ＴＲ－ＳＬＴ－００３７

機械学習を用いた発話スタイル依存音響モデル自動選択の検討

Automatic Selection of Speaking-Style-Dependent Acoustic Models Using Machine Learning

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

2003 年 3 月 28 日

概要

本研究は，朗読発話と自然発話から学習された 2 つの音響モデルを，単語単位で選択的に利用することにより音声認識性能を向上させることを目的としている．その予備実験として，言語尤度，品詞等の言語情報に基づき，2 つの音響モデルの分析・比較を行った．その結果，音響尤度は言語情報に依存し，適切な音響モデルが異なることが示唆された．そこで，それらの知見を用いて機械学習することにより，使用する音響モデルの自動選択を試みている．

（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

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

1．はじめに

音声翻訳システムや音声対話システムは，会話調の音声を処理する必要がある．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[1]．また，発話スタイルの異なる複数の音響モデルを用いて，発話単位で最尤となる結果を自動選択する実験によれば，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[2]．そこで本研究では，発話スタイルの異なる朗読発話により学習された朗読発話音響モデルと自然発話により学習された自然発話音響モデルの二つの音響モデルを，発話より小さい単位である単語単位で自動選択することにより，音声認識性能の向上を試みる．

まず，そのための予備実験として，単語単位で二つの音響モデルを比較することにより，言語尤度と発話スタイル，品詞と発話スタイルの関係について調査を行った．次に，それらの知見と機械学習を用いることにより，単語単位での自動選択を試みている．

2．品詞と発話スタイル，言語尤度と発話スタイルの関係

本研究では，単語単位による音響モデルの自動選択を目的としていることから，単語の持つ情報である，品詞や言語尤度と，発話スタイルとの関係を調べている．

2．1 使用した資料

2．1．1 音響モデル

発話スタイルとしては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．自然発話としては旅行会話を模擬した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読み上げ音声を用いた．音響分析の条件を Table 2.1 に，学習に用いた音声データの概要を Table 2.2 に示す．

2．1．2 分析用音声データ

分析に用いた音声データは，日本人同士の対話音声，男性 17 名，女性 25 名，のべ 551 発話と，通訳を介した日本語－英語の対話音声（日本語側のみ）男性 8 名，女性 15 名，のべ 330 発話である．なお，日本人同士の対話音声を直接対話データと呼び，通訳を介した対話音声を間接対話データと呼ぶことにする．また，間接対話データは，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[2].

2．2 発話スタイルの比較方法

一般に，音声認識において，正解系列の音響尤度が高くなることが好ましいことから，ここでは，音響尤度の大小比較を行っている．具体的には，朗読発話と自然発話の各音響モデルを用い，分析用データについての単語単位の音響尤度を求め，比較･分析を行っている．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大きくなる単語の割合により比較する．この割合を自然発話音響モデル優位率と呼ぶことにする．この値が高ければその発話スタイルは自然発話に近く，低ければ朗読発話に近いことを意味する．

2．3 品詞と発話スタイルの関係

Fig.2.1 は音響尤度の比較結果を，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の優位率を表している．図中の白い棒グラフは直接対話データでの結果を，黒い棒グラフは間接対話データでの結果を表している．

Fig.2.1 を見ると，ほぼ全ての品詞について，直接対話データの自然発話音響モデル優位率が高くなっているが，特に内容に関する重要な情報を伝達する形容詞類，名詞類や，名詞類に伴う接尾辞，接頭辞では，直接対話データと間接対話データに顕著な差があること

が分かる．直接対話データに対し，自然発話音響モデルが優位になる傾向は，発話単位での分析においても生じることが報告されている[2]．また，品詞間の比較では，品詞によって朗読発話に近いものと自然発話に近いものとに大きく分かれる．自然発話特有の品詞である感動詞や，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くなり，逆に，内容に関する重要な情報を伝達する名詞類では低くなっている．この傾向は，直接対話データと間接対話データに共通して生じているが，特に間接対話データの場合に顕著に生じている．

2．3 名詞類に関しての自然発話音響モデル優位率

Fig.2.2 は Fig.2.1 に示した名詞類を，より詳細に分類し，集計した結果である．図中の縦軸は，Fig.2.1 同様，自然発話音響モデル優位率を表している．

