音声言語情報処理 7－6
（1995．7．20）

音声対話システムにおける プロンプト音声再送タイミング制御法

西 宏之　　北井 幹雄
NTT ヒューマンインタフェース研究所
〒238-03, 神奈川県横須賀市武 1-2356, TEL 0468-59-8771,
FAX 0468-55-1021, e-mail nishi@nttspch.ntt.jp

あらまし　従来の音声対話システムでは、システムのプロンプト音声の後に一定時間ユーザの発話が開始されない場合に再プロンプトを行う処理が一般に用いられているが、具体的にどのようなタイミングで再プロンプト音声を送出すればよいかについては明確に示されておらず、現実のシステムではシステム設計者が経験的に設定しているのが実状である。本報告では、始めに、音声メールや留守番電話などで対象とされる伝言メッセージ録音時の対話を例に、システムプロンプト音声送出後のユーザ発話までのポーズ時間長のデータの収集結果およびその分析結果を示す。次に前記ポーズ時間長データの統計的性質を用いて、ユーザが発話できない状態あるいは発話を躊躇している状態を検出し、分かり易く言い換えたプロンプト音声を送出するタイミングを合理的に設定する方法を提示する。

和文キーワード　音声対話，プロンプト、再送、発話促進，タイミング

Control Method of Re-Prompting timing for Speech Dialog Systems

H. NISHI, M. KITAI
NTT Human Interface Laboratories.
1-2356, Take, Yokosuka-shi, Kanagawa 238-03, Japan

Abstract　New designing method of re-prompting timing for speech dialog systems is described. In order to establish the better human interface of speech dialog systems, the timing of the system prompt are important. Espetially, the re-prompting timing for the users who hesitate or can not begin to utter after the system prompt is very important because the situation often causes the fatal problems to complete the dialog. Firstly, this paper shows the experimatal result of the pause length between dialog system's prompts and user's utterances. Secondly, the statistical characteristics of the data are analysed. Finally, using the result of the previous analysis, the new method that the system detect wheather the user will begin the utterance or not.

英文 key words　dialogue, utterance promoting, utterance timing, end point detect

1 はじめに

音声をメディアとした通信サービスの対話処理およびヒューマンインタフェースの検討を進めている [1]-[4].

音声をメディアとして用いた対話システムでは、システムのガイダンスまたはプロンプト音声に対してユーザが発話を行うという手順を繰り返す方法が一般に用いられる。このような系では、ユーザがシステムのガイダンス音声の意味を十分に理解できることや、システムに対する要求を十分に理解した上で対話を開始することを前提にしていることが多い。従って、ユーザがプロンプト音声がよく聞き取れなかったり、プロンプト音声の意味内容を十分理解できない場合、ユーザが発声を開始できずに、無言状態に陥るという問題が生じる。

このような状況に対応するため、システムのプロンプトに対して、種々の理由でユーザが発話できない状態あるいは発話を躊躇している状態を検出し、分かり易く言い換えたプロンプト音声を送出する手法が提案されている [5]。

従来の研究手法では、システムの最初のプロンプト音声の後に一定時間のポーズを検出した後、再プロンプトを行う手順が提案されているが、具体的にどのようなタイミングで再プロンプト音声を送出すればよいかについては明確に示されておらず、現実のシステムではシステム設計者が経験的に閾値を設定しているのが実状である。

本報告は、音声対話システムにおいてユーザがシステムのプロンプト音声に続いて発声を開始するまでのタイミングデータに基づいて、どのくらいの無音時間長が検出されればユーザが無言状態であると判断でき、再プロンプトを送出してよいかの判定基準となる値を求めることを主たる目的とする。

まず始めに、対話型のメッセージ伝言形音声対話システムを対象として、ユーザがシステムのプロンプト音声に続いて伝言メッセージを録音するタイミングデータを収集した結果を明らかにする。その際、プロンプトの内容や属性によってタイミングデータの統計的な性質に有意な差が存在するか否かを明確にする。次に、実際の系でシステムが処理すべき手順に従い、ユーザの無言状態を検出するための閾値と前記データとの関係を明確にする。

最後に、実際のユーザの振る舞いの観察結果に基づき、本提案手法の問題点を考察するとともに今後の検討課題を整理する。

2 発声タイミングの定義

始めに本報告で使用する発話タイミングに関連する用語の定義を行う。用語の定義および対話における位置づけを図1に示す。

これらは以下のように整理される。

(1)T_{pre}： システムがプロンプト音声を送出し終わってから検出される無音時間長である。

(2)$T_{prethre}$： 上記T_{pre}に対する閾値で、T_{pre}が$T_{prethre}$に達した時点でユーザが発声を開始してなければシステムは再プロンプトを送出する。

