一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2015-44,SP2015-13,WIT2015-13(2015-06)

構音障害者音声認識のための混合正規分布に基づく音素ラベリングの検討

高島　悠樹† 　中鹿　亘†† 　滝口　哲也††† 　有木　康雄†††

† 神戸大学大学院システム情報学研究科 〒657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 電気通信大学大学院情報システム学研究科 〒182-8585 東京都調布市調布ヶ丘 1-5-1
††† 神戸大学自然科学系先端融合研究環 〒657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †y.takashima@me.cs.scitec.kobe-u.ac.jp, ††nakashika@uec.ac.jp, †††{takigu,ariki}@kobe-u.ac.jp

あらまし　本報告では，アテトーゼ型脳性麻痺による構音障害者の音声認識の検討を行う．従来研究として，畳み込みニューラルネットワークを用いた特徴量抽出法が提案され，その有効性が示されてきた．ニューラルネットワークの学習には教師信号として HMM による強制アライメントの結果が用いられているが，構音障害者の音声は発話毎にスペクトルの変動が大きいため，正確なアライメントを取ることは極めて難しい．誤った教師信号ではネットワークは十分に学習できないと考えられるため，上述の手法はその性能を十分に発揮できていないと考えられる．しかし，構音障害音声の音素境界は，健常者に比べて非常に曖昧であり，正確な音素境界を与えることは根本的に難しい．本研究では，音素ラベルを正規分布を用いた確率表現で与えることにより，構音障害音声特有の音素境界の曖昧性を考慮したラベリングを新たに提案する．本稿では，提案手法より得られる音素ラベルを教師信号としてネットワークを学習し，そのネットワークから抽出された特徴量を用いた単語認識実験の結果を報告する．

キーワード　構音障害, 特徴量抽出，畳み込みニューラルネットワーク, ボトルネック特徴量，音素ラベリング

Phone Labeling Based on Gaussian Mixture Model for Dysarthric Speech Recognition

Yuki TAKASHIMA†, Toru NAKASHIKA††, Tetsuya TAKIGUCHI†††, and Yasuo ARIKI†††

† Graduate School of System Informatics, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
†† Graduate School of Information Systems, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan
††† Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: †y.takashima@me.cs.scitec.kobe-u.ac.jp, ††nakashika@uec.ac.jp, †††{takigu,ariki}@kobe-u.ac.jp

Abstract　We investigate in this paper speech recognition for a person with an articulation disorder resulting from athetoid cerebral palsy. As our previous work, the feature extraction method using a convolutional neural network is proposed, and showed its effectiveness. The neural network needs the teaching signal to train the network using back-propagation, and the previous method uses forced alignment using HMMs from speech data for the teaching signal. However, because the dysarthric speech fluctuates every utterance, it is difficult to obtain the correct alignment. It is considered that the network is not adequately trained due to the wrong alignment. However, phone boundaries for dysarthric speech are ambiguous, and it is difficult to give the correct alignment and it is difficult to give the correct alignment. Therefore, we propose a phone labeling method using the Gaussian distribution. In this paper, we report our experimental results of speech recognition using the networks trained by the phone alignments calculated by our proposed method.

Key words　Articulation disorders, feature extraction, convolutional neural network, bottleneck feature, phoneme labeling

1. はじめに

近年，音声認識技術は広く普及し，人々の生活の助けとなっている．スマートフォンを例に挙げると，端末に対して発話を行うことで通話やメールを行うことができる．さらに，子供や高齢者などの発話スタイルが成人と異なる人物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，使用される機会が増加している [1]〜[3]．しかし，これらは言語障害などのない人々を対象としており，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．言語障害には様々な種類の症状があるが，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 [4], [5] 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．そのような音声認識が実現すれば，発話が困難で手足が不自由な人でも環境制御装置を用いることで，テレビやパソコンといった周辺機器の操作を自らの力で行うことが可能になる．さらに，音声認識を用いることで，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 [6]〜[8]) を用いた発話変動にロバストな音声特徴量抽出法 [9] が提案されてきた．CNN 特有の畳み込み操作とプーリング操作により，構音障害者の発話変動によるスペクトルの微小な変化に対して頑健な特徴量抽出を行うことができる．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．

