

非言語パターンの自動抽出による 状況説明会話における言い淀みシーンの分析

岡田 将 吾^{†1} 坊 農 真 弓^{†2,†5} 角 康 之^{†3}
高 梨 克 也^{†5,†4} 新 田 克 己^{†1}

本研究では多人数会話において表出される非言語パターンの有無を自動でアノテーションする枠組みを述べ、この手法によりアノテーションされた各非言語パターンの時系列遷移を分析する。本研究では状況説明タスクを実験協力者に行ってもらい、各種モーションセンサを用いて腕・頭部のジェスチャに関する時系列データを取得し、説明者のジェスチャ、聞き手の頷きの有無を自動抽出した。生起する各非言語パターンの遷移を分析した結果、説明者、聞き手の全ての非言語チャンネルでパターンが生起しない特異な区間が存在する事を確認し、この区間に説明者の言い淀みと、もう一方の説明者への説明補助を要求するシーンが多く含まれることを示した。さらに、聞き手の頷きの有無の利用が言い淀みシーンの検出に有効である事を示した。

Analysis of Falts in Conversational Interaction Using Algorithms to Detect Nonverbal Activity

SHOGO OKADA,^{†1} MAYUMI BONO,^{†2} YASUYUKI SUMI,^{†3}
KATSUYA TAKANASHI^{†2} and KATSUMI NITTA^{†1}

This paper proposed an algorithm to classify to non motion patterns (such case that hand is set on home position) and motion patterns including gestures automatically by using Hidden Markov Model. To understand the structure of conversational interaction, we analyze the transition of gesture patterns classified by the algorithm and nod patterns used in an explanation task. In the task, two participants explain the contents of animation to one participant cooperatively. The result of the analysis shows that a listener tend to stop nodding when explainers hesitate in speech. Thus, it is useful that we use the features of timing of nod for detection of hesitation patterns in speech.

1. はじめに

我々は空間を伴う状況を他者に説明する場合に、言語による情報だけでなく、直感的に状況を伝える事の出来るジェスチャを効果的に用いている。また説明を聞く側は、頷く事で説明者に理解していることを伝え

る。こうした会話における非言語情報の役割は重要であることが認識されているにも関わらず、言語のように明確な文法規則が存在しないため、非言語情報の役割や構造を定義することは難しい。この問題解決を目指し、情報・デバイス技術を利用して人間同士のインタラクションを機械可読なデータとして収録し、蓄積された会話データを分析し、会話構造の理解を目指す研究が盛んに行われている。¹⁾⁸⁾²⁾⁶⁾

本研究では会話中に表出されるジェスチャの有無や頷きの自動検出アルゴリズムを用いて、会話参加者の非言語パターンの遷移を分析し、機械的に認識可能な会話におけるパターンを抽出することを目的とする。従来、対面会話における視線や頭部動作に関する自動検出・認識アルゴリズム⁴⁾や会話における非言語行為のルール発見手法⁷⁾は提案されているが、会話中に用いられるハンドジェスチャの抽出・分析に焦点を当てた研究は少ない。

^{†1} 東京工業大学大学院 総合理工学研究科
Dept.of Interdisciplinary Graduate School of Science
and Engineering,Tokyo Institute of Technology
^{†2} 国立情報学研究所
National Institute of Informatics
^{†3} はこだて未来大学 システム情報科学部
Faculty of Systems Information Science, Future University Hakodate
^{†4} 京都大学学術情報メディアセンター
Academic Computing and Communications Center,
Kyoto University
^{†5} 科学技術振興機構さきがけ
PRESTO, Japan Science and Technology Agency (JST)

