音声インタフェース評価マニュアル

Rev.1.1

2009 年 3 月

早稲田大学大学院　人間科学学術院

菊池　英明　研究室

目次

1. マニュアルの目的 3
2. ユーザビリティ評価ガイドラインの動向 4
3. 被験者選定基準 6
3.1 利用者として想定されるユーザ層を対象 6
3.2 タスク経験をファクターとして被験者を選別 7
3.3 音声インタフェース利用経験では被験者を選別しない 7
4. タスク達成率 8
4.1 タスク達成の考え方 8
4.2 タスク達成率の定義におけるタスク達成時間の扱い 9
4.3 タスク誤解の最小化 9
4.4 実験指示文設計 10
4.5 音声による実験指示 10
4.6 モダリティ利用バランスの制限 10
4.7 途中の認識誤り、操作誤りはタスク達成の判断の材料としない 10
5. 事前学習 11
5.1 発声方法の学習 11
5.2 インタフェース操作方法の学習 11
6.1 事前学習との関連 13
6.2 実験回数および間隔 13
6.3 トークスイッチ操作の習熟 13
7. 詳細評価方法 14
7.1 NASA-TLX 14
7.2 二重課題法 16
7.4 ストレス発話分析 20
8. 音声インタフェース評価におけるポイント 21
謝辞 26
参考文献 26

1. マニュアルの目的

音声認識技術や音声合成技術などの音声言語処理技術の発展にともなって、車載カーナビゲーションシステムや航空券予約電話自動応答システムなどの音声インタフェースシステムの実用化が進んでいる。しかしながら、音声メディアには他のメディアにない数多くの利点がある[1]ことや、近年の音声認識・合成技術の発展に対して、音声インタフェースの普及は十分とはいえない。石川ら[2]が指摘するように、要素技術の性能向上とは別に、音声インタフェースの使用性(ユーザビリティ)を正しく評価し向上させることが今後重要になってくる。

一般に、ユーザインタフェースのユーザビリティを正しく評価する手段として、ユーザビリティ評価のガイドラインが用いられる。ユーザインタフェース設計の標準規格である ISO 9241, ISO 13407（JIS Z 8521, JIS Z 8530）においてもユーザビリティ評価の指針が示され、広く参照されている。また、Shneiderman や Nielsen など専門家によって提唱され実際に利用されているガイドラインもある。これらの規格やガイドラインはユーザインタフェース全般を対象にしており、音声インタフェースにも適用して有効に利用できる部分が多い。しかし、言語そのものを入出力のメディアとする音声インタフェース特有の問題も多い。石川[3]は以下のように指摘している。

> GUI が基本的には、システムの機能をユーザに提示し、選択させるのに対し、音声認識では「ユーザの要求」を直接入力させるため、ユーザの想定するシステムの機能自体に誤解が生じる場合がある。特に複雑な機能を有するシステムでは、「なんと言えばいいのか」以前に「なにができるのか」が問題になる場合がある。（中略）タスクを与えての評価、すなわち、被験者に機能を既知とした評価だけが行われる場合も多い。また、インタフェース評価では、可学習性の観点から、被験者やインストラクションに大きく結果は依存する。

音声インタフェースの評価については、これまでに様々な研究者・機関によってそれぞれの目的に応じて適切な方法が検討され知見が得られている。また本研究室においても、音声インタフェースの評価に際してのいくつかの問題について実験を通じて予備的な検討を行ってきた。本マニュアルでは、広範に支持され利用されている規格やガイドラインにのっとりながら、先行研究から得られる知見を導入して、音声インタフェースの評価手法を整理し、解説する。

以下には、2 章にユーザビリティ評価のガイドラインの動向を概説し、評価において特に留意すべき被験者選定基準、タスク達成率、事前学習、習熟による影響の考慮についてそれぞれ 3, 4, 5, 6 章で概説する。

2. ユーザビリティ評価ガイドラインの動向

1999 年に“Human-centred design processes for interactive systems”と題された ISO 13407（「人間工学 -インタラクティブシステムの人間中心設計プロセス」JIS Z 8530）が国際規格として制定されている。この規格において、ユーザビリティは「ある製品が、指定されたユーザによって、指定された利用の状況下で、指定された目標を達成するために用いられる際の、有効さ、効率、満足の度合い。」と定義されている（図 1)。その際、利用の状況として、ユーザ、仕事、装置、環境があげられている。

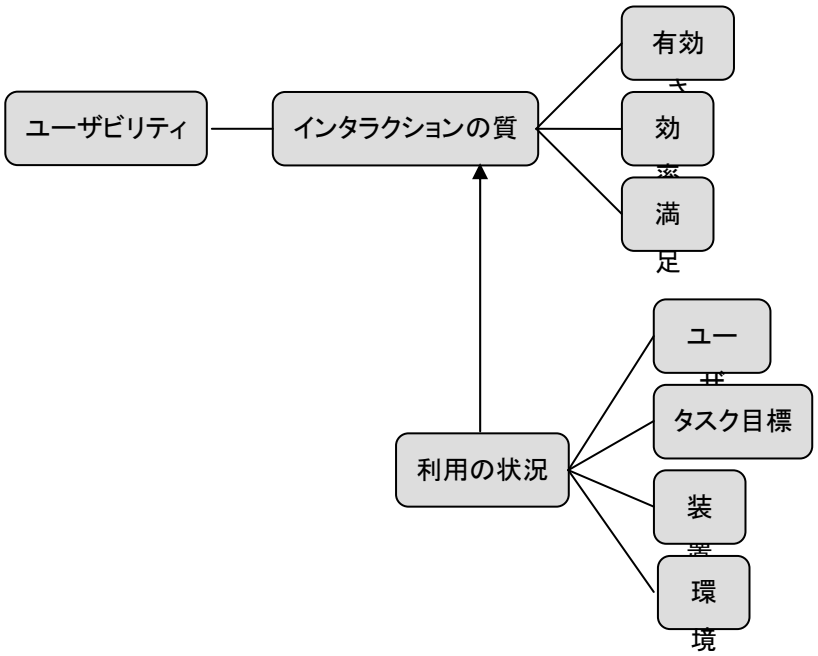

図 1. ISO 13407 における「ユーザビリティ」の定義

この規格で示されている人間中心設計の４つの活動のうち、評価に関わる「利用の状況の把握と明示」「要求事項に対する設計の評価」について以下に概説する。詳細は規格を参照されたい。