図 1 は，健常者発話/ikioi/のスペクトル，図 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

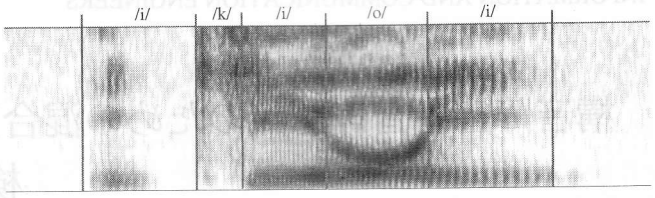

図 1　健常者発話/ikioi/のスペクトル

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

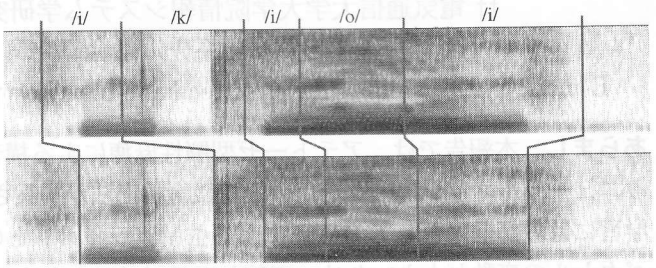

図 2　構音障害者発話/ikioi/のスペクトル

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

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．図 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，図 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．文献 [9] において，教師信号として用いられている，入力に対応する音素ラベルは 2 値のハードラベルとなっている．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．本研究で提案する正規分布を用いた音素ラベリングは，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で構成される混合正規分布 (Gaussian mixture model ; GMM) で表現する．各時間における音素ラベルは，GMM の事後確率により与えられる．これにより，音素境界のラベリングに対して曖昧性を考慮して行うことが可能になる．

以下，第 2 章で CNN を用いた特徴量抽出手法を述べ，第 3 章で本稿の提案手法を説明する．第 4 章でこれまでの HMM による強制アライメントを用いた手法と比較し，第 5 章で本稿をまとめる．

2. CNN ベースの特徴量抽出

近年，音声認識の分野において，音声信号から時間-周波数の 2 次元特徴を抽出し，画像処理の分野で広く用いられてきた CNN を用いて音声認識を行うアプローチが研究されている [8], [10], [11]．本研究では，構音障害音声の特徴量抽出に，CNN と multi layer perceptron (MLP) が階層的に接続されたネットワークを使用する．MLP 層は中間層のユニット数が隣接層に比べて小さくなっており，このような構造を持つネットワークを convolutive bottleneck network (CBN [12]) という．

2.1 CNN

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，細かい位置ずれに対する不変性を実現するプーリング層を交互に繰り返す構造をとる．

$(k-1)$ 層の特徴マップ $\{\mathbf{h}_1^{k-1}, \cdots, \mathbf{h}_i^{k-1}, \cdots, \mathbf{h}_I^{k-1}\}$ が与えられたとき，k 層における畳み込み演算による j 番目の特徴マップ $\mathbf{h}_j^k \in \mathbb{R}^{N_n^{h^k} \times N_m^{h^k}}$ は以下の式で計算される．

$$\mathbf{h}_j^k = f\Big(\sum_i^I \mathbf{w}_{j,i}^k * \mathbf{h}_i^{k-1} + b_j^k \mathbf{E}\Big) \qquad (1)$$

ここで，$\mathbf{w}_{j,i}^k \in \mathbb{R}^{N_n^w \times N_m^w}$ と b_j^k はそれぞれ，$k-1$ 層の i 番目の特徴マップから k 層の j 番目の特徴マップへの畳み込みフィルタと k 層の j 番目の特徴マップのバイアスを表す ($N_n^{h^k} \equiv N_n^{h^{k-1}} - N_n^w + 1$, m も同様)．記号 $*$，$\mathbf{E}$ はそれぞれ，畳み込み演算と各要素が 1 の配列と定義する．$f(\cdot)$ は活性化関数を表し，本稿ではシグモイド関数を用いる．畳み込み層の各ユニットは前層の特徴マップ内の $N_n^w \times N_m^w$ の大きさの領域と局所的に接続しており，前層から与えられた入力の局所的な特徴を捕えることができる．

