複合的言語制約に基づくキーフレーズ検出を用いた汎用的なデータベース検索音声対話プラットフォーム

駒谷和範[†] 鹿島博晶[†,☆]
田中克明[†,☆☆] 河原達也[†]

ドメインに依存しないで汎用的に用いることのできるデータベース検索音声対話プラットフォームを提案する．様々なドメインにおける音声入力インタフェースの開発には，文法の記述もしくは当該ドメインのコーパス収集など多大な作業が必要である．本プラットフォームでは，ドメイン固有の語彙をデータベースから抽出し，キーフレーズ部分の文法をテンプレートに基づき生成する．この文法と類似タスクドメインコーパスから学習された統計的言語モデルを融合して複合的言語制約とし，これを用いたキーフレーズスポッティングにより柔軟な音声認識・理解を実現する．本プラットフォームを用いてホテル検索システムと文献検索システムを作成し，ホテル検索システムでの 24 名の発話データを用いて評価実験を行った．複合的言語制約に基づくキーフレーズスポッティングにより，記述文法を言語制約とする従来手法と比較して，意味理解誤り率は 15.5%削減された．

Domain-independent Spoken Dialogue Platform for Database Query Using Key-phrase Spotting Based on Combined Language Model

Kazunori Komatani,[†] Hiroaki Kashima,[†,☆] Katsuaki Tanaka[†,☆☆] and Tatsuya Kawahara[†]

We present a domain-independent platform of spoken dialogue systems for database query. Conventional development of speech interfaces involves much labor cost in either describing task grammars or collecting domain corpora. Our platform generates a lexicon and a language model of key-phrases based on task description. The generated grammar is combined with the word 2-gram model trained with similar domain corpora. Flexible speech understanding is realized by spotting key-phrases based on the combined language model. We applied this platform to hotel search and literature search tasks. The experimental evaluation on the generated hotel search system shows that the key-phrase spotter using the combined language model improves the semantic accuracy by 15.5% compared with decoding the whole utterance using a fixed grammar.

1. はじめに

近年の音声認識技術の進歩により，様々な機関で音声対話システムの研究・開発がさかんに行われている．これまでに研究されてきたシステムは多岐にわたるが，実用化が有望なタスクの 1 つにデータベース検索タスクがあげられる．データベース検索タスクにおける主要な操作は検索条件の入力であるが，音声には多数の選択肢からの選択を効率良く行えるという利点[19] がある．そのためデータベース検索タスクは，マウスなどのポインティングデバイスと比較して，音声インタフェースが有効に働くタスクの 1 つである．

しかしながら，現在の実用的な音声対話システムの構築には，認識辞書や文法の作成，インタフェースの設計など，人手による膨大な手間を要している．また作成したシステムの他のドメインへの移植性[1),6),9)] もきわめて低い．音声対話の分析やモデルの学習のためには当該タスクドメインにおける大量の対話データが必要であり，Wizard of Oz 法を用いたデータ収集がしばしば行われるが，人間が対話システムと同等の能力を持ったシステム役を演じなければならないなど，非常に労力を要する作業である．したがって音声対話システムを簡単に作成できるラピッドプロトタイピング技術は重要である[11)]．

† 京都大学大学院情報学研究科知能情報学専攻
Graduate School of Informatics, Kyoto University
☆ 現在，日本アイ・ビー・エム株式会社
Presently with IBM Japan, Ltd.
☆☆ 現在，ヤマハ株式会社
Presently with Yamaha Corporation

本稿では，このようなプロトタイプシステムの作成を容易にするための，汎用的なデータベース検索音声対話プラットフォームを提案する．タスクを定型的なデータベース検索に限定することにより，ユーザ発話を定型的なキーフレーズの組合せでモデル化する．この際に，キーフレーズ文法だけでは言語モデルとして不十分なため，類似タスクドメインコーパスから学習された統計的言語モデルを融合し，複合的言語制約を構成する．これを用いてキーフレーズスポッティングを行うことにより，定型的な発話の認識率を損なうことなく，非定型な発話に対しても柔軟な音声認識・解釈を実現する．プロトタイプシステムにおいて，事前に当該ドメインの学習データを十分に用意するのは困難であるため，本稿ではタスクドメインが完全に一致したコーパスの存在を前提とすることなく，柔軟な言語制約を実現する．

2. 汎用性を備えた音声対話システム

2.1 データベース検索タスクのモデル

本システムでは，属性とその値の組で定義される**図 1** のようなデータベースに対して，条件の入力・削除を行う定型的なデータベース検索を対象とする．検索項目間の関係は AND のみを扱う．

このようなデータベース検索タスクにおけるユーザ発話は，各検索フィールドに対応するスロットへの値の追加/削除であると見なせる．この検索フィールドとその値（キーワード）の組をキーフレーズとして定義し，文法をキーフレーズ単位で生成する．ドメインに依存した語彙は当該ドメインのデータベースから自動的に抽出する．キーフレーズに対する文法は，図 6 に示される例のように生成される．これらによりドメイン固有の言語制約を作成する．

本プラットフォームにおける意味解釈の目的は，発話中のキーワードを，対応する検索スロットに変換することである．あるキーワードと検索スロットとの対応は，当該データベース中でキーワードが属していたフィールドから得られる．解釈に複数の可能性が生じる場合，すなわちある単語がデータベース中で複数のスロットに存在する場合には，後段の対話処理部へその情報を渡し，対話的に曖昧性を解消する．

このようにキーフレーズを単位として発話をモデル化し，それを発話中からスポッティングすることにより，音声認識・理解を行う．キーフレーズは非定型な発話においてもその構文が保持されることが多く，かつ意味解釈の単位と合致するため，頑健な理解が実現できる[2]．

名称	***ホテル
タイプ	ビジネス
所在	三重県 津市
交通	近鉄津駅から徒歩で 1 分
シングル料金下限	5,000
シングル料金上限	6,600
立地	駅周辺
付帯施設	レストラン カフェ
周辺レジャー	ゴルフ 海水浴

図 1 データベースの具体例
Fig. 1 An example of database.