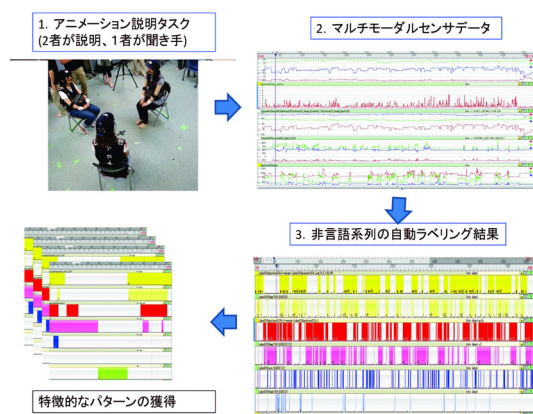


図 1 本研究のアプローチ
Fig.1 Proposed approach

まず分析対象として、空間を伴う状況を他者に説明するアニメーション説明課題タスクを行い、得られた各種センサデータから非言語パターンの検出を行う。本タスクでは、二人の説明者が事前に見たアニメーションの内容を聞き手に協調的に説明することが要求される。二人の説明者間のジェスチャを伴う説明方法や、聞き手の頷きを含む会話参加者全員の非言語パターンの共起関係などが分析対象となる。

本タスクにおいて説明者が説明に用いるジェスチャと聞き手の頷き、また説明者・聞き手の発話区間を自動抽出する手法を導入し、説明者・聞き手の非言語パターンの遷移から、説明者の言い淀みパターン、聞き手が説明者の話しに割り込むような特異なパターンが分析の結果、抽出できることを報告する。本研究のアプローチを図1に示す。

本研究の貢献は、ジェスチャパターンの有無を含む非言語情報の自動抽出に基づく分析方法を提案することと、説明者・聞き手双方の非言語パターンの遷移を利用する事で、言い淀みシーンを高精度に抽出出来ることを示す事である。

2. 非言語データの取得環境

本研究では会話者の言語・非言語の両面から、多人数会話におけるインタラクションを分析することを最終目的としている。会話者の発話から言語情報と韻律情報を取得するために、指向性無線マイクと録音機材を、会話者の非言語情報として視線、ハンドジェスチャ、頭部ジェスチャをセンシングするために、モーションアナリシス社製の光学式モーションキャプチャシステム：Mac3D、ナックイメージテクノロジー社製のアイマークレコーダ：EMR9、ATR-promotions 社製の加速度センサ：WAA-010、をそれぞれ用いる。各種センサを実験協力者が装着している様子を図2に示す。

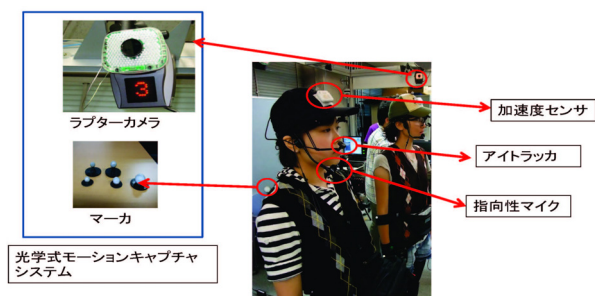


図 2 データ収録で用いた各種センサ（ハンドジェスチャのデータマイニングには腕に設置したマーカの三次元座標値を用いる。）
Fig.2 Sensors used in this research

これら複数のセンシング機器を有するマルチモーダルデータ収録環境（IMADE ルーム⁶⁾）でデータの取得を行った。

3. 非言語パターンの自動抽出

本節では、説明者、聞き手の表出する非言語パターンを抽出し、それら複数種類の非言語パターンの共起性に関して分析を行う。本論文では自動抽出可能なパターンの組み合わせから、定義可能な構造を探るため、言語内容に踏み込まず、抽出可能な非言語パターンの生成頻度に関して分析を行う。以下に分析対象として抽出した説明者・聞き手の非言語チャンネルを列挙する。

- (1) **説明者のジェスチャ区間**：手の動作が検出された区間を説明者のジェスチャ区間とする。
- (2) **説明者と聞き手の発話区間**：零点交差法によって発話区間の検出を行う。振幅と交差回数のパラメータに関しては予備検討の後決定した。
- (3) **聞き手の頷き動作**：文献¹¹⁾に従い、頷き動作の抽出を行った。具体的には、頭部に設置した加速度センサによって取得したY軸方向とZ軸方向の加速度の絶対値に対して、600msの区間で分散値を計算し、その区間の平均と偏差から計算した閾値に基づいて頷きの候補となるパターンを抽出した。

