実環境福祉ロボット用音声認識システムの設計

Design of Speech Recognition System for Welfare Robot in Real Environment

中込　亮介
Ryosuke Nakagome
法政大学情報科学部ディジタルメディア学科
E-mail: ryosuke.nakagome.cw@cis.hosei.ac.jp

Abstract

Recently, development of a supplementary robot of a guidance person of nursing rehabilitation physical exercise is advanced. Speech recognition technology among communication functions was developed. In this research, it aimed to recognize word that user intended, and to return the pertinent reply. And it examined it. The mike used speaker phone and selected it from recognition rate, frequency, and cost. The number of recognition words prepared 98 words. The recognition word gave some flexibility. Grammar is made for task domain. Moreover, this robot is used on a real environment. Therefore, voice of five men and women were recorded. They are in their sixties. And, voices were evaluated from experiment that Rosenberg had advocated. As a result, obtaining the recognition rate of 90% needed to make dictionary which have 36 words.

1 まえがき

ロボットの開発は進化していて，様々な用途を持ったロボットが出始めている．その一環として，発達が著しい音声認識システムとの融合である，ロボットとの音声認識技術を用いた対話システムも普及し始めている．

今日までに，様々な音声認識技術を用いたコミュニケーション機能の開発が進められているロボットが存在する．パートナーロボット呼ばれる次世代のロボットで，「PAPERO[1][2]」は会話や散歩をするホームロボット，「ifbot(イフボット)[1][3]」は感情を入れた言葉で対話をする会話ロボットとして開発されている．これらの仕様としては，「PAPERO」は約 650 種類の音声を認識し，約 3000 種類の言葉を話すことができる．「ifbot」は，5 歳児並みの知能を持ち，相手の言葉と同時に感情も判断することで，目やまぶた，口の動きなどを駆使した 40 種類の喜怒哀楽を表現した言葉を返すことができる．

現在，介護リハビリ体操の指導士の補助として，老人ホームなどで体操のお手本を見せるような福祉ロボット「たいぞう」(図 1)の開発が進んでおり，音声認識技術を核としたロボットとのコミュニケーション機能を搭載する予定である．

本研究では，このコミュニケーション機能のための音声認識システムを研究する．本ロボットの体操機能を持つプロトタイプの開発はほぼ完成しており，去年の 11 月に茨城で行われた全国健康福祉祭「ねんりんピック」では実際にデモも行われた．しかし，現段階では体操を行うことしかできない状況であり，音声認識技術による音声で日常会話をする機能の開発が進められている．

現時点，このロボットと体操を行う高齢者は，前でロボットが動いているという印象しか受けないように思われる．しかし，最終的に高齢者が気兼ねなく世間話程度の話ができ，なおかつ孫に話しかけるかのような，単なるロボットではない，人間のような存在になることを目指す．

そのために，本ロボットに適した，対象を高齢者を中心とした不特定話者のためのシステム開発を行う．

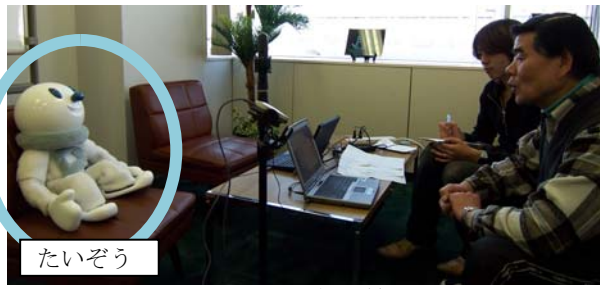

図 1　収録風景

2 ロボット用音声認識技術の現状

ロボットにおける音声認識機能の開発では，一般的に多くのやるべきことがある．特に実環境上であると，雑音や音声区間検出などの問題が出てくる[4]．次に，現在行われている一般的な研究について述べる．

(1)雑音除去

音声認識において，雑音による影響は大きいとされている．そこで，これらを除去する研究が盛んに行われている．信号処理段階での方法としては，①対象音源を強調する②雑音成分を除去する(スペクトルサブトラクション法など)③目的音の抽出などがある．また，音声認識における対策として④低周波成分カット⑤音響モデルを雑音混じりに対応などがある．