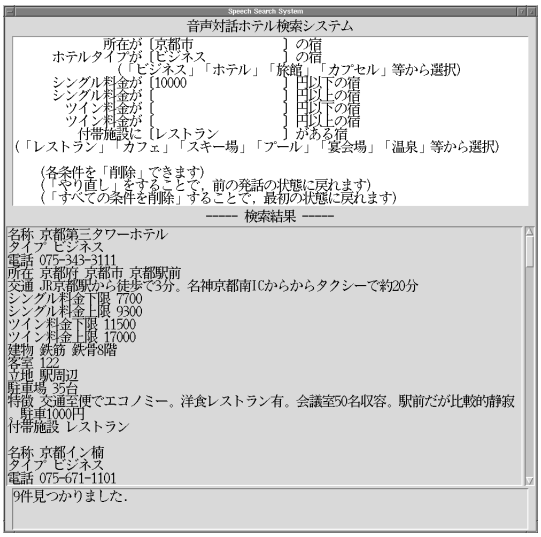

図 2 システムの GUI（ホテル検索）
Fig. 2 Outlook of GUI (hotel query system).

2.2 GUI の利用

システムの想定していない語彙外・文法外などの発話による音声認識誤りは，音声対話システムにとって重大な問題である．認識誤りによる誤解の発生に対しては，マルチモーダルな応答によりシステムの状態をユーザに提示することが有効であると知られている[20]が，音声認識誤りを防ぐためには，システムの受理可能なキーフレーズパターンをユーザに知らせるのが効果的である．そこで本プラットフォームでは，**図 2** に示すような GUI によりシステムの状態を逐次ユーザに提示するとともに，典型的なキーフレーズパターンをユーザに示し，ユーザ発話をシステムの受理可能な範囲内に誘導する．典型的なキーフレーズパターンは，アプリケーション開発者が，後述するタスク規定ファイル群生成ツールを用いて規定する．

検索は発話ごとに実行し，結果を表示する．そのためシステムは，検索条件の確認や検索実行の許可などの多くの対話処理を省略できる．検索条件の追加が行われると，追加された検索条件が画面上の対応する検索項目に表示される．これによりユーザは逐次的に検索結果とともに現在の検索条件，つまりシステムの状

態を知ることができる．

2.3 音声認識部における柔軟な言語制約

音声認識部の言語モデルには，当該タスクドメインのコーパスが大量に得られる場合には単語 N-gram モデルが用いられる[3),4)]．N-gram モデルにより，非定型な発話に対しても柔軟に認識を行うことが可能となるが，個々のタスクドメインにおいて十分な学習データを用意することは容易でなく，プロトタイプシステムの作成には向かない．一方，文法を用いて認識を行う場合[7)]は，タスクに特化した知識を容易に導入でき，語彙などの変更も簡単に行えるが，受理可能な発話が限定されてしまうためユーザの多様な発話に対処できない．

このような問題を解決するために，キーフレーズ部分にはドメインごとに用意される語彙と文法，それ以外の部分には類似タスクドメインのコーパスから学習した統計的なモデルを用いた，複合的言語制約を提案する．ドメインに特化した語彙（固有名詞など）が含まれるキーフレーズ部分に対しては，語彙はデータベースから抽出し，文法はフレーズ単位のテンプレートから生成する．キーフレーズ以外の部分に対して類似タスクドメインのコーパスから得られる統計的言語モデルを用いることで，汎用性を損なうことなく，柔軟な音声認識部が実現できる．

3. 音声対話プラットフォームの構成

本プラットフォームは，アプリケーション開発者によるタスク仕様に基づき単語辞書や文法を生成するタスク規定ファイル群生成ツールと，タスク規定ファイル群に従ってユーザとのインタフェースとして動作する汎用的音声対話エンジンから構成される（**図 3**）．以下 3.1 節でタスク規定ファイル群生成ツールについて，3.2 節で汎用的音声対話エンジンについて述べる．

3.1 タスク規定ファイル群生成ツール

本ツールはアプリケーション開発者により指定されるタスク仕様に基づいて，データベースから語彙を抽出し文法を生成することで，タスク規定ファイル群を作成する．処理の流れを**図 4** に示す．タスク規定ファイル群には，キーフレーズ部分に対する音声認識用の文法・単語辞書のほかに，意味解釈用の文法や同音異表記語列挙ファイル，タスク仕様を保存した検索項目設定ファイルが含まれる．

タスク規定ファイル群生成ツールの GUI を**図 5** に示す．アプリケーション開発者は，検索項目の名称とその別名，キーワードの後に用いられる助詞とその別名，キーワードの単位・敬称などを指定し，キーフレーズパターンを定義する．さらに，キーワードの抽出方法やフレーズ文法タイプ（項目名称の省略を認めるか否か），検索タイプ（キーワードが入力された際の検索方法）を GUI 中の項目から選択し，指定する．キーワードの抽出方法は，データベースから自動抽出（分かち書きされているエントリをそのまま抽出，または形態素解析して抽出）するか，後述するように数値や年月・日付といった概念を指定して生成するかを選択する．指定されたこれらの内容（タスク仕様）に基づいて，データベースから各項目ごとに語彙が自動的に抽出される．

