音声言語情報処理 45－6
（2003．2．7）

生駒市コミュニティセンター音声情報案内システムの開発と運用

西村 竜一[†]　西原 洋平[†]　鶴身 玲典[†]　李 晃伸[†]　猿渡 洋[†]　鹿野 清宏[†]

あらまし　生駒市北コミュニティセンターの音声情報案内システム「たけまるくん」を開発した．本システムでは，大語彙連続音声認識を利用した一問一答形式の音声対話により，同センターや生駒市に関する案内を行うことが可能である．実用化を目指した本システムは，2002年11月6日からセンター内に常設され，開館時は誰でも自由に愛嬌のあるエージェントとのコミュニケーションを楽しむことができる．また，システムの改良に必要な対話記録を実際の運用を通じて収集し，発話内容の書き起こし等のデータの整備もすすめている．本稿では，主に本システムの構成および発話音声データ収集の状況について報告する．また，成人による比較的クリーンな発話をテストセットにした本システムの評価実験を行い，84%の単語正解率と70%の応答正解率を確認した．

キーワード 音声対話システム, ソフトウェアエージェント, Julius, 発話データの収集

Development and Operation of Ikoma Community Center Speech-oriented Information System

Ryuichi NISIMURA[†]　Yohei NISHIHARA[†]　Ryosuke TSURUMI[†]
Akinobu LEE[†]　Hiroshi SARUWATARI[†]　Kiyohiro SHIKANO[†]

Abstract　We implemented a practical speech guidance system for public use. It is called "Takemaru-kun", and located daily at the entrance hall of Ikoma Community Center to inform visitors about the center and around Ikoma city via speech human-machine interface and funny animating agent of Takemaru. This system aims to promote a field test for robust speech recognition in practical environment, and to collect actual utterance data in the framework of human-machine speech dialogue. The system has been running everyday since November 6, and a large number of user utterances have been collected. Classification and transcription of the data is also undertaken. This paper reports the outline of this system and current status of the data collection. In a recognition experiment with extracted samples of adult voices, word accuracy of 84% and answer rate of 70% was obtained.

Key words　Speech dialogue system, Software agent, Julius, Collection of utterance data

1　はじめに

ヒューマンロボットコミュニケーション [1] や擬人化エージェントシステム [2] など音声インタフェースの応用に対する社会からの注目はますます高くなってきている．しかし，現状では，実用的であり利用価値の高いシステムの実例は少ない．また，一般の人々が気軽に利用できる音声インタフェースの応用システムでさえ少数である．

著者らは，これまで N-gram 言語モデルを用いた大語彙連続音声認識を基盤とした音声対話機能を持つ受付案内ロボット ASKA（アスカ）[3] の開発を通じて，音声インタフェースの実用化の検討を行ってきたが，ロボットのハードウェア保守などの問題から ASKA を日常的に運用することは困難であった．このため，実際に一般のユーザが利用できる時間は，本学のオープンキャンパスのデモンストレーションなどのごく少ない機会に限られていた．また，ASKA の運用時には説明員が必要など，ASKA は実環境下で日常的に利用できるシステムとは言い難い．このため，ユーザとシステムとのインタラクションにおける記録データの収集が不十分であった．実用化を目指したシステムの改良をすすめていく上で，フィールドテストによる実際の記録データは必要不可欠であり，その収集のための実環境での日常的なシステムの運用は，実現すべき重要な課題である．

そこで，われわれは，生駒市の協力のもと「生駒市北コミュニティセンター ISTA はばたき」（奈良県生駒市上町 1543 番地）の音声情報案内システム「たけまるくん」を開発した．

「生駒市北コミュニティセンター ISTA はばたき（以下，北コミュニティセンターと略）」は，2002年11月6日にオープンした市民向け多目的コミュ

[†] 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

図 1: システム外観

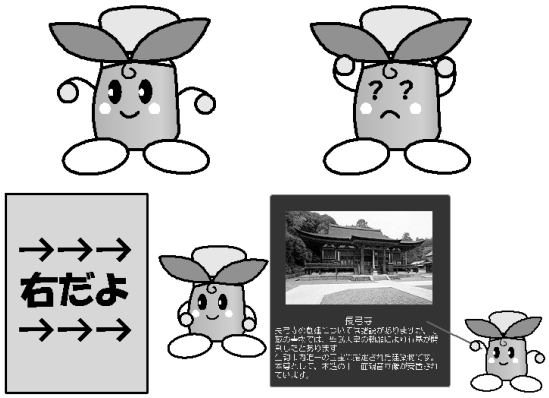

図 2: たけまるエージェントの例

ニティセンターであり，456 人収容できる大ホール（はばたきホール）や小ホール，セミナー室，和室，図書館などを備える．また，市民サービスコーナーでは，市役所業務の一部も行っている．

