音声言語情報処理 7−21
（1995. 7. 22）

休止を区切りとした対話処理

伊藤克亘　秋葉友良　上條俊一†　田中和世

電子技術総合研究所，†東京工業大学

音声対話システムにおいて、利用者に発話をできるだけ制限させなくてすむ対話処理の枠組について検討をおこなった。本稿では、対話処理の枠組を、1) 認識の単位 2) 理解の単位 3) 応答のタイミングの三つの視点から検討する。その結果、音響的・言語的に意味のある区切りとなることが知られている休止を共通の単位として対話を処理する方法を提案し、今後の研究課題について考察した。

A Dialog Processing Method Based on Interpausal Utterance

ITOU Katunobu, AKIBA Tomoyosi, Kamijo Syunichi†, and TANAKA Kazuyo

Electrotechnical Laboratory, †Tokyo Inst. of Tech.

We proposed a dialog processing method for a speech dialog system, which made users restrict as little as possible. In this paper, we studied a design of a dialog processing method from three viewpoints as follows: 1) the timing for sending a speech recognition result to the natural language processing (NLP) module 2) the timing for generation of an NLP result 3) the timing for generation of a spoken response. As a result, we proposed the dialog processing method which used a pause as a period of an utterance.

1　はじめに

近年、さまざまな音声対話システムが研究されている。ここでいう音声対話システムとは、機械 (計算機) と人間が音声言語をやりとりの手段の一つとして用いるシステムであるとする。これらのシステムは、なぜ、音声言語を用いて人間とやりとりしなければならないのだろうか？

人間と機械のコミュニケーションを考えてみる。自動販売機での切符の購入やビデオの予約、自動車の運転など、これまで、人間は機械とある程度高度なコミュニケーションを言葉以外の媒体を通じておこなってきた。これらは、言葉以外の媒体を通じておこなわれたというよりは、個々の事例では、言葉を使うよりも適切な媒体であるともいえるだろう。D. Norman は、「たとえ相手が秘書や友達でも、要求やものの使い方などを言葉だけで伝えるのはむずかしい。これは、言語というものが性格さを来した伝達手段としてデザインされていないからだ。」[1]と主張しているほどである。確かに、言語にはこういった側面がある。電話や会合は時間をいたずらに浪費するものである、という主張もよく聞く。会合をせずに電子メールを使えとか、電話のかわりに FAX を使えというのは、要するに、音声言語はついつい時間を浪費してしまうものである、ということである。

このように話をすすめると、電話のように音声しか使えない状況や、手などを使っているときに同時に使えるものとして声が有力だ、という理由で、音声対話システムの存在を正当化する意見が聞こえてきそうである。確かに、そのような限定された状況では音声は有効だろう。しかし、音声が有効であることと、音声言語が

[1]MacPower 95 年 6 月号 148 ページのインタビューより。他のインタビューや著作でも同様の主張を展開しているはずである。

有効であることは別問題である。

さて、多くの人間が、音声言語の効率の悪さという問題点に気づいているにもかかわらず、多くの人間が多くの時間を費やして音声言語で対話している。人間は、情報を伝える以外の側面を、話し言葉による対話に感じている、もしくは、求めているのではないだろうか。この側面を意識的/無意識的にとらえた研究として、音声対話システムの「顔」の存在に着目した研究がいくつかあげられる [1–5]。また、対話を円滑にすすめるという点からシステム側からのフィードバックに着目した研究 [6, 7] も、同じような方向性を持った研究であると考えられる。

本稿では、このような「利用者が対話をつづける気にさせる」、もしくは、そのレベルが無理ならば、せめて「利用者が対話の途中で話すのを止めてしまいたくならない」ようなシステムのふるまいについて考える。なかでも、まず、利用者の発話を処理してシステムの発話を生成するまでの対話処理手法について検討していく。

2 話し言葉の理解

2.1 話し言葉を理解する「単位」

本稿で対話処理とよぶ処理は、普通のシステムでは、次の三つの部分に分けられる。

1. 音声認識
2. 発話理解 (構文解析, 意味処理など)
3. 発話生成

これらの要素は、従来、別々の要素技術として研究がすすめられてきた。そのため、ひとつにまとめようと思ってもなかなか難しい問題がある。ここでは、それぞれの処理がどういう単位に基づいておこなわれるかに着目して、話し言葉の理解に必要な条件を検討する。

まず、音声認識部で文法を用いるアプローチの場合、その文法がそもそも、「文」を前提とする場合が多い。この手法の場合、発話理解部には、書き言葉の自然言語で用いられるような手法を用いることが多いので、そのあとの処理は「文」を単位に進められる。

