情報処理学会研究報告
IPSJ SIG Technical Report

複数の対話エージェントを扱う音声対話システムの構築と評価

西 村 良 太[†1]　中 川 聖 一[†1]

現在の音声対話システムは，一人のユーザに対して一つのエージェントが対応する 1 対 1 の対話を扱っているが，本論文ではシステム側のエージェントを複数にすることで多人数対話を実現するシステムを構築する．今回は，2 つのエージェントを扱った，三者対話システムの開発を行った．本システムは，我々がこれまでに構築した 1 対 1 対話の音声対話システムを拡張しており，ユーザに対する応答のタイミングや，応答の種類（あいづちなど）の制御を決定木を用いて行っている．また，ユーザからの非流暢な発話に対しても頑健に応答することが可能になっている．エージェントは，2D のキャラクタを，それぞれ 2 台のディスプレイに 1 つずつ表示し，出力音声も別々のスピーカから出力されるようになっている．被験者実験の結果，被験者は三者対話による内容の幅の広がりや，対話の自然性を感じていることが示された．

Development and Evaluation of Spoken Dialog System using Multiple Dialog Agents

Ryota Nishimura[†1] and Seiichi Nakagawa[†1]

Almost all present spoken dialog systems have treated dialog that one user talks with one agent. In this paper, to achieve the multiparty conversation (polylogue, many participates conversation), the number of system agents is increased. Three person's conversation system that treats two agents was developed. This system is extended from the spoken dialog system of two person's conversation that we have developed so far. The response timing to the user and response type are controlled by using the decision tree. The system also reacts robustly to the user's disfluencies. The agent is displayed by the 2D character on two displays respectively one by one. Their speech outputs are also output from two different loud-speakers. As a result of the experiment, the subject felt that the content by multiparty conversation become wider, and that the conversation was natural.

1. はじめに

近年，音声認識技術を用いたインターフェースの需要が高まっており，それに伴って音声対話システムの開発も行われてきている．我々も，これまでに音声対話システムの開発を行ってきており，より自然な対話を実現することが重要であると考え，人間同士の雑談対話中にて生じる種々の対話現象を模倣する音声対話システムを構築した[1)]．このシステムでは，応答として，あいづち，復唱，共同補完などを扱っており，決定木を用いて応答種類と応答タイミングを決定している．また，このシステムは，ユーザからのオーバーラップ発話（バージイン）やユーザからの非流暢な発話に対しても頑健に応答することが可能になっている．

本研究では，ユーザを対話に引き込み，より楽しく対話ができる環境の構築を目指す．その為に，これまでのユーザ対システムという 1 対 1 の対話を，1 ユーザ対多エージェントとの対話に拡張した[2)]．これにより，新しい形態の対話システムを構成することができ，これまで実現不可能であった対話を実現させることが期待される．例えば，エージェント間の上下関係や，ユーザ専属のエージェント，エキスパートエージェントなど知識の差別化を図りることや，考えの異なるエージェントとの対話に発展させることによってユーザに新たな考えをうながす効果が期待できる．

多人数対話の先行研究として，Dielmann ら[3)] は，多人数対話での DialogAct を自動で付与するためのモデルの学習を行っている．Ginzburg ら[4)] は，二話者対話プロトコルを，多人数対話にスケールアップする方法についての研究を行っている．多人数対話では，質問に対する応答発話や確認発話などが，二者対話に比べて遠い距離で（3 発話以上あとに）現れる場合が多くある．これに対応する為に，スタックを用いた対話処理を行っている．

浅井ら[5)] は，複数の人間と複数の対話エージェントによる多人数対話において，対話エージェントが状況に応じた働きかけを行うことで，全体のコミュニケーションを活性化させている．対話はテキストベースの対話システムで行われており，2 名のユーザと，2 つのエージェントが対話に参加している．対話ドメインは，人物当てクイズである．2 つのエージェントは，出題エージェントと回答エージェントに分かれており，両方が共感的発言や自己中心的発言を行う．対話実験の結果，ユーザの満足度やユーザの発言数を増加させる効果があ

