Vol.2010-SLP-82 No.22
2010/7/24

# 音声認識の信頼度・複数候補を利用した WFST 対話システムの評価

木村直人[†,††] 堀智織[†] 翠輝久[†] 大竹清敬[†]
柏岡秀紀[†] 中村哲[†]

我々は，拡張可能で適応可能な対話システムを構築する枠組みとして，重み付き有限状態トランスデューサ（Weighted Finite-State Transducer：WFST）に基づく対話システムを提案した．このプラットフォームでは，対話制御のための全てのルールを WFST 形式で表すことができ，マルチモダリティを用いた様々な対話タスクの対話制御に WFST を用いる．更に，WFST に基づく対話制御の枠組みを用いて，ラベル付けされた人間対人間の音声対話コーパスから統計的対話システムを構築した．WFST の入出力ラベルとして，コーパスから得られたユーザのコンセプトタグとシステムアクションタグを用いた．音声で入力されたユーザの発話をユーザの発話意図タグに変換する言語理解 WFST，およびユーザの発話意図タグをシステムのアクションタグに変換する対話シナリオ WFST を導入した．言語理解 WFST とシナリオ WFST は合成され，ユーザの入力に応じてシステムの次のアクションを決定する対話制御 WFST となる．これまでの研究では，我々は IF（Interchange Format）タグ付きホテル予約の音声対話コーパスを用いて言語理解 WFST を自動構築し，シナリオ WFST のタグ列の確率を推定した．これまで人手による書き起こし文を用いて対話戦略を評価してきたが，本稿では音声認識結果をシステムの入力とした場合の WFST に基づく対話制御の性能について示す．音声認識結果を入力とした場合，音声認識誤りによる対話制御の性能が劣化するという問題に対して，音声認識結果の音響的・言語的信頼性を示す信頼度および音声認識結果の複数候補(N-best)を処理できるよう WFST に基づく対話制御（WFSTDM）を拡張した．評価実験では，言語理解結果の精度と WFST 対話制御によって選択されたシステムアクションの適切さの評価を行った．実験結果より，信頼度および N-best を考慮する事により WFST パスの中からより適切なアクションの選択が可能となり，対話制御の性能が向上する事を確認した．

# Evaluation of WFST-based Dialog System using Confidence Score and Multiple Hypotheses of Speech Recognition

Naoto Kimura[†,††] Chiori Hori[†] Teruhisa Misu[†]
Kiyonori Ohtake[†] Hideki Kashioka[†] Satoshi Nakamura[†]

We proposed a weighted finite-state transducer-based dialog manager (WFSTDM) which is a platform for expandable and adaptable dialog systems. In this platform, all rules and/or models for dialog management are represented in WFST form, and the WFSTs are used to accomplish various tasks via multiple modalities. Based on this framework, we constructed a statistical dialog system using user concept and system action tags as input and output labels of the WFST where the tags were acquired from an annotated corpus of human-to-human spoken dialogs. We introduced a spoken language understanding (SLU) WFST converting user utterances to user concept tags and a dialog scenario WFST converting user concept tags to system action tags. The tag sequence probabilities of the scenario WFST were estimated with a spoken dialog corpus for hotel reservation. The SLU and scenario WFSTs were then combined to be a dialog management WFST which determines system's next action in response to user input. In our previous research, we evaluated its dialog strategy by referring to the manual transcription of the dialog for hotel reservation. In this paper, we present the performance of WFST-based dialog management when speech recognition hypotheses are input to the manager. To alleviate degradation of the dialog management performance due to speech recognition errors, we expand the WFSTDM for handling multiple hypotheses of speech recognition and confidence score indicating acoustic and linguistic reliability of speech recognition. The accuracy of SLU results and the correctness of system actions selected by the dialog management WFST were evaluated. We confirmed that the performance of dialog management was enhanced by choosing the optimal action among all the WFST paths for multiple hypotheses (N-best) of speech recognition and taking consideration of confidence score.

