社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49　(43)
2003／12／19

ユーザ評価と達成度との相関に基づく音声対話システムの品質評価の予備的検討

白勢 彩子　原 直　藤村 浩司　伊藤 克亘　武田 一哉　板倉 文忠

名古屋大学大学院情報科学研究科／CIAIR　〒464-8603 愛知県名古屋市千種区不老町
E-mail: shirose@cog.human.nagoya-u.ac.jp

あらまし　本研究は，音声対話システムの利用に必要な知識，技術の学習過程および問題点を明らかとし，これらがシステム評価に与える影響を検討するため，実システムを用いた観察実験を行ない，それによって得られた結果に基づいて，ユーザの達成度とアンケートによるシステム評価との関連性に関する基礎的な議論を行なった．分析の結果，発話認識率と発話数とに相関がある評価項目はほとんどなく，むしろ，会話満足度，システム理解度と多く関連することが明らかとなった．従来，認識率とシステム評価とがよく一致することが知られているが，他の観点からの評価も考慮する必要があることが示唆された．今後は，被験者数を増大させて条件を統制した実験を行ない，より詳細な議論をしていきたい．

キーワード　音声対話システム，品質評価，ユーザ達成度，利用技術，相関性

Preliminary study on the evaluation of a quality of spoken dialogue system in terms of user factors

Ayako SHIROSE　Sunao HARA　Hiroshi FUJIMURA　Katsunobu ITO

Kazuya TAKEDA　and　Fumitada ITAKURA

Department of Information Science/ CIAIR, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan
E-mail: shirose@cog.human.nagoya-u.ac.jp

Abstract　This study aims to describe user problems and process of learning skill in using spoken dialogue systems and to reveal how these impact on the evaluation of the system usefulness. For this aim, we designed a new dialogue system and carried out a field test for a large number of subjects and asked them to evaluate the usefulness of the system. The results showed that the evaluation of the system did not correlate a recognition rate but user satisfaction and comprehension. This suggested that the spoken dialogue systems should be evaluated in terms of user factors. Controlled experiments are needed to discuss in detail.

Keyword　Spoken Dialogue System, Evaluation of a Quality of System, Task Completion, User Skill, Correlation

1. はじめに

本研究は音声対話システムの利用に必要な知識，技術の学習過程および問題点を明らかとし，ユーザの知識，技術がシステム評価とどのように関わるかを検討してユーザの利用技術に関するガイドライン、ユーザの利用技術も考慮したシステム評価方法の提案することを目標としている．

従来，対話システムは，システムの認識精度といったシステムそのものの性能，あるいは，ユーザから見たシステムの使い勝手，いわゆる「ユーザビリティ」といった点から評価がなされてきている．これに対し，ユーザが対話システムを利用するにあたって持つべき知識や技術，教示方法の評価といった視点を踏まえてのシステム評価はなされてこなかった感がある．しかしながら，高い利用技術を持つユーザはシステム利用に困難さを感じにくいために高くシステムを評価をするなど，ユーザの持つ知識や技術の多寡がシステムの評価と高い関連があると推測される．そこで，本研究では，大規模な被験者を対象に，音声対話システムを用いたフィールドテストを行ない，システム利用のユーザファクタを明らかにしようとしている．

本稿では，観察実験によって得られた結果に基づいて，ユーザの達成度とアンケートによるシステム評価との相関関係を明らかとして議論を行なう．

2. 音声対話インタフェース概要

2.1. システム概要

実験に用いる音声対話インタフェースとして，楽曲検索システムを構築した．これは，ユーザが対話によって聞きたい楽曲を検索し試聴できるというシステムである．このようなインタフェースによる課題は，日常的な行動との乖離が少なく，被験者の主体的な参加が見込まれるとして採用した．さらには，車内での音声対話システムとしての実用化も考えられ，実験課題に適切であると考えた．