ジェスチャ区間の抽出にはモーションキャプチャMac3Dにより取得される腕のマーカの座標値 (x, y, z) の3次元データを用いる。両腕にマーカを装着し、両腕のジェスチャを観測するため、合計6次元の時系列データを取得する。なおサンプルデータは120フレーム毎秒で取得される。頷きは頭部に装着した加速度センサを用いた。加速度センサについてサンプルデータは60

フレーム毎秒で取得される。音声についてはサンプリング周波数 16kHz でデータの収録を行った。紙面の都合上、以降では説明者のジェスチャ区間抽出に関するアルゴリズムの説明と評価実験のみを記載する。

3.1 ジェスチャ区間の抽出

3.1.1 ノイズ除去のための前処理

手の動作に関係ないノイズデータの除去を行った。ノイズデータとは、手の動作として意味のある時間幅に比べて非常に小さいピークを指す。時系列を走査して、 x フレームの間にピークが存在する場合、このピークに対し、ピークの開始点と終了点の間を線形近似する。この処理をピークが無くなるまで繰り返す。この前処理を、手の動作に関わる系列データの各次元に対して適用する。実験では $x = 100$ と設定し、すなわち 0.83 秒以内のピークの除去を行った。

3.1.2 探索区間の絞り込み

一般的に会話中では話者は意識的であれ、無意識的であれ手を動かす場面と、手がホールド状態やホームポジションにある、手が動かない場面が存在する。まず大量の時系列データから手の動きを検出出来れば、ジェスチャを行っている可能性のある領域を特定する事が可能となる。またこの結果、分析者は手の動きがあるところだけを抽出して観察することが出来るため、分析の負荷を軽減できるというメリットもある。

手の動いていない領域では時系列データの変化量が極端に少ない。この変化のない時系列データをパターン発見・探索する区間から除去する。本研究では Hidden Markov Model (HMM)⁵⁾ を用いて、無変化の時系列データの除去を行う。HMM への入力には腕のマーカの三次元位置座標の差分を用いた。まずセッションの開始後 3 秒までのユーザが動作を行っていない状態をホームポジションと定義し、3 秒間の系列データを HMM の学習データとした。ホームポジションにおける時間的変位が少ない時系列データが HMM に学習され、学習後にはホームポジションにある場合の系列データが入力されると、HMM からは高い確率値が出力され、ジェスチャを含む手の動作が始まると、HMM からは低い確率値が出力される。この性質を利用し、時系列データから学習した HMM の出力確率が大きいフレームを変化のない部分（ホームポジション）と認定した。

3.2 ジェスチャ区間の自動抽出精度の評価

ジェスチャ区間の抽出精度を評価した。評価のために、ジェスチャ区間を著者以外のラベラーによりラベリングしてもらい、これを正解データとした。

4 章で後述する会話データにおける説明者の左手の

表 1 再現率・適合率

Table 1 Recall and precision

	再現率: R	適合率: P
y 座標の差分による手法	0.83	0.72
HMM による手法	1.00	0.94

ジェスチャから得られる時系列データをベンチマークとして、正解データとシステムによるジェスチャの検出区間との一致度を算出した。3.1.2 節で紹介した Hidden Markov Model を用いる手法、そして y 座標のフレーム間の座標値の変化量を閾値に従い分節化する手法とを比較した。後者の手法では、ホームポジションから手が離れた場合に動作が始まるという前提の下、腕のマーカの y 座標が大きく変化したらジェスチャ区間と判定する。

