

デジタルサイネージとしての ロボット音声対話システムの開発

山本 一公

Development of Robot Chat Spoken Dialogue System as Digital Signage

Kazumasa Yamamoto

We are developing a spoken dialogue system with two robots that talk to the user during chatting between the robots for natural information acquisition. In this paper, we developed a framework of the whole system as a first step of this research.

Keywords : *Spoken dialogue system, Robot, Chi-Chat, Distant speech recognition*

1. はじめに

近年の音声認識技術の発展に伴い、「AI スピーカー」「音声 AI アシスタント」として音声対話システムの実用化が進んできている。これらの代表的なものとして、Google 社の「Google Home」「Google アシスタント」、Apple 社の「Apple HomePod」「Siri」、Amazon 社の「Amazon Echo」が挙げられ、徐々に活用場を広げている。一方で、これらの製品における対話相手はスマートホンやスピーカー（多くは円筒形）であり、人間同士の対話とはかなり異なるスタイルのものである。またこれらの製品は、ユーザの質問に対してシステムが答えるものであり、且つ文脈の情報はほとんど利用されないことから、基本的に「1 問 1 答対話」と呼ばれる形式の対話しか行えず、複雑な対話を行うことは難しい。

現在の音声対話システム研究において、考慮すべき大きなポイントが 2 点ある。1 点目は、システムが扱う対話が「タスク指向対話 (Task-Oriented Dialogue)」なのか「雑談対話 (Chit-Chat)」なのかという点である。従来の音声対話システムの研究は、ユーザが解決すべき問題（タスク）を持っており、その問題を解決するためのシステムであった。具体的なタスクとしては、航空券予約タスク¹⁾、レストラン検索タスク²⁾等がタスクとして設定され、それを解決するための対話システムとしてシステムが構築されてきた。AI スピーカーとの対話も基本的にタスク指向対話である。タスクがはっきりしている対話システムの場合、問題を解決するためにユーザから獲得しなければならない情報もはっきりしているため、基本的にはその情報を順番にシステムから質問し、ユーザがそれに答えること（スロットフィリング）³⁾を繰り返し、全ての情報スロットが埋まった段階でシステムがそれに応じた回答をユーザに行うことでタスクを達成することが可能である。この種の対話システムは、対話制御の理論の進歩等と合わせて、精度良く動作することができるようになっている⁴⁾。一方、最近の対話研究のトピックとして盛んに研究が行われるようになってきたのが「雑談対話」である。ユーザが特に解決すべき問題を持っていない、

例えば最近の天気の良いし悪し等、一般に「雑談」と呼ばれる種類の対話をどのようにシステムに行わせるか、その場合にシステムが扱える話題の範囲から逸脱しないようにどのように話題の制御を行うべきか、どのようにシステムが応答すると雑談として自然だとユーザは感じるのか、というようなことが研究されている⁵⁾。さらに、人間同士の対話であれば、雑談と言っても無意味なものではなく、その中にタスク指向的な情報のやり取りも含まれることから、タスク指向対話と雑談対話の融合を目指して研究が進められている。考慮すべき2点目は、音声対話で話しかける対象である。システムがどのような形態であれば人間が話しやすいと感じるかという研究は古くから行われており、単なる情報画面よりも擬人化エージェントが映し出されている方が話しやすく、擬人化エージェントよりも実体を持つロボットの方が話しかけやすいと感じられるという実験結果が我々の研究からも得られている⁶⁾。ロボットにもロボット然とした形態を持つものから、より人間に近づけた外観のものまで様々な種類があり、その外観によってロボットから受け取る情報の信頼性や対話のカジュアルさ等が変化すると考えられることから、用途に応じて様々なセッティングを行う必要があると考えられる。

筆者の所属してきた研究室においても従前より、富士山周辺の観光名所・施設案内を行うことを想定したタスク指向対話システムである「富士山観光案内システム」⁷⁾、人間同士の雑談対話中にて生じる種々の対話現象（あいづち、復唱、共同補間等）をシステムが模倣することで人間同士の対話のようなリズムを生み出すことを目指した雑談対話システムである「天気情報雑談音声対話システム」⁸⁾、複数の対話エージェントを置くことで話題を広げてより楽しい雑談を行うことを目指した「複数エージェント雑談音声対話システム」⁹⁾の開発を行ってきた。

