

# 音声対話によるテキスト内容の伝達方法

杉山 聡<sup>†</sup>, 堂坂 浩二<sup>†</sup>, 川端 豪<sup>†</sup>

本論文では、テキストの内容を音声で伝達するユーザ・インタフェースの1つとして、話しことば対話の特徴を利用してテキスト内容を伝達するインタフェースを実現するための方法を提案し、そのインタフェースの有効性を実験により示す。本方法の有効性を検証するために、話しことば対話の特徴を利用して伝達するシステムと、文単位で読み上げて伝達するシステムを作成し、伝達効率と伝達の達成率を比較する実験を行った。その結果、話しことば対話の特徴を利用して伝達する方法は、読み上げて伝達する方法と比較して、より効率的に必要な情報を伝達できることが分かった。

## A Method for Conveying the Contents of Written Texts by Spoken Dialogue

AKIRA SUGIYAMA,<sup>†</sup> KOHJI DOHSAKA<sup>†</sup>, and TAKESHI KAWABATA<sup>†</sup>

This paper proposes a method for conveying the contents of written texts by spoken dialogue and shows the effectiveness of this method based on the result of the experiment in which a system using this method was compared with a system using the method of reading aloud individual sentences one by one. In this experiment, we compared two systems in terms of the efficiency of conveyance and the task achievement. The result of the experiment shows that the method for conveying the contents by spoken dialogue can convey the information the user requires more efficiently than that of reading aloud individual sentences.

### 1. はじめに

音声対話で情報を授受できるインタフェースは、計算機のユーザ・インタフェースの選択肢の1つとして実現が期待されている。話しことばで行われる音声対話は、人間同士が情報を交換する際に日常用いている方法であり、話しことば対話で情報交換ができるインタフェースには、ユーザが使う際に新たに習熟せずに利用できるという利点がある。

計算機を利用して引き出すことのできる情報の知識源として、電子化された新聞記事や NetNews で配信されている記事や Web コンテンツとして公開されている文章などの大量のテキストが利用できるようになってきた。様々なユーザが、計算機を通じてこれらのテキストにアクセスできるようになり、単にテキストをディスプレイに表示するのではなく、テキストの

内容を知的にユーザに提示できるシステムの必要性が高まっている。

テキストの内容を利用者に伝達するシステムの1つとして、音声を伝達メディアとしてテキストの内容を伝達するシステムが考えられる。音声による情報の伝達には、車の運転中などのディスプレイに表示されたテキストを読むことが困難な状況や、視覚障害を持つ方に対しても有効であるなどの重要な役割がある。

音声による伝達では、聞き取りや理解の失敗などが発生することが避けられないため、一度の伝達に必要な情報を受け手に理解させられるとは限らない。必要な情報を確実に伝達するためには、ラジオのように情報の与え手が一方的に情報を伝達する方法は適さない。確実な情報の伝達には、情報の受け手が与え手に働きかけることができる仕組みが必要である。

人間は、情報の与え手と受け手が同時に発話できるような場面では、両者が協調しながら伝達を行っている。たとえば、情報の与え手は受け手の状態を確認しながら情報を伝達し、受け手は、あいづちなどを打つことによって、理解状態を情報の与え手に示すことが行われる。また、受け手が理解に失敗した場合には、問い返しを行うことで理解に失敗したことを示し、与

<sup>†</sup> NTT 基礎研究所

NTT Basic Research Laboratories

現在、NTT コミュニケーション科学基礎研究所

Presently with NTT Communication Science Laboratories

現在、NTT サイバースペース研究所

Presently with NTT Cyber Space Laboratories

え手はその問い返しに対して応答する。このような話しことば対話による情報伝達は、日常生活の中で自然に行っている方法であり、音声による情報伝達を確実に行う有効な手段の1つである。

テキストの内容を伝達するシステムを構築する際にも、読み上げシステムを利用して、ラジオのように伝達するシステムが考えられる。この場合、伝達を確実に行うためには、ユーザが理解に失敗した場合に対応できる仕組みを用意する必要がある。

本論文では、話しことば対話を持つ、情報伝達を確実に行うことに役立つ機能に注目し、話しことば対話の特徴を利用して、音声でテキストの内容を伝達する方法を提案する。2章で、情報の伝達に関する研究を概観し、本研究の位置づけを行う。3章で、本論文で利用した話しことば対話の特徴について説明して、4章でそれらの話しことば対話の特徴を利用した情報の伝達を実現する方法を説明する。本論文で提案する方法で情報を伝達するシステムと、テキストの読み上げによって伝達するシステムを実装し、情報の伝達効率と伝達の達成率を比較する実験を行った。5章でその実験方法について説明し、6章で実験結果を示して考察を行う。

## 2. 関連研究

これまでに多くの音声対話システムに関する研究が行われている<sup>1)~3)</sup>。本研究は、音声対話システムにおけるシステムの言語生成に焦点をあてた研究である。本研究と、従来の対話システムにおける言語生成の研究との違いは以下の点である。従来の対話システムの代表的な言語生成方法では、発話内容をプランニングの手法で決定し、言語表現を文法やテンプレートなどを利用して生成する。システムの知識は、システム用に定義した表現形式で記述しておく。これらの研究では、旅行計画の立案や経路の案内などのように、必要なすべての知識を論理的な表現形式で記述できるドメインを対象として行われている。それに対して本論文では、新聞記事などのテキストをシステムの知識源として、テキストの内容を伝達するための言語表現の生成方法を提案する。従来の言語生成の研究とは、ユーザに伝達する内容を記述した、システムの持つ知識の表現形式が異なる。

テキスト内容の提示に関する研究には、テキストのハイパーテキスト化に関する研究<sup>4)</sup>と自動要約に関する研究<sup>5)~8)</sup>がある。前者に関して、ユーザは、ハイパーテキスト化されたテキストを専用のブラウザを利用して読むことによって、プレーンなテキストを順番

に読むよりも容易に情報を獲得できることがある。たとえば、ある用語とそれに関連する部分とが関連付けられていれば、その用語をマウスなどで指定することによって関連した部分を読むことができる。テキストのハイパーテキスト化の研究の主な観点は、このようなテキストあるいはテキスト群の構造の抽出と構造化を、自動的あるいは半自動的に行うための方法を確立することにある。また、自動要約に関する研究の主な観点は、談話構造を解析する方法の確立や、重要な節や文を適切に抽出する方法を確立することにある。テキスト内容の提示は、要約した結果をディスプレイに表示することで行う<sup>9)</sup>。それに対して本論文では、音声対話において、システムとユーザが同時に発話することが可能な状況での、テキストの内容の提示方法を提案する。ユーザが即時的に反応してくる状況で適切な伝達を行うためには、伝達の単位の決定方法や、伝達途中に発生するユーザの要求への対応方法などが問題となる。

