2005－SLP－55　(4)
2005／2／4

音声対話インタフェースの長期利用時における学習効果の評価

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 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構築している．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．以前の報告ではユーザが約 1 時間の間システムを利用する実験を行った．この実験において被験者によって度合は異なるものの，習熟することにより認識性能が向上するという知見が得られた．そこで，本報告では，被験者がシステムに十分慣れるように，1 時間のセッションを 5 回繰り返す実験を行った．実験により収録した 12 名の音声を分析した結果，最終日において初日の約 60%の誤り改善率を得た．

キーワード　音声認識，音声対話インタフェース，楽曲検索，学習効果

Evaluation of training effects by long-term use of a spoken dialogue interface

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 We are developing a music retrieval system for in-car use based on a spoken dialogue interface. The system can retrieve and play musics that a user wants to listen to. We have previously conducted experiments where each subject uses the system for one hour. In the experiments, we have found that the speech recognition performance is improved as the subjects get used to the system, although the degree of training depends on the subject. In this paper, we conduct extended experiments where each subject uses the system over five one-hour sessions. Experimental results for twelve subjects show that the system achieves about 60% relative improvement in recognition performance at the fifth session compared to the first session.

Keywords　speech recognition, spoken dialog interface, music searching, training effect

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，例えば，カーナビゲーションシステムが応用例として挙げられる．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が，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．

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ている [1]．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．システムとの対話例を図 1 に示す．

以前の報告では，被験者が約 1 時間インタフェースを利用する実験を行った [1] が，被験者によって習熟度合は異なるが，インタフェースの操作に習熟することで認識性能が向上するという知見が得られた．実環境においてユーザがインタフェースを利用する場面を想定すると，ユーザはインタフェースを何度も利用する中で，よりインタフェースの操作に慣れることが考えられる．

そこで本報告では，ユーザが複数回インタフェースを利用した際の音声収録の概要を説明する．またインタフェースの評価にどのような影響を与えるのかを報告する．

System:	何か聞きたい曲はありませんか？
User:	サイモン＆ガーファンクル
System:	サイモン＆ガーファンクルの曲を検索しますか？
User:	はい
System:	サイモン＆ガーファンクルの曲を検索します． (… 検索中 …)
System:	60 曲見つかりました． I am a rock，明日に架ける橋 …
User:	その曲
System:	サイモン＆ガーファンクルの明日に架ける橋をダウンロードします． (… 楽曲再生 …)

図 1: インタフェースとの対話例

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以下の手順で行われる．まず，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

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 [2] 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 2 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 3 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．

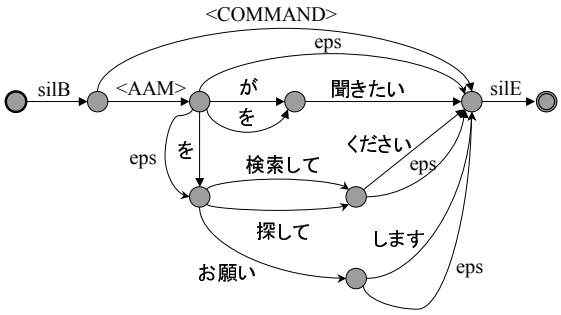

図 2: メイン文法

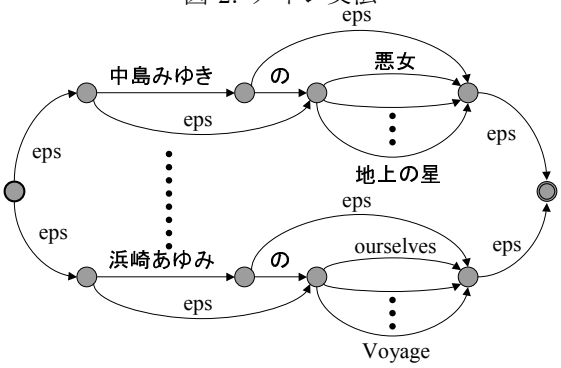

図 3: <AAM>　アーティスト名及び楽曲

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．

3　複数日利用時の対話音声収録

3.1　収録概要

被験者は 12 名とし，各被験者は 5 回インタフェースを使用した（のべ 60 名）．このとき，各被験者は同じ日に 2 回以上の実験は行っていない．

室内での収録は名古屋大学工学部 7 号館可変残響室 (図 4) にて行った．暗騒音レベルは 21.2dB(A) であった．使用した機材を表 1 に示す．

車内での収録にはトヨタ レジアス [3] を用いた．

一人の収録は次の手順で行われる．

1. 新聞記事の読み上げ (10 分)
2. 室内での楽曲検索 (20 分)
3. シミュレータ運転時の楽曲検索 (15 分)