(3)T_{post}： ユーザの発声開始後に検出される無音時間長である。

(4)$T_{postthre}$ 上記T_{post}に対する閾値で、T_{post}が$T_{postthre}$に達した時点で、システムは次のプロンプトを送出する。

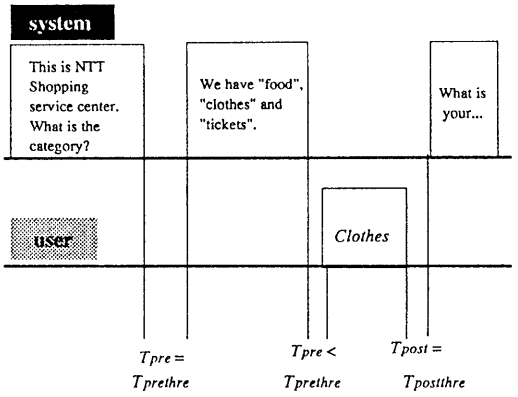

図 1: タイミング関連用語の定義

上記 (3) および (4) の制御方法すなわち、ユーザの発声終了を検出し適切なタイミングでシステムプロンプト音声を送出するタイミング設計法 [6][7] およびユーザにとってどのようなタイミングが好まれるか [8] については既に報告した。

一方、上記 (2) のような状況はユーザがプロンプト音声をよく聞き取れなかったり、プロンプト音声の意味内容を十分理解できない場合に発生し、最悪の場合ユーザはその後、一度も発声することなく対話を終了させてしまうこともある。これに対処するためには、システムはユーザの発声状況を慎重に吟味し、ユーザがパニック状態に落ち入る前に、適切なタイミングで再プロンプトを送出しなければならない。本報告の主目的は、この$T_{prethre}$を、T_{pre}の統計的な性質から求めることにある。

前記無音時間長T_{pre}が$T_{prethre}$に達した時点でユーザが発声を開始してなければシステムは再プロンプトを送出するという手順は、極めて単純なものである。しかしながら、ユーザの音声の有無を検出する手法については過去に多くの研究がなされてきているが [9][10][11]、本稿では音声の検出手法については触れず、検出されたオンオフパターンの取り扱いに焦点を絞って論ずることとする。音声の有音区間と無音区間の統計的性質につい

ては種々の報告がなされているので[12][13][14][15]、これらの研究結果を踏まえるとともに、本報告で述べる対話型のメッセージ伝言形音声対話システムに対象を絞ったデータの収集結果の報告も含めて以下の議論を進める。

3 T_{pre} の統計的な性質

3.1 実験条件

表1に実験条件を示す。また、実験に用いた系を図2に示す。

表 1: T_{pre} データ収集の実験条件

項目	内容
対話場面	電話の伝言メッセージの録音
発声者	48名 (うち男性24名)
対話内容 (パターン1)	(1)S: はい, 鈴木です. ただ今不在です 伝言がございましたら ピーという音の後にどうぞ (2)U: 用件を言う (発声内容自由) (3)S: どうも有難うございました
対話内容 (パターン2)	[S:System, U:User] (1)S: はい, 鈴木です ただ今留守にしております 恐れいりますが、 どちら様でしょうか? (2)U: 名前を名乗る (発声方法は自由) (3)S: 戻りましたら電話させますので 電話番号をどうぞ (4)U: 電話番号を言う (発声内容自由) (5)S: 伝言がございましたら ピー音の後にお話し下さい (6)U: 用件を言う (発声内容自由) (7)S: どうも有難うございました
測定内容	S:System 発声終了後、U:User が発声を開始するまでの時間長
分析条件	サンプリング周波数: 8kHz フレーム周期: 10ms フレーム長: 10ms 有音/無音閾値: パワー最小値+ 3dB
データ数	パターン1: 96 (48 × 2) パターン2: 144 (48 × 3)

対話の混乱を避けるため、発声者には、予めデータ収集の主旨を説明した。発声内容についてはプライバシー保護のため内容に規制を設けず、仮名を用いることなどを許した。ただし、できるだけ自然なデータを収集できるよう日常の電話使用時と同じように振る舞うよう依頼した。

