柔軟な話者交代を行う音声対話システム DUG-1

中野 幹生* 堂坂 浩二* 宮崎 昇* 平沢 純一* 田本真詞*
川森 雅仁* 杉山 聡* 川端 豪**
*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
**NTT サイバースペース研究所 メディア処理プロジェクト

1 はじめに

近年の音声認識技術の向上により, 不特定のユーザが使える音声対話システムの実現が可能になってきている. しかし, 音声対話システムが実際にインターフェースとして普及するためには, 単にタスクが達成できるだけではなく, システムがユーザにとって使いやすいものでなくてはならない. そのためには, 後で詳述するように, ユーザが話している間にもあいづちや復唱によってシステムの理解の状態をユーザに示すことや, システムが話している間にもユーザの割り込みを受け付けて生成内容を変えること, すなわち柔軟な話者交代を行うことが重要である.

従来の音声対話システムでは, 長いポーズ, 特殊なキーワード, マウスボタンなどを用いて, ユーザが自分の発話の終了を明示的にシステムに伝えなければ, システムは話し始めることができなかった. また, システムの発話が終了しなければ, ユーザの発話を認識, 理解することができなかった. このため, 柔軟な話者交代を行うことができず, ユーザにとってシステムは使いやすいものではなかった.

本稿では, 我々が構築した, 柔軟な話者交代を行う音声対話システム DUG-1 について報告する. DUG-1 では, ユーザ発話の逐次理解法とシステム発話の逐次生成法を用いること, および, 言語理解部と言語生成部を並行に動作させることにより, 柔軟な話者交代を実現している.

2 柔軟な話者交代の必要性

従来の音声対話システムの多くは, ユーザの発話とシステムの発話が交互に切り替わる, いわばトランシーバ型のシステムであった[2), 3), 16)]. そのため, ユーザが話している途中にはシステムが応答したり割り込んだりすることができなかった. また, システムが発話している途中にユーザが割り込んでもシステムは反応することができなかった. これらのシステムでは, 話者交代のタイミングは, ある程度の長さのポーズを検出するか, または, 発話の終了を示す特別な語をユーザに話させる[19)] ことによって行っていた.

人間同士の対話では, 相手の発話に重なる発話を行うことが頻繁に起こることが知られている[15)]. このような話者交代の仕方を, 本稿では 柔軟な話者交代と呼ぶことにする. 言うまでもなく, 音声対話システムをユーザフレンドリなインタフェースの一形態として実現を試みる立場では, システムとユーザの対話が人間同士の対話と同じである必要はない. 人間同士の時よりもユーザにとって快適な会話が行えることを目標にしなくてはならない. したがって, 人間同士の会話で発話が重なるからといって, それと同じことをシステムに求める必要はない.

それでも, 柔軟な話者交代が行えることはシステムの使いやすさにとって非常に重要である. それは, 次のような理由による. まず, ユーザの発話にシステムの発話が重なる場合を考える. ユーザの発話中にシステムがあいづちをうつことにより, ユーザが気軽に話すことができる. これは, システムがユーザの発話を聞いているということや, 理解しているということを, ユーザが認識できるからである[10), 17)]. 自分の発話中にシステムがあいづちを打つのを快適に感じるユーザと不快に感じるユーザがいることが実験により判明しているが[8)], 人間同士の会話では, 相手のあいづちを特に不快に感じているわけではないことから, 適切なタイミングでのあいづちは有効であると考えられる. また, ユーザの話していることをシステムが理解できないときに, ユーザの発話をインタラプトすることができる. もしインタラプトしなければ, ユーザは, システムが理解できていないのにもかかわらず無駄に話し続けてしまい, 結果とし

"DUG-1: A Spoken Dialogue System with Flexible Turn-Taking" by Mikio Nakano, Kohji Dohsaka, Noboru Miyazaki, Jun-ichi Hirasawa, Masafumi Tamoto, Masahito Kawamori, Akira Sugiyama, and Takeshi Kawabata (NTT Laboratories)

て対話が非常に非効率的になる.