利用の状況の把握と明示

利用の状況の定義

対象とするユーザの特性:知識、技能、経験、教育、訓練、身体的特性、習慣、好み、能力など

ユーザが行う仕事:システムの利用に対する全体的な目標を盛り込む。仕事の特性（例えば発生回数及び作業の持続時間）は記述されることが望ましい。

ユーザがシステムを利用する環境:ハードウェア、ソフトウェア、資料。

物理的及び社会的環境:より広い意味での技術的な環境、物理的環境（仕事場）、周囲の環境（気温、湿度）、法制上の環境（法律、条例）、社会的文化的環境（仕事の習慣、組織の構造及び風土）

要求事項に対する設計の評価

一般

評価は次の目的のために実施する。

a）設計を改善するために利用されるフィードバックの提供。

b）ユーザ及び組織の目的が達成されてきていることの確認。

c）製品及びシステムの長期的な使用のモニター。

評価計画

次の事項に関連する評価計画を作成することが望ましい。

a）人間中心設計の目標

b）評価責任者

c）評価されるべきシステムの構成要素とその評価方法

d）評価の実施方法と試験の手順

e）結果の評価及び分析に必要な資源、及びユーザとのやりとり

f）評価活動のスケジュールとプロジェクトの日程との関連

g）他の設計活動への評価結果のフィードバックと利用

目的達成度の評価

妥当性のある結果を得るためには、評価は、現実的な仕事を行っているユーザの代表を用いた適切な方法を使用することが望ましい。

評価目的は、主目標、下位目標、二次的目標に関係づける。

フィールドにおける妥当性検証

長期間のモニタリング

ここで指摘されている要件のうち、被験者実験における被験者選定基準、目的達成度の評価尺度、事前学習、習熟による影響といった事項について、特に音声インタフェース特有の問題を考慮しながら、次章以降に記す。なお、次章以降に記すある種の指針は、経済産業省情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発「音声認識基盤技術の開発」における議論を通じて形成した。こうした指針の有効性は目的や考え方によって異なり、それらを示したうえで指針を述べ、有効範囲を明確にすることが必要と考える。また、継続して指針の修正や追加などの運用が必要である。

3. 被験者選定基準

前述したように、ISO 13407(JIS Z 8530)では、対象とするユーザの特性として知識、技能、経験、教育、訓練、身体的特性、習慣、好み、能力などがあげられている。
山岡[4]は、ユーザビリティテストにおける被験者特性について考慮すべきポイントを以下のようにあげている。

> 情報処理の過程に影響を与えるのは、年齢、教育程度、文化的背景、経験、日本語や外国語の読解力、モチベーションなどである。一方、行動に影響を与えるのは、年齢、身体面での障害、スキル、経験などである。これらの要因により反応の相違となり、個人差となっていく。

一方、音声インタフェース特有の問題として、音声認識や音声合成といった基本技術に対する経験や親和性を被験者特性としてどのように考慮すべきかという点があげられる。また、音声メディアは日常のコミュニケーションで利用されるものであり、幅広い利用者層に適用され得るとの期待がある。こうした問題を含めて、被験者実験の被験者を選定する際に考慮すべき条件を以下に記す。なお、被験者実験に際して収集すべき被験者属性については被験者選定基準とは別に議論すべきである。

a. 利用者として想定されるユーザ層を対象 … 3.1
b. タスク経験をファクターとして被験者を選別 … 3.2
c. 音声インタフェース利用経験では選定しない … 3.3

3.1 利用者として想定されるユーザ層を対象

音声メディアは日常のコミュニケーションで利用されるものであり、幅広い利用者層に適用できる音声インタフェースの登場が期待されている。一方で、現状の音声インタフェースの多くは発話表現が有限の単語に限定される、決められたタイミングでの発声しか受け付けないなどの制約を設けて、利用方法や利用者層を制限せざるを得ない。したがって、現状では、一般的なユーザインタフェース評価[Nielsen1999](p.137)と同様に、評価対象の利用者として想定される層を代表するように被験者を選定すべきである＜パタンランゲージ：**「代表ユーザ被験者選定」**＞。選定に際して例えば、以下のような属性を考慮する必要がある。

- 機器操作に対する親和性や関心
  操作方法の習得が必要な場合など。。
- 記憶能力
  操作に際して音声コマンドの記憶が必要な場合など。

これらの属性は概ね、年齢、性別、職業などのバランスを取ることで網羅される。

3.2 タスク経験をファクターとして被験者を選別

自動車運転免許を持たない者をカーナビゲーションシステム評価実験の被験者にするわけにはいかない。つまり、実験で行うタスクに対して現実的に対象となる者を被験者として選定すべきである。そのうえで、実験中にタスクを誤解した、タスクに関心を持たなかったなどは実験による不備と考えるべきである。なお、実験で行うタスクそのものの経験がなくても類似したタスクの経験があれば良い場合がある＜パタンランゲージ：**「タスク経験に基づく被験者選定」**＞。

3.3 音声インタフェース利用経験では被験者を選別しない

Nielsen は、あるインタフェースから次の新しい世代に移行する場合にテスト前にユーザをトレーニングしなければならないことがあると述べている[5]。音声インタフェースの利用経験を積極的に調査する場合を除き、音声インタフェースデザインの評価を目的とするなら、やはり音声インタフェースの利用方法を事前にトレーニングする必要がある。詳しくは 5 に述べる。

