ICOT Technical Memorandum: TM-0391

TM-0391

韻律情報を利用した構文推定および ワードスポットによる会話音声理解方式

小松　昭男，大平栄二，市川熹
（日立）

September, 1987

©1987, ICOT

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

韻律情報を利用した構文推定および ワードスポットによる会話音声理解方式

正員　小松昭男[†]　　非会員　大平栄二[†]　　正員　市川熹[†]

Prosodical Sentence Structure Inferrence and Word Spotting for Conversational Speech Understanding

Akio KOMATSU,[†] Eiji OOHIRA[†] and Akira ICHIKAWA[†]
Member　　Nonmember　　Member

† (株)日立製作所中央研究所，東京都
Central Research Laboratory, Hitachi Ltd., Tokyo, 185 Japan

Prosodical Sentence Structure Inferrence and Word Spotting for Conversational Speech Understanding

ABSTRACT

The development of a system capable of understanding spoken language is a highly desirable objective, because speaking is the most natural form of communication.

Analyzing natural conversational speech, it became clear that human speech comprehension appears to mainly involve use of prosodic information for syntactic structural analysis and that speech content seems to be understood as a set of keywords, not as a string of phonemes.

The algorithm proposed here for conversational speech understanding consists of the following steps:

(1) Syntactic structural inference from prosodic information processing -- Analyzing the fundamental frequency contour pattern of speech, sentence structures are inferred.

(2) Semantic content inference from phonetic information processing -- Utilizing word spotting processing, with use of partial word patterns as the unit of recognition, speech content is inferred.

The performance of our algorithm is analysed by computer simulation experiments using actual spoken sentences within a conversational model.

あらまし　人と計算機との知的な対話機能の実現のためには、人間にとって極めて自然な情報伝達手段である音声による会話が必須の要件である。しかし、従来の音声認識・理解の研究開発においては、比較的丁寧に、協力的に発声された音声を対象とした場合が多く、自然な会話環境との隔たりが少なくない。

本稿では、実際の会話環境に近い状態で収録した自然な会話音声を解析・検討し、会話音声理解における問題点や課題を明らかにした。更に、その検討結果をベースにして、自然な会話音声の理解方式を提案した。これは、会話での情報伝達において音声が持つ音韻情報と韻律情報とが果たす機能を考慮したもので、音韻、韻律、言語の各レベルにたいして一貫した推論方式をもつアプローチである。特に、①会話音声に固有な韻律情報を積極的に利用し、音声会話文の文構造や重要単語の位置を推定すること、②標準単語音声の部分パターンを認識の単位として、文意伝達の核となる重要単語を推定・検出（ワードスポット）すること、に特徴がある。

1．　まえがき

音声は、人間にとって極めて自然な情報伝達手段である。音声による会話によって自然な情報交換がなされていることを考えれば、人と計算機との自然で知的な対話機能を実現するためには、音声による会話が必須の要件であるといえる。しかしながら、音声認識・理解の分野では、多様なニーズに答えるべき技術力はいまだ十分ではないため、話者、単語数、発声内容、発声方法、発声環境などに対する制限がついた状態で音声が利用されている。

現在、離散的に発声された単語音声の認識装置は製品化のフェーズにある。しかし、離散単語音声認識装置を利用したインタフェースが、余り知的でないことは明らかで、語数拡大や認識性能向上と共に、連続単語音声認識へと研究開発が進められている。[1]-[5] 更に、音声タイプライターを指向した連続音韻認識の研究がなされている。[6]-[10] しかし、これらの研究開発では、実際の会話環境のなかでの自然な音声を扱ったものは少なく、比較的丁寧に、協力的に発声された音声を対象とした場合が多い。即ち、利用者に、何らかの意味で拘束条件を与えたものであり、自然な音声による実際の会話とは隔たりがある。

本稿においては、利用者に発声に関する制約を課さない状態の自然な会話音声を解析・検討し、その結果をベースに、音声・言語に関する知識を利用して文意を推定する会話音声理解方式を提案する。これは、①会話音声に固有な韻律情報を利用して文構造を推定し、②標準単語音声と入力音声とに共通する部分単語を認識の単位にして重要な単語を推定・検出（ワードスポット）する、点に特徴がある。

2．会話音声理解方式の基本検討

会話音声理解において対象とする会話音声を、実際の会話環境に近い状態で収録し、その解析をベースに、会話音声理解方式についての基本的な検討を行う。

2．1　会話音声の収録

一般に、会話音声理解システムにおいては、或る目的を持った会話を対象とする。ここでは、PBX(Private Branch Exchange) での電話交換業務をタスクとして想定し、構内電話を用いた模擬的な会話実験を行って、会話音声を収録した。代表的な会話例は、次の通りである（交nは、交換手役のn番目の発声を示し、利mは利用者役のm番目の発声を示す）。[11][12]

交1: こちらは...でございます
利1: おはようございます
交2: おはようございます
利2: ...と申しますが
交3: はい
利3: 内線3138の...をお願いしたいんですが
交4: はい、お待ち下さいませ

この会話例の音声サンプルを図１に示す。このような会話サンプルを70例収録した。

2．2　会話進行に関する解析

音声による会話の進行に関して次のことが分かった（これらのすべてが会話音声理解に直接的に役立つものではないが、理解方式の枠組みを考える上で基本となるものである)。

(1) 音声会話文は文末まで聞かないで理解されている。

上記の会話例において、利１の挨拶が終わる前に交２の挨拶が始まっているし、交２の挨拶が終わる前に利２の発声が始まっている。このような会話のオーバラップは随所に見られるし、また、特定の話者に限らず、一般的な現象であった。このことから、音声による会話では、相手の発声内容は、文末まで聞きおえて理解されているのではなく、会話の状況や、会話の進み方などから、相手の発話内容を予測・推定しながら会話が進められているといえる。

(2) 相槌などの応答で会話がスムーズに流れている。

ある程度まとまった発声の終わり近くでは、「はい」などの応答が多い。この相槌の様に新しい情報のない応答を時間的にオーバラップしながら会話することにより、話したり聞いたりすることを容易にしている。これは、音声には「消しゴム」がなく実時間性が強いという特質と相まって、会話の自然性を高めるのに役だっていると思われる。一般に、相手への情報伝達が部分的な時点での応答は、会話の流れをスムーズにする上で重要な役割を果たすといえる。

