グループ会話対応型会話エージェントにおける受話者推定システム

Identifying the Addressee in Multiparty Human-Agent Conversations

馬場　直哉 [1*]　　黄　宏軒 [2]　　中野　有紀子 [3]

Naoya Baba[1], Hung-Hsuan Huang[2], and Yukiko Nakano[3]

[1] 成蹊大学大学院理工学専攻理工学研究科
[1]Graduate School of Science and Technology, Seikei University
[2] 立命館大学情報理工学部情報コミュニケーション学科
[2] Department of Information & Communication Science, Ritsumeikan University
[3] 成蹊大学理工学部情報科学科
[3] Department of Computer and Information Science, Seikei University

Abstract: In multiparty human–agent interaction, the agent should be able to properly respond to a user by determining whether the utterance is addressed to the agent or to another person. This study proposes a model for predicting the addressee by using the acoustic information in speech and head orientation as nonverbal information. First, we conducted a WOZ experiment to collect human–agent triadic conversations. Then, we analyzed whether the acoustic features and head orientations were correlated with addressee-hood. Based on the analysis, we propose an addressee prediction model that integrates acoustic and bodily nonverbal information using SVM.

1. はじめに

ショッピングモールや博物館等の公共施設では，複数人から構成されるグループで情報提供端末を利用することが多い．そこでは，グループ構成員が互いに相談しながら，必要に応じて情報を得るための操作を行う場合が一般的である．このようなグループユーザが利用できる情報キオスクとして会話エージェントを実現するには，多人数会話特有の機能を実装することが不可欠である．例えば，あるユーザがエージェントに対して話しかけている場合には，エージェントは，その発話に対して，正確に反応するべきである．一方，あるユーザがもう 1 人のユーザに話しかけている場合には，エージェントは，ユーザ同士の会話に敢えて介入する必要はないだろう．このような振る舞いのできる会話エージェントを実現するためには，ユーザ発話の受話者を推定し，発話がエージェントに向けられているか否かを判別する機能が必要となる．そこで，本研究では，2 人のユーザとエージェントとの 3 人会話において，ユーザからの問いかけに適切に応答できるグループユーザ対応型会話エージェントの実現を目指し，受話者推定方式を提案する．先行研究[1]では，受話者推定において，顔向きを含めた視線情報が有用であることが既に報告されているが，本研究では，これに加え，ユーザの音声発話の韻律的特徴を用いることにより，受話者推定の精度向上を狙う．

本研究のアプローチとして，まず，WOZ(Wizard of Oz)実験により，人間 2 人とエージェントとの 3 人会話のデータを収集する．次に，受話者推定に有用なパラメータを決定するために，音声情報と顔向き情報を分析する．最後に，この分析結果に基づき，受話者推定システムを構築し，方式評価を行った結果について報告する．

2. 関連研究

ターン交代において視線が重要であることが，コミュニケーション研究で既に知られているが[2]，Takemae ら[3]は，発話者の視線は受話者に向けられ，

*連絡先：成蹊大学大学院理工学研究科
〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1
Email:dm116233@cc.seikei.ac.jp