Fig.2.2 を見ると，固有名詞，数詞では朗読発話音響モデルが優位であるのに対し，代名詞，サ変名詞では自然発話が優位となっている．この原因として，固有名詞，数詞は対話中で，電話番号，名前，日時といった重要な情報を表現していることが多く，聞き手にはっきりと聞き取れるよう明瞭に発話されており，朗読発話に近い発話スタイルになっていると考えられる．この傾向は間接対話データの場合に顕著に表れており，特に数詞は，直接対話データと間接対話データとの間に大きな差が見られる．一方，代名詞，サ変名詞については，聞き手にとって聞き取りづらい状況であっても，対話進行への影響は小さいため，あまり明瞭に発話されていないと考えられる．また，直接対話データと間接対話データとの差は固有名詞，数詞ほど見られない．

自然発話においては，発話速度の分散が大きいとの報告[3][4]もあることから，今後発話速度と，ここで得られた知見との関係についても明らかにしたい．

2．4 言語尤度と発話スタイルの関係

Fig.2.3 は自然発話音響モデル優位率を，言語尤度の値を用いて集計した結果である．ここでは，分析用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．Fig.2.3 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内の単語の言語尤度が低くなっている．縦軸は自然発話用音響モデルの優位率を表している．なお，言語尤度にはマルチクラス複合バイグラム[5]を用いている．

Fig.2.3 を見ると，言語尤度が低いほど，自然発話用音響モデル優位率が低く，また，言語尤度が高いほど，自然発話用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，明瞭に発話され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．

3．トライフォンのカバレージ

2．3では，品詞に依存して適切な音響モデルが異なり，特に名詞類に関しては，朗読発話音響モデルに適合しやすく，逆に感動詞に関しては，自然発話音響モデルに適合しやすいことが分かった．このことは，発話スタイルの違いによるものであると考えているが，それ以外にも，分析データと自然発話音響モデルの学習データがともに旅行会話であり，タスクが一致するといったことから，トライフォンのカバレージによる影響の可能性も考えられる． そこで，発話スタイルが朗読発話に最も近かった名詞類，自然発話に最も近かった感動詞，それ以外の品詞の3種類に関して各音響モデルの作成に用いた学習データのカバレージを調べた．その結果を Fig.3.1 に示す．ここでは，トライフォンのカバレージを，分析データで出現するトライフォンが横軸の回数以上学習されている割合により表している．なお，縦軸はトライフォンのカバレージを，横軸はその閾値を表している．また，2つの音響モデル間には学習データ数の割合が約 4.6 倍であることから，比較が行えるよう，横軸の縮尺を変えている．

この図から，両モデルとも感動詞とその他の品詞の間には，さほどの違いが見られず，名詞類に関しては，トライフォンのカバレージが他の品詞よりも低いことが分かる．つまり，朗読発話音響モデルと自然発話音響モデルとのカバレージには，学習データ数に大きな差があるが，データ数の割合を合わせた場合に，品詞間の関係は両モデルの間でさほどの違いはないことが分かる．このことから，トライフォンのカバレージが品詞間に与える影響は少ないと考えられ，2．3で，品詞により，適合する音響モデルが異なった理由としては，発話スタイルによる違いが大きく関わっていると考えられる．

4．認識実験

2．で示した知見を用いて，朗読発話音響モデルと自然発話音響モデルの各音響モデルを用いた場合の各認識結果から，正しいと思われる認識結果を単語単位で自動選択することにより，認識精度の向上を試みた．なお，自動選択には，Support Vector Machine[6]による機械学習を用いている．また，機械学習には，2．1．2で示した直接対話音声データを用いて，クローズドの実験を行っている．

4．1 学習方法

2．で示したように適切な音響モデルは，言語尤度，品詞が関係していることから，機械学習には以下の27次元のパラメータを用いた．

- 各音響モデルを用いた場合の音響尤度の大小関係 (1次元)
- 各音響モデルを用いた場合の言語尤度 (2次元)
- 各音響モデルを用いた場合に出現する品詞 (24次元)

今回の実験では，単語単位による比較を目的としているが，音響尤度，言語尤度等は，時間的区間が同一でなければ，比較を行うことができないため，対応関係の取れた複数の単語同士を比較して学習を行う．具体的には，Fig.4.1 に示すように，正解系列と朗読発話音響モデル，自然発話音響モデルの各音響モデルを用いた場合の認識結果の 3 系列について比較を行う．図から分かるように，比較対象の単語が 1 対 1 で対応しているとは限らないので，対応の取れた箇所について，正解系列とのＤＰ距離が最小となる結果が自動選択されるよう機械学習を行った．また，その学習データ数は185箇所であった．

