複数人会話における振り向き動作と発話動作解析

Analysis of Head Turning Motion and Utterance Timings in Conversation

小山　大幾[1]，水本　武志[2]，中村　圭佑[2]，中臺　一博[2]，今井　倫太[1]
Daiki KOYAMA[1]，Takeshi MIZUMOTO[2]，Keisuke NAKAMURA[2]，
Kazuhiro NAKADAI[2]，Michita IMAI[1]
[1] 慶應義塾大学
[1]Keio University
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2]Honda Research Institute Japan Co., Ltd.

Abstract: This paper discusses a relationship between head turning motions and utterances in conversation. We constructed an audio-visual conversation recording system. The recorded data was manually annotated in terms of head turning motion and utterance timings. We also investigated automatic annotation of these timings by extracting a posture, face orientation, speech direction, and speech activity

１．緒言

ロボットが会話に参加する場合，ロボットの反応タイミング遅れなどの影響で不自然なインタラクションとなり，違和感を覚えてしまうという問題がある．

人同士の会話では，タイミングの良い振舞いは，よい会話のリズムを生むものとされている[1]．したがって，ロボットの反応タイミングを適切に生成することができれば，より円滑な人とのインタラクションを実現できると考えられる．

我々は，その手始めとして，複数名による実際の人同士の会話における人の振舞いに着目した．これまで，複数人の会話における振舞いの研究はなされているが，誰が発話したか，あるいはグループの中で誰と誰が話しているかなどの状態に対して視線変化などの振舞いがどの程度の割合で起こるといった研究が多い．しかし我々は，ロボットの振る舞いのタイミングを検討するため，会話における一つ一つの振舞いについてさらに時間的に詳細な分析を行った．具体的には，まず，会話収録システムの作成を行った．次に，作成したシステムを用いて収録した会話データをもとに，発話の開始時刻と振り向きの時刻について手動でアノテーションを行い，これらの相関を分析した．

また，アノテーションを自動で行うアルゴリズムの検討を行った．このためのキューとして，深度センサ用いて得られる顔・骨格情報，マイクロホンアレイを用いて収録した多チャンネル音響信号に対してロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [2]を適用することで得られる音源方向，およびその音源方向に対応する分離音を用いた．得られた情報に対して，振り向きの時刻，および発話の開始時刻を推定し，手動で行ったアノテーション結果との比較を行った．

２．関連研究

本研究では，エージェントの複数人会話における聞き手としての振る舞いを検討するために会話分析を行う．本章では，複数人会話のエージェントで検討する意義，そして会話分析では何に着目するべきか，それぞれ関連研究を踏まえ，2.1 節 2.2 節に示した．

2.1 ロボットの振る舞いのタイミング

ロボットの振る舞いのタイミングを考慮した関連研究として，身体的リズムに着目し，発話音声からうなずきを自動生成するモデルをエージェントに導入する研究[3]，社会学における人の会話研究から得たモデルを導入することで，あたかもロボットが人とインタラクションしているかのように見せる研究[4]など，聞き手としての振る舞いの検討が行われている．しかし，これらの研究は，1 対 1 の対面コミュニケーションを対象とした研究であり，複数人のコミュニケーションを対象とした研究では，聞き手としての振る舞いの検討はほとんど行われていない．例えば，発話後の空白時間によってエージェントが発話権を獲得するタイミングを検討する研究 [5] など，会話の主導権の移り替わりがより複雑な複数人会話では，如何にユーザの会話に適切なタイミングで介入し，発話するかということが重要視されるため，多くは話し手としての振る舞いの検討や，音声対話の研究を扱っている．

本研究では，複数人の会話に参加するエージェント

の聞き手としての振る舞いを検討する．その手始めとして，聞き手として重要である振り向きと，話し手の発話に着目して，実際の人の会話を分析した．

2.2 会話研究

複数人の会話における，頭部方向と発話の関係について，マルコフモデルを使用して，発話の状態から，参加者の頭部方向を推定する研究[6]などがあるが，ロボットの振舞いのタイミングを検討するためには，振舞いの時間的な相関が重要となる．そこで本研究では振り向き時刻と話し手の発話開始時刻に着目し，両者の相関を分析した．さらに，会話を効率よく分析するために，会話の中で自動的に振り向き時刻と発話時刻を検出するアルゴリズムの検討を行った．

