2009－SLP－75 (9)
2009／2／7

複数の言語モデル・言語理解方式を用いた音声理解の高精度化

勝丸真樹† 中野幹生‡ 駒谷和範† 成松宏美†† 船越孝太郎‡
辻野広司‡ 高橋徹† 尾形哲也† 奥乃博†

† 京都大学大学院 情報学研究科 知能情報学専攻
‡ (株) ホンダ・リサーチ・インスティチュート・ジャパン
†† 津田塾大学 学芸学部 情報数理科学科

音声対話システムでは，学習データや発話によって適した言語モデル・言語理解方式が異なる．そのため最適なモデル・手法を選び音声理解部を構築することは容易でない．本稿は，複数の言語モデルと言語理解方式とを用いて複数の理解結果を得ることにより，それらから最も良い結果を選択したり，文脈理解部で複数の結果を扱える枠組みを提案する．本枠組みの一つの実装として，言語モデルは文法モデルと単語 N-gram モデルの 2 種類，言語理解方式は FST と WFST，キーフレーズスポッティングの 3 種類を用いて，それらの任意の組合せを用いて音声理解を行い，それらの結果から，発話ごとに適した理解結果を動的に選択し，最終的な理解結果を得るような音声理解システムを構築した．評価実験の結果，単一の言語モデル・言語理解方式を用いたときと比較して言語理解精度が向上することが確かめられた．

キーワード 音声対話システム，複数の言語モデル，複数の言語理解方式，音声理解結果の選択

Improving Speech Understanding Accuracy by Using Multiple Language Models and Language Understanding Methods

Masaki Katsumaru†, Mikio Nakano‡, Kazunori Komatani†,
Hiromi Narimatsu††, Kotaro Funakoshi‡, Hiroshi Tsujino‡, Toru Takahashi†,
Tetsuya Ogata† and Hiroshi G. Okuno†

†*Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University*
‡ *Honda Research Institute Japan Co., Ltd.*
†† *Tsuda College*

Optimal language models (LMs) and language understanding (LU) methods for spoken dialogue systems vary depending on available training data or utterances to handle. Finding their optimal combination is difficult because much data and expertise are required. We developed a framework for improving speech understanding accuracy under various situations by using multiple LMs and LU methods. As its experimental evaluation, We used two LMs such as grammar-based and statistical models, and three LU methods such as finite states transducer (FST), weighted FST (WFST) and keyphrase-spotting. Six speech understanding results are obtained by combining these models and methods, and the most appropriate one was dynamically selected by a decision tree for each utterance. We showed that our method improved speech understanding accuracy compared with those obtained from any combination of single LM and LU method.

Index Terms: spoken dialogue systems, multiple language models, multiple language understanding methods, select an appropriate speech understanding result

1. はじめに

音声対話システムでは，ユーザの発話から意味表現を生成する音声理解部が重要な働きをする．本稿では，意味表現はコンセプトの集合で表されたものを用いる．音声理解は，音声を単語列に変換する音声認識と，単

語列から意味表現を抽出する言語理解の二つのプロセスからなる．音声認識には音響モデルと言語方式が必要であるが，音響モデルは音声対話システムのタスクドメインには依存しないので，システム開発者は，ドメインに応じて，言語モデルと言語理解方式を用意する必要がある．ユーザの多様な発話に対して頑健な音声理解を行うには，音声認識用の言語モデルと言語理解方式の適切な選択・組合せが必要である．

これまで用いられてきた言語モデルとしては，ネットワーク文法（以下文法）や，コーパスから統計的に学習する言語モデルである N-gram モデルなどがあるが，それぞれ一長一短がある．文法は人手で記述する場合が多く，その場合学習データが不要である．しかし，複雑なタスクでは，高いカバレージと高い予測性能を両立する文法規則の構築にはスキルが必要である．また，想定外の発話に対して頑健でない．N-gram モデルは，文法ベースの言語モデルと比較すると，局所的な制約であり，未登録語や認識誤りが生じても回復が容易であるという利点がある．しかし N-gram モデルを十分推定できるだけの訓練データが必要となる．

