2003－SLP－47　(9)
2003／7／19

複数モデルを選択的に用いる音声対話システムにおける ドメイン切り替え尺度の検討

磯部 俊洋 [1,5]　早川 昭二 [2,5]　村尾 浩也 [3,5]
水谷 龍治 [4]　武田 一哉 [4,5]　板倉 文忠 [4,5]
1 ㈱NTT データ　〒104-0033 東京都中央区新川 1-21-2
2 ㈱富士通研究所　〒674-8555 明石市大久保町西脇 64
3 三洋電機㈱　〒573-8534 大阪府枚方市走谷 1-18-13
4 名古屋大学大学院工学研究科　〒464-8603 名古屋市千種区不老町 1
5 名古屋大学統合音響情報研究拠点(CIAIR) 〒464-8603 名古屋市千種区不老町 1
E-mail:　isobet@nttdata.co.jp

あらまし　本稿では、ドメインに依存した言語モデルを用いた音声認識器を複数駆動させる音声対話システムについて述べる。我々は、自動車内での使用を想定し、異なる 3 つのドメイン(レストラン情報、天気予報、ニュース検索)に関する情報検索を行う実験システムを構築した。入力された発話のドメインは、それぞれの音声認識器のスコアに基づき決定される。異なるドメインの言語モデルからの尤度を適切に比較するため、各ドメインの音声認識器の言語重みと単語挿入ペナルティについて調整した結果、95%のドメイン識別率を得ることができた。

キーワード　ドメイン識別　言語モデル　音声認識　音声対話システム

A Study of Domain Switching Measure for the Spoken Dialogue System using Multiple Statistical Language Models Selectively

Toshihiro ISOBE[1,5]　Shoji HAYAKAWA[2,5]　Hiroya MURAO[3,5]
Tatsuji MIZUTANI[4]　Kazuya TAKEDA[4,5]　Fumitada ITAKURA[4,5]
1 NTT Data Corp.　1-21-2 Shinkawa, Chuo-ku, TOKYO, 104-0033 Japan
2 Fujitsu Laboratories LTD.　64 Nishiwaki, Ohkubo-cho, Akashi-shi, HYOGO, 674-8555 Japan
3 SANYO Electric Co., Ltd.　1-18-13, Hashiridani, Hirakata-shi, OSAKA, 573-8534 Japan
4 Graduate School of Engineering, Nagoya University 1 Furo-cho, Chikusa-ku, AICHI,464-8603 Japan
5 Center for Integrated Acoustic Information Research 1 Furo-cho, Chikusa-ku, NAGOYA,464-8603 Japan
E-mail:　isobet@nttdata.co.jp

Abstract　In this paper, we present a multi-domain spoken dialogue system equipped with the capability of parallel computation of speech-recognition engines that are assigned to each domain. The experimental system is set up to handle three different domains (restaurant information, weather report, and news query) in an in-car usage. All of these tasks are of information retrieval nature. The domain of a particular utterance is determined based on the likelihood of each speech recognizer. Experimental evaluation has yielded 95 percent recognition accuracy in selecting the task domain based on a specially designed scoring method.

Keyword　Domain Discrimination, Statistical Language Model, Speech Recognition, Spoken Dialogue System

1　はじめに

音声対話の書き起こしテキストコーパスの整備が進むに従って、音声対話システムにも統計的言語モデルが用いられるようになり、ドメインやタスクを限定すれば比較的自由な発話による音声対話システムの実現が可能となってきている。しかし、これまでの音声対話でよく用いられていた有限状態オートマトン(FSA)と異なり、統計的言語モデルは、モジュール性が高くないという欠点がある。例えば、二つ以上の N-gram 言語モデルを組み合わせるためのシステマチックな方法は確立されておらず、特に学習コーパスの大きさが違うときは組み合わせることが難しい。これは、複数の話題を扱うような対話システムを構築する際、重要な問題となる。

単純に予め決められたシナリオにしたがって言語モデルをスイッチすることは、複数の話題の対話戦略としては効果的だが、話題が頻繁に切り変わるような対話の場合には、この戦略を適用することは難しい。自動車を運転しながらの交通情報やナビゲーション検索、ドライバー補助などに対する Q&A セッションはこのような対話を必要とする状況であると想定される。