３． システムによる会話収録

3.1 システム概要

非言語情報を伴う会話の分析を行うことを目的として，会話における人の振る舞いを収録するシステムを製作した．Fig.1 にシステム構成図を示す．本システムでは，顔・骨格情報を伴う画像データと，音源定位・分離用に多チャンネル音声データを収録することが可能である．なお，収録対象は，参与者が所定の位置に座っている会話である．

3.2 画像データ

画像データを得るために，RGB・深度センサとして Microsoft 社の Kinect™ を用いた．得られた RGB・深度情報から，OpenCV によりカラー画像を生成し，さらに人の顔・骨格情報を画像に描画する．画像は 1280×1024 の画素数で毎秒約 5 フレームで PC に保存される．Fig.2 に得た画像例を示す．なお，解析の際に他の Kinect から得た画像，および音声データとの同期を図るために，保存された画像にはそれぞれタイムスタンプが押されている．また，画像を保存する際には，描画された顔・骨格情報の画像上の座標が逐一記録される．

3.3 音声データ

音声の収録用デバイスには，マイクロホンアレイ処理による音源定位・音源分離を行うことができるよう 8 チャンネルマイクロホンアレイを使用した．収録データは 16 ビット, 16 kHz でサンプリングした．収録した音声データに対して，前述の HARK の Software as a Service (SaaS) 版である HARK-SaaS を用いて，音源定位・分離を行い，分析に使用した．

４． 会話分析の予備実験

4.1 実験概略

会話における身体動作と音声の時間的な関連性を分析するために，収録システムを使用して，3 人による会話を収録し，分析を行った．

4.2 実験環境

Fig.3 に実験環境を示す．参与者をそれぞれ A, B, C として Kinect は A,B,C に対して 1 台ずつ正面に設置した．マイクロホンアレイは 3 人の中心部に設置した．PC は各 Kinect に対して 1 台，またマイクロホンアレイ用に 1 台，計 4 台を使用した．

4.3 被験者の情報

被験者 3 名は全員男性であり，年齢は 1 名が 21 歳，他 2 名は 22 歳である．なお，3 名は慶應義塾大学理工学部の 4 回生として在籍している．3 名の関係は，理工学部情報工学科における同じ研究室の同期であり，日常的に会話をする機会が多く，3 名の人間関係は良好である．

4.4 実験方法

主に「学校の食堂について」および「ラーメン屋について」という話題で行われた 3 名による 10 分間の自由

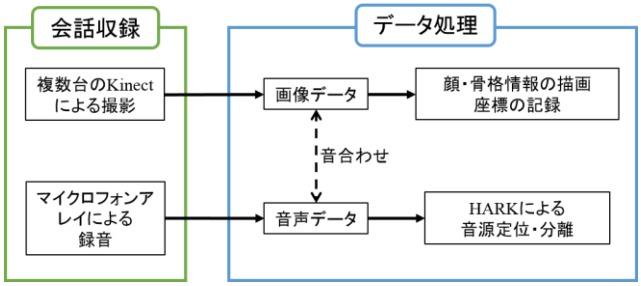

Fig.1 System flow

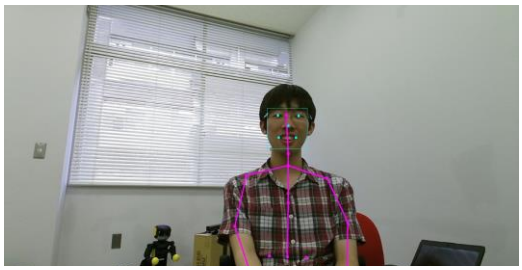

Fig.2 A captured image with Face tracking and Skelton information

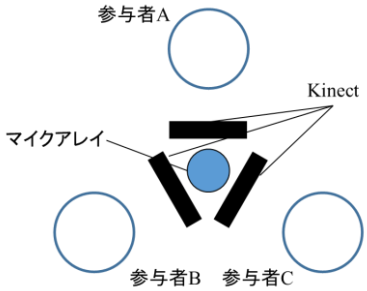

Fig.3 Experimental environment

会話を分析対象とした．なお，事前に会話の話題について，「話題は特に制限は設けず，自由に会話を行うこと，また，会話の話題は途中で変更されてもよいものとする」という教示を行った．また，各画像データと音声データの頭出しを獲得し，データ間の同期を図るために，被験者 3 名には実験の開始と終了時に順番に手を叩いてもらった．