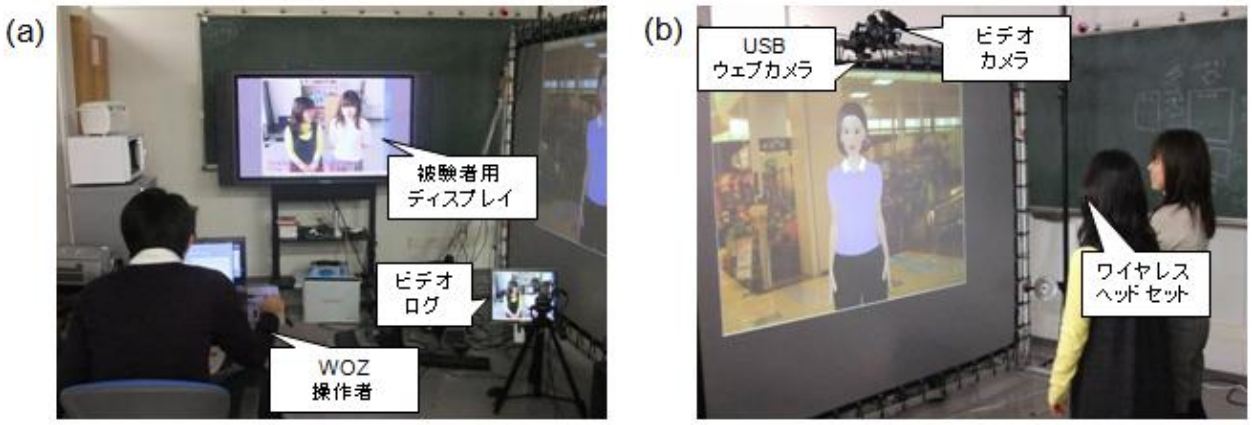

図 1：実験概要図

ターン管理を制御する機能を有していることを示している．Frampton らは[4]，音声認識と画像処理により自動的に抽出された，言語，音声，視覚的情報を用いて，人間同士の多人数会話において，受話者を決定するモデルを提案している．このモデルについては，60%の分類精度であると報告されている．

また，より本研究と関連性の高い研究として，Katzenmaier らは[5], 音声情報と視覚的情報の統合により，2 人のユーザとロボット間の会話において受話者を特定する方式を提案している．彼らは受話者推定において，音声認識から得られる言語情報を利用しており，発話がロボットに向けられていることを認識する精度として, F 値 0.72 を報告しているが，システム実装までは行われていない．

3. 会話コーパスの収集

人間とエージェントとの多人数会話において，受話者の分析に用いるビデオコーパスを収集するために，ペアの被験者に 2 種類のタスクを課した WOZ 実験を実施した．

3.1. 実験の概要

実験では，図 1 に示すように，ペアの被験者がスクリーンから約 1.5m 離れて立ち, 等身大の女性のバーチャルキャラクターとインタラクションした．各ペアの被験者には，以下の 2 つのタスクに取り組んでもらった．

履修登録タスク：被験者 2 人には，来学期に一緒に出席する 12 の授業のうち 3 つを選ぶよう教示をした．被験者は，チューターの役割を持つエージェントに質問することにより，授業に関する情報を得ることができる, なお, 被験者間の議論を活発にするため，各被験者に，週のうちある半日を忙しい日とし，一緒に授業に参加できない時間帯を制約として設定した．

旅行計画タスク: 京都の観光スポット 14 ヶ所のうち 3 ヶ所を自由に回ることができる旅行クーポンを入手したという設定で，旅行代理店のスタッフとなったエージェントから情報を得ることにより，訪問する場所を 2 人で話し合って決めるよう被験者に教示した．

3.2. 収集データ

本実験の被験者は，同性の友達同士の 21 組，計 42 人の大学生，大学院生であり，平均年齢は，20.1 歳であった．21 組のうち，男性ペアは 14 組，女性ペアは 7 組であった．不必要に笑いが起きるなど，インタラクションの質が十分ではないセッションを除外したため，本研究では，男性 10 組，女性 7 組の計 34 人を分析対象とした. エージェントと被験者とのインタラクションについては，前方と後方からの二台のビデオカメラによって録画した．これらのビデオカメラに加えて，図 1(b)に示すように，スクリーン上部に USB ウェブカメラを設置した. 被験者には，音声データの収録のため Bluetooth のワイヤレスヘッドセットを身につけてもらった．

4. コーパスの分析

4.1. データの切り出しとアノテーション

3 節において収集した音声データを発話単位で分析するために，発話データの切り出しを行った．音声認識エンジン Julius[1]を用い，200ms の無音区間が検出された場合，そこを発話の区切りとして，自動で音声の切り出しを行った．この処理により，全 1,830 発話を抽出した．

[1] http://julius.sourceforge.jp/

また，ビデオデータのアノテーションには，ビデオアノテーションツール Anvil4.7.7[2]を用い，発話の話し手を発話者，発話の受け手を受話者と定義し，発話者と受話者について，発話ごとにラベリングを行った. ラベリング結果を表 1 に示す. これにより，エージェントに対する発話，863 発話，もう一人のユーザに対する発話，967 発話が収集され，これらの各発話について韻律的分析を行った.

表 1：受話者がエージェントとユーザの発話数

受話者 / 性別	Agent	Partner	合計
男性ペア	509	522	1031
女性ペア	354	445	799
合計	863	967	1830

4.2. 音声情報の分析

受話者推定の最も重要な韻律情報[6]として，ピッチ，パワー，話速に着目するとともに，発話継続長も分析対象とし，エージェントに話しかけている時と，ユーザ同士で話している時の差について分析を行った.

