2003－SLP－47　(10)
2003／7／19

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による対話音声認識

青野邦生[†,‡]，安田圭志[†,‡]，竹澤寿幸[†]，山本誠一[†]，柳田益造[‡]

[†]ATR 音声言語コミュニケーション研究所　[‡]同志社大学

対話音声認識の精度向上を目的とした，発話スタイル依存音響モデルの自動選択手法を提案する．提案手法では，まず，発話スタイルごとに作成された音響モデルを用いて，複数の認識結果を作成する．次に，これらの認識結果の中から，言語的な情報をもとに，信頼性の高い認識結果を単語単位で自動的に選択する．本論文では，まず，自然発話音響モデルと朗読発話音響モデルを用いた予備実験により，言語尤度，品詞といった言語的な情報と，各音響モデルの認識性能との関係について示す．次に，これらの知見を用いた認識実験の結果が示されている．認識結果を自動選択するためのモジュールは SVM を用いて作成した．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，1.06 ポイント低減できている．

Dialogue Speech Recognition
Based on Speaking-Style-Dependent Acoustic Model Selection

Kunio Aono[†,‡], Keiji Yasuda[†,‡], Toshiyuki Takezawa[†], Seiichi Yamamoto[†], Masuzo Yanagida[‡]

[†]ATR Spoken Language Translation Research Labs.　[‡]Doshisha University

A speaking-style-dependent acoustic model selection scheme is proposed, aiming at improving dialogue speech recognition performance. The proposed scheme generates multiple recognition results by adopting various speaking-style-dependent acoustic models. The most reliable recognition result is then selected based on language information. In this paper, pilot experiments using a spontaneous speech acoustic model and a read speech acoustic model show the relationship between recognition reliability and the word information, including word category and language likelihood. Recognition experiments were carried out using the knowledge gained from the pilot experiments. In this set of experiments a SVM-based selection module is used. The proposed method reduced word error rate by 1.06 point over the baseline system.

１．はじめに

新聞記事の読み上げのような音声については，認識精度が向上してきたが，人間同士の日常的な会話のような対話音声については，まだ十分な認識精度に至っていない．音声翻訳システムや音声対話システムは，そのような会話調の音声を処理する必要があるが，その対象は機械を意識した対話音声であるので，その特性を考慮した性能向上を検討する．例えば，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]．対話音声の特性を分析した先行研究に[3,4]があるが, 音声認識の性能向上に関する実践はしていない．そこで，本研究では，まさに対話音声を対象とし，発話より小さい単位で，言語尤度や品詞といった情報と発話スタイルの関係を調べ, 得られた知見に基づき，音声認識性能が上げられることを認識実験により検証する．

2．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読み上げ音声を用いた．音声の分析条件を表 1 に，学習に用いた音声データの概要を表 2 に示す. 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, 自然発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 1 音声の分析条件

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

表 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

3．品詞，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．そのため, 単語の持つ情報として，品詞および言語尤度について，発話スタイルとの関係を調べた[5].

3．1　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．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，音響尤度の大小比較を行う．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比較を行う. 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．

$$WR_s = N_{s-model} \Big/ N_{total} \qquad (1)$$

ただし, N_{total} は全単語数を, $N_{S\text{-}model}$ 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.

3．2　分析データ

分析に用いた音声データは, 録音スタジオでの通訳者を介した日本語－英語の対話音声（日本語側のみ），男性 8 名，女性 15 名，延べ 330 発話である．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[2].

3．3　言語モデル

本実験で使用する言語モデルの学習データの概要を表 3 に示す．(a)は録音スタジオで収録した日本人同士の対話音声データを切り出しと書き起こしを行ったものであり，SDB/TRA と呼ぶ[6]. (b)は録音スタジオでの通訳者を介した対話音声データの切り出しと書き起こしを行ったものであり, SLDB と呼ぶ[7]. (c)はオフィスルームでの通訳者を介した対話音声データを書き起こしたものであり，LDB と呼ぶ[7]．なお，言語モデルには，多重クラ

ス複合バイグラム[7]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) 25,548 発話，約 39 万単語
日本語－英語の対話（日本語側のみ）
(b) 15,777 発話，約 21 万単語
(c) 85,765 発話，約 72 万単語