名古屋大学統合音響情報研究拠点(CIAIR)[1]では、自動車内での運転者の情報検索や運転支援を目的とした音声対話システムの実現に向け、試作機の構築と共に研究開発を実施してきた[1,2,5]。今回我々は試作機をマルチドメインに拡張することを検討した。本試作機では、入力に対して複数のドメインに対応する音声認識器を同時に動作させ、それぞれの音声認識器からの結果を選択的に用いる。

本稿では、3 つの独立の対話サブシステムからなるマルチドメイン音声対話システムについて示す。3 つのサブシステムのタスクドメインとして、「レストラン情報」、「天気予報」、「ニュース」をそれぞれ扱う。また、同システムのドメイン識別部におけるドメイン選択方法についてシミュレーション実験した結果について報告する。

2　マルチドメイン音声対話システム

2.1　システム構成

提案するマルチドメイン音声対話システムの構成を図１に示す。本システムは 3 つの音声認識器、3 つのタスク処理部、音声合成器、対話管理部、応答文格納部およびナビゲーションイベント通知部からなる。音声認識器とタスク処理部は三種類のドメイン（レストラン情報、天気予報及びニュース）にそれぞれ割り当てら

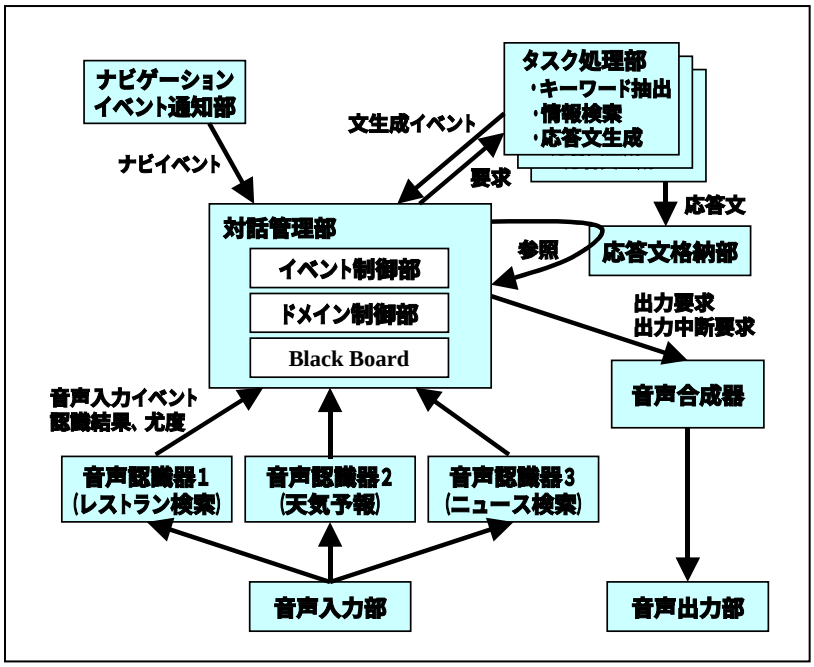

図１．システム構成

れている。

ユーザの発話はそれぞれのドメイン依存の音声認識器によって認識される。ドメインごとに音声認識器を駆動させることにより、統計的言語モデルにおけるモジュール性の問題に対応している。すなわち、ドメインの追加・削除は、駆動する音声認識器の選択により実現される。次にタスク処理部においてドメインごとの音声認識器からのスコアに基づき発話のドメインを選択し、システムの応答文が生成され音声合成器・音声出力部から合成音声が出力される。

また、これら 3 つの情報検索対話に加えて、ナビゲーションイベント通知部より道案内メッセージ（カーナビゲーションガイダンス）が送られ、対話処理に割り込むようになっている。