(3) 簡潔な省略文により効率よく会話の指導権が交代されている。

普通、「...をお願いします」によって、問題解決（内線接続）に必要な情報が、利用者主導型で提供される。しかし、情報が不十分な場合や、不確実な場合には、交換手（システム）主導型で情報の補充・確認を行う。例えば、発信人の名前を知るために、「どちら

様です？」と簡単に質問している。また、「内線1234番」の下２桁が不明瞭な場合には、単に「12」と（特有のイントネーションで）下２桁の再入力を促している。このように、状況に応じた簡潔な省略表現で会話の指導権が効率良く制御されつつ、利用者に適度な自由度を与えて会話が進められている。

２．３　会話音声に関する解析

(1) 発声内容

収録した70例の会話音声の文型は、ほぼ限られた種類のものであった。たとえば、自分の名前を言う場合(自己紹介)には、大半が、「...ですが　...をお願いします」の文型で、受信人名を指示する前に、自己紹介を行っている。また、受信人名の指示方法では、[所属、名前]、例えば、「<所属>の<名前>をお願いします」、が49例(70%)で最も多く、残りは、[内線番号、名前]が17例、[内線番号、所属、名前]が4例であった。これは、タスクが簡単なため、多様な言い回しが実質上は不必要であること、及び、会話においては定型的な言い回しが多用されること、の理由によるものと考えられる。

また、このように簡潔な表現が多いため、不整文は少なく、「...さんを　おみえでしょうか」と、格助詞が間違ったもののみであった。これは、「...さんを　お願いします」の途中で、更に丁寧に表現しようとして文型が混乱したためと解釈でき、音声が思考の中枢に直結したメディアであることの一つの証とも言える。

会話音声に固有な、「あのー」や、「えー」などの無意味語は、殆どすべての会話中に表れていた。このような無意味語が無いものは、数例のみで、いずれも、「本社です 1部の鈴木をお願いします」のように、簡潔な文型のときのみであった。

(2) 発声方法

典型的な会話例(図1)を音響処理し、詳細な解析を行い、次のことが分かった。

①重要な単語は丁寧に発声されている。

数字が比較的聞き取りにくいことを考えれば、「内線」という単語は、相手の情報処理の予測を促す重要な単語であるため、自ずと、丁寧に発声されている。この部分の発声速度は、７~８音節／秒であり、会話のなかで比較的ゆっくり発声されている。また、その部分の基本周波数は高い。

②新しい情報のない部分の発声は雑である。

「．．．をお願いしたいんですが」の文末に近い部分は、定型的な表現であり、会話の双方にっとって新しい情報を含んではいない。発声は丁寧ではなく、相手も注意して聞いてはいない。発声速度は約１５音節／秒で、部分的ではあるが極めて早口な発声であるといえる（基本周波数は低く最低の値に近い）。この部分の音韻認識は極めて困難であることは明らかであろう。従って、音韻を１つづつ正しく認識することよりも、新しい情報のない定型的な表現であることを理解することが必要である。会話音声における基本周波数の変化は、伝達すべき情報の価値（重要度）を理解する上で、一つの重要な指標であるといえる。

２．４　会話音声理解の課題

以上に示した実際の状況に近い会話音声の解析・検討結果から、会話音声理解における主要な課題は、次のようにまとめられる。

(1)会話音声に固有な韻律情報の処理

音声パワー、アクセントやイントネーションなどの韻律は、①相手が理解しやすいように自然に、標準的なルールに従って表現されたもの、②発話者の状況・意図などに依存した省略や強調などが、意識的または無意識的に表現されたもの、とに分けられる。いずれにしろ、韻律情報に関した暗黙の了解によって、効率良く会話が進められている以上、会話音声理解において、韻律情報の利用は不可欠である。

(2)音声パターンの曖昧性の処理

自然な会話においては、発声速度などの話し方に関するルールは殆どなく、話者が自由に制御できる。例えば、名前をいう場合、「す、ず、き」と丁寧に離散的に発声する場合も有れば、早口に発声する場合もある。また、部分的には、発声速度が数倍のオーダで変化しうる。このため、調音結合の影響による音声パターンの変形が不均一になり、曖昧性が増大する。このような調音結合の影響は、音韻レベルの場合もあるし、単語レベル、さらには、文節レベルまで及ぶ場合もある。しかも、そのレベルを前もって一意的に定めておくことは不可能である。このような、広いレンジに及ぶ音声パターンの変形を動的に吸収することが必要である。

(3)多様な言い回しを含んだ会話的表現の処理

話し言葉では、書き言葉に比べて、表現の自由度が高い。このため、音声による会話文は、主に書き言葉を対象にした文法からみれば、必ずしも整った形式にはなっていない。さらに、思考の進み具合によって、「あのー」などの無意味語が挿入されたり、逆に、表現の一部が省略されたりする。このように多様な表現では、文法主導型ではなく、意味主導型で処理することが必要である。

２．５　会話音声理解の基本方式の提案

これまでの解析・検討の結果をベースに会話音声理解の基本方式を提案する(図2)。これは、会話での情報伝達において音声がもつ音韻情報と韻律情報とが果たす機能を考慮したもので、音韻、韻律、言語の各レベルにたいして一貫した推論方式をもつアプローチである。特に、①会話音声に固有な韻律情報を積極的に利用し、音声会話文の文構造や重要単語の位置を推定すること、②標準単語音声の部分パターンを利用して、文意伝達の核となる重要単語を推定・検出（ワードスポット）すること、に特徴がある。本方式は、意味主導型で会話音声を理解するものであり、多様な会話的表現にも対処しうるものである。

このような会話音声理解の基本的な枠組みを基にして開発した、具体的な処理内容の概略は、次の通りである（各々の詳細については、３章と４章とで述べる）。

(1)韻律情報を利用した文構造推定

会話音声の基本周波数の形状を折線で近似し、その形状から、会話文を意味的な纏まりに分割すると共に、各分割区間の結合度（分割点前後のフレーズの係受け関係の強さ）を推定し、フレーズの結合関係から、音声会話文の文構造を推定する。

