2004－SLP－53　(6)
2004／10／22

音声対話による楽曲検索システム

原 直[†], 白勢彩子[‡], 宮島 千代美[†], 伊藤 克亘[†], 武田 一哉[†]

[†] 名古屋大学大学院情報科学研究科, [‡] 独立行政法人理化学研究所

〒 464-8603 名古屋市千種区不老町 1
[†]{hara,miyajima,itou,takeda}@sp.m.is.nagoya-u.ac.jp, [‡]shirose@brain.riken.jp

あらまし　近年，音声認識技術を用いた様々なアプリケーションが考えられている．例えば，ハンズフリーで利用できるという特徴を生かして，カーナビゲーションシステムなどに利用されている．また，インターネット経由の楽曲ダウンロードシステムが増えており，使いやすい楽曲検索インタフェースへの期待が高まっている．そこで本研究では車内での利用を想定した音声対話による楽曲検索システムを作成した．このシステムは，ユーザが対話によって聞きたい楽曲を検索し再生するというシステムである．本報告ではシステムの詳細及びシステムを用いた音声対話の収録について述べる．プロトタイプを用いて約 150 名の被験者実験を行った結果，室内で約 80%，実車運転時で約 76%の単語正解率を得た．

キーワード　音声認識，音声対話インタフェース，楽曲検索，ユーザビリティ

A music searching system by spoken dialogue

Sunao HARA[†], Ayako SHIROSE[‡], Chiyomi MIYAJIMA[†],
Katsunobu ITOU[†] and Kazuya TAKEDA[†]

[†] Graduate School of Information Science, Nagoya University
[‡] RIKEN

Furo-cho 1, Chikusa-ku, Nagoya 464-8603, JAPAN
[†]{hara,miyajima,itou,takeda}@sp.m.is.nagoya-u.ac.jp, [‡]shirose@brain.riken.jp

Abstract　Recently, various applications equipped with speech recognition are developed. For example, it is used for the car-navigation systems with handsfree operation. There are some systems of music download via the Internet, so a music search interface which is easy to use is expected. Then we create a music search system supposing use in the car was used by spoken dialogue. This system can search and play the music what the user want to listen. In this paper, we discuss a detail of the system and spoken dialogue recording with the system. Experimental results of 150 subjects with a prototype system show that the system could achieve about 80% word correct indoor environment, and about 76% word correct in car environment.

Keywords　speech recognition, spoken dialog interface, music searching, usability

1　はじめに

近年，音声認識技術はハンズフリーな入力手段として注目されており，様々な場面での実用化が考えられている．実際に製品化されている例としてカーナビゲーションシステムが挙げられる．車の運転中にカーナビゲーションシステムを操作していると，脇見運転・片手運転になってしまい危険である．しかし，音声認識技術を用いることで，機器に触れることなくカーナビゲーションシステムを操作することが可能である．カーナビゲーションシステムは主に地図情報を検索するシステムであり，車内で利用する情報検索システムとしてはさまざまなものが考えられる．例えば，レストラン検索・案内システム，ニュース検索システム，そして楽曲検索システムなどである．

また近年，インターネット経由での楽曲ダウンロードシステムが増えており，使いやすい楽曲検索インタフェースへの期待が高まっている．ダウンロードのための楽曲検索としてはWWWページ上でテキスト検索が行われている．一般に用いられている検索手法は，アーティスト名や曲名などのキーワード検索やジャンルからの絞込み検索などである．

そこで本研究では，車内で利用する情報検索システムとしてインターネットを利用した音声対話による楽曲検索システムを作成した．これは，ユーザが対話によって聞きたい楽曲を検索しストリーミング再生するというシステムである (図 1)．また，このシステムは音声対話インタフェースのユーザビリティの評価に適している．車運転時には多くの人がラジオや楽曲を聞いている．そのためユーザは普段行っている行動を音声によって行うことになり，システムよりもインタフェースそのものの評価を行いやすい．

