音声言語情報処理 34－24
(2000. 12. 22)

音節強調発声に頑健な自然発話音声の認識法

奥田 浩三　　松井 知子　　中村 哲

ATR 音声言語通信研究所
〒619-0288 京都府相楽郡精華町光台 2 丁目 2 番地 2
E-Mail:{kokuda, tmatsui, nakamura}@slt.atr.co.jp

あらまし 自然発話音声には，正しく意図を伝えるための強調発声や言い直し，感情表現など，さまざまな発話様式が含まれている．より良いヒューマンインタフェースとして音声認識システムを考えた場合，これらの発話様式の変動に頑健な音声認識システムを構築することは非常に重要である．特に現在の音声認識システムでは，誤認識の発生は避けられず，その言い直しに対して頑健にする必要がある．言い直し発話では，より明瞭に発声する，音素継続時間長が増加するという変化が生じるとともに，音節強調発声の出現頻度が増加するという傾向がある．本稿では，言い直し発話における音節強調発声に有効な音声認識手法について検討したので報告する．音節強調発声は，発話様式が孤立音節発声に近くなるとともに，音節間の音響的特徴が変形する．本手法では，後続音素環境が無音の triphone 母音モデルと，先行音素環境依存 biphone 母音モデルをマルチモデル化して用いることにより，上記の音節強調発声の問題に対処する．デコードの際，音素ごとに尤度の高いモデルを選択することで，認識辞書の拡張や音響モデルの切り替えを行うことなく，音節強調発声に対する認識率を向上することができた．

キーワード 音声認識，発話様式，音節強調発声，言い直し，マルチモデル

Robust speech recognition for stressed Japanese speech

Kozo OKUDA, Tomoko MATSUI, Satoshi NAKAMURA
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 JAPAN
E-Mail:{kokuda, tmatsui, nakamura}@slt.atr.co.jp

Abstract Speaking styles in spontaneous speech often vary in order to convey stress, emotion or error recovery. In a speech recognition system for an intelligent human-machine interface, it is crucial to achieve robustness against speaking style variations. The system must be especially robust with regard to error recovery speech because current speech recognition systems always exhibit errors. Speech in error recovery tends to be uttered in a syllable-stressed way and to have longer phoneme durations. This paper investigates a robust method to recognize syllable-stressed speech for error recovery. In syllable-stressed speech, each syllable is uttered in an isolated way and the acoustic characteristics between syllables are changed. To cope with these problems, we propose a new recognition method. For isolated syllables, we use a vowel model which is succeeded by silence with a conventional triphone model. For the change in acoustic characteristics between syllables, we use a left context dependent biphone vowel model with a conventional triphone model. During decoding, the model which shows the highest likelihood for input speech is implicitly selected for each phoneme. This method improves the performance for syllable-stressed speech without any need to expand the recognition dictionary or explicitly select the models.

keywords speech recognition, speaking style, stressed speech, error recovery, multi-model

1. はじめに

近年の音声認識技術は，統計的手法の導入と大規模なデータベースの構築により，不特定話者連続音声認識においても，飛躍的に認識率が向上している．しかしながら，現行の音声認識システムでは認識しやすい発話様式と，認識しにくい発話様式が存在し，認識しにくい発話様式では認識率が急激に劣化する傾向にある．自然発話音声には，正しく意図を伝えるための強調発声や言い直し，感情表現など，さまざまな発話様式が含まれている．より良いヒューマンインタフェースとして音声認識システムを考えた場合，これらの発話様式の変動に頑健な音声認識システムを構築することは非常に重要である．特に現在の音声認識システムでは，誤認識の発生は避けられず，言い直しによる発話様式の変動に対して頑健にする必要がある．