図 4: 収録風景

表 1: 収録使用機材

PC 1	Sony PCG-GRT99V/P
PC 2	Sony PCF-V505R/PB
USB Audio 1	M-Audio Mobile Pre USB
USB Audio 2	EDIROL UA-1000
Mixer	Sony SRP-X1008
Microphone S	SENNHEISER HMD 410
Microphone O	Sony F-740
Speaker S,O	YAMAHA MS101 II
DVCam	Sony DCR-TRV900

4. 実車運転時の楽曲検索 (15 分)
5. 事後アンケート (10 分)

読み上げに用いたテキストには asahi.com の 2004/5/4〜2004/5/8 の記事を利用した．これには，191 記事，1714 文が含まれている．

3.2　統制条件

被験者への統制は 2 種類設定した．それぞれの条件について統制を行う，または統制を行わないとすることで，全 4 群 (fl, fL, Fl, FL) の被験者群となる．

- フィードバック条件
  記事読み上げの際に被験者の発声に関する情報をフィードバックする
  - `F` フィードバック有り
  - `f` フィードバック無し
- オペレータ指導条件
  楽曲検索システムの利用時にオペレータが被験者のそばについて指導を行う
  - `L` 指導有り
  - `l` 指導無し

F 群の被験者へは以下の情報を提示した．

- フレーム内の二乗平均パワー
- フレーム内の二乗平均パワーの変動
- ピッチの変動
- ピッチの傾き
- 話速 (音素継続長)

被験者へのフィードバックは一発話毎に行わる．被験者はモニタで確認しながら新聞記事の読み上げを行った．

4　時系列での分析結果

収録データを発話環境や時系列により分類し分析する．環境の表記は室内収録を indoor，シミュレータ運転時収録を simulator，実車運転時収録を incar と表記する．評価は「システムとの会話の満足度」と「システムの使用方法の理解度」の主観評価と成功率による客観評価を行う．また，音響的特徴として平均 SNR の標準偏差と話速 (音素継続長) の平均を，言語的特徴として未知語率の平均と一発話の単語数と平均 Perplexity を用いて評価する．

なお，成功率とは，音声認識結果と発話文の意味が同じであるかどうかで評価した基準である．

4.1　収録データ概要

12 名の被験者それぞれが 5 日間の実験を行っており，また各実験において 3 環境で実験を行ったので，計 180 セッションのデータを収録した．総文数は 11915[文] であり，全単語数は 12988[語] であった．従って，一文あたりの単語数は 1.09[語/文] である．

インタフェース使用時と同様のモデルを用いたとき，平均単語認識率は 69.0[%]，平均単語正解精度は 54.9[%] であった．また，未知語率は 17.3[%] であり，Perplexity は 7.29 であった．全被験者の各日における平均単語認識率を図 5 に示す．

次に，被験者の発話内容の種別を図 6 に示す．図より，コマンド文の発声が多いことがわかる．時間経過やセッションの違いによる変化の違いは見られなかった．

4.2　発話文の言語的特徴

時系列での未知語率の平均，一発話あたりの単語数，検索戦略の変化，平均 Perplexity をそれぞれ図 7，図 9，図 10，図 8 に示す．未知語率には下降傾向が見られ，単語数には下降傾向が見られ，Perplexity には上昇傾向が見られる．Perplexity はアーティスト名や曲名が多く発声されることで増加する．これらより，被験者が習熟することによって無駄な発話を減らし，検索できないとわかっているアーティスト名や曲名以外を発声していたと考えられる．特に，未知語率の低下と Perplexity の上昇は認識不可能なアーティストや曲名を被験者が学習したことを示していると考えられる．図 9 や図 10 より，発話のほとんどが単語発声となっており，2 日目には 95%の発話が単語発声となっていることがわかる．

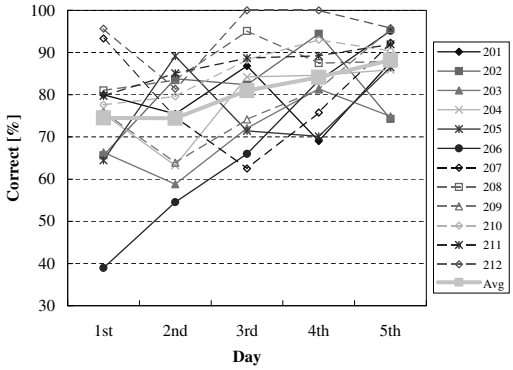

図 5: 室内収録の認識率

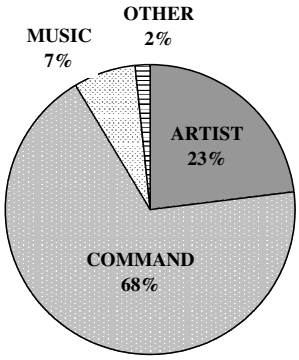

図 6: 発話種別の分類

また，以前の報告でシミュレータ運転時に，Perplexity の低下が見られた [1] が，図 8 を見るとシミュレータ運転時に特に Perplexity が上昇していることがわかる．これはシミュレータ運転に慣れたことからシステムにより意識を向けることができたためと考えられる．