システム処理の流れを図1に示す．まず，ユーザの音声によるシステムへの検索要求を認識させる．例えば，「中島みゆきの曲が聞きたい」等である．音声認識の結果によって，音楽配信ポータルサイト上の情報の検索が開始される．音楽配信ポータルサイトには「レーベルゲート」（http://www.labelgate.com）を用いた．

検索結果は，プレイリストとして取得され，合成音声器によりユーザへ呈示される．音声合成器には富士通 FineSpeech を用いた．ユーザが楽曲を選択するには，再度，音声によりシステムへの検索要求を認識させることによって，認識結果と一致した項目が選択，再生される．

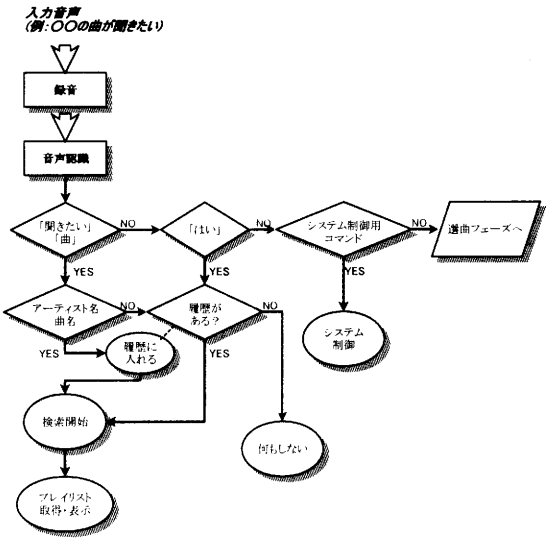

図1. 処理ブロック図

2.2. 認識処理

音声認識エンジンには大語彙音声認識エンジン Julius の Windows DLL 版である，Juliuslib 3.1p2-sp4 を用いた．音響モデルは，CSRC の標準日本語音響モデル[1]より，状態数 3000，性別非依存，64 混合，triphone モデルを用いた．認識に用いる辞書の語彙サイズは 7710 単語（1601 アーティスト，6071 曲）である．アーティスト名・曲名の辞書データはオリコン（http://www.oricon.co.jp/）の週間ランキング（2002 年 10 月第 1 週から 2003 年 9 月第 2 週までの計 86 週）及びレーベルゲートの登録曲（2003 年 9 月 24 日時点で 1404 アーティスト，5862 曲）を用いた．上記 2 つのデータを重複なしに結合して，辞書を作成した．

2.3. 検索手続

システムに対し，ユーザが「～の曲」「～の曲が聞きたい」と検索要求を認識させると，該当のアーティスト名ないし曲名が履歴に保存され，WWW 上で検索される．検索結果はすべてプレイリストに表示される．例えば，「中島みゆきの曲が聞きたい」との要求では，“中島みゆき”が履歴に保存され，検索される．検索の結果，「レーベルゲート」のサイト上で合致した中島みゆきの曲の情報がプレイリストに取得され，表示（本実装ではさらに音声出力）される．

次いでユーザはプレイリストから曲を選び，再度，検索要求を認識させることによって，「絞込検索」としての目的曲選択を行なうことができる．例えば，上記手続に引き続いて「地上の星」との要求がなされるとプレイリスト上の一致項目が選択され，音楽再生アプリケーションが起動し再生される．

検索に用いることのできるキーワードを示す．

- アーティスト名
- 曲名
- アーティスト名＋曲名
- 「最新の曲」
- 「人気の曲（ヒット曲）」

3. 実験概要

本研究では，観察実験と統制実験の 2 種を行なう．観察実験では，主に被験者の行動を観察し，ユーザの持つ問題点を発見，整理することを目的としている．統制実験では，ユーザの持つシステムに関する知識，技術を統制し，グループ間での印象評定の差異を検討することにより，ユーザの持つ知識，技術とシステム評価との相関性を具体化することを目的としている．以下，これら 2 種の実験について詳細を述べる．