言語理解方式としては，音声認識結果に現れるタスク達成に重要な句を単純にすべてコンセプトに変換するキーフレーズスポッティング (以下スポッティング) や，意味文法に基づいて認識結果を解析する Finite State Transducer (FST)，FST に重みの概念を与えた weighted FST (WFST) があり，これも一長一短がある．スポッティングによる言語理解は，構築が簡単であるが，音声認識結果に誤りを含む場合に誤ったコンセプトに変換してしまう．FST による言語理解では，音声認識結果を，意味文法を表現した FST に入力してコンセプトに変換する．言語理解のための意味文法は人手で記述される場合が多い．WFST に基づく言語理解では，音声認識誤りなどで文法構造から外れた音声認識結果に対しても頑健に言語理解結果を出力できる．しかし，WFST では重みの推定に多くのデータが必要である．どのような言語理解方式が適しているかは，音声認識に用いる言語モデルによっても異なるため，言語モデルと言語理解の組を適切に選択することが必要である．

従来の多くの音声理解研究では，ATIS のような，あらかじめ与えられた訓練データとテストデータを用いて，最も高い理解性能を得られる方法の提案が行われてきた．しかしながら，実際の音声対話システムの構築では，統計的言語モデルや言語理解のパラメータ推定に使える訓練データの量や，文法構築にかけられる労力などが，人員や予算などシステム開発の制約によって異なる．したがって，そのような制約の中で，どのようにして高い理解性能を持つ音声理解部を構築するかが課題である．

従来のシステム開発においては，多くの場合，システム開発者が試行錯誤に基づき言語モデルと言語理解方式の選択を行ってきた．しかしながら，上で述べたように，各言語モデルと言語理解方式は一長一短があるため，非常に大量の訓練データや十分なシステム開発人員がある場合を除き，単一の手法で高い性能を出すことは難しい．したがって，複数の音声理解方式を用いることが有効だと考えられる．各理解方式がうまく理解できる発話は異なるので，正しい理解結果が含まれる可能性が高くなるからである．今までに，複数の言語モデルや言語理解方式を使う方法が提案されているが，言語モデルと言語理解方式のどちらかのみを複数用いる方法であり，さまざまな開発の制約下で，十分な性能が得られるとは考えにくい．

本稿では，システム開発時の制約の下でできるだけ高精度な音声理解を行うための音声理解フレームワーク MLMU (Multiple Language models and Multiple Understanding models) を提案する．MLMU では，複数の言語モデルと複数の言語理解方式を用いてユーザ発話を理解し，複数の理解結果を出力する．そのような複数の理解結果から，一つの理解結果を自動的に選択したり，理解結果の信頼度を計算したりする．どのような言語モデル・言語理解方式を用いるかは，開発の制約に応じ，対話のタスクドメイン毎にシステム開発者が指定できる．たとえば，文法開発に労力をかけることができれば，文法ベースの言語モデルを指定し，統計言語モデルを作ることができれば，統計言語モデルを指定する．

我々は，このフレームワークの一実装として，2 種の言語モデルと 3 種の言語理解方式を用いた音声理解を，マルチドメイン対話システム構築ツールキット RIME-TK (Robot Intelligence based on Multiple Experts Tool Kit)[1] 上に構築した．また，複数出力される音声理解結果から，適切な理解結果の選択を行う識別器を実装した．この識別器は，音声認識時の特徴と言語理解時の特徴を用いる．この識別器を用いることにより，単一の言語モデル・言語理解方式を用いたときと比較して，高精度な音声理解ができる．

2. 関連研究

異なる言語モデルを用いた音声認識を複数用いる研究として，ROVER 法[2] がある．ROVER 法では，複数の音声認識の結果から多数決を用いて音声認識結果を選択する．また，安田ら[3] は，二つの文法を用いて音声認識を用い，どちらの認識結果を用いるかを識別する決定木の学習方法を提案している．これらの研究で

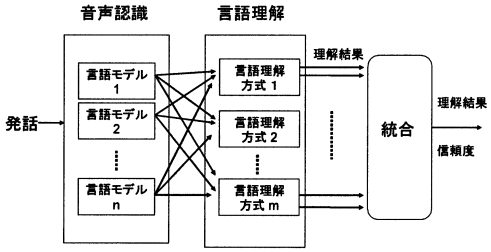

図 1 本手法における音声理解の流れ

```
<slu id = "grammar-fst" decoding="grammar1"
      lu="fst1" />
<slu id = "classlm-wfst" decoding="classlm1"
      lu="wfst1" />
<integration selected="utter-veri"
      slu1="grammar-fst" slu2="calsslm-wfst" />
```

