キーワードラティスのLR解析による自由発話理解

坪井宏之[†] 竹林洋一[††] 橋本秀樹[†††]

キーワードスポッティングに基づく自由発話の理解方式について述べる．この方式はtask-orientedな自由発話の理解を目的とし，新たに開発した時間的に離散なキーワードの系列を解析する拡張LRパーザにより，入力の自由発話から得られたキーワードラティスを解析し，入力発話の意味表現を生成する．解析処理は，キーワードスポッティングのイベントが発生するごとに駆動され，実時間対話システムに適している．文法には構文意味的な制約とともにキーワード間の時間的制約を記述し，キーワードラティスを効率良く始終端フリーに実時間で解析して意味表現を得ることができる．パーザは，検出キーワードの文始端の可能性，解析済の部分文候補との接続可能性，解析済の部分文候補の文可能性の判定をLRパーズ表に基づいて逐次行い解析を進める．キーワード数49単語のファーストフードに関する注文をタスクとする認識実験で，男性2名がそれぞれ発声した350文を評価した結果，3位までの累積文理解率は82.2%，単語認識率は90.7%が得られた．

Spontaneous Speech Understanding Method Based on LR Parsing of Keyword Lattice

HIROYUKI TSUBOI,[†] YOICHI TAKEBAYASHI[††]
and HIDEKI HASHIMOTO[†††]

This paper describes a task-oriented spontaneous speech understanding method, which extracts semantic content from the spotted keyword lattice using a newly developed generalized LR parser. Whenever a set of hypothesized keywords are spotted, the parser is driven by events to construct a semantic structure using a semantic keyword grammar. This efficient parsing enables real-time initial-state-free speech understanding for spontaneous speech. The parser comprises the following functional components: initial-state processing to check if a keyword can be an initial keyword, connection processing to check if a current keyword can connect with a hypothesized sub-sentence, accept processing to check if a hypothesized sub-sentence can be accepted as a sentence. An experiment was carried out on conversational 350 sentences from a 49-word vocabulary fast-food ordering task by 2 males. Accuracies of sentence understanding for top three candidates and word recognition for top candidate, were 82.2% and 90.7%, respectively.

1. はじめに

音声メディアの自然でユーザーフレンドリーな特長を活かした音声対話の研究が活発になってきている[1)~9)]．音声対話システムは，テキスト入力のような文字面の入力とは異なり，ユーザと計算機の間の意図，意志の伝達が目的であるため，単語音声や文音声の認識率よりも発話の意味内容の理解が重要となる．また，ヒューマンインタフェースの観点から，音声対話システムは発話の制約を極力少なくすることが望ましく，自由発話（spontaneous speech）の音声理解処理が必要になる．

隠れマルコフモデル（HMM）による高性能の大語彙連続音声認識[10),11)]が内外の研究機関で開発され，オフィスシステム[1)]や国際会議予約システム[2)]などについて高い認識性能が得られたことが報告されている．さらに，音声処理と言語処理との統合化の検討として，単語ラティスに対する拡張LR解析[12)]，島駆動による双方向のLR解析[13)]，フレーム同期構文解析[14)]，N-best候補の解析[15)]，談話など動的な制約を考慮した解析[16)]などがあり，音声認識結果の曖昧性を言語処理により効率良く解消するための検討がなされている．しかし，これらのシステムでは，実際の対話場面での自由発話への対処や耐雑音性能は必ずしも十

† 株式会社東芝関西研究所
Kansai Research Laboratory, Toshiba Corporation
†† 株式会社東芝研究開発センター
Research & Development Center, Toshiba Corporation
††† 東芝ソフトウエアエンジニアリング株式会社
Toshiba Software Engineering Co., Ltd.

分ではなく，発話の文型や形式に関する制約が強いという問題がある．

一方，自由発話を対象として，意味構文解析を用いたフレーズスポッティングによる航空旅行情報サービスシステム[17]や意味を考慮した2段LRパーザによる電話番号案内タスクの自由発話音声認識[18]，キーワード抽出と意味解析による音声対話システム[19]などの研究が報告されている．このような自由な発話を対象とするシステムでは，不要語，言い直し，省略，ポーズ，環境の雑音，対象外の単語などに対処する必要がある．たとえば，国際会議の問合せのタスクにおける自由発話の解析[20]では50%の発話に冗長な表現が含まれていると報告されている．しかし，これらの問題に対処する際に実時間で動作するシステムの構築に必要な自由発話理解方式の検討は十分にはなされていない．

これに対して，筆者らは応用分野を限定し，ユーザに何ら制約を設けないというコンセプトのもとに音声対話の研究を行っており，これまでに雑音免疫学習に基づく不要語を含む連続音声中からのキーワードスポッティング[21]を検討し，キーワードスポッティングを実時間で行うために必須のリアルタイム処理用のDSPアクセラレータを開発し[22]，ファーストフードの注文をタスクとした小規模な語彙による実時間音声対話システムを構築した[23]．本論文では，上述した音声対話システムに組み込んだユーザの一発話の理解方式について述べる．自由発話には，たとえば「冗長語」,「無意味語」などのさまざまな現象が含まれるが，本論文では自由発話を「冗長語，無意味語，言い淀み（以後，これらを不要語と呼ぶ）を含んだ発話」，キーワードを「自由発話を理解して意味表現を得るために必要なあらかじめ定めた特定の単語」，文を「自由発話中の構文的，意味的に妥当なキーワードの系列」と定義し，自由発話理解の実現のために，キーワードのみの文法的意味的関係と時間接続関係を用いてユーザの発話を実時間で理解するアプローチを検討した．また，本方式で対処できない問題，すなわち「文中の長いポーズ」,「言い直し」,「未知語」などには，対話により対処することとした．以下では，キーワードラティスに基づく構文意味解析のアプローチと，構文意味解析の方式を述べ，さらにファーストフードにおける注文をタスクとした場合のキーワード，意味表現，文法と具体的な解析の流れを示す．最後に，評価の結果を示し，本方式の有効性，問題点と今後の課題について述べる．

