発話制限された音声に対する音声認識

大槻　典行*

Speech Recognition of the Voice which is Restricted

Noriyuki OHTSUKI

Abstract - This report presents a speech recognition method of the voice which is restricted. In this report, the speech recognition of the voice which was restricted according to impediment was examined. This speech recognition method uses a multi-layerd neural network which has two different structures. The neural network consists of a self-organized input layer and the perceptron type of layers. The self-organized input layer extracts some characteristics of speech in spectrum domain and compresses them. This self-organized layer affects restricted voice effectively.
Key words : restricted voice, speech recognition, neural network, self-organization

１．はじめに

近年，音声認識装置は，特殊な装置を用いることなく実現可能で，容易に入手可能なパーソナルコンピュータ上で音声認識システムを構築することが可能である[1]．このような状況の中で，誰もが利用可能な音声認識システムの構築が期待されている．特に，特殊な状況にある場合の利用が有用視されている．本報告では，障害などにより音声の発話に制限を受けている音声の認識について検討した．

現在，主流となっている音声認識システムは，あらかじめ学習と呼ばれる利用者の音声を登録する作業を必要としないにも関わらず，誤認識が起きにくいシステムになっている．これは音声認識の中心となるパターンマッチング部にＨＭＭ法と呼ばれる確率モデルが用いられ，一般的な多くの音声学習データからそのモデルが構築されているためである．このモデルを用いたシステムは，利用者に合わせた学習も追加で可能であり，その学習は，モデル構築時よりも非常に少ない学習時間で終了する．しかし，あらかじめ用意されたモデルと全く特徴の異なるモデルが必要となった場合，その学習には，非常に多くの学習データと時間を必要とし，それらを用いない場合，認識率が極端に低下する．従って，あらかじめ用意されたモデルから外れた音声に対する高い音声認識率は期待できない．つまり，障害などにより，音声の発話に制限を受けている状態で，発声された音声は，制限されない音声に比べ音響モデルが異なっていると考えられるため，高い認識率を得ることは難しい．そこで，本報告では，従来より提案していた自己組織化入力層を持つニューラルネットワークを用いた音声認識手法[2]を利用することで比較的少ない学習で，発話制限された音声に対する音声認識が実現できる可能性について検討した．

*釧路高専 情報工学科

２．発話制限された音声

ここでは，音声認識の対象とする発話制限された音声について考察する．

音声の生成器官は，図１のようになっている．声帯で音声の音の源となるほぼ周期的で定常な音が作られる．これを音源と呼ぶ．この音は，声の高低の特徴は持つが，「何を話しているか」の特徴は持っていないと考えてよい．「何を話しているか」は，舌，顎，口，唇および鼻の運動によって作られる喉から唇までの形状が変化する管状の部分によって決められる．これを声道と呼ぶ．これらの一部でも運動が制限されると，発声する音声も制限されることになる．しかし，声道の形状を決める部分によってはほとんど発声する音声に影響を与えない部分もある．鼻などは特に影響の少ない部分であると考えられている．

本報告では，舌の一部の運動が制限されていると考えられる音声に対する認識を試みる．

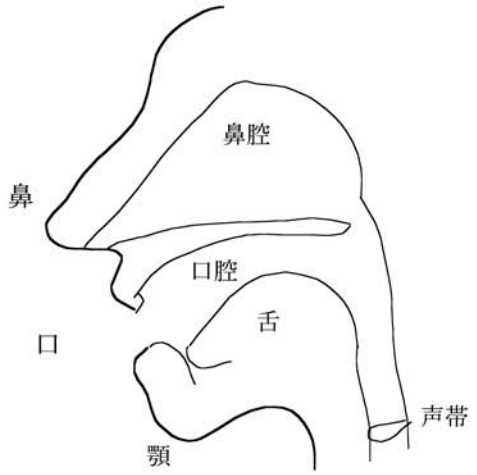

図１：音声の生成器官（声道・声帯）

３．音声認識アルゴリズム

ここでは，本報告で用いる音声認識アルゴリズムについて解説する．

３．１　ニューラルネットワークを用いた音声認識

音声認識の手順としては，音声の分析，特徴抽出，カテゴリーの分類，認識となる．この手順を図２に示す．音声認識では音声の特徴を抽出した後，この特徴を使って認識を行う．本音声認識手法は音声の分析に適応信号処理を用いて正確な音声の特徴を抽出している．認識はニューラルネットワークを用いている．ニューラルネットワークは入力に与えられたパターンを学習という操作によって，与えられた入力の特徴を自ら獲得し，各カテゴリに適切に分類する能力を持つ．この学習によって得られる分類能力は外部から知識として与える必要がないため，分類の対象となるものの特徴を利用者が知っている必要はない．これは明確なカテゴリを形成できない音声認識，特に発話制限された音声の特徴が不明確になるような場合，非常に有利と考えられる．