文法はテンプレートに従って生成され，定義されたキーフレーズのパターンは図 2 のユーザインタフェース上に明示される．たとえば，図 5 の画面からホテ

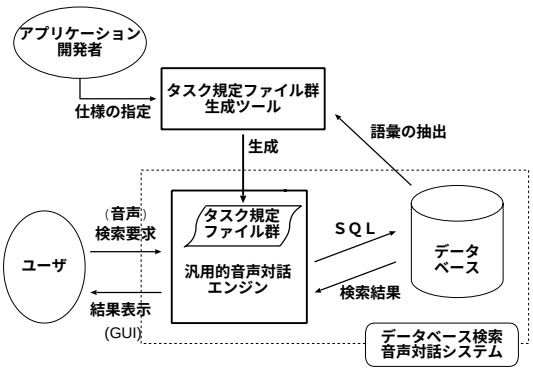

図 3 ドメインに依存しない音声対話プラットフォームの構成
Fig. 3 Overview of domain-independent platform of spoken dialogue interface.

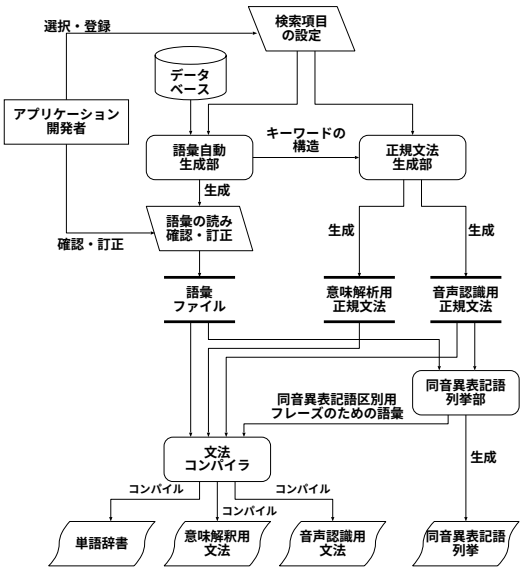

図 4 タスク規定ファイル群生成ツールの構成
Fig. 4 Overview of generation of task description files.

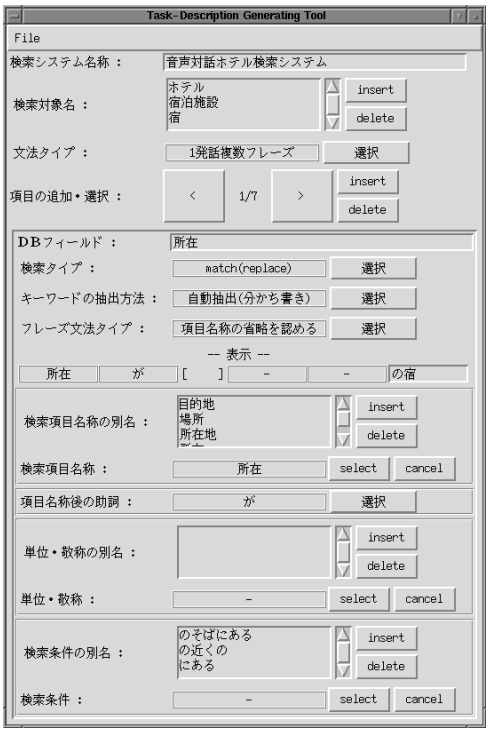

図 5 タスク規定ファイル群生成ツールの GUI
Fig. 5 Outlook of GUI for task description.

```
KEY_PHRASE : ITEM0 KEYWORD0
KEY_PHRASE : KEYWORD0
ITEM0 : NAME0 ZYOSHI0
KEYWORD0 : $所在$
KEYWORD0 : $所在$ END0
NAME0 : 目的地|場所|所在地|所在
ZYOSHI0 : が
END0 : のそばにある|の近くの|にある
```

図 6 所在フィールドに対して生成される文法の例
Fig. 6 An example of generated grammars for the location.

ルの所在に対するキーフレーズの文法が**図 6** のように生成される．$所在$ は地名を表す非終端記号であり，この語彙は検索に用いるホテルデータベースから自動的に抽出する．フレーズ文法タイプとして「検索項目名称の省略を認めない」を選択した場合には，KEY_PHRASE : KEYWORD0 が生成されず，「所在が」の部分が必須となる文法が生成されるなど，キーフレーズの文法タイプを選択することもできる．生成された文法は各検索項目ごとに分けておき，認識結果がどの検索項目に属するかを判定するための意味解釈用文法とする．このようにして音声認識用文法と意味解釈用文法，単語辞書がタスク規定ファイル群として生成される．その後，キーワード中の同音異表記語が文法と単語辞書から抽出され，同音異表記語列挙ファイルが生成される．

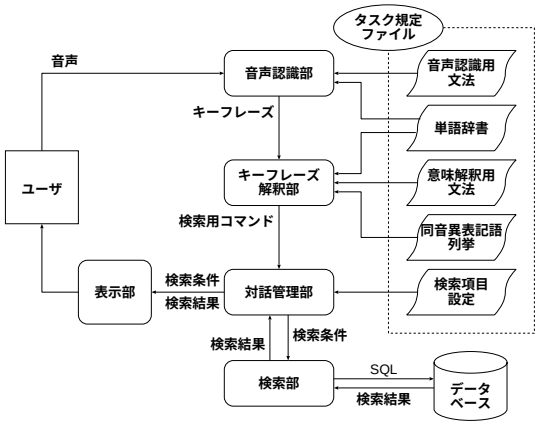

図 7 データベース検索のための汎用的音声対話エンジンの構成
Fig. 7 Overview of domain-independent spoken dialogue engine for database query.