4.5 会話分析

4.5.1 会話内容の詳細

会話はAが「学校の食堂について」という話題を提案したことにより，その話題で始まった．内容はキャンパスによる食堂の違いや，食堂への要望であった．この話題では，特に主導権を持つ者は存在しなかった．会話の中で，Bが日吉駅付近の飲食店について発言したことにより，話題は「ラーメン屋について」へと切り替わった．内容は好きなラーメン屋，および参 B のラーメン屋に行った際の体験談であった．こちらでは，Bが体験談を語ったことにより，途中でBが会話の主導権を握っている時間があった．

なお，全体を通して，AはBとCと比較すると発言する頻度は低かった．

4.5.2 手動アノテーション

はじめに，リファレンスを作成するため，取得した画像データと音声データに対して，手動でアノテーションを行った．画像データのアノテーションについては，参与者ごとに他の参与者へ顔を向けた時刻と誰に対して顔を向けたかを逐一記録した．また，音声データのアノテーションについては，参与者ごとに各発話の開始時刻を逐一記録した．ただし，「ああ」や「うん」などのあいづち，および笑い声は発話記録の対象から除外した．各参与者の他二名への振り向き回数，および発話の回数を Table 1 に示した．

Table 1 より，振り向きの頻度は A, B, C の順で高く，また発話の頻度は C，B，A の順で高いことがわかる．両者の記録された時刻について，関連性を把握するために，時間軸上の合致割合を算出した．合致の判定は，記録された各発話の開始時刻に対して，最も近い振り向きの時刻との差が 1 秒以内であれば合致するとみなした．これを各参与者の発話と，振り向いた相手によってそれぞれ算出し，発話の数で除したものを Table 2 に示した．Table 2 では，例えば B の発話に対して，A が振り向いた割合は 65%であるといった具合に，話し手の各発話に対して，聞き手が振り向きで注意を示した割合がわかる．A は，Table 1 において，B と C に対して振り向いた回数が 136 回と 139 回で同等であった．しかし，Table 2 で発話との関連性を見ると，B，C の発話の総数に対して振り向いた回数の割合は 65%，40.3%と，差異が認められる．これは，4.5.1 項で示したように会話の中で B が体験談を話し，会話の主導権を握る時間が存在したため，A は B の発話に対して振り向く割合が C より高くなったことが，一因として考えられる．

また，B の A と C に対する振り向き回数は，77 回，113 回と C への振り向きが多いが，Table 2 における割合がそれぞれ 34.1%，34.5%と同等の割合であることが

Table 1 Number of Face turn and Speech

振り向き				発話	
	A(回)	B(回)	C(回)		回数
A		136	139	A	44
B	77		113	B	80
C	36	49		C	87

Table.2 Ratios of head turning motions which is considered as triggered by utterances