３．４　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．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く，つまり，その発話スタイルは自然発話に近いのに対し，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自然発話に近いことが分かる．

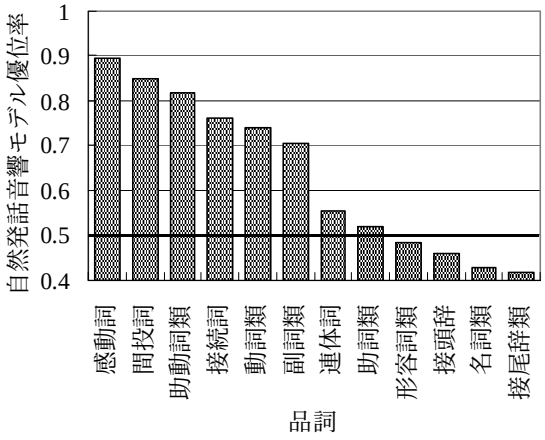

図 1 各品詞に対する自然発話音響モデル優位率

３．５　名詞類についての詳細

図 2 は図 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，図 1 同様，自然発話音響モデル優位率を表している．

図 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．一方，サ変名詞，代名詞については，聞き手にとって聞き取りづらい状況であっても，冗長性が大きく，聞き落としても補完しやすいため，あまり明瞭に発話されていないと考えられる．

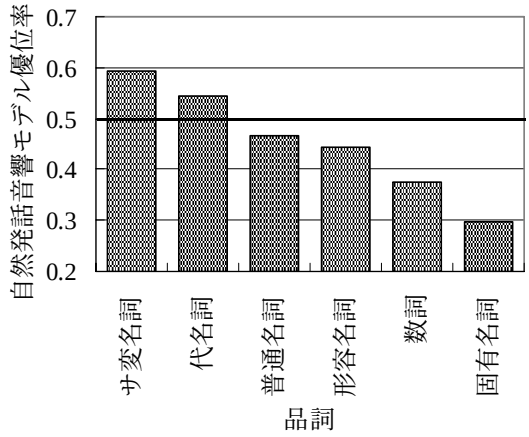

図 2 名詞類に対する自然発話音響モデル優位率

３．６　言語尤度と発話スタイルの関係

図 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，

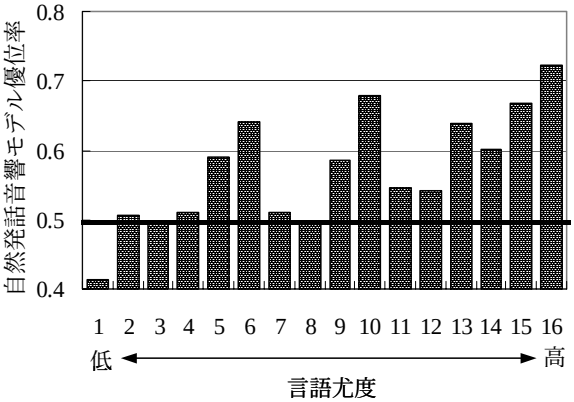

図 3 言語尤度に対する自然発話音響モデル優位率

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．図 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優位率を表している．

図 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．

４．認識実験

３．では，品詞，言語尤度の値により，適切な音響モデルが異なることが分かった．そこで，それらの知見から，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認識結果から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[9]による機械学習を用いており，使用する言語モデルは，学習データに３．３の表 3 の(a)，(c)を用いている．

４．１　自動選択手法

本実験では，単語単位による自動選択を目的としているが，時間的区間が同一でなければ，比較を行うことはできず，また，必ずしも 1 対 1 で比較ができるとも限らない．そこで，時間的に対応の取れた複数の単語同士を比較し，自動選択する必要がある．具体例を図 4 に示す．このように，朗読発話音響モデルを用いた場合の認識結果である「ワタナベケンタ」と自然発話音響モデルを用いた場合の認識結果である「また歩いて三日」を比較し，自動選択しなければならない．

なお，本手法では機械学習用データ作成と自動選択の 2 段階により構成されている．

まず，機械学習用データ作成の流れについて説明する．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 系列について比較することにより作成する．具体的には，図 5 に示すように，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と正解系列の 3 系列につ