語彙をデータベースから抽出する際に，地名や数値表現に対して拡張を行う．たとえば金額の場合，データベースに「8,000 円」というエントリがたまたま含まれていないと「8,000 円」を含まない単語辞書ができてしまう．そこで，「金額」や「距離」などの数値フィールドや「年」「日付」「時間」「曜日」といった年月・日付に関するフィールドに対する単語辞書を用意する．数値フィールドの場合，アプリケーション開発者が「1,000 円から 20,000 円まで 1,000 円刻み」のようなタスク仕様を指定すると，プラットフォームはこの仕様に従い単語辞書を生成する．これらの拡張は，異なるドメインでも共通に存在する年月や数値という概念に依存するもので，ドメインには依存しない．導入した概念はフライト予約[4] や列車の検索[3),5)] でも一般的に用いられるため，音声対話プラットフォームに用意することで汎用性の向上が期待できる．

3.2 データベース検索のための汎用的音声対話エンジン

汎用的音声対話エンジンはタスク規定ファイル群に従って，そのドメインのデータベースに対する検索システムとして動作する（**図 7**）．

音声認識部ではキーフレーズを単位としてスポッティングを行い，意味解釈部では認識されたキーフレーズがどの検索項目の文法により受理されるかを判定する．キーフレーズごとに信頼度を計算し，効率良く確認を行うことも可能である[21)]．また相対的に日付を表す表現（昨年，…年前）に対して，現在の日付から解釈し具体的な数値に変換する．

複数の解釈が存在するキーフレーズに関しては，対話管理部においてユーザに質問を行うことにより曖昧性を解消する．本プラットフォームでは，(1) 検索キーがどの検索項目に対応するか一意に判定できない，(2) 検索キーに同音異表記語が存在する，の 2 種類の曖昧性が存在する．(1) はたとえばホテル検索システムにおいて金額だけが発話されたときに，それがシングル料金の金額かツイン料金の金額か分からない場合である．(2) は文献検索システムで人名の「あべ」に対して，データベースに「安部」「阿部」の 2 つが存在する場合が例としてあげられる．システムはいずれの場合もユーザに候補を示し選択を促すことで，対話的に曖昧性を解消する．

得られた意味解釈結果に従って検索条件が更新され，検索が行われる．検索条件と検索結果は GUI を通じて逐次ユーザに提示される．検索結果として，該当した項目数とそのすべての内容がユーザに提示される．

4. 複合的言語制約に基づくキーフレーズ検出

対話音声のような多様な発話を認識・理解するには，文全体を認識するのではなく，必要な部分のみを認識するキーフレーズスポッティングに基づく手法が有効であり，非定型な発話に対する頑健性の向上が示されている[2]．しかし，キーフレーズとフィラー☆が任意に接続するような緩い言語制約を用いる場合には，特に短いキーフレーズの湧き出し誤りが多く発生する．

そこで，タスクドメインに特化したキーフレーズ部分に記述文法を，タスクドメインに依存しないフィラー部分に対しては類似タスクドメインのコーパスから得られる単語 2-gram を与える，複合的言語制約に基づくキーフレーズスポッティングを提案する．フィラー部分に単語 2-gram を適用することにより，キーフレーズとフィラーが任意に接続する言語制約に比べて文全体に対する言語制約が強くなる．この単語 2-gram はドメイン固有の表現を含まないフィラー部分に用いるため，タスクドメインの完全な合致を要求せず，類似タスクドメインの大規模コーパスを利用できる．

タスクドメインが完全に合致したコーパスを用意できる場合には，それから統計的言語モデルを学習したり[8]，大規模なコーパスから学習された言語モデルを適応したり[15] することができる．このような場合，記述文法を用いるよりも統計的言語モデルを用いる方が優れていることが示されている[10),14),16]．これに対して本研究では，主にプロトタイプシステムの作成など，統計的モデルを学習するための当該タスクの学習データがまったくない場合においても，柔軟な言語制約を構成する方法を提案する．

☆ キーフレーズ部分以外をフィラー部分とする．

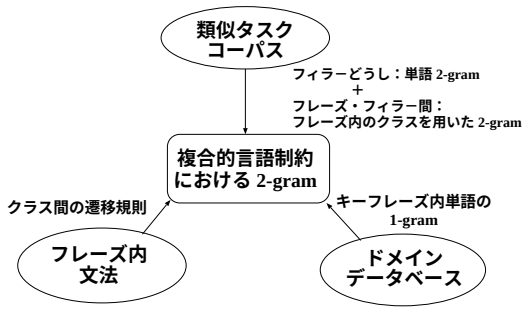

図 8 2-gram 構築における複合的言語制約の適用
Fig. 8 Concept of combined language model.