開発した本システムは，同センター内に常設され，一般の来訪者がいつでも気軽に利用できる音声インタフェースを持った案内システムである．また，日常的に運用することにより，実際の来訪者の発話音声を収集することができ，収集データはシステムの改良や問題点の検討に利用することができる．さらに，本システムの開発や運用を通じて，音声インタフェースの応用システムの実用化に向けての様々な知見の獲得が期待できる．

本稿では，主に音声情報案内システム「たけまるくん」の構成および来訪者による実際の発話音声データの収集状況について報告する．また，収集データから整備したテストセットを用いて実験した本システムの評価についても述べる．

2　システムの概要と構成

開発した音声情報案内システム「たけまるくん」は，北コミュニティセンターの館内施設や生駒市の観光情報，周辺情報などの各種案内を行うためのシステムである．本システムは，一問一答形式の音声による対話機能を持ち，来訪者の質問に対して，合成音声とアニメーションによるソフトウェアエージェントの応答を用いてガイドを行う．また，同時に World Wide Web（WWW）を利用した関連情報の提示が可能である．

本システムが想定している質問の内容は，以下のようなトピックに関するものである．

1. 北コミュニティセンター館内の案内（部屋や施設の場所など）
2. 業務の内容（手続きの方法や開館時間の案内など）
3. 周辺の案内（駅，バス停，郵便局の場所など）

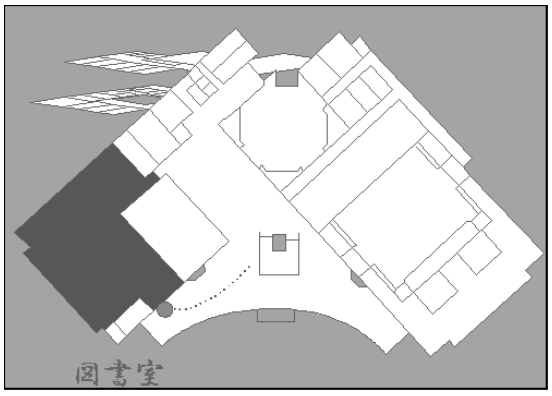

図 3: 場所案内図（図書館の例）

4. 奈良・生駒の観光情報
5. たけまるくん自身に関する情報

本システムの外観を図 1 に示す．ユーザは，机上のマイクに向かって発話した後，応答は，2 個のディスプレイとディスプレイ背面のスピーカから出力される．左のディスプレイに表示されるのは，生駒市のイメージキャラクタ「たけまる」のソフトウェアエージェントである．たけまるエージェントは，図 2 に示すようなアニメーションにより，ユーザにジェスチャを用いた案内を行う．同時に右ディスプレイには，応答内容に関連する Web ページを提示する．図 1 の例では，システムは新聞の Web ページを表示しているが，ユーザは備え付けのマウスを用いて Web ブラウジングを行い，提示された Web ページの中から詳細な情報を得ることもできる．また，右ディスプレイには，図 3 のような案内図や時刻表などを表示することもできる．

本システムのハードウェアは，イーサネットによって接続された 2 台の Linux を OS とする PC で構成されている．また，ソフトウェアは，図 4

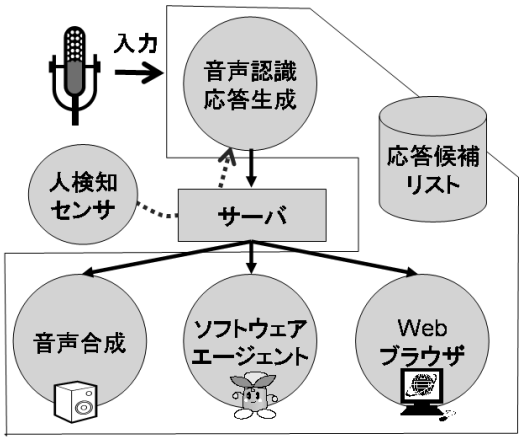

図 4: ソフトウェアモジュールの構成

に示す TCP/IP によって互いに通信するプログラムモジュールによって構成される．サーバは，それぞれのモジュールの出力および動作状態を保持するプログラムである．各モジュールは，サーバに保持されている情報によって他のモジュールの出力結果と動作状態を知ることができる．