2. キーワードを基本とした音声理解

2.1 キーワードに基づく理解のアプローチ

従来の自由発話に対処する方式は，探索における評価を音声区間全体にわたって行うため，不要語やタスク対象外単語を認識語彙としたり，それらをGarbageモデルで表現するなどして音声区間全体に対する認識結果を求める必要があった．さらに，不要語やタスク対象単語を加えた文法を記述しなければならないため，文法が複雑となり，認識対象語彙の増加のために言語処理における探索空間の広がりが問題となっていた．

一方，自由発話の理解にキーワードスポッティングを利用するアプローチは，タスクに関連したキーワードを定義し，キーワードのみを認識対象とすることにより，不要語などを認識する必要がない．したがって，音声処理部に続く言語処理部では，文法に不要語などを記述する必要がなく，キーワード間のみの関係を記述すればよいため，文法が簡潔になり言語解析部の処理の高速化が期待できる．

このワードスポッティングを利用するアプローチでは，時間的に離散な単語ラティスを解析して，発話内容の理解を行う必要がある．単語ラティスを解析する方法としては，これまで，LRパーザを利用した解析方式が提案されているが，音声区間全体に対して単語の連結による文としての評価を行う方法であるため，不要語などの挿入は文法中に記述する必要があり，自由発話を対象とした場合には解析の探索空間が広がるという問題があった[12]．また，島駆動で解析する方法も提案されているが，島となる単語を決定するために単語ラティスが求まった後に解析を開始する必要があり，実時間処理に適した方式ではなかった[13]．

そこで，筆者らはこれらの問題に対処し，自由発話音声から発話内容を理解する方式としてワードスポッティングで得られた時間的に離散な単語ラティスから発話内容を理解する実時間処理向きの音声理解方式を提案する．

次節では理解部の入力となるキーワードのスポッティングの方式について説明する．

2.2 連続音声中からのワードスポッティング

不要語を含む連続音声中から単語をスポッティングして認識するために，単語の存在を仮定して単語特徴ベクトルと認識辞書の間で尤度を求め，連続的に始終端自由なパターン照合を行う．

ワードスポッティング処理を図1に示す．入力と単語lとの照合は，入力音声の分析フレーム周期ごとに単語終端候補点t_jを仮定し，あらかじめ定めた認識

対象単語 l の最長継続時間 dmax(l)，最短継続時間 dmin(l) から単語終端候補点 t_j に対する単語始端候補区間を求める．次に候補区間内のすべての始端候補点 $t_i, t_{i+1}, \cdots, t_{i+n}$ と，前述した終端候補点 t_j とで定まる音声パターンを線形にサンプルし複数個の固定次元の単語特徴ベクトルを抽出する．そして，固定次元の単語特徴ベクトルと認識辞書との複合類似度を求める．すべての認識対象単語について類似度演算を行った後に尤度に変換し，尤度がしきい値を超えた単語を検出結果とする．検出されたキーワードはキーワード名，始端時刻，終端時刻，尤度の4項目で表現され，**図2**に示すような時間的に離散なキーワードラティスの形式で理解部に順次入力されていく．

2.3 キーワードラティスに基づく音声理解

上述したキーワードスポッティング処理により順次出力される時間的に離散な単語ラティスを実時間で解析するために，バックトラックなしに一意に決定しながら解析を効率良く高速に行う拡張LR法を基本方式として採用し，キーワードが文を覆う割合を考慮した評価関数，キーワード間の時間関係，フレームを基本とした意味解析の3点について拡張した．

本方式ではキーワード系列の評価において，キーワードのスコアに加えて発声区間内でキーワード全体が占める割合を考慮するようにし，発声長の長いものも評価が高くなるようにした．また，不要語によりキーワードが時間的に離れた場合でも対処できるように，キーワード間の時間的な接続条件を判定し，文法に記述された範囲のキーワードとの接続を可能にした．さらに，文法に記述された構文情報と意味解析手続きに従って構文的な接続関係を判定するだけでなくフレームを基本とした意味的な関係を同時に解析できるようにし，探索空間の増加を抑えるようにした．文法では，キーワードのみの関係を記述することにより不要語の解析を行う必要がなくなり，不要語を文法に記述して全区間を解析する方式に比べて解析の処理量が抑えられる．また，始端となりうるキーワードが検出されるごとに，逐次に新たな解析スタックを生成するようにし，キーワードラティスの始終端フリーな解析を可能にした．