振り向いた人 ＼ 発話者	A(%)	B(%)	C(%)
A		65.0	40.3
B	34.1		34.5
C	11.4	23.75	

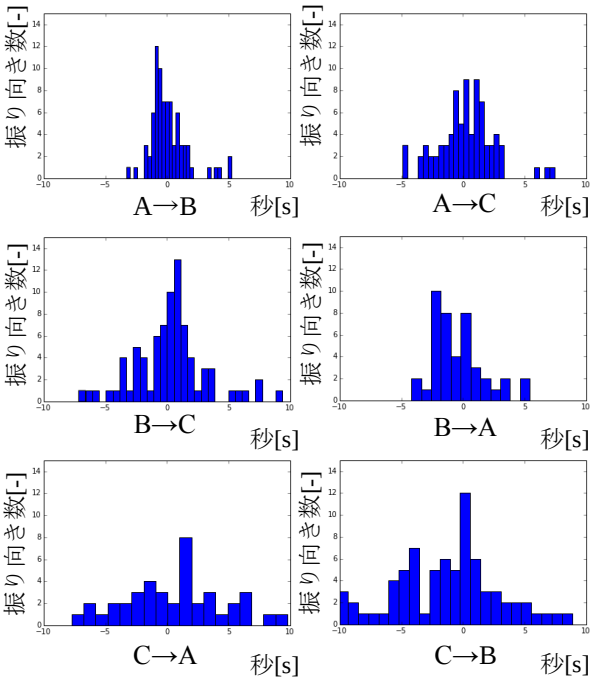

Fig.4 Histogram of the difference between head turning and utterance timings

認められる．また，いずれの場合も C の割合が本来の発話数から期待される割合よりも低いことがわかる．この結果の解釈として，Table.1 において，B はＡよりもＣへ振り向く頻度が高いが，これは，単にＣの発話によって振り向いているのではなく，発話回数の多いＣが発話することを期待して振り向いているものが，この中に含まれている可能性があると考えられる．

また，発話者の発話時刻とその発話時刻の直近で他の参与者がその発話者へ振り向いた時刻との差を全発話に対して算出し，ヒストグラムとしてまとめたものを Fig.4 に示した．横軸は時刻差（秒）で，振り向きが発話より早い場合は正、遅い場合は負の値となっている．また，A→B は A が B に振り向いた場合のグラフであることを示している．Fig.4 は Table 2 と相関があり，Table 2 における割合が高いほど，Fig.4 の対応するヒストグラムはより狭い範囲に収束している．また，ヒストグラムを見ると，発話時刻よりも以前にその人へ振り向いている場合がある．つまり，振り向きが発生する際，発話するであろうことを期待して振向くなど，必ずしも発話が行われた後に反応して振り向くのではないことがわかる．

これは，発話以外のキューによって振向きが誘発されている可能性も示唆しており，そのようなキューを見つけるため，現在，画像上に記録された他の骨格座標を使用した分析を検討している．

4.6 自動アノテーション

4.6.1 顔方向推定

画像上の鼻の座標と首の座標を利用して，いつ，どの参与者に顔を向けたのかを自動的に推定できるアルゴリズムの構築を試みた．具体的には，正面から人を見た際，Fig.5 に示すように首に対する鼻の位置が水平方向で右側にあるとき，その人は右側を向き，逆に左側にあれば，左側を向いていると判定するものとした．アノテーションによる実際の結果との比較を行い，推定の精度を求めた．まず，フレームごとに鼻と首の x 座標の差分を計算した．この際，微小な変動を取り除くために，5 フレーム毎の移動平均処理を行った．次に，差分の時間変化における極値を求め，その極値をとる時刻を，顔を他者に向けた時刻とした．向けられた人の判定は，極値の極大・極小を使用した．なお，Kinect より取得した画像は，鏡像であるため，顔を向けた相手は，極大の場合は実世界で，本人に向かって左側の人間，極小の場合は右側の人間となる．手動アノテーション結果と比較し，推定の精度を求めた．振向き時刻の推定結果の１秒以内の時刻に，対応する振向きが手動アノテーション中に存在する場合に推定成功とした．推定によって得た各参与者の振り向き回数を Table.3 に示した，また，推定成功回数を推定された振向きの総回数，および手動アノテーション中の振向きの総回数で割り，再現率および適合率を求めた．Table 4，5 に得られた再現率と適合率を示した．

Table.3 と 4 章 Table.1 における手動アノテーションによって得た振り向き数を比較すると，いずれも Table.3 における回数の方が多いことがわかるが，特にＣの回数が飛躍的に増加している．これについて，会話の映像を確認したところ，定性的ではあるが C は A,B に対して振り向き以外で頭部を動かす機会が比較的多かったことを確認した．そのため頭部の移動によって，振り向きでない箇所でも極大，極小が発生し，それが振り向きとして判定され，推定された C の振り向き回数は多くなったと考えられる．また，Table 4 と Table 5 を比較すると，いずれの場合も再現率は最低でも 70%弱と高い

