4B−2−01

情報科学研究の実環境プラットフォームとしての受付案内ロボット ASKA

西村 竜一　怡土 順一　李 晃伸　松本 吉央
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

近年，エージェントやロボットが人間と対話するシステムの研究が盛んである．そのようなシステムの構築には，様々な情報科学・情報工学の技術が必要となるため，研究プラットフォームとしても適していると考えられる．奈良先端科学技術大学院大学・情報科学研究科では，「受付案内ロボット」を開発するという共通の目標となるタスクを設定し，研究室の枠を越えて様々な研究成果を統合しながらロボットシステムを構築する試みをはじめている．本稿では，受付ロボットシステム “ASKA” の開発目的，現状でのシステム構成および音声対話機能について述べる．

2 研究プラットフォームとしてのロボット

情報科学あるいは情報工学の研究には，画像処理，音声処理，知識処理，ヒューマンインタフェース，VR，データベース，通信，CG，シミュレーションなど，様々な研究分野がある．従来，これら情報分野において共通して用いられる研究機器としては，コンピュータのみであった．コンピュータとユーザの間のインタフェースとして，画面上にCGで描かれたエージェント（人物像）が用いられることは多いが，あくまでコンピュータ上の存在であり，実世界で物理的に人間とインタラクションできる存在ではない．

一方，従来のロボット研究においては，機構，制御をはじめとした主に機械系の要素技術が研究されてきたが，近年の研究の広がりにより，上記に挙げたような情報分野との重なりが増えている．しかし，実際には情報分野の研究者がロボットを用いた実験や応用システムの構築を行うというよりは，ロボット（主に機械系）の研究者が情報分野の成果を独自にコンピュータ上に実装，もしくは情報分野の成果として公開されたソフトウェアをユーザとして利用しているというケースがほとんどである．このように，ロボット研究では情報科学・情報工学の研究成果を非常に多く利用しているが，逆にロボット研究の成果をほとんど還元していない．これは，現在のロボット研究が様々な要素技術を集めて実装するというシステム統合化中心の研究であることを考えると仕方がない部分もあるが，今後ロボットが社会に広まっていくためには，ロボットはまず最も身近な情報分野の研究者に広く受け入れられる必要があるはずである．

受付案内ロボットは，このような背景からロボットを情報科学・情報工学の分野における「共通の研究プラットフォーム」として用いる試みとして，本学情報科学研究科において開発が始められた．ベースとなる1台のロボットシステムに「受付案内」という共通の目標となるタスクを設定し，研究室の枠を越えて様々な研究成果を統合しながらロボットシステムを構築することを目指している．このタスクには人間との知的なコミュニケーションが必要であり，その実現には以下に挙げるように様々な情報分野の要素技術が必要となる．

- 画像処理による人や顔の発見，認識
- 音声処理による音声認識と発話（音声合成）
- 自然言語処理による質問の意味理解
- 知識処理やデータ検索処理による質問に対する回答の生成
- センサー情報の学習による環境の認識と行動生成
- 効率のよい並列・分散コンピューティング

これらの要素技術は，いずれも従来から本学情報科学研究科の研究室において研究されているテーマであるため，情報科学研究科で開発するロボットの題材として適していると考えられる．

ロボットを，他分野の研究のためのツールとして用いるという考え方はこれまでにもあった．例えばMITのヒューマノイドCOG[1]は認知科学の研究を行うために開発されたものであり，またERATO川人学習動態脳プロジェクトにおいては脳科学研究のためにヒューマノイドロボットDB[2]を用いている．これらサイエンスの追求のために用いられるロボットと異なり，我々の目指しているシステムは情報分野の成果を工学的に実装するための人間型ロボットシステムである点が異なる．

3 ハードウェア構成

3.1 ボディ部

図1に，開発された受付案内ロボットASKAの外観を示す．ベースとなっているのは，人間型遠隔操作ロボット・テムザックIV（テムザック社[3]）である．テムザックIVのシステムはWindows上で動作し，PHS経由での専用コックピットを用いた遠隔操縦に特化されているため，自律ロボットとしての利用はできない．そこで搭載されているコンピュータを

