情 報 メ デ ィ ア 15－5
（1994．3．11）

音声自由対話システム TOSBURG II における データ収集と評価

舘森三慶　瀬戸重宣† 金沢博史　竹林洋一
(株) 東芝 研究開発センター

本報告では，不特定ユーザに対して何ら制約を設けない音声自由対話システム TOSBURG II の対話データ収集と評価について述べる．TOSBURG II の評価システムでは，対話音声データだけでなく，音声認識・理解結果，対話の履歴等の対話データ収集ができ，その解析により認識・理解や対話処理の性能および使い勝手の向上が可能である．また，TOSBURG II ではキーワードに基づき自由発話を理解するというアプローチをとっており，対話音声の詳細なトランスクリプションをせずにキーワード列だけからユーザの発話内容を記述することにより，対話データベースを効率的に作成する．評価システムを用いて収集した対話データを解析したところ，システムの性能が話者の熟練度や対話条件に影響されること，人間同士の対話中ではほとんど現れない発話が出現することが確認された．

Dialogue Data Collection and Evaluation for Spontaneous Speech Dialogue System TOSBURG II

Mitsuyoshi TACHIMORI, Hiroshi KANAZAWA,
Shigenori SETO and Yoichi TAKEBAYASHI
Toshiba Corporation, R & D Center
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210 Japan

This paper describes the dialogue data collection and evaluation for a spontaneous speech dialogue system TOSBURG II. The evaluation system of TOSBURG II is designed so as to record not only speech data and the final results of speech understanding, but also its intermediate results which include a keyword lattice, sentence candidates and dialogue history. These data are used to improve the system's performance. TOSBURG II employs the keyword-based spontaneous speech understanding approach, thus its performance can be evaluated from a sequence of correct keywords, without full transcriptions of the user's utterances. The experimental evaluation has shown that the system's performance is affected by the user's skill, and the utterances in human-computer dialogue are different from the ones in human-human dialogue.

†現在，日本電子化辞書研究所に勤務．

1. まえがき

音声は人間の最も基本的かつ重要なコミュニケーション手段であり，計算機との対話手段としても有望視され，世界各国で研究開発が盛んに行なわれている．音声入力はキーボード，マウス，ペン入力とは異なり，ハンドフリーという特徴があり，ハードウェアの低コスト化により，携帯用の情報機器やマルチモーダルインターフェースとしてその重要性が高まってきている．しかし，これまでに開発された連続音声理解システムは，実験室レベルでは高い認識性能が実現されているものの，環境騒音に対して脆弱であり，また，いい淀みや不要語等を含む話し言葉に対するロバスト性が不十分であるため [1, 2]，音声入力インタフェースが広範に応用されるには至っていない．

人間と計算機とが自由に対話できる実用的システムを構築するためには，大量の対話データの蓄積とシステムの性能・使い勝手の評価が必須である．従来，対話データの収集は，人間同士の模擬対話，あるいは人間が応答操作する疑似対話システム (Wizard システム) とユーザとの対話により行なわれた [3, 4]．しかし，人間は相手や状況により話し方を変えるので，対話システムと人間の対話では，人間同士の対話あるいは疑似対話システムとの対話とは異なる言語現象が起こると考えられる [5]．そのため，実システムを用いた対話データ収集とシステム評価・改良が必要となる．

これに対して筆者らは，日常的なファーストフードの注文タスクを選定し，ユーザに何ら制約を設けることなく，自由にシステムと対話が可能な実時間対話システムの構築を目指した．ロバストなシステムの実現のため，雑音免疫法によるキーワード検出 [6]，キーワードに基づく話し言葉の理解，ユーザ主導型の対話，マルチモーダル応答，音声応答のキャンセル機能などを統合し，不特定ユーザを対象とした音声自由対話システム TOSBURG II を開発した．さらにシステムの評価，改良のために，実対話システムによる対話データ収集・評価システムを作成した [7, 8]．

本報告では，TOSBURG II の特徴について述べ，次に TOSBURG II をベースに構築した対話データ収集・評価システムについて説明する．さらに本システムを用いて行なった評価実験について述べる．

