音声言語情報処理 14－16
(1996. 12. 13)

構文規則と前終端記号バイグラムを併用する対話音声認識手法の高速化と高性能化

竹澤 寿幸　　森元 逞

ATR 音声翻訳通信研究所

〒619-02 京都府相楽郡精華町光台 2-2
電話番号 (代表): (0774) 95 1301
E-mail: {takezawa, morimoto}@itl.atr.co.jp

あらまし

文脈自由文法形式の統語的な制約を用いて，部分木系列をスコア付きの仮説として出力する，音声パーザの検討を行なっている．自然な発話を扱うために，文法は部分木を単位として記述する．品詞を細分化したものに相当する前終端記号のバイグラムを構文規則と併用することで高い性能が得られることを，ポーズで区切られた区間 (ポーズ単位) を対象とする対話音声認識実験により既に確認している．辞書引きの実装方法とビーム探索の手法を改善することにより，高速化と高性能化が達成できたので，その内容を報告する．さらに，複数のポーズ単位からなる発話を対象に効率的な対話音声認識を行なう手法を検討する．

キーワード

連続音声認識，音声言語処理，音声言語統合処理，構文規則，部分木，統計的言語モデル．

Performance Improvement of Dialogue Speech Recognition Method Using Syntactic Rules and Preterminal Bigrams

Toshiyuki Takezawa, Tsuyoshi Morimoto

ATR Interpreting Telecommunications Research Laboratories

2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
Telephone (International): +81 774 95 1301
Telephone (Domestic): (0774) 95 1301
E-mail: {takezawa, morimoto}@itl.atr.co.jp

Abstract

We are studying a CFG-based speech parser that outputs scored subtree sequences. In order to deal with spontaneous speech, the syntactic rules are written to parse subtrees rather than sentences. Bigrams of preterminal symbols are used during beam search. We previously conducted pause unit-based experiments and have confirmed that our method yields good performance. In this paper, we report on two kinds of performance improvement; i.e., the dictionary access improvement and the improvement of beam search. We also discuss ways to deal with more than one pause unit.

key words

Continuous speech recognition, spoken language processing, integrated processing of speech and language, syntactic rules, partial trees, statistical language modeling.

1 まえがき

文脈自由文法形式の統語的な制約を用いて，部分木系列をスコア付きの仮説として出力する，音声パーザの検討を行なっている．自然な発話を扱うために，文法は部分木を単位として記述する [8, 9]．品詞を細分化したものに相当する前終端記号のバイグラムを構文規則と併用することで高い性能が得られることを，ポーズで区切られた区間 (ポーズ単位) を対象とする対話音声認識実験により既に確認している [9]．

「それでは，鈴木和子様」という発話があった場合，仮に「それでは」と「鈴木和子様」の二つの文節に分けたとしても，断片的な発話なので，文としての構造を持っているとは必ずしも言えない．このような背景から，部分的な構造を表現することが必要となり，我々はそれを部分木と名付けた．

このアプローチの考え方は，まず部分木に基づく文法を採用することで，文法の被覆率を高める．そして，その部分木に基づく文法と前終端記号のバイグラムを併用することで，文法単独やバイグラム単独より，さらに性能が上げられるという主張である．なお，前終端記号バイグラムは語[1]のバイグラムより移植性がよいという長所がある．

さらに，この音声認識部から出力される構造を，音声翻訳や音声対話システム [3] の言語処理部で利用することにより，全体として効率的な音声言語統合処理の実現を目指している．

辞書引きの実装方法とビーム探索の手法を改善することにより，高速化と高性能化が達成できたので，その内容を報告する．さらに，複数のポーズ単位からなる発話を対象に効率的な対話音声認識を行なう手法を検討する．

第 2 章で高速化と高性能化のための改善点について述べる．**第 3 章**で提案手法による対話音声認識実験の結果を示し，その効果を論ずる．**第 4 章**で複数のポーズ単位からなる発話を対象とする対話音声認識手法について検討する．最後に全体をまとめる．

