論文 / 著書情報
Article / Book Information

論題(和文)	並列処理型計算機による混合主導型対話音声認識システムの構築
Title(English)	
著者(和文)	田熊 竜太, 森山 達裕, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2002年秋季講演論文集, Vol. , No. 2-9-9, pp. 79-80
Citation(English)	, Vol. , No. 2-9-9, pp. 79-80
発行日 / Pub. date	2002, 9

並列処理型計算機を用いた混合主導型対話音声認識システムの構築 *

◎田熊 竜太　森山 達裕　岩野 公司　古井貞熙　(東工大)

1　はじめに

発話内容ごとに作成された複数の言語モデルを並列処理型計算機を用いて同時並行に駆動し、尤度を基準に発話内容を同定し最適な認識結果を出力する混合主導型の対話システムを構築した [1]。本システムは、発話内容ごとの言語モデルからの出力を尤度を基準に選択することにより発話内容を同定する。同時にその発話のキーワードを抽出する。いくつかの発話内容同定法を提案し、単一の言語モデルを用いたシステムとの比較を行う。

2　対話タスク

本稿では飲食店検索をタスクとした。利用者は希望の場所と料理を発声し、さらに希望条件を発話して、店舗データベースから好みに合った店の情報を得ることができる。場所としては東京都圏主要駅 1,069 駅が登録されている。飲食店データは NTT 番号情報案内 (株) の提供する「インターネットタウンページ」[3] のデータベースを利用した。このデータベースは 80 業種、4,491 店舗分の営業時間、平均予算、利用可能サービスなどから構成されている。

3　システムの内部状態

システムはマイク入力と画面出力を持つ。画面表示に対応した以下の 4 つの内部状態を持ち、この内部状態を利用者の発話内容に応じて遷移させ利用者の要求を実現する。

- **内部状態 a** (駅名・料理受付)
  利用者の利用したい駅名と食べたい料理の種類を受け付ける。
- **内部状態 b** (番号発声による確認)
  内部状態 a での駅名および料理を確認する。選択肢には番号が振られており、利用者はこの番号を発声することで選択をする。駅名・料理のいづれかが間違いだった場合は内部状態 a に戻る。両方が正しかった場合は内部状態 c へ進む。
- **内部状態 c** (検索条件絞込み)
  駅名・料理から絞られる店舗候補数は画面に表示するには多すぎることがある。その場合、この内部状態で利用者に更なる絞込み条件を発声してもらう。画面表示可能な候補数まで絞られた場合は内部状態 d へ遷移する。
- **内部状態 d** (店舗詳細情報表示)
  画面に候補となる店舗の名前を表示し、利用者にその店舗名を発声してもらうことで店舗の詳細情報を表示する。

任意の内部状態で利用者はシステムを直前の状態に戻したり、検索条件を変更したり、システムを初期化や終了することができる。内部状態の遷移の様子を図 1 に示す。

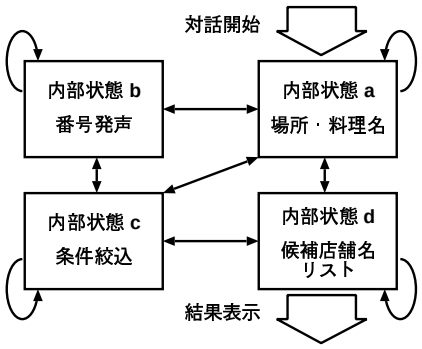

図 1. システムの内部状態

4　発話内容と言語モデル

システムを使用する利用者からの想定される発話をその内容ごとに 5 つに分類し、各発話内容ごとに以下の 5 つの言語モデルを作成した。

- **LM-P** : 場所・料理の種類用言語モデル
  本システムでは対話開始時に利用者が食べたい料理と最寄の駅を指定する。料理、駅名のいづれか一方でも含む発話を発話内容 **P&F** とした。
- **LM-R** : 検索条件絞込み用言語モデル
  「3,000 円以内で食べられるところ」「駅から近いところで」など、店舗候補を絞り込むための発声を発話内容 **RES** とした。
- **LM-S** : 店舗名用言語モデル
  画面に表示された候補店舗名を発声することで利用者はその店舗の詳細情報を見ることができる。店舗名を含む発話を発話内容 **STR** とした。
- **LM-C** : システム制御コマンド用言語モデル
  利用者はシステムを制御するためのコマンドを含む発話を発声することができる。コマンドを含む発話を発話内容 **COM** とした。
- **LM-N** : 番号発声用言語モデル
  画面に表示された選択肢には番号が振られており、その番号を発声することで利用者はシステムの提示する候補を選択することができる。番号発声を含む発話を発話内容 **NUM** とした。

5　発話内容同定

音声信号 x が入力されたとき、発話内容 c の単語列 w が出力される確率 $P(c, w|x)$ は、字式のように表される。

$$P(w, c|x) = \frac{P(x|w, c)P(w, c)}{P(x)} = \frac{P(x|w, c)P(w|c)P(c)}{P(x)}$$

∗ Mixed-initiative spoken dialogue architecture based on parallel-computing

By Ryuta Taguma, Tatsuhiro Moriyama, Koji Iwano and Sadaoki Furui, Department of Computer Science, Tokyo Institute of Technology