(2)マイクロフォン選定

認識精度の向上には，マイクの選定も重要となってくる．マイクロフォンによっても入力周波数域や性能などがことなってくるためである．例として，「PAPERO」では胴体正面に 4 個，左右にそれぞれ 1 個，背面に 1 個の全指向性マイクロフォン及び頭部に 1 個の単一指向性マイクロフォン，「ifbot」では音声認識マイクと音声検知マイクをそれぞれ 1 個搭載している．

Supervisor: Prof. Katunobu Itou

(3)音響的要因への対策
雑音環境や収録系が音響モデルの学習データと異なる場合，認識精度が劣化する．また，不特定話者音響モデルでは，学習データの話者の平均から外れるような話者の場合，認識率が著しく劣化する場合がある．このような場合は，音響モデルを対象データに対して適応する必要性がある．
(4)言語的要因への対策
自然な発話には書き言葉に比べて話速の変化や特有の表現，未知語など多くの問題が存在する．単一モデルでは，これらすべてのカバーは難しい．よって認識精度を上げるにはシステムごとに現れうる語彙や表現を過不足なくカバーしたタスクドメイン用の文法の作成し，文候補を言語的に絞り込むことが重要である．

3 本研究での開発計画

本研究でのこのロボットは，約 80cm の身長のロボットであり，体操の腕の可動範囲が約半径 40cm である．マザーボードにはサウンドカードがなく USB 端子の搭載しか想定されていない．よって，USB 端子の数の制約からスピーカーフォンを使用することとした．また，認識距離は人がロボットに手を伸ばし，触れることができる程度の距離を想定する．
本研究では，本ロボットの技術開発に対して以下のような研究を行った．なお，雑音影響に関しては今回の想定は老人ホームの体操の場面などであり，認識の距離は近く雑音の影響があまりないと考えられるので，今回の計画からは外した．

3.1 スピーカーフォン評価

本研究では先にも述べた通り，一般的に使用されるマイクロフォンではなくスピーカーフォンを使用する．今回は次の 3 つのスピーカーフォン(図 2)が候補としてあがった．

・miniVox MV100
Portable USB Speakerphone (以下 mvox)
・ELECOM Skype 会議用ハンズフリーフォン
MS-CO95USV (以下 elecom)
・Polycom Skype 専用ハンズフリースピーカーフォン
Polycom communicator(以下 polycom)

それをスペック(仕様・機能など)，周波数特性，認識精度，コスト面から評価し，選定する．

図 2　スピーカーフォン候補
(左から mvox，elecom，polycom)

3.1.1 周波数特性の評価実験

マイクの周波数特性を評価する場合，各周波数特性のフラットなインパルス応答を用いる．しかし，インパルスは一瞬の音なので，雑音などの影響を受けやすいため，それを録音することが困難である．そこで，TSP を用いる．
TSP(Time-Stretched Pulse)とはインパルスのエネルギーを時間軸上に分散させた波形であり，TSP 応答を録音後，TSP 逆フィルタと畳み込み演算をし，分散したエネルギーを元に戻したインパルス応答を得る．
本実験では，0～8kHz までの周波数成分を含んだ TSP をスピーカー(TDK フラットパネル・アコースティックサウンドシステム「Xa-10」)から再生し，3 つのスピーカーフォンと通常のマイク(SONY DYNAMIC MICROPHONE F-V320)で，サンプリング周波数 16kHz，離散化ビット 16bit，モノラルで同時に録音する．距離は 10cm．そして，それぞれをインパルス応答に再構築し，FFT して求めたパワースペクトルを対数パワースペクトルに変換し，その図から周波数特性の比較を行う．