テキストの内容を音声で伝達することに関する研究には、テキストの読み上げに関する研究がある<sup>10)</sup>。読み上げに関する研究における主な観点は、音声合成の質を向上することや、構文解析などを利用して自然な読み上げを実現する方法を確立することにある。現在の読み上げシステムは、読み上げる部分をユーザに指定されることによって動作する。読み上げ箇所は、ディスプレイに表示されたテキストの一部分をマウスなどで指定するなどの方法で指定する。また、読み上げの繰返しなども、マウスなどを使って指定されることで動作する。それに対して本論文の伝達方法では、伝達すべきテキストが与えられた後は、システムが自動的に、テキストの中から次に発話する部分を決定する。また、ユーザは、繰返しなどの要求を音声でシステムに伝達する。システムが次に発話する部分を決定する方法や、音声によるユーザの要求を処理する方法が問題となる。

## 3. 話しことば対話の特徴

テキストの内容を音声で効率良く伝達するためには、話し手と聞き手の協調が必要である。たとえば、テキストの内容を知りたい人が能動的にテキストを読む場合には、理解に困難が生じても必要なところを自分で読み返せばよいが、システムがユーザに音声でテキストの内容を伝達する場合には、システムに対して理解に困難が生じたことを伝え、必要な部分を聞き直すことが必要になる。テキストの内容を音声で伝達するインタフェースには、このようなシステムとユーザの協

NTT では、超高速ホログラフィ動画記録を可能にする新しい結晶材料を発見しました。この新材料を用いると、極低温において、書き込みレーザー光の周波数を変えて記録媒体に情報を重ね書きする「周波数多重記録」が可能となり、静止画像なら約 1000 万枚（テレビ映像換算 100 時間分）が記録できます。さらに光反応速度が速いため、周波数を連続的に変化させることにより実際に動いている物体の記録も実現でき原理的には 10 億分の 1 秒の高速記録も期待されます。NTT では、光源として精密制御色素レーザーを用いた実験を行い、約 20 秒のホログラフィ動画の記録/再生に成功しました。

図 1 テキストの例

Fig. 1 Example of written text.

調作業を効率良く行うことができる仕組みが必要である。人間は、話しことば対話によって、音声のみでも効率的に情報を伝達することができる。本論文では、話しことば対話には、話し手と聞き手の協調作業を効率良く行う機能を持った特徴があると考え、話しことば対話の特徴を利用してテキストの内容を音声で伝達する方法を提案する。本章では、本論文で利用した話しことば対話の特徴をまとめる。

本論文では、1) 発話の単位が短い、2) 聞き手が任意のタイミングで問い返すことができる、3) 情報を段階的に詳細化する、という 3 つの話しことば対話の特徴を、音声によるテキストの内容の伝達に利用した。それぞれの特徴が話し手と聞き手の協調作業の効率化に果たす役割と、特徴を利用することによって可能となる伝達方法を説明する。

### 3.1 発話単位

話しことば対話において情報の与え手は、与え手と受け手の相互理解を確認しながら伝達を進める<sup>11)</sup>。受け手の理解力には限界があるため、一度に多くのことを伝達しても、そのすべてが理解されるとは限らない。受け手の理解力を考慮した伝達を行い、受け手が理解したことを確認しながら漸次的に伝達することが必要である。即時的な反応が可能な話しことば対話では、理解したことを相手に伝える手段の 1 つとしてあいづちが用いられる。あいづちの研究では、あいづちがうたれた発話には、終助詞などをともなう文末表現が多いことや<sup>12)</sup>、音調的な特徴がみられることが指摘されている<sup>13)</sup>。与え手は、これらの特徴を明示することで、受け手に相互理解を確認する機会を与えていると考えられる。

本論文では、ポーズや対話相手の発話で分割された発話のまとまりを発話単位と定義する。このように発話単位を定義すると、書きことばに見られる文よりも短い単位が発話単位となりうる。人間同士の話しことば対話に用いられる発話単位は、書きことばに見られる文よりも短いことが指摘されている<sup>14),15)</sup>。

話しことば対話のように短い発話単位でテキストの内容を伝達する例として、図 1 に示したテキストの内容を文よりも短い発話単位で伝達すると図 2 のよう

システム: NTT では <P>  
 ユーザ: はい  
 システム: 超高速ホログラフィ動画記録を可能にする <P>  
 新しい結晶材料を <P>  
 ユーザ: はい  
 システム: 発見しました <P>

図 2 短い発話単位で伝達する対話例

Fig. 2 Example of dialogue where information is conveyed in small utterance unit.

な対話となる。この対話例では、情報の与え手であるシステムは「NTT では」「可能にする」「結晶材料を」「発見しました」の後で、発話単位の切れ目を明示しながら伝達し、受け手であるユーザは「NTT では」と「結晶材料を」の後で理解を示すあいづちを発話している。

### 3.2 問い返し

話しことば対話では、受け手が発話するタイミングに制限が加えられていないため、受け手は任意のタイミングで問い返すことができる。情報の与え手が、受け手にとって分かりやすいと思われる発話単位で発話し、相互に確認をとりながら伝達を進めたとしても、理解の失敗が発生することは避けられない。受け手は、理解困難な状態になった場合、問い返しを行うことで与え手にそのことを伝える。与え手は、問い返しを受けた時点で伝達を中断し、要求に対する応答を行う。受け手は理解に困難を感じた時点でそのことを与え手に伝えることができるため、理解できないまま情報の伝達が進むことが避けられる。

任意のタイミングで問い返しができるという話しことば対話の特徴を利用することで可能となる、テキスト内容の伝達の例を図 3 に示す。例では「超高速ホログラフィ動画記録を可能にする」とシステムが発話したところで、ユーザが問い返しを行っている。システムはこの問い返しを受け、伝達を先に進めずに、問い返しに対する応答を行っている。また、問い返しに対する応答も、文の先頭から繰り返すのではなく、ユーザにとって必要だと思われる部分のみを推定して応答している。

システムが明示する発話単位の切れ目を <P> で示す。

システム: NTT では <P>  
 ユーザ: はい  
 システム: 超高速ホログラフィ動画記録を 可能にする <P>  
 ユーザ: もう一度  
 システム: 超高速ホログラフィ動画記録を 可能にする <P>  
 ユーザ: どうぞ  
 システム: 新しい 結晶材料を 発見しました

図 3 問い返しの例

Fig. 3 Example of dialogue where the user requests the system to repeat systems utterances.

システム: NTT では <P>  
 ユーザ: はい  
 システム: 新しい 結晶材料を <P>  
 ユーザ: 結晶材料  
 システム: 結晶材料というのは <P>  
 ユーザ: はい  
 システム: 超高速ホログラフィ動画記録を  
 可能にするというものです

図 4 情報の詳細化の例

Fig. 4 Example of elaboration on contents.

