発話スタイルの変動に頑健な
音響モデル構築法に関する研究

平成１７年６月

おく だ こうぞう
奥田浩三

内容梗概

計算機能力の向上や統計的手法の導入，大規模音声データベースの構築などにより，飛躍的に性能が向上した音声認識であるが，発話スタイルの違い（文章を読み上げる読み上げ音声，対話音声，講演などの独話音声といった違いや，誤認識時の言い直し発話と通常の発声との違いなど）により，認識性能が劣化するという課題がある．これらの課題は，大量の学習データを収集しモデルを構築する統計的手法とモデル適応手法だけでは十分に改善することが難しいと考えられる．このため，学習データの統計的な特徴のみからモデルを構築するのではなく，それぞれの発話スタイルが有する音響的特徴を十分に分析し，その特徴を表現できるようモデル構造や認識手法を見直す必要がある．

本論文は，より自然で自由に発話された音声を高精度に認識する，発話スタイルの変動に頑健な音響モデルを構築することを目的とし，各発話スタイルの音響的特徴の分析を通して，既存の音響モデルおよび認識手法の課題を明確にし，それぞれの発話スタイルにおける認識性能の劣化を改善する手法を研究したものである．

第１章では，本論文における研究の背景である，大語彙連続音声認識技術の現状と課題についてまとめる．

第２章では，隠れマルコフモデル（hidden Markov model；HMM）を用いた音響モデルの構築方法を中心に連続音声認識の概要をまとめるとともに，発話スタイルの変動が現状の音声認識システムのどの部分にどのような影響を与えるのかについて明らかにする．

第３章では，多数話者音声データベースを用い，地域・年齢の違いによる音響的特徴の違い，発話スタイルの違いによる音響的特徴の違いについて分析する．また，本データベースを用いて構築した音響モデルの認識性能を評価し，地域や年齢の違いが認識性能に対して影響を与えることを明らかにする．

第４章では，発話スタイルの一つである講演音声の認識について，研究結果を述べる．文章を読み上げない自発的な発話である講演音声は，文章を読み上げた音声や対話音声とは異なった音響的特徴を有しており，特に発話速度が大きく変動する傾向にある．第４章では，講演音声の発話スタイルが認識性能に与える影響を調べるとともに，ゆう度基準による分析周期・窓長の選択手法を提案し，発話速度の変動を吸収し認識性能が改善できることを明らかする．

第５章では，音響的特徴が通常の発話と大きく異なる誤認識時の言い直し発話に着目し

た研究結果を述べる．現状の認識システムでは誤認識の発生は避けられず，誤認識時にシステム利用者は言い直しを余儀なくされる．しかしながら，言い直し発話は通常の発話と比較し，音響的特徴が大きく変化するため認識性能がかえって劣化するという現象が生じる．第５章ではこの問題を解決するため，誤認識時の言い直し発話が認識性能に与える影響とその音響的特徴を分析するとともに，この音響的特徴の違いを吸収する音響モデル構築手法を提案する．

最後に第６章において本研究の総括を行うとともに，音声認識における本研究の位置付け，および重要性についてまとめ，本研究の結論とする．

目　　次

第１章

緒論

1.1 大語彙連続音声認識技術の現状

近年, 大語彙連続音声認識技術の認識性能は, 飛躍的に向上している[1]. この背景には, コンピュータ技術の劇的な進歩，効率よい探索を行うためのデコーディング手法の研究[2][3]や，探索空間の削減と文章として成り立たない認識結果の排除による認識精度の向上のための言語モデルの研究[4]に加え，統計的手法の導入や大規模音声データベースの構築・整備による音響モデルの性能向上[5]がある．音声認識をパターンマッチングの一つと捕らえた場合，音響的特徴パターンである音響モデルの性能向上が，大語彙連続音声認識の認識性能向上において重要となる．

初期の音声認識では，認識したい単語をあらかじめ読み上げた音声の音響的特徴をそのままモデルとして登録し，認識対象音声の音響的特徴に時間・周波数領域において最も近いモデルを認識結果として出力するものであった．この方法は，モデルとして登録した音声と認識対象音声の，時間・周波数的特徴が同じであると仮定する特定話者単語音声認識として有効な手法である．しかしながら同一の話者が同一の単語を発声した場合でも，その発話速度は変動しており，発話速度の違いによる時間的特徴の違いが認識性能に大きく影響するという問題を含んでいた．また，発話速度は局所的にもゆらいでいるため，時間長を線形に伸縮するだけでは発話速度の変動を十分に吸収することは困難であった．この問題を解決する手法として，モデル登録時と認識時の音声の時間的特徴が異なった場合でも，時間・周波数領域で最適な経路を探索する，動的計画法（dynamic programming ; DP）による時間軸伸張マッチング（dynamic time warping ; DTW）が導入された．時間軸伸張マッチングの導入により，認識時における発話速度の変動に対して，認識性能を向上することが可能となった[6].

特定話者単語音声認識は，認識単語数の増加に伴い，モデル登録のための利用者の負担が大きくなるとともに，利用者が代わるたびに音声を登録しなおさなければならないという課題があった．このため，利用者交代による周波数的特徴の変動が生じても十分な認識性能が得られる不特定話者音声認識技術の研究[7]が進められた.

周波数領域における変動を吸収する手法として，入力された音声の音響的特徴から直接モデルを構築するのではなく，多くの音声から統計量を抽出しモデルを構築する隠れマルコフモデル（hidden Markov model ; HMM）の導入[8]が提案された．HMM を用いた音響モデルは，その統計的な学習能力の高さと，1980 年代後半から大量の音声データベースの利用が可能となったことで，音声認識において広く用いられるようになった．より多くの話者，発話内容から HMM を用いて構築した音響モデルによって，同一話者であっても発話時刻の違いにより生じる周波数的特徴の変動や，利用者交代による周波数的特徴の変

動に対して認識性能を改善することが可能となった．また通常，HMM による音響モデルは認識対象単語を直接モデル化するのではなく，単語よりも短い音節や音素単位のサブワードモデルとして構成，これらのモデルを結合して利用するため，認識語彙数が増加した場合でも，モデルそのものを追加する必要がなくなる．このようにして，不特定話者連続音声認識においても，認識性能が飛躍的に向上することとなった．

1.2 発話スタイルの変動が認識性能に及ぼす影響

音響的特徴のモデル化に統計的手法の一つである HMM を導入したことで，音声認識の認識性能は大きく向上したが，全ての話者に対して十分な性能が得られたわけではない．利用者によってはモデル構築の際に用いた音声データの統計的な分布から外れた音響的特徴を有する話者も存在し，こういった話者の場合，HMM を用いた場合でもその認識性能は急激に劣化する．これは統計的手法が，モデル構築の際に用いた学習用音声データと実際の認識対象音声の統計的な特徴が同一であるとの仮定のもと，認識しているために生じる問題である．この問題を解決する手法として，話者適応化の線形変換（アフィン変換）行列を最ゆう推定する MLLR（Maximum Likelihood Linear Regression）[9]や最大事後確率推定法（Maximum a Posteriori Estimation ; MAP）[10][11]に代表されるモデル適応手法が提案されている．これらの手法は，既に構築されている音響モデルのモデルパラメータを，認識対象話者の音響的特徴に近づけるものである．例えば教師あり話者適応として MAP によるモデル適応を行う場合，適応対象話者が発声した既知の発話内容からその音響的特徴を抽出し，現在使用している音響モデルのモデルパラメータをその特徴空間へ移動する．モデル適応の十分な効果を得るためには，適応対象話者の音響的特徴をより多く取得する必要があったが，効果的に特徴空間の移動を行う移動ベクトル平滑化手法（vector field smoothing ; VFS[12]）などの導入により，比較的少ないデータからでも十分な適応効果を得ることが可能となった．モデル適応の導入により，既に構築されている音響モデルの統計的な分布から外れた話者の認識性能を改善することに成功している．

しかしながら，音響的特徴は話者の違いだけではなく，同一話者であっても，その発話スタイルの違いによって大きく変動する[13]．発話スタイルの違いとは，文章を読み上げた発話と読み上げでない自発的な発話の違いや，自発的な発話においても対話と独話の違いなどを指す．広く捕らえた場合，怒りや悲しみなどの感情の違いも発話スタイルの違いと考えることができる．また，システムが誤認識を発生した場合の言い直し発話なども，１回目の発話と比較し，発話スタイルが異なるものである．発話スタイルが異なった場合，発声内容が同一であっても，その音響的特徴が変化し，統計的手法では十分にその変化を

吸収することができず，認識性能が劣化することが報告されている[14].

モデル適応により発話スタイルの変動を吸収することも考えられるが，発話スタイルによっては適応データの収集が困難であること，誤認識時の言い直し発話のように音響的特徴が大きく変化する誤認識の発生そのものを検出する必要があることなどから，十分な効果を得ることは難しい．また，モデル適応は既に構築されている音響モデルのモデルパラメータを適応するものであり，HMM で構築された音響モデルのモデル構造そのものがミスマッチを起こしている入力音声に対しては，その効果を期待することができない.

1.3 本研究の概要

以上のように，発話スタイルの変動による認識性能の劣化は，大量の学習データを収集しモデルを構築する統計的手法とモデル適応手法だけでは十分に改善することが難しい．このため，学習データの統計的な特徴のみからモデルを構築するのではなく，それぞれの発話スタイルが有する音響的特徴を十分に分析し，その特徴を表現できるようモデル構造や認識手法を見直す必要がある.

本研究の目的は，より自然で自由に発話された音声を高精度に認識する，発話スタイルの変動に頑健な音響モデルを構築することである．各発話スタイルの音響的特徴の分析を通して，既存の音響モデルおよび認識手法の課題を明確にし，それぞれの発話スタイルにおける認識性能の劣化を改善することを目指している．具体的には，講演音声に代表される独話音声の認識性能の改善，及び誤認識発生時の言い直し発話のおける認識性能の劣化改善について，研究したものである.

本論文の構成を図１．１に示す．まず第２章において，HMM を用いた音響モデルの構築方法を中心に連続音声認識の概要をまとめると共に，発話スタイルの変動が現状の音声認識システムのどの部分にどのような影響を与えるのかについて述べる.

第３章では，多数話者音声データベースを用い，地域・年齢の違いによる音響的特徴の違い，発話スタイルの違いによる音響的特徴の違いについて分析した結果を述べる．また，本データベースを用いて構築した音響モデルの認識性能を評価することで，第２章で記述した現状の認識システムの問題点を，音響的な観点で検証する.

以上の結果を踏まえ第４章では，発話スタイルの一つである講演音声の認識について，研究結果を述べる．文章を読み上げない自発的な発話である講演音声は，文章を読み上げた音声や対話音声とは異なった音響的特徴を有しており，特に発話速度が大きく変動する傾向にある．第４章では，講演音声の発話スタイルが認識性能に与える影響を調べるとともに，発話速度の変動を吸収する認識手法，及びモデル構築手法についてまとめる.

第５章では，音響的特徴が通常の発話と大きく異なる誤認識時の言い直し発話に着目した研究結果を述べる．現状の認識システムでは誤認識の発生は避けられず，誤認識時にシステム利用者は言い直しを余儀なくされる．しかしながら，言い直し発話は通常の発話と比較し，音響的特徴が大きく変化するため認識性能がかえって劣化するという現象が生じる．第５章ではこの問題を解決するため，誤認識時の言い直し発話が認識性能に与える影響とその音響的特徴を分析するとともに，この音響的特徴の違いを吸収する音響モデル構築手法を提案する．

最後に第６章において本研究の総括を行うとともに，音声認識における本研究の位置付け，および重要性についてまとめる．

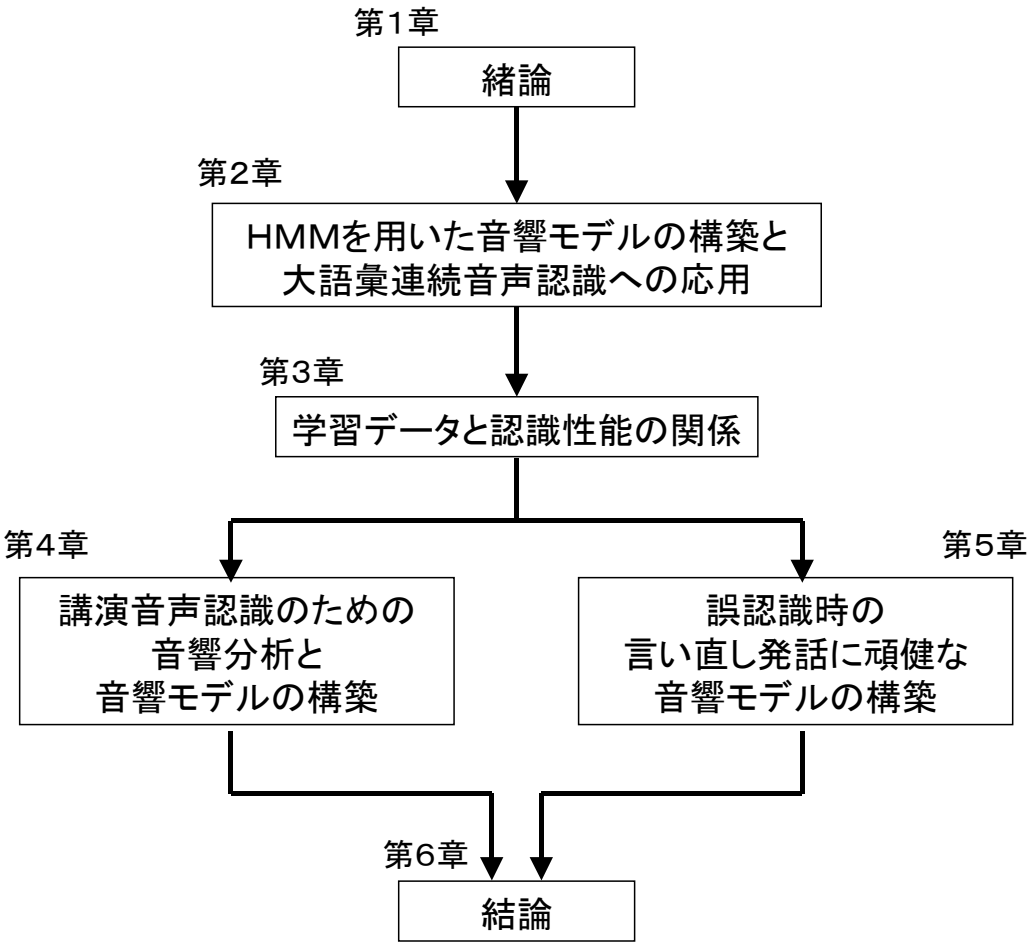


図１．１　本論文の構成

参考文献

[1]中川聖一,“サーベイ論文／音声認識研究の動向,”電子情報通信学会論文誌, vol.J83-D-II, no.2, pp.433-457, 2000.

[2] 李 晃伸, 河原達也, 堂下修司,“単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識,”電子情報通信学会論文誌, vol.J82-D-II, no.1, pp.1-9, 1999.

[3] 河原達也, 加藤一臣, 南條浩輝, 李晃伸,“話し言葉音声認識のための言語モデルとデコーダの改善,”情報処理学会研究報告, SLP36-3, 2001.

[4] 山本博史, 匂坂芳典,“接続の方向性を考慮した多重クラス複合 N-gram 言語モデル,”情報処理学会研究報告, SLP24-7, 1998.

[5] Tomoko Matsui, Masaki Naito, Yoshinori Sagisaka, Kozo Okuda and Satoshi Nakamura,“Analysis of acoustic models trained on Large-Scale speech database,”Proc. of ICSLP2000, vol.2, pp.503-506, 2000.

[6] 鹿野清宏, 中村 哲, 伊勢史郎,“音声・音情報のデジタル信号処理,”昭晃堂, 1997.

[7] 管村, 鹿野, 好田,“SPLIT 単語マルチテンプレート法による不特定話者単語音声認識,”電子情報通信学会論文誌, vol.J67-D, no.10, pp.1210-1217, 1984.

[8] Lawrence Rabiner, Biing-Hwang Juang 共著, 古井貞熙 監訳,“音声認識の基礎（上）（下),”NTT アドバンステクノロジ株式会社, 1995.

[9] M.J.F.Gales and P.C.Woodland,“Mean and variance adaptation within the MLLR framework,”Computer Speech and Language, vol.10, pp.249-264, 1996.

[10] 中川聖一, 越川 忠,“最大事後確率推定法を用いた連続出力分布型 HMM の適応化,”日本音響学会誌, vol.49, no.10, pp.721-728, 1993.

[11] J-L.Gauvain and C.H.Lee,“Maximum a posteriori estimation for multi variate Gaussian mixture observations of Markov chains,”IEEE Trans. Speech and Audio Process., vol.2, pp.291-298, 1994.

[12] 宮沢康永, 大倉計美, 嵯峨山茂樹,“全音素エルゴディク HMM を用いた教師なし話者適応,”電子情報通信学会論文誌, vol.J77-A, no.12, pp.112-119, 1994.

[13] 山本一公, 岩井直美, 中川聖一,“発話スタイルの違いが音声認識に及ぼす影響についての検討,”電子情報通信学会技術研究報告, SP99-31, 1999.

[14] S.Oviatt,“The CHAM model of hyperarticulate adaptation during human-computer error resolution,”Proc. of ICSLP'98, pp.2311-2314, 1998.

第 2 章

ＨＭＭを用いた音響モデルの構築と大語彙連続音声認識への応用

２．１　緒言

発話スタイルが変動することにより認識性能が劣化する現象は，現状の音声認識手法では吸収することができない要因をその発話スタイルが含んでいるためと考えられる．発話スタイルの変動に頑健な音響モデルの構築方法を検討するには，これらの要因について調査・分析を行う必要があるが，そのためにはまず統計的手法の一つである隠れマルコフモデル（hidden Markov model ; HMM）を用いた音響モデルの構築方法[1]や，大語彙連続音声認識への適用方法について理解する必要がある．本章では， HMM を用いた音響モデルの構築方法を中心に連続音声認識の概要をまとめると共に，発話スタイルの変動が現状の音声認識システムのどの部分にどのように影響を与えるのかについて述べる．

まず２．２節において，入力音声から音響的特徴を抽出するための，音響分析と特徴量抽出についてまとめる．次に２．３節において，HMM を用いた音響モデルの構築方法として，音声のモデル化，および HMM の学習方法をまとめる．２．４節では，HMM を用いた音声認識の具体的な方法として，大語彙連続音声認識手法について述べる．最後に２．５節において，発話スタイルの変動が認識システムに与える影響を，認識システムの全体構成を踏まえながらまとめる．

２．２　音響分析と特徴量抽出

音声認識においてはまず，入力音声から音響的特徴を抽出するため，音響分析を行う．入力音声は，音声の特徴が多く含まれている８kHz 以下の周波数が利用できるよう，シャノンの定理より 16kHz でサンプリングされる場合が多い．図２．１に短時間スペクトルの例を示す．音声は声帯の振動により生成された音源が，声道を通過する過程で共振することにより発声される．短時間スペクトルには，声帯の振動による周期的な微細構造（ハーモニックス構造）と，声道の共振によるスペクトルの大局的な概形（スペクトル包絡）が含まれており，これらの特徴を抽出することが音声の分析においては重要となる．

スペクトル分析方法はいくつか存在するが，離散フーリエ変換（discrete Fourier transform ; DFT）を用いた周波数分析が一般的に広く用いられており，入力音声に対し比較的短時間の時間窓を掛け，一定周期で周波数分析を行う（この処理単位をフレームと呼ぶ）．時間窓に方形窓を用いた場合，窓の両端部分における不連続性から高調波成分が現れるため，図２．２に示すような Hamming 窓や Hanning 窓が多く用いられる．多くの認識システムでは，認識性能と計算機コストのバランスから，分析周期は８msec や 10msec の一定間隔で，分析窓長は分析周期の倍にあたる，16msec や 20msec の長さとなっている．

音声は，声帯が振動することにより生成された音源が声道を通過する過程で共振することで発せられる．このことから音声波形 x は，声帯の振動による音源 g と声道のインパルス応答 v の畳み込みとして表現することができる．

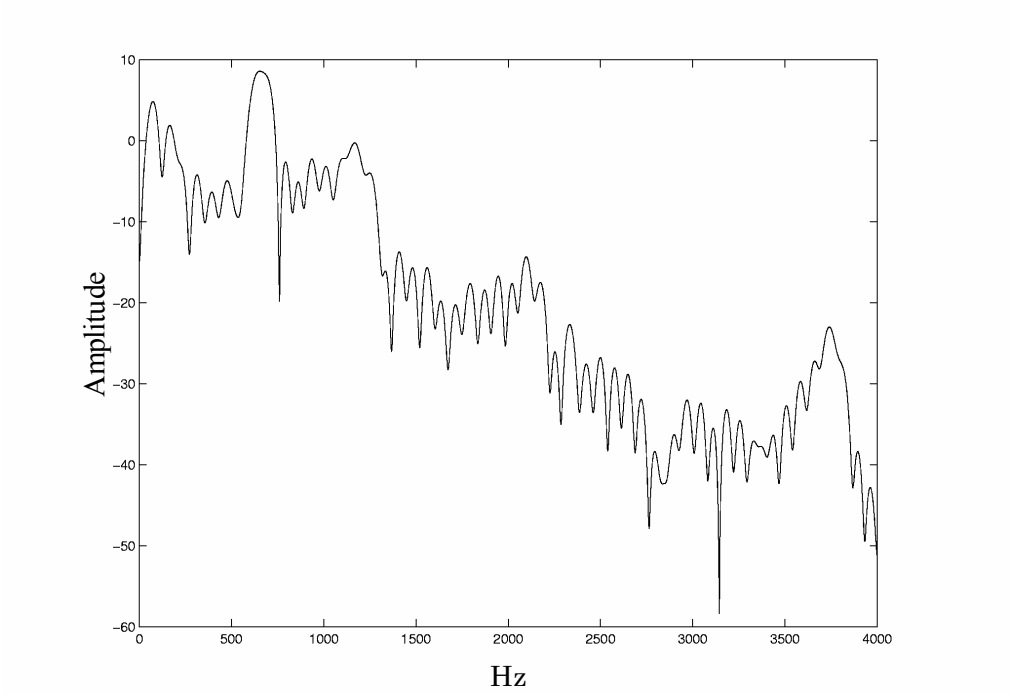


図２．１　母音“あ”の短時間スペクトルの例

$$x_n = \sum_{k=-\infty}^{\infty} g_k \cdot v_{n-k} = g_n * v_n \tag{2.1}$$

音声波形x_nのフーリエ変換は,

$$X(\omega) = G(\omega) \cdot V(\omega) \tag{2.2}$$

となり，そのパワースペクトル $S(\omega)$も

$$S(\omega) = |X(\omega)|^2 = |G(\omega)|^2 \cdot |V(\omega)|^2 \tag{2.3}$$

となる．この両辺の対数をとることで

$$\log|X(\omega)| = \log|G(\omega)| + \log|V(\omega)| \tag{2.4}$$

と表現することができる．このようにスペクトル強度を対数とすることで，音声のスペクトルを $G(\omega)$と $V(\omega)$の対数の和，つまり声帯の振動によるスペクトルの微細構造と，声道

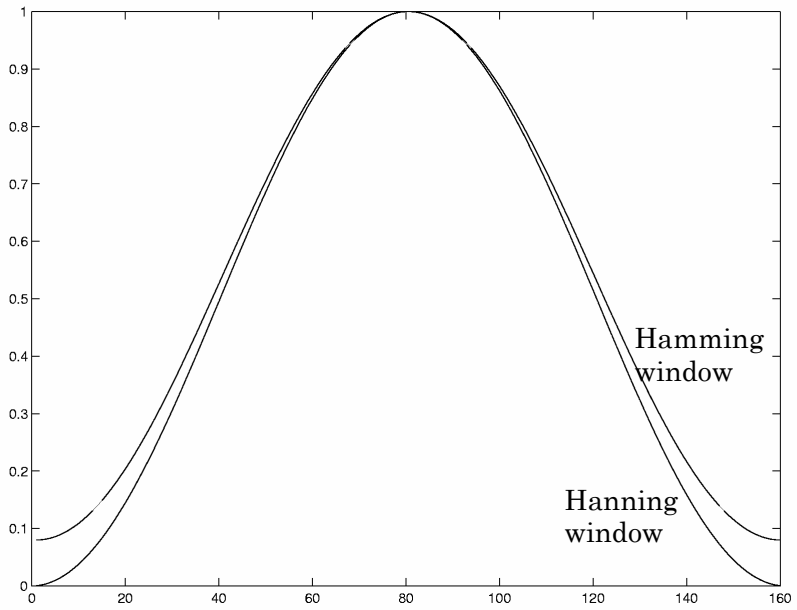


図２．２　短時間スペクトル分析のための時間窓

における共振に対応するスペクトル包絡との和の形として得ることができる．

こうして得られたスペクトルを逆 DFT により時間領域に変換することで Cepstrum 係数が定義される．図２．３に Cepstrum 係数の例を示す．Cepstrum 係数では，スペクトル包絡は低次に，スペクトルの微細構造は高次に集中する．一般に音声認識では，人間の声道フィルタの特徴が現れる共振周波数（フォルマント）の位置や高さの情報を利用しており，声の高さであるピッチ，つまり声帯振動の情報は用いていない．そこで，Cepstrum 係数を適当な次数で打ち切ることで，ピッチ情報を削除する．ピッチ情報を削除した Cepstrum 係数を再び DFT することで，図２．４に示すように，ピッチとその高調波成分であるハーモニックス構造が取り除かれ，スペクトル包絡のみが残されたパワースペクトラムを得ることができる．音声認識では，Cepstrum 係数を音響特徴量として用いる場合が多い．また人間の聴感特性に合うよう，周波数軸を Mel スケールに周波数ワーピングする[2]ことが多い. Mel スケールに周波数ワーピングを行った Cepstrum 係数を MFCC（Mel Frequency Cepstrum Coefficient）と呼ぶ.

このようにして抽出した短時間スペクトルを用いて音声の特徴パラメータを構成するが，スペクトルの動きにも重要な情報が含まれていることが知られている．音声認識においてはスペクトルの動きによる動的な特徴，つまり時間的な変化の情報を利用することで認識性能が向上することが報告されており，現状の音声認識システムでは積極的に利用されている．一般には，前後２フレーム程度の MFCC より線形回帰係数を求めることで，動的な

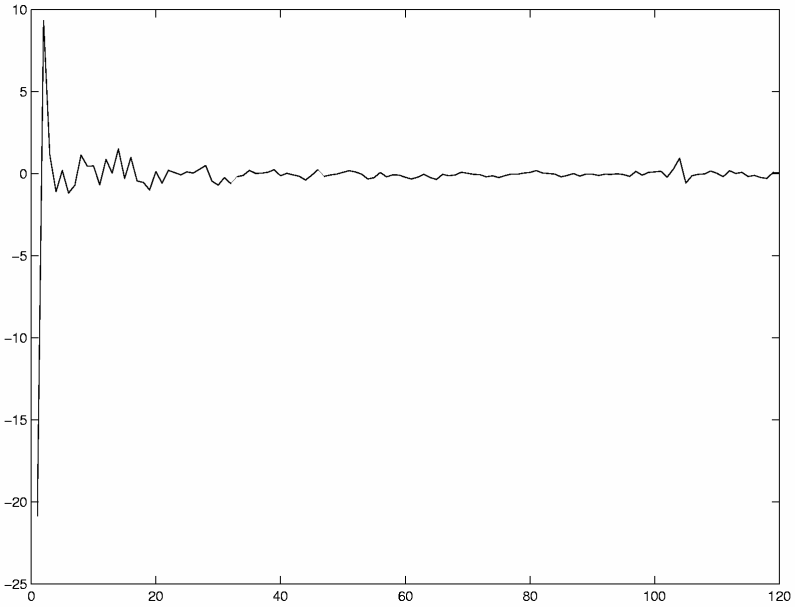


図２．３　Cepstrum 係数の例

特徴である ΔMFCC を求める．この際，対数パワーの変化量も特徴パラメータとして用いることが多い．このようにして，MFCC＋ΔMFCC＋Δlog power を結合したベクトルを特徴パラメータとして用いている．

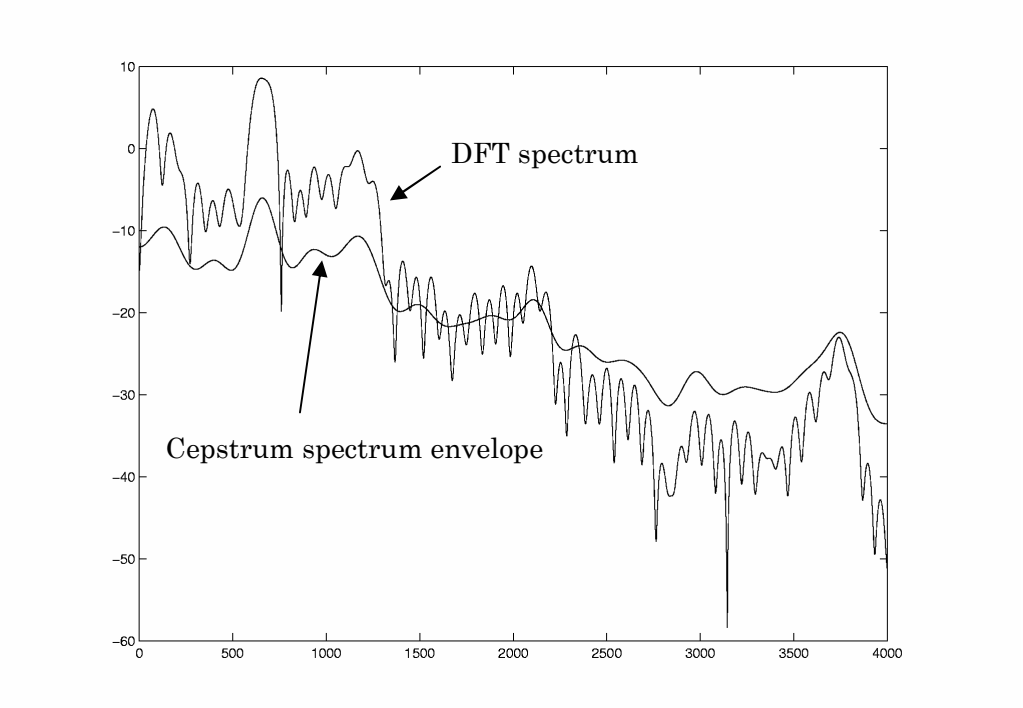


図２．４　DFT スペクトルとケプストラムスペクトル包絡の比較

２．３　ＨＭＭを用いた音響モデルの構築

２．３．１　ＨＭＭによる音声のモデル化

音声認識もパターンマッチングの一つと捕らえることができ，入力音声と認識システム内に保持されているパターンとの間でマッチングを行い，最も適したパターンを認識結果として出力する．最も適したパターンの探索方法としては，入力音声の特徴パラメータ列と，モデル化されている特徴パラメータ列のスペクトルマッチング尺度計算などがあり，モデル登録時の音声と認識時の音声の特徴が同一であると仮定できる特定話者音声認識で良好な結果を得ることができる．しかしながら音声認識を不特定の話者に拡張した場合や，特定話者であっても発話時刻の違いにより音響的特徴が変動する場合などは，十分な認識性能を得ることができない．このような場合，認識単語ごとにモデルを 1 つ用意するのではなく，複数個のモデルを用意しなければならなくなるが，認識単語数が増加するにつれ，莫大な量のモデルを保持する必要が生じる．そこで，認識単語ごとに複数のモデルを準備するのではなく，大量の学習データから得た統計量をそのままモデルとして利用する手法が提案されている．代表的なものが HMM である．HMM は特徴パラメータの分布の表現方法により，離散分布 HMM と連続分布 HMM に大きく分類される．離散分布 HMM は，出力ベクトルをベクトル量子化により離散化したものであり，連続分布 HMM は統計量である平均値と分散値をモデルパラメータとして音響モデルを構築したものである．連続分布 HMM には，出力ベクトルの確率分布を単一の分布で近似する単一分布型 HMM と，複数の分布の和で近似する混合分布型 HMM がある．また，ベクトルの各要素の平均値と分散値のみをモデル化し，共分散の成分を零と仮定する対角行列（diagonal matrix）のものと，要素間の共分散値までをモデル化した全角行列（full matrix）のものがある．一般に，連続単一分布 HMM では全角行列が用いられ，連続混合分布 HMM では対角行列が用いられる．

音声認識を統計的に捕らえ，入力音声のパターンを I 個のフレームの時系列として表した場合，入力音声 X は，

$$X = x_1, x_2, ..., x_I \tag{2.5}$$

となる．音声認識は，入力音声 X を観測し，最もよくマッチングする単語列，

$$W = w_1, w_2, ..., w_N \tag{2.6}$$

を探索する問題と考えることができる．つまり，

$$P(W \mid X) = P(w_1, w_2, ..., w_N \mid x_1, x_2, ..., x_I) \tag{2.7}$$

を最大にする単語列 W を探索すればよい．ここで，$P(W|X)$はベイズの定理により，

$$P(W \mid X) = \frac{P(X \mid W) \cdot P(W)}{P(X)} \tag{2.8}$$

と表現することができる．$P(X)$は，入力音声そのものの生起確率のため，探索においては一定である．つまり，音声認識では，

$$P(X \mid W) \cdot P(W) \tag{2.9}$$

が最大となる単語列 W を探索することとなる．認識システムにおいては，言語データベースから学習することにより，単語列 W の生起確率をモデル化した統計的言語モデルを用いている．このことから $P(W)$は単語列の事前確率を，

$$P(W) = P(w_1, w_2, ..., w_N) = P(w_1) \prod_{i=2}^{N} P(w_i \mid w_{i-1} \cdots w_1) \tag{2.10}$$

と定義できる．
後は，$P(X|W)$が計算できれば $P(X|W) \cdot P(W)$が最大となる単語列 W を探索することが可能となる．$P(X|W)$は，

$$P(X \mid W) = P(x_1, x_2, ..., x_I \mid w_1, w_2, ..., w_N) \tag{2.11}$$

となり，認識の音声パターンの部分系列

$$x_i, x_{i+1}, ..., x_{i+n} \tag{2.12}$$

と単語w_kに対して，

$$P(x_i, x_{i+1}, ..., x_{i+n} \mid w_k) \tag{2.13}$$

が計算できればよい．この計算のためのモデルを音響モデルと呼び，そのモデルとしてはサブワード単位の Left-to-right 型 HMM が多く用いられている．このモデルは図２．５に示すように，１つのサブワードを複数の状態で表現するものである．Left-to-right 型 HMM では，入力端と出力端は固定であり，定常信号源の連鎖として表現した確率モデルとなっている．確率モデルのため，時間軸伸張マッチング（dynamic time warping；DTW）のように登録されているモデルとの距離尺度で単語を探索するのではなく，入力音声の特徴パラメータに対してモデルが出力する確率を用いて探索することができ，入力音声の揺らぎを吸収することが可能となる．各状態は，ガウス分布に代表される確率密度関数を用いてモデル化されている．通常，確率は確率密度関数の面積から計算されるが，認識においては入力音声の特徴パラメータにおける確率密度関数の値を，一種の確率であるゆう度として出力する．

単語 HMM はサブワード HMM を連結することで実現することができるため，例えばサブワードに音素を用いた場合は，各音素に対する音素 HMM を構築すれば，全ての単語に対する HMM を構築することが可能となる．しかしながら言語的に同じ音素であっても，前後につながる音素によって音響的特徴が変化することがある（調音結合）．例えば，音素/k/では，後ろに続く音素が/a/の場合と/i/の場合で音響的な特徴が異なることが知られている．後続する音素が/a/の場合は破裂音になり，後続する音素が/i/の場合は摩擦性の音になる．このような変化に対してより精密なモデルを構築することを目的として，音素の前後環境

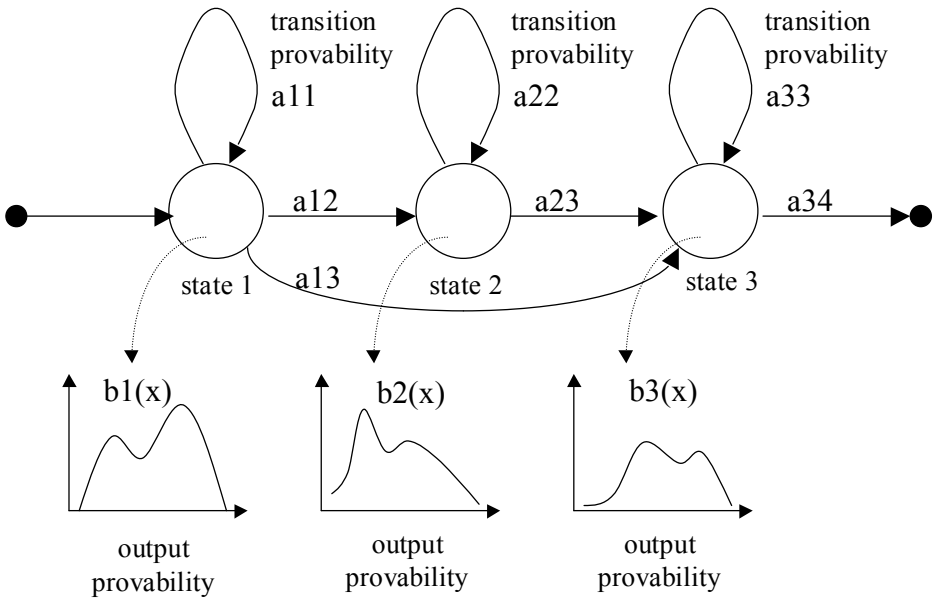


図２．５　HMM の構造

を考慮したモデリングも行われている．先行する 1 音素，または後続する 1 音素の影響を考慮して構築したモデルを biphone 音響モデルと呼び, それぞれ先行音素環境依存 biphone 音響モデル，後続音素環境依存 biphone 音響モデルと呼ぶ．また，先行する 1 音素と後続する 1 音素の両方の影響を考慮して構築したモデルを triphone 音響モデルと呼ぶ.

