第3章　無言検出法に関する検討

3.1　はしがき

ユーザが無言状態となった場合に対処するため，システムのガイダンス後，一定長の無音区間が検出されれば無言状態と判断し，わかりやすい表現に直したガイダンスを送出する手順が提案されている [1]．しかし，無言状態であると判断するしきい値の設定法については明確に示されていない．本章では，システムのガイダンス後，ユーザが発話を開始するまでの無音区間長の実測値統計に基づき，できるだけ短いしきい値で無言状態を検出するための検討を行う [18][19]．

3.2　無言検出の考え方と従来法の問題点

図 3.1 に対話中の無音区間と無言検出の例を示す．同図に用いた用語を下記に定義する．

(1)T_u：ガイダンス送出後，ユーザ発話前の無音区間長

(2)T_r：上記 T_u に対するしきい値．T_u が T_r に達した時点で無言状態と判断して，システムは再ガイダンスを送出する．

(3)T_p：ユーザの発話開始後に観測される無音区間長．

(4)τ_{end}：上記 T_p に対するしきい値．T_p が τ_{end} に達した時点でユーザの発話終了を検出し，システムは次のガイダンスを送出する．

対話内容によって思考時間が異なり，ユーザ発話前の無音区間長 T_u の分布も異なる．T_u の分布例およびしきい値 T_r の設定例を図 3.2 に示す．対話 A は回答にあたって思考をあまり必要とせず，発話前の無音区間長が短い対話の例，対話 B は回答にあたって相当の思考を必要とし，発話前の無音区間長も長くなる対話の例である．仮に，無言検出しきい値として図の T_{r1} を設定した場合を考える．しきい値 T_{r1} は，対話 A に対しては，ユーザの発話前に再ガイダンスを送出するという問題を生じないが，対話 B ではユーザに発話の

