論文 / 著書情報
Article / Book Information

論題(和文)	自由発話を対象とした対話システムの構築と評価
Title(English)	Development and Evaluation of a Spoken Dialog System Using Spontaneous Speech
著者(和文)	岩野公司, 斉藤朗, 貞苅 宏宣, 田熊竜太, 古井貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	情報処理学会研究報告, 2001-SLP-36-12, Vol. 2001, No. 55, pp. 79-86
Citation(English)	, Vol. 2001, No. 55, pp. 79-86
発行日 / Pub. date	2001, 6
権利情報 / Copyright	

自由発話を対象とした対話システムの構築と評価

岩野公司　　斉藤　朗　　貞苅宏宣　　田熊竜太　　古井貞熙

東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
Email: {iwano, asaito, sada, rtag, furui}@furui.cs.titech.ac.jp

ユーザにとってより使いやすい対話システムを目指し，我々は自由発話による音声入力が可能な情報検索を目的とした対話システムの構築を進めている．そこで，効率よく改良を進めるために，ユーザの主観評価を反映する数値（コスト）を定義し，それらのコストを削減する方針で改良を行った．改良後のシステムについて被験者実験を行ったところ,「システムが結果を表示してからユーザが次の動作を実行するまでの時間の総和（反応コスト）」がユーザの主観評価を強く反映していることがわかり，それによって効果的にシステムが改善されたことが示された．さらに，被験者実験で得られた音声の書き起こしデータから作成した言語モデル（N-gram）を用いた認識デコーダを用いることで，自由発話入力時のキーワード認識率が 約 13 % 改善された．これにより，さらなるシステムのコスト削減・使用感改善の可能性が示された．

Development and Evaluation of a Spoken Dialog System Using Spontaneous Speech

Koji Iwano, Akira Saito, Hironori Sadakari, Ryuta Taguma, Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: {iwano, asaito, sada, rtag, furui}@furui.cs.titech.ac.jp

We have been developing a spoken dialog system for information retrieval, which can accept spontaneous utterances to provide user-friendly interface. In this paper, we report our dialog system has been improved efficiently by decreasing objective costs which reflect users' subjective evaluation. As a result of subjective experiments, "user response costs" most effectively work on improvement of the system; this cost means expected sum of time from system output to next user's action for a task. Furthermore, 13 % improvement of the keyword recognition rate in spontaneous utterances was observed, using a language model (N-gram) built from many transcriptions obtained through the subjective experiments. Since the cost is expected to be decreased further, our system will be more user-friendly using the stochastic language model.

1　はじめに

近年，音声認識技術の発展に伴い，多くの音声対話システムの構築が進められている．我々も，情報検索を目的とした対話システムの構築を行い [1][2]，その発展，改良を進めてきた [3]．これは，東京近郊の店舗情報を検索するための対話システムであり，ユーザの発話中から場所・業種といったキーワードを抽出し，尤度の高い順に候補を画面に出力，ユーザに選択・確認を促すといった流れでタスクが進行していく．

ユーザにとってより使いやすいシステムを構築するためには，自由発話による入力が望ましいと考えられる [4]．我々のシステムにおいても，自由発話中に現れる不要語をガーベージモデルで吸収するようなネットワーク文法を利用し，自由発話による音声入力を実現している．しかし一方で，自由発話の認識率やシステムの構成方法によっては，孤立単語認識を基本とした，いわゆる「一問一答」形式のシステムの方が，ユーザの使用感が良いという結果になることを多く耳にする．詳細は後述

