音声言語情報処理 14－3
(1996. 12. 12)

発音ネットワークに基づく発音辞書の自動生成

深田 俊明　　匂坂 芳典

ATR 音声翻訳通信研究所

〒619-02 京都府相楽郡精華町光台 2-2

Tel.: 0774-95-1387　e-mail: {fukada,sagisaka}@itl.atr.co.jp

あらまし　自然発話音声では，読み上げ発声では起こらないような，大きな発声変形を生じることがある．このような発声を音声認識しようとした場合，標準的な読みが付与された発音辞書を用いても，正しい認識結果は得られない．つまり，標準的な発音系列と実際に発声される発音系列のミスマッチを緩和する機構が必要である．本稿では，発声内容を標準的な辞書表記に基づいて書き起こした音素系列を標準発音系列とし，この発声を音素認識した結果の音素系列を修正発音系列と見なし，これらの対応関係を発音ネットワークとして構築し，この発音ネットワークを利用して発音辞書を自動的に生成する方法について述べる．本手法は，(1) 学習データ中の発声数が少ない語彙に対しても信頼性の高い発音記号列が得られる，(2) 任意の認識対象語彙の追加に対して，発音辞書を生成することができるなどの特徴をもつ．自然発話音声認識実験から，この発音ネットワークに基づいて自動生成した発音辞書は，認識性能，認識時間の両面において，従来の標準発音列に基づいた辞書よりも優れていることが分かった．

キーワード　発音辞書，ニューラルネットワーク，自然発話音声，HMM，音声認識

Automatic Generation of Pronunciation Dictionary Based on Pronunciation Networks

Toshiaki Fukada　　Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02
Tel. 0774-95-1387　e-mail: {fukada,sagisaka}@itl.atr.co.jp

Abstract　In spontaneous speech, word pronunciations vary more than in read speech, while only standard pronunciations in citation form are used in most Japanese recognition systems disregarding real pronunciation variations. To cope with pronunciation variations in spontaneous speech, it is important to build a pronunciation dictionary from real speech data. In this paper, we propose an automatic generation method of a pronunciation dictionary based on pronunciation networks which give the most plausible pronunciations (henceforth correct pronunciations). First, phoneme recognition results are taken as the correct pronunciations. Then pronunciation networks are trained using multi-layer perceptron neural networks to predict these correct pronunciations from the canonical symbol sequences. Finally, pronunciation dictionaries are automatically generated from the pronunciation networks. Experimental results on spontaneous speech show that the automatically-derived pronunciation dictionaries give consistently higher recognition rates and require less computational time for recognition than the conventional dictionary.

key words　pronunciation dictionary, neural networks, spontaneous speech, HMM, speech recognition

1 まえがき

話者，単語，コンテキストなどの違いによる発声変形を考慮した発音辞書の構築は，不特定話者音声認識システムにおける認識精度の向上を実現する上で重要な課題である．これまで，音声学的な知識を利用する方法や，ルールや人手による発音辞書の作成により，認識性能を改善できるという報告がなされている [1] ～ [3]．しかしながら，ルールによる辞書の作成は，エキスパートが必要となる，実際の発声変形に対する正確なモデリングができている保証がないことが問題であり，人手で辞書作成を行う場合には，更に，一貫性を欠く可能性がある，作業時間が膨大になるなどの欠点がある．

このため，データベースを利用して，人手を介することなく発音辞書を自動的に生成しようとする試みも近年数多く検討されてきている [4] ～ [9]．これらの研究は，データベース中に存在する複数発声された認識対象語いを用いて，音素認識や仮説に対するゆう度比較を行うことにより，ラチス表現された発音ネットワークや N-best の単語発音辞書を自動的に生成している．これらの方法は，各認識対象語いごとの発声変形が考慮できるという特長がある反面，(1) 発声数が少ない単語に対しては，自動生成された発音辞書の信頼性は低くなる，(2) Switchboard コーパスのような自然発話音声では，収録データ量を増やしても一定数以上の認識対象語いの発声が得られる保証がない，(3) 任意の認識対象語いの追加に対する発声変形を考慮した発音辞書の生成が困難になるなどの問題点もある．

本稿では，自然発話音声認識システムのための発音辞書の自動生成法を提案する．まず，発声内容を標準的な辞書表記に基づいて書き起こした音素系列を標準発音系列とし，この発声を音素認識した結果の音素系列を修正発音系列と見なす．次に，標準発音系列と修正発音系列の対応づけを行い，標準発音系列から修正発音系列を予測する発音ネットワークをニューラルネットワークにより構築する．この発音ネットワークを用いて，認識対象語いに対する発音辞書を自動的に作成する．つまり，発音辞書を学習データ中の認識対象語いごとの複数発声の認識結果のみから決定するのではなく，学習データ全体を用いて構築した一つの発音ネットワークから決定する．従って，先行研究で生じていた前述の 3 つの問題点が解消でき，学習データ中の認識対象語いの発声数の多少によらない頑健な発音辞書が生成できると考えられる．これは，とりわけ自然発話音声を対象とした音声認識システムにおいて効果があると期待できる．