誤認識時の言い直しにおいて，システム利用者はシステムが認識しやすいよう，発話様式を変える場合が多い．言い直し発話では，より明瞭に発声する，音素継続時間長が増加する，などの変化が生じるとともに[1][2]，音節強調発声の出現頻度が増加するという傾向がある．通常の言い直し発話に対しては，言い直し発話データを収集しモデル適応に用いることで，認識率を改善できることが報告されている[3]．しかしながら，言い直し発話の中でもその音響的特徴の変形が大きい音節強調発声までは考慮されていない．

本稿では，言い直し発話における音節強調発声に着目し，音節強調発声に対する認識率を向上する音声認識手法について検討したので報告する．本手法は，従来の母音モデルと共に，音節強調発声における発話様式の変動を表現するための母音モデルを複数用意し，デコードの際，音素ごとに尤度の高いモデルを選択するものである．これにより，認識辞書の拡張や音響モデルの切り替えを行うことなく，音節強調発声に対する認識率の改善を達成した．

本稿ではまず，使用したベースラインの認識システムについて述べ，次に言い直し発話における音節強調発声の出現頻度とその音響的特徴についてまとめる．更にベースラインシステムを用いた認識実験を行い，音節強調発声の認識率への影響について述べる．以上の結果を元に，音節強調発声に対応したデコード手法を提案するとともに，評価実験を通して本手法の有効性を確認する．

2. ベースラインシステム

本報告における認識実験では，当研究所で開発した連続音声認識用の音声認識エンジン，ATRSPREC を使用した[5]．評価実験においてベースラインとなるシステムの概要は以下の通りである．音響特徴パラメータは，サンプリングレート 16kHz，プリエンファシス 0.98，窓長 20msec，フレームシフト 10msec で抽出した 25 次元の特徴ベクトル（12 次メルケプストラム，12 次Δメルケプストラムと Δlog power）を用いている．ベースライン音響モデルは性別依存モデルであり，5 混合ガウス分布（pause model は 10 混合），1400 状態（pause model は 3 状態）の状態共有化 HMM（HMnet）で表現されている．学習データは本研究所で収集した旅行対話タスクデータベースより，男性 167 話者（約 2 時間），女性 240 話者（約 3 時間）のデータを使用した[6]．

言語モデルに関しては，ベースライン音響モデルと同じ学習セットを用いて学習した，多重クラス複合 N-gram モデルを用いた[7]．多重クラス複合 N-gram は，クラス N-gram を基本として，直前直後の単語の接続性を考慮し，各単語を先行単語として用いる場合と，後続単語として用いる場合とで，複数の異なるクラスを割り当てるモデルである．本言語モデルにおけるクラス数は，先行単語が属するクラス（from クラス）700，後続単語が属するクラス（to クラス）700 となっている．認識辞書は 27k 単語である．

3. 言い直し発話における音節強調発声

音声認識の誤認識時における言い直し発話では，より明瞭に発声する，音素継続時間長が増加する，無音区間の出現頻度と継続時間長が増加する，ピッチが上昇するなどの変形が生じる．これらの変形に加え，各音節を強調して発声する，音節強調発声の出現頻度が増加するという傾向がある．音節強調発声とは例えば，「あした」という発声が，音節ごとに強調され，「あ＿し＿た」という具合に，孤立音節発声に近い発声になることである．

本稿ではまず，音節強調発声の出現頻度を調査するため，誤認識時の言い直し発話音声データを収録した．調査にあたり，単語発声の方がその傾向がより明確に現れるため，発声は単語発声とし，210 単語を収録した．収録は，誤認識をシミュレートする収録装置を用いて行い，被験者にはディスプレイに表示される単語を音声で入力してもらった．また，誤認識時には認識が成功するまで音声入力を繰り返してもらった．この際，認識ができなかったことのみを画面に表示し，どのように誤認識したのかなどの情報は，被験者に一切与えなかった．連続音声認識において，我々の認識システムの平均認識率は 80％を超えるが，認識率の悪い話者に対しては 60％程度となるため，全体の 40％の単語に対して誤認識を発生させた．また，誤認識する単語の 50％が 1 回目，25％が 2 回目，12.5％が 3 回目，12.5％が 4 回目の言い直しで正しく認識するようにした．このようにして 5 名の被験者から，言い直し発話データを含む音声データを収録した．