4．2 自動選択による認識性能

認識実験は，朗読発話音響モデル，自然発話音響モデルをそれぞれ単独で用いた場合と，機械学習により，自動選択した場合，ベースラインとして音響尤度と言語尤度の値から最尤選択を行った場合の4通りを示す．なお，使用した学習データは，自然発話を用いており，総単語数は4990単語であった．

次に，実験の結果得られた単語誤り率を Table 4.1 に示す． このように，単語単位で音響モデルを自動選択した場合の単語誤り率は，朗読発話音響モデルを単独で使用した場合よりも約1.7ポイント，自然発話音響モデルを単独で使用した場合より約0.7ポイントと，ともに改善が見られる．また，機械学習による自動選択で改善の見られたのは，66.5%（185箇所中 123 箇所）であった．このように，品詞，言語尤度の情報をパラメータとして用いて，単語単位で自動選択することにより，認識性能向上が期待できる．

5．まとめ

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，分析用データについて単語単位の音響尤度を求め，言語情報を用いた比較・分析を行った．その結果，品詞に依存して適切な音響モデルが異なることが示唆された．名詞類について詳細に調べたところ，固有名詞，数詞といった会話中の重要な情報の表現に用いられやすい品詞ほど明瞭に発話されており，その発話スタイルは朗読発話に近いことが示唆された．また，この傾向は特に間接発話データの場合に顕著に表れて見られる．また，音響尤度と言語尤度の関係を調べたところ，言語尤度が高いほど，自然発話用音響モデル優位率が高くなる傾向が見られた．このように，音響尤度は言語情報に依存し，適切な音響モデルが異なることが示唆された．

次に，これらの知見と機械学習を用いることで，音声認識性能の向上を試みた．その結果，単語誤り率は，単語単位で音響モデルを自動選択した場合では単独で音響モデルを使用した場合よりも，改善されることが分かった．

今回は，機械学習に用いる学習データ数が少なかったため，機械学習に対しては，クローズドな実験となっているが，今後は，機械学習用のデータを増やしたオープンでの有効性を検証したい．

謝辞

本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，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] 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.

[2] Toshiyuki Takezawa, Fumiaki Sugaya, Masaki Naito, 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.

[3] 山本一公，中川聖一：“発話スタイルによる話速･音韻間距離･尤度の違いと音声認識性能の関係”，信学論，Vol.J83-DⅡ，No11，pp.2438-2447，2000.

[4] 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-DⅡ，No12，pp.1741-1749，1995.

[5] 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-DⅡ，No.22，pp.2146-2151，2000.

[6] http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

Table 2.1 音響分析の条件

サンプリング周波数	16[ksamples/sec]
分析窓長	20[msec]（Hamming 窓）
シフト間隔	10[msec]
特徴パラメータ	MFCC(12 次元)+ΔMFCC(12 次元) +Δパワー

Table 2.2 学習用音声データの概要

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
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