Receptionist Robot ASKA – Real World Research Platform for Information Science Studies
Ryuichi NISIMURA, Junichi IDO,
Akinobu LEE, Yoshio MATSUMOTO
Graduate School of Information Science,
Nara Institute of Science and Technology

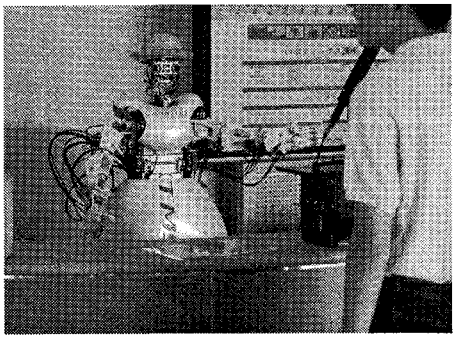

図 1: 受付案内ロボット “ASKA”

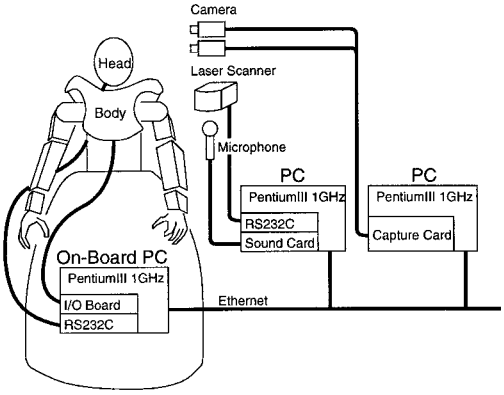

図 2: ASKA ハードウェア構成図

表 1: テムザック IV の仕様

寸法	全高	1200[mm]
	全幅	600[mm]
	全長	750[mm]
重量		約 100[kg]
自由度	胸部	1[DOF]
	腕部	7×2=14[DOF]
	手部	3×2=6[DOF]

Linux ベースの組み込み PC に替え，通常の LAN 環境に接続した．また，画像処理の研究のためにステレオカメラ組み込み型のロボットヘッドを製作し搭載した．さらに，音声認識のためのマイク，および人検出処理のためのレーザースキャナ [5] を取り付けた．図 2 に ASKA のハードウェア構成図を示す．

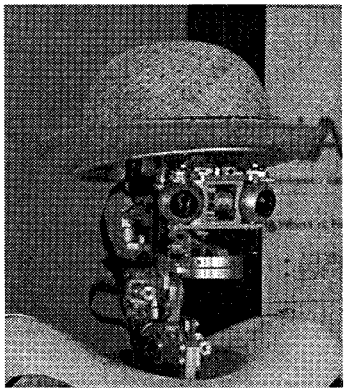

図 3: ASKA 頭部

表 2: 頭部の仕様

自由度	首部	Pan×1, Tilt×2[DOF]
	目部	Pan×2, Tilt×1[DOF]
	口部	2[DOF]
カメラ		Wide×2, Tele×2

3.2 頭部

人間と対話したり，画像による環境認識を行うために，カメラを組み込んだロボットヘッド（図 3）を製作した．このシステムは，Infanoid Robot[4] とほぼ同一のものである．表 2 にその主な仕様を示す．全ての自由度には小型 DC モータ（Maxon 社 RE シリーズ）を用い，ケーブル駆動を行っている．首部はパン・チルトするだけでなく，2 自由度のチルトを用いて人間の首のように前後に動くことができる．目部は，首部とは別にパン・チルトが可能である．両眼球それぞれにパンの自由度があるため輻輳（寄り目）も可能となっている．両眼球には超小型 CCD カラーカメラ (ELMO 社 QN42H) が 2 個ずつ（広角と望遠）が組み込まれている．また，口部は開閉と上げ下げの自由度を持っており，人とコミュニケーションする時に重要となる発話や感情表現の生成に用いられる．