２．３．２　HMMの学習方法

HMM を用いた音響モデルの学習は，大別するとモデル構造の決定と，モデルパラメータ学習の２つのステップからなる．

まず，モデル構造の決定について述べる．サブワードとして音素を用いた場合，その音素数がモデルの複雑さに影響する．例えば国際電気通信基礎技術研究所（ATR）では 26 音素，情報処理推進機構（IPA）の研究テーマ「日本語ディクテーション基本ソフトウェアの開発」[3]では 43 音素となっているが，音響モデルを triphone で構築した場合，ATR 音素セットでは 26×26×26=17576 個，IPA 音素セットでは 41×41×41=68921 個の論理的なモデルを必要とする．それぞれのモデルが Left-to-right 型 HMM で複数の状態から成るため，総状態数で言えば非常に多くの状態を有するモデルとなる．また，いかに多くの学習データを収集したとしても，十分な統計量が得られない状態も存在し，かえって認識性能が劣化する恐れがある．そこで HMM では，音響的特徴が類似した triphone をグループ化することによって，モデルの数を削減する場合が多い．例えば先行音素が a，後続音素が k の該当音素 u（HTK[4]では，a-u+k と表記される）と，先行音素が i，後続音素が k の該当音素 u(HTK 表記の i-u+k)に関して見た場合，先行音素は異なるが，後続音素と該当音素が同じなため，モデル後半の状態を共有することで効率よくモデルを構築することが可能となる．このようにして音素環境クラスタリングを行うが，問題は共有する状態の決定，つまりモデル構造（トポロジー）の決定方法である．状態共有の決定方法は大別して，ボトムアップ方式とトップダウン方式がある．ボトムアップ方式は，学習データから得ることのできる全ての triphone を作成し, 作成された triphone からよく似たモデルをマージしていくものである[6]．トップダウン方式は中心音素が同じ triphone を，前後の音素環境の違いによる影響が大きい要因から分割していくものである．ゆう度最大化基準により時間方向・音素環境方向に分割していく最ゆう逐次状態分割法(Maximum Likelihood Successive state splitting ; ML-SSS)[5]や，学習データに従い音素環境をツリー状に成長させる生成木に基づいた状態分割法[7]がある．音素セットや学習データ量にもよるが，トップダウン方式はボトムアップ方式に比較し，学習データに出現しなかった triphone をとり扱うことができる長所がある．

このようにして決定されたモデル構造の HMM に対して，次にモデルパラメータの学習を行う．モデルパラメータの学習方法としては，最ゆう推定(maximum likelihood estimation)[1]が広く用いられている．通常の HMM における最ゆう推定では図２．６に示すように，与えられた学習データの集合 Y に対して，状態 i から状態 j への遷移を直接数え上げることで遷移確率を求めることができる．生起確率を計算する方法としては，初期状態から前向きに確率を計算するフォワードアルゴリズムと，最終状態から後ろ向きに確率を計算するバックワードアルゴリズムがあり，これらを組み合わせたフォワード・バックワードアルゴリズムが広く用いられている．しかしながら音声データを用いた学習を行う場合，学習データの特徴パラメータ列のみが観測され，状態間の遷移を直接観測することができない．そこで一旦初期モデルを仮定し，そのモデルを用いて算出した状態遷移回数から遷移確率，及び出力確率を最ゆう推定する．次に推定したモデルパラメータを初期モデルとしてこの処理を繰り返す．この２段階の過程を収束するまで繰り返すことでモデルパラメータを推定するアルゴリズムは一般に EM アルゴリズム（expectation-maximization algorithm）[1]と呼ばれる．

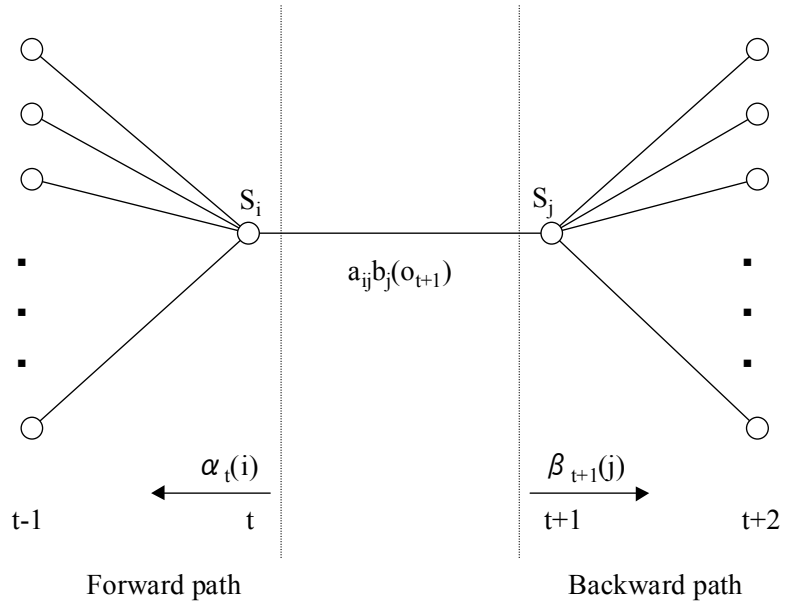


図２．６　状態 i と状態 j におけるフォワードパスとバックワードパス

２．４　大語彙連続音声の認識

大語彙連続音声認識は，式(2.9)で示された確率を最大とする単語列wを探索する問題ととらえることができることは既に述べた．音響モデルを用いることで，式(2.11)は算出することができるため，式(2.10)を算出することができれば，式(2.9)を計算することが可能となる．式(2.10)を算出するための確率モデルとしては，多量のテキストデータから単語列の生起確率を，数え上げによってモデル化することが可能である．ただし，全ての単語列が現れるテキストデータを準備することは困難なことから，統計的パラメータの数を減らすため，マルコフ過程を導入する．単語w_iが単独で生起する確率は，テキストデータから容易に求めることができる．単語w_{i-1}の次に単語w_iが生起する確率を$P(w_i|w_{i-1})$とすると，単語列Wが生起する確率は単純マルコフ過程として，

$$P(w) \cong \prod_{i=1}^{n} P(w_i \mid w_{i-1}) \tag{2.14}$$

と近似することができる．同様に，２重マルコフ過程を考えた場合，

$$P(w) \cong \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \tag{2.15}$$

と近似することができる．単語w_{i-1}の後に単語w_iが生起する確率や，単語列w_{i-1}，w_{i-2}の後に単語w_iが生起する確率は，テキストデータから数え上げによって求めることができるため，効率よくモデル化することが可能となる．言語モデルでは，単純マルコフ過程をバイグラムと呼び，２重マルコフ過程をトライグラムと呼ぶ．しかしながら，多くのテキストデータを準備しても，論理的に連接が可能な全てのバイグラム，トライグラムを算出することはできない．そこで，トライグラムを補間する手法として，ユニグラム（単語そのものが生起する確率)，バイグラムとの線形補間がある．

$$\hat{P}(w_i \mid w_{i-2}, w_{i-1}) = \lambda_1 P(w_i \mid w_{i-2}, w_{i-1}) + \lambda_2 P(w_i \mid w_{i-1}) + \lambda_3 P(w_i) \tag{2.16}$$

線形補間パラメータλ_1，λ_2，λ_3は，その和が１に成るよう，与えられる．このようにして構築した言語モデルと音響モデルを利用し，入力音声に対して式(2.9)が最大となる単語列を探索する．

実際の探索は図２．７に示すような枠組みとなる．図２．７におけるデコーダ（認識エンジンとも呼ばれる）は，音響モデルによる確率と言語モデルによる確率の積（対数スケ

ールの場合は和）が最大となる単語列を探索する．ただし，音響モデルが出力する確率と言語モデルが出力する確率のダイナミックレンジが異なるため，一般には言語重みと呼ばれる重み係数を掛けた状態で確率の積が算出される．

大語彙連続音声認識におけるデコーダは，膨大な探索空間を効率よく探索するため，近似，評価値算出，枝刈りを繰り返すことで認識結果を出力する[8]．枝刈りは，探索の過程において有望でない仮説を評価対象からはずす操作を行うものである．評価値算出における近似精度や，音響モデル・言語モデルの精度が高いほど，正しい枝刈りを行うことができるが，その分計算機コストが多く必要となる．こういった問題を回避するため，探索を複数回に分割するマルチパスデコーディングなどもよく用いられる[9]．例えば２パスデコーダなどでは，ある程度の精度の音響モデル・言語モデルで一旦探索を行い，その結果をもとに高精度のモデルで再び探索を行うものである．一般には，最初の探索では処理の簡便なバイグラム言語モデルを用い，２度目の探索でより制約の強いトライグラム言語モデルを用いる場合や，最初の探索と２度目の探索で言語重みを変更する，といった手法が用いられる．

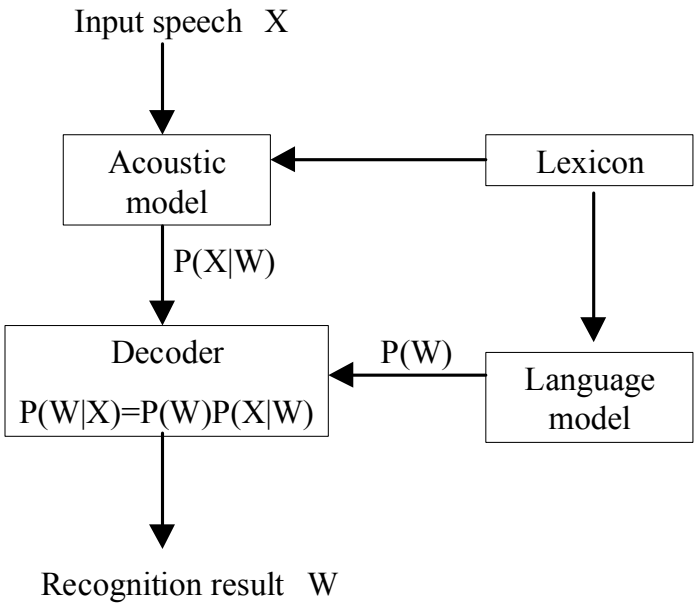


図２．７　確率的単語認識の枠組み

２．５　発話スタイルの変動が認識システムに与える影響

図２．７で示した確率的単語認識の枠組みを，実際に音声認識システムの構築から入力音声を認識するまでの流れで記述したものを図２．８に示す．一般的な認識システムでは，事前に収集された音声データ，テキスト情報を用いて，音響モデル，言語モデルを学習する．この際，実際の認識対象となるドメインで収集された音声データ，テキスト情報を利用することが，高い認識性能を実現するためには必要となる．このようにして構築された音響モデル，言語モデルを用い，実際の入力音声に対して確率的単語認識を行い，認識結果を出力する．認識性能は一般に，次に定義される単語正解精度，もしくは単語誤り率で評価される．

$$\text{単語正解精度} = \frac{N - I - D - S}{N} \times 100 \quad (\%) \tag{2.17}$$

N：正解の総単語数
I：挿入誤り数
D：脱落誤り数
S：置換誤り数

$$\text{単語誤り率} = 100 - \text{単語正解精度} \quad (\%) \tag{2.18}$$

音響的に見た場合，図２．８からわかるように，学習データとして利用される音声データと，実際に認識対象となる入力音声との間で，統計的な性質が同一であるとの仮定のもと，探索が実行される．また，言語モデルに関しても，認識対象となるドメインで用いられる単語の生起確率などを学習したものである．このため，統計的性質が同一であるという仮定が成り立たない場合，認識性能は急激に劣化する．自然な発話では，言語的には間投詞やフィラー（「えー」「あー」等）の挿入，言い直しなどの言語現象が発生し，音響的には発音の怠けによる発声変形や発話速度の局所的な変化などによる音響モデルとのミスマッチが多く発生する．特に音響的に見た場合，図２．８の音響分析，特徴ベクトル抽出，HMM モデル構造，HMM モデルパラメータが影響を受けることとなり，認識性能が急激に劣化する原因となる．このことから，音響的特徴の変動や発話速度の変化を伴う発話スタイルの変動は，既に構築されている音響モデルとのミスマッチを引き起こすとともに，通常の発話速度に最適化された音響分析部やデコーダでは十分な性能が得られない原因の一つとなっている[10]．テキストの読み上げ音声と比較し，自然な発話（人間同士の対話や，講演発

表などの独話）の認識が難しいのはこのためである．

これらの問題を解決するためには，より多くの話者から自然に発話された音声データを収集し，その音響的特徴について調査することが必要となる．ATR における多数話者音声データベースの構築[11]や，「話し言葉工学」プロジェクト[12]による CSJ（Corpus of Spontaneous Japanese）の整備などはこのためであり，既に自然発話音声認識を対象とした研究が進められている．実際，音響モデル構築の際に自然発話音声データを利用することで，自然発話音声の認識性能がある程度改善されることも報告[13]されている．

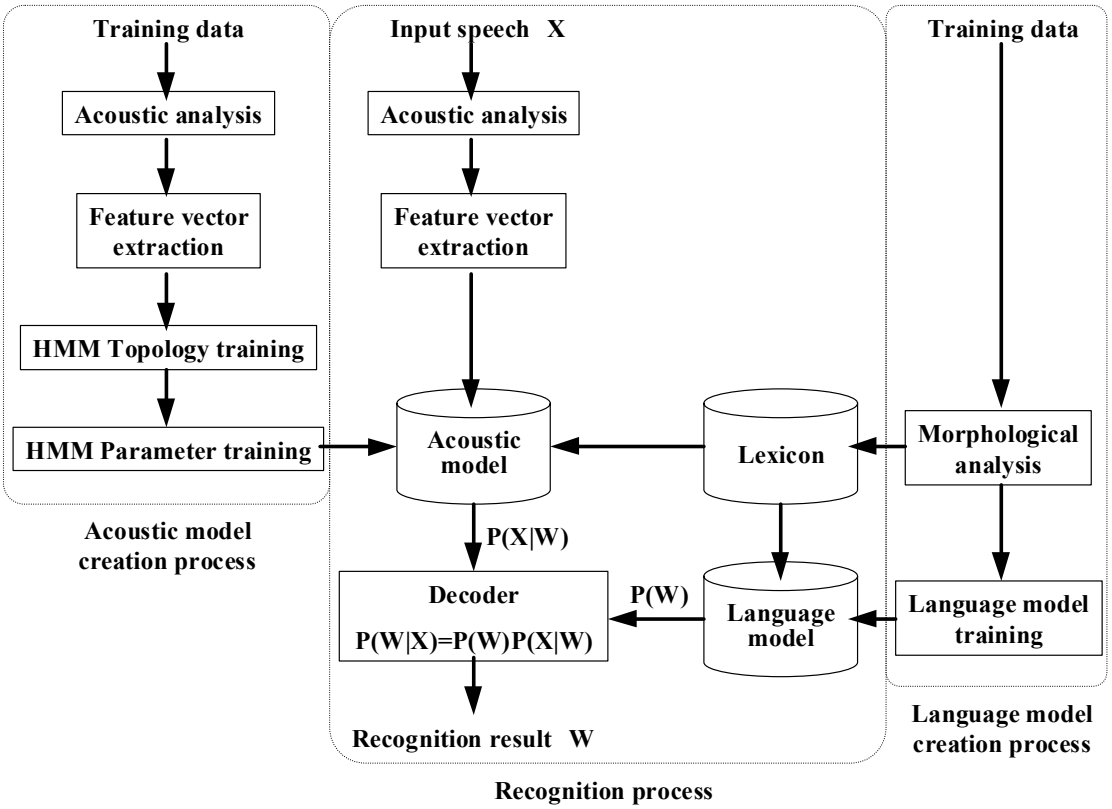


図２．８　大語彙連続音声認識における処理の流れ

２．６　結言

HMM を用いた音響モデルの構築方法と大語彙連続音声認識への応用，および発話スタイルの変動が認識システムに与える影響についてまとめた．多くの学習データから音声の統計的特徴を抽出しモデル化する統計的手法の導入により，発話時刻の違いや利用者交代などによる音響的特徴の変動に対して頑健な音響モデルを構築することが可能となった．しかしながら，多くの音声データベースに収録されている学習データは，その発話スタイルが限られており，認識対象音声の発話スタイルが異なる場合，認識性能が劣化する．これは入力音声の時間・周波数的特徴が，学習データの統計的特徴と異なることにより，音響モデルのモデル構造やモデルパラメータにミスマッチが生じるためである．音響モデルの性能が劣化した場合，近似と枝刈りを基本としたデコーディングにおける枝刈り精度の劣化などに大きく影響し，結果として認識性能が急激に劣化する．

HMM を用いた音響モデルでこの問題を解決するためには，全ての発話スタイルを学習データに含む必要があり，非現実的である．学習データの量を増やすだけではなく，認識対象となる発話スタイルの音響的特徴を分析することにより，発話スタイルの変動を吸収する手法を研究する必要がある．第３章では，ATR で構築された多数話者音声データベースの分析を通して，発話スタイルの変動が認識性能に与える影響について，詳細に述べる．

参考文献

[1] Lawrence Rabiner, Biing-Hwang Juang　共著, 古井貞熙　監訳, “音声認識の基礎（上）（下),” NTT アドバンステクノロジ株式会社, 1995.

[2] 鹿野清宏, 中村 哲, 伊勢史郎, “音声・音情報のデジタル信号処理,” 昭晃堂, 1997.

[3] 河原達也, 李　晃伸, 小林哲則, 武田一哉, 峰松信明, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田　篤, 宇津呂武仁, 鹿野清宏, “日本語ディクテーション基本ソフトウェア（97 年度版）,” 日本音響学会誌, vol.55, no.3, pp.175-180, 1999.

[4] Entropic Ltd. The HTK Book (for HTK Version 2.2), 1999.

[5] M.Ostendorf and H.Singer, “HMM topology design using maximum likelihood successive state splitting,” Computer Speech and Language, vol.11, pp.17-41, 1997.

[6] K.F.Lee, “Automatic Speech Recognition: The development of the SPHINX system,” Kluwer Academic Publishers, 1989.

[7] 堀　貴明, 加藤正治, 伊藤彰則, 好田政紀, “音素決定木に基づく逐次状態分割による HM-Net の検討,” 電子情報通信学会論文誌, vol.J80-D-II, no.10, pp.2645-2654, 1997.

[8] 河原達也, “探索アルゴリズムーA*探索を中心にー,” 電子情報通信学会技術研究報告, SP92-36, 1992.

[9] A.Lee, T.Kawahara and S.Doshita, “An Efficient Two-pass Search Algorithm using Word Trellis Index,” Proc. of ICSLP'98, pp.1831-1834, 1998.

[10] 奥田浩三, 川原達也, 中村 哲, “ゆう度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデル構築,” 電子情報通信学会論文誌, vol.J86-D-II, no.2, pp.204-211, 2003.

[11] 奥田浩三, 松井知子, 内藤正樹, 匂坂芳典, 中村 哲, “大規模日本語音声データベースの構築と評価,” 日本音響学会誌, vol.58, no.9, pp.569-578, 2002.

[12] 古井貞熙, 前川喜久雄, 井佐原 均, “科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく『話し言葉工学』の構築,” 日本音響学会誌, vol.56, no.11, pp.752-755, 2000.

[13] T.Matsui, M.Naito, Y.Sagisaka, K.Okuda and S.Nakamura, “Analysis of acoustic models trained on a large-scale Japanese speech database,” Proc. of ICSLP2000, vol.2, pp.503-506, 2000.

第 3 章

学習データと認識性能の関係

３．１　緒言

本章は，論文“大規模日本語音声データベースの構築と評価”[1]に関するものである．

音声研究における音声データベースの重要性は古くから認識されており，これまでも単語や文音声に関するデータベースが作成されてきた[2]～[5]．一方，連続音声認識技術の発展に伴い，テキストを読み上げない自発的な音声（以降，自由発話音声と呼ぶ）に関する音声認識への関心も高まっている．このような背景の下，国際電気通信基礎技術研究所（以下，ATR）では各種研究目的に共同利用可能な，連続発話音声データベースの構築が進められてきた．

ATR では，1986 年より研究用音声データベースの構築を開始し，1987 年に研究用 ATR 日本語音声データベースを公開した[6]～[8]．このデータベースは，(1)単語音声データベース，(2)連続音声データベース，(3)不特定話者用音声データベース，(4)英語音声データベースの４種類のデータベースから構成されている．単語音声データベースは言語処理との整合性を考慮し，名詞だけではなく通常の文章や会話に現れる語彙も意識した単語の選定により抽出された 8,500 単語を，男女各 10 名（アナウンサ各８名，ナレータ各２名）が読み上げたものである．連続音声データベースは，研究用途での利用を考慮し音韻バランスのとれた 503 文を新聞，雑誌などから選択，男性６名（アナウンサ３名，ナレータ３名），女性４名（アナウンサ１名，ナレータ３名）が読み上げている．不特定話者用音声データベースは，不特定話者連続音声認識研究における利用を目的とし，最重要語 520 単語，音韻バランスを考慮した数字, 連続発声, 文節発声を計 240 名から収録したものとなっている．

連続音声認識技術の発展に伴い，ATR では自由発話音声の研究に利用可能な研究用自然発話音声データベースの構築を開始した．そのひとつとして「自然発話音声・言語データベース」が公開されている[9]．このデータベースは旅行会話をタスクとし，ホテルのフロントと顧客の電話を通じた非対面の対話を想定して構築されたものであり，日本語での対話 892 会話と，日本語と英語での対話 681 会話が収録されている．日本語での対話においては，顧客側は音声の収録に関して経験のない，もしくは少ない発話者が，ホテル側には発声に対して経験を持つナレータや音声研究者が発声している．

隠れマルコフモデル（hidden Markov model ; HMM）に代表される統計的手法を用いた不特定話者音声認識技術の研究を考えた場合，地域や年齢に対して偏りの少ない評価データやモデル学習用データが望まれる．しかしながら，従来の自由発話音声を対象とした音声データベースについて見ると，その多くは首都圏や大都市などの一部の地域で，年齢層が 20～40 代の話者を対象に収集したものが多い．一方で，話者の多様性を主眼においた日本語データベースも収集されているが，この場合，収集された音声は予め指定したテキ

ストを読み上げた朗読音声が多い[10]. 発話スタイルの違いにより音声の音響的な特徴が変化し，認識性能に影響を与えることが報告されている[11][12]ことから，これらのデータベースを用いて自由発話音声を対象とした認識技術を検討することは難しい.

そこで ATR では，さらに多くの話者を含む大規模音声データベースの構築を目指し，テキストを読み上げない自由発話音声，テキストを読み上げた朗読発話の両方を含む，3,700 人規模の多数話者音声データベースを構築した[13]. 本章では，このデータベースの概要について述べると共に，地域・年齢の違いによる音響的特徴の違い，地域・年齢の違いが認識性能に与える影響，学習データ量と認識性能の関係について論じる.

まず３．２節において，ATR 研究用多数話者音声データベースの概要について，収録条件やデータベースの構成，収録方法について述べる．次に３．３節において，本データベースに収録されている音声データの音響的特徴の解析として，対話と朗読の違いによる音響的特徴の違い，年齢・地域の違いによる音響的特徴の違いについて述べる．３．４節では認識実験を通じて年齢や地域の違いが認識システムに与える影響を明らかにし，最後に３．５節で，実際に本データベースを用いて構築した音響モデルの認識性能を評価することで，学習データ量が認識性能に与える影響などについて述べる.

３．２　ＡＴＲ研究用多数話者音声データベースの概要

３．２．１　収録条件

音声収録は，遮音室等の環境雑音の影響が極めて少ない録音専用スタジオで行われている．マイクロフォンはコンデンサーマイク（SONY C-355）を使用し，話者の音声はマイクロフォンを通じてサンプリング周波数 48kHz，16 ビットの量子化精度で DAT へと録音されている．

３．２．２　データベースの構成

幅広い地域の話者の音声を収録するため，図３．１に示す全国 18 都市において収録が行われた．このデータベースは， 3,771 名（女性 2,390 名，男性 1,381 名）の話者の音声を含んでいる．

話者については，職業，音声収録の経験の有無に関する条件はなく，多くの発話者は音声収録の経験のない，もしくは少ない話者となっている．図３．２に地域別の話者数，図３．３に年代別の話者数を示す．また，表３．１に，各地域別の平均年齢を示す．図３．２において，地域は現在居住している地域を指し，継続的に居住と記されている話者は，小学校，中学校，高校も同地域であった話者を示している．地域別では近畿，中部，関東に偏りがあるものの，北海道から九州までカバーしたデータベースとなっている．年代別

図３．１　音声データベースの収録都市

の分布に関しては，10 代から 60 代（14 歳から 65 歳）と広がりは大きいものの，話者の集めやすさから，やや 20 代に集中している．

各話者は表３．２に示す３種類の音声の発声を行っている．これらの音声は，収録後 16kHz にダウンサンプリングした上で，専門家により発話単位での音声波形ファイルへの切り出し，音声区間の時刻情報の付与，音素書き起こしが行われている[7]．加えて，模擬対話音声については，日本語仮名漢字まじりの書き起こし，一部については形態素情報が付与されている．

これらのデータ種別毎のデータ量を表３．３に示す．既存の対話音声データベースの中で最大規模のものとしては，米国 LDC が扱う英語音声を対象とした ATIS や Switchboard コーパスが挙げられる．これらのデータベースの規模は，ATIS が話者約 600 名，約 25,000 発声（マイク音声），Switchboard が話者 543 名，2,400 対話（電話音声）を含む．日本語を対象とした対話音声データベースとしては，ATR 自然発話音声・言語データベースがあり，話者約 500 名，のべ約 900 対話，約 23,000 発話を有する．本データベースの規模は，特に話者数に関しては，これら既存のデータベースと比較して桁違いに大きい．

一方，話者数の多い日本語音声データベースとしては，電話を通じて 8,866 名の音声を収録した「Voice Across Japan（VAJ）データベース」がある[10]．本データベースに含まれる話者は VAJ データベースの約４割と少ないが，VAJ は朗読調の音声を対象としたデータベースであり，一人あたりの発声数も少ないため，データベース全体のデータ量では本データベースが上回る．また，VAJ は電話回線を通じて収録したデータベースのため，電話回線の特性が重畳したデータになっていることも，本データベースとの違いである．

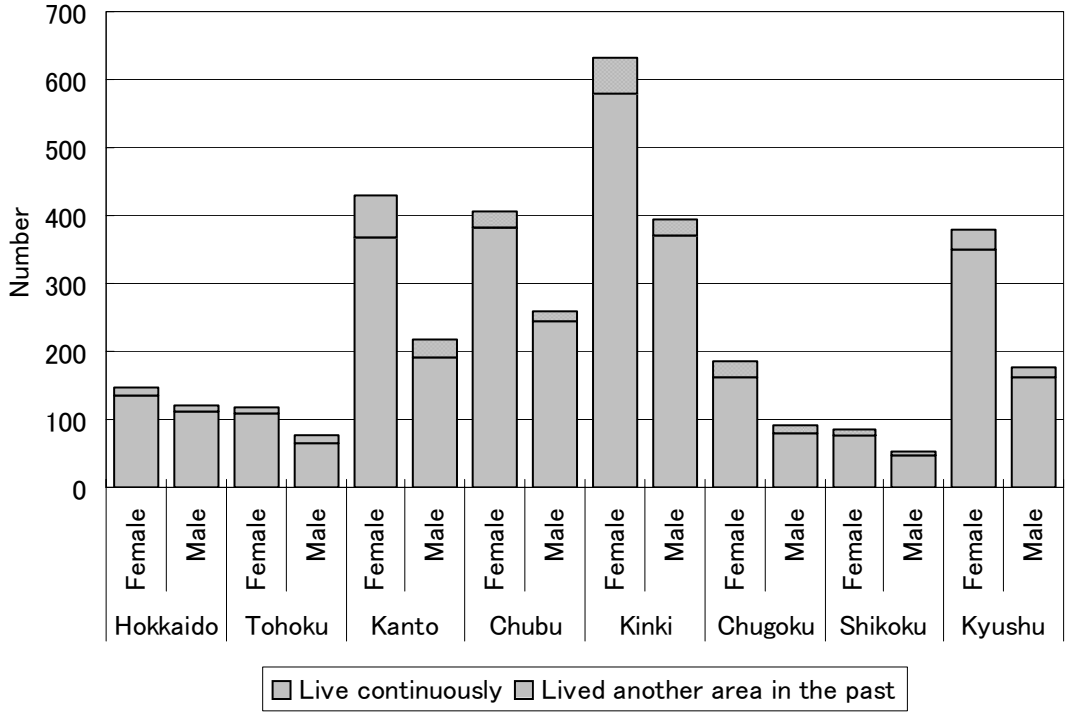


図３．２ 地域別話者数

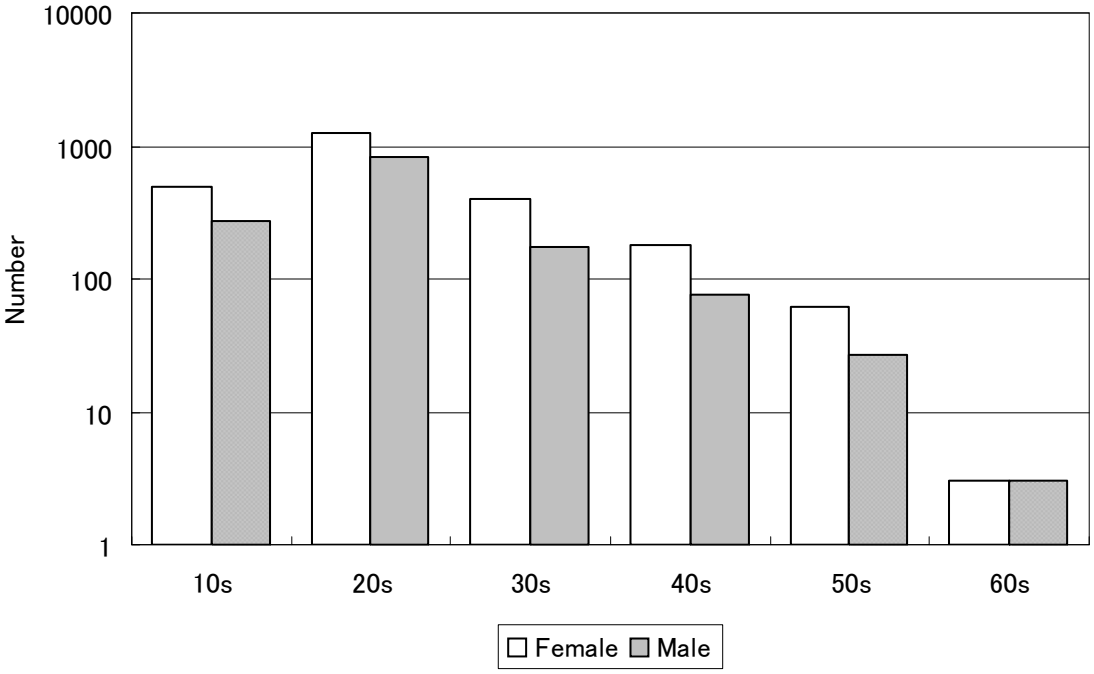


図３．３ 年代別話者数

表３．１ 地域別平均年齢

出身地方	年齢（男性）		年齢（女性）	
	平均	標準偏差	平均	標準偏差
北海道	27.1	8.1	25.8	7.2
東北	28.9	11.1	30.1	10.3
関東	28.9	10.3	30.1	11.1
中部	25.4	8.7	25.5	8.4
近畿	23.5	5.9	24.6	6.5
中国	23.4	6.4	25.9	9.6
四国	24.5	7.5	24.3	7.7
九州	23.0	5.3	27.1	9.0