表 1. システムの内部状態と想定発話内容

内部状態 a	**P&F, COM**
内部状態 b	**NUM, COM**
内部状態 c	**RES, COM**
内部状態 d	**STR, RES, NUM, COM**

したがって、次式で表される尤度 P を最大化する w, c を求めることで、発話内容 c の同定と単語系列 w が決定される。

$$P = P(x|w,c)P(w|c)P(c) = P(x|w,c)P(w|c)^{\alpha}P(c)^{\beta}$$

ここで、$P(x|w,c)$ は音響尤度、$P(w|c)$ は言語尤度、$P(c)$ は発話内容尤度、α は言語重み、β は発話内容重みとなる。

発話内容同定法として以下の 4 つを提案する。

- **AL：** 複数の認識器からの結果を比較する尤度として言語尤度と音響尤度のみを利用する。発話内容尤度は考慮していない。
- **ALC：** 発話内容尤度を用いた手法。発話内容尤度は各内部状態における発話内容の出現頻度から計算した。計算に用いたデータは評価データとは別の対話音声である。
- **ALCR：** 利用者はシステムの提示に従った発声をすると仮定する。この場合、利用者の発声する発話内容は各内部状態において、表 1 に示す発話内容に限定される。そこで、発話内容尤度を次のルールで定める。「利用者は等確率で表中の発話内容を発声すると仮定し、その発話内容に等しい尤度を与え、表中に含まれない発話内容は尤度 0 とする。」
- **KNOWN：** 利用者が発声する発話内容をシステムがあらかじめ知っていると仮定した理想的な手法で、複数の言語モデルからの出力を比較することはなく、利用者の発声した発話内容の言語モデルのみを用いて認識する。

また、単一プロセッサにより単一の言語モデルを利用したシステムの例として以下の二つの手法による評価実験を行った。

- **ALM：** 作成した 5 つの言語モデルを等しい重みで混合した言語モデルを用いて認識する。これは全ての内部状態であらゆる発話内容を受理できるという意味で並列処理型計算機を用いた手法の **AL** に対応する。
- **CLM：** 各内部状態で、表 1 中の発話内容を受理する言語モデルのみを等重みで混合する。各内部状態でシステムが想定した発話内容のみを受理することができるという意味で並列処理型計算機を用いた手法の **ALCR** に対応する。

6 実験

提案する 4 つの発話内容同定法による結果と単一プロセッサによるシステムを想定した 2 つの手法の結果を図 2 に示す。評価セットには収録した対話音声 1,209 文を用いた。

発話内容尤度を利用しない **AL** と比較して発話内容尤度を用いた **ALC**、**ALCR** は発話内容同定率、

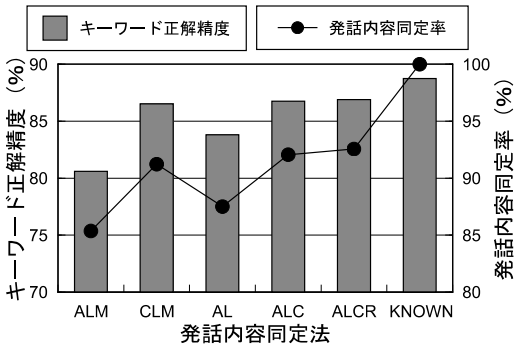

図 2. 認識結果

キーワード正解精度とも上昇した。また、発話内容同定率が 100% と仮定した理想条件 **KNOWN** では、キーワード正解率がさらに上昇している。これにより、発話内容同定率の改善によりキーワード正解精度がさらに上昇することを示している。

発話内容尤度をルールで定めた **ALCR** と学習データから算出した **ALC** を比較すると、発話内容同定率、キーワード正解精度ともに **ALCR** の結果が僅かばかり勝っている。しかし、システムの各内部状態で、あらゆる発話内容を受理できるような場合、**ALCR** の手法は **AL** の手法と同じになる。そのような場合には **ALC** の手法が有利である。

単一プロセッサでの結果と並列処理型計算機を用いた手法を比較する。対応する **ALM** と **AL**、**CLM** と **ALR** を比較するとどちらも提案手法の性能が勝っている。これにより、混合主導型対話システムにおいては発話内容に応じた複数の言語モデルを用意し同時並行に駆動する本手法が有効であることがわかる。

7 まとめ

飲食店検索をタスクとする混合主導型音声対話システムを構築した。利用者からの自由な発話を受理するために、想定される発話内容ごとに言語モデルを用意した。並列処理型計算機を用いてそれぞれの言語モデルを持つ認識器を同時並行に駆動し、発話内容の認識結果を尤度に基づくスコアを用いて選択する。このようなシステムを 2 つ構築し、単一言語モデルのみからなるシステムと発話内容同定率とキーワード正解精度を比較した。混合主導型のシステムにおいて提案手法の有効性を示した。

謝辞 店舗情報データベースを提供していただいた NTT 番号情報株式会社に感謝する。

参 考 文 献

[1] 田熊 他：“並列処理型計算機を用いた音声対話システムの検討 “ , 人工知能学会 言語・音声理解と対話処理研究会, pp.21-26 (2000-6).

[2] J. G. Fiscus：“A post-processing system to yield reduced word error rates： Recognizer output voting error reduction (ROVER),“ Proc. IEEE ASRU, pp.347–354, (1997-12).

[3] http://www.itp.ne.jp/