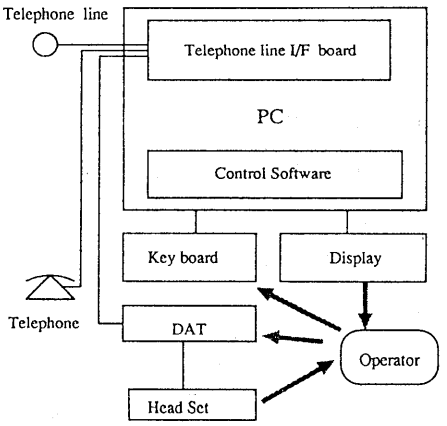

図 2: 実験システムの構成

実験システムはＰＣに電話回線インタフェースボードを挿入し、本実験のためのソフトウエアを作成して使用した。電話回線インタフェースボードは、応答メッセージの登録の他、電話着信の検出、回線の閉成、応答メッセージの送出、伝言メッセージの録音、回線の開放などの機能を持つ。

実験システムはプロンプト音声送出後、ピーという音の直後から録音を開始し、次のプロンプト音声の送出開始とともに録音を終了するという動作を繰り返す。従って、各録音された音声ファイルには先頭部分に無音状態があり、その後に伝言等のメッセージ音声が収録されていることになる。即ち、各ファイルの先頭部分に存在する無音状態区間の長さを計算機を用いて測定することにより、目的とする統計データを機械的に収集することができる。表に示したように、各音声ファイルは8kHz、10msのフレーム長、フレーム周期でパワーを計算される。求められたパワー値の系列から、各音声ファイルごとの最小値が求められ、これに3dBを加えられた値がその音声ファイルにおける有音/無音の閾値として設定される。その閾値を元に先に述べたファイルの先頭の無音区間長が計測される。

3.2 実験結果

全データを対象としたしたポーズ長の分布を図3に示す。平均974ms、標準偏差639msであり、3σ の長さをもって、無言状態と判断する場合、判断閾値は約2900msとなることがわかった。

次にパターン1、パターン2個別のポーズ長データを図4および5に示す。

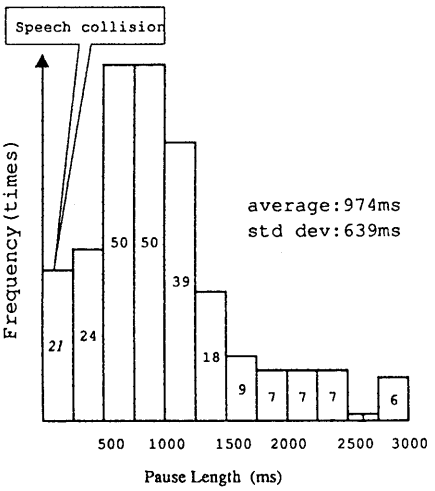

図 3: パターン１、２混合のポーズ長

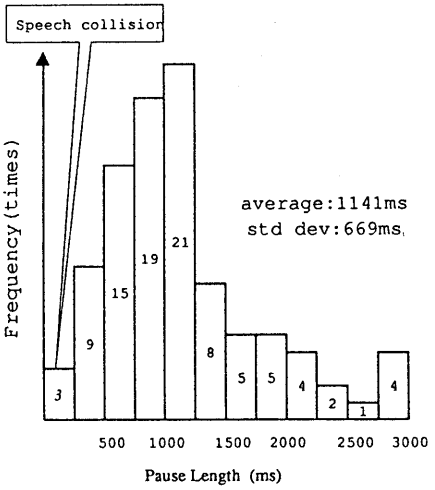

図 4: パターン１のポーズ長

パターン１では平均1141ms、標準偏差669ms、パターン２では平均863ms、標準偏差592msであり、3σで無言状態と判断する場合、判断しきい値はパターン１で約3150ms、パターン２では約2640msとなり、しきい値に500ms程度の差が生じることがわかった。

本結果から、以下のことが明らかになった。

(1) パターン１の方が、パターン２に比べてユーザが発声を開始する前のポーズ時間長が長い。

(2) パターン１の方が、パターン２に比べてガイド音とユーザ発声が衝突する確率が低い。

上記(1)項の理由として、以下の状況が考えられる。

パターン１では、一往復の対話しかないため、ユーザは自分の名前や、用件などを一度に頭の中で整理する必要があるため、発声開始までの時間長が長くなるものと考えられる。一方、パターン２では、名前、電話番号、用件など、システムが必要な情報を細切れにして質問してくるので、ユーザは一度に複数の情報を整理する必要がないことから、パターン１に比較して短い時間で発声を開始できるものと考えられる。

また、(2)項の理由としては、上記(1)の状況と同様の理由で、ユーザにとって思考時間をほとんど要しないため、逆にピー音を待ちきれずに発声を開始してしまうものと考えられる。