2. 実時間音声自由対話システム TOSBURG II

2.1 TOSBURG II の構成

現在音声を含むマルチメディア入力に対する期待が高まっているが，大半のシステムは単にマルチメディアデータを信号レベルで入出力 (記録 / 再生) したり，入力信号を記号に変換 (文字認識等) しているに過ぎない．本来のマルチメディア / マルチモーダルインターフェースに必須なのはメディア変換機能に加えてユーザの意図理解，状況理解であり，そのためには意味・文脈レベルの処理が不可欠である．

筆者らはまず，音声対話システムの高度化のため要素技術を開発し，それらを統合して不特定話者の自由な発話を理解する実時間音声対話システム TOSBURG を試作した [9]．さらに，TOSBURG に音声キャンセル機能を組み込み，システムの応答中でもユーザの割り込み入力を可能とする TOSBURG II を開発した [7]．そのシステム構成を図 1に示す．

単語検出部では，ユーザの発声した連続音声の中から，TOSBURG II のタスク達成に重要な 49 語のキーワードを検出する．キーワード検出は雑音免疫ワードスポッティング法を用いて，雑音下でも頑強なキーワード認識を可能としている．

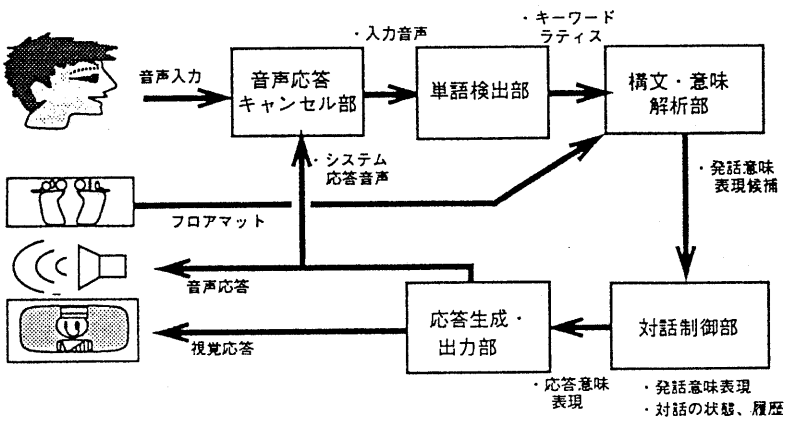

図1. 音声対話システムTOSBURG IIのシステム構成

構文・意味解析部は，単語検出部で検出されたキーワード系列をもとに，ユーザの発話文の内容を理解する．ユーザの発話の内容を捉えるには全ての言語や音節を詳細に聞きとる必要はなく，発話中の重要な意味を持つキーワードを組合せればよいという立場に基づいている．これにより，自由発話に頻繁に現れる不要語や言い間違い等に対して，ロバスト性を向上させている．キーワード列の構文意味解析は拡張LRパーザにより行ない，発話意味表現候補を出力する．

対話制御部では，対話の履歴や対話の状況に基づいて，発話意味表現候補の中から最も確からしい候補を対話音声理解結果として選択する．対話の文脈を利用し，不完全な発話，省略表現を含む自然な発話にも対応することができる．さらに対話制御部では，対話の状況とシステムの発話理解に基づき，応答意味表現を出力する．

応答生成出力部では，応答意味表現とシステムの内部状態に基づき，アニメーション，テキスト，音声によるマルチモーダル応答を生成する．

音声応答キャンセル部では，システムの応答中にアクティブ騒音制御技術を用いてユーザの発話に混入するシステム音声応答を引き去る．これによって音声応答中でも随時ユーザの割り込み発話が可能となる．

以上に述べたように，TOSBURG IIでは，発話文を一字一句認識するのではなく，キーワードに着目して発話の意味をとらえ，自然な発話に含まれる不要語，いい淀み，語順の逆転に対処し，自然な発話の意味内容を理解できるようになった．このシステムの試用を通じてマルチモーダル応答や音声認識の際のユーザの反応に関する知見を得ることができた[10].