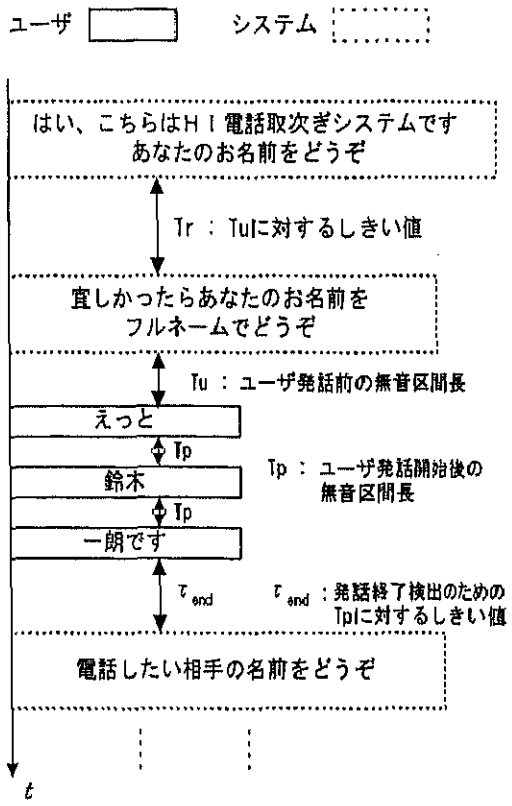

図 3.1: 対話中の無音区間と無言検出

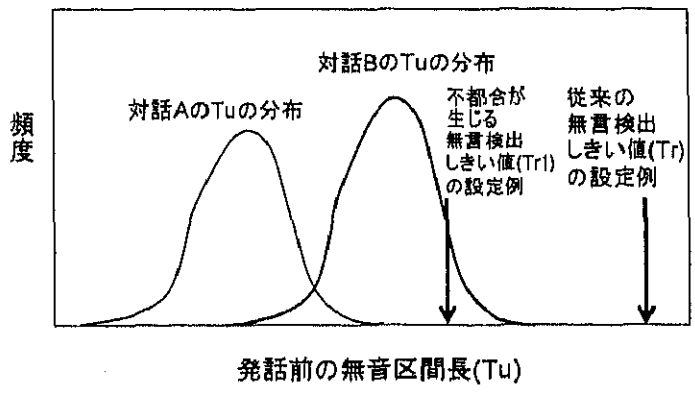

図 3.2: 従来の無言検出しきい値設定の考え方

意思があるにも関わらずシステムが再ガイダンスを送出する場合が生じ，ヒューマンインタフェース上大きな問題となる．このような状況を避けるため，従来法では，しきい値を図3.2のT_rに示すように，十分に長い一定値(例えば5秒程度)に設定していた．その結果，システムが無言状態を検出する前にユーザが回線を切断する場合があり，対話完了率が低下するという問題が生じていた．

3.3 発話前無音区間長の統計的性質に基づく無言検出しきい値設定法

3.3.1 実測値統計に基づく無言検出の考え方

しきい値として，十分に長い一定値を用いる従来法の問題点を改善するため，対話内容に対応した発話前無音区間長の実測値統計に基づき，より短いしきい値を設定する方法を検討する．図3.3の対話AのT_uの分布を例に，本手法の考え方を説明する．対話Aの発

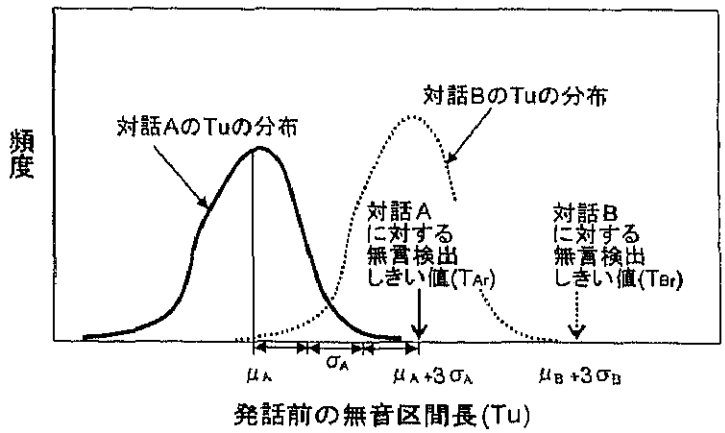

図 3.3: 実測値統計に基づく無言検出しきい値設定の考え方

話前無音区間長の平均値をμ_A，標準偏差をσ_Aとする．平均値に標準偏差のn倍を加えた値$(\mu_A + n\sigma_A)$の時間以後に発話を開始するユーザの割合が，十分に小さくなるようにnを選べば，$\mu_A + n\sigma_A$を無言検出しきい値に設定することができる．例えば，T_uが正規分布ならば，$n = 3$，すなわち$\mu_A + 3\sigma_A$以後に発話を開始するユーザの割合は0.135 %であり十分に小さい．本論文では，$n = 3$，すなわちしきい値を

$$T_{Ar} = \mu_A + 3\sigma_A$$

とするが，nはシステム運用上許容される誤判断の程度に応じて設定することができる．なお，nは必ずしも整数である必要はない．

対話Bについても，同様の考え方に基づき，発話前無音区間長の平均値をμ_B，標準偏差をσ_Bとして，

$$T_{Br} = \mu_B + 3\sigma_B$$

をしきい値として設定する．このように統計的な根拠に基づいてしきい値を設定することにより，ユーザ発話前にシステムが再ガイダンスを行うことを防ぐとともに，従来より短い時間で無言状態を検出することができるものと期待できる．

以下，発話前の無音区間長が比較的長い対話 (図3.3の対話Bに相当) と，思考をあまり必要とせず発話前の無音区間長も短い対話 (対話Aに相当) とを対象に，ユーザ発話前の無音区間長T_uの実測値統計を求める．その結果から対話内容によるT_uの分布の違いを確かめるとともに，上記考え方に従ってしきい値を求め，本設定法の有効性を明らかにする．