4. タスク達成率

ISO 13407 においてユーザビリティを構成する３つの要素のうちの一つ「有効さ」は「利用者が、指定された目標を達成する上での正確さ及び完全さ」として定義される。一般に、音声インタフェース研究でよく用いられる「タスク達成率」は、指定された条件下でタスク（仕事）をいかに正確あるいは完全に達成できたかを測る尺度であり、ISO 13407 における有効さの尺度に対応するものといえる。本章では、タスク達成率を計測するうえで考慮すべき点を整理する。

4.1 タスク達成の考え方

音声インタフェースにおいては、要求の実現においてユーザの発声が必要であり、その際、発声内容の表現はユーザに任される場合が多い。例えば、評価実験において実験者が「新宿区新大久保」と入力することを想定してタスクをデザインした場合、被験者がそのことを正しく理解しながら「新宿区西大久保」と勘違いして入力することは十分にあり得る。今、図 x のような状況があったとしよう。
このとき、以下のような立場があり得、立場に応じてタスク達成の判断結果が異なることがある。

(1) A と C が一致していなければならないという立場
　(a) A と B のズレを許容しない
　(b) A と B のズレは許容するが、B と C が一致しているかはわかり得ないと考える
(2) B と C が一致していれば A と C が一致していなくても良いとする立場

A. 実験者が設定したタスク ex.)「新宿区新大久保と入力」

B. 被験者が理解したタスク ex.)「新宿区大久保と入力」

C. 被験者が実施したタスク ex.)「新宿区と入力」

図 4.1 タスク達成における状況の例

(1)-(a)
タスクについての説明が適切であれば A と B のズレは生じないはずと考えればこの立場となる。そうすると A と C が一致した場合のみをタスク達成と判断することになる。しかし、本マニュアル筆者らは、タスクについての説明を充分に行うことの重要性を認めつつ、実験においてタスクについての完璧な説明は経験上難しいと考える。

(1)-(b)
(1)-(a)に対して、A と B のズレを許容するが、被験者がどのようにタスクを理解したか

を知ることは難しく、したがってBとCが一致しているかはわかり得ないと考えればこの立場になる。そうすると(1)-(a)と同様に、AとCが一致した場合のみをタスク達成と判断することになる。
本マニュアル筆者らは、被験者が理解したタスクを達成できたかどうかを答えさせることができると考える。しかし、タスク理解の基準が被験者によってまちまちだったり、同じ被験者でも気分によって基準が変わる可能性は否めない。ただし、現実の場面でもタスク設定は堅固とはいえないため、実験の動機付けをきちんと高めることを前提として、現実に即してタスク理解は被験者の判断にゆだねるべきと考える。

(2)
本マニュアル筆者らは、タスク達成を判断する際にAとCの一致を見るのではなく、BとCの一致を見ればよいという立場をとる。当然ながら、その際、タスクについての説明を充分に行い、実験の動機付けを高めることを前提とする。そのうえで、Cは観察できてもBを客観的に知ることは難しいため、自分が理解したタスクを達成できたかどうかを答えさせることで、タスク達成の判断を行うという方法が最も現実的と考える。

4.2 タスク達成率の定義におけるタスク達成時間の扱い
有効さの尺度は、時間のような効率の尺度と関連づけられる。つまり、タスク達成時間の制限によってタスク達成率は変化する。音声インタフェース評価実験においては、慣れない音声入力に手間取り音声誤認識や語彙外発話が頻発して一つのタスクの遂行に思わぬ時間がかかることがある。一方、被験者実験に先立ってタスクごとに適切な制限時間を設けることは困難である。そこで、タスク達成時間の制限を設けずに＜パタンランゲージ:「**時間無制限タスク**」＞，実験結果の分析時に制限時間を変数としたT-A(Time - Achievement rate)グラフを用いて評価する。これにより、タスクの性質と被験者の作業負荷の関係が明らかになることが期待できる。
その際、タスクが困難であれば被験者の意思でギブアップを行ってもらう．被験者によっては，すぐにギブアップしてしまうことがあるが，その際にはギブアップが可能になるまでの時間を設ける＜パタンランゲージ：「**ギブアップ制限**」＞などの工夫をする。また、被験者には「効率良く操作」の努力目標を設定する。
なお、現実的には、制限時間を設けなければ実験のスケジュール遂行が困難になることもあるので、その場合は実験者からギブアップができることを伝えて、あくまでも被験者の意思で継続・終了を判断させる。

4.3 タスク誤解の最小化＜パタンランゲージ:「**タスク誤解最小化**」＞
先に引用したように、音声インタフェースではユーザの要求を直接入力させる形に近いため、ユーザの想定するシステムの機能自体に誤解が生じる場合がある[3]。

そうした原因などにより、被験者実験において設定と異なるタスクを遂行する被験者がいる。これはタスク達成率を算出する際にタスク達成を判断するうえで問題となる。誤解をしていて指摘されれば誤りに気づく場合、指摘されても理解できない場合がある。そこで、まず、タスク内容あるいはそれに類似する行動に対する経験で被験者を事前に絞り込む。実験前には充分にタスクの設定を理解させる。
それでも誤解が生じた場合には実験後にゴールを理解できていたかどうかを確認する。その結果、実験の準備段階で誤解を極力減らせる。実験後の確認により残った誤解を明確にできる。

4.4 実験指示文設計＜パタンランゲージ：「実験指示文抽象化」＞

音声インタフェース評価実験において、発声内容については被験者に委ね、どのような発話表現が被験者によって発せられたかを分析の対象にすることがあるが、実験の指示文が発話表現を誘導してしまうことがある。実験の指示においては、期待しない誘導を避けるように注意しなければならない。

4.5 音声による実験指示

被験者実験における指示を統一するために音声ガイダンスを用いることがあるが、音声インタフェース実験においては音声メディアを主な情報伝達手段としているため、インタフェース操作に必要な音声ガイダンスと実験遂行に必要な音声ガイダンスが被験者によって混同される恐れがある。実験遂行に必要なガイダンスの設計において注意が必要である。
その際、被験者の注意を促すために合成音声よりも録音音声の方が効果的であることがある。合成音声によるメッセージを繰り返し発していると注意が著しく減ることがある。