精度評価の指標には再現率・適合率を用いる。正解データの抽出区間とのフレームの差が ± 10 以内であれば、本手法が判定した区間は正解であると判定する。再現率 R 、適合率 P は以下の式で算出される。

$$R = \frac{\text{正解と判定された区間の個数}}{\text{正解データの区間の個数}}, (0 \leq R \leq 1) \quad (1)$$

$$P = \frac{\text{正解と判定された区間の個数}}{\text{抽出された区間の個数}}, (0 \leq P \leq 1) \quad (2)$$

y 座標の差分による手法と HMM の最良なパラメータについては、探索的に決定した。表 1 にジェスチャ区間の抽出精度の比較を示す。HMM の精度は y 座標の差分による手法を上回っていた。この理由として、y 座標の差分による手法は無動作区間に含まれるデータの観測ノイズに弱いことが挙げられる。この結果を踏まえて、本分析では HMM による手法をジェスチャ区間の抽出に採用した。

図 3 は、説明者の左手のセンサから得られた 3 次元の位置座標を 80 秒間取得した時系列データから、HMM により無動作区間の抽出を行った結果を示している。無動作区間と判定されたフレームでは 1000、動作と判定されたフレームでは 0 の値をとる。左手が動くときセンサは反応するが、左手が無動作の状態ではマーカの観測ノイズを除いて変化はない。図において点線は動作判定の結果を示している。マーカの観測ノイズを含む無動作区間が正しく抽出出来ている事がわかる。また無動作区間でない部分としてジェスチャ区間を抽出出来ている事が併せてわかる。

4. 分析対象

本章では分析対象として選定した多人数会話データについて説明する。

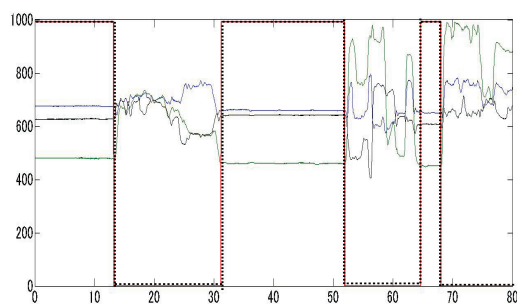


図 3 無動作区間およびジェスチャ区間の抽出例（図において点線は動作判定の結果を示している．無動作と判定されれば 1000，動作と判定されれば 0 を出力する．）

Fig.3 Example of detection result of non motion interval and motion interval

4.1 分析対象のデータ

本研究ではジェスチャ分析研究で頻繁に用いられてきた，“Canary Row”というアニメーションの内容・情景を説明するタスク³⁾を分析対象とした．このタスクでは，アニメーションの情景を説明するために，ハンドジェスチャを伴う発言が多く含まれる傾向にある．以後，このタスクを説明タスクと呼称する．

McNeill は 1 人の説明者が 1 人の聞き手に説明を行う説明タスクを分析した³⁾が，本研究は将来的に多人数会話分析を目標としているため，複数人が協調して説明を行う場面を分析対象とした．具体的には，2 人の説明者が 1 人の聞き手に説明を行う場面の分析を行う．

図 4 に示すように，3 人は着座状態で会話を行う．人材派遣会社を通じて計 3 名の実験協力者を募集した．募集した 3 名はいずれも 20～22 歳の女性とした．3 人 1 グループで説明タスクを行い，センサデータを収集した．

4.2 非言語パターンの自動抽出に基づく分析の目的

本タスクでは説明者と聞き手が明示的に分かれており，聞き手の非言語行為としてハンドジェスチャはほとんど観測できない．その代り，聞き手はレスポンスとして頷くしぐさを頻繁に行っている事を観測した．この結果より，聞き手の頷きを含む頭部動作を抽出し，説明者のハンドジェスチャと頭部の動作パターンの時系列に沿った生起頻度を分析する．分析の目的は，説明者・聞き手の非言語パターンの生起頻度と時系列遷移から，会話における特徴的なパターンを見つけることである．