プーリング層では，前層 (畳み込み層) の特徴マップ内の小領域の応答をまとめ，1 つのユニットで表現する．この小領域は重複させないため，プーリング層における特徴マップのサイズは前層に比べて小さくなる．このことは入力データの低解像度化に対応し，細かい位置ずれに対する不変性を獲得することができる．

2.2 CBN

CBN は，図 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 層は図 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．また，入力特徴量を小さい次元数で表現することから，MFCC，linear discriminant analysis (LDA) や PCA と同様の次元圧縮処理の振る舞いをすると考えられる．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

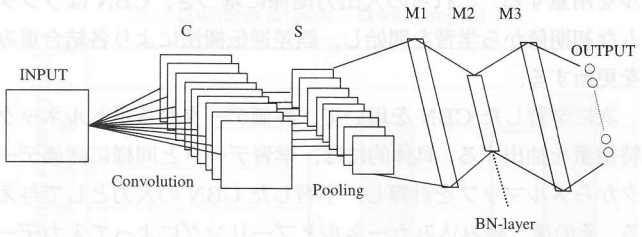

図 3　畳み込みボトルネックネットワークの構造
Fig. 3　Convolutive Bottleneck Networks (CBN)

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 [10]．

2.3 ボトルネック特徴量抽出

図 4 に特徴量抽出のフローを示す．学習データには，構音障害者 1 名が発話する 5 回連続発話データのうち，比較的安定している 2 回目以降の発話を用いる．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を割り当てる．具体的には，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 1，他のユニットは 0 の値を持つバイナリベクトルとなる．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を用いた強制アライメントによって求める．この強制アライメントは，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

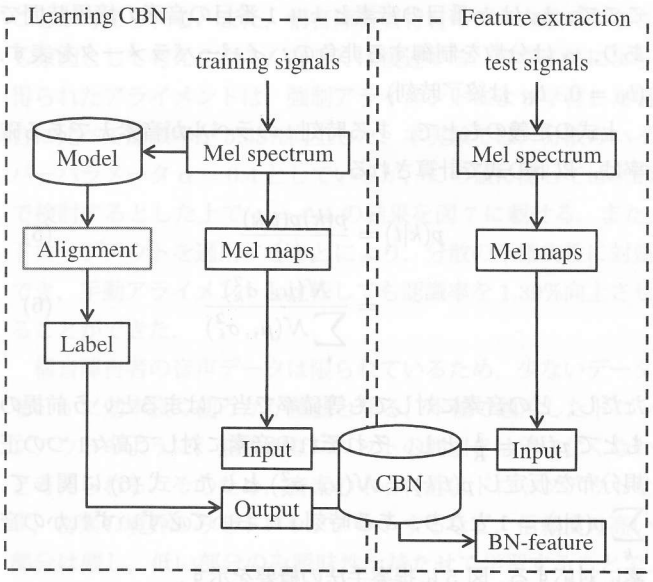

図 4　CBN を用いた特徴量抽出の流れ
Fig. 4　Flowchart of the bottleneck feature extraction using CBN

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．この操作は，入力特徴量に対する各音素の事後確率を計算していることに近い．このとき，ボトルネック層は，事後確率を計算するために必要な情報を低次元でよく表現していると考えられる．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．