本システムの主な処理の流れは以下のようになる（図 4）．まず，音声認識および応答作成モジュールが入力音声の音声認識結果をもとに応答を作成する．応答文の候補リストが，モジュール間で共通のファイルの中でインデックス番号付きで定義されており，音声認識および応答作成モジュールは，生成した応答に対応するインデックス番号をサーバに送信する．音声合成，ソフトウェアエージェント，Web ブラウザの各モジュールは，その動作が応答内容に対してあらかじめ定義されており，サーバ経由で取得したインデックス番号に応じて処理を実行する．

ここで，音声合成モジュールは，Text-To-Speech による応答音声の生成と生成ファイルの再生を行うプログラムである．ソフトウェアエージェントは，たけまるエージェントの表示，Web ブラウザは，応答内容に関連する Web ページの提示を行うモジュールである．たけまるエージェントのアニメーションの作成には，インターネット上で広く普及しているマクロメディア社 [4] の Flash を利用し，Web ブラウザと Flash プレイヤで再生している．なお，本システムは，現在，38 のたけまるエージェントの動作パターンを持つ．

各モジュールは，他のモジュールの動作状態をサーバ経由で取得することによって，自らの動作を決定することができる．例えば，本システムは，赤外線による人検知センサを実装しており，人検知センサモジュールの動作状態を音声認識および応答作成モジュールは，音声認識の録音開始のトリガに利用することができる．

また，ソフトウェアエージェントは，応答のアニメーションのみでなく，音声認識が録音中のたけまるエージェントのうなずき動作を担っている．うなずきのタイミングは，ソフトウェアエージェントが音声認識および応答作成モジュールの動作状況を取得して決定しており，音声認識の録音開始と同時にうなずきを開始する．

3 音声認識プログラムの構成

本システムの音声認識には，大語彙連続音声認識エンジン Julius[5] を用いた．また，本システムの認識タスクに適した言語モデルを得るため，以下に述べる手順でバックオフ N-gram 言語モデルを作成した．

まず，下記の 2 種類のテキストからそれぞれ学習した 2-gram および逆向き 3-gram モデルを作成した．

1. 検索エンジンを用いて収集した生駒市関連および生駒市ホームページ内の Web ページテキスト（統計テキストフィルタ [6] を用いて整形），1,080,272 文，総単語 31,265,487 個，異なり単語 218,723 個
2. 本システムを想定して人手で収集した質問文テキスト，6,488 文，総単語 56,108 個，異なり単語 3,231 個

各モデルの学習に用いた単語は，「1. Web ページテキスト」に関しては，出現頻度の高いものから上位 4 万語であり，「2. 想定質問文テキスト」に関しては，全ての単語（3,231 個）である．具体的なバックオフ N-gram 言語モデルの作成手順は，「日本語ディクテーション基本ソフトウェア（99 年度版）」[7] のものに準じた．ディスカウントには Witten-Bell 法を用いた．

次に，生成されたそれぞれのモデルを N-gram モデルの融合ツール [8] を用いてマージした．融合の割合は，1:1 である．この結果生成された 2-gram と逆向き 3-gram 言語モデルを以後はベース言語モデルと呼ぶ．

続けて，ベース言語モデルにネットワーク文法による接続制約を適用した文法適用言語モデルの作成を行った．適用に用いた文法は，本システムのタスクのために人手で作成したものであり，異なり単語数は 441 である．今回の適用は，ベース言語モデル中の N-gram エントリに含まれる 2 単語の対が，用意した文法により受理可能な場合，そのエントリの持つ確率値を強制的に上げることで実現した [9]．具体的には，2-gram エントリに関しては，その単語対が文法で受理可能なもの，3-gram エントリは後ろの単語対が受理可能なものに対して，それらの対数尤度値を 0.55 倍した．1-gram（単語）に関しては，文法にその単語が定義されているものを対象とした．

また，文法では定義されているが，ベース言語モデルには含まれない 43 単語に関しては，単語

101 こんにちは。
208 今は、<<hour>>時<<min>>分です。
212 バスの時刻表を表示します。
301 トイレは、左の奥か、はばたきホール、入り口の近くにあります。
305 図書館の入り口は、市民サービスコーナーの横です。