本研究では、これまで筆者が関わってきた音声対話システム開発の研究を受けて、デジタルサイネージとして利用可能な音声対話システムの構築を目的とする。このタイプの音声対話システムの典型的な例として、名古屋工業大学の正門に設置されている「メイちゃん」が挙げられる¹⁰⁾。メイちゃんは画面に表示されている擬人化エージェントであり、画面前に設置されているマイクを通じて話しかけることで大学の情報等を教えてくれるタスク指向音声対話システムになっている。画面の前に非常に分かりやすくマイクが設置されているため、「マイクに話しかければ何かが起きそうだ」ということがユーザに伝わりやすいが、逆にマイクが設置されていないならば「話しかけることができるかどうか分からない」という状況になりやすいのがこの種のシステムの問題点となる。この問題点を解決するため、ロボットを2体設置し、ロボット同士での雑談にユーザが割り込んだり、ロボット同士の雑談中にロボットからユーザに話しかけたりすることで、ユーザが自然に対話システムに話しかけることができ、その対話の中から情報を取得できるようなシステムを構築することを本研究では目指していく。本稿では、その研究の最初の段階として、システム全体の枠組みの構築と、音声入力部分である遠隔発話（マイクから離れた位置からの発話）音声認識を頑健に行うための初期評価について述べる。

2. システムの全体構成

システムは大きく分けて、マイクロホンアレイにより遠隔発話音声を受音・処理して音声認識結果を求める「音響分析・音声認識部」、ユーザ発話（音声認識結果）を解析しシステムの内部状態を更新して応答文を生成する「対話管理部」、対話管理部が生成した応答文に合わせてロボットに発話・身振り手振りをさせる「応答出力部」から成る。各部の関係の概略を図1に示す。各部について、以下により詳しく述べていく。

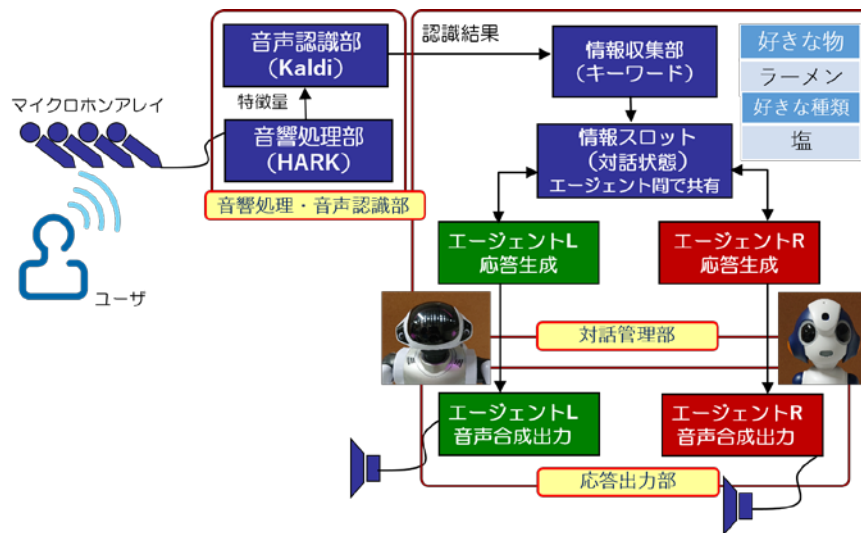


図 1. システムの全体構成

2.1. 音響処理・音声認識部

本研究では、雑音や残響が存在する環境でユーザ（人間）が複数のロボットと対話することを想定している。ユーザの音声はユーザから離れた位置にあるマイクホンによって受信され、その後信号処理部に渡され残響除去等の信号処理が行われるが、この信号処理を高精度に行うためにはシステムが置かれる環境が事前に分かっている必要があるため、大学の実験室の1室を対話システムを設置するための実験環境として固定した。図2に実験室の配置図を示す。