2 高速化と高性能化のための改善点

前終端記号バイグラムの評価を予測的に行なうために，辞書引きの実装方法を変更した．また，ビーム探索の枝刈りの条件と，スコアの評価式を改良した．

2.1 辞書引きの実装方法

2.1.1 目的

文献 [9] の手法では，前終端記号バイグラムを利用する場合の言語スコアは予測された音素系列の文法履歴から

(<前終端記号> -> 終端記号)

という形式の構文規則を取り出して計算していた．つまり，予測された音素系列の中で確定した語についての言語スコアを計算していた．

前終端記号バイグラムの評価を語候補が確定する前に行なうほうが効率的な探索が実現できると期待できる．しかしながら，語彙項目と構文規則を一緒にして LR 構文解析表を作成してしまうと，前終端記号バイグラムを予測的に評価する探索を実現しにくい．そこで，語彙項目のみからなる LR 構文解析表と構文規則のみからなる LR 構文解析表の二つに分離する方針とした．

(1)	<S>	→	<PP> <S>
(2)	<S>	→	<V>
(3)	<PP>	→	<N> <P>
(4)	<PP>	→	<S> <P>
(5)	<V>	→	k i t a
(6)	<V>	→	t u t a w a q t a
(7)	<N>	→	k i t a
(8)	<N>	→	b u ng k a
(9)	<P>	→	k a r a
(10)	<P>	→	g a

図 1: 文法規則の記述例

(1)	<S>	→	<PP> <S>
(2)	<S>	→	V
(3)	<PP>	→	N P
(4)	<PP>	→	<S> P

図 3: 前終端記号までの文法規則

2.1.2 例

例を使って説明する．**図 1**に簡単な文法の記述例を示す．これから LR 構文解析表を作成すると**図 2**のようなものが得られる．この表には前終端記号の情報は含まれていないので，何らかの方法で元の構文規則を参照しなければならない．

そこで，**図 1**の文法を，**図 3**のような前終端記号までの文法規則と，**図 4**のような語彙規則に分離する．**図 3**の文法規則では元の文法の前終端記号が終端記号となっている．**図 3**から LR 構文解析表を作成すると，**図 5**が得られる．先読み可能な記号が前終端記号なので，次につながる可能性のある前終端記号を簡単に予測することができる．つまり，音声認識過程で前終端記号バイグラムの評価を予測的に活用することができる．

さらに，**図 4**の語彙規則に対して，北らの提案するカテゴリ到達可能性検査を使って LR 構文解析表を作ると，**図 6**のようなものが得られる[2]．シフト動作のところに到達可能なカテゴリ (元の文法の前終端記号) の情

[1] 本稿では，単語や形態素を語と呼ぶ．

	b	k	t	u	g	i	ng	a	r	w	q	$	<N>	<PP>	<V>	<P>	<S>
0	s1	s4	s5										g2	g3	g6		g7
1				s8													
2		s9			s10											g11	
3	s1	s4	s5										g2	g3	g6		g12
4						s13											
5				s14													
6		r2			r2							r2					
7		s9			s10							acc				g15	
8							s16										
9								s17									
10								s18									
11	r3	r3	r3														
12		s9,r1			s10,r1							r1				g15	
13			s19														
14			s20														
15	r4	r4	r4														
16		s21															
17									s22								
18	r10	r10	r10														
19								s23									
20								s24									
21								s25									
22								s26									
23		r7,r5			r7,r5							r5					
24										s27							
25		r8			r8												
26	r9	r9	r9														
27								s28									
28											s29						
29			s30														
30								s31									
31		r6			r6							r6					

図 2: LR 構文解析表の例

(1)	<preterm>	→	<V>
(2)	<preterm>	→	<N>
(3)	<preterm>	→	<P>
(4)	<V>	→	k i t a
(5)	<V>	→	t u t a w a q t a
(6)	<N>	→	k i t a
(7)	<N>	→	b u ng k a
(8)	<P>	→	k a r a
(9)	<P>	→	g a