(2)部分単語パターンを利用したワードスポット

標準単語音声と入力音声とが共通する部分単語パターンをボトムアップで検出し、この部分単語パターンの組合せによって、トップダウンで予測される単語の検証を行う。前半の処理は、連続ＤＰマッチングの手法を用い、後半の処理は、未知語を含む構文解析手法を用いる。

３．文構造推定法

会話音声の韻律情報を利用[13][14]した会話文の文構造推定法を、処理手順に沿って述べる。先に提案した文構造推定法[15]のタスク依存性をなくして、一般的に利用できる方法に改良したものである。

３．１　音響処理

入力音声を、12KHz、符号付き12ビットでAD変換する。さらに、60m秒(720サンプル点)の矩形窓を掛けた後、自己相関法を用いて基本周波数を求める(窓は20m秒毎にシフトする)。後述のスペクトル分析に用いる窓長(20m秒)よりも大きな窓を利用して、基本周波数を安定に求められるようにした。さらに、基本周波数の連続性を利用して、倍ピッチ/半ピッチの抽出エラーを修正している。

３．２　基本周波数形状の折線近似

基本周波数の変化の様子を簡潔に表現するために、以下の手順で基本周波数の形状を直線の組(折線)で近似する。

①音声の始終端を始終点として、直線近似の対象区間とする。

②始終点を結ぶ1っの直線で近似する。この近似直線と実際の基本周波数との誤差の平均が一定値(3-4Hz/20m秒)以下ならば、当該区間の直線近似を終えて、③に行く。誤差が大きい場合には、④に行く。

③直線近似を終えた当該区間の終点を新たな始点として、後続区間の終点までを新たな対象区間として②に行く。新たな対象区間がない(当該区間の終点が音声の終端と一致した)場合、直線近似の処理を終わる。

④当該区間の始点と、近似直線との誤差が最も大きい点を終点とする区間を、新たな対象区間として②に行く。

３．３　結合度の定義

音声区間を、係受け関係(結合)の弱い点で分割して、構文木を推定する。折線で表現された基本周波数形状の極小値点を分割点とする(これは、文末、文節末、単語境界、単語内音節境界である)。分割点前後の2つの音声区間の基本周波数の最小二乗直線を求め、これを各々の(仮想的な)フレーズ成分とみなす。このフレーズ成分の傾きや長さ、および、フレーズ成分の立ちあげの大きさを考慮して、その分割点におけるフレーズ間の結合度を定義する。結合度の大きさは、定性的には、次のように推論できる。

①分割点前のフレーズの傾きが、フレーズ成分の標準的な値（-25Hz/秒とする)に近い程、フレーズの終点である可能性が高く、結合度は小さい。

②分割点前のフレーズの時間長が短い程、単語内分割点である可能性が高く、結合度は大きい。

③分割点前後のフレーズ成分の差（基本周波数の差）が大きい程、新しく大きないフレーズの立ちあげである可能性が高く、結合度は小さい。

結合度を、上述の３つの変数からなる関数と考え、分割点iにおける、前後のフレーズ(AとB)の結合度 Ri(A,B)を、次式により定義する(図3a)。

Ri(A,B) = Wd*フレーズAの傾きd (Hz/秒：-25Hz/秒との差の絶対値)

+ Wl*フレーズAの時間長l（秒：無音区間は除く）

+ Wg*フレーズBの立ちあげg（Hz：Bの始点とAの終点との差）

ここで、Wd,Wl,Wgは重み係数である。

３．４　重み係数の決定

重み係数は、実際には、多くの実データの解析に基づいて前もって定める。しかし、現段階では、ヒューリスティックな処理ではあるが、会話音声サンプルから、次のように、学習して決める(図3b)。

学習用に発声した1つの会話音声サンプル、「おはようございます　中研　６部の　小松ですが　8部の　鈴木さんお願いします」、に対し、5個の分割点が生じた。この音声サンプルから、[[おはようございます][[[[中研][6部の]][小松ですが]][[8部の][鈴木さんお願いします]]]]、のような正しい構文木が生成されるためには、各分割点での結合度は、図3cに示すように、四つの制約条件(c1~c4)を満足する必要がある。各分割点での実測値に基づいて、この制約条件を満たす解の空間を求め、その内の任意の点から、重み係数(Wd,Wl,Wg)を決定する(図3d)。

このような方法においては、学習サンプルを増す(制約条件を増す)ことによって、より最適な解の範囲に絞り込むことができる。しかし、解が存在しなくなる場合がありうる、これは、制約条件が厳しくなり過ぎたためか、結合度の定義が基本的に間違っていたかの、いずれかであり、場合に応じて、詳細な検討を要する。上の例では、十分に広い解の空間が得られており、結合度の定義や制約条件は、一応、妥当であったと言える。

３．５　構文木生成

発話内容が未知の入力音声にたいして、基本周波数形状の直線近似を行い、分割点を求め、各分割点での結合度を計算する。この結合度は、分割点前後のフレーズ間の結合の強弱を推定し、相対的な値で表したものである。従って、各分割点の結合度の弱い順に音声会話文の分割を進めることにより、最終的には、フレーズ間の係受け関係を反映させた、一つの構文木を得ることができる。

なお、ここで述べた構文木は、文法的な拘束条件を反映させたものではなく、会話音声の発声方法から推定できるフレーズの間の係受け関係を表現したものである。従って、この構文木は、意味主導型の理解において、より高次な知識を利用した構文・意味解析の予測値として利用できる。

３．６　構文推定の実験

以下のような会話音声サンプル（接話マイク、成人男子１名）について構文推定処理の実験を行った。各々3回づつの発声ではあるが、3回共に同じ構文木が得られ、安定に動作することが確認できた(図4)。

(1)重み係数を学習したのと同型の会話文に対して、同じ形の構文木が得られた(図4a)。

(2)同一タスク(PBX電話交換業務)のやや複雑な会話文、「資材課 内線 3611の 佐藤さんお願いします」、に対して妥当な構文木が得られた(図4b)。

(3)無意味語を含んだ会話文(「あの－鈴木さんをお願いします」)と、指示語を含んだ会話文(「あの鈴木さんをお願いします」)とが、得られた構文木の形から区別できた(図4c)。

