構音障害者音声認識のための確率表現に基づく音素ラベリングの検討*

☆高島悠樹 (神戸大), 中鹿亘 (電通大), 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，筋肉の不随意運動が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．音声認識技術を用いることで，発話内容を聞き取ることが困難な健常者とのコミュニケーションが円滑になり，障害者の就業機会の増加や講演時の補助等への活用などが期待される．

構音障害者の発話スタイルは，筋肉の付随意運動により健常者と大きく異なるため，従来の不特定話者音響モデルは役に立たず，認識精度が著しく低下する．そのため，構音障害者特有のモデルを用意する必要がある．また，発話内容が同一であっても，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．従来研究として，CNN (convolutional neural network [1]) を用いた発話変動にロバストな音声特徴量抽出法 [2] が提案されてきた．この手法はネットワークの学習に誤差逆伝播法を用いており，教師信号として HMM (hidden Markov model) による強制アライメントの結果を用いている．しかし，構音障害音声のスペクトルは変動が大きいため，精度の良いアライメントをとることができない．そのため，ネットワークの学習に用いる教師信号は誤りを含むことになり，より有効な特徴量抽出を阻害していると考えられる．さらなる構音障害者音声認識精度向上のために，より精度の高いアライメント情報を得る必要がある．

Fig. 1 は，健常者発話/ikioi/のスペクトル，Fig. 2 は，障害者発話/ikioi/の強制アライメント結果と手動アライメント結果の比較を示す．構音障害者はアテトーゼと呼ばれる筋肉の不随意運動を伴うため，常に意図した構音ができるとは限らない．つまり，同一話者の同一発話であっても，そのばらつきは健常者と比べて大きくなるという傾向がある．Fig. 2 より，強制アライメントの精度が悪いことや，健常者発話に

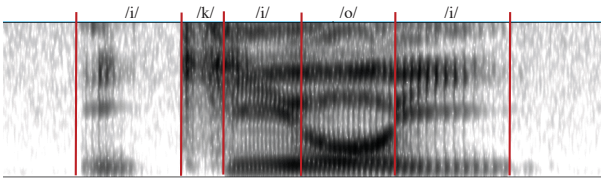

Fig. 1　Example of a spectrogram spoken by a physically unimpaired person /ikioi/

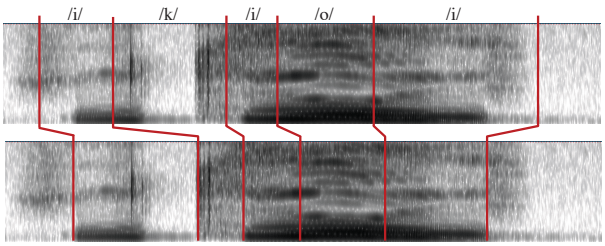

Fig. 2　Example of a spectrogram spoken by a person with an articulation disorder /ikioi/

比べて障害者発話はフォルマントがはっきりせず，特徴を掴みづらいことが確認できる．また，音素の境界も曖昧であり，Fig. 2 に示した手動アライメントも容易ではなく，誤りを含んでいる可能性が残っている．そのため，特徴量として MFCC を用いて強制アライメントを行なった場合，その結果は必ずしも期待したものになるとは限らない．

本研究では音素境界の曖昧性を考慮し，正規分布を用いた確率表現による音素のソフトラベリング法を提案する．viterbi アルゴリズムによる強制アライメントでは必ずハードラベリングになるが，隣接音素間には渡りの部分が存在し，その部分へのハードラベリングはネットワークの学習の教師信号として用いる場合には好ましくないことが考えられる．さらに，構音障害者の場合にはその隣接音素間の曖昧性は健常者と比べてより顕著であり，音素境界を考慮したラベリングが必要であると考えられる．ある発話において各音素区間の中心を平均とする正規分布を構成し，これを用いた存在確率により各時間における音素ラベルを与えることにより，音素境界のラベリングを曖昧性を考慮して行うことが可能になる．

*Phone Labeling based on the Probabilistic Representation for Dysarthric Speech Recognition, by Yuki Takashima (Kobe University), Toru Nakashika (UEC), Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

2 CNN ベースの特徴量抽出

2.1 CBN

Convolutive bottleneck network (CBN [3]) は，Fig. 3 に示すように，入力層，畳み込み層，プーリング層，フル接続の MLP 層および出力層により構成される．C,S および M はそれぞれ畳み込み層，サブサンプリング層および MLP 層を表す．畳み込み及びプーリング操作により，構音障害音声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) に対処できると考えられる．MLP 層は Fig. 3 に示すように 3 層設けられており，中間層 (M2) がボトルネック特徴量を表す．各層のユニット数は第 4 章で説明する．ボトルネック層のユニット数は，隣接層に比べて極端に小さくなっており，各ユニットには集約された情報が表現されると期待できる．本稿では，メル周波数スペクトラムを数フレーム統合したメルマップを CBN への入力とし，抽出されたボトルネック特徴量を音声認識に用いる．また，プーリングには広く用いられている max pooling を用いる [4].