図 4: 語彙規則

	N	V	P	$	<PP>	<S>
0	s1	s3			g2	g4
1			s5			
2	s1	s3			g2	g6
3			r2	r2		
4			s7	acc		
5	r3	r3				
6			s7,r1	r1		
7	r4	r4				

図 5: 前終端記号を終端記号とする LR 構文解析表

報が埋め込まれているため，不必要な音素照合を削減することができる．

2.1.3 関連研究との比較

伊藤らは未登録語を語レベルで扱うために，語彙項目の LR 構文解析表と構文規則の LR 構文解析表に分離している [2]．語彙項目の LR 構文解析表は構文カテゴリ (前終端記号) 毎に複数用意している．我々の目的は前終端記号バイグラムを予測的に評価することである．また，我々の実装方法では，語彙項目の LR 構文解析表は一つである．語彙項目のみからなる LR 構文解析表と構文規則のみからなる LR 構文解析表をつなぐ際に不必要な予測が起こらないように，北らの提案するカテゴリ到達可能性検査 [5] を導入した．

文節間文法と文節内文法に分離して利用する目的で，北らはカテゴリ到達可能性検査を提案している [5]．我々の文法は文節を特別扱いせず，部分木を単位として文法を記述している．我々が文法を 2 段階に分けたのは，前終端記号のレベルであり，目的は前終端記号バイグラムを予測的に評価するためである．

我々の実装方法の利点を要約する．

- 語候補が確定する前に前終端記号バイグラムを評価することが簡単にできる．
- 人名などの新語登録を簡便に実現することができる．
- 未登録語の扱いも語レベルで行なえる [2].

2.2 ビーム探索

文献 [9] のビーム探索手法では，一定の個数を残して枝刈りする方針を採用していた．関連研究 ([2] 等) で閾値による枝刈りが有効であると報告されているので，閾値による枝刈りを導入した．いくつかの条件下で実験を試みたところ，閾値による枝刈り方式のほうが性能がよかったので，次章では閾値による枝刈り方式の結果のみを示す．

3 対話音声認識実験

提案手法の効果を確認するために，ポーズ単位の対話音声認識実験をいろいろな条件の下で行なった．

3.1 実験条件

ATR で収集中の旅行会話データベース [6] から選んだ対話音声を対象に実験を行なった．ポーズの自動検出を行なって分割した音声区間を認識対象とした．対数パワーとゼロ交差数の二つの特徴量を用い，300 ms より長いものを選べば促音と区別してポーズを検出できた[3]．音素モデルとしては，音素バランス 50 文により VFS 法で話者適応を行なったモデル (状態数 401，混合数 5) [11] を利用した．音声の分析フレーム長は 10 ms とした．音声認識の探索手法はフレーム同期方式を採用した．なお，実験に利用したマシンは HP9000/735 である．

文法の諸元を**表 1**に示す．小さい文法は大きい文法の部分集合となっている．旅行会話データベースからテストセットとは異なる 50 対話 (1959 文) を選び，前終端記号のバイグラムを求め，削除補間法により平滑化したところ，前終端記号のみによるテストセットに対する語パープレキシティは 29.2 であった．**表 1**を見ると，いずれの文法の場合であっても，併用時の語パープレキシティのほうが，元の文法のみの値，前終端記号のみの値いずれと比べても小さいことがわかる．

3.2 評価尺度の検討

かな漢字文字列に変換した表記により，正解ラベルと音声認識候補の間でどの程度一致しているかを評価した．ポーズ単位認識率は，ポーズ単位全体が正解ラベルとすべて一致したものの全体に対する割合である．部分的に正解が含まれることがあるため，語認識率も求めた．語認識率は正解ラベルに対して音声認識候補の語が一致している割合を DP マッチングにより求めた．上位候補