2.2 音声認識器

ユーザからの発声を検出すると、音声入力イベントと、それぞれのドメインからの音声認識結果および評価値（スコア）が”Blackboard”に書き込まれる。音声認識エンジンとしては、”大語彙連続音声認識エンジン Julius3.1”[3]を用いた。音響モデルはそれぞれのドメインに共通で性別非依存 PTM モデル[4]を用いた。言語モデルはドメインごとに学習した。学習コーパスとして CIAIR において収集された車内音声対話の文と FSA により自動生成した人工文とを合わせて使用した。

2.3 音声合成器

音声合成エンジンには波形接続方式である株式会社アニモの音声合成製品”FineSpeech[1]”を使用した。利用者側の確認の容易さを考慮し、ドメインごとに合成音の性別や声の高さを変えている。

2.4 対話処理部

対話処理部は、”Blackboard”、ドメイン制御部およびイベント制御部からなる。それぞれのドメインに対する音声認識結果やスコアは”Blackboard”に保存される。

対話処理部は、”Blackboard”に記録されている音声認識結果のスコアとユーザ発話無しの継続時間を用いてドメイン間のスイッチング処理を管理する。この対話処理部は 4 つの状態からなる状態遷移図として図 2 のように表される。各状態は、「レストラン情報」「天気予報」「ニュース」および「ニュートラル」にそれぞれ割り当てられている。ニュートラル状態は、上記 3 つの状態のいずれにもあてはまらないケースであることを表す。

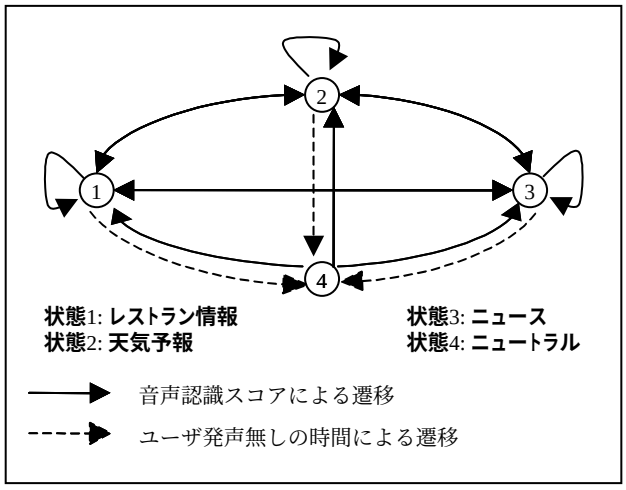

図 2 ：状態遷移図

ニュートラル状態への遷移以外の遷移においては、最も評価値(スコア)が大きいドメインが選択される。一方、ユーザの発声が一定時間中に無かった場合には、ニュートラル状態への遷移が起こる。

イベント制御部は、イベントを受理したり、モジュール間に要求を送ったりすることにより、対話システムを上位レベルから管理する。運転中における安全性の観点から、ナビゲーションイベントは最も優先度を高く設定してある。この場合、イベント管理部は他のすべてのプロセスに対して割り込みを行い、ナビゲーションに関する合成音を出力する。対話制御部の GUI を図 3 に示す。選択されたドメインの”Blackboard”が音声認識結果とタスク処理部からの応答文とともに、最前面に表示される。ドメイン制御部では、音響モデル、言語モデルのスコア、単語数から計算される各ドメインでの入力に対するスコアが表示される。

2.5 タスク処理部

システムは 3 つのドメインを持ち、3 種類のタスク処理部はドメインごとに駆動される。対

[1] “FineSpeech”は富士通株式会社の登録商標である

話管理部によって選択されたドメインのタスク処理部は、音声認識結果からキーワードを抽出し、スロットテーブルにそれらを埋め込み、情報検索に対してクエリーを発行する。要求された情報を得た後は、適切な応答文が生成され、応答文格納部に送られる。

2.6 応答文格納部

対話中に、ナビゲーションイベントが発生した場合には、対話システムの処理は中断され、ナビゲーションに関する合成音が優先的に出力される。応答文出力中に中断された場合には、中断された応答文を再度先頭から出力するために、応答文格納部に一時的に応答文がバッファリングされている。