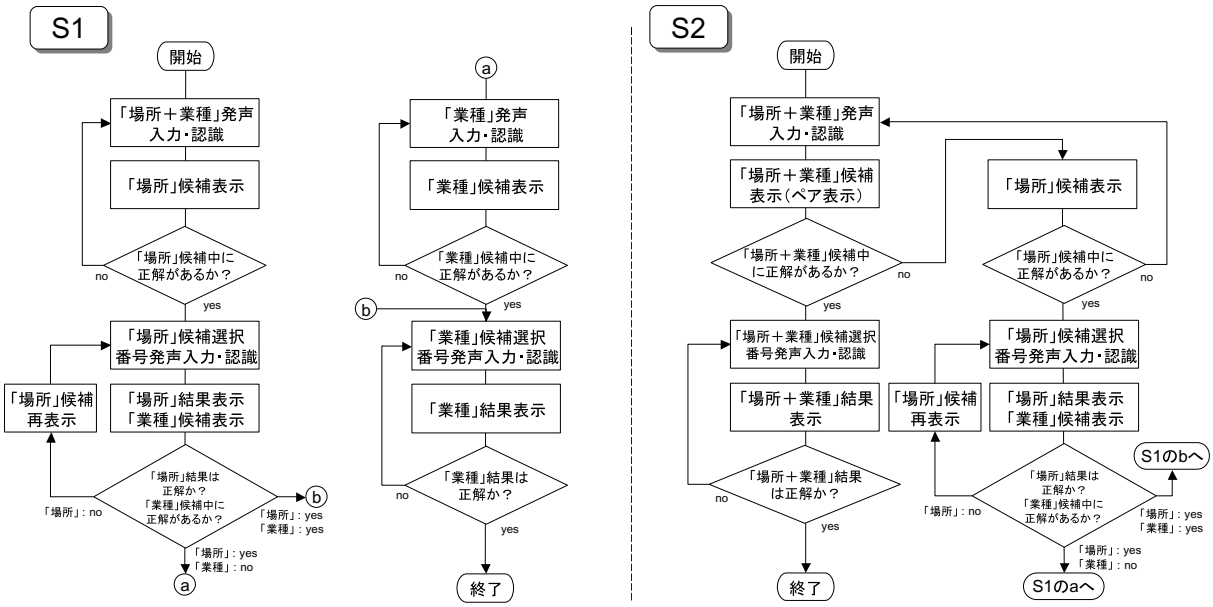

図 1: 店舗情報検索のための対話システムの流れ

するが，実際に，初期の我々のシステムと一問一答形式の入力を用いたシステムとの主観評価実験による比較では，一問一答形式の方が高い評価結果を得ている．ユーザにとってより自然な自由発話よりも一問一答のシステムの方が評価値が高くなるのは，自由発話システムのユーザインタフェースに問題があるためと考えられる．

そこで，本研究ではシステムを効率的に改良し，ユーザの使用感の向上を試みる．システムの改良を進めるにあたって，ユーザの主観評価を強く反映する客観的数値（コスト）を定義することができれば，被験者を要する評価実験を不必要に行わずにすみ，改良の方針も立ちやすい．本稿では，このコストを改良前のシステムによる被験者実験を通して定義し，コストを削減する方針で行ったシステムの改良について述べる．さらに，改良前後のシステムについて，コストと被験者による主観評価結果を比較し，コストの有効性・改良の妥当性を示すものとする．

また，上記コストを削減するための一つの方法として，自由発話の認識率の向上があげられる．現在のシステムは，ネットワーク文法を認識に用いているが，大量のデータから構築された統計的言語モデルを用いることにより認識率の向上が期待される．そこで，上記の被験者実験より得られた発話データを書き起こし，それから構築された N-gram を用いることによる認識率の向上についても検討を行った．その結果についても，本稿の後部で報告する．

2　初期のシステム構成

改良前のシステム構成と比較対象となる一問一答形式のシステムの構成について説明する．改良前の 2 種類のシステムを **S1**, **S2**，比較用の一問一答形式のシステムを **S0** とする．**S1**, **S2** のシステムの流れを図 1 に示す．フローチャート中の条件分岐部分は，マウスによるボタン入力で行うようになっている．また，発声の入力前にはシステムからの音声プロンプトが提示される（システム主導）．

システム主導型のシステムは音声プロンプトの与え方によってユーザの発話のバラエティーをある程度抑えることが可能である [5]．本システムにおいても,「番号入力」や「場所入力」「業種入力」といった場合には，単語発声に近い音声が多く入力される傾向にある．一方,「場所＋業種入力」では，両方のキーワードが含まれるような一括発声を促しており，自由発話の性質が強い音声が入力・認識されるため，前者に比べ認識率が劣る傾向にある．