単語ごとの認識結果を必要とすることなく，発声変形を考慮しようとする試みは，音素の認識誤り傾向をコンフュージョンマトリックスとして保持することにより，発音辞書の音素系列を確率的に表現する方法 [10] や，音素認識をベースとした認識システムにおける単語候補の予備選択部分への適用 [11]，音素認識結果から得られる HMM の状態系列からコンフュージョンマトリックスを利用して語い候補を追加する方法 [12] などの研究によって既になされている．しかしながら，これらの研究は，音素ごとの誤り傾向をコンフュージョンマトリックスとして保持しているため，左右環境より長い環境 (例えば前後 2 音素ずつの音韻環境) を考慮することは，パラメータ数が非常に膨大となるうえ，個々の誤り傾向の信頼性が低くなると考えられる．このように，長い音韻環境を考慮することが困難であるという点は，とりわけ調音結合が激しい自然発話を対象とする場合に不利になると考えられる．これに対して，提案法では，全ての誤り傾向を一つの発音ネットワークで構築しているため，信頼性を失うことなく，少量のパラメータでより長い音素環境が考慮できる．

以下，**2.** では，発音ネットワークの構築法，およびこれを用いた発音辞書の生成方法について述べ，**3.** では，自然発話音声を用いた実験の結果を報告し，**4.** では，本手法により作成した発音辞書，音声認識結果の誤り傾向の分析，認識時間に対して考察を行う．

2 発音ネットワークに基づく発音辞書の自動生成

発音ネットワークに基づく認識用発音辞書は，(1) 修正発音列の生成，および標準発音列との対応づけ，(2) 発音ネットワークの構築，(3) 発音ネットワークを利用した認識語いに対する発音辞書の作成，の 3 つの手順により作成される．以下，これらについて説明する．

2.1 修正発音列の生成

修正発音列の生成，および標準発音列との対応付けは，次のように行う．

1. 学習データに対して音素認識を実行し，認識結果の音素系列を修正発音列とする．
2. 書き起こし読み系列 (標準発音列) と修正発音列との間で文字列レベルの DP をとる．例えば，「あらゆる (標準発音列 /a r a y u r u/, 修正発音列 /a w a u r i u/)」が，

   a r a y u r u (標準発音列)
   a w a u r i u (修正発音列)

   と対応付けられた場合，標準発音列 /a/ は修正発音列 /a/ になり (/a/ に置換と考える)，/r/ は /w/ に置換し，/a/ は /a/ に，/y/ は脱落となり，/u/ は /u/ に，/r/ は /r i/(/i/ が挿入) に，/u/ は /u/ になるとする．

音素認識は，日本語の 26 音素 (無音を含む) を用い，日本語の音素の接続規則による制限下で行った [13]．

2.2 発音ネットワークの構築

発音ネットワークは，図1に示す構造をもつmulti-layer perceptron型のニューラルネットワークを用いて構築した(この発音ネットワークは，英語の綴りから中心文字に対する発音を予測するNETtalk[14]に類似している)．入力は，前後2音素ずつのコンテキストを考慮した5音素からなる標準発音列$L(m-2),\ldots,L(m+2)$であり，出力は，中心音素$L(m)$に対応する**2.1**で得られた修正発音列$A(m)$である．ここで，入力層129ユニット(先々行音素26，先行26，当該25，後続26，後々続26)出力層53ユニット(置換26，挿入26，脱落1)とした．入力層の当該音素のみが25音素となっているのは，認識対象語いを標準発音系列で表現した場合，無音が含まれることはなく，標準発音系列の当該音素が無音となる部分のデータは利用しなかったためである．また，標準発音列と修正発音列が一致している音素，例えば，/a/が/a/に対応している場合，/a/が/a/に置換したとして利用した．学習時の入力層と出力層のデータは，該当するユニットに1を与え，それ以外には0を与えた．例えば，標準発音列/r a y u r/，および修正発音列“”(脱落)の場合，入力層$L(m-2)$の/r/，$L(m-1)$の/a/，$L(m)$の/y/，$L(m+1)$の/u/，$L(m+2)$の/r/のユニットにそれぞれ1を与え，それ以外のユニットには0を与える．このとき，出力層の脱落のユニットに1を与え，それ以外のユニットには0を与える．