3 ドメイン識別実験

3.1 音声認識スコアに基づくドメイン識別

提案システムにおけるドメイン識別は、複数の音声認識器が出力する複数の評価値のうち最も良いものを選択することにより行われる。単語数nで構成される音声認識仮説hの評価値は(1)式で表される。

$$f(h) = AC(h) + LM(h) \cdot WLM + n \cdot PLM \Lambda \quad (1)$$

$AC(h)$: 仮説hに対する音響モデルの対数出力確率
$LM(h)$:仮説hに対する言語モデルの対数出現確率
WLM: 言語モデル重み
PLM: 単語挿入ペナルティ

言語モデル重みや単語挿入ペナルティは評価値に対する音響モデルと言語モデルの影響の違いや、単語数による言語モデルの影響を補正するために用いられ、実験的に最適値に設定されることが多い。複数の音声認識器からの評価値を用いてドメイン識別を行う場合には、異なる音響モデルや言語モデルからの評価値を比較するため、ドメインごとにモデルの影響が異なることを考慮する必要がある。

実験では、次の3つの条件でのドメイン識別結果を比較した。なお音響モデルはドメイン独立とした。

1) 各ドメインの音声認識器の言語モデル重みと単語挿入ペナルティを音声認識率最適の基準で設定した場合
2) 1)と同じパラメータを、ドメイン識別率最適の基準で設定した場合
3) 言語モデルのエントロピー[6]に基づいて評価値を補正した場合

3.2 実験条件

「レストラン情報」、「天気予報」、「ニュース」の3ドメインについて、それぞれに対応させた認識器からの評価値の比較により識別した。評価データには男性12名による540発話を使用した。実験条件を表1に示す。

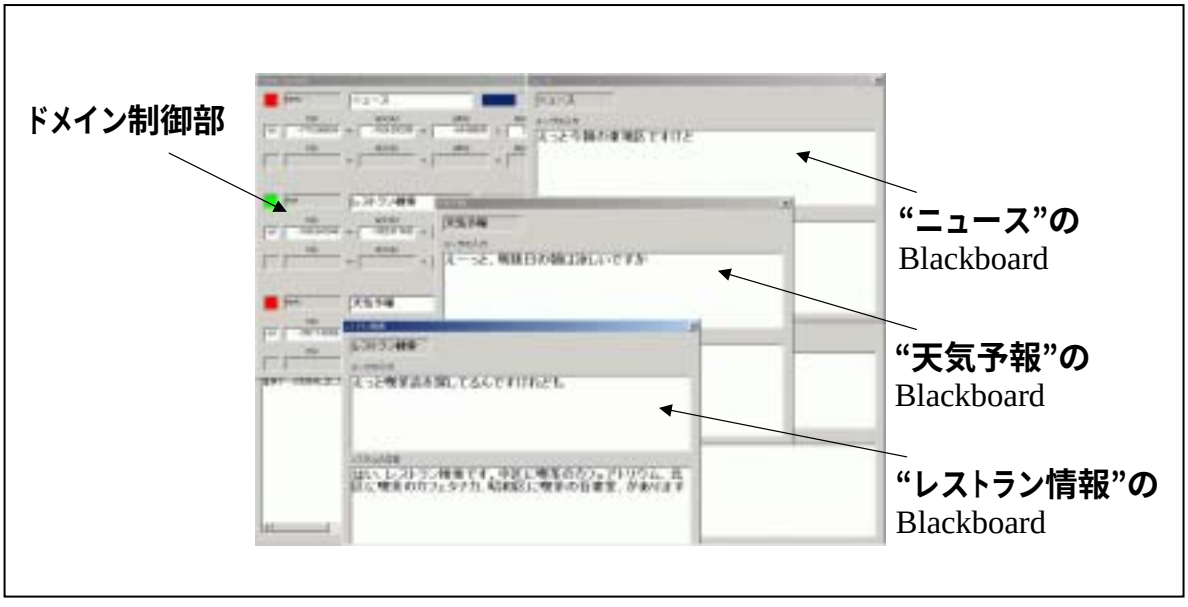

図３：対話制御部の画面

表1. 実験条件

認識エンジン	Julius v3.3
音響モデル	性別非依存 PTM モデル (3000 状態)（ドメイン独立）
言語モデル (3-gram)	学習文数: RS=6K　WT=3.3M　NW=13M 語彙ｻｲｽﾞ： RS=1940　WT=396　NW=515 OOV： RS=12　WT=30　NW=73 ﾃｽﾄｾｯﾄ PP： RS=29.2　WT=94.9　NW=165.0