4.6 モダリティ利用バランスの制限＜パタンランゲージ：「モダリティバランス制限」＞

音声以外のモダリティが利用できる場合に、被験者によっては日常的に慣れているモダリティを多用することがある。音声モダリティの導入による効果を知りたい場合には、何らかの方法でモダリティ利用バランスを制限することも必要である。

4.7 途中の認識誤り、操作誤りはタスク達成の判断の材料としない

タスク達成の判断において、途中の音声誤認識や操作誤りは考慮するべきでない。音声誤認識の中には、稀に、被験者が言い誤ったが望ましい単語に誤認識したというケースもあるが、これも同様にタスク達成かどうかとは無関係とする。(誤認識しても最終的にタスクが達成されればタスク達成ととらえる)

5. 事前学習

新しいユーザインタフェースを導入する際に、ユーザが事前のトレーニングを受ける機会を持てることがある。音声インタフェースの評価に際しても、状況に応じて事前学習を前提にすべきである。音声インタフェースの事前学習において、大きく以下の二つのフェーズがあり得る。

5.1 発声方法の学習

音声は日常のコミュニケーションに用いられるメディアだが、音声インタフェースの利用に際しては、必ずしも日常の発声方法が許容されないことがある。ユーザに、個々の音声インタフェースに応じたある種の発声方法を習得させることが音声インタフェースのユーザビリティ向上につながる。習得させるべき発声方法としては以下の観点が考えられる。

- 個々の音韻の構音
- 連続する音韻の調音
- 発話速度
- 語彙
- 文法

事前に学習する方法として、適切な発声方法を提示するだけの事前教示と、ユーザの練習に対して何らかのフィードバックを与える事前訓練がある。いずれも様々な方法が考えられる。

筆者らによる実験からは、音響的明瞭性を向上するには教示だけでなく訓練が必要であることが示されている[6]。なお、事前訓練によって音声認識率にして平均 10 ポイント程度の向上が認められることもある[6]。

音声インタフェースの評価を発声方法の習得の問題とは切り離したい場合、発声方法の事前学習を行うのが良い＜パターンランゲージ：「**発声方法の事前学習**」＞。

5.2 インタフェース操作方法の学習

音声インタフェースの導入に際して、発声方法の習得と切り離して、音声インタフェース操作方法についても習得が必要である＜パターンランゲージ：「**インタフェース操作方法の事前学習**」＞。例えば、現状の音声インタフェースではユーザが発話を行う際に指定されたボタン（「トークスイッチ」などと呼ばれる）の押下を求められるものが多いが、ボタンをいつ押せばよいか、発話中にボタンを押す必要があるか、などは必ずしも理解が定まっていない。

なお、インタフェース操作方法はそれ自体が音声インタフェース評価の重要な対象であり、事前学習でなく自然に習得することが望まれることもある。

6. 習熟

ユーザビリティ評価において、長期間のモニタリングは重要である。先行研究において、被験者が音声インタフェースの利用回数を重ねるごとに音声認識性能が向上し、それに伴ってタスク成功率と満足度が向上することが報告されている[7]。筆者らの実験によっても(1)習熟を考慮する場合には利用回数と利用間隔の両方を考慮する必要があること、(2)トークスイッチ操作の認知的負荷にも習熟の影響があり得ることがわかっている[8]。以上の知見を踏まえて、音声インタフェース評価において習熟に関して考慮すべき点を以下に整理する。

6.1 事前学習との関連

5 に述べた事前学習を行わない場合、実験中に学習が大きく進むことがある。その結果、実験データには初期段階とそれ以降で被験者行動や評価が大きく変化することに注意されたい。特に音声認識精度については習熟の初期段階で大きな変化が生じることがあり、それがユーザビリティ評価に大きく影響を与えることもある。

6.2 実験回数および間隔

音声インタフェース評価において習熟を考慮する場合には、音声インタフェースを利用する回数と間隔の両方を考慮する必要がある。一回の実験で利用する回数が多くなればそれだけ習熟が早い一方、利用する間隔が長くなれば習熟は遅くなる。

6.3 トークスイッチ操作の習熟

トークスイッチを押下してから発話が許可されるタイプの音声インタフェースは多い。習熟によってトークスイッチ押下後発話までの時間が短縮する傾向が、筆者らの実験により確認されており[8]、トークスイッチ操作の認知負荷が習熟にともなって軽減するものと推測できる。一方で、経験的にはトークスイッチ操作の習熟が一向に進まない被験者層もいる。トークスイッチ操作を音声インタフェース評価の問題から切り離したい場合、操作の習得に充分に時間をかける必要がある。

7. 詳細評価方法

ユーザビリティ評価の重要な 3 つの要素である「有効さ」「効率」「満足度」では捉えきれない観点について、音声インタフェース評価に有効とされる方法を以下に紹介する。

7.1 NASA-TLX

(1) 手法

NASA-TLX(National Aeronautics and Space Administration Task Load Index)はメンタルワークロードの測定を目的として開発された手法であり、航空機操縦に始まり、自動車運転時のユーザインタフェース操作などの状況にも適用されている.測定項目は6つの尺度から成り立っており,さらに6つの尺度に対して重みを与える.

評価手法は厳密に定められていて、信頼性を高めるにはこれに従う必要がある。具体的な評価方法は例えば以下のように定められている。

- 被験者は作業終了直後に,説明文を参考に作業を振り返り,6項目の評価を行う.
- 評価は両端が低い/高い,または良い/悪いとされた12cmの線分に印を付ける.
- 線分上に記された位置を0〜100の数値として読み取り,素点とする.

