ＴＲ－ＳＬＴ－００６１

対話音声における発話スタイルの分析とそれに基づく認識率の改善

Improvement of Speech Recognition Rate Based on Analysis of Speaking Styles in Dialogue Speech

青野　邦生† ‡
Kunio Aono
安田　圭志† ‡　　竹澤　寿幸†
Keiji Yasuda　　Toshiyuki Takezawa

２００４年３月２日

概要

音声翻訳システムや対話音声システムは，会話調の音声を処理する必要がある．現在の音声認識技術は，大規模音声コーパスの開発や計算機の高速化の後押しを受け，読み上げ調の音声に対しては高性能の音声認識システムが開発されている．しかし，会話調の音声の認識については多くの課題が残されており，認識率を向上させる必要がある．本研究では，まず，予備実験として単語の持つ情報である品詞，言語尤度に注目し，発話スタイルとそれらの情報との関連性を調べる．そして，それにより得られた知見に基づき，発話スタイルの異なる二つの音響モデルを単語単位で自動選択することにより，音声認識率の改善を行う．その結果，単独の音響モデルを用いた場合よりも，単語誤り率が，1.62 ポイント低減できている．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２　TEL：0774－95－1301
† ATR 音声言語コミュニケーション研究所　‡ 同志社大学工学部

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone　+81-774-95-1301
Fax　+81-774-95-1308

目次

第1章 序論

音声翻訳システムや対話音声システムは，会話調の音声を処理する必要がある．現在の音声認識技術は，大規模音声コーパスの開発や計算機の高速化の後押しを受け，読み上げ調の音声に対しては高性能の音声認識システムが開発されている．しかし，会話調音声の認識については多くの課題が残されている．

本論文では，朗読発話と自然発話の 2 種類の発話スタイルに注目する．朗読発話とは，テキストを読み上げた朗読音声のような明瞭な発声である．それに対し自然発話とは，対話音声のような自発的発声である．朗読音声と対話音声を比較すると，言語的な面から見た場合，対話音声では朗読音声とは異なり，言い直し，省略，間投詞といった書き言葉には見られない現象が多く出現する[1]．また，音響的な面から見た場合，朗読音声は明瞭に発話され，発話速度の分散が小さい．それに対して，対話音声は，朗読音声に比べ乱雑に発声されており，その発話速度の分散が大きいといった特徴が挙げられる[2]．

音声翻訳システムや音声対話システムといったシステムを介した対話を想定すると，会話調の音声を処理しなければならないため，朗読音声ではなく対話音声で音響モデルを学習させることが望ましい．しかし，雑音の少ない環境下で大量の対話音声データを集めるのは困難である．そのため，現状の音声データベースでは朗読音声を収録したものが圧倒的に多い．対話音声を対象とした音声認識では，数少ない対話音声を用いて音響モデルを構築するのではなく，双方を同等に扱うことにより高精度な音声認識を行うことが望まれる．

現在の音声認識技術では，適切な音響モデルを発話単位で動的に選択することが可能である．また，システムを介した対話実験によれば，システムに慣れた話者と不慣れな話者では発話スタイルが異なることが知られている[3]．そして，同一話者においても発話内容に応じて発話スタイルが変化することが知られている[4]．そのため，従来の解決策として，朗読音声により学習された音響モデルと対話音声により学習された音響モデルを動的に選択することにより，音声認識を向上できるといった研究が行われている[4]．

発話より小さい単位で発話スタイルが変化するのであれば，発話より小さい単位で複数の音響モデルを自動選択することによって，音声認識性能の改善が期待できる．そこで本論文では，発話より小さい単位である単語単位により，発話スタイルの違いを分析する．なお，単語の持つ情報である品詞，言語尤度に注目し，発話スタイルとそれらの情報との関連性を調べる．そして，それにより得られた知見に基づき，音声認識率の改善を行う．

第2章 発話スタイルに依存した音声データベース

2.1 緒言

本論文では，朗読音声により学習された音響モデルと対話音声により学習された音響モデルを扱っている．そのため，その音響モデルの学習のために，朗読音声と対話音声を収録した音声データベースが必要である．しかし，朗読音声に関する音声データベースは多く集められているが，対話音声に関する音声データベースは数少ない．そこで，本章ではそのような発話スタイルに依存した音声データベースの現状について述べる．

2.2 音声データベースの必要性

現在の音声言語処理技術は数理統計処理にその基盤を置いている．音声処理の研究・開発を進めるためには，正確な統計量を測る必要があり，量的に少なくては正しい統計量を測ることができない．そのため，大量の音声・言語素材が必要である．

音声認識システムにおいては，音声データベースを主に音響モデルの学習に利用する．音響モデルは，主に音素を単位としており，音素の音響的な特徴は，前後の音素の影響を受けて大きく変化する．このような現象は，調音結合と呼ばれ，音声の認識を困難にしている大きな要因の 1 つである．調音結合に対する最も直接的な対処方法は，前後の音素を考慮した 3 つの音素の組み合わせ（トライフォン）を認識の処理単位として用いることである．トライフォンによるモデル化では，処理単位の数が音素の 3 乗という膨大な数に上る．例えば，日本語の場合，音素の種類数を 27 とすると，トライフォンの種類数は 3,590 である．そのため，トライフォンに対して十分な学習データを確保するには，大量の音声データが必要であり，本研究で使用する音響モデルでは音素情報をタグ付けした音声を計 28 時間用いている．なお，音響的特徴が類似したトライフォンをグループ化することによってモデルの数を削減することが広く行われている．

また，音声認識システムの目的や研究内容によって要求されるデータはさまざまであり，例えば，日本語の音声認識を行うには，日本語の音声データベースが必要であるし，音声対話システムを構築する場合には，対話を収録した音声データベースが必要となる．

2.3 音素バランス文

音響モデルを学習することを考えると，発話内容に関して何らかの制限を行設けた方が効率がよい．例えば，目的とする音響モデルの単位（音素や音節）や組み合わせが万遍なく出現していれば，どの単語も偏りなしに良いモデルが学習できる．音素に関してこのよ

うに設計された単語または文の集合を音素バランス単語，あるいは音素バランス文と呼ぶ[5]．なお，音素バランス文は，音響モデルの学習を目的として選定されるため，文としての自然さを特に考慮する必要はない．

2.4 朗読音声データベース

朗読音声のデータベースとしては，ATR 音声言語コミュニケーション研究所で作成された「ATR 音素バランス文連続音声コーパス」（64 名，約 113,000 発話）や日本音響学会音声データベース調査研究委員会と情報処理学会・音声言語情報処理研究連絡会大語彙連続音声認識研究用データベースワーキンググループにより作成された「新聞記事読み上げ音声データベース：JNAS」（306 名，約 47,000 発話）[6]が代表的である．

現在の連続音声認識システムのほとんどは，音素や音節を基本認識単位として設定し，その組み合わせによって文章を認識している．また，その基本単位の認識には HMM に代表される統計的手法を用いている．このため，基本認識単位の学習用として，多種大量の連続発声データが必要である．できるだけ広範な音声現象が，効率よくコーパスに含まれることが必要である．このコーパスでは，ATR 音素バランス 503 文を収集対象として選び，データを収集している．以下に 503 文の例を示す．このように本コーパスでは，さまざまな音素の組み合わせが出現するよう意図的に収録に用いる文を選定されている．

[ATR 音素バランス 503 文の例]

01) あらゆる現実をすべて自分のほうへねじ曲げたのだ。
02) 一週間ばかりニューヨークを取材した。
03) テレビゲームやパソコンでゲームをして遊ぶ。
04) 物価の変動を考慮して給付水準を決める必要がある。
05) 救急車が十分に動けず、救助作業が遅れている。
・・・・・・・・・・

2.5 対話音声データベース

近年では，対話音声を収集したデータベースが盛んに作成されている．その代表としては，重点領域研究「音声対話」による音声対話コーパス（83 名，93 対話）[8]や「ATR 旅行会話データベース」（500 名，約 21,000 発話）[9]がある．しかし，対話音声では，相手の発話中に同時に発言（Simultaneous Talk）したり，あいづちや相手の発話に割り込んだりすることがよく起きる．このように 2 つの音声が混じっているデータは音声認識の学習用データとしては利用しにくい．また，収録の際に防音室等を使うと日常的環境と異なることやマイクが前にあるとそれを意識するので自然発話でなくなるなどの問題がある．他にも，