注) RS: レストラン情報　WT: 天気予報　NW: ニュース

3.3 実験結果

1) 音声認識率

ドメイン既知の条件で、各音声認識器の単語認識率が最適となるように言語モデル重みを設定した場合の音声認識率と言語モデル重みを表2に示す。認識率が最適になる言語モデル重みは「レストラン情報」が他のドメインと比較して高くなった。

表2. ドメインごとの単語認識率

	ACC	COR
RS	77 %　(9.0)	82 %　(7.0)
WT	72 %　(3.0)	77 %　(2.0)
NW	78 %　(3.0)	84 %　(2.0)

注) ACC: 単語正解精度　COR: 単語正解率
() 内は言語モデル重み
単語挿入ペナルティは-2.0 に固定

2) ドメイン識別率

ドメイン識別に使用する評価値を計算するための言語モデル重み(*WLM*)や単語挿入ペナルティ(*PLM*)の設定法について、以下の5種類の条件を設定し、各条件に対して識別率、*WLM* 及び *PLM* を測定した。(表3参照)

a) ドメインごとに単語正解率が最適となる *WLM*、*PLM* を使用して仮説を求め、そのときのスコアをそのままドメイン識別に使用
b) a)で求めた仮説において、ドメイン識別率が最適になるように *WLM* だけを再設定（つまり、$AC(h)$、$LM(h)$、*PLM* は a)と同じ値を使用）
c) a)で求めた仮説において、ドメイン識別率が最適になるように *WLM*、*PLM* を再設定（つまり、$AC(h)$、$LM(h)$は a) b)と同じ値を使用）
d) ドメイン識別率が最適になるように *WLM* をドメインごとに再設定(再度認識（仮説の再推定）を行うため、仮説は音声認識時と異なる可能性あり)
e) 評価値（音声認識スコア）を、言語モデルのエントロピーを用いて補正

条件a)～d)では、単語正解率を最適化するなど、事後的な知識を用いてパラメータの設定を行っており、実際のシステム構築に際しては適当な方法ではない。事前知識のみを利用したドメイン識別の最適化手法について検討する必要がある。事前知識の一つとして、言語モデルのエントロピーを利用したスコア補正が、異なる言語モデル間のスコア補正に有効であることが報告されている[6]。そこで今回、条件e)を設定し、言語モデルのエントロピーを用いたスコアの補正について検討した。条件e)において、評価値 $f(h)$ は(2)式により計算する。ここで、$LM(h)$ は使用した3ドメインに対応する言語モデルのエントロピーの平均値 $\overline{H}$ によって補正される。表4に、各言語モデルのエントロピーを示す。エントロピーにより、(2)式における言語モデルに関する項（$LM(h)-n(H_i-\overline{H})$）は補正されるが、言語モデル項と音響モデルに関する項（$AC(h)$）の間の関係はここでは補正されていないため、これらを補正するパラメータである言語モデル重み *WLM* に関しては事後知識を用いることとし、ドメイン識別率が最適になるように再設定した。

$$f(h)=AC(h)+WLM\cdot\{LM(h)-n(H_i-\overline{H})\}\Lambda \quad (2)$$

$$\overline{H}=\frac{1}{m}\sum_{i=1}^{m}H_i \quad m:\text{ドメイン数}$$

$$H_i=-\frac{1}{3}\sum_{w_{t-2},w_{t-1},w_t\in\Omega_i}P_i(w_t\mid w_{t-2},w_{t-1})\log_{10}P_i(w_t\mid w_{t-2},w_{t-1})$$