4.1 記述文法と類似タスクド メインコーパスによる 2-gram の構築

以下，単語 2-gram を複合的言語制約に基づいて構築する手法について説明する．ここでは，類似タスクドメインのコーパスから得る 2-gram と，検索に用いるデータベースによる 1-gram，キーフレーズ内の記述文法の 3 つの言語制約を組み合わせて 2-gram を構成する（**図 8**）．

まず，フィラーどうしの遷移時に用いる 2-gram は，コーパスによる推定値をそのまま適用する．これは，特に文末表現などの言語制約として働く．次に，フィラーとキーフレーズ間の遷移時に与える 2-gram には，コーパスによる推定値を近似的に用いる．キーフレーズ内の単語はドメインに特化した語彙であり，類似タスクドメインのコーパス中に存在することは期待できないため，コーパス中の名詞（主に普通名詞）のクラス 2-gram を用いる．また，比較的ドメインに独立な概念（金額，日付，地名など）を持つキーワードは，コーパス中の同概念の単語をクラス化し，そのクラス 2-gram を用いる．クラス 2-gram を単語 2-gram に展開する際，検索に用いるデータベースの項目に対応するクラスに対しては，そのデータベースにおける単語の出現頻度を反映した 1-gram を利用する．たとえば，同じ地名クラス中でも「京都市」と「宇治市」では，前者の方が検索に用いるデータベースでの出現回数が多く，発話される可能性も高いため，そのデータベース中の出現回数に応じて確率を付与する．キーフレーズ内の遷移は，記述文法に基づき設定される．文法カテゴリとクラス 2-gram のクラスは同じになるように設計する．したがって文法ルールはクラス間の遷移の有無を規定する．検索に用いるデータベースから自動抽出したキーワードに対しては，そのデータベースに

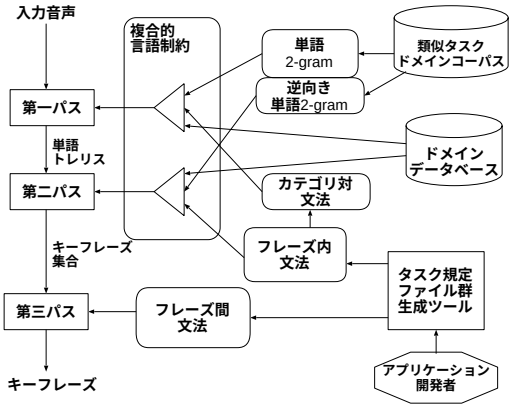

図 9 複合的言語制約を用いたキーフレーズ検出
Fig. 9 Overview of key-phrase spotting method based on combined language model.

より推定された 1-gram に基づいて確率を付与する．

以上をまとめると以下のような式になる．

- フィラー部分

$$p(w_2|w_1) \quad = \quad p_{co}(w_2|w_1)$$

- フィラー・キーフレーズ間

$$p(w_2|w_1) = \begin{cases} p_{co}(c_2|w_1) \cdot p_{db}(w_2|c_2) \\ \quad \text{(フィラーからキーフレーズへの遷移)} \\ p_{co}(w_2|c_1) \\ \quad \text{(キーフレーズからフィラーへの遷移)} \end{cases}$$

- キーフレーズ内

$$p(w_2|w_1) = p_{gr}(c_2|c_1) \cdot p_{db}(w_2|c_2) = \begin{cases} p_{db}(w_2|c_2) & (\text{if } c_1c_2 \text{ is defined}) \\ 0 & (\text{otherwise}) \end{cases}$$

ただし，w_i，c_i はそれぞれ単語，クラスを表し，$p_{co}()$，$p_{db}()$，$p_{gr}()$ はそれぞれ類似タスクドメインコーパス，検索に用いるデータベース，フレーズ内文法から得られる確率である．フレーズ内文法に与えられる確率（$p_{gr}(c_2|c_1)$）は，文法で c_1c_2 の遷移が定義されていれば 1，そうでなければ 0 となる．実際には上式に加えてバックオフスムージングが行われる．

4.2 複合的言語制約に基づくキーフレーズ検出

音声認識においては，キーフレーズに着目した段階的探索を行う．これはキーフレーズ部分に対して，段階的に強い言語的制約を適用していくものである．処理の流れを**図 9** に示す．

まず第 1 パスでは，複合的言語制約に基づき前向きにフレーム同期のビーム探索を行い，入力全体として単語候補を絞り込む．次に第 2 パスでは，第 1 パスの結果をヒューリスティックとし，スタックデコーディングによる後向きの best-first 探索を行う．その際に，通常の N-best 探索では一部の機能語が置換された候補しか得られない場合が多い．そこで多様なキーフレーズ候補を効率良く求めるために，仮説のマージを行う．すなわち既出のフレーズ列と同じフレーズ列を導く展開を行う仮説は破棄する．具体的には，キーフレーズあるいはフィラーとして受理された仮説にさらに 1 単語接続したときに，その接続境界時刻を保存する．この時刻がすでに保存されている時刻と一致すれば，その仮説の展開はそこで中止される．ここで受理された仮説がキーフレーズであればその始終端時刻やその仮説のスコアとともに出力する．

第 3 パスでは，スポッティングされたキーフレーズを組み合わせて，文として認識・理解する．キーフレーズ候補はそのスコアと出現制約に従って接続される．出現制約は図 5 の GUI でアプリケーション開発者により指定された，文法タイプが「1 発話複数フレーズ」か否かから得られる．この認識も A*探索により実現される[2]．

表 1 ホテル検索システムと文献検索システムの諸元
Table 1 Specifications of two generated systems.