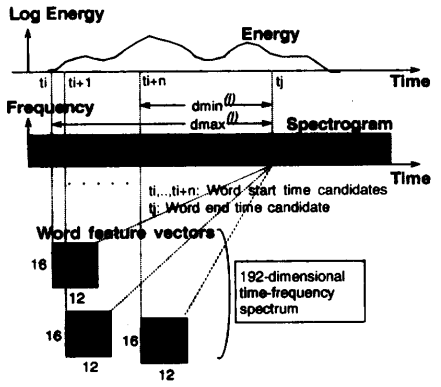
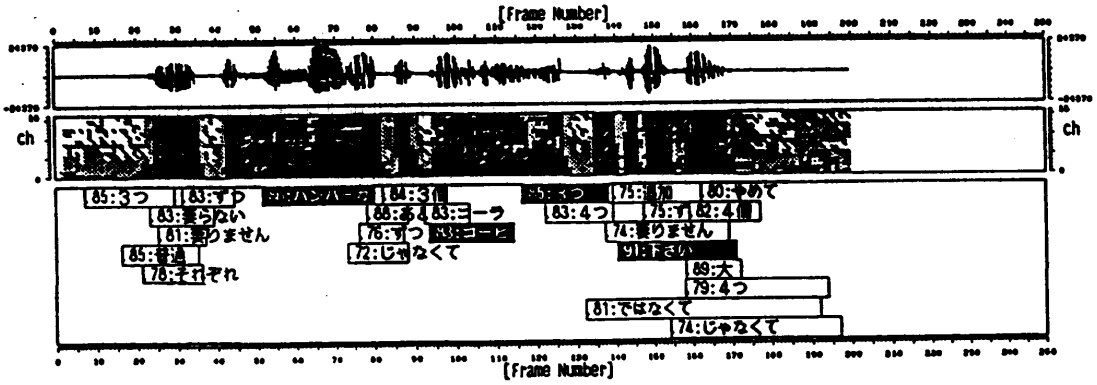


図1 ワードスポッティングによる単語検出
Fig. 1 Keyword-spotting process based on time-frequency word pattern.

3. キーワードラティスに基づく音声理解方式

3.1 理解処理

2章で述べた点を考慮したキーワードに基づく音声理解部の構成を図3に示す．キーワードラティスの構文意味解析は文始端判定，文候補解析，文終端判定からなり，キーワードをスポッティングするごとに解析を駆動して，文始端判定，文候補解析，文終端判定を順次行う．さらに，解析途中の文またはその一部分の



「えーと、ハンバーガーと、うんーと、コーヒー3つ下さい。」
("Eh... One hamburger and... uhm... three coffee, please.")

図2 ワードラティスの例
Fig. 2 An example of keyword lattice.

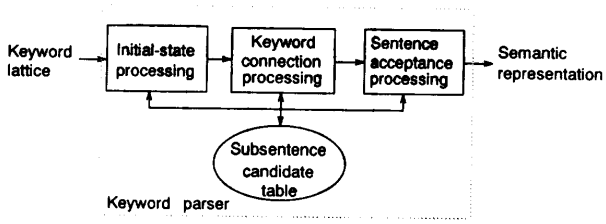


図3 音声理解部の構成

Fig. 3 Block diagram for speech understanding system using keyword-spotting.

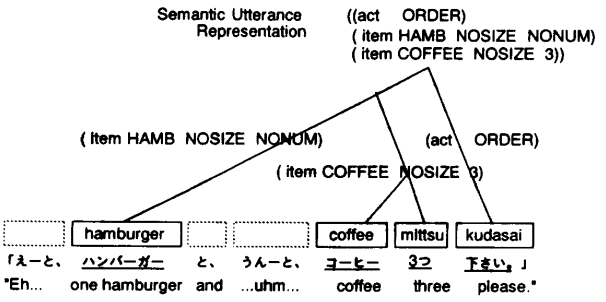


図4 意味表現と解析木の例

Fig. 4 An example of tree structure and semantic representation.

候補（以降，部分文候補と呼ぶ）を保持するための部分文候補バッファを持つ．

解析により得られる意味表現はフレーム形式であり，図4に「えーと、ハンバーガーと、うんーと、コーヒー3つ下さい.」(下線はキーワードを表す）の発生の意味表現と解析木の例を示す．ここで，actとは発声者の意図するアクションであり，この例では「下さい」の意味を表すORDER,「追加」の意味を表すAPPEND,「要らない」の意味を表すDELETEなどがある．itemは品目を表し，単語名，サイズ，個数の組からなる．意味表現は1つのactと複数のitemからなる．解析でアクション，サイズ，個数がない場合はNOACT, NOSIZE, NONUMと表記し，それぞれのキーワードが解析で得られなかったことを表現する．

解析に用いる文法と辞書を図5の例に示す．文法は，拡張文脈自由文法であり，文法の拡張項にフレーム生成mf()，スロット生成ms()，スロット追加as()などの意味解析手続きを記述して，構文解析と同時に意味処理を行う．また，辞書は単語名，終端記号，接続可能範囲からなる．*は文法の終端記号を表す．

解析法は拡張LR法[12]を基本としており，キーワードが文を覆う割合を考慮した評価関数，キーワード間の時間関係，フレームを基本とした意味解析の3点について拡張した．文法情報はパーザジェネレータによりLRパーズ表，構文意味解析処理手続き，接続判定処理手続きにあらかじめ変換される．図5の文法を変

Semantic keyword grammar

<S>	::= <NP> <VP>	{X0=mf(X2, X1)}
	| <NP>	{X0=mf(?, X1)}
<NP>	::= <NP> <ITEM>	{X0=as(X1, X2)}
	| <ITEM>	{X0=as(X1)}
<ITEM>	::= <FOOD> <NUM>	{X0=ms(item, X1, X2)}
	| <FOOD>	{X0=ms(item, X1, ?)}
	| <DRNK> <NUM>	{X0=ms(item, X1, X2)}
	| <DRNK>	{X0=ms(item, X1, ?)}
<VP>	::= *kudasai	{X0=ms(act, ORDER)}
	| *tsuika	{X0=ms(act, APPEND)}
	| *iranai	{X0=ms(act, DELETE)}
<FOOD>	::= *hamb	{X0=X1}
	| *cheeseb	{X0=X1}
	| *potato	{X0=X1}
<DRNK>	::= *coffee	{X0=X1}
	| *cola	{X0=X1}
<NUM>	::= *one	{X0=1}
	| *two	{X0=2}
	| *three	{X0=3}

Dictionary

hamburger	*hamb	S1 : E1
coffee	*coffee	S2 : E2
ikko	*one	S3 : E3
hitotsu	*one	S4 : E4
kudasai	*kudasai	S5 : E5
onegai	*kudasai	S6 : E6
:	:	:

(mf() : make frame, ms() : make slot, as() : append slot
? : Undefined term, X0 : Semantic representation of left term
X1,X2 : semantic representation of 1st and 2nd right term)

図5 意味キーワード文法と辞書の例

Fig. 5 An example of semantic keyword grammar and dictionary.

