うなずき・相槌による音声対話システムの理解状態開示

平沢純一　中野幹生　川端 豪

ＮＴＴ基礎研究所

1　はじめに

人間同士で対話するのと同じ感覚で快適に対話できる音声対話システムの実現を目指している。快適な対話を実現するには多くの要因があるが、共有すべき情報を伝達するのとは独立に、対話進行の途中で情報の共有状態を確認・促進する機能 (**対話の調整** [5]) を持つことが快適な対話の実現に重要である。

つまり、うなずき・相槌を始めとする対話の調整行動を通じて、システムの理解状態をユーザに開示していく機構が必要になる。さらに実時間システムでは、人間同士の対話のように必要に応じてユーザ発話の途中にもうなずき・相槌を返せる [11] 必要がある。

しかし従来の音声認識の手法は、発話の終了を待ってから認識結果を出力するため、相手発話の途中に重ねて応答するシステムのためには不十分である。そこで音声認識の中間認識結果を利用して、ユーザ発話の途中にも応答可能な実験システムを作成する。実験システムは実時間で理解状態を開示しながら対話できるが、音声認識の遅れや誤りへの対処が今後の重要な課題である。

2　問題

従来の音声認識の手法は、認識結果の精度を最重視する動機で考案されている。そのため発声区間の終了を待ってから最適解 (最終認識結果) を出力する。つまり音声認識より後続の処理モジュールは、発話 (発声区間) が完了するまで認識結果を得られない。

しかし人間同士の対話においては、話し手発話に対する聞き手の応答が、話し手発話に頻繁に重なることが観察されている [8]。音声対話システムが、人間同士の対話と同様にユーザ発話の途中にも応答するためには、発話の終了を待たずに音声認識の中間結果を得ていく必要がある [1, 2]。

3　実験システム

ユーザ発話の途中にもシステムが応答できるように、音声認識の中間結果を利用して相槌などの応答 (理解状態開示) を行う実験システムを作成した [3]。

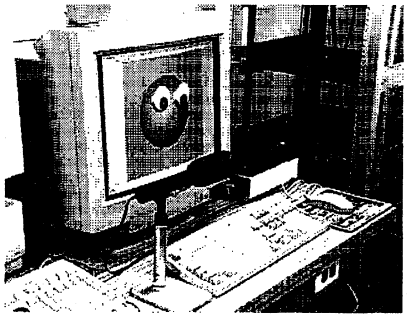

図 1: 音声対話システム「うなずき君」(外観)

3.1　タスクの選定

スロットフィリングのタスク構造を持つ「会議室の予約」をタスクとする。会議室予約は最も単純な構造のスロットフィリングタスクで、拡張により他のさまざまなタスクにも応用できる。また、タスクは単純であるが対話として十分に興味深い現象が現れる。

3.2　システムの入出力インターフェース

入力:　マイクからの音声入力に限る。

出力:　システムからの出力は「録音編集合成音声」と「CG による顔画像の顔向の動き [6]」のみである。これらを用いて、よそ見・注視・相槌 (うなずき)・聞き返し (首かしげ) ・確認発話生成などのふるまいを行い、理解状態を開示する。

3.3　システムの構成要素

実験システムは図2に示す通り、**音声認識モジュール**・返事生成を担当する**単語処理モジュール**・タスク遂行と確認発話生成を担当する**予約処理モジュール** から構成される。

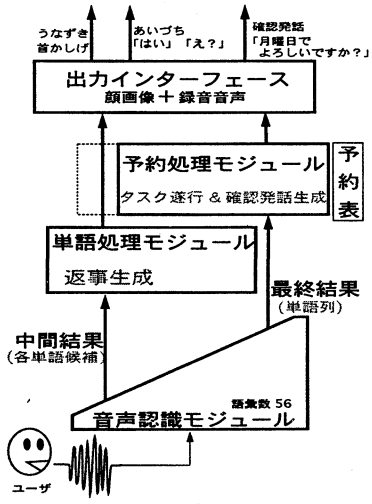

図 2: システムの構成

音声認識モジュール　音響学会の連続音声データベース [7] を用いて音素ごとに連続分布型の隠れモルコフモデルを学習する。文法は機能語・接辞・間投詞を含めてノード数 57 の単語ネットワークで記述する。認識結果は、単語候補が生成された時点で当該単語 (中間結果) が単語処理モジュールに出力される (図 3)。またユーザの発声が完了すると認識の最終結果 (単語列) が予約処理モジュールに出力される。