さらに、パターン２の中で、項目ごとのポーズ長を、図6、図7および図8に示す。

発呼者名、電話番号、用件のうち電話番号が最もポーズ時間長が短い。すでに自分の名前を名乗った後であるので、発声への抵抗が少ないものと考えれられる。

発呼者名は着信直後の最初の発声であり、緊張感も手伝ってやや長くなったと理解できる。また、用件は内容を整理する必要があることから、さらにポーズ時間長が長くなるものと考えられる。各パターンのポーズ時間長の平均値を表2に整理した。

表 2: パターン別ポーズ長平均値 (ms)

パターン名	ポーズ長平均値 (標準偏差)
全データ	974(639)
パターン１	1141(669)
パターン２	863(592)
発呼者のみ	820(597)
電話番号のみ	797(638)
伝言のみ	970(521)

4 対話制御への応用と問題点

前節で求められた無言状態の検出の閾値をT_i (iは第i番目のプロンプト音声であることを示す) とすると、システムのプロンプト音声送出後、検出される無音時間長がT_iに達してもユーザが発話を開始しなければ再プロンプトを送出するという手順になる。

ここで、T_iは対話の内容によって可変であり、実験結果の項で例を示したように、必要以上の無音検出をしないように最適設計することが可能である。これにより、ユーザが心理的な不安感やパニック状態に陥ることから対話を異常終了する以前に再プロンプト音声を送出し、対話を継続させることが期待できる。

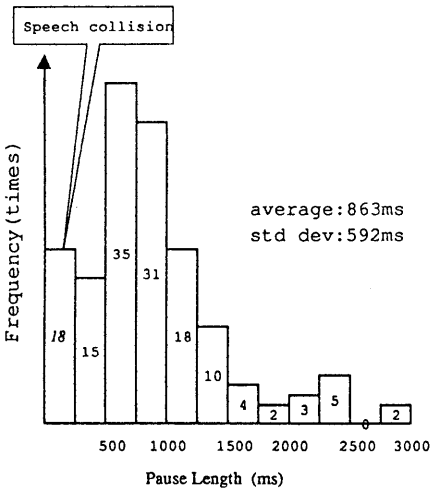

図 5: パターン２のポーズ長

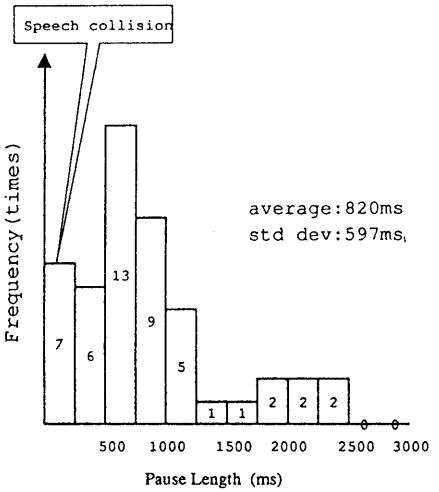

図 6: 発呼者名発話におけるポーズ時間長

以上、本報告ではユーザの発声前の無音時間長の統計的性質からユーザの無言状態を検出し、再プロンプトを送出することでパニック状態を回避するという立場で、議論を進めてきた。この手法は、一旦無言状態に陥ったユーザはシステムから再プロンプトが送出されるまで無言を継続するという仮定の上に成り立っているが、実際はユーザは「えっと」、「どうしよう」、「困ったな」など、躊躇、困惑、不安などの間投詞を発声してしまうことも多く、このような場合には、本報告の手法は直接的には適用できない。上記、間投詞などを音声認識し、その内容を理解した上で対話を進めるという極めて困難な処理に挑まなければならない。

これらを解決するためには、ユーザが心理的に困惑した状況でどのような発話行為を行うのかについて慎重なデータの収集と分析が必要であり、同時にそれらの分析結果をユーザの心理状態を認識するという工学的手法にまで展開する研究が必要となる。

5 まとめと今後の検討課題

音声対話システムにおいて、システムのプロンプトに対して、種々の理由でユーザが発話できない状態あるいは発話を躊躇している状態を検出し、分かり易く言い換えたプロンプト音声を送出するタイミング設計法を提案した。

始めに、音声メールや留守番電話などで対象とされる伝言メッセージ録音時の対話を例に、システムプロンプト音声送出後のユーザ発話までのポーズ時間長のデータを収集し分析した結果、

- 伝言など比較的長いメッセージを発声する場合は平均 1141ms、標準偏差 669ms、名前・電話番号など比較的短いメッセージの場合は平均 863ms、標準偏差 592ms であり、ユーザの思考時間を要するほどポーズ時間長も長くなる
- 長いメッセージの場合は短いメッセージの場合に比べてプロンプト音声後のピー音とユーザ発話が衝突する可能性が小さい
- 3σ で無言状態と判断する場合、判断閾値は長いメッセージの場合で約 3150ms、短いメッセージの場合では約 2640ms となる