換したLRパーズ表の一部を図6に示す．LRパーズ表のSnは状態番号nにシフト操作すること，rmは文法規則mを利用してリデュース操作すること，accは文としてアクセプトされることを示す．

解析と同時に入力単語の尤度から部分文候補の尤度を評価し，ビームサーチによる部分候補の削減を行っている．部分文候補の尤度は次式により求める．

$$TLF = (1-\alpha)\left(\sum_{i}^{N} LF_i\right) \Big/ N + \alpha \left(\sum_{i}^{N} t_i\right) \Big/ T \quad (1)$$

ここで，TLF は文の尤度，N は文を構成するキーワードの個数，LF_i は第 i キーワードの尤度，T は文の継続時間，t_i は第 i キーワードの継続時間，α は係数を表す．第2項は文の区間をキーワードが覆う割合を表し，係数 α が0では尤度のみを評価するが，0でない場合は文の区間をキーワードが覆う割合が高い候補を高く評価する．

単語間の時間的な接続可能性は，単語の終端点と別の単語の始端点の時間的関係から判断するため，図7に示すように単語間の時間的な隔たりと重なりとを許す接続可能範囲Lを各単語ごとに設定している．図

	*kudasai	*tsuika	*iranai	*hamb	*cheeseb	*potato	*coffee	*cola	*one	*two	*three	$
0				s6	s7	s8	s9	s10				
1												acc
2	s13	s14	s15	s6	s7	s8	s9	s10				r2
3	r4	r4	r4	r4	r4	r4	r4	r4				r4
4	r9	r9	r9	r9	r9	r9	r9	r9	s17	s18	s19	r9
5	r11	r11	r11	r11	r11	r11	r11	r11	s17	s18	s19	r11
6	r12	r12	r12	r12	r12	r12	r12	r12	s12	s12	s12	r12
7	r13	r13	r13	r13	r13	r13	r13	r13	s13	s13	s13	r13
8	r14	r14	r14	r14	r14	r14	r14	r14	s14	s14	s14	r14
9	r15	r15	r15	r15	r15	r15	r15	r15	s15	s15	s15	r15
10	r16	r16	r16	r16	r16	r16	r16	r16	s16	s16	s16	r16
11												r1
12	r3	r3	r3	r3	r3	r3	r3	r3				r3
13												r5
14												r6
15												r7
16	r8	r8	r8	r8	r8	r8	r8	r8				r8
17	r17	r17	r17	r17	r17	r17	r17	r17				r17
18	r18	r18	r18	r18	r18	r18	r18	r18				r18
19	r19	r19	r19	r19	r19	r19	r19	r19				r19
20	r10	r10	r10	r10	r10	r10	r10	r10				r10

	S	NP	VP	ITEM	FOOD	DRINK	NUM
0	1	2		3	4	5	
1							
2			11	12	4	5	
3							
4							16
5							20
...							
20							

図6 LRテーブルの例
Fig. 6 An example of LR table.

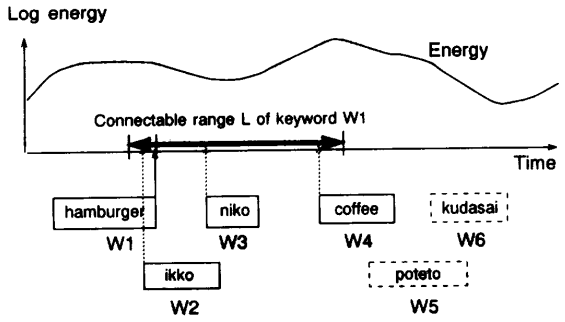


図7 接続可能範囲の適用例
Fig. 7 An example of keyword connection check.

単語名	終端記号	始端	終端
hamburger	*hamb	0	24
coffee	*coffee	0	67
ikko	*one	-7	44
hitotsu	*one	-5	37
kudasai	*kudasai	0	16
onegai	*kudasai	0	15

始端、終端は基準点からのフレーム数
(1フレームは8 msec.)

図8 接続可能範囲の例
Fig. 8 Examples of keyword connection range.

