Vol.2011-SLP-87 No.17
2011/7/23

# 住空間音環境コーパスを活用した適応型音声インタフェースデザイン

池谷謙吾[†]　柴田健一[†]
竹林洋一[††]　北澤茂良[†††]　桐山伸也[†††]

住空間における家電・音響機器の多様化・高機能化が進んでいるが，核家族化と高齢化に伴い多様なユーザや利用環境への適応が課題となっている．筆者らは，個人や空間の特性の差異が特に大きい「音」に着目し，音環境の包括的デザインという観点から，快適な住空間の実現のための音声インタフェース開発を進めている．実世界における音環境の特徴を複数の観点から捉えることができる音環境コーパスを基軸に，適応型音声インタフェースの設計支援システムを開発した．システムに搭載したマルチモーダル行動分析機能，音声認識誤りの状況分析機能，時系列行動パターン分析機能が，音声インタフェースのユーザ環境適応デザインに役立つことを示した．

# Adaptive Speech Interface Design for Living Space Utilizing Environment Acoustic Corpus

Kengo Ikeya[†]　Kenichi Shibata[†]
Yoichi Takebayashi[††] Shigeyoshi Kitazawa[†††] and Shinya Kiriyama[†††]

Home appliances and acoustic devices in living space have evolved into diverse and sophisticated applications. The trends toward nuclear families and aging brought us problems for adapting them to various users and their environments. Acoustic environments largely differ among themselves in personal traits and spatial features. From the viewpoint of comprehensive acoustic environment design, we have been developing speech interfaces to make living spaces comfortable. Based on the acoustic environment corpus which allows us to investigate acoustical features in the real world from multiple viewpoints, an adaptive speech interface design support system has constructed. The functions of analyzing user behaviors from multimodal viewpoints, detailed situations for each misrecognition, and time-series behavior patterns have proved to be useful in designing speech interfaces to adapt for each user environment.

## 1. はじめに

家電機器の多様化・高機能化により，ユビキタスホーム環境における機器連携の研究も進展している．たとえば，コモンセンス知を活用したインテリジェントなユーザサポートエージェント ROADIE に関する研究[1], 蓄積されたセンサ情報から家庭で暮らす家族の生活パターンを推測し，生活支援を行う研究[2]などがある．核家族化と高齢化に伴い多様なユーザや利用環境への適応が課題となっている．また，計算機の高速化･大容量化により人間の行動や発言を記録するセンシング技術が高度化している．音声認識は統計モデルをベースに開発されることが多く，計算機の能力向上により，高い認識率を得ることが可能となってきている．そんな中，より高度な音声インタフェース，音声対話システムの開発に向けて利用者の意図を察する音声対話システムの研究[3]や人間の多様な振る舞いに対応した音声 UI[4]など様々な研究が行われている．しかし，利用者や利用環境の多様な状況を考慮し，カスタマイズして実世界に適応して動作する音声インタフェースの開発事例は少ない．適応型の音声インタフェースの開発を行うためには，各利用者の特徴の把握や状況理解を行う必要がある．現状では音声・音響・音楽の環境デザインはそれぞれ個別に検討が行われている．実際の環境ではそれぞれが個別に作用することはなく，相互に影響し合っているため，音環境の包括的なデザインを行うことが求められる．

筆者らは，個人や空間の特性の差異が特に大きい「音」に着目し，音環境の包括的デザインという観点から，快適な住空間の実現のために音声インタフェース開発を進めてきた[5]．これは，気の利いた人間支援システムに不可欠な常識知のモデル化[6]のため，発達段階の子どもの行動映像事例を蓄積し，エビデンスに基づく多視点観察で思考の発達を捉えるコーパスベースの方法論[7]に基づくものである．

音にフォーカスしたこれまでの検討から，実際の住空間で音声インタフェースを利用することを想定し，家電・家具などを配置した実験室を設けて音声インタフェース利用データを蓄積してきた．これらの実験データに対して発話・行動・音声インタフェースの動作などの多様な注釈を付与した「住空間音環境コーパス」を構築した．また，住空間音環境コーパスを基盤として，認識誤りやシステムの動作などの特定場面の抽出，関連事例の検索などを用いることで様々な観点から状況分析が可能な「音環境測定支援システム」を開発した[8]．本稿では音環境コーパスの構築技術と，音環境測定支援システムを用いた分析について述べる．