表３．２　データベース内のデータ種別

模擬対話	会議のスケジュール調整をタスクとした２人の話者による模擬対話を収録した音声データ
朗読：文	音韻環境がバランスするように，ATR音素バランス文５０３文[6]の数文を読み上げた音声データ
朗読：単語	現在の日本語で使用されている外来音を含む音韻を幅広くカバーするため，国語辞典，ことわざ辞典，地名辞典，外来語辞典などから抜粋された文章，及び単語を読み上げた音声データ

表３．３　データベースの規模

データ種別	話者数［対話数］	発話数	発話時間
擬似対話	3,771[1,888]	32,914	53.4h
朗読（文）	3,770	112,660	127.6h
朗読（単語）	3,770	31,589	13.8h

３．２．３　音声データ収録方法

ここでは，音声データの収録方法について述べる．模擬対話は二人の話者が非対面で，日本語で行われている．会話はスケジューリングに関するタスクであり，異なる会社に所属する二人が会議のスケジュールを決定するため，電話を通じて対話する（非対面の対話）という設定となっている．実際の収録は電話回線ではなく，マイクロフォン，ヘッドフォンを通じて行われている．話者には事前にそれぞれの役割に応じた会話プロットが渡されており，プロットには自分のスケジュール，会議室の利用予約状況，会社の地図が記載されている．記載は，より自然な会話となるよう短い言葉やシンボルのみとなっている．このプロットをもとに，お互い相手のスケジュール等は知らない状態で対話が開始される．

対話は会議のスケジュールの決定を目的として行われ，会議の時間調整が主となるが，会議開催場所の確認のための道案内等の会話もなされている．発話表現は自由であり，自由発話特有の間投詞の挿入，言い淀み，言い直し等を許している．発話者に対する事前の注意として，(1)各発話者の音声は重ならないようにすること，(2)極端に長い発話を避けること，という２点についてのみ指示が与えられている．発話権の移動は外部からは与えず，会話の状況に応じて発話者間で暗に決定される．従って，一回の発話権の間に一人の話者が複数の発話を連続して行う場合もある．表３．４に，本データベースにおける対話例を示す．

文，単語の朗読音声の収録は，模擬対話の収録と同一の場所で行われている．各話者は

表３．４　本データベースにおける対話の例

話者B＞はい，こちら水野金属＜mizunokiNzoku＞資材部でございます．
話者A＞［あ］もしもし，こちら山田物産＜yamadabussaN＞のわたくし岡田＜okada＞ と申しますけれども．大角＜oosumi＞さん，お願いできますでしょうか．
話者B＞［あ］はい，あたしが大角＜oosumi＞ です．いつもお世話になっております．
話者A＞［あ］こちらこそいつもお世話になってます．
話者A＞［えーっとですね，あのー］来週に前からお話ししてました件で一度会議を持ちたいと思うんですけれども，ご都合のほうとかどうなってらっしゃいますでしょうか．
話者B＞［え］来週ですね．
話者B＞［えー］来週は＜wa＞ ［えー］あたくしのほうは＜wa＞ ，月曜日火曜日の午前中と，あとは＜wa＞ ［えー］六月二日＜futsuka＞金曜日の午後からでしたら都合がつくんですけれども．
話者A＞［あ］そうですか．［えーと］金曜日でしたらこちらも好都合ですので，［えーと］じゃあ，時間のほうは＜wa＞ 一時ぐらいからでよろしいですか．
話者B＞［あ］はい，結構です．

発声リストに記載された文・単語を読み上げている．発声に際しては，発話速度，文内での息継ぎ，休止の位置などの指定は行われておらず，発声リストを誤って発話した場合にのみ再発話を行うよう，指示されている．

３．３　多数話者音声データベースの音響的特徴の解析

本節では，収録されている多数話者の音声データを用いて，対話と朗読の発話スタイルの違いによる音響的特徴の違い，地域・年齢の違いによる音響的特徴の違いについて分析する．なお分析において，基本周波数は各発話中の母音部（ビタビアライメントにより判定）の平均基本周波数とした．また発話速度のばらつきは，各音声セグメントに含まれる母音の，中央点の時間間隔の標準偏差として求めた．またフォルマント周波数は，各音声セグメントに含まれる母音の中央点の周波数を抽出している．

３．３．１　対話と朗読の発話スタイルの違いによる音響的特徴の違い

対話音声と朗読音声の音響的特徴の違いを分析，比較する．図３．４に全話者による，それぞれの発話スタイルにおける各母音の第１，第２共振周波数（フォルマント周波数）F1-F2 の分布を示す．男性では第１フォルマントの平均周波数が 15.3Hz，第２フォルマントの平均周波数が 50.2Hz 変化しており，女性では第１フォルマントの平均周波数が 11.9Hz，第２フォルマントの平均周波数が 72.3Hz 変化している．図３．５に，男女各１話者における各母音のフォルマント周波数 F1-F2 の分布を示す．図中の縦横に伸びる線は，各母音における F1-F2 の標準偏差を表している．これらの図より同一話者においても，フォルマント周波数の平均値の変化が読み取れる．これらの結果より，対話音声などのテキストを読み上げない音声を認識する場合，朗読音声を用いて学習した音響モデルではミスマッチが大きく，認識性能が劣化すると考えられる．表３．５に，それぞれの発話スタイルにおける，母音の平均基本周波数，母音継続時間長の平均値，標準偏差を示す．基本周波数に関しては，朗読音声と比較し，対話音声において上昇する傾向があり，特に女性にその傾向が強いことがわかる．母音継続時間長の平均値に関しては，対話，朗読の間に大きな差は見られないが，標準偏差については対話音声の方が大きくなっている．これは対話音声の方が，発話内における発声速度の変動が大きいことを示している．

表３．５ 対話音声・朗読音声（文）の母音基本周波数と母音継続時間長

データ種別	平均基本周波数(Hz)	継続時間長	
		平均値(msec)	標準偏差(msec)
女性　対話音声	246.2	84.6	56.4
朗読音声（文）	230.4	86.3	40.0
男性　対話音声	140.7	78.3	58.7
朗読音声（文）	135.7	78.4	38.8

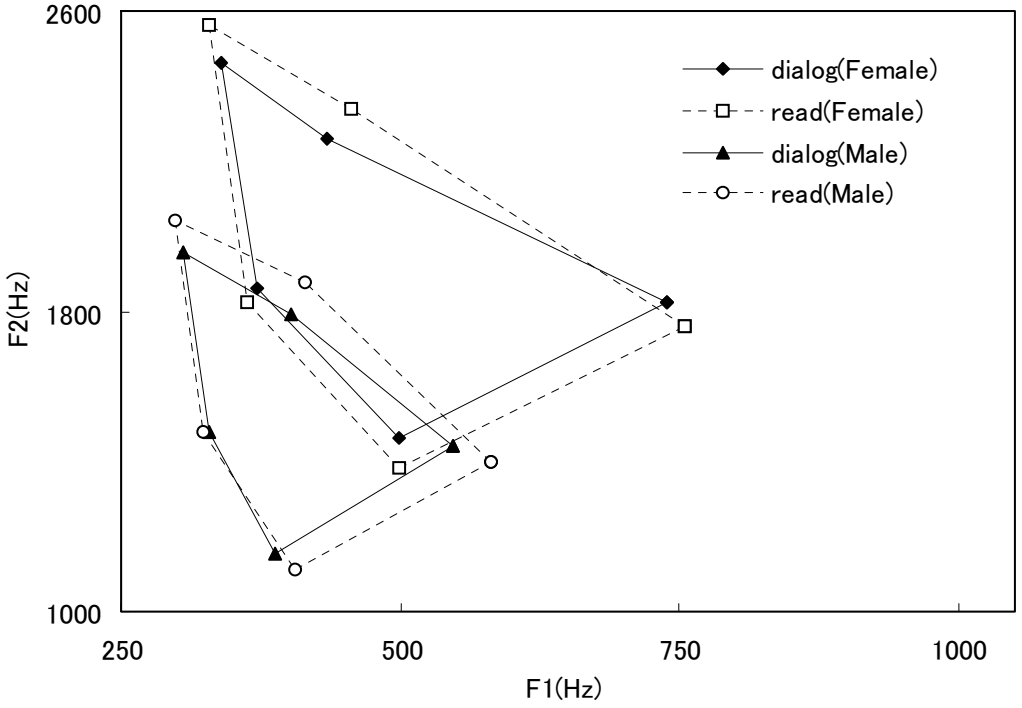


図３．４対話音声，朗読音声中の母音平均フォルマント周波数 F1-F2 の分布（全話者の平均）

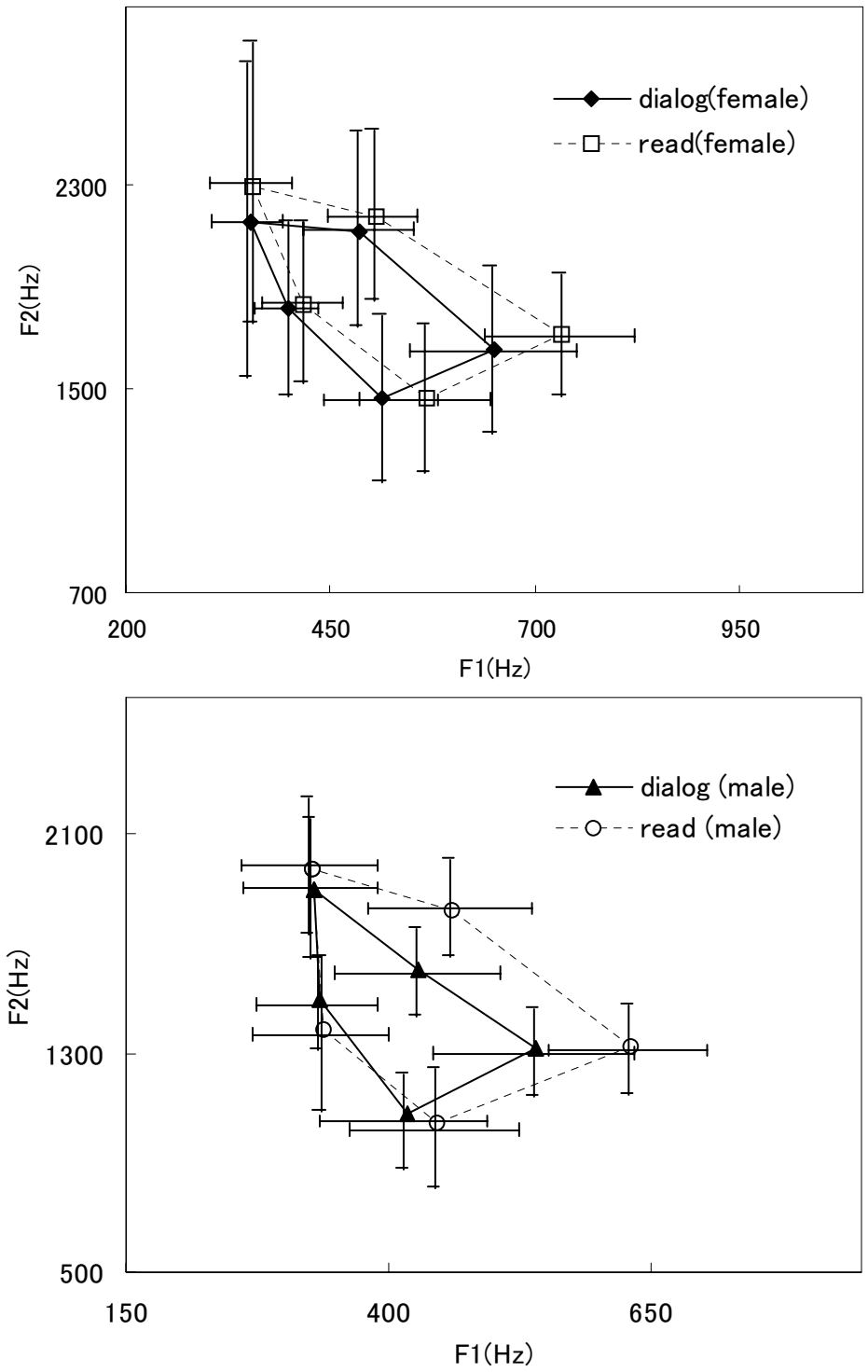


図３．５ 男女各１話者における対話音声，朗読音声中の母音平均フォルマント周波数 F1-F2 の分布

３．３．２　年齢・地域の違いによる音響的特徴の違い

ここでは対話音声を分析することで，年齢・地域の違いによる音響的特徴の違いを分析する．

まず，年齢別の音響的違いを分析する．図３．６に，年代別のフォルマント周波数 F1-F2 の分布を示す．年代の違いによる平均フォルマント周波数の最大値と最小値の差は，男性では第１フォルマントで 23.3Hz，第２フォルマントで 54.2Hz であり，女性では第１フォルマントで 33.6Hz，第２フォルマントで 93.9Hz となっているが，有意水準 0.05 において年代別フォルマント周波数の変動に有意な差は認められなかった．表３．６に，それぞれの年代における母音の平均基本周波数，母音継続時間長の平均値，標準偏差を示す．基本周波数に関しては，年齢が上がるにつれて低くなる傾向にある．母音継続時間長に関しては，年齢が上がるほど長くなり，標準偏差に関しては，年齢が上がるほど大きくなる傾向にある．このことより年齢が上がるほど，発声速度がゆっくりになるとともに，発話内での変動が大きくなる傾向にあることがわかる．

表３．６ 年代別対話音声の母音基本周波数と母音継続時間長

年代	平均基本周波数(Hz)	継続時間長	
		平均値(msec)	標準偏差(msec)
女性 10代	254.1	84.6	52.1
20代	248.9	84.5	57.0
30代	241.6	84.9	57.5
40代	231.7	84.5	58.7
50代	221.8	85.8	56.7
男性 10代	146.1	78.0	54.3
20代	141.1	77.8	57.4
30代	137.6	79.0	63.8
40代	131.9	80.7	64.9
50代	140.6	81.1	66.5

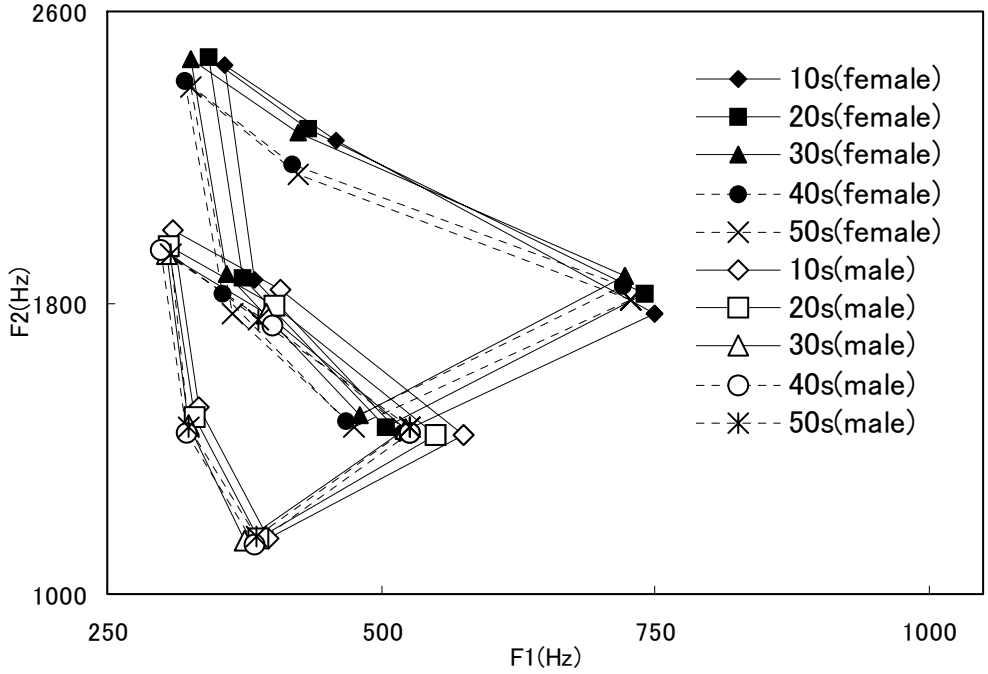


図３．６ 年代別の母音のフォルマント周波数 F1-F2 の分布

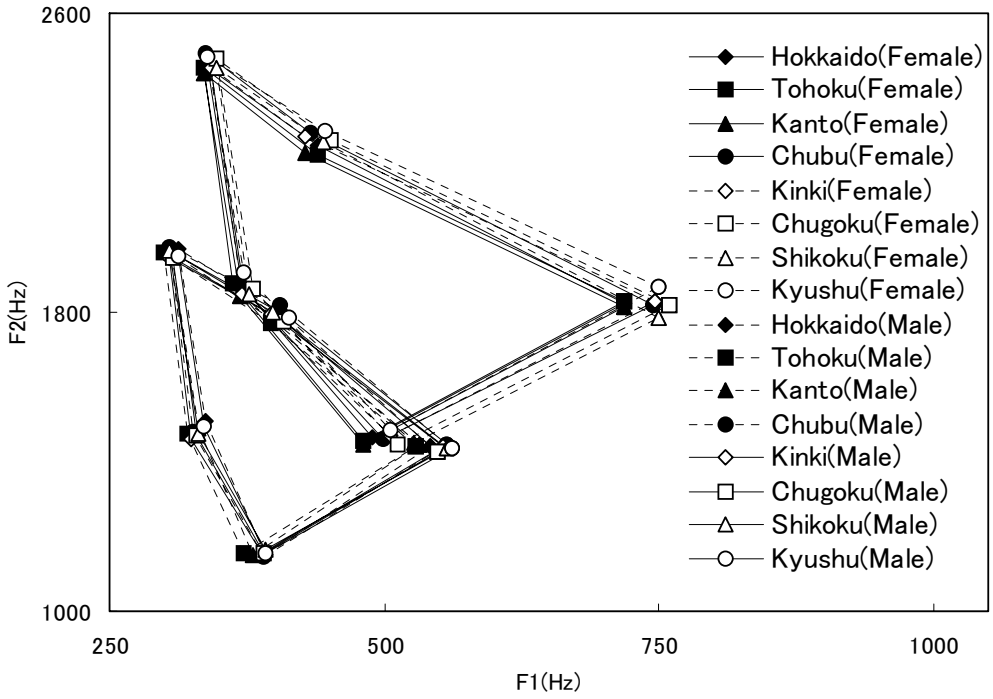


図３．７ 地域別の母音のフォルマント周波数 F1-F2 の分布

次に地域別の音響的違いを分析する．図３．７に継続的に居住している話者に関する，地域別のフォルマント周波数 F1-F2 の分布を示す．地域別の平均フォルマント周波数の最大値と最小値の差は，男性では第１フォルマントで 19.6Hz，第２フォルマントで 34.5Hz となっており，女性では第１フォルマントで 28.2Hz，第２フォルマントで 59.1Hz となっている．このことより，地域差によるフォルマント周波数 F1-F2 の変化は，年齢による変化と比べると男性で第１フォルマントが 3.7Hz，第２フォルマントが 19.7Hz，女性で第１フォルマントが 5.4Hz，第２フォルマントが 34.8Hz 小さいことがわかる．地域別のこれらの変動は，各地域における方言や，データベースにおける各地域の年齢分布の偏りなどが影響している可能性もあるが，有意水準 0.05 においてこれらの変動に有意な差は認められなかった．表３．７に，それぞれの地域における，母音の平均基本周波数，母音継続時間長の平均値，標準偏差を示す．母音継続時間長に関しては，地域により平均値で男性が 2.8msec，女性が 4.6msec，標準偏差で男性が 6.5msec，女性が 6.4msec 変動していることがわかる．また，それぞれの地域において女性より男性の方が，標準偏差が大きくなる傾向にあることもわかる．

表３．７ 地域別対話音声の母音基本周波数と母音継続時間長

地域	平均基本周波数(Hz)	継続時間長	
		平均値(msec)	標準偏差(msec)
女性 北海道	245.8	85.2	56.3
東北	239.9	84.2	58.4
関東	237.8	83.7	56.0
中部	248.7	85.2	57.1
近畿	248.3	83.7	56.0
中国	248.8	85.8	54.7
四国	251.9	88.3	61.1
九州	253.1	85.0	54.7
男性 北海道	144.2	79.9	59.0
東北	135.6	78.0	61.1
関東	134.8	77.9	59.0
中部	142.6	79.1	59.4
近畿	142.8	77.1	57.9
中国	134.2	79.9	58.8
四国	142.5	77.3	54.6
九州	142.6	78.1	59.0

３．４　年齢差・地域差が認識性能に与える影響

不特定話者音声認識を実現するためには，地域や年齢の違いによる発声の特徴の違いが，音声認識に及ぼす影響について分析する必要がある．３．３節で述べたように，スペクトル情報を表す特徴の一つである第１，第２フォルマントには，年齢や地域差による有意な差は見られなかったが，音声認識ではスペクトル情報を表す特徴量としてケプストラムが用いられる．本節では，本データベースを用いた認識実験を通して，ケプストラム特徴量に基づく年齢や地域の違いが認識性能に与える影響を調査する．音響的違いによる影響を調査するため，子音が連続することのみを制限した音素バイグラムを用い，音素認識実験を行った．

３．４．１　実験条件

本実験では，ATR で開発されたデコーダ，ATRSPREC を用いた[14][15]．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，サンプリング周波数 16kHz，プリエンファシス 0.98，分析周期 10msec，窓長 20msec で抽出した 25 次元の特徴ベクトル（12 次 MFCC，12 次Δ MFCC とΔlog power）を用いている．音響モデルは，最ゆう逐次状態分割法（Maximum Likelihood Successive State Splitting ; ML-SSS）[16]を用いて学習した隠れマルコフ網（hidden Markov networks ; HMnet）[17]を使用しており，性別依存モデル，５混合ガウス分布（無音モデルは 10 混合），1,400 状態（無音モデルは３状態）の状態共有化 HMM となっている．学習データには，ATR で収集された，東京や神奈川出身の 20 代の話者を多く含む，旅行対話をタスクとした自然発話音声・言語データベースを用いた[9]．この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．

３．４．２　認識実験結果

図３．８に継続的に居住している話者の地域別の音素正解精度を，図３．９に年代別の音素正解精度を示す．これらの図において，各方形領域の上下辺は四分位範囲を，中央の水平線は中央値を表し，方形領域の幅は各地方，年代に含まれる話者数の平方根に比例している．男女性別の平均認識率は，女性が 70.4%（最低 30.7～最高 87.7%），男性が 67.1%（最低 38.4～最高 85.4%）であった．

図３．８より，地域により音素正解精度が若干ではあるが変動していることが読み取れる．有意水準 0.05 においてこれらの変動に検定を行った結果，有意な差が認められた．

若干ではあるが地域別の音素正解精度に差が見られたことより，地域別音響モデルを構

築することで認識性能の向上が期待できる．そこで，地域別に音響モデルを構築した場合の認識実験を行った．学習データ量の認識性能への影響を抑えるため，各地域男女それぞれ 100 話者を用いて，地域別音響モデルを構築した．音響モデルの構築方法としては，(1) 全ての話者を用いて学習した音響モデルのトポロジーに対して，地域ごとにそのモデルのパラメータ推定を行ったもの，(2) 地域ごとに個別に学習したトポロジーに対して，地域ごとにパラメータ推定を行ったもの，の２種類を用意した．男女各 100 名で音響モデルを構築するため，東北，中国，四国の３地域に関しては話者数が足りず，本評価実験では用いていない．

図３．１０にそれぞれの音響モデルを用いた場合の地域ごとの音素正解精度を示す．評価話者は，それぞれの地域より継続的に居住している話者を男女それぞれ 10 名ずつ選択して用いている．この図より，地域ごとにトポロジーを学習した音響モデルの方が，良好な結果が得られていることがわかる．

次に，教師あり話者適応を行った場合の音素正解精度について調査を行った．話者適応には MAP-VFS[18]～[20]を用い平均値を，Baum-Welch 法[21]を用いて状態遷移確率を適応している．また，適応には各話者全ての発声を用いている．同様に図３．１０に話者適応を行った場合の音素正解精度を示す．話者適応を行った場合でも，地域ごとにトポロジーを学習した音響モデルの方が良好な結果が得られている．地域ごとにトポロジーを学習することで認識性能が改善することから，方言などの影響により地域ごとの音韻環境が異なっている可能性もあると考えられる．

図３．９より，年代により音素正解精度が変動していることがわかる．地域による変動と同様，有意水準 0.05 において，これらの変動に有意な差が認められた．年代に関しては，30 代をピークに年齢が上がるに連れて，認識率が低下することがわかる．この理由としては，評価における音響モデルの学習話者セットが 20 代の話者を多く含んでいたことが挙げられる．また表３．６より，年齢が上がるほど発声速度の変動が大きくなる傾向があり，音素ごとに安定した発声が得られにくいということも，認識性能の低下に関連している可能性がある．

以上の結果より，地域・年代により認識性能が影響を受けることがわかる．この変動に頑健な音響モデルを構築するには，これらの変動を十分に含む学習データが不可欠である．しかしながら大量の学習データを用いてパラメータ推定を行う，例えば HMM で各音素を表現する場合，各状態が表現する分布が広がることにより，各音素の識別性能が低下し，かえって認識性能が劣化することも考えられる．地域別・年代別による変動を吸収する方法や，十分な量の学習データがある場合は，地域別・年代別に音響モデルを構築する方法

も検討する必要がある．

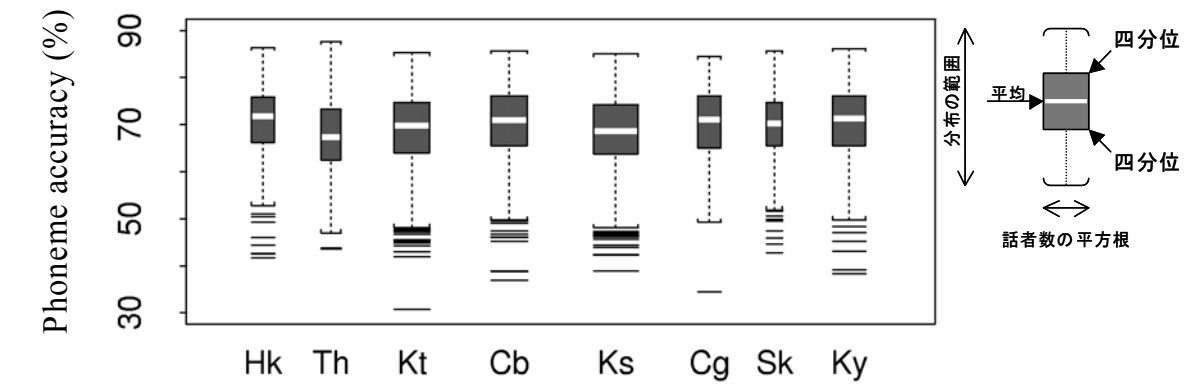


図３．８ 地域別の音素正解精度の分布（Hk：北海道，Th：東北，Kt：関東，Cb：中部，Ks：近畿，Cg：中国，Sk：四国，Ky：九州）

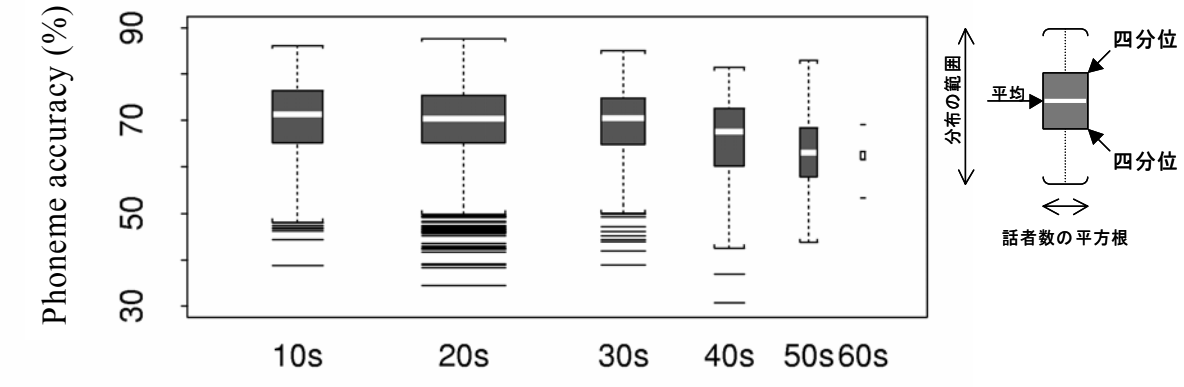


図３．９ 年代別の音素正解精度の分布

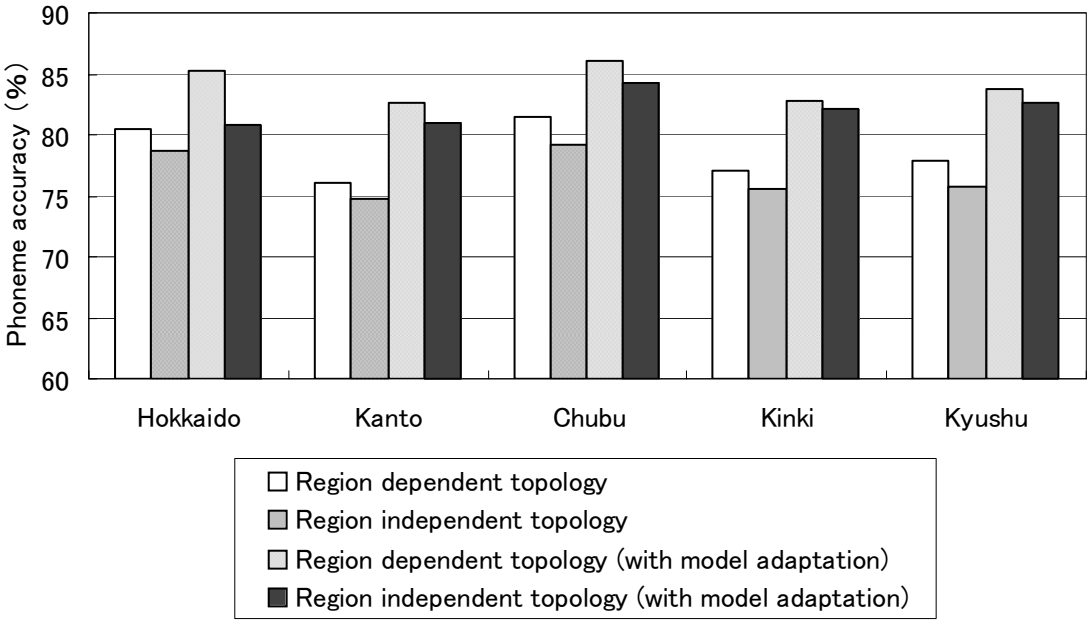


図３．１０ 全話者で構築したトポロジー，地域別に構築したトポロジーを用いた音響モデルによる地域別の音素正解精度

３．５　多数話者音声データベースを用いた音響モデルの構築

前節までの結果から，発話スタイルの違いや地域・年齢の違いにより，音響的な特徴が変動し，これらの変動が認識システムの性能に影響することが明らかになった．一般に不特定話者音声認識における音響モデルの構築には，できるだけ多くの話者の音声を使用した方が良いとされているが，その理由の一つがこれらの変動の影響である．しかしながら，その上限の話者数などについての十分な検討は行われていない．