3.1. 被験者

150 名を対象とする．年齢は 20 歳～50 歳とし，性別比は 5:5 とした．これらの被験者は 50 名ずつの 3 グループに分けられ，うち 50 名が観察実験，残り 100 名が統制実験に参加した．

3.2. 実験条件

より詳細な観察，議論を行なうため，観察実験，統制実験ともにいくつかの実験条件を設定した．

観察実験では，被験者のさまざまな行動を引き出せるよう，複数種の実験設備を用意した．特に，マイクロフォンによる行動の相違を観察するため，表 2 に示す 4 種類を用意した．マイクロフォン 1 と 2 は同種で

表 2. 実験機器類一覧

PC 1	DELL Inspiron 5150
PC 2	Sony VAIO PCF-V505R/PB
USB Audio 1	M-Audio Mobile Pre USB
USB Audio 2	EDIROL UA-5
ミキサー	Sony SRP-X1008
ﾏｲｸﾛﾌｫﾝ 1, 2	Sony ECM-77B
ﾏｲｸﾛﾌｫﾝ 3	SENNHEISER HMD 410
ﾏｲｸﾛﾌｫﾝ 4	Sony F-740
スピーカ	YAMAHA MS101 II
カメラ 1	Sony DCR-TRV 900
カメラ 2	Panasonic NV-GS100

あるが，1 は被験者が装着し，2 は実験設備の一部であるドライビングシュミレータ機に設置した．加えて，システム解説資料であるマニュアルを参照しながら検索する条件と参照せずに検索する条件とを設けた．

統制実験では，システムを使用するにあたって必要とされる知識，技術とシステム評価との相関を明確にするため，これらを統制するような条件を設けた．具体的には，システム使用に関する知識を提供するマニュアルについて，詳細な解説のものと簡略な解説のものとを用意した．さらに，それぞれ，映像資料，文字資料の 2 種類が作成された．すなわち，統制実験では計 4 種類のマニュアルを用意した．

3.3. 手続

実験は，名古屋大学 IB 情報館 4 階の実験室内にて 1 名ずつ行なわれた．被験者は着席してシステムとの対話を行なった．実験機器類および接続についてそれぞれ表 2，図 2 にまとめた．

システムの使用に先立ち，被験者はフェースシート項目に筆記回答した．このフェースシート項目は，被験者個人の情報（性別，年齢等）を尋ねるものと，日常の生活における音楽への馴染みの程度，電子機器類（音声認識システム含む）の使用技術の程度を尋ねるものとがある．後者は例えば「インターネットを日頃使っているか」等である．

引き続き，被験者はシステム仕様に関する知識を提供された．観察実験では，被験者は概要を箇条書き程度にまとめた簡略なマニュアルを黙読した．統制実験では，3.3 に述べた実験条件に応じて作成されたマニュアルを黙読ないし視聴した．

実験では，2 セッションを設けた．指定曲リストの 5 曲を検索するセッションと自由に曲を検索するセッションである．観察実験では，自由に曲を検索するセッションでマニュアルを用いずに検索するよう指示した．概ね，1 セッションの所要時間は 15 分程度であった．原則的に，システムの動作画面は被験者に呈示しなかった．

実験中の被験者の音声，行動は PC ハードディスクおよびデジタルビデオカメラにて記録された（表 3 参照）．ビデオカメラでは，着席した被験者の正面および俯瞰位置から撮影した．観察実験では，被験者の行動は実験者が記録シートへ記入することによっても記録された．この記録では，実験設備システムのふるまい（認識，誤認識等）とそれに対する被験者の行動およびこれらの発生時刻が主な対象である．