(4)構文的に曖昧な表現(「庭には 鶏がいる」と「2羽 庭には 鳥がいる」)が、得られた構文木の形から区別できた(図4d)。

４．ワードスッポット法

表現が多様な会話音声を理解するためには、連続音声中から文意伝達の核となる単語を切り出すワードスポット[15][17][18]が必須である。本章では、標準単語音声と入力音声との共通する部分パターンを求め、その組合せで単語音声を推定・検出するワードスポット法について、処理手順に沿って述べる。

４．１　音響処理

標準単語音声、未知の入力音声ともに、１２ＫＨｚ、１２ビットで標本化した後、差分処理を行い、２０ｍ秒（２４０サンプル点）のハニング窓をかけ、分析次数１２でＬＰＣ分析を行う（窓は２０ｍ秒毎にシフトする）。

４．２　量子化

スペクトル分析の結果得られるスペクトル情報を量子化して，記号の列で表す。これにより、後続する論理的な処理との親和性がよくなり、記号列処理をベースにした統一的な論理処理を実現できる。ベクトル量子化[19]の考え方を基本にした処理であるが、量子化によって失われるスペクトル情報を少なくするため、量子化誤差がスペクトルの分析誤差の程度になるようにする。具体的には、定常母音を１サンプル点毎にシフトして分析した時の隣接フレーム間のスペクトル誤差の最大値付近(累積頻度90%のところ)を分析誤差とみなし、同程度のスペクトル誤差を許容範囲としてコードブックを作成する。入力フレーム毎にコードブックとの参照を行い、最小となる距離が許容範囲内であれば該当するコードで入力フレームを表し、許容範囲以上の場合には、入力フレームをコードブックに追加登録する。予備実験の結果では、誤差の許容範囲は、COSH尺度で、0.35程度であった。

４．３　クラスタリング

スペクトル情報を細かく量子化しているため(誤差の許容度が少なく)、同じ音韻に対応していると思われるスペクトル情報も異なったコードで表されてしまい、パターンの認識(類別)ができない。このため、何らかの同値関係を定めたクラスタリングが必要である。この時、厳しい同値関係では、パターンの類別ができないし、逆に、同値関係をルーズにすると、本来異なったクラスターに入るべきものが同じクラスターに含まれてしまい、パターンの誤認識の原因となる。

このような、クラスタリングの閾値決定の問題は、整合的な知識ベース構築の問題[20]として扱うことができる。すなわち、音韻に関する知識ベース(記号列と音韻との対応を示すルール)において、①証明可能性、②冗長性、③矛盾性、④独立性、の概念をチェックしながら、矛盾が生じる直前まで同値関係(異なる記号間のスペクトル情報が同一とみなす関係)の閾値を拡大する。これは、知識の量や内容に応じてクラスタリングの閾値を動的に調整する方法である。人間のパターンの類別と同様、知識が多くなるに従って類別のための閾値が自動的に厳しくなる。

しかし、実際のワードスポットの処理過程においては、このような動的なクラスタリングを行うことは、処理量の面から現実的ではない。従って、予備的な実験において上述の処理を行い、クラスタリングの閾値を前もって定めておくことにした。5人名の単語音声(成人男子1名)を対象とした場合、cosh尺度で0.55程度の閾値が求められたので、これを

クラスタリングの閾値としてコードブックを再編成し、標準単語音声や入力音声をコード化して、記号列で表現する。

４．４　部分パターンマッチング

入力音声と各標準単語音声との間で記号列のマッチングを行い、両者に共通する部分単語パターン(部分単語音声)を抽出する。このような部分パターンマッチングでは、次のような機能を実現する。

①入力パターンの任意の時点の記号と、標準パターンの任意の時点の記号とが一致した時点をマッチングパスの始点とし、一致する記号の列が最長となる区間をマッチングパスとして求める。

②時間軸上での非線形な伸縮を許す(±30%程度の伸縮を許す)。

③記号の不一致を一定の割合で許す(約1割のミスマッチを許す)。

以上の部分パターンマッチングにより、或る程度の曖昧性を吸収しながら、入力音声と標準単語音声とで共通する部分単語パターンを検出することができる(図5a)。

なお、この処理は、概念的には、連続DPマッチング[2]の始点開放の機能を、入力パターンと標準パターンとに2重に適用した方式となっている。

４．５　音韻推定

部分パターンマッチングで得られた部分単語パターンは、入力音声と標準単語音声との音響的に似た部分を示している。従って、発声内容が既知の標準単語音声中の位置関係などから、部分単語音声の音韻情報を推定できる(図5b)。例えば、部分単語パターンが、「satou」という標準単語音声の先頭の20%程度と対応していたとすれば、この部分単語パターンの音韻内容は、/s/であると推定でき、入力音声中の部分単語パターンに対応する部分が、/s/であるという仮説を生成することができる。

この処理内容から明らかなように、音韻推定の処理は、入力音声の音響的な性質から音韻を直接推定しているのではなく、その音響的な性質がいかに表記されるかを、(標準単語音声の表記から)推定しているのである。従って、調音結合などの影響によって音響情報と音韻情報との対応を厳密に規定できないような環境においても、共通する音響的な性質から、それに対応した部分単語を推定することができる。例えば、単語「日立」が、実際には[ʃitachi]と発声されようと、[i]が極端に無声化されて発声されようとも、学習した単語音声と音響的性質が似ていれば、「日立」という単語、または、部分単語を推定できる(それが、[hi]か[ʃi]かの判定をする必要はない)。

なお、部分単語パターンの仮説には、矛盾するものが含まれる場合がある(例えば、図5bにおいて、/ai/と/e/との仮説が入力音声の同じ場所に生じている)。これは、調音結合の影響による対立仮説に相当するもので、複雑な調音結合の問題に間接的に対処した方式であるといえる。

４．６　単語推定(ワードスポット)