図 5: 応答候補の例

トイレ+トイレ+2 は+ワ+65 、+、+79 どこ+ドコ+14 です+デス+74/56/1 か+カ+70 ?+?+77#301
食堂+ショクドー+2 は+ワ+65 、+、+79 あり+アリ+47/17/5 ます+マス+74/58/1 か+カ+70 ?+?+77#332

図 6: 用例テキストの例

辞書中の未知語クラスのエントリに対して，43 単語の出力表記と読み（音素記号）を与えることで文法適用言語モデルに追加した．

4 応答の生成

本システムでの応答の生成は，音声認識結果の形態素列とデータベース内の用例テキストの形態素列とのマッチングを行い，その一致数のカウントによるスコア計算によって，対応する応答候補を選択することで行われる．

図 5 は，応答候補ファイルの例である．各行の行頭三桁の数字は，2 節で述べたモジュール間の通信に用いるインデックス番号である．なお，応答候補には，定型文の他にスロットにパラメータを代入できるものが作成可能である．図 5 の 2 行目，括弧（<<>>）で囲まれた箇所がスロットの例である．現在，202 の応答候補が本システムに登録されている．また，そのうち，スロット型の応答候補は，時間や日付などに関する応答の 3 つである．

用例テキストは，あらかじめ想定されたシステムへの質問文の形態素列である（図 6）．全ての用例テキストには，前述の応答候補の中からその質問に対する応答として最も相応しいものが定義されている．具体的には，図 6 中の # の後に書かれた数字が，その用例テキストに対応づけられた応答候補のインデックス番号である．なお，現在は，アンケートや市役所業務の記録などから作成した形態素解析済みの 2,309 文を用例テキストとして登録している．

図 7 は，用例テキストベースのスコア計算の概略である．入力された音声認識結果に対して，全ての用例テキストについて形態素の一致数を求め，対応する応答候補にスコアとして加算する．この時，カウントに用いるのは，質問の意図理解に重

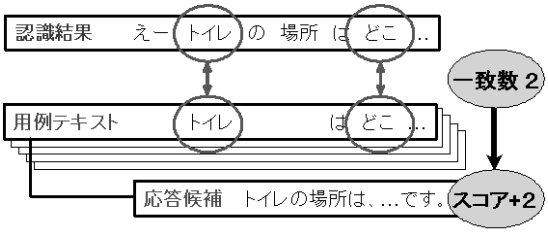

図 7: 用例テキストベースのスコア計算

要な名詞，動詞，形容詞などの自立語形態素のみである．全ての用例テキストに対して一致数のカウントを行い，その結果，最も高いスコアを持つ応答候補を応答結果とする．もし複数の応答候補のスコアが同じ場合，応答はそれら応答候補の中からランダムに選択される．

本手法では，用例テキストの数を増やすことで，発話の言い回しや語句の違いなどの様々な表現様式に対して柔軟に対処できるのが特徴である．

また，入力である音声認識結果には，N-best 出力結果を用いた．本来の正解単語が 1-best 結果では出力されない場合でも，2-best 以降に出力される本来の正解単語をスコア計算に含めることができる．実際の運用では，100-best 結果を使用した．

なお，本システムでは，応答候補ごとにキーワードを定義することもできる．音声認識の結果に含まれるキーワードの数は，前述の応答候補のスコアとして加算される．このキーワードリストは，応答候補に対して用例テキストの数が少ないの時に必要であり，不十分を補うことができる．

5 音声データの収集

本システムは，北コミュニティセンターのオープン日である 2002 年 11 月 6 日から運用を開始した．11 月 8 日からは，Julius 3.3p2 以降に含まれる入力音声の録音機能を用いて，ユーザの発話音声は発話開始時間の情報付きで記録している．雑音のみの入力や正しく音声が収録されていないものも含めて，2002 年 12 月 26 日までの休館日を除く 40 日間に記録されたデータは 22,613 個であり，そのファイルの総容量は 1,327,364 キロバイト（約 1.3GB）になる．これは単純計算で，約 708 分の長さの音声に相当する．

同時に，人手による収集データの整備をすすめている．この作業では，削除可能な雑音の切り出しによる除去，発話内容のテキストへの書き起こし，音声のみからの主観による話者の性別および年齢層のラベル付けを行っている．2002 年 12 月 26 日現在で整備済みのデータは，12 月 6 日分までの 16,070 個である．このうち，雑音が多少含まれるが，発話を明瞭に聞き取れる比較的クリーンな音声データは，8,285 個であった．このデータの年齢層および性別の分布を表 1 に示す．この結