## 1. はじめに

我々は，人間と機械が人間と人間が対話するように自由に対話できる頑健な音声対話システムの構築を目指している．従来の音声対話システムでは，ユーザに対してシステムの質問に即した応答を求めている事から，受理されるユーザの応答はシステムの質問に制限されていた．そのため，ユーザは対話を通してほとんど柔軟な対話を行うことができなかった．ただし，このようなシステムの質問にユーザが回答する形式で駆動される対話では，ユーザの自由な発話入力を回避することができるので，より精度の高い音声認識結果を得ることが可能となる．現在，最先端の音声認識技術では，100M を超える単語辞書を用いてリアルタイムで自然発話を認識することができる事から[1]，ユーザの自由度の高い対話行為を受理する対話システムを構築する事ができる．

本研究では，ユーザの自由度の高い対話システムを構築するため，人間対人間の対

† 情報通信研究機構
National Institute of Information and Communications Technology (NICT)
†† 奈良先端科学技術大学院大学
Nara Institute of Science and Technology (NAIST)

話コーパスに基づく対話モデルを利用する．人間対人間の対話には典型的なパターンがあり，タスクが明確な対話では特に決められたパターンがある事から，統計的対話シナリオモデルを対話コーパスから学習することができるものと考えられる. さらに，様々な言語表現とそれが示す意図をより多く網羅するため，コーパスに基づく言語理解とシステム応答生成が必要となる．

我々は，様々なタスクやマルチモダリティを扱う拡張可能で適応可能な対話システムを構築するため[2]，ユーザの発話意図タグとシステムのアクションタグをそれぞれトランスデューザの入出力とした重み付き有限状態トランスデューサ(WFST)を用い，効率的に対話制御を行う枠組みを提案している．更に，WFST に基づく対話制御を用いて，人間対人間の対話コーパスに付与されたユーザの発話意図タグとシステムのアクションタグを WFST の入出力とする統計的対話制御を提案している[3][4].

これまでの研究では，インターチェンジフォーマット(IF)タグが付与された人間対人間のホテル予約の音声対話コーパスを用いて，WFST に基づく対話制御(WFSTDM)を構築した．IF タグとは，機械翻訳用の中間言語として開発されたタグである．我々は，IF タグをユーザの発話意図タグとシステムのアクションタグとして適用し，言語理解(SLU)WFST と対話シナリオ WFST をコーパスより自動学習した[4]．自動獲得されたそれらの統計的対話制御モデルの性能を評価するため，ユーザ入力の書き起こしをシステムに入力し，次のシステムアクションの推定精度の評価を行った．これまで音声の書き起こしを用いて対話制御をおこなってきたが，本稿では音声認識結果を入力とした場合の WFST に基づく対話制御の性能について示す.

音声認識結果を入力として対話制御を行った場合，言語理解モデルは対話コーパスの書き起こし文から学習していたため, 音声認識誤りは対話制御の性能を劣化させる. 先行研究では，対話制御の音声認識誤りによる劣化を軽減するため，音声認識結果の複数候補や音声認識の音響的・言語的信頼性を表す信頼度を対話システムの言語理解に適用した[5][6][7]．具体的には，各音声認識結果の信頼度が閾値より低い場合，ユーザ入力は対話システムによって棄却される，または，システムによって冗長的な確認対話が生成するのを防ぐ事で, より効率的に対話戦略を行う方法が提案されている. 別のアプローチでは，信頼度で理解候補に重みづけを行い，最も尤度の高い理解を選択することで言語理解を行う方法が提案されている.