そこで本節では，本データベースを用いて音響モデルを構築することにより，学習データ量と音響モデルの性能の関係を調査した．

３．５．１　実験条件

実験条件は，基本的に３．４節と同じである．音響モデルの学習は，話者数の効果を調べるために，本データベースのうち，後述する評価セットを除いたすべてのデータを用いた場合（S），その半分の話者数のデータを用いた場合（S／2），その４分の１の話者数のデータを用いた場合（S／4），その８分の１の話者数のデータを用いた場合（S／8）について行った．また，比較のために，本データベースのタスクとは異なる，旅行対話タスクの自然発話音声・言語データベースを用いた場合（T）についても行った．各学習セットの話者数と発話時間を表３．８に示す．

評価では，話者の違いによる影響を調べるために，本データベースからは，ランダムに選択した男女各 20 名（S[標準]），予備実験で高い認識率を示した話者から選択した男女各 12 名（S[高い]），低い認識率を示した話者から選択した男女各 12 名（S[低い]）を用いた．また，自然発話音声・言語データベースからは，42 名（T[標準]）を選択して使用した．

表３．８ 各学習セットの話者数と発話時間

学習	性別	話者数（人）	発話時間（h）
S	F	2310	32.0
	M	1321	19.3
S/2	F	1155	15.8
	M	661	9.8
S/4	F	578	8.2
	M	331	4.9
S/8	F	289	4.0
	M	166	2.4
T	F	240	3.0
	M	167	2.0

言語モデルに関しては，多重クラス複合 N-gram モデルを用いた[22]．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．言語モデルの学習は，本データベース，および自然発話音声・言語データベースの両方のデータを用いて行い，まとめて一つのモデルを作成した．認識辞書は 27k ワードである．評価には，本データベースの対話音声を使用した．

３．５．２　混合数，状態数の影響

表３．９に，学習セットＳを用いた場合に，混合ガウス分布数を{５，10}，状態数を{1,400，2,100，2,800} に変化させた場合の単語正解精度を示す．分布数に関しては 10 の方が良好な結果となった．これは，学習データ量の増加により，その音響的な広がりが大きくなったことによるものと考えられる．また状態数に関しては認識性能に差は少なく，2,800 の場合にはむしろ性能が低下する場合が多いことがわかる．本データセットを用いた場合でも，2,800 状態では過分割されているものと判断できる．総合的には，この実験条件では混合ガウス分布数 10，状態数 1,400 が適当である．以降，本節ではこの設定をデフォルトの設定として使用する．

３．５．３　学習データ量の影響

表３．１０に，学習にＳ，Ｓ／２，Ｓ／４，Ｓ／８のデータセットを用いた場合の単語

表３．９ 混合ガウス分布数，状態数による単語正解精度(%)の違い（[ ] : realtime factor）

性別	混合分布数	状態数	評価					
			Ｓ［標準］		Ｓ［低い］		Ｓ［高い］	
F	5	1,400	89.7	[1.2]	74.5	[1.4]	94.8	[0.8]
		2,100	89.9	[1.2]	76.5	[1.4]	94.4	[0.8]
		2,800	90.0	[1.3]	75.1	[1.3]	94.5	[0.8]
	10	1,400	**90.8**	[1.3]	**78.1**	[1.8]	95.4	[1.0]
		2,100	90.5	[1.3]	77.2	[1.6]	95.0	[1.0]
		2,800	90.1	[1.3]	77.0	[1.6]	**95.8**	[1.0]
M	5	1,400	85.2	[1.7]	71.7	[3.0]	93.3	[1.2]
		2,100	85.6	[1.7]	72.9	[2.7]	93.7	[1.2]
		2,800	85.4	[1.7]	72.0	[2.6]	93.4	[1.2]
	10	1,400	85.5	[2.0]	73.2	[3.3]	**94.1**	[1.6]
		2,100	86.3	[1.9]	**73.3**	[3.2]	93.7	[1.6]
		2,800	**86.6**	[1.9]	72.9	[3.6]	93.7	[1.6]

正解精度と，Ｓの単語正解精度を標準とした時のＳ／２，Ｓ／４，Ｓ／８の誤り増加率を示す．学習データ量が少なくなるに従い，誤り増加率が急激に大きくなる傾向が見られた．また特に男性話者に関しては，学習セットＳで認識性能が飽和傾向にあった．これは，本評価実験環境ではこれ以上学習データ量を増加しても，認識性能の大幅な改善は期待できないということを示している．しかしながら，評価セットＳ[標準]，Ｓ[低い]の認識性能は十分ではない．これらの話者に対しては，単に多くの学習データを集めるのではなく，それぞれの話者の音響的な特徴を十分分析した上での学習データの選び方，音響モデルの構築方法の検討が必要である．

さらに，女性話者に比べ男性話者の方が全体的に認識性能が低くなっている．発話時間 19.3 h の学習データＳで構築した男性用音響モデルよりも，発話時間 2.4h の学習データＳ／８で構築した女性用音響モデルの方が，良好な認識率を得ている．３．３節で分析した内容から，女性話者に比べ男性話者は母音継続時間長の標準偏差が大きく，発声速度が変動していることが読み取れる．これらの変動を含む音響的特徴の偏りが，認識性能に影響を与えた可能性がある．

３．５．４　クロスタスクの影響

表３．１１にクロスタスクによる単語正解精度をまとめる．また図３．１１に評価セットＳ[標準]を用いた場合の学習データ量と単語正解精度を，図３．１２に評価セットＴ[標準]を用いた場合の学習データ量と単語正解精度を示す．なお，自然発話音声・言語データベースの学習セットＴに関しては，予備実験から HMnet により構築した混合ガウス分布数５，状態数 1,400 の状態共有化 HMM を使用した．

図３．１１より，評価セットＳ[標準]に関して見た場合，タスクの異なる学習セットＴ

表３．１０ 学習データ量による単語正解精度(%)（[ ]：Ｓを基準とした時の誤り増加率%）

性別	学習セット	評価					
		S[標準]		S［低い］		S[高い]	
F	S	90.8	－	78.1	－	95.4	－
	S／２	90.5	[3.3]	76.1	[9.1]	94.6	[17.3]
	S／４	89.9	[9.9]	75.9	[10.0]	94.6	[17.3]
	S／８	88.8	[21.7]	73.6	[20.5]	93.7	[37.0]
M	S	85.5	－	73.2	－	94.1	－
	S／２	85.4	[0.7]	73.3	[-0.3]	94.2	[-1.7]
	S／４	83.5	[13.8]	70.7	[9.3]	93.0	[18.6]
	S／８	81.5	[27.6]	70.4	[10.4]	92.2	[32.2]

の音響モデルより，同一タスクの学習セットで構築した音響モデルの方が認識率が高く，学習データ量の一番少ないＳ／８の場合でも，学習セットＴより良好な結果となっていることがわかる．

これに対して図３．１２より，評価セットＴ[標準]に関して見た場合，Ｓ／８の音響モデルより学習セットＴの音響モデルの方が良好な結果となっているが，学習データ量が増

表３．１１ クロスタスクの単語正解精度(%)

学習	性別	評価	
		S[標準]	T［標準］
S	F	90.8	89.7
	M	85.5	85.7
S／2	F	90.5	88.8
	M	85.4	84.8
S／4	F	89.9	88.5
	M	83.5	83.9
S／8	F	88.8	86.0
	M	81.5	81.3
T	F	83.1	87.7
	M	76.6	82.5

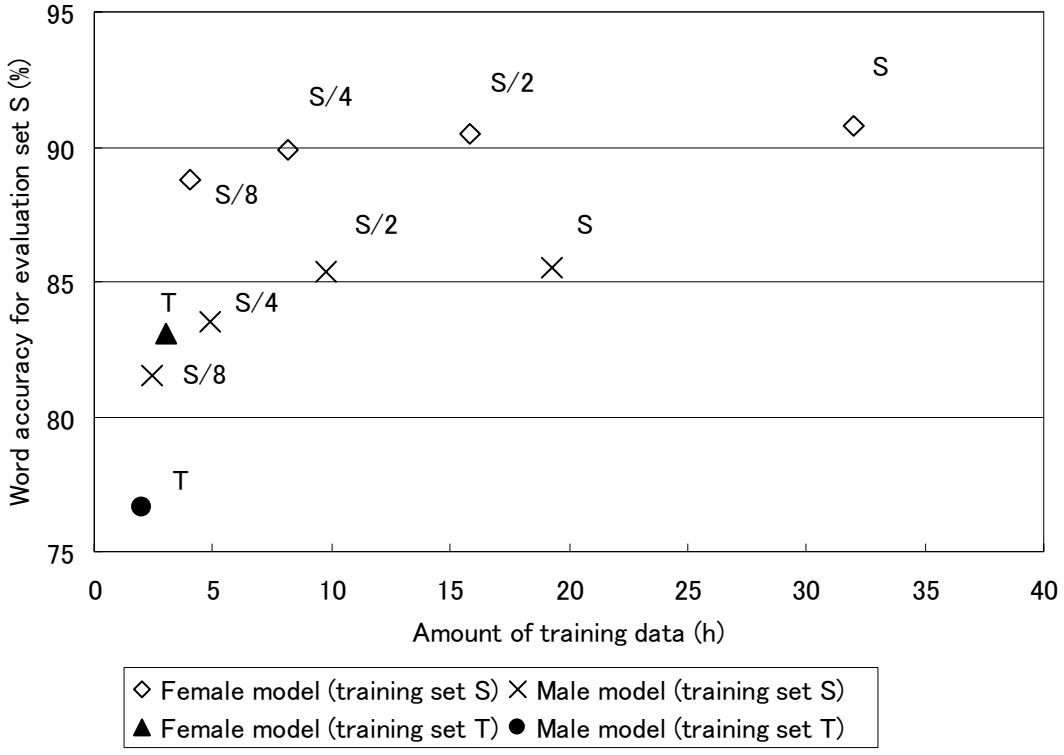


図３．１１ クロスタスクの単語正解精度（評価セットＳ[標準])

えるにつれ，本データベースを用いた音響モデルの方が良好な結果となっていることがわかる．学習セットＳを用いた場合，女性で 89.7%，男性で 85.7%の単語正解精度が得られており，同一タスクの学習セットで構築した音響モデルにおけるに単語正解精度より，女性で 2.0%，男性で 3.0%改善している．

自然発話音声・言語データベースは模擬対話データであり，発声内容も予約内容や時刻の確認が多く含まれ，異なるタスクとは言え本データベースと比較的近いタスクである．上記の結果は，話者数を確保することにより，本データベースと自然発話音声・言語データベースのように近いタスク間であれば，タスクに依存しない音響モデルを構築することが可能であることを示している．なお，（図３．１２から）Ｓ／４の 909 名で学習した場合と，Ｔの 407 名で学習した場合とで，評価セットＴ[標準]に対して，ほぼ同程度の性能が得られた．

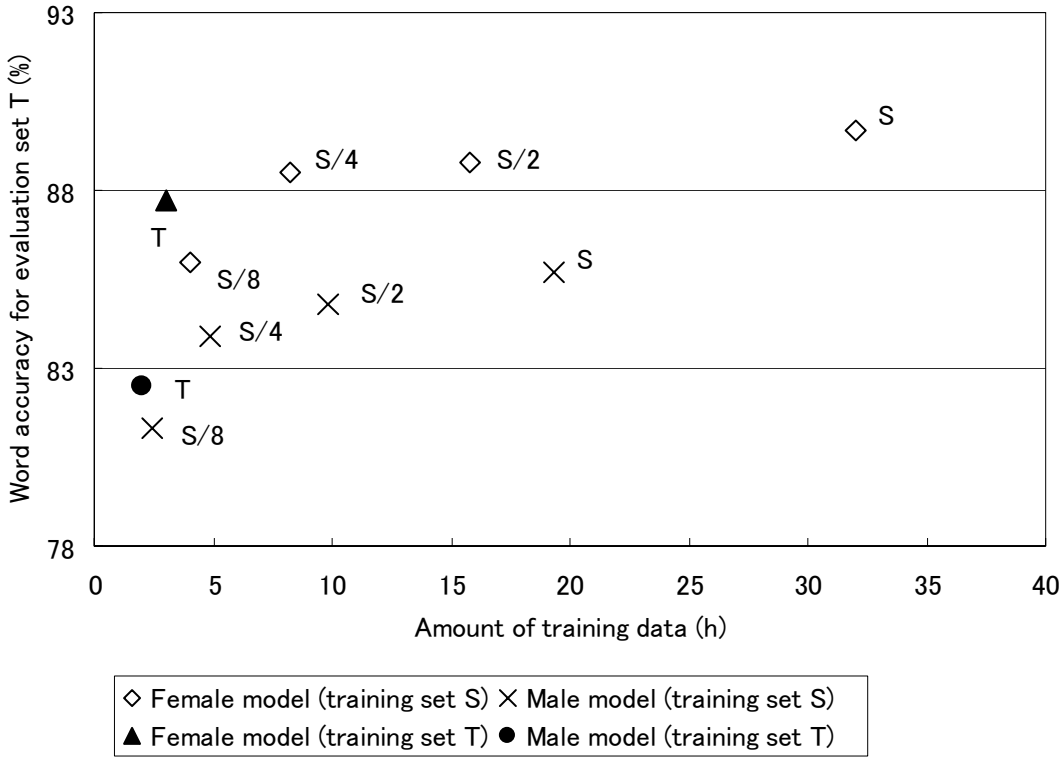


図３．１２ クロスタスクの単語正解精度（評価セットＴ[標準]）

３．６　考察

本章におけるデータベースのフォルマント特徴量の分析からは，地域別，年代別でその音響的特徴の違いに有意な差は見られなかったが，ケプストラム特徴量を用いた認識実験を通して見た場合，地域・年代の差が認識性能に影響を与えることがわかった．これは，音素認識実験の際に用いた音響モデルの学習セットと，評価データセットの話者の違いによる認識性能への影響からも読み取ることができる．

不特定話者音声認識を実現するためには，音声認識システムで扱う音響的特徴の変化を吸収できる音響モデルの構築や認識手法の開発が必要であり，本データベースを用いた音響モデルの構築も，その一手法と言える．学習セットの話者数を増やすことで，全体の認識性能が向上していることからも，本手法は不特定話者音声認識に有効である．しかしながら３．５．３で述べたように，ある一定以上のデータ量を越えた時点で，認識性能の大きな改善が期待できなくなる可能性がある．またデータ量を増加することで，音響的な広がりが大きくなり，かえって認識性能が劣化する場合もある．この場合，地域別・年齢別に個別に音響モデルを構築することも必要になるが，どの程度のデータ量でそれぞれの音響モデルの性能がピークに達するのかの調査が必要となる．

本章での，学習データ量による認識性能の評価実験からは，学習データ量の上限を導き出すには至っていないが，図３．１１における認識性能の飽和傾向から，同一タスクによる評価では今回使用したデータ量でその上限にかなり近い量になっていると判断できる．

３．７　結言

ATR において収録された 3,700 人規模の多数話者音声データベースを用い, 地域や年齢, 発話スタイルの違いによる音響的特徴の違い，学習データ量と認識性能の関係について分析した.

年齢の違いによる音響的特徴の違いとしては，周波数的には有意な差は無いが，高齢になるにつれ平均基本周波数が下がるとともに，発話速度がゆっくりとなり，発話内での発話速度の変動が大きくなる傾向があることがわかった．地域の違いによる音響的特徴の違いとしては，年齢の違いと同様に，周波数的には有意な差は認められなかった．発話速度についても大きな差は見られなかった．読み上げ音声と対話音声の間においては周波数的に変動するとともに，対話音声では基本周波数が上昇し，発話内での発話速度の変動が大きくなることがわかった.

認識性能については地域，年齢の違いで有意な差が認められた．これは，有意な差が認められなかった音響的特徴の違いや発話速度の違いでも，認識性能に影響を与える要因になることを示している．また全ての話者から構築したトポロジーよりも，地域別に構築したトポロジーを用いた方が認識性能が良く，話者適応の効果が大きく現れることも明らかとなった．発話スタイルの変動などによりトポロジーのミスマッチが生じる場合にも，認識性能が大きく劣化すると考えられ，この改善は難しい.

学習データ量と認識性能の関係については, 学習データ量が増加するにつれ認識性能が改善することを確認した．また，発話スタイルやタスクが比較的近い場合，より多くの学習データを利用することにより認識性能が改善できることも確認した.

以上の結果より，ATR 多数話者音声データベースは音声認識研究において有効であること，音声認識システムで扱う音響的特徴の違いが認識性能に大きな影響を与えることが明らかになった．発話スタイルが大きく異なる場合，音響モデルのモデルパラメータだけではなく，音響分析手法やトポロジーまでを含んだ検討が必要となる．次章以降，発話スタイルの変動に頑健な音声認識手法についての研究結果について述べる.

参考文献

[1] 奥田浩三，松井知子，内藤正樹，匂坂芳典，中村 哲，“大規模日本語音声データベースの構築と評価，”日本音響学会誌，vol.58，no.9，pp.569-578，2002.

[2] 田中和世，速水 悟，“電総研の研究用音声データベース，” 日本音響学会誌，vol.48，no.12，pp.883-887，1992.

[3] 小林哲則，板橋秀一，速水 悟，竹沢寿幸，“日本音響学会研究用連続音声データベース，” 日本音響学会誌，vol.48，no.12，pp.888-893，1992.

[4] 板橋秀一，“文部省「重点領域研究」による音声データベース，”日本音響学会誌，vol.48，no.12，pp.894-898，1992.

[5] 牧野正三，二矢田勝行，真船裕雄，城戸健一，“東北大-松下単語音声データベース，”日本音響学会誌，vol.48，no.12，pp.899-905，1992.

[6] 桑原尚夫，匂坂芳典，武田一哉，安部匡伸，“研究用ATR日本語音声データベースの作成（別冊Ｉ連続音声テキスト)，”TR-I-0086，1989.

[7] “研究用日本語音声データベースの作成，”ATR 自動翻訳電話研究所テクニカルレポート，TR-I-0086，1989.

[8] A.Nakamura，S.Matsunaga，T.Shimizu，M.Tonomura and Y.Sagisaka，“Japanese Speech Database for Robust Speech Recognition，” Proc. of ICSLP'96，pp.2199-2202，1996.

[9] T.Matsui，M.Naito，H.Singer，A.Nakamura and Y.Sagisaka，“Japanese spontaneous speech database with wide regional and age distribution，” Proc. of Eurospeech'99，vol.5，pp.2251-2254，1999.

[10] 工藤育男，中間崇夫，“Voice Across Japan データベース，”情報処理学会論文誌，vol.40，no.9，1999.

[11] H.Soltau and A.Waibel，“On the influence of hyperarticulated speech on recognition performance，” Proc. of ICSLP'98，pp.229-232，1998.

[12] 奥田浩三，松井知子，中村 哲，“音節強調発声に頑健な自然発話音声の認識法，” 電子情報通信学会技術研究報告，vol.100，No.523，pp.19-24，2000.

[13] T.Matsui，M.Naito，Y.Sagisaka，K.Okuda and S.Nakamura，“Analysis of acoustic models trained on a large-scale Japanese speech database，”Proc. of ICSLP2000，vol.2，pp.503-506，2000.

[14] 内藤正樹，山本博史，シンガー ハラルド，中嶋秀治，中村 篤，匂坂芳典，“対話音声を対象とした連続音声認識システムの試作と評価，”電子情報通信学会論文誌，vol.J84-D-II，no.1，pp31-40，2001.

[15] 山本博史, シンガー ハラルド, リーブス ベン, 匂坂芳典,“日英音声翻訳システム「ＡＴＲ－ＭＡＴＲＩＸ」における音声認識部分の構造と制御方法,”日本音響学会 1998 年春季研究発表会, 2-Q-21, 1998.
[16] M.Ostendorf and H.Singer,“HMM Topology Design using Maximum Likelihood Successive State Splitting,” Computer Speech and Language, vol.11, no.1, pp.17-41, 1997.
[17] 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,”電子情報通信学会論文誌, vol.J79-D-II, no.10, pp.2155-2164, 1993.
[18] M.Tonomura, T.Kosaka and S.Matsunaga,“Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation,”Computer Speech and Language, vol.10, pp.117-132, 1996.
[19] J.Takahashi and S.Sagayama,“Vector-field-smoothing Bayesian learning for fast and incremental speaker/telephone-channel adaptation,”Computer Speech and Language, vol.11, pp.127-146, 1997.
[20] 大倉計美, 杉山雅英, 嵯峨山茂樹,“混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式,”電子情報通信学会技術研究報告, SP92-16, 1992.
[21] Lawrence Rabiner, Biing-Hwang Juang 共著, 古井貞熙 監訳,“音声認識の基礎（上）(下),”NTT アドバンステクノロジ株式会社, 1995.
[22] 山本博史, 匂坂芳典,“接続の方向性を考慮した多重クラス複合 N-gram 言語モデル,”情報処理学会研究報告, SLP98-7, 1998.

第 4 章

講演音声認識のための音響分析と音響モデル構築

４．１　緒言

本章は，論文“ゆう度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデルの構築”[1]に関するものである．

第３章で明らかになったように，異なる発話スタイルの間では，その音響的特徴が大きく変化する．この音響的特徴の変化は音声認識システムの認識性能に大きく影響する．このため音声認識システムを利用する場合，利用者は認識しやすい発話スタイルでの発声を求められる場合が多く，音声認識技術の利用範囲を制限する原因となっている．

現状の音声認識システムの多くは，読み上げに近い発話様式を対象としており，講演音声などの書き起こしや，自然な話し言葉を対象としたヒューマン・インタフェースを考えた場合，必ずしも十分な認識性能が得られているわけではない．発話様式の違いが認識性能に大きく影響することからも[2]，話し言葉の認識を対象とした音響モデルや言語モデル，デコーダ等の研究開発は不可欠である．

これに対し平成 11 年度より，話し言葉音声の分析と認識を目的とした開放的融合研究「話し言葉工学」プロジェクト[3][4]が実施されている．このプロジェクトでは，話し言葉音声を対象とした大規模なコーパスの構築と，話し言葉の音声認識，理解，要約などの技術基盤の確立を目指している．既に，講演音声を中心とした「日本語話し言葉コーパス（Corpus of Spontaneous Japanese ; CSJ）」[5][6]の構築も進められており，このコーパスを用いた研究も開始されている[7]～[10]．

講演音声のような話し言葉を対象とした音声認識においては，発話速度の変動が認識性能に大きく影響することが報告されている[7]．具体的には，発話速度の速い音声においては脱落誤りや置換誤りが多く発生し，発話速度の遅い発声においては挿入誤りが多く発生すると報告されている[8]．これに加えて，講演音声においては，講演者のスキルや講演内容，講演のスタイル（原稿を読み上げた場合や暗記して講演する場合，原稿を準備しない場合など）により，発話速度にばらつきが多く見られるとともに，同一講演内においても講演の前半と後半では発話速度に差が生じる傾向が観測されている[11]ことから，発話速度の変動に対して頑健な音声認識手法の検討が重要となる．

これらの背景のもと，本章では CSJ を用い，講演音声の音響的特徴を分析するとともに，分析周期・窓長を変更することによる発話速度の補正を用いたデコーディング手法，ならびに音響モデルの構築方法を提案する．提案するデコーディング手法は，発話ごとに複数の分析周期・窓長を用いて認識した結果に対して，ゆう度基準で発話ごとに適した分析周期・窓長を選択することにより，発話速度補正の効果を得るものである．また，同様の手法を音響モデルの学習時にも適用することで，発話速度を正規化した音響モデルを構築す

る．

まず４．２節において，認識実験を通して講演音声が認識性能に与える影響について述べる．次に４．３節で音素継続時間長，周波数的特徴，認識性能を対話音声，読み上げ音声と比較することで，講演音声の音響的特徴をまとめる．４．４節では，講演音声認識のための音声認識手法の提案として，音響モデルにおける発話速度変動のモデル化，分析周期・窓長の最適化を提案し，教師なし話者適応の導入などによる効果を述べる．これらの結果を踏まえ，４．５節では発話速度に合わせた分析周期・窓長選択手法の一般化について述べ，４．６節で実際のデコーディング手法とその効果をまとめる．さらに４．７節で，本手法をモデル学習時に適応した場合の効果について述べる．

４．２　講演音声が認識性能に与える影響

講演音声に対して頑健な音響モデルの構築方法を検討するにあたり，本節では対話音声により構築した音響モデルを用いた認識実験を行うことで，講演音声の発話スタイルが認識性能に与える影響について調査する．

４．２．１　実験条件

本節における認識実験では，国際電気通信基礎技術研究所（ATR）で開発されたデコーダ，ATRSPREC[13]を用いている．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，分析周期 10msec，窓長 20msec で抽出した 25 次元の特徴ベクトル（12 次 MFCC，12 次ΔMFCC とΔlog power）を用いた．

音響モデルには，HMnet（hidden Markov networks[14]）により構築した状態共有化 HMM で，各音素モデル３状態，10 混合ガウス分布，総状態数 1,400 で表現された性別依存モデルを用いている．音響モデルの学習データには，ATR で収集された多数話者音声データベース[15][16]より，男性話者 1,321 人の模擬対話音声データ（約 19.3 時間）を使用した．この音響モデルを対話音響モデルと呼ぶこととする．

単語認識実験においては，2000 年度に配布された「話し言葉工学」プロジェクトのモニタ版に付属されている前向き単語バイグラムと認識辞書を使用した．ただし，認識辞書における音素表記が，情報処理推進機構（IPA）の研究テーマ「日本語ディクテーション基本ソフトウェアの開発」[17]で定義されている音素セットで表記されており，ATR で使用している音素セットと異なるため（IPA では 41 音素，ATR では 26 音素），ATR 音素セットに変換し利用している．音素認識実験の際には，連続する子音など，日本語表記においてあり得ない組み合わせのみを制限した音素バイグラムを使用した．

評価データには，2000 年度に配布された「話し言葉工学」プロジェクトのモニタ版に付属されている，４名分の講演音声データからなる評価セットを用いた．評価セットを表４．１に示す．

４．２．２　認識実験結果

音素認識実験，単語認識実験の結果を表４．２にまとめる．ここで示した単語正解精度は，言い誤りは除外し，間投詞は含んだ状態で算出している．

比較として，音響モデルの学習セットと同じ会議予約タスクの評価セット（T set）を用いた場合の認識結果も示す．T set の単語認識には，ATR にて収録された自然発話音声・言語データベース[18]を用いて構築した認識辞書，言語モデルを使用した．融合研究プロジ

ェクトのテストセットは，音素正解精度で平均 64.4%，話者によっては 70.3%となっているが，T set と比較すると平均で 10%程度劣化している．単語正解精度においては，テストセットは T set の認識率より平均で 20%程度劣化している．また，T set の認識結果は音素正解精度より単語正解精度の方が向上しているのに対し，テストセットでは単語正解精度が大きく劣化する．以上の結果より，テストセットは音響モデル・言語モデル共にミスマッチが生じていることがわかる．

表４．１　「日本語話し言葉コーパス(CSJ)」モニタ版(2000)に含まれる評価セット

話者	時間（分）	単語数	略称
A01M0035	28	6127	AS22
A01M0007	30	4302	AS23
A01M0074	12	2486	AS97
A05M0031	27	5305	PS25

表４．２　模擬対話音声による音響モデルを用いた場合の講演音声の認識率

話者	音素正解精度	単語正解精度
A01M0035	64.6%	50.3%
A01M0007	60.5%	61.5%
A01M0074	70.3%	59.5%
A05M0031	65.1%	52.2%
average	64.4%	54.7%
T set	74.3%	76.6%

４．３　講演音声と対話音声，読み上げ音声の音響的な違い

４．２節において，対話音声により構築した音響モデルを用いた場合，講演音声に対しては認識性能が劣化することを明らかにした．本節では，講演音声が対話音声や読み上げ音声などと比較し，音響的にどのような違いがあり，その違いが認識性能に対してどのように影響するのかを調査する．

４．３．１　評価データ

講演音声の音響的特徴が対話音声や読み上げ音声とどのように異なるかを調査するにあたり，テストセットの話者 A01M0035 について，講演音声の発話内容と同一の読み上げ音声データを同一話者 A01M0035 から収録した．話者 A01M0035 の講演音声の書き起こしデータの内，講演内容の前半（3,272 単語）を使用し，間投詞，言い誤りもそのまま読み上げていただいた．読み上げ単位は，「話し言葉工学」プロジェクトより配布された書き起こしデータ内に記述されている時間情報を元に，人手で文らしくつなげた単位としている．

対話音声に関しては，講演内容と同一の対話音声を収録することが難しいため，ATR で収録された多数話者音声データベースの模擬対話音声データを用いることとした．

４．３．２　認識実験結果から見た特徴の違い

ここでは認識結果における認識誤りの傾向を調査することで，講演音声と読み上げ音声の違いについて考察する．

（1）実験条件

デコーダには，ATRSPREC を用いた．音響特徴パラメータは４．２．１と同じであり，音響モデルには対話音響モデルを用いている．

単語認識実験を行う場合，音響的要因と言語的要因の両方が含まれた認識結果が得られることとなる．しかしながら講演音声の発話様式に頑健な音響モデルを検討するためには，言語的要因による認識性能の劣化を極力排除することが望ましい．そこで本節ではテストデータ A01M0035 を対象にした音響的な分析を行うため，A01M0035 の発話内容のみで構築した単語バイグラムを言語モデルとして用いることとした．

認識辞書に関しては，2000 年に配布された「話し言葉工学」プロジェクトのモニタ版に付属されているものを用いるが，未知語による影響をなくすため，言い誤り以外の，A01M0035 の発話内容に出現する全ての単語を追加登録した．

（2）単語認識実験結果

講演音声データと読み上げ音声データについて，単語認識実験を行った．言語モデルが対応できないため，タスクの異なる対話音声についての認識実験はここでは行っていない．講演音声と読み上げ音声の単語認識実験結果を表４．３に示す．この結果から，講演音声より読み上げ音声の方が良好な認識率が得られていることがわかる．これは講演音声と比較し，読み上げ音声と音響モデルとのミスマッチが小さいためである．講演音声は自由発話音声であるが，音響モデルの学習データである模擬対話音声と比較するとその音響的特徴が大きく異なっている．

次に，話者適応を用いた場合の単語認識実験を行った．読み上げ音声，講演音声それぞれに対して，全ての発話データを用いたクローズドデータでの教師あり適応を，平均値についてはMAP-VFS[19]～[21]により，状態遷移確率についてはBaum-Welch法[22]により行っている．話者適応を用いた単語認識実験の結果を表４．３に示す．

話者適応を行った場合，読み上げ音声で91.8%の認識率が得られている．これに対し講演音声は，話者適応を行った場合でも認識率が 81.2%であり，読み上げ音声と比較すると劣化している．音響モデルが良好に働いた場合，本実験環境においてはその認識率の上限が 91%付近であると考えることができ，講演音声は話者適応だけでは十分な性能が得られていない．読み上げ音声で適応した音響モデルによる講演音声の認識，講演音声で適応した音響モデルによる読み上げ音声の認識においても，認識率の改善はほとんど見られない．話者性を適応するだけでは，十分な認識性能は得られないということであり，講演音声には発話様式の違いによる認識性能劣化の要因が含まれていると考えられる．

（3）音素認識実験結果

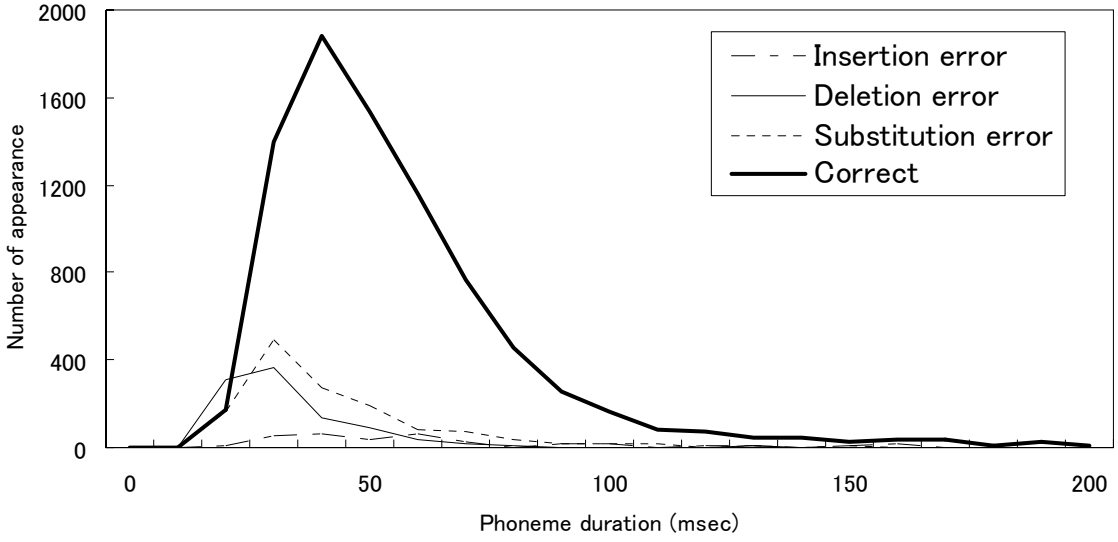
次に講演音声の音素認識実験を通して，音素継続時間長と認識誤りの関係を調査した．音響モデルのミスマッチによる要因を極力減らすため， 単語認識実験で用いたクローズド