3.3.2 実験条件

表3.1にユーザ発話前の無音区間長データ（T_u）を収集するための実験条件を，図3.4に実験系の構成を示す．対話内容によりT_uの分布が異なるか否かを検討するため，図3.3の対話Bを想定してパターン1 (伝言内容を一度に発話させる従来型) を，対話Aを想定してパターン2 (名前, 電話番号など短いメッセージに分割して発話させる対話録音型) の対話を用いた．実験システムはガイダンス送出後，ピーという音の直後から録音を開始し，次のガイダンス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．したがって，録音された音声ファイルには先頭部分に無音区間があり，その後に伝言等のユーザ側メッセージ音声が収録される．各ファイルの先頭部分に存在する無音区間の長さを測定することにより，目的とするデータ (T_u) を機械的に収集する．表に示したように，10msのフレーム長とフレーム周期でパワーを計算する．背景雑音を含む音声から音声区間のみを検出する方法には種々の報告がある [11][20][21][22]．本章では背景雑音レベルが比較的小さく一定していることから，各音声ファイルごとのパワー最小値を求め，これに3dBを加えた値をその音声ファイルにおける有音／無音のしきい値とした．設定されたしきい値をもとに，ファイルの先頭の無音区間長 (T_u) を計測した．

表 3.1: ユーザ発話開始前無音区間長の実験条件

項目	内容
対話場面	電話の伝言メッセージの録音
被験者	48 名 ・男性：24 名 ・女性：24 名 ・20 歳〜39 歳：26 名 ・40 歳〜59 歳：22 名
電話回線	内線電話回線
対話内容 (従来型) (パターン 1)	[S:システム, U:ユーザ] (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 [ピー] (2)U:用件を言う (発話内容自由) (3)S:どうも有難うございました
対話内容 (対話録音型) (パターン 2)	(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？[ピー] (2)U:名前を名乗る (発話方法は自由) (3)S:戻りましたら電話させますので 電話番号をどうぞ [ピー] (4)U:電話番号を言う (発話内容自由) (5)S:伝言がございましたら ピー音の後にお話し下さい [ピー] (6)U:用件を言う (発話内容自由) (7)S:どうも有難うございました
測定内容	システム発話終了後, ユーザが 発話を開始するまでの時間長:T_u
パワーによる 有音／無音 測定条件	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／無音しきい値：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)

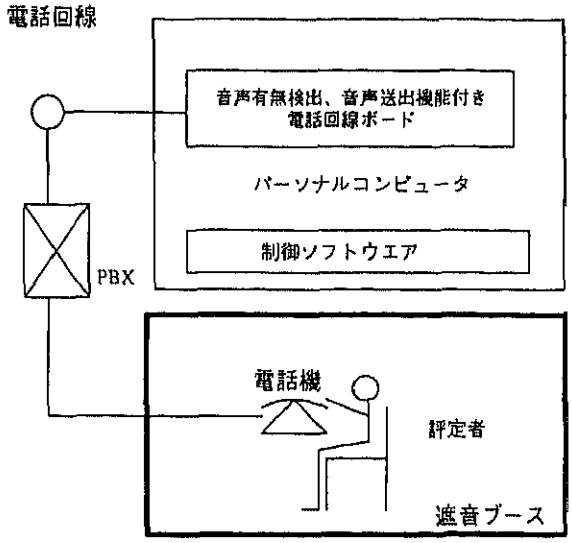

図 3.4: 実験系の構成

3.3.3 実験手順

被験者は次のような手順でシステムと対話を行う.

(1) **システムへの電話** 被験者は指定された番号に電話をする.

(2) **オペレータからの説明** オペレータが実験の手順, 注意事項等, 簡単な説明をした後, 一旦電話を切る.

(3) **パターン 1 の対話の収録** 被験者が再度電話するとシステム (従来型, パターン 1) が応答する. ガイダンスに従ってメッセージを録音し, 終了後電話を切る.

(4) **パターン 2 の対話の収録** 被験者が電話するとシステムは対話録音型 (パターン 2) で応答する. ガイダンスに従ってメッセージを録音する.