Vol.2011-SLP-87 No.17
2011/7/23

## 2. 音環境コーパスの構築

これまで筆者らは人間の思考行動モデルの研究として，船舶ブリッジにおける船員の音声コミュニケーション行動コーパスを構築した[9]．また，子どもの行動観察は人間の認知・行動モデルの研究に適していると考え，幼児教室を開催し子どもの行動コーパスを構築してきた．発話・行動・思考・感情などの観点から設計されたコーパスを用いることで，内面的特徴に関わる発達モデルを産出できることを実証した[7].

これらの基盤技術を確立した上で，音環境コーパスを設計した．音環境コーパスでは人間の行動と思考の関係や利用環境を含めて設計を行った．具体的な項目は発話・認識モデル・環境音などの音声認識に直接関わる要素に加え，ユーザの位置・動作・環境状態などの記述項目である．音環境コーパスを構築して分析を行うことで，環境・ユーザのプロファイルを作成し，各ユーザ環境へ適応可能な音声インタフェースの開発を行う．

コーパスに蓄積するデータを取得するために対話型音声インタフェース BalloonNavi$^{TM}$（図 1）に大語彙連続音声認識エンジン Julius[10]を用いた音声認識機能を拡張した音声インタフェースを用いる．BalloonNavi$^{TM}$ は SilverLight 環境の Web 上で動作するアプリケーションである．ユーザはシステムが提示する選択肢を選び，状態遷移を行うことで目的の動作にたどり着く．ユーザの選択に応じて，システムから次の選択肢を提示できるので，対話的な応答が可能である．音声認識を組み合わせることで提示された選択肢に加えて，それ以外の発話から動作を行うことも可能である．また，システムから提示した選択肢によりユーザの発話を誘導しやすいため，音声認識との相性が良いのも利点である．また，実験室として住空間を模した部屋を構築した（図 2）．ソファー，テーブル，TV など一般的なリビングを想定した構成としている．このような実験環境を用いた実環境での利用実験から映像・音声データを蓄積しており，複数の観点から注釈をつけてコーパスを構築している．