S1 と **S2** との大きな違いは，**S1** は一括入力の

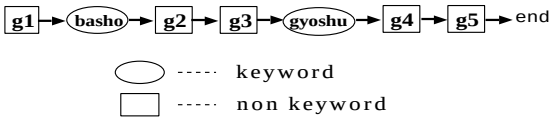

図 2: 店舗情報検索システムの認識で用いるネットワーク文法

後，それぞれのキーワードを個別に表示・確認していくように設計されているが，**S2** は一括入力の後にキーワードの組（ペア）を表示し，一括して確認させるような設計である，という点である．なお，**S0** は,「場所＋業種」発声による一括入力は行わず,「場所」「業種」の順に個別に入力・確定させていく．**S1**, **S2** の後半の「業種」発声の入力・認識以降（図 1 の@以降）の流れを，対象を「場所」としてまず行い,「場所」確定後に対象を「業種」として行う，といった構成である．

一括入力音声の認識に用いるネットワーク文法を図 2 に示す．**g1**, **g2**, **g5** は文頭・文中・文末に挿入されるフィラーや言い回しを登録したノンキーワードモデル，**g3**, **g4** は業種名の前後につく「お」や「屋さん」といった表現を吸収するためのノンキーワードモデルであり，合計で 88 種類のノンキーワードが登録されている．また，**basho**, **gyoshu** で示される場所・業種のキーワードモデルにはそれぞれ 204, 206 種類の単語が登録されている．

各システムの認識器には，NTT-AT 社の「音声認識ソフトウェア REX」の開発キットを用いており，100 best で認識結果を得た後，キーワードに違いのある認識結果を最大 10 個，上位から選んで表示する．その中に入力したキーワードがない場合（認識誤り）は，再発声を促すこととした．なお，認識はある程度速度を優先するように設定しているため，実験結果は最高性能を示してはいない．

3　初期システムによる被験者実験

前節にあげた改良前の 2 つのシステムと，一問一答形式のシステムの主観評価の比較，および認識率や時間といった数値データの収集のため，35 名（男性 27 名，女性 8 名）による被験者実験を行った．35 名うち約半分は情報系以外を専攻している学生である．残りの半分は，情報系を専攻している学生・社会人で，音声研究に携わっている者も 9 名含まれている．

主観評価は「認識性能」「認識時間」「操作性」「メッセージのわかりやすさ」について 5 段階（1: 非常に良い〜 5: 非常に悪い）で行った．また，システムの総合的な使用感について 3 つのシステムの順位付けを行ってもらった．

同時にシステムは各イベントの時間ラベルや認識結果を記録しており，最終的に，

- ユーザが発声し結果が返ってくるまでの平均時間（発声・認識時間）: U
- 認識結果の候補をユーザが見てから次の行動（条件分岐先を確定させるためのボタン操作）を起こすまでの平均時間（反応時間）: C
- 発話のキーワード認識率 : P

といった物理的評価値を算出する．

3.1　実験条件

各ユーザはシステムの提示する音声・表示の指示にしたがって，場所と業種を入力・確定するタスクを行う．タスクは，2 つのキーワード入力が正しく行われる（正常終了）か，同じ場面での再発声の回数が 3 回を超える（強制終了）ときに完了とし，それまではユーザにタスクの変更を認めないものとした．実験にあたり，ユーザにはシステムの目的を説明し，あらかじめ問い合わせを行う場所と業種を 10 種類考えてもらい，紙に記録しておいてもらう．各システムを使用する際にはその紙を参考に発声を行ってもらうため，各被験者は，どのシステムに対しても同じ 10 種のタスクを行っていることになる．発声の形態については，特に指示を出さず自由に発声してもらった．また，ユーザの慣れによる主観評価の偏りをなくすため，システムの使用順はユーザごとに異なるようランダムで与えた．