†1 豊橋技術科学大学　情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology

ることが示され，エージェントからの共感的発言がユーザ満足度を更に向上させ，対話を活性化させている．

このように，複数のエージェントとの対話はユーザ満足度の向上や対話の活性化に繋がることが示唆されている．しかし，浅井らの実験はテキストベースのシステムで行われており，音声対話システムでの効果は分からない．

岡本ら[6] は，複数エージェント対話システムを構築する際の，エージェント同士の自然な対話を実現するために，どのような非言語動作をどの時点で取るべきかを明らかにしようとしている．分析には漫才を用いている．この理由としては身体動作への制約が最小限であり，対話のみで情報伝達が行われているからである．分析の結果，対話全体として，エージェントの視線が相方，姿勢が観客である場合が多かった．動作に制約がない漫才においても，観客への姿勢配分が大きくなることから，姿勢（ポスチャ）に注目する必要性がある．

岡本らの指摘からは，エージェントの表示と，姿勢・視線の制御が必要であることが示されている為，複数エージェントの対話システムを構築する際には，この条件を満たすエージェント表示部も必要になる．

これらのことをふまえ，我々は，複数の対話エージェントを扱う音声対話システムの開発を行う．ユーザを対話に引き込み，ユーザの満足が得られる対話システムの構築を目指す．

2. 対話システム

これまで我々が開発してきた音声対話システムは，ユーザ対システムの 1 対 1 の対話を扱ったものであったが，これを，”性格の異なる 2 つのエージェント (システム) とユーザとの 3 人対話”に拡張した[2]．エージェント間では，実際に発話した内容以外にも，すべての情報が共有できる為，様々な対話制御が可能となり，広い応用が考えられる．今回構築した三者対話用の音声対話システムの概略図をを図 1 に示す．このシステムでは，音声認識した結果から，テンプレートマッチングによって応答文を生成している．また，韻律素性を決定木に入力することで，応答の種類とタイミングを決定している．詳細については，以下の節で述べる．

2.1 対話ドメイン

システムとの対話内容としては，誰でも対話ができ，また，三者対話において，ユーザの引き込みを実現させることができるものが好ましい．このことから，2 つの物/事柄の好き嫌い/賛成反対の話を扱う．今回は,「うどんとラーメンのどちらが好きか」といった話題で対話を行うようにした．