2.2 ユーザ主導による自由対話

TOSBURG IIでは円滑な対話を実現するために，従来のシステム主導型の「穴埋め形式」ではなく，発話の制約を極力少なくし，ユーザが自由に発話できるユーザ主導型の対話管理を行なっている[11].

図2に本システムの対話モデルを，図3に対話処理制御部での処理を示す．対話状態は，対話の流れに即して発話を理解するユーザ状態と，タスクを管理しユーザ発話に応じた応答を返すシステム状態からなる．対話はシステムの注文要求から始まり，ユーザの注文を理解し，マルチモーダル応答により理解結果をユーザに確認するという形式で進行する．また，システムの理解結果とこれまでの対話の流れが合致しない場合には，

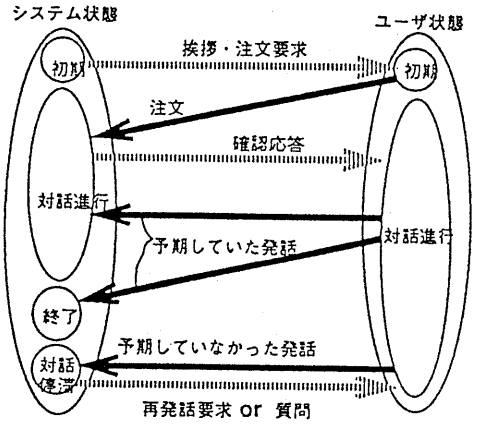

図 2: TOSBURG IIの対話モデル

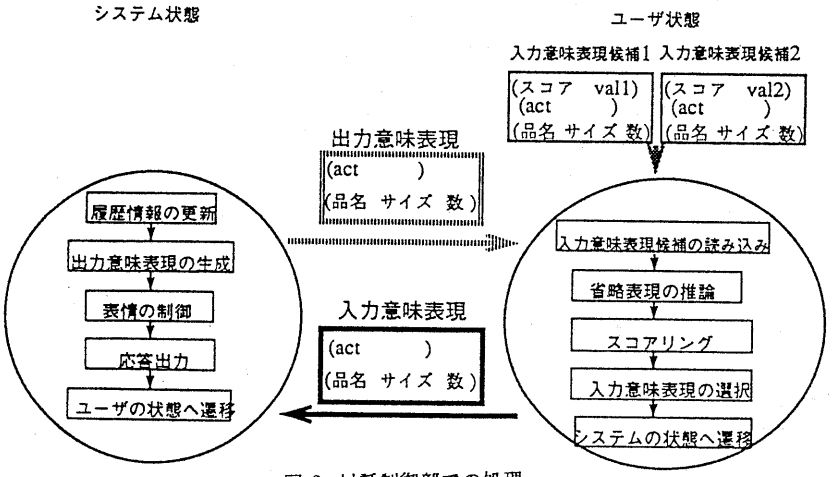

図 3: 対話制御部での処理

再発話を要求するか，これまでの注文を再確認するなどして対話混乱を防いでいる．

対話処理制御部において，ユーザ状態での発話理解とシステム状態での応答生成は次のように行なわれる．ユーザ状態では，自由発話でしばしば起こる省略表現に対処するため，最新のユーザ発話を利用して省略部分を推論する．その後，対話履歴に基づいて，複数の発話意味表現候補のスコアリングを行なう．入力候補の中からスコアの最も高いものをユーザの発話意味として選択し，システム状態に遷移する．

システム状態では理解した発話内容に従って対話履歴を更新し，ユーザに対して理解内容の確認を行なうための応答を生成する．ここで，キーワードの検出誤りや文理解結果の誤りに起因する発話の曖昧性に対処するために，必要に応じ曖昧な部分に関する質問や再発話要求を生成する．これらの応答は，合成音声とテキスト(音声応答文)，店員の表情，注文品目のグラフィックス表示から成るマルチモーダル応答によってなされる．これにより，ユーザフレンドリーなシステム応答を実現し，円滑な対話進行をサポートしている．