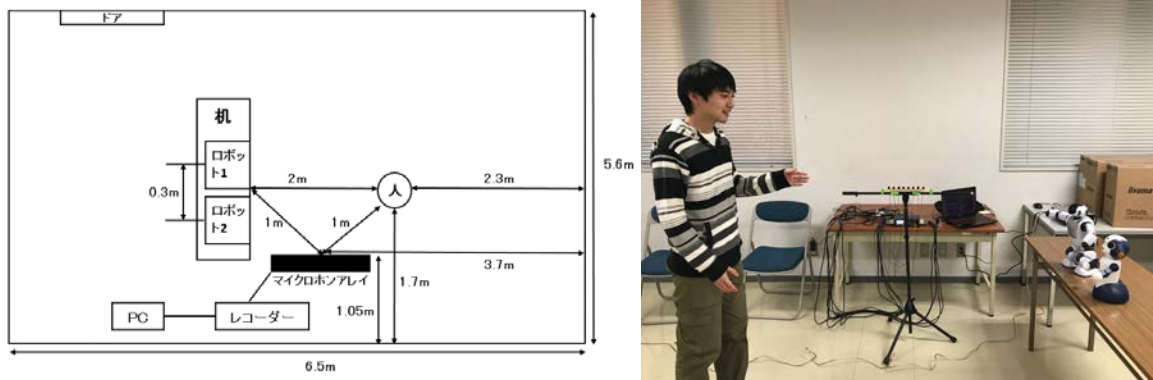


図 2. 実験室内の配置図と実際の対話の様子

2.1.1. HARK

HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) は、「ロボットの耳」を作することを目的に京都大学で開発されたオープンソース・ソフトウェアである¹¹⁾。HARKには、音がどこから来るのかを推定する複数音源定位、特定方向の音に耳を傾ける音源分離、耳を傾けた音を音声認識する同時発話音声認識の機能が含まれている。本研究では、音声認識部の音声受音部分をHARKを用いて構築する。HARKによる音声認識処理の流れはおおまかに、(1) マイクロホンアレイによる音声の受音、(2) 音源定位、(3) 音源分離、(4) 雑音抑圧、(5) 音声認識の順で行われる。

2.1.2. マイクロホンアレイ

マイクロホンアレイとは、複数のマイクロホンを幾何学的に配置したもので、各マイクロホンに入力された音声信号を処理することで単一のマイクロホンでは得られない音の空間的な情報を取得できるようにしたものである。本研究では、各マイクロホンに到達する音声信号の時間差から音声の到来方向を推定することで、人間の音声を強調したり、ロボットの発話を音声認識しないようにするために用いる。

本研究では線形マイクロホンアレイを用いる。線形マイクロホンアレイでは、音声の到来方向として判別できるのは $\pm 90^\circ$ の範囲に限られる（線形マイクロホンアレイの正面を 0° とする。線形マイクロホンアレイの場合、音波形の到達時間差からはマイクロホンアレイの前後を判別できない）が、図2のようにマイクロホンアレイを設置することで、人間の発話かロボットの発話かを到来方向から判別できるようにしている。

2.1.3. 空間の伝達関数の測定

音源定位、音源分離を行う際に空間の伝達関数を利用するため、これを事前に測定しておく必要がある。マイクロホンアレイから見て、 θ 方向にある音源 $S(\theta)$ から i 番目のマイクロホンまでの周波数領域での伝達関数を $h_i(\theta, \omega)$ （ ω は周波数）とすると、マルチチャンネルの伝達関数ベクトルは以下のように表せる。

$$H(\theta, \omega) = [h_1(\theta, \omega), \dots, h_M(\theta, \omega)] \quad (1)$$

ここで M はマイクロホンアレイのチャンネル数である。伝達関数 $h(t)$ はインパルス信号 $\delta(t)$ をマイクロホンで測定することで求めることができるが、実際にはインパルス信号は非常に短い時間に非常に大きなパワーを出す必要があるため、スピーカーから正確に出すことはできない。そのため、インパルス信号の代わりに TSP (Time-Stretched Pulse) 信号を測定する¹²⁾。TSP 信号はインパルス信号を時間軸方向に引き伸ばしたもので、TSP 信号に逆 TSP 信号を畳み込むとインパルス信号になることから、TSP 信号により測定した信号に逆 TSP 信号を畳み込むことで空間の伝達関数を得る。

2.1.4. 音源定位

音源定位部では、マイクロホンアレイで受信された複数チャンネルの音声波形データに対して、MUSIC (Multiple Signal Classification) 法¹³⁾を用いて、マイクロホンアレイ座標系における水平面での音源方向を推定する。MUSIC 法は、どの方向にどのぐらいの大きさの音があるかを推定するアルゴリズムであり、大きなパワーを持つ方向を各分析フレームで検出することで、音源の方向や音源数、発話区間などをある程度知ることが可能である。音声の入力角度を θ 、入力周波数を ω 、入力分析フレームのインデックスを f 、空間の伝達関数を H 、 H のエルミート転置を H^* 、雑音に対する固有ベクトルを e_i とした場合、音声のパワー $P(\theta, \omega, f)$ は

$$P(\theta, \omega, f) = \frac{|H^*(\theta, \omega)H(\theta, \omega)|}{\sum_{i=N+1}^M |H^*(\theta, \omega)e_i(\omega, f)|} \quad (2)$$

