音声言語情報処理 2－8
（1994．7．8）

対話理解に対する抑揚情報の役割

市川熹　佐藤伸二

千葉大学工学部情報工学科

概要

音声に含まれる抑揚情報が、対話の制御に対してどの様な役割を持つかを解析した。対話音声理解システムを実現するうえでの課題の一つに、システムと利用者の協調的対話の進行がある。我々の日常対話には省略や言い直しなどの曖昧な表現が多数含まれているにも関わらず、内容を即座に理解し話しのポイントが簡単に聞きとれたり、また対話を行なっている二人の間の発話の自然な交代が可能なのは、抑揚の効果であると考える。本研究では、対話データの解析結果から対話の制御に対して抑揚情報が持つ役割を示し、それが対話音声理解システムにどの様に利用できるかを検討した。

Roles of Prosodies in Dialogue

Akira Ichikawa　Shinji Sato

Dept. of Information and Computer Sciences, Chiba University

Abstract

In the situation of the dialogue, prosody has some essential functions. The first is to indicate the semantic structure of the utterance. The second is the real-time control information for dialogue. The third is as the information of speaker's psychological situations. In this paper, we report about some characteristics of the prosody from the viewpoints of the second and third function in a spontaneous spoken Japanese dialogue. These results will be useful to deverop the spoken dialogue human-machine interface system which should not obstruct the user's consideration for his task.

1 まえがき

近年情報処理技術の発達は著しく、それに伴い社会の情報化も急速に進んでいる。しかしながら、我々が普段コンピューターに接する際に使用するインターフェースは、キーボード、マウスといったある程度の知識と訓練を必要とするものである。これらは、高齢者や障害者などを含め誰もが自由に利用できるインターフェースとは言い難く、伝達できる内容にも限界がある。そのため自分の意志をより自由に、より簡単にコンピューターに伝えることが出来るインターフェースの実現が望まれる [4]。

誰もが自由に自分の意志を伝達することが出来る手段として音声などの対話型自然言語に着目する。特に我々が日常会話で用いている「話し言葉」は、自分の意志を自由に相手に伝えることが可能である。「話し言葉」を用いてコンピューターと対話できるようなシステムが実現できれば、普段人間と会話するかのように、自由に自分の意志をコンピューターに伝達することが可能となる。

しかし、「話し言葉」を認識できる対話音声理解システムの実現には様々な課題がある。

本研究では、その中でも主要な課題の一つであるシステムと利用者の協調的対話の進行を実現するために、音声固有の情報である抑揚情報が対話の制御に対してどの様な役割を持つかを解析し、対話音声理解システムにおけるその利用について検討する。

2 対話音声と抑揚情報

2.1 対話音声の特徴

対話音声とは、二人の話者の間で音声により行われる会話のことであり、話し言葉によって行われる。

話し言葉とは、我々が日常的な音声コミュニケーションの手段として用いている言葉であり、書き言葉に対して様々な違いを持つ。書き言葉は、文章を書く時など、文字として表される言葉であり、定められた文法規則に従った表現である。これは、書き言葉は、人間の思考の伝達をリアルタイムではなく、一旦文字化という意識的、理論的作業を介してから行うためである。

これに対し、話し言葉による意志伝達は人間の思考に対してリアルタイムである。またそれ故に、省略、倒置、言い直し、言い淀みなどの文法規則に反した表現や、文意に影響しない単語 (付加語) を多く含む。これは話し言葉が、音声という人間の思考に直結したメディアによって表されるためであり、発話中の話の意図の変化、忘却等の影響を直接に受けるためである。

さらに対話音声では、一方の話者の発話中にもう一方の話者が発話を開始する現象（割り込み）や、あいづちなど、二人同時に発話するという現象が見られる。これは、書き言葉ではありえない対話音声独特の現象である。

これらの特徴から、我々が日常的コミュニケーション手段として行っている対話は、書き言葉と比較して曖昧かつ複雑であり、理解しずらいもののように思われる。確かに、対話音声をそのままテキスト化した場合、単語の意味や、発話の意図、文の構造などが掴みにくく、理解が困難なことが多い。それにもかかわらず、我々が普段の対話で不自由を感じること無く、スムーズに意志の伝達を行えるのは、音声に含まれる抑揚情報の働きによると思われる。