3. 対話データの収集とシステム評価

3.1 実システムによる対話データ収集

人間同士の対話では，聞き手の表情，応答，また対話の状況に合わせて，話し手の言葉遣いや発話速度，抑揚の変化が観察される．計算機との対話においては，ユーザはシステム応答やシステムの対話制御方法，システムに対する印象に基づいて，対話状況に応じた話し方をする．そのため，システムとの対話では人間同士の対話で起こる言語現象とは異なった現象が起こるので，実システムを用いた対話データ収集を行なう必要がある[10, 12]．

また，実システムによる対話データ収集は，システムの改良・評価に必須である．先に試作したTOSBURGではユーザはシステム応答中にユーザ発話を受け付けないように設計したが，多くのユーザはシステムの応答終了を待たずに割り込んで発話する傾向があった．より自由な対話を実現するため，システムを改良し，音声応答中でも割り込み入力が可能であるTOSBURG IIを試作した．

このように，システムを実際に使って初めてわかる事象があるので，リアルタイムで動作する実システムの構築と，それを用いた対話データの収集が重要である．

3.2 音声と対話処理データ

従来の音声対話データの収集は，ユーザの発話文，システム応答文、システムの理解結果を記録しているに過ぎない[13]．しかし，音声理解や対話処理のシステム中間データ(システムの内部処理結果)を対話データと対応づけてデータベース化すれば，より多面的な評価・改良が効率的に行なえる．図4に示すTOSBURG IIの対話データの収集・評価システムにおいては，以下に示すような，システムが発話理解の過程で生成する中間データも対話データとして記録する．さらにシステムの内部状態や発話の種類(注文，追加など)を記録し，システムの発話理解過程，内部状態まで含めた対話の忠実な再現を可能としている．

- 入力音声波形データ
- システム応答音声波形データ
- キーワードラティス
- 発話意味表現候補
- 発話意味表現(対話処理結果)
- 対話の状態，履歴
- 応答意味表現

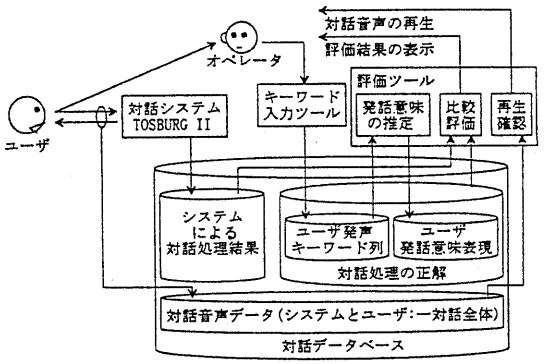

図 4: TOSBURG IIの評価システム

3.3 キーワード入力ツール

収集された対話文のデータベース化の際には，人間が対話文を聞きながら文字に起こす作業(トランスクリプション)が一般に行なわれている．トランスクリプションには多くの人手と時間を必要とするため，大量の対話データ収集が困難となり，音声対話や音声ヒューマンインタフェースの研究開発の大きな障害となっている．

これに対して，TOSBURG IIではユーザの発話内容をキーワードによって捉える方式を採用しているため，発話文の詳細は必要としていない．この点に着目し，対話データのラベリングに際しては，キーワードを効率的に入力するのため，キーワード入力ツールを作成した．図5に示すツールでは，対話データ収集中にオペレータが発話中に含まれるキーワードを聞きとり，該当するボタンをクリックすることにより，自動的に

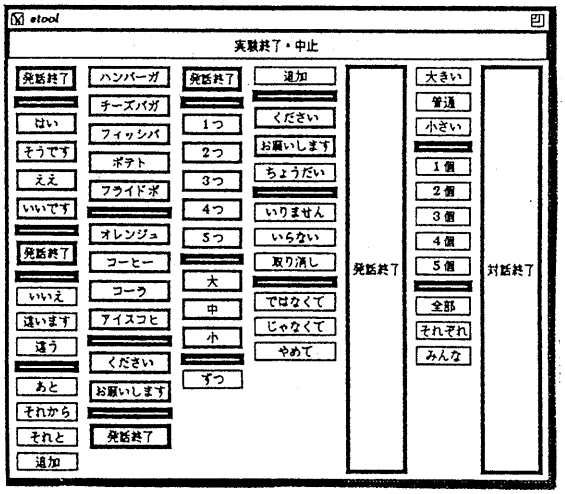

図 5: キーワード入力ツールの表示画面

キーワードファイルを生成する．これらの作業はリアルタイムで処理可能であるので，一般の方法に比べ作業効率が大幅に向上する．また，既に入力された対話音声データのラベリングのためオフライン処理機能も用意している．