4.3　収録環境別の分析結果

以下の分析は未知語を全て登録した言語モデルを使用することで，未知語の影響を排除した評価を行った．未知語登録済言語モデルを使用したとき，平均単語認識率は 80.8[%]，平均単語正解精度は 72.4[%] であった．

環境別での成功率の平均を図 11 に，また誤り改善率の平均を図 12 に示す． 図より，時系列後半にな

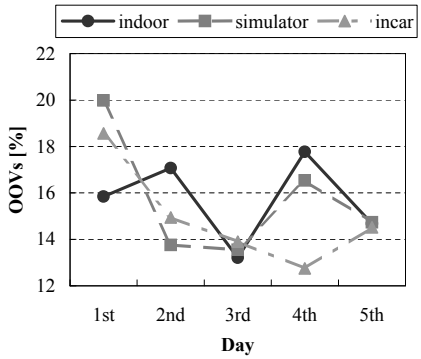

図 7: 未知語率

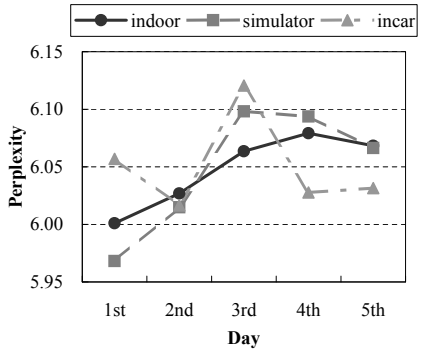

図 8: Perplexity の変化

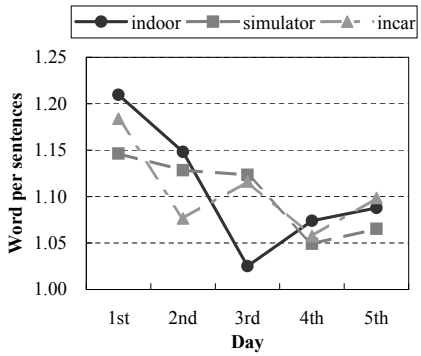

図 9: 一文あたりの単語数の変化

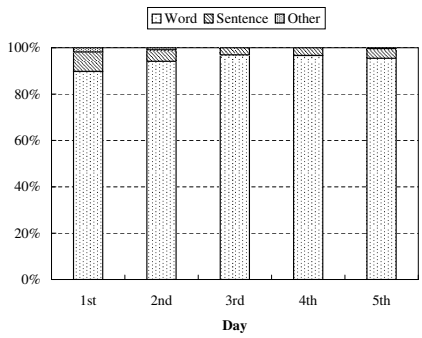

図 10: 検索戦略の変化

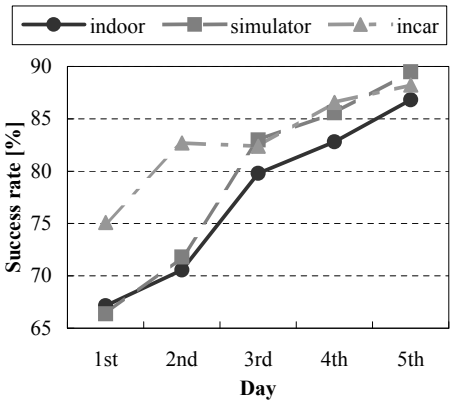

図 11: 各環境毎の成功率

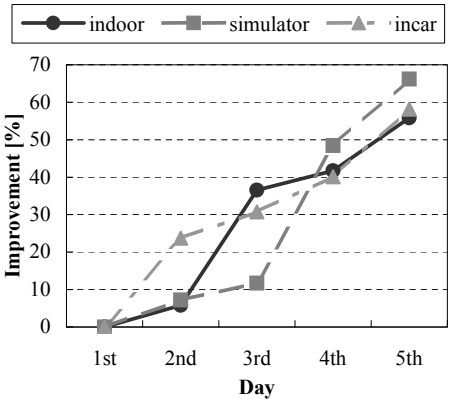

図 12: 各環境毎の誤り改善率

るにつれて成功率が向上しており，誤り改善率は 3 環境全てを平均したとき最終日において 60.1[%] となっている．実車運転時 (incar) において全体的に成功率が高くなっている．しかし，時系列後半になると実車運転時も室内 (indoor) もシミュレータ運転時 (simulator) もほぼ同じ成功率となっていることから，実車運転時収録が一日の実験で最後の実験なので一時間程度で被験者の習熟が見られたと考えられる．