部分単語パターンは、標準単語音声の一部と似た部分が、入力音声のどの部分に存在しているかを示す情報である。従って、このような部分単語パターンを時間軸に沿って並べ、種々の組合せを検討することより、入力音声中の単語音声を検出することができる(図5c)。これは構文解析法と類似の処理内容で、部分単語パターンを形態素とみなした場合、その

組合せから、文に対応する単語音声を推定する方法であるといえる。但し、単語音声を構成するすべての部分単語パターンが推定できている保証はないので、対応する部分単語が存在しない未知部分をスキップして構文解析を進める機能が必要である(図5cでは、未知部分を＊で示してある)。また、生成された単語仮説の評価関数として、単語音声中の部分単語の一致度を用いる。これは、認識対象単語と一致した音韻の割合であり、母音(1.0)、子音(0.5)、調音位置や調音方法が同じ子音(0.3)で異なった重みとする。例えば、「tanaka」(4.5)に対する単語仮説[*,a,n,a,t,a](3.8)の評価値は，0.84となる。

このような処理手順から明らかなように、認識対象の単語音声の音響的な内容がすべて既知でない場合でも、単語の推定が可能である。例えば、入力音声の一部が、「中村」という標準単語音声の前半と一致し、それに続く部分が、「村山」という標準単語音声の後半と一致したとすれば、「中山」に対応する標準単語音声がなくても、「中山」という単語を推定することができる。

４．７　ワードスポットの実験

標準単語音声として、頻度の高い10人の人名単語(鈴木、佐藤、田中、山本、渡辺、高橋、小林、中村、伊藤、斎藤)を各1回学習した後、以下の入力音声に対して実験(成人男子1名、3回づつ発声)を行った結果、学習単語、部分的に学習した単語、未知単語を、いづれも正しくスポットできることが分かった(図6)。

(1)入力音声「えー田中さんお願いします」から、学習した単語(「田中」)を、第1位で推定できた(図6a)。

(2)入力音声「えー山中さんお願いします」から、部分的に学習した単語(「山中」という単語は学習していないが、その部分単語、「山」と「中」、が学習単語に含まれている)を推定できた(図6b)。

(3)入力音声「おはよう」に対応する単語音声が未知の場合(学習していない場合)でも、「おはよう」か「今日は」かの2つの仮説にたいして、「おはよう」を一位で推定できた(図6c)。

5．実験結果

5．1　実験条件

特定話者(成人男子)1名が、接話マイクを使用した会話音声を対象とする。学習用の標準単語音声は、出現頻度の高い人名10単語(4.7参照)を各々1回発声したもので、評価用の入力音声は、「えー○○○をお願いします」と、会話調に、20回発声したものを用いた。

5．2　音響処理

10語の単語音声の総フレーム数は、278フレームであり、平均6.5音節/秒の発声速度であった。これらをベクトル量子化(閾値のCOSH尺度は0.55)した結果、58個のコードで量子化できた。評価用の入力音声は、この閾値以内のスペクトル距離の場合には、対応する記号(2つ以上ある場合には、それらすべての記号の組)で表現し、その他の場合には、未知の記号を割り当てる(この未知の記号を含む部分は、標準単語音声と共通になる可能性はない)。

5．3　構文推定

20会話音声例の推定構文木の形は、すべて同じで、2つのセグメントに分割された構文木であった(図4cの無意味語を含んだ場合と同じ構文木)。これは、無意味語の「えー」の部分が、本文の「○○○をお願いします」と分離されて、重要語（相手の名前）を提示している部分が切り出されていることから、２０会話音声すべてに対して、構文推定と重要単語位置の推定とが正しくできたといえる。

5．4　ワードスポット

重要単語が含まれていると推定される本文の部分(数音韻分のずれを考慮して、前後10フレーム、200m秒分を含む)に対してワードスポットを行った結果、20名中17名は正しく単語検出できた。誤認識した3例は、/saitou/を/itou/と誤認識したもの(2例)と、/watanabe/を/tanaka/と認識したもの(1例)とであった。これらはいずれも、標準単語の包含関係の考慮が不十分なためであった。すなわち、/saitou/の/s/に一致する部分単語パターンが得られず、/aitou/に対して、/saitou/より/itou/に対する仮説のスコアが良くなったためである。同じく、/watanabe/の場合、/tana/の部分単語パターンのみが推定された結果、/tanaka/と誤認識した。これに対し、単語音声の判定法を改良し、推定された部分単語音声の長さをも考慮(一致した音韻の個数で重み付けして長さを優先)することにより、誤認識は、/watanabe/の1例のみとなり、95%の理解率が得られた。

5．5　結果の検討

上記の実験結果から、次のことがいえる。

①提案した会話音声理解方式の正当性　：　本論で提案した理解方式が、「えー」などを含んだ自然な会話音声を理解するのに有効な方式であり、基本的な機能が正しく動作している。また、パラメータなどの最適化は不十分であるにもかかわらず、一定の性能が得られたことから、会話音声の安定した理解が期待できる。

②タスク拡大、汎用化の見通し　：　サンプル実験では10単語のみを認識の対象としたが、タスク拡大に伴う語彙拡大に対して、認識性能の劣化は少ないと思われる。これは、本方式が、予測検証型の理解方式であるためである。例えば、「金沢」と「神奈川」のように

その一部のみが異なった単語対にたいしては、同じ仮説に対しては同じスコアが得られ、異なった部分のスコアが強調されるためである。しかし、包含関係にある単語対の認識精度には問題があり、対判定などの後処理が必要である。なお、本方式は汎用的な処理手順で構成されているので、タスク汎用化の可能性は高い。

6．むすび

実際の会話環境に近い状態で収録した自然な会話音声を解析・検討し、会話音声理解における問題点や課題を明らかにした。更に、その検討結果をベースにして、自然な会話音声の理解方式を提案した。これは、会話での情報伝達において音声が持つ音韻情報と韻律情報とが果たす機能を考慮したもので、音韻，韻律、言語の各レベルにたいして一貫した推論方式をもつアプローチである。特に、①会話音声に固有な韻律情報を積極的に利用し、音声会話文の文構造や重要単語の位置を推定すること、②標準単語音声の部分パターンを利用して、文意伝達の核となる重要単語を推定・検出（ワードスポット）すること、に特徴がある。本方式は、意味主導型で会話音声を理解するものであり、多様な会話的な表現に対処しうるものである。