図 1：BalloonNavi$^{TM}$

図 2：住空間実験室

## 3. 音環境測定支援システム

次に挙げる特徴を持つ，音環境コーパス構築・分析ツールである音環境測定支援システム（図 3）を開発した．

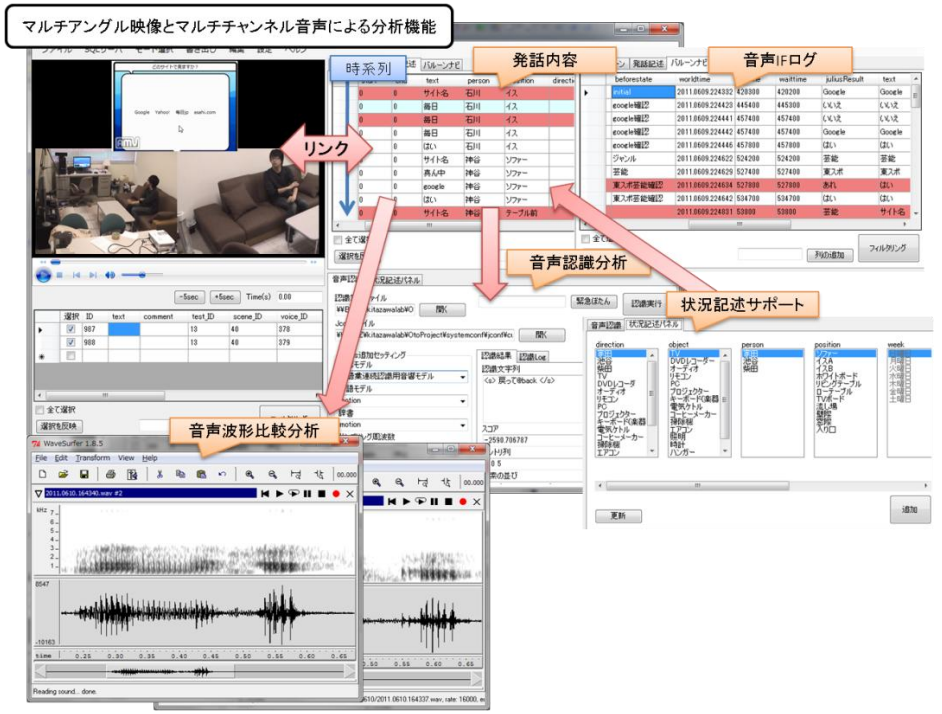


図 3：音環境測定支援システム

**マルチアングル映像とマルチチャンネル音声**

音環境測定実験では複数台 Web カメラを設置しマルチアングルで撮影を行い，音声インタフェースのキャプチャ映像を記録する．これらの映像データに加え，音声認識器で切り出された音声ファイル，実験室の複数箇所に設置されたマルチチャンネルマイクを用いて音声を記録する．マルチチャンネルマイクにより実験室の環境音収録に加え，各設置場所での音声認識時の分析も可能である．このようなマルチソース環境で収録された実験データを活用するため，音環境測定支援システムには多視点観察機能を実装した．映像データを複数表示可能であり，記録された発話内容やシーンの該当箇所に簡単にアクセス可能である．また，選択した発話に関連する音声認識結果や音声インタフェースのログ，収録されたマルチチャンネルマイクの音声を表示する．この機能により，観察者は複数収録された映像，音声を切り替えつつ，気になる分析箇所の比較が可能である．

多視点観察機能を利用することで，単一の映像，音声だけでは分かりづらいユーザの感情や意図などを表情やしぐさから分析可能である．また，音声インタフェースの状況などを分析しやすく，それに伴うユーザの動作と関連して状況理解を深めることができる．

**認識誤りの区別**

音声インタフェースに認識誤りはつきものである．文法認識をベースに音声認識を行う音声インタフェースにとって,認識結果で重要なのは発話に含まれる重要単語(キーワード）である．ユーザが発話に含めたキーワードを認識器が認識できなかった場合が認識誤りとなり，次のパターンに分類される．

- 置換誤り：別のキーワードとして認識した場合
- 挿入誤り：未発話のキーワードを認識した場合
- 脱落誤り：発話したキーワードを認識できなかった場合

これらの分類はユーザの発話と音声インタフェースの認識結果を用いることで，半自動的に行うことが可能である．置換誤りは，音声認識器の認識結果と，ユーザの発話内容を比較することで判別が可能である．また，認識ログが存在するが，それに該当するユーザ発話が存在しないとき，挿入誤りと判別可能である．逆に脱落誤りではユーザ発話が存在するが，それに該当する認識ログは存在しないことから判別可能である．

また，これまで蓄積してきた音声インタフェース利用における認識誤りの傾向から，次のような分類から誤り原因を分類している.誤りが発生した原因を分類することで，後に詳細な分析を行う際に役立つ.

- 未知語
- 似た音素の単語
- 同じ文字数
- 先頭と末尾の母音が一致
- 他者の発話
- システムの応答を認識
- 環境音

また，音声認識の波形レベルでの分析を行うため，Julius，Wavesurfer[11]との連携機能を実装した.音声にたいしてパラメータを様々に変更して音声認識を実行できる．認識結果のスコア，音素区間分割などの Julius 認識結果を Wavesurfer で利用可能なデータ形式へと変換して，波形分析へと利用可能な機能を実装した．これらの認識誤り分類，原因分類，波形分析機能により，認識結果のみに限らず様々な観点から認識誤りについて分析を行うことが可能である．

**ユーザ発話と音声インタフェースログの時系列表示**

ユーザの発話とそれに伴う音声インタフェースの認識結果，状態遷移状態やユーザの感情，意図，発話内容の変化などを時系列表示する機能を実装した．

ユーザの意図を一つの区切りとし，ターンとしてカウントする．ユーザがある意図を持って発話し，その発話の意図が通ったか通ってないかに関わらず，ユーザの意図が切り替わった時にターン数は増加する．これとは別に，ユーザの意図が通ったか通ってないかは情報を付与しておく．