(5) **パターン 1 の対話の収録** 被験者が電話するとシステムは再度従来型 (パターン 1) で応答する. ガイダンスに従ってメッセージを録音する.

収録された音声は前記実験システムによって機械的に処理され, 発話前の無音区間長が測定される. なお, メッセージの内容は特に指示をせず, 被験者の任意とした.

3.3.4 実験結果

パターン1 (従来型) の無音区間長の分布を図3.5に，パターン2 (対話録音型) の分布を図3.6に示す．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，

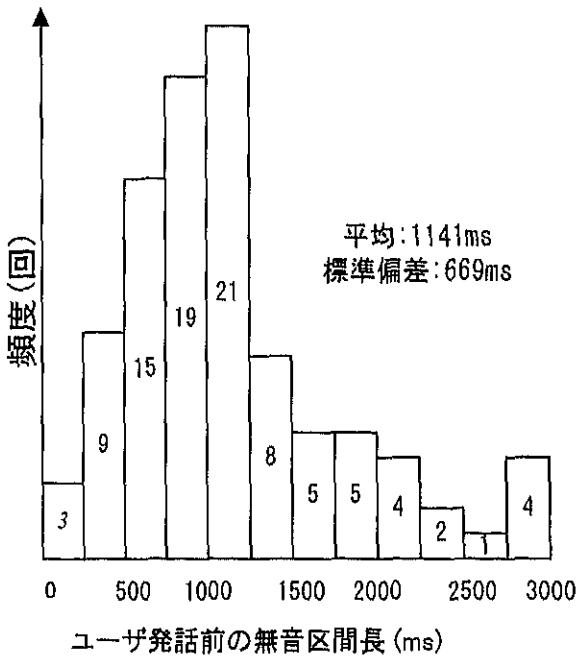

図 3.5: ユーザ発話前の無音区間長 (Tu：パターン1)

標準偏差592msであった．χ 2乗分布を用いて検定したところ，危険率1%で分布の差が有意であった．平均値μに標準偏差σの3倍を加えた値$(\mu+3\sigma)$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であった．

図中，横軸の値が0のデータは，ガイダンス後のピー音に重畳してユーザが発話を開始した場合の頻度を示す．パターン2の方がパターン1より「ピー音」とユーザ発話が重畳する確率が高いことがわかった．

パターン2のうち，発話項目ごとの結果を，図3.7，図3.8および図3.9に示す．各パターンの無音区間長T_uの平均値μ，標準偏差σおよびしきい値設定例$(\mu+3\sigma)$を表3.2に示す．発呼者名 (パターン2-1)，電話番号 (パターン2-2)，用件 (パターン2-3) のうち電話番号の場合が最も無音区間長の平均値が短く，続いて発呼者名，用件の順であった．しかし，パターン2-1，2-2，2-3の分布の差は小さく，有意な差は認められなかった．

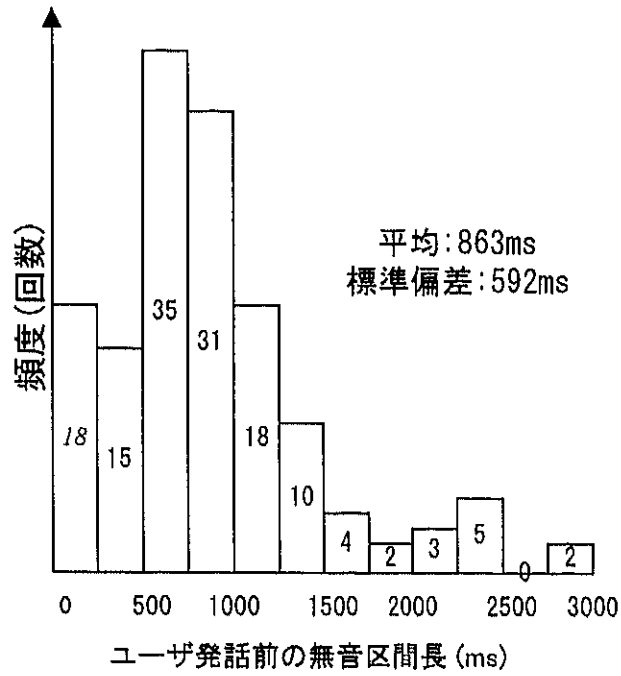

図 3.6: ユーザ発話前の無音区間長 (Tu : パターン 2)