3.1.2 認識実験

それぞれのマイクと比較用のマイクでの音声認識性能の違いがあるかを調べる実験を行う．
方法としては，各マイクと比較用のマイクで同じ文章を同時に録音し，それぞれの音声データをフリーの音声認識ソフトJulius-v3.0[1]で認識させる．
録音する文章は，政治やスポーツ，芸能などの様々なジャンルのニュースから，適当な長さの 10 文を選んだ．
評価には単語正解精度[5]を使用した．

3.2 タスクドメイン用文法の構築

本研究では老人ホームなどの福祉施設でのコミュニケーション促進のための対話を想定しているため，タスクドメイン用の文法を作成することが必要となる．

3.3 想定性能の評価

実環境データを使用した認識実験を行う．本研究では実際に収録を行い，それをタスクドメイン用の文法を構築した記述文法音声認識実行キット Julian での認識を行い，結果を評価した．

3.3.1 実環境データの収録

今回収録は，つくばシルバーリハビリ指導士会の方にご協力を頂き，実環境で収録を行った．対象者は 60 代の男女計 19 名．用意した認識語 98 語をロボットに話しかけて頂き，その音声を収録した(図 1)．使用スピーカーフォンは polycom．また，比較実験用に mvox でも収録した．一人当たり 2 回収録した．
部屋は会議室を使用し，個人がロボットに話しかける距離 60~80cm 位にマイクを立てて行った．
収録したデータは，サンプリング周波数 16kHz，16bit，モノラルの raw データである．これを，波形編集ソフトを使用し音声部分のみを切り出した．

[1] Julius: http://julius.sourceforge.jp/index.php?q=juliuskit.html

3.3.2 辞書サイズの評価実験

安定して音声対話を遂行するには，一定の音声認識性能が必要である．本研究では期待正解率という尺度を用いて辞書サイズと認識性能を評価する．

Rosenberg[6]は孤立単語認識において，平均順位の分布が混合 2 項分布に従うなら，エラー率が辞書サイズの関数になるという確率モデルを提唱した．

このモデルでは，正解単語以外の単語のスコアが高くなる確率を定め，単語音声認識をシステム辞書に登録された単語数分のベルヌーイ試行とみなす．この仮定に基づくと，ある発話においての正解単語の順位が 2 項分布となる．単語の認識の難しさが複数ありえる場合は，混合 2 項分布となる．このモデルを用いると，
単語誤り率は，

$$E\{E_v\}=1-\sum_{m=1}^{M}h_m(1-p_m)^{N-1},\quad \overline{P}_v=\sum_{m=1}^{M}h_m p_m \qquad (1)$$

と定義される．

実験方法としては任意に定めた辞書サイズに対して，用意した認識語からランダムに選択する．それぞれの辞書に対しての各被験者の平均順位と標準エラー率を求め，被験者全員の平均から理論値を導出する．

4 スピーカーフォンの評価結果

4.1 スペック・価格例による評価

それぞれのスピーカーフォンのスペックを表 1 に示す．

mvox は，本音声認識システムで必要とされるサンプリング周波数 16kHz での録音ができないことが欠点であるといえる．しかし，コスト面は低価格である．

elecom は，やはり Echo Cancellation 機能がないことが 1 番の欠点である(常にマイクからの音がスピーカーから出力され，さらに，それがマイクに入力されてしまうため，録音すると Delay のかかったような音声になってしまう)．本ロボットでは，常にどちらの機能もなくてはならないので，この問題は大きいと考えられる．

polycom は，サンプリング周波数も 16kHz で録音が可能で，3 つの中で 1 番性能が良いと思われる．しかし，一番高価でコストが最もかかる．

表 1　スペック・価格

	mvox	elecom	polycom
入力	サンプリング周波数 8kHz	25Hz~16kHz	200Hz~22kHz
価格	4,980 円	7,033 円	16,756 円
主な機能・特徴	Echo Cancellation		Echo Cancellation