音声インタフェースを利用している間にもユーザの感情や意図は変化している．ユーザが音声インタフェースの利用を終えた後に，音声インタフェースに対する感想や目的の達成度を調査するのでは不十分である．状況理解を深めるに当たり，ユーザの発話，音声インタフェースの動作によって，ユーザの感情や意図がどのように変化したのか分析を行うことが必要である．ユーザの発話と音声インタフェースの動作を時系列的に閲覧し，分析を行うことでユーザの発話の流れ，音声インタフェースの状態を考慮した分析などの基本的な分析に加え，音声インタフェース利用時におけるユーザの意図がどのように変化したのか分析可能である．音声インタフェースの利用に対するユーザの印象，ユーザの要求が満たされたのかどうかなどの他，音声インタフェースをデザインするにあたりサービス内容の検討が深められる．

## 4. 音環境コーパスの実践的評価

### 4.1 多視点観察基づく分析

映像と音声を切り替えつつ，ユーザの表情やしぐさ，発話内容の調子を観察しながらシステムに対する印象を注釈として付与した．各ユーザに音声インタフェースを利用してもらい，そのデータをまとめたものを示す(図 4)．音声インタフェースに対して positive な親切，自然と，negative な不満，苛立ちの 4 項目に分類し，ユーザ毎にどのような印象を持つのか分析を行った．ユーザ A,B,C,D が比較的 positive な要素が多く，音声インタフェースを利用することに対して好印象であることが分かる．ユーザ E の音声インタフェースに対する印象は不満，苛立ちといった negative な要素が多いことが明らかである．ユーザ E がこのような印象を持った原因を分析するため，多視点観察機能を利用し，これまでの実験データを多方面から観察を行った．分析を行った結果，ユーザの意図を最終的に決定する発話に関して認識率が低いことが分かった．分析結果については，後述する各機能を連携して分析を行った例として 4.4 で詳細に述べる．

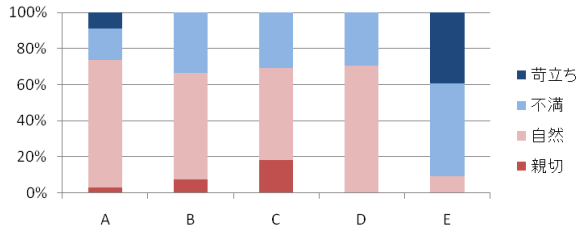


図 4：ユーザの音声インタフェースに対する印象

### 4.2 認識誤り発話の詳細分析

認識誤りが発生している箇所について分析を行った．最も簡単な分析方法として，発話内容と認識結果を比較する方法が挙げられる．発話内容を書き起こし，音声認識器の認識結果と照らし合わせることで誤り原因を探す．しかし，この方法では簡単な原因しか分析を行うことができない．認識結果の文字列から原因が分からない場合，さらに詳細に分析を行う必要がある．音環境測定支援システムには発話内容を波形レベルで分析可能であり，この機能により詳細な原因分析が可能である．

ユーザが「スポーツ」と発話を行ったが，音声認識器が「あの」と誤認識した事例を用いて音環境測定支援システムの認識誤り分析機能の利用例を示す（図 5）．まず分析を行いたい発話を選択し，Julius を用いて音声を再認識させることで音素区間の検出を行う．波形分析では Wavesurfer を用いる．Julius の-palign オプションによる音素ラベルデータを音環境測定支援システムから生成し，正解ラベルとの比較分析が容易に可能である．図 5 に分析例を示す．ユーザの発話内容「スポーツ」は子音[sp]が続き，ユーザの発話音量も小さいことが分かる．それに続く母音[o]は音量が大きいことが見て取れる．ここで，誤認識結果の「あの」の音素と比較すると，母音[o]との一致区間が長いことが分かる．この結果，認識辞書に含まれている単語に最も似ていたため，「あの」と認識が行われたことが分析できる．