として推定できる。ここで、 M はマイクロホンアレイのチャンネル数、 N は音源数である。

2.1.5. 音源分離

音源分離では、GHDSS (Geometric High-order Decorrelation-based Source Separation) 法¹⁴⁾を用いて音源分離処理を行う。 N 個の音源から発せられた音 $S(k_i)$ は、その空間の伝達関数 $H(k_i)$ の影響を受け、 M 個のマイクロホンを通じて、

$$\mathbf{X}(k_i) = \mathbf{H}(k_i)\mathbf{S}(k_i) + \mathbf{N}(k_i) \quad (3)$$

のように観測される．ここで， i はマイクロホンアレイのチャンネル番号， k_i はチャンネル i の周波数ビン， $\mathbf{N}(k_i)$ は各マイクロホンに作用する加法性雑音である．このとき，分離音の複素スペクトル行列 $\mathbf{Y}(k_i)$ は，分離行列 $\mathbf{W}(k_i)$ を用いて，次のように表せる．

$$\mathbf{Y}(k_i) = \mathbf{W}(k_i)\mathbf{X}(k_i) \quad (4)$$

分離行列は，

$$\mathbf{W}(k_i) = \frac{\mathbf{H}^H(k_i)}{M} \quad (5)$$

として初期化され（ここで， $\mathbf{H}^H$ はエルミート作用素である），分離信号 $\mathbf{Y}(k_i)$ の高次相関行列の対角成分以外を 0 にすることと，分離行列 $\mathbf{W}(k_i)$ と直接音の空間の伝達関数 $\mathbf{H}(k_i)$ の積が単位行列になるようにすることを同時に満たすようにフレーム毎の繰り返し更新を用いて推定される．

2.1.6. 雑音抑圧

音源分離した音に対して，スペクトル減算による雑音抑圧を行う．定常雑音スペクトルの推定には HRLE (Histogram-based Recursive Level Estimation) 法¹⁵⁾を用いる．HRLE は，入力スペクトルのレベル頻度分布を計算し，正規化累積頻度分布からノイズレベルを推定する．

ここまでの処理により，音源（人間）方向の信号を強調した音声信号を得ることができる．この後，得られた音声信号から音響特徴量を抽出して，音声認識処理を行う．

2.1.7. 音声認識

音声認識に用いる音響特徴には，MSLS (Mel-Scale Log-Spectrum)を用いる．MSLS を求めるために，まずフレーム単位で切り出された音声信号に対して離散フーリエ変換を行い，複素スペクトルを得る．これの絶対値を取って振幅スペクトル化したものに対して，メルスケールフィルタバンク（周波数軸を人間の対数的な聴覚特性に合わせてメルスケールとし，メル周波数軸上に等間隔に三角フィルタを配置したフィルタバンク）を通して得たフィルタバンクエナジーの対数を取ったものが MSLS である．MSLS 特徴量は，現在のニューラルネットワークベースの音声認識エンジンで広く用いられている特徴量である．

音声認識デコーダでは，DNN-HMM (Deep Neural Network – Hidden Markov Model)音響モデルと N-gram 言語モデルを用いて，WFST (Weighted Finite State Transducer)を用いて単語列を推定する．音響モデル，言語モデルの学習と実際の認識には，Kaldi 音声認識ツールキット¹⁶⁾を用い，HARK と Kaldi 間の接続には HARK で用意されている Unix Socket による通信を用いる．

音声認識部の出力は，音声認識結果（文字列）である．これを対話管理部に送って，対話処理を行う．音声認識部と対話管理部は Unix Socket によって通信を行う．また，現在はまだこのシステムには実装されていないが，音声の韻律情報を対話管理部に逐次送信することによって，相槌や共同補完等の自然な対話を行うために必要な対話のリズム制御を行う予定である．

2.2. 対話管理部

対話管理部に，「複数エージェント雑談音声対話システム」⁹⁾で開発した対話管理部を用いる．「複数エージェント雑談音声対話システム」では，ディスプレイに映った 2 体の擬人化エージェントとの対話を行っていたが，本システムでは 2 体のロボットに置き換えることで，ロボット対話を実現する．

対話管理部の処理の流れは図1に示されている。音声認識部による音声認識結果は、情報収集部へと送られ、情報スロットに必要な情報を格納する。情報スロットに格納された情報は、応答生成部へと送られ、応答生成に用いられる。応答生成部では、音声認識結果と情報スロットを用い、ELIZA¹⁷⁾型のテンプレートによる手法によって応答が生成される。ロボット毎に独立な対話制御を行うことも考えられるが、自律制御によるエージェント同士の対話を実現することはまだ難しいため、単一の対話管理部で2体のロボット両方の制御を行って、適切に対話が進行するようにしている。