3.2　実験結果

被験者による主観評価の平均を表 1 に，物理的評価値を表 2 に示す．

表 1: 初期のシステムにおける主観評価結果

	認識性能	認識時間	操作性	メッセージ	総合順位
S0	2.47	2.75	2.42	2.31	1.33
S1	2.92	3.39	2.58	2.42	2.17
S2	3.03	3.31	2.89	2.64	2.47

表 2: 初期のシステムにおける物理的評価値

	U（秒）	C（秒）	P（%）
a. 場所＋業種	14.72	5.45	48.9
b. 場所のみ	6.30	3.09	77.0
c. 業種のみ	7.62	3.05	71.0

各物理的評価値はそのイベントの出現場面が「**a.** 場所＋業種」「**b.** 場所のみ」「**c.** 業種のみ」の場合について分けて示す．U，C は一回あたりの平均の時間で示してある．認識率 P はある発話に対して，出力された最高 10 件の候補中に正解キーワードが含まれている確率を表している．例えば，「**a.** 場所＋業種」の入力をした際の P（以後 P_a とする）は，発話中の 2 つのキーワードが共に候補中に含まれる確率を示している．なお，発声したキーワードが認識語彙の中に含まれておらず，未知語となってシステムが強制終了する場合が，各話者について平均で約 3 回存在したが，このような場合は評価値の算出に用いていない．

表 1 をみると，「場所＋業種」の自由発話による一括入力を許すシステム **S1**, **S2** が，一問一答形式の **S0** に比べ総合的に低い評価になっており，ユーザが一番使いやすいシステムとして **S0** を選んでいることがわかる．また，表 2 を見ると，U, C, P 全てについて「場所＋業種」発声時の評価値が劣っており，この評価値の改善が自由発話を利用したシステム **S1**, **S2** の使用感を改善するために必要であると考えられる．

4 システムの改善

前節の結果を踏まえ，システム **S1**, **S2** の改善を行う．そこで，U, C, P を一元的に扱うことのできるコストを 2 種類考案し，それを削減する方針でシステム改良を進めた．

4.1 コストの定義

主観評価を反映する数値として，以下の 2 種類のコストを定義した．

発声・認識コスト $\hat{U}_{all}$： 1 タスクを完了するまでに要する（番号発声による候補選択以外の）発声・認識時間 U の総和の期待値

反応コスト $\hat{C}_{all}$： 1 タスクを完了するまでに要する反応時間 C の総和の期待値

発声・認識コスト $\hat{U}_{all}$ は話しかけてからシステムの反応が返ってくるまでの待ち時間の合計に相当するため，このコストが減ることはシステムがより迅速・正確に応答していることを意味する．一方，反応コスト $\hat{C}_{all}$ はシステムの応答に対してユーザが次の行動を考える時間に相当するため，このコストが減ることはユーザがより迷わず迅速に反応できる，すなわちユーザにとって操作がわかりやすくなるということを意味している．また，両者とも認識誤りによる言い直しを繰り返すことで増加するため，認識率の向上によってもコストは減少する．

コストの計算の際には，誤認識による言い直しは無限回行うものとした．例としてシステム **S1** におけるコスト $\hat{U}_{all}$, $\hat{C}_{all}$ の式を以下に示す．ここで，U_a は「**a.** 場所＋業種」発声時の，U_c は「**c.** 業種のみ」発声時の U を示している．C_b, C_c, P_c についても同様に，添字が認識対象の発声内容を示している．ただし，P_{ab}, P_{ag}, P_{an} は「**a.** 場所＋業種」の一括入力を行ったとき，表示する候補の中に「場所のみ正解が含まれる確率」「業種のみ正解が含まれる確率」「場所・業種とも正解が含まれていない確率」を示しており，観測された値は，$P_{ab} = 26.9\%, P_{ag} = 10.4\%, P_{an} = 13.7\%$ であった（$P_a + P_{ab} + P_{ag} + P_{an} = 1$）．$m, n$ はそれぞれ「場所＋業種」「業種のみ」発声の（誤認識によって生じる）言い直しの回数を示している．なお，候補選択時の番号音声の認識率は非常に高いことから 100 % として近似した．

$$\hat{U}_{all} = \lim_{m,n\to\infty} \left\{ \sum_{k=1}^{m} (P_{ag} + P_{an})^{k-1} P_a \cdot kU_a \right.$$