3. 確率表現に基づく音素ラベリング

本節では，音素ラベリングを正規分布に基づく確率表現で与える方法を提案し，前章で述べた CBN の学習に用いる音素ラベルへ適用する手法について述べる．

3.1 混合正規分布による音素ラベリング

ある発話において，時刻 t における音素の存在確率を GMM を用いて以下のように定義する．

$$p(t) = \frac{1}{K}\sum_{k}\mathcal{N}(\mu_k, \sigma_k^2) \tag{2}$$

ただし，$\mathcal{N}(\mu, \sigma)$ は平均 μ，分散 σ^2 の正規分布である．ここで，K は発話に含まれる音素数，k はそのインデクスである．本稿では，μ_k，σ_k はそれぞれ k 番目の音素の中心とその分散を表し，以下のように定義する．

$$\mu_k = \frac{b_{k-1} + b_k}{2} \tag{3}$$

$$\sigma_k = \alpha(|\mu_k - b_{k-1}| + |\mu_k - b_k|) \tag{4}$$

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)

上式の定義のもとで，ある時刻 t のラベルが音素 k である確率は，以下の式で計算される．

$$p(k|t) = \frac{p(k)p(t|k)}{p(t)} \tag{5}$$

$$= \frac{\mathcal{N}(\mu_k, \sigma_k^2)}{\sum_i \mathcal{N}(\mu_i, \sigma_i^2)} \tag{6}$$

ただし，どの音素に対しても等確率で当てはまるという前提のもとで $p(k) = \frac{1}{K}$ とし，それぞれの音素に対して高々1 つの正規分布を仮定し $p(t|k) = \mathcal{N}(\mu_k, \sigma_k^2)$ とした．式 (6) に関して，$\sum_k p(k|t) = 1$ となり，ある時刻 t において必ずいずれかの音素に対応する．図 5 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) より GMM を構成する．最後に，式 (6) から全フレームに対して事後確率を計算し，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．

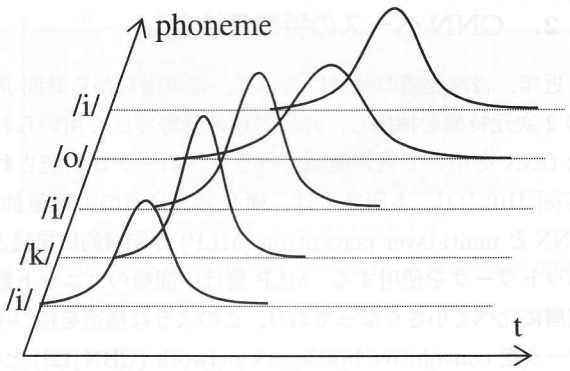

図 5　/ikioi/に対する提案ラベリング手法の概要
Fig. 5　Illustration of Gaussian labeling /ikioi/

3.2 ドロップアウトの適用

提案手法の問題点として，子音は母音と比較して短いなどの音素固有の継続長が考慮されていない点が挙げられる．それらの情報は，本来分散として考慮されるべきだが，音素継続長は発話内容で異なるものであり，また，構音障害音声における音素継続長は健常者のものと傾向が異なるため，使える情報がない．そのため，分散の信頼度は低く，依然として音素ラベルは誤りを含んでいる可能性があり，CBN の教師信号として正規分布によるラベルを用いた場合，その誤りを学習する可能性がある．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 [13] 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 50%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．

$$f'(\mathbf{x}) = f(\mathbf{x}) \circ \mathbf{m} \tag{7}$$

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 c%の確率で 1，$100 - c$%の確率で 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，$c = 50$ として評価を行う．

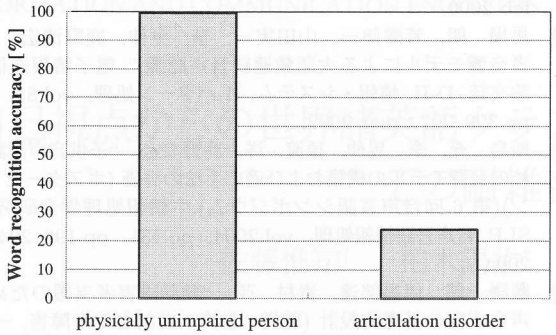

図 6　不特定話者音響モデルを用いた単語認識実験結果

Fig. 6　Word recognition accuracy for each speaker using the speaker-independent acoustic model