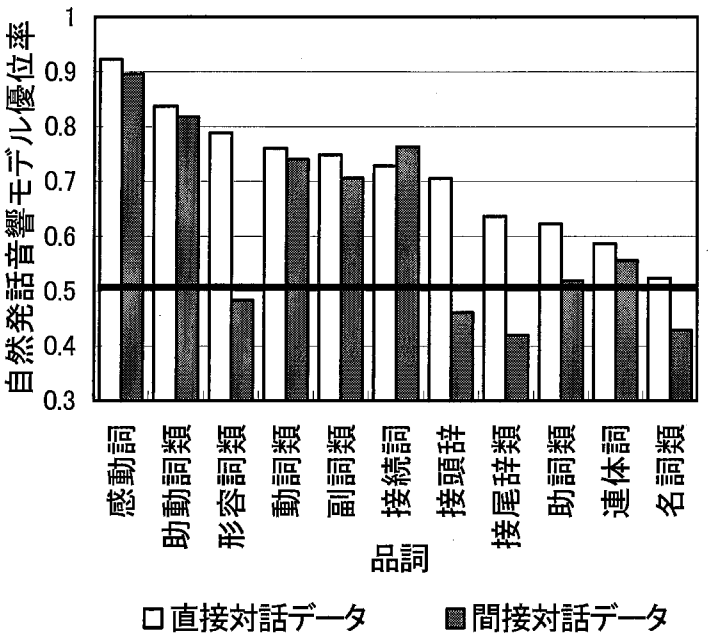


Fig.2.1 品詞と発話スタイルの関係

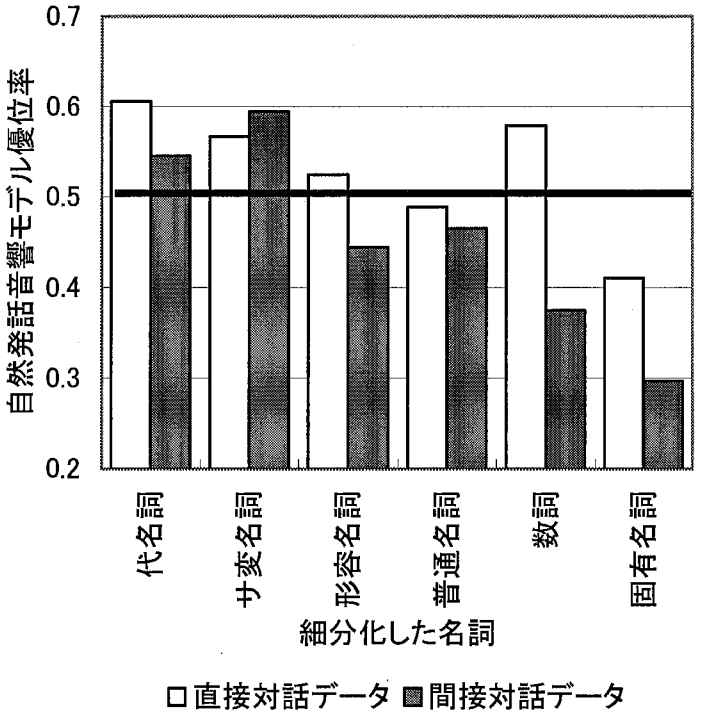


Fig.2.2 名詞類に対する自然発話音響モデル優位率

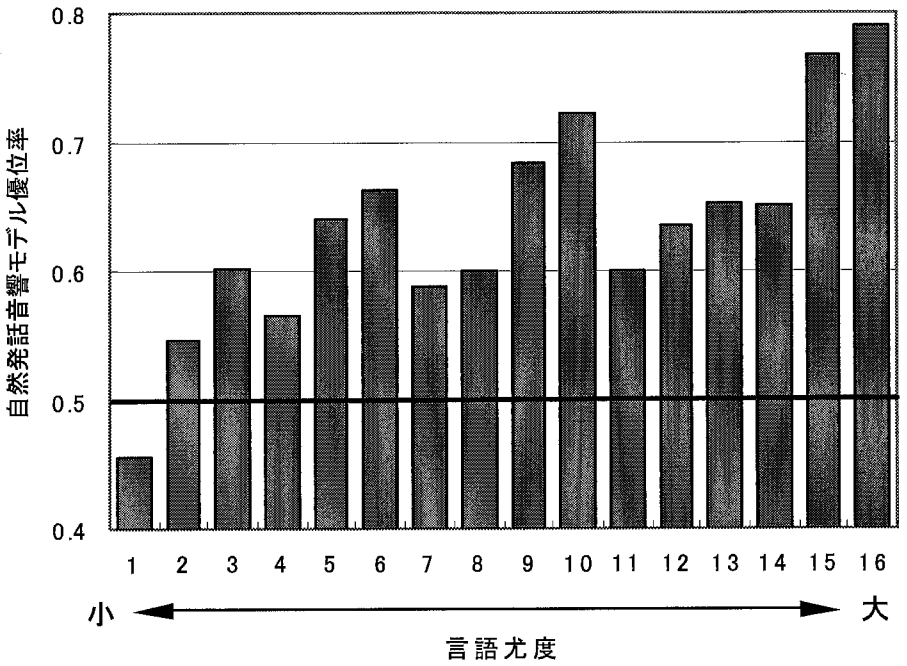
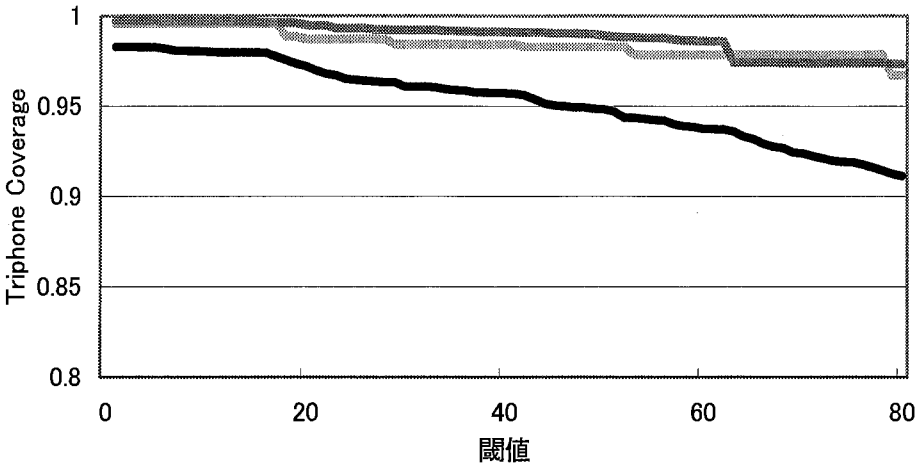
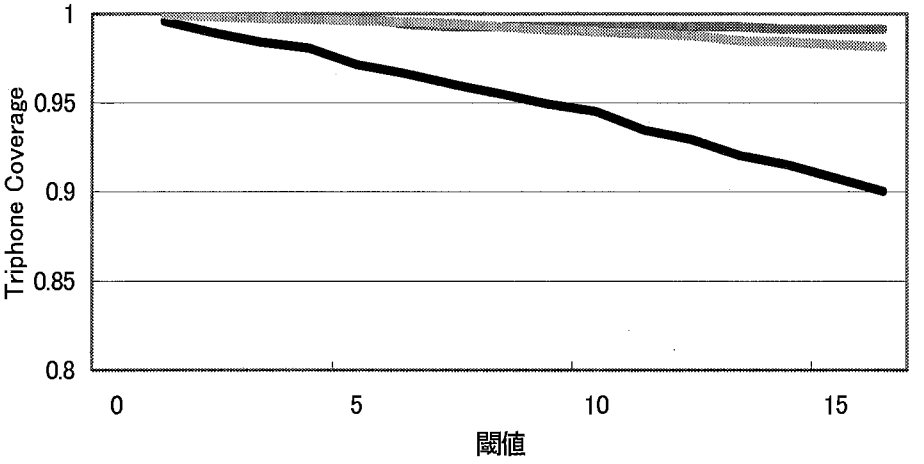


Fig.2.3 自然音響モデル優位率と言語モデルの関係