項目	ホテル検索	文献検索
データベースのサイズ	850 KB	333 KB
データの件数	2,040	1,913
検索項目数	7	5
語彙サイズ	903	4,158
キーフレーズ部分の文法ルール数	837	160

5. 実装と評価実験

5.1 ホテル検索システムと文献検索システムの実装

提案した音声対話プラットフォームをホテルデータベースと文献データベースに適用し，ホテル検索システムと文献検索システムを作成した（**表 1**）．ホテル検索システムは，関西地区の 2,040 件のホテルが収録されているデータベースを検索対象として，「所在」「シングル料金上限」「付帯施設」などの 7 項目に対して条件を追加・削除できる．1 発話で複数の検索項目を指定・削除することも可能である．文献検索システムは音声の研究に関する 1,913 件の文献を対象とし，「タイトル」「著者」「掲載誌」「掲載年」など 5 項目に関して検索を行える．これらを実装し動作実験を行い，動作を確認した．

このうちホテル検索システムを用いて評価実験を行った[13]．この段階での音声認識エンジンは Julian [12] を使用し，言語制約はフレーズ単位の文法の繰返しに，文末表現やフィラーを接続したものを用いた．ユーザ

表 2 GUI の有無におけるユーザ発話の分類
Table 2 Classification of user utterances with/without GUI.

	対話条件 1	対話条件 2
システムの想定内のフレーズ	231 (44.6%)	328 (76.5%)
語彙外のフレーズ	113 (21.8%)	49 (11.4%)
文法外のフレーズ	11 (2.1%)	5 (1.2%)
タスク外のフレーズ	163 (31.5%)	47 (11.0%)
合計	518	429

は音声対話システムを使用したことのない，主として工学部以外の学生 28 名（男性 19 名，女性 9 名）である．以下のように対話条件 1 と対話条件 2 を設定し，対話条件 1，対話条件 2 の順にシステムを使用してもらった．

対話条件 1 事前にシステムの検索対象である項目（所在・ホテルタイプ・料金・付帯施設）をマニュアルにより通知した．GUI（図 2）は使用しない．

対話条件 2 GUI（図 2）を使用して，キーフレーズパターンと認識結果をユーザに示す．

対話条件 1 では，ユーザ発話の認識結果，その認識結果から抽出したキーワード，検索結果の 3 つを，画面（文字端末）にテキストで表示した．ユーザは自分の希望する条件を考えたうえで，それを満たす結果が得られるまで 5〜6 分間を目安にシステムを使用した．得られたユーザ発話の中から，検索条件を指定するフレーズを，システムの想定内のフレーズと想定外のフレーズに分類した．ここでは検索条件を指定するフレーズのみを扱い，検索条件を削除するフレーズなどは含めていない．「語彙外のフレーズ」は語彙を追加すれば検索可能となるもの，「文法外のフレーズ」は語彙および文法規則を追加すれば検索可能となるもの，「タスク外のフレーズ」は語彙・文法を追加するだけでは検索可能とはならないものである．この定義に基づいて発話を分類すると**表 2** のような結果が得られた．

対話条件 1 と対話条件 2 を比較すると，対話条件 2 ではシステムの想定内のフレーズの割合が大きく増加している．対話条件 1 と対話条件 2 の差には，ユーザの慣れに起因するものも含まれているが，GUI の併用によりシステムの想定外のフレーズが減少する傾向が示されている．

5.2 複合的言語制約の効果

上記の対話条件 2 で収集したデータを用いて，複合的言語制約に基づくキーフレーズスポッティングの評価実験を行った．ここでは質問に対する肯定・否定や検索条件を削除する発話などもすべて含めて認識実験を行う．発話数は 665，含まれるキーフレーズ数は 797 である．音響モデルは，性別依存で学習された不特定話者向け triphone HMM [18] で，総状態数は 2,032，1 状態あたり 16 混合分布を持つ．

表 3 2-gram 学習用コーパスの仕様
Table 3 Specification of training corpora for 2-gram.

コーパス	テキストサイズ	語彙サイズ
ATR-SDB	136,051	3,430
ATR-SLDB	37,552	2,015
RWC	34,762	2,418
合計	208,365	5,432

比較対象として，文単位の記述文法に基づき得られた文候補を意味解釈する方式[21] と，言語制約をキーフレーズ・フィラーの連接のみとして 2-gram を与えずにキーフレーズスポッティングを行う方式[17] を用いる．文単位の記述文法では，文頭あるいは文末へのフィラー挿入を許している．

単語 2-gram 学習用の類似タスクドメインコーパスとして，ATR 自然発話音声データベース（ATR-SDB, ATR-SLDB）と RWC 音声対話データベース（RWC）より，客の発話部分を用いた．コーパスのサイズはのべ 21 万単語（＝形態素），語彙サイズは 5,432 単語である（**表 3**）．キーフレーズ内文法の語彙サイズは 712 単語であり，2-gram 学習用コーパスの語彙と組み合わせて，合計 6,124 単語の辞書が構成された．2-gram 学習時にはバックオフ平滑化を行っており，バックオフ係数の推定には Witten Bell ディスカウンティングを用いている．カットオフは行っていない．

対話音声理解を目的とした評価基準として，キーワード（＝スロット）の誤受理率（False Acceptance：FA）と理解誤り率（Slot Error：SErr）の和を用いる．

$$FA = \frac{\text{受理した中で誤っていたスロット数}}{\text{受理したスロット数}}$$

$$SErr = 1 - \frac{\text{受理した正解スロット数}}{\text{実際の正解スロット数}}$$

評価用サンプルは，以下の 3 タイプに分類した．ただし，ここでの語彙と文法は，文単位の記述文法におけるものである．

- 文法内発話：用意された語彙と文法に完全に従っている．
- 準文法内発話：助詞の省略，言い淀み，文頭末のフィラー挿入を含む．
- 文法外発話：未知語，文法外表現，文中のフィラー挿入がある．

ホテル検索における各タイプの発話例を**図 10** に示す．

表 4 種々の言語モデルによる意味理解誤り率の比較（FA+SErr）
Table 4 Comparison of various language models in speech understanding error rate (FA+SErr).