テキストを伝達する方法に文単位で伝達する方法もあるが、文単位の伝達では、このような文の途中の問い返しに対応できない。この場合、ユーザは unnecessary 情報を聞かされることになる。

### 3.3 情報の詳細化

話しことば対話で用いられる談話戦略の 1 つに、ある事柄に関する情報を一度に伝達するのではなく、受け手の反応に応じて少しずつ情報を詳細化しながら伝達するという談話戦略がある。情報を詳細化しながら伝達することによって、受け手の理解力を考慮した伝達が可能となる。

また、話しことば対話では、受け手の反応を観察しながら伝達することができる。このため、与え手は受け手が興味を示すまでは詳細情報を伝えずに伝達を進める戦略をとることが可能である。詳細情報を省略しながら伝達し、受け手が必要とする詳細情報のみを伝達することによって、効率的な情報伝達が可能となる。

テキストの内容伝達においても、詳細情報を省略しながら伝達することによって、図 4 に示すような伝達が可能となる。この例では、システムは最初、「結晶材料」の詳細情報である「超高速ホログラフィ動画記録を可能にする」という部分を省略して伝達している。この例ではユーザが「結晶材料」と問い返すことによって、「結晶材料」に対して興味を示したので、省略した部分の伝達を行っている。ユーザが興味を示さない場合は、そのまま次の文の伝達に移る。

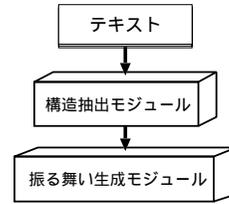


図 5 処理過程

Fig. 5 Flow of processing.

## 4. 話しことばへの変換方法

本章では、3 章で説明した話しことば対話の 3 つの特徴を利用して、テキストの内容をユーザに伝達する方法を説明する。本論文で用いる方法は、話しことば対話の特徴を利用した伝達を計算機で実現する 1 つの方法であり、人間が実際に処理している過程をシミュレートすることを意図したものではない。本方法は、テキストから発話単位を決定したり、問い返しに対する応答を生成する際に、人間が行っているであろうすべての要因を加味したものではなく、容易にモデル化できる一部の要因のみを利用したものである。

本論文で提案する伝達方法の処理過程を図 5 に示す。処理過程は 2 つのプロセスに分けられる。1 つ目のプロセスは、話しことば対話の特徴を利用した発話の生成に必要な情報を、テキストから抽出するプロセスで、構造抽出モジュールで処理する。構造抽出は、テキストの内容伝達の進行とは無関係にあらかじめ抽出しておくことが可能であるため、ユーザへの伝達を始める前に行う。2 つ目のプロセスは、対話の進行に応じてシステムの次の振舞いを決定するプロセスで、振舞い生成モジュールで処理する。

### 4.1 構造抽出

構造抽出モジュールは、プレーンテキストを入力とし、話しことば対話の特徴を利用した発話の生成に必要な構造を抽出して、抽出した情報をテキストに付加した構造化テキストを出力する。本論文で構造化テキストと呼ぶものは、形態素情報、文節の切れ目、文節間の係り受けに関する情報を付加したテキストである。たとえば、図 1 で示したテキストを構造抽出モジュールに入力すると、図 6 のような構造化テキストが出力される。

話しことば対話の発話単位は、テキストに現れる文よりも短い。スムーズに伝達を行うためには、ユー

実際はリスト形式で表現しているが、ここでは分かりやすいように木構造で示す。また、形態素情報は省略し、テキストの 1 文目のみを示す。

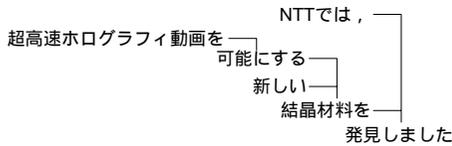


図 6 構造化テキスト

Fig. 6 Example of structured text.

ザにとって理解しやすい発話単位を決定する必要がある。本方法では、意味的なまとまりを利用して発話単位を決定する。意味的なまとまりをテキストから抽出するために、本方法では文節の係り受け関係を利用する。文節の係り受け関係は、ユーザから問い返しがあつた場合の適切な応答を推定する際にも利用する。

情報を段階的に詳細化しながら伝達するという特徴を利用してテキストの内容を伝達するためには、ある事柄の詳細情報を与えている箇所を抽出することができなければならない。本方法では、テキストからある事柄とその詳細情報とを抽出する手段として、連体修飾節の構造を利用する。被修飾語が“ある事柄”であり、連体修飾節が被修飾語に対する詳細情報を与えると仮定している。

以上をまとめると、本方法で発話を生成する際に利用する構造は、文節構造と係り受け構造と節の構造である。本方法では、文節構造や係り受け構造などを抽出するツールとして KNP<sup>16)</sup> を利用した。また、KNP を利用するための形態素解析ツールとしては JUMAN<sup>17)</sup> を用いた。

#### 4.2 振舞い生成

話しことば対話の特徴を利用してテキストの内容を伝達するシステムの振舞いを生成する方法を説明する。

振舞い生成モジュールは、構造化テキストとユーザの発話意図を入力として、テキストを発話単位に分割して生成することと、つねにユーザの問い返しを受け付け、問い返しに適切に回答することと、詳細情報は省略しながら伝達し、ユーザから要求があつた場合には詳細情報を伝達することを行う。以降これらの振舞いを生成する方法を、3章で説明した3つの話しことば対話の特徴ごとに説明する。

##### 4.2.1 発話単位

話しことば対話で観察される発話単位は文よりも短いという分析がある。文を発話単位としたのでは、文の途中で理解できないなどの障害が発生したときに

即座に対応できない。また、問い返しに対する応答もつねに文単位で行ったのでは冗長な対話になる。したがって、句点のみを手がかりとして発話単位を決定したのでは不十分である。逆に、たとえば、文節を発話単位だとして、文節ごとにポーズを置きながら発話したのでは、対話が間延びしてしまう。本項では、意味的なまとまりを手がかりとして、いくつかの文節をまとめて発話単位を決定する仕組みを説明する。

意味的なまとまりやすさの尺度として文節の係り受け関係を利用する。係り受け関係を木構造で表したときに、同じ枝にある文節は意味的にまとまりがあるため同じ発話単位で発話した方がよく、異なる枝にある文節は別の発話単位とした方がよいという方針で発話単位を決定する。具体的には、隣り合う2つの文節の分かれやすさを表す値  $dist$  を次のように定義し、 $dist$  が閾値を超えたところで分割する。

$$dist(b_n, b_{n+1}) = \begin{cases} \alpha \times (depth(b_n, b_{n+1}) - 1) + weight(b_n) & (n = k) \\ dist(b_{n-1}, b_n) + \alpha \times (depth(b_n, b_{n+1}) - 1) + weight(b_n) & (n > k) \end{cases}$$