認識誤り分析機能により，認識結果の文字列だけでは誤認識の原因が分かりづらい事例に対し，波形レベルで分析を行うことで誤認識結果を分析可能である．


図 5：認識誤り発話の波形分析例

### 4.3 時系列行動パターン分析

ユーザの発話，音声インタフェースの動作によって，ユーザの感情や意図がどのように変化するのか，実例を用いて分析を行った．感情や意図の変化が現れやすい事例を探すため，誤認識を繰り返した時にユーザがどのような対応をとるか分析を行った．ユーザの対応によって，語彙を変えた，ゆっくり発話した，音量を上げた，イントネーションを変えた，繰り返した，の 5 つに分類しユーザ毎にまとめたものを示す（図 6）．

ユーザ A は語彙を変える，音量を上げるといった短調な対応をとっていることが分かる．しかし，ユーザ C は対応が 5 つに分散され，多様な対応をとっていることが分かり，特徴的な事例が発見できる可能性が考えられる．この方針より，ユーザ C が音

声インタフェースを利用する際に，認識誤りが発生した事例の一部を時系列的に表したものを示す（図 7）．

ユーザの発話を時系列的に見ると，「サイト名」という発話がうまく認識されず，ユーザは意図を変え，「毎日」という発話を行っていることが分かる．この「毎日」というのは，インターネットサイト毎日.jp を指している略語である．しかし，これも期待通りに認識せず，ユーザは繰り返し，音量を上げるという対応をとっている．誤認識に対して柔軟に対応していることが確認できる．また，ユーザの感情を時系列的に分析すると，システムに意図が通らない場合に negative な感情を持っていることが分かる．ユーザ C に限らず，ユーザの意図が連続して通らない場合には音声インタフェースへの印象が悪くなる．しかし，軽度の誤り程度ならば，次の認識時に印象はある程度回復することが分かった．

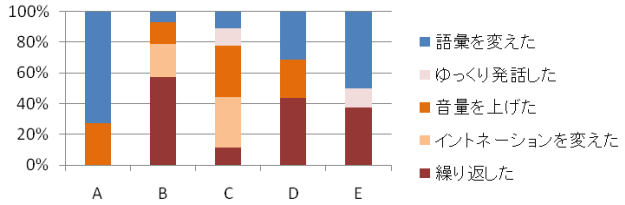


図 6：ユーザによる誤認識時の対応の違い

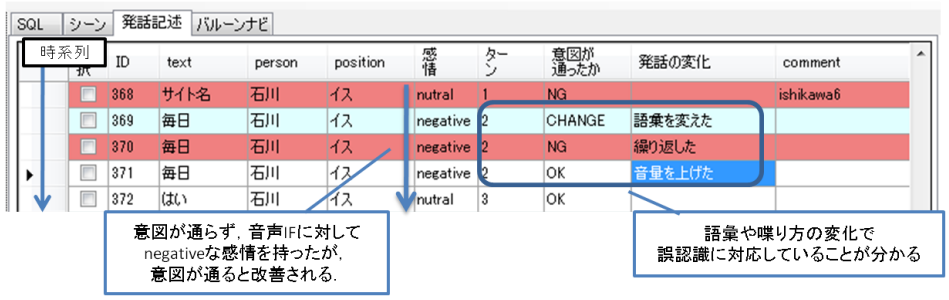


図 7：ユーザ C の誤認識時の対応例

### 4.4 各機能を用いた分析

これまで 4 章で述べてきた各機能を連携してユーザの特徴分析を行う．例として 4.1 で述べたユーザ E が音声インタフェースに対して negative な印象を持っている原因分析の詳細を取り上げる．

これまで実験で蓄積したデータをもとに，ユーザの発話と音声インタフェースの認識誤りの観点から分析を行った．各発話内容について認識率を比較すると，意思決定の意味を持つ「はい」という発話に関して認識率が低いことが明らかとなり，それが原因でユーザは音声インタフェースに悪い印象を持っている可能性があることが明らかとなった(図 8)．これらの発話音声に対して，波形レベルでの分析を行った．1 回目の誤認識の「はい」と正常認識の「はい」の波形を比較したものを図 9 に示す．各音声について改めて音声認識を行い，その音素分割と手作業によるラベルファイルを作成した．誤認識だった「はい」は，先頭の子音である[h]の発声が小さく，「a i」のように聞こえていたことが分かった．認識辞書に含まれている単語の中で，最も似ていた「あの[a n o]」と認識されていることが確認できた．ほかの 2 回目，3 回目の「はい」に関しても同様の傾向が得られ，ユーザ E は「はい」という発話の際に[h]を小さく発声することが分析できた．この分析により，ユーザ E の適応例として辞書ファイルの変更や，音響モデル学習の方針を立てることができた．