真の自然発話を録音するために，マイクを隠して収録すると人権侵害になる等の問題もある．特の感情音声（喜び，悲しみ，怒りなどの感情を持って発声した音声）の収録は難しい．このように，対話音声を収集した音声データベースは，朗読音声と異なり非常に集めづらい．近年では，多くの対話音声が集められており，「ATR 旅行会話データベース」では約 21,000 発話と大規模なコーパスが作成されているが，朗読音声を収録した「ATR 音素バランス文連続音声コーパス」では約 113,000 発話であり，朗読音声に比べると量的に少ない．

2.6 結言

本章では，音声データベースの必要性，および朗読音声・発話音声を収集した音声データベースについて述べた．上記で述べたように，朗読音声の収集は容易であり，さまざまな音素の組み合わせが出現するよう計画的に収集されている．それに対して，対話音声は，非常に集めづらく，量的にも十分であるとはいえない．そのため，より高性能な対話音声認識システムを構築するには，対話音声を用いて学習された音響モデルだけではなく，朗読音声を用いて学習された音響モデルも用いる必要がある．

第3章 品詞，言語尤度と発話スタイルの関係

3.1 緒言

本章では，対話音声について発話スタイルの違いを単語単位で調べる．ただし，本論文では朗読発話的発声と自然発話的発声を発話スタイルと呼ぶことにする．具体的には，品詞，言語尤度といった単語の持つ情報と発話スタイルとの関係を調べる．分析方法としては，朗読発話，自然発話それぞれで学習された音響モデルを用いて，どういった場合にどちらの音響モデルが適切であるのかを調査する．

3.2 分析条件

3.2.1 音響モデルの構築

本研究では，発話スタイルとして自然発話と朗読発話を選び，男女別に音響モデルを準備した．自然発話としては日本人同士の対話音声，朗読発話としては音素バランス文の朗読音声を用いた．音声の分析条件を表 1 に示す．このように，音声データから 12 次元のメル周波数ケプストラム係数（MFCC）とその一次差分（ΔMFCC）とパワーの一次差分（Δpower）を計算し，それら 25 次元を音響モデルを構築するための特徴パラメータとした．

表 1 音声の分析条件

サンプリング速度	16k samples/sec
分析窓長	20m sec（Hamming 窓）
窓間隔	10m sec
特徴パラメータ	MFCC（12 次元） ＋ΔMFCC（12 次元）＋Δpower
HMnet	音声 1,400 状態 3 混合 無音 3 状態 10 混合

次に，学習に用いた音声データの概要を表 2 に示す．朗読発話学習用音声データセットには，「ATR 音素バランス文連続音声コーパス」を用いている．このコーパスは 2.4 で説明したものと同じである．そして，自然発話学習用音声データセットには，旅行会話を収録した日本人同士の対話音声[4]を用いている．この音声データの収録の際には，被験者に空港やホテルのフロントといった状況を想定してもらい，録音スタジオで自由に会話を行ってもらっている．一度の収録には 2 人の被験者が会話を行い，アイコンタクトやジェスチャーなどの音声以外の情報伝達を避けるために，被験者と被験者の間には仕切りを立てて

いる．また，収録に用いるピンマイクは被験者に取り付けている．そのため，これらの音声は真の自然発話とはいえない．

なお，朗読発話音声により学習された音響モデルを「朗読発話音響モデル」と呼び，対話音声により学習された音響モデルを「自然発話音響モデル」と呼ぶことにする．

表 2 学習用音声データの概要

朗読発話学習用音声データセット（音素バランス文）
男性：165 話者，総発話時間 約 9 時間
女性：235 話者，総発話時間 約 14 時間
自然発話学習用音声データセット（日本人同士の旅行対話）
男性：167 話者，総発話時間 約 2 時間
女性：240 話者，総発話時間 約 3 時間

3.2.2 言語モデルの構築

言語モデル構築に用いた学習データの概要を表 3 に示す．表に示すように，言語モデル構築には 2 種類の学習データを用いている．一方が日本人同士の対話[4]であり，自然発話音響モデルの学習に用いた音声を書き起こしたものである．もう一方が通訳者を介した対話[9]である．この学習データの収録条件は日本人同士の対話と同じであり，2 人の被験者と通訳者の間には仕切りを立てている．また，この 2 つの学習データの言語的特性の違いはない．なお，言語モデルには，多重クラスバイグラム[10]を用いている．

表 3 言語モデルの学習データ

日本人同士の対話
(a) 約 26,000 発話，約 39 万単語
日本語－英語の対話（日本語側のみ）
(b) 約 85,000 発話，約 72 万単語

3.2.3 評価データ

分析に用いた音声データは，通訳者を介した日本語－英語の対話音声（日本語側のみ），話者数 71 人，約 16,000 発話である．評価データは，音響モデルおよび言語モデルで用いている学習データとは異なるデータを用いている．この評価データは，自然発話音響モデルに用いる学習データと同じ収録条件で収集されている．なお，通訳者を介した対話音声の発話スタイルは，対話システムを介した場合の発話スタイルと類似しているという報告がある[4]．

3.3 分析方法

3.3.1 音響尤度の比較

一般に，音声認識においては，正解系列の音響尤度が高くなることが好ましい．そこで，朗読発話と自然発話の各音響モデルを用い，評価データについての単語単位の音響尤度を求め，モデル間の優位性の比較を行う．そして，自然発話音響モデルを用いた場合の音響尤度が，朗読発話音響モデルを用いた場合よりも高くなる単語の割合によりモデルの優位性を評価する．この割合のことを「自然発話音響モデル優位率（WR_S）」と呼ぶことにし，次式のように定義する．

$$WR_S = 100 \times {N_s}/{N_{total}} \qquad \text{(式 1)}$$

ただし，N_{total} は全単語数を，N_S は自然発話音響モデルを用いた場合の音響尤度が朗読発話音響モデルを用いた場合よりも高くなる単語の数を表す．なお，WR_S が 50 より大きい場合は，自然発話音響モデルが朗読発話音響モデルよりも認識に適していることになる．

3.3.2 単語正解率の差による比較

本分析では，各音響モデルを用いて認識実験を行い，品詞，言語尤度に基づき単語正解率の集計を行う．ここで，朗読発話音響モデルを用いた音声認識を「R-system」，自然発話音響モデルを用いた音声認識を「S-system」と呼ぶことにする．R-system，S-system には，ともに同一のデコーダと言語モデルを用いている．そして，品詞，言語尤度といった情報に基づき集計された R-system と S-system の単語正解率の差により比較する．なお，単純に差を比較するのではなく，次式により単語正解率の差を正規化している．

$$D_{correct} = \frac{S_{correct} - R_{correct}}{100 - \max\{R_{correct}, S_{correct}\}} \qquad \text{(式 2)}$$

ただし，$D_{correct}$ は正規化された単語正解率の差を，$R_{correct}$ は R-system の単語正解率を，$S_{correct}$ は S-system の単語正解率をそれぞれ表す．

3.4 分析結果

3.4.1 品詞と発話スタイルの関係

図 1 は，自然発話音響モデル優位率 WR_S を品詞ごとに集計した結果である．図中の縦軸は，自然発話音響モデル優位率を表し，横軸は品詞を表している．また，評価データにつ

いての各品詞の単語数を図 2 に示す．図 1 から分かるように，品詞によって朗読発話に近い品詞と自然発話に近い品詞とに大きく分かれている．具体的には，自然発話特有の品詞である感動詞，間投詞や，文末表現である助動詞では自然発話音響モデル優位率が高いので，その発話スタイルは自然発話に近いことが分かる．それに対して，内容に関する重要な情報を伝達する名詞類では自然発話音響モデル優位率が低く，その発話スタイルは朗読発話に近いことが分かる．なお，自然発話音響モデル優位率が品詞ごとに有意な差があるかどうかの検定（χ^2 独立性の検定）を行っている．その検定結果を表 4 を示す．表の値は，行と列の品詞の自然発話音響モデル優位率の差を表している．

