音声言語情報処理 39－16
(2001. 12. 20)

Juliusを用いた学内案内ロボット用音声対話システムの作成

西村 竜一† 内田 賢志† 李 晃伸† 猿渡 洋† 鹿野 清宏†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0101 奈良県生駒市高山町 8916–5

E-mail: †{ryuich-n,takasi-u,ri,sawatari,shikano}@is.aist-nara.ac.jp

あらまし ASKA（アスカ）は，大学の受付案内システムを目標として開発中の頭部や腕のジェスチャ機能を持つ人間型音声対話ロボットである．音声対話機能は，大語彙連続音声認識エンジン Julius と学内案内タスク向け N-gram 言語モデルを基礎としたキーワード検索による音声認識理解部と音声合成部によって構成されており，対人センサやジェスチャ生成などの他のモジュールと状態を通信しながら分散的な動作を行なう．本ロボットは，奈良先端大における学内共同プロジェクトで開発されており，エージェントシステムにおける様々な要素技術の実環境での検証プラットフォームと位置付けられている．今後も新たな要素技術を採り入れながら開発を続ける予定である．本稿では，音声対話機能の実装方法を中心に現在の ASKA の概要および今後の予定について述べる．

キーワード 音声対話ロボット，大語彙連続音声認識エンジン Julius，キーワード検索，N-gram 言語モデル

Development of Julius-based Speech Dialogue System for Campus Receptionist Robot

Ryuichi NISIMURA†, Takashi UCHIDA†, Akinobu LEE†, Hiroshi SARUWATARI†, and Kiyohiro SHIKANO†

† Graduate School of Information Science, Nara Institute of Science and Technology
8916–5 Takayama-cho, Ikoma, Nara, 630–0101, Japan

E-mail: †{ryuich-n,takasi-u,ri,sawatari,shikano}@is.aist-nara.ac.jp

Abstract ASKA is a speech-orinted humanoid robot system to realize an automated receptionist agent of a university. The speech dialogue function is composed of speech understanding module and text-to-speech synthesis module, each communicates to other parts such as object sensing module and genture generation modules to accomplish an distributed system integration. The recognition scheme is based on LVCSR engine Julius, with task-dependent N-gram language model and keyword-based meaning extraction algorithm. The development is a collaborative work of several laboratories in NAIST, and is aimed to be a common verification platform of various agent technologies and engineerings in real environment. In this paper, we describe the overview of ASKA, the implementation method of the speech dialog functions on ASKA and the future schedules.

Key words Speech dialogue robot, LVCSR engine Julius, Keyword search, N-gram language model

1. はじめに

現在，エンターテイメントロボットや二足歩行ロボットが高い注目を浴びている．商品化された高機能なロボットの中には，音声認識や音声合成技術をベースとする人間との音声対話機能を持つものもある．現状では，まだロボットの目新しさのみが注目されている感があり，実用的に人間にとって役に立つものは少ない．しかし，音声対話機能を持つロボットは，実世界で人間に役に立つエージェントシステムとして活躍することが期待されている．音声対話システムのアプリケーションは，実用化が近いと言われながらも，キラーアプリケーションと呼べるような高い有用性を示したものはいまだかつて登場していない．音声対話ロボットは高い有用性の可能性を持つため，今後，音声対話ロボットの実装及びその評価は重要性を増すことになる．

しかし，実環境で動作する音声対話ロボットの開発には，音声情報処理以外の様々な要素技術が必要であり，実装は困難である．そこで関連技術を持つ学内の他の講座との共同プロジェクトを作り協調作業により開発を進めている．このロボットは，大学の受付案内タスクをターゲットとしたエージェントシステムであり，音声情報処理のみならず様々な要素技術の実環境での検証プラットフォームとなることが期待されている．

本報告では，音声対話機能を中心に開発中のロボットの現状及び今後の予定について述べる．

2. 受付案内ロボット ASKA

2.1 ASKA の概要

本プロジェクトで開発するロボットは，奈良先端科学技術大学院大学（NAIST）情報科学研究科の研究科棟一階フロアの入り口に設置し，来客の応対をタスクとする人間型受付案内ロボットである．音声による対話及び頭部や腕などのジェスチャによるインタフェースを用いて受付案内を行なう．名前は，ASKA（アスカ）と命名した（図1）．