メンタルワークロード値を求めるには,いくつかの方法が考案されているが、その中でも, CSTLX(card-sort TLX)は簡易的な手法として有効である[9].本来のNASA-TLXのメンタルワークロード値の算出方法であるWWL(Weighted Workload)が一対比較を前提としているため,被験者に多大な負担をかける恐れがあること,また一対比較をするだけの時間がかかるなど短時間の実験の中で複数回実施するには不向きであるためである.この煩雑な手続きを簡便化する方法が三宅ら[10]により提案されている. WWLと相関が高いCSTLX[9]を用いる事で,被験者の負担の軽減,及び時間の節約を図ることができる.

CSTXLではNASA-TLXの6つの評価項目を被験者ごとに重要度順に並べてもらい,その順に6〜1の重みをつける.各重みを係数とし,それぞれの素点と掛け合わせ,合計を重み付け係数の総和である21で割る事で求められる.得られたメンタルワークロードの点数が高いほど,被験者の疲労度は高いということになる.

野田ら[11]は、被験者が各項目の評価尺度を充分に理解できるように日本語版 NASA-TLX を改善し、音声対話システムにあわせて説明を具体的かつ簡潔にする工夫を行った。図 7.1 にその例を示す。

音声インタフェース評価シート

1(1 / 27)
それぞれの項目について、10段階の評価で選んでください。

精神的要求:考えたり覚えたりするのは大変でしたか？
低い (単純だった。覚えきれた) ○1 ●2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 高い (複雑だった。覚え切れなかった)

身体的要求:身体は疲れましたか？
低い (動作がらくだった。休み休みできた) ●1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 高い (動作がきつかった。動きっぱなしだった)

時間的圧迫感:あせりましたか？
低い (余裕があった。ゆっくりできた) ○1 ○2 ●3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 高い (余裕がなかった。速かった)

作業達成度:上手くやり遂げられましたか？
良い (満足している) ●1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 悪い (満足していない。十分は達成できず)

努力:一生懸命になりましたか？
低い (努力するほどではなかった) ○1 ●2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 高い (かなり努力した)

フラストレーション:イライラしましたか？
低い (リラックスしていた) ●1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 高い (ストレスを感じた)

図 7.1 NASA-TLX を改良した音声インタフェース評価シート例

(2) 結果

得られた数値に基づいて心的負荷値を算出することができる。また、総合値だけでなく 6 つの尺度ごとの値を用いて詳細に見ることもできる。

(3) 注意事項

・説明文は定められているものを用いる。具体的には[9]を参照されたい。

・評価をする作業の単位は認知過程を振り返られる程度にする。

7.2 二重課題法

(1) 手法

インタフェース操作時の作業負荷の量を間接的に計測する方法として、二重課題法がよく用いられる。二重課題法は 1960 年代から使われている方法であり、音声インタフェースについては近年特によく使われている。西本は音声インタフェースを対象にした二重課題法の適用について「音声インタフェースの利用を第一課題とし、ゲームなどの客観的に負荷が測定可能な第二課題を被験者に並行して行わせ、第二課題の成績が低い状況において音声インタフェースの負荷が大きい、と判断する」としている[7]。音声インタフェースはハンズフリーやアイズフリーを利点とするため、インタフェース利用時に並行して別の作業を行うことを前提とすることが多い。その点で、NASA-TLX とあわせて作業負荷の量を評価することは有意義である。二重課題法でワークロードを測定する場合には、音声インタフェース操作を第一課題とし、第二課題に何を選択するかが非常に重要になる[12]。以下には[7]にならって筆者らが実施した手法について説明する。

副課題には音声 I/F を操作するときの心的負荷を反映しやすく、被験者のなれの効果などの影響を受けにくいなどの理由により早押しゲームの課題を使用する。図 7.2 にゲーム画面を示す。

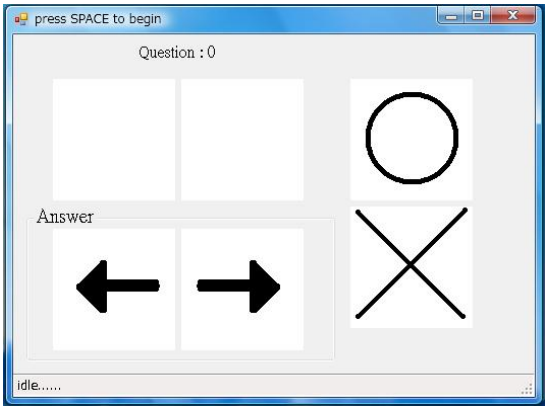

図 7.2　副課題のスタート画面

画面の左側には上に２つの空欄と下に２つの矢印、右側には上下に○と×が並んでいる。ゲームが始まると、左上の２つの空欄の一方に左右いずれかの矢印（←、→）がランダムに現れる。現れた矢印を下の空欄の矢印と照合して同じ向きの場合（図 7.3 を参照）は○、違う向きの場合（図 7.5 を参照）は×とし、それぞれキーボードにある上矢印（↑）キーと下矢印（↓）キーを押すことになる（図 7.4 と図 7.5 を参照）。キーを押すと、正解不正解に関わらず１秒後に次の矢印が現れる。ゲームの開始と終了のタイミングは主課題に合わせる。