ここで，発音ネットワークに要するパラメータ数と，音素の誤り傾向をコンフュージョンマトリックスとして作成したときのパラメータ数を比較する．例えば，前述の入力層129ユニット，出力層53ユニットとし，中間層のユニット数を100とした場合の重み総数は18,200個となる．これは，文献[11]の前後1音素ずつのコンテキストを考慮したコンフュージョンマトリックスとした場合($26^3 = 17,576$)と同程度である．更に，提案法と同様に，前後2音素ずつのコンテキストを考慮した場合($26^5 = 11,881,376$)と比較すると，提案法の方がはるかにパラメータ数は少なく効率的である．すなわち，図1の構造をもつニューラルネットワークを用いる方が，発音変形をより頑健にとらえることが可能であると考えられる．

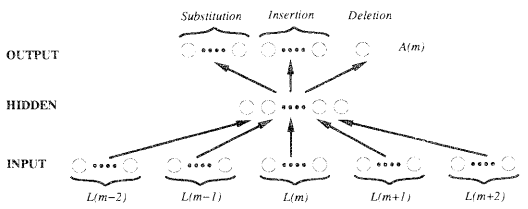

図 1: 発音ネットワークの構造

2.3 発音辞書の作成

前述の発音ネットワークを用いて，標準発音列を入力としたときの出力値に基づいて決定される修正発音列から，認識対象語いに対する発音辞書を自動的に作成する．このとき，以下の3種類の方法で辞書を作成した．

1. 発音ネットワークの出力値が最大のものから得られる修正発音列を登録(network derived single pronunciation: **N(single)**)
2. 1. の系列が標準発音列と異なる場合，標準発音列も辞書に登録(N(single) + canonical pronunciation: **N(single)+C**)
3. 発音ネットワークの出力値を基に，最大N個の候補を複数発音として登録(network derived multiple pronunciations: **N(multi)**)

つまり，方法1は方法3において$N = 1$とした場合と等価である．ここで，方法3におけるNの値は，一般に，語いの音素数が多くなると，複数発音の数も多くなると考えられるため，語いの音素数が5 ～ 9場合$N = 2$，10 ～ 14の場合$N = 4$，15以上の場合$N = 8$とした．但し，発音ネットワークの出力値が小さい候補(当該コンテキストに対する最大出力値で除した値が0.03未満)は登録しない．方法2における各語いに対する発音系列の登録数は，修正発音列と標準発音列が全く同じ場合は1つとなり，異なる場合は2つとなる．方法1から方法3の全ての辞書作成は，5音素以上の音素数Mからなる語いの3音素目から$M-2$番目の音素に対して行う(すなわち1, 2, $M-1$, M番目の音素に対しては標準発音列をそのまま適用する)．このとき，それぞれ前後2音素ずつの音素コンテキストを考慮した合計5音素の音素系列を一音素ずつずらしながら発音ネットワークの入力層に入力し，それぞれの音素系列に対する出力層の値に基づいて修正発音列を決定した．方法1から方法3に対する発音辞書作成のためのフローチャートをそれぞれ図2から図4に示す．

2.4 自然発話音声を用いた発音辞書生成

まず，“ATR Travel Arrangement Corpus”[15]の中の男性1名を用いて，表1に示す分析条件で，逐次状態分割法[16]によって総状態数400，各3混合のHMnet(音素環境依存HMM)を作成し，これに1状態10混合の無音モデルを付加したものを音響モデルとして使用した．この音響モデルを用いて，同一話者の1,530発声，約10万音素の発声データを音素認識した結果を修正発音列とした．このとき，標準発音列(音素書き起こし文字列)を正解として見なした場合の音素正解率は83.80%であった．発声データは，発音ネットワークの学習用(1,489発声)と評価用(41発声)の2つに分けて利用した．発音ネットワークの中間層は一層とし，ユニット数は100個とした．