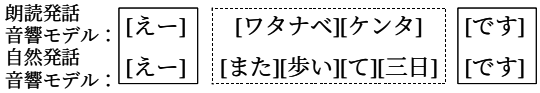

図 4 対応の取り方

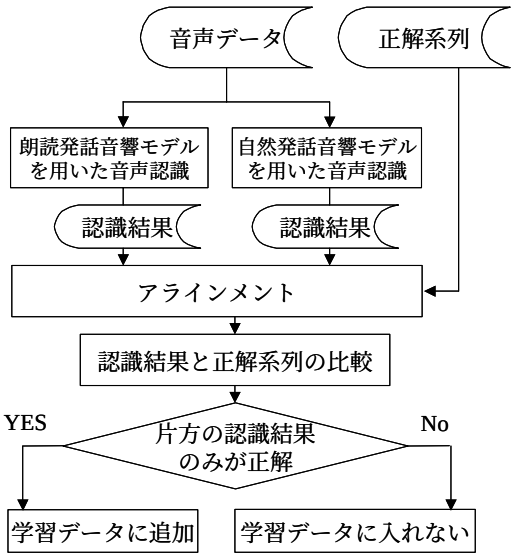

図 5 学習データ作成の流れ

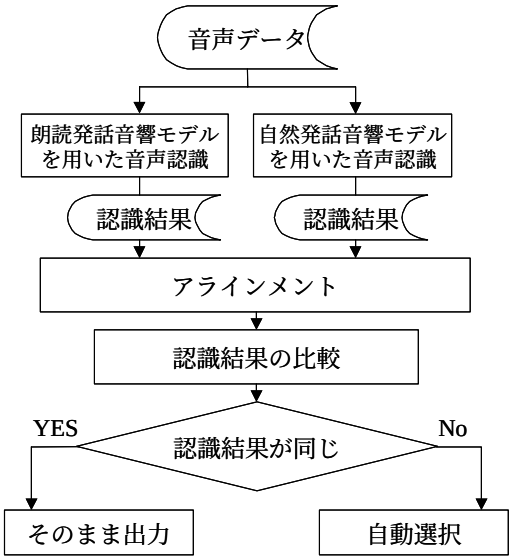

図 6 自動選択の流れ

いて, DP を用いることでアライメントをとる. そして, 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追加する. なお, ともに誤っている場合や, 正解する場合は, 学習データには追加しない.

次に, 自動選択方法の流れについて説明する. 具体的には, 図 6 に示すように, 各音響モデルを用いた場合の認識結果について, DP を用いることで 2 系列のアライメントをとる. そして, 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, 自動選択する.

４．２　実験条件

４．２．１　機械学習データおよび評価データ

本実験の機械学習および評価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), 延べ 15,788 発話である. なお, 機械学習に対してオープンの実験を行うため, 上記のデータを 10 分割し, 9 割で機械学習, 残り 1 割で評価実験を行うといった, 10 クロスバリデーションの実験を行っている.

４．２．２　自動選択の種類

・品詞を用いた自動選択

３．４では, 品詞により適切な音響モデルが異なることが分かった. そこで, 品詞情報をパラメータとして自動選択することにより, 音声認識精度の改善を試みた. また, 自動選択を行う際の機械学習には SVM を用い, そのパラメータには各音響モデルを用いた場合に出現する品詞の出現単語数, 計 24 次元（12 品詞×2 音響モデル）を用いている. なお, SVM のカーネル関数には Gaussian 関数を用いている.

・言語尤度を用いた自動選択

３．６では, 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た. そこで, 言語尤度をパラメータとして自動選択することにより, 音声認識精度の改善を試みた. また, 自動選択を行う際の機械学習には SVM を用い, そのパラメータには各音響モデルを用いた場合の言語尤度を一単語当たりの言語尤度に正規化した値, 計 2 次元を用いている. なお, SVM のカーネル関数には線形関数を用いている.

・最尤選択

本実験では, 品詞, 言語尤度の情報を用いた自動選択以外にも, 言語尤度と音響尤度の積がより高くなる結果を選択する, 最尤選択による実験を行っている. なお, 最尤選択による実験では, 品詞, 言語尤度を情報とした自動選択の効果を示すためのベースラインとして用いている.