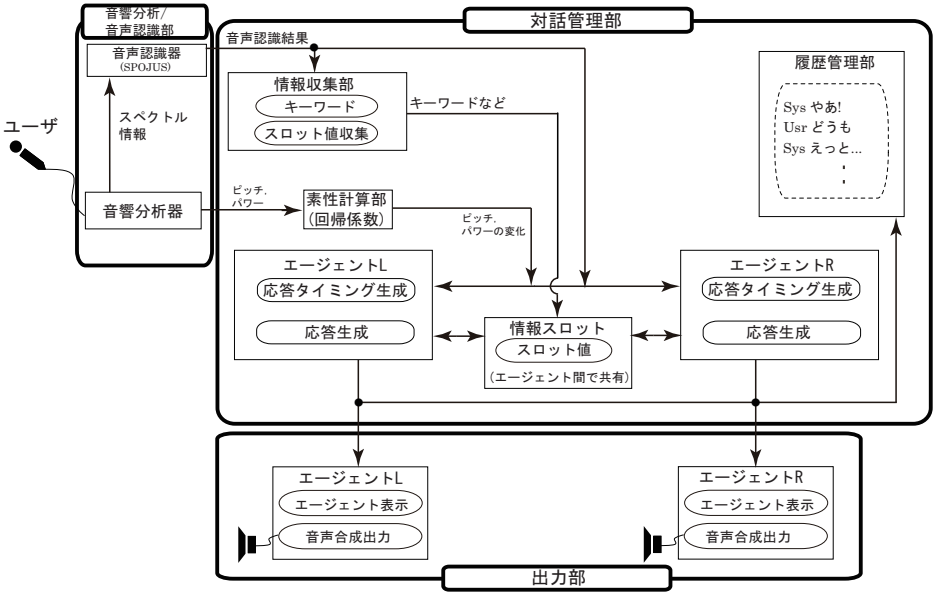

図 1 三者対話システムの概略図

二人のエージェントが，うどんとラーメンについてそれぞれ良い点・悪い点を示して対話を進めていく．この際，エージェント間で意見を対立させ, ユーザをどちらかの意見に引き込むことや，エージェント間の意見を揃えて，ユーザを特定の意見に引き込むといった戦略も考えられる．

好き嫌いデータベースは Web から収集する予定であるが，現段階では人手で作成している．

2.2 音響分析・音声認識部

本システムで用いる音声認識器には，本研究室で開発された SPOJUS[7),8)] を用いる．SPOJUS には，2 つのバージョンがあり，一つは n-gram を用いた大語彙連続音声認識用のもの，もう一つは CFG (Context Free Grammar) を用いたものがあり，今回は，CFG 版の SPOJUS を用いている．

音声認識と同時に，本システムでは，音響分析として韻律情報の抽出も行っており，ピッチ・パワー情報を抽出して応答タイミング生成部へ送信している．これは，決定木の素性と

して用いている．

2.3 対話管理部

対話管理部は，以下に示すサブコンポーネントから構成されている．

2.3.1 素性計算部[1)]

ここでは，音響分析器から得られた音響分析結果を元に，韻律素性を計算している．素性としては，ピッチ (F0)，パワーの回帰係数を求め，これを応答タイミング・応答種類制御をする決定木の入力として用いる．

2.3.2 情報収集部

ここでは，音声認識器からの認識結果から，必要な情報を抽出し，スロットに格納している．スロットに格納された値は，応答生成に用いられる．これにより，ある程度文脈を考慮した対話が可能になっている．また，名前やエージェントの一人称などを保持しておくことで，応答テンプレートの汎用性を高めている．

今回は，対話ドメインが「うどんとラーメンについての話」であることから，スロットの例としては,「ユーザが好きなもの」「その食べ物が好きな理由」「もう一方の食べ物が嫌いな理由」などの情報を認識結果から抽出し，対話を行う．

2.3.3 情報スロット

対話中の重要な情報がスロットに格納されており，これらについては，エージェント間で情報を共有している．この情報を参照して，ユーザの嗜好に合わせた共感発話を行い，対話を盛り上げる方向に進めるようにしている．また，共有している情報を元に，対話の流れ（シナリオ）を変化させ，情報を応答に盛り込み，結論の誘導を行うことが出来る．

2.3.4 応答生成部

本システムでの各エージェント内の応答生成には，テンプレートマッチングを用いている．入力された音声を音声認識し，その結果と応答用テンプレートとのマッチングを行って，マッチするものに対して，それに対応した応答文を出力として用意する．出力文を生成する際には，スロット情報も用いて，文脈を考慮した応答文生成を行うことができる．

また，応答戦略として，サブタスク（サブシナリオ）を定義することで，文脈を考慮した対話が可能になっている．以下に，テンプレートの例を示す．以下の例では,「きつねうどん」「とんこつラーメン」についての応答のテンプレートも示してあるが，これらは好き嫌いに関する具体的な例であり，これについては，好き嫌データベースから読み込んで生成される．

```
[first prompt]
@（.*)
＝ ; うどんとラーメンだったらどっちが好き？;subtask:1,sentence:1;
[topic]
@（food)
＝ subtask:1,sentence:1,nowTopic:.+; じゃあ、$2だったら好きな種類は?;sentence:2;
@（きつね)
＝ subtask:1,sentence:2,nowFood:.+;$2 は、揚げがおいしいよね。;sentence:3;
@（とんこつ)
＝ subtask:1,sentence:2,nowFood:.+;$2 は、こってりがおいしいよね。;sentence:3;
@（.*)
＝ subtask:1,sentence:0,nowTopic:.+;$2 もおいしいよ。$2 ではなにが好き?;sentence:2;
```