3.4 評価ツール

対話データの評価ツールの画面表示例を図 6に示す．このツールは収集された対話データを読み込み，対話文，キーワードラティス，キーワード文候補，発話意味表現を表示する．マウス操作により，発話文単位での音声の再生と，指定したキーワードの再生が可能である．このツールでは，対話データの検証のみならず，他の必要な情報をデータに付加し，対話データベースの作成を支援する．また，ここで作られた対話音声データベースを利用して，自動的に単語検出率，発話文認識・理解率を集計することができ，単語検出部，構文意味解析部，対話処理部といった構成要素ごとに評価できる．さらに，システムの内部状態データを検証することによって，システムの対話の流れの制御の評価も行なえる．

このように，システムの最終理解結果のみならず，中間処理結果も含めた対話データの収集を行なうことにより，多面的なシステム評価ができる．

4. システムの評価実験

以下では，音声対話システムの試用を通じて明らかとなった知的音声インタフェースの評価の際の問題点と実験結果について述べる．

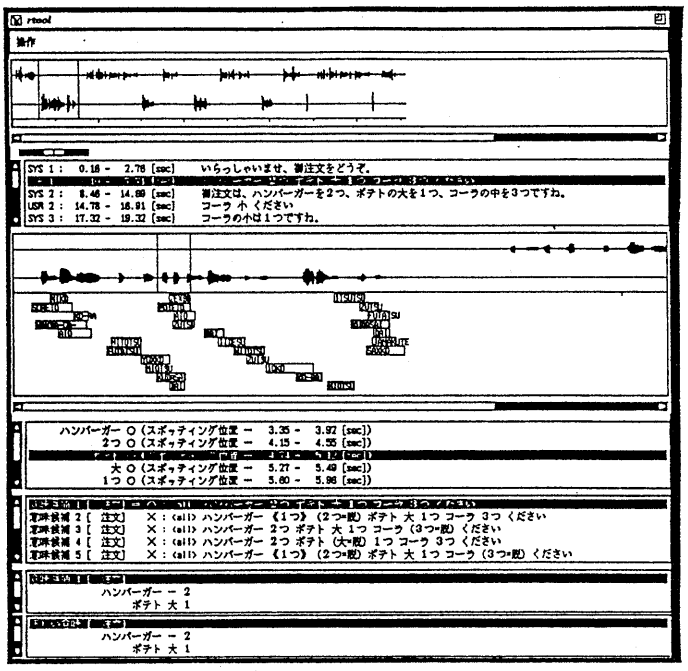

図 6: 評価ツールの表示画面

4.1 性能評価の問題点

対話システムの性能評価においては，被験者の選び方，対話条件の設定の仕方が問題となる．

システムを初めて利用するような初心者を対象とした時には，ユーザの不慣れに起因する認識誤りを起こしやすくなり，システムの性能が過小に評価される．逆にシステムに習熟したユーザの場合には，性能が過大に評価される傾向がある．

また，対話条件の設定もシステムの評価に大きく影響する．TOSBURG II はファーストフードの注文をタスクとしているが，対話実験を実際の場面で行なっているわけではない．そのため，ユーザは必要以上に注文の訂正，追加を行なったり，システムの性能を試すために故意に認識困難な発話をすることがあった．また，「普通のハンバーガー店にいるつもりで話してください」とお願いした実験では，注文数の多いユーザ，あるいは逆に少ないユーザと，ユーザの対話の長さにばらつきがでた．さらに，別の対話条件を設定したところ，大幅に性能が向上する場合もあった．