表４．３ 模擬対話音声を用いて構築した音響モデル，及びクローズド教師あり話者適応を用いた場合の単語正解精度

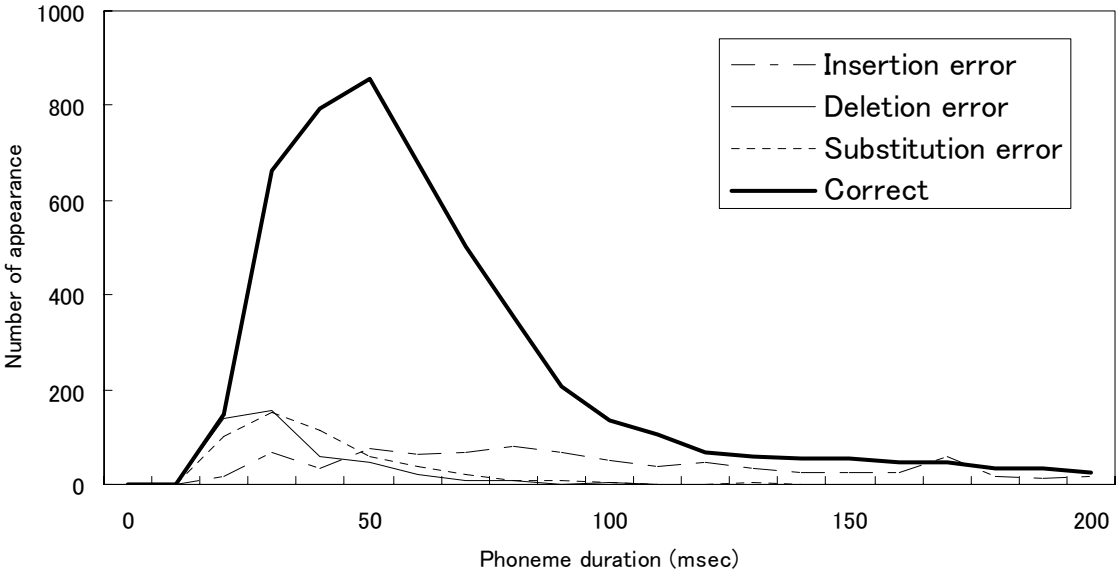
評価音声	対話 音響モデル	教師あり 適応モデル （読み上げ）	教師あり 適応モデル （講演）
読み上げ	81.9%	91.8%	59.8%
講演	74.3%	76.6%	81.2%
※言語モデルにはクローズドモデルを使用			

教師あり話者適応モデル（講演）を使用している．

図４．１に講演音声において誤りが生じた音素の，継続時間長ごとの分布を示す．この図から，脱落・置換誤りは音素継続時間長の短い領域で多く発生していることがわかる．挿入誤りについては，母音に関しては全体的に生じているが，子音に関しては継続時間長



― 子音 ―



― 母音 ―

図４．１音素継続時間長ごとの誤りの分布

の短い領域で多く発生している．

継続時間長の短い領域に誤りが多く発生する原因としては，音響モデル学習データである模擬対話音声の音素継続時間長分布が，講演音声の分布とずれており，継続時間長の短い音素に対する学習データが不足していることが挙げられる．また，音響モデルが 3 状態の left-to-right モデルであり, 分析周期が 10msec という制約上, 音素継続時間長が 30msec 未満の音素に対しては，モデルの構造そのものがミスマッチを生じていると考えられる．さらに，分析窓長が 20msec のため，継続時間長が 20msec 程度の音素に対しては，周波数分析の時点で分析精度が劣化している可能性がある．

４．３．３　音素継続時間長における違い

講演音声データは，読み上げ音声や模擬対話音声と比較し，発声速度が速くなる傾向がある．そこで，講演音声の音素継続時間長が，読み上げ音声や音響モデル学習データである模擬対話音声と比較し，どの程度異なっているかを分析した．分析は，ビタビ・アライメントによりそれぞれの音素区間を判定し，継続時間を算出している．音響モデルとのミスマッチによるアライメントのずれを極力減らすため，読み上げ音声，講演音声に対しては，全ての発話データを用いた教師あり話者適応を行った．話者適応は，MAP-VFS を用いて平均値を，Baum-Welch 法を用いて状態遷移確率を適応している．

図４．２に講演音声，読み上げ音声，模擬対話音声それぞれに含まれる各音素の継続時間長分布を示す．この図より講演音声は，音素継続時間長のピークが 30～40msec 付近に存在するとともに，模擬対話音声の分布から左（継続時間長が短くなる方向）にずれていることがわかる．読み上げ音声，模擬対話音声の音素継続時間長のピークはそれぞれ 60msec，50msec とずれているが，講演音声と比較すると，継続時間長分布は比較的近いと言える．

表４．４に講演音声，読み上げ音声，模擬対話音声それぞれの音素継続時間長の平均値を示す．講演音声は，読み上げ音声や模擬対話音声と比較し，各音素の平均継続時間長が短くなっていることがわかる．

表４．４ 講演音声，読み上げ音声，及び模擬対話音声における
平均音素継続時間長

評価音声	母音	子音	全体
講演音声	65.7msec	54.3msec	60.3msec
読み上げ音声	87.6msec	73.0msec	80.7msec
模擬対話音声	79.2msec	60.2msec	70.4msec

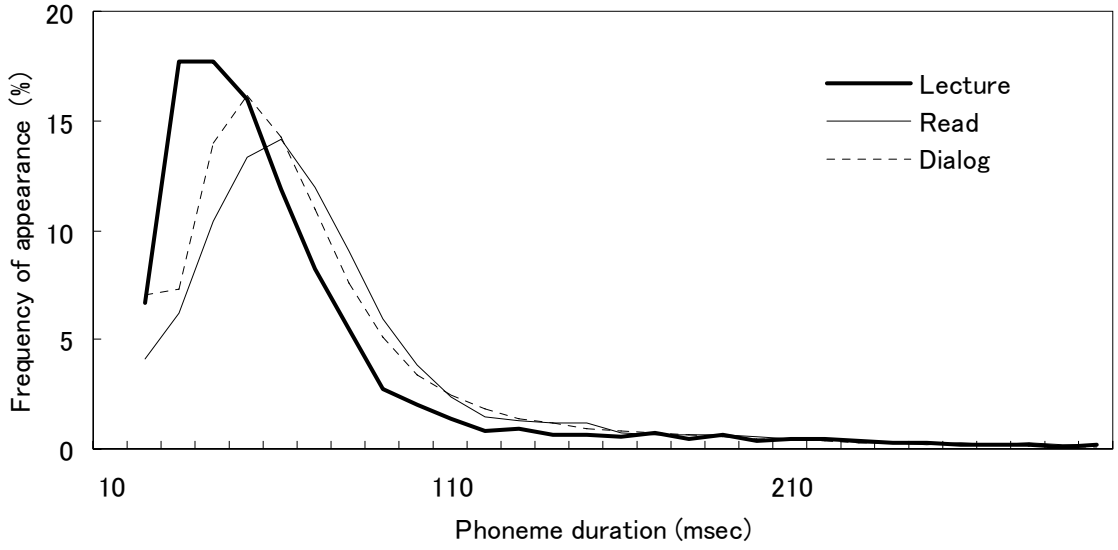


図４．２ 各音声データの音素継続時間長の分布

４．３．５　周波数領域における違い

ここでは周波数領域において，講演音声が読み上げ音声や模擬対話音声と，どの程度異なっているかを分析した．分析は，A01M0035 の読み上げ音声，模擬対話音声との比較により行なった．分析において第１，第２共振周波数 F1，F2 は各母音セグメント（ビタビ・アライメントにより判定）の中央点において算出した．

図４．３に講演音声, 読み上げ音声, 模擬対話音声における母音の F1-F2 の分布を示す．F1-F2 の分布は同一話者にも拘わらず，講演音声と読み上げ音声の間で変動しており，また音響モデル学習データである模擬対話音声とも大きく異なっていることがわかる．４．３．２の認識実験において，講演音声による話者適応モデルの認識性能が，同一話者の読み上げ音声に対して大きく劣化している要因も，これらの周波数的特徴の変動によるものである．この変動は発話様式の違いによって生じており，話者適応による話者性だけではなく発話様式への適応も必要となる．

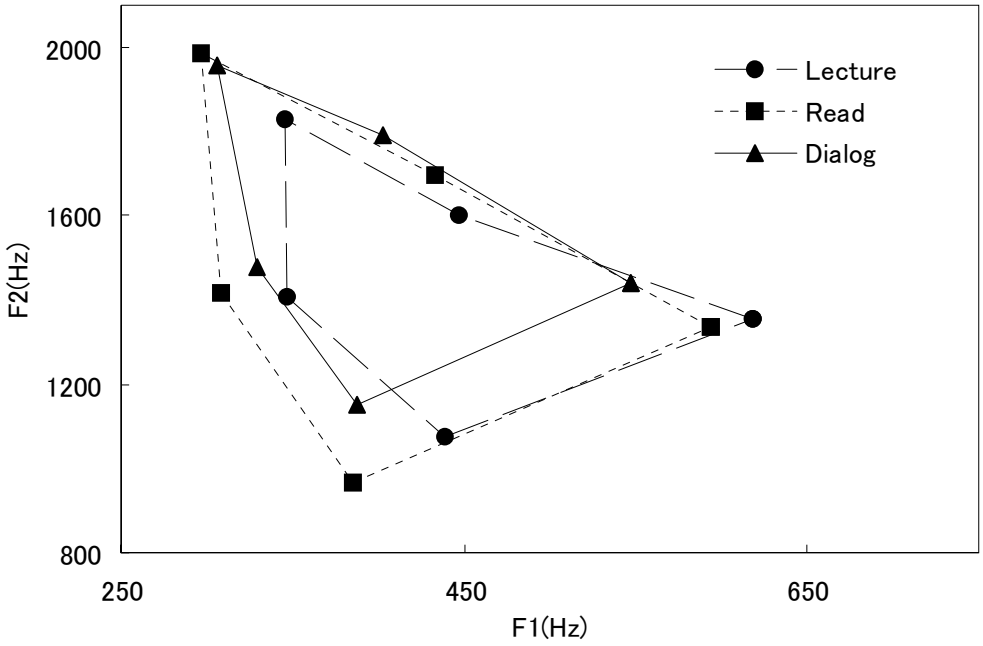


図４．３　講演音声，読み上げ音声と模擬対話音声の母音平均フォルマント周波数 F1-F2 の分布

４．４　講演音声認識のための音響モデル

前節までの結果より，A01M0035 の講演音声は，(1)発話速度が速くなるため，音素継続時間長の分布が学習データとずれていると共に，継続時間長の短い音素が多く出現する，(2)周波数領域における特徴空間において，音響モデルの学習データとのずれが存在する，といったことが明らかになった．これらを踏まえ，A01M0035 を用いて講演音声に頑健な音響モデルの構築方法を検討する．

４．４．１　音響モデルによる発話速度変動のモデル化

４．３節の結果より，講演音声では継続時間長が短い音素において，脱落・置換誤りが多く発生していることがわかった．これは学習データにおいて，継続時間長の短い音素データが量的には十分出現しているが，全体的には出現頻度が低いため，統計的アプローチで構築する HMM では，その領域の認識性能が劣化したためと考えられる．そこで，継続時間長の短い音素セグメントのみを用いた音響モデルの構築を検討する．認識の際，継続時間長の短い音素の出現位置はわからないため，対話音響モデルと音素継続時間長の短い音響モデルを一つの音響モデルに統合する[23]．図４．４に示すように，一つの音素環境に対して２つのモデルを定義し，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ものとした．デコードの結果，ゆう度の高い経路が選択されるため，継続時間長の短い音素の出現位置を事前に知る必要がなく，また認識辞書の変更も不要となる．

図４．２より，脱落誤り，置換誤りの分布のピークが 30msec に存在するため，継続時

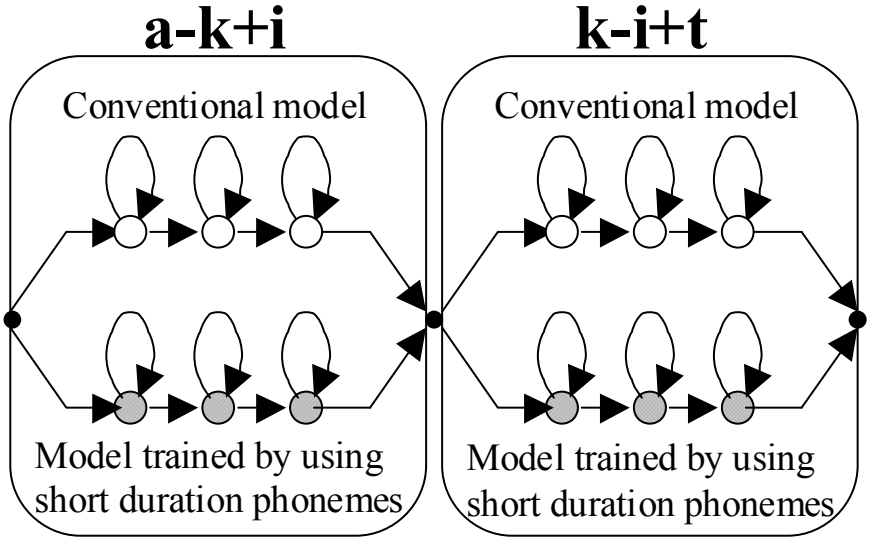


図４．４複数のモデルを統合した音響モデル

間長が 40msec 以下の音素セグメントを用いて音響モデルを構築した. 構築した音響モデルは，総状態数 1,400，５混合ガウス分布とした．このモデルと対話音響モデルを統合したモデルを，音響モデルＳと呼ぶこととする．

表４．５に，音響モデルＳによる単語正解精度を示す．継続時間長の短い音素セグメントで構築したモデルを統合することで，単語正解精度が 3.1%向上した．提案手法では挿入誤りは増加するものの，脱落，置換誤りが減少していることから，継続時間長の短い音素セグメントで構築したモデルが効果的に働いていることがわかる．

４．４．２　発話速度に合わせた分析周期・窓長の最適化

講演音声の発話速度を正規化し, 学習データとの音素継続時間長分布を近づけると共に, 前処理やモデルの構造に適した発話速度にすることで，認識性能の改善が期待できる．発話速度の正規化に関しては，発話速度に応じて分析周期を変更する方法[24]や，特徴パラメータの間引き，相関による補間を用いた方法[25]などが提案されている．しかしながら前者は発話速度が遅い発声に対しては効果が得られているが，発話速度の速い発声に対しては逆に劣化するという結果が得られている．また後者の手法は，発話速度の速い発声に対しても効果が得られているが，講演音声のように分析窓長より短い音素が多く出現する場合，前処理における周波数分析の精度が低くなるため，十分な性能が得られない可能性がある．前者の手法で，発話速度の速い発声に対して認識率が劣化する原因としても，周波数分析精度の劣化が影響している可能性がある．

そこで，分析周期と分析窓長を合わせて変更することとした．発話速度が速い場合，分析周期を短くすることで時間方向の分解能は向上する．しかしながら継続時間長の短い音素に対しては，分析窓長が長すぎる場合，一つの音素内の周波数変化に対する周波数分析精度が劣化する．このため，分析周期と合わせて分析窓長も短くする必要がある．表４．６に，分析周期，分析窓長をそれぞれ変更した場合の単語正解精度を示す．この結果より，分析周期８msec，分析窓長 16msec としたものが最も良い結果となっている．これは分析周期 10msec，分析窓長 20msec と比較するとそれぞれ 1.25 倍となっており，学習データ

表４．５ 音響モデルＳによる単語正解精度と誤りの傾向

	単語正解精度	挿入誤り	脱落誤り	置換誤り
対話音響モデル	74.3%	3.2%	8.1%	14.4%
音響モデルS	77.4%	3.7%	5.9%	13.0%

と講演音声データの音素継続時間長平均値の比率が 1.17 (= 70.4 msec/60.3msec)であることから，全体的にマッチしたものと考える．分析周期９msec，分析窓長 18msec より認識率が向上しているのは，継続時間長が短い音素に対する周波数分析精度が向上したためである．

次に，音響モデルＳと分析周期・窓長の変更を併用した場合の認識実験を行った．表４．７に結果を示す．さらに認識率が向上し，対話音響モデルと比較し単語正解精度で 5.7%，認識誤り率で 22.2%改善している．

４．４．３　教師なし話者適応の導入

４．３節において，学習データと講演音声の間で，F1-F2 の分布が異なっていることを述べた．この結果を踏まえ，教師なし話者適応を用いた評価実験を行った．教師なし話者適応は，講演音声を認識後，その認識結果を正解発話内容として，全てのデータを用いた

表４．６ 分析周期，分析窓長を変更した場合の単語正解精度

フレーム周期	分析窓長	単語正解精度
10msec	20msec	74.3%
9msec	20msec	77.1%
8msec		76.2%
7msec		76.2%
6msec		75.0%
10msec	18msec	74.9%
	16msec	75.3%
	14msec	75.5%
	12msec	73.9%
9msec	18msec	75.9%
8msec	16msec	**77.3%**
7msec	14msec	75.9%
6msec	12msec	76.7%

表４．７ 音響モデルＳと分析周期・分析窓長変更の併用による単語正解精度

フレーム周期	分析窓長	単語正解精度
10msec	20msec	77.4%
9msec	18msec	79.2%
8msec	16msec	**80.0%**
7msec	14msec	79.0%
6msec	12msec	77.5%

MAP-VFS による適応を行うものとした. 評価実験は, 分析周期・窓長を変更しないものと, それぞれ８msec，16msec に変更したものを用い，対話音響モデル，音響モデルＳそれぞれに対して行った.

結果を図４．５に示す．比較として，４．３．２で得られた教師あり話者適応の結果も示す．分析周期・窓長を８msec，16msec に変更し音響モデルＳを用いたものが最も良く，教師あり話者適応よりも良好な結果となっている．これらの結果より講演音声を認識するには，発話速度を考慮した前処理，音響モデルの構築が有効であり，話者適応によりさらに認識率が改善されることが示された.

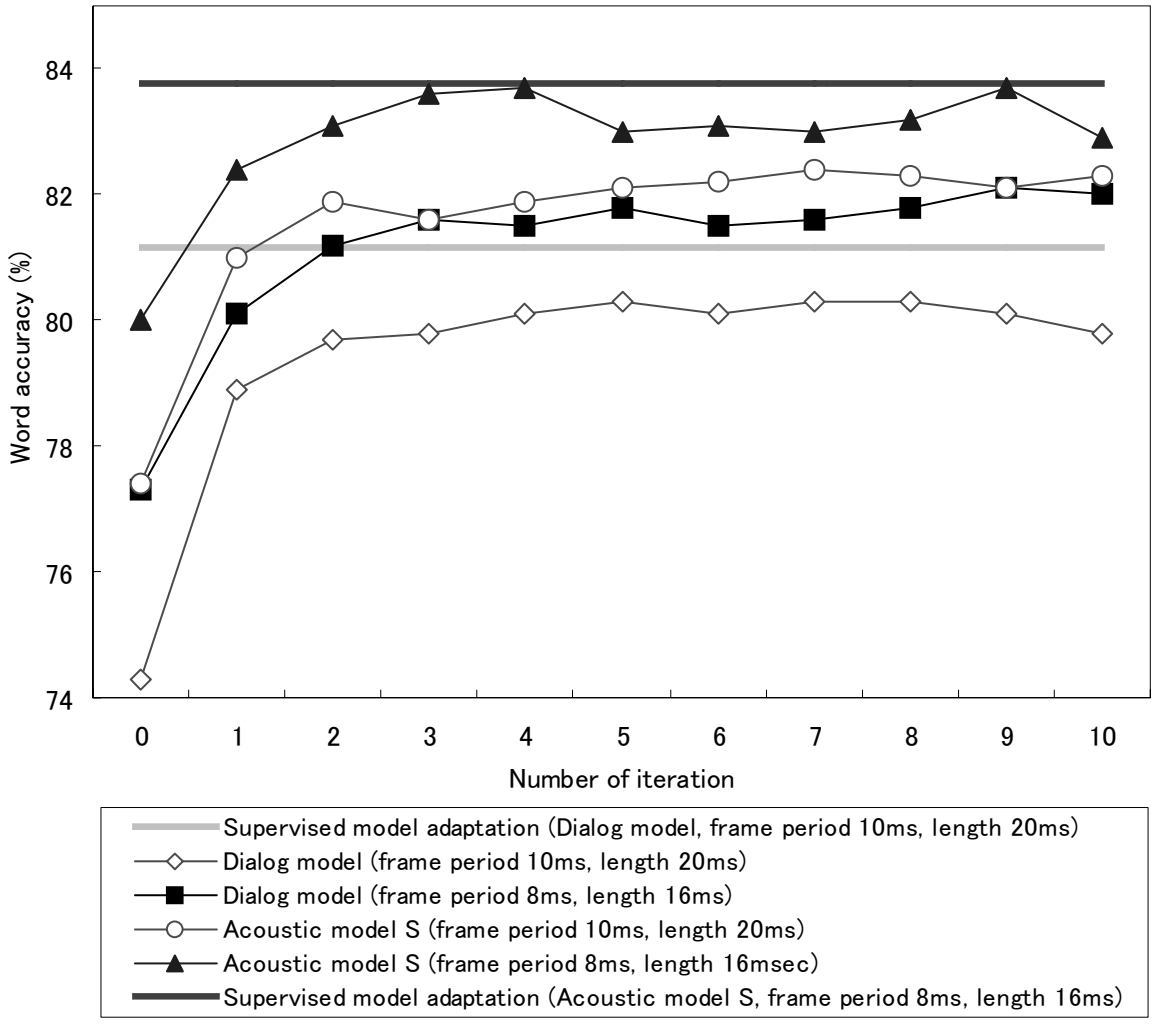


図４．５ 教師なし話者適応を用いた場合の，各分析周期・窓長における
対話音響モデル，音響モデルＳの単語正解精度

４．５　発話速度に合わせた分析周期・窓長選択手法の一般化

４．５．１　実験条件

ここで用いる評価セットには，「話し言葉工学」プロジェクトより配布されている「日本語話し言葉コーパス（CSJ）」モニタ版（2001）[11]から，表４．８に示す男性話者 10 名（最後の３名は，モニタ版に含まれていない.）を選択した．評価実験の際には，モニタ版に含まれる書き起こしデータに記述されている時刻情報をもとに，発話単位にファイルを分割している．

ここで用いたベースラインの認識システムの概要は，次の通りである．音響特徴パラメータは，16kHz サンプリング，分析周期 10msec，分析窓長 20msec の条件で抽出した 25 次元の特徴ベクトル（12 次 MFCC，12 次Δ MFCC，1 次Δlog power）を用いている．

音響モデルは，状態共有化 HMM（HMnet）により構築された性別依存モデルであり，各音素モデルは３状態（状態の飛び越し遷移なし），10 混合ガウス分布，総状態数 1,400 で表現されている．なお，本章で用いた評価話者は全て男性話者のため，モデルは男性モデルのみを用いている．

音響モデルの学習データには，モニター版のうち評価話者を除く全ての学会講演，模擬講演男性話者データ 200 名（約 34 時間）のデータを用いた．なお音響モデルの学習は，配布された書き起こしデータに記述されている時刻情報，カタカナ書き起こし情報を元に行っている．本節において，この音響モデルをベースライン音響モデルとする．

言語モデルは京都大学で作成され，モニタ版と共に配布された講演音声用言語モデル[9]の前向き単語バイグラム，後ろ向き単語トライグラムを用いている．認識辞書に関しても同様で 19k 単語のものを用いている．デコーダには Julius3.2[12]を使用した．

表４．８「日本語話し言葉コーパス(CSJ)」モニタ版(2001)に含まれる評価セット

話者	時間（分）	単語数	略称
A01M0035	28	6127	AS22
A01M0007	30	4302	AS23
A01M0074	12	2486	AS97
A05M0031	27	5305	PS25
A02M0117	57	9858	JL01
A03M0100	15	2161	NL07
A06M0134	23	4467	SG05
KK99DEC005	42	6557	KK05
YG99JUN001	14	2764	YG01
YG99MAY005	15	2939	YG05

４．５．２　評価セットにおける発話速度と認識性能の関係

発話速度と認識性能の関係を調査するために，ベースライン音響モデルを用いた認識実験を行った．図４．６に，各話者の単語誤り率と平均発話速度をまとめる．図中における各話者の平均発話速度は，文ごとに算出した１秒あたりのモーラ数の平均を示している．１秒あたりのモーラ数は，ビタビ・アライメントにより算出した音声区間の時間長で文中のモーラ数を割った値である．講演音声においてはフィラーの出現頻度が多いため，このようにして算出した発話速度が必ずしも正確な値とはなっていないが，各話者の傾向を知ることはできる．

単語誤り率は，話者によるばらつきが大きいが，平均で 33.9%となっている．図４．６より発話速度の速い話者ほど，単語誤り率が増加する傾向にあることがわかる．図４．７に，各話者の単語誤り率と平均発話速度の分布を示す．単語誤り率と平均発話速度は，相関係数 0.78 と高い相関を示している．また，ベースライン音響モデルは，各音素モデルを３状態で表現しているため，状態の飛び越し遷移を許さない場合，30msec 未満の音素に対しては十分な認識性能が得られない. 図４．８に, 各話者の単語誤り率と継続時間長 30msec 以下の音素セグメントの，出現頻度の分布を示す．30msec 以下の音素出現頻度は，状態飛び越し遷移を許したビタビ・アライメントにより抽出した各音素の継続時間長より算出した．単語誤り率と 30msec 以下の音素出現頻度は，相関係数 0.83 とさらに高い相関を示している．

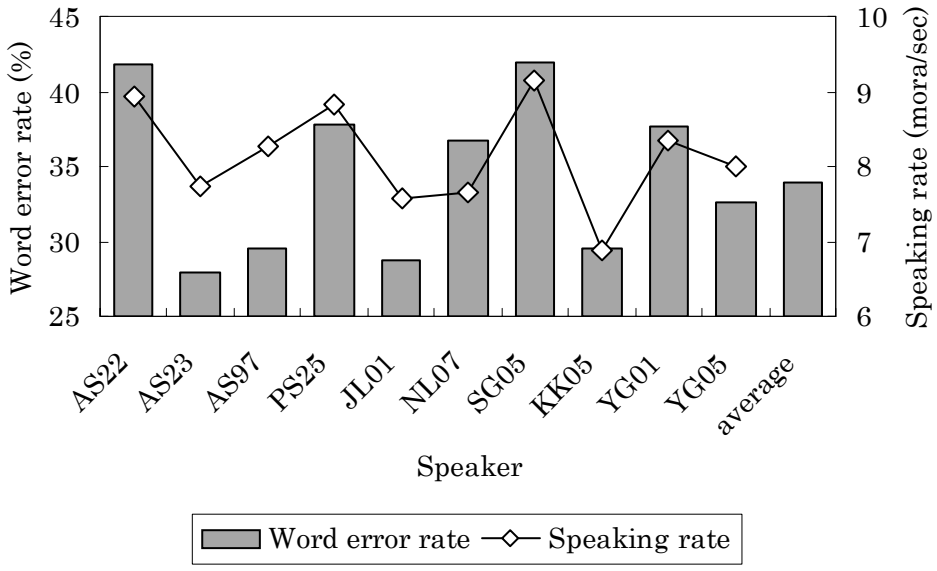


図４．６ ベースライン音響モデルによる単語誤り率と平均発話速度

以上の結果より，発話速度の速い発声への対応が重要であり，音響的特徴の変動への対応や，発話速度の正規化や補正が必要であることがわかる．本節では特に後者の問題について取り扱う．

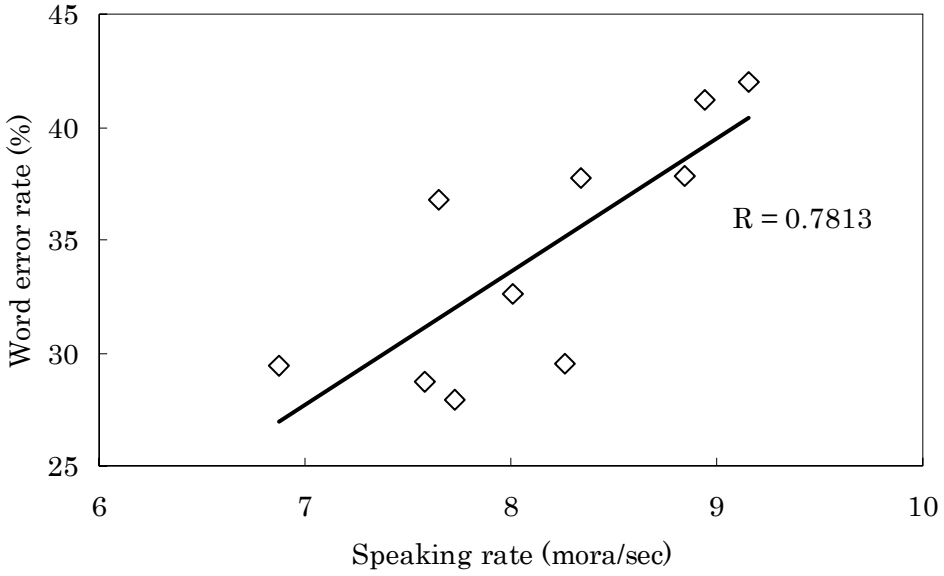


図４．７ 単語誤り率と平均発話速度の関係

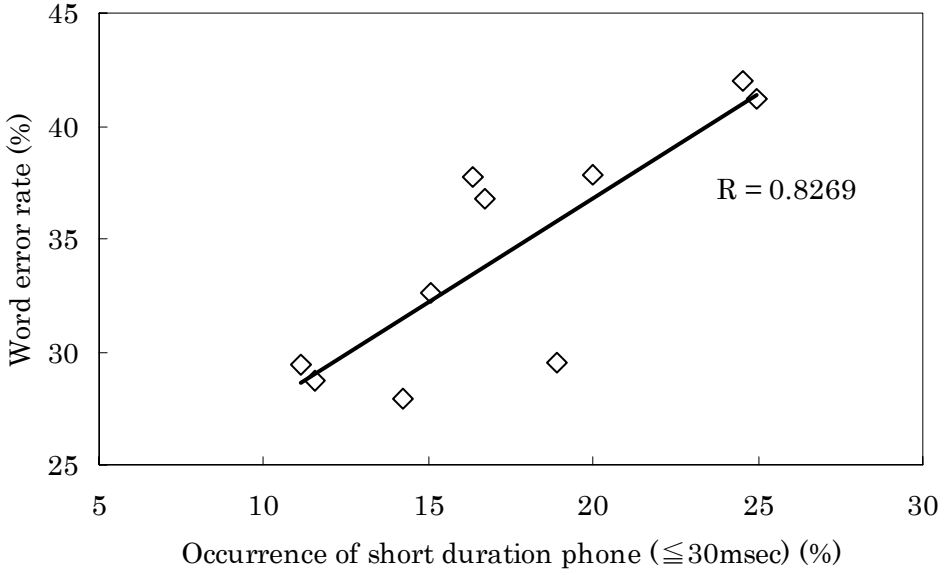


図４．８ 単語誤り率と 30msec 以下の音素出現頻度の関係

４．５．３　分析周期・窓長の変更による発話速度の補正

発話速度の正規化や補正を行う手法として，フレームの削除や相関を用いた挿入を行う手法[25]が提案されている．この研究では，継続時間長の長い音素に対してフレームを間引くとともに，継続時間長の短い音素に対しては Missing Feature Theory[26]により前後のフレームとの相関を用いて合成したフレームを挿入することで，1 音素当りのフレーム数を正規化している．ただし，音素境界が既知の場合には効果が得られているものの，フレーム挿入位置の推定精度に性能が大きく影響されるため，音素境界が未知の場合には効果が得られていない．

他の手法として，発話速度に応じて分析周期を変更する方法[24]が提案されている．この研究では，基準となる分析周期の認識結果に対して，分析周期を変更した場合の音響ゆう度を算出・比較することで最適な分析周期を決定し，その分析周期により再び認識を行うことで，最終的な認識結果を得ている．分析周期を変更すると分析フレーム数が異なるため，各分析周期の音響ゆう度を直接比較することはできない．そこでこの研究では，基準となる分析周期で学習した GMM（Gaussian Mixture Model）による音響ゆう度との差が最も大きいものを選択することとしている．ただし，発声速度の遅い発話に対して効果が得られているものの，発話速度の速い発声に対しては認識精度の改善が得られていない．この要因としては，基準となる分析周期での認識結果を分析周期の決定に用いていること，分析周期決定の処理と実際の認識処理が一貫していないことに加えて，分析窓長が固定であることが挙げられる．

分析周期の変更に関しては，ケプストラムの変化量に応じて発話内で分析周期を可変とする手法[27]や，特徴量抽出間隔より短い間隔の静的特徴量を用いて動的特徴量を推定する手法[28]が提案されているが，いずれも雑音に対する認識性能の改善を目的としたものである．そこで本研究ではまず，発話速度に応じて分析周期と分析窓長の両方を変更することを検討する．