音声対話システムでは各コンポーネントが完全に独立して動作しているわけではない事から，言語理解の頑健性や対話制御の次の発話の予測精度を向上させることができれば, 音声認識誤りによる対話制御への影響を軽減することができる. 本稿では，音声認識結果の複数候補や信頼度を扱うために WFSTDM を拡張する．音声認識結果を入力とした場合の対話制御の性能を評価するため，言語理解結果の精度や対話制御 WFST によるシステムアクションの推定の精度を評価した．我々は，信頼度および N-best を考慮する事により WFST パスの中からより適切なアクションの選択が可能となり，対話制御の性能が向上する事を確認した.

## 2. WFST に基づく対話システム

### 2.1 システム概要

WFST による対話システムでは，ユーザの発話文を対応する発話意図タグ列へと変換する言語理解 WFST と，発話意図タグ列をシステムのアクションタグ列に変換する対話シナリオ WFST，アクションタグ列より応答文を生成する文生成 WFST とを合成した対話制御 WFST を構成し，発話文を入力記号列として対話制御を行う．対話システムは入力としてユーザの発話文を受理し，出力として対応するシステムアクションや文を返す．WFST は様々な知識やモデルに基づき制約やコストを埋め込むことができるため，WFSTDM では多様な対話制御が可能である．本研究では，コーパスから自動学習した言語理解 WFST と対話シナリオ WFST，文生成 WFST を合成することにより対話制御を行う[3][4].

### 2.2 言語理解 WFST

言語理解 WFST は，入力文から発話意図を表す特徴的なパターンを検出し，発話意図タグへ変換するパターン検出器である．コーパスより同一タグに対応する発話文集合を抽出し，その中で相対頻度の高い 1 ～M 個組の単語列をその発話意図タグの特徴的な表現パターンとして言語理解 WFST で表現し，組み込んだ．具体的には，入力単語列が言語理解 WFST で表現されたキーフレーズにより長くマッチ（最長パターン一致）するパスが低いコストとなるように，WFST の遷移重みを決定した.

本稿ではホテル予約コーパスより発話意図タグに対応するキーフレーズ (1～6 単語)を用いて言語理解 WFST を構成した.

### 2.3 シナリオ WFST

統計的対話シナリオはコーパス中の IF タグ列を用いて学習される. ユーザへの応答を選ぶ際に，いくつかの選択肢があるが，対話システムのシナリオ WFST は，対話の各状態に従って，ユーザ入力に応答するどのシステムアクションを選択するか決定することができる．本稿では，コーパスから得られる IF タグ列の 3-gram を用いてシナリオ WFST を構成した.

### 2.4 文生成 WFST

文生成(SG)WFST はシステムのアクションタグ列を対応する自然文に変換する．SG WFST もまたコーパスから得ることができる．我々は，これをタグのバックオフバイグラムモデルとしてそれを設計した．また，現在の対話のコンテキストを考慮するので，SG WFST は現在の対話状況で適当なアクションタグを選択し，コーパス中で同一のアクションタグの付与されている発話文から最も頻度の高い発話文を選

択する[4]．ただし，本稿では文生成 WFST を合成しないで評価を行う．

### 2.5 対話制御 WFST

言語理解 WFST，シナリオ WFST と文生成 WFST は WFST の演算を使用して合成されたのち，最適化される．最終的に合成された WFST は，対話制御 WFST を意味する．ユーザ発話の書き起こしや音声認識結果は対話制御 WFST に入力され，次のシステムアクションが出力される．

### 2.6 音声認識結果の複数候補を用いた対話制御

これまで対話制御の有効性を評価するため，WFST の入力として対話のユーザ発話の書き起こし文を用いていた．本研究では，音声認識誤りを含む音声認識結果を WFST の入力として利用する．我々は，対話制御への認識誤りの影響を軽減するため，音声認識の音響的・言語的信頼性を表す信頼度付きの音声認識の複数候補を扱うために WFSTDM を拡張した．新しい WFSTDM は N-best を考慮した複数候補の全ての WFST パスの中から尤もらしいアクションを選択する．ここで，各パスは対応する候補の信頼度で重みづけされている．このアプローチを用いる事により，最尤仮説よりも誤り率の少ない仮説を含む N-best リストを考慮でき，さらに最尤スコアで求められない適切な仮説を対話のコンテキストを用いる事で選択することができる．その結果，音声認識誤りの影響を軽減し，次のシステムアクションの推定精度の向上を図ることができる．