対話中の重要な情報は対話管理部が管理する情報スロットに格納されている。この情報を参照して、ユーザの嗜好に合わせた共感発話を行い、対話を盛り上げる方向に進めるようにしている。また、共有している情報を元に、対話の流れ（シナリオ）を変化させ、情報を応答に盛り込み、結論の誘導を行うことが出来る。具体的な対話状態の制御を図 3 に示す。

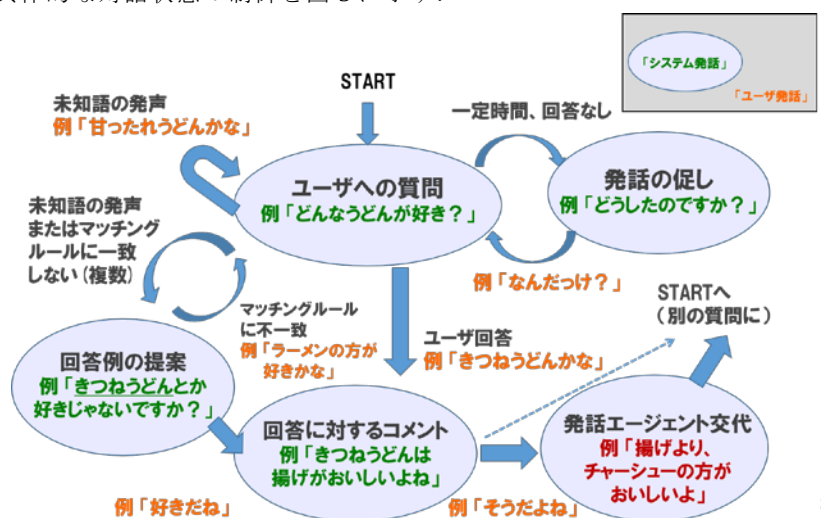


図 3. 対話の状態遷移



図 4. 本研究で用いるロボット（左が PALRO, 右が Sota）

2.3. 応答出力部

本研究で使用するロボットを図 4 に示す。本研究では、FUJISOFT 社製「PALRO」とヴイストン社製「Sota」の 2 台を用いてシステムを構築する。異なる 2 体のロボットを用いることによって、個々のロボットに個性をもたせることができ、対話エージェント同士のやり取りにおけるバリエーションを増

やすことが容易になる。

PALRO, Sota 共に音声合成を行うことが可能であり（本体内またはインターネットクラウドに音声合成エンジンを持つ）、対話管理部の応答生成部が生成した応答文を本体で受け取り、それを音声合成 API を用いて音声として発することができる。また、両者ともにジェスチャーが可能であるため（PALRO は 22 自由度（内 10 自由度は脚部）、Sota は 8 自由度（内 1 軸は胴体回転用））、応答内容に合わせた適切なジェスチャーにより、より楽しく対話を行うことができる。

3. 空間伝達関数の測定

音響処理部の処理精度を上げるために、実験室の空間伝達特性の測定を行う必要がある。本節では、空間伝達特性測定のための具体的な手法と測定結果について述べる。

使用したマイクロホンアレイを図 5 に示す。JTS/CX-500 コンデンサマイクロホンを 8 素子使用し、3cm 間隔で直線上に配置した線形マイクロホンアレイである。



図 5. 使用したマイクロホンアレイ

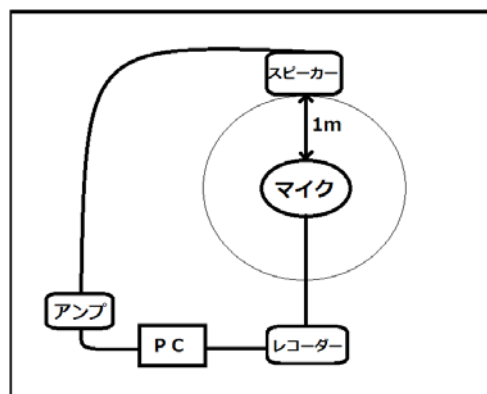


図 6. 伝達関数測定時の実験室配置

伝達関数測定時の実験室の配置図を図 6 に示す。実験室は図 2 に示したものと同一で、部屋の高さは 2.6m である。マイクロホンアレイの中心から 1m 離れた同心円上に置いたスピーカーから同心円の中心（マイクロホンアレイの中心）に向かって TSP 信号を流し、マイクロホンアレイを通してマルチチャネル収録を行う。同心円上を 1 周 360° に対して 5° ずつずらしながら測定することで、それぞれの角