発話速度の速い発声においては，分析周期を短くすることで音響モデルが表現できる時間構造に近づけることができる．しかしながら，周波数領域における時間変化が大きい場合，分析窓長が長すぎるとその区間の特徴を十分に表現できない．４．４節の結果より，分析周期に合わせて分析窓長を変更することで認識性能の改善が確認された．ここでは分析周期・窓長のみを変更し，音響特徴パラメータ抽出の際の，他の条件は変更していない．動的特徴を示すΔパラメータに関しても，明示的な補正は行っていない．分析周期を短くすることによりΔパラメータ抽出の時間間隔が短くなり，補正の効果が得られるためである．

図４．９に，ベースライン音響モデルを用いて，認識の際の分析周期・窓長をそれぞれ，（10msec・20msec），（９msec・18msec），（８msec・16msec）の組み合わせで変更した場合の認識結果を示す．比較として，分析周期 10msec，窓長 20msec において状態の飛び越し遷移を用いた場合の結果も示す．この図より，平均発話速度の比較的速い話者においては，分析周期 8msec，窓長 16msec で分析することで認識性能が改善され，比較的遅い話者においては，分析周期，窓長の長い方が，認識性能が良くなる傾向にあることがわかる．状態の飛び越し遷移は，話者によっては効果があるが，平均認識率で見た場合では，特に性能の改善は見られなかった．

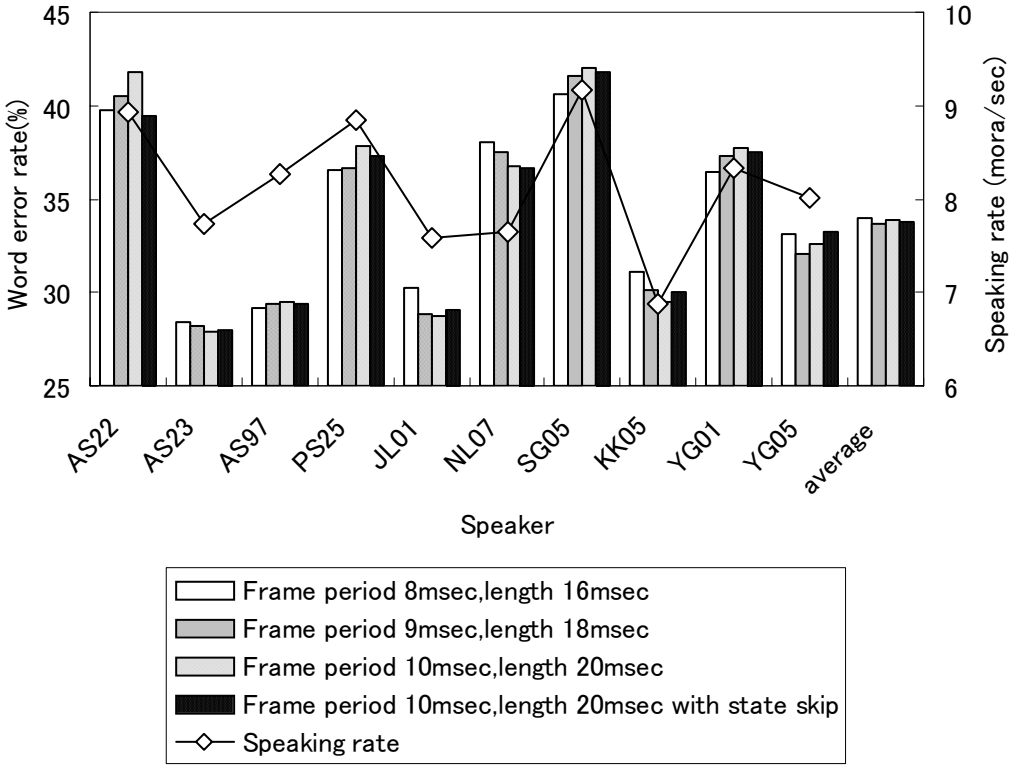


図４．９　認識時における分析周期・窓長を変更した場合の単語誤り率と平均発話速度

４．６　ゆう度基準による分析周期・窓長の自動選択を用いたデコーディング手法

前節の結果より，発話速度に応じて分析周期・窓長を変更することで，発話速度の補正の効果があることが確認できた．しかしながら，自動認識においては最適な分析周期・窓長を自動選択する必要がある．また，発話速度は話者による違いだけではなく，同一話者の講演内においても変動しているため，発話ごとに補正を行う必要があるが，各発話の発話速度は事前にはわからない．発話速度の推定を行うことも考えられるが，発話速度の正規化や補正の効果がその推定精度に大きく影響される．

そこで本節では，発話速度の推定を行うことなく，ゆう度基準により発話ごとに適した分析周期・窓長を選択するデコーディング手法を提案する．具体的には，複数の分析周期・窓長を用いていったん認識を行い，得られた認識結果における音響ゆう度・言語ゆう度を比較することで，認識対象となる発話に適した分析周期・窓長を事後的に選択するものである．異なる分析周期・窓長により認識を行うため，異なるマッチングパスが得られることとなり，各認識結果における音響ゆう度，言語ゆう度もそれぞれ異なったものとなる．以下，分析周期・窓長の選択に音響ゆう度のみを用いた場合と，音響ゆう度・言語ゆう度の両方を用いた場合について述べる．

４．６．１　音響ゆう度を用いた選択手法

４．５．３で述べたように，分析周期を短くした場合，一発話当りのフレーム数が多くなるため，音響ゆう度をそのまま比較することはできない．そこで，分析周期により正規化した音響ゆう度を比較する．認識においては対数音響ゆう度を用いており，認識結果の対数音響ゆう度は各フレームの対数音響ゆう度の総和として得られる．そこで対数音響ゆう度の正規化は，式(4.1)に示すように，フレーム数の比で除する（＝分析周期の比を乗じる）ことにより行う．式(4.1)において分母が 10 となっているのは，分析周期 10msec を基準としているためである．

$$AM' = \frac{AM \times Frame\_period(m\sec)}{10} \tag{4.1}$$

AM : 発話ごとの対数音響ゆう度

AM': 分析周期により正規化した対数音響ゆう度

発話ごとに，式(4.1)により算出した対数音響ゆう度 AM' を比較し，最も AM' の大きい分析周期・窓長の認識結果を選択する．ベースライン音響モデルを用いた場合の，各分

析周期の単語誤り率と，音響ゆう度を用いた選択手法による単語誤り率を図４．１０に示す. 音響ゆう度を用いた選択デコーディング手法により, ベースライン（分析周期 10msec, 窓長 20msec）から，平均単語誤り率が 0.4%しか改善していない．話者ごとに見た場合，発話速度の比較的速い話者においては，単語誤り率が改善されているが，発話速度の比較的遅い話者においては逆に低下していることがわかる.

４．６．２　音響ゆう度・言語ゆう度を用いた選択手法

連続音声認識においては，音響ゆう度に言語ゆう度を加えたスコアを用いて認識結果を探索する．このため，探索過程において音響ゆう度のみが大きくなる仮説を選択した場合，言語ゆう度が低くなっている認識結果も存在する．そこで，分析周期で正規化した音響ゆう度に言語ゆう度（挿入ペナルティを含む）を加えたスコアを再計算し，この値が最も高い分析周期・窓長の認識結果を選択することとした．ベースライン音響モデルを用いた場合の，各分析周期の単語誤り率と音響・言語ゆう度を用いた提案手法の単語誤り率を図４．１１に示す.

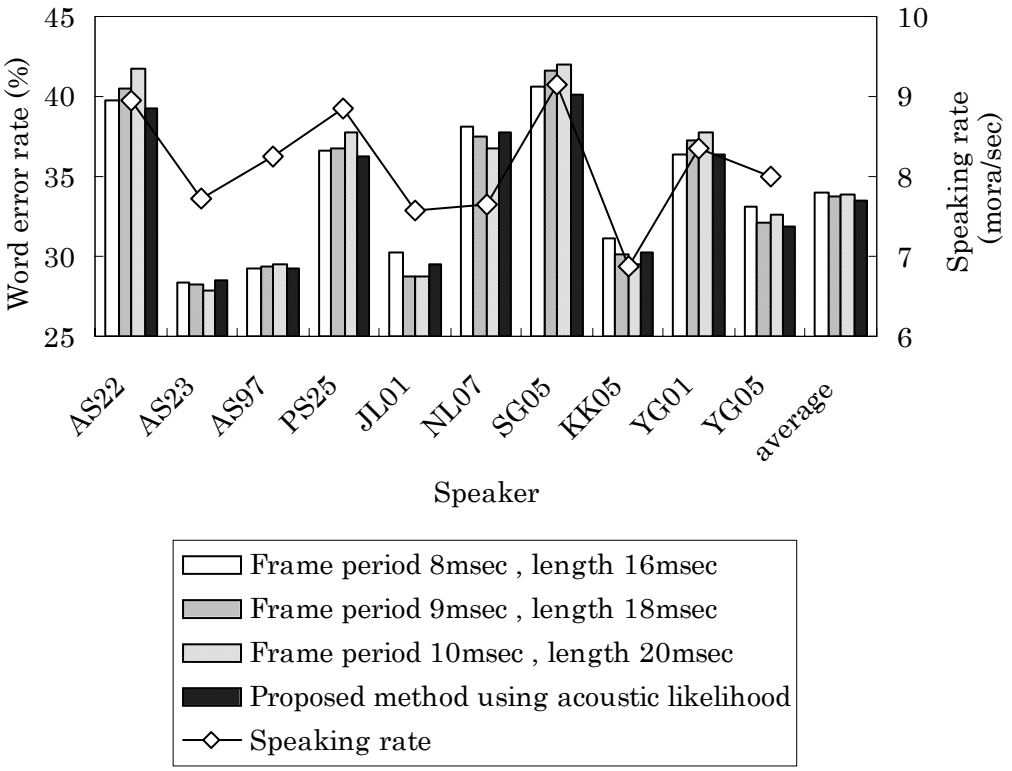


図４．１０　音響ゆう度を用いた選択デコーディング手法の単語誤り率

この結果より，平均単語誤り率はベースラインとなる分析周期 10msec，窓長 20msec の場合と比較して 0.8%改善した．この改善の度合いは，有意水準１％で有意である．発話速度の速い３名の話者 （AS22, PS25, SG05） に関しては，それぞれ 2.1%，2.1%，1.9%，平均で 2.0%認識性能が改善している．また，発話速度の遅い話者に対する認識性能の低下も小さくなっている．以上より，音響・言語ゆう度を用いた提案デコーディング手法は，発話速度の変動の大きい講演音声認識に対して有効である．以降，音響・言語ゆう度を用いた提案デコーディング手法を音響・言語ゆう度ベース選択手法と呼ぶ.

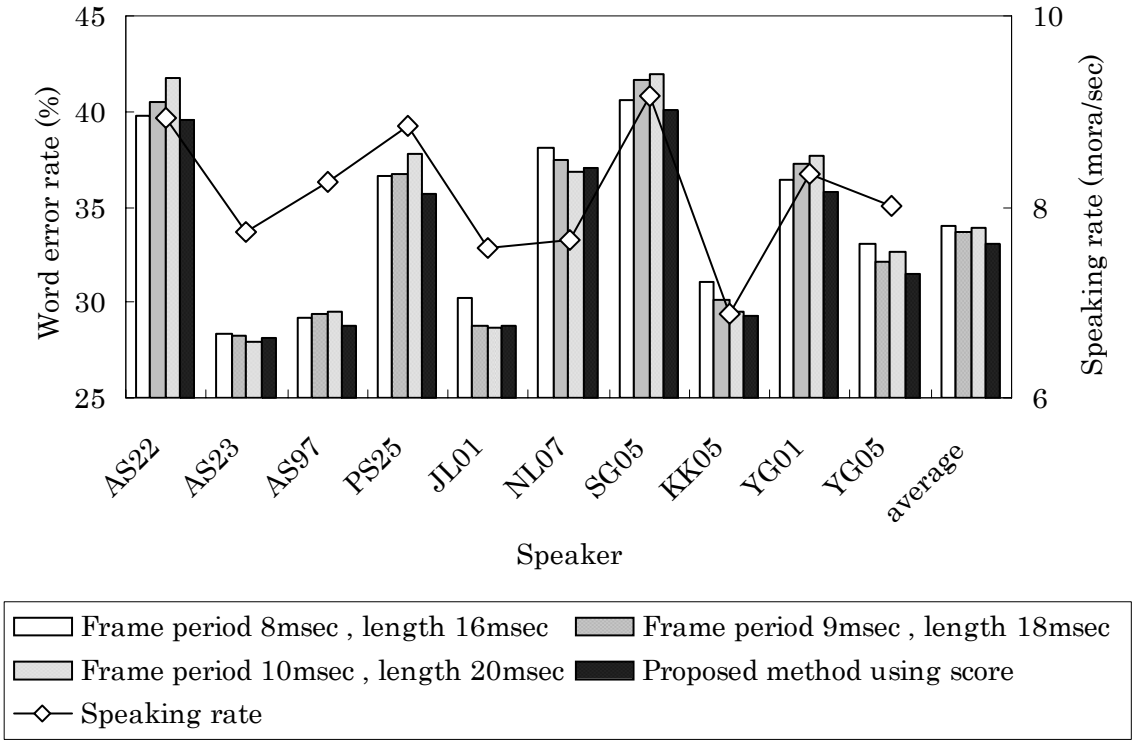


図４．１１　音響・言語ゆう度を用いた選択デコーディング手法の単語誤り率

４．７　音響モデル学習時における発話速度の補正

ここではさらに，ゆう度基準による分析周期・窓長の自動選択手法を，音響モデルの学習時に適用することを提案する．

４．７．１　音響モデル学習データの分析周期・窓長の選択方法

認識時においては，分析周期で正規化した対数音響ゆう度と言語ゆう度を組み合わせて選択する手法が効果的であった．これに対して，音響モデル学習データは発話内容が既知であるため，それぞれの分析周期・窓長により抽出した特徴パラメータに対してビタビ・アライメントを行い，算出された対数音響ゆう度を用いて選択することとした．ビタビ・アライメントはベースライン音響モデルを用いて行い，分析周期選択の際の音響ゆう度正規化は，認識時と同様に式(4.1)により行った．図４．１２に，本手法により選択された分析周期に対する学習データ量（発話数）を示す．この学習データにおいては，分析周期９msec，窓長 18msec の分析条件が最も多く選択されていることがわかる．

このようにして選択した分析周期・窓長を用いて，分析周期・窓長以外はベースライン音響モデルと同一の条件で音響特徴パラメータを抽出し，音響モデルを構築する．選択された分析周期・窓長ごとに音響モデルを構築することで，認識性能の改善が期待できるが，一発話内で発話速度が変化するような場合，逆に認識性能が低下することも考えられる．

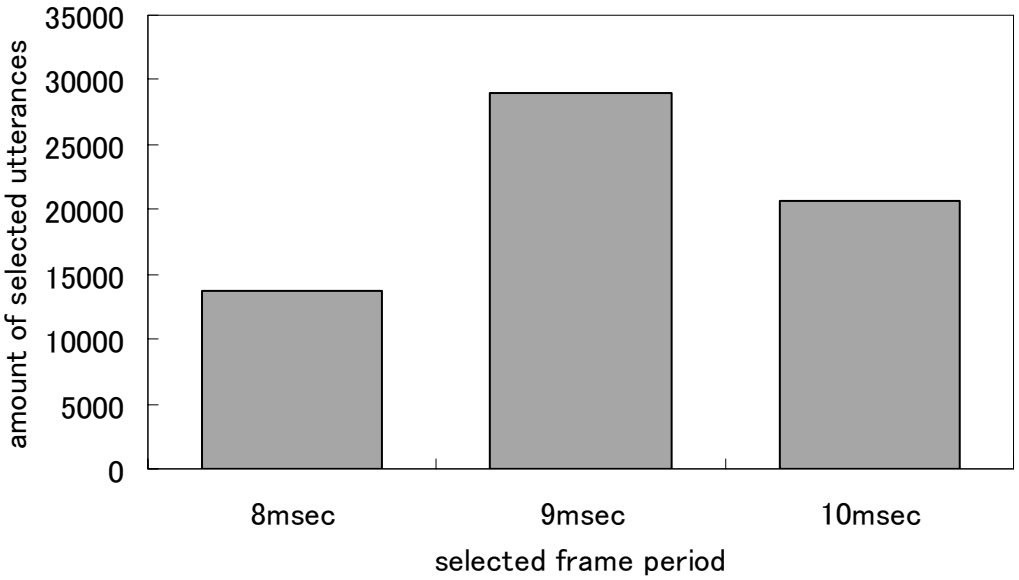


図４．１２　分析周期により正規化した対数音響ゆう度により選択された
分析周期・窓長の発話数

そこで，以下の２種類のモデルを構築した．

(1)抽出した全ての特徴パラメータを用いて一つの音響モデルを構築（single model）
(2)選択された分析周期・窓長ごとに音響モデルを構築（multiple models）

音響・言語ゆう度ベース選択デコーディングを行う際，single model は全ての分析周期・窓長で用いるが，multiple models は，各分析周期・窓長に対して該当するモデルを用いる．

４．７．２　認識実験結果

図４．１３に，ベースライン音響モデルによる音響・言語ゆう度ベース選択手法，発話速度の補正を行った音響モデルによる音響・言語ゆう度ベース選択手法の単語誤り率を示す．

この結果より，発話速度の補正を音響モデル学習時にも行うことで，認識性能がさらに改善されることがわかる．また，発話速度を補正した学習データで一つの音響モデルを構築する場合(single model)より，選択された分析周期・窓長ごとに音響モデルを構築した場

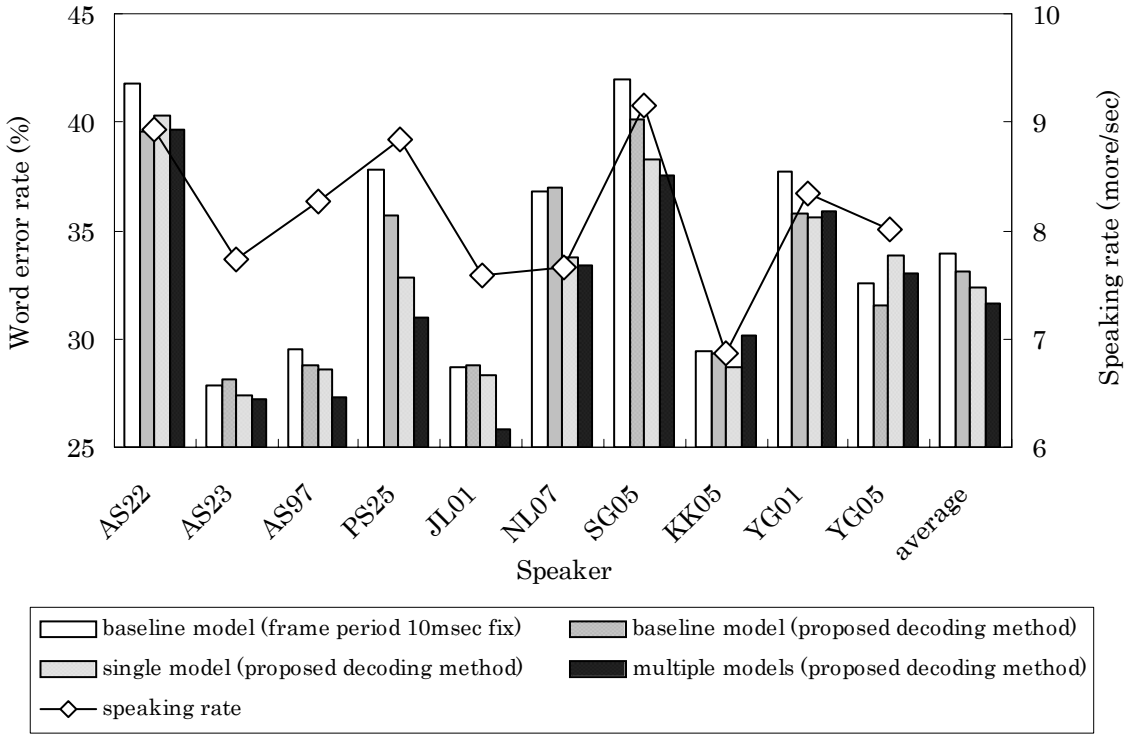


図４．１３　各音響モデルを用いた音響・言語ゆう度ベース選択手法による単語誤り率

合（multiple models）の方が，認識性能が高いことがわかる．この場合で，ベースライン音響モデルで提案手法を用いない場合（分析周期 10msec, 窓長 20msec 固定）と比較し, 2.3%の改善となった.

４．８　考察

ここでは，発話速度の補正を行った音響モデルの，提案デコーディング手法における効果について考察する．

表４．９に，発話速度の補正を行った音響モデル（single model）を用いた場合の，各分析周期の単語誤り率を示す．発話速度を補正した学習データで一つの音響モデルを構築した場合は，各分析周期における認識性能の差が小さい．

表４．１０に，発話速度の補正を行った音響モデル（multiple models）を用いた場合の，各分析周期の単語誤り率を示す．表４．９と比較すると，各分析周期における認識性能の差が大きくなっていることがわかる．multiple models は選択された分析周期ごとにモデルを学習することで，発話速度の違いによる時間構造以外の音響的特徴の違いもある程度モデル化できるためである．

図４．１２からわかるように，multiple models の各モデルは single model と比較し学習データ量が少なくなっているにも拘わらず，認識性能が向上したことからも，分析周期ごとに音響モデルを構築することは有効である．また，分析周期ごとに音響モデルを構築することで，各分析周期における音響ゆう度差が大きく現れるため，音響・言語ゆう度ベースの選択手法も効果的に働いている．

次に，single model と multiple models の両方を用いて，音響・言語ゆう度ベース選択手法を行った場合の評価実験の結果を図４．１４に示す．この図より，single model と multiple models の両方を用いた提案手法において，最も認識性能が良く，ベースラインと

表４．９　発話速度補正音響モデル（single model）の各分析周期の単語誤り率

略称	分析周期			音響・言語ゆう度ベース選択手法
	8msec	9msec	10msec	
AS22	40.5%	41.9%	43.1%	**40.3%**
AS23	28.2%	28.4%	28.7%	**27.4%**
AS97	29.0%	29.2%	30.9%	**28.6%**
PS25	**32.8%**	33.7%	34.7%	32.8%
JL01	29.4%	28.9%	38.9%	**28.3%**
NL07	34.6%	34.3%	34.3%	**33.8%**
SG05	**38.1%**	39.9%	41.8%	38.2%
KK05	30.2%	29.3%	29.2%	**28.7%**
YG01	35.7%	36.4%	37.1%	**35.6%**
YG05	33.7%	**33.6%**	33.7%	33.9%
average	32.9%	33.2%	33.8%	**32.4%**

なる分析周期 10msec，窓長 20msec 固定の場合と比較して，平均単語誤り率が 2.9%改善された．

表４．１０　発話速度補正音響モデル（multiple models）の各分析周期の単語誤り率

略称	分析周期			音響・言語ゆう度ベース選択手法
	8msec	9msec	10msec	
AS22	40.3%	41.3%	44.0%	**39.7%**
AS23	28.8%	27.8%	31.0%	**27.2%**
AS97	29.8%	28.2%	31.7%	**27.4%**
PS25	36.4%	31.9%	35.2%	**31.0%**
JL01	31.6%	27.6%	35.5%	**25.9%**
NL07	37.0%	33.7%	38.8%	**33.4%**
SG05	41.0%	38.6%	41.4%	**37.6%**
KK05	34.8%	30.6%	30.6%	**30.1%**
YG01	37.3%	36.5%	39.5%	**35.9%**
YG05	37.1%	33.1%	33.5%	**33.1%**
average	35.2%	32.6%	36.1%	**31.6%**

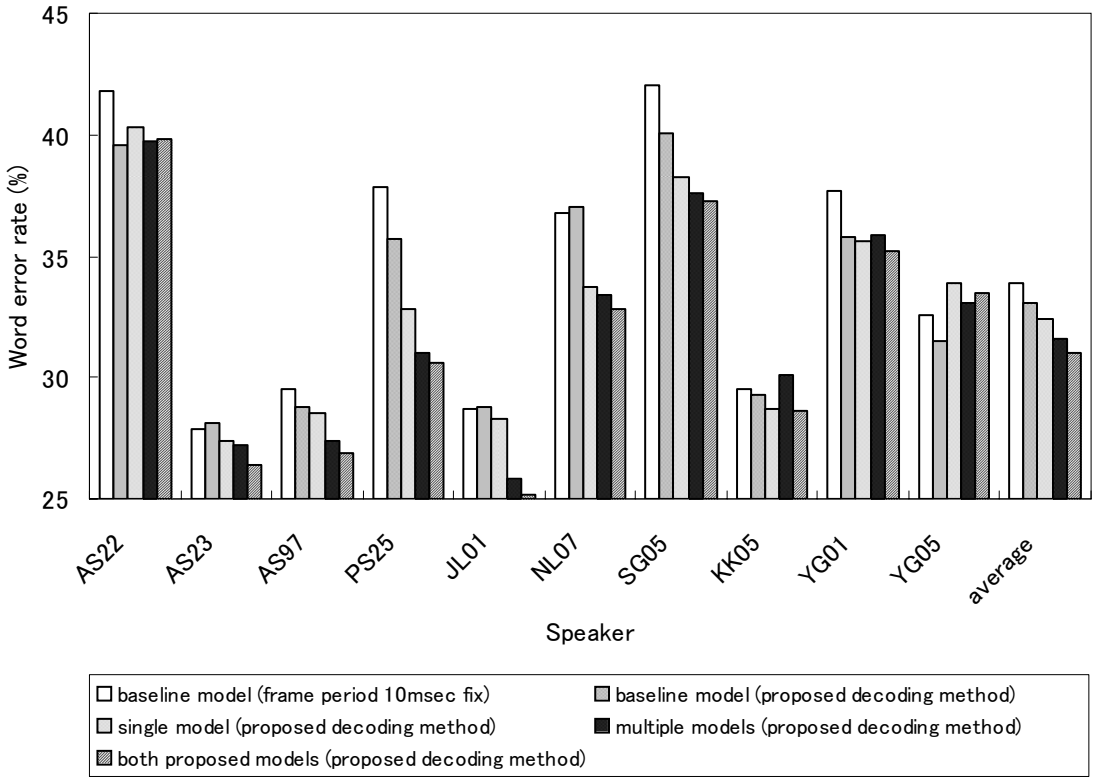


図４．１４　single model と multiple models を併用した場合の単語誤り率

表４．１１に，分析周期ごとに single model，multiple models が選択された発話に含まれる全音素の継続時間長の平均と標準偏差を示す．選択される分析周期が短くなるにつれて，音素継続時間長の平均値は小さくなっていることから，発話速度の速い発声に対して短い分析周期が選択されていることがわかる．各分析周期において，single model が選択された発話と multiple models が選択された発話を比べた場合, single model の方が標準偏差が大きくなる傾向にあることがわかる．これはつまり，single model が選択された発声においては，発話内での発話速度の変動が大きくなっていることを示している．音響・言語ゆう度ベース選択手法は，発話内での発話速度の変動に関しては基本的に考慮していない．このため，発話内で発話速度が変動した場合，分析周期ごとに構築した multiple models ではミスマッチが生じるが，各分析周期における音響的特徴を全て用いて構築した single model によりある程度対応が可能である．

表４．１１　各分析周期において single model，multiple models が選択された発話の音素継続時間長

分析周期	音響モデル	平均(msec)	標準偏差(msec)
10msec	single model	84.4	107.5
	multiple models	69.1	69.1
9msec	single model	69.9	67.3
	multiple models	69.3	60.1
8msec	single model	65.2	50.9
	multiple models	67.1	52.0

４．９　結言

自由発話のひとつである講演音声について，その音響的特徴を分析するとともに，発話速度の変動に頑健な音声認識手法を提案した．

音響的特徴の分析より，講演音声は対話音声や読み上げ音声と比較し，周波数的特徴が変化すると共に，音素継続時間長が短くなる傾向にあることを示した．また，読み上げ音声と比較した場合，教師あり話者適応を用いても単語正解精度は劣化する傾向にあり，音素認識においては継続時間長の短い音素の脱落誤りが増加する傾向にあることが明らかになった．このことから講演音声の発話スタイルには，モデルパラメータの適応だけでは吸収できない，認識性能劣化の要因が含まれていることを示した．

継続時間長の短い音素に対する認識性能の改善手法として，継続時間長の短い音素により構築した音響モデルをマルチパスの手法で統合した音響モデルを提案した．しかしながらこのモデルだけでは不充分であり，話者適応の利用に加え，分析周期・窓長の変更が有効であることが明らかになった．つまり，モデルパラメータの違いだけではなく，発話速度の変化に伴う音素継続時間長の変動による影響を考慮しなければならないということである．しかしながら，発話速度の変動具合は話者により異なるため，何らかの方法による発話速度の推定，もしくは分析周期・窓長の決定が必要となる．

そこでゆう度基準により，発話ごとに分析周期・窓長を選択する手法を提案し，発話速度が未知の講演音声に対して有効であることを示した．これは，複数の分析周期・窓長でいったん認識を行い，音響ゆう度の最も高い分析周期・窓長の認識結果を選択するものである．分析周期・窓長を変更した場合，分析周期の短い認識結果の音響ゆう度が大きくなる傾向にあるため，音響ゆう度は分析フレーム数で正規化を行うこととした．更に，言語ゆう度を加えたものを比較することで，認識性能が改善することを確認している．

デコード時に分析周期・窓長を変更することの有効性は明らかになったが，音響モデルの学習時とは条件が異なるため，音響モデルとのミスマッチは解消されていない．そこで，音響モデル学習時の分析周期・窓長の選択に，本提案手法を利用することで更なる認識性能の改善が可能であることを示した．以上のことから，発話速度が変動する講演音声認識においては，分析周期・窓長の変更も含んだ認識手法の検討が有効であると言える．また，同様の手法を音響モデル構築時に用いることで，さらなる認識性能の改善を得た．これらにより，発話速度の補正効果を示した．

本提案手法は，発話ごとに分析周期・窓長を選択するため，single model と multiple models を併用した場合でも，発話内での発話速度の変動に対して十分な性能が得られているわけではないが，発話内での動的な分析周期・窓長の選択手法の検討は可能である．ま

た，発声のなまけなどによる音響的特徴の変形に対しては，なまけを吸収するために提案されている音響モデル構築手法と併用することなどで，認識性能の改善が期待できる．

参考文献

[1] 奥田浩三，川原達也，中村 哲，“ゆう度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデル構築,”電子情報通信学会論文誌， vol.J86-D-II， no.2, pp.204-211，2003.

[2] 山本一公，岩井直美，中川聖一，“発話スタイルの違いが音声認識に及ぼす影響についての検討,”電子情報通信学会技術研究報告，SP99-31，1999.

[3] 古井貞熙，前川喜久雄，井佐原 均，“科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく『話し言葉工学』の構築,”日本音響学会誌， vol.56， no.11，pp-752-755，2000.

[4] 龍宮隆之，菊地英明，小磯花絵，前川喜久雄，“大規模話し言葉コーパスにおける発話スタイルの諸相 --書き起こしテキストの分析から--,”日本音響学会 2000 年秋季研究発表会，2-Q-9，2000.

[5] 前川喜久雄，“言語研究における自発音声,”日本音響学会 2001 年春季研究発表会，1-3-10，2001.

[6] 小磯花絵，前川喜久雄，“『日本語話し言葉コーパス』の概要と書き起こし基準について,”情報処理学会研究報告，SLP36-1，2001.

[7] 篠崎隆宏，古井貞熙，“話し言葉認識における決定木を用いた誤り要因の分析,”日本音響学会 2001 年秋季研究発表会，1-1-9，2001.

[8] 南條浩輝，河原達也,“発話速度に依存したデコーディングの検討,”日本音響学会 2001 年秋季研究発表会，1-1-6，2001.

[9] 河原達也，加藤一臣，南條浩輝，李晃伸，“話し言葉音声認識のための言語モデルとデコーダの改善,”情報処理学会研究報告，SLP36-3,2001.

[10] 南條浩輝，加藤一臣，李晃伸，河原達也，“大規模な日本語話し言葉データベースを用いた講演音声認識,”電子情報通信学会論文誌，vol.J86-D-II，no.4，pp.450-459，2003.

[11] 前川 喜久雄，“「日本語話し言葉コーパス」の構築,”話し言葉の科学と工学ワークショップ講演予稿集，p.7-12，2001.

[12] 李 晃伸，河原達也，堂下修司，“単語 N-gram と段階的探索に基づく大語彙連続音声認識エンジン JULIUS,”日本音響学会 1998 年春季研究発表会，1-6-24，1998.

[13] 山本博史，シンガー ハラルド，リーブス ベン，匂坂芳典，“日英音声翻訳システム「ＡＴＲ－ＭＡＴＲＩＸ」における音声認識部分の構造と制御方法,”日本音響学会 1998 年春季研究発表会，2-Q-21，1998.