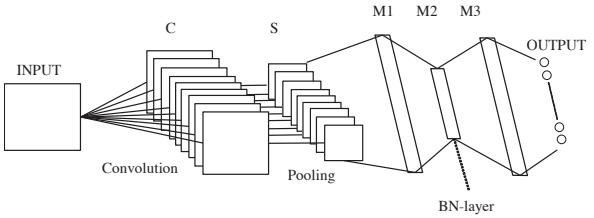

Fig. 3 Convolutive Bottleneck Networks (CBN)

2.2 ボトルネック特徴量抽出

まず，学習データの各音声信号は短時間フーリエ変換 (STFT) を施しメルフィルタバンクを掛けることにより，メル周波数スペクトルに変換される．そして，得られたスペクトルの任意の時刻における前後数フレームのスペクトルから 2 次元特徴を得る．この 2 次元特徴をメルマップと呼び，各メルマップをネットワークの入力層として与える．出力層の各ユニットには，入力層のメルマップ (数フレームのうち中央のフレーム) に対応する音素ラベル (バイナリベクトル) を割り当てる．この音素ラベルとして強制アライメントによるものが考えられるが，前章で述べたように信頼性が低いため，本研究では正規分布による確率表現を用いて音素ラベルを用意する．これらの入出力関係に基づき，CBN はランダムな初期値から学習を開始し，誤差逆伝搬法により各結合重みを更新する．

次に学習した CBN を用いて，評価データからボトルネック特徴量を抽出する．具体的には，学習データと同様に評価データからメルマップを計算し，学習した CBN の入力として与える．その後，畳み込みカーネルとプーリングによって入力データの局所特徴を抽出し，ターゲットとなる音素候補へと非線形に変換される．本稿では，このボトルネック層のユニットで表現される情報 (ボトルネック特徴量) を評価データの音響特徴量とし，従来の GMM/HMM により認識を行う．

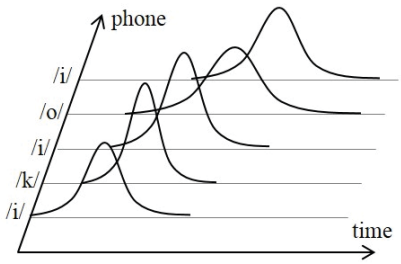

Fig. 4 Illustration of Gaussian labeling /ikioi/

3 確率表現に基づく音素ラベリング

3.1 正規分布による音素ラベリング

K 個の音素を含む発話において，時刻 t の音素を表す確率変数を $X_t \in \{1, 2, \cdots, K\}$ とする．例えば $X_t = k$ は，時刻 t における音素ラベルは発話中の k 番目の音素であることを表す．本稿では，確率 $p(X_t = k)$ を以下の式で定義する．

$$p(X_t = k) = \frac{\mathcal{N}(\mu_k, \sigma_k)}{\sum_{k'=1}^{K} \mathcal{N}(\mu_{k'}, \sigma_{k'})} \tag{1}$$

ただし，$\mathcal{N}(\mu, \sigma)$ は平均 μ，分散 σ^2 の正規分布である．すなわち，k 番目の音素は時刻 μ_k において最も高い確率となり，時刻 μ_k から遠ざかるにつれて減衰する．本稿では，μ_k，σ_k をそれぞれ以下のように定義する．

$$\mu_k = \frac{b_{k-1} + b_k}{2} \tag{2}$$

$$\sigma_k = \alpha(|\mu_k - b_{k-1}| + |\mu_k - b_k|) \tag{3}$$

ここで，b_k は k 番目の音素と $k+1$ 番目の音素の境界時間であり，α は分散を制御する非負のハイパーパラメータを表す ($b_0 = 0$，b_K は終了時刻とする)．

Fig. 4 に提案手法の概要を示す．音素ラベリングを行う場合，まず各発話に対して各音素の境界時間 b_k を与える必要がある．ただし，正確な境界を与える必要はなく，おおよその境界を与えるだけで良い．このことは，音素境界が曖昧な構音障害音声に対して非常に有利な点であり，手動アライメントによる負担を軽減することができる．次に，与えられた音素境界から，式 (2),(3) より平均及び分散を計算する．最後に，式 (1) から全フレームに対して存在確率を計算し，得