単語処理モジュール (返事生成)　音声認識の中間結果で得られる単語候補を入力とする。その単語が曜日・時間・部屋名に関する単語で、スコアが一定以上なら「ハイ」とうなずき、スコアが不十分だと「エッ？」と首をかしげて、システムの理解 (認識) 状態を開示する。

予約処理モジュール (タスク遂行と確認発話生成)　ユーザ発話の完了後に音声認識の最終認識結果 (単語列) を受け取る。その単語列中に予約に関する単語 (曜日・時間・部屋名) が含まれていれば予約表を埋める。3 つのスロットがすべて埋まった時点で、予約内容に関して確認発話を生成する。

さらに、システムの確認内容が間違っていると、真っ当なユーザなら情報を訂正して言い直す。そのユーザ発話の認識結果 (単語列) 中に予約に関する単語が含まれていれば、システムは予約表を修正する。

4　考察

実験システムの到達点　実験システムでは音声認識の中間結果を用いて、ユーザの発話途中の応答 (相槌 / 問い返し) を可能にした。その結果、システムの理解状態を実時間で開示できるようになり、人間と対話しているような感覚の対話 (図 4) を行うことができる。

これは伝達すべき用件だけをやりとりする従来の「質問＝応答型 (一問一答型) システム」とは異なり[1]、システムの理解状態開示によってユーザが情報の共有状態を把握できるようになったためである。しかし現状で開示している理解状態は単語認識のタイミングと尤度だけであり、実際は相槌のハイに限ってもさまざまな機能がある [9]。

システムの用途　第一に音声対話を通じてスロットフィリングタスクを遂行するシステムとしての用途が考えられる。あるいは、人間同士の対話データから得られる対話モデルをシステムに実装することで、モデル検証のプラットフォームとしての利用も可能である。但し、使用に際しては以下に述べる問題点を考慮する必要がある。

5　実験システムにより明らかになった課題

ここでは実験システムによって明らかになったさらなる問題点と、解決への指針について述べる。

5.1　音声認識に起因する問題点

システムの遅れ　近年の計算機性能の向上と発話終了を待たずに認識結果を得る手法 [2] により、実時間に近い感覚で対話できるようになった。それでも処理の進行次第で認識が遅れる可能性は必ずある。

人間同士の対話では「聞き取り」の遅延はありえない[2]。そのためシステムの反応の遅延に対して人間がいかにふるまうかは明らかでない。システムの遅延 (沈黙) に対する人間の解釈の仕方は (a)「遅延 (沈黙)」を「遅延」として解釈する (b) 人間同士では「沈黙」を「了解」と解釈する [4] ため「遅延による沈黙」が「了解」と解釈されてしまう (ユーザとシステムの誤解の溝が深まる) (c) タイミング次第では別の対象に対する応答と

[1] 一問一答型システムと対話の調整を行うシステムの違いは [10] に詳しい。従来型システムの対話がボール 1 つの「キャッチボール」や「ピンポン」だとすれば、本システムは複数のボールをお互いにぶつけあう「ノック合戦」に近い。

[2] 聞き取れない時は「聞き取れない」ということがその場 (実時間) で判明するのであって、遅延するのではない。

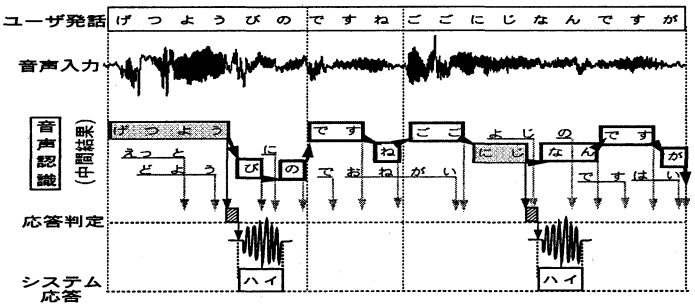

図 3: 音声認識の中間結果の利用

誤解されてしまう (図5) などの可能性が考えられ、人間の受け止め方は明らかでない。

音声認識の誤り　いかに音声認識の精度が上がっても認識結果が誤りである可能性は最後まで残る。そのためシステムの認識結果をユーザに確認する必要がある。しかし一字一句を復唱して確認するのには無理がある。またシステムの確認発話に対するユーザの返答発話を、さらに誤認識する「対話のパラドックス [4]」も指摘されている。