[14] 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,”電子情

報通信学会論文誌, vol.J79-D-II, no.10, pp.2155-2164, 1993.
[15] 松井知子, 内藤正樹, シンガー　ハラルド, 匂坂芳典,“大規模な日本語音声データベースによる音響モデルの分析,”日本音響学会 2000 年春季研究発表会, 1-Q-28, 2000.
[16] 奥田浩三, 松井知子, 内藤正樹, 匂坂芳典, 中村　哲,“大規模日本語音声データベースの構築と評価,”日本音響学会誌, vol.58, no.9, pp.569-578, 2002.
[17] 河原達也, 李　晃伸, 小林哲則, 武田一哉, 峰松信明, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田　篤, 宇津呂武仁, 鹿野清宏,“日本語ディクテーション基本ソフトウェア（97 年度版）,”日本音響学会誌, vol.55, no.3, pp.175-180, 1999.
[18] A.Nakamura, S.Matsunaga, T.Shimizu, M.Tonomura and Y.Sagisaka,“Japanese speech database for robust speech recognition,”Proc. of ICSLP'96, pp.2199-2202, 1996.
[19] M.Tonomura, T.Kosaka and S.Matsunaga,“Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation,”Computer Speech and Language, vol.10, pp.117-132, 1996.
[20] J.Takahashi and S.Sagayama,“Vector-field-smoothing Bayesian learning for fast and incremental speaker/telephone-channel adaptation,”Computer Speech and Language, vol.11, pp.127-146, 1997.
[21]大倉計美, 杉山雅英, 嵯峨山茂樹,“混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式,”電子情報通信学会技術研究報告, SP92-16, 1992.
[22] Lawrence Rabiner, Biing-Hwang Juang　共著, 古井貞熙　監訳,“音声認識の基礎（上）（下),”NTT アドバンステクノロジ株式会社, 1995.
[23] 奥田浩三, 松井知子, 中村　哲,“音節強調発声に頑健な自然発話音声の認識法,”電子情報通信学会技術研究報告, SP2000-98, vol.100, no.523, pp.19-24, 2000.
[24] S.Tsuge, T.Fukada and K.Kita,“Frame-period adaptation for speaking rate robust speech recognition,”Proc. of ICSLP2000, vol.3, pp718-721, 2000.
[25] Jon P.Nedel and Richard M.Stern,“Duration normalization for improved recognition of spontaneous and read speech via missing feature methods,”Proc. of ICASSP2001, vol.1, pp.313-316, 2001.
[26] Martin Cooke, Andrew Morris and Phil Green,“Missing data techniques for robust speech recognition,”Proc, of ICASSP'97, vol.2, pp863-866, 1997
[27] Q.Zhu and A.Alwan,“On the use of variable frame rate analysis in speech recognition,”Proc.ICASSP, vol.3, pp1783-1786, 2000.
[28] 小早川健, 世木寛之, 松井淳, 尾上和穂, 佐藤庄衛, 今井亨, 安藤章男,“音声認識

における高精度な動的特徴量計算方法の提案,”電子情報通信学会技術研究報告,
SP2001-84, 2001.

第 5 章

誤認識時の言い直し発話に頑健な音響モデルの構築

５．１　緒言

本章は，論文“誤認識時の言い直し発話における発話スタイルの変動に頑健な音響モデル構築法”[1]に関するものである．

前章までにおいて，学習データ量と認識性能の関係や，学習データと認識対象のタスクの違いによる認識性能の関係を示した．また，読み上げ音声，対話音声や講演音声などの，発話スタイルの違いによる音響的特徴の違いと，その違いが認識性能に与える違いについて，分析を行った．本章では，発話スタイルの一つである誤認識時の言い直し発話が認識性能に与える影響と，その改善方法について述べる[2]．

近年，音声認識技術の認識性能は飛躍的に向上したとはいえ，依然として誤認識の発生は避けられず，システム利用者は正しく認識させるために再発話を行う必要が生じる．ところが，誤認識した発声内容を再発話しても，全く認識できないという様な現象は，一向に解決されていない．再発話の際，システム利用者はシステムが認識しやすいよう発話スタイルを変える場合が多く，音響的特徴が変化することによって，かえって認識性能が劣化するという現象が生じる[3][4]ためである．誤認識時の言い直し発話に対する頑健性が小さい場合，音声対話システムや音声翻訳システムにおいては，タスクそのものが達成できなくなる可能性があり，またシステム利用者の負担も大きくなる．このため，誤認識時の言い直し発話に対する頑健性が求められている．

話者の違いによる音響的特徴の違いを吸収する手法として，MLLR（Maximum Likelihood Linear Regression）[5]や最大事後確率推定法（Maximum a Posteriori Estimation ; MAP）[6]に代表される話者適応の手法が提案されている．これらの手法は，既に統計的手法によりモデル化した複数の学習話者の音響的特徴を，適応対象話者の音響的特徴に近づくよう，モデルパラメータの値を適応するものである．話者適応の導入により，認識対象話者の音響的特徴が学習データの音響的特徴からずれたことによる認識性能の劣化を改善することに成功している．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性能を改善できることが報告されている[7]．しかしながら，発話スタイルの違いによる音響的特徴の違いは，モデルパラメータの値に加えて，モデル構造そのものにも大きく影響するため，モデル構造の違いを考慮しない MLLR や MAP では，認識性能の改善にも限界があると考えられる．また，多くの言い直し発話を収集することは一般に難しいため，言い直し発話の音響的特徴を調べて対処することが望まれる．

以上の背景のもと，本章では言い直し発話における発話変形に着目し，話者適応では対応が難しい発話変形に対して認識性能を改善する音声認識手法について提案する．具体的

には，言い直し発話の音響的特徴を分析することで，あらかじめその特徴を表現できるモデル構造を用意するものである．これにより，モデル構造のミスマッチによる認識性能の劣化を抑えるとともに，話者適応が効果的に働くようにする．まず５．２節において，言い直し発話を収録，分析することで，言い直し発話の音響的特徴とその出現頻度について論じる．５．３節では，収録した言い直し発話を用いた認識実験を通して，発話変形が音声認識システムに与える影響を明らかにする．以上の結果を踏まえ，５．４節では言い直し発話における発話変形に頑健な音響モデル構築手法を提案し，５．５節にて本手法の有効性を明らかにする．

５．２　言い直し発話における発話変形の特徴

本節では，言い直し発話における発話変形の音響的特徴とその出現頻度についてまとめる．

５．２．１　言い直し発話音声の収録方法

言い直し発話の音響的特徴を調査するにあたり，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，収録データは単語発声とした．より自然な言い直し発話を収録するため，データの収録には誤認識をシミュレートする収録装置を用いた．本収録装置は，入力すべき単語を画面に表示し，被験者から入力された音声に対する認識結果を表示するものである．誤認識が発生した場合，つまり画面に表示した単語と認識結果が異なる場合は，正しく認識されるまで同一単語の音声入力を促すようになっている．認識結果については，正しく認識された場合は認識結果の単語を表示する．誤認識の場合は，その認識結果によって言い直し方が影響されないよう，誤認識が発生したことのみを画面に表示し，どのように誤認識したかなどの情報は，被験者に一切与えないものとした．

本章では国際電気通信基礎技術研究所（ATR）で開発された連続音声認識用のデコーダ，ATRSPREC[8]を使用した．ATRSPREC の平均単語正解精度は，ATR で収録した自然発話音声・言語データベース[9]においては 80%を越えるが，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ている．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，４回目の言い直しで正しく認識するように設定した．全ての被験者に対して全く同じ誤認識を発生するようにしている．本システムを用い，５名の被験者から 210 単語，言い直し発話を含めると，のべ 378 単語の音声データを収録した．

５．２．２　発話変形の音響的特徴

言い直し発話においては，再発話によって認識結果が正しくなるよう，より明瞭に発声する傾向にあり，この場合，音素継続時間長が変化するとともに，特に日本語では，各音節を強調した発声（本稿では，この発声を音節強調発声と呼ぶ）の出現頻度が増加するという傾向にある．図５．１に，１回目の発話と，音節強調発声と判断される同一話者の言い直し発話のスペクトログラムを示す．それぞれ，／jizai／（自在）と発声しているものである．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共に，音響的特徴が大きく変化している

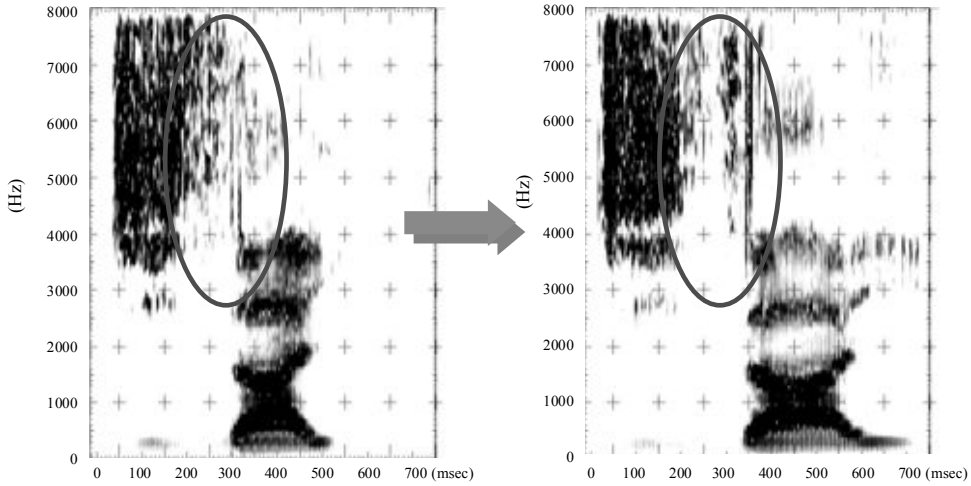


図５．１　／jizai／と発声した場合の，通常の発声と音節強調発声のスペクトログラム

ことがわかる．特に音節強調発声においては，意図的に挿入することが難しい 50msec 程度の無音区間が存在していることがわかる．スペクトログラムの比較より，音節強調発声は次の特徴を有することがわかる．

① 音節間の音響的特徴，連続性が変形する
② 音節間に無音が存在する孤立音節発声に向けて，発話スタイルが変形する

音節強調発声データを収集し，音響モデルを構築することにより，認識性能を改善することも考えられるが，音節強調発声は通常の発声とその音響的特徴が大きく異なるため，通常の発声に対して認識性能が劣化する可能性がある．通常の発声に対する認識性能を劣化させずに，音節強調発声に対して認識性能を向上する手法を検討する必要がある．

５．２．３　言い直し発話における音節強調発声の出現頻度

収録した音声データ内の言い直し発話（各話者 168 単語）において，音節強調発声の出現頻度を調査した．

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，現時点ではその強さを評価する指標が存在しない．そこで本章ではまず評価者１名ではあるが，聴感的な主観評

価を行った．評価は，１回目の発話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段階で行った．本評価結果を図５．２に示す．

本結果を踏まえ，次に音響ゆう度を基準とした客観評価を行った．評価方法は，１回目の発話と言い直し発話それぞれの音響ゆう度を算出し，１回目の発話の音響ゆう度からの劣化の割合を元に行った．音響ゆう度の算出に用いた音響モデルは性別依存モデルであり，５混合ガウス分布，1,400 状態の状態共有化 HMM（hidden Markov networks；HMnet[10]）で表現されている．学習データには ATR で収録された自然発話音声・言語データベースより，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．

１回目と言い直し発話の，音響ゆう度の変化の割合は，

$$R = \frac{L_r - L_1}{L_1} \times 100 \qquad (5.1)$$

の式により算出したRを用いている．ここでL_1は１回目の発話の音響ゆう度，L_rは言い直し発話の音響ゆう度である．このようにして算出したRを元に，音節強調発声の出現頻度を決

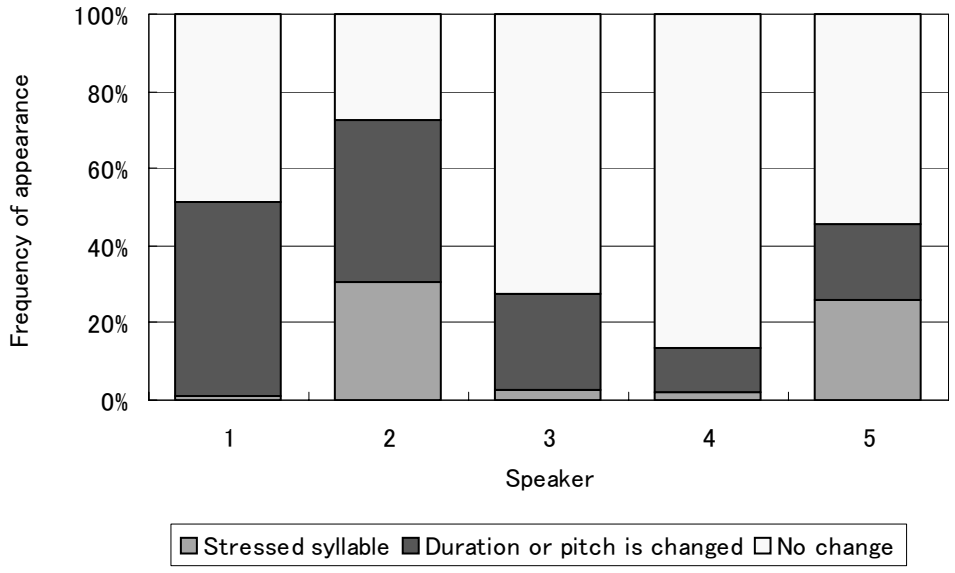


図５．２　言い直し発話における音節強調発声の出現頻度（主観評価）

定するにあたり，前述の主観評価結果による出現頻度と自乗誤差が最小になるよう閾値を選択した．図５．３に，音響ゆう度の変化率Rと自乗誤差の関係を示す．この図より，R＝－20.0 つまり１回目の発話と比較し音響ゆう度が 20%以上劣化する発声の出現頻度が，主観評価結果における音節強調発声の出現頻度と自乗誤差が最も小さくなり，音節強調発声の可能性が高いと判断できる．図５．４に言い直し発話における $R \geqq 0$ ， $0 > R \geqq -20$， $-20 > R$ の発声の出現頻度を示す．この結果より，多い話者で 25%程度の割合で，音響ゆう度が大きく劣化し音節強調発声の可能性がある言い直し発話が出現していることがわかる．

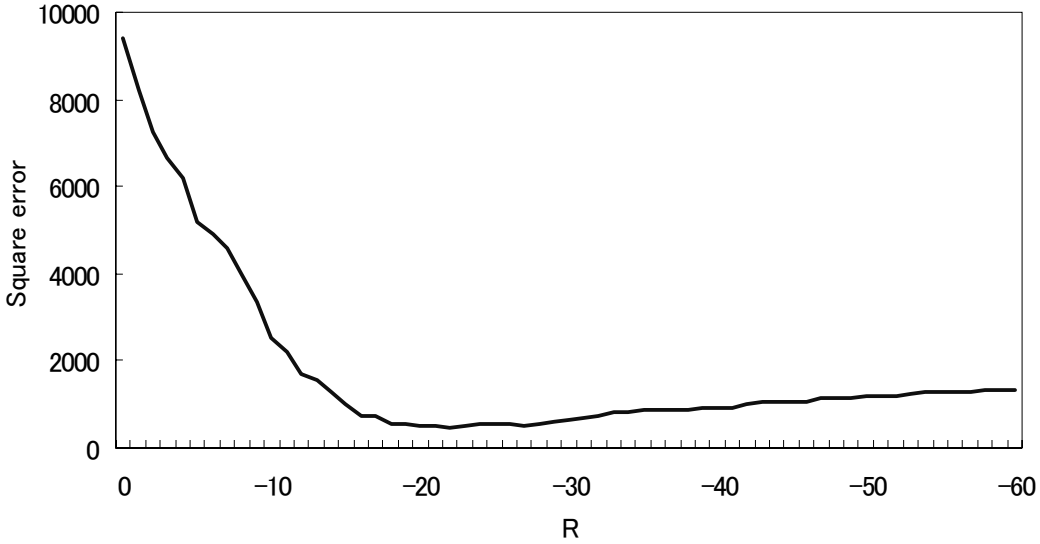


図５．３　音響ゆう度の変化率 R と２乗誤差との関係

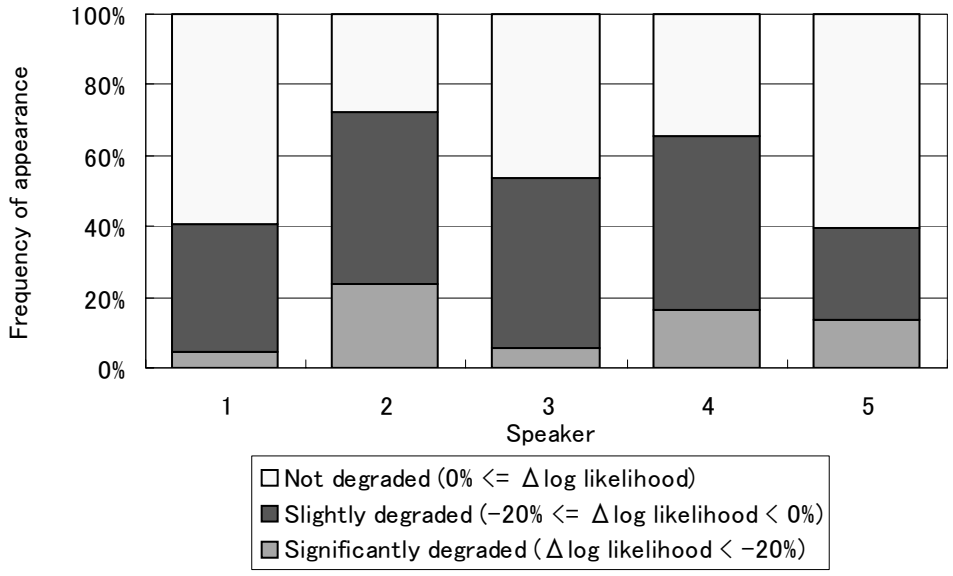


図５．４　言い直し発話における音声強調発声の出現頻度（音響ゆう度差）

５．３　音節強調発声が認識システムに与える影響

本節では，音節強調発声が認識性能にどの程度の影響を与えるのかについて調査した．

５．３．１　実験条件

本節における認識実験では，デコーダに ATRSPREC を使用した．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次の通りである．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，分析窓長 20msec，分析周期 10msec で抽出した，25 次元の特徴ベクトル（12 次元 MFCC，12 次元ΔMFCC とΔlog power）を用いている．ベースラインとなる音響モデルには，５．２．３で音響ゆう度算出に用いた音響モデルと同一のものを使用した．

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[11]を用いた．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．認識の際には HTK[12]同様，スキップ可能な無音モデルを単語間の接続部分で用いている．

５．３．２　認識実験結果

ベースラインシステムを用いた評価実験を行った．ベースラインシステムは，連続音声認識用に開発されたものであり，また孤立単語認識では音響的な特徴の差による認識性能の差が判断しにくいため，ここでは連続発声音声を収録して認識実験を行っている．評価データは，５．２．１とは異なる男性話者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，ATR で用いられている旅行対話タスクの中から選択した 20 文章である．データの収録は，収録装置の画面に表示される漢字かなまじりの文章を読み上げる形で行った．被験者に対しては，「表示される漢字かなまじりの文章を，各音節を強調して読み上げてください」という指示のみを与えて収録している．このため話者によっては，音節を完全に区切って読み上げる孤立音節発声や，各単語の１音節目のみを強調した発声などが含まれている．

収録した評価データに対するベースライン音響モデルでの認識結果を表４．１に示す．この表から，連続音声認識用に学習した音響モデルを用いた場合，通常の発声に対しては

80%程度の認識率が得られているのに対し，音節強調発声に対しては認識性能が大きく劣化することがわかる．単語正解精度に負の値が出現しているが，これは挿入誤りが多く発生したことにより，単語正解精度の計算式（2.17）の結果が負の値になったためである．認識性能劣化の主な要因としては，音節強調発声と音響モデルとの音響的な違いが挙げられる．

そこで，話者適応を用いた場合の認識実験を行った．話者ごとに，MAP-VFS[13]～[15]を用いて平均値を，Baum-Welch 法[16]を用いて状態遷移確率を適応したモデルを用いた．適応データには評価データである音節強調発声 20 文章全ての音声データを用い，本評価実験はクローズドの評価となっている．

表５．１に話者適応を用いた場合の音節強調発声に対する認識率を示す．話者適応を行うことにより，認識性能は改善されるが，通常発声に対する認識性能と比較すると，大きく劣化している．

以上のことより，音節強調発声においては音響的特徴の変化が大きいため，話者適応などでは十分な認識性能の改善が得られないことがわかる．

表５．１　ベースラインシステムを用いた場合の単語正解精度

話者	通常発声	音節強調発声	
	ベースライン音響モデル	ベースライン音響モデル	ベースライン音響モデル＋話者適応
1	75.8%	18.6%	63.1%
2	78.6%	-28.9%	40.4%
3	87.3%	-44.0%	-2.1%
4	78.6%	-86.9%	-65.5%
5	80.6%	-51.9%	-8.7%
6	81.4%	-27.1%	5.7%
7	81.3%	-12.0%	44.4%
8	73.8%	6.2%	38.6%
9	74.4%	-2.5%	22.5%
10	75.0%	32.1%	62.1%
平均	78.4%	-15.9%	28.0%

５．４　音節強調発声に頑健な音声認識手法

ここでは，音節強調発声に頑健な音声認識手法について検討する．本手法では，５．２．２で述べた音節強調発声の音響的特徴をモデル化することで，認識性能を改善する．音節強調発声の音響的特徴は次の通りである．

① 各音節間の音響的特徴，連続性の変形：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特徴になる．このため，各音節において後続音素環境との結びつきが弱くなり，各母音において後続音素環境に依存していた音響的特徴が変形する．

② 孤立音節発声に近い音節の出現：各音節間に無音が存在する，孤立音節発声に近い音響的特徴になる．

通常の triphone 音響モデルではこれらの音響的特徴を十分に表現することができず，認識性能が劣化したと考えられる．

図５．５にこの様子の例を示す．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

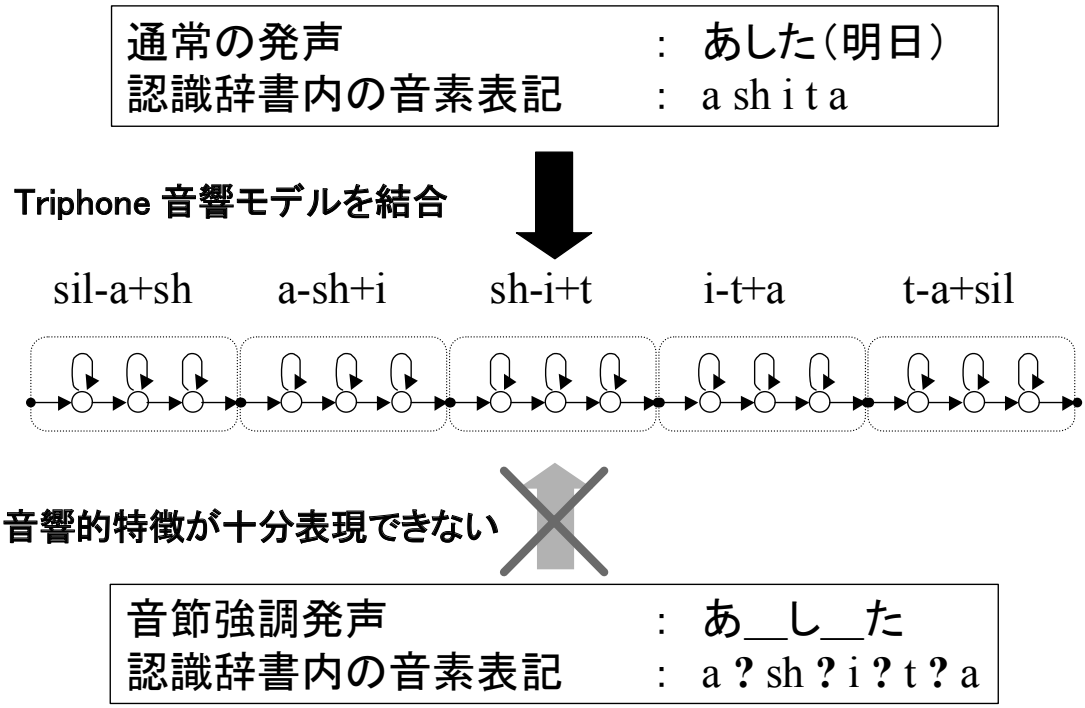


図５．５　音節強調発声における音素環境の変化

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モデル化することで，連続発声音声に対する認識性能を向上している．しかしながら，音節強調発声の場合，認識辞書に記述された音素表記に従い音素モデルを結合すると，音節間の音響的特徴の変形を表現することができず，認識性能が劣化する．完全な孤立音節発声に対しては，認識辞書内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，ある程度対応可能であるが，孤立音節発声の出現位置は事前にわからないため，全ての組み合わせに対して音素表記を併記しなければならなくなる．また，連続発声と孤立音節発声の中間に位置する音響的特徴を有する音節強調発声に対しては，効果が期待できない．

そこで本章では，一つの音素環境に対して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 既存の triphone 音響モデル

② 各音素間の音響的特徴，連続性の変形に対応するモデルとして，先行音素環境依存 biphone 母音モデル

③ 孤立音節発声に近い発声に対応するモデルとして，後続音素環境が無音の triphone 母音モデル

を用いる．また音響的特徴の変形だけではなく，完全な孤立音節発声にも対応できるよう，それぞれのモデルにはスキップ可能な１状態の無音モデルを追加する．また，追加した母

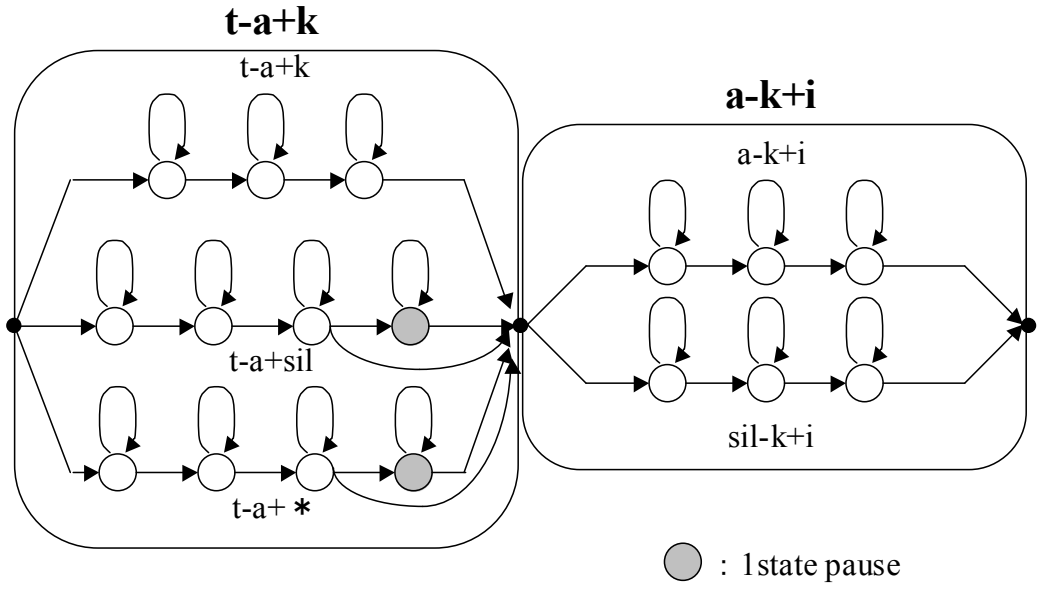


図５．６　提案手法における音響モデルの例

音モデル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追加する．音節強調発声の出現位置は事前にわからないため，これらのモデルと通常の triphone 音響モデルを図５．６に示すように, 各音素環境に対して定義する．この図の例では，先行音素が t，後続音素が k の中心音素 a のモデルと，先行音素が a，後続音素が i の中心音素 k のモデルの例である．

音節強調発声の音響的特徴を，既存の音響モデルを利用してモデル化することが可能となるため，学習データを追加する必要がなくなる．また全体としては状態数を増やすこととなり，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共に，デコードの際，ゆう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スタイルごとの音響モデルの切り替えも不要となる．

５．５　評価実験

本節では評価実験を通して，提案手法の有効性を確認する．

５．５．１　実験条件

評価実験で実際に用いた音響モデルは次のようにして構築した．本手法では基本的にベースライン音響モデルを使用した．提案手法で用いる後続音素環境が無音の母音モデルは，ベースライン音響モデルに含まれる該当母音モデルをそのまま利用した．また，先行音素環境依存 biphone 母音モデルについては，ベースライン音響モデルの学習に使用した学習データを用いて構築した，状態数 1,400，ガウス混合分布数５の性別依存，先行音素環境依存 biphone 音響モデルを新たに構築し，その中の該当母音モデルを使用した．

評価データには，５．５．２から５．５．４の評価実験に関しては，５．３．２で収録した通常発声データと意図的に発声した音節強調発声データを使用している．５．５．５では，誤認識をシミュレートする収録装置により新たに収録した，誤認識時の言い直し発話を使用している．

５．５．２　認識実験結果

本提案手法を用いた場合の，音節強調発声に対する認識結果を図５．７に示す．この結果より，話者適応を行ったベースライン音響モデルと比較し，話者適応を行わない提案手法の方が良好な結果が得られていることがわかる．しかしながら，提案手法を用いた場合でも，話者 5 のように単語正解精度が 10%未満の話者が含まれている．この原因を分析するため，提案手法で用いたベースライン音響モデルによる音節強調発声の音響ゆう度を調査した．調査は，ベースライン音響モデルを用いて音素アライメントをとり，話者ごとに音響ゆう度を算出し，そのゆう度を比較することで行った．音素アライメントによる音響ゆう度と，提案手法による認識率の関係を図５．８に示す．この図より，音響ゆう度の比較的大きい，すなわちベースライン音響モデルとの音響的な特徴空間が比較的近い話者ほど，提案手法による認識性能が良いことがわかる．音響ゆう度が小さい話者５は，提案手法で用いた音響モデルとのスペクトル的特徴のずれが大きすぎるため，提案手法が効果的に働かなかったと考えられる．話者 1 や話者 10 は音響モデルとのスペクトル的特徴のずれが小さく，提案手法が効率よく働いていることから，これらの話者はスペクトル的特徴の変化は少ないものの，ベースライン音響モデルの有するモデル構造では十分に表現することのできなかった発話スタイルになっていると判断できる．

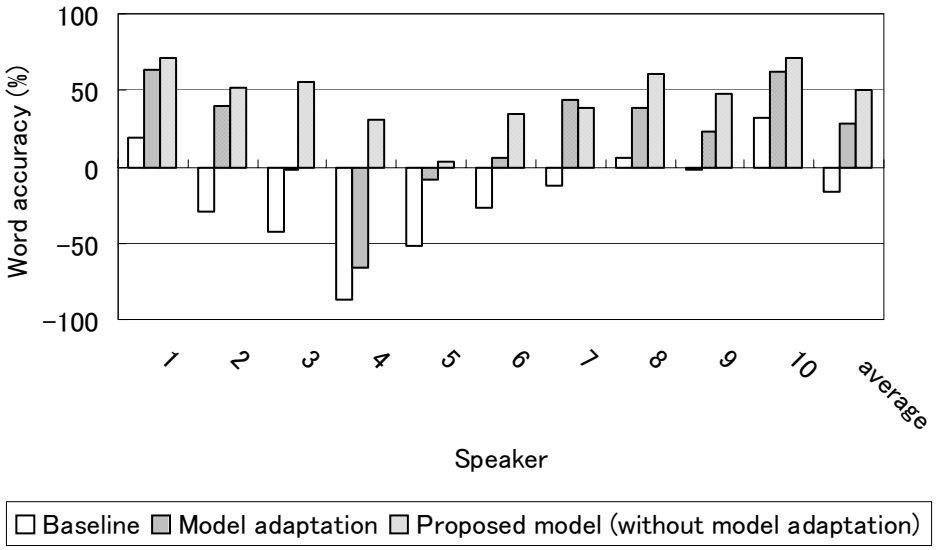


図５．７　提案手法を用いた場合の音節強調発声の単語正解精度

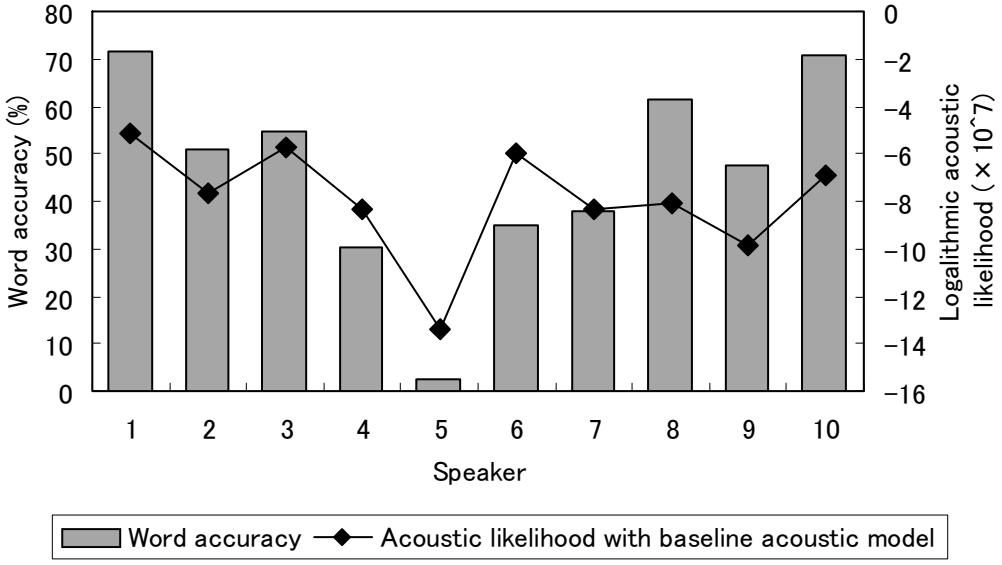


図５．８　ベースライン音響モデルを用いた場合の音響ゆう度と，提案手法による単語正解精度

５．５．３　話者適応との併用による提案手法の認識性能

各話者と提案手法で用いるベースライン音響モデルの間の音響的特徴空間を近づけることで，提案手法がより効果的に働くことが期待できる．そこで，話者適応を行った音響モデルを用いて提案手法を行った場合の評価実験を行った．ベースライン音響モデルとして