記述方法は，以下の通りである．

@ マッチングルール
＝ スロット条件; 出力文; スロット書き換え; アニメーションコマンド

マッチングルールは，正規表現で記述する．一つのマッチングルールに対して，出力文（「＝」行）はいくつでも記述することができる．アニメーションコマンドは,「nod」でうなずきを行うなど，エージェントの動作を記述することができる．

上記の例から生成される対話例を以下に示す．

システムＬ：うどんとラーメンだったらどっちが好き？
ユーザ　　：えっと，うどんが好きだよ．
システムＬ：じゃあ，うどんだったら好きな種類は？
ユーザ　　：きつねうどんかなぁ．
システムＬ：きつねうどんは，揚げがおいしいよね．
ユーザ　　：そうだね〜．
システムＲ：ラーメンもおいしいよ．ラーメンでは何が好き？
ユーザ　　：とんこつラーメンとか良いな〜．
システムＲ：とんこつは，こってりがおいしいよね．

マッチングルールにマッチすると,「＝」行のスロット条件が判定され，条件を満たして

いる場合には応答文を出力し，スロット値を書き換え，エージェントアニメーションをさせる．上記のマッチ文の例は，1 つのサブタスクを示している．最初にシステムを起動すると，[first prompt] の中から，文を選択し出力する．今回の場合は，システムから「うどんとラーメンだったらどっちが好き？」と発話し，その際に，右に記した 2 つのスロットの値を書き換えている．次に，ユーザの発話がなされたあと，その発話とのマッチングをとる．[topic] では，ユーザ発話内容と，サブタスク番号，サブタスク内の文番号によって，シナリオの展開を記述している．このようにすることで，システム同士の掛け合いが可能となる．

現在のシステム同士の対話の発話タイミングは，相手のエージェントの発話が終了した直後になっている．ユーザ入力に対するエージェントからの応答については，決定木でタイミングを決定しているので，将来的にはエージェント間の対話のタイミングの制御も行いたいと考えている．

2.3.5 応答タイミング生成部[1)]

今回構築したシステムで用いる応答タイミング生成の手法は，我々が先行研究で用いていた手法と同じものである[1)]．このシステムでは，ユーザの発話中・ポーズ中に関わらず，全てのセグメント (100ms 毎) に対して，応答するかどうかの判定を行っており，ユーザ発話にオーバーラップする応答を返すことが出来る．

応答タイミング生成器は，決定木にて韻律素性を用いて応答タイミングを生成する．また同時に，応答生成器にて生成された応答の中から適切な応答を選択する．応答タイミングの決定に関しては，以下の素性を用いている．

- ユーザ発話開始時点からの経過時間 (F1)
- ユーザ発話終了時点からの経過時間 (F2)
- システム発話終了時点からの経過時間 (F3)
- ユーザ発話の最後の 100ms 区間のピッチ・パワーの傾き (F4)
- ユーザ発話の最後の 500ms 区間のピッチ・パワーの傾き (F5)
- 認識結果（途中結果も含む）の最終単語の属性 (F6)

図 2 に，各素性の関係を図示する．

決定木では，応答生成器にて応答が準備できているかどうかも素性として用いる．各応答種類毎に一つの応答が準備される．各素性は，100ms 毎に決定木に入力され，応答すべきかどうかの判定と，応答する場合には適切な応答種類の判定を行う．選択される応答の種類には,「あいづち・復唱・一般的な応答・待ち」がある.「待ち」の場合には，応答を出力しない．応答の回数は，1 回のユーザ発話に対して 1 回のシステム応答に制限されているが，あ