4.2 周波数特性による評価

それぞれのスピーカーフォンと評価用のマイクの周波数特性を図 3 に示す．

mvox は，自動的にローパスフィルタがかかり，周波数 4kHz 以上をカットすることがわかる．

elecom の入力は 16kHz であるが，結果はみてのとおり mvox と似たような波形となってしまっている．明らかにローパスフィルタがかかっているように思われる．

polycom は，全体的には通常のマイクとあまり変わらない結果であるといえる．しかし，このマイクはパワーが全体的に低めとなっている．

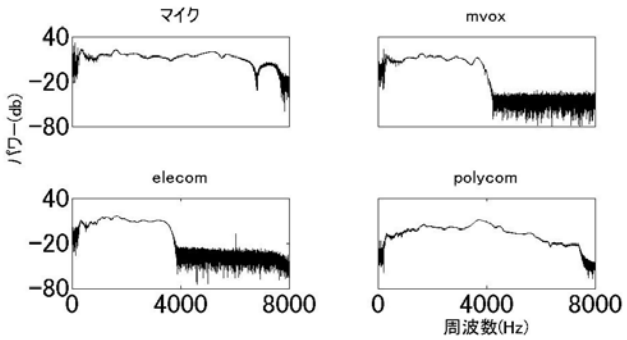

図 3　周波数特性

4.3 認識実験による評価

それぞれと通常のマイクの認識精度を比較した(表 2).

表 2　認識実験結果

	平均単語正解精度
mvox:比較マイク	0.673 : 0.810
Elecom:比較マイク	0.609 : 0.879
polycom:比較マイク	0.841 : 0.850

mvox は，入力のサンプリング周波数が半分であることが，正解精度に影響を及ぼしたと思われる．

elecom は，mvox より低い結果となっている．本スピーカーフォンのサンプリング周波数は 16kHz なので特性による影響が考えられる．

polycom は，かなり高い認識率を得ることができた．文によってはマイクより高いものもあった．

4.4 評価の考察と結果

それぞれの評価実験から，様々な長所短所が考察された(表 3)．考察の結果，本実験ではコスト面に問題が残るが，認識率を重視して polycom を使用することとする．

表 3　スピーカーフォンの長所短所

	長所	短所
mvox	・コストパフォーマンスがよい	・サンプリング周波数に問題あり
elecom	・コストパフォーマンスがよい	・認識率が悪い ・エコーキャンセレイションがない
polycom	・認識率がよい	・コストが高い

5 文法構築の詳細

本研究では，記述文法音声認識実行キット Julian を使用し，認識辞書の文法を作成する[7]．認識語は 98 語．それぞれに適切な応答をロボットが行う．今回は，適切な応答をする結果の語(帰結語)のさまざまなバリエーションである認識語を，ある程度の柔軟性を持たせ考慮した．また，認識語はすべて単語として辞書に登録するのではな

く，単語と文法規則に分けて作成し，汎用性を持たせた．結果，文法規則は 67 個，辞書サイズは 84 個である．

<認識の流れの例>

認識語 : 何歳?or 何歳ですか?or いくつ?or いくつですか?
↓
帰結語 : 何歳?
↓
応答例 : やっつ

6 想定性能の評価実験

6.1 実験

3.3.1 で述べた音声データを使用し 3.3.2 で述べた手法を用いて本タスクにおける想定性能を評価した．

本実験では，辞書サイズを 2，5，10，20，40，80，98 とし，それぞれの辞書サイズに合わせた認識語は用意した 98 フレーズの認識語からランダムに選択した．各サイズの辞書に含まれるエントリの音声データを用いて，Julian を使用して認識を行った．その結果を用いて次の 2 点について評価・考察する．

(1)標準エラー率を算出し，被験者男女各 5 名の平均の値から理論値を求める．結果から期待正解度を導出する．

(2)また，2 つのスピーカーフォンを，標準エラー率の理論曲線から評価する．

6.2 実験評価・考察

(1)の結果を図 4 に示す．結果は，ほぼ Rosenberg の述べた関係となった．理論曲線のパラメータは表 4 のようになり，これを(1)式に当てはめることでエラー率に対する辞書サイズの理論値が求まる．今回の結果から辞書サイズと認識率の関係は，辞書サイズ 132 語のときエラー率 20%，66 語で 15%，36 語で 10%となる．