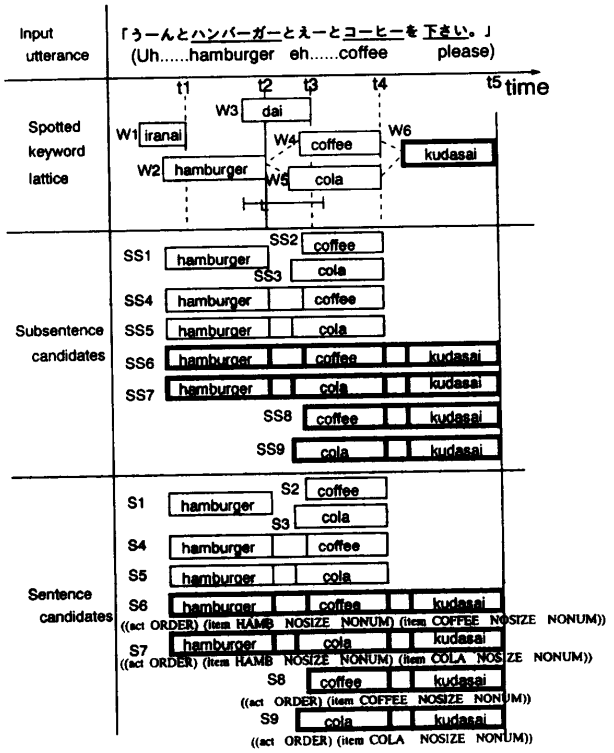


図9 解析処理過程の例
Fig. 9 Examples of parsing process.

7ではW1「ハンバーガー」の終端を基準として接続可能範囲Lに W2「1個」, W3「2個」, W4「コーヒー」の始端があり, それぞれがW1「ハンバーガー」に接続可能である. 接続可能範囲は, 単語ごとの接続可能範囲の最大範囲を用いている. 図8には接続可能範囲の例を示す. 始端, 終端の数字は基準点からのフレーム数 (1フレームは8 msec) で表している.

以下, 文始端判定, 文候補解析, 文終端判定の処理を図9の「うーんとハンバーガーとえーとコーヒーを下さい」(下線はキーワードを示す) の発声例に基づいて説明する. W1からW6はスポッティングされた単語名, SS1からSS9は部分文候補, S1からS9は文候補を表す.

3.2 文始端判定

文始端判定は, LRパーズ表に基づいて入力された単語が文の先頭となる単語であるか否かの判定を行う. 入力単語が文の先頭となりうる場合には意味表現を含む部分文候補を生成し部分文候補バッファに記憶する. LRパーズ表の状態番号0に動作が記述されている単語が文の先頭単語となりうる. 図9の例では, 時刻 t_1 においてW1「いらない」は先頭となりえないので, それを文頭とする新しい文候補は生成されないが, 時刻 t_4 において, W4「コーヒー」とW5「コーラ」は文頭となりうるので, それぞれを文頭とする部分文候補SS2とSS3を生成する.

3.3 文候補解析

文候補解析は, すでに解析された部分文候補と新たに入力された単語候補の組合せに対する解析処理を行う. すなわち, LRパーズ表を参照しながら, 部分文候補と単語の組合せに対して構文的, 意味的に接続可

能であるかの判定を行い，さらに，両者の時間接続関係を判定する．接続可能ならば入力単語と接続統合した部分文候補を新たに部分文候補バッファに加え，意味表現の生成，部分文候補の尤度の評価を行う．

図9ではW6「下さい」の始端点はW2「ハンバーガー」の接続可能範囲Lに入っていないので「ハンバーガー，下さい」の部分文候補は生成されないが，W4「コーヒー」，W5「コーラ」と接続可能であり部分文候補SS6からSS9までが生成される．

3.4 文終端判定

文終端判定は，部分文候補バッファの部分文候補のすべてに対して，各部分文候補が文として成立しているか否かをLRパーズ表を利用して判定する．つまり，文の終端を表す記号（$）を部分文候補に仮想的に接続し，文として受理されるかの判定を行う．図9の例では，部分文候補SS6，SS7，SS8，SS9についてW6「下さい」が文終端となりうるので，文候補S6，S7，S8，S9とその意味表現が出力される．このように，本解析方式により入力音声に対する文スポッティング処理，すなわちキーワード系列のスポッティングを行い，キーワードラティスの解析結果は文ラティスの形式で出力される．

4. 評価実験

連続音声中からのキーワード検出性能およびキーワードラティスを用いた音声理解の性能を評価するためにキーワード検出，連続文認識，連続文理解の実験を表1の音声資料を用いて行った．キーワードは表2に示すファーストフードの注文に関する49単語である．学習および評価用の連続音声の発話の文は同一であり，対話システムを想定して人間同士で実際に対話した発話文に基づいている．表3に発話文の例を示す．実験に用いた連続文のそれぞれの発声データは，話者が対話システムを想定しながら発声した音声を収集したものである．実験では計算機室で収集した雑音N1を音声に加え，表4に示す条件で分析した．実験のS/Nは音声区間内のエネルギと同区間の雑音のエネルギの対数比で定義した．

4.1 キーワード検出

キーワード辞書の訓練データは表1の孤立単語S1と連続文発声S2を用い，雑音免疫法により学習を行った．連続音声データは1文中に存在する単語ごとに単語区間の前後に30フレーム（240 msec）を付加して切り出し，雑音を重畳させて学習に用いた．不要語の付加に対してもワードスポッティング性能を向上させるため，孤立発声データに対しては不要語の付加と雑音の重畳の両方を行った[24]．付加した不要語は表1の音声資料N2で「えーと」，「あのー」，「ああ」，「いや」，「ん」の5種類である．不要語の終端と単語音声の始端の間隔は80 msecから240 msecまで40 msec刻みで変化させた．学習の際のS/Nは∞dB，20 dB，15 dBと段階的に低下させた．評価用データは連続文発声S3を用いた．なお，学習用音声データ中の各単語の基準となる始終端点は，孤立発声のデータについ

表1 音声資料とその用途
Table 1 Speech data set.