図 2 RIME-TK における音声理解の指定例

は音声認識結果を一つに選択してしまうので，異なる音声認識結果に対して，別々の言語理解を適用することができない．

また，発話検証のために複数の言語モデルを用いる方法が提案されている[4),5)]．言語理解用の言語モデルとは異なる言語モデルで音声認識を行い，音響尤度などを比較することで認識結果の信頼性などを計る．これらの方法でも，言語理解のために用いる音声認識結果は単一の言語モデルに基づく結果だけである．

複数の言語理解方式を用い，最適な理解結果を選択する研究もある[6),7)]．これらの方法では，大規模コーパスを前提としており，そのコーパスから学習した単一の言語モデルで音声認識を行っている．現実のシステム開発ではそのような大規模コーパスはいつも用意できるとは限らない．訓練データがあまりなく，有効な言語モデルが明らかでないときでも高精度な言語理解を実現するには，複数の言語モデルを用いる必要がある．ドメイン毎に異なる音声理解方式を用いる研究[8)]もあるが，各ドメイン内での音声理解方式は単一のものなので，頑健性がない．

3. 複数の言語モデル・言語理解方式を用いる発話理解フレームワーク MLMU

本稿で提案する MLMU は，言語モデルと言語理解方式の組み合わせである音声理解方式を複数用いることができるフレームワークである．システム開発者は，タスクドメイン毎に，用意できる言語モデルと言語理解方式を列挙することによって，複数の音声理解方式を用いて発話理解を行わせる．さらに，各音声理解が出力する理解結果から，一つの理解結果を自動的に選択したり，理解結果の信頼度を計算したり，理解結果のランキングを行ったりすることが可能なフレームワークである (図 1)．

複数の音声理解方式を用いることで，単一の音声理解方式よりも高精度な音声理解が可能である．たとえば，文法でカバーされる発話に対しては，文法と FST による音声理解の性能が高い場合が多いが，カバー率と予測性能の高い文法が常に構築できるとは限らない．文法でカバーできない発話に対しては，統計的言語モデルとスポッティングや WFST が有効である場合が多い．このように，音声理解方式毎に得意とする発話が異なるため，複数の音声理解を適用することにより，理解結果の中に正解が含まれる可能性が高まる．したがって，これらの理解結果の中から適切な理解結果を選択したり，理解結果を統合したりすることにより，理解精度の向上が可能となる．選択時には，音声認識・言語理解時の特徴だけでなく，神田ら[9)]や Abdou ら[10)]のように対話管理部から得られる文脈情報を用いることも効果的であると考えられる．また，複数の音声理解結果から，ある一つの音声理解結果に絞ることは必須ではない．複数の候補を保持しておき，ユーザとの対話を通して最終的な理解結果を同定することも考えられる．さらに，理解結果の一致度などを調べたり，ランク付けや多数決を行うことで信頼度の算出が可能となる．言語理解に信頼度を付与することで，その信頼度に基づいた効率的な対話管理[11)]が可能となる．

4. 複数の音声理解方式に基づく音声理解の実装とその評価

我々は複数の言語モデルと言語理解方式を組み合わせて音声理解を行う枠組みを RIME-TK[1)] 上に実装した．システム開発者があらかじめ言語モデルと言語理解方式を用意することで，それぞれの組み合わせを簡便に指定できる．図 2 に指定例を示す．slu のタグで音声理解方式を記述する．slu の属性の id で音声理解方式の ID を記述し，decoding で音声認識に用いる言語モデルを，lu で言語理解方式をそれぞれ指定する．また，統合手法を用意することで，複数の理解結果からの統合手法を integration タグで指定できる．図 2 の例では，一行目の slu タグ内で，音声認識用言語モデルに文法モデル，言語理解方式に FST を指定している．

MLMU の有効性を示すため，この実装を用いてユーザ発話の音声理解実験を行い評価した．実験に用いたシステムはレンタカー予約システムと人間の対話コー

パス[12]である．言語モデルを2種類，言語理解方式を3種類用いた．また，これらに加え発話検証のための言語モデルを1種類用いた．2種類の言語モデルによる音声認識結果のそれぞれを，3種類の言語理解方式に入力し，6種類の音声理解結果を得た．複数の音声理解結果から識別器に基づき選択することで，最終的なシステムの言語理解結果とする．

4.1 言語モデル・言語理解の種類

本実験では，言語モデルは以下のモデルを用いた．

言語モデル (1) 文法ベース言語モデル

言語モデル (2) ドメイン依存統計言語モデル

検証用言語モデル ドメイン非依存大語彙統計言語モデル

文法ベース言語モデルで用いる文法は，言語理解で用いる意味文法と同じものを用いる．ドメイン依存統計言語モデルは，言語理解文法から自動生成した10000文から単語N-gramを学習し作成した．レンタカー予約システムにおける言語モデルの語彙サイズは，文法ベース，ドメイン内統計言語モデルいずれも257である．検証用のドメイン非依存大語彙統計言語モデルには，連続音声認識コンソーシアム配布の，Web文章から学習した単語N-gramモデルを用いた[13]．語彙サイズは60,250である．