４．３　実験結果

各手法による認識実験の結果得られた単語誤り率を表 4 に示す. なお, 表の「多数決」は, 言語尤度を情報とした自動選択, 品詞を情報とした自動選択, 最尤選択の 3 通りの結果を多数決することによる自動選択を表している. また, 表の「上限」は, 自動選択箇所すべてについて, 改善されるよう選択した場合を示しており, どのような選択手法を用いても「上限」より高い精度を得ることはできない. 逆に, 表の「下限」は, 自動選択箇所すべてが最悪の選択をした場

表 4 認識率結果

	誤り単語数	単語誤り率(%)
(a) 朗読発話音響モデル単独	35893	17.52
(b) 自然発話音響モデル単独	36041	17.59
(c) 品詞情報による自動選択（SVM）	34383	16.78
(d) 言語尤度による自動選択（SVM）	35104	17.13
(e) 最尤選択	34433	16.80
(f) c，d，e の多数決	33719	16.46
(g) 上限	28293	13.81
(h) 下限	43787	21.37

合を示しており, どのような手法で選択しても「下限」の精度は確保できることになる. 表に示すように, 品詞を情報として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用いた場合よりも約 0.74 ポイント, 自然発話音響モデルを単独で用いた場合よりも 0.81 ポイントの改善が見られた. また, 言語尤度を情報として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用いた場合よりも約 0.39 ポイント, 自然発話音響モデルを単独で用いた場合よりも約 0.46 ポイントの改善が見られた. そして, 最尤選択では, 朗読発話音響モデルを単独で用いた場合よりも 0.72 ポイント, 自然発話音響モデルを単独で用いた場合よりも 0.79 ポイントの改善が見られた. このように, 品詞, 言語尤度を情報として自動選択することの有効性が認められた. また, 品詞を情報とした自動選択は, 最尤選択と同等以上の効果が見られたが, 言語尤度を情報とした自動選択は, 最尤選択ほどの効果は見られなかった. しかし, 言語尤度を情報とした自動選択, 品詞を情報とした自動選択, 最尤選択の 3 通りの選択結果の多数決では, 最尤選択よりも, より高い改善が得られており, 朗読発話音響モデルを単独で使用した場合よりも約 1.06 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も約 1.13 ポイントの改善が見られた.

5. 終わりに

本論文では, 発話スタイル依存音響モデルの自動選択手法を提案した. 提案手法では, まず, 発話スタイルごとに作成された音響モデルを用いて, 複数の認識結果を作成する. 次に, これらの認識結果の中から, 品詞, 言語尤度の情報をもとに, 信頼性の高い認識結果を単語単位で自動的に選択する. その予備実験として, 朗読発話音響モデルと自然発話音響モデルの 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, 品詞・言語尤度を用いた比較・分析を行った. その結果, 品詞・言語尤度の値によって, 適切な音響モデルが異なることを示した.

次に, これらの知見から, 品詞・言語尤度を情報として, 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した. その結果, 品詞を情報とした自動選択では, 最尤選択と同等の精度を改善することができ, 言語尤度に関しては, 最尤選択ほど精度は上がらなかったが, 優位な改善が認められた. また, 品詞を情報とした自動選択, 言語尤度を情報とした自動選択, 最尤選択の 3 種類の自動選択の結果を用いた多数決により, 単独の音響モデルを用いた場合よりも, 単語誤りを 1.06 ポイント改善することができた.

謝辞：本研究は通信・放送機構の研究委託により実施したものである. また, 本研究の一部は, 同志社大学学術フロンティア事業の援助を受けた.

参考文献

[1] 菅谷史昭他：“ 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, et al. :“ 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] 青野邦生他 :“ 言語情報を考慮した発話スタイル依存音響モデル自動選択の予備検討 ”, 音講論, pp.89-90, 2002-09.

[6] 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.

[7] Toshiyuki Takezawa, et al. :“ Speech and Language Databases for Speech Translation Research in ATR ”, Proceeding of Oriental COCOSDA Workshop , pp.148-155, 1998.

[8] 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.

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