次にこの言い直し発話中に，音節強調発声がどの程度出現しているかを評価した．音節強調発声の出現頻度は，

強調の強さを評価する必要があり，現時点でその強さを評価する指標は存在しない．そこで本章では評価者1名による聴感的な主観評価を行った．評価は，1回目の発声と言い直し発話を比較し，1）特に変化のなかったもの，2）ピッチの上昇や母音継続時間長の増加のみの変化であると判断されるもの，3）発声の変形が音節強調発声にまで至っているもの，の3段階評価で行った．結果を表1に示す．主観評価ではあるが，多い人で30%程度の音節強調発声が含まれていることがわかる．

通常の発声と音節強調発声の音響的な違いとして，図1に／jizai／と発声した1回目の発話と，音節強調発声と判断される言い直し発話のスペクトログラムを示す．円で囲んでいる部分は，/i/ と/z/ の接続部分であるが，1回目の発話と比較し音節強調発声では，音節間の連続性が崩れるとともに，音響的特徴が大きく変化していることがわかる．

言い直し発話に対しては，言い直し発話データのみを収集し音響モデルを構築することが有効である [3][4]．しかしながら，音節強調発声は言い直し発話においても出現頻度が低く，収集が困難である．また，音節強調発声データを含んだデータによりモデルを構築した場合，音響的な分布が広がるため，モデルそのものの性能が劣化する可能性がある．音響モデルの性能を劣化させずに，音節強調発声に対して認識率を向上する手法を検討する必要がある．

4. ベースラインシステムを用いた評価実験

音節強調発声が認識システムに与える影響を調査するため，ベースラインシステムを用いた認識実験を行った．男性話者5名，女性話者5名から，通常の連続発声音声と，意図的に発声した音節強調発声音声の文章データを収録した．発声内容は，当研究所で用いている旅行対話タスクの中から選択した 20 文章をそれぞれの発声で用いた．収録したデータに対するベースライン音響モデルでの認識結果は表2の通りである．

表1　言い直し発話における音節強調発声の出現頻度

話者	変化なし	母音継続時間長の変化やピッチの変化のみ	音節強調発声の傾向あり
1	48.7%	50.0%	1.3%
2	27.4%	42.0%	30.6%
3	72.6%	24.8%	2.5%
4	86.6%	11.5%	1.9%
5	54.4%	19.6%	25.9%

表2　ベースライン音響モデルでの認識結果

話者	通常発声	音節強調発声
1	75.81%	18.95%
2	78.55%	-28.93%
3	87.32%	-43.96%
4	78.62%	-86.90%
5	80.62%	-51.88%
6	81.43%	-27.14%
7	81.30%	-11.97%
8	73.79%	6.21%
9	74.38%	-2.50%
10	75.00%	32.14%
平均	78.39%	-15.94%

この結果からわかるように，連続音声認識用に学習した音響モデルでは，通常の発声に対して 80%程度の認識率が得られているのに対し，音節強調発声では認識率が大きく劣化する．この原因の一つとして，音節強調発声時のストレスによる音響的な特徴量のずれと音素継続時間長の変化による，音響モデルとのミスマッチが考えられる．そこで，音節強調発声に対して，話者ごとに MAP 推定による話者適応を行った．収録したデータ数が少ないため，全データを適応に用い，評価実験はクローズドデータによるものとした．適応は平均値のみ，平均値と状態遷移確率の両方，の2種類を行った．適応前の音響モデルと，適応後の音響モデルによる認識実験の結果を図2に示す．話者適応を行うことにより，認識率は改善される．また音節強調発声に対しては，状態遷移確率の適応も効果があることがわかる[3]．しかしながら話者適応を行っても，通常発声の認識率より大きく劣化している．

