた音声認識を行った．また，言語障害者 (構音障害者) の音声認識 [2] において認識精度の向上がみられている CNN(Convolutional Neural Network)[3, 4, 5] のボトルネック特徴量を用いた音声認識を行った．

2 CNN のボトルネック特徴量

2.1 Convolutive Bottleneck Network

提案手法では，Fig. 1 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 [6]，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．

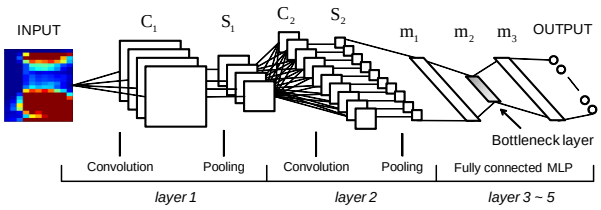

Fig. 1 Convolutive Bottleneck Network (CBN).

2.2 ボトルネック特徴量の抽出

まず，聴覚障害者が発話した音声データを用いて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ルを，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．

3 評価実験

3.1 実験条件

実験として聴覚障害者の音声データを用いた孤立単語認識実験を行った．評価対象として，聴覚障害者の男性 1 名が発話する ATR 音素バランス単語 (216 単語) を用いた．CBN および音響モデルの学習データとして，同じ聴覚障害者が発話する ATR 音素バランス単語 (1310,2620 単語) を用いた．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 30 のネットワークを用意し，そこで得られたボトルネック

*A preliminary demonstration of speech recognition for a hearing disorder. by Yasuhiro KAKIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University), Nobuyuki MITANI, Kiyohiro OMORI (Hyogo Institute of Assistive Technology)

特徴量 (30 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 (30 次元) をベースラインとし，提案手法との比較を行う．また，事前の実験として健常者男性 4 名 (各 2620 単語) の発話を学習した健常者の音響モデルを用いて評価を行った．その結果を Table 1 に示す．

Table 1　Word recognition accuracy using HMMs for physically unimpaired persons.

Test data	Word recognition rate [%]
Hearing impaired	3.82
Physically unimpaired	79.1

3.2　ネットワークのサイズ

本稿では，Fig. 1 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 層とする) 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ルをフレーム幅 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 2 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 2　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

3.3　ネットワークの学習方法

各学習データについて，メル周波数スペクトル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．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 [7] で初期化した．

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

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．

3.4　評価結果

Fig. 2 に (i)MFCC+ΔMFCC(30 次元) を用いた場合と，(ii) ボトルネック特徴 (30 次元) を用いた場合の評価結果を示す．特定話者モデルを用いた音声認識において，学習単語数が 1310 単語の場合に約 1.9%，2620 単語の場合に約 1.4% の認識精度の改善が見られた．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，聴覚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．

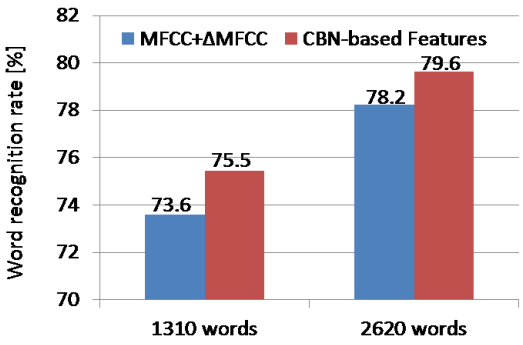

Fig. 2　Word recognition accuracy using HMMs for a hearing impaired person

4　おわりに

聴覚障害者の発話スタイルは健常者と大きく異なり，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低下 (79.1%→3.8%) する．本論文では特定話者モデルを用いて，発話に不自由のある聴覚障害者の発話音声認識の検討を行った．ベースラインと比べて，ボトルネックの構成を持つ CNN (CBN) を用いた特徴量抽出を行った場合，認識性能の改善が見られた．今後は，健常者と発話スタイルの異なる聴覚障害者と健常者間のコミュニケーションの改善のため，音響特徴以外の唇の動きや，手の動き，表情の変化などを捉える画像特徴量を音声認識に取り入れたい．

参考文献

[1] 舩坂宗太郎,“聴覚障害と聴覚補償,” コロナ社, 2007.
[2] 吉岡利也 他, “Convolutive Bottleneck Network 特徴量を用いた構音障害者の音声認識,” 日本音響学会 2014 年春季研究発表会, 3-Q5-20, pp.237-240, 2014-03.
[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] C. Plahl *et al.*,“Hierarchical bottle neck features for LVSCR,” in Interspeech, pp. 1197-1200, 2010.
[7] 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.