開始 - 終了	発話	
S 0.00 - 0.40	beep	(対話開始)
U 1.64 - 2.36	あのー	[予約フェーズ]
U 3.28 - 5.26	会議室の [1] 予約をしたいんですけど	↓
S 3.95 - 4.19	[1] はい	↓
S 5.52 - 5.72	[2] はい (○)	↓
U 6.71 - 7.65	金曜日の [3]	
S 7.65 - 7.81	[3] はい	
U 8.40 - 9.45	午後四時か [4] ら	
S 9.25 - 9.47	[4] はい (○)	
U 9.94 - 11.27	第二会議室 [5] を [6]	
S 10.71 - 10.93	[5] はい	
S 11.27 - 11.51	[6] はい (○)	
U 11.66 - 12.48	お願 [7] いします	
S 11.70 - 12.03	[7] え？	
S 12.80 - 13.46	[8] はーい (○)	
S 13.48 - 15.18	わかりました えーとー	(スロット埋まる)
S 15.20 - 18.23	水曜日午後四時から第二会議室	(最初の確認発話)
S 18.25 - 19.73	ということでよろしいでしょうか	[確認フェーズ]
U 20.37 - 21.40	金曜日の	↓
U 21.62 - 22.69	午後四時か [9] ら [10]	↓
S 22.16 - 22.37	[9] はい	↓
S 22.47 - 22.66	[10] はい (○)	
U 23.01 - 24.16	第二かい [11] ぎしつ	
S 23.48 - 23.69	[11] はい	
U 24.22 - 25.38	を [12] お願いします [13]	
S 24.24 - 24.43	[12] はい (○)	
S 25.06 - 25.24	[13] はい	
S 25.66 - 26.34	[14] はーい (○)	
S 26.39 - 30.08	金曜日午後四時から 第二会議室ですね	(確認発話 2回目)
U 30.08 - 30.38	はい	
U 30.42 - 31.03	そうです	
S 31.16 - 31.38	[15] はい	
U 31.81 - 32.96	ありがとうございました	(終了表現)
S 33.32 - 33.44	[16] は？	
S 33.46 - 34.39	どういたしまして	(対話終了)

図 4: **ユーザ (U) とシステム (S) の対話例.** 時間の単位は sec. (○) は正認識による応答. [n] は相槌のタイミングを示すタグ.

5.2　システム実装からの解決法

遅延に対して　音声対話システムの実装を目指す立場からは、システムの応答に対して時間的な統制を行うことが必要である。システムの応答に遅延が生じた時には、

1. そのまま応答する
2. 遅延を反映させて応答の形態を修正する
3. 時間切れなので応答を断念する

のいずれかの対処を行うべきである。システムによるあらゆる応答 (ふるまい) に対して、人間は必ず解釈を加えてしまう。遅延による沈黙にも意味が生じてしまうため、ユーザの解釈まで考慮に入れてシステムを設計しなければならない。

誤認識に対して　システムは音声認識結果の誤りの可能性を前提として設計することが必要である[3]。ユーザへの確認を通じて誤りを訂正するのであれば、

- 「確認の必要な対象」と「確認しなくてもよい (多少の誤認識が致命的でない) 対象」を判別する
- 適切な確認の仕方を選ぶ
  (相槌・復唱・確認質問など)
- 対話のパラドックス [4] を防止する

などの問題を解決する必要がある。

[3] 音響処理だけで入力音声を文字列に変換する音声タイプライターの実現は難しい。なぜなら人間も言語知識・領域知識がなければ聞き取れない。音声タイプライター的な認識結果 (誤認識を含まない文字列) を入力に仮定するシステム研究は、音声対話システムの実現にあまり貢献しない。