	発声内容	単語数	話者	用途
S1	孤立49単語	3855	男女計102名	学習
S2	連続350文	31842	男29名	学習
S3	連続350文	843	男3名	しきい値設定
S4	連続350文	1098	男5名	単語評価
S5	連続350文	2196	男2名	文評価
N1	計算機室雑音	-	-	音声に重畳
N2	不要語	250	男10名	音声に付加

表2 キーワードのリスト
Table 2 Examples of Japanese keywords.

ハンバーガー チーズバーガー フィッシュバーガー ポテト
フライドポテト コーヒー アイスコーヒー コーラ
オレンジ ジュース 1個 2個 3個 4個 5個
1つ（ひとつ） 2つ（ふたつ） 3つ（みっつ）
4つ（よっつ） 5つ（いつつ） はい ええ そうです
いーです いいえ 違います 違う 要らない 要りません
取消し あと それから それと 追加 ではなくて
じゃなくて やめて 下さい お願いします ちょーだい
ずつ 全部 それぞれ みんな 大（だい） 中（ちゅー）
小（しょー） 大きい（おーきい） 普通（ふつー）
小さい（ちーさい）

表3 連続音声の文例（下線部が認識対象単語）
Table 3 Examples of input speech.

・ハンバーガーを下さい
・コーラはやめて，んー，コーヒー
・いやいや，えー，コーヒーはよっつです
・チーズバーガーをひとつ下さい
・じゃ，コーヒーをみっつお願いします
・コーヒーとポテトを三個ずつ下さい
・えーと，あとポテト，んーと，ふたつ下さい

表4 分析条件
Table 4 Experimental conditions.

サンプリング	12 kHz
FFTフレーム長	24 msec
フレーム周期	8 msec
フィルタ分析	16チャネル
単語特徴ベクトル次元数	192次元（16ch × 12フレーム）

表5 単語の検出性能
Table 5 Results for keyword-spotting.

S/N	∞	20
正解	1076	1044
脱落	22	54
付加	9634	8894
FA/H/W	38.7	35.7
検出率（%）	98.0	95.1
1位検出率（%）	92.2	85.3

ては，自動検出により求め，連続発声のデータについては視察により求めた．

認識結果を表5に示す．ここでは，入力単語と同一の単語が検出され，視察で求めた単語区間と検出された区間とが70%以上オーバーラップした場合を正解，入力単語が検出されなかった場合を脱落，上記以外の場合を付加誤りとした．また，正解である率を検出率，正解であり順位が1位である率を1位検出率とした．単語検出のためのしきい値は辞書作成データおよび評価用データとは別の表1の音声資料S3を用いて，脱落誤りと付加誤りに20:1の重みを付けて単語ごとに誤りを最小化するように定めた．付加誤りは発声音声の総時間長およびキーワード数に依存するため，FA（False Alarms: 付加誤り数）ではなく，FA/H/W（False Alarms/Hour/Word: 単位時間に発生する1単語あたりの付加誤りの数）を示す．評価データS5の総発声時間は489.2秒であった．

FA/H/Wが35.7～38.7の幅であり，雑音や音韻環境の影響による変化は少なく，雑音免疫学習法による効果が現れている．また，雑音による1位検出率の低下に比べて検出率の低下は少なく，キーワードラティスとして，しきい値以上の複数の候補を扱うことが有効であることを示している．

4.2 連続文認識

評価用連続文発声の正しいキーワード列と構文・意味解析を行って得たキーワード列が完全に一致する率を文認識率として評価した．たとえば，属性が省略されている品名が入力された場合には認識結果として同様に省略されている場合のみを正解とする．また，1位のキーワード列中の各単語が正解の単語である率を単語認識率として評価した．認識辞書の訓練データは実験1と同じものである．評価用データは表1のS5を用いた．認識に用いた文法は81ルールからなる．文法の作成は実験に用いた発声文に基づいて作成した．この文法の静的平均分岐数は9.5である．キーワードの接続可能範囲はしきい値設定用データS3を用いて各キーワードに接続するキーワードの最小，最大範囲を求め設定した．処理において保持する部分文候補のビーム幅は単語検出実験の累積検出率結果に基づき20で行った．式(1)の有効性を検討するために，S/N20 dBの評価用データに対して式(1)の係数αを0.0から0.3まで0.1きざみで変化させた1位文認識率は，27.1%，41.0%，35.0%，26.1%となった．実験結果を解析したところ，αが0.0の場合に受理されたキーワード列は正しいキーワード列に比較して短い傾向にあった．これは，キーワードの尤度だけを評価すると高い尤度を持つ単語の短いキーワード列に文として高い評価値を与えるためである．評価に部分候補区間を覆う割合を加えることにより，尤度と発声時間長を考慮した評価が可能となり文認識率が向上したといえる．この結果に基づき，以降の実験では係数αを0.1に固定して実験を行った．

表6 連続文認識率（%）と単語認識率
Table 6 Results for sentence and word recognition.

S/N	累積順位 1	累積順位 2	累積順位 3	単語認識率
∞	56.0	70.8	76.9	90.7
20	41.0	55.1	61.8	81.4

S/Nが∞，20 dBのときの1位から3位までの累積文認識率，単語認識率の結果を表6に示す．誤った場合を調べると，たとえば，「アイスコーヒー、もお3つ追加」（アイスコーヒー，3つ，追加）の入力に対して（はい，3つ）の文が，「それと、フライドポテトは要りません」（それと，フライドポテト，要りません）の入力に対して（それと，ポテト，要りません）の文が1位となるような場合が多かった．これは，アイスコーヒーの「アイ」と「はい」や「フライドポテト」と「ポテト」など入力音声中の正解キーワードの一部分に継続時間の短い単語が一致して高い尤度を得ているためである．

また，S/Nが低い場合の文認識率の低下の原因は，発声の短い単語の検出率が低いことであった．また，キーワードの接続可能範囲の設定の有効性を検討するために異なる接続可能範囲の文認識率を求めた．評価用データは表1のS3を用い，キーワードの接続可能範囲を評価用データS3の各キーワードに接続するキーワードの最小，最大範囲を設定した場合とすべてのキーワードの最小，最大範囲を−10，+40に固定した場合について評価した．固定した条件は，最小，最大範囲の最大になるようにしたものであり，接続可能範囲を大幅に緩めた場合に相当する．その他の実験条件は同じである．実験の結果，キーワードごとの最大，最小範囲の場合の文認識率は，44.1%であり，固定した場合には31.0%であった．この結果からキーワード

表7 連続文理解率（%）
Table 7 Results for sentence understanding.