5. 説明者・聞き手の非言語パターン遷移の分析結果

3 章の方法に従い，説明者の発話区間・ジェスチャ

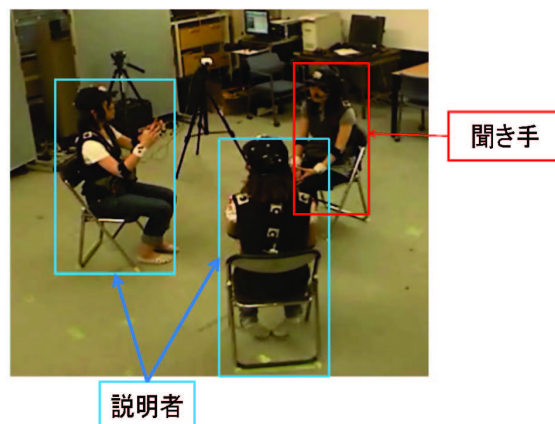


図 4 三者のアニメーション説明タスク

Fig.4 Explain task by three participants

区間と聞き手の頷きパターンを抽出した．それぞれのパターンをラベル化し，時系列上に並べた様子を図 5 に示す．1，2 行目に説明者 1 のジェスチャ有無のラベルと発話区間ラベルを，3，4 行目に説明者 2 のジェスチャ有無のラベルと発話区間ラベルをそれぞれ記載する．また 5，6 行目に聞き手の頷きパターンの有無と発話区間ラベルをそれぞれ記載する．この図より聞き手は，不定期に頷きを行っている事がわかる．また 2 人の説明者の発話区間とジェスチャ区間は交互に生起している．

各時刻でいずれかの時系列データのパターンが生起していることがわかる．一方で，1 から 6 行目のいずれの時系列データでもパターンが生起しない区間が存在する．この区間を無ラベル区間と定義し，7 行目に記載する．説明者と聞き手の役割が分かれている本タスクにおいて，3 人の非言語パターンが生起しない無ラベル区間は，特異であると言える．本分析ではこの特異な無ラベル区間について分析を行う．

会話データから，無ラベル区間は 11 回抽出された．11 の区間で起きたイベントをビデオ分析により確認し，その会話場面を 3 タイプに分類した．11 の無ラベル区間が検出された時間と，その区間の会話場面のタイプを表 2 に記載する．無ラベル区間のタイプについて，タイプ 1 は説明者が言い淀み，もう一方の相手に説明の助けを求める場面であり，タイプ 2 は聞き手が説明者の説明に割って入る場面であり，タイプ 3 は短い区間で発話に間が空いた場面であった．

タイプ 1 の代表的な無ラベル区間の周辺におけるパターンの様子を図 6 に示す．説明者にとってアニメーションの暗記が完全でない場面において，説明者は言

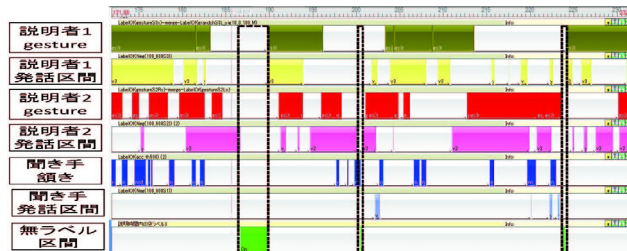


図 5 説明者と聞き手が表出する非言語パターンの時系列遷移図 (171.88 秒から 232.39 秒の間の会話において表出されたパターンの時系列遷移を示している。)

Fig.5 Transition of nonverbal patterns used by explainers and a listener

表 2 無ラベル区間の生じた時間と会話場面のタイプ (タイプ 1: 説明者が言い淀み, もう一方の相手に説明の助けを求める。タイプ 2: 聞き手が説明者の説明に割って入る。タイプ 3: 発話に間が空く)

Table 2 Time when non label interval exists and types of conversation scene

No.	開始時間 [s]	終了時間 [s]	区間の長さ [s]	タイプ
1	34.52	34.75	0.23	1
2	48.06	48.69	0.63	2
3	117.08	119.33	2.25	1
4	154.66	157.04	2.38	1
5	186.63	189.69	3.06	1
6	200.48	200.65	0.17	3
7	223.66	223.95	0.29	1
8	273.20	274.07	0.87	1
9	332.33	333.26	0.93	1
10	362.96	363.90	0.94	1
11	382.80	383.00	0.20	3