5.3 人間同士の対話研究からの解決法

遅延に対して 通常の人間同士の対話では反応の遅延が生じにくいので、遅延に対する人間の対処が含まれている対話を調べる必要がある。遅延に対する人間の対処について何らかの知見を得られれば、ユーザが自然に受け入れられる「遅延の開示方法」をシステム設計に活かすことができる。例えば、処理の遅延を正直に示すのがよいのか、遅延を取り繕って示す方が対話を円滑にできるのか、実験的に検討できる。

誤認識に対して システムほどではないにせよ人間にも聞き誤りは存在するので、誤りの検出・復旧についてより詳しく調べる必要がある。人間同士の対話において人間が (a) どの時点で自らが「誤っている」ことに気付くか (b) 誤りをいかに修正するのか (開示するのか) (c) 相手に確認することで修正する時はどのように確認するのか に関しての知見があれば、システムの誤認識への対処法の参考にすることができる。さらにシステムの誤認識の傾向も考え合わせることで、適切に確認する対話システムを実現できる。

6 まとめ

人間同士の対話と同じ感覚で対話できるシステムを実現するため、音声認識の中間結果を用いてユーザ発話の途中にも応答できる実験システムを作成した。この実験システムは、従来の一問一答型システムとは異なり、システムの理解状態開示を実時間で行う。しかし音声認識の「遅れ (システムの反応の遅延)」と「誤り」という問題が明らかになり、対話システム実装の面からと人間の対話研究の面からの解決の指針を示した。

遅延と誤認識の問題に共通しているのは、理解主体 (人間 or システム) が (a)**for what:** どんな内部状態に関して開示 (応答) するのか (b)**when:** どの時点で開示するのか (どの時点を過ぎたら異なる対応が必要か) (c)**how:** どのような形態で開示するのか を明らかにする必要がある、ということである。

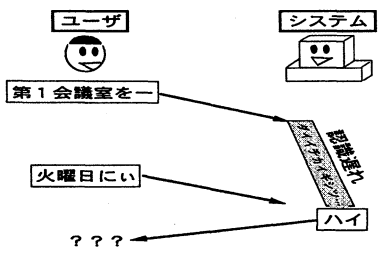

図 5: 反応の遅れが応答の意味を誤解させる

謝辞 日頃よりご指導いただく NTT 基礎研究所情報科学研究部 石井健一郎部長、有益な示唆をいただく対話理解研究グループの諸氏、システムの実装にご協力いただく NTT-AT 社の木間良子さん、久保田哲也さんに感謝いたします。

参考文献

[1] Görz,G., Kesseler,M., Spilker,J. and Weber,H. :Research on Architectures for Integrated Speech/Language Systems in Verbmobil. *COLING96*, pp.484-489 (1996)

[2] 平沢純一, 川端豪: わかってうなずくコンピュータの試作. 信学技報 NLC97-54, SP97-87, 情処研報 SLP97-19 (1997)

[3] 平沢純一, 川端豪: 音声対話システム Noddy– ユーザ発話途中でのうなずき・相槌生成 —. 情処研報 SLP98-20 (1998)

[4] 片桐恭弘: 対話調整の分散処理モデル. 情処研報 SLP 94-2 (1994)

[5] 片桐恭弘, 川森雅仁, 島津明: あいづちの分散システムモデル. 言語処理学会第 1 回大会, pp.33-36 (1995)

[6] 川端豪: 音声理解システム JUNO における対話マスコット. 平成 9 年春季 音響学会講演論文集 2-Q-2, pp.143-144 (1997)

[7] 小林哲則, 板橋秀一, 速水悟, 竹沢寿幸: 日本音響学会研究用連続音声データベース. 日本音響学会誌 48 巻 12 号, pp.888-893 (1992)

[8] 小坂直敏: あいづちを中心とした会話音声の呼応関係の分析. 信学技報 SP87-107 (1987)

[9] 島津明, 川森雅仁, 小暮潔: 対話の分析 – 間投詞的応答に着目して –. 情処研報 NL93-95 (1993)

[10] 島津明, 小暮潔, 川森雅仁, 堂坂浩二, 中野幹生: 対話処理システムにおける内的コミュニケーション. 言語処理学会第 2 回大会, pp.26-27 (1996)

[11] Ward,Nigel. :Using Prosodic Clues to Decide When to Produce Back-channel Utterances. *ICSLP96*, pp.1728-1731 (1996)