られた確率値を音素ラベルとして割り当てる．正規分布の特性上，平均 (音素の中心) 付近ではその音素である確率は高く，遠くなればなるほど確率が低くなるため，音声における音素の定常状態と渡りの状態を模したラベリングが行なえると期待できる．

3.2 ドロップアウトの適用

構音障害者は単語を健常者と同じ音素系列で発話することができない場合がある．例えば，/magane/という単語は，音素表記すると”/m/ /a/ /g/ /a/ /n/ /e/” となるが，構音障害者はしばしば，”/m/ /e/ /g/ /a/ /n/ /e/” と構音してしまい，この差異を含んだままネットワークを学習してしまう．本研究では，この問題に対するアプローチとして，CBN の学習時にドロップアウト [5] の適用を行う．ドロップアウトとは，学習時に入力・中間層の一部のユニットを入力データごとにランダムに使わないようにして曖昧な学習を行なうことで，ネットワークに汎用性を持たせる技術である．本研究では，ドロップアウトを出力層の各ユニットに対して適用することで，教師信号に対して過学習してしまうことを防ぐ．具体的には，ある入力データ $\mathbf{x}$ が与えられたとき，その入力データに対する出力信号 $f(\mathbf{x})$ に対して以下のようなドロップアウトマスクをかける．

$$f'(\mathbf{x}) = f(\mathbf{x}) \circ \mathbf{m} \tag{4}$$

ただし，$\circ$ は要素同士の積を表す演算子である．ここで，$\mathbf{m}$ は $f(\mathbf{x})$ と同じ次元のベクトルで，各ユニットは確率 c で 1(確率 $1-c$ で 0) の値をとるバイナリベクトルである．これより，教師信号との距離が縮まりにくくなり，ネットワークの曖昧な学習が実現される．本研究では，$c = 0.5$ として評価を行う．

4 評価実験

4.1 不特定話者音響モデルによる認識実験

構音障害者の発話スタイルは健常者とは大きく異なるため，広く用いられている不特定話者音響モデルを用いた音声認識は困難である．

実験用データとして，構音障害者男性 1 名と ATR 研究用日本語データベース (A set) [6] から選択した男性 1 名の発話する ATR 音素バランス単語 (216 単語) を用いた．音響モデル及び認識実験にはオープンソース音声認識ソフトウェア Julius[1]を使用した．Fig. 5 に不特定話者音響モデルを用いた音声認識実験結果を示す．不特定話者音響モデルは構音障害者音声認識に対しては有効でないことが確認でき，特定話者音響モデルの必要性があると考えられる．

[1]http://julius.sourceforge.jp

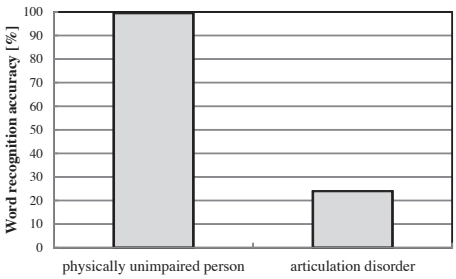

Fig. 5 Word recognition accuracy for each speaker using the speaker-independent acoustic model

4.2 単語認識実験結果

4.2.1 実験条件

評価対象として構音障害を患う男性 1 名 (話者 A), 女性 1 名 (話者 B) が発話する ATR 音素バランス単語を用いた．話者 A に関して，各単語は連続で 5 回発話されており，実験では各発話を切り出した合計 1,080 単語を使用する．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した．話者 B に関して，単語は連続で 3 回発話されており，実験では各発話を切り出した合計 600 単語を使用する．各単語の第 1 発話 (200 単語) を評価データ，残りの第 2〜3 発話 (400 単語) を CBN および音響モデルの学習データとした．

本章で用いる音響モデルは，54 音素の monophone-HMM で，各 HMM の状態数は 3，状態あたりの混合分布数は 8 である．

本実験では CBN の入力層に 2 次元特徴であるメルマップを与える．メルマップは，学習データの各音声データから 39 次元のメル周波数スペクトラムを計算し，任意の時刻において前後 6 フレームのスペクトルを統合したものを使用する．また，出力層には入力に対応する音素ラベルベクトル (54 次元) を与えた．音素ラベルとして，強制アライメントによるハードラベル (“forced”)，手動でおおよその境界を与えたハードラベル (“manual”)，手動ラベル対して提案法適用して得られたソフトラベル (“gaussian”) との間で，音声認識実験結果の比較を行なった．

実験に用いた CBN のアーキテクチャは Fig. 3 に示すとおりで，特徴マップ数，カーネルサイズ及びプーリングサイズはそれぞれ，13，7×11 及び 3×3 とし，MLP 層のユニット数は M1 から順に 108, 30, 108 とした．