S/N	累積順位		
	1	2	3
∞	64.6	77.8	82.2
20	57.0	67.3	73.2

ごとに接続可能範囲を設定することにより，不要語が含まれる自由発話中のキーワード間の時間的な関係を考慮した接続判定が有効であると考えられる．

4.3 連続文理解

評価用連続文発声の意味表現と構文・意味解析を行って得た意味表現が一致する率を文理解率として評価した．ここでいう「意味の一致」とは，「下さい」，「お願いします」のように異なる単語でも同じ意味の単語は正解であるとし，さらに品名のサイズ，個数の省略は，デフォルト値と同一と解釈し，脱落があった場合でも実際の発声がデフォルト値と同一であれば正解とすると定義した．

キーワード辞書の訓練データ，評価用データおよび文法は連続文認識実験と同じものである．また，接続可能範囲，ビーム幅の設定，部分文候補評価も連続文認識実験と同じである．

結果を表7に示す．同じ意味の候補を正しいとすることにより理解率は認識率に比較して向上しており，対話システムでの性能を示している．さらに，4.2節で述べたと同様に，実際の対話システムでは，品名のサイズ，個数に加えてactも対話の文脈を利用して正しく理解することが可能となるので，より高い理解率が音声対話システムで得られると考えられる．

4.4 実時間対話システムの音声理解

実験で用いたタスクは語彙数が49単語，静的平均分岐数が9.5であり，小規模なものである．一方，応用場面を意識したtask-orientedなシステムには，タスクが限定された小規模なシステムでも利用者が自由に発話できることが必要である．特に，不特定のユーザが利用するシステムの場合には，システムの規模にかかわらずシステムに精通していない不慣れなユーザを考える必要がある．そのため，対話管理をシステム主導ではなく，ユーザー主導で行い，ユーザが何を話しても対話を自然に進行するという指針のもとに対話システムの構築する必要がある．実験の検討から，小規模なタスクの自由発話理解が可能であることを示した．自由発話理解の大語彙のタスクへの拡張は，丁寧な発声を対象にする場合に比較してキーワード検出の付加，脱落の増加による誤りが多くなる．そのため，対話システムの対話管理機能によって認識対象語彙が小語彙あるいは中語彙になるように対話を進行し，語彙数を限定する必要がある．また，意味表現のactionに相当するキーワードの検出は，対話管理の点から重要であり誤りが少ないことが必要である．そのため，キーワードを基本とした自由発声の理解方式を利用するシステムは，登録，指示など利用目的が特定されたタスクへの利用に適しているといえる．

本論文で述べた解析方式の処理量は語彙数さらに文法の複雑さの関数となるが，実際の解析処理においては，処理量を削減するために，保持する部分文候補のビーム幅を設定している．また，2.3節で述べたように，解析処理はキーワード検出イベントが発生するごとに行う．そのため，処理量は，語彙数や文法規模にかかわりなく，部分文候補のビーム幅とキーワード検出イベントの発生頻度に比例する．

実時間で処理を実現するためには，平均キーワード検出イベント発生時間内に検出キーワードとビーム幅数の部分文候補との解析が終了すればよい．実験では平均キーワード検出イベント発生時間は15 msecであり，AS3160（Sun3と同等）のワークステーションによる1キーワード検出イベントに対する解析処理は7 msecであった．したがって，語彙数や文法規模が増加しても，50～100語のキーワードをセットとしたタスクであれば，実時間で動作するといえる．本方式では，キーワード間の文法，時間関係，文の評価値に基づいて解析した複数の意味表現を出力することができるため，対話制御部は，音声理解部の複数の意味表現候補から対話の文脈や履歴情報をもとに適切な意味表現を解釈，選択することが可能となり，文理解率の向上を図ることができる．

5. まとめ

本報告では，task-orientedな対話システム実現のためのキーワードラティスに基づく連続音声理解の方式について述べ，49語のキーワードからなるファーストフードの注文のタスクを想定した自由発話の認識理解実験の結果を示し，その有効性を示した．

本方式はキーワードの検出ごとに構文意味解析を進めるため，リアルタイム処理に適している．上記のタスクのリアルタイム連続音声理解が4枚のDSPアクセラレータと1台の汎用ワークステーションにより実現し，音声対話システムTOSBURG（Task-Oriented System Based on Speech Understanding and Response Generation）を構築した[23]．