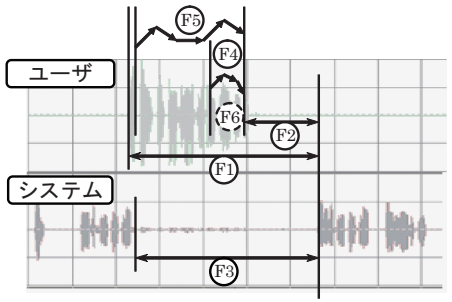

図 2 決定木で用いる素性

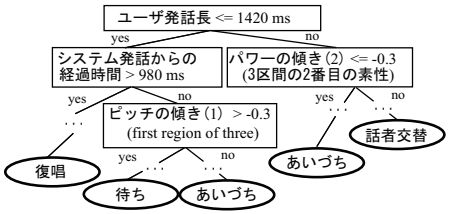

図 3 決定木の一部

いづちと復唱に関してはこの制限はない．つまり，1 回のユーザ発話に対して，共同補完と一般的な応答は 1 回応答することができ，あいづち・復唱は何度も応答することができる．

決定木の学習には，RWC コーパス[9)] を用いており，このコーパスには，対話ドメインとして「自動車販売」「海外旅行計画」の 2 種類の対話が合計 48 対話含収録されている．対話は二者対話である．データ量としては，コーパス全体で 6.5 時間分あり，各対話の時間は 10 分程度である．また，発話数は，16,399 発話である．一方の話者は，実際の自動車販売員，または旅行代理店員であり，もう一方の話者は，12 人の一般人である．このコーパスを用いて決定木の学習を行った．決定木の学習器には，C4.5[10)] を用いた．

2.3.6 履歴管理部

対話履歴を保存しておき，後に参照して文脈を考慮した対話戦略を実現する為の部分である．この部分については現段階では対話に利用していないが，今後は，対話履歴の情報を活用する対話を行いたいと考えている．

2.4 出 力 部

出力部では，対話管理部から送られてくる出力結果を，各エージェントから出力する．対話管理部から送られてくる出力結果には，エージェントの発話内容，アニメーション内容の情報が記述されており，それに基づいて映像，音声にて出力する．

各エージェントはそれぞれ別々の画面（PC）に表示される．また，音声も別々のスピーカ（PC）から出力される．以下に詳細を述べる．

2.4.1 エージェントの表示方法

今回は，エージェントの表示方法としては，2 つの画面に個別に表示する手法を用いる．また，エージェントの表示には，NHK 放送技術研究所にて開発された TVML(TV program

Making Language)[11] を用いた．表示するエージェントについては，アニメキャラクターのような 3D モデルを用いた（TVML オプションパック内の「abeno(男性)」と「suyama(女性)」）．TVML は，元々はスクリプトを記述することで TV 番組を簡単に作成できるということが特徴であるが，外部からリアルタイムに制御することができる．今回は，外部制御プログラムを改良し，TCP/IP 通信にて，発話内容とアニメーション内容を受信できるようにした．

待ち状態の場合には，体が少し揺れたりするなどのアニメーションを行うことも可能になっている．また，音声出力を行っている間は，音声合成器から発話時間を取得し，その時間に合わせて口をパクパクと動かして，喋っていることを表現することもできる．この場合のアニメーションは，厳密なリップシンクではないが，出力音声の大きさに応じて，口を開く大きさが変化するようになっている．

岡元ら[6] は，エージェントを制御する際のポイントとして，以下のことを挙げている．

- ユーザに明示的に語りかける場面においては非言語チャネルの指向性を同調させる（内部指向性）
- 対話相手のエージェントに向けた発話では基本的に視線を同調させ（対話相手へ向け）つつ姿勢をユーザに向ける（外部指向性）

この指摘から，複数エージェントによる対話では，エージェントの視線と姿勢の制御が重要であり，これらを表示・制御できるものが好ましい．頭部のみを表示するもの（Galatea Toolkit に含まれる顔合成ソフトなど）では，体が表示されていないため，姿勢の制御が出来ない．これらのことから，今回は，本システムの目的に合ったエージェント表示部として，TVML を用いた．