音声認識部でスポッティングの手法を用いるアプローチの場合、スポッティングの段階では「文」をあまり意識していないが、スポッティングされた複数の候補 (ラティスなど) から正解を選ぶ発話理解の段階で、明示的/非明示的に「文」を前提とするような制約を使うことが多い。

ここまで、「文」と何の定義もなく書いてきたのだが、果たして、「文」とはそれほど明確な単位なのだろうか。そもそも、書き言葉であっても、句点と読点は可換なものがあり、句点から句点までという定義すら形式上の意味しかない。話し言葉の場合、その句点すら明確でないため、文の認定自体が難しい問題であるといわれている [8]。

さらに、音声理解の多くの手法では、認識誤りを訂正するために、対話をすすめる上で必要な情報が過不足なく出現する候補に対して有利なスコアをつけることが多い [9]。つまり、「文」として、書き言葉的に整った発話を仮定しているのである。これでは、話し言葉らしい整っていない発話 (たとえば、強調するために一部の語句を繰り返すような発話) などは正しく認識されると意味処理時に逆にスコアが悪くなることもありうる。

このようなアプローチは、言語の情報を効率よく伝える側面ばかりに注目した発話理解手法であるとはいえないだろうか。そのような立場に立てば、繰り返しによる強調や補足による説明の詳細化など、話し言葉に特徴的な「わかりやすく」するための発話がやっかいな現象になってしまう。本稿では、音声認識結果の訂正能力は低いかもしれないが、話し言葉に特徴的な現象をとらえることのできる対話処理の枠組を考えることにする。そのためには、まず「文」に変わる処理の単位を考えなければならない。

2.2 話し言葉的な現象と休止

休止に着目して認識用の文法を構築する試みは、文献 [?] で述べられている。しかし、この文法は目標を翻訳電話においてあるため、機械相手の対話システムに対する前提とは異なる部分もあると考えられる。竹沢らの「部分木」という仮定を離れて、休止と言語現象との対応をまとめると、以下のようになる [8, 10]。

- いわゆる倒置, いいよどみ — いい足す前に休止が入ることが多い。
- つなぎ語 — 休止に囲まれることも多い。すくなくとも、前後どちらか片方には、休止がある。
- 語間 — 非活用語 (名詞, 副詞, 感動詞, 助詞) の後と助動詞列の後に休止が入る。だいたいが、いわゆる「文節」の区切りだが、名詞と助詞の間にも入りうる。

このように、休止で区切られた部分は、構文的にも意味的にも、それなりに意味のあるまとまりであることがわかる。

したがって、音声認識と言語処理を統合するときには、それなりによい目安になる単位であると考えられる。本来なら、音声認識時に、あらゆる言語的な知識も導入して完全に統合した形で音声理解をおこなうのが理想的である。しかし、そういった形態をとるには、たとえば、音声認識のフレームシフトである 10ms ごとに言語処理をおこなう必要がある。しかし、現実的には、処理量の面でも無理であろう。また、言語処理から音声認識にトップダウン情報として役立つ制約も一文内のレベルでは考えにくいため、休止ごとに音声認識部と言語処理部がおたがいに情報をやりとりするのが現実的な解法であると考えられる。

3 休止を単位とした対話処理

本稿で提案する処理の流れは以下の通り。

音声認識 -(認識結果)-> 構文解析
-(QLF)-> 文脈処理 -> 応答生成

簡単に、各部分での処理について述べる。

3.1 音声認識

息つぎ程度の休止 (試行実験では、250ms 程度の継続時間) で区切られた区間を認識する。区間ごとに構文解析部に上位 N 個の結果を送る。

3.2 構文解析

単語列を擬似論理式 (QLF) に翻訳する。QLF は、(<cnf>, <inf>,<props>) の三つ組であらわされる。<cnf> は、はい/ いいえ・副詞的な情報など発話に伴う付随的な情報をあらわす。<inf> は伝達内容を表す論理式で、表層発語内行為 (surface illocutionary acts) に相当するものである。<props> は <inf> 内の変数に対する制約である。たとえば、「東急ハンズの場所を教えて下さい」という発話は、以下のような QLF に翻訳される (本来は、もう少し厳密であるが、説明に不要な部分は省略している)。ここで、IP はシステム、IS は利用者をあらわす。

```
cnf   -
inf   want(IS, inform_ref(IP, IS, R))
props inst(R, place),place(X, R),
      name_of(X, N),value(N, 東急ハンズ)
```