4 ソフトウェア構成

現在，ASKA には図 4 に示すソフトウェアモジュールが組み込まれており，サーバプロセスを介してソケット通信により相互に情報をやりとりしている．このうちの音声対話関連のモジュール（音声認識理解，音声合成）については音情報処理学講座が，またそれ以外のモジュール（胴体ジェスチャ，頭部ジェスチャ，レーザースキャナによる人検出など）はロボティックス講座が開発を担当した．

これらのモジュールは，定められたプロトコルを

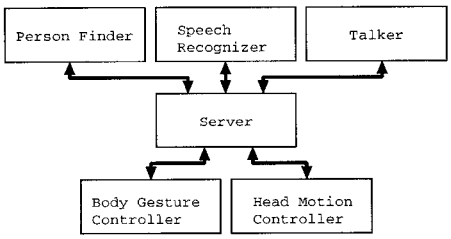

図 4: ASKA ソフトウェア構成図

用いてサーバプロセスとのメッセージ交換することで，他のモジュールとの協調作業を実現している．モジュールが互いに独立に動作するため，複雑な同期処理には向かないが，システムの開発を容易にできる．また，モジュール内に閉じた開発を行なえるため，新たな技術の組み込みを手軽に行なえるメリットがある．

5　対話機能の概要

ASKA の対話は，ユーザ（来客）による

- 教官および研究室の場所と内線番号
- 学内および周辺の施設
- ASKA に関する事柄
- その他，いくつかのあいさつ

のような質問を音声認識の結果を用いて意味理解した後，適切な応答を作成，音声とジェスチャを交えた案内をすることで実現している．実際の ASKA の対話の様子を図 5 に示す．

ASKA の対話の処理の流れは，おおまかに以下のようになる．

1. レーザースキャナによって ASKA の前に立つユーザを検知し，ASKA からの距離と角度を得る．
2. ユーザ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．同時に顔をユーザに向け，質問を受け付けることができる状態であることを示す．
3. ユーザが ASKA に質問をする（音声入力）．
4. 音声認識理解部が入力音声に対する応答文を作成，結果をサーバに送信する．
5. 音声合成部は，応答文から合成音声を作成，発話待ちの状態で待機する．
6. ボディと頭部のジェスチャ部は，応答文などの必要パラメータがサーバに入力されたのを検知し，ジェスチャの動作パターンに基づいて動作を開始する．
7. 音声合成部は，ジェスチャーと同時に発話を開始する．
8. ユーザからの発話待ち状態に戻る．