音声の特徴－何を話しているか－は声道の形状とその時間変化によって決まる．声道の形状は音声のスペクトルに対応し，声道の形状の時間変化は音声のスペクトルの時間変化に対応する．この二つの特徴を同時に捉えることが重要であるが，発話制限された音声では，スペクトルの状態およびその時間変化が乱れる可能性がある．また，階層構造を持つニューラルネットワークにとって，この二つの特徴を同時に捉えることは困難な問題である．そこで，本システムでは層構造を持つニューラルネットワークの入力部分に自己組織化層を設けることによりネットワーク全体で二つの特徴を同時に捉えることを可能にしている．自己組織化層はスペクトルの時間変化の瞬時値を捉える働きがある．則ち，発声された音声の１時刻毎の特徴の変化に反応する．自己組織化層に続く多層ネットワーク部分はスペクトルの時間変化を特徴として捉え認識出力を得る．

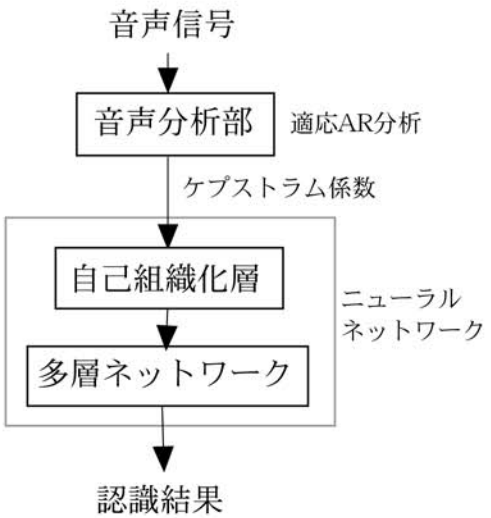

図２：音声認識システム

３．２　自己組織化入力層

ここではこの音声認識手法の特徴となっている自己組織化入力層の動作を簡単に説明する．この自己組織化層にはベクトルが入力として与えられる．この入力ベクトルと自己組織化層を構成するユニットが持つ特徴ベクトルとの距離が求められる．この距離は総てのユニットに対して求められ，その中から最も距離が近いユニットが選び出される．この選び出されたユニットと入力ベクトルとの距離の大きさによって，新たにユニットを生成したり，ユニットが持つ特徴ベクトルを更新する．この動作によって入力ベクトルが持つ特徴を自己組織化層の中に特徴として取り込む．このようにして自己組織化入力層は与えられる入力に対して似た特徴を持つユニット同士が集まり，これらがカテゴリを形成する．このとき入力以外に外部から与えられる情報はない．つまり，この入力層自身に与えられる入力のみの情報から，その特徴を分類し複数のカテゴリを形成するのである．

３．３　多層ニューラルネットワーク

自己組織化入力層に続くニューラルネットワークはパーセプトロン型のニューラルネットワークで学習には教師付きのバックプロパゲーショントレーニングアルゴリズムを用いた多層ネットワーク[3]である．

自己組織化入力層で音声のスペクトルをベクトル量子化していると考えて良いので，ベクトル量子化されたスペクトルの時間変化を特徴として，多層ニューラルネットワークで分類することになる．ここで行なう学習は教師付きとなるので，与えられた入力の音声に対し出力層でその音素に対する正解の出力を与えることになる．

４．実　験

発話制限された音声が，音声認識に与える影響を検討するため，スペクトルの状態を制限されない音声と比較した．また，発話制限された音声とされない音声の認識実験は，母音認識に関して行った．

４．１　スペクトルの比較

発話制限された音声と発話制限されない音声の母音部分のスペクトルの時間変化を比較した．音声の母音部分は，短時間では定常な状態と考えて良いのでスペクトルも安定しておりその特徴も把握しやすい．

用いた音声は，8kHz，16bitサンプリングしたものである．スペクトル分析の分析窓幅は，32msec(256ポイント)，フレームシフトは2.5msec(20ポイント)とした．分析は適応AR分析を行い，14次の適応ARラチスパラメータを計算した後，このパラメータからケプストラム係数を求めスペクトル包絡を表