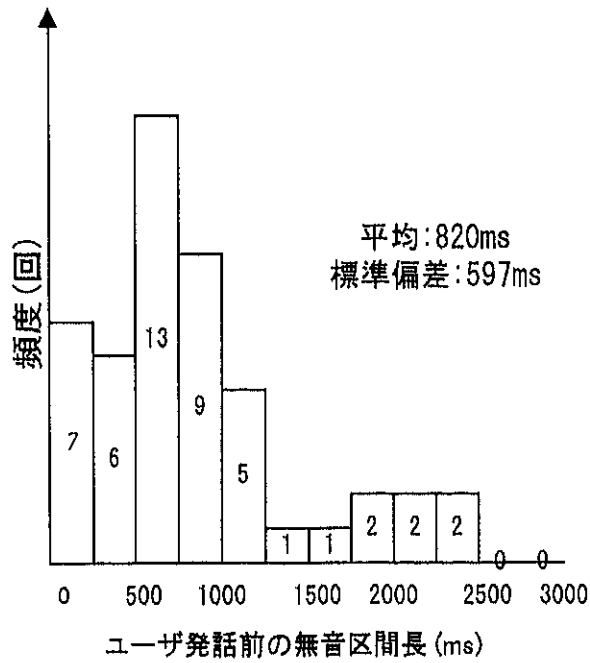

図 3.7: ユーザ発話前の無音区間長 (Tu : パターン 2-1, 発呼者名の発話)

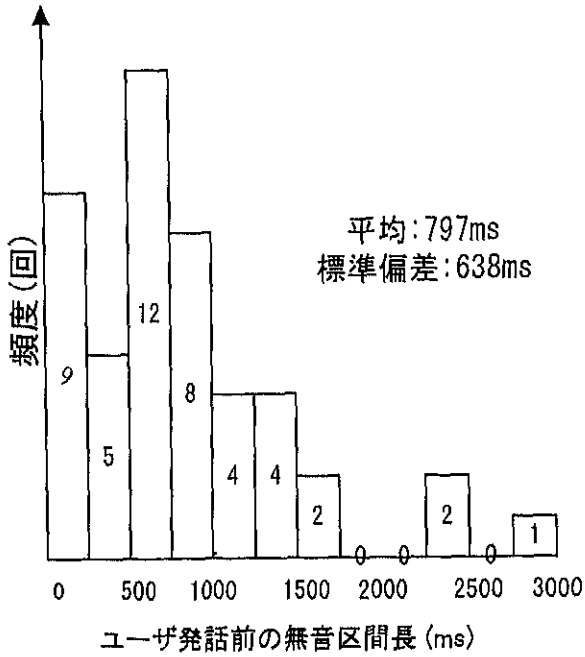

図 3.8: ユーザ発話前の無音区間長 (Tu : パターン 2-2, 電話番号の発話)