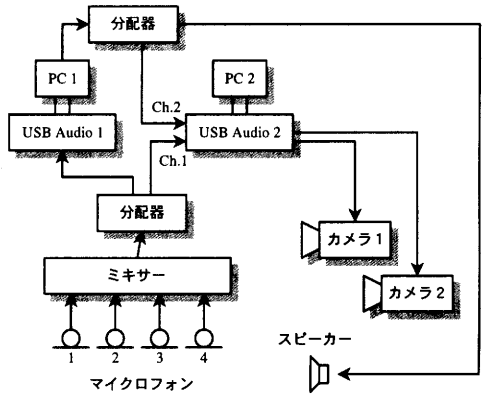

図 2. 実験機器の接続

4. 評価アンケート

被験者は実験後にアンケートに回答するよう求められた．アンケートは 32 項目を含む 15 問として作成されている．うち，選択肢回答項目が 27，自由回答項目が 5 である．アンケートの作成にあたっては，関連する先行研究[2],[3]を参照した．この研究は音声対話システムのサービス品質の評価基準を示そうとするものであり，具体的な評価項目とそれらの分類がなされている．評価項目の分類基準として「全体的な印象」「システムとのコミュニケーション」「システムのふるまい」「システムとの会話」「得られた情報の正確さ」が示され，具体的な項目が挙げられている．本研究でもこの分類に従い，前掲研究の評価項目を参考に評価アンケートを作成した．以下に詳細を述べる．

a. 全体的な印象（11 項目）

「丁寧さ」「声の印象」「総合的な印象」について，7 段階（大変良い～ふつう～大変悪い）で評定を求めた．「システムの良かった点」「改善点」など 6 項目について自由記入欄を設け評価を求めた．

さらに，「会話の満足度」について 100 点満点のうちの点数を尋ねる項目を含めた．

b. システムとのコミュニケーション（2 項目）

「声の聴き取りやすさ」を上記と同一の 7 段階で評定を求め，「認識の適切さ」を尋ねた．

c. システムのふるまい（5 項目）

「親しみやすさ」について上記と同一の 7 段階で評

定を求めた．「動作の適切さ」について尋ね，適切でないと回答した場合には具体的な点を自由記入で回答を求めた．

さらに，「システムの回答のタイミング」「音楽のかかるタイミング」「システムからのはたらきかけ」を5段階（早過ぎる～ちょうど良い～遅すぎる，ないし多過ぎる～ちょうど良い～少な過ぎる）で評定を求めた．

d. システムとの会話（7項目）

「言葉の長さ」「会話の長さ」「会話の自然さ」「会話のスムーズさ」「表現のわかりやすさ」について上記と同一の7段階で評定を求めた．

「システムへの話し方」「システムに話しかけるタイミング」の理解の程度，すなわち自己の評価を5段階（わかった～わからなかった）で評定を求めた．

e. 得られた情報（2項目）

「情報の正しさ」について上記と同一の7段階で評定を求め，さらに「聴きたい曲が聴けたか」について尋ねた．

f. システムの理解度（1項目）

システムの使い方が理解できたかどうかについて，100点満点のうちの点数で自己評価を求めた．

g. 誤認識率の印象（1項目）

システムの誤認識の程度についての印象を，100%のうちのどのくらいであるかを尋ねた．

なお，先行研究ではYes/Noの2段階での評定が多くなされていたのに対し，本研究では5段階ないし7段階，あるいは連続的な数値軸上で回答を求めるように改良した．加えて，「システムへの話し方」「システムに話しかけるタイミング」「システムの理解度」「誤認識率の印象」といったいわば自己評価ともいえる項目を追加していることが本研究の特徴である．

5. 結果と分析

本稿では，観察実験に参加した15名の被験者の結果について分析する．被験者の内訳は，20歳代4名，30歳代5名，40歳代4名，50歳代2名，男女それぞれ8名，7名である．