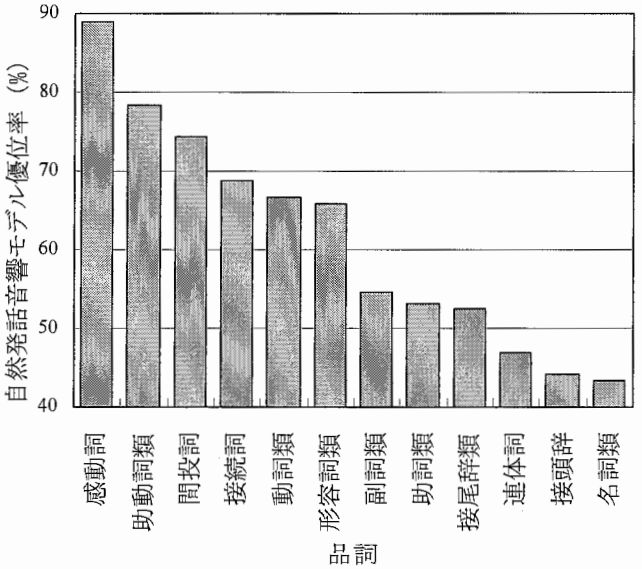

図 1 品詞ごとに集計した自然発話音響モデル優位率

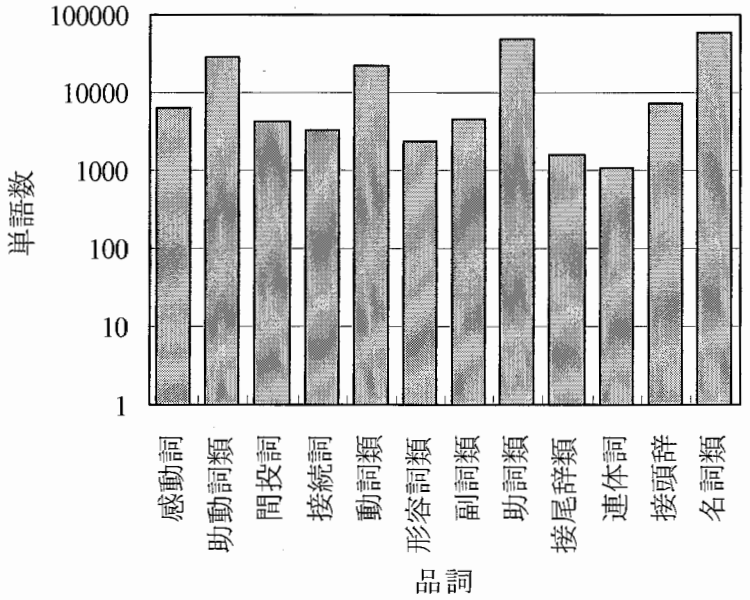

図 2 各品詞の単語数

表 4 品詞ごとに集計した自然発話音響モデル優位率についての差の検定

	感動詞	助動詞類	間投詞	接続詞	動詞類	形容詞類	副詞類	助詞類	接尾辞類	連体詞	接頭辞	名詞類
感動詞	-	10.6***	14.6***	20.2***	22.3***	23.1***	34.4***	35.9***	36.5***	42.2***	44.9***	45.6***
助動詞類	10.6***	-	4.0***	9.5***	11.7***	12.4***	23.7***	25.3***	25.9***	31.5***	34.2***	34.9***
間投詞	14.6***	4.0***	-	5.6***	7.8***	8.5***	19.8***	21.3***	21.9***	27.6***	30.3***	31.0***
接続詞	20.2***	9.5***	5.6***	-	2.2*	2.9*	14.2***	15.7***	16.3***	22.0***	24.7***	25.4***
動詞類	22.3***	11.7***	7.8***	2.2*	-	0.7	12.0***	13.5***	14.2***	19.8***	22.5***	23.2***
形容詞類	23.1***	12.4***	8.5***	2.9*	0.7	-	11.3***	12.8***	13.4***	19.1***	21.8***	22.5***
副詞類	34.4***	23.7***	19.8***	14.2***	12.0***	11.3***	-	1.5*	2.1	7.8***	10.5***	11.2***
助詞類	35.9***	25.3***	21.3***	15.7***	13.5***	12.8***	1.5*	-	0.6	6.3***	9.0***	9.7***
接尾辞類	36.5***	25.9***	21.9***	16.3***	14.2***	13.4***	2.1	0.6	-	5.7**	8.4***	9.1***
連体詞	42.2***	31.5***	27.6***	22.0***	19.8***	19.1***	7.8***	6.3***	5.7**	-	2.7	3.4*
接頭辞	44.9***	34.2***	30.3***	24.7***	22.5***	21.8***	10.5***	9.0***	8.4***	2.7	-	0.7
名詞類	45.6***	34.9***	31.0***	25.4***	23.2***	22.5***	11.2***	9.7***	9.1***	3.4*	0.7	-

*p<0.05, **p<0.01, ***p<0.001

また，図 3 に品詞ごとに集計した単語正解率および $D_{correct}$ を示す．図中の横軸である品詞は図 1 と同様の並びである．このように，品詞により $D_{correct}$ の値は大きく異なることが分かる．このように，自然発話音響モデル優位率による比較と同様に，感動詞，助動詞に関しては，自然発話音響モデルを用いた方が高い認識率が得られている．逆に，名詞類に関しては，朗読発話音響モデルを用いた方が高い認識率が得られている．なお，各品詞についてR-systemとS-systemの単語正解率に有意な差があるかどうか検定(χ^2独立性の検定)を行っている．その検定結果を表 5 に示す．表の値は R-system と S-system の単語正解率の差を表している．

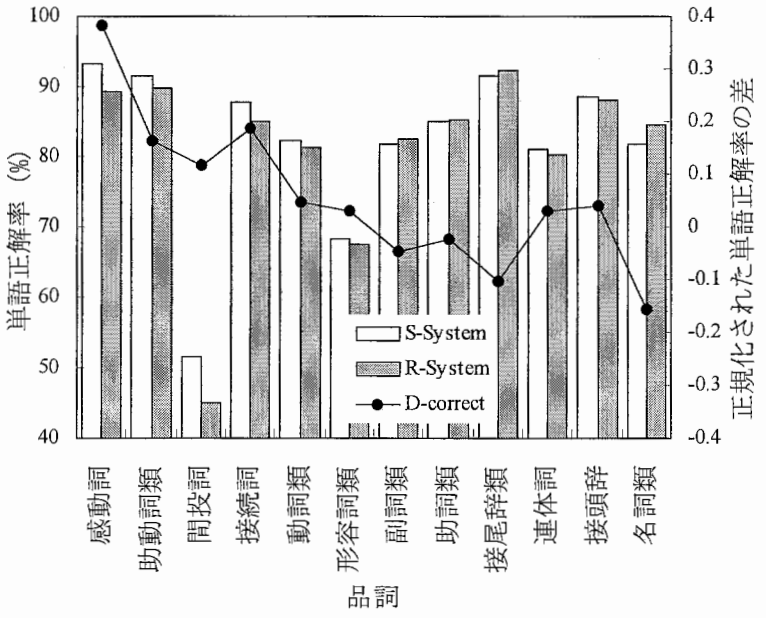

図 3 品詞ごとに集計した単語正解率

表 5 品詞ごとに集計した単語正解率についての差の検定

品詞	単語正解率の差
感動詞	4.1***
助動詞類	1.6***
間投詞	6.4***
接続詞	2.8***
動詞類	0.8*
形容詞類	0.9
副詞類	0.8
助詞類	0.4
接尾辞類	0.9
連体詞	0.6
接頭辞	0.5
名詞類	2.9***

*p<0.05, **p<0.01, ***p<0.001

3.4.2 言語尤度と発話スタイルの関係

図 4 は自然発話音響モデル優位率を，言語尤度の値によって集計した結果である．ここでは，評価データを言語尤度の値によりソートし，単語数が均一になるよう 16 グループに分割している．図中の横軸は，各グループ番号を表しており，その値が小さいほどそのグループに属する単語の言語尤度が低くなっている．縦軸は自然発話音響モデル優位率を表している．なお，図 5 は，評価データを言語尤度の値により集計した頻度分布であり，図中の 1 から 16 までの数字は，各グループ番号を表している．