副課題の操作が主課題への影響をできるだけ避けるため、被験者には、主課題を優先し、副課題は正確さを重視しつつできるだけ早く応答するように指示する。

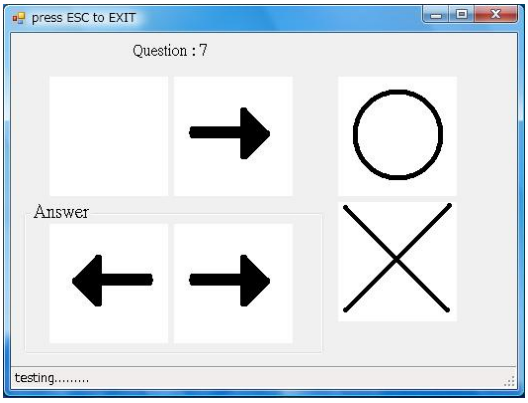

図 7.3　問題例１

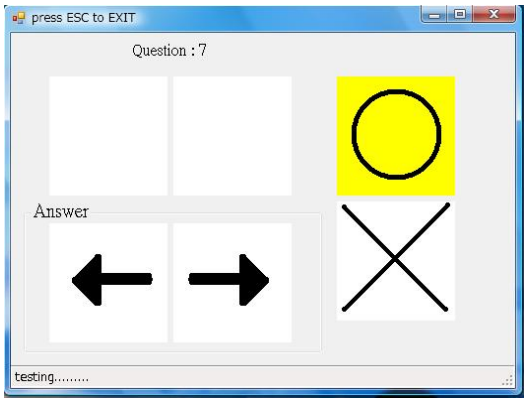

図 7.4　回答例１

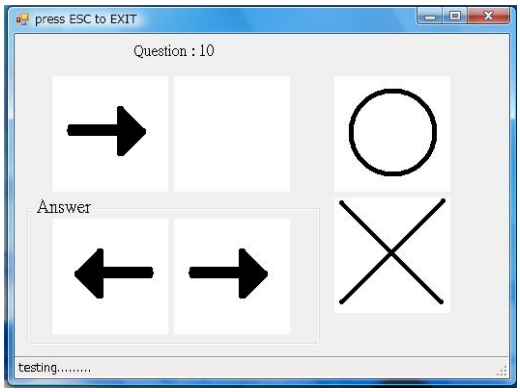

図 7.5　問題例２

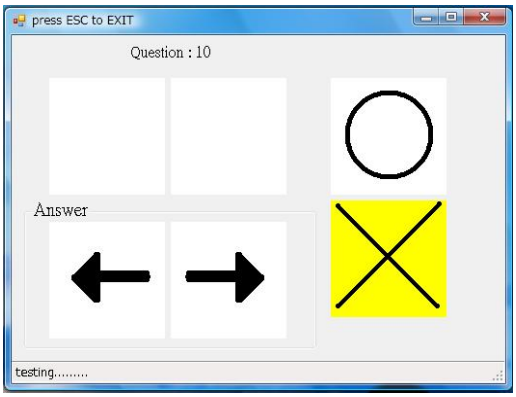

図 7.6　回答例２

心的負荷を評価する際に、上の欄に矢印が出てから回答するまでの時間をミリ秒単位で測り、違う刺激を盛り込まれた主課題の試行の間で比較する。応答時間が短くなるほど、同時に行われた主課題の負荷が小さいと考えられる。

(2) 結果

主課題の種類ごとに副課題の成績や応答時間から主課題の心的負荷を求めることができる。

(3) 注意事項

・主課題と副課題に注ぐ認知的な資源のバランスが途中で大きく変わらないよう、できるだけ主課題を達成するように心がけさせる。

7.3 生体信号計測

ユーザインタフェース操作時の心理的な負荷、いわゆるストレスを物理的に計測しようという生体信号計測の試みが近年盛んに行われている。筆者らはこれまでに音声インタフェース評価実験に生体信号計測を利用できる可能性を検討し、限定的ではあるが有意義な情報が得られる感触を得てきた。以下には生体信号計測方法の一例を示す。

計測には，PC にオンライン接続した計測生体情報計測装置 ProComp（プロコンポ）を使用する （図 7.7 参照）．また，収集したデータの表示・解析には専用ソフト BioGraph INFINITI（バイオグラフインフィニティ）を使用する．

図 4.3　生体情報計測装置 ProComp

▣ 実験で使用する指標

◎皮膚コンダクタンス（Skin Condudtance：SC）

発汗には，精神性発汗と，温熱性発汗の２種類がある．これらのうち精神性発汗を電気的に捉えたのが皮膚電気活動（Electrodermal Activity ： EDA）である．

手指に装着した一対の電極間に微弱な電流を流し，皮膚の見かけ上の抵抗変化を調べる通電法（exosomatic method）を用いる．通電法で測定される反応には，皮膚抵抗（Skin Resistance ：SR）と，皮膚コンダクタンス（Skin Conductance ： SC）があり，SR の逆数が SC である．

心理的に緊張・興奮すると交感神経の働きで発汗し，心が落ち着いてくると発汗は止まる．これらは皮膚の電気抵抗の変化となって現れる．ストレスが溜まった状態，もしくは解消された状態になると，SC はそれらに比例するように増加もしくは減少する．

◎皮膚温（Temperature：TEMP）

皮膚温は，常に一定範囲に維持されている核心部分とは異なり，環境温や各種要因に

よって常に変動をしている．交感神経系（SNS）の賦活によって血液収縮，つまり末梢の動脈の直径が減少することによって皮膚温は低下し，SNS の弛緩によって血液拡張つまり末梢の動脈の直径が増加することによって皮膚温は高まる．

ProComp による皮膚温測定の原理は，サーミスタ（thermistor）と呼ばれる温度センサを皮膚表面に直接装着して皮膚温を測定し,それを電流の変化に変換するものである．

従来から皮膚温と情動の関係が報告されてきた．「不安」「困惑」「怒り」などの情動によって手指の皮膚温が低下し，安堵・弛緩で回復したことが示されている．

これら２つの指標は，いずれもセンサを指先へ装着するだけで波形が得られるため生体信号測定を行うことによる被験者への負担がより少なく済む．センサによっては、胸部や頚部などに装着するものもある。

生体信号計測を行う場合には以下のように様々な点に注意する必要がある。

・被験者募集の際にセンサを体に装着しての計測を行う旨を伝える
・被験者が女性のときには特にセンサ装着を行う者の性別、環境などについて配慮する
・装着したセンサが実験中にはずれないように、必要以上に動かないよう指示する一方で、自然なインタフェース操作が阻まれないように指示方法を配慮する
・センサに異常がないか、常にセンサからの信号をモニタできるようにする
・音声インタフェースの場合、特に口の周囲に装着するセンサにノイズが加わる可能性や、センサが発声行動に及ぼす影響を注意する必要がある。