これまで書き起こし文を入力として対話制御を評価した際には発話境界が既知であると仮定していたが，音声認識結果では発話境界は曖昧であり，発話境界を正しく検出することは難しい．そこで，音声認識結果を用いる際には，複数の文をまとめて受理できるように WFSTDM を修正し，発話境界を考慮せずに入力文を変換し発話意図タグを連続的に出力した．つまり，入力シンボル列の文境界の異なる複数パスを考慮した．

## 3. 評価実験

### 3.1 評価データ

本研究では，評価データとして，英語話者と日本語話者間のホテル予約の模擬対話コーパス 25 対話を用いた[4]．25 対話の平均ターン数は 11 ターンであった．英語話者と日本語話者間の対話は通訳を介して行われた．コーパスには音声翻訳システム用の中間言語であるインターチェンジフォーマット(IF)タグが付与されていたが，本来 IF は対話制御用に設計されたものではないため，タグの一貫性に欠けるといった問題があった．したがって，IF タグを対話制御のための一貫したタグに改良し，ユーザの発話意図タグと，システムのアクションタグを修正した．付与されたタグは，58 種類の発話意図タグと 88 種類のシステムアクションタグの計 146 種類である．本稿では，評価データに日本語話者の音声を利用した．

### 3.2 音声認識

本研究では，Julius を用いて音声認識を行った[8]．音響モデルとして日本語の男女非依存のトライフォンモデル，言語モデルとしてホテル予約を含む旅行対話コーパス 2206 対話 87194 発話から学習した 2-gram・3-gram モデルを用いた．評価データ 25 対話に対する，3-gram の言語モデルのテストセットパープレキシティは 31.97，未知語は 111 単語，0.81%であった．これらを用いて音声認識結果を行った結果，単語正解精度は 76.36%であった．

我々は，複数候補の中から最尤の理解と最尤のシステムアクションを選択するために WFSTDM に対して，Julius から得られる音声認識の単語信頼度を適用した．単語信頼度の精度を評価するために，不正解受理率(FA)と正解棄却率(FR)を下記のように定義し算出した[8].

$$FA = \frac{All\quad of\quad incorrectly\quad accepted\quad words}{All\quad of\quad accepted\quad words} \tag{1}$$

$$FR = \frac{All\quad of\quad incorrectly\quad rejected\quad words}{All\quad of\quad rejected\quad words} \tag{2}$$