本報告ではインタフェースの仕様及びインタフェースを用いた音声収録の概要を示す．また，収録した音声の認識を行いインタフェースの評価を行う．

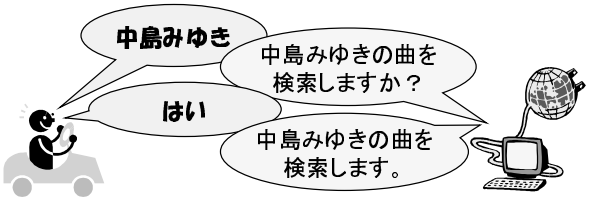

図 1: 音声対話による楽曲検索の概要

2　システム概要

2.1　音声対話による楽曲検索

本インタフェースを用いた楽曲検索は図 2 のような流れで行われる．すなわち，システムに音声による検索要求を認識させ，認識結果に基づきインターネット上の検索サービスより楽曲一覧を取得し，一覧から楽曲を選択すると曲が流れる，という手順で

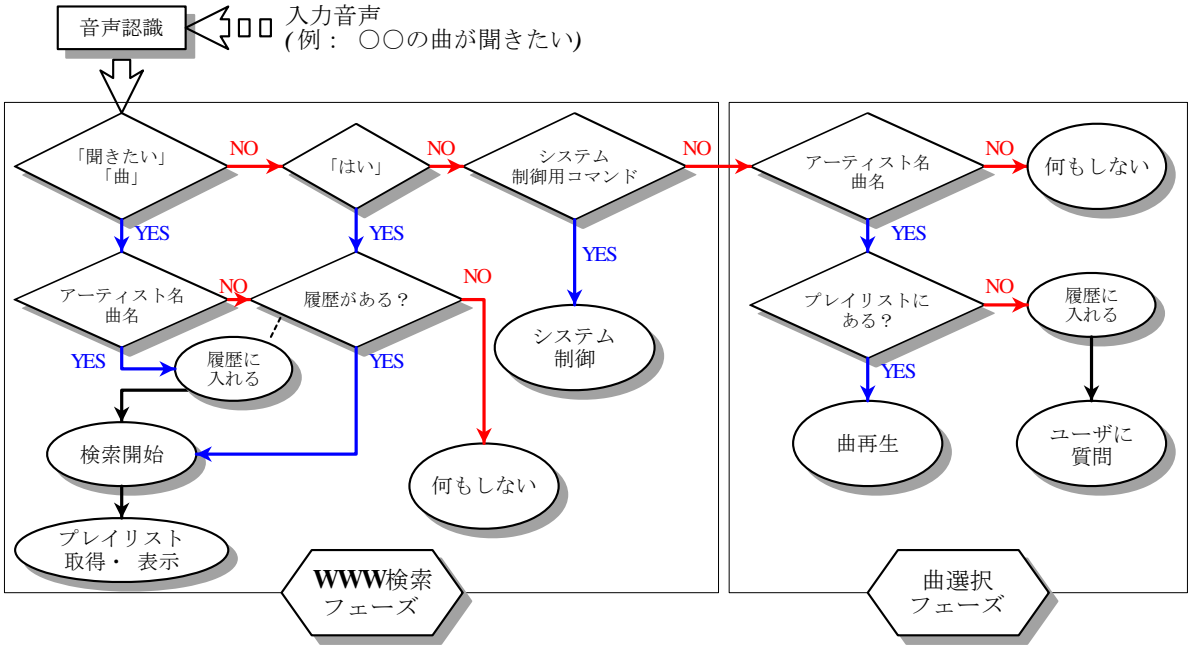

図 2: 音声による楽曲検索の流れ図

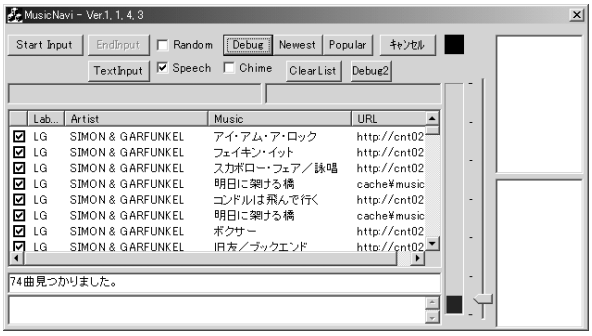

図 3: 音声対話インタフェース画面

ある．このとき，インタフェースからの指示や楽曲リスト提示は合成音声によって行われる．ユーザはインタフェースとの対話をすることで楽曲を選択することができる．

2.2 インタフェースの仕様

インタフェースは図 3 のような画面となっている．ただし，ユーザは利用時にインタフェースの画面を見る必要はなく，認識結果や検索結果などはすべて合成音声でユーザに提示する．

