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 ―

平沢 純一　川端 豪
{jun,kaw}@idea.brl.ntt.co.jp
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。ユーザが自由なタイミングで快適に話しかけることのできるシステムであるには、システムは自身の理解状況を開示するため、適切なタイミングで応答（相槌やうなずきを含む）することが必要である。システムは認識・理解が十分でなくとも、対話の進行に応じて実時間で応答しなければならない。本発表では、ユーザの発話途中でも応答可能な音声対話システムについて、ビデオプレゼンテーションを中心に紹介する。

2　システムの概要

従来の音声対話システムでは、ユーザの発話が完了してから音声認識の結果を得て、応答生成などの後続の処理を行っていた。将来、驚異的に認識処理の速度が上がったとしても、この枠組みを採用する限り、ユーザの発話中にシステムが応答することは不可能である。そこで、我々は連続音声認識アルゴリズムの中間結果を用いる[1]ことにより、ユーザの発話中にも応答を返せるシステムを試作した[1]。以下でこのシステムの概要を紹介する。

システム構成：　音声対話システム Noddy(図 1) は 3 つのモジュール (音声認識モジュール・単語処理モジュール・予約処理モジュール) と出力インタフェース (録音音声・顔画像) から構成される (図 2)。

[1]中間結果をどのタイミングでどのように用いるかについては、まだ検討するべき事項が山積みである。

図 1: 音声対話システム Noddy(外観)

タスク：　システムは会議室の予約タスクを行う。システムは、曜日・時間・部屋名の 3 つのスロットを埋める。

入力：　システムへの入力は音声入力のみである。

出力：　システムからの出力は「録音再生音声」と「顔画像マスコットの視線 (顔向) の動き [2]」のみである。このふたつのモーダルの組み合わせにより、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、システムの理解状態を開示する。

音声認識モジュール：　音素ごとに連続分布型の隠れマルコフモデルを用いた。音素モデルの学習には音響学会の連続音声データベース [3] を用いた。文法は機能語・接辞・間投詞を含めて、ノード数 57 のネットワーク文法を記述した。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送られ (図 3)、ユーザ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送られる。

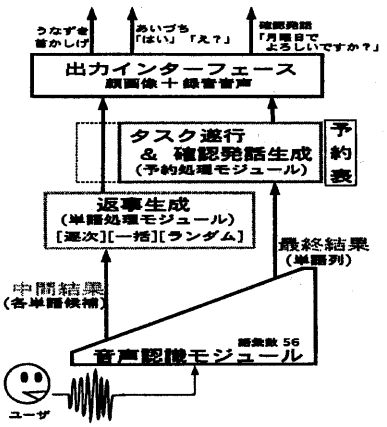

図 2: システムのモジュール構成

単語処理モジュール： 音声認識の中間結果で得られる単語候補を受け取る。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、スコアが不十分だと「エッ？」とふるまい、システムの理解状況を開示する。

予約処理モジュール： ユーザの発話完了後に音声認識の最終認識結果 (単語列) を受け取る。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ると、予約内容に関して確認発話を生成する。

3 問題点

現時点でのシステムの問題点 (今後の課題) は主にふたつ考えられる。ひとつは、まだ有効に利用できていない対話履歴情報などを用いて、対話管理の機能を向上させる点。ふたつめは、システムの出力 (タイミング) を

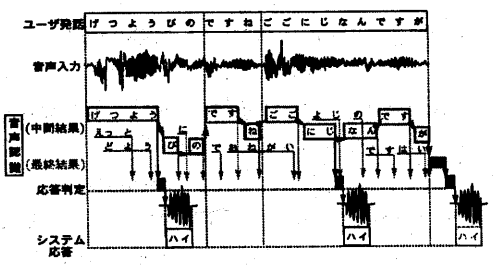

図 3: 音声認識の中間結果の利用

	開始 - 終了	発話	
S	0.00 - 0.40	beep	(対話開始)
U	1.64 - 3.36	あのー	[予約フェーズ]
U	3.28 - 5.28	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	[6] はーい (○)	
S	13.48 - 15.16	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 [13] お願いします [13]	
S	24.24 - 24.43	[13] 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	30.05 - 30.36	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: 対話例 (○は適切な「はい」)

もっと厳密に統制するべきである。現在は、ユーザ発話の途中での応答が可能になったとは言え、あくまでもシステム側の都合でふるまいの生成タイミングが決まるが、ふるまいのタイミングは対話の状況に応じて定めなければならない。今後これらの点を改良していく予定である。

謝辞 日頃よりご指導いただく NTT 基礎研究所情報科学部 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、システム実装にご協力いただいた NTT-AT 社の久保田哲也氏に感謝いたします。

参考文献

[1] 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)

[2] 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 音講論 2-Q-2, pp.143-144 (1997)

[3] 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)