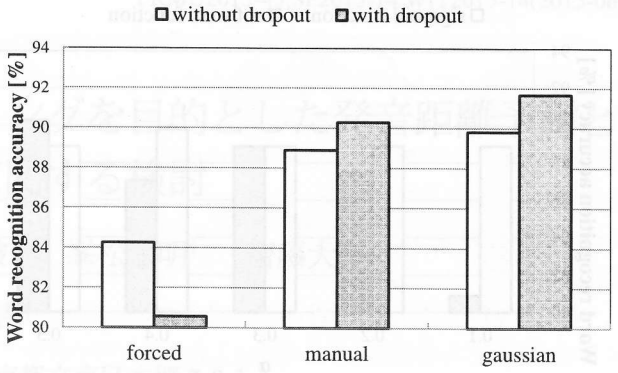

図 7　音素ラベリングの違いによる単語認識実験結果の比較

Fig. 7　Word recognition accuracy of each phone labeling method

4. 評価実験

本章ではまず，従来の音声認識で用いられている不特定話者音響モデルを用いた認識実験について述べ，その後，提案手法に関する実験について述べる．

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．そのため，構音障害者固有の特定話者モデルが必要となる．

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) [14] 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(注1) を使用した．図 6 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．

4.2 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，実験では各発話を切り出した合計 1,080 単語を使用する．本実験では，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した．

音声の標本化周波数は 16kHz であり，音響分析にはハミング窓を用いている．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 13 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")，手動ラベルから提案法を用い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図 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．CBN のカーネル及び各ユニット間の結合重みは一様乱数を用いて初期化され，各パラメータは誤差逆伝搬法を用いて更新される．

4.3 実験結果と考察

図 7 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．誤りを多く含んでいる強制アライメントは他 2 つの手法に比べて精度が低いことが分かる．また，強制アライメントを用いる場合，ドロップアウトを適用すると，精度が下がってしまっている．この理由として，誤りを多く含んだ強制アライメントに対して曖昧な学習をしたため，十分な学習が行なえなかったことが考えられる．また，構音障害者特有のデータ量の少なさも原因として考えられる．詳しくは後述する．提案手法により得られたアライメントは，強制アライメントにより学習した場合に比べて認識率が 5.55%向上した．本実験では，分散のハイパーパラメータ $\alpha = 0.4$ としているが，この値に関しては次節で検討するとした上で $\alpha = 0.4$ の結果を図 7 に載せる．また，ドロップアウトを適用することにより，分散の不確実性に対処でき，手動アライメントと比較しても認識率を 1.39%向上させることができた．

構音障害者の音声データは限られているため，少ないデータをいかに効率良く使うかが重要となる．本稿で適用したドロップアウトのように，無差別なユニットの間引きは望ましくないと考えられる．そのため，音素の中心付近はドロップアウトせず，音素の境界のみをドロップアウトするなど，信頼度の高い部分は残し，低い部分のみ曖昧性を持たせて学習することで，更なる精度の向上が期待できる．

(注1)：http://julius.sourceforge.jp

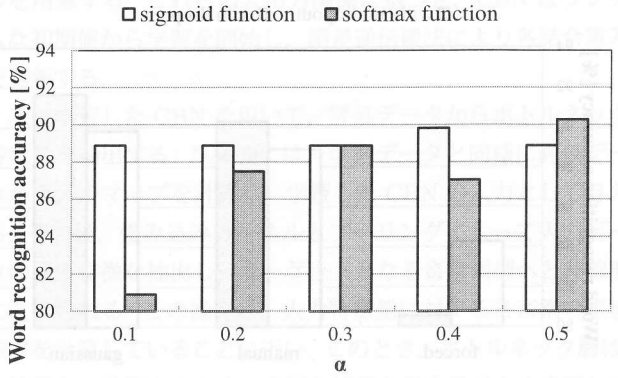

図 8　提案手法における α と識別関数による単語認識実験結果の比較
Fig. 8　Word recognition accuracy of the proposed method when changing the value of α and the activation function