本インタフェースでは，楽曲検索サービスに，音楽配信ポータルサイト「Mora(`http://mora.jp`)」を利用した．Mora はアーティスト名・曲名・キーワードによる部分一致検索と，アーティスト・曲名・レーベル名の各頭文字による絞り込み検索という 2 通りの検索方法が可能である．本実装では，前者のうちキーワードによる検索を利用している．

音声合成エンジンには FUJITSU FineSpeech を用いた．

音声認識エンジンには大語彙音声認識エンジン Julius の WindowsDLL 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 3000/129，性別非依存，64 混合，PTM triphone モデルを用いた．

言語モデルは，次の図 4 に示す文法から学習文を生成し，bigram, 逆向き trigram を作成した．ここで,「eps」はヌル遷移,「silB」は文頭,「silE」は文末を表すシンボルである．文法中の <AAM> はアーティスト名や曲名を表すシンボルであり，図 5 の構造を持っている．また，<COMMAND> はシステム制御用のコマンドを表すシンボルである．

認識に用いる辞書の語彙サイズは 7710 単語 (うちアーティスト 1601 名，曲数 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 Mora 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．上記二つのデータを重複なしに結合して，辞書を作成した．

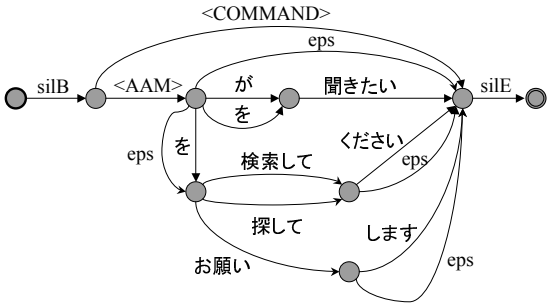

図 4: メイン文法

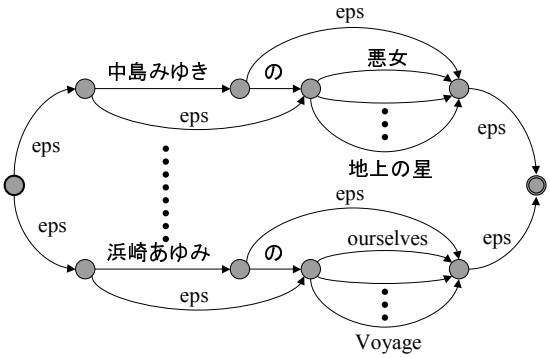

図 5: <AAM>　アーティスト名及び楽曲

本インタフェースでは，検索した楽曲が WWW サイトでもダウンロードできないときやネットワーク接続の不具合が生じたときに楽曲のダウンロードに失敗してしまう．インタフェースはダウンロードに失敗したとき,「曲のダウンロードに失敗しました」と言う．ダウンロードに時間がかかっている場合にも,「ダウンロードに時間がかかっています」と言う．これらのメッセージは誤認識によって誤動作しているとユーザが勘違いしないようにするためのものである．

2.3 検索例

図 1 のようにユーザはインタフェースと対話をすることで楽曲検索を行う．また，キーワードのみによる検索も可能である．以下にキーワード検索の例を挙げる．

(1) アーティスト名で検索　　認識結果が「<アーティスト名>」＋「曲」，または「<アーティスト名>」＋「曲」＋「<動詞>」という組み合わせのとき，<アーティスト名>で検索を行う．例えば,「中島みゆきの曲」,「中島みゆき

の曲が聞きたい」と認識すると，プレイリストには検索結果全てが表示され，＜曲名＞でのみヒットした項目は検索誤りとしてチェックマークがはずされる．

(2) 曲名で検索　　認識結果が,「＜曲名＞」＋「＜動詞＞」という組み合わせのとき，＜曲名＞で検索を行う．例えば,「地上の星が聞きたい」と認識すると，プレイリストには検索結果全てが表示され，＜アーティスト名＞でのみヒットした項目は検索誤りとしてチェックマークがはずされる．