計算によって得られた FA と FR を図 1 に示す．

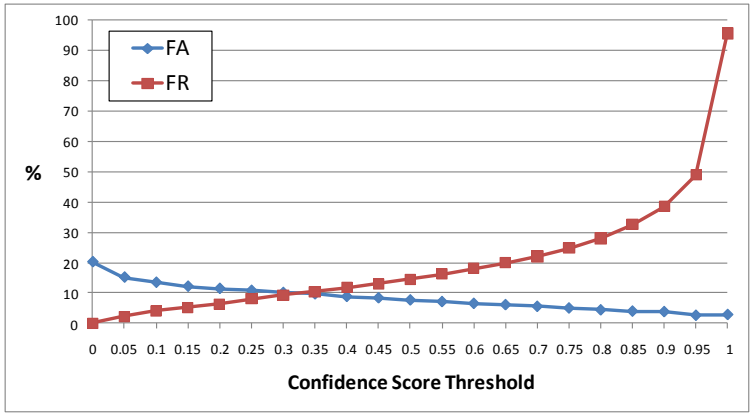


図 1　信頼度の性能

また，不正解受理率と正解棄却率が等しくなる等誤り率は 9.81%であった．

### 3.3 実験方法

システムを評価するため，テストセットを正解対話としてシミュレーションを行った．IF タグ付き模擬対話コーパス 25 対話に対し，各対話をテストセットとし，それ以外の 24 対話を学習セットとするヘルドアウト法を用いた. 正解対話系列のシステムの各ターンで，ユーザの発話を入力とし，言語理解部の理解精度とシステムアクションの予測精度を評価した．システムの言語理解の精度とシステムの次のアクションの予測精度を評価するために，テストセットの正解対話に従って WFST に正確な遷移をさせ，各ターンでユーザからの入力を与え，その入力に対する言語理解とシステムの次のアクションタグ列を予測させた．システムの次のアクションの予測精度の評価基準には平均逆順位(Mean Reciprocal Rank, MRR)を用いた MRR は次式のように定義される．

$$MRR = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{R_i} \qquad (3)$$

ここで，$R_i$ は適切なアクションタグ列のランク，$M$ はターン数を表す．MRR は正解が推定候補の上位にある場合に値が大きくなる評価で，正解が少ない場合の評価に適している．

## 4. 実験結果

### 4.1 言語理解性能の評価

言語理解 WFST はコーパスから学習され，最長パターン一致に基づいてユーザ発話を発話意図タグへと変換することで動作する。そのため、ユーザの発話意図タグに置き換えられる単語 n-gram で表現されたフレーズが学習データに存在しない場合，あるいは音声認識結果に誤認識がありフレーズが一致しない場合，対話システムはユーザ発話を正しく理解することはできない．

言語理解 WFST の性能の上限を調べるため，ユーザ発話から抽出された最長フレーズパターンから推定される発話意図タグの複数候補に正解の発話意図タグがどれぐらい含まれているかを評価した. 書き起こし文, 1-best の音声認識結果, 複数候補(5-best, 10-best)の音声認識結果を入力とした場合の正しい発話意図タグの網羅率を評価した結果を図 2 に示す．

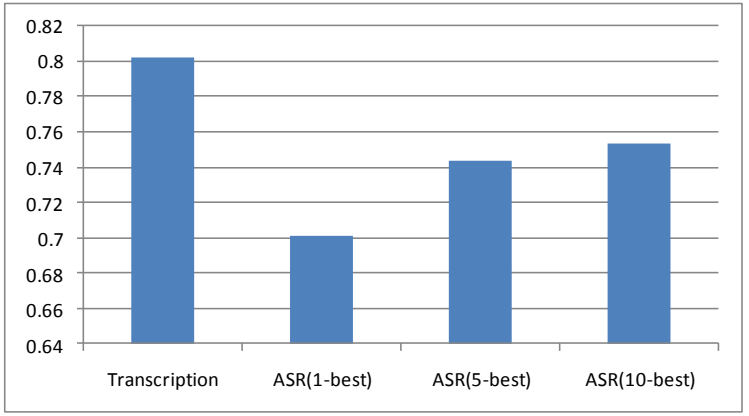


図 2　言語理解部のユーザコンセプトタグ候補中の正しいユーザコンセプトタグの網羅率（Transcription : 書き起こし文，ASR(N-best) : 音声認識結果の N-best，N= 1, 5，10）

書き起こし文では，発話境界は既知である．音声認識結果では，発話境界は未知である. 図 2 より, 言語理解の複数候補の中に正しい発話意図タグが網羅される割合は，書き起こし文を入力した場合は 80.2%，音声認識結果では，1-best で 70.1%，5-best で 74.3%，10-best で 75.3%であった．この結果から，書き起こしと同等にユーザ発話を正しく認識できれば，言語理解 WFST ではユーザの発話意図を最大で 80%正しく理解できる事が分かる．また，書き起こしの代わりに音声認識結果を対話システムの入力とした場合でも，最大でユーザの発話意図の 75.3%を理解できる可能性がある．音声認識結果を入力とした場合，1-best を用いた場合と比べて複数候補を用いた方が性能改善している事から，音声認識結果の複数候補を考慮することにより言語理解の性能向上が期待できる．