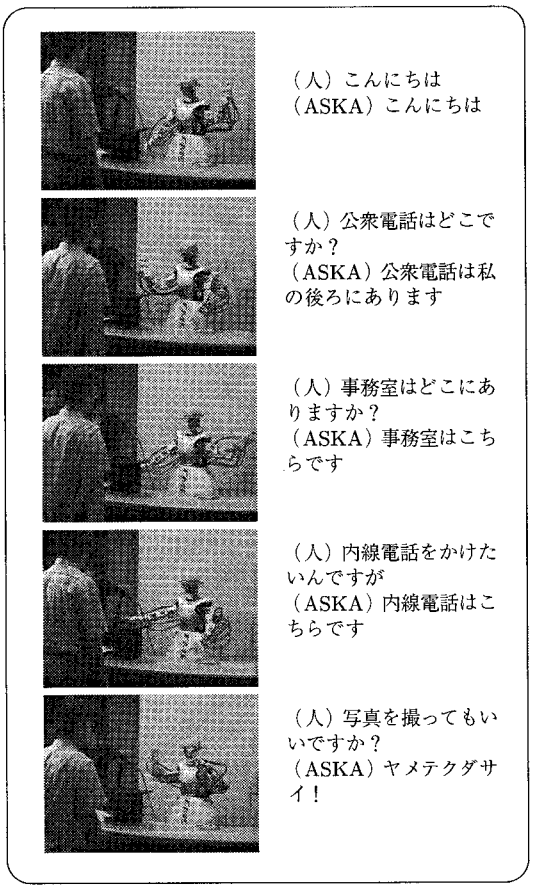

図 5: ASKA と人の対話の例

6　ジェスチャ生成システム

ASKA のジェスチャは，頭部制御部およびボディー制御部により生成される．頭部制御部は，通常はレーザスキャナーで検出されたユーザ（複数いる場合にはそのうちの最も近いユーザ）の方向を向くように頭部を制御している．また，ASKA がユーザへ応答する場合には，音声対話システムの出力に同期して口の開閉を行う．

また，ボディー制御部には，あらかじめ前後左右をはじめとした様々な方向を指し示す動作等，約 30 種類の動作が登録されている．これらの動作から，音声対話システムの出力内容に合うものが選択され，音声出力に同期して実行される．

7　音声対話システム

7.1　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と音声合成の各モジュールによって構成される．音声認識理解部は，入力音声の音声認識，意味理解，応

答の生成を担当する．意味理解及び応答の生成は，後述するように，あらかじめ用意した応答文のテンプレートの中から，認識結果とキーワードの一致回数を用いて，適切な応答文の選択をすることで実現している．音声合成部は，TTS（Text To Speech）プログラムを用いた応答音声の生成を行う．この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」[6]を用いた．

以下では，音声対話システムの核となる音声認識理解部を中心に解説を行なう．

7.2 大語彙連続音声認識と言語モデル

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の記述が必要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，統計言語モデルは，大量の学習用テキストからコンテキストに依存した単語の出現確率を統計的に学習し，その確率を用いて単語の出現を推定するため，柔軟に様々な発話を受理することが可能となる．

ASKAでは，自由な発話を柔軟に受理するために，広く使われている統計言語モデルであるN-gramモデルを用いた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で開発されているJulius [7, 8]を用いた．

7.2.1 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なる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言語モデルでは，十分な性能を得ることができない．そこでASKAのための言語モデルを新たに作成した．

言語モデルの統計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）を持つサイト上のWebページを収集ロボット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年間の学内の学生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を元に作成したテキストである．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[9]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．

表3: 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

情報科学センターはどこにありますか？
木戸出先生の研究室はどこにありますか。
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。

図6: 評価用テキスト（受付での質問）の例

収集の結果，作成した学習用テキストの緒元を表3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．今回収集できたテキストの容量は新聞記事1年半弱に相当する．

学習用テキストからの言語モデル構築には，Palmkit[10]及び「連続音声認識コンソーシアム2000年度版ソフトウェア」[11]に収録の各種ツールを用いた．作成したモデルは，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，学習用テキスト中の出現頻度の上位2万語である．

7.2.2 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．評価用テキストとして，図6に示すような奈良先端科学技術大学院大学の受付案内での質問を想定したテキスト（150文，総単語数1697個）を作成した．

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，雑音の少ない状態でのASKA設置予定の受付である．また，実際の使用状況に合わせたPCの音声入力による録音を行なった．

音声認識エンジンには，Julius 3.2[7, 8]，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

100 おはようございます。
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。
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 7: 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 8: is-staff ファイルの例

ウェア」[11] に含まれる日本音響学会新聞記事読上げ音声コーパス（JNAS）から学習した PTM（Phonetic Tied Mixture） triphone の HMM（Hidden Markov Model）音響モデル（男性用, 64 混合, 3000 状態）を用いた．

実験の結果，3-gram 単語パープレキシティは 29.44，未知語率は 0.30%，認識率（単語正解精度）は 89.80% であった．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は 90.55%であった．

実験結果より，全ての評価尺度において高い性能を確認した．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%以上の高い認識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回作成した言語モデルの組合わせによって，ASKA の音声対話システムの構築は可能であると判断した．