言語理解方式は以下の3種類を用いた．

(1) Finite State Transducer (FST)

(2) Weighted FST (WFST)

(3) スポッティング

FSTによる言語理解では，言語理解文法からFSTを生成し，そこに音声認識結果を入力することで言語理解結果を得る．WFSTによる言語理解は福林らの手法に基づく[14]．WFSTはMITToolkit[15]を用いて構築し，WFSTのパラメータは評価用データとは異なる一名の105発話から推定した．スポッティングによる言語理解では，音声認識結果においてコンセプトに変換できる単語のまとまりを，全てコンセプトに変換する．

4.2 音声認識と言語理解における特徴に基づく理解結果の選択

複数出力される理解結果から最適な理解結果を選択するため，音声認識時と言語理解結果から得られる特徴を入力とする識別器を構築する．音声認識時に得られる特徴を用いることで，複数の言語モデルに基づく音声認識結果のうち正しい結果を推定できる．さらに，言語理解結果から得られる特徴も考慮することで，適切な言語理解結果を推定する．これらの特徴により決定木を構築する．決定木の構築にはC5.0[16]を用いた．以下では，用いた特徴について詳しく述べる．

音声認識時に得られる特徴を表1に示す．ここで音響スコアは発話時間で正規化したものを用いる．S1からS3は，使用した言語モデルに基づく音声認識時の音響スコアである．これらから，それぞれの言語モデルを基づく音声認識時の尤度を特徴として用いることができる．S4，S5は検証用言語モデルと他の言語モデルとの関係を表した特徴である．大語彙の言語モデルとの音響尤度を比較することで，音声認識結果の信頼性を検証できる．S6は言語モデル間の音声認識時の尤度の関係を考慮するために導入した．S7は，発話長によってそれぞれの言語モデルによる認識性能が変化する可能性を考慮して導入した．

表1 音声認識時の特徴

S1:	検証用言語モデル使用時の音響スコア
S2:	言語モデル (1) 使用時の音響スコア
S3:	言語モデル (2) 使用時の音響スコア
S4:	検証用言語モデル使用時と言語モデル (1) 使用時との音響尤度差
S5:	検証用言語モデル使用時と言語モデル (2) 使用時との音響尤度差
S6:	言語モデル (1) 使用時と言語モデル (2) 使用時との音響尤度差
S7:	発話時間 [秒]

表2 言語理解結果から得られる特徴

L1〜L6:	音声理解結果1〜6の事後確率に基づくコンセプトの信頼度の相加平均
L7:	L1からL6の相加平均
L8 〜 L13:	信頼度の相加平均の比 (= L1, ... L6 / L7)
L14 〜 L19:	音声理解結果1から6のコンセプト数
L20:	L14からL19の相加平均
L21 〜 L26:	コンセプト数の比 (= L14, ... L19 / L20)

言語理解結果から得られる特徴を表2に示す．L1からL6では，コンセプトの信頼度を考慮した．コンセプトの信頼度は，音声認識結果の10bestから得られる事後確率[11]に基づき算出した．L7からL13は理解結果どうしの信頼度の関係を表す．これらはある音声理解結果の信頼度が他の理解結果より相対的に高いか低いかを表す．L14からL19は理解結果に含まれるコンセプトの個数である．コンセプトの個数に応じて，理解方式の性能が変化する可能性を考慮した．L20からL26は各理解結果に含まれるコンセプトの個数と，コンセプトの個数との平均との関係をあらわす特徴である．この値が大きい場合，ほかよりコンセプトを多く出力しており，挿入誤りが多い可能性が高い．

4.3 評価対象発話データ

本実験では，レンタカー予約システムと人間の対話データ (22名× 8対話) 中のユーザ発話3,086発話を用いた．音声認識器はJulius (ver. 4.0.2) を用い，音響モデルは話者非依存PTMトライフォンモデルを用い