示した．母音/u/の部分のスペクトルの時間変化を図３に示す．発話制限されない音声のスペクトルの時間変化は図３．(a)である．発話制限された音声のスペクトルの時間変化は図３．(b)に示した．発話制限されない音声のスペクトルは，フォルマントがはっきりと確認でき，時間的な変化も少なく安定したスペクトルである．一方，発話制限された音声のスペクトルは，フォルマントが(a)程，はっきりせず，時間的にも不定期に変化している．また，図４に発話制限された音声で比較的はっきりと聴こえる母音音声/a/のスペクトルの時間変化を示した．発話制限された音声は，母音として明確に聞こえるにも拘らず特徴となるフォルマントが特定しにくい状態になっている（図4(b)）．音声の母音の特徴は，フォルマントにあり，この周波数を人間が聞き分けていると考えられている．音声認識においても，母音の認識に関しては，フォルマント周波数が重要であると考えられている．フォルマントが明確ではなく，安定していない場合，音声認識率の低下は避けられないと考えられる．また，発話制限された音声では，安定している筈の母音のスペクトルの乱れを考えると，子音のスペクトルの時間変化は，より特徴を捉えずらい状態になっていることが予想される．

４．２　認識実験

発話制限された音声のスペクトルが安定したものではなく，また，特徴も明確なものではないことが明らかになった．しかし，そのスペクトルに全く特徴がない訳ではないのであれば，その特徴を捉え，認識に利用することが可能である．