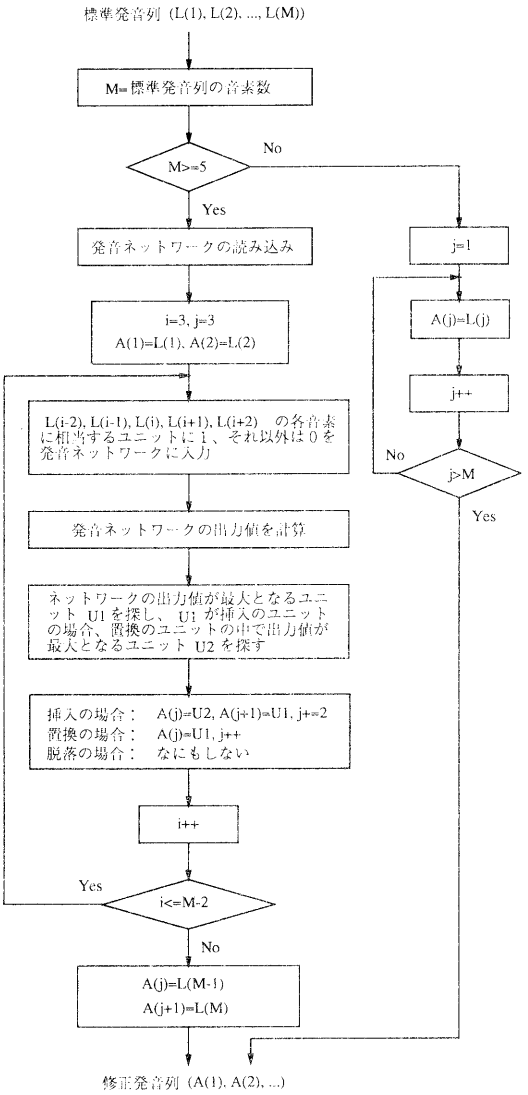

図 2: 発音ネットワークの最大出力値を登録 (方法 1)

ここで，発音ネットワークを構築する際に，不特定話者のデータを用いず，特定話者から作成した理由は，以下の音響モデルに要求される条件と，今回使用したデータベースの状況を鑑みたためである．修正発音系列の生成のための音響モデルに要求される点として，

- 音素認識の性能が悪い音響モデルから作成した発音ネットワークは，発声変形を吸収するものにはならないため，音響モデルの認識性能をなるべく高くする

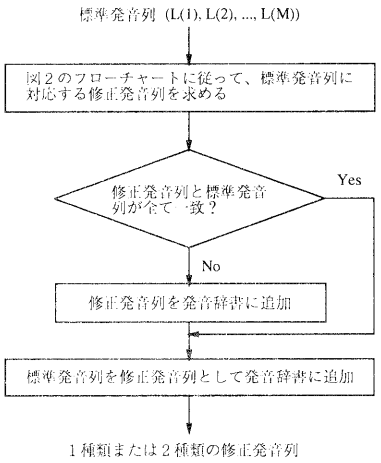

図 3: 発音ネットワークの最大出力値，および標準発音列を登録 (方法 2)

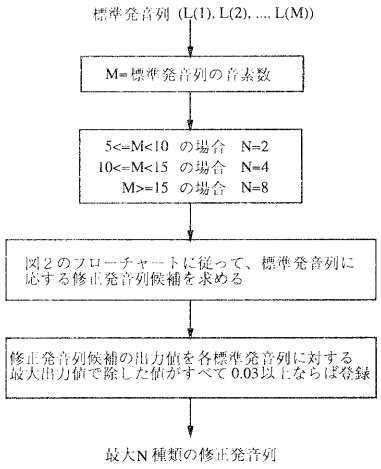

図 4: 発音ネットワークの出力値を基に N 個の候補を登録 (方法 3)

- 複数話者の音響モデルを使用する場合，モデルに大きな性能差を生じない

ことなどが挙げられる．しかしながら，現状の音響モデルは，

- 自然発話音声データに対して不特定話者モデルを用いた場合，高い認識性能を示さない
- 複数話者の話者ごとの自然発話音声のデータ量は少なく，音素コンテキストも話者ごとに不均一なため，

表 2: 発音ネットワークから得られる置換，挿入，脱落の例

	標準発音列 (入力)	修正発音列 (出力)
置換	k a ng *z* e e (関税)	k a ng *d* e e (かんでえ)
挿入	w a r *i* a i (割合)	w a r *i* *y* a i (わりやい)
脱落	g o *y* o y a k u (御予約)	g o o y a k u (ごおやく)

表 1: 修正発音列生成のための分析条件

サンプリング周波数	16 kHz
プリエンファシス	1 - 0.98 z^{-1}
フレーム長	25.6 msec (ハミング窓)
フレーム周期	10.0 msec
特徴ベクトル	10 次 MFCC,
	10 次 Δ MFCC,
	正規化 log パワー,
	Δ 正規化 log パワー

話者適応化手法を用いて複数話者の音響モデルを作成しても，個々の音響モデルに性能差を生じる

ことが確認されている．一方，本稿で用いた自然発話音声データベースは，

- 特定話者 (男性 1 名) については，大量の自然発話音声が利用可能

である．以上の点を考慮して，本稿では，この特定話者に対して作成した音響モデルを用いて，特定話者の音声データに対する修正発音列を求め，発音ネットワークを学習した．

図5 は，学習用データを用いて発音ネットワークを学習した後，評価用データに対して 5 音素 (先行 2 音素，当該，後続 2 音素) からなる標準発音列を入力した時の第一位の出力結果が，音素認識を行った結果に対して，どの程度一致しているかを表す適合率 (%) である．この図から，繰り返し回数が増えるにつれて，徐々に修正発音列 (音素認識結果) に近い発音列を生成する発音ネットワークが構築されていく様子がわかる．