胴体のハードウェアにはテムザック社[1]によって開発された「テムザック4」を用いた．テムザック4は，遠隔操作用に開発されているが，ASKA では自律動作が求められるため，内蔵計算機を十分な演算能力を持つ計算機に載せ替えている．また，車輪による移動能力を持つが，前述の計算機の載せ替えなどによるハードウェアの制限により現在は利用できない．

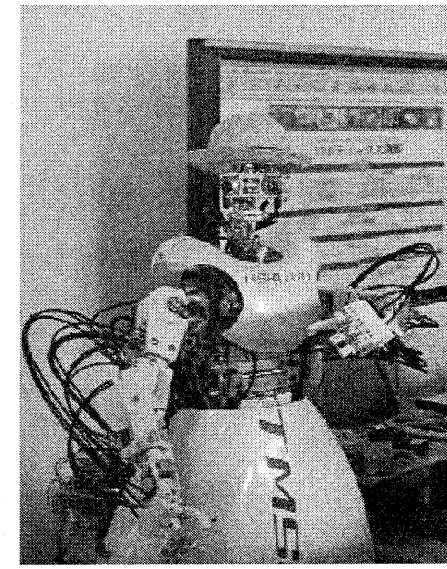

図1 NAIST 受付案内ロボット ASKA

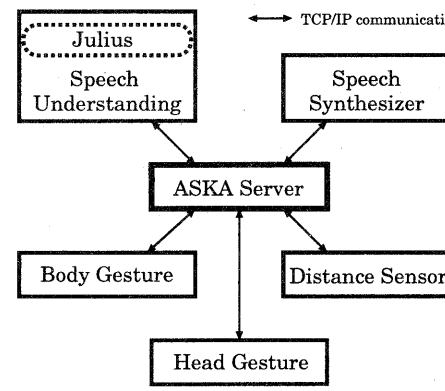

図2 システム概要

頭部ハードウェアには，通信総合研究所によって開発された「Infanoid」[2]の頭部を用いた．この頭部では，目や口及び首振りの動作をジェスチャとして利用できる．また，目に埋め込まれたカメラの画像処理による応用が可能である．

プログラムは，機能ごとにわかれたモジュールと TCP/IP によって各モジュールと通信してパラメータを記憶するサーバによって構成される．現在のモジュールには，音声認識理解部，音声合成部，位置センサ部，胴体ジェスチャ部，頭部ジェスチャ部がある（図2）．モジュールは，自分以外の他のモジュールの動作 ON/OFF 状況及びモジュールの動作結果により得られた戻値をサーバとの通信でモニタしながら，互いが独立に動作する．この方法は，モジュール間

の複雑な同期処理には向かないが，システム全体の開発が容易になる利点を持つ．基本的にモジュール内の開発を開発グループごとに独立に行なえるため，新たな技術の組み込み及び実環境での検証を手軽に行なうことができる．

本システムで用いる計算機のOSは，全てLinuxを利用している．C言語及びPerlのライブラリを用意することにより，モジュールはC言語またはPerlで記述することができる．計算機は，主にジェスチャを担当する胴体部内蔵のものと主に音声認識理解，音声合成を担当するイーサーネットによって接続した外部のものの2台を利用している．計算機の数，モジュールの数などは任意の数に増減が可能である．

位置センサ部のセンサハードウェアは，SICK社[3]の「レーザ測定システムLMS400」を利用した．

2.2 ASKAの機能及び処理の流れ

最初に，来客がASKAの前に立つと，位置センサ部によって客のASKAからの距離と角度を得る．来客の立ち位置が設定された距離以内に入ると，音声認識理解部が音声の入力を開始する．同時に，頭部を客の方に向けることによりASKAは，質問の受け付けができる状態であることを示す．