$b_i$  は文の  $i$  番目の文節を表し、 $dist(b_i, b_{i+1})$  は  $i$  番目の文節と  $i+1$  番目の文節の分かれやすさの値を表す。式中の  $k$  は、現在計算している発話単位の先頭の文節の番号を表す。 $depth(b_i, b_{i+1})$  は係り受け構造の  $b_i$  と  $b_{i+1}$  の間の枝の数を表す。直接係り受け関係のある文節間の場合は1で、たとえば図6の“NTTでは”と“超高速ホログラフィ動画記録を”の  $depth(b_1, b_2)$  は4である。第3項は、文節  $b_i$  での分かれやすさの重みを表す。たとえば、 $b_i$  が提題の八格や読点があるなどの言語的に分かれやすい文節である場合や、4.2.3項で定義する重要語である場合には、 $weight(b_i)$  に正の値が与えられる。本論文では、 $\alpha$  の値を10として計算し、閾値は50とした。 $weight(b_i)$  に関しては、たとえば、 $b_i$  が提題の八格であるか、接続詞、副助詞の「など」「も」「でも」「なら」のいずれか、あるいは読点を含んでいた場合には、50とした。

発話の生成過程では、最後に伝えた文節を  $b_i$  とし、次の文節  $b_{i+1}$  を発話する前に、発話単位の切れ目としてポーズを置くか置かないかを決定するために上記の関数を用いる。閾値を超えない場合は連続して発話する。閾値を超えた場合にはポーズを置き、 $i+1$  番目の文節は、次の発話単位の先頭の文節とする。

たとえば、図6の文を伝える場合で、ユーザから

ユーザの発話に即時に対応するため、ユーザの発話の開始も監視している。

システム: NTT では <P>  
 システム: 超高速ホログラフィ動画記録を 可能にする <P>  
 システム: 新しい 結晶材料を <P>  
 システム: 発見しました <P>

図 7 発話単位の例

Fig. 7 Example of utterance unit.

のあいづちや問い返しがない場合は、図 7 のように、<P> のところでポーズを置きながら発話する。“NTT では”と“超高速ホログラフィ動画を”の  $dist(b_1, b_2)$  は 1 項は 0 で、2 項は  $\alpha \times 3$  である。本論文では、 $\alpha$  の値に 10 を用いたので、2 項の値は 30 になる。3 項の  $weight(b_1)$  には“NTT では”が八格であるので 50 が与えられる。 $dist(b_1, b_2)$  の値は 80 であり、閾値の 50 よりも大きいため、この 2 つの文節間で発話単位が分かれてポーズが置かれる。“超高速ホログラフィ動画を”と“可能にする”の  $dist(b_2, b_3)$  は、“NTT では”で発話単位が分かれているので 1 項は 0 で、2 項が  $\alpha \times 0$  で 3 項も 0 であるため、合計が閾値より低く、同じ発話単位としてまとめられる。

本方法では、文節ごとに伝達済か否かを管理する。システムの意図した発話単位の途中であいづちや問い返しが発話された場合は、あいづちが挿入された文節までを伝達済として記録する。そして、伝達済として記録された文節までを発話単位とし、次の文節から新たに発話単位を形成する。

#### 4.2.2 問い返しの扱い

任意のタイミングで問い返すことを許し、問い返しに対して適切に応答するための方法を、以下の 2 つの観点から説明する。1 つは、ユーザの問い返しを即座に認識し応答できるようにすることで、もう 1 つは、問い返しの応答として適切な箇所を特定することである。

前者に関する問題点の 1 つは、ユーザの発話の認識が遅れることによって、ちぐはぐな対話になってしまうことである。たとえば、ユーザの発話の認識を音声認識システムの認識結果の出力のみで行うと、図 8 のような対話になる可能性がある。この例では、「可能にする」を発話し終えた時点で、ユーザが「もう一度」という発話を開始しているにもかかわらず、この時点では音声認識システムが出力を出していないため、次の文節の「新しい」を発話している。「新しい」を発話した時点で先ほどの「もう一度」が認識されたため、「新しい」を応答している。

この問題点に対する解決方法として、音声認識システムとは別に音声区間を監視する装置を用意し、音声区間の始まりを検知した時点で振舞い生成モジュール

システム: 超高速ホログラフィ動画記録を 可能にする <P>  
 ユーザ: もう一度  
 システム: 新しい  
 システム: 新しい <P>

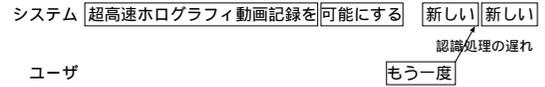


図 8 認識の遅れ

Fig. 8 Example of interpretational delay.

に信号を送ることとした。振舞い生成モジュールは、音声区間の始まりの信号を受けた場合、発話を中断する信号を振舞い実行モジュールに送り、音声認識システムの出力を待つ。音声区間の始まりを検知することにより、図 8 のようにユーザが発話をしているのに次の発話をする事なく、即座に応答を開始できる。

もう 1 つの観点である、問い返しの応答として適切な箇所を推定する方法を説明する。たとえば、「もう一度」という問い返しに対してつねに文頭から繰り返したのでは、冗長な伝達となってしまう適切な応答とはいえない。ユーザが必要としている箇所を含み、どこから繰り返したのかが分かりやすいところから繰り返すことが必要である。

本方法ではこの問題に対する解決方法として、問い返しに対する適切な応答部分の推定に、係り受け構造を利用する。問い返しが発話されたときにシステムが発話していた文節（あるいは最後に発話した文節）をもとにして、その文節に係る伝達済の文節を係り受け構造の葉の方向にたどる（複数の文節に係る場合には、最も近い文節をたどる）。係り受け構造の葉にあたる文節からシステムが発話していた文節までが、問い返しに対応する部分だと推定し、葉にあたる文節から逐次発話する。たとえば、図 6 の「発見しました」の文節を伝達しているときに問い返された場合には、「新しい」から「発見しました」までが応答部分であると推定し、「新しい」から逐次発話する。

#### 4.2.3 情報の詳細化

テキストから、ある事柄の詳細情報を与えている箇所を抽出することができれば、詳細情報が記述された部分は省いて伝達し、必要になったらその部分を伝達することが可能となる。この方法で伝達することができれば、テキストに記述されたすべての内容を伝達せずすみ、効率の良い伝達が期待される。

詳細情報を記述した部分の特定は、構造抽出モジュールで抽出した節構造を利用して行う。連体修飾節の部分が被修飾語の詳細情報を記述した部分であると仮定

システム：NTT では <P>  
 システム：新しい 結晶材料を 発見しました <P>  
 ユーザ： はい  
 システム：この 新材料を 用いると <P>

システム：NTT では <P>  
 システム：新しい 結晶材料を  
 ユーザ： なんです  
 システム：結晶材料というのは <P>  
 ユーザ： はい  
 システム：超高速ホログラフィ動画記録を  
 可能にするというものです <P>  
 ユーザ： はい  
 システム：この結晶材料を 発見しました <P>

図 9 情報の詳細化の例

Fig. 9 Example of elaboration on contents.