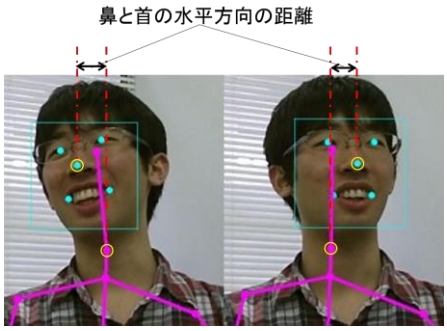

Fig.5 Horizontal distance between nose and neck

Table 3 Number of head turning estimate

振り向いた人 ＼ 振り向いた相手	A(回)	B(回)	C(回)
A		161	155
B	151		155
C	138	130	

Table 4 Recall of head turning timing estimate

振り向いた人 ＼ 振り向いた相手	A(%)	B(%)	C(%)
A		86.1	84.3
B	80.8		75.2
C	71.4	68.8	

Table 5 Precision of head turning timing estimate

振り向いた人 ＼ 振り向いた相手	A(%)	B(%)	C(%)
A		73.3	76.1
B	41.7		54.8
C	18.1	25.4	

一方で，適合率は，再現率と比較すると低いことがわかる．これより，推定によって算出された振り向きには，実際の振り向きは高い割合でカバーできているものの，上述したとおり，振り向きでない頭部の動きも振り向きとして含まれてしまっていると考えられる．

さらに推定の精度を向上させるためには，極大，極小の点に対して，何かしらフィルタリングをかけ，振り向きでない点を排除するなどして，再現率を向上させる必要がある．また，極大，極小を使用せずに，Kinect センサから，奥行き情報を取得し，3 次元的に首のねじれを検出し，直接顔の向きを特定することなどが精度向上の方法として考えられる．

4. 6. 2 HARK-SaaS による音源定位・分離

HARK-SaaS では，多チャンネル音声データに対して，音源定位・分離を行い，更に分離した各音の開始時刻をクラウドサービスとして提供するため，会話における発話開始時刻とその発話者の特定に有効である．今回は定位結果に対して，Fig.5 に示すように，70°から 110°を参与者 A の発話，-170°から-130°を参与者 B の発話，-10°から-50°を参与者 C の発話として分類した．分類によって得た発話回数を Table.6 に示した．また，分類した各音源の開始時刻について，アノテーションによって得た各参与者の発話開始時刻と比較し，その精度を算出した．具体的には，HARK-SaaS から得た，各参加者の発話の開始時刻に対して，対応する手動アノテーション結果との差が 1.6 秒以内である場合に発話検出成功とした．1.6 秒以内であるのは，HARK-SaaS では，音源定位の際に 1 フレーム 32msec，計 50 フレームのミュージックスペクトルを平均して計算するため，最大で約 1.6 秒定位結果が遅れる可能性があるからである．検出成功回数を， HARK-SaaS によって得た全発話数と，手動アノテーションで得られた総発話数でそれぞれ割り，各参与者の発話に対する再現率，および適合率を算出した．Table.7 に算出結果を示した．

Table.6 では，4 章 Table.1 における手動アノテーションによって得た発話数と比較すると，A,B,C でいずれも発話数が多くなっていることがわかる．これは，HARK-SaaS によって得られる定位結果に，余計に発話が検出される”挿入誤り”が多く含まれるためであると考えられる．これは，定位結果に，笑い声や相槌といったアノテーション対象外の音源が含まれていることが一因ある．

Table. 7 における再現率と適合率を比較すると，A,B,C で，いずれの場合も適合率より再現率が高いことがわかる．これは，上述した”挿入誤り”である，アノテーション対象外の音源が発話に含まれているため，再現率が低くなったと考えられる．より精度の高い分析を行うためには，HARK-SaaS によって分離された音声を発話か，笑い声か，相槌か，といった分類を行う必要がある．