表 3: 初期システムにおけるコスト

	$\hat{U}_{all}$（秒）	$\hat{C}_{all}$（秒）
S0	18.9	8.31
S1	23.2	8.65
S2	23.2	11.9

$$
\begin{aligned}
&+\sum_{k=1}^{m}\sum_{l=1}^{n}(P_{ag}+P_{an})^{k-1}P_{ab}\\
&\cdot(kU_a+(1-P_c)^{l-1}P_c\cdot lU_c)\Big\}\\
=\ &\frac{U_a+P_{ab}\cdot\frac{U_c}{P_c}}{P_a+P_{ab}}
\end{aligned}
$$

$$
\begin{aligned}
&\hat{C}_{all}\\
=\ &\lim_{m,n\to\infty}\Bigg\{\sum_{k=1}^{m}(P_{ag}+P_{an})^{k-1}P_a\cdot(kC_b+C_c)\\
&+\sum_{k=1}^{m}\sum_{l=1}^{n}(P_{ag}+P_{an})^{k-1}P_{ab}\\
&\cdot((kC_b+C_c)+(1-P_c)^{l-1}P_c\cdot lC_c)\Big\}\\
=\ &\frac{C_b+P_aC_c+P_{ab}(C_c+\frac{C_c}{P_c})}{P_a+P_{ab}}
\end{aligned}
$$

表 2 に示した U, C, P の実測データを，定義したコストの式に代入したときの結果を，それぞれのシステムごとに表 3 に示す．どちらのコストについても，表 1 の主観評価結果の順位との関連性が見て取れる．

4.2 コスト削減によるシステムの改良

コストの削減によって自由発話入力を許すシステム **S1**, **S2** の改良を図る．図 3 に，定義式より計算される，P_a を変化させた時の各システムのコスト $\hat{U}_{all}, \hat{C}_{all}$ の変化の様子を示す（図中の点線で示される **S3** については次項 4.2.1 で言及する）．この時，P_{ab}, P_{ag}, P_{an} の値は，被験者実験で得られた比の関係を保持しつつ，$P_a+P_{ab}+P_{ag}+P_{an}=1$ を満たすように設定した．他の値は固定している．この図から得られる知見を用いて，以下に述べる改良を行った．

4.2.1 システム構成の改善

$\hat{C}_{all}$ の変化を見ると，システム **S2** の方が **S1** に比べ，コストの減り方が急であることがわかる．これは，**S2** が「場所＋業種」の一括入力結果をまとめて確認できることから，P_a が向上すればするほど，余分な経路を通ること無くタスクを終了できることに起因している．それに対し，**S1** は特に低認識率の範囲でコストが抑えられている．これは，**S2** では場所候補中に正解がない場合，「場所＋業種」「場所のみ」の二段の確認を経て発声に戻る流れになっているため，P_a の低認識率部でコストが劣化するのに対し，**S1** では「場所」確認のみに移ることでその劣化を回避しているためである．そこで，1)「場所＋業種」発声の一括入力認識・一括確認を行い，2) それが成功しなかった場合は「場所」発声・認識・確認に移るシステムを作成した．これをシステム **S3** とする．なお，「場所＋業種」入力の認識結果を見ると，一括入力が成功しなかった場合は，場所よりも業種が間違っている可能性が約 16 % 高い（$P_{ab} > P_{ag}$）ため，システム **S3** では一括入力が成功しなかった場合については業種の認識結果は保持せず，必ず業種発声をやり直すようにした．

このシステム **S3** に関し，他のシステムと同様にコストの式を定義し，P_a を変化させた時のコスト $\hat{U}_{all}, \hat{C}_{all}$ の変化を，図 3 にあわせて示してある．これをみると，$\hat{U}_{all}$ では **S1**, **S2** より改善が見られ，$\hat{C}_{all}$ に関しては，約 90 % を境として，認識率の低い箇所では **S2** より改善が見られ，認識率の高い箇所では **S1** より改善が見られる．