ユーザインタフェース評価への生理学的アプローチの詳細については、たとえば[13][14][15]などを参考にされたい。

7.4 ストレス発話分析

一般的なユーザインタフェースの評価においては、プロトコル分析法（思考発話法）が実践的な手法としてよく利用される。しかしながら、音声インタフェースに対しては、操作のために発話を行う必要があり、プロトコル分析法を適用することが困難である。一方で、音声は話者の感情や情動といった内的な状態を表現するため、音声インタフェース評価において、被験者が発した入力音声から内的な状態を推定して評価の手がかりにできる可能性がある。以下には、筆者らが試みたストレス発話推定について紹介する。

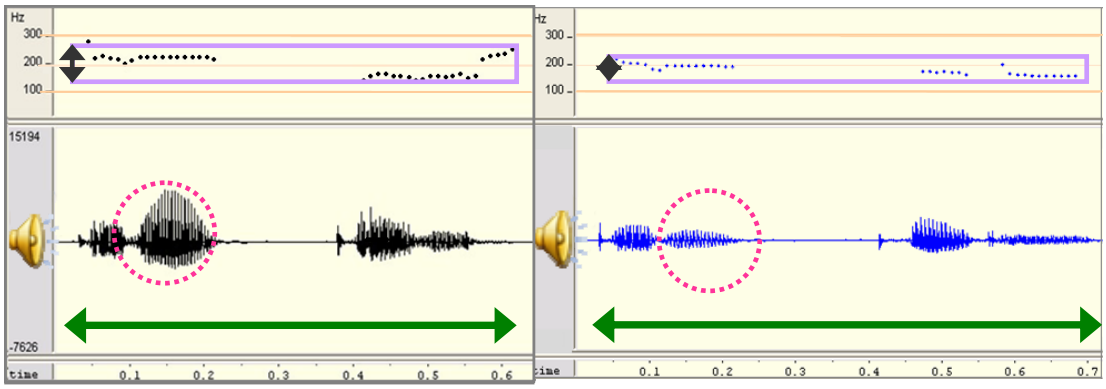

図 5　平常時（左）と不安時（右）の発話の比較（上部が F0、下部が音声波形）（同一話者がいずれも「ブロッコリー」と発声）

図 5 は、平常時と、意図的に不安な状況を作り出した時の発話の音響的特徴を比較したものである。ストレス発話の中でも、不安を感じているという印象を受ける発話（不安発話とする）は、音圧が小さい、基本周波数のレンジが狭い、発話速度が遅い、疑問調になる、などの特徴が比較的よく観察できる。そこで不安発話の判別モデルを構築し、音声インタフェース評価実験の被験者音声を対象に印象評定を行い、半数の評定者が高い「不安」の印象を受けた発話を対象に評価したところ、90%の判別精度が得られた[10]。さらに判別精度を向上する必要があるが、このようにしてストレス発話を推定することができれば、プロトコル分析法のように思考過程を言語化してもらうかわりに、「わかりづらい」「いらいらする」などの内的状態の変化を検出する手がかりが得られる。

8. 音声インタフェース評価におけるポイント

〈設計指針と関係付けながら、パタンランゲージにより記述〉

代表ユーザ被験者選定

文脈

- **評価実験の被験者を一定の基準で選定する**

問題

- **現状の音声インタフェースでは操作に際して様々な制約を設けざるを得ず、ユーザ層を制限せざるを得ない。評価実験においても被験者層を一定の基準で制限する必要がある。**

解決策

- **評価対象のユーザ層を明確にし、それを代表するように被験者層を選定する。**
- **語彙を制約したり、発話タイミングを制約したり、操作方法に制約を設ける場合には、特に音声インタフェースだからといって、幅広いユーザ層を対象にするのは得策ではない。**

結果

- **現状の音声インタフェースの性能に応じた被験者層による評価実験が実施できる。**

関連するパターン

- **パターン**

タスク経験に基づく被験者選定

文脈

- **評価実験の被験者を一定の基準で選定する**

問題

- **実験で行うタスクに対して、経験がない、関心がない、理解できない、といった被験者が存在する。**

解決策

- タスク経験を考慮して、タスクに対して現実的に対象となる者を被験者として選定**する。**
- そのうえで、実験中にタスクを誤解した、タスクに関心を持たなかったなどは実験による不備と考えるべきである。
- 実験で行うタスクそのものの経験がなくても類似したタスクの経験があれば良い。

結果

- **タスクに関心を持たない被験者をあらかじめ除外できる。**

関連するパターン

- **パターン**

時間無制限タスク

文脈

- タスク達成率の観点からシステムの客観評価を行っている．設計者の主観は可能な限り除きたい。

問題

- 各々のタスクの制限時間は必ずしも自明ではない。

解決策

- 複数のタスクをまとめて 1 セッションを定義する．
- 各々のタスクの制限時間を定めず，実験結果の分析時に制限時間を変数とした T-A グラフを用いて評価する．
- タスクが困難な場合は被験者の意思でギブアップを行ってもらう．
- 「効率良く操作」をするよう、努力目標を設定する。
- 被験者によっては，すぐにギブアップしてしまうことがあるが，その際にはギブアップが可能になるまでの時間を設ける．

結果

- 被験者によっては，ひとつのタスクに時間をかけすぎて，実験が時間通りに終了しないことがある．
- 実験者による介入によって，ギブアップができることを伝えることで，被験者の意思でのタスク達成率が得られる．

関連するパターン

- **パターン**

タスク誤解最小化

文脈

- 被験者実験において、タスクを設定してそれを被験者に遂行させる。

問題

- 設定と異なるタスクを遂行する被験者がいる。タスク達成率を算出する際にタスク達成を判断するうえで問題となる。誤解をしていて指摘されれば誤りに気づく場合、指摘されても理解できない場合がある。