(3) アーティスト名と曲名で検索　　認識結果が,「＜アーティスト名＞」＋「＜曲名＞」＋「＜動詞＞」という組み合わせのとき，＜アーティスト名＞で検索を行う．例えば,「中島みゆきの地上の星が聞きたい」と認識すると，プレイリストには＜アーティスト名＞による検索結果が表示され，＜曲名＞がヒットした項目にチェックマークがつけられる．

(4) おすすめの曲を聴く　　認識結果が「最新の曲」または「人気の曲」であったとき，それぞれ Mora のサイトで最近登録された曲，人気タイトルとして紹介されている曲がプレイリストに表示される．

3　インタフェースを用いた対話音声収録

3.1　収録概要

プロトタイプシステムを評価するために被験者実験を行った．被験者数は 143 人である．

被験者は観察実験 (47 人)，統制実験 (96 人) に分類される．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する．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する．

統制実験のうち 48 名は教示方法によって分類する．このときマニュアル (紙 or ビデオ) や認識訓練 (行う or 行わない) によって 4 群に分けた．残りの 48 名は発話環境 (室内・シミュレータ・車内) によって分類する．

室内での収録は名古屋大学 IB 情報館 417 号室で行った (図 6)．部屋は四方にカーテンが掛けられている．暗騒音レベルは 28.7dB(A) であった．使用した機材の名称及び接続図をそれぞれ表 1，図 7 に示す．

図 6: 収録風景 - 室内

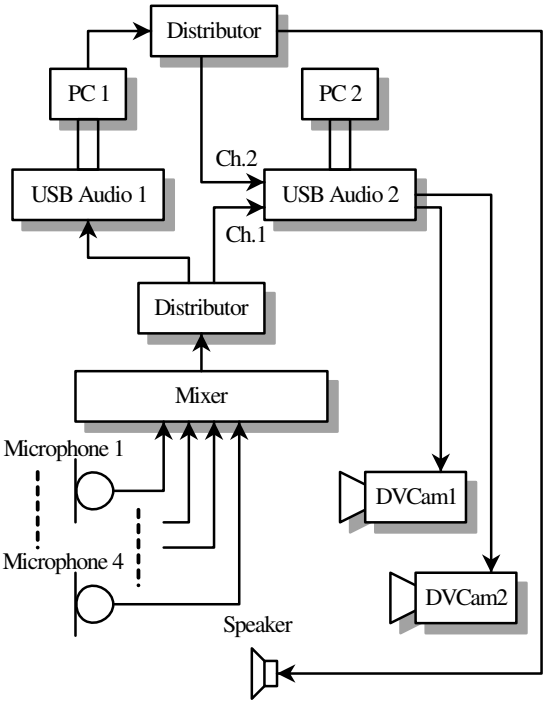

図 7: 室内機材接続図

表 1: 室内収録使用機材

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
Mixer	Sony SRP-X1008
Microphone 1,2	Sony ECM-77B
Microphone 3	SENNHEISER HMD 410
Microphone 4	Sony F-740
Speaker	YAMAHA MS101 II
DVCam 1	Sony DCR-TRV900
DVCam 2	Panasonic NV-GS100