図 4 を見ると，言語尤度が低いほど，自然発話音響モデル優位率が低く，また，言語尤度が高いほど，自然発話音響モデル優位率が高くなっている．この理由としては，言語尤度の低い単語ほど，対話中でその単語が持つ情報量が大きいため，話者が明瞭に発話し，朗読発話に近い発話スタイルとなっていると考えられる．なお，各グループにより自然発話音響モデル優位率に有意な差があるのか検定（χ^2 独立性の検定）を行っている．その検定結果を表 6 に示す．表の値は行と列のグループ間の自然発話音響モデル優位率の差を表している．

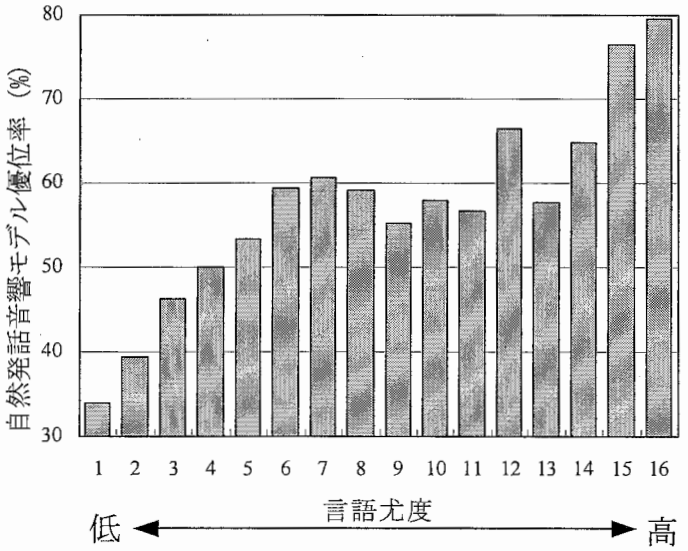

図 4 言語尤度の値により集計した自然発話音響モデル優位率

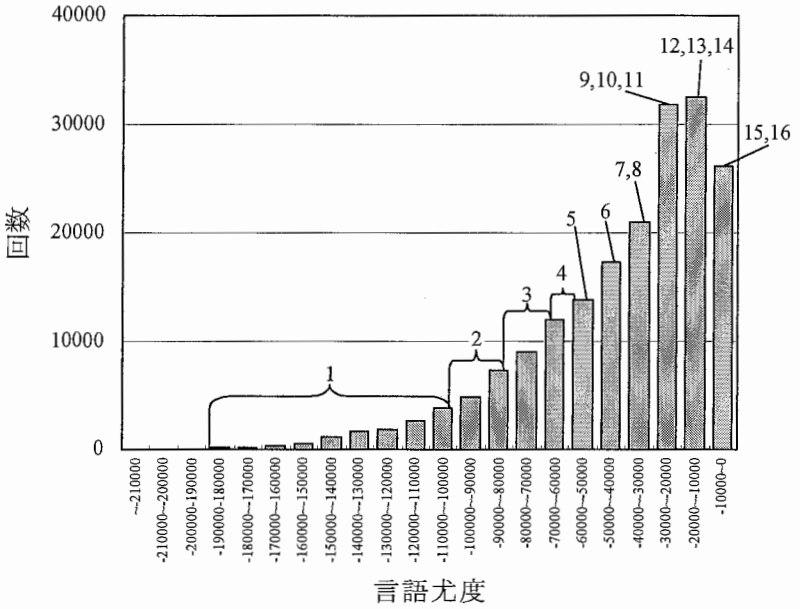

図 5 言語尤度で集計した頻度分布

表 6 言語尤度の値により集計した自然発話音響モデル優位率についての差の検定

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	5.3***	12.2***	16.0***	19.2***	25.3***	26.7***	25.2***	21.1***	23.8***	22.7***	32.4***	23.6***	30.7***	42.4***	45.6***
2	5.3***	-	7.0***	10.7***	14.0***	20.0***	21.4***	19.9***	15.9***	18.6***	17.4***	27.1***	18.4***	25.4***	37.1***	40.3***
3	12.2***	7.0***	-	3.7***	7.0***	13.0***	14.4***	12.9***	8.9***	11.6***	10.4***	20.1***	11.4***	18.5***	30.2***	33.3***
4	16.0***	10.7***	3.7***	-	3.3***	9.3***	10.7***	9.2***	5.1***	7.9***	6.7***	16.4***	7.7***	14.7***	26.4***	29.6***
5	19.2***	14.0***	7.0***	3.3***	-	6.0***	7.4***	5.9***	1.9**	4.6***	3.4***	13.1***	4.4***	11.5***	23.2***	26.3***
6	25.3***	20.0***	13.0***	9.3***	6.0***	-	1.4*	0.1	4.1***	1.4*	2.6***	7.1***	1.6*	5.4***	17.1***	20.3***
7	26.7***	21.4***	14.4***	10.7***	7.4***	1.4*	-	1.5*	5.6***	2.9***	4.0***	5.7***	3.0***	4.0***	15.7***	18.9***
8	25.2***	19.9***	12.9***	9.2***	5.9***	0.1	1.5*	-	4.0***	1.3*	2.5***	7.2***	1.5*	5.5***	17.2***	20.4***
9	21.1***	15.9***	8.9***	5.1***	1.9**	4.1***	5.6***	4.0***	-	2.7***	1.5*	11.3***	2.5***	9.6***	21.3***	24.4***
10	23.8***	18.6***	11.6***	7.9***	4.6***	1.4*	2.9***	1.3*	2.7***	-	1.2	8.5***	0.2	6.9***	18.6***	21.7***
11	22.7***	17.4***	10.4***	6.7***	3.4***	2.6***	4.0***	2.5***	1.5*	1.2	-	9.7***	1	8.0***	19.7***	22.9***
12	32.4***	27.1***	20.1***	16.4***	13.1***	7.1***	5.7***	7.2***	11.3***	8.5***	9.7***	-	8.7***	1.7**	10.0***	13.2***
13	23.6***	18.4***	11.4***	7.7***	4.4***	1.6*	3.0***	1.5*	2.5***	0.2	1	8.7***	-	7.1***	18.8***	21.9***
14	30.7***	25.4***	18.5***	14.7***	11.5***	5.4***	4.0***	5.5***	9.6***	6.9***	8.0***	1.7**	7.1***	-	11.7***	14.9***
15	42.4***	37.1***	30.2***	26.4***	23.2***	17.1***	15.7***	17.2***	21.3***	18.6***	19.7***	10.0***	18.8***	11.7***	-	3.2***
16	45.6***	40.3***	33.3***	29.6***	26.3***	20.3***	18.9***	20.4***	24.4***	21.7***	22.9***	13.2***	21.9***	14.9***	3.2***	-

*p<0.05, **p<0.01, ***p<0.001

また，図 6 に言語尤度の値ごとに集計した単語正解率および $D_{correct}$ を示す．図から分かるように，言語尤度が高いほど $D_{correct}$ は高く，言語尤度が低いほどその値は低くなっている．このように，自然発話音響モデル優位率による比較と同様に，言語尤度の値が高いほど，自然発話音響モデルを用いた方が高い認識率を得られている．逆に，言語尤度の値が低いほど，朗読発話音響モデルを用いた方が高い認識率を得られている．なお，各グループについて R-system と S-system の単語正解率に有意な差があるかどうかの検定（χ^2 独立性の検定）を行っている．その検定結果を表 7 に示す．