[2] 表の一部は省略している．

[3] 今回実験に用いた対話音声データに限る．我々の集めている旅行会話データベース全体の特徴という主張ではない．

	t	b	k	g	a	...	$	<V>	<N>	<P>	<preterm>
0	<s1{V}>	<s3{N}>	<s5{P,N,V}>	<s6{P}>		...		g2	g4	g7	g8
1						...					
2						...	r1				
3						...					
4						...	r2				
5					<s12{P}>	...					
6					<s13{P}>	...					
7						...	r3				
8						...	acc				
9	<s14{V}>					...					
10						...					
11	<s16{N,V}>					...					
12						...					
13						...	r9				
14					<s18{V}>	...					
15			<s19{N}>			...					
16					<s20{N,V}>	...					
17					<s21{P}>	...					
18						...					
19					<s23{N}>	...					
20						...	r6,r4				
21						...	r8				
22					<s24{V}>	...					
23						...	r7				
24						...					
25	<s26{V}					...					
26					<s27{V}>	...					
27						...	r5				

図 6: 到達可能なカテゴリ情報のついた LR 構文解析語彙辞書の例

表 1: 文法の諸元

文法名	語数	規則数	前終端記号数	語パープレキシティ	
				文法のみ	前終端記号バイグラム併用時
2S	317	1395	184	18.6	10.4
2M	561	1567	247	39.1	22.2
2L	1010	1809	291	71.2	25.9

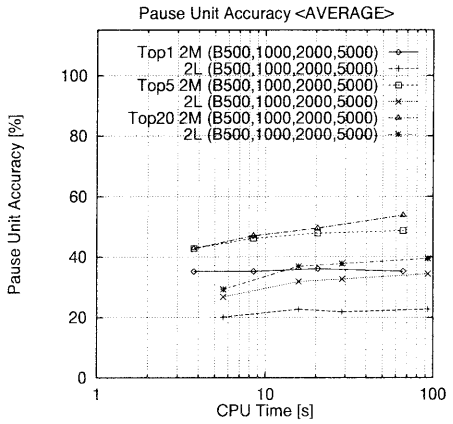

図 7: 従来手法の実装方式による対話音声認識実験結果．個数によるビーム探索方式．ポーズ単位認識率による評価

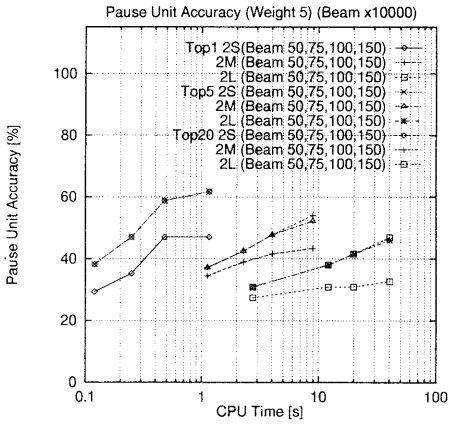

図 9: 提案手法による前終端記号バイグラム併用時の対話音声認識実験結果．閾値によるビーム探索方式．語数による正規化なし．ポーズ単位認識率による評価

に対し個別に語認識率を計測した時の最大値を累積の語認識率とした．

3.3 ポーズ単位の対話音声認識実験結果

5 対話，4 話者，2 話題 (ホテルの予約とホテルでのサービス)，66 発話，119 ポーズ単位，845 語を対象に実験を行なった．「あのキャンセルしたいんですが」のように間投詞 (この例では「あの」) も随所に挿入されている．「はい」のような感動詞 1 語や「え」のような間投詞 1 語で一つのポーズ単位となることもあるし，「あいにくですがシングルが満室となっておりますが」という比較的長いポーズ単位もある．なお，ポーズ単位の平均時間は 1874 ms であった．

従来手法 [9] の実装方法による実験結果を**図 7**, **図 8**に示す．文法のみを利用し，個数によりビーム探索を制限している．**図 7**がポーズ単位認識率，**図 8**が語認識率である．

提案手法による実験結果を**図 9**, **図 10**に示す．閾値によるビーム探索方式の条件で，前終端記号のバイグラムを併用した場合の結果である．**図 9**がポーズ単位認識率，**図 10**が語認識率である．