以後の評価実験では，将来的な認識率の向上を期待して，**S1** の代わりに **S3** を改良システムとして加えることとした．

4.2.2 認識率・認識時間・操作性の改善

図 3 からは，自由発話である「場所＋業種」発声の認識率を向上させることでシステム **S1**, **S2** のコストは削減されていくが，両者が **S0** のコストより小さくなるためには，$\hat{U}_{all}$ で約 70 %，$\hat{C}_{all}$ で約 75 % 以上という高い認識性能を確保しなければならないことがわかる．したがって，認識率 P の改善だけでなく，U, C の改善も重要である．

そこで，**S2**, **S3** において P, U, C の値の改善を

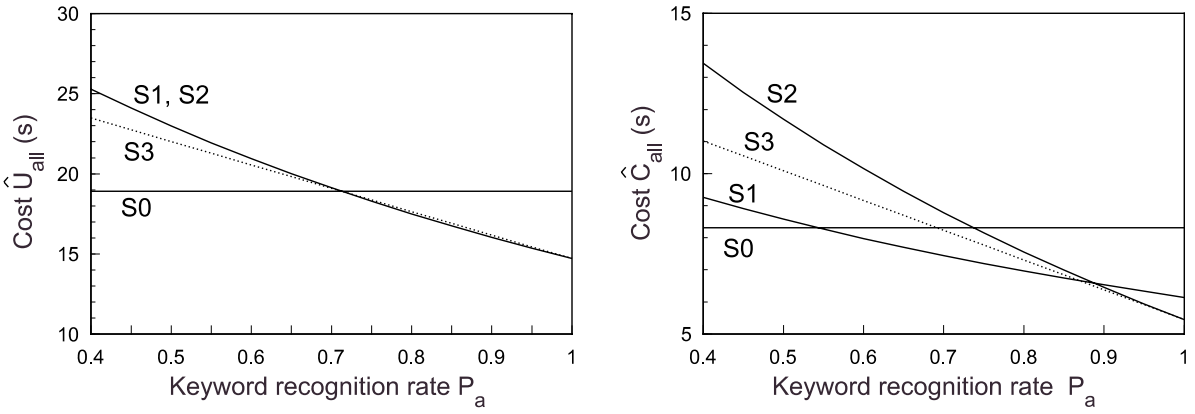

図 3: 「場所＋業種」発声時のキーワード認識率 P_a とコスト $\hat{U}_{all}$（左），$\hat{C}_{all}$（右）の関係

行った．まず，認識ソフトウェアのバージョンをあげ認識率 P の改善と，認識時間 U の短縮を行った（**S2-R**, **S3-R**）．さらに C を改善するため，システムのボタン表示方式の簡素化や，ボタン数の削減といったインタフェースの改良を行った（**S2-RI**, **S3-RI**）．

5　改良システムによる被験者実験

3 節と同様の方法で改良システムについても被験者による評価実験を行った．被験者の数は 11 名（男性 10 名，女性 1 名）で，この中の 8 名は初期システムの被験者実験にも参加していた．比較のための **S0** を含めて，5 つのシステム（**S0**, **S2-R**, **S3-R**, **S2-RI**, **S3-RI**）の評価実験を行った．表 4 に得られた主観評価結果を，表 5 に改良後のシステムで実測された物理的評価値（C については **S2-RI**, **S3-RI** での数値）を示す．表 6 には，この被験者実験で得られた物理的評価値を用いて新たに計算し直したコストを示す．したがって，改良を加えていない **S0** のコストであっても表 3 の値と一致していない．

表 5 より，改良を行ったシステムでは，P_a について 20 % 近くの改善がなされ（テストセットが異なるため厳密な比較ではない），U_a, C_a についてはそれぞれ約 6 秒，2 秒近く短縮されていることがわかる．