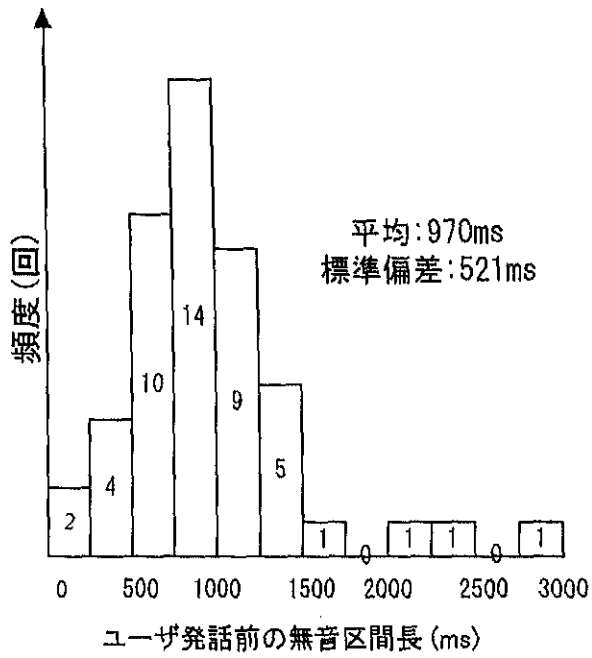

図 3.9: ユーザ発話前の無音区間長 (Tu : パターン 2-3, 用件の発話)

表 3.2: パターン別発話前無音区間長 (Tu) の平均値等 (ms)

パターン名	T_u の平均値 (標準偏差)	しきい値の設定例 (μ +3 σ)
パターン 1 (従来型)	1141(669)	3148
パターン 2 (対話録音型)	863(592)	2639
・発呼者名 (2-1)	820(597)	2611
・電話番号 (2-2)	797(638)	2711
・用件 (2-3)	970(521)	2533

3.4 考察

パターン 1(伝言内容を一度に発話させる対話) の発話前無音区間長 (T_u) は，パターン 2(名前, 電話番号など短いメッセージに分割して発話させる対話録音型の対話) より長くなることがわかった．パターン 1 では，名前や用件など発話内容を整理する思考時間が必要であるためと考えられる．一方，パターン 2 では，名前，電話番号など，発話内容を順次指示されるので，発話内容が簡単になり，短時間で発話を開始できるものと考えられる．平均値 μ に標準偏差 σ の 3 倍を加えた値 ($\mu + 3\sigma$) で無言状態と判断する場合，しきい値に 500ms 程度の差が生じることがわかった．以上の結果から対話内容に応じてしきい値を制御することにより，従来より短い時間で無言状態を検出できるものと期待できる．

なお，ガイダンス後の「ピー音」とユーザ発話が重畳する確率は，パターン 2 の方がパターン 1 より高いことが明らかになった．パターン 2 ではパターン 1 に比べて発話の開始が早くなることにより,「ピー音」とユーザ発話が重畳する確率も高くなるものと考えられる．

3.5 むすび

無言検出前にユーザが回線を切断するという問題点に対処するため，無言検出しきい値 T_r を，ユーザ発話開始前の無音区間長 T_u の実測値統計から求め，できるだけ短いしきい値で無言を検出する方法を検討した．その結果，以下の項目が明らかになった．

- 一回の発話に名前，電話番号，伝言内容など複数の項目を含む場合 (パターン 1) は，各項目ごとに分けて発話を行う場合 (パターン 2) に比べて，発話前の無音区間長 T_u

が長くなる

- 平均値に標準偏差の3倍を加えた値を無言検出しきい値とする場合，パターン1のしきい値は，パターン2より約500ms長くなる
- ガイダンス後のピー音とユーザの発話とが重畳する確率は，パターン1の方が小さい

上記第1項および第2項より，無言検出の無音区間長しきい値を設定する際，対話内容によって必要とする思考時間が異なり，しきい値の長さを変える必要があることがわかった．この結果，十分長い一定値をしきい値としていた従来法に比べ，対話内容に応じてしきい値を制御することにより，より短い時間で無言状態を検出でき，対話完了率の向上に資する見通しを得ることができた．

認識対話の場合は，ディクテーションサービスを除けば，ユーザが長文のメッセージを発話することは少ない．したがって，蓄積対話における伝言メッセージに比べ，無言検出のしきい値を短く設定することができると考えられる．しかし，思考時間をどの程度必要とするかにも影響されるので，本章で述べた手法を用い，対話内容に対応した実測値データから無言検出しきい値を設定する必要がある．