4. 6. 3 推定結果同士のマッチング

4. 4. 1 項 Table.2 と同様に，推定によって得た振り向き時刻と発話時刻について，発話数に対する両者の合致割合を算出し，Table.8 に示した．Table.8 を，Table.2 と比較すると，特に C の A,B の発話に対する振り向きの割合が高くなってしまっている．これは，4. 4. 1 項 Table.1 および 4. 4. 2 項 Table.3 に示すように，推定で得た振り向き数が実際の振り向き数より多く，判定では

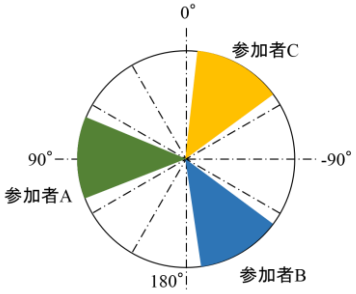

Fig.5 Sound location analysis using HARK-SaaS

Table.6 Number of sound source detection

	回数
A	56
B	105
C	139

Table.7 performance of sound source detection

発話者	Recall(%)	Precision(%)
A	70.5	55.4
B	81.5	62.9
C	89.7	33.8

Table.8 Ratios of head turning motions which is considered as triggered by utterances on estimation

振り向いた人 ＼ 振り向いた相手	A(%)	B(%)	C(%)
A		50.5	45.3
B	37.5		57.6
C	44.6	43.8	

推定された発話開始時刻に対して，振り向きでない頭部の動きが最も近い振り向きとして，選択されてしまい，割合が高くなったと考えられる．こちらの精度を上げるには，やはり 4.5.1 項および 4.5.2 項と同様に，振り向き時刻と発話開始時刻の推定における適合率を向上させることが必要であると考える．

５．結言

本稿では，非言語情報を伴う会話を分析することを目的として，視聴覚会話収録システムの開発およびシステムを使用した分析を行った．振り向き，および発話タイミングを手動でアノテーションし，分析を行ったところ，１）発話後１秒以内に振り向く割合は，振り向く頻度が高い人ほど高い，２）振り向きは発話後に発生するとは限らないことが分かった．特に２）に関しては，振り向きを誘発する別のキューが存在する可能性があり，骨格座標情報を利用したより詳細な解析を検討している．

またアノテーションデータの自動作成に向けて，Kinect から得られる顔や骨格の座標に対して OpenCV を用いて振り向きを検出するアルゴリズムの検討を行った．また，音源定位・分離情報を得るため，ロボット聴覚ソフトウェアの SaaS 版である HARK-SaaS の使用を検討した．振り向き時刻推定では再現率は高いが適合率が低かった．精度を向上させるには極大，極小による推定結果に対して振り向きでない頭部の動きを排除したり，Kinect センサから奥行きの把握や，関節のねじれの検出によって直接顔の向きを特定するなど，アルゴリズムの改良が必要である．発話時刻推定についても再現率は高いが適合率は低くなった．より精度を向上させるには，発話，あいづち，笑い声といった分類を行う機能が必要であることが分かった．

今後はシステムを改良し，より多くの会話を収録・分析を行い，振り向きと発話タイミング生成モデルを構築し，聞き手としてのロボットを検討していきたい．

参考文献

[1] Watanabe, T., and Yuuki, N.: A voice reaction system with a visualized response equivalent to nodding, Advance in Human Factors/Ergonomics,12A, Vol. 1, pp. 396-403, (1989).

[2] Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H.: Design and Implementation of Robot Audition System'HARK'— Open Source Software for Listening to Three Simultaneous Speakers, Advanced Robotics, Vol.24, pp. 739-761. (2010)

[3] Watanabe, T., Okubo, M., Nakashige, M., and Danbara, R.: Interactor: Speech-driven embodied interactive actor. International Journal of Human-Computer Interaction, Vol. 17, No.1, 43-60, (2004).

[4] Breazeal, C., and Scassellati, B.: A context-dependent attention system for a social robot, rn, Vol.55, No.3, (1999).

[5] Bohus, D., and Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech. In International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, p. 5, (2010)

[6] Otsuka, K., Takemae, Y., and Yamato, J.: A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In Proceedings of the 7th international conference on Multimodal interfaces, pp. 191-198, (2005)