アニメーションコマンドは，TVML スクリプト言語で記述する．例として「character: gaze(name=abeno, yaw=80)」と記述した場合には，キャラクター（名前：abeno）の視線を 80 度移動させることができる（首と胴体を回転させる）．なお，現在のエージェントは，いつも発話している相手の方を向くようになっている．エージェント L は，エージェント R が喋ればエージェント R の方を向き，ユーザが喋ればユーザの方を向く．発話しているエージェントは，発話内容に応じて，呼びかける相手の方を向くようになっている．

2.4.2 音声出力部

音声出力は，音声合成器を用いて行う．音声合成には，TVML インストールプログラムに含まれている GalateaTalk（擬人化音声対話エージェントのツールキット Galatea Toolkit[12] に含まれる音声合成器）を用いている．この音声合成器は，発話者タイプ（男女など）の変更や，抑揚・話速を自由に変更できる．

本システムでは，対話エージェントを２つ扱っており，差別化を図るために，エージェントは，それぞれ男と女のエージェントとしており，出力音声もそれにあわせて変更している．

表 1 質問１に対するアンケート結果

評価点	1	2	3	4	5
人数	0	1	1	3	0

表 2 質問２に対するアンケート結果

評価点	1	2	3	4	5
人数	0	1	2	2	0

3. 被験者実験

3.1 実験内容

開発した三者対話システムを用いて，被験者対話実験を行った．被験者は 5 名の男性であり，音声関連の研究室の学生である．全員，筆者が以前に開発した二者対話システム[1] を使用した経験がある（天気に関する雑談対話）．被験者は 1 名毎に三者対話システムと対話を行い，その後アンケートに記入をした．アンケート項目については，対話前に確認を行った．アンケートは以下の項目で行われた．

- 質問１：対話は楽しかったか（5 段階評価）
- 質問２：またこのシステムを使いたいと思ったか（5 段階評価）
- 質問３：三者対話システムの良かった点（自由筆記）
- 三者対話システムの悪かった点（自由筆記）
- 以前の二者対話システムと比較してどのように感じたか（自由筆記）
- その他の感想（自由筆記）

3.2 実験結果

実験結果として，被験者からのアンケートの結果を示す．質問１に対するアンケート結果を表 1 に示す．５段階評価の平均点は，3.4 点である．被験者５名の内，３名は４点を付けており，比較的に楽しく対話が出来ていた．質問２に対するアンケート結果を表 2 に示す．５段階評価の平均点は 3.2 点である．使いたくない（１点）と評価した被験者はいなかったが，使いたい（５点）と評価した被験者もいない．

これらのアンケート結果に対して，質問３～６への自由筆記形式の回答を参照することで，被験者の意見を分析する．質問３の三者対話システムの良かった点については，「２つ

のエージェントとやりとりするので，対話内容の幅が広がっているように感じる」,「一方のエージェントが他方のエージェントを見るなど，１対１対話にはない動きがあり自然に感じた」などの意見が挙がった．今回構築した三者対話システムによって，被験者が対話をより広く，自然に感じることができることが示されている．

質問４の三者対話システムの悪かった点については,「エージェント間のやりとりがなく三者による議論に感じなかった」,「エージェント間の会話のテンポが悪かった」などの意見が挙がった．前者については，対話の流れで，エージェント同士の対話に繋がらない場合があったためであり，これについては，対話シナリオの拡充などによって対話制御の幅を広げ，エージェント間対話を活発に行う必要がある．後者については，現在のシステムではエージェント間の発話タイミングについては，固定値を用いているためである．これについては，被験者とエージェントの間の対話のリズムを，エージェント間の対話に取り入れ，対話全体のリズムを制御する必要がある．

質問５の二者対話と比較してどのように感じたかについては,「エージェントにそれぞれ個性があり，対話が楽しく感じた」などの意見が上がり，エージェントの数が増えたことによって，被験者のシステムに対する印象が良くなっている．

質問６のその他の感想については,「音声認識がうまくいかず，対話が失敗する」などの意見が挙がっていた．これについては，誤認識した場合に対しても，自然な対話になるように対話制御を行うなどの対策が必要である．