また、本アルゴリズムを、実際の会話音声サンプルを用いて評価した結果、基本的な機能が正しく動作し、ワードスポットによって、95％程度の理解率が得られることが分かった。

本方式の提案により、各種の知識を利用した会話音声理解方式の基本的な枠組みが整ったといえ、汎用的な会話音声理解システム開発の手掛かりが得られた。

謝辞

本研究の推進を積極的に支援して頂いた、当所、堤善治第６部長、および、江尻正員主管研究長に感謝する。また、音声･言語について多くの有益な御意見、御討論を頂いた、北原義典企画員、淺川吉章研究員、畑岡信夫研究員、および、当社基礎研究所、新田義彦主任研究員に感謝する。

なお、本研究は、第５世代コンピュータ・プロジェクトの一環として、(財)新世代コンピュータ技術開発機構(ICOT)からの委託により行ったものである。

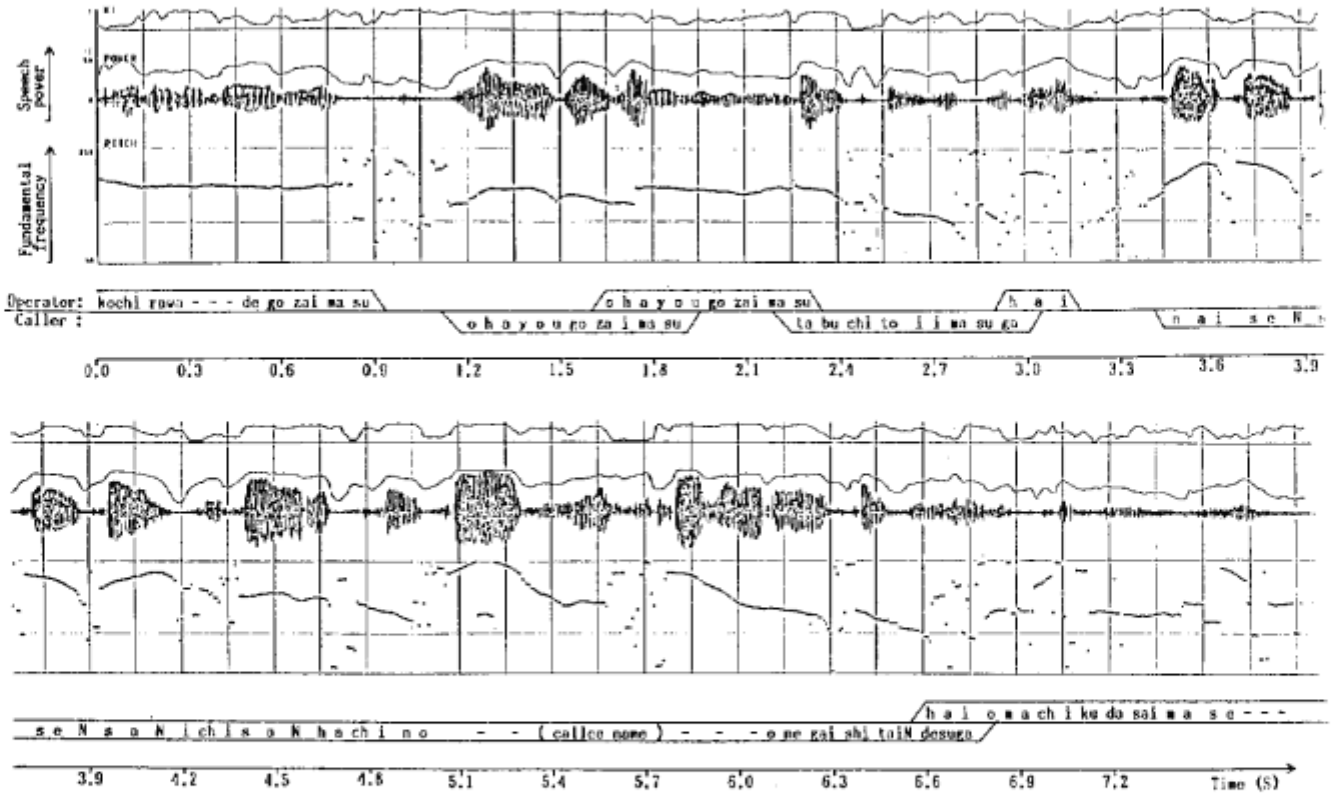


図1　自然な会話音声の例
Fig.1　An example of natural conversational speech.

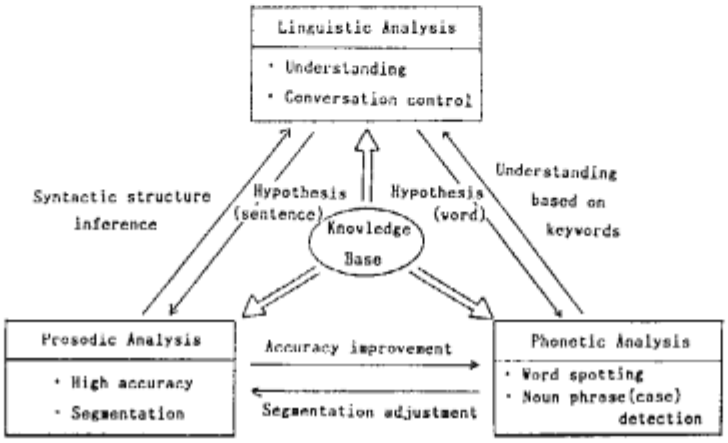
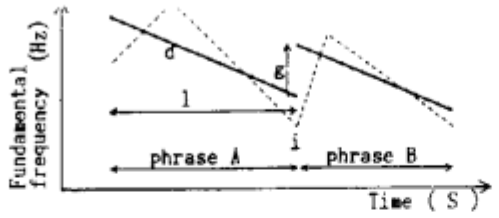