今後は，キーワードの継続時間長を考慮したスコアリング方式が課題である．

謝辞 本論文のワードスポッティングと雑音免疫学習に関して協力いただいた東芝関西研究所金沢博史氏と，日頃ご討論いただく東芝関西研究所の諸氏に感謝いたします．

参考文献

1) Rudnicky, A.I., Lunati, J.M. and Franz A.M.: Spoken Language Recognition in an Office Management Domain, *Proc. ICASSP '91*, pp.829–832 (1991).
2) 竹沢寿幸，大倉計美，森元 逞，嵯峨山茂樹，樽松 明：日英音声言語翻訳実験システム SL-TRANS2，音講論，1-5-24，pp.47–48 (1991).
3) 山本哲也，尾崎 弘，堀 雅洋，溝口理一郎：音声理解システム SPURT-I のための対話管理機構，音講論，3-1-13，pp.101–102 (1989).
4) 新美康永，小林 豊：音声対話システムにおける文脈解析とその単語予測への応用，第16回東北大応情研シンポジウム予稿集，pp.151–158 (1990).
5) 山岡孝行，飯田 仁：文脈を考慮した音声認識結果絞りこみ手法，情処研報，91-NL-85-4, pp.121–128 (1991).
6) 村山秀樹，天野明雄，北原義典，市川 熹：連続音声入力による情報検索システム，音講論，3-6-11，pp.131–132 (1991).
7) 畑崎香一郎，渡辺隆夫，磯谷亮輔，塚田 聡，野口 淳，坂井信輔，古賀真二，吉田和永：半音節を認識単位とする不特定話者連続音声認識システム，信学技報，SP90-83, pp.45–52 (1990).
8) 大黒慶久，橋本泰秀，中川聖一：構文解析駆動型日本語音声理解システム，信学技報，SP88-87, pp.55–62 (1988).
9) 伊藤克亘，速水 悟，田中和世：拡張LR構文解析法を用いた連続音声認識，信学技報，SP90-74, pp.49–56 (1990).
10) Lee, K.: *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston (1989).
11) 北 研二，川端 豪，斎藤博昭：HMM 音韻認識と予測LRパーザを用いた文節認識，信学技報，SP88-88, pp.63–69 (1988).
12) Tomita, M.: An Efficient Word Lattice Parsing Algorithm for Continuous Speech Recognition, *Proc. ICASSP '86*, pp.1569–1572 (1986).
13) Saito, H.: Bi-directional LR Parsing from an Anchor Word for Speech Recognition, *Proc. COLING-90*, pp.237–242 (1990).
14) 中川聖一：文脈自由文法のフレーム同期型構文解析法による連続音声認識，信学論 D, Vol.J70-D, No.5, pp.907–916 (1987).
15) Schwartz, R., et al: The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses, *Proc. ICASSP '90*, S2.12, pp.81–84 (1990).
16) 岡田美智男：文脈自由な句構造文法からのミニマル・サブネットワークの生成について，信学技報，SP91-27, pp.73–80 (1991).
17) Ward, W.: Understanding Spontaneous Speech: The Phoenix System, *Proc. ICASSP '91*, pp.365–367 (1991).
18) 南 泰浩，山田智一，吉岡 理，鹿野清宏：自由発声音声認識における意味を考慮した2段LRパーザ，音講論，3-4-10, pp.69–70 (1993).
19) 荒木雅弘，河原達也，西田豊明，堂下修司：キーワード抽出に基づく意味解析による音声対話システム，信学技報，SP91-94, pp.25–32 (1991).
20) 村上仁一，嵯峨山茂樹：自由発話の音声認識における問題点の検討，音講論，2-P-26, pp.189–190 (1991).
21) 金沢博史，坪井宏之，竹林洋一：雑音下の連続音声中からのキーワード検出，信学論，J76-D-II, No.3, pp.427–435 (1993).
22) 坪井宏之，金沢博史，竹林洋一：音声認識研究用リアルタイム処理システムの開発，信学技報，SP90-37 (1990).
23) Takebayashi, Y., Tsuboi H., Kanazawa H., Sadamoto Y., Hashimoto H. and Shinchi H.: A Real-time Speech Dialogue System Using Spontaneous Speech Understanding, *IEICE Trans. Inf. & Syst.*, Vol E76-D, No. 1, pp.112–120 (1993).
24) 金沢博史，坪井宏之，竹林洋一：騒音環境下でのワードスポッティングによる音声認識における不要語の影響，音講論，2-8-8，pp.61–62 (1990).

(平成6年 5 月 27 日受付)
(平成8年 12 月 5 日採録)

坪井 宏之（正会員）

昭和51年名古屋工業大学工学部電気学科卒業．同年東北大大学院工学研究科博士課程前期了．同年(株)東芝入社．以来，音声認識理解，音声対話の研究に従事．昭和62～平成元年(株)日本電子化辞書研究所出向．現在，同社研究開発センター関西研究所主任研究員．平成5年日本音響学会技術開発賞受賞．電子情報通信学会，日本音響学会，人工知能学会，ASA 各会員．

竹林 洋一（正会員）

昭和49年慶応義塾大学工学部電気学科卒業．昭和55年東北大大学院工学研究科博士課程了．同年(株)東芝入社．以来，音声情報処理，知的インタフェースの研究に従事．現在，同社研究開発センター情報・通信システム研究所ヒューマンインタフェース技術センター長．昭和60～62年MITメディア研究所客員研究員．平成4～5年(株)日本電子化辞書研究所第5研究室室長．工学博士．平成4年人工知能学会全国大会優秀論文賞，平成5年日本音響学会技術開発賞，平成6年情報処理学会山下記念研究賞受賞．電子情報通信学会，日本音響学会，人工知能学会各会員．

橋本 秀樹（正会員）

昭和62年新潟大学経済学部経済学科卒業．同年東芝ソフトウエアエンジニアリング(株)入社．以来，音声処理システム，自然言語処理システムの研究開発に従事．平成5年日本音響学会技術開発賞受賞．