５．３．２で用いた話者適応モデルを使用した．また，後続音素環境が無音の母音モデルには，同じ話者適応モデル内の該当母音モデルをそのまま使用した．先行音素環境依存 biphone 母音モデルは，５．５．１にて構築した biphone 音響モデルを５．３．２の話者適応モデルと同一の条件で適応したモデルを使用した．

図５．９に話者適応と併用した提案手法の認識結果を示す．話者適応を併用することにより提案手法が効果的に働き，認識性能の改善の小さかった話者５に対しても認識性能を大きく改善することができた．また話者１や話者 10 に関しては，単語正解精度が 80%まで改善した．

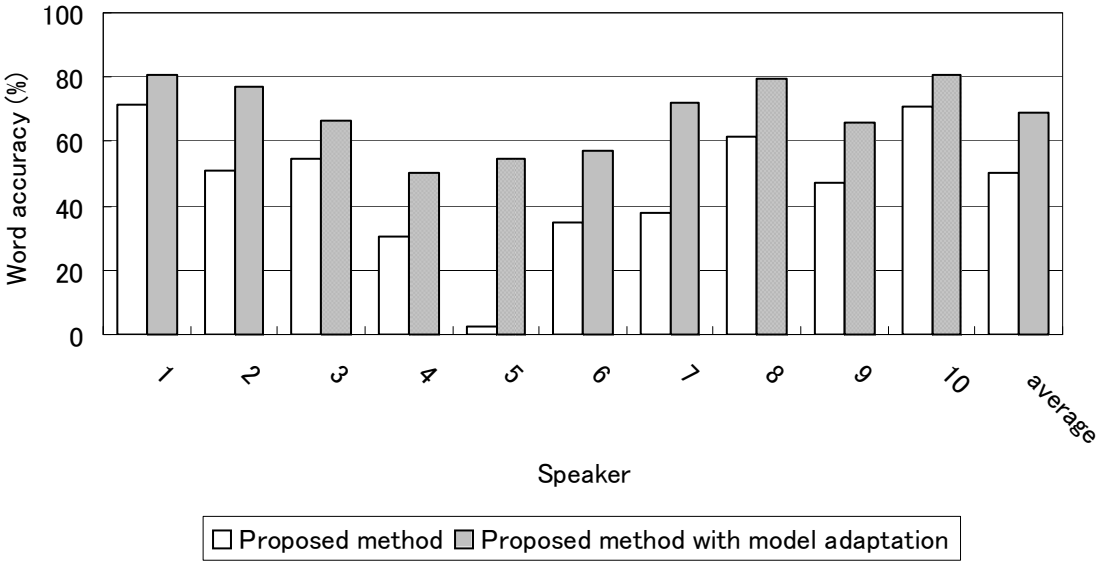


図５．９　話者適応を併用した提案手法を用いた場合の音節強調発声の単語正解精度

５．５．４　通常発声に対する効果

提案手法を用いることで，音節強調発声に対する認識性能の改善が確認できたが，通常発声に対して逆に認識性能が劣化する可能性がある．本節では，通常発声に対する提案手法の効果を調査した．音響モデルは，５．５．１で用いた，話者適応を行っていないモデルを用いた．表５．２に話者ごとのベースライン音響モデルと提案手法を用いた場合の単語正解精度を示す．この表より，提案手法はベースライン音響モデルと比較し，平均単語正解精度で２%改善している．表５．３に，後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルのみをマルチパス化した場合の通常発声と音節強調発声の単

語正解精度を示す．これらの結果より，提案手法において追加した後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルは，ベースライン音響モデルの効果を妨げていないということがわかる．また通常発声においても後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルの方が，マッチングが良い区間が存在するため，全体として若干ではあるが認識性能が改善している．

表５．２　提案手法を用いた場合の通常発声の単語正解精度（%）

話者	ベースライン音響モデル	提案手法
1	75.8%	**80.8%**
2	78.6%	**79.3%**
3	87.3%	**88.0%**
4	78.6%	**84.8%**
5	**80.6%**	80.0%
6	**81.4%**	80.0%
7	81.3%	**85.2%**
8	73.8%	**75.9%**
9	**74.4%**	**74.4%**
10	75.0%	**77.9%**
平均	78.4%	**80.4%**

表５．３　モデルごとの単語正解精度

	ベースライン音響モデル	後続音素環境が無音の母音モデル	先行音素環境依存biphone母音モデル
通常発声	78.4%	79.4%	79.4%
音節強調発声	-15.9%	41.9%	29.0%

５．５．５　言い直し発話に対する効果

前節までの結果より，意図的に発声した音節強調発声に対して提案手法が有効であることを示した．ここでは実際の使用場面における，より現実的な言い直し発話を用いて評価を行う．５．２．１にて用いた収録装置を用いて，５．２．１と同じ５名の被験者より文章の言い直し発話を収録した．収録方法も５．２．１と同様で，単語ではなく連続音声認識でデータを収録している．入力すべき文章を画面に表示し，被験者から入力された音声に対する認識結果を表示する．文章全体が正しく認識されるまで，同一文章の音声入力を

促すようになっている．認識結果については，正しく認識された場合は，その認識結果の文章を表示する．誤認識の場合は，誤認識が発生したことのみを画面に表示し，どのように誤認識したかなどの情報は一切与えないものとした．

また，認識誤りについても全体の 40%の文章に対して発生するようにしている．誤認識する文章の 50%が１回目，25%が２回目，12.5%が３回目，４回目の言い直しで正しく認識するようにしている．このシステムを用い，５．２．３でデータを収録した５名の被験者から 50 文章，言い直しを含めると 88 文章の音声データを収録した．

図５．１０に，収録したデータにおける言い直し発話に対する単語正解精度を示す．この図より話者適応を用いない場合，提案手法はベースラインと比較して平均で 4.0%，話者によっては単語正解精度が 8.3%改善しており，実際の言い直し文章発話に対しても，提案手法が有効であることがわかる．話者１においては話者適応を行うよりも，話者適応を併用しない提案手法の方が高い認識率を示しており，話者適応だけでは十分な認識性能が得られない発話スタイルであることがわかる．話者適応を併用した場合は，話者適応のみの場合と比較し，若干認識性能が低下する話者もいるが，提案手法が効果的に働いていることがわかる．特に話者適応だけでは認識性能の改善が小さい話者において，提案手法との併用による効果が大きく現れていることがわかる．

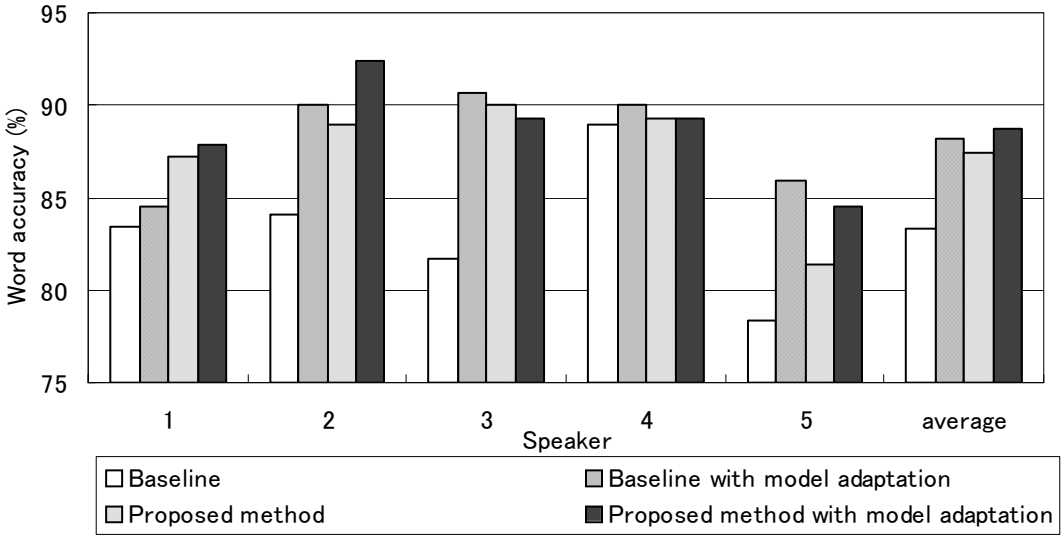


図５．１０　提案手法を用いた言い直し発話に対する認識性能

５．６　考察

本章では，意図的に発声した音節強調発声のデータを用い，提案手法における後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデル，１状態無音モデルの効果について考察する．

５．６．１　提案手法における各モデルの効果

ここではマルチパスモデルとして追加した，後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルの効果について調査した．調査の方法としては，音節強調発声をデコードする際の，それぞれのモデルの選択比率と単語正解精度との関係を調べることで行った．この際，後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルそれぞれの影響を区別するため，どちらか一方のモデルのみをマルチパス化した音響モデルを用いて行った．

表５．３の結果から，音節強調発声に対しては，先行音素環境依存 biphone 母音モデルより後続音素環境が無音の母音モデルの方が効果があることがわかる．

図５．１１に，話者ごとに後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルのみをマルチパス化したモデルを用いた場合の単語正解精度を示す．この結果は，表５．３の音節強調発声に対する認識性能を，話者ごとに示したものである．図５．１１より，話者４や話者５に関しては，先行音素環境依存 biphone 母音モデルの効果は少なく，後続音素環境が無音の母音モデルが効果的に働いていることがわかる．このことより，話者４や話者５は，音節間に無音区間が多く含まれるような発話スタイルになっていることがわかる．これに対して他の話者は，後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルそれぞれ効果的に働いており，音節間の連続性が変形していることがわかる．

図５．１２に後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルの選択比率と，認識改善率を示す．認識改善率は，各母音モデルをマルチパス化した場合の単語正解精度と，ベースライン音響モデルによる単語正解精度の差で示している．図中における各点は話者を表したものである．この図より，後続音素環境が無音の母音モデルに関しては，その選択比率が高くなるほど，認識改善率が大きくなっており，音節強調発声において音節間に無音区間が発生する発話スタイルが認識システムの認識性能に大きな影響を与えていると考えられる．

先行音素環境依存 biphone 母音モデルに関しては，各話者同程度の選択比率となっており，その改善率にも大きな差はなかった．

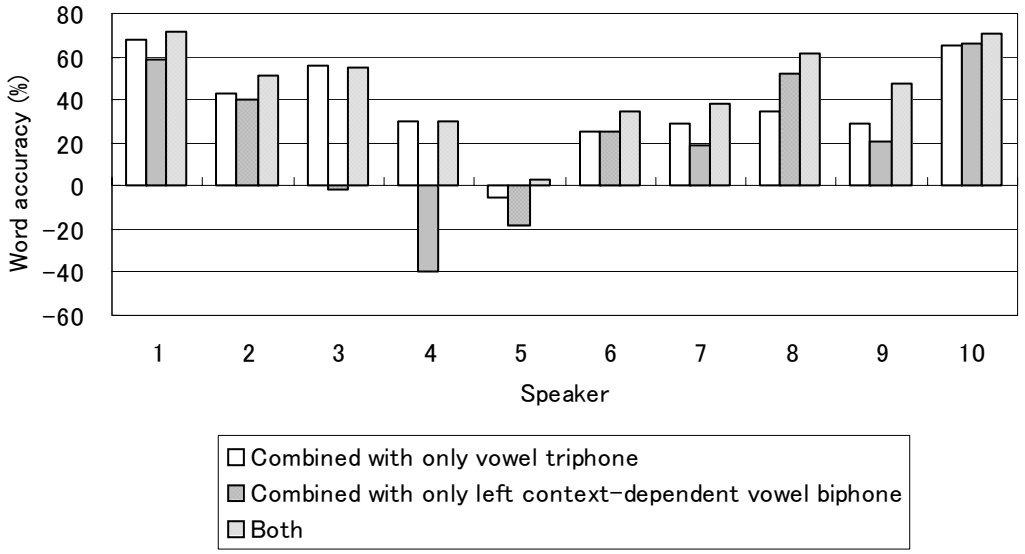


図５．１１　音節強調発声における各母音モデルの効果

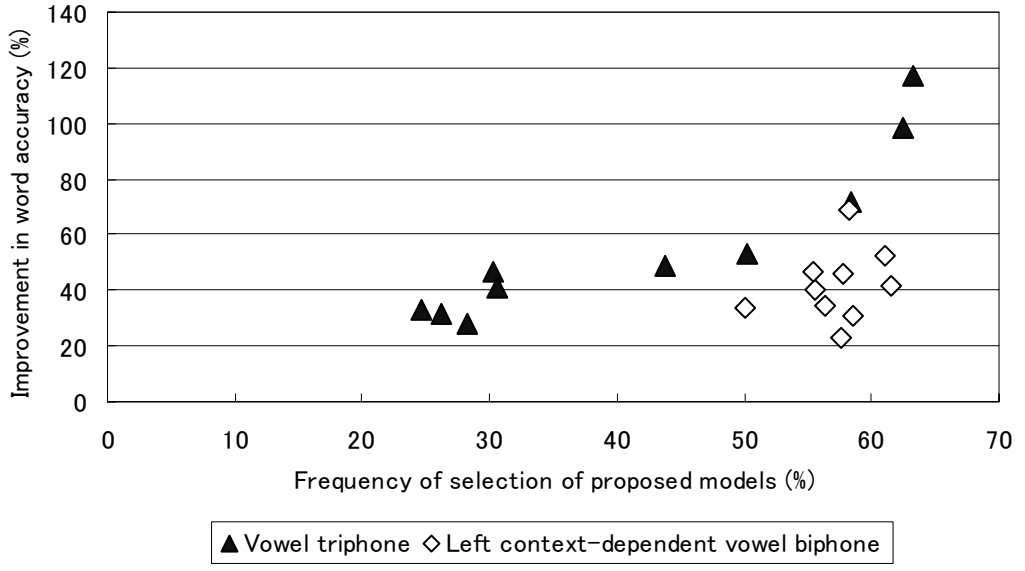


図５．１２　モデル選択比率と認識改善率

５．６．２　１状態無音モデルの効果

HTK[12]などにおいても，スキップ可能な１状態の無音モデルを用いることができる．しかしながら HTK では単語間の接続部分に限定されているため，提案手法における１状態

の無音モデルとは効果が異なる．ここでは，１状態の無音モデルの効果を調査する．

図５．１３に，後続音素環境が無音の母音モデル，先行音素環境依存 biphone 母音モデルの両方に，音節間の１状態の無音モデルを用いた場合と用いない場合の認識性能を示す．この図より，１状態の無音モデルを追加することで，各話者認識性能が改善しており，平均で 7.3%単語正解精度が改善していることがわかる．特に話者４に関しては大きく改善しており，この話者の発声において，音節間に無音区間が多く挿入されていることがわかる．

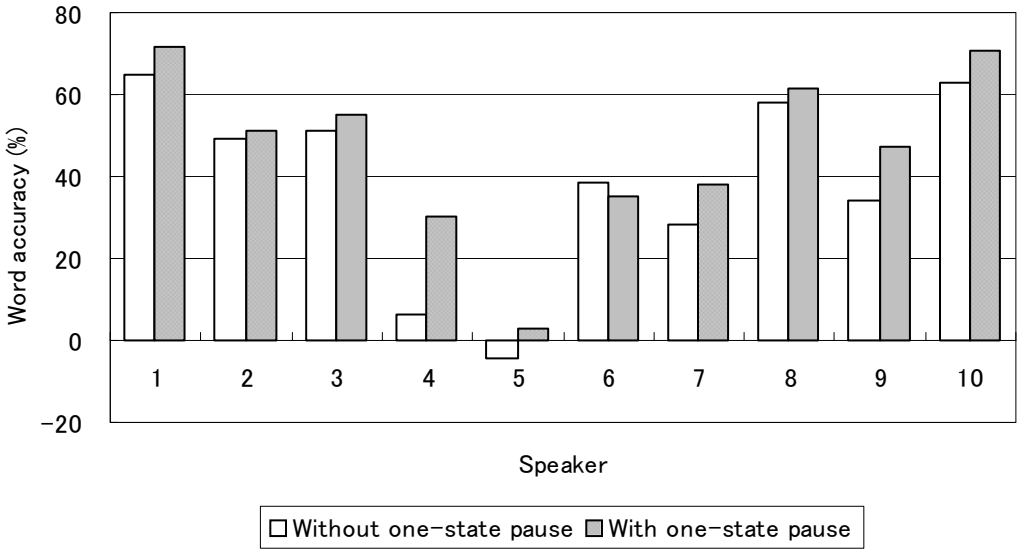


図５．１３　無音モデルの有無による単語正解精度

５．７　結言

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，その有効性を確認した．

音響的特徴の変化について調査した結果，誤認識時の言い直し発話においては，音節間の音響的特徴，連続性が変形するとともに，音節間に無音が存在する孤立音節発声にむけて発話スタイルが変形する，音節強調発声になることが明らかになった．また多い人で，言い直し発話の 25%程度が，音節強調発声になっていることも明らかになった．

収録した言い直し発話を用いた認識実験を行った結果，通常の発話と比較し，音節強調発声は認識性能が著しく劣化することを示した．音節強調発声は音響的特徴の変化が大きすぎるため，話者適応などでは十分な認識性能の改善は得られなかった．

以上の結果を踏まえ本章では，音節強調発声の音響的特徴を考慮したモデル構造を有する音響モデルの構築を提案した．この手法は，音節強調発声の音響的特徴である音節間の音響的特徴，連続性の変化に対しては先行音素環境依存 biphone 母音モデルを利用することで，孤立音節発声に近い発声に対しては後続音素環境が無音の triphone 母音モデルをマルチパス化することで，それぞれの音響的特徴を吸収するものである．それぞれの母音モデルは，既存の音響モデルの学習データを利用して構築できるため，新たに学習データを収録する必要はない．

認識実験を行った結果，学習データの追加や認識辞書の拡張等を行うことなく，ほとんど認識できなかった，意図的に発声した音節強調発声に対して平均 50%，話者適応を併用した場合で平均 70%の単語正解精度を得ることができた．また，通常の発声においても若干ではあるが認識性能が改善した．これは通常の発声においても，局所的には後続音素環境が無音の母音モデル，先行音素環境依存 biphone モデルの方がマッチングの良い区間が存在ためである．以上の結果より提案手法は，音節強調発声の出現を検出することなく，通常発声，音節強調発声両方に対して認識性能を改善可能であることを示すことができた．

本章を通じて，モデルパラメータだけではなく，認識対象音声の音響的特徴が表現できるモデル構造の検討などが，発話スタイルの変動に対しては重要であることを示した．統計的手法の導入により飛躍的に性能が向上した音声認識であるが，認識対象の音響的特徴分析を通じたモデル構造の検討など実験的アプローチも，今後の音声認識研究においては重要であることを示した．

参考文献

[1] 奥田浩三, 松井知子, 中村 哲, “誤認識時の言い直し発話における発話スタイルの変動に頑健な音響モデル構築法,” 電子情報通信学会論文誌, vol.J86-D-II, no.1, pp.42-51, 2003.

[2] Kozo Okuda, Tomoko Matsui, Satoshi Nakamura, “Towards the Creation of Acoustic Models for Stressed Japanese Speech,” Proc. of Eurospeech2001, vol.3, pp.1653-1656, 2001.

[3] S.Oviatt, “The CHAM model of hyperarticulate adaptation during human-computer error resolution,” Proc. of ICSLP'98, pp.2311-2314, 1998.

[4] H.Soltau and A.Waibel, “On the influence of hyperarticulated speech on the recognition performance,” Proc. of ICSLP'98, pp.229-232, 1998.

[5] M.J.F.Gales and P.C.Woodland, “Mean and variance adaptation within the MLLR framework,” Computer Speech and Language, vol.10, pp.249-264, 1996.

[6] 中川聖一, 越川 忠, “最大事後確率推定法を用いた連続出力分布型 HMM の適応化,” 日本音響学会誌, vol.49, no.10, pp.721-728, 1993.

[7] H.Soltau and A. Waibel, “Specialized acoustic models for hyperarticulated speech,” Proc. of ICASSP2000 , pp.1779-1782, 2000.

[8] 山本博史, シンガー ハラルド, リーブス ベン, 匂坂芳典, “日英音声翻訳システム「ＡＴＲ－ＭＡＴＲＩＸ」における音声認識部分の構造と制御方法,” 日本音響学会 1998 年春季研究発表会, 2-Q-21, 1998.

[9] A.Nakamura, S.Matsunaga, T.Shimizu, M.Tonomura and Y.Sagisaka, “Japanese speech database for robust speech recognition,” Proc. of ICSLP'96, pp.2199-2202, 1996.

[10] 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,” 電子情報通信学会論文誌, vol.J79-D-II, no.10. pp.2155-2164, 1993.

[11] 山本博史, 匂坂芳典, “接続の方向性を考慮した多重クラス複合 N-gram 言語モデル,” 情報処理学会研究報告, SLP98-24, 1998.

[12] Entropic Ltd. The HTK Book (for HTK Version 2.2), 1999.

[13] M.Tonomura, T.Kosaka and S.Matsunaga, “Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation,” Computer Speech and Language, vol.10, pp.117-132, 1996.

[14]J.Takahashi and S.Sagayama, “Vector-field-smoothing Bayesian learning for fast and incremental speaker/telephone-channel adaptation,” Computer Speech and Language,

vol.11, pp.127-146, 1997.
[15]大倉計美, 杉山雅英, 嵯峨山茂樹, “混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式,” 電子情報通信学会技術研究報告, SP92-16, 1992.
[16] Lawrence Rabiner, Biing-Hwang Juang　共著, 古井貞熙　監訳, “音声認識の基礎 (上) (下),” NTT アドバンステクノロジ株式会社, 1995.

第 6 章

結論

本論文では，音響モデルの構築方法を中心に，発話スタイルの変動に頑健な音声認識手法について研究を行った．

まず第２章で，隠れマルコフモデル（hidden Markov model ; HMM）を用いた音声認識の基本的な手法についてまとめた．音声認識技術の進展や計算機能力の向上により，認識性能が飛躍的に向上したとは言え，現状の音声認識には多くの課題が残されている．発話スタイルの変動による認識性能の劣化もその一つである．第２章は，発話スタイルの変動に頑健な音声認識手法の研究を行うにあたり，現状の音声認識の基本的な手法と，発話スタイルの変動により認識システムがどのように影響を受けるかについてまとめたものである．本論文の研究対象である音響モデルの構築方法が，認識システムにおいてどのように位置付けられるかを明確にしている．

第３章は，国際電気通信基礎技術研究所（ATR）で収録された 3,700 人規模の多数話者音声データベースを用いた研究内容に関するものである．自然発話音声を高精度に認識するためには，自然発話研究のためのデータベースの整備が不可欠であり，本データベースもこのような観点から構築されたものである．第３章ではこのデータベースの概要についてまとめると共に，このデータベースを用いた音響モデル構築について論じた．具体的には，地域や年齢の違いによる音響的特徴の違いや，これらの違いが認識性能に与える影響について，音響分析や認識実験を通じて明らかにした．地域や年齢の違いによる音響的特徴の違いや，地域別，年齢別音響モデルの構築により認識性能が向上することについては，古くから言われていることであるが，実際に多数話者音声データベースを用いてその違いを明らかにしたことは，音声研究において大きな意味を持つものである．さらに発話スタイルの違いによる音響的特徴の違いについても調査することができたことから，本データベースが研究用大規模音声データベースとして有効であるとも，明らかにすることができた．

第４章では，自由発話の一つである講演音声認識において認識性能を向上するためのデコーディング手法，および音響モデル構築手法について研究し，発話速度の変動を考慮した，より精度の高い音響分析の重要性を示した．現状の音声認識システムは，認識時における計算機コストと認識性能のバランスから，分析周期・窓長は実験的に決定された値で固定される場合が多いが，発話速度の変動による認識性能の劣化を考えると，高い精度での音響分析を行うための，発話速度に依存した分析手法を検討する必要がある．この意味で，ゆう度基準により分析周期・窓長を選択する提案手法は有効である．本研究で提案した手法は，１発話内では発話速度が一定であるとの仮定のもとでの認識手法であるが，１発話内での発話速度変動までを考慮した音響分析手法や音響モデル構築方法へも展開が可

能である．

第 5 章においては，誤認識時の言い直し発話の音響的特徴を分析することで，多くの学習データを収集することなく，誤認識時の言い直し発話に対する認識性能を改善する手法を提案した．誤認識時の言い直し発話を分析することで，言い直し発話において音節強調発声の出現頻度が増加することを明らかにするとともに，その音響的特徴の変化の傾向をモデル化することで認識性能が改善できることを示した. HMM を用いた音声認識では現在まで，より多くのデータを収集することで認識性能の改善を図ってきたが，発話スタイルの変動に対する認識性能の向上のためには，多くのデータを収集するだけではなく，その発話スタイルを効率良くモデル化するための手法の検討が有効であることを，本研究を通じて示すことができた．誤認識の発生を避けることができない音声認識システムにおいては，言い直し発話に対する認識性能の劣化をいかに防ぐかが，重要な課題であることは言うまでも無く，本研究成果は単語正解精度だけでは表すことのできない認識システムの性能（音声対話システムでは，タスク達成率など）に対して大きく貢献するものである．

統計的手法の導入と大規模音声データベースの構築により，飛躍的に性能が向上した音声認識技術ではあるが，いまだシステムの認識対象となる発話スタイルは限られている．より多くの発話スタイルに対して認識性能を向上するため，それら全ての発話スタイルに対応する音声データベースの構築を待っていては，音声認識技術の利用範囲を拡大することはできない．限られたデータを分析し，各発話スタイルを効率よく分析，モデル化できる音声認識手法を研究することが，音声認識技術の利用範囲の拡大において重要であることを，本論文を通じて明らかにした．

業績一覧

学術論文

[1] 奥田浩三, 松井知子, 内藤正樹, 匂坂芳典, 中村 哲, “大規模日本語音声データベースの構築と評価,” 日本音響学会誌, vol.58, no.9, pp.569-578, 2002.

[2] 奥田浩三, 松井知子, 中村 哲, “誤認識時の言い直し発話における発話スタイルの変動に頑健な音響モデル構築法,” 電子情報通信学会論文誌, vol.J86-D-II, no.1, pp.42-51, 2003.

[3] 奥田浩三, 川原達也, 中村 哲, “ゆう度基準による分析周期・窓長の自動選択手法を用いた発話速度の補正と音響モデル構築,” 電子情報通信学会論文誌, vol.J86-D-II, no.2, pp.204-211, 2003.

国際学会

[1] Tomoko Matsui, Masaki Naito, Yoshinori Sagisaka, Kozo Okuda, Satoshi Nakamura,” Analysis of Acoustic Models Trained on a Large-Scale Speech Database,” ICSLP2000 (6th International Conference of Spoken Language Processing), vol.2, pp.503-506, 2000.

[2] Kozo Okuda, Tomoko Matsui, Satoshi Nakamura,” Towards the Creation of Acoustic Models for Stressed Japanese Speech,” Eurospeech2001, vol.3, pp.1653-1656, 2001.

[3] Kozo Okuda, Tatsuya Kawahara, Satoshi Nakamura,” Speaking Rate Compensation Based on Likelyhood Criterion in Acoustic Model Training and Decoding,” ICSLP2002 (7th International Conference on Spoken Language Processing), vol.4, pp.2589-2592, 2002.

国内学会

[1] 奥田浩三, 松井知子, 中村 哲, “自然発話音声における音節強調発声に頑健な音響モデルの構築法,” 日本音響学会 2001 年春季研究発表会, vol.1, pp.41-42, 2001.

[2] 奥田浩三, 中村 哲, “スペクトルの時間変化量に基づく可変分析フレームを用いた発話スタイルの変動に頑健な音響モデルの構築,” 日本音響学会 2001 年秋季研究発表会, vol.1, pp.9-10, 2001.

[3] 奥田浩三，南角吉彦，中村 哲，“音響特徴パラメータの相関を利用した音素継続時間長の正規化，”日本音響学会 2002 年春季研究発表会，vol. 1，pp. 19-20，2002.

国内研究会

[1] 奥田浩三，松井知子，中村 哲，“音節強調発声に頑健な自然発話音声の認識法，”電子情報通信学会技術研究報告，vol. 100，no. 523，pp. 19-24，2000.
[2] 奥田浩三，中嶋秀治，松井知子，河原達也，中村 哲，“講演音声認識のための音響モデル構築方法の検討，”ワークショップ 「話し言葉の科学と工学」，pp. 109-116，2001.
[3] 奥田浩三，中嶋秀治，河原達也，中村 哲，“講演音声の音響的特徴分析と音響モデル構築方法の検討，”情報処理学会研究報告，vol. 2001，no. 68，pp. 73-78，2001.
[4] 奥田浩三，河原達也，中村 哲，“講演音声認識における発話速度の変動を考慮した音声認識手法，”電子情報通信学会技術研究報告，vol. 101，no. 523，pp. 13-18，2001.
[5] 奥田浩三，河原達也，中村 哲，“発話速度の補正を用いた音響モデルの構築，”ワークショップ 「話し言葉の科学と工学」，pp. 65-72，2002.

謝辞

本論文をまとめるにあたって，研究室の立ち上げという忙しい時期であったにも拘わらず，終始変わらぬご指導とご鞭撻を賜った大阪市立大学大学院工学研究科 會田田人教授に心より感謝いたします.

同じく本論文をまとめるにあたり，有益な御教示を賜った大阪市立大学大学院工学研究科 濱 裕光教授, 辻本浩章教授, 平井 誠助教授, 高橋秀也助教授に厚く御礼申し上げます.

大学時代の指導教官であり，本論文をまとめるにあたりさまざまなアドバイスをいただきました，宝塚造形芸術大学　志水英二教授に心より感謝いたします.

本研究は，筆者が（株）ＡＴＲ音声言語コミュニケーション研究所において行ったものである．本研究の機会を与えていただいた，（株）ＡＴＲ音声言語コミュニケーション研究所 山本誠一前所長（現 同志社大学 教授）に感謝いたします．また本研究を進めるにあたり多大なご指導を賜りました，(株)ＡＴＲ音声言語コミュニケーション研究所 中村 哲所長に深く感謝いたします．講演音声認識に関する研究において多岐にわたるご指導を賜りました(株)ＡＴＲ音声言語コミュニケーション研究所第１研究室 河原達也教授（現 京都大学学術情報メディアセンター 教授）に厚く御礼申し上げます．また，多数話者音声データベースを用いた研究，誤認識の言い直し発話に頑健な音響モデルの研究に関して，多大なご指導を賜りました(株)ＡＴＲ音声言語コミュニケーション研究所第１研究室 松井知子助教授（現 統計数理研究所 助教授）に深く感謝いたします．多数話者音声データベースを用いた研究，講演音声認識に関する研究において，多岐にわたるご指導を賜りました(株)ＡＴＲ音声言語コミュニケーション研究所第２研究室 匂坂芳典教授（現 早稲田大学 教授）に深く感謝いたします.

多数話者音声データベースを用いた研究に関して，様々なデータを提供いただいたＫＤＤＩ研究所の内藤正樹氏，言語モデルに関する技術的なご指導，ご支援を賜った（株）ＡＴＲ音声言語コミュニケーション研究所第２研究室 山本博史氏，さまざまな言語モデルを提供いただいた元（株）ＡＴＲ音声言語コミュニケーション研究所第２研究室 中嶋秀治氏（現 ＮＴＴ）に深く感謝いたします．実験環境の構築からデータ収録に至るまで，心強い支援をいただきました（株）ＡＴＲ音声言語コミュニケーション研究所テクニカルサポートグループの皆様，データ収録からデータの整備に至るまで，音声認識の研究を行う上でなくてはならない部分をサポートいただきました（株）ＡＴＲ音声言語コミュニケーション研究所ラベラーの皆様に深く感謝いたします.

（株）ＡＴＲ音声言語コミュニケーション研究所への出向という機会を作っていただき

ました，三洋電機(株) 研究開発本部 デジタルシステム技術開発センター 虎沢研示前所長，企画室 片山 立室長，ヒューマンインタフェース研究部 大橋秀紀前課長，藤本光男前課長に厚く御礼申し上げます．本論文の執筆するにあたり多大なご支援をいただきました，三洋電機（株） 研究開発本部 ヒューマンエコロジー研究所 安田昌司所長，奥田泰生企画課長に深く感謝いたします．（株）ＡＴＲ音声言語コミュニケーション研究所出向中多岐にわたるご支援と，三洋電機復帰後さまざまなご指導を賜りました三洋電機(株) 研究開発本部デジタルシステム研究所 大西宏樹部長，三洋電機に復帰後の研究遂行にあたり様々なご指導を賜りました三洋電機(株) 研究開発本部経営企画室 船造康夫室長に，深く感謝いたします．本研究をまとめるにあたり，多岐にわたるアドバイスをいただいた三洋電機(株) 研究開発本部ヒューマンエコロジー研究所ホームアメニティ研究部 蚊野 浩部長，大倉計美課長に厚く御礼申し上げます．音声合成の視点から，さまざまな技術ディスカッションを行っていただいた平井啓之主任をはじめ，三洋電機(株) 研究開発本部ヒューマンエコロジー研究所ホームアメニティ研究部の皆様に深く感謝いたします．

最後に，仕事を持ちながら家庭を守り，且つ精神的な支えとなってくれた妻 綾子，一緒に遊びたいのを我慢し，小さいながらに協力してくれた長女 雪乃に心から感謝いたします．