7.3 応答文の作成

ASKA では，来客の質問に対する応答文のリストをあらかじめ用意している．学内アンケートにより ASKA に受け付けてほしい事項を調査し，その結果，必要性が高いと思われる応答文を作成している．また，応答文の追加削除は容易であり，必要に応じて適時追加を行なっている．

現在，ASKA に登録されている応答文の数は，61 個である．図 7 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，定型なものと他のデータベースからデータを挿入できるもの（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

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 9: キーワードリストの例

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．

以下の文章は，挿入型応答文の例であり <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 8 に示す．先頭のインデックス番号は 0 番目として数える．

7.4 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 9 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 9 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．

8 今後の開発方針

このシステムは，本年 7 月 31 日に奈良先端科学技術大学院大学で開催されたイベント「ロボフェスタ生駒」においてお披露目された．この時の様子を図 10 に示す．

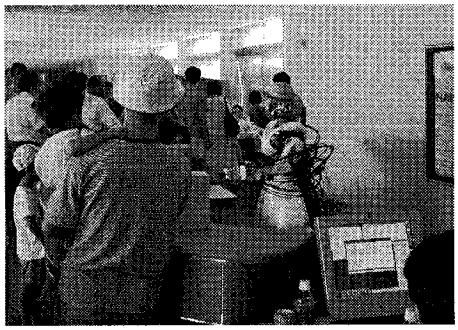

図 10: ロボフェスタ生駒でのデモンストレーション

デモンストレーションを行った場所は本学の情報科学研究科受付である．デモ当日は観客が多かったための背景雑音が多く，音声認識の誤認識が生じた．この対策として，現在は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，今後はマイクロホンアレーによる音声入力デバイスの実装を予定している．また，オープンキャンパスには子供が多数来場するが，現在の成人音声から学習した音響モデルでは子供の声の認識は困難である．子供も ASKA を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．さらに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．

また，画像処理の応用として，レーザースキャナを利用した話者の検知の代わりに，頭部のカメラを用いたアイコンタクト [12] を用いることを検討している．

9 おわりに

現在の ASKA は，情報科学研究科の 2 つの研究室が共同でシステムを構築し，デモができるようになったところである．今後はさらに多くの研究室に参加してもらい，得意な技術を持ち寄ってシステム構築を進めていく予定である．具体的には，まず画像処理で人や顔を認識してコミュニケーションをはじめる機能を組み込みたい．また，音声認識で分かったように，実環境でデモを行うとシステムの問題点が明らかになってくる．それを各研究者が持ち帰って研究として取り組み，改良された技術をさらに統合するというフィードバックを常にかけられるような研究プラットフォームとして，この受付案内ロボットを利用していきたい．

謝辞

ロボット頭部の設計図および制御回路を提供して下さった通信総合研究所 小嶋秀樹氏，テムザック IV 用の Linux 用デバイスドライバを提供して下さった産業総合研究所 原功氏，およびテムザック IV の内部情報を開示して下さった（株）テムザックに感謝いたします．

奈良先端科学技術大学院大学・情報科学研究科 植村 俊亮研究科長，鹿野 清宏教授，小笠原 司教授をはじめとする本開発プロジェクトをご支援いただいているみなさまに感謝いたします．

参考文献

[1] R. A. Brooks et al.:"The Cog Project: Building a Humanoid Robot," In C.L.Nehaniv, ed., "Computation for Metaphors, Analogy and Agents," Vol. 1562 of Springer Lecture Notes in Artificail Intelligence, Springer-Verlag, 1999.

[2] S. Kotosaka et al.: "Humanoid robot DB;" In Proc. of Int. Conf. on Mach Automat (ICMA2000), pp.21-26, 2000.

[3] http://www.tmsuk.co.jp/

[4] H. Kozima, H. Yano: "A Robot that Learns to Communicate with Human Caregivers," The First International Workshop on Epigenetic Robotics, 2001.

[5] http://www.sick.co.jp/

[6] http://www.createsystem.co.jp/linux.html

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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.