図6 は，**3.** の実験で使用した Travel Arrangement をタスクとする 6,635 語の辞書を用いた場合の，辞書中の語いの表記 (標準発音列) と，発音ネットワークの出力結果の最大値からなる系列の一致度 (%) を学習時の繰り返し回数に対してプロットしたものである．ここで，一致度とは，ある語いに対する標準発音列と，これを発音ネットワークに入力した場合の出力系列が，全く同じ場合は一致したとし，1 音素でも異なる場合は不一致として，一致した数を総語い数 (6,635) で除した割合である．この図から，学習が進むにつれて，辞書の標準発音列との一致度が低下していく，すなわち，元の標準発音列とは異なる表記をもつ発音辞書が生成されていく様子がわかる．

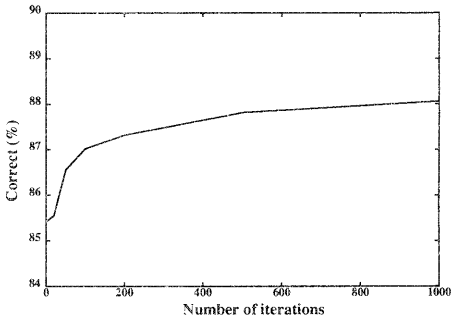

図 5: 繰り返し回数に対する修正発音列への適合率 (%)

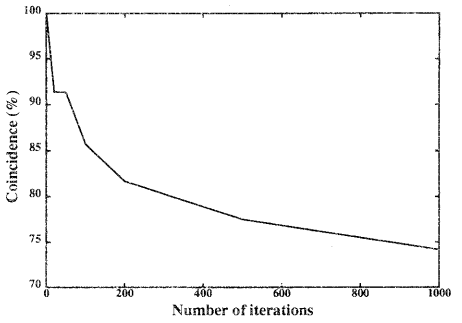

図 6: 繰り返し回数に対する標準発音列との一致度 (%)

表 3: 発音辞書の作成例

辞書	表記
N(single)	k a m o a a r y o k a ng
N(single)+C	k a m o a a r y o k a ng
	k a m o g a w a r y o k a ng
N(multi)	k a m o a a r y o k a ng
	k a m o a w a r y o k a ng
	k a m o a m a r y o k a ng

繰り返し回数 200 回の発音ネットワークに対する置換，挿入，脱落の例を表2 に，「鴨川旅館 (標準発音列 k a m o g a w a r y o k a ng)」に対する作成結果を表3 に

表 4: 不特定話者モデル作成のための分析条件

サンプリング周波数	12 kHz
プリエンファシス	1 - 0.98 z^{-1}
フレーム長	20.0 msec (ハミング窓)
フレーム周期	10.0 msec
特徴ベクトル	16 次 LPC ケプストラム, 16 次 Δ LPC ケプストラム, log パワー, Δ log パワー

それぞれ示す.

3 音声認識実験

従来法として利用している標準発音列の辞書 (canonical: **C**), および発音ネットワークから得られた 3 種類の発音辞書 (**N(single)**, **N(single)+C**, **N(multi)**) の合計 4 種類の辞書を評価するために, 単語グラフに基づく自然発話音声認識システム [17] を用いた認識実験を行った.

3.1 実験条件

発音ネットワークを作成した話者と同一話者の特定話者モデル (SD), 不特定話者モデル (SI), 不特定話者モデルから話者適応を行ったモデル (ASI) に対して実験を行った. SD は **2.4**で作成したものと同一の HMnet である. SI は表4 に示す条件で, 不特定話者音声データに対して逐次状態分割法を用いて作成した 400 状態, 各 5 混合の HMnet に 1 状態 10 混合の無音モデルを付加したモデルを用いた. ASI は評価データ以外の同一話者による発声データを用いて, SI から移動ベクトル場平滑法 [18] による話者適応を行い作成した. 言語モデルは品詞・単語可変長 n-gram[19] により得られたものを全ての音響モデルに対して使用した. 評価データは, 発音ネットワークの作成, 音響モデル, 言語モデルの全てに対してオープンであり, SD は 41 発声, SI, および ASI は 7 話者からなる 98 発声を用いた.