図 2 に示された言語理解の性能の上限は，推定候補の中に正解が含まれる網羅率を示しているだけである事から，必ずしも言語理解 WFST の性能を反映するとは限らない．そこで，言語理解 WFST で変換されたユーザの発話意図タグを出力し，正解タグとの一致率を用いて言語理解 WFST の性能を評価した．ユーザ発話の書き起こし文と音声認識結果を言語理解 WFST の入力とし，ユーザの発話意図タグへと変換した際の言語理解の性能を図 3 に示す．

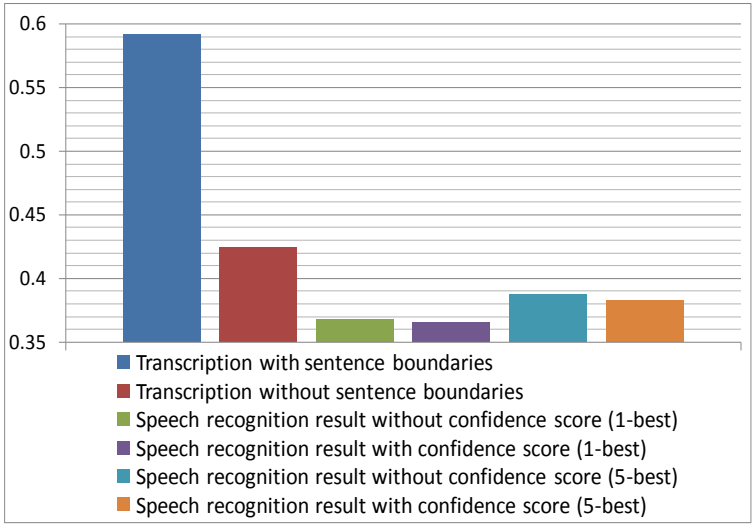


図 3　言語理解 WFST を用いた言語理解の性能

図 3 より，発話境界既知の書き起こし文を入力とした場合の言語理解 WFST による発話理解の精度は 59.2%，発話境界未知の書き起こし文の場合は 42.5%であった．音声認識結果には明確な発話境界がないため，音声認識結果の場合の精度は発話境界が未知である事によっておよそ 30%劣化している事が分かる．信頼度を考慮せずに音声認識結果を用いた場合，1-best では 36.8%であったが，5-best では 38.8%に改善した．この事から，音声認識結果の複数候補を用いる事で言語理解の性能を改善できる事を確認した．音声認識結果を入力として信頼度を考慮した場合，1-best では 36.5%，5-best では 38.3%であった．この事から，音声認識結果の複数候補を用いる事で言語理解の性能が改善される事が分かる．他方，信頼度の考慮の有無を比較した場合，1-best，5-best 共に言語理解性能が改善されていない事が分かる．この結果から，必ずしも信頼度が言語理解 WFST の性能に貢献しない事が示された．信頼度によって言語理解性能が改善しない理由として，正しく音声認識できているにも関わらず信頼度が低い事によりペナルティが課せられてしまい，正しい音声認識が選ばれない事により間違った言語理解が行われたためと考えられる．このように，信頼度が必ずしも正しくない場合，信頼度による言語理解 WFST の性能改善は期待できない．

### 4.2　システムアクションタグの予測性能

言語理解 WFST とコーパスのタグ列の確率から学習されたシナリオ WFST を合成して対話制御 WFST を構築し，WFSTDM 全体性能を評価した．ユーザの入力に対するシステムのアクションタグの予測性能を MRR に基づき評価した．WFSTDM の入力として，書き起こし文，音声認識結果の 1-best，5-best を用いた．WFSTDM に対する信頼度や音声認識結果の複数候補の効果について図 4 に示す．