分析にあたっては，ユーザの「達成度」とシステム評価の相関関係を検討した．「達成度」として以下の5点を扱った．

a. 発話認識率

発話認識率として，実験中の被験者の行動観察の記録データから，1回ごとの発話が認識された率を扱った．ここでは単語ごとの認識率ではなく，複数の単語を含む発話を分析対象としている．

b. 5曲中の発話数

指定5曲を検索するセッションでの発話数を測った．指定曲検索セッションでは，すべての被験者が同一の5曲を検索しており，一定の基準を満たすためにどの程度被験者がシステムに話しかけたかを示す値とみなすことができる．

c. 会話満足度

d. システム理解度

e. 誤認識率の印象

評価アンケートから「会話の満足度」「システム理解度」「誤認識率の印象」を「達成度」の指標として扱った．これらはいずれも，個別的なシステムの性能の評価というより，自己およびシステムのふるまい全体を点数化して評定する項目であり，分析の視点として適切であると考えた．これら3項目の評定は100を最高値として与えられている．

さらに，システムを評価した項目として，評価アンケートから以下の16項目を取りあげる．（ ）内に，相関値を算出するために用いた評定値を表示する．例えば，(a)では，7段階評定項目であり，「大変良い」を7，「大変悪い」を1として相関値を算出したことを意味する．

- 「丁寧さ」「声の印象」「総合的な印象」
  （「大変良い」：7，「大変悪い」：1）
- 「声の聴き取りやすさ」
  （「大変良い」：7，「大変悪い」：1）
- 「親しみやすさ」
  （「大変良い」：7，「大変悪い」：1）
- 「システムの回答のタイミング」「音楽のかかるタイミング」
  （「ちょうどよい」：3，「早い」「遅い」：2，「早過ぎる」「遅過ぎる」：1）
- 「システムからのはたらきかけ」
  （「ちょうど良い」：3，「多い，少ない」：2，「多すぎる，少なすぎる」：1）
- 「言葉の長さ」「会話の長さ」「会話の自然さ」「会話のスムーズさ」「表現のわかりやすさ」
  （「大変良い」：7，「大変悪い」：1）
- 「システムへの話し方」「システムに話しかけるタイミング」（「よくわかった」：1，「わからなかった」：2）
- 「情報の正しさ」（「大変良い」：7，「大変悪い」：1）

5.1. 達成度間の相関関係

達成度を測る5つの度数間での相関係数を算出し，相互の関連性を検討する．表3に達成度間相関係数の結果を示す．±.4以上の相関が見られたのは，「発話認識率」と「システム理解度」(-.54)，「発話認識率」と「誤認識率の印象」(-.66)，「会話満足度」と「誤認識率の印象」(-.54)の3点である．「5曲中の発話数」については，他の達成度との相関は得られなかった．よ

表 3. 達成度間の相関

	発話数	会話満足度	理解度	誤認識率の印象
a. 発話認識率	-. 02	. 15	-. 54	-. 66
b. 発話数		-. 15	-. 24	-. 26
c. 会話満足度			. 36	-. 54
d. ｼｽﾃﾑ理解度				. 14

り高い相関が得られた「発話認識率」と「誤認識率の印象」との結果について，図を例示する．

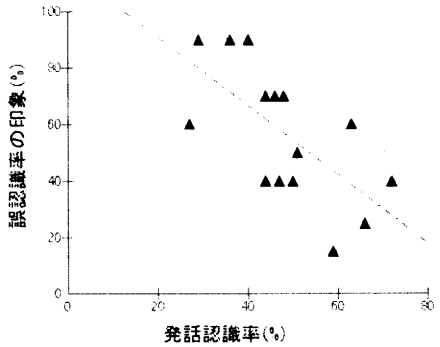

図 3.「発話認識率」と「誤認識率の印象」との相関

「発話認識率」と「システム理解度」については，発話認識率が高くなるにつれ自己のシステム理解の印象が低くなるという傾向がある．実際の認識の程度がよい場合であっても，システムの使い方をよく理解したと評価するとは限らないことをこの結果は示唆している．

「発話認識率」と「誤認識率の印象」については，発話認識率が高くなるにつれ誤認識率の程度の印象が低くなるという結果である．すなわち，実際の認識率が高い被験者は誤認識の率が低い，つまり正しく認識しているとの印象を持ち，逆に実際の認識率が低い被験者は誤認識の率が高い，つまり誤って認識しているとの印象を持っているということができる．被験者は，ある程度正確にシステムの認識精度を測ることができているということを示す結果である．