4.2.1. 音声特徴量の抽出

ピッチ，パワー，話速の抽出には，音声分析ツール Praat[3]を用いた. ピッチとパワーについては，Praat スクリプトを記述し，それぞれの発話から 0.01 秒ごとにピッチとパワーの値が出力されるようにした. また，話速に関しては，Praat から算出した音節数と発話継続長から，1 秒あたりの音節数を求め，これを話速とした. 発話継続長は，Julius によって自動で切り出された発話区間の開始と終了時間との差を取ることにより算出した.

4.2.2. 音声情報と受話者の関連性

図 2 に，ある発話においての実際に算出される韻律情報の値の例を示す. 発話内容欄の(a)の発話は，エージェントに向けられたものであり，(b)の発話は，ユーザに向けられたものである. 発話内容の下に示されているグラフは，ピッチとパワーの値のプロットである. 上段のピッチのグラフより，発話の F_0 はもう一人のユーザに対して話しかける場合(平均値:120Hz)よりも，エージェントに対して話しかける場合(平均値:167Hz)においてより高いことがわかった. 同様に，パワーにおいてもユーザに対して話しかける(平均値:64dB)より，エージェントに対して話しかける(平均値:61dB)方がより大きいことが分かった. この結果に基づき，ピッチとパワーの平均値，話速，発話継続長について，全被験者の平均値を算出し分析を行った. なお，ピッチに関しては性別により大きな違いがあるため，別々に分析した. 分析結果を図 3 に示す. これらの結果は図 2 の結果と一致するものであり，ピッチとパワーについては，エージェントに対する発話は，もう一人のユーザに対する発話より高くなり，発話継続長においても，エージェントに対して発話する方が長いという結果となった. 一方，話速に関しては，エージェントに対して発話する方が遅いという結果になった. t 検定を行ったところ，女性のピッチを除くすべての分析に