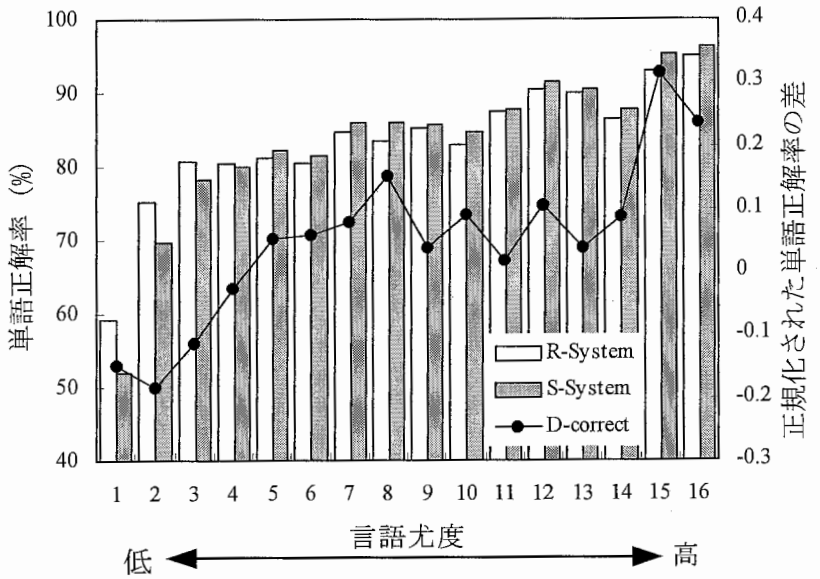

図 6 言語尤度ごとに集計した単語正解率

表 7 言語尤度ごとに集計した単語正解率についての差の検定

グループ番号	単語正解率の差
1	7.1***
2	5.6***
3	2.5***
4	0.5
5	1.0*
6	1.1*
7	1.2**
8	2.5***
9	0.6
10	1.5***
11	0.2
12	1.0**
13	0.4
14	1.2**
15	2.2***
16	1.2***

*p<0.05, **p<0.01, ***p<0.001

3.5 結言

本章では，品詞と発話スタイル，言語尤度と発話スタイルの関係を調べた．その方法としては，朗読発話と自然発話から学習された 2 種類の音響モデルを用い，品詞，言語尤度に基づき比較・分析を行った．その結果，品詞，言語尤度によって望ましい音響モデルが異なることが示唆された．次章では，本章で得られた知見に基づいて，音声認識システムの改善を試みる．

第4章 機械学習を用いた認識結果の自動選択

4.1 緒言

第3章では，品詞，言語尤度の値により，望ましい音響モデルが異なることが示された．本章では，それらの知見から，朗読発話音響モデルを用いた場合と自然発話音響モデルを用いた場合の各認識結果を単語単位で自動選択することによって，認識率の向上を試みる．なお，自動選択には，Support Vector Machine[11]による機械学習を用いる．

4.2 実験条件

本実験で使用する音響モデルは3.2.1と同様である．言語モデルに関しては，3.2.2と同様の学習データを用いるが，本実験では2種類の言語モデルを用意し，認識実験では多重クラス複合バイグラム[10]を，自動選択では多重クラスバイグラム[10]を用いている．その理由は以下の通りである．多重クラス複合バイグラムは出現頻度の高い単語ペアを連鎖語とみなすことで，多重クラスバイグラムを改良した言語モデルである．そのため，多重クラス複合バイグラムを用いる方が，多重クラスバイグラムよりも高い認識率が得られるため，認識実験では多重クラス複合バイグラムを用いる．しかし，多重クラス複合バイグラムは，出現頻度の高い単語ペアを連鎖語とみなすため，自動選択を行う際には，自動選択の比較箇所が減ることになる．そのため，自動選択の際には，連鎖語を用いない多重クラスバイグラムを用いている．なお，両言語モデルとも共通の学習データを用いている．

そして，本実験で使用する機械学習データおよび評価データには，3.2.3と同じものを用いている．

4.3 自動選択手法

本手法は，単語単位による自動選択を目的としているが，比較対象の時間区間が同一でなければ，比較を行うことはできず，認識結果が異なる場合は，必ずしも単語対単語で比較できるとは限らない．そこで，時間的に対応の取れた複数の単語間で比較し，自動選択を行う．本手法は機械学習用データ作成と自動選択の2段階で構成されている．

4.3.1 機械学習データ作成

まず，機械学習用データ作成の手順について説明する．機械学習用データは，各音響モデルを用いた場合の認識結果とその正解系列の3系列を比較することにより作成する．そこで，この3系列の単語単位の対応関係を取らなくてはならない．本手法では，DP (Dynamic

Programming)を適応している．DP は二つのパターンの要素間の対応づけ（整列化）を行い，それによって類似度を計算する方法であり，本手法では，その距離（DP 距離）が最小となる対応づけを採用する．次に，各音響モデルを用いた場合の認識結果と正解系列との比較を行い，どちらか一方が正解で，もう一方が誤りである場合のみを機械学習の学習データとして用いる．ともに誤っている場合や，ともに正解である場合は，学習データには用いない．

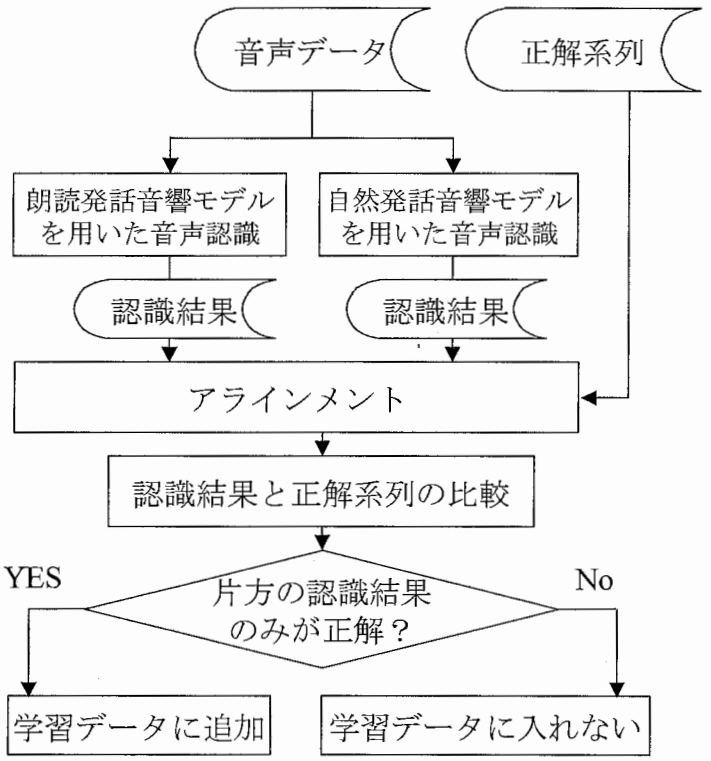

図 7 機械学習データ作成の流れ

4.3.2 自動選択

次に，認識結果の自動選択の方法について説明する．自動選択を行う際には，各音響モデルを用いた場合の認識結果を比較する．この 2 系列の対応づけには， 4.3.1 と同様に DP を用いる．そして，朗読発話音響モデルと自然発話音響モデルによる認識結果出力の単語単位の対応関係を求めて置き，認識結果がともに同じ箇所については，そのまま出力し，異なっている箇所については，自動選択を行う．選択の具体的な方法は次節に述べる．

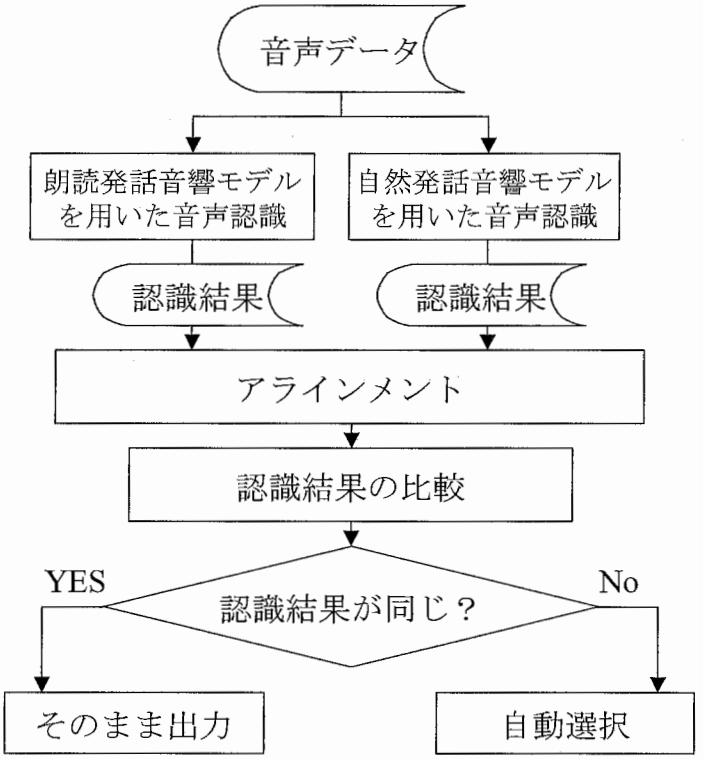

図 8 自動選択の流れ

4.4 自動選択の種類

4.4.1 品詞を用いた自動選択（SVM-1）

品詞により適切な音響モデルが異なることが分かっているので，品詞を素性として自動選択することによる音声認識率の改善を試みる．また，機械学習の素性には各音響モデルを用いた場合の認識結果を品詞ごとに集計し，その単語数を用いた．そのため，素性の数は計 24 次元（12 品詞×2 音響モデル）である．なお，SVM の核関数には Gaussian 関数を用いている．核関数の選択には，最も認識率の高い結果が得られる関数を用いている．SVM-2，SVM-3 についても同様の選び方をしている．