解決策

- まず、タスク内容あるいはそれに類似する行動に対する経験で被験者を事前に絞り込む。
- 実験前には充分にタスクの設定を理解させる。

- それでも誤解が生じた場合には実験後にゴールを理解できていたかどうかを確認する。

結果

- 実験の準備段階で誤解を極力減らせる。実験後の確認により残った誤解を明確にできる。

関連するパターン

- **パターン**

ギブアップ制限

文脈

- 評価実験に際してタスク完了が困難な状況に陥ることがある。

問題

- あらかじめ被験者の意思でギブアップをすることを許可すると、被験者によってはすぐにギブアップしてしまうことがある。

解決策

- ギブアップは一定時間経過後のみ許容することにする。
- ギブアップボタンを設け、一定時間を経過しないとボタンが押下できないようにするなどの工夫をする。

結果

- 被験者はタスク完了が困難な状況に陥っても、完了を目指して努力する。。

関連するパターン

- **パターン**

実験説明文抽象化

文脈

- 実験についての説明を文章で表現したり、口頭で伝えたりする。
- 被験者が発する音声発話表現を広く収集したい。

問題

- 説明文が、被験者の操作における音声発話表現を誘導してしまうことがある。

解決策

- 説明文を抽象的に表現する。

結果

- 説明文によって音声発話表現が誘導されず、被験者が想起した表現が収集できる。

関連するパターン

- **パターン**

モダリティバランス制限

文脈

- 音声以外のモダリティが利用できる。

問題

- 被験者によっては日常的に慣れているモダリティを多用することがある。
- 音声モダリティの導入による効果を知りたい。

解決策

- 音声モダリティは必ず使用するように制限する。

結果

- 被験者は音声モダリティを使用するように努力する。

関連するパターン

- **パターン**

発声方法の事前学習

文脈

- 音声インタフェース評価に際して、音声認識性能とは異なる視点の評価を行いたい。

問題

- 被験者の音声認識技術に対する知識や習熟の程度は様々であるが、それが音声インタフェース評価を決定付けてしまうことが多い。
- 音声認識技術に対する知識や習熟の程度は、短時間にも変化し得るため、実験中に評価が大きく変化することがある。

解決策

- 評価の際、事前学習によりある程度被験者の音声認識技術に対する知識や習熟の程度を統制する。
- 音声認識技術を実際に使用させ、認識結果をフィードバックしたり、発声方法のアドバイスを行ったりすることによって、良く習熟させることができる。

結果

- 被験者の音声が認識されやすくなり、音声認識性能とは独立した音声インタフェース設計の評価が可能になる。

関連するパターン

- **パターン**

インタフェース操作方法の事前学習

文脈

- 被験者実験によってインタフェース設計の適切さを評価する。

問題

- 被験者が初めてインタフェースを利用する場合に、実験中に習熟することにより大きく実験中に評価が変化することがある。
- 安定した評価結果を得るために必要な実験時間が長くなる。
- 最初に実施するタスクの難易度によって習熟のプロセスが異なる。

解決策

- 事前にインタフェース利用の訓練を行い、ある程度習熟させたうえで実験を始める。その際、できるだけ被験者間で訓練の種類を共通にする。

結果

- 被験者個人内（タスク間）の評価の習熟に伴うばらつきが減る。

関連するパターン

- **パターン**

謝辞

本書は経済産業省情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発「音声認識基盤技術の開発」における「音声インタフェース評価技術の開発」課題の一環で作成されました。メンバーの皆様に感謝致します。

参考文献

[1] 中川聖一他, ″岩波講座 言語の科学 2 音声、″ 岩波書店, pp.179, 1998.
[2] 石川泰他, “音声インタフェースの評価,” 日本音響学会誌, vol.61, no.2, pp.79-84, 2005.
[3] 石川泰,“音声認識の実用化の阻害要因と課題:音声インタフェースのユーザビリティ評価(音声認識の実用化の阻害要因と課題” 情報処理学会研究報告, 2006-SLP-63, pp.45-46, 2006.
[4] 山岡俊樹, “ユーザー優先のデザイン・設計”, 共立出版, 2000.
[5] Nielsen, J., “Usability Engineering” (ヤコブ・ニールセン, 「ユーザビリティエンジニアリング原論」), 東京電機大学出版局, pp.138, 1999.
[6] 網田泰裕他, “音声認識における事前教示・訓練の影響″, 日本音響学会秋季研究発表会, 2-7-9, 2008.
[7] 原直他, ″音声対話インタフェースの長期利用における学習効果の評価,″ 情報処理学会研究報告, 2005-SLP-55, pp.17-22, 2005.
[8] 菊池英明他, ″音声インタフェース評価における慣れの影響の分析″, 情報処理学会研究報告, 2007-SLP-67, pp.97-102, 2007.
[9] 芳賀繁,水上直樹,”日本語版 NASA-TLX によるメンタルワークロード測定-各種室内実験課題の困難度に対するワークロード得点の感度-”,人間工学, Vol.32, No.2, pp.71-80, 1996.
[10] 三宅晋司,神代雅晴,”メンタルワークロードの主観的評価方法-NASA-TLX と SWAT の紹介および簡便法の提案-”, 人間工学, Vol.29, No.6, pp.399-408, 1993.
[11] 野田幸志, 西田昌文, 堀内靖雄, 市川熹, ″心的負荷における車載情報機器のための音声対話戦略分析,″ 情報処理学会研究報告, 2006-SLP-64, pp.149-154, 2006.
[12] 芳賀繁, “メンタルワークロードの理論と測定”, 日本出版サービス, 2001.
[13] 吉川榮和他, “ヒューマンインタフェースの進路と生理 第２回情報行動計測アプローチ”, ヒューマンインタフェース学会誌, vol.1, no.3, pp.3-12,

1999.
[14] 吉川榮和他，“ヒューマンインタフェースの心理と生理”，コロナ社，2006.
[15] 飯田健夫他，″特集記事 インタフェースと生理計測″，ヒューマンインタフェース学会誌，vol.6，no.1，2004.