以上のことより音節強調発声は，連続発話音声データで構築した音素環境依存 triphone 音響モデルでは表現できない音響的特徴を有するとともに，その変形が大きいため，話者適応では十分な性能が得られないと考えられる．

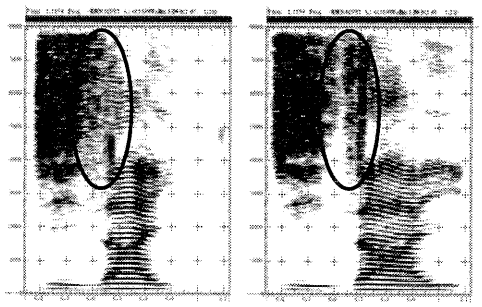

図1　／jizai／と発声した場合のスペクトログラム

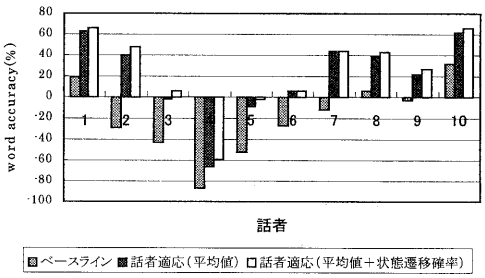

図2　話者適応による音節強調発声の認識率

5. 音節強調発声に対応した認識法

通常の大語彙連続音声認識システムでは認識の際，認識辞書に登録されている音素表記に従い，HMMで表現された音素モデルを結合，デコードを行う．特にtriphone音響モデルを用いた音声認識システムでは，各音素間で連続的に変化する音響的特徴も表現することで，連続発話音声に対して認識率を向上している．しかしながら，音節強調発声では各音節を強調して発声するため，次のような変形が生じる．

1) 孤立音節発声に近い発話様式になる
各音節間に無音が存在する，孤立音節発声に近い発話様式になるため，各母音の音響的特徴が，後続音素環境が無音の母音モデルに近くなる．
2) 各音節間の音響的特徴が変形する
各音節間の音響的特徴が変形し，連続発声と孤立音節発声の中間的な発話様式になるため，後続音素環境との結びつきが弱くなり，後続音素環境に依存していた音響的特徴が変形する．

これらの変形により，通常のtriphone音響モデルでは十分にその特徴を表現できず，認識率が劣化したと考えられる．図3に，この様子の例を示す．完全な孤立音節発声の場合は，認識辞書に連続発声用と孤立音節発声用の音素表記を併記することで認識率の向上も期待できるが，連続発声と孤立音節発声の中間に位置する音響特性を有する音節強調発声にはあまり効果が期待できない．そこで本稿では，1つの音素表記に対し複数のモデルを用意し，デコードの際，それぞれのモデルに対して仮説を展開する音響モデルを提案する．具体的には，

1) 孤立音節発声に近い発話様式に対応したモデルとして，後続音素環境が無音のtriphone母音モデル
2) 各音節間の音響的な変形に対応したモデルとして，先行音素環境依存biphone母音モデル