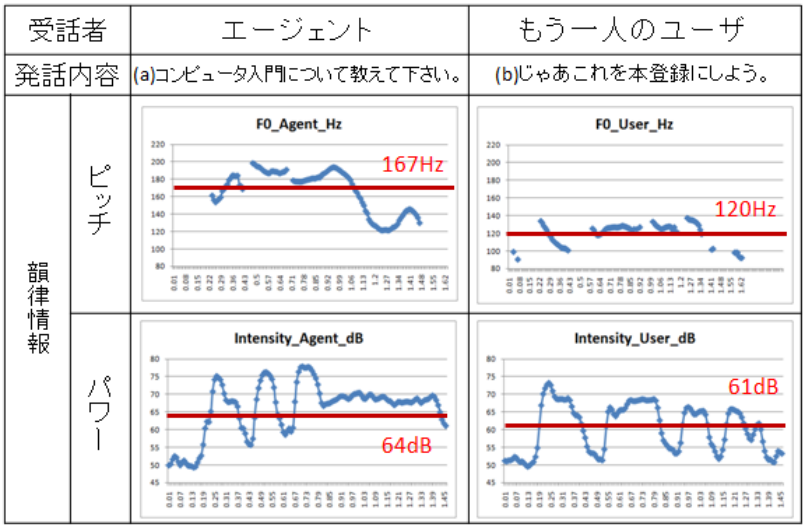

図 2：ピッチとパワーの出力例

[2] http://www.anvil-software.de/
[3] http://www.fon.hum.uva.nl/praat/

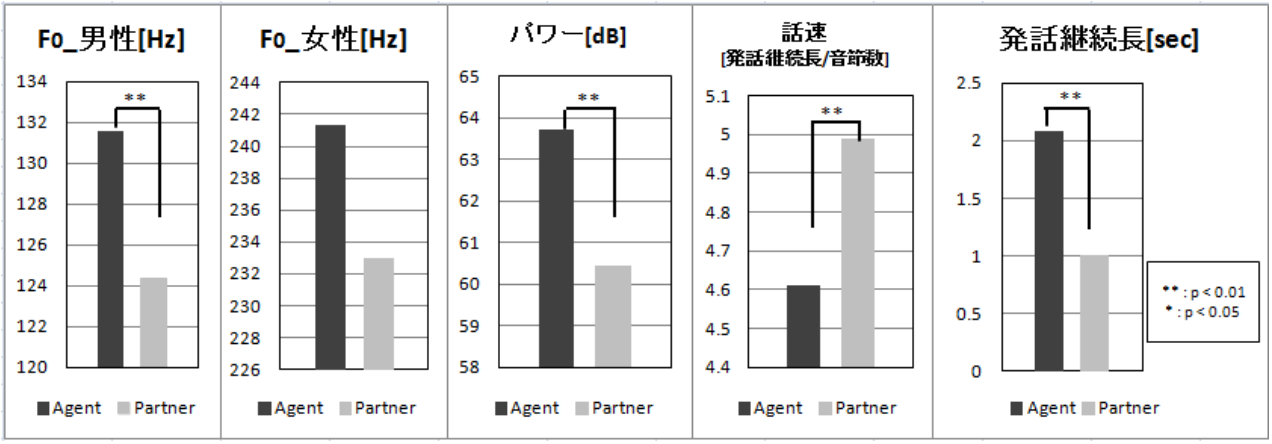

図 3：音声情報分析結果

おいて有意差がみられた(男性 F_0: t(19)=4.4, p<0.1, 女性 F_0: t(13)=1.5, p<0.8, パワー: t(33)=12, p<0.1, 話速: t(33)=-4.6, p<0.1, 発話継続長: t(33)=16.5, p<0.1). このことから，ピッチとパワーについては，各発話の平均値が，受話者推定に有効であることが確認できた．そして，ユーザがエージェントに対して話しかける時は，比較的高い声で，ゆっくりと大きく話す傾向があり，これらの音声情報が受話者推定に有用であることが示唆された．

4.3. 顔向き情報の分析

4.3.1. 顔向きの自動判定

次に，実験で収集したビデオデータに対し，顔認識ソフト FaceAPI[4]を用い，分析を行った．FaceAPI により，三次元の頭部の位置と回転角度が取得できる(30fps). 被験者の顔向きを自動で判別するために，データマイニングツール Weka[5]の J48 による決定木学習を行った．顔向きの教師データには，正面(Agent)，左(Left)，右(Right)の 3 種類のラベル付けを 4 ペア計 8 人のビデオデータについて行った結果を用いた．トレーニングデータには，FaceAPI の頭部 3 次元位置，3 軸回転角度，頭部姿勢測定の信頼度の計 7 種類を用いた. 決定木学習の結果を表 2 に示す. 10 回の交差検定における分類精度は，97.2%であり，十分な精度が得られたため，このモデルを用い，残りのビデオデータに対し顔向きの自動ラベリングを行った．

4.3.2. 顔向きと受話者の関連性

ユーザは発話時に受話者の方向を必ず見るといった，高い相関性がある場合には，顔向きのみで受話者を推定することが可能である．そこで，顔向きラベルと受話者ラベルとの一致度を算出した．発話中のビデオデータから抽出した計 49,170 フレーム中，受話者ラベルと顔向きラベルが一致しているフレーム数は 36,867 フレームあり，不一致の割合は約 25%であった．この結果より，顔向きは受話者の推定には有用であることは確かであるが，それのみで完全に推定できるものではないことがわかった．

表 2：顔向き自動編別の評価結果

	適合率	再現率	F 値
Agent	0.982	0.984	0.983
Left	0.953	0.949	0.937
Right	0.929	0.900	0.914
分類精度	97.2%		

5. 受話者推定方式

5.1. 特徴量の設定

音声情報と顔向き情報を統合した受話者推定モデルを確立するために機械学習を行った．

5.1.1. 音声情報の特徴量

音声情報に関する特徴量として, 4.2 節で分析したピッチ，パワー，話速，発話継続長に関し，発話ごとに以下の特徴量を設定した．

(1) ピッチの平均値
(2) パワーの平均値
(3) 話速
(4) 継続時間
(5) (1)と全被験者のピッチの平均値との差
(6) (2)と全被験者のパワーの平均値との差

[4] http://www.seeingmachines.com/product/faceapi/
[5] http://www.cs.waikato.ac.nz/ml/weka/

表 3：10 回の交差検定による受話者推定モデルの評価結果

		ピッチ	パワー	話速	音声	顔向き	パワー+顔向き	音声+顔向き
F値	エージェント	0.702	0.715	0.687	0.717	0.759	0.798	0.799
	パートナー	0.785	0.776	0.778	0.781	0.656	0.809	0.806
分類精度		75.0%	74.9%	74.0%	75.3%	71.6%	80.3%	80.3%