次に, システムの発話中にユーザから割り込む場合を考える. システムの発話している内容がユーザにとって必要のない情報である時, システムの発話をインタラプトすることにより, ユーザは時間を節約することができる. また, システムの発話が理解出来なかったときには, ユーザはシステムにもう一度言い直すよう要求することができるので, うまく理解しあえていないのに対話が先に進んでしまい, もう一度初めからやりなおすということがなくなる. これらのことから, 柔軟な話者交代が音声対話システムの使いやすさにとって重要である.

システム発話中にユーザの割り込み発話を受け付けることができるシステム[6), 9)]や, ユーザの発話中に割り込んで応答するシステム[1), 7), 20)]が作られているものの, これらの両方を行うことができるシステムは作られていない. しかしながら, ユーザとシステムの間で主導権が交代する, いわゆる相互主導型 (mixed-initiative) 音声対話システムでは, ユーザ発話に割り込むことと, ユーザからの割り込みに対処することの両方が必要である.

3 音声対話システム DUG-1

3.1 アーキテクチャ

我々が作成したDUG-1は, ユーザ発話を逐次的に理解する発話理解部[13)]と, システム発話を逐次的に生成する発話生成部[6)]を並行に動作させる[18)]ことによって, 柔軟な話者交代を実現する音声対話システムである. DUG-1は音声認識部, 言語処理部, 音声生成部からなる. 図1にこれらの関係を示す.

音声認識部は, 連続分布型の音素HMMによる不特定話者の連続音声認識器で[14)], ISTAR (Incremental Structure Transmitter And Receiver) プロトコル[7), 12)]を用いて音声認識結果の逐次出力が行えるようにしたものを用いている. これは, 各音声フレーム毎に, 最もスコアの高い探索パスの単語仮説を出力する (単語の途中の場合でも出力する) ものである. 認識用の文法はネットワーク文法で記述している. この文法の制限はあまり強くなく, 各音声区間が, 辞書中の単語からできる任意の文節の任意個の繰り返しからなるとしている. 文法はいくつか用意し, 対話の場面に応じて切り替えることができるようにしている. これは, 対話の場面の情報を利用して, ユーザ発話の予測を行うことに相当する. 対話のどの場面にいるかの判断は, ユーザ発話理解の結果とシステム発話の内容を用いて後述の言語生成プロセスで行っている. 各文法中に含まれる単語は100語未満であり, 十分にリアルタイムで認識が可能である.

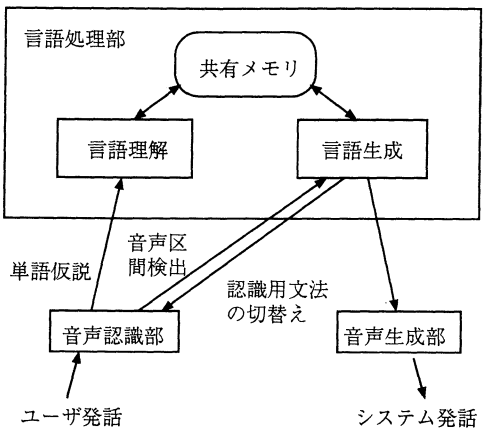

図 1: DUG-1 のアーキテクチャ

言語処理部では, 言語理解プロセスと言語生成プロセスが並行して動作する. 主導権がユーザにあるとき (正確には, 主導権がユーザにあるとシステムが考えているとき) には, ユーザ発話の逐次理解法[13)]を用いて, 単語単位で逐次的にユーザ発話を理解する. このときには構文規則を用いて理解を進める. 理解の結果はフレームで表現する[4)]. 構文解析の結果として得られる意味表現に応じてフレームを書き換える. フレームは共有メモリに書き込まれるので, 言語生成プロセスからも参照できる. 言語生成プロセスは, フレームの状態やポーズの出現に応じてあいづち, 問い返し, 確認発話などを行う. 音声認識と構文解析は, ユーザの発話と同時に進めているため, これらの応答をユーザの発話中に行うことが可能である.