し、この部分を省略して伝達する。被修飾語は“重要語”であると見なし、重要語に対する問い返し(「なんです」など)を受けた場合には応答として、連体修飾節の部分を逐次発話する。「なんです」などの重要語に対する問い返しを受けた場合には、まず、どの部分に対する問い返しかを特定し、その部分に重要語が含まれている場合には連体修飾節の部分を逐次発話する。問い返しに対応する部分の特定には、4.2.2 項で説明したように、伝達済の文節を対象として係り受け構造を葉の方向にたどり、問い返しを文節から葉の部分までを問い返しに対応する部分だとする。

たとえば図 6 を、この方法を利用して伝達すると図 9 のような伝達になる。図 9 の 1 番目の伝達例はユーザから問い返しを受けなかった場合で、詳細情報を記述した部分を伝達しないまま対話が進行している。2 番目の伝達例では、ユーザの問い返しを受け、省略した箇所を伝達している。

## 5. 実験

音声のみを利用してテキスト内容を伝達する方法の中で、話しことば対話の特徴を利用して伝達する方法の特性を評価することを目的とした実験を行った。

### 5.1 実験方法

本実験では、3 章で説明した話しことば対話の特徴を利用した伝達方法と、それらの特徴を持たない伝達方法とを比較することで、話しことば対話の特徴を利用した伝達の特性を調べた。比較実験のために、話しことば対話の特徴を利用した伝達を行う話しことば対話システムを 4 章で説明した方法で作成し、話しこと

節の終わりが「できる、である、ような、用いた、かけた、きた、可能にする、可能とする」のいずれかであるという条件も加えてある。

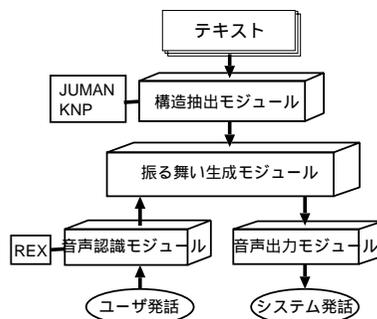


図 10 構成図

Fig. 10 System architecture.

ば対話の特徴を持たない伝達を行うシステムとして、テキストを文単位で伝達する読み上げシステムを作成した。それぞれのシステムについて説明する。

#### 5.1.1 話しことば対話システム

3 章で説明した 3 種類の話しことば対話の特徴を利用してテキストの内容を伝達する仕組みを、4 章で説明した方法で実装したシステムを話しことば対話システムと呼ぶ。

話しことば対話システムの構成は図 10 のようになっている。構造抽出モジュールは、テキストを入力として、4.1 節で説明した構造化テキストを出力する。テキストの解析には JUMAN と KNP を利用した。振り舞い生成モジュールは、構造化テキストとユーザの発話意図を入力として、4.2 節で説明した方法で次のシステムの振り舞いを決定する。システムの振り舞いには、1) テキストの一部を発話する、2) ポーズを置く、3) 発話を中断するという 3 種類がある。音声認識モジュールでは、音声認識システムと音声区間を検出する装置を利用している。音声認識システムとしては、NTT ヒューマンインターフェイス研究所で開発された REX<sup>18)</sup> を使った。REX には単語認識用の文法を与えてあり、ユーザが単語を発話した後に認識結果が出力される。音声認識モジュールは、ユーザの発話と発話意図を対応させた変換テーブルを利用して、REX が出力したユーザの発話からユーザの発話意図を推定して、振り舞い生成モジュールに送る。変換テーブルを表 1 に示す。また、音声区間を検出する装置を利用して、ユーザの発話の開始を検知した時点で振り舞い生成モジュールに信号を送る。音声出力モジュールは、システムの発話の実行と中断を行う。システムの発話は、聞き取りやすさの観点から、規則合成を利用した音声合成システムを利用せず、人間が読み上げた音声をハードディスク上に記録しておいた音声データを再生することで行った。この音声データは文節ごとに記録しており、振り舞

表 1 ユーザの発話意図と発話  
Table 1 User's intentions and utterances.

発話意図	発話
あいづち	はい, どうぞ
繰返し要求	もう一度, 文の初めから, 前の文 一番最初から
詳細要求	なんですか (システムの発話した語)
繰返し位置調整	戻って, もっと前
一時中断	待って

い生成モジュールが指定した文節に対応する音声データを逐次再生する。また、音声データの再生中に中断命令が入力された場合は、再生の中断も行う。

システムは、4.2.1 項の方法で求めた発話単位ごとにポーズを置きながらテキストに記述されたとおりに発話する。ただし、4.2.3 項で説明したように、連体修飾節の部分は省略する。あいづちを認識した場合には、それ以上ポーズを置かずに次の発話を出力する。「もう一度」という繰返し要求を認識したら 4.2.2 項で説明した方法で応答を発話する。表 1 に示した、繰返し位置調節を認識した場合には、4.2.2 項と同じ方法を用いて、繰返し要求で応答したよりも前の文節から応答部分を計算して応答を生成する。詳細要求を認識した場合には 4.2.3 項で説明したように、連体修飾節の部分を伝達する。システムの発話した語としては、テキスト中に現れる名詞を認識語彙として登録した。重要語 (被連体修飾語) でない部分に対して詳細要求を認識した場合には、繰返し要求と同じ応答をする。一時中断を認識した場合には、その時点で発話を止め、次のユーザ発話があるまで伝達を停止する。「文の初めから」「前の文」「一番最初から」は、読み上げシステムが認識するユーザ発話と同じであり、これらに対するシステムの応答は 5.1.2 項で説明する。

### 5.1.2 読み上げシステム

話しことば対話システムとの比較のため、読み上げシステムは、3 章で説明した話しことば対話の特徴を利用しないでテキストの内容を伝達するシステムとして以下のように設定した。1) システムの発話単位は話しことばの特徴である短い単位ではなく、書きことばで明示されている文を単位とする。2) システムが伝達している最中のユーザからの問い返しには応じない。3) ユーザの応答に応じた情報の詳細化は行わず、テキストの内容をそのまま伝達する。

読み上げシステムは、話しことば対話システムと同様の 4 つのモジュールから構成される。構造抽出モジュールは句点を利用して文ごとに分割することのみを行い、振舞い生成モジュールは次に発話する文を決定する。音声認識モジュールは REX を利用してユー