抑揚情報とは、音韻情報や言語情報といった言語的情報と共に音声に含まれる音声固有の情報であり、声の高さや大きさ、間 (ポーズ)、発話の早さなどがある。対話音声を、そのままテキストに書き起こした場合に理解が困難になるのは、これらの抑揚情報が欠落するためである。このことから、対話音声理解システム実現のためには、抑揚情報の利用が必要不可欠であるといえよう。

2.2 対話音声における抑揚情報の役割

抑揚情報には、以下のような機能があると考えられる [4]。

1. 単語、文節レベル

   同音意義語の区別

2. 文レベル

   文の種類の区別、文の構造の明確化、一つの意図を伝える範囲の明確化

3. 話題の展開、転換の表示

話題内の起承転結、異なった話題への転換を明確にする。

4. 心理状態の表現

喜怒哀楽や強調、付加語とあいまって心理状態などを表す。

5. 対話の制御

発話権移動のコントロールやあいづちなどにより、対話をスムーズにする。

本研究では、これらの内4と5を中心として検討を行なう。

3 あいづち、発話権の制御、付加語

3.1 あいづち

あいづちは、自分が相手発話を認識していることを明示する行為であり、これにより相手の不必要な確認をなくし、相手の発話を促進するという機能があると考えられる。したがって、あいづちは対話をスムーズに行なううえで不可欠なものだと言える。

対話音声理解システムにおける対話においても、システムが利用者発話中に適切なあいづちを打つことにより、よりスムーズな対話が可能になることが予想される。

3.2 発話権の制御

発話権とは、対話において、発話する権利のことを言う。対話の制御を行なうということは、発話権の移動を適切に行なうということである。通常は二人の話者の内どちらか一方が発話権を持ち、その発話が終了した時点で、もう一方の話者に発話権が移る。しかし我々の日常対話においては、相手の発話中に発話を開始したり(割り込み)、相手の発話に対してあいづちを打ったりと、発話権の移動は単純ではない。それにもかかわらず、我々は対話を行なっている時、特に意識することなく発話の開始、終了を繰り返し、スムーズに意図の伝達を行なうことができる。これは、無意識のうちに抑揚情報によって発話権の制御が行なわれているためであると考えられる。

発話権を制御する行為としては以下のようなものが考えられる[2]。

1. 発話権の譲渡

自分の発話を終了するとともに、相手に発話権を渡す(相手の発話を求める)。

2. 発話権の放棄

発話を終了することにより、発話権を放棄する。

3. 発話権の獲得

相手が発話権を持っている(発話中である)とき、そこに割り込むことにより、発話権を獲得する。

4. 発話権の維持

発話中に、相手のあいづちを促すような話し方をすることによって発話権を維持する。

5. 相手の発話権の維持

相手の発話中にあいづちを打つことにより、相手の発話権を維持する。

3.3 付加語

対話音声をテキスト化した場合、削除しても特に意味が変わらない語句がある。ここではこれらを総称して「付加語」と呼ぶ。具体的には「えーと」や「あのー」といった間投詞、文末に現れる「ね」や「よ」などの終助詞、「はい」や「うん」などのあいづちに用いられる感動詞などがある。これらは文の意味構成には直接関与せず、又、書き言葉には通常現れないため、これまでその機能があまりはっきりとはとらえられず、「不要語」などと呼ばれ不必要なものとされていた。

しかし、どの様な対話音声にも付加語は含まれることから、本研究では付加語にも何らかの役割があると考え、付加語を積極的に利用した対話音声理解を考える。

本研究では、特に出現頻度の高かった「ね」に対象を絞り検討を行なった。

4 対話音声データベースの作成

本研究では、対話音声データの解析結果より以下の情報を含むデータベースを作成し、これをもとに後に示す検討を行なった。解析の対象としたのは、2名の女性がテーマや発声条件などの制限を全く課せられずに、非対面で行なった約40分の対話の一部である。

1. 音韻情報

   聞き取りにより、音韻区間を切り出し時間軸上に記述した。

2. 抑揚情報

   (a) ピッチパターン
   始端ピッチ、終端ピッチ、最大ピッチ

   (b) ポーズ
   位置、長さ

   (c) 音声パワー

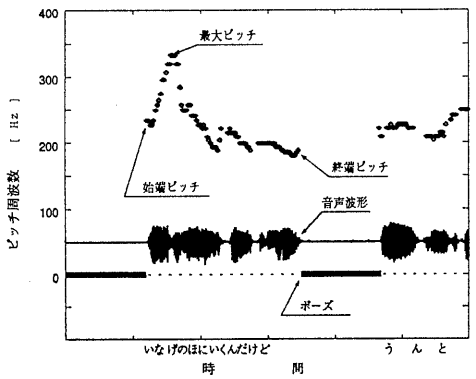

図1. 抑揚情報の抽出結果の例.