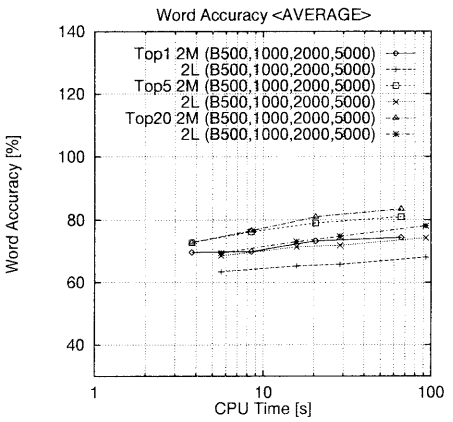

図 8: 従来手法の実装方式による対話音声認識実験結果．個数によるビーム探索方式．語認識率による評価

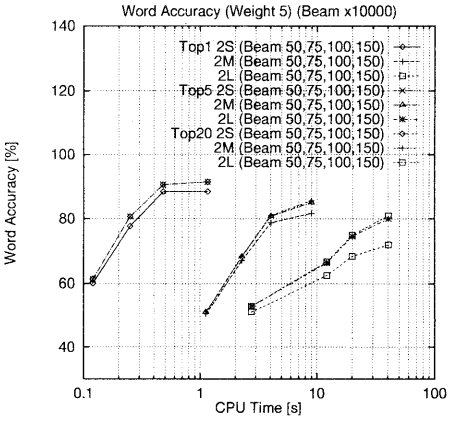

図 10: 提案手法による前終端記号バイグラム併用時の対話音声認識実験結果．閾値によるビーム探索方式．語数による正規化なし．語認識率による評価

3.4 議論と要約

前終端記号バイグラムを予測的に評価する，効率的な探索手法を実現し，その効果を確認した．従来手法はCPU 時間で計測して実時間のほぼ2倍ないしそれ以上であったが，提案手法は中小語彙サイズであれば実時間処理をほぼ達成した．

ビーム探索において一定の個数を残す手法と閾値による枝刈り手法を比較した結果，閾値による枝刈り手法のほうが効率的であることが確認できた．

ビーム探索過程で利用する，尤度スコア ($score$) の計算は次の二つの式を試みた．

$$score = \log P_A + weight \times \frac{\log P_L}{N} \qquad (1)$$

$$score = \log P_A + weight \times \log P_L \qquad (2)$$

ここで，P_A は音響スコア，P_L は言語スコアである．N は音素系列を構成する語数である．$weight$ は重み係数である．音響スコアと言語スコアの対数の底を揃えた上で予備実験を行ない，$weight$ は 5.0 とした．

文献 [9] より以前に行なった，認識候補を後処理的に並べ換える予備実験では，語数で正規化したほうが正規化しない場合よりよい結果が得られていた．しかし，実際に認識過程で併用する実験を行なうと，いずれもほぼ同程度の性能向上が確認できたが，(1) 式は正規化に要する計算量の処理時間が増加した．

要約すると，前終端記号バイグラムを (2) 式の評価方法で予測的に併用する探索手法で，閾値によるビーム探索を行なう場合がよい．

4 複数のポーズ単位からなる発話を対象とする対話音声認識手法の検討

多くの発話は我々の日本語文法で文として取り扱えるので，ポーズ単位ではなく，発話 (文) の文脈自由文法を用いた対話音声認識を実現したい．ただし，一つの発話は複数のポーズ単位から構成されることがあるため，ポーズをスキップしながら複数のポーズ単位を構文解析できる仕組み [7] を用いた対話音声認識手法を検討する．