を通常の母音モデルとマルチモデル化して用いた．音

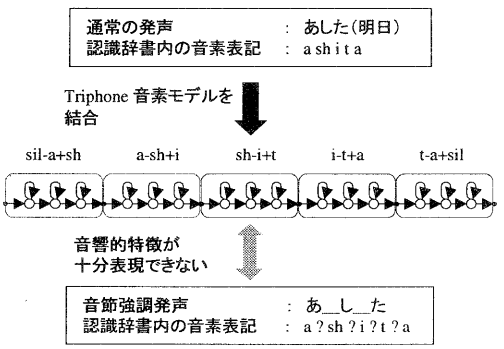

図3　音節強調発声における音素環境の変形

響的特徴の変化だけではなく，完全な孤立音節発声にも対応できるよう，各モデルにはスキップ可能な1状態の無音モデルを追加した．また，マルチモデル化した母音に後続するモデルに対しても，孤立音節発声に対応できるよう，先行音素環境が無音のモデルをマルチモデル化した．本提案例として図4に，先行音素がt，後続音素がkの該当音素aと，先行音素がa，後続音素がiの該当音素kに関するモデルを示す．音節強調発声の音響的な特徴に近い通常のモデルを流用することで，学習データを追加する必要はなくなる．また，マルチモデル化するため，各状態が表現する分布を広げることなく，音節強調発声に対応できるとともに，デコードの際，尤度が最も高くなる経路が選択されるため，認識辞書の拡張や，発話様式ごとの音響モデルの切り替えが不要になる．以下，ベースとなる母音モデルを第1モデル，後続音素環境が無音の母音モデルを第2モデル，先行音素環境依存biphone母音モデルを第3モデルと呼ぶ．

6. 提案手法による評価実験

提案手法による評価実験を行った．ベースとなる音響モデルには，ベースライン音響モデルを使用した．提案手法で用いる第2モデルは，ベースライン音響モデルに含まれる該当母音モデルを流用した．また，第3モデルには，ベースライン音響モデルの学習に使用した学習セットを用いて構築した状態数1400，ガウス混合分布数5の性別依存 biphoneモデルの母音モデルを使用した．

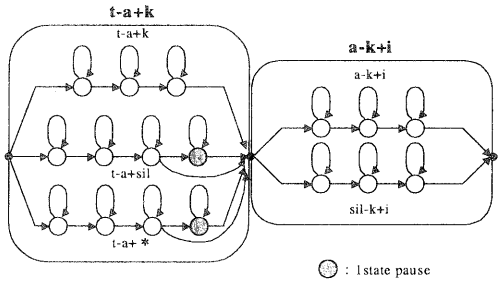

図4　マルチモデル化した音響モデルの例

表3　通常発声に対する提案手法の認識率

話者	ベースライン	提案手法
1	75.81%	80.80%
2	78.55%	79.03%
3	87.32%	88.03%
4	78.62%	84.83%
5	80.63%	80.00%
6	81.43%	80.00%
7	81.30%	85.21%
8	73.79%	75.86%
9	74.38%	74.38%
10	75.00%	77.86%
平均	78.39%	80.37%

6－1．提案手法による認識実験

本手法による通常発声の認識率を表3に示す．この結果より本提案手法は，ベースライン音響モデルと比較し，通常発声に対して認識率の劣化がないことがわかる．次に本提案手法を用いた，音節強調発声に対する認識率を図5に示す．この結果より，話者適応を行ったtriphone音響モデルと比較しても，話者適応を行っていない本提案手法の方が良好な認識率を得ていることがわかる．

ここで，提案手法による話者ごとの認識結果について分析する．図5における話者5のように，提案手法でも認識率が10%未満の話者が含まれている．この原因を分析するため，提案手法で用いた音響モデルによる音節強調発声の音響尤度を調査した．調査の方法として，提案手法を用いる際のベースとなった音響モデルを用いて音素アライメントを行い，話者ごとに音響尤度を算出，その尤度比較を行うことを試みた．音素アライメントによる音響尤度と，提案手法による認識率の関係を図6に示す．この図より，音節強調発声とベースとなる音響モデルのミスマッチが小さい程，提案手法による認識率が良いことがわかる．話者5に関しては，発声音声と音響モデルとのミスマッチが大きすぎるため，提案手法でも十分な認識率が得られなかったと考えられる．逆に話者1や話者10は，音響モデルとのミスマッチは小さく，提案手法が効率よく働いたと考えられる．言い換えれば，これらの話者は，音響的特徴の変化は少ないものの，ベースライン音響モデルでは十分に表現することのできなかった発話様式になっていたと考えられる．