音声認識理解部は，入力音声に対する応答文を作成した後，結果をサーバに送信する．音声合成部は，応答文からTTS（Text To Speech）プログラムにより合成音声を作成する．作成が完了した後，合成音声の発話時間をサーバに送信し，発話待ちの状態で待機する．

胴体と頭部のジェスチャ部は，動作に必要な応答文や合成音声の発話時間などのパラメータを受け取ると，あらかじめ用意されたジェスチャの動作パターンに基づいて動作を開始する．ジェスチャに利用できる動作は，胴体部の両手，頭部の目，口及び首の振りである．合成音声の発話時間は，唇の動作時間の決定に用いる．音声合成部は，ジェスチャと同時に発話を開始することによって，簡単な同期を行なえるようになっている．

図3は，実際にジェスチャ動作中のASKAである．このジェスチャは，場所を案内している時のものであり，案内音声と共に手の動作と頭と目の向きでその場所の方向を指し示している．なお，現在のASKAには27通りのジェスチャが登録されている．

3. 音声対話システムの実装

3.1 音声対話システムの構成

ASKAの音声対話システムは，前述の音声認識理解部と音声合成部モジュールによって構成される．

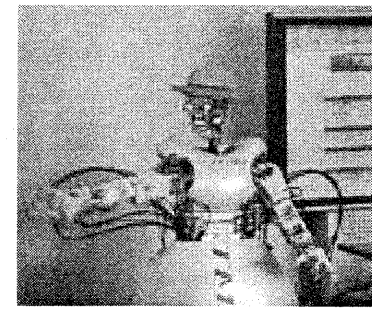

図3 ジェスチャの様子

音声認識理解部の処理では，まず，入力音声の音声認識を行なう．そして，あらかじめ用意した応答文のテンプレートの中から認識結果を基に適切なものを選択する．応答文の選択は，認識結果とキーワードとの一致回数をカウントすることで行なう．キーワード検索の詳細は，3.6節で後述する．

以下では，具体例を用いて音声対話システムの核となる音声認識理解部の実装について述べる．

なお，音声合成部のTTSプログラムには，クリエートシステム開発社の「Linux版日本語音声合成ライブラリー」[4]を用いた．

3.2 大語彙連続音声認識

通常，タスクを限定した音声対話システムの音声認識には，文法記述型の音声認識を用いることが多い．しかし，文法記述型音声認識では，システムはあらかじめ想定された文法内の発話のみしか受理できず，発話内容や語尾などの発話様式が限定されてしまう．また，認識対象語彙が少なくなる事や複雑な文法を開発者が記述する必要があるなどの問題点が知られている．一方，統計言語モデルを用いた大語彙連続音声認識では，認識結果を開発者があらかじめ想定することは難しく，一見すると音声対話システムに組み込むのには向かない．しかし，柔軟に様々な発話を受理することが可能となる．

ASKAでは，以上のような検討をふまえた上で，自由なユーザの発話を柔軟に受理するためにN-gram言語モデルによる大語彙連続音声認識を利用する．音声認識エンジンには，奈良先端科学技術大学院大学及び京都大学によって開発されたJulius[5],[6]を用いた．

3.3 言語モデルの作成

本システムでは，固有名詞（教官名，講座名，場所など）や専門用語などのキーワードとして利用す

表 1　学習用テキストの緒元

	学習用テキスト	参考 新聞記事 1 年分
ファイル容量	129MB	92MB
文章数	277 万文	97 万文
異なり語彙数	17 万個	16 万個

ることの多い特有な単語を音声認識できなければならない．よって，新聞記事などから学習した既存の N-gram 言語モデルでは，十分な認識性能を得ることができない．そこで本学の受付案内のタスクのための N-gram 言語モデルを新たに作成した．