4.1 基本的な考え方

試作済みの音声認識用日本語文法において，部分木で扱う現象を大まかに分類すると，次のようになる．

1. 間投詞
   文頭 (発話の開始時点) にある間投詞のみ句構造規則で扱うことができる．その他の場合は部分木になる．

2. 呼びかけ
   文頭 (発話の開始時点) にある呼びかけ語のみ句構造規則で扱うことができる．その他の場合は部分木になる．

3. 助詞の省略
   話者自身を指す「こちら」や「わたくし」以外の名詞句が助詞を伴わずに動詞を修飾している場合，句構造規則では扱わず，部分木として出力する．こうした部分木は意味処理部において併合することを想定している [1]．

4. 倒置
   倒置は部分木として出力する．

5. 言い直し
   言い直しを扱う句構造規則はない．ポーズで区切られていれば，原理的に何らかの部分木出力は得られる[4]．

6. 述語の省略
   述語がない場合，部分木で扱う．

7. 融合文
   ねじれた文やいわゆる「非文」は部分木で処理する．

8. 箇条発話
   現在の日本語文法では並立する表現は，一部の例外を除いて，部分木としている．

我々の旅行会話データベースを調査すると，対話音声の多くは文らしい発話である．例えば，**第3章**で扱った対話音声データの 66 発話中，77% に相当する 51 発話は我々の日本語文法では一つの構文木にまとまる．なお，残りの部分木で扱う必要があるものは，主に箇条発話と，格助詞あるいは述語の省略であった．そこで，文献 [7] の手法を導入して，複数のポーズ単位からなる発話を対象とする，効率的な探索手法を検討する．

4.2 予備実験

予備実験として，文献 [7] の手法により，ポーズをスキップしながら，複数のポーズ単位を受理する対話音声認識実験を行なった．ポーズは語の間であれば任意の箇所に挿入できるように設定した．つまり，名詞と助詞の間にポーズが挿入されてもよい．

まず，その効果と課題を調査することを目的に，最終的に一つの構文木にまとまる発話のみを実験対象とし，部分木を出力するようには設定しなかった．そのような観点から**第3章**で扱った対話音声データの一部を除外したところ，5 対話，4 話者，2 話題 (ホテルの予約とホテルでのサービス)，51 発話，74 ポーズ単位，766 語となった．ポーズ単位の平均時間は 1818 ms，発話全体の平均時間は 2637 ms であった．

発話全体の認識率を**図11**に，語認識率を**図12**に示す．閾値によるビーム探索を採用し，言語スコアの重み ($weight$) を 5.0 として，(2) 式の評価方法とした．

[4] 語のバイグラムに基づく手法を採用してもこれと同程度以上に言い直しが扱えるとは考えにくい．

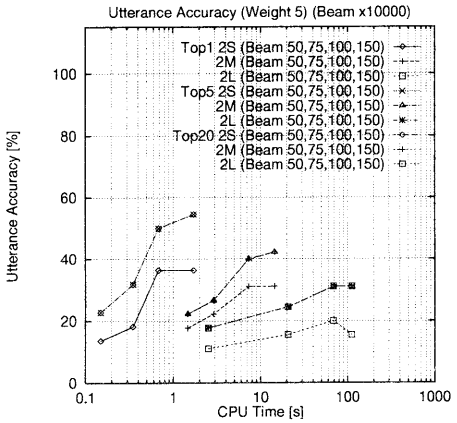

図 11: 複数のポーズ単位を受理する予備実験結果．発話認識率による評価

なお，この予備実験で対象とした発話はおおむね文に対応すると言ってもよいので，**図11**の発話認識率はおおむね文認識率に相当すると考えてもよい．音声翻訳や音声対話システム [3] の言語処理部に渡すことを考えれば，まずは文認識率が高いことが望ましい．

4.3 議論と今後の課題

一つの構文木にまとまる発話を対象に，複数のポーズ単位を受理する対話音声認識の実験を実施したところ，**図9**のポーズ単位の結果を組み合わせて発話全体を構成するのに比べて，発話全体の認識率 (**図11**) は上昇する傾向が見られる．しかしながら，**図12**に示すように，