4.4　識別関数の検討

前節の実験は，ネットワークの識別関数としてシグモイド関数を使用していた．しかしながら，提案手法の場合，教師信号が確率値のため出力信号も確率値であることが望ましいと考えられる．そこで，ネットワークの識別関数としてソフトマックス関数を使用した実験を行なった．また，式 (4) における α の値についても同時に検討した．図 8 に示すように，最適な α を用いた場合，識別関数をソフトマックス関数にした方が精度が向上した．教師信号を確率値として表現する場合，出力信号も確率値 (ユニットの値を総和すると 1 になる) である方が自然であると考えられる．なお，$\alpha = 0.5$ として，識別関数にソフトマックスを用いてドロップアウトを適用すると認識率が 87.50%となり精度が低くなった．この理由として，ドロップアウトを適用することで出力ベクトルの値の総和が 1 でなくなることが挙げられる．これにより，ソフトマックス関数の制約から外れることになり，性能が改善しなかったと考えられる．前節のシグモイド関数の場合，出力ベクトルに対する制約はないため，問題がなかったと考えられる．

5. おわりに

本研究では，音声に対する音素ラベルを正規分布を用いた確率表現により与える手法を提案した．構音障害者の音声は，健常者に比べて音素境界が極めて曖昧であるため，ハードラベルによる音素ラベリングは非常に難しい．孤立単語認識実験により，確率値によるソフトラベリングをすることにより音素境界の曖昧性を考慮でき，認識精度の向上を確認した．音素間の渡りの存在は健常者の場合も同様であるため，本手法は健常者に対しても有効であると考えられる．正確でなくともおおよその音素境界が与えられれば計算できることが本手法の強みであるが，同時に，人手で境界を与える作業は避けられない．今後は，おおよその音素境界を自動で推定できる手法について検討していきたい．

文　献

[1] 中川聖一，“音声認識研究の動向，” 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理, vol.83, no.2, pp.433–457, feb 2000.
[2] 馬場　朗，芳澤伸一，山田実一，李　晃伸，鹿野清宏，“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. D-II, 情報・システム, II-パターン処理, vol.85, no.3, pp.390–397, mar 2002.
[3] 鮫島　充，李　晃伸，猿渡　洋，鹿野清宏，“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 (ポスターセッション)(第 6 回音声言語シンポジウム)，” 情報処理学会研究報告. SLP, 音声言語情報処理, vol.2004, no.131, pp.199–204, dec 2004.
[4] 藪謙一郎，伊福部達，青村　茂，“発話障害者支援のための音声合成器の基礎的設計 (聴覚・音声・言語とその障害, 一般)，” 電子情報通信学会技術研究報告. SP, 音声, vol.105, no.686, pp.59–64, mar 2006.
[5] C. Veaux, J. Yamagishi, and S. King, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders.,” INTERSPEECH, pp.967–970, 2012.
[6] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” The Handbook of Brain Theory and Neural Networks, pp.255–258, MIT Press, Cambridge, MA, USA, 1998.
[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Intelligent Signal Processing, pp.306–351, IEEE Press, 2001.
[8] H. Lee, P.T. Pham, Y. Largman, and A.Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks.,” NIPS, pp.1096–1104, 2009.
[9] T. Nakashika, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, “Dysarthric speech recognition using a convolutive bottleneck network,” ICSP, pp.505–509, 2014.
[10] O. Abdel-Hamid, A. rahmanMohamed, H.J. 0001, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition.,” ICASSP, pp.4277–4280, 2012.
[11] T.N. Sainath, A. rahmanMohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for LVCSR.,” ICASSP, pp.8614–8618, 2013.
[12] K.V. etal., “Convolutive bottleneck network features for LVCSR,” ASRU, pp.42–47, 2011.
[13] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” 2012.
[14] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR japanese speech database as a tool of speech recognition and synthesis.,” Speech Communication, vol.9, no.4, pp.357–363, 1990.