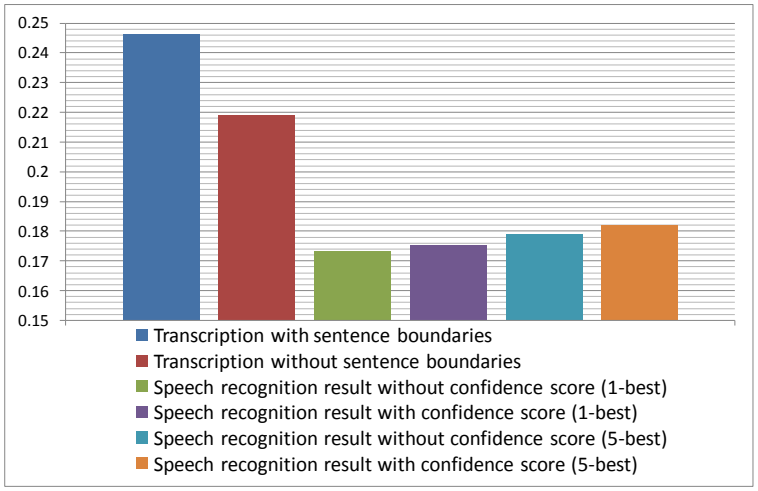


図 4　MRR に基づく WFSTDM を用いたシステムアクションタグの予測性能

図 4 より，発話境界既知である書き起こし文を入力とした場合，MRR は 0.246 であったのに対し，発話境界未知の書き起こし文を入力とした場合の MRR は 0.219 と劣化した．次に，入力を音声認識結果の 1-best とした場合の MRR は 0.173，5-best とした場合の MRR は 0.179 であった．この結果より，図 3 で得られた言語理解の性能が WFSTDM の全体の性能に直接反映されていることが分かる．信頼度は言語理解の性能に必ずしも寄与するとは限らなかったが，システム全体の性能としては，図 4 に示すように 1-best の場合は 0.173 から 0.175，5-best の場合は 0.179 から 0.182 へと MRR が改善した．我々は，信頼度付きの複数候補を用いることでシステムアクションタグの推定性能が向上することを確認した．この結果は，より長いフレーズを表す発話意図タグが音声認識結果の複数候補によってカバーされるため正解の網羅率が向上し，結果的に言語理解性能が向上する事に因る．さらに，音声認識の信頼度は，システムアクションタグの複数候補の中でより正解のアクションタグのランクを向上させ，その結果 MRR を改善したものと考えられる．

図 5 に音声認識性能と，WFSTDM の言語理解および対話制御性能の相関を示す．ただし，対話制御性能はシステムアクションタグの予測性能を MRR で評価した値である．25 対話の平均の単語正解精度は 61.74%～76.3%の間であった．