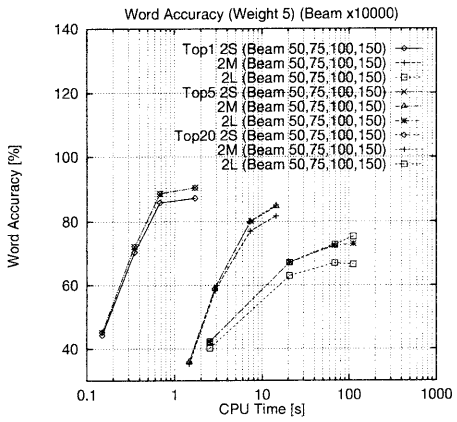

図 12: 複数のポーズ単位を受理する予備実験結果．語認識率による評価

語認識率は上昇しなかった．この結果は，文らしい発話が多いとしても，部分木出力を併用したほうがよいことを示唆している．文法がそもそも部分木として出力することを前提としている現象については，部分木出力を許さない限り，原理的に受理できない．将来的には，ポーズをスキップしながら文らしい候補を探索し，同時に部分木をも出力できるような，効率的な対話音声認識手法がよいと考えられる．

5 むすび

文脈自由文法形式の統語的な制約を用いて，部分木系列をスコア付きの仮説として出力する音声パーザにおいて，辞書引きの実装方法とビーム探索の手法を改善することにより，高速化と高性能化が達成できた．

今後は，複数のポーズ単位からなる発話を対象に効率的な対話音声認識を行なう手法に関する検討と実験を追加するとともに，部分木による言語解析部とのインタフェース手法 [10, 1]，対話・文脈処理部との協調動作 [4] に関してさらに研究を進める予定である．

謝辞

本研究を進めるにあたり，適切な助言と支援をいただいた荒川直哉氏に感謝します．さらに，実験を進めるうえで支援していただいた林輝昭，大槻直子 両氏に感謝いたします．

参考文献

[1] 荒川直哉，竹澤寿幸，森元逞，“統計的手法による部分木併合，” 信学技報，NLC-96-8, 1996.

[2] 伊藤克亘，連続音声認識システムに関する研究，東京工業大学，1993.

[3] 巌寺俊哲，竹澤寿幸，田代敏久，加藤直人，石崎雅人，森元逞，“ポーズ単位に基づく音声言語統合処理と発話状況管理の統合 — 音声対話システムの試作 —，” 信学総大，SD-4-3, pp. 331–332, 1996.

[4] 巌寺俊哲，竹澤寿幸，石崎雅人，森元逞，“次発話予測による音声認識候補の再順序付け，” 情報処理学会第 *53* 回 (平成 *8* 年後期) 全国大会，7N-5，分冊 **2**, pp. **2**-355–**2**-356, 1996.

[5] K. Kita, T. Morimoto, and S. Sagayama, “LR Parsing with a Category Reachability Test Applied to Speech Recognition,” *IEICE Trans. Inf. & Syst.*, Vol. **E76-D**, No. 1, pp. 23–28, 1993.

[6] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki, “A Speech and Language Database for Speech Translation Research,” *Proc. of ICSLP '94*, pp. 1791–1794, 1994.

[7] T. Takezawa, and T. Morimoto, “An Efficient Predictive LR Parser Using Pause Information for Continuously Spoken Sentence Recognition,” *Proc. of ICSLP '94*, pp. 1–4, 1994.

[8] 竹沢寿幸，田代敏久，森元逞，“自然発話の言語現象と音声認識用日本語文法，” 情処研報，95-SLP-6-5, 1995.

[9] 竹澤寿幸，森元逞，“部分木に基づく構文規則と前終端記号バイグラムを併用する対話音声認識手法，” 信学論 *D-II*, Vol. **J79-D-II**, No. 12, 1996.

[10] 田代敏久，竹澤寿幸，森元逞，“音声言語処理のための部分木併合手法，” 情処研報，95-NL-109-4, 1995.

[11] 外村政啓，小坂哲夫，松永昭一，門田暁人，“MAP-VFS 話者適応法における平滑化係数制御の効果，” 音学講論，2-5-6, 1995.