表 1: 収集音声の年齢層と性別ごとの発話数

年齢層		男性	女性	性別不明	合計
a)	幼児	22	240	100	362
b)	低学年子供（小学校 3 年生ぐらいまで）	1100	2165	762	4027
c)	高学年子供（中学生ぐらいまで）	122	115	15	252
d)	成人	2869	718	15	3602
e)	高齢者	7	1	0	8
x)	年齢層の判断ができなかったもの	0	5	29	34
合計		4120	3244	921	8285

（性別不明: 声から性別を判断できなかったもの）

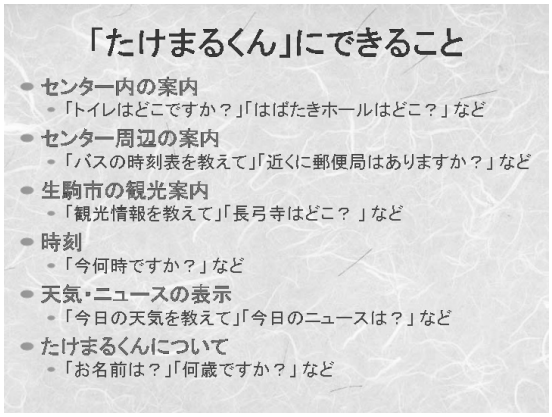

図 8: ユーザへのガイト

果，本システムでは，男女を問わず広い年齢層から発話音声データの収集ができることがわかる．

6 テストセットの作成

本システムの性能の指標を調べるため，収録音声からテストセットデータを作成した．テストセットには，前述の比較的クリーンな音声データ 8,285 個の中から，成人および高齢者の男女の発話による計 500 文（男性 280 文，女性 220 文）を選択した．なお，本システム周辺には，ユーザへの発話を促すため，発話例を示した図 8 のような紙のガイドを掲示しているが，実際のユーザの発話内容はこのガイドに大きく影響されており，発話内容に偏りが生じた．掲示する発話例を適時変更して対処しているが，今回の収集データでも発話内容の偏りを確認した．テストセット作成時には，この偏りを防ぐため，書き起こしテキストをキーにしてソートを行い，同じ発話内容のデータは間引いた後，データをランダムに選択している．作成したテストセットに含まれる発話内容の例を図 9 に示す．

なお，このテストセット（500 文，総単語数 3,223 個）に対する 3 節で述べたベース言語モデルの 3-gram テストセット単語パープレキシティは 16.0，

はばたきホールはどこにありますか？
トイレに行きたいんですけど。
スカイランド生駒に行きたいんだけど。
長弓寺について教えてください。
たけまる君は何歳ですか？
えー今何時ですか？
今日のお天気は？

図 9: 発話内容の例

未知語率は 0.81%であった．また，ベース言語モデルに含まれない 43 単語を追加した文法適用言語モデルにおける未知語率は 0.34%であった．

7 実験

評価実験では，作成したテストセット 500 文に対する音声認識性能と，本システムがどの程度正しく応答できるかの指標を調べた．

7.1 音声認識実験

Julius[5] による大語彙連続音声認識実験を行った．実験条件は，実際に本システムを運用時と同様のものを用いた．音響モデルは，日本音響学会新聞記事読み上げ音声コーパス（JNAS）[10] のクリーン音声に 25dB SNR で電子協騒音データベース [11] の展示会場の雑音を重畳した音声から学習した PTM[12] triphone の性別非依存 HMM モデルである．言語モデルには，実際の運用には 3 節で述べた文法適用言語モデルを利用しているが，比較のためにベース言語モデルの場合の実験結果も示す．

単語正解率と単語正解精度を表 2 に示す．文法適用言語モデルを使用することで，ベース言語モデルよりも高い認識精度を示すことがわかる．また，今回は雑音の少ないクリーンな音声をテストセットに用いて，80%以上の認識率を得ることができたが，実際の運用を考慮すると，さらに雑音のまじった音声などでの評価が必要である．さらに本システムのユーザには子供も多いことから，子供音声での評価も重要な検討事項である．