表 3　本手法による音声理解結果選択の Confusion Matrix

正解 \ 識別結果	(1)	(2)	(3)	(4)	(5)	(6)	棄却	計 (再現率 [%])
(1) 文法 + FST	0	0	0	0	0	0	0	0
(2) 文法 + WFST	0	3	5	1	0	45	1	55 (5.5)
(3) 文法 + Spotting	0	0	111	2	2	139	0	254 (43.7)
(4) 統計 + FST	0	0	2	1	0	39	0	42 (2.4)
(5) 統計 + WFST	0	0	9	0	2	85	0	96 (2.1)
(6) 統計 + Spotting	0	1	53	7	6	2529	4	2600 (97.3)
棄却	0	0	8	0	0	31	0	39 (0.0)
計 (適合率 [%])	0 (0.0)	4 (75.0)	188 (59.0)	11 (9.1)	10 (20.0)	2868 (88.2)	5 (0.0)	3086 (85.7%)

た[13]．文法ベース，単語 Ngram ベースを用いたときの音声認識精度はそれぞれ，68.7%と 76.2%だった．評価尺度は Concept Error Rate (CER) である．CER は，(システムが誤ったコンセプト数) / (発話に含まれるコンセプト数) で定義される．理解結果を選択する決定木の構築のため，学習データとして発話ごとに音声理解方式の正解ラベルを付与した．正解ラベルは，6 つの音声理解方式のいずれかもしくは，棄却かである．発話ごとに CER が最も低くなる音声理解方式を付与した．棄却のラベルは，挿入誤りが多く，6 つすべての音声理解で CER が 1 を越える場合に付与した．CER が最も低くなる音声理解方式が複数存在する場合，全学習データに対する精度が最も良い理解方式を正解ラベルとした．

4.4 実験結果

10-fold クロスバリデーションで発話データから決定木の構築と理解結果の選択を行なった．音声理解選択時の confusion matrix を表 3 に記す．また，単一の言語モデル・言語理解方式を用いたときの言語理解精度と，本手法の言語理解精度を表 4 に示す．(1) から (6) の番号付けされた条件が，単一の言語モデルと単一の言語理解方式で音声理解を行なった場合である．表中で，文法ベースの言語モデルとドメイン依存統計言語モデルを音声認識に用いたことを，それぞれ" 文法" と" 統計" で表す．今回 (1) 文法 + FST の理解結果が (3) 文法 + Spotting の理解結果と全く同じ結果となった．これは音声認識文法が言語理解用文法から作成されていることで，(1) と (3) では音声認識結果が全て言語理解文法に沿ったものであったためである．まず，表 3 について考察する．本手法による音声理解選択の精度は 85.7%であった．これは，最も音声理解精度が高い (6) 統計 + Spotting に対する再現率が最も高くなるように決定木を学習した結果である．その代償として，他の音声理解方式が正解となる場合でも (6) に誤って識別される理解結果が多かった．

次に表 4 について考察する．言語モデルとしてドメイン依存統計言語モデルを用いた場合の方が文法ベースの言語モデルを用いたときより全体的に高い精度になっている．これは音声認識率が統計言語モデルを用いたときのほうが高いことに起因する．(4) 統計 + FST が統計的言語モデルを用いているにも関わらず，最も低い精度となったのは，FST では受理できない認識結果が統計言語モデルに基づく認識結果に含まれていたからである．また，WFST による言語理解の精度が，単純なスポッティングより低いのは，WFST のパラメータが不適切であったことが原因として上げられる．(1) から (6) で最も精度が良かったのは (6) である．本手法での音声理解精度は，(6) の精度より 1.9 ポイント高い．これは複数の言語モデルと言語理解方式を考慮した結果である．6 つの候補から人手で正しい理解結果を選んだ場合の CER は 16.5%となった．これは (6) より 11.4 ポイント高い精度である．これは，複数の言語モデルと言語理解方式を用いた場合の性能の上限を示しており，今回の実験での性能とは 9.9 ポイントの差がある．性能の向上には，まず特徴量の検証が必要である．

表 4　各音声理解方式ごとの CER

音声理解方式 (言語モデル + 言語理解方式)	CER[%]
(1) 文法 + FST	32.3
(2) 文法 + WFST	34.2
(3) 文法 + Spotting	32.3
(4) 統計 + FST	36.8
(5) 統計 + WFST	31.6
(6) 統計 + Spotting	27.9
(1) ～ (6), 棄却から選択 (本手法)	26.0

5. おわりに

本研究では，音声対話システム開発時の限られた訓練データや労力などの制約のもとで高精度な音声理解を実現するために，複数の言語モデルと言語理解方式を用いた音声理解について述べた．まず，発話ごとに適した音声理解ができるように，言語モデルと言語理解方式のすべての組合せを用いた音声理解を行った．