い淀み, もう一方の説明者に助けを求める傾向にある。この助けに応じて説明者が交代し, もう一方の説明者が発話権を得るという場面が 11 回の内, 8 回観測された。図 6 では, 説明者 2 が説明中に言い淀み, 無ラベル区間を経て, 説明者 1 に発話権が移っている事がわかる。

タイプ 2 は説明中に聞き手が割り込みをかけて再説明を要求するシーンであった。タイプ 2 の無ラベル区間の周辺におけるパターンの様子を図 7 に示す。この場面は 11 回の内, 1 回観測された。図 7 では, 聞き手が説明者 2 の説明に割り込んだ後, 説明者が説明を止める (無ラベル区間)。その後に聞き手が質問をしている。タイプ 3 は説明中の非常に短い休止区間であり, 特質すべき特徴を有していなかった。

5.1 分析結果の考察

説明者, 聞き手が表出する非言語情報に関する, 全てのチャンネルでパターンが現れない無ラベル区間の存在を確認した。無ラベル区間では言い淀みと, 説明の助けを求める行為, また聞き手が割り込みをかけて再説明を要求する行為が行われていることを確認した。言い淀みと説明の助けを求める行為は 11 回の無ラベル区間の内 8 回であった。シーン全体でこの行為の頻



図 6 言い淀みシーン

Fig.6 Scene where explainers say uncomfortably

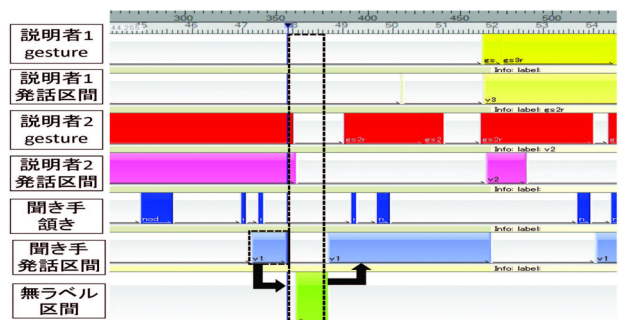


図 7 聞き手が割り込みをかけて再説明を要求するシーン

Fig.7 Scene where listener requires re-explanation

度を調べたところ 8 回であったため, 無ラベル区間に全て含まれていることがわかった。

次に説明者・聞き手のいずれの非言語チャンネルが言い淀みシーンの検出に寄与したかを考察する。ここでは, 3 人の発話区間に含まれていない区間を無ラベル区間として抽出した場合と, 3 人の発話区間・ジェスチャ区間に含まれていない区間を無ラベル区間として抽出した場合と, さらに聞き手の頷き区間を含めた上で無ラベル区間を抽出した場合とで, そららの無ラベル区間にどの程度言い淀みシーンが含まれているかを検証する。比較するための評価基準には再現率・適合率を用いた。抽出した無ラベル区間と, 言い淀みシーンにオーバーラップがあれば正解とする。比較結果を表 3 に示す。表 3 より, 聞き手の頷き, 説明者のジェスチャを加える事により適合率が向上する事を示した。

従来, 言い淀みの検出には, 発話に含まれる音韻・

表 3 言い淀み抽出精度に及ぼした各非言語チャンネルの影響 (再現率・適合率の計算を行う。)

Table 3 Influence of non-verbal channels on accuracy of falters

手法	再現率: R	適合率: P
(1): 3 者の音声区間	1.00 (8/8)	0.05 (8/146)
(2): (1) + 説明者のジェスチャ区間	1.00 (8/8)	0.15 (8/52)
(3): (2) + 聞き手の頷き	1.00 (8/8)	0.73 (8/11)

韻律情報を利用する研究⁹⁾が多かった。これに対し本研究の結果から、聞き手の非言語チャンネルの情報も言い淀みシーンの検出に有効であることを示唆した。

井上¹⁰⁾ は本研究で用いたデータと同じデータを用いて、説明者同士のジェスチャの引き取り方法について、詳細な分析を行っている。文献¹⁰⁾ の 4.4 節では、要求された引き取りとして、表 2 の 4 番のシーンに関する分析を行っている。本研究の結果はこのシーンを自動的に抽出出来ることを示唆した。