「会話満足度」と「誤認識率の印象」については，誤認識率の程度の印象が低くなるにつれ，満足度が高くなるという傾向が示されている．正しく認識しているとの印象があると会話に満足し，誤って認識しているとの印象があると会話に満足しなくなりやすいことを示唆している．実際の発話認識率と会話満足度との間に相関が見られなかったにも関わらず，被験者から見た認識の程度の印象と会話に満足したかどうかの間には相関が見られたことは興味深い．実際のシステムの認識率を向上させるだけでなく，高い認識精度を有しているとの印象を与える要因を明らかにし，システム構築に反映させることが，ユーザの満足度を高めるために必要であることを示唆するものと考えられる．

5.2. 達成度と評価項目との相関関係

5.2.1. 発話認識率

「発話認識率」と評価 16 項目との相関係数を算出したところ，いずれの項目とも相関が得られなかった．総合的な印象や会話のスムーズさ，声の聴き取りやすさといった，システム評価で一般的と見られる項目も含まれているが，これらにおいても認識率との間に相関はなかったのである．従来，認識率とシステム評価とがよく一致することが知られているが，必ずしも関連を持っているわけではないことを予想させる結果である．

5.2.2. 5 曲中の発話数

評価 16 項目のうち，「5 曲中の発話数」と相関があった項目は，「話しかけるタイミング」の 1 項目のみであった（-.43）．発話数が多くなるほど話しかけるタイミングの理解が低いとの印象を持ったという結果である．言い換えると，より多く発話が生起した被験者では，どのタイミングで話せばよいかの理解が不充分であったとの印象を持っている．話しかけるタイミングがわからないためにより多く発話したことを表わしていると考えられる．

前項で，5 曲中の発話数については発話認識率とも相関が見られなかったことが示されている．発話が多く生起するのは，誤認識率が高いことが要因ではなく，ここで指摘したように，システムへの話しかけるタイミングがわからないためであると考えられる．

5.2.3. 会話満足度

「会話満足度」については，評価 16 項目のうち 8 項目と相関が見られた．うち，まず 7 項目を示す．

- 「総合的な印象」(.69)
- 「声の聴き取りやすさ」(.61)
- 「親しみやすさ」(.69)
- 「会話の長さ」(.41)，「会話の自然さ」(.47)，「会話のスムーズさ」(.46)
- 「システムへの話し方」(.52)

これらの項目では，満足度が高くなるにつれ，システムに対してよい印象を持つ傾向が高くなる結果と見ることができる．中でも，特に「総合的な印象」との間に相関が見られているのが会話満足度の結果の特徴である．会話に満足した場合にはシステム全体を評価した総合的な印象が高くなる結果が得られている．「会話に満足する」ということがどのような要因によってもたらされるものか，より詳細に検討することにより，総合的な印象を向上させることが可能となると考えられる．今回の結果では，会話の自然さ，スムー

ズさとの相関関係も認められており，これらの点を詳しく分析していきたい．

上記に加え，相関が見られた項目として「音楽のかかるタイミング」がある（.53）．満足度が高い場合では，音楽のかかるタイミングという点でのシステムからの反応が適切であるとの印象を持つ傾向があると見ることができる．

5.2.4. システム理解度

「システム理解度」については，評価 16 項目のうち 8 項目と相関が見られた．うち，まず 7 項目を示す．

・「丁寧さ」（.62）

・「親しみやすさ」（.47）

・「言葉の長さ」（.52），「会話の長さ」（.52），「表現のわかりやすさ」（.51）

・「システムへの話し方」（.48），「システムに話しかけるタイミング」（.48）

上記の項目では，理解度が高くなるにつれ，大変良いとする印象が高くなるという結果が示されている．

相関が見られた項目の分類を見ると，システムとの会話に関する評価項目が多いことがシステム理解度に関する結果の特徴である．すなわち，システムとの会話に関して良い印象を持った場合には，システムの使い方をよく理解したと感じる傾向があるということを示唆している．この点は，ユーザにとって適切な会話を定義することにより，システムを使いこなせるようになったとの理解度が促進される可能性があることを指摘するものである．