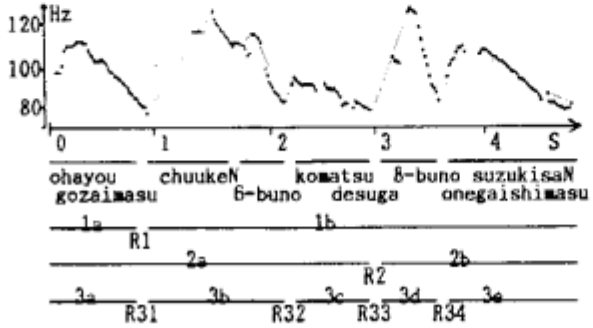
図２　会話音声理解の基本方式
Fig.2 Basic approach of conversational speech understanding.



$$Ri(A,B) = Wd * d + Wl * l + Wg * g$$

Wd,Wl,Wg : weighting coefficients
d : decline of approximation line (Hz/Sec)
l : length of speech (Sec)
g : gap of fundamental frequencys (Hz)

(a) Definition of connection rate Ri(A,B), at i, between two phrases, A and B.

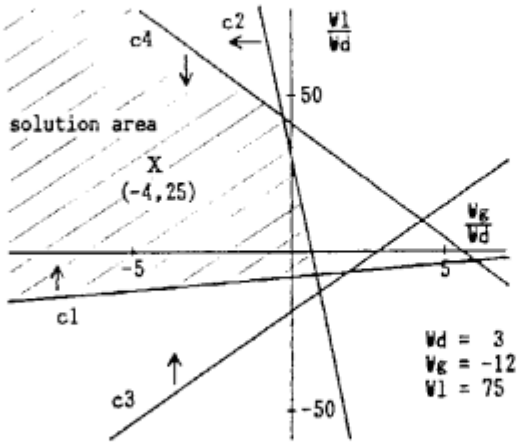


(b) Analysis of sample speech for deciding weighting coefficients.

c1 : constraint 1 -- R1 (1a,1b) < R2 (2a,2b)
c2 : constraint 2 -- R31(3a,3b) < R32(3b,3c)
c3 : constraint 3 -- R32(3b,3c) > R33(3c,3d)
c4 : constraint 4 -- R33(3c,3d) < R34(3d,3e)

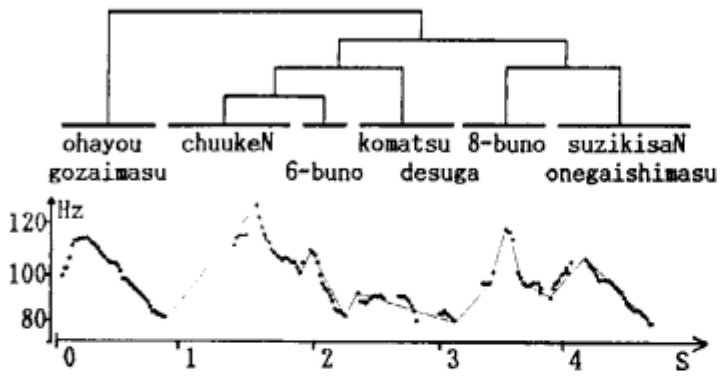
proper parsing tree of sample speech
[[ohayou][[[[chuukeN][6-buno]][komatsudesuga]]
[[8-buno][suzukisaNonegaishimasu]]]]

(c) Constraints on weighting coefficients to obtain proper parsing tree.

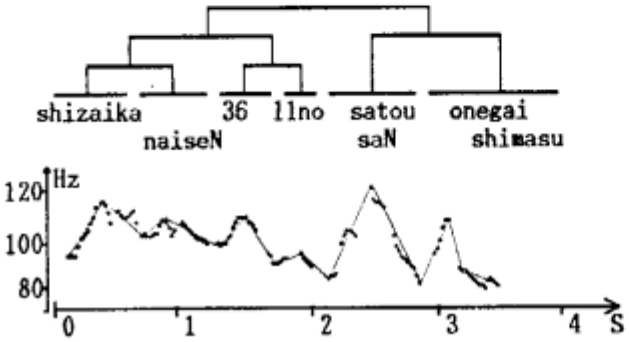


(d) Decide weighting coefficients from a point X within a solution area.

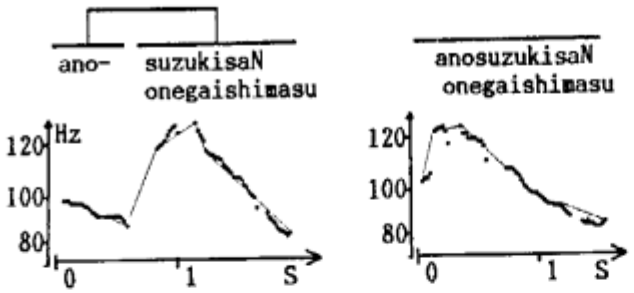
図3 フレーズ間結合度の係数の決定手順
Fig.3 Steps to decide weighting coefficients of connection rate between phrases.



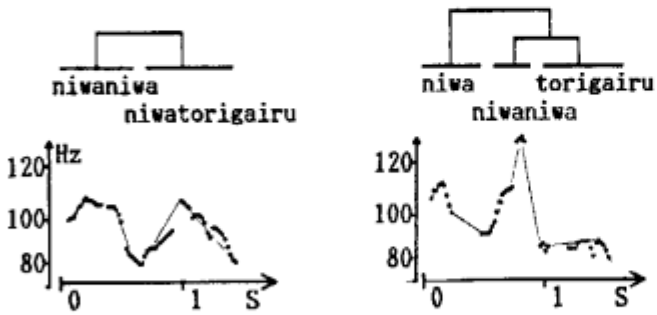
(a) Sample speech which is a same sentence structure as learned speech (Fig.3b).



(b) Sample speech in PBX task.