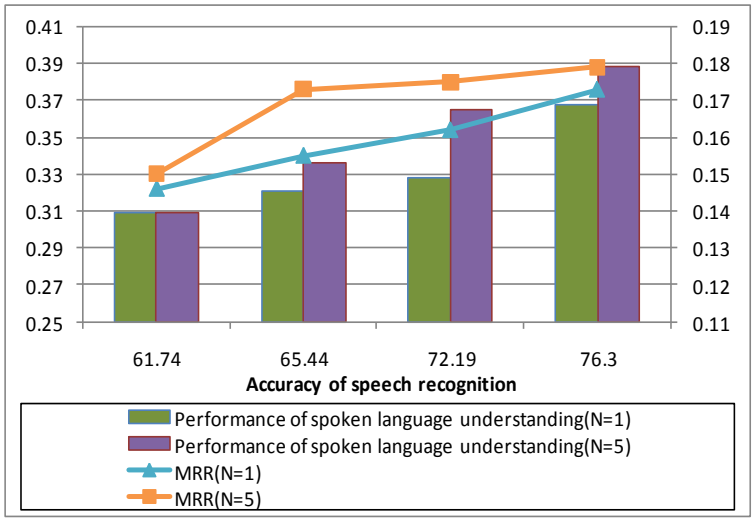


図 5　音声認識性能の変化に基づく言語理解 WFST による言語理解性能と WFSTDM による対話制御性能の変化

図 5 より，単語正解精度が 61.74%，65.44%，72.19%，76.3%の各場合の言語理解の性能は 1-best ではそれぞれ 0.309, 0.321, 0.328, 0.368, 5-best ではそれぞれ 0.309, 0.336, 0.365，0.388 と，認識率が改善するにつれて性能が改善した．次に，システムアクションタグの予測性能を示す MRR は，各音声認識精度に対して 1-best ではそれぞれ 0.146，0.155，0.162，0.173，5-best ではそれぞれ 0.15，0.173，0.175，0.179 となり，こちらも認識率が改善するにつれて性能は改善した．また，1-best と 5-best を比較すると，複数候補を考慮することで全体的に性能が改善している事が分かる．単語正解精度が上昇するにつれて，言語理解の精度と MRR が改善している．音声認識性能と言語理解性能の相関係数は，1-best で 0.907，5-best で 0.995 という強い相関を示している．更に，音声認識性能とシステムアクション予測性能の相関係数は，1-best で 0.983，5-best で 0.84 と強い相関を示している．

## 5.　おわりに

本稿では，ホテル予約の音声対話コーパスを用いて WFST に基づく対話制御に基づく統計的対話システムを構築し，ユーザ発話の書き起こしと音声認識結果を入力として対話制御の性能評価を行った．音声認識結果で発話境界が明確に与えられないという問題を解決するため，対話の各ターンで複数発話から発話意図タグを繰り返し出力できる枠組みを構築した．評価実験では，言語理解 WFST による言語理解の性能，さらに，言語理解とシナリオ WFST を合成して作成した対話制御 WFST を用いて入力に対するシステムアクションの選択の精度を評価した．実験結果より，音声認識結果の複数候補(N-best)を考慮する事により，より適切なシステムアクションが WFST のパスから選択することができ，言語理解や対話制御の性能が向上することを確認した．更に，音声認識結果を入力として信頼度を考慮した場合，言語理解性能を直接向上させることは示されなかったが，対話制御全体として性能が改善される事を確認した．本稿では，文生成 WFST モジュール[4]を行わなかったことから，音声入力を用いた文生成の評価は行っていない．今後は，音声をシステムの入出力として，人間による対話制御の性能評価を行う予定である．

## 参考文献

1)　T. Hori, C. Hori, Y. Minami, and A. Nakamura, “Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition,” IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, pp. 1352—1365, 2007.
2)　K. Kayama, A. Kobayashi, E. Mizukami, T. Misu, H. Kashioka, H. Kawai and S. Nakamura, “Spoken Dialog System on Plasma Display Panel Estimating Users' Interest by Image Processing”, 1st International Workshop on Human-Centric Interfaces for Ambient Intelligence, July 2010. (to appear)
3)　C. Hori, K. Ohtake, T. Misu, H. Kashioka, S. Nakamura, “Dialog management using weighted finite-state transducers,” Interspeech, 2008.
4)　C. Hori, K. Ohtake, T. Misu, H. Kashioka, S. Nakamura, “Weighted Finite State Transducer Based Statistical Dialog Management,” ASRU 2009.
5)　K. Komatani and T. Kawahara, “Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output”, In Proc, 467-473, 2000.
6)　T. J. Hazen, S. Seneff and J. Polifroni, “Recognition confidence scoring and its use in speech understanding systems,” Computer Speech and Language, 49-67, 2002.
7)　河原，荒木，“音声対話システム，” オーム社，2006.
8)　A. Lee, K. Shikano, T. Kawahara, “Real-time word confidence scoring using local posterior probabilities on tree trellis search,” ICASSP, 2004.