	語彙サイズ	文法内発話	準文法内発話	文法外発話	合計
正解数		561	116	120	797
文単位 記述文法	942	14.8%	51.4%	175.2%	45.8%
フレーズスポッティング （連接モデルのみ）	1,290	16.8%	34.2%	154.2%	37.9%
フレーズスポッティング （複合的制約）	6,124	**10.8%**	**24.7%**	**140.9%**	**30.3%**

文法内発話
所在が京都市の宿
ホテルタイプは旅館でお願いします
レストランとバーのあるホテル

準文法内発話
所在、京都市の宿
旅館、旅館で
えっとー、レストランとバーのあるホテル

文法外発話
所在が三条の宿　（「三条」が未知語）
旅館タイプでお願いします（「旅館タイプ」が文法外表現）
レストランと、えーと、バーのあるホテル

図 10 ホテル検索における発話例
Fig. 10 Example utterances at the hotel task.

従来方式との誤り率（FA+SErr☆）の比較を**表 4** に示す．文法内発話において，複合的言語制約を用いた提案手法が最高の理解率を示している．特に言語制約の緩いキーフレーズ・フィラーの連接モデルを用いたスポッティングによる方式[17] と比較して大きな向上が確認され，フィラー部分に対する 2-gram 適用の有効性が示された．連接モデルは，本タスクのようにキーフレーズ部分の語彙サイズが大きい場合には高い性能を得られない．また提案手法により，文単位の記述文法と比較しても高い理解率が得られている．

準文法内，文法外発話に対しても提案手法が最高の理解率を得ている．これらの発話では，文単位の記述文法に基づく方式に比べて，キーフレーズスポッティングに基づく方式が大幅に高い理解率を得ており，非定型な発話に対する頑健性が示された．図 10 に示されるような，文中にフィラーが挿入されている文法外発話は，従来の記述文法では受理不可能であるが，提案手法ではキーフレーズ単位でモデル化しているため，そのような場合でもキーフレーズを抽出できる☆☆．

また，提案手法とキーフレーズ・フィラー連接モデルによるスポッティングについて，誤りの内訳を**表 5** で比較した．提案手法により理解率は 7.6%向上したが，FA（誤受理率）の削減による向上が 6.7%を占めており，キーフレーズの沸き出し誤りの削減の効果が大きい．提案手法では，当該タスクドメインに合致していないコーパスを用いて語彙が大幅に増加しているにもかかわらず，定型・非定型の両方の発話に対して大きな改善が得られている．

表 5 複合的言語制約における 2-gram の効果
Table 5 Effect of 2-gram in the combined language model.

言語制約	FA	SErr	合計
連接モデルのみ	19.3%	18.6%	37.9%
複合的言語制約	12.6%	17.7%	30.3%

さらに，汎用性を検証するために，本手法をソフトウェアサポートサービスタスクへ適用した．ここでは主に，アプリケーション製品に関するユーザ発話の意図理解を目的としている．本タスクはデータベース検索タスクではないが，項目名とその値の組でキーフレーズが定義できるため，提案手法を適用できる．フレーズ内文法として，「アプリケーション名＋バージョン名＋付属語」や「名詞＋助詞＋動詞（サ変名詞）」などを定義した．フレーズ内の語彙サイズは 889 である．このタスクに対しても実装を行い，良好に動作することを確認した．

6. おわりに

本稿ではドメインに依存しないデータベース検索音声対話プラットフォームを提案し，その実装と評価について述べた．本プラットフォームは，アプリケーション開発者が定めた仕様に基づいて，データベースから語彙を自動抽出し，キーフレーズ部分に対する文法をテンプレートから生成する．キーフレーズ部分の文法と類似タスクドメインコーパスの 2-gram をともに用いて複合的言語制約とし，キーフレーズスポッティングを行う．また，典型的なキーフレーズパターンや認識結果を GUI を通じて明示することでユーザ発話を

☆ 置換誤りは FA と SErr の両方で計数されている．
☆☆ ただし，キーフレーズ中にフィラーが挿入されている場合は提案手法でも扱えない．

誘導する．評価実験の結果，複合的言語制約に基づくキーフレーズスポッティングにより，言語制約を文法で与えて認識する場合と比較して，意味理解誤り率が15.5%削減されることが確認された．