3. ラベル

   対話音声データにおいて、何らかの特徴がある部分に対して、聞き取りによる判断に基づき以下のようなラベルをつけた。

   (a) 発話開始部

   cut : 割り込み (cut in)
   c-res : 割り込み的あいづち
   res : あいづち (responce)
   sta : その他の発話開始 (start)

   ここで割り込み的あいづちとは、意味的にはあいづちと同様であるが、割り込みの様に一旦発話権を獲得して発話を行なうものである。我々の対話ではあいづちと同様に扱われるが、対話音声理解システムでは、一旦発話権が移動することから割り込みとして扱う方が適切であると考えられる。

   (b) 発話終了部

   tra : 発話権を譲渡しての終了
   ren : 発話権を放棄しての終了
   c-end : 割り込まれての終了
   end : その他の終了

   (c) 発話継続部

   cont : 発話権を維持しての継続

   (d) 助詞「ね」の種類

   A1 : 相手の同意を求めるもの (返答要求型)
   A2 : あいづちに使われるもの (あいづち型)
   B : 相手の返答を期待しないもの (非返答要求型)

   (e) ピッチパターン

   助詞「ね」についてピッチパターンを急上昇型、緩上昇型、下降型に区別した。

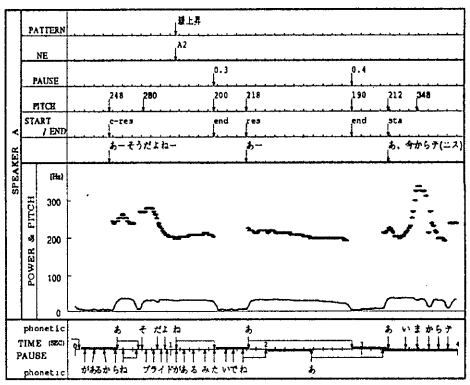

図2. データベースの例.

図はspeaker Aについての抑揚情報、ラベルのみを示したものであるが、実際は下半分にspeaker Bについての抑揚情報、ラベルが同様に記されている。

5 抑揚情報の解析

5.1 あいづちと抑揚情報

対話中におけるあいづちは、どの様な場合にどの様なタイミングで打たれているかを解析した。解析の対象としたのは、対話音声データベースにおいて、あいづちのラベル (res) がついている発話 (27 例) と、割り込み的あいづちのラベル (c-res) がついている発話 (11 例) である。

解析の結果、あいづちが打たれるのは、相手発話中に (1) 発話権の維持を示す発話、(2) 対話中のキーワードとなる単語、のいずれかが現れた直後であることが解った。発話権の維持と抑揚情報の関係については後で述べる。対話中のキーワードとなる単語と抑揚情報の関係については、抑揚情報を積極的に利用して発話中の重要単語を抽出する方法 (ワードスポッティング法 [1]) についての研究がすでになされているので、本研究では検討を行なっていない。

相手発話中に、あいづちを打つべき発話が現れてから、あいづちが打たれるまでの時間を図 3 に示す。図 3 より、あいづちを打つべき発話が現れてから約 0.4 秒以内にあいづちが打たれていることがわかる。これは、システムがあいづちを出力する際のタイミングの目安になるとともに、ユーザーが遅さを感じないシステムを実現する際の一つの条件を示しているともいえる。

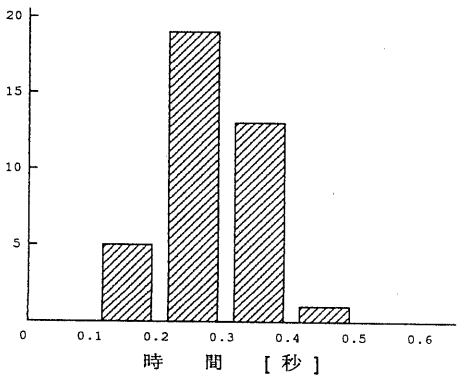

図 3. あいづちが打たれるまでの時間 .