システムとの会話に関する評価項目の中でも，「会話の自然さ」，「会話のスムーズさ」の 2 項目は，理解度との相関が得られなかった（.27，.22）．一方，これらの項目は 5.2.3.で述べたように，会話満足度との相関が得られている．これらの結果は，類似した評価項目であってもユーザの達成度の反映の有り様が異なることを示しており，システム評価項目を洗練させるための手がかりとなる結果であると考えられる．

上記の 7 項目に加え相関の見られた項目は「システムからのはたらきかけ」（.60）である．この結果は，理解度が高くなるにつれシステムからのはたらきかけが適切であったとの印象が高くなることを示している．

5.2.5. 誤認識率の印象

「誤認識率の印象」については，評価 16 項目のうち 4 項目と相関が見られた．相関が見られたのは，「声の聴き取りやすさ」（-.42），「親しみやすさ」（-.55），「音楽のかかるタイミング」（-.60），「情報の正しさ」（-.45）である．これらの項目では，いずれも，誤認識の程度の印象が高くなるにつれ，「大変悪い」と回答する傾向が高くなるという結果である．言い換えると，正しく認識しているとの印象を持つ被験者ではこれらの項目に関して大変良いとする傾向があり，逆に，誤って認識しているとの印象を持つ被験者ではこれらの項目に関して大変悪いとする傾向があることが示されている．これらの中でも，特に「情報の正しさ」との間に相関が見られているのが誤認識率の印象に関する結果の特徴である．高い相関ではないが，誤認識率の印象が高いほど得られた情報が正しいとする印象が高くなるのである．5.1 において，ユーザの満足度を高めるためには高い認識精度を有しているとの印象を与える要因を明らかにすることが必要であると指摘したが，情報が正しいとの印象をもたらすためにもこの点を把握することが重要であると考えられる．

6. まとめ

楽曲検索システムを構築し，システムを用いた対話実験を行なった．観察実験の結果から，達成度間および達成度とシステム評価項目との相関関係を検討した．分析の結果，以下の点が明らかとなった．

「発話認識率」と「誤認識率の印象」との相関関係から，被験者がある程度正確にシステムの認識精度を予測できることがわかった．「会話満足度」は「発話認識率」と相関がない一方で「誤認識率の印象」とは相関があることから，実際のシステムの認識率を向上させるだけでなく，高い認識精度を有しているとの印象を与える要因を明らかにし，システム構築に反映させることがユーザの満足度を高めるために必要であることが示唆された．

システム評価項目との相関に関しては，「発話認識率」，「5 曲中の発話数」と関連する項目はほとんどなく，むしろ，「会話満足度」，「システム理解度」と多く関連することが明らかとなった．従来，認識率とシステム評価とがよく一致することが知られているが，必ずしも関連しないことを予想させる結果であり，他の観点からのシステム評価を考慮する必要があることを示唆している．

今後は，今回の観察実験の結果を踏まえ，被験者数を増大させて条件を統制した実験を行ない，ユーザの持つ知識，技術とシステム評価の関連性についてより詳細に明らかにしたい．

文　献

[1] 河原達也，李晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，山本幹雄，山田篤，宇津呂武仁，鹿野清宏，“日本語ディクテーション基本ソフトウェア(98 年度版)”，日本音響学会誌，56，4，pp.255-259，Apr.2000.

[2] S. Möller and J. Skowronek, “Quantifying the Impact of System Characteristics on Perceived Quality Dimensions of a Spoken Dialogue Service,” Proc. 8th European Conf. on Speech Commun. Tech. (Eurospeech 2003)，pp.1953-1956，Geneva，Switzerland，Sept.2003.