4. ま と め

本論文では，これまでに我々が開発してきた二者音声対話システムを拡張した，１ユーザ対２システムエージェントによる三者対話が可能な音声対話システムの開発を行った．本対話システムでは，ユーザの嗜好（うどんとラーメン）についての話題を通して，ユーザを対話システムに引き込む戦略をとっている．システムは，ユーザ入力から重要な情報を抽出（スロットフィリング）して，それを応答に組み込み，対話を行うことができる．また，このスロットフィリングを行うことによって，ユーザ入力に対して頑健に応答を返すことが可能になっている．ユーザ入力に対するシステム応答については，応答タイミングを決定木で制御しており，適切なタイミングであいづちを返すことも可能になっている．エージェント表示に関しては，3D のアニメーションにより，うなずきや簡単なリップシンクなどを表現する．これを２台のディスプレイに表示させることで，それぞれのエージェントを別のものとしてユーザが識別できるようにした．また，出力音声についても，別々のスピーカ（PC）から出力している．被験者実験の結果，被験者は三者対話による内容の幅の広がりや，対話の自然性を感じているが，まだ多くの課題がある．

本システムでは，ユーザとシステムとの間の対話においては，応答タイミングが考慮されているが，システム同士の掛け合いのタイミングについては考慮されていない．今後は，全体的なリズムの制御も行いたいと考えている．また，より詳細な被験者実験とその分析を行い，ユーザ満足度や，ユーザの引き込まれ度など，三者対話の有効性について，調査・分析を行いたい．

参 考 文 献

1) 西村良太，中川聖一：応答タイミングを考慮した音声対話システムとその評価，音声言語情報処理（SLP）研究報告，Vol.2009-SLP-77, No.22 (2009).
2) 西村良太，中川聖一：複数の対話エージェントを扱う音声対話システムの開発，音声言語情報処理（SLP）研究報告，Vol.2010-SLP-080, No.7 (2010).
3) Dielmann, A. and Renals, S.: DBN Based Joint Dialogue Act Recognition of Multiparty Meetings, Proceedings of ICASSP ’07，pp.133–136 (2007).
4) Ginzburg, J. and Fernández, R.: Scaling up from Dialogue to Multilogue: Some Principles and Benchmarks, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pp.231–238 (2005).
5) 浅井亮太，堂坂浩二，東中竜一郎，　南泰浩，前田英作：多人数対話における対話エージェントのコミュニケーション活性効果，言語処理学会第 15 回年次大会発表論文集 (2009).
6) 岡本雅史，大庭真人，榎本美香，　飯田仁：対話型教示エージェントモデル構築に向けた漫才対話のマルチモーダル分析 (< 特集 > ソーシャルインテリジェンス)，日本知能情報ファジィ学会，Vol.20, No.4, pp.526–539 (2008).
7) 甲斐充彦，中川聖一：日本語連続音声認識システム SPOJUS-SYNO の改良と評価，電子情報通信学会技術報告, SP93-20 (1993).
8) Zhang, J., Wang, L. and Nakagawa, S.: LVCSR based on context dependent syllable acoustic models, *Asian Workshop on Speech Science and Technology, SP2007-200*, pp.81–86 (2007).
9) 田中和世，速水　悟，山下洋一，鹿野清宏，板橋秀一，岡　隆一：RWC 計画における音声対話データベースの構築，情報処理学会音声言語情報処理１１－７ (1996).
10) J.Quinlan, R.: C4.5:Programs for machine learning, *Morgan Kaufmann* (1992).
11) `http://www.nhk.or.jp/strl/TVML/`.
12) 嵯峨山茂樹，川本真一，下平　博，新田恒雄，西本卓也，中村　哲，伊藤克亘，森島繁生，四倉達夫，甲斐充彦，李　晃伸，山下洋一，小林隆夫，徳田恵一，広瀬啓吉，峯松信明，山田　篤，伝　康晴，宇津呂武仁：擬人化音声対話エージェントツールキット Galatea，情報処理学会研究報告 (2002-SLP-45-10) (2003).