N-gram 言語モデルの学習に必要な学習用テキストには，以下のリソースから取得したテキストを結合したものを用いた．

- Web ページ
- 学内メーリングリスト
- 学内教職員データベース
- ATR 自然発話音声データベース（旅行対話）

「Web ページ」は，奈良先端科学技術大学院大学のインターネットドメイン（`aist-nara.ac.jp`）を持つ Web サイト上の Web ページを収集ロボットプログラムを用いて集めたものである．「学内メーリングリスト」は，過去約 2 年間に学内連絡用メーリングリストに流れたメールを集めたものである．「学内教職員データベース」は，学内の教職員名，講座名，研究テーマなどが収録されている公開データベースにデータの追加削除などの独自の修正を行なったものである．これら三つのテキストに関しては，HTML のタグやメールのヘッダなどの学習に不要な定型部分を削除した後に，本文に対して統計的テキスト整形フィルタ [7] を用いて整形処理を行ない，シグネチャや単語羅列文，絵文字などの不定形な不要部分の削除を試みた．ATR 自然発話音声データベースに関しては，ヘッダ情報などの定型部分の削除のみを行なった．

収集の結果，作成した学習用テキストの緒元を表 1 に示す．Web ページからの収集テキストが学習用テキストのおおよそ半分を占めた．

学習用テキストから N-gram 言語モデルの作成には，Palmkit [8] を用いた．日本語形態素解析には，ChaSen 2.2 を使用した．さらに，「連続音声認識コンソーシアム 2000 年度版ソフトウェア」[9] に含まれる読み変化プログラム ChaWan および数字読み付与プログラムを使用して読み仮名の付与を行ない，語彙辞書を作成した．作成した N-gram 言語モデルは，Julius 向けの 2-gram 及び逆向き 3-gram モデルである．学習に使用した語彙は，学習用テキストの中の出現頻度が高いものから上位 2 万語である．カットオフはいずれも 1 で，Witten Bell ディスカウンティングを適用している．

情報科学センターはどこにありますか？
小笠原研究室の松本先生の部屋番号を教えてください。
木戸出先生の研究室はどこにありますか。
湊教授の部屋はどこですか？
高山サイエンスプラザに行きたいのですが。
学園前まで行く方法を教えて下さい。
煙草を吸いたいので、喫煙所を探しています。
近くにコンビニはありますか？

図 4　評価用テキスト（受付での質問）の例

表 2　予備評価実験の結果

PP			29.44
OOV (%)			0.30
	話者 A	話者 B	合計
Acc (%)	90.61	88.99	89.80
n-Acc (%)	91.49	89.60	90.55

3.4　予備評価実験

作成した言語モデルの評価のための予備実験を行なった．予備実験のための評価用テキストとして，奈良先端科学技術大学院大学の受付での質問を想定したテキスト（150 文，総単語数 1697 個）を学内の学生へ行なったアンケートを元にして作成した．評価用テキストの例を図 4 に示す．

音声認識実験に使用する評価音声は，男性話者 2 名による上記評価用テキストの読み上げ音声を利用した．収録場所は，ASKA 設置予定の一階フロア入り口の受付である．実際の使用状況に合わせるため，PC の音声入力機能による録音を行なった．

音声認識エンジンには，Julius 3.2 [5], [6]，音響モデルには，「連続音声認識コンソーシアム 2000 年度版ソフトウェア」[9] に含まれる PTM triphone，64 混合，3000 状態の男性用 JNAS モデルを用いた．

3-gram 単語パープレキシティ（PP），未知語率（OOV），単語正解精度（Acc）の結果を表 2 に示す．また，認識結果から名詞のみを抜き出して算出した単語正解精度（n-Acc）の結果も示す．

実験結果より作成した N-gram 言語モデルは，全ての評価尺度において受付案内タスクに対して高い性能を示した．特にキーワードとして利用することの多い名詞のみを抜き出した場合の認識率は，90%以上の高い精度を得ることができた．

この予備実験の結果から大語彙連続音声認識と今

100 おはようございます。
103 ご用件はなんでしょうか？
200 私の名前は、アスカです。
204 施設の案内や、先生方のお部屋の案内ができます。
302 <is-staff:3>先生の部屋は、<is-staff:5>です。
303 <is-staff:3>先生の内線番号は、<is-staff:8>です。
404 内線電話は、こちらにあります。
405 公衆電話は、私の後ろにあります。
415 バス停は、そこの玄関を出てまっすぐ道路へ出て左側にあります。