車内での収録にはトヨタ レジアス [2] を用いた

図 8: 収録風景 - 車内

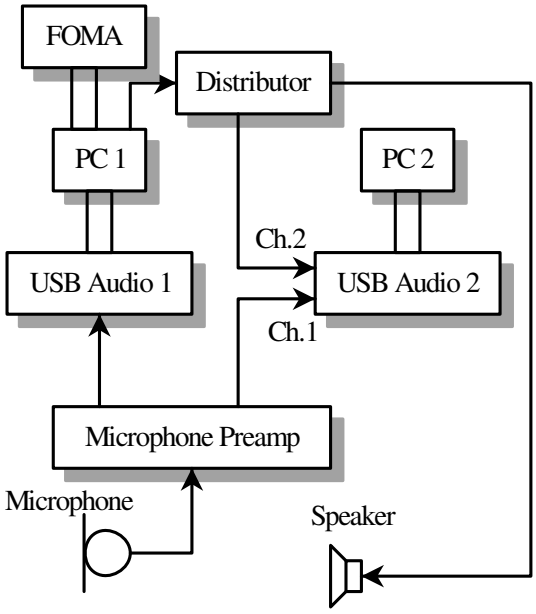

図 9: 車内機材接続図

表 2: 車内収録使用機材

PC 1,2	SUMICOM P4-S600
FOMA	NTT DoCoMo F2102V
USB Audio 1	M-Audio Mobile Pre USB
USB Audio 2	EDIROL UA-5
Microphone Preamp	audio-technica AT-MA2
Microphone	SENNHEISER HMD 410
Speaker	YAMAHA MS101 II

(図 8)．使用した機材の名称及び接続図をそれぞれ表 2，図 9 に示す．

各被験者に対して，以下の課題 A,B を与えて，最後にインタフェースに関するアンケートを実施した．
A) 指定曲検索：指定された 5 曲をダウンロードできるまで検索
B) 自由曲検索：自由に検索 (約 15 分)

表 3: システム反応速度 [sec]．括弧内は実時間比．

環境	発声 → 認識	認識 → 検索結果
室内	5.05 (2.76)	0.58 (0.042)
シミュレータ	4.95 (2.80)	0.85 (0.481)
実車運転時	10.4 (5.86)	6.36 (3.58)

3.2 収録されたデータ

観察実験では，インタフェースによる音声のほかにビデオカメラ 2 台を用いて映像を収録した．このとき，一台は被験者正面に配置し，もう一台は被験者斜め方向からの俯瞰撮影を行った．

続いての統制実験では，室内については音声と正面映像のデータを収録した．シミュレータ運転時ではさらにシミュレータの動作ログデータを収録した．実車運転時は接話マイク (SENNHEISER HMD 410) による音声と，ドライバの顔映像 (2 視点)，遠隔マイク (9 箇所)，運転行動記録を収録した．

4 インタフェースの性能評価

4.1 システム反応速度

収録時のシステムの反応時間を表 3 に示す．室内・シミュレータ環境では実時間の約 3 倍で動作しているが，車内環境では実時間の 9.4 倍である．発声から認識までの速度差は，車内の PC と室内の PC との性能の違いによるものであり，認識から検索結果表示までの速度差は，ネットワークの違いによるものである．

4.2 収録音声の認識結果

143 名，計 216 セッションの音声を用いて音声認識を行い性能を評価する．

音声認識には大語彙音声認識エンジン Julius 3.4.1 を用いた．言語モデルにはデータベース収録に使用したインタフェースと同様のものを使った．語彙数は 7710 単語である．音響モデルには，CIAIR の車内音声データベースより学習した音響モデル (性別非依存,32 混合,2000 状態,triphone) での認識結果を用いて話者適応 (教師無し MLLR 適応) を行ったモデルを用いた．

実験 A,B における 143 人の音声対話より音声区間を手動で切り出し，17,869 発話を得た．この全発話のテストセットパープレキシティは 23.2 であり，未知語率は 5.79%であった．また，この全発話に対して音声認識を行った結果，単語正解率 (%Cor-