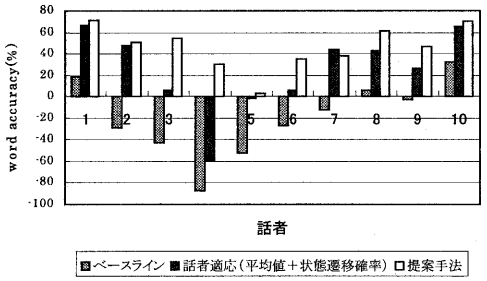

図5　音節強調発声に対する提案手法の認識率

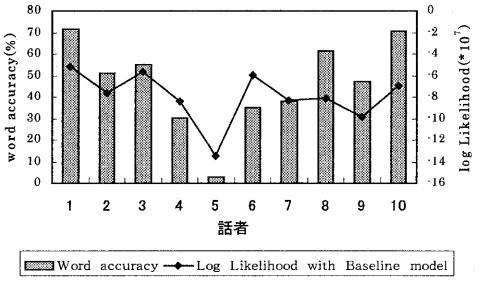

図6　ベースライン音響モデルによる音節強調発声の対数尤度と、提案手法による認識率との関係

6－2．提案手法における各モデルの効果

次に，本提案手法が音節強調発声に対して，どのように働いたかを分析するため，マルチモデルとして追加したモデルがデコードの際，どの程度選択されたかを調査した．調査は第2モデル，第3モデルそれぞれの影響を区別するため，どちらか1つのモデルのみをマルチモデル化して行った．結果を表4に示す．通常発声では，第2モデル，第3モデルともに，ベースライン音響モデルとほぼ同程度の認識率となっている．これに対して，音節強調発声に対しては，第2モデル，第3モデルともに認識率が向上し，第2モデルが最も良い結果となっている．通常発声と音節強調発声における各モデルの選択率を比較すると，第3モデルが1.2倍の増加に対し，第2モデルは4.7倍と大きく増加している．

話者ごとのモデル選択率と，ベースライン音響モデルに対する改善認識率（ベースライン音響モデルでの認識率と，各モデルでの認識率の差）の関係を図7に示す．第3モデルは，認識率の改善量に関わらず，どの話者でも同程度の比率で選択されている．これは，音響尤度的に通常のtriphoneモデルよりもマッチしているが，選択されたモデルが必ずしも認識率の向上に貢献しないということを意味する．これに対して，第2モデルと認識率の改善量には何らかの関係があり，マルチモデルとして追加したモデルがより多く選択された話者は，認識改善量も大きくなっている．また，選択比率が少なく認識改善量が小さい話者は，ベースとなる音響モデルと音節強調発声との音響的距離が大きな話者となっている．言い換えれば，話者適応などにより音響モデルとの音響的距離を改善すれば，これらの話者に対しても提案手法が効果的に働くことが期待できる．

表4　モデルごとの認識率とモデル選択率

	ベースライン	第2モデル (モデル選択率)	第3モデル (モデル選択率)
通常発声	78.39%	79.35% (10.5%)	79.40% (47.4%)
音節強調発声	−15.94%	41.91% (49.7%)	28.95% (57.0%)

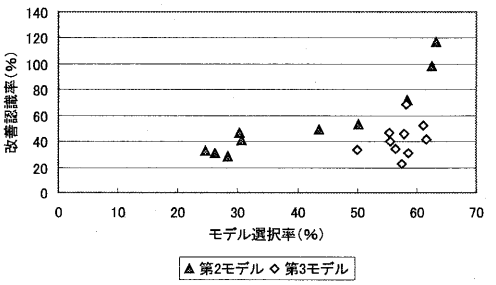

図7　第2モデル、第3モデル選択比率

6－3．　話者適応を併用した認識実験

MAP 推定による話者適応を行ったモデルをベースとなる音響モデルに使用し，本提案手法を実施した結果を図8に示す．ベースとなる音響モデルの適応には全データを用い，平均値と状態遷移確率の両方を適応した．第3モデルの適応も同様の条件で行った．結果，適応しないモデルを用いた提案手法ではあまり改善できなかった話者5に対しても，認識率が大きく改善した．また，話者によっては認識率が80%まで改善した．