参 考 文 献

1) Kaspar, S. and Hoffmann, A.: Semi-Automated Incremental Prototyping of Spoken Dialog Systems, *Proc. ICSLP* (1998).
2) Kawahara, T., Lee, C.-H. and Juang, B.-H.: Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification, *IEEE Trans. Speech and Audio Processing*, Vol.6, No.6, pp.558–568 (1998).
3) Lamel, L., Rosset, S., Gauvain, J.-L. and Bennacef, S.: The LIMSI ARISE System for Train Travel Information, *IEEE Int'l Conf. Acoust., Speech & Signal Processing* (*ICASSP*) (1999).
4) San-Segundo, R., Pellom, B., Ward, W. and Pardo, J.: Confidence Measures for Dialogue Management in the CU Communicator System, *IEEE Int'l Conf. Acoust., Speech & Signal Processing* (*ICASSP*) (2000).
5) Sturm, J., den Os, E. and Boves, L.: Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System, *Proc. ESCA workshop on Interactive Dialogue in Multi-Modal Systems* (1999).
6) Sutton, S., Novick, D.G., Cole, R., Vermeulen, P., de Villiers, J., Schalkwyk, J. and Fanty, M.: Building 10,000 Spoken Dialogue Systems, *Proc. ICSLP* (1996).
7) 中野幹生，堂坂浩二，宮崎 昇，平沢純一，田本真詞，川森雅仁，杉山 聡，川端 豪：TV 番組の録画予約を受け付ける実時間音声対話システム，情報処理学会研究報告，98-SLP-22-8 (1998).
8) 小暮 悟，堀 賢史，中川聖一：音声対話システムのための未知語の登録を考慮した言語モデルの構築，情報処理学会研究報告，SLP-31-6 (2000).
9) 小暮 悟，伊藤敏彦，中川聖一：音声対話システムの移植性に関する考察—観光案内システムとデータベース検索システム，情報処理学会研究報告，SLP-25-3 (1999).
10) 小暮 悟，伊藤敏彦，廣瀬良文，甲斐充彦，中川聖一：CFG/bigram を使用した対話音声認識における意味理解の比較検討，第 57 回情報処理学会全国大会講演論文集，4R-4, pp.87–88 (1998).
11) 新田恒雄，神尾広幸，雨宮美香，松浦 博，内山ありさ，田村正文：マルチモーダル UI とラピッドプロトタイピング，情報処理学会研究報告，95-SLP-7-5 (1995).
12) 李 晃伸，河原達也，堂下修司：文法カテゴリ対制約を用いた A*探索に基づく大語彙連続音声認識パーザ，情報処理学会論文誌，Vol.40, No.4, pp.1491–1498 (1999).
13) 安達史博，駒谷和範，河原達也：音声対話情報検索システムにおける想定外の発話の分析とその対処，人工知能学会研究会資料，SIG-SLUD-A001-2 (2000).
14) 甲斐充彦，廣瀬良文，中川聖一：単語 N-gram 言語モデルを用いた音声認識システムにおける未知語・冗長語の処理，情報処理学会論文誌，Vol.40, No.4, pp.1383–1394 (1999).
15) 伊藤彰則，好田正紀：N-gram 出現回数の混合によるタスク適応の性能評価，電子情報通信学会論文誌，Vol.J83-D-II, No.11, pp.2418–2427 (2000).
16) 中川聖一，大谷耕嗣：Bigram の使用による話し言葉用確率文脈自由文法の自動学習，情報処理学会論文誌，Vol.39, No.3, pp.575–584 (1998).
17) 河原達也，石塚健太郎，堂下修司：発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化，情報処理学会論文誌，Vol.40, No.4, pp.1491–1498 (1999).
18) 河原達也，李 晃伸，小林哲則，武田一哉，峯松信明，嵯峨山茂樹，伊藤克亘，伊藤彰則，山本幹雄，山田 篤，宇津呂武仁，鹿野清宏：日本語ディクテーション基本ソフトウェア（99 年度版），日本音響学会論文誌，Vol.57, No.3, pp.210–214 (2001).
19) 河原達也，田中克明，堂下修司：音声言語を用いた仮想空間との対話による試着システム，情報処理学会論文誌，Vol.39, No.5, pp.1267–1274 (1998).
20) 竹林洋一：音声自由対話システム TOSBURG II—ユーザ中心のマルチモーダルインタフェースの実現に向けて，電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1417–1428 (1994).
21) 駒谷和範，河原達也：音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理，情報処理学会論文誌，Vol.43, No.10, pp.3078–3086 (2002).

(平成 14 年 5 月 28 日受付)
(平成 15 年 3 月 4 日採録)

駒谷 和範（正会員）

1998 年京都大学工学部情報工学科卒業．2000 年同大学院情報学研究科知能情報学専攻修士課程修了．2002 年同大学院博士後期課程修了．同年より京都大学情報学研究科助手．京都大学博士（情報学）．言語処理学会会員．

鹿島　博晶

1999 年京都大学工学部情報学科卒業．2001 年同大学院情報学研究科知能情報学専攻修士課程修了．現在，日本アイ・ビー・エム株式会社在籍．

田中　克明

1997 年京都大学工学部情報工学科卒業．1999 年同大学院工学研究科情報工学専攻修士課程修了．現在，ヤマハ株式会社在籍．

河原　達也（正会員）

1987 年京都大学工学部情報工学科卒業．1989 年同大学院修士課程修了．1990 年同博士後期課程退学．同年京都大学工学部助手．1995 年同助教授．1998 年同大学情報学研究科助教授．現在に至る．この間，1995 年から 96 年まで米国ベル研究所客員研究員．1998 年から ATR 客員研究員．1999 年から国立国語研究所非常勤研究員．2001 年から科学技術振興事業団さきがけ研究 21 研究者．音声認識・理解の研究に従事．京都大学博士（工学）．1997 年度日本音響学会粟屋賞受賞．2000 年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表．電子情報通信学会，日本音響学会，人工知能学会，言語処理学会，IEEE 各会員．