主導権がシステムにあるときには, 言語生成プロセスが, 階層的プランニングを用いて, 節以下の細かいフレーズで逐次的に発話を生成する. このときには, ユーザの発話の理解は文節毎に行い, 構文解析は現在のところ用いていない. システムは, ユーザの割り込みやあいづちに応じて説明の内容を変化させる[6)]. これは, 人間同士の対話の分析から得られた協調的対話原則[5)]に従って行う.

主導権の移動は次のようにして行う. ユーザに主導

権があるとき，システムはフレームの内容を見て，ユーザが主導権を譲ろうとしているかどうかを判断し，譲ろうとしている場合にはシステムが主導権を取る．そして，フレームの内容に応じて，ユーザに伝達すべきことを生成する．ユーザの割り込みの内容から，ユーザが主導権を取ろうと判断できたとき，または，ユーザに伝達すべきことをすべて伝達し終わり，ユーザの発話を待つ状態になったとき，システムは，主導権をユーザに渡す．

音声生成部は，あらかじめ録音された音声を文節毎に再生することによって音声生成を行う．ユーザが気軽に話せるように，ディスプレイ上に擬人エージェント[11]を表示し，生成内容に合わせてうなずいたり首をかしげたりすることができる．

以上のアーキテクチャにより，DUG-1は，ユーザの発話にシステム発話を重ねることと，システム発話中のユーザ割り込みに対処することが可能であり，したがって，柔軟な話者交代を行うことができる．

本稿の主題は柔軟な話者交代の実現であるが，その他にもDUG-1には従来の音声対話システムにはない特長がある．たとえば，ユーザに主導権があるときに，ユーザ発話の逐次理解により，いつでも，ユーザ発話の理解の途中結果を保持していることから，ポーズを検出すると，その途中結果を用いて発話生成モジュールが即座に応答を生成することができる．また，逐次理解を用いていることから，ユーザが1文を一気に話さなくても少しづつ理解していくことができる[13]．

3.2 タスク

本稿では具体例として，TV番組の録画予約の受付タスクを説明する．これは，以下のような状態からユーザの予約したい番組の予約を達成するタスクである．

- ユーザは，予約したい番組のタイトルは知らないが，その番組のジャンル，出演者，時間帯などの断片情報を知っている．
- システムは完全な番組表を持っているがユーザの予約したい番組は知らない．

したがってこのタスクは，

> ユーザとシステムとに分散された情報を合わせることによりタスクを遂行する共同意思決定タスクである

といえる．

S1 はいどうぞ
U2 えーとですね
S3 はい
U4 スピードが出る朝じゃなくて夜の〈はい〉歌番組を予約したいんですけれどもー
S5 スピードが出てる夜の歌番組ですか
U6 はい
S7 えーっと ポップジャムという番組が〈はい〉あります
S8 夜の歌番組で〈ポップジャムでしたっけ〉スピードが
S9 そうです
U10 はい
S11 ポップジャムという番組があります
U12 はい
S13 スピードが出演します
U14 はい
S15 ポップジャムという番組を予約しますか
U16 はい予約します
S17 かしこまりましたー ほかにございますか
U18 えーと〈はい〉それだけでいいです
S19 ありがとうございました

図 2: 対話例

DUG-1はこのタスクを次のように遂行する．最初はユーザが主導権を持ってこれらの断片情報をシステム側に伝えると，システムが主導権を取って候補の番組のタイトルや特徴をユーザに伝える．再び主導権はユーザに移り，ユーザが候補の番組の中から予約したい番組を選ぶか，予約をとりやめるかを決定すると，次の番組を予約するフェーズに戻る．システムの設計に際しては，同じタスクを用いて人間同士の対話を収録し，その分析を行った[5]．

3.3 対話例

図2にユーザとシステムの対話例を示す．図中でUはユーザ発話，Sはシステム発話である．〈〉は聞き手の割り込み発話を示す．S7でシステムが主導権をとり，S15の発話終了後に主導権をユーザに渡している．

話者交代に注目すると，まず，U4の発話中にシステムがあいづちを打っている．これは，「夜の」というフレーズをとらえて理解した結果に基づくあいづちである．あいづちを打つタイミングは，ユーザ発話の意味内

容や, あいづちの頻度を考慮しながら決定している. また, S7 中のユーザのあいづち「はい」は, ユーザの理解を示すものとしてとらえ, システムはそのまま説明を進行させている. S8 中の, 「ポップジャムでしたっけ」というユーザの割り込みを検出すると, 「スピードが」と, 次の説明を言いかけていたのをやめ, S9 「そうです」と反応し, 説明しなおしている. この例からわかるように, DUG-1 では柔軟な話者交代が可能である.