5.2 割り込みとあいづちの判別

システムの発話中にユーザーが発話を開始した場合、それが割り込みであるかあいづちであるかの判別を、抑揚情報を用いることにより行なうことができるか、という可能性について検討した。

対象としたのは、作成した音声対話データベースにおいて、割り込みのラベル (cut) が付いている発話 (15 例) と、あいづちのラベル (res) が付いている発話 (27 例)、割り込み的あいづちのラベル (c-res) が付いている発話 (11 例) である。

5.2.1 始端ピッチと最大ピッチ

割り込み、あいづち、割り込み的あいづちそれぞれについて、始端ピッチと最大ピッチの平均を求めた。結果を図 4 に示す [3]。

図 4 より、始端ピッチでは 3 つの場合ともさほど大きな違いは見られないが、最大ピークピッチにおいては、割り込みは大きくあいづちを上回っていることが解る。割り込み的あいづちはその中間の値をとっているが、それでもあいづちに比べ、かなり大きな値といえる。これより、話者は、割り込みによって発話権を獲得しようとするとき、高いピッチによってその意図を表しているといえる。また、あいづちによって相手の発話を促進しようとするときは、相手の発話を妨げないようピッチを低く抑えるといえる。さらに、割り込み的あいづちは、高いピッチにより一時的に発話権を獲得し、相手の発話に対する反応を積極的に伝えようとすると考えられる。

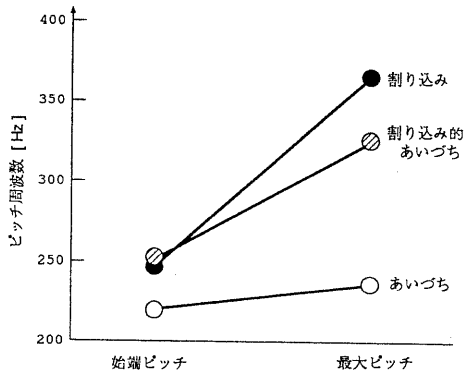

図 4. 始端ピッチと最大ピッチ .

5.2.2 始端ピッチと最大ピッチの分布

5.2.1 で、割り込みや、割り込み的あいづちによって発話権を獲得しようとするとき、高いピッチがその意図を表すことが示された。この結果を、対話音声理解システムにおいて、割り込みの検出に利用しようとした場合、発話の開始からどれほどの時間で最大ピッチが出現するかということは、割り込みの検出に要する時間に直接関係するため重要である。そのため、割り込みの開始から、どれほどの時間で最大ピッチが出現するかを分析した。結果を図 5 に示す。対象としているのは、割り込みのラベル (cut) が付いている発話 (15 例) と、割り込み的あいづちのラベル (c-res) が付いている発話 (11 例) である。

図 5 より、分析の対象とした発話では、発話の開始から最大ピッチ出現までの時間は、最長でも 1 秒以内であることがわかる。これより、発話開始から 1 秒以内の最大ピッチを調べることにより、割り込みや割り込み的あいづち、あいづちが区別できる可能性が示された。そこで、始端ピッチを横軸に、割り込み開始から 1 秒以内の最大ピッチを縦軸に取り、割り込みと割り込み的あいづち、あいづちがどのように分布するかを分析した。結果を図 6 に示す。

図 6 より、割り込み、割り込み的あいづち、あいづちでは明らかに分布が異なり、始端ピッチと最大ピッチにより、これらの判別を行なうことができる可能性が示された。

5.3 発話の終了と継続

発話中にポーズが現れた時に、それが発話の終了である場合と、発話継続中の文章の区切りである場合では、抑揚的にどのような差異があるかについて分析を行った。

対象としたのは、作成したデータベースにおいて、発話権の譲渡、放棄のラベル (tra、ren) が付いている部分 (27 例) と、発話継続のラベル (cont) が付いている部分 (25 例) である。

発話の終了の場合と継続の場合では、ポーズが現れる直前のピッチに何らかの特徴があると考えられる。そこで、終端ピッチをパラメータとして解析を行なった。

5.3.1 終端ピッチの分布

発話終了部と発話継続部のそれぞれについて、終端ピッチがどのように分布するかについて分析を行った。結果を図 7 に示す。

図 7 より、発話の終了の場合と継続の場合では、明らかに終端ピッチの分布に違いが見られる。しかし、同一の範囲に分布している部分も多く、終端ピッチにより発話の終了と継続を明確に判断するのは困難であるといえる。

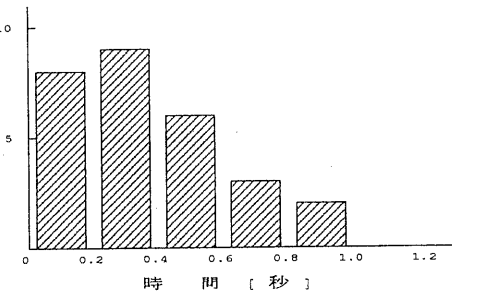

図 5. 始端から最大ピッチ出現までの時間 .