4.2.2 実験結果と考察 (話者 A)

まず，分散のハイパーパラメータ α について検討を行なった．α を 0.1 から 0.5 まで変化させ実験を行い，最も良い精度が得られた $\alpha = 0.4$ のものを提案

手法として用いる．

Fig. 6 に音声認識実験の結果を示す．各ラベリング手法について，ドロップアウトを適用した場合としない場合で比較した．CBN を用いた場合，従来の MFCC よりも良い精度が得られることが分かる．提案手法により得られた音素ラベルを用いることで，強制アライメントを用いた場合に比べて精度が大きく向上することが分かる．さらに，教師信号にドロップアウトを適用することで，より精度を向上させることができた．

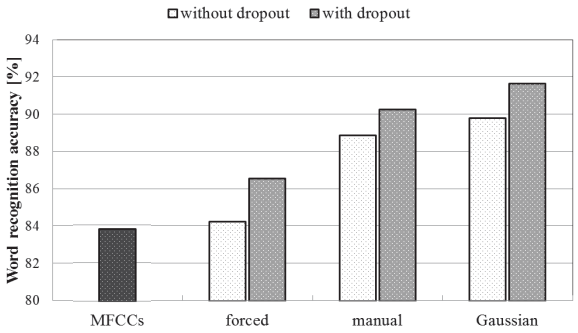

Fig. 6　Word recognition accuracy of each phone labeling method

4.2.3　実験結果と考察 (話者 B)

話者 A と同様にまず，分散のハイパーパラメータ α に関する検討を行なった．α を 0.1 から 0.6 まで変化させ実験を行なった結果，$\alpha = 0.5$ の時に最も良い精度が得られたが，話者 A と条件を揃えるために，次いで精度の良かった $\alpha = 0.4$ のものを提案手法として用いる．

Fig. 7 に話者 B の単語認識実験結果を示す．Fig. 7 に示すように，話者 A とは異なる傾向を示していることが分かる．1 つの原因として，データ量の少なさが考えられる．構音障害者の音声データは限られているため，CBN や HMM のパラメータを適切に学習できなかったと考えられる．さらに，手動のアライメントを用いた場合に，強制アライメントに比べ精度が悪化してしまっている．これは，強制アライメントが手動アライメントに比べて精度良くラベリングが行なえていたためだと考えられる．

提案手法にドロップアウトを適用しても，精度が向上しなかった．これは，話者 B の場合，少ないデータ量に対して、全体における適切なラベルの占める割合が大きくなったため，ドロップアウトを適用することで適切なラベルデータを多く間引いてしまったことが原因だと考えられる．

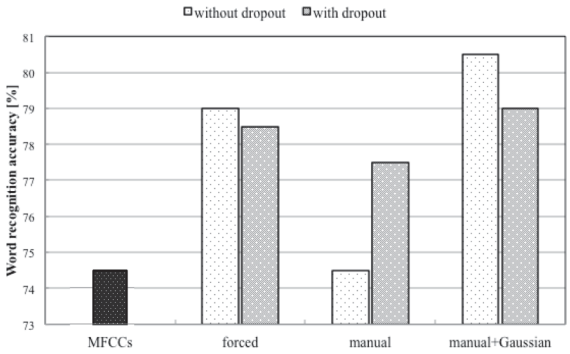

Fig. 7　Word recognition accuracy for utterances of “speaker B” using each phoneme labeling method and conventional MFCC features

5　おわりに

本稿では，正規分布を用いた音素のラベリング手法を提案し，CBN の教師信号として用いて単語認識実験により，その有効性を示した．ソフトな音素ラベルを用いることで，構音障害者の曖昧な音素境界に対処でき，音声認識精度を向上させることができた．本研究では人手により与えた音素境界に対して提案法を適用したが，手作業による負担軽減のため，半教師学習の導入などを検討する．

参考文献

[1] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” in *Intelligent Signal Processing*. 2001, pp. 306–351, IEEE Press.

[2] T. Nakashika *et al.*, “Dysarthric speech recognition using a convolutive bottleneck network,” in *ICSP*, 2014, pp. 505–509.

[3] K. V. et al., “Convolutive bottleneck network features for LVCSR,” in *ASRU*, 2011, pp. 42–47.

[4] O. Abdel-Hamid *et al.*, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition.,” in *ICASSP*, 2012, pp. 4277–4280.

[5] G. E. Hinton *et al.*, “Improving neural networks by preventing co-adaptation of feature detectors,” 2012.

[6] A. Kurematsu *et al.*, “ATR japanese speech database as a tool of speech recognition and synthesis.,” Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.