表 4, 6 をみると，ユーザの評価順位と発声・認識コスト $\hat{U}_{all}$ には相関がみられない．一方，反応コスト $\hat{C}_{all}$ と評価順位はよく合致しており，「操作性」「メッセージのわかりやすさ」の評価値とも強い関連が見られる．このことから，$\hat{C}_{all}$ がユーザの主観評価を反映する値として有効であると考えられる．

そこで，$\hat{C}_{all}$ と順位の相関について検定を行った．反応コスト $\hat{C}_{all}$ と主観評価の順位には，被験者間で違いがあるため，各被験者ごとに $\hat{C}_{all}$ と順位についてスピアマンの順位相関係数を求め，その相関係数に関する検定を行った．帰無仮説を「両者の間に相関が無い」とした検定の結果，有意水準 5 % で 11 名中 10 名，有意水準 1 % で 11 名中 9 名について帰無仮説が棄却された．

$\hat{C}_{all}$ に注目すると，認識性能のみを改善したシステム **S2-R**, **S3-R** のコストは改良前の初期システムに比べれば，それぞれ確かに小さくなっているが，一問一答形式のシステム **S0** よりは小さくなっておらず，結果的に主観評価でも **S0** を超えるに至っていない．これは，初期システムの被験者実験で得られたコストによる予想（図 3 中の $P_a = 67.8\%$ 付近での $\hat{C}_{all}$ の順位）と一致している．

最終的には，インタフェースの改良を施すことで，コストをさらに 2 秒程削減することができたため，**S0** よりも小さいコストが実現され，主観評価においても **S2-RI**, **S3-RI** の方が優れているという結果を得ることができた．

表 4: 改良システムとシステム **S0** における主観評価結果

	認識性能	認識時間	操作性	メッセージ	総合順位
S0	1.82	2.00	2.45	2.27	3.00
S2-R	2.27	2.18	3.55	3.55	4.73
S3-R	1.82	2.18	3.00	3.18	3.82
S2-RI	1.82	2.09	1.73	2.18	1.91
S3-RI	2.00	2.27	1.73	2.09	1.55

表 5: 改良後のシステムにおける物理的評価値

	U（秒）	C（秒）	P（%）
a. 場所＋業種	8.37	3.15	67.8
b. 場所のみ	5.65	2.41	85.0
c. 業種のみ	5.86	2.61	87.5

表 6: 改良システムとシステム **S0** のコスト

	$\hat{U}_{all}$（秒）	$\hat{C}_{all}$（秒）
S0	15.7	6.26
S2-R	11.2	8.11
S3-R	11.5	7.51
S2-RI	11.2	5.75
S3-RI	11.7	5.32

6 統計的言語モデルを用いた認識性能の改善

前節までに，本システムの使用感の改良には，自由発話の認識性能の改善，認識時間の短縮，インタフェースの改善などが重要であることを述べた．ここでは，自由発話の認識性能のさらなる改善について検討する．

これまで使用したシステムは，ネットワーク文法を認識に用いているが，被験者実験によって大量の自由発話データが得られたことから，その書き起こしを用いて統計的言語モデルを作成することが可能である．そこで，この言語モデルを用いることによる「場所＋業種」発声時の認識率（P_a）の向上を試みる．

初期システムの被験者実験で得られた 35 名分の「場所＋業種」発声部分について書き起こしを行い，これを学習データとして言語モデルの構築を行う．認識誤りによってやり直した再発声分も含まれるため，得られた書き起こしは約 1100 文となった．このデータ中に出現しない語彙（キーワード）が存在するため，場所と業種を表す単語についてはそれぞれ ⟨**basho**⟩, ⟨**gyoshu**⟩ と書き起こしてクラスのモデルとし，N-gram を作成した（クラス N-gram）．辞書中には，各クラスに対して，ネットワーク文法時に用いられていた語彙（キーワード）を同様に登録しておく．

比較する対象は，5 節で述べた改良システムでの認識性能（$P_a = 67.8$ %）であるから，その時の被験者実験で得られた音声データを実験データとした．また，実験データの話者が学習データ中に含まれている場合があるので，その際には学習データからその話者のデータを除いて 34 名分のデータで言語モデルを構築し，各話者ごとに認識結果を算出することとした．