システム: NTT では <P>  
システム: 新しい 結晶材料が <P>  
システム: 発見しました  
ユーザ: なんですか  
システム: 結晶材料というのは <P>  
システム: 超高速ホログラフィ動画記録を可能にするというものです <P>  
ユーザ: なんですか  
システム: 超高速ホログラフィ動画記録を可能にするというものです <P>  
ユーザ: はい  
システム: この 新材料を用いると <P>  
システム: 極低温において <P>  
システム: 周波数多重記録が <P>  
システム: 可能となり <P>  
システム: 静止画像なら  
システム: 約 1000 万枚が 記録できます <P>  
ユーザ: 戻って  
システム: 約 1000 万枚が  
ユーザ: もっと前  
システム: 静止画像なら  
ユーザ: もっと前  
システム: 周波数多重記録が 可能となり <P>  
ユーザ: はい  
システム: 静止画像なら <P>  
システム: 約 1000 万枚が 記録できます <P>  
ユーザ: はい  
システム: さらに

図 11 話しことば対話システムの実行情例

Fig. 11 Dialogue between the Spoken Dialogue System and the user.

ザ発話を認識するが、音声区間の検出は行わない。音声生成モジュールは、文ごとに読み上げた音声データを持ち、指定された文の音声データを再生する。

システムは 1 文ずつ発話し、1 文の発話が終了したらユーザが発話するまで待つ。ユーザが発話できる単語は「はい」「どうぞ」「文の初めから」「前の文」「一番最初から」の 5 種類である。「はい」「どうぞ」を認識したら、次の文を発話する。「文の初めから」を認識したら、最後に発話した文を文頭から繰り返す。「前の文」を認識した場合には 1 つ前の文を文頭から発話する。「一番最初から」を認識した場合には、テキストの 1 文目を文頭から発話する。システムが文を発話している最中にユーザの発話を認識しても、システムの発話を途中で止めることはしない。その文を文末まで発話した後、ユーザの発話をもとに次に発話する文を決定して発話する。

### 5.1.3 実行例

話しことば対話システムの実行情例として、図 1 のテキストを伝達した対話の一部を図 11 に示す。1 文目の伝達では、“超高速ホログラフィ動画記録を可能にする”という部分が“結晶材料”に係る連体修飾節であるため、省略して伝達している。また、“NTT では”が提

システム: NTT では, 超高速ホログラフィ動画記録を可能にする新しい結晶材料を発見しました。  
 ユーザ: 文の始めから  
 システム: NTT では, 超高速ホログラフィ動画記録を可能にする新しい結晶材料を発見しました。  
 ユーザ: はい  
 システム: この新材料を用いると, 極低温において, 書き込みレーザー光の周波数を変えて記録媒体に情報を重ね書きする「周波数多重記録」が可能となり, 静止画像なら約 1000 万枚(テレビ映像換算 100 時間分)が記録できます。

図 12 読み上げシステムの実行例

Fig. 12 Dialogue between the Read Aloud System and the user.

題の八格であり, 読点で区切られているので, “NTT では” を発話した後にポーズを置いている。“新しい” と “結晶材料を” の間はポーズを置かずに連続して発話され, “結晶材料” が重要語であるため, “結晶材料を” の後でポーズを置いている。ユーザによる 1 つ目の「なんですか」という詳細情報の要求に対しては, 問い返しに対応する部分として “新しい 結晶材料を発見しました” が求まり, この中に重要語 “結晶材料” が含まれているので, 連体修飾節の部分を詳細情報として逐次発話している。また, 2 つ目の「なんですか」に対しては, 問い返しに対応する部分 “超高速ホログラフィ動画記録を可能にする” の部分に重要語がないので, この部分を繰り返し発話している。また, 2 文目の伝達では, 繰り返し位置調整が行われている。“静止画像なら” と “約 1000 万枚が” が係り受け構造の葉の文節であるため, これらの文節から応答が始まっている。

読み上げシステムの実行例を図 12 に示す。読み上げシステムでは, システムは文を発話の単位として伝達する。話しことば対話システムとは異なり, 連体修飾節の部分を省略せずに, テキストに記述されているとおりに伝達している。

## 5.2 実験手続き

これらのシステムを用いて, テキストから必要な情報を獲得することを課題とした実験を行った。被験者は, 獲得すべき知識を質問形式であらかじめ与えられ, 実験者によって指定されたシステムを使って, テキストから情報を獲得することを行う。実験では被験者にヘッドホンを使用させ, システム発話はヘッドホンから出力した。このことにより, システムの発話が誤ってユーザ発話として入力されることが防がれる。

以降の説明では, テキストから必要な情報を獲得して質問に解答するまでを 1 サイクルと呼ぶ。実験では 4 つのテキストを用いた。用いたテキストは「NTT

R&D 報道発表」で公表されているもので, 専門的な内容を一般向けに記述した記事である。被験者は, 1 つのテキストに対して 1 サイクルのみを行う。したがって, 被験者 1 人あたり, 話しことば対話システムを用いた 2 サイクルと読み上げシステムを用いた 2 サイクルの合計 4 サイクルの実験を行った。

1 サイクルは以下の手順で行った。ただし, システムの動作の教示はそのシステムを初めて使うときのみ行った。

- (1) そのサイクルで使用するシステムの教示を行う。システムが受け付ける単語と, その単語に対するシステムの働きを教示し, 教示用のテキストを利用した実演を行う。また, システムが受け付ける単語とその働きを簡単にまとめた表は, 実験中に見ることができる場所においておく。
- (2) そのサイクルで伝達するテキストの内容に関する質問事項を記述した質問用紙を被験者に渡す。
- (3) 被験者に質問事項を読んでもらい, 獲得すべき情報について把握させる。
- (4) システムを使って, 必要な情報を獲得させる。システムと対話しながら解答を記述してもよいとした。
- (5) 被験者自身がシステムの利用の終了を合図することでサイクルが終了する。

## 5.3 評価基準

情報伝達システムに求められる能力として, 情報伝達の効率性と, 必要な情報をどれだけ理解させることができるのかが重要だと考える。本実験では, 伝達終了までにシステムの発話した発話モーラ数と課題達成率を 2 つのシステムの比較の基準とした。システムの発話モーラ数は, 課題終了までに被験者に伝達された情報量の近似になると仮定した。また, 課題達成率は, 質問事項への解答の正答率で測った。

本実験では, 課題終了までに要した時間を評価基準として含めなかった。システムを利用している最中に質問への解答を筆記していた時間が課題終了までに要した時間の多くを占めていたことと, 筆記の時間とそれ以外の時間を分離して計測することが困難であったため, 評価基準としては不適切であると判断した。

## 6. 実験結果と考察

実験で用いたテキストの内容に関する知識を持たない大学生 32 人を被験者として実験を行った。1 つのテキストにつき, 話しことば対話システムを使って実験を行った被験者が 16 人で読み上げシステムを使った被験者が 16 人となる。

6.1 実験結果

課題達成までにシステムが発話した発話モーラ数と課題達成率の平均と標準偏差をテキストごとに計算した結果を表 2 に示す．平均を各欄の左側に示し，標準偏差を右側に示した．各テキストの発話モーラ数の平均の列の 1 行目に示した値は，テキストをひととおり読み上げるのに要する発話モーラ数である．表中の

表 2 実験結果  
Table 2 Experimental results.