言語モデルと音響モデルの重み付けのバランス, ビーム幅は, まず, 従来の辞書で最高性能を与える重み付け値, およびビーム幅を SD, SI, ASI のそれぞれに対して決定し, 提案法の 3 種類の辞書を用いる場合も, これらの値で固定して音声認識を行った. また, 発音辞書は繰り返し回数 200 回の発音ネットワークから作成したものを用いた. これは, 繰り返し回数 50, 100, 200, 500, 1,000 回の 5 種類の発音ネットワークから得られた発音辞書を利用して音声認識実験を行った結果, 全ての辞書 ($5 \times 3 = 15$ 通り) は, 従来の辞書 **C** よりも認識性能が良かったが, 繰り返し回数 200 回を過ぎると認識性能が変化しない, あるいは若干低下することが確かめられたためである.

表 5: 辞書サイズ (延べ語い数)

辞書	C	N(single)	N(single)+C	N(multi)
サイズ	6,635	6,635	7,854	14,324

表 6: 認識結果 (単語 accuracy %)

辞書	C	N(single)	N(single)+C	N(multi)
SD	19.98	20.82	21.07	24.46
SI	12.19	12.89	16.20	19.37
ASI	27.39	28.16	32.41	32.56

る. この性能低下は, 過学習が原因であると考えられる.

3.2 認識結果

実験に用いた辞書サイズ (延べ語い数), および認識結果をそれぞれ表5 , 表6 に示す. 表6 の結果から, 特定話者 (SD) では, 従来の辞書 **C** に比べて, 発音ネットワークから得られた同じ辞書サイズの辞書 **N(single)** の方が, 若干認識率が向上していることがわかる. また, 辞書 **N(single)+C**, 辞書 **N(multi)** の場合も, 辞書サイズは辞書 **C** よりも増えるが, 辞書 **C** を用いた場合と比較して, 更に認識率が改善されていることがわかる.

次に, 不特定話者 (SI), 話者適応モデル (ASI) の結果においても, 特定話者の結果と同様に, 提案法の辞書の方が認識率が高いことがわかる. この結果より, これらの認識実験で用いた発音辞書は, 特定話者の発音ネットワークを基に作成されているにもかかわらず, 不特定話者や話者適応モデルに対しても有効であることがわかる. また, 辞書 **C** と辞書 **N(multi)** の認識率の違いを話者ごとに調べた結果, 7 話者中, 不特定話者モデルでは 5 話者, 話者適応モデルでは 6 話者について, 辞書 **N(multi)** の方が認識率が高かった. これは, 今回用いた特定話者の発音ネットワークが, この話者固有な発声変形の傾向のみを保持しているのではなく, むしろ話者に不偏な自然発話における発声変形を捉えていることを示していると考えられる. しかし, 他の特定話者から作成した発音ネットワークを用いても, 同様なことが言えるかどうかは結論づけられず, 今後の検討課題である.

4 考察

ここでは, 発音辞書, 音声認識結果の分析, 認識時間の観点から考察を行う.

4.1 発音辞書の分析

3.2 の認識結果より, 提案法により自動生成された辞書 **N(single)**, 辞書 **N(single)+C**, 辞書 **N(multi)** は, 従来辞書 **C** よりも認識性能が改善できることが示された.

表 7: 認識結果の分析 (挿入 / 脱落 / 置換数)

辞書	C	N(single)	N(single)+C	N(multi)
SD	274/12/375	265/10/379	264/10/379	230/13/381
SI	261/88/789	248/91/790	233/106/747	223/107/715
ASI	228/87/626	234/89/608	197/101/578	184/95/595

表 8: 認識時間の比較 (単位秒. 括弧内は辞書 C を 1 とした場合の比率)

音響モデル	発声時間	C	N(single)	N(single)+C	N(multi)
SD	195.5	104.1 (1)	103.8 (1.00)	106.7 (1.02)	104.1 (1.00)
SI	320.7	3,650 (1)	2,932 (0.80)	3,021 (0.83)	2,530 (0.69)
ASI	320.7	1,497 (1)	1,530 (1.02)	1,196 (0.80)	1,138 (0.76)

ここでは，その要因を見つけることを目的として，辞書 **N(single)** の修正発音列と辞書 C の標準発音列の差異について分析を行った．但し，ここで行った分析は，当該音素のみに着目した結果であり，前後の音素については考慮していないことに注意されたい．ここで，辞書の総語い数 6,635 語のうち，提案法により変更対象となる語いは，音素数 5 以上の語い 5,029 語であり，以下の分析はこの 5,029 語に対して行った．このときの変更対象の音素数は，全部で 21,513 個である．