QLF は、ひとつの発話行為にひとつの式が対応するようになっている。たとえば「東急ハンズは」という発話でも、疑問文になりえるので、ひとつの QLF に翻訳できる。

現在、発話パターン から QLF への翻訳規則はドメインの制限なども考慮にいれながら、人間が作成している。この規則を使う場合には、発話の中に規則に関係ない単語があっても無視することもできるように設計してある。したがって、QLF に変換できないような語を含む発話でも、それを理由に認識結果の候補からはずされることはない。

3.3 文脈処理 — QLF 列の評価/解釈

QLF は、文脈によって、その真の内容が変化する。たとえば、「東急ハンズは」という発話は、文脈によっては「東急ハンズはどこですか」の意味になることもあれば、「東急ハンズは百貨店ですか」の意味になることもある。QLF が照応表現を含むような場合は、まず、照応表現を推定する必要がある。

本システムでは、対話の履歴を用いて、文脈を考慮した解釈と照応表現の解消をおこなう。ここで、文脈的な正当性の評価もおこない、解消できない照応表現を持つ候補などにはペナルティをあたえる。

3.4 応答の生成

QLF が完成したら、それにもとづいて応答を生成する。したがって、休止があっても、応答が生成される場合とされない場合がある。たとえば、次のような発話について考えてみる。

わかりました (休止) 駅から (休止) は
どのくらいの距離ですか (休止)

たとえば、この発話では、「わかりました」の後の休止のところで、「わかりました」に対応した QLF が作られるが、別に応答するほどの意味はないので、何も答えない。次に、「駅から」の後の休止のところでは、十分に推論できないため応答を生成できない。しかし、「駅からはどのくらいの距離ですか」を使った QLF は十分に応答を生成できる。

現在は、応答が生成され、長い休止が検出されたときに、システムは応答文を発声する。しかし、この応答のタイミングも、応答の内容や対話の文脈によって変える必要があるだろう。

たとえば、相手が勘違いしたことをいっていることに気づいたら、相手がいい終る前にとがめるためにも相手の発話に割り込んだ方がいい場合もある。また、応答が生成されないほどの漠然とした発話がおこなわれた場合、相手がそれに気づいて補足するまで黙って待った方がいい場合もあれば、早期に必要な情報を発話するように促すような発話をした方がいい場合もあるだろう。

4 まとめと今後の課題

音声対話システムにおいて、休止を区切りとして音声認識・言語処理をおこない応答を生成する対話処理手法について提案した。今後、この手法を WOZ 方式で収録した実対話データを利用して評価していく予定である。

今後は、音声理解だけでなく、やりとり全体として「話し言葉」を意識した手法を対話システムに導入していく必要があるだろう。話し言葉のやりとりに特徴的な現象として「割り込み」や「あいづち」があげられる。こういった現象に対しては、もちろん、韻律の認識や合成面であいづちらしさを実現して総合的に対処していくことが不可欠になっていくだろう。

参考文献

[1] A. Takeuchi and K. Nagao. Communicative facial displays as a new conversational modality. In *INTERCHI-93*, pp. 187–193, 1993.

[2] 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—. 信学論 (D-II), Vol. J77-D-II, No. 8, pp. 1417–1428, 8 1994.

[3] 安藤ハル他. インテリアデザイン支援システムを対象としたマルチモーダルインタフェースの評価. 信学論 (D-II), Vol. J77-D-II, No. 8, pp. 1465–1474, 8 1994.

[4] 伊藤克亘他. 音声・視覚・画像をもつインタラクションシステム. 情報処理学会音声言語情報処理研究会資料, 1995.

[5] K. Watanuki, K. Sakamoto, and F. Togawa. Analysis of multimodal interaction data in human communication. In *Proc. of ICSLP*, pp. 899–902, 1994.

[6] 菊池英明他. 音声対話インタフェースにおける発話権管理による割込みへの対処. 信学論 (D-II), Vol. J77-D-II, No. 8, pp. 1502–1511, 8 1994.

[7] 天野明雄他. 階層的フィードバックに基づいた音声対話システムの高度化. 情報処理学会研究グループ資料, Vol. 93-SLP-3, pp. 51–54, 2 1994.

[8] 国立国語研究所 (編). 話しことばの文型 (1), 国立国語研究所報告, 第 18 巻. 秀英出版, 1960.

[9] 永井明人他. 概念素に基づく音声対話理解方式における意味スコアリング法. 日本音響学会講演論文集, pp. 151–152, 1994.

[10] 上條俊一他. 休止を処理の区切りとした自由発話理解. 情報処理学会第 50 回全国大会講演論文集, pp. 3-91–3-92, 1995.