4.4.2 言語尤度を用いた自動選択（SVM-2）

言語尤度が低いほど自然発話音響モデルに適合しやすく，言語尤度が高いほど朗読発話音響モデルに適合しやすいことが分かっている．そこで，言語尤度を素性として自動選択することによる音声認識率の改善を試みる．また，機械学習の素性には各音響モデルを用いた場合の比較区間に対しての言語尤度，計 2 次元を用いている．なお，SVM の核関数には線形関数を用いている．

4.4.3 品詞，言語尤度，音響尤度を用いた自動選択（SVM-3）

品詞，言語尤度をともに素性とした場合の自動選択を試みる．なお，素性には，品詞，言語尤度以外にも，各音響モデルを用いた場合の音響尤度をフレーム単位で正規化し，その音響尤度の差を素性に追加しており，計 27 次元を用いている．なお，SVM の核関数には ANOVA 関数を用いている．

4.4.4 発話単位での最尤選択（baseline）

提案手法を評価するには，従来法との比較を行わなければならない．そこで，従来法として発話単位での最尤選択を行ったものを採用する[4]．それは，各音響モデルによる認識結果のうち，以下の式で示すように，発話単位全体での音響尤度と言語尤度の対数尤度の荷重和が大きい方を選択する．

$$\hat{m} = \arg\max_{m=R,S} \left\{ L_{acoustic}(m) + w \times L_{language}(m) \right\} \qquad \text{(式 3)}$$

ただし，m は R-system もしくは S-system を，$L_{acoustic}$ は音響尤度を，$L_{language}$ は言語尤度をそれぞれ表す．なお，言語尤度に対して w で重み付けを行っている．この値としては，評価データに対して最も認識率の高くなる値を用いるが，本実験では経験的にその値を 13 としている．

4.5 評価実験

本手法の有効性を検証するため，評価実験を行った．本実験では，機械学習および評価には共通のデータを用いているが，機械学習に対してオープンな評価実験を行うため，10 クロスバリデーションの実験を行っている．つまり，データを 10 分割し，そのうち，9 割で機械学習データを作成し，残り 1 割で評価実験を行うといったことを 10 通りについて行っている．

評価実験により得られた単語誤り率を図 9 に示す．図中の「Baseline selection」は，発話単位での音響尤度と言語尤度の対数尤度の荷重和が大きくなる認識結果を選択した結果である．「Upper bound」は，自動選択箇所すべてが正しい選択をした場合を示しており，どのような選択手法を用いても「Upper bound」より高い認識率を得ることはできない．逆に，「Lower bound」は，自動選択箇所すべてが誤った選択をした場合を示しており，どのような手法で選択しても「Lower bound」の認識率は確保できることになる．また，各手法について認識率改善の有効性を検証するために，一標本 t 検定を行っており，その結果を表 8 に示す．検定には，評価データを 10 セットに分け，その 10 通りの評価結果をサンプルとし，その各サンプルについて各手法と手法との単語誤り率の差について検定を行っている．な

お，表の値は，列に示された手法と，行に示された手法の単語誤り率の差を表している．

その結果，SVM-1，SVM-2，SVM-3，ベースラインともに，音響モデルを単独で用いた場合に比べ，有意な改善が得られている．また，SVM-1 および SVM-2 は，ベースラインと比較して改善はできなかった．SVM-3 については，ベースラインよりも有意に誤りを削減することができた．

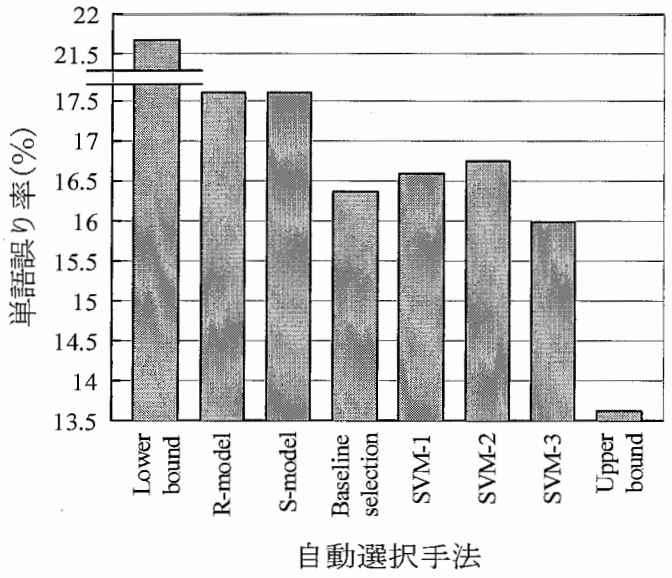

図 9 単語誤り率

表 8 単語誤り率の差の検定

	R-System	S-System	SVM-2	SVM-1	Baseline selection	SVM-3
R-Sytem	-	0.00	0.84 *	1.01 *	1.23 *	1.62 *
S-System	-	-	0.84 *	1.01 *	1.23 *	1.62 *
SVM-2	-	-	-	0.16	0.39	0.78 *
SVM-1	-	-	-	-	0.23	0.62 *
Baseline selection	-	-	-	-	-	0.39 *
SVM-3	-	-	-	-	-	-

*p<0.01

4.6 評価結果の検討

4.5 では，品詞，言語尤度を素性として機械学習することにより，認識率を改善できることが示された．そこで，第 3 章で得られた知見が，自動選択結果にどのような影響を与えているのかを検討する．

4.6.1 評価指標

評価結果の詳細を調べるため，4.5 で示した評価結果を，品詞，言語尤度の値に基づき集

計する．そして，改善の見込まれる自動選択箇所のうち，正しく選択された単語の割合により比較を行う．この割合のことを「正解単語選択率」と呼ぶことにし，次式のように定義する．

$$S_{correct} = \frac{C_{select} - C_{lower}}{C_{upper} - C_{lower}} \qquad \text{(式 4)}$$

ただし，$S_{correct}$ は正解単語選択率を表している．また，C_{select} は当該場合の単語正解率を，C_{lower} は自動選択箇所すべてを誤って選択した場合の単語正解率を，C_{upper} は自動選択箇所すべてを正しく選択した場合の単語正解率をそれぞれ表している．

なお，$S_{correct}$ は 0 から 1.0 までの値を取り，その値が 0.5 を超えた場合に認識率が改善されたことを意味する．

4.6.2 品詞が認識率向上に与える影響

図 10 に各品詞について集計した単語正解率を，図 11 に各品詞について集計した正解単語選択率を示す．図中の横軸は品詞を表しており，縦軸は単語正解率，正解単語選択率をそれぞれ表している．図中の棒の種類は各手法をそれぞれ表している．

図 11 から分かるように，品詞を素性とした自動選択である SVM-1，SVM-3 では，品詞により正解単語選択率の値が大きく異なっている．特に，自然発話音響モデルに適合しやすかった品詞である感動詞や助動詞，朗読発話音響モデルに適合しやすかった品詞である名詞に関しては，大きな改善が見られている．逆に，副詞，助詞などのどちらの音響モデルにも適合しなかった品詞に関しては，さほどの改善は見られない．また，間投詞に関しては，自然発話音響モデルに適合しやすく改善が見込まれていたが，ほとんど改善されていない．しかし，間投詞は冗長な表現であり，さほどの情報はないため，認識誤りをしてしまった場合でも，他の品詞に比べてリスクは少ないといえる．