	発話モーラ数		達成率 (%)	
	平均	標準偏差	平均	標準偏差
テキスト 1	377			
話しことば	1040	377	91.5	8.5
読み上げ	1383	425	88.8	13.4
テキスト 2	513			
話しことば	2039	437	91.7	7.8
読み上げ	2550	736	92.6	8.2
テキスト 3	344			
話しことば	1370	337	95.5	5.6
読み上げ	1870	687	77.7	18.4
テキスト 4	561			
話しことば	1207	302	95.1	5.5
読み上げ	1673	561	90.6	14.5

“話しことば” は話しことば対話システムを用いた伝達の結果であり，“読み上げ” は読み上げシステムを用いた結果を示している．

各テキストごとにシステムの発話モーラ数を横軸にとり，課題達成率を縦軸にとってプロットしたグラフを図 13 に示す．話しことば対話システムで伝達した結果を ○ で示し，読み上げシステムで伝達した結果を × で示した．

6.2 考察

表 2 から，いずれのテキストでも，話しことば対話システムで伝達した方が，読み上げシステムで伝達した場合よりも課題達成に要した発話モーラ数の平均が小さく，標準偏差も小さいことが分かる．この傾向は，図 13 から，話しことば対話システムのプロットである ○ の方は分布の幅が小さく，グラフの左側にまとまりがあるのに対し，× の方はグラフ全体に広がっていることから読み取れる．

課題達成率に関しては，表 2 から，話しことば対話システムで伝達した場合でも，読み上げシステムと同じ程度かそれ以上の課題達成率が得られたことが読み取れる．また，話しことば対話システムで伝達した場

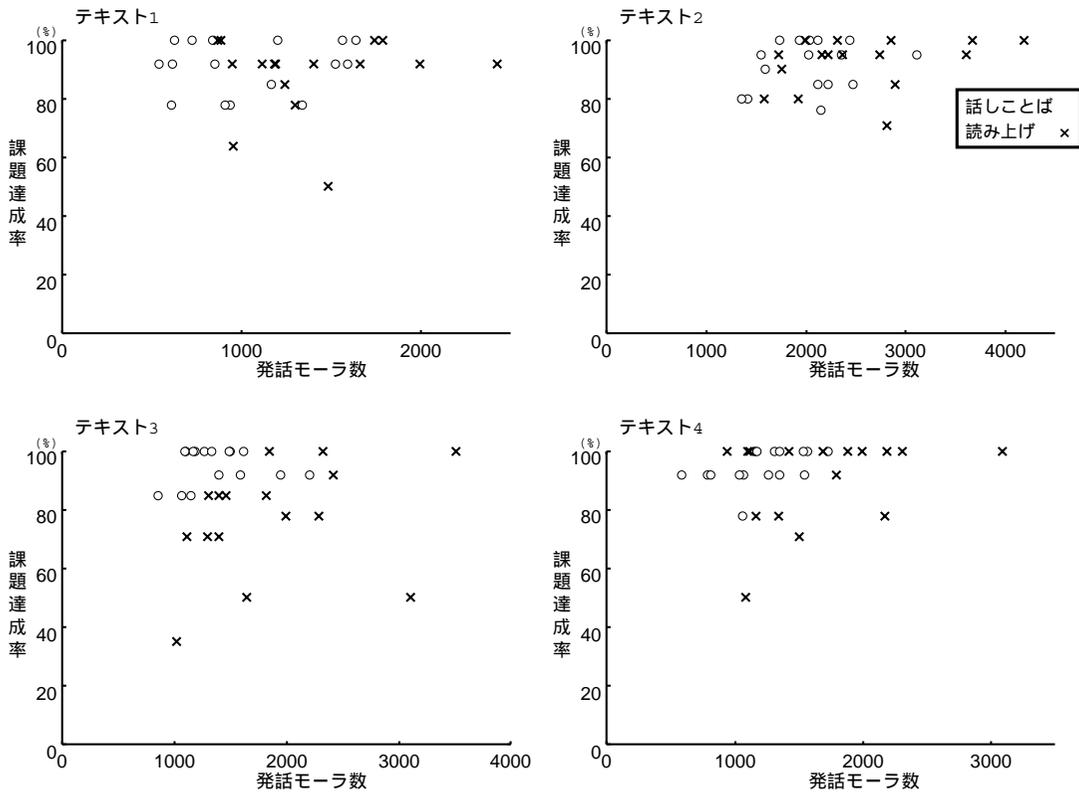


図 13 課題達成率と発話モーラ数との関係  
Fig. 13 Distribution of the number of morae and task achievement for each subject.

合の方が標準偏差も小さいことが分かる。このことは、図 13 では、印が×印に比べてグラフの上の方にまとまっていることでも示されている。

以上の分析から、テキストに記述された内容を音声のみを使って伝達する方法として話しことば対話の特徴を利用した方法は、文を単位に読み上げる方法と比べて以下の特徴があることが分かった。

- 課題達成に要する発話モーラ数のばらつきが少なく、要する発話モーラ数も少ない。
- 課題達成率のばらつきが少ない。

このことから、テキストに記述された内容を音声のみを使って伝達する方法として話しことば対話の特徴を利用した方法は、文を単位に読み上げる方法と比べて、使う人によるばらつきが少なく、効率的に必要な情報を伝達することができることが示された。

音声認識を利用したシステムでは、音声認識システムの誤認識に起因する問題を考慮する必要がある。今回の実験では、システム発話をヘッドホンから出力することでシステム発話と被験者の発話を明確に分けたこと、実験を騒音の少ない部屋で行ったこと、システムの認識する語彙が限られており単語認識であることなどの理由から、誤認識の数は少なかった。音声認識に関して最も多かった問題は、被験者の「はい」という発話が認識されないことで、多くの被験者で発生した。この問題は、教示の際にあらかじめ指摘しておいたため、被験者は特に混乱することなく「はい」あるいは「どうぞ」を再発音することができていた。また、話しことば対話システムの方が読み上げシステムよりも若干誤認識の数は多かった。認識する語彙の数が話しことば対話システムの方が多いことが一因と思われる。読み上げシステムと比較して誤認識の数が多くても、対話全体で見た場合の課題達成率やシステムの発話モーラ数の成績では、話しことば対話システムの方が良かったという結果は、本方法の優位性を示すものであるとも考えられる。