5.1.2. 顔向き情報の特徴量

顔向きに関する特徴量として, 4.3 節で行った顔向きの自動判定結果から，エージェントの方を向いている(agent)，もう 1 人のユーザの方を向いている(user)，それ以外の方を向いている(elsewhere)の 3 種類の顔向きを設定した．そして，これらの各発話中の比率と, agent→user, agent→elsewhere, user→agent, elsewhere→agent の 4 種類の顔向き遷移バイグラムの回数(但し，user→elsewhere，elsewhere→user は途中に agent を必ず向くことになるので除外)の計 7 種の特徴量を算出し，もう 1 人のユーザの顔向き特徴量 7 種類を加えた計 14 種類の特徴量を顔向きに関するものとして設定した．

5.2. 機械学習の結果

音声情報に関する特徴量を 6 種類，顔向きに関する特徴量を 14 種類，計 20 種類の特徴量を設定し，さらに，性別を特徴量に加え，SVM(support vector machine)による機械学習を実施した．モデルには，音声特徴量の 6 種類を用いた音声モデル，顔向きの 14 種類の特徴量を用いた顔向きモデル(システム実装では不使用)，音声特徴量と顔向き特徴量のすべてを用いた統合モデルの 3 種類を構築した．また，ピッチ，パワー，話速の韻律情報の特徴量の有用性についてさらに詳細に調べるため，5.1.1 節の(1),(4),(5)の特徴量を用いたピッチモデル，(2),(4),(6)を用いたパワーモデル，(3),(4) を用いた話速モデルを構築した．発話継続長については，4.2 の t 値が最も大きかったことから，最も有効な特徴量であると考え，全てのモデルに含めた．さらに，システムを構築するにあたり，ピッチと話速は，ユーザの発話の音量が小さいなど，取得できない場合があるので，確実に計測することができるパワーモデルと顔向き情報を統合したパワー+顔向きモデルも構築した．

音声モデルの構築には，全 1,830 発話を用い，顔向きモデル及び，統合モデルには，FaceAPI からの三次元の頭部の位置と回転角度が取得できている 1,237 発話を用いた．それぞれのモデルにおいて，受話者がエージェントの場合ともう一人のユーザ(パートナー)の場合の 10 回の交差検定による評価結果を表 3 に示す．ピッチモデル，パワーモデル，話速モデルを比較すると，これら 3 つのモデル間の精度にはほとんど差が見られなかった．この 3 つを統合した音声モデルにおいては，若干の精度向上が見られたため，音声特徴量が全て揃っている場合には，音声モデルを採用した．

6. 受話者推定システムの構築

前節において構築したモデルを組み込み，リアルタイムの受話者推定システムを構築した．

6.1. システムアーキテクチャ

本研究において，提案するシステムアーキテクチャを図 4 に示す．

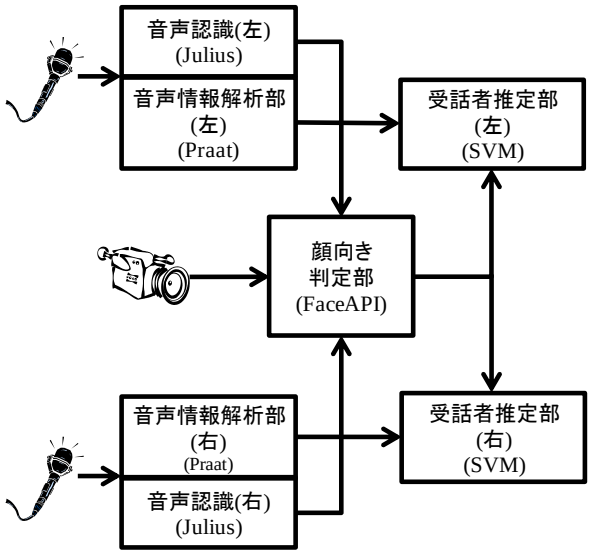

図 4：提案するシステム構成図

(1) 音声認識・音声情報抽出部：

現在の実装システムでは，音声認識は行っておらず，発話開始情報と終了情報のみを顔向き判定部に送っている． また，音声情報抽出部では，マイクから入力された音声に対して, 4.2 節で行った手法に基づき，ピッチとパワーの平均，話速，発話継続長の算出を行う．算出された各特徴量は，受話者推定部に送られる．

(2) 顔向き判定部：

顔向き判定部では，FaceAPI において，カメラから撮影されたユーザ 2 人の頭部の位置と回転角度を

取得し, 4.3 節で構築した顔向き判定の決定木を用いることにより，リアルタイムでユーザの顔向きを判別する．そして，発話をしている場合のみ，5.1.2 節で提案した各特徴量を算出し, 受話者推定部に送る．