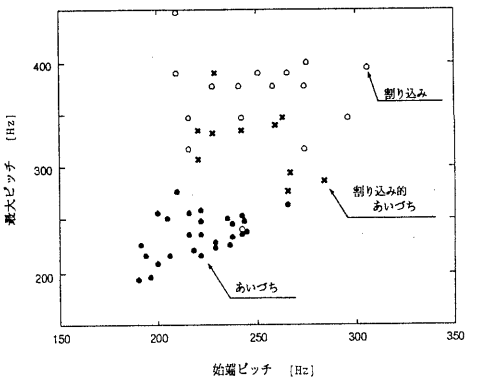

図 6. 始端ピッチと最大ピッチの分布 .

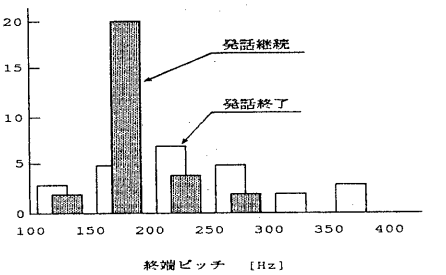

図 7. 終端ピッチの分布 .

5.3.2　終端ピッチと発話終了の割合

5.3.1 により、終端ピッチは、発話の終了と継続の場合である程度の違いを持つが、それにより発話の終了と継続を明確に判別するのは困難であることが示された。そこで、発ｖ話の終了と継続に対して、終端ピッチがどの様な関係を持つかについて解析を行なった。終端ピッチに対する発話終了の割合を図 8 に示す。

図 8 より、発話終了の割合は終端ピッチが高くなるほど大きく、また終端ピッチが極端に低い場合にも大きいことが解る。これは、発話の終了には、発話権の放棄と発話権の譲渡の 2 通りがあるためであり、発話権の放棄の場合は低いピッチで終了し、発話権の譲渡の場合は高いピッチで、相手に問いかけるように終了するためだと考えられる。

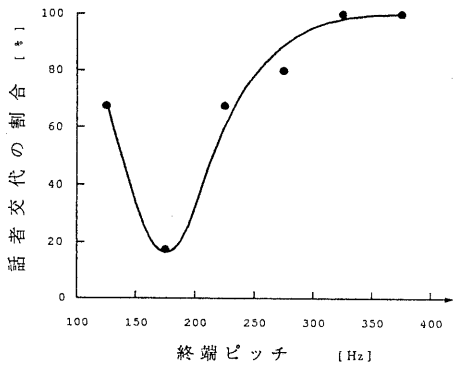

図 8. 終端ピッチに対する発話終了の割合．

5.3.3　終端ピッチと次発話開始までの時間

終端ピッチと次発話開始までの時間、さらには発話終了の割合がどの様な関係にあるかについて解析を行なった。ここで次発話開始までの時間とは、一方の話者の発話の終了から、もう一方の話者が発話を開始するまでの時間である。

図 8 に次発話開始までの時間のグラフを加えたものを、図 9 に示す。

図 9 より、発話の終了か継続かの判断がしにくいところでは、次発話の開始が非常に早いことが解る。これは、いい変えれば、次発話が非常に早く始まった場合に限り発話の終了になる、ということである。つまり、終端ピッチに発話の終了や継続を表す明確な情報がなかった場合、それが発話の終了か継続かという判断は、相手話者の発話を行なう意志の有無によって決定されるといった、曖昧なものであると考えられる。

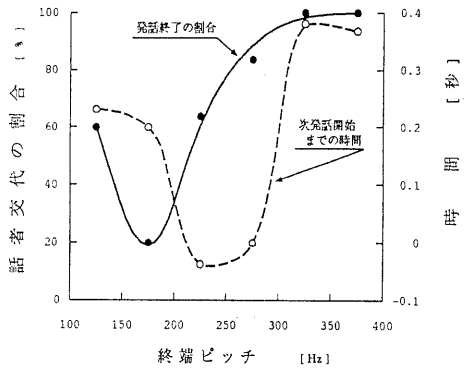

図 9. 終端ピッチに対する次発話開始までの時間と発話終了の割合．

5.4　付加語「ね」

対話中に現れた「ね」を、返答要求型 (A1)、あいづち型 (A2)、非返答要求型 (B) に分類し、それぞれのピッチパターンが、急上昇型、緩上昇型、下降型のどれにあたるかを調べ、整理した結果を表 1 に示す。