語いを構成する音素系列全体として評価した場合，5,029 語のうち，提案法により生成した修正発音列が標準発音列と同一であったものは，3,810 語 (75.8%) であった．つまり，語い数の 4 つに 1 つは何らかの変化が生じていることになる．また，音素ごとに評価した場合，5,029 語の総音素数 21,513 個のうち 20,057 音素 (93.2%) は，修正発音列と標準発音列が同一であった．つまり，約 7% の音素に対して変化が生じていた．

脱落となった音素数の合計は 804 個であり，辞書 C に対して脱落が多かった音素ラベル (当該音素の全体数に対する脱落数の割合が 10% 以上) としては，/q(促音)/(46%), /w/(36%), /p/(17%), /g/(11%) があった．次に，置換で目立ったものには，/b/ が /g/ または /d/ へ，/g/ が /n/ へ，/w/ が /m/ または /y/ へ，/z/ が /d/ へなどがあった．挿入された音素は，全体で 67 個と少なかったが，その中では，/y/(12 個)，/sil(無音)/(10 個)，/g/(9 個) が多かった．一方，変化を受けにくい音素もあり，辞書 C と辞書 **N(single)** の表記の 95% 以上が同じであった音素は，/a/, /ch/, /o/, /r/, /s/, /ts/, /zh/ があった．

以上の分析結果の中で，例えば，/g/ の鼻濁音化や閉鎖部の脱落・挿入などは，文献 [2] で述べられている規則と同じである．このことからも，発音ネットワークは，特定話者に固有な音素変形規則のみを保持しているというよりは，むしろ，話者に不偏な変形ルールが構築できていると考えられる．

4.2 音声認識結果の分析

次に，表 6の音声認識結果の挿入，脱落，置換数に対する分析を行う．結果を表 7に示す．

まず，特定話者認識 (SD) に対しては，提案法の発音辞書を用いることで挿入誤り数が減少していることがわかる．特に，辞書 **N(multi)** を用いた場合，16% も挿入誤りが改善されている．これは，従来法の発音辞書 C では，実際の発声に対して適切なものが少なく，その結果，音素数の多い語いの入力，例えば，「ありがとうございます」(実際の発声は，/arigatoozaimasu/ に近い) が，「有」+「り」+「が」+「東西」+「ま」+「す」のように，短い語いの連続として認識されてしまうのに対し，提案法では，実際の発声により近い発音辞書 (/arigatoozaimasu/) が構築されているため，このような認識誤りが起こらず，結果として挿入誤りが低減したと考えられる．逆に，短い語いの連続が一つの長い語いとなる，すなわち，脱落数が増加することも考えられるが，表 7からわかるように，このような現象はほとんど生じていないと考えられる．

次に，不特定話者 (SI)，話者適応モデル (ASI) に対する傾向は，SD で見受けられた挿入誤りの低減に加え，辞書 **N(single)+C** や辞書 **N(multi)** を用いることで，置換誤り数が大幅に低減しており，これが全体の認識性能の改善に大きく寄与していることがわかる．しかしながら，SI や ASI に対しての提案法の発音辞書 **N(multi)** は，辞書 C に比べて脱落誤り数が増加している．特に，SI では，辞書 C と比較して，22% も脱落誤り数が増加していることがわかる．これは，ASI の脱落誤り数の増加が 9% 程度であることから察するに，発音辞書の系列に不具合を生じているというよりは，むしろ，音響モデル SI の性能が悪いことに起因していると考えられる．

4.3 認識時間の比較

3.2より認識性能の面では，**2.** で述べた方法で作成し

た発音辞書は有効であることがわかったが，辞書N(single)+Cや辞書N(multi)の場合，辞書サイズが増加するため認識時間が増大する可能性がある．そこで，Hewlett Packard社のHP9000/735ワークステーション(135SPECint92)を用いた場合の認識時間を測定した．結果を表8に示す．この表より，辞書サイズが増加しているにもかかわらず，認識時間はほとんどの場合増加せず，SIやASIにおける辞書N(single)+Cや辞書N(multi)においては，辞書Cと比べて約20%〜30%程度の認識時間の高速化が達成できている．これは認識対象の発声に対して，適切な表記が辞書中に含まれている場合，音響ゆう度が辞書Cによる表記に比べ高くなり，ビーム中に含まれる候補が言語的な辞書を用いた場合よりも減少したためであると考えられる．

5 むすび

本稿では，自然発話における発声の揺れや変動を統計的にとらえることを目的として，音素認識結果を利用した発音ネットワークの構築方法，およびこれを用いた発音辞書の自動作成法を提案した．この発音ネットワークは，音素の置換，脱落，挿入を取り扱うことができる，長いコンテキスト(本稿では前後5音素)を少量のパラメータで考慮できるなどの特徴をもつ．この発音ネットワークに基づいて音声認識用辞書を作成し，自然発話音声認識実験を行った結果，従来の標準発音列で表記された発音辞書を用いるのに比べて，5〜7%程度単語accuracyが向上し，更に，認識時間も最大30%程度低減できることが確かめられ，本手法の有効性が示された．