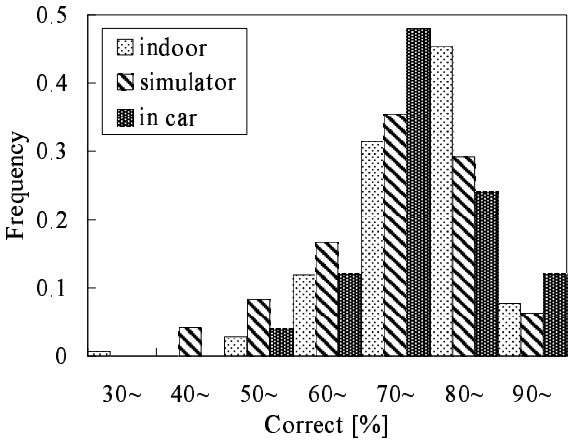

図 10: 発話環境別の単語正解率のヒストグラム

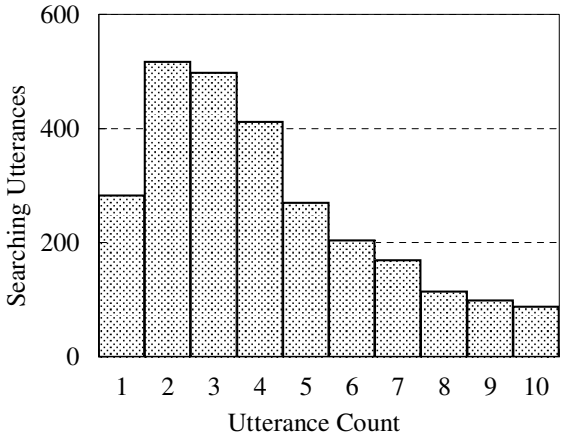

図 11: 検索要求あたりの発話数のヒストグラム

rect)=79.1[%], 単語認識精度 (Accuracy)=67.4[%] であった.

発話環境ごとに算出した単語正解率のヒストグラムを図 10 に示す. 平均単語正解率は室内 (indoor) で 79.7[%], シミュレータ運転時 (simulator) で 76.1[%], 実車運転時 (in car) で 75.9[%] であった. 室内での認識率は高い傾向にあり, シミュレータ運転時や実車運転時ではやや認識率が落ちている. 特にシミュレータ運転時は低い認識率にも分布しており, 被験者への負荷が大きかったと考えられる.

環境ごとに算出した一発話あたりの単語数の平均は, 室内で 1.68 語, シミュレータ運転時で 1.46 語, 実車運転時で 1.36 語となった. 室内での単語数がもっとも多く, シミュレータ運転時, 実車運転時にはやや少なくなっている. これは, 車の運転に気をとられているために発話の内容が単純になっていると考えられる. このことからも車の運転が音声対話に影響を与えていることがわかる.

ある曲を聞くための一連の対話 (検索要求) に含まれる発話数を図 11 に示す. 検索要求は 2967 セットとなり, 1 セットの平均発話数は 5.49 文であった. 発話数は 2,3 文が頻度が高い. これはシステムに質問し, その回答に対して返答して終わるというケースが多いためである.

なお, 本インタフェースを全く使えないという人はおらず本システムの有用性が示された.

5 まとめ, 今後の課題

本報告では, 楽曲検索のための音声対話インタフェースの構築, ならびに本インターフェースを用いた収録実験について説明した. また, 収録データを用いてデータ分析を行った. 本インタフェースで収録した対話の音声認識を行った結果, 室内で約 80%の単語正解率を得た. しかし, シミュレータ運転時・実車運転時では単語正解率と一発話あたりの単語数がともに室内に比べ低下していることから, インタフェースの利用がユーザにとってやや負荷が大きいものであったと考えられる. 特に負担が大きいのは楽曲数が多くなったときであることが観察により確認された. これは楽曲数が多くなったときにも検索された楽曲をすべて読み上げているためである. また, 本インタフェースでは明確なシステムの状態というものが定義されていないため, 誤認識時にユーザの意図しない結果が返ってくることが多かった.

そこで楽曲数の多いときにも対応できるよう楽曲の絞込み検索を実装すること, 各状態でのシステムの動作を容易にユーザに推測できるようシステムの状態を定義することが今後の課題となる.

謝辞

本研究の一部は文部科学省「e-Society 基盤ソフトウェアの総合開発」によるものである.

参考文献

[1] 河原, 李, 小林, 武田, 峯松, 伊藤, 山本, 山田, 宇津呂, 鹿野, "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 巻 4 号, pp.255-259, 2000

[2] 河口, 松原, 武田, 板倉, 稲垣, "実走行車内音声対話データベース", 情報処理学会研究報告, SLP39-24, pp.141-146, 2001

[3] 白勢, 原, 藤村, 伊藤, 武田, 板倉, "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", 信学技報 SLP49-43, 2003.

[4] 原, 白勢, 伊藤, 武田, 板倉, "音声対話による楽曲検索インタフェースを用いたユーザビリティ評価", 音講論 3-Q-31, pp.205-pp.206, Mar.2004.