図 5　応答文の例

073 李 晃伸 リ アキノブ B613 B 6 5282 音情報処理学
鹿野 オトジョウホウショリガクコウザ シカノケン

図 6　is-staff ファイルの例

回作成した言語モデルの組合わせによって，学内案内受付タスクの音声対話システムを構成することが可能であることがわかった．

3.5　応答文の作成

ASKA では，来客の質問に対する応答文をあらかじめ用意している．この応答文は，ASKA に必要な機能のアンケートを行ない，その結果から必要性が高いと思われる質問を選びだし，その質問に応じた応答を用意するという手順で作成している．

現在，ASKA に登録されている応答文の数は，61 個である．図 5 に登録されている応答文の例を挙げる．文頭の三桁の数字は，応答文ごとに付けられたインデックス番号であり，他のモジュールとの通信には，このインデックス番号を用いてメッセージの交換を行なう．

応答文は，完全に定型なものと他のデータベースからデータを検索して挿入できるもの（挿入型）の 2 種類がある．挨拶や場所の案内には，定型のものを利用する．教職員の居室や内線番号案内などの応答は，名前，番号などのデータは別に記憶しておき，そこから検索して挿入型応答文にデータ挿入することで生成する．その結果，応答文のエントリを一つにまとめることができる．

図 5 の例の中で，以下の文章は挿入型応答文の例で，<is-staff:3>と<is-staff:5>がデータの挿入箇所を示す．

<is-staff:3>先生の部屋は、<is-staff:5>です。

<is-staff:3>は，is-staff という名前のデータファイルから，認識結果を元に検索して，三番目のデータ（この例では名字の読み）を挿入するという意味である．is-staff ファイルの内容例を図 6 に示す．なお，先頭のインデックス番号は，0 番目として数える．

表 3　応答文作成のためのアンケートの結果
（回答人数: 40 人）

投票数	質問内容
37	バスの時刻
25	近くの駅の電車の発車時刻
24	研究内容での質問に対する講座の案内
23	教授・助教授・助手の居室と内線番号
23	学内および周辺の施設の場所
20	最新ニュース及び今日の出来事
20	天気予報
20	その日の講義及び休講情報
17	今月の学内イベント情報
17	近くのタクシー会社の電話番号
17	付近のお店の情報
17	学生の所属研究室名
12	テニスコート・グランドなどの使用予約状況

応答文は，容易に追加削除が可能であり，適時追加を行なっている．また，応答文の拡充のために学内の学生に対するアンケートを継続中である．アンケートは，ASKA に答えて欲しい質問事項を挙げてもらう形で行なっている．表 3 にこれまでのアンケートの結果をまとめる．有効回答数は 40 である．40 人中希望が多かった事項の中には，来客者よりも学内の人間が求むようなものも多く，当初想定した ASKA のタスクとは若干異なる．しかし，要望が多いものに関しては，ASKA を学内の人間にも有用なエージェントシステムとしても利用できるように採用していく予定である．

3.6　キーワードリストと応答文の選択

認識結果から応答文の選択に用いるキーワードリストは，図 7 に例を示すように応答文ごとにキーワードを定義して作成する．行頭の番号は，前述の応答文のインデックス番号である．

図 7 の例中で，インデックス番号に続いて，"k" が指定されている文字列が登録するキーワードであり，形態素単位に記述する．応答文の選択は，形態素に分割された認識結果とキーワードとの一致の数をカウントして，最も一致数の多いものを選ぶことで行なう．高い選択性能を得るために，音声認識の結果出力には N-best を用いる．

キーワードリスト中でインデックス番号に続いて，"p" が指定されている文字列は，パターンマッチワードとして登録する文字列である．認識結果の一部がこの文字列と一致する場合は，キーワードより優先

```
100 p おはよう
302 k 先生
302 k 教授
302 k 助教授
302 k 助手
302 k 部屋
302 k 番号
```