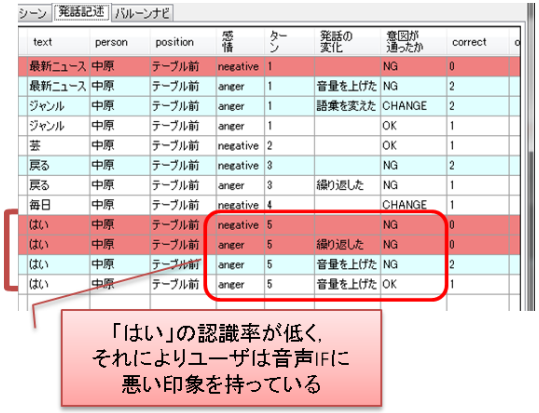


図 8：ユーザ E の発話内容と認識状態


図 9：「はい」に対する波形分析

## 5. まとめ

ユーザや利用空間に適応可能な音声インタフェースの開発について，コーパスによる問題解決方針について述べた．音環境コーパスの構築技術と音環境測定支援システムを用いた分析を行った．実環境を想定した実験室を用意し，ユーザに音声インタフェースを利用してもらうことでデータを蓄積し，音環境コーパスを構築した．構築した音環境コーパスを音環境測定支援システムの各機能を用いることで，ユーザの意図や感情，音声インタフェースの振る舞いなどを分析できることを具体例と共に示した．また，各機能を連携して用いることによりユーザごとの特徴を抽出し，適応型音声インタフェースにおけるユーザ適応方針を検討できることを示した．

今後はさらに実験データを増やし，音環境測定支援システムの機能拡張に加え，状況理解の常識知のモデル化を検討していきたい．

## 参 考 文 献

1 ) Henry Lieberman, José Espinosa : A goal-oriented interface to consumer electronics using planning
2 ) 美濃導彦: ユビキタスホームにおける生活支援， 人工知能学会誌， Vol.20， No.5, pp.579-586， (2005).
3 ) 翠輝久，堀智織，香山健太郎，大竹清敬，小林亮博，水上悦雄，柏岡秀紀，河井恒，中村哲，ユーザの意図を察する音声対話システム，人工知能学会研究会資料，Vol.SIG-SLUD-B200，39-42, (2010)
4 ) 岡本淳，庄境誠，人間の多様な振る舞いを考慮した音声 UI の必要性，情報処理学会研究報告，Vol.2009-SLP-78，No.10，(2009)
5 ) 桐山伸也，本間永愛，石川翔吾，秋川友宏，竹林洋一，"ユーザ環境に適応した住空間音声インタフェースの検討，" 音講論，3-1-16, 2009-9.
6 ) 竹林洋一，桐山伸也，"工学的視点からの幼児の行動観察とコーパス構築―認知・行動モデルの深化がもたらすもの―，" 日本音響学会誌，vol.65, no.10, pp.544-549, 2009-10.
7 ) 石川翔吾，桐山伸也，大谷尚史，北澤茂良，竹林洋一： マルチモーダル幼児行動コーパスに基づく指示表現の発達分析とモデル構築，チャイルド・サイエンス，vol.5, pp.68-72, (2009).
8 ) 柴田健一，池谷謙吾，立蔵洋介，北澤茂良，竹林洋一，桐山伸也，住空間サービス実現に向けた音環境測定支援システム，日本音響学会 春季研究発表会，2-P-55(a)，(2011)
9 ) 桐山伸也，鈴木敦志，青島大悟，本間永愛，竹林洋一，安全航行のための船員の音声コミュニケーション分析，音講論，1-1-22, (2008).
10 ) 大語彙連続音声認識エンジン Julius: http://julius.sourceforge.jp/
11 ) Wavesurfer:: http://sourceforge.net/projects/wavesurfer/