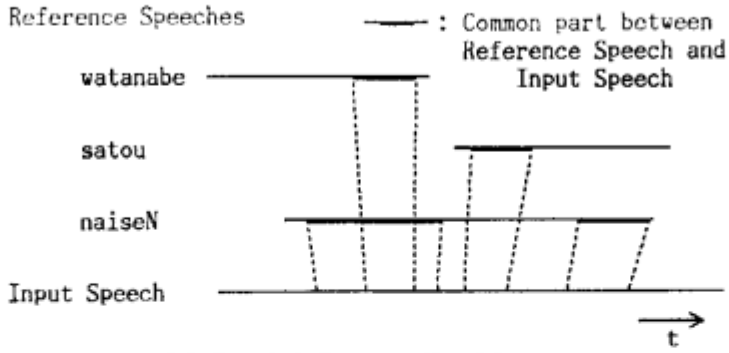


(c) Sample speeches with a meaningless word (ano-) or demonstrative pronoun (ano).

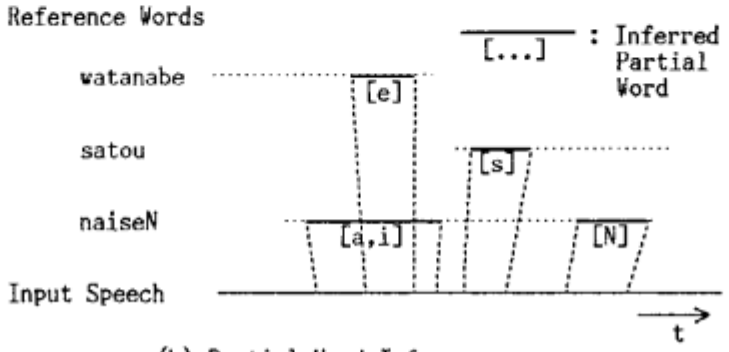


(d) Sample speeches of ambiguous sentence structure.

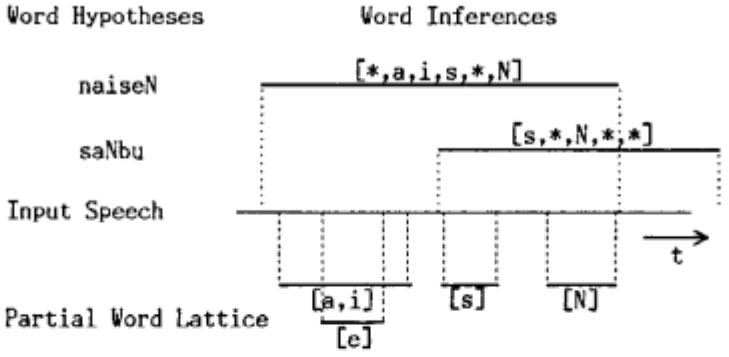
図4 韻律情報を利用した構文推定の例
Fig.4 Examples of parsing tree of speeches by prosodical sentence structure inference.



(a) Partial Pattern Matching.

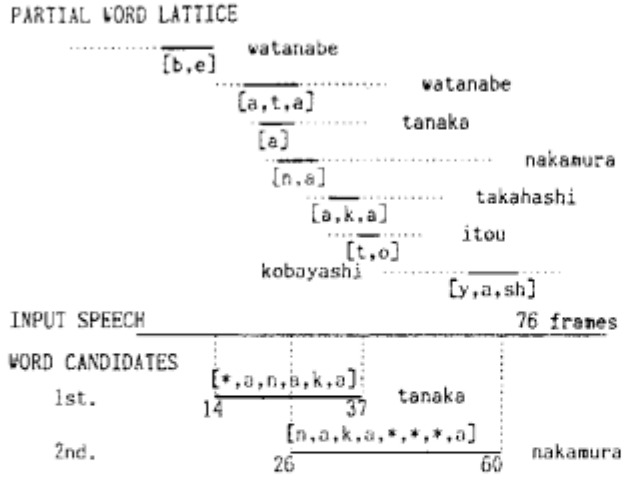


(b) Partial Word Inference.

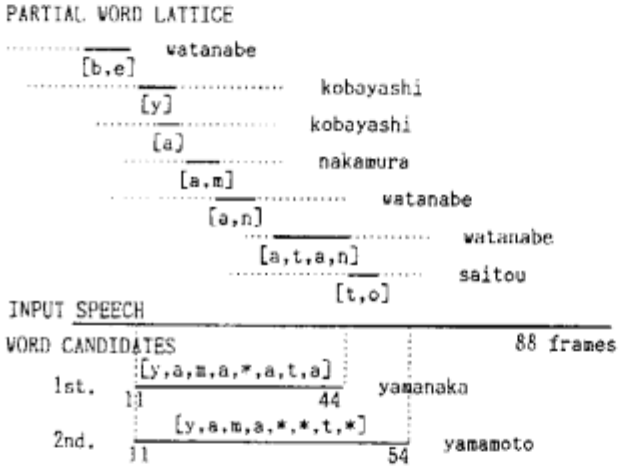


(c) Word Inference.

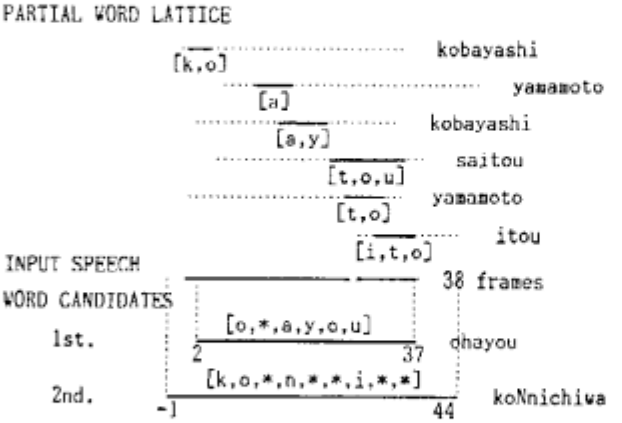
図5　部分単語パターンを利用したワードスポッティングの処理手順
Fig.4　Steps of word spotting utilizing partial word patterns.



(a) Input speech (/e- tanakasaN wo onegai shimasu/) with a known word (tanaka).



(b) Input speech (/e- yamanakasaN wo onegaishimasu/) with a patially known word (yamanaka).



(c) Input speech (/ohayou/) with a unknown word (ohayou).

図6 部分単語パターンを利用したワードスポッティングの例
Fig.6 Examples of word spotting utilizing partial word patterns.