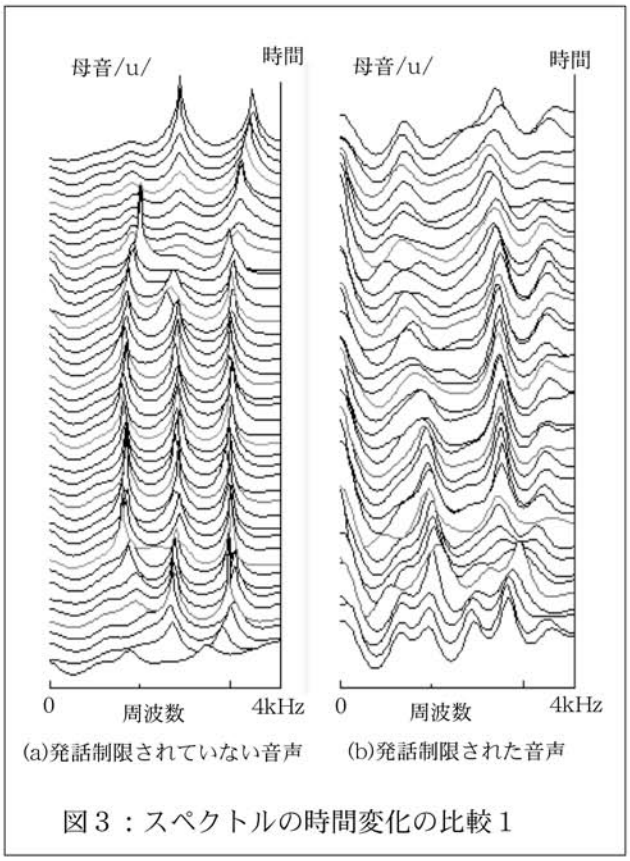

図３：スペクトルの時間変化の比較１

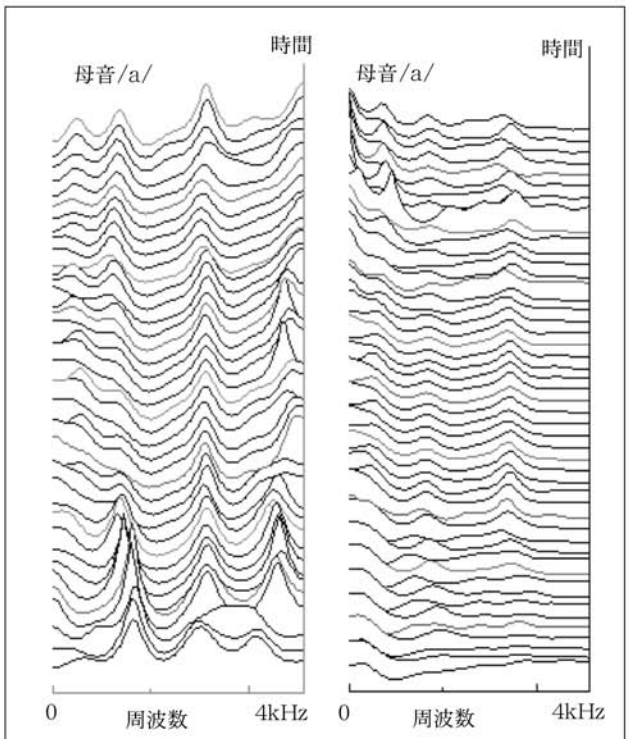

図４：スペクトルの時間変化の比較２

本報告で用いる音声認識手法では，音声の特徴を自ら捉える手法であるため，発話制限された音声スペクトルからでもその特徴を捉えることが期待できる．

音声認識実験は母音認識について行った．発話制限されていない人と発話制限された人，それぞれ一名の音声データから母音部を切り出し，認識実験に用いた．この実験では，認識部の自己組織化層のユニットの生成状態を比較した．音声分析の条件は，音声は，8kHz，16bitサンプリング，分析の分析窓幅は，32msec，フレームシフトは5msecである．認識に用いた特徴パラメータは，適応AR分析で得た14次の適応ARラチス系数から14次のケプストラム係数を計算し，認識部へ与えた．認識部では，各母音５パターンを学習データとして各１回ずつ自己組織化層へ与え自己組織化層の学習を行った．多層ネットワーク部は各学習データに対して1000回の学習をバックプロパゲーション法[3]を用いて行った．この結果，自己組織化層の生成ユニット数を比較すると，発話制限されていない音声は各母音に対して2～3個のユニットが生成されるのに対して，発話制限された音声では，ほぼフレームシフト毎にユニットが生成される状態であった．つまり，各母音で20～30程度のユニットが生成された．発話制限されない音声に対してほぼ10倍のユニットが生成されたことになる．ただし，この自己組織化層のユニット生

表１：音声認識結果

	発話制限されない音声		発話制限された音声（自己組織化層のユニット生成制限）		発話制限された音声（自己組織化層のユニット生成制限なし）	
	出/入	認識率	出/入	認識率	出/入	認識率
a	77/77	100%	25/30	83.3%	29/30	96.7%
i	100/100	100%	20/30	66.7%	29/30	96.7%
u	94/96	98%	18/30	60%	25/30	83.3%
e	53/53	100%	25/30	83.3%	28/30	93.3%
o	71/71	100%	28/30	93.3%	30/30	100%
合計	395/397	99.5%	116/150	77.3%	141/150	94%

成には制限を設けて一定の数以上に増大しないように設定している．認識実験は，学習に用いない母音音声データを用いた．認識結果を表１に示す．この表からも解るように発話制限されない音声は，母音に関してほぼ100%の認識率を得ている．一方，発話制限された音声に対しては，認識率が低下している．これは，母音においてもスペクトルが安定しない部分で学習パターンとの不一致が多く生じてしまうためと考えられる．なお，自己組織化層のユニットの生成数を制限しない状態では，認識率の向上がみられる．

実験の結果より，母音認識に関して，発話制限された音声は，スペクトルが安定しない状態が認識率への影響を与えていると考えられる．本報告で用いた認識手法は，この安定しないスペクトルの変化に自己組織化層のユニット生成が追従すると認識率の向上が得られる．

５．まとめ

本報告では，発話制限された音声を認識する手法について検討した．まず，音声のスペクトルおよびその時間変化の比較を行った．発話制限された音声のスペクトルは，発話制限されない音声のスペクトルに比べフォルマントが明確になっていない，定常な状態が見られないなど，スペクトル構造が異なるものであった．これは，想定する音響モデルが発話制限されない音声の場合と大きく異なるものとなり，HMM法などを用いた音声認識では，モデルの構築，つまり学習を始めから行う必要が有ることを示している．本報告で用いた自己組織化入力層を持つ音声認識手法は，学習時に音声の特徴を自ら取得する手法であり，特定の音響モデルを想定していない．音声認識実験では，自己組織化層で生成されるユニットの数が発話制限されていない音声に比べ増大した．これは，発話制限された音声のスペクトル構造が，前述した通り「フォルマントが明確ではない・定常なスペクトルではない」ため，特徴を捉える自己組織化層のユニットが異なるスペクトルに応じて逐次生成された為である．このように，安定したスペクトル構造を持たない音声に対して，本手法は自己組織化層のユニットの生成で追従しようとする．従って，認識率が高すると，自己組織化層のユニットの生成が膨大になり，全音素の認識を目指すと容量およびその処理に現実性を見いだせない状態である．今後，これらのユニット数増大に対する制限条件を検討する必要がある．

本研究は，平成14・15年度科学研究費補助金(萌芽研究14655144)による研究の一部として行われた．

参考文献

[1]鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，“音声認識システム”，オーム社

[2]N.Ohtsuki,Y.Miyanaga,K.Tochinai,"Continuous Speech Recognition using A New Neural network with Two Different Structures",EUSIPCO-96 signal processing VIII theories and applications,vol.1,pp.109-112,1996.

[3]D.E.Rumelhart,J.L.McClelland, and so on,"Parallel Distributed Processing Explorations in the Microstructure of Cognition",The MIT Press Cambrige,1986.