このように，システム評価においては，ユーザの熟練度や対話条件の設定によってシステムの性能は大きく変わるので，実験に際しては条件を明確にしなければならない．

4.2 ユーザおよび対話条件の影響

対話システムの評価については，対話時間や発話回数という尺度もあるが[10]，ここでは単語検出率，文認識率，対話文認識率という尺度で評価を行なう．

単語検出率は，連続音声中のキーワード検出率であり，文認識率は1発話中の全てのキーワード系列を正しく認識した場合に正解とする．また，対話文理解率は対話処理を経た音声理解結果であり，必ずしもキーワード系列を認識していない場合でも，システムが発話の意味内容を正しく理解した時に正解とする．

＜実験 1＞

初心者18人，熟練者8人を被験者とし，「品目指定なし」と「品目指定あり」の2通りの対話条件を設定して，ユーザの熟練度および対話条件の性能への影響を調べた．品目指定がない場合には，対話に先だって被験者が注文品目を自由に3，4品目決めてから対話を行なった．品目指定ありの場合には「ハンバーガー1個，チーズバーガー2個，フライドポテト小1個，オレンジジュース中3個」を指定した．表1，表2に評価結果を示す．

品目指定のない場合，単語検出率が熟練者で95.9%であり，雑音免疫法の効果により高い認識率を示している．初心者の場合も大差がない．熟練者の場合の文認識率は62.5%であり，キーワードの付加，脱落，置換エラーが1つでもあると不正解となるため，キーワード検出率よりも低くなっている．対話文理解率は70.8%と文認識率より約8%高くなっている．対話文理解率は，さほど重要でないキーワードの検出誤りに影響されず，また，対話履歴により意味内容の補間が可能なためである．熟練者と初心者の文認識率，対話文理解率を比較すると，初心者の場合には咳ばらいや照れ笑い，あるいは躊躇した発声などのため熟練者に比べて正解率が低くなっている．

表 1: 品目指定なしの場合の評価結果

被験者	単語検出率 (%)	文認識率 (%)	対話文理解率 (%)
熟練者	95.9	62.5	70.8
初心者	94.5	57.1	62.8

表 2: 品目指定ありの場合の評価結果

被験者	単語検出率 (%)	文認識率 (%)	対話文理解率 (%)
熟練者	95.6	72.3	76.5
初心者	94.4	56.7	61.9

品目指定の影響を調べるため，表1と表2を比較すると，初心者，熟練者ともに，単語検出率の有意差はない．また，初心者については文認識率，対話文理解率の差異は見られなかった．一方，熟練者については，品目指定により文認識率が72.3%，対話文理解率が76.5%と大幅に向上している．この理由は，初心者の場合には品目指定の有無にかかわらず発話のバリエーションが少なかったためであり，熟練者の場合には自由発話を楽しむ傾向があったため，キーワードの文法として記述されていないような発話が含まれたからである．

＜実験 2＞

被験者20人(そのうち，初心者17人，熟練者3人)について，「注文の際には品名を指定し，サイズおよび個数は指定しない」という対話条件を課し，実験を行なった．

結果を表3に示す．誤認識の多い数詞が含まれていないため，単語検出率は98.9%と，表1の結果に比べ大幅に向上している．同様に文認識率，対話文理解率も88.7%，90.7%となり，10発話に対して1度だけ誤り訂正のための発話が必要な高い性能が得られている．対話に制限のない実験1と比較して，認識の難しい数詞，サイズを指定しないことにより文認識率，対話文理解率が大幅に向上したことがわかる．

一般に，多くの音声応用システムでは数詞の入力が必要であるが，これまでのデモ用の連続音声理解シス

テムでは,都市名や駅名などの入力を扱う場合が多かった.上述したように,数詞の入力が必要な場合,音声理解や音声対話の性能が大幅に劣化するので,システムの評価の際にはタスクの内容について考慮しなくてはならない.

以上の結果から,システムの性能は話者の熟練度や対話条件により左右されることが確認された.また,TOSBURG IIは自由発話理解に基づくユーザ主導型のシステムであり,単語認識率ではなく,使い勝手や対話文理解率で評価しなくてはならないことが明らかとなった.