## 7. おわりに

本論文では、テキストの内容を音声のみを用いて伝達する方法の1つとして、話しことば対話の特徴を利用して伝達する方法を提案し、文単位で伝達する方法との比較実験を行うことで本手法の有効性を示した。

本研究では、計算機が情報を与えることのできる知識源として、大量に存在する計算機可読なテキストに注目した。また、情報を伝達する媒体としては音声に注目し、音声を用いてテキストの内容を効率的に伝達する方法を構築することを目指している。音声を用い

た情報伝達として、我々が日常使っている話しことば対話に注目し、話しことば対話の特徴を利用することが音声による伝達に有効であると考えた。本論文では、話しことば対話の3つの特徴を利用してテキストの内容を音声で伝達する方法を提案した。

本手法の有効性を確認するために、文単位で伝達するシステムを比較の対象にして、課題達成率と情報の伝達効率を比較する実験を行った。実験の結果、話しことば対話の特徴を利用して伝達する方法は、文単位で伝達する方法と比べて、課題達成に要する発話モーラ数の平均が少なく、被験者によるばらつきも小さいという結果が得られた。また、課題達成率に関しても、話しことば対話の特徴を利用して伝達する方法は、文単位で伝達する方法と同程度の達成率が得られ、文単位で伝達する方法よりも人によるばらつきが小さいという結果が得られた。

今後の課題として、本論文で検討しなかった話しことば対話の特徴の分析と、話しことば対話の特徴を利用する方法の再検討を考えている。特徴を利用する方法に関しては、本論文では主に文節の係り受け関係を利用して発話単位を決める方法を提案したが、発話単位に含まれる情報の量など、他の要因も考慮にいった方法が必要だと考えている。

謝辞 日頃ご指導いただく NTT コミュニケーション科学基礎研究所、東倉洋一所長、石井健一郎部長、萩田紀博部長、相川清明グループリーダーおよび、討論していただいたマルチモーダル対話研究グループの皆様、北陸先端科学技術大学院大学島津明教授に感謝します。

## 参考文献

- 1) Allen, J.F., Miller, B.W., Ringger, E.K. and Sikorski, T.: A Robust System for Natural Spoken Dialogue, *Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pp.62-70 (1996).
- 2) Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goodine, D., Goddeau, D. and Glass, J.: PEGASUS: A Spoken Dialogue Interface for On-Line Air Travel Planning, *Speech Communication*, Vol.15, pp.331-340 (1994).
- 3) 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, pp.1417-1428 (1994).
- 4) Nakagawa, H., Mori, T., Omori, N. and Okamura, J.: Hypertext Authoring for Linking Relevant Segments of Related Instruction

- Manuals, *Proc. 17th COLING and 36th ACL*, pp.929–933 (1998).
- 5) Kurohashi, S. and Nagao, M.: Automatic Detection of Discourse Structure by Checking Surface Information in Sentences, *Proc. 15th International Conference on Computational Linguistics*, pp.1123–1127 (1994).
  - 6) Fukumoto, J. and Tsujii, J.: Breaking Down Rhetorical Relations for the purpose of Analysing Discourse Structures, *Proc. 15th International Conference on Computational Linguistics*, pp.1177–1183 (1994).
  - 7) Ono, K., Sumita, K. and Miike, S.: Abstract Generation Based on Rhetorical Structure Extraction, *Proc. 15th International Conference on Computational Linguistics*, pp.344–348 (1994).
  - 8) Marcu, D.: From Discourse Structures to Text Summaries, *Proc. 35th ACL and 8th EACL Workshop on Intelligent Scalable Text Summarization*, pp.82–88 (1997).
  - 9) 亀田雅之: 疑似キーワード相関法による重要キーワードと重要文の抽出, 言語処理学会第2回年次大会発表論文集, pp.97–100 (1996).
  - 10) Dutoit, T.: *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers (1997).
  - 11) Clark, H.H. and Schaefer, E.F.: Contributing to Discourse, *Cognitive Science*, Vol.13, pp.259–294 (1989).
  - 12) 島津 明, 川森雅仁, 小暮 潔: 対話の分析—問投詞的応答に注目して, 情報処理学会自然言語処理研究会, Vol.95, No.9, pp.65–72 (1993).
  - 13) Ward, N.: Using Prosodic Clues to Decide when to Produce Back-channel Utterances, *Proc. 4th International Conference on Spoken Language Processing*, pp.1728–1731 (1996).
  - 14) 堂坂浩二, 島津 明: タスク指向型対話における漸次的発話生成モデル, 情報処理, Vol.37, No.12, pp.2190–2200 (1996).
  - 15) Halliday, M.A.K.: *An Introduction to Functional Grammar*, Edward Arnold (1994).
  - 16) Kurohashi, S. and Nagao, M.: KN Parser: Japanese Dependency/Case Structure Analyzer, *Proc. International Workshop on Sharable Natural Language Resources (SNLR 1994)*, pp.48–55 (1994).
  - 17) 松本裕治, 黒橋禎夫, 山地 治, 妙木 裕, 長尾 真: 日本語形態素解析システム JUMAN 使用説明書, version 3.11 (1996).
  - 18) 野田喜昭, 山口義和, 山田智一, 今村明弘, 高橋 敏, 松井知子, 相川清明: 音声認識エンジン REX の開発, 1998 年電子情報通信学会総大会, 情報・システム 1, D-14-9, p.220 (1998).

(平成 10 年 9 月 28 日受付)

(平成 12 年 3 月 2 日採録)



杉山 聡 (正会員)

1971 年生. 1994 年東京工業大学工学部情報工学科卒業. 1996 年同大学院修士課程修了. 同年 NTT 入社. 現在, NTT コミュニケーション科学基礎研究所メディア情報研究部勤務. 音声対話, 自然言語処理に関する研究に従事. 言語処理学会会員.



堂坂 浩二 (正会員)

1962 年生. 1984 年大阪大学基礎工学部情報工学科卒業. 1986 年同大学院修士課程修了. 同年 NTT 入社. 現在, NTT コミュニケーション科学基礎研究所メディア情報研究部勤務. 音声対話システム, 自然言語生成, 文脈理解の研究に従事. 情報処理学会平成 9 年度論文賞受賞. 電子情報通信学会, 人工知能学会, ソフトウェア科学会, 言語処理学会, ACL 各会員.



川端 豪

1978 年東北大学工学部電子工学科卒業. 1983 年同大学院博士課程修了. 工学博士. 同年日本電信電話公社入社. 1986–1989 年, ATR 自動翻訳電話研究所に勤務. 1990 年, NTT 基礎研究所に復帰, 1999 年, NTT サイバースペース研究所に勤務, 現在に至る. 音声認識・理解・対話に関する研究に従事. 電子情報通信学会, 日本音響学会, IEEE 各会員.