本手法により自動的に作成された辞書N(single)+Cや辞書N(multi)は，音響モデルの再学習のために利用することもでき[7][9]，今後はこの再学習による効果を調べていく予定である．この他にも，(1) 不特定話者による発音ネットワークの構築，(2) Nベスト候補の再評価による単語間，および単語の始端，終端2音素に対する発音ネットワークの適用，(3) 複数発音候補の確率値付きネットワークによる表現[6]などに対する展開が考えられる．

謝辞

日頃熱心に討論頂くATR音声翻訳通信研究所の皆様に感謝します．特に，ニューラルネットワークの学習プログラムを御提供頂いたMike Schuster研究員，および実験方法についての有益なコメントを頂いた清水徹主任研究員に深謝いたします．

参考文献

1) 杉山 雅英, 相川 清明, 鹿野 清宏：“音声認識システムにおけるTop-down的音響処理”，音声研資, S82-62, pp. 489–496 (1982-12).

2) S. Kimura and Y. Nara : “Extraction of phonemic variation rules in continuous speech spoken by multiple speakers,” *Proc. ICASSP-87*, pp. 825–828, 1987.

3) L. Lamel and G. Adda : “On designing pronunciation lexicons for large vocabulary, continuous speech recognition,” *Proc. ICSLP-96*, pp. 6–9, 1996.

4) P. Schmid, R. Cole and M. Fanty : “Automatically generated word pronunciations from phoneme classifier output,” *Proc. ICASSP-93*, pp. II-223–II-226, 1993.

5) 今井 亨, 安藤 彰男, 宮坂 栄一：“発声変形ルールの自動生成に基づく音声認識辞書のマルチエントリー化”，音響学会講論集, 1-R-11 (1994-10).

6) C. Wooters and A. Stolcke: “Multiple-pronuncition lexical modeling in a speaker independent speech understanding system,” *Proc. ICSLP-94*, pp. 1363–1366, 1994.

7) T. Sloboda : “Dictionary learning : performance through consistency,” *Proc. ICASSP-95*, pp. 453–456, 1995.

8) J. Humphries, P. Woodland and D. Pearce : “Using accent-specific pronunciation modelling for robust speech recognition,” *Proc. ICSLP-96*, pp. 2324–2327, 1996.

9) E. Fosler : “Automatic learning of word pronunciation from data,” *Proc. ICSLP-96*, pp. 28–29 (addendum), 1996.

10) 川端 豪, 三輪 譲二, 城戸 健一, 牧野 正三：“音素の信頼度を利用した単語音声認識”，音声研資, S80-18, pp. 141–148 (1980-6).

11) 田中 信一, 伊藤 彰則, 牧野 正三, 曽根 敏夫, 城戸 健一：“日本語Dictationシステムにおける文節検出の高速化”，信学技報, SP90-70, pp. 17–24 (1990-12).

12) 脇田 由実, ハラルド シンガー, 匂坂 芳典：“複数音素にわたるHMMの誤認識特性を用いた語彙候補の追加”，信学技報, SP95-30, pp. 41–47 (1995-06).

13) 大脇 浩, ハラルド・シンガー, 鷹見 淳一：“音素配列構造の制約を用いた音素タイプライタ”，信学技報, SP93-133, pp. 71–78 (1993-12).

14) T. Sejnowski and C. Rosenberg: “NETtalk: a parallel network that learns to read aloud,” The Johns Hopkins Univ. Electrical Engineering and Computer Science Tech. Report JHU/EECS-86/01, 1986.

15) A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka: “Japanese speech databases for robust speech recognition,” *Proc. ICSLP-96*, pp. 2199–2202, 1996.

16) 鷹見 淳一, 嵯峨山 茂樹：“逐次状態分割法による隠れマルコフ網の自動生成”，信学論(D-II), **J76-D-II**, 10 pp. 2155–2164 (1993-10).

17) 清水 徹, 山本 博史, 松永 昭一, 匂坂 芳典：“単語グラフを用いた自由発話音声認識”，信学技報, SP95-88, pp. 49–54 (1995-12).

18) K. Ohkura, M. Sugiyama and S. Sagayama : “Speaker adaptation based on transfer vector field smoothing with continuous mixture density,” *Proc. ICSLP-92*, pp. 369–372, 1992.

19) H. Masataki and Y. Sagisaka : “Variable-order n-gram generation by word-class splitting and consecutive word grouping,” *Proc. ICASSP-96*, pp. 188–191, 1996.