$P_i()$: 3*gram* 確率

Ω_i: i番目のドメイン用言語モデルの全学習データ

表3. ドメイン識別率

	識別率	WLM RS / WT /NW	PLM RS / WT / NW
a	91 %	7.0 / 2.0 / 2.0	-2.0 / -2.0 /-2.0
b	94 %	8.0 / 3.0 / 3.0	-2.0 / -2.0 /-2.0
c	94 %	7.0 / 2.0 / 2.0	-2.0 / -4.0 /-2.0
d	95 %	9.0 / 2.0 / 3.0	-2.0 /-2.0 /-2.0
E	94%	7.0 / 3.0 / 3.0	---

表4. 各言語モデルのエントロピー

	Entropy H_i	$\overline{H}$	$H_i - \overline{H}$
RS	1.576	1.236	0.34
WT	1.076		-0.16
NW	1.056		-0.18

単語正解率を最適にする言語モデル重みにおけるスコアを用いてドメイン識別をした場合、識別率が最も低く91%であった。言語モデル重みと単語挿入ペナルティの設定の自由度を増加させたり、仮説を再推定したりすることにより、識別率が向上する傾向がみられた。言語モデルのエントロピーに基づくスコア補正を行うと、識別率は 94%となった。

4 考察

今回の実験では音響モデルをドメイン独立としているため、各ドメインの音声認識器の評価値は言語モデルの影響の違いのみが現れていると考えられる。実験では、言語モデル重みや単語挿入ペナルティを最適化することにより、ドメイン識別率の向上がみられたが、ドメイン識別率を最適化する場合と、音声認識率を最適化する場合とでは最適なパラメータの値は異なることがわかった。

音声認識時の言語モデル重みや単語挿入ペナルティの値は、仮説の評価値に与える音響モデルと言語モデルの影響の違いを補正するのに効果的であり、ドメイン識別時におけるこれらのパラメータの値は、言語モデル同士の影響の違いの補正に効果的であったといえる。

言語モデルのエントロピーによるスコア補正手法を用いた場合のドメイン識別性能は、理想的なパラメータを使用した場合と比べ同等な結果となった。よって、平均的な統計的言語モデルのエントロピーの値と追加しようとする統計的言語モデルのエントロピーの値が事前知識として使用できれば、言語重みの調整だけで補正が可能であると言える。

5 まとめ

「レストラン情報」、「天気予報」、「ニュース」の3つのドメインについて検索を行う音声対話システムを紹介した。本システムでは、複数の音声認識器からの評価値を比較してドメイン識別を行う。ドメイン識別実験を行った結果、最適なドメイン識別性能を得るためには、言語モデル重みと単語挿入ペナルティのパラメータを音声認識時とは異なる値に制御する必要性が確認された。また、パラメータを最適化することにより、約95%の入力発話に対して正しいドメインを識別できることがわかった。

今後は言語モデルの複雑度を考慮した評価値の制御法や、音響モデルの切り替えも含め、ユビキタス環境下での音声対話を目指した総合的な環境変動への対応手法等について検討していく予定である。

6 謝辞

本研究は文部科学省科学研究補助金 COE 形成基礎研究費(課題番号 11CE2005)の補助を受けて行われた。

参考文献

[1] http://www.ciair.coe.nagoya-u.ac.jp/

[2] 早川, 磯部, 河口, 武田, 板倉, “音声対話システムを用いた車内対話の収集”, 音講論集, 3-P-25, pp.213-214 (2001.3)

[3] Lee, A. et al., “Julius - an Open Source Real-Time Large Vocabulary Recognition Engine”, in Proceedings of Eurospeech2001, pp.1691-1694

[4] 李晃伸, 河原達也, 武田一哉, 鹿野清宏, “Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識”, 信学論 J83-DⅡ, 12, pp.2517-2525 (2000)

[5] N. Kawaguchi, S. Matsubara, K. Takeda and F. Itakura, “Multi-Dimensional Data Acquisition for Integrated Acoustic Information Research”, in Proceedings of 3rd International Language Resources and Evaluation Conference (LREC-2002), pp. 2043-2046, Canary, Spain, May. (2002).

[6] 水谷隆治, 武田一哉, 板倉文忠, “音声認識における音響モデル言語モデルの切り替え方法に関する検討”, 音講論集, 3-Q-31, pp.213-214. (2003.3)