7. 考察

以上の結果より，本提案手法は音節強調発声に対して効果があることが確認できた．提案手法は，音節強調発声による音響的な変形に対し，既存モデル内に存在する音響的に近いモデルを流用したものである．この結果，話者適応によりモデルとの音響的な距離を小さくすることで，提案手法による認識率がさらに向上したと考えられる．

本提案手法を用いた場合，音節強調発声の傾向が強くなるにつれ，第2モデルの選択比率が増加するという結果が得られている．この結果を踏まえ，収録した言い直し発話データにおける第2モデルの選択比率を調査することにより，音節強調の強さを調べた．第3章での主観評価結果と，各話者の2回目以降の言い直し発話に対する第2モデルの選択比率の関係を図9に示す．主観評価した結果と比較すると，第2モデルの選択比率が概ねその傾向を表現していることがわかる．

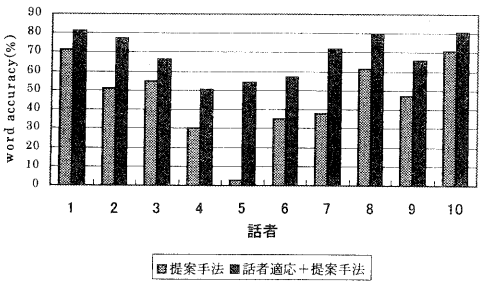

図8　話者適応による提案手法の効果

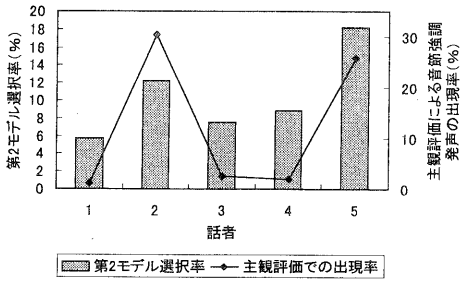

図9　音節強調発声の主観評価結果と、第2モデル選択比率の関係

8. まとめ

誤認識時の言い直し発話に含まれる，音節強調発声に頑健な音声認識手法について提案した．提案手法により，多量の音節強調発声音声の収集や，音響的な分布の拡大による音響モデルの性能劣化を招くことなく，その音響的な変形を吸収することができた．しかしながら，話者適応を行ったモデルを用いて提案手法を実施した場合でも，認識率が50%程度までしか回復しない発話様式も存在し，本提案手法でもカバーできない要因が含まれていると考えられる．

今後は，これらの要因について調査を進め，言い直し発話における認識率の更なる向上を目指すとともに，言い直し発話以外の発話様式の変動に対しても頑健な，音声認識手法について研究を進める．

謝辞　本研究の機会を与えてくださった，ATR 音声言語通信研究所 山本誠一社長に感謝する．

文　献

[1] S.Oviatte. The CHAM model of hyperarticulate adaptation during-human-computer error resolution. In Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[2] H.Soltau and A.Waibel. On the influence of hyperarticulated speech on the recognition performance. In Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[3] H.Soltau and A.Waibel. Specialized acoustic models for hyperarticulated speech. IEEE International Conference on Acoustic, Speech, and Signal Processing, Istanbule, Turkey, 2000

[4] H.Soltau and A.Waibel. Acoustic model for hyperarticulated speech. In Proceedings of the International Conference on Spoken Language Processing, Beijing,China, 2000.

[5] "ATRSPREC Home Page," http://www.itl.atr.co.jp/sprec.

[6] A.Nakamura, S.Matsunaga, T.Shimizu, M.Tonomura and Y.Sagisaka. Japanese speech databases for robust speech recognition. In Proceedings of the International Conference on Spoken Language Processing, Philadelphia, USA, 1996.

[7] 山本，匂坂，単語の方向性を考慮した多重クラス複合 N-gram 言語モデル，信学技報，SP98-102, pp,49-54, (1998-12)