図 7　キーワードリストの例

して応答文の選択に用いられる．

なお，キーワードリストの追加はシステムを止めることなく随時行なうことができる．

4. 今後の予定

今後の検討事項としては，主に音声認識の基本性能と対話処理の改良が考えられる．

音声認識の基本性能の改良としては，大学の一階フロア入り口という環境を考慮した雑音処理の導入が挙げられる．現在は，雑音対策として雑音信号の畳み込み音声から作成した音響モデルを用いる程度だが，マイクロホンアレーを応用した音声入力デバイスの実装を行なう．また，オープンキャンパスなどのデモンストレーションに来る子供でも利用できるようにするため，子供の声のための音響モデルの整備も重要な課題事項である．N-gram 言語モデルの改良も引き続き行なう．

対話処理は，本稿で述べた現在の音声認識理解部のキーワード検索による応答文の選択方式では，一問一答式の受け答えしか行なうことができない．このままでは不十分なので，対話の流れを管理することで，より複雑な受け答えができるプログラムの開発を進めている．しかし，現在の音声認識理解部のままのシステムの改良も引き続き行ない，キーワードリストの充実による応答文選択性能の向上も目指す．

また，音声認識理解部以外では，現在レーザ測定システムを利用している話者位置の測定の代わりに頭部のカメラを用いたアイコンタクト [10] を用いることを検討している．

5. おわりに

本稿では，音声対話機能を中心に NAIST 受付案内ロボット ASKA について述べた．ASKA のシステムは，互いに通信しながら独立して動作する音声認識理解部，音声合成部，位置センサ部，胴体ジェスチャ部，頭部ジェスチャ部のモジュールとそれらと通信するサーバから構成される．音声対話機能は，大語彙連続音声認識エンジン Julius をベースとしたキーワードマッチによる音声認識理解部と音声合成部で構成する．

ASKA の開発は，実装に必要な様々な要素技術を持つ奈良先端科学技術大学院大学情報科学研究科の複数の講座による共同プロジェクトにより進めている．本プロジェクトでは，ASKA を情報科学分野の様々な要素技術を持つエージェントシステムの実環境での検証プラットフォームとして位置付けている．今後はさらに多くの講座の開発参加を予定しており，新たな要素技術を用いた改良を継続して行なう予定である．

謝辞　ASKA の実装は，奈良先端科学技術大学院大学情報科学研究科の松本 吉央先生，怡土 順一さんをはじめとするロボティックス講座との共同開発により行なわれた．音声認識理解部の実装では自然言語処理学講座，音情報処理学講座，ロボティックス講座のみなさんにご協力いただいた．また，情報科学研究科長の植村 俊亮先生には，本プロジェクトに対し多大なるご支援をいただいた．本プロジェクトをご支援いただいているみなさまに深く感謝いたします．

文　献

[1] http://www.tmsuk.co.jp/
[2] H. Kozima, H. Yano: "A Robot that Learns to Communicate with Human Caregivers," The First International Workshop on Epigenetic Robotics, 2001.
[3] http://www.sick.co.jp/
[4] http://www.createsystem.co.jp/linux.html
[5] 李, 河原, 堂下: "単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識," 信学論, J82-D-II No.1, pp.1–9, 1999
[6] A. Lee, T. Kawahara, K. Shikano: "Julius — An Open Source Real-Time Large Vocabulary Recognition Engine," In Proc. of 7th European Conference on Speech Communication and Technology (EUROSPEECH2001), pp.1691–1694, 2001
[7] R. Nisimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari, K. Shikano: "Automatic N-gram Language Model Creation from Web Resources," In Proc. of 7th European Conference on Speech Communication and Technology (EUROSPEECH2001), pp.2127–2130, 2001
[8] 伊藤, 好田: "単語およびクラス n-gram 作成のためのツールキット," 信学技報, SP2000-106, pp.67–72, 2000
[9] 河原, 住吉, 李, 武田, 三村, 伊藤, 伊藤, 鹿野: "連続音声認識コンソーシアム 2000 年度版ソフトウェアの概要と評価," 情処学研報, 2001-SLP-38-6, pp.37–42, 2001
[10] Y. Matsumoto, A. Zelinsky: "An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement," In Proc. of IEEE Fourth International Conference on Face and Gesture Recognition (FG2000), pp.499–505, 2000.