そして，複数出力される結果から，発話ごとの特徴に応じて音声理解結果を動的に選択した．

評価実験では，単一の言語モデル・言語理解方式を用いたときと比較して，本手法により音声理解精度の向上を確認した．これにより，複数の言語モデルと言語理解方式を用いることの有効性を示した．

本枠組みの有効性の検証のためには，他ドメインでの実験や，言語モデルの性能を変化させた実験が必須である．また，学習データが少ない状況においても高精度な音声理解を実現するには，決定木を学習する上でのデータ量を少なくする必要がある．また，今回は理解精度のみを評価基準としたが，効率的な対話管理のためには理解結果への適切な信頼度付与が重要となる．複数の言語モデルと言語理解方式を利用した信頼度算出も今後の課題として挙げられる．

参 考 文 献

1) Mikio Nakano, Kotaro Funakoshi, Yuji Hasegawa, and Hiroshi Tsujino. A framework for building conversational agents based on a multi-expert model. In *Proc. 9th SIGdial Workshop on Discourse and Dialogue*, pp. 88–91, 2008.
2) Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Automatic Speech Recognition and Understanding*, 1997.
3) 安田宜仁, 堂坂浩二, 相川清明. 2つの認識文法を用いた主導権混合型対話制御. 情報処理学会研究報告, 2002-SLP-40-22, pp. 127–132, 2002.
4) 西田昌史, 寺師弘将, 堀内靖雄, 市川薫. ユーザの発話の予測に基づく音声対話システム. 情報処理学会研究報告, 2004-SLP-12-22, pp. 307–312, 2004.
5) Kazunori Komatani, Yuichiro Fukubayashi, Tetsuya Ogata, and Hiroshi G. Okuno. Introducing utterance verification in spoken dialogue system to improve dynamic help generation for novice users. In *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, pp. 202–205, 2007.
6) Bogdan Minescu, Geraldine Dammnati, Frederic Bechet, and Renato De Mori. Conditional use of Word Lattices, Confusion Networks and 1-best string hypotheses in a Sequential Interpretation Strategy. In *Proc. Interspeech*, pp. 1617–1620, 2007.
7) Stefan Hahn, Patrick Lehnen, and Hermann Ney. System Combination for Spoken Language Understanding. In *Proc. Interspeech*, pp. 236–239, 2008.
8) Mikio Nakano, Atsushi Hoshino, Johane Takeuchi, Yuji Hasegawa, Toyotaka Torii, Kazuhiro Nakadai, Kazuhiko Kato, and Hiroshi Tsujino. A robot that can engage in both task-oriented and non-task-oriented dialogues. In *humanoids06*, pp. 404–411, 2006.
9) 神田直之, 駒谷和範, 中野幹生, 中臺一博, 辻野広司, 尾形哲也, 奥乃博. マルチドメイン音声対話システムにおける対話履歴を利用したドメイン選択. 情報処理学会論文誌, Vol.48, No.5, pp. 1980–1989, 2007.
10) Sherif Abdou and Michael Scordilis. Integrating multiple knowledge sources for improved speech understanding. In *Proc. Eurospeech*, pp. 1783–1786, 2001.
11) 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理. 情報処理学会論文誌, Vol.43, No.10, pp. 3078–3086, 2002.
12) Mikio Nakano, Yuka Nagano, Kotaro Funakoshi, Toshihiko Ito, Kenji Araki, Yuji Hasegawa, and Hiroshi Tsujino. Analysis of user reactions to turn-taking failures in spoken dialogue systems. In *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, pp. 120–123, 2007.
13) Tatsuya Kawahara, Akinobu Lee, Kazuya Takeda, Katsunobu Itou, and Kiyohiro Shikano. Recent progress of open-source LVCSR Engine Julius and Japanese model repository. In *Proc. ICSLP*, pp. 3069–3072, 2004.
14) 福林雄一朗, 駒谷和範, 中野幹生, 船越孝太郎, 辻野広司, 尾形哲也, 奥乃博. 音声対話システムにおけるラピッドプロトタイピングを指向した言語理解. 情報処理学会論文誌, Vol.49, No.8, pp. 2762–2772, 2008.
15) Lee Hetherington. The MIT finite-state transducer toolkit for speech and language processing. In *Proc. ICSLP*, pp. 2609–2612, 2004.
16) J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993. http://www.rulequest.com/see5-info.html.