時系列での SNR の標準偏差を図 13 に示す．SNR の算出には一発話の対数フレームパワーの分布を，音声と雑音による 2 混合のガウス分布として推定する手法を用いた [**?**]．図より SNR の標準偏差について下降傾向が見られる．認識率の向上と関連づけて考えると，SNR のばらつきが減ることによって認識率が向上したと考えられる．

話速の平均を図 14 に示す．図より SNR の標準偏差同様に下降傾向が見られる．話速が音響モデルとして使用した HMM に対する最適な話速となることで認識率が向上したと考えられる．

満足度と成功率を時系列に並べた図を図 15 に示

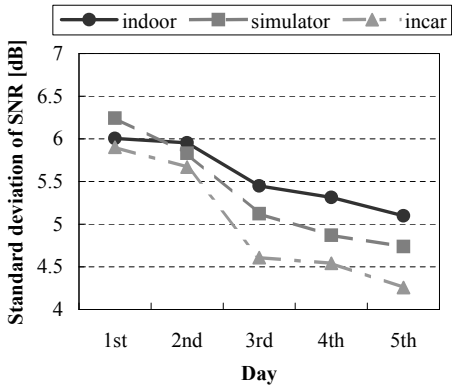

図 13: SNR の標準偏差の変化

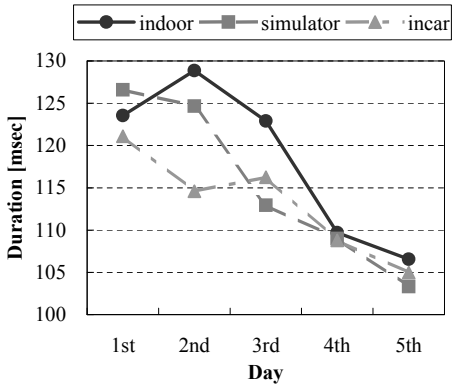

図 14: 話速の変化

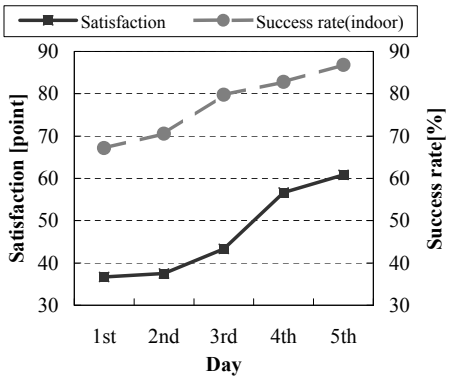

図 15: 満足度の変化

す．満足度は成功率の増加と同様に増加しており，成功率の向上が満足度の向上に貢献していると言える．

4.4 統制条件別の分析結果

各統制条件での室内収録 (indoor) 音声の成功率を図 16(フィードバック条件) に示す．図 16 よりフィードバック有り (F) では初日から成功率が高いことがわかる．これは事前訓練の効果の現れであると考えられる．オペレータによる指導の有無によって成功率の目立った違いは見られなかった．このことから，

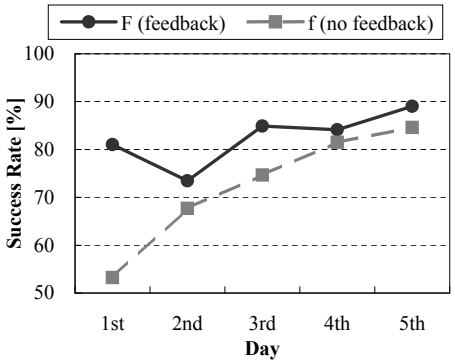

図 16: フィードバック条件有無による成功率

オペレータによる指導はユーザの習熟に影響を与えていないと言える．

5 まとめ，今後の課題

本報告では，複数回の音声対話インタフェース利用実験の概要を説明し，その収録データの分析を行った．客観的評価及び主観的評価共にインタフェースの使用回数を重ねる毎に向上するという結果を得た．特に成功率については最終日において初日に対して約 60[%] の誤り改善がみられた．また，統制条件の違いによる成功率の変化より，オペレータにより指導を行うことによる影響は小さく，事前にフィードバックを与えて訓練を行うことで成功率を高くすることができるという結果を得た．

今後の課題としては，事前訓練としてより効果のある方法を検討する必要がある．

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] 原, 白勢, 宮島, 伊藤, 武田, ”音声対話による楽曲検索システム”, 情報処理学会研究報告 SLP-53-6，2004.

[2] 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

[3] 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

[4] T.H Dat, K. Takeda and F. Itakura, “Robust SNR estimation of noisy speech based on Gaussian mixtures modeling on log-power domain,” COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction (Robust2004, Norwich), (CDROM Proceedings), 2004.