なお，ここでの認識には 2 パスデコーダである「大語彙連続音声認識デコーダ Julius 3.1[6]」を用いた．使用する言語モデルは 2-gram と 逆向き 3-gram である．比較対象のネットワーク文法を用いたシステムの認識は，速度をある程度優先させるような設定を用いているため，この実験においても速度優先の設定を用いて実験を行った．

実験の結果，P_a は 81.1 % となり，約 13 % の認識性能の向上が確認された．ただし，両者は言語モデルだけでなく，音響モデル，探索手法も異なるため，単純に言語モデルによる効果と断定することはできない．したがって，厳密な言語モデルの違いによる性能比較ではないことを強調しておく．認識性能が 81.1 % まで改善することを考えると，図 3 より反応コスト $\hat{C}_{all}$ が **S2** で約 1.5 秒，**S3** で約 1 秒改善されるため，さらにユーザの使用感は高まるものと考えられる．

7 まとめ

自由発話による音声入力を用いた店舗情報検索を目的とする対話システムの改良のため，被験者実験に基づいてシステムのコストを定義し，それによるシステムの改善，検証の比較実験を行った．その結果，システムの提示に対してユーザが反応するまでの時間を表す反応コスト $\hat{C}_{all}$ がユーザの主観評価結果を強く反映することがわかった．そのコストを自由発話部の認識率やシステムのインタフェースを改善することで削減し，システムの使用感を改善することができた．このコストは,「ユー

ザがシステムの提示をみて，次の動作を判断するまでの時間・手間」を表しており，これがシステムのわかりやすさとして，ユーザの使用感に大きな影響を与えていることを示唆している．結果として，当初は自由発話を許したシステムが一問一答形式のシステムより使用感で劣っていたが，改良後のシステムはそのシステムより使用感の優れたものとなった．

また，さらなる改善手法の一つとして，統計的言語モデル（クラス N-gram）を用いた自由発話の認識性能の向上を検討したところ，約 13 % のキーワード性能率の向上が確認され，これによってさらなるコストの削減，使用感の向上が可能であることが示された．

今回の報告では，コストの定義に際してユーザの「慣れ」は考慮していない．実際には，タスクを繰り返すことで本来は短くなる時間（判断時間など）があり，より厳密にコストを定義するためには，このような影響を考慮する必要がある．今後は，これら今回扱っていない要因についても検討する必要がある．また，システム自体の改良として，クラス N-gram を用いた認識部の実装を検討している．

謝辞

ご助言を頂いた NTT サイバースペース研究所の大附克年氏に感謝致します．また，被験者実験に協力して下さった多くの方々に感謝致します．

参考文献

[1] 山口晃一郎，古井貞熙，“音声対話を用いた情報検索システムの検討，” 音講論集，3-6-19, pp.119–120 (1998-3).

[2] S. Furui and K. Yamaguchi, “Designing a multimodal dialogue system for information retrieval,” *Proc. ICSLP*, vol.3, pp.1191–1194 (1998-12).

[3] 本田征嗣，斉藤　朗，古井貞熙，“信頼度尺度を用いた音声対話システムの検討，” 音講論集，3-8-11, pp.87–88 (2000-3).

[4] 中川聖一，“小特集に寄せて–音声対話システム構築の課題，” 音響誌，vol.54, no.11, pp.783–790 (1998-11).

[5] 中川聖一，山本誠治，“音声対話システムの構成法とユーザ発話の関係，” 信学論（D-II），vol.J79-D-II, no.12, pp.2139–2145 (1996-12).

[6] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano, “Free software toolkit for japanese large vocabulary continuous speech recognition,” *Proc. ICSLP*, vol.4, pp.476–479 (2000-10).

[7] 斉藤　朗，古井貞熙，“情報検索のための音声対話システムの構成法と評価，” 音講論集，1-3-22, pp.47–48 (2001-3).