(3) 受話者推定部：

受話者推定部では，音声情報抽出部と顔向き判定部から送られる各特徴量を受話者推定モデルに適用し，受話者の推定を行う．実装システムでは，表 3 に示される構築したモデルのうち以下の 4 種類のモデルを状況によって使い分けた．

◆ 統 合 モ デ ル : 全ての特徴量の取得成功時
◆ パワー+顔向き: F_0，話速の取得失敗時
◆ 音 声 モ デ ル : 顔向きの取得失敗時
◆ パ ワ ー モ デ ル : 顔向き，F_0，話速の取得失敗時

このように特徴量の取得状況に応じてモデルを適宜切り替えられることで，より頑健性の高いシステムを構築した．

6.2. 受話者推定システムの評価

6.1 節で構築したシステムを用い, 受話者推定システムの評価実験を行った．エージェントが受話者と判定された場合のエージェントの応答については，前回同様 WOZ により，実験者が発話を選択・実行した．

6.2.1. 実験概要

評価実験における被験者に与えられた課題，教示は，3 節と同じである．被験者は全員，大学生，大学院生であり，6 組計 12 人分のデータを収集した．そのうち，男性が 4 組，女性が 2 組であり，平均年齢は，22.7 歳であった．

6.2.2. 評価結果

評価実験において，受話者推定システムから，全 602 発話が出力されたが，そのうち 108 発話(受話者がエージェントであると推定された 3 発話，受話者がユーザであると推定された 105 発話)は不正な音声入力であった．最も多かったのは，2 人の被験者の立ち位置が近いことにより，もう一人のユーザの音声が入ってきてしまう問題であった．他にも，鼻のすすりや咳払いなどがあった．

分析方法は，4.1 節と同様，自動で区切られた発話区間に対し, 人手で受話者のアノテーションを行い，システムが出力した結果との一致率を計算した．上記のエラーを除外しない場合には，分類精度は 68%と下がったが，想定外の音声入力によるエラーを除外した場合は，83%と良好な結果となった．

7. おわりに

本稿では，音声情報と顔向き情報に基づく受話者推定方式を提案した．また，より頑健性の高いシステムを構築するため, 4 種類のモデル(パワーモデル，音声モデル，パワー+顔向きモデル，統合モデル)を SVM により構築した. 全ての特徴量がそろった統合モデルでの分類精度は，約 80%であり，先行研究[5]の 72%を上回る結果となった．最後に，システムを実装し評価実験を実施した．不正な音声入力を除外すると，分類精度が 80%以上であることから，音声入力エラーを防ぐ手法を提案できれば，現実場面においても推定が可能であることが示唆された．エラーを防ぐ方法として，骨伝導マイクの使用や，音量の違いによる発話者の決定などがあげられる. また，キーワードスポッティングを用いることにより，受話者推定において言語情報を考慮することが可能になり，さらに推定精度を向上させることができると考える．今後は，実装した受話者推定機構を対話システムに組み込むことにより，グループユーザ対応型会話エージェント全体を実装する予定である．

謝辞

本研究の一部は科研費基盤(S)(課題番号:19100001)と科研費若手(B)(課題番号：23700183)の助成による．

参考文献

[1] Vertegaal, R., et al. Eye gaze patterns in conversations: there is more the conversational agents than meets the eyes. in *CHI 2001*, 2001.

[2] Argyle, M. and M. Cook, *Gaze and Mutual Gaze*, Cambridge: Cambridge University Press, 1976.

[3] Takemae, Y., K. Otsuka, and N. Mukawa. Video cut editing rule based on participants' gaze in multiparty conversation. in *the 11th ACM International Conference on Multimedia*. 2003.

[4] Frampton, M., et al. Who is “You”? Combining Linguistic and Gaze Features to ResolveSecond-Person References in Dialogue. in *the 12th Conference of the European Chapter of the ACL*,2009

[5] Katzenmaier, M., R. Stiefelhagen, and T. Schultz. Identifying the Addressee in HumanHumanRobot Interactions based on Head Pose and Speech. in *international Conference on Multimodal interfaces (ICMI04)*, 2004.

[6] Rodriguez, H., et al. Audio Analysis of Human/Virtual-Human Interaction. in *IVA 2008*, 2008.