表 2: 音声認識実験の結果

言語モデル	単語正解率 (%)	単語正解精度 (%)
ベース	81.5	76.9
文法適用	84.4	78.9

7.2 対話実験

次に，応答性能のついて調べた．4 節で述べた本システムの応答生成のプログラムに，音声認識の 100-best 認識結果を入力し，結果を集計した．結果の応答内容が満足なものかは人の主観により判別した．なお，実験に用いた実験条件は，7.1 と同様である．ただし，言語モデルには文法適用言語モデルを用いた．

実験の結果，満足な応答を得られたものは，500 文のテストセット中 348 文で，69.6%の応答正解率であった．

8 まとめと今後の課題

本報告では，生駒市北コミュニティセンターの音声情報案内システム「たけまるくん」について述べた．また，本システムを用いて，実際のユーザの発話データを収集することができることを示した．実験では，収集データから整備したテストセットを用いた本システムの評価を行った．

前述のように，今回の実験に用いたのは雑音の少ない比較的クリーンな音声データによるテストセットである．収集したデータには，雑音のはげしいもの，不明瞭な発話や音声がオーバフローしているもの，子供の声など様々な実際の発話が含まれる．本システムの現状では，これらの発話に対する対処は十分ではなく，正しい動作を期待することはできない．収集した音声からデータの整備をさらにすすめ，様々な条件のテストセットを作成し，それらの評価を行うことが当面の今後の課題である．その結果を検討し，必要な雑音対策や子供声認識などの音声認識技術の導入を行うことを予定している．

また，収集したデータは，音声認識性能の向上だけではなく，対話処理の改良にも利用する予定である．特に，応答のバリエーションが限られていると，次第にユーザからの関心を失うことになり，結果としてデータの収集が出来なくなる．今後も発話データの収集を行うためには，収集した発話データをもとにした応答内容の多様性の充実は必要不可欠である．

最後に，最も重要なことは，継続的な本システムの運用であると考えている．一般のユーザのシステムに対する扱いは非常に雑であり，サービスを提供し続けることは，多大なる労力を必要とする．しかし，運用を通じて，有用なシステムの実現を目指した様々な観点からの検討を行っていきたいと考えている．

謝辞 本システムを開発，運用する上で生駒市市役所 白本 和久氏，北コミュニティセンター館長 米田 秀一氏をはじめとする生駒市職員のみなさまに多大なるご支援をいただいた．ご協力いただいた関係各位に深く感謝する．

参考文献

[1] 小林: “マルチモーダルインタフェースを持つ会話ロボット,” 2001 年電子情報通信学会総合大会講演論文集（基礎・境界）, pp.504–505, SA-7-3, 2001-3

[2] 川本, 下平, 新田, 西本, 中村, 伊藤, 森島, 四倉, 甲斐, 李, 山下, 小林, 徳田, 広瀬, 峯松, 山田, 伝, 宇津呂, 嵯峨山: “カスタマイズ性を考慮した擬人化音声対話ソフトウェアツールキットの設計,” 情報処理学会論文誌, Vol.43, No.7, pp.2249–2263, 2002

[3] R. Nisimura, T. Uchida, A. Lee, H. Saruwatari, K. Shikano, Y. Matsumoto: “ASKA: Receptionist Robot with Speech Dialogue System,” In Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2002), pp.1314–1317, 2002

[4] http://www.macromedia.com/jp/

[5] 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

[6] 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

[7] 河原, 李, 小林, 武田, 峯松, 嵯峨山, 伊藤, 伊藤, 山本, 山田, 宇津呂, 鹿野: “日本語ディクテーション基本ソフトウェア (99 年度版),” 日本音響学会誌, Vol.57, No.3, pp.210–214, 2001

[8] 長友, 西村, 小松, 黒田, 李, 猿渡, 鹿野: “相補的バックオフを用いた言語モデル融合ツールの構築,” 情報処理学会論文誌, Vol.43, No.9, pp.2884–2893, 2002

[9] 鶴身, 李, 猿渡, 鹿野: “タスク文法による N-gram 確率の部分強化を用いた認識アルゴリズムの評価,” 情報処理学会研究報告, 2003-SLP-45-13, 2003

[10] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi: “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” The Journal of the Acoustical Society of Japan (E), Vol.20, No.3, pp.199–206, 1999.

[11] 板橋: “騒音データベースと日本語共通音声データ DAT 版,” 日本音響学会誌, Vol.47, No.12, 951–953, 1991

[12] 李, 河原, 武田, 鹿野: “Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識,” 電子情報通信学会論文誌, J83-D-II No.12, pp.2517–2525, 2000