表 3: 数詞・サイズなし,品目自由の評価結果

単語検出率 (%)	文認識率 (%)	文理解率 (%)
98.9	88.7	90.7

4.3 誤りの原因

対話条件を設けずに初心者13人について対話実験を行ない,収集した159発話について発話理解誤りの原因となる言語現象を調べた。

発話理解誤りを起こした67発話のうち,40発話が表4に示す発話であった.ユーザのいい淀みによる誤認識が28発話あり,この場合,システムは発話を途中で打ち切り,処理を行なうため対話が混乱する.「えー」「えーっと」などの間投詞によるいい淀みも認識率の低下の原因となる.

また,このシステムでは49単語のキーワードを設定したが,「(注文は)以上です」や「(コーヒーは)なし」などのキーワードとすべき語句が出現した.このようなキーワード不足による発話理解誤りが8発話あった.これらは,TOSBURGの作成にあたって行なった模擬対話や実際のファーストフード店での人間同士の対話では現れなかった語句であり,対話システム実験にのみ見られた.この結果は実システムによる対話データ収集の必要性を示している.

表 4: 誤りの原因と内訳

・いい淀み	28発話
・キーワード不足	8発話
・上記2種の混合した発話	4発話
(誤り総数 / 総発話数	67/159発話)

以上のすべての実験において,ユーザの協力的な発話に対するタスク達成率は100%であり,TOSBURG IIは不特定話者に対して充分にロバストなシステムであることが確認された.

5. むすび

本報告では,音声自由対話システムについて論じた.TOSBURG IIの対話データ収集機能と評価ツールを付加し,実対話音声データの収集を行ない,実時間システムによるデータと中間処理結果を収集し,評価・改良に用いてその有効性を確認した.システム評価においては,ユーザのシステムに対する熟練度と対話条件により,システムの性能が大きく変動することがわかった.システムとの対話においては人間同士の対話では見られない発話が観察され,性能低下の要因となることが明らかとなった.

今後はユーザの主観評価も加えてTOSBURG IIの改良を行なっていく予定である.また,本報告の対話システムのデータ収集・評価環境を新しいマルチモーダルインタフェースの研究に活用していきたい.

参考文献

[1] 嵯峨山, 他 : "自動翻訳電話実験システム ASURA の概要," 音講論, 3-4-17, pp. 83–84 (1993.3).

[2] 吉岡, 他 : "電話番号案内を対象としたマルチモーダル対話システムの作成," 音講論, 1-8-19, pp. 37–38 (1993.10).

[3] M. Bates, et al. : "The BBN/HARC Spoken Language Understanding System," *IEEE ICASP93*, Vol II, pp. II-111–II-114 (1993).

[4] V. Zue, et al : "Pegasus: A Spoken Dialog Interface for On-Line Air Travel Planning," *ISSD'93*, pp. 157–160 (1993).

[5] 速水, 他 : "音声対話システムの構築とそれを用いた会話音声収集," 信技報, SP91-101, pp. 79–86 (1991).

[6] 竹林, 他 : "ワードスポッティングによる音声認識における雑音免疫学習," 信学誌, J74-D-II, No.2, pp. 121–129 (1992).

[7] 竹林, 他 : "音声自由対話システム TOSBURG II – マルチモーダル応答と音声応答キャンセルの利用 –," 情報メディア研究発表会資料, pp.93–100 (1992).

[8] 瀬戸, 他 : "音声自由対話システム TOSBURG II による実対話データ収集," 音講論, pp. 121–122 (1993.10).

[9] 竹林, 他 : "不特定話者音声対話システム TOSBURG の開発," 音講論, 1–P–16, pp. 135–136 (1992.3).

[10] 新地, 他 : "音声自由対話システムにおける対話データ収集," 情処全国大会, Vol. 2, pp. 235–236 (1993.3).

[11] 竹林, 他 : "不特定ユーザを対象とした音声対話システムの試作," 人工知能学会研究会資料, SIG-SLUD-9201-4(4/15), pp.27–36 (1992.4).

[12] S. Oviatt : "Predictiong Spoken Disfluencies During Human-Computer Interaction," *ISSD'93*, pp. 53–56 (1993.11).

[13] J. Polifroni, et al : "Experiments in Evaluating Interractive Spoken Language System," *Proc. Speech and Natural Language Workshop*, DARPA, pp. 28–33 (1992.2).