6. 結 論

アニメーション説明課題タスクにおける話者の用いるハンドジェスチャ区間抽出を行い、ジェスチャパターンの有無ラベルの自動アノテーションを行った。また聞き手の頷き検出と、発話区間検出を行い、これらに関してのアノテーションも行った。

この結果を利用して説明者、聞き手が表出する非言語パターンの同期的生起パターンを分析し、どのモダリティでもパターンが現れない無ラベル区間を抽出した。この無ラベル区間は、言い淀みによる発話権の交代場面で見られる事を確認した。また聞き手の頷きの有無パターン、説明者の発話有無パターン、ジェスチャ有無パターンの全てを統合的に用いる事で高精度に、言い淀みによる発話権の交代場面を特定できる事を示した。今後は、多人数会話における言い淀みシーン検出アルゴリズムの開発と、ハンドジェスチャの種類や役割を考慮した分析を行う予定である。

謝 辞

本研究は、国立情報学研究所公募共同研究「多人数インタラクション理解とその応用に関する研究」、科研費 若手研究 (B)「インタラクションマイニングによる対話コンテキスト依存型ジェスチャ認識モデルの構築」の補助を受けています。

参 考 文 献

1) Carletta, Jean, Ashby, Simone, Bourban, Sebastien, Flynn, Mike, Guillemot, Mael, Hain, Thomas, Kadlec, Jaroslav, Karaiskos, Vasilis, Kraaij, Wessel, Kronenthal, Melissa,

Lathoud,Guillaume, Lincoln, Mike, Lisowska, Agnes, Mc-Cowan, Iain, Post, Wilfried, Reidsma, Dennis, Wellner, Pierre . The AMI meeting corpus: A pre-announcement. In Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005), Vol.3869 of Lecture Notes in Computer Science, pp.28-39, Springer,(2005)

- 2) Chen, Lei, Rose, Travis R., Qiao, Ying, Kimbara, Irene, Parrill, Fey, Welji, Haleema, Han, Tony, Tu, Jilin, Huang, Zhongqiang, Harper, Mary, Quek, Francis, Xiong, Yingen, McNeill, David, Tuttle, Ronald, Huang, Thomas , VACE multimodal meeting corpus. In Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005), Vol.3869 of Lecture Notes in Computer Science, pp.40-51, Springer,(2006)
- 3) David McNeill, Hand and Mind, The University of Chicago Press,(1992)
- 4) Otsuka,Kazuhiro, Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-face Meetings. 14th International Conference on Human-Computer Interaction (HCI International 2011), 171-179, Springer,(2011)
- 5) Lawrence. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, ACM, New York, NY, USA, pp.257-286,(1989)
- 6) 角 康之, 矢野 正治, 西田 豊明 : マルチモーダルデータに基づいた多人数会話の構造理解 社会言語科学会論文誌 (採録決定) (2011),
- 7) 中田篤志,・角康之・西田豊明 (2011). 非言語行動の出現パターンによる会話構造抽出. 電子情報通信学会論文誌, Vol.J94-D, No.1, pp.113-123
- 8) Waibel, Alexander, Stiefelhagen, Rainer (Ed.) , Computers in the Human Interaction Loop. Springer (2009)
- 9) 後藤真考, 北山広治, 伊藤克亘, 小林哲則, 音声スポット : 人間同士の会話中に音声認識が利用可能な音声入力インターフェイス, 情報処理学会論文誌, Vol.48,No.3, pp.1274-1283, (2007)
- 10) 井上卓, 角康之, 高梨克也, 状況説明会話における説明者間の発話とジェスチャの引き取り, インタラクション 2011, (2011)
- 11) 齊賀弘泰, 角 康之, 西田豊明, 多人数会話におけるうなずきの会話制御としての機能分析, 情処学研報 (ユビキタスコンピューティングシステム), vol.2009-UBI-23, 2009