度の伝達関数を得ることができる。また、安定した伝達関数を得るために、ひとつの角度に対して 10 回 TSP 信号を収録し、同期加算を行うことで測定精度を上げている。TSP の再生にはスピーカーとしてアンブレラカンパニー社製 AURATONE、アンプとして DENON 社製 PMA-50 プリメインアンプを用い、PMA-50 を PC に USB 接続して TSP 信号を再生している。レコーダーには ZOOM 社製 F8 マルチトラックレコーダーを使用した。

この条件で測定したマイクロホン 0 (図 5 の左端のマイクロホン) の伝達関数の振幅周波数応答を図 7 に示す。測定は $0^{\circ} \sim 355^{\circ}$ までの 72 個の角度で行っており、角度毎に伝達関数が得られていることが図 7 から分かる。図 7 の横軸 [Hz] は振幅応答の周波数を、縦軸 Src [No] は測定した 72 個の角度の番号 (0~71) を、高さ [Amp] は応答の振幅を表している。なお、角度は 0° がマイクロホンアレイの正面 (図 5 を撮影した方向)、 $\pm 90^{\circ}$ がマイクロホンアレイのマイクが配置されている直線方向 ($+90^{\circ}$ が図 5 の右手方向) を示しており、図 7 では 0° が Src 0 に、 90° が Src 18 に対応する。円周上で測定を行うため、Src 0 と Src 71 の角度の違いは 5° であり、連続的な変化になっていることが図 7 から分かる。この伝達特性を用いて、音声認識部に渡す音声信号の強調を行うことで、音声認識性能を向上させる。

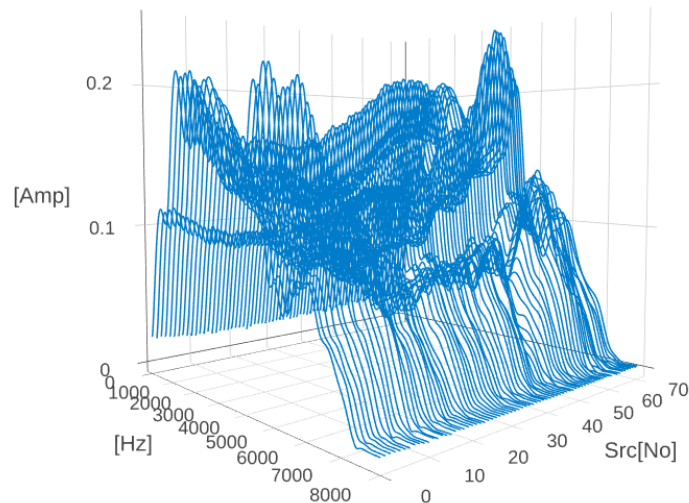


図 7. 測定した部屋の伝達関数

4. 音声認識部の改善

4.1. 比較手法

音響モデルを学習するデータの違いや、HARK の雑音抑圧の除去、話者の位置によってどの程度音声認識率に影響が現れるかを調査する。まず、音響モデルを学習する際、雑音や残響を含まないクリーンな音声を使用する場合と、実験室の伝達関数を畳み込んだ音声を使用して学習したモデルで比較を行う。その後、伝達関数を畳み込んで学習した音響モデルを使用して、HARK の雑音抑制の有無による比較、マイクロホンアレイと話者の距離による比較を行う。

4.2. 実験条件

実験に使用するデータは、日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ)¹⁸⁾ に含まれる全ての音声 (学会講演, 模擬講演, 朗読音声, その他) である。学習データは 417,384 発話の約

650 時間であり，話者は男女混合の 1,417 名である．テストデータは学習データに含まれない男性話者 10 名の 1,274 発話（27,651 単語）を用いた．

クリーン音声で音響モデルを学習する実験では，CSJ に収録されているデータをそのまま使用して，音響モデルの学習を行った（すなわち，テスト環境とはミスマッチとなる）．実験室の伝達関数を畳み込んだ音声を使用して音響モデルを学習する実験では，マイクロホンアレイ正面から 0° ， -30° ， -45° ， -60° ， -90° の 5 種類の伝達関数を，CSJ の学習データをランダムに 5 等分した音声データにそれぞれ畳み込んで（すなわち，学習データの総量はクリーン学習の場合と同じである）実験室の音声データを擬似的に生成し，HARK で音源分離・雑音除去を行ってから，学習を行った．認識実験は，図 8 のようにマイクロホンアレイ，スピーカーを配置し，実験室でテストデータを再収録したものにより行った．図 8 はマイクロホンアレイの中心から 1m の距離にスピーカーを配置したものだが，距離が 1.5m の場合も同様に実験を行った．