この結果から A1 は急上昇型、A2 は緩上昇型、B は下降型がそれぞれ発生件数が多いことがわかる。特に A2 は約９０％を緩上昇型が占めている。

またＡ類をまとめて考えると、急上昇型は約３９％、緩上昇型は５３％となり上昇型が全体の約９３％を占めていることから聞き手の知識状態が関与している時は上昇型となることが推測される。

以上のことから整理すると、「ね」のピッチ周波数は、

1. 聞き手に返答を求めるときは上昇する
2. あいづちを打つときは緩やかに上昇する
3. 聞き手の返答を期待しないときは下降する

ことが推測でき、これに注目することにより、ユーザーがシステムに対して発話を要求しているか否かの判断がある程度可能であることが示された。

ラベル	急上昇型	緩上昇型	下降型	計（件）
A1	10	5	2	17
A2	1	10	0	11
B	0	4	8	12

表 1: ピッチパターンごとの発生件数（計40件）

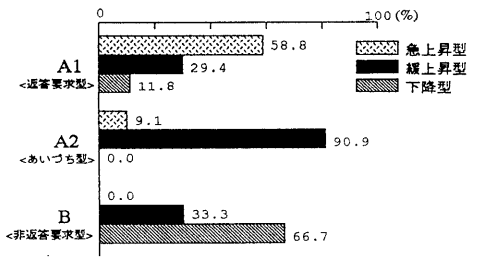

図 10.「ね」の種類別ピッチパターンの発生割合.

6 対話理解の処理系案

対話音声を理解するシステムの構成として、分散協調問題解決型のエージェントシステムの構成を提案する (図 11)。対話音声の理解には話題の展開経過が重要なことを考慮し、文節的情報の処理、" かぶせ素性" の処理、話題の情報の提供、対話の経過情報の提供、次入力の予測などの各機能を持ったエージェントによる協調処理方式である。各エージェントには、さらに学習機能を持たせ、過去と類似の入力に対しては迅速に応答させることにより実時間性に優れた理解システムの実現が図れると思われる。

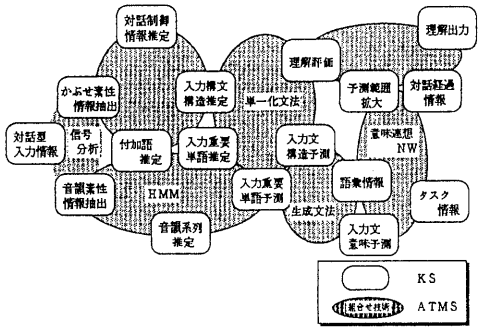

図 11. 対話音声理解の処理系案.

7 あとがき

本研究では、音声に含まれる抑揚情報が、対話音声理解にどの様に利用できるかについて検討を行なった。しかし、1つの対話データについてのみ解析を行なったため、割り込みとあいづちを区別する際のいき値の設定などの、定量的な検討は行なっていない。今後個人差にどの様に対処するかなども含め、定量的な判断の方法の研究が必要である。

また、本研究で解析に用いた対話データは、特に目的を持たない自由対話であった。しかし、対話音声理解システム上で行なわれる対話は、何らかの目的を持った対話であると考えられる。より実際的な結果を得るためには、対話音声理解システム上で行なわれる対話を想定したデータを解析する必要がある。

謝辞

本研究は、文部省科学研究費重点領域研究" 音声対話" 及び旭硝子財団の助成による。また実験においては、当研究室の石田祐子さんに協力してもらった。

参考文献

[1] 小松 明男、" 会話音声理解によるマンマシンインターフェースに関する研究 "、早稲田大学 博士論文 (1991)

[2] 菊池 英明・小林 哲則・白井 克彦、" 音声対話におけるタイミングの自由度と割り込みの扱い "、文部省重点領域研究「音声対話」D班会議資料 (1993)

[3] 中嶌 信弥・塚田 元、" 協調的対話における発話パタンの特徴分析"、人工知能学会研究会資料 (1993)

[4] 市川 熹、" 日本語ガーデンパス文の聴読理解比較 "、音学講論 I-Q-17(1994)

[5] 松永 隆雄・北澤 茂良、" 対話音声の音響的非言語情報の分析 "、日本音響学会講演論文集 (1993)