Vol.2013-HCI-151 No.9
Vol.2013-SLP-95 No.9
2013/2/1

ロボット自身のマイクを介したNaoとの音声対話

駒谷 和範　中島 大一　杉山 貴昭

概要：我々の研究室で，デモ用に開発している音声対話システムを紹介する．このシステムは，ヒューマノイドロボット Nao と，音声を用いて対話するものである．具体的には，ユーザに回文の冒頭部分を指定させ，その回文を読み上げる．本システムの特長は以下の 3 点である．(1) ロボット自身に備え付けられたマイクを用い，接話マイクを使用しない．(2) 音源定位により，話者の方向を向くことができる．(3) ロボットが選択肢を列挙し，それに対する割り込みタイミングにより，ユーザの意図を解釈するモードがある．

1. はじめに

本稿では，我々の研究室で開発したデモシステムを紹介する．本システムは，アルデバランロボティクス社で開発されたヒューマノイドロボット Nao[*1] と，音声対話を行うものである．本システムの特長のひとつとして，接話型マイクを用いず，Nao 自身に備え付けられたマイクを用いる点が挙げられる．音声対話システムは，電話を介したシステムが歴史的に数多く開発されてきたことから，話者の口元にマイクがあることが前提とされることが多い．

接話型マイクを用いない場合，対象とするユーザの音声に限らず，周囲の雑音がマイクに混入する．このため，これを念頭においたシステム設計が必須である．具体的には，以下の 2 点が必要である．

(1) 話者からの音響信号のみを認識対象とするフロントエンド処理
(2) 音声認識性能が高くない場合のバックアップとなる対話戦略

本システムでは，前者として，ロボット聴覚ソフトウェア HARK[1] を利用し，音源定位機能と，その結果に対するしきい値処理によって，雑音による誤動作の回避を狙う．後者は，常に受け身にユーザの発話を認識・解釈するだけでなく，システム側から選択肢を提示し，それに対するユーザの割り込み（バージイン）タイミングを用いて選択対象を指定させる，というモードを備えている．これにより，音声認識性能が低い場合でも，ユーザが意図を伝達可能となる手段を用意している．

さらに，身体性を持ったロボットが対話を行う際に，話者の方向を向くのは必須の機能である．これも，HARK による音源定位結果を用いて実現している．

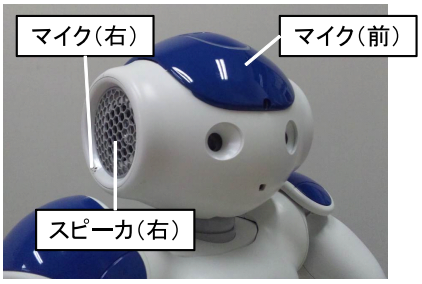

図 1　Nao のマイクとスピーカの位置

2. システム構成

2.1 音源定位，音声認識

音響信号は，Nao の頭部にある 4 個のマイクを介して入力される．Nao のマイクとスピーカの位置を図 1 に示す．図の写真では見えないが，頭部の左側と後ろ側にも同様にマイクが内蔵されている．スピーカとマイクが近接していることから，Nao のスピーカで音を再生している間に，正しく音声認識を行うのは困難であると予想できる．

この 4 個のマイクから入力される 4 チャンネルの音響信号に対して，ロボット聴覚システム HARK[1] により音源定位を行う．この音源定位は，MUltiple SIgnal Classification (MUSIC) 法に基づいており，1 フレーム (0.01 秒) ごとに，定位角度とそのパワーが出力される．MUSIC 法で用いる伝達関数の作成のために，インパルス応答を 1m 離れた点から 36 点計測した（10 度間隔）．したがって，音源定位の角度分解能は 10 度である．音源分離部での伝達関数の作成にも，音源定位部と同様に測定したデータを利用した．