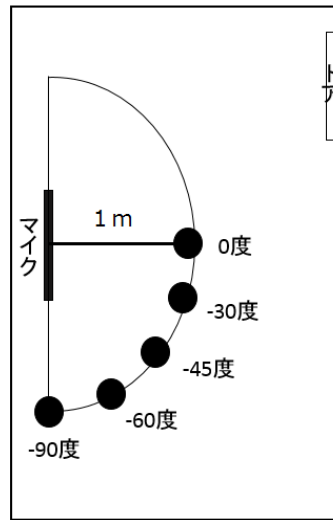


図 8. 認識実験のスピーカー配置

音響特徴量には，MSLS 特徴を用いる．分析条件は，窓長 25ms，10ms シフトで，窓関数にはハミング窓を使用，FFT サイズは 512 点とし，メルファイルバンクの次数は 40 とした．MSLS 特徴量を DNN に入力する際は前後 5 フレーム，合計 11 フレームの特徴を合わせて使用した．そのため，DNN の入力層は 440 次元となる．DNN の隠れ層は 6 層でユニット数は 2048，出力層のユニット数は DNN-HMM の元となる GMM(Gaussian Mixture Model)-HMM と同じ 4,960 で，HMM の単位は状態共有 triphone（前後の音素環境によりモデルを場合分けした音素モデル）である．DNN の活性化関数として，隠れ層では Rectified Linear Unit (RLU)，出力層では Softmax を使用した．言語モデルは CSJ 付属の短単位形態素解析結果から学習した trigram で，語彙サイズは 71,942 である．

4.3. 実験結果

実験の結果を「単語誤り率 (Word Error Rate; WER)」によって示す．WER は以下の式によって算出される．

$$WER = \frac{S + D + I}{N} \times 100 \text{ [%]} \quad (6)$$

ここで、S, D, I, N はそれぞれ、置換誤り単語数、削除誤り単語数、挿入誤り単語数、正解文の単語数を表し、WER が低ければ低いほど認識性能が高いことを示す。

表 1 に音声認識実験結果を示す。容易に予想されることだが、クリーン音声で学習した音響モデルに比べて、伝達関数を畳み込んだ音声で学習した音響モデルを用いる方が、音声認識性能が高くなっている。また、HARK の雑音抑圧を使用することで数ポイント程度の WER の改善が得られることが分かる。実験環境は背景雑音が大きい環境ではないものの、音声には残響の影響も含まれるため、雑音抑圧手法が有効に働いたと考えられる。実際の対話ではユーザは -45° の位置付近に立つことが多いと考えられるが、認識実験では $-45^\circ \sim -60^\circ$ の位置での認識性能が高くなっており、実用的なセッティングなっていると考えて良い。しかし、スピーカーをマイクロホンアレイから 1.5m まで離して音声を再生すると大幅に音声認識性能が低下している。これは距離減衰により音声の受音時のパワーが低下することで S/N 比が低下したことによるものと考えられるが、人間とロボットの位置関係は今後被験者実験を行う等して、適切な位置を探っていく必要がある。

さらに、この実験で得られた WER は、接話マイクロホン条件の結果と比較して 15 ポイント以上悪い結果となっているため、より適切な音声認識条件を考えていく必要がある。

表 1. 音声認識実験結果 (WER [%])

モデル学 習音声	スピーカー 距離	HARK 雑音抑圧	角度				
			0°	−30°	−45°	−60°	−90°
クリーン	1.0m	有り	47.1	45.8	45.0	43.3	48.2
伝達関数 畳み込み		無し	48.2	43.9	42.1	38.6	43.3
		有り	38.9	37.5	35.6	36.4	35.7
	1.5m	有り	46.8	44.6	44.8	43.0	45.7

5. まとめ

本稿では、ロボットとの雑談を交えた音声対話を通して情報提示を行う、デジタルサイネージとしてのロボット音声対話システムの開発を目指して、その最初の段階として、システム全体の枠組みの構築と、音声入力部分である遠隔発話（マイクから離れた位置からの発話）音声認識を頑健に行うための実験室における測定、および単語誤り率による音声認識性能の評価を行った。結果として、システム全体の枠組みを構築することはできたが、音声認識性能が十分ではなく、人間同士の対話のようにロボットと対話を行うことはまだ難しい。