ことがわかった。

一方、本報告では、一旦無言状態に陥ったユーザはシステムから再プロンプトが送出するまで無言を継続するという仮定の上に検討したが、実際は躊躇、困惑、不安などを表す間投詞を発声してしまうことがあるので、今後は上記問題にも対応できる手法について検討する。

日頃ご指導いただく NTT ヒューマンインタフェース研究所北脇音声情報研究部長、西野音声サービス方式研究グループリーダに感謝いたします。

参考文献

[1] 北井, 西:" 電話取り次ぎシステムにおける実対話分析", 音講論,1-7-15(1994.3).

[2] 北井, 西:" ワードスポッティングを用いた電話音声対話システムにおける対話手順の評価", 信学論 A Vol.J 77-A No.2 pp.251-258,1994 年 2 月.

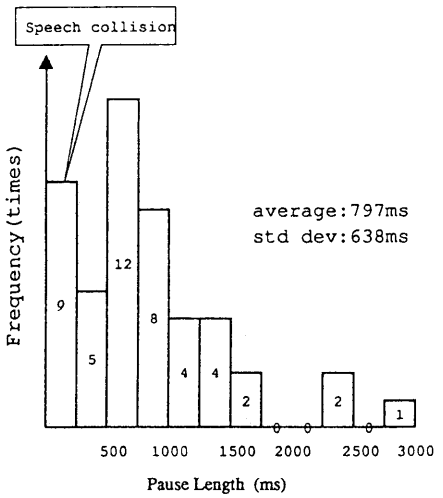

図 7: 電話番号発話におけるポーズ時間長

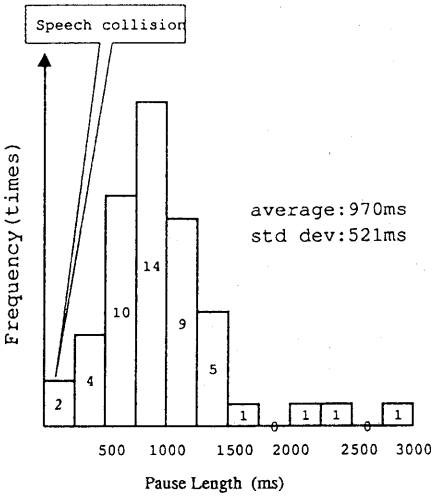

図 8: 伝言発話におけるポーズ時間長

[3] 西, 北井:" 尤度と正解率の統計的関係に基づく認識結果確認対話戦略", 信学論 A Vol.J 77-A No.2 pp.259-266,1994 年 2 月.

[4] 西, 小島:" 電話会話における「相づち」の発言促進効果", 第 3 回ヒューマンインタフェースシンポジウム,3132(1987.10).

[5] Rabinar,:"Applications of voice processing to telecommunications",Proceedings of the IEEE,Vol.82. No.2,February,1994

[6] 西, 小島, 五味:" 人間機械間の音声対話におけるタイミング", 第 2 回ヒューマンインタフェースシンポジウム,2212(1986.10).

[7] 西, 五味, 小島:" 音声対話における確率的発声終了検出法", 信学論 D Vol.J 70-D No.11 pp.2108-2114,1987 年 11 月.

[8] 西, 北井:" 蓄積形音声対話システムにおける発話促進要因の分析と評価", 情処研報, Vol.95, No.16, 95-SLP-5-8(1995 年 2 月).

[9] 八塚陽太郎:" 極性系列に着目した高感度音声検出器", 信学論 (A),J63-A,7,pp.413-420(昭 55-07).

[10] 佐藤, 新田, 小原:" 音声パケット通信のための有音検出方式", 信学技報,CS89-23-37,(1989).

[11] 白木, 鈴木, 野口, 庄司:" 適応型零交差しきい値を用いた音声検出方式", 信学春期全大, 予稿集 3,pp.79(1989).

[12] P. T. Brady:"A statistical analysis of on-off patterns in 16 conversations",BSTJ,47-1,pp.73-89,(Jan. 1968).

[13] 比企, 金森, 大泉:" 連続音声中の音韻区分の持続時間の性質", 信学誌, 50-5,pp.849-856(昭 42-05).

[14] 藤崎, 大村:" 連続音声における発声区間および休止の性質について", 秋季音講論,2-1-9,pp.221-222(昭 46).

[15] 高木, 保浦, 板橋:" 模擬対話音声における各種区分の持続時間の性質" 信学技報,SP92,pp.63-70(1992).