4 おわりに

本稿では, 柔軟な話者交代を行う音声対話システムのアーキテクチャを提案し, 実験システム DUG-1 について報告した.

謝辞

日頃ご指導いただく NTT CS 基礎研 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に感謝いたします. また, 本研究に際し, NTT サイバースペース研究所メディア処理プロジェクトで開発した音声認識エンジン REX を使用しました. 関係各位に感謝致します.

参 考 文 献

1) G. Aist. Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption authors. In *ICSLP-98*, pp. 928–931, 1998.

2) J. F. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In *ACL-96*, pp. 62–70, 1996.

3) H. Aust, M. Oerder, F. Seide, and V. Steinbiss. The Philips automatic train timetable information system. *Speech Communication*, 17:249–262, 1994.

4) D. Bobrow, R. Kaplan, M. Kay, D. Norman, H. Thompson, and T. Winograd. GUS, a frame driven dialog system. *Artificial Intelligence*, 8:155–173, 1977.

5) 堂坂, 川端, 島津. 複数の対話ドメインにおける協調的対話原則の分析. 電子情報通信学会技術研究報告 NLC-97-58, pp. 25–32, 1998.

6) 堂坂, 島津. 協調的な話し言葉生成. 電子情報通信学会技術研究報告 NLC-96-32, pp. 9–16, 1996.

7) J. Hirasawa, N. Miyazaki, M. Nakano, and T. Kawabata. Implementation of coordinative nodding behavior on spoken dialogue systems. In *ICSLP-98*, pp. 2347–2350, 1998.

8) 平沢, 中野, 川端. 音声対話システムの相槌応答タイミングによるユーザの印象への効果. 言語処理学会第 5 回年次大会論文集, 1999.

9) C. Kamm, S. Narayanan, D. Dutton, and R. Ritenour. Evaluating spoken dialog systems for telecommunication services. In *Eurospeech-97*, pp. 2203–2206, 1997.

10) 片桐, 川森, 島津. あいづちの分散システムモデル. 言語処理学会第 1 回年次大会論文集, pp. 33–36, 1995.

11) 川端. 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.

12) 川端, 宮崎, 平沢. 逐次的音声認識・理解のための ISTAR アーキテクチャ. 平成 10 年秋季 音響学会講演論文集 1-1-15, pp. 29–30, 1998.

13) 中野, 宮崎, 平沢, 堂坂, 川端. 多重文脈を用いた逐次的な発話理解. 情報処理学会研究報告 SLP-22, pp. 21–26, 1998.

14) 野田, 山口, 山田, 今村, 高橋, 松井, 相川. 音声認識エンジン REX の開発. 1998 年電子情報通信学会総合大会, 情報・システム 1, D-14-9, pp. 220, 1998.

15) 小坂. あいづちを中心とした会話音声の呼応関係の分析. 電子情報通信学会技術研究報告, SP87-107, 1987.

16) J. Peckham. A new generation of spoken language systems: Results and lessons from the SUNDIAL project. In *Eurospeech-93*, pp. 33–40, 1993.

17) 島津, 川森, 小暮. 対話の分析 – 間投詞的応答に着目して –. 電子情報通信学会技術研究報告, NLC-95-9, 1993.

18) 島津, 小暮, 川森, 堂坂, 中野. 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 回年次大会発表論文集, pp. 333–336, 1996.

19) R. W. Smith and D. R. Hipp. *Spoken Natural Language Dialog Systems.* Oxford University Press, 1994.

20) N. Ward. Using prosodic clues to decide when to produce back-channel utterances. In *ICSLP-96*, pp. 1728–1731, 1996.