それに対して，ベースラインである最尤選択は，品詞により多少のばらつきはあるが，平均的に改善されている．

本手法の SVM-3 とベースラインを比較すると，感動詞や助動詞，名詞といったどちらかの音響モデルに適合しやすかった品詞に関しては，SVM-3 のほうが大幅に改善されている．逆に，副詞，助詞などのどちらの音響モデルにも適合性しなかった品詞に関しては，ベースラインのほうが改善されている．

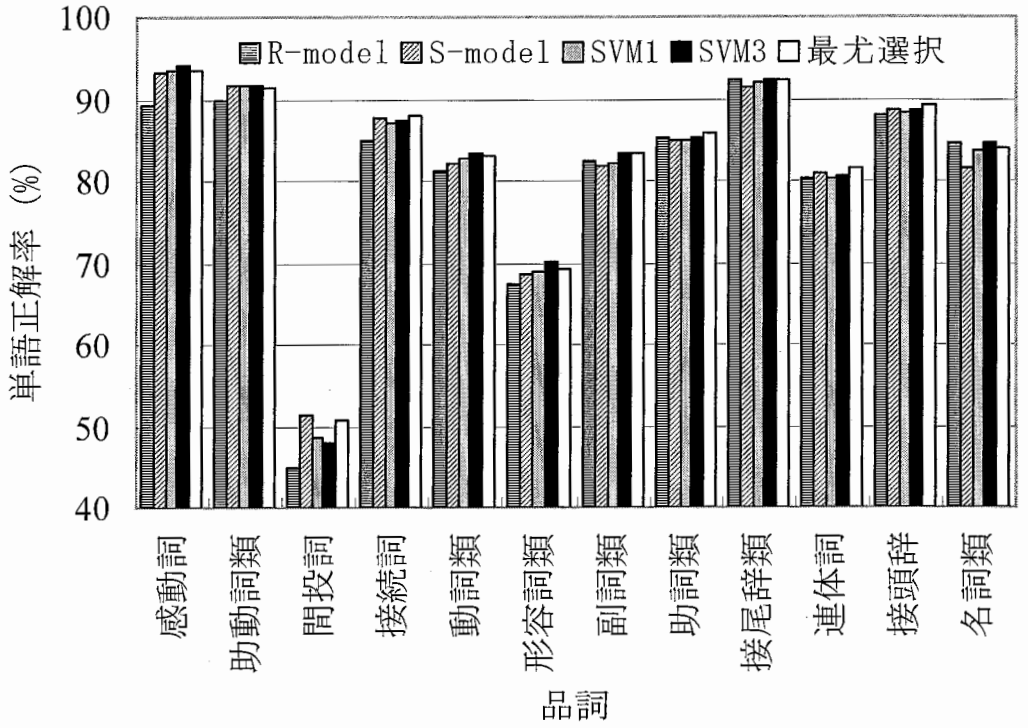

図 10 品詞ごとに集計した各手法の単語正解率

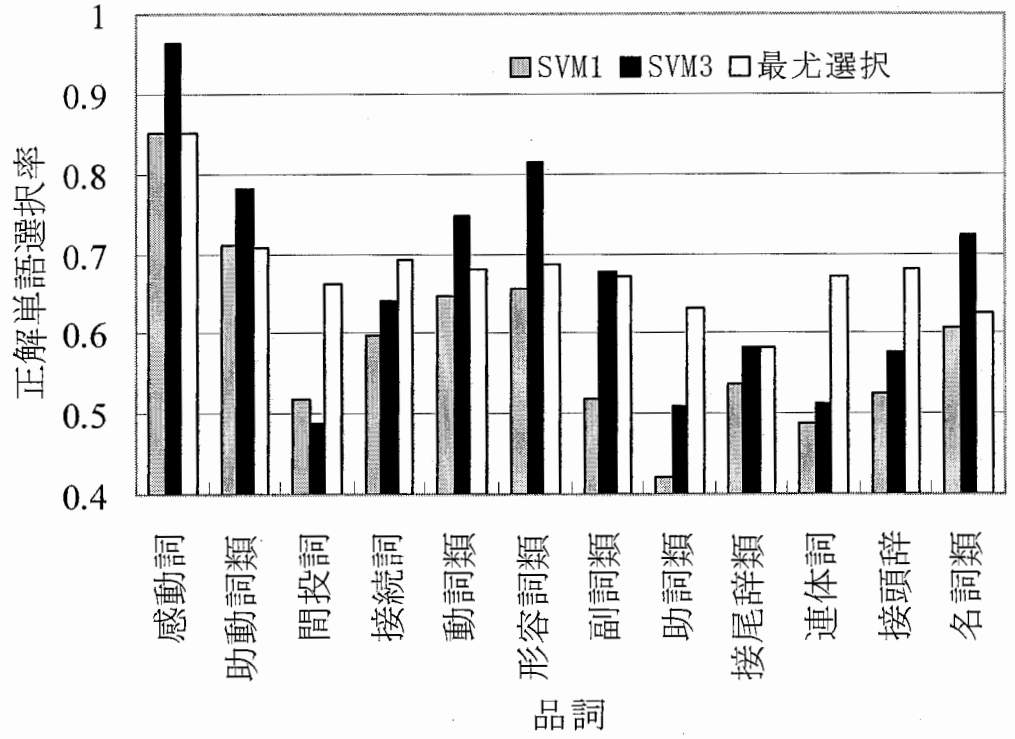

図 11 品詞と正解単語選択率の関係

4.6.3 言語尤度が認識率向上に与える影響

図 12 に言語尤度の値より集計した単語正解率を，図 13 に各品詞について集計した正解単語選択率を示す．ここでは，図 4 同様，評価データを言語尤度の値によりソートし，単語数が均一となるよう 16 グループに分割している．図中の横軸は，各グループ番号を表しており，その値が小さいほどそのグループに属する単語の言語尤度が低くなっている．縦軸は単語正解率，正解単語選択率をそれぞれ表している．図中の棒の種類は各手法をそれぞれ表している．

図 13 から分かるように，言語尤度を素性とした自動選択である SVM-2，SVM-3 では，言語尤度の値により正解単語選択率の値が大きく異なっている．特に，言語尤度の値が高い場合と低い場合に，大きな改善が見られる．それに対し，言語尤度の値が中間の場合にはさほどの改善は見られない．このことは，0 で示したように，言語尤度の値が低い場合は，朗読発話音響モデルに適合しやすく，高い場合は自然発話音響モデルに適合しやすいため，適切な音響モデルを用いた認識結果が選択されているためであると考えられる．

それに対して，ベースラインである最尤選択は，言語尤度が大きいほど，正解単語選択率の値は高く，認識精度の改善が見られる．これは，言語尤度が高ければ，正解である認識結果は誤りである認識結果よりも言語尤度が高くなりやすい．そのため，最尤選択では，言語尤度が高い場合に，正解である認識結果が選ばれやすく，認識率が改善されやすかったと考えられる．

本手法の SVM-3 とベースラインを比較すると，言語尤度が低い場合では SVM-3 のほうが大幅に改善されており，言語尤度が高い場合は最尤選択の方が認識率が良いものの，ともに大幅な改善が見られる．逆に，音響モデルの適合性に差がみられなかった言語尤度の中間の場合では，ベースラインのほうが改善されている．