音声認識部は，タスクに必要な語彙が小さく，適当な学習コーパスが存在しないことから，記述文法を言語制約と

[1] 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464–8603, Japan

[*1] http://www.aldebaran-robotics.com/ja/

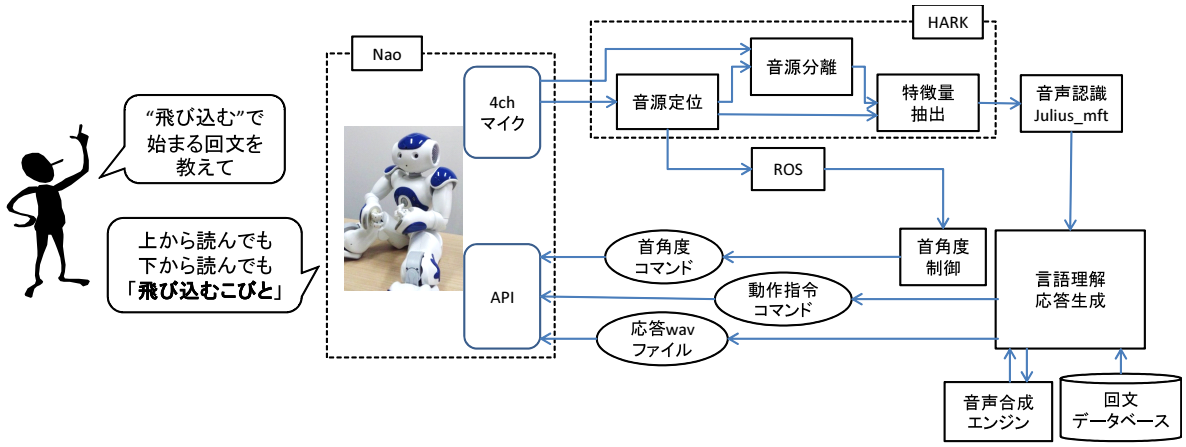

図 2 システム構成図

している．語彙サイズは 81 である．音響モデルには，2012 年 3 月の HARK 講習会で配布されたモデルを用いている．特徴量には MSLS (Mel-Scale Log Spectrum) を用いた．

2.2 言語理解・応答生成

言語理解や対話管理は，基本的に一問一答形式で行われる．つまり，入力である音声認識結果に含まれるキーワードに対応づけられた応答文を生成する．また基本的に対話状態は仮定していない．直前のシステム発話の内容のみ保持しており，「もう一度言ってください」や「それを逆から読んでください」というユーザ要求に対応する．一部の応答では，発話とともに，立ち上がる動作や座る動作，手を振る動作，見上げる動作，うなずく動作を行う．また，システムの発話中にユーザが割り込んで発話した場合に，音声合成を止める機能（バージイン機能）も実装している．

システムのタスクは，回文の読み上げである．回文とは，例えば「うどん噛み感動」「報告聞く候補」のような，文頭と文末のいずれから読んでも，同じ読みになる文のことである．本システムでは，登録されている回文の先頭部分をキーワード（キーフレーズ）としてユーザに入力させ，その回文をシステムが読み上げる．現時点では，2 文節または 3 文節から成る回文 48 個が登録されている．これらは，文節集合から網羅的に回文を生成する研究 [2] において，その途中段階で得られた回文のごく一部分である．

さらに，読み上げ可能な回文の例をシステムが列挙している最中にユーザが発話した場合に，そのタイミングによって，ユーザの意図を解釈するモードを備えている [3]．現在のシステムでは，登録している回文の数は多くないが，回文を多く登録し，音声認識辞書の語彙サイズが増大した場合には，音声認識性能が低下することが予想される．このような場合に，音声認識結果だけではなく，発話のタイミングを用いることで，ユーザの意図する対象を指定できる．この点は，接話型マイクを使わないことによる音声認識性能の低下にも，同様に有効であると考えている．

3. 今後の展開

本稿で述べた音声入力部分は，複数のロボットを使用して，複数のユーザと会話を行うシステム [4] と共通である．このシステムでは現在，音源定位結果と顔検出結果をそれぞれ蓄積することで，会話におけるユーザの状態を推定し，それに応じた応答を行う手法を開発している．本稿で述べたシステムでは，得られた入力に対して反射的に応答を行うのみであるが，このように対話やユーザの状況を蓄積して，システムから多様な発話を生成することが考えられる．

さらに，ロボット自身の動作音や合成音声による誤動作も避ける必要がある．ユーザの発話とこれらの音とを判別する GMM を，分離音から作成する研究も進めている [5]．これを使用することによって，さらに雑音に対して頑健なシステムとなることが期待できる．

謝辞 ロボット聴覚ソフトウェア HARK の作成，保守に関わる各位に感謝する．Nao と HARK を接続するプログラムは，京都大学の水本武志氏と協力して作成した．本研究の一部は，JST 戦略的創造研究推進事業さきがけの支援を受けた．

参考文献

[1] 奥乃 博，中臺一博：ロボット聴覚オープンソフトウエア HARK，日本ロボット学会誌， Vol. 28, No. 1, pp. 6–9 (2010).
[2] 鈴木啓輔，佐藤理史，駒谷和範：文頭固定法による効率的な回文生成，言語処理学会第 17 回年次大会発表論文集，pp. 826–829 (2011).
[3] 駒谷和範，松山匡子，武田 龍，高橋 徹，尾形哲也，奥乃 博：発話行為レベルの情報をユーザ発話の解釈に用いる音声対話システム，情報処理学会論文誌， Vol. 52, No. 12, pp. 3374–3385 (2011).
[4] 中島大一，駒谷和範，佐藤理史：複数人会話におけるロボットによる視聴覚情報に基づくアクティブユーザの推定，情報処理学会研究報告，2013-SLP-95-20 (2013).
[5] 杉山貴昭，駒谷和範，佐藤理史：ロボットとの音声対話における発話の重なりを含む入力音の判別，情報処理学会第 75 回全国大会講演論文集，(to appear) (2013).