今後は、

- ・ 音響処理部の性能改善（マイクロホンアレイの改善、音声強調アルゴリズムの改善）
- ・ 音声認識性能の改善（実環境音声認識に適した音響モデルの開発）
- ・ パラ言語情報の利用（ピッチ情報を利用したあいずち等の生成）
- ・ 対話管理部の改善（雑談用の自動的な話題獲得とその場合の適切な対話制御）
- ・ ロボットが持つカメラ映像の活用
- ・ デジタルサイネージとしての情報提示用ディスプレイの活用

といった課題に取り組むことで、人間のような対話を行えるロボット音声対話システムを開発していく予定である。

謝辞

本研究は中部大学総合工学研究所 平成 30 年度プロジェクト 4 の援助を受け遂行されたものであり、

ここに謝意を表します。また、本研究の遂行にご協力いただいた、中部大学工学部情報工学科・中川聖一教授と、情報工学科4年生の山本直哉君、宮部昂太朗君に謝意を表します。

参考文献

- 1) V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goodine, D. Goddeau, J. Glass, “PEGASUS: A spoken dialogue interface for on-line air travel planning”, *Speech Communication*, Vol.15, No.3-4, pp.331-340, 1994.
- 2) D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, E. Fosler, N. Margan, “The Berkeley restaurant project”, *Proc. ICSLP 94*, pp.2139-2142, 1994.
- 3) S. Seneff, J. Polifroni, “Dialogue management in the Mercury flight reservation system”, *Proc. 2000 ANLP/NAACL Workshop on Conversational systems*, Vol.3, pp.11-16, 2000.
- 4) T. Zhao, M. Eskenazi, “Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning”, *Proc. SIGDIAL 2016*, pp.1-10, 2016.
- 5) 東中竜一郎, “雑談対話システムの評価とその問題点”, *人工知能学会基本問題研究会, SIG-FPAI*, 101, pp.21-23, 2016.
- 6) 玉川聡, 山本一公, 中川聖一, “音声対話インタフェースにおけるロボットエージェントの優位性の評価”, *情報処理学会音声言語情報処理研究会, SIG-SLP*, 97, 11, 8 pages, 2013.
- 7) 伊藤敏彦, 小暮悟, 中川聖一, “協調的応答を備えた音声対話システムとその評価”, *情報処理学会論文誌*, Vol.39, No.5, pp.1248-1257, 1998.
- 8) R. Nishimura, S. Nakagawa, “A spoken dialog system for spontaneous conversations considering response timing and response type”, *IEEJ Trans.*, Vol.6, No.S1, pp.S17-S26, 2011.
- 9) 藤堂祐樹, 西村良太, 山本一公, 中川聖一, “複数の対話エージェントを用いた雑談指向の音声対話システム”, *電子情報通信学会論文誌*, Vol.J99-D, No.2, pp.188-200, 2016.
- 10) 大浦圭一郎, 山本大介, 内匠逸, 李晃伸, 徳田恵一, “キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム”, *人工知能学会誌*, Vol.28, No.1, pp.60-67, 2013.
- 11) 中臺一博, 奥乃博, 水本武志, 中村圭佑, “ロボット聴覚オープンソフトウェア HARK の開発とその応用”, *シミュレーション Journal of the Japan Society for Simulation Technology*, Vol.35, No.1, pp.32-38, 2016.
- 12) 守谷直也, 金田豊, “インパルス応答計測のための最適信号の検討”, *電子情報通信学会応用音響研究会, EA*, 104, 715, 6 pages, 2005.
- 13) R. Schmidt, “Multiple emitter location and signal parameter estimation”, *IEEE Trans. Antennas and Propagation*, Vol.34, No.3, pp.276-280, 1986.
- 14) K. Sekiguchi, Y. Bando, K. Nakamura, K. Nakadai, K. Itoyama, K. Yoshii, “Online simultaneous localization and mapping of multiple sound sources and asynchronous microphone arrays”, *Proc. IEEE/RSJ IROS*, pp.1973-1979, 2016.
- 15) H. Nakajima, G. Ince, K. Nakadai, Y. Hasegawa, “An easily-configurable robot audition system using histogram-based recursive level estimation”, *Proc. IEEE/RSJ IROS*, pp.958-963, 2010.
- 16) D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, “The Kaldi speech recognition toolkit”, *Proc. ASRU2011*, 4 pages, 2011.
- 17) J. Weizenbaum, “ELIZA – a computer program for study of natural language communication between man and machine”, *Communication of the ACM*, Vol.9, No.1, pp.36-45, 1966.
- 18) 小磯花絵, 前川喜久雄, “『日本語話し言葉コーパス』の概要と書き起こし基準について”, *情報処理学会自然言語処理研究会, SIG-NL*, 143, pp.41-48, 2001.