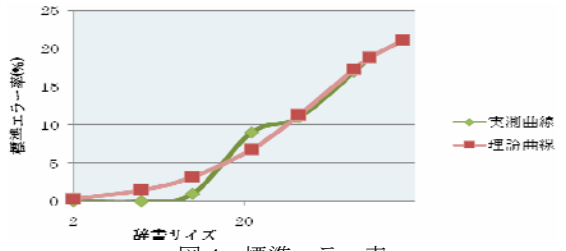

図 4　標準エラー率

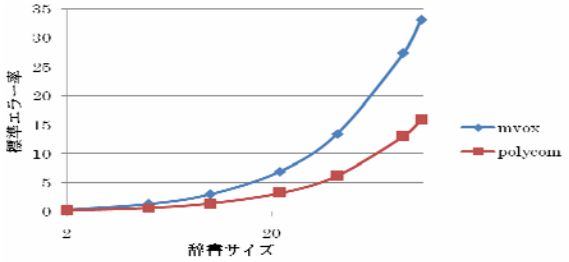

図 5　スピーカーフォン比較

表 4　理論曲線のパラメータ

h1	h2	h3	p1	p2	p3
0	0.021	0.979	0.017157	0.017157	6.85E-07

(2)の結果を図 5 に示す．結果はエラー率 15%を得る辞書サイズは mvox47 語に対し polycom102 語，10%のときは 32 語に対し 68 語，5%のときは 16 語に対し 34 語となった．結果として，polycom は mvox の約 2 倍の辞書サイズを得ることができることがわかる．

7 むすび

本研究では，実環境におけるロボットの音声認識システム確立のために必要な検討を行った．

マイクは polycom を使用することで通常のマイクと変わらない認識率を得ることができた．また，mvox でも音声認識機能をサンプリング周波数 8kHz に対応することができれば認識率が向上し，本ロボットに適する可能性があった．

文法構築においても，認識語から様々な柔軟性をもったものが作成でき，認識実験においても認識率を向上することができた．

また，実環境で収録した評価用データを認識し，エラー率から期待正解率を算出する本実験から，90%という認識率を得る辞書サイズは 36 語であった．これで最低限の音声認識対話を実現することができる[8].

今後の課題としては，マイク選定では，認識率が良いことが最も重要ではあるが，コスト面なども考慮にいれることが開発には重要である．

認識語は，期待正解率に対する認識語の影響を考慮する必要がある．

今後の展望として，本機能の実用化にはさらなる評価データによる認識実験や実環境コーパス作成による他コーパスとの評価，音響モデル構築・評価，認識間違い時の対処，雑音対処，実装・実験等まだまだやらなければならないことがある．

8 謝辞

本研究を行うに当たりご指導・貴重な御意見を下さった JST 藤井洋之氏，また音声収録に御協力頂いたつくばシルバーリハビリ指導士会の方々に心から感謝致します．

文　献

[1]”ロボタリゼーションが日本を活性化する”
経済界,Vol.38,No.22,pp56-63,November 2003

[2]藤田”チャイルドケアロボット PaPeRo”
日本ロボット学会誌,Vol.24,No.2,pp162-163,March 2006

[3]木村憲次”ロボットで拓く②”
週刊東洋経済,No.5891,臨時増刊,pp5,May 2004

[4]李ほか”ロボットにおける音声認識技術”
計測と技術 第 46 巻 第 6 号，pp441-446

[5]堀ほか”単語抽出による音声要約文生成法とその評価”
信学論 D-II Vol.J85-D-II No.2 pp.200-209 Feb2002

[6]A.E.Rosenberg “A Probabilistic Model for the Performance of Word Recognizers” AT&T BELL LABORATORIES TECHNICAL JOURNAL Vol.63, NO.1, 1984

[7]李ほか“記述文法に基づく高性能連続音声認識エンジン Julian”音響 講演論，2001 年 10 月，pp111-112

[8]佐藤ほか “パーソナルロボット PaPeRo における音声インターフェース”日本音響学会誌 62 巻 3 号,pp173-181,2006