(a)朗読発話音響モデル



(b)自然発話音響モデル

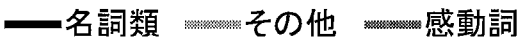


Fig.3.1 トライフォンのカバレージ

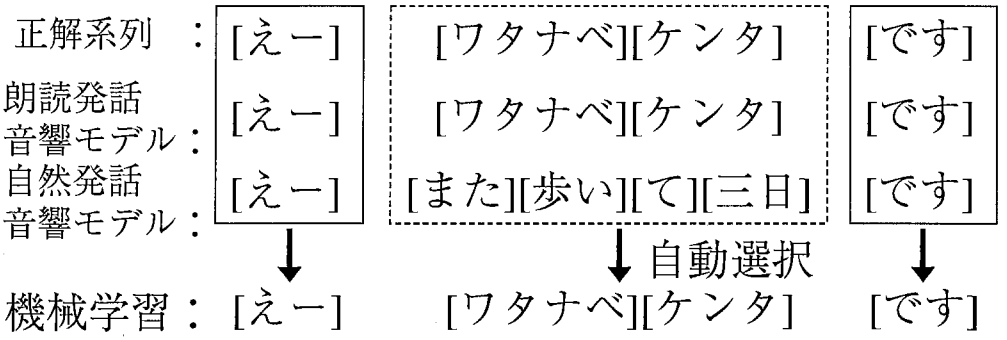


Fig.4.1 自動選択の例

Table 4.1 単語誤り率

	誤り単語数	誤り単語率（%）
朗読発話音響モデルのみ	693	13.9
自然発話音響モデルのみ	646	12.9
SVM による自動選択	608	12.2
最尤選択法	631	12.6
（上限）	（533）	（10.3）

（総単語数 4990 単語）