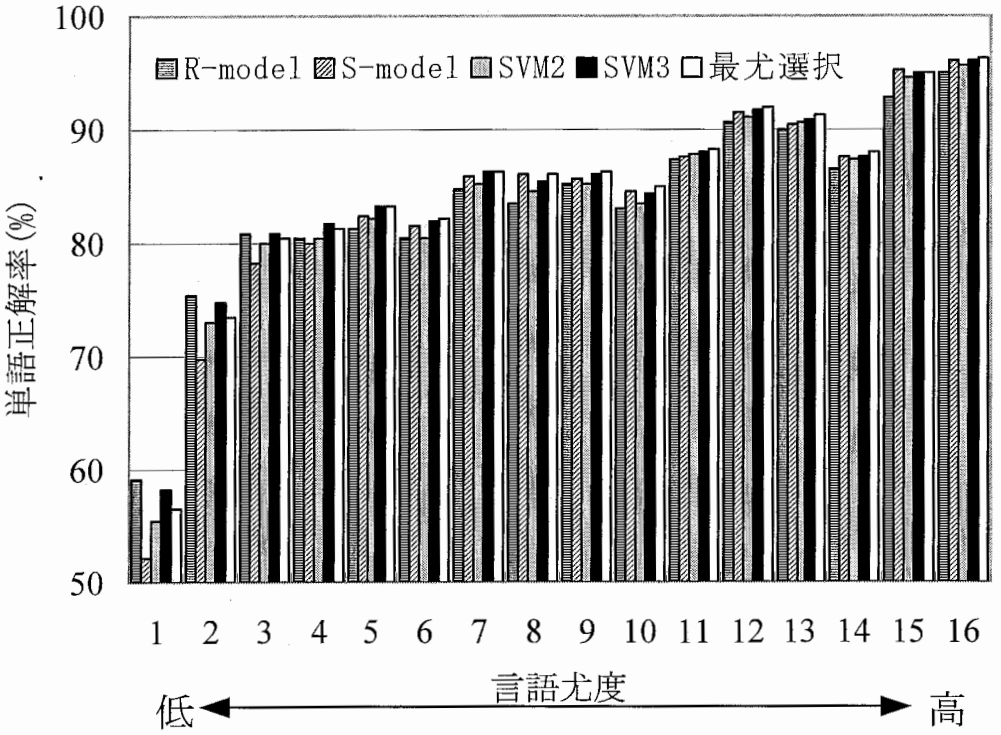

図 12 言語尤度の値により集計した各手法の単語正解率

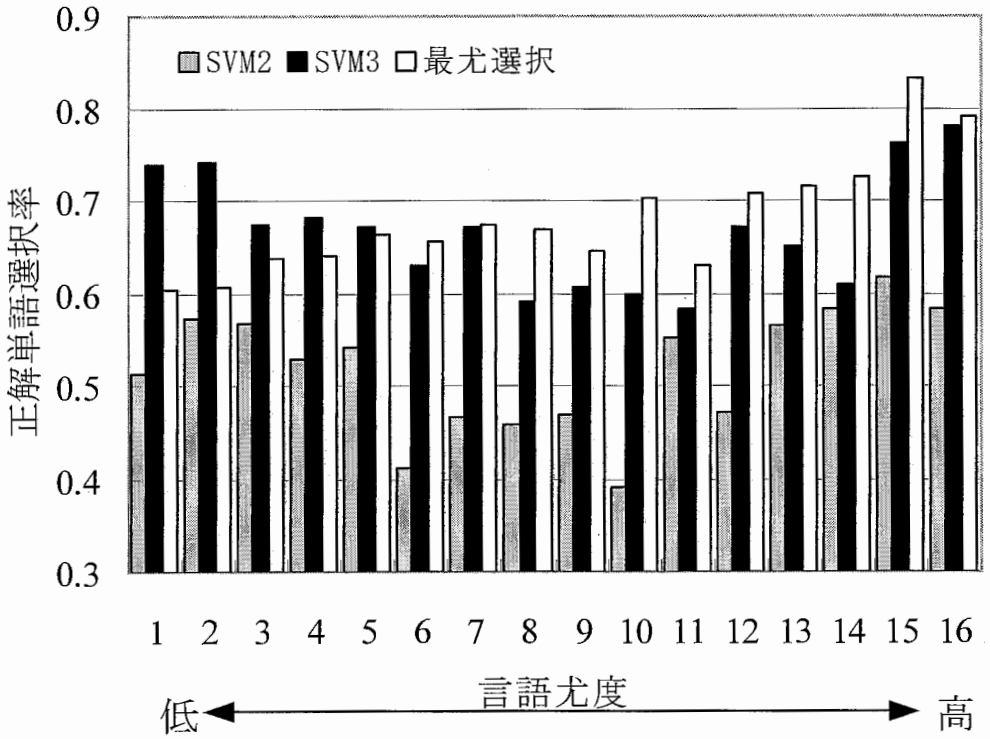

図 13 言語尤度と正解単語選択率の関係

4.7 結言

本章では，品詞・言語尤度を素性として，朗読発話音響モデルを用いた場合と自然発話音響モデルを用いた場合の各認識結果を単語単位で自動選択する方法を検討した．その結果，品詞を素性とした自動選択，言語尤度を素性とした自動選択では，最尤選択ほど認識率は上がらなかったが，音響モデルを単独で用いた場合よりも有意な改善が認められた．また，品詞，言語尤度，音響尤度を素性とした自動選択では，最尤選択よりも有意に認識率を改善することができ，音響モデルを単独で用いた場合よりも，単語誤りを 1.62 ポイント改善することができた．

第5章 結論

本論文では，朗読発話音響モデルと自然発話音響モデルの発話スタイルの異なる音響モデルを用いて，対話音声の発話スタイルを単語単位で分析を行った．具体的には，品詞，言語尤度といった単語の持つ情報と発話スタイルとの関係を調べた．その結果，言語尤度が低い場合や名詞の場合などでは，朗読発話音響モデルを用いるのが望ましく，言語尤度が高い場合や感動詞，助動詞の場合は自然発話音響モデルを用いるのが望ましいことが分かった．このように，品詞，言語尤度によって適切な音響モデルが異なることが示唆された．

次に，これらの知見から，品詞・言語尤度を素性として，各音響モデルを用いた場合の認識結果を単語単位で自動選択する方式を検討した．その結果，品詞を素性とした自動選択，言語尤度を素性とした自動選択では，ベースラインである最尤選択ほど精度は上がらなかったが，有意な改善が認められた．また，品詞，言語尤度，音響尤度を素性とした自動選択では，最尤選択よりも有意に認識率を改善することができ，音響モデルを単独で用いた場合よりも，単語誤りを 1.62 ポイント改善することができた．

今後の課題としては，朗読音声と対話音声を混合して音響モデルを学習し，朗読音響モデル，自然発話音響モデル，混合音響モデルの 3 系列を用いて分析および認識率の改善を試みたい．

謝辞

研究を行う機会を与えてくださいました，音声言語コミュニケーション研究所　山本誠一所長，菊井玄一郎第二研究室室長，同志社大学工学部　柳田益造教授に感謝いたします．また，熱心に議論に応じてくださいました第二研究室の皆様に感謝いたします．

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

参考文献

[1] 白井克彦："音声対話コーパスと対話課程のモデル"，人工知能学会誌，Vol.12，No.1，pp.30-35 (1997).

[2] 村上仁一，嵯峨山茂樹："自由発話音声における音響的な特徴の検討"，信学論，Vol.J78-D-II，No.12，pp.1741-1749，(1995).

[3] 菅谷史昭，竹澤寿幸，隅田英一郎，匂坂芳典，山本誠一："音声翻訳システム：ATR-MATRIX の開発と評価"，情処学論，Vol.43，No.7，pp.2230-2241 (2002).

[4] Toshiyuki Takezawa，Sugaya Fumiaki，Masaki Naito and Seiichi Yamamoto："A Comparative Study on Acoustic and Linguistic Characteristics Using Speech from Human-to-human and human-to-machine conversations"，ICSLP2000，Vol.3，pp.522-525 (2000).

[5] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄："音声認識システム"，オーム社 (2001).

[6] 板橋秀一，山本幹雄，竹沢寿幸，小林哲則："日本音響学会新聞記事読み上げ音声コーパスの構築"，音講論，2-Q-36，Vol.I，1997-9.

[7] Atsushi Nakamura，Shoichi Matsunaga，Tohru Shimizu，Masahiro Tonomura and Yoshinori Sagisaka："Japanese speech databases for robust speech recognition"，Proceedings of International Conference on Spoken Language Processing，pp.2199-2202 (2000).

[8] 板橋秀一，山本幹雄，河原達也："重点領域研究「音声対話」における音声コーパス"，人工知能学会研究会資料 SIG-SLUD-9701-5，pp.25-30 (1997).

[9] Toshiyuki Takezawa，Tsuyoshi Morimoto and Yoshinori Sagisaka："Speech and Language Databases for Speech Translation Research in ATR"，Proceeding of Oriental COCOSDA Workshop，pp.148-155 (1998).

[10] 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"，信学論，Vol.J83-D II，No.22，pp.2146-2151 (2000).

[11] V.N.Vapnik："Statistical Learning Theory"，Wiley-Interscience (1998).