1D2-5

会話エージェントにおける非言語行動の異文化適応

Enculturating Nonverbal Behaviors in Conversational Agents

山岡　雄治[*1]　　Afia Akhter Lipi[*1]
Yuji Yamaoka

[*1] 東京農工大学 工学府 情報工学専攻
Tokyo University of Agriculture and Technology

Matthias Rehm [*2]　　中野　有紀子[*3]
Yukiko Nakano

[*2]Multimedia Concepts and Applications
University of Augsburg,Germany

[*3]成蹊大学
Seikei University

Aiming at building culturally adapted conversational agents, this paper collects comparative multimodal conversation corpus, analyzes nonverbal behaviors focusing on posture shifts, and proposes an agent behavior generation mechanism. First, we collect conversation corpus in Germany and Japan using exactly the same experimental setting and materials. Then, we analyze the frequency and the distribution of posture shifts for each culture, and compares the difference between the two countries. As a result, we found that usage of head and leg postures are quite similar between the two cultures. However, as for arm postures, frequently used posture types are very different to each other. Finally, based on the empirical results, we will propose a mechanism that takes text input and generates synchronized speech and character animations as agent's nonverbal behaviors that are natural and appropriate in a given culture.

1. はじめに

言語や文化が異なる人とのコミュニケーションにおいて，相手の表情やしぐさが自分の文化と若干異なっていることに気付き，会話がスムーズでないと感じることがある．我々は，このようなコミュニケーションにおける文化的な特徴をユーザインタフェースに取り入れることを目指し，文化に適合した振る舞いができる会話エージェントの開発を進めている．その第一段階として，ドイツと日本において，同じ実験手続きを用いて比較可能な会話コーパスを収集し，基礎的な分析を行った[1]．本研究では，姿勢やジェスチャの非言語行動の違いを，さらに詳細に統計的に分析した結果を報告するとともに，分析結果に基づき，各々の文化に適応した会話エージェントの非言語行動を自動生成する機構を提案する．

2. 分析データ

ドイツ人同士 20 組，日本人同士 22 組から各組 3 対話 (合計約 25 分間) の会話を収録し[1]，そのうち，ドイツ人同士 8 組，日本人同士10組の対話について詳細な分析を行った．3 対話の内訳は，1 対話目：初対面の会話，2 対話目：ある目的に対しての議論，3 対話目：目上の人との会話である．そのうち，初対面の対話についてのみ分析を行った．

但し，本研究では，会話参加者の一方 (以後，参加者 B と呼ぶ) の行動を比較的均一にし，それに対して，他方 (以後，参加者 A と呼ぶ) の行動を比較できるように，話者 B として男性 2 名，女性 2 名の俳優，あるいは演劇経験者が交代で収録に参加した．発話数は，ドイツ：1094　(A：639，B：455)，日本：1556 (A：833，B：723)) であった．

連絡先：山岡雄治，東京農工大学　工学府　情報工学専攻，
50007646208@st.tuat.ac.jp

2.1 アノテーション

対話参加者の非言語行動として，話者 A の姿勢（頭，腕，足），ジェスチャの発生箇所と表現形態（繰り返し性，流暢さ，力強さ，速さ，長さ）をアノテーションツール anvil[2] を用いて，30frames/秒の精度でデータ化した．尚，姿勢変化の分類には [4]を，ジェスチャ表現形態の分類には [3] を用いた．本コーパス中の姿勢とジェスチャのデータ数は以下の通りである．

- ドイツ姿勢変化数：644 回(足：76，頭：274，腕：294)
- 日本姿勢変化数：561 回(足：118，頭：248，腕：195)
- ジェスチャ総数（ドイツ：184 回，日本：70 回）

さらに，これらの非言語行動と発話の書き起こしデータを 1 つの anvil ファイルに統合し，言語・非言語情報の共起関係を視覚的にとらえられるコーパスデータを作成した．図 1 にその一例を示す．会話の書き起こしでは，ポーズ長が 0.3秒以上である場合に発話の区切りとした．

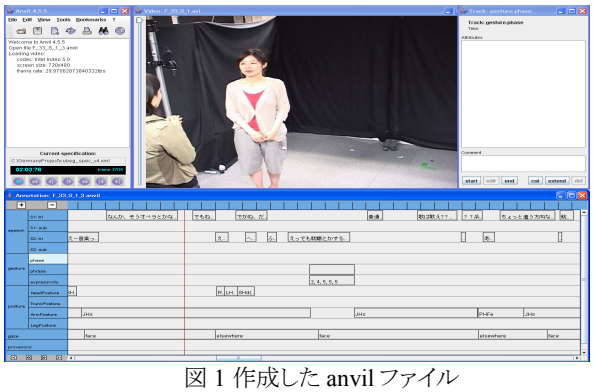

図 1 作成した anvil ファイル

以下の分析では，特に会話中の姿勢変化に着目し，詳細な統計的分析を行う．

2.2 対話のターン定義

対話とは，発話の番(ターン)を交互に交替することにより成り立つものである．参加者 A の非言語行動と発話の関係性を調査するため，分析データ内で，だれが話し手かを判定し，参加者 A が話し手である場合を話者ターン，参加者 B が話し手である場合を聞き手ターンとした．

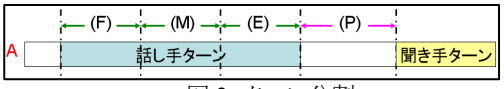

図 2 ターン分割

2.3 ターンの分割

姿勢の変化がターン内のどこで起きるのかを調査するために，ターンの開始 0.5 秒前から終了後 0.5 秒後までの時間を 3 分割し，開始から最初の 1/3 を First(F)，中間部分を Middle (M)，最後の 1/3 を End(E)とした．また，ターン交替時のポーズ区間を Pause (P)とした (図 2)．

ターンの総数と平均継続長のデータを次の表 1に示す．

表 1 ターン数と継続長

	ドイツ	日本
ターン総数	453	913
話し手ターンの平均時間	5.86	4.24
聞き手ターンの平均時間	5.03	3.92

3. 分析

Aの発話中(話者ターン中)において，各区間 (F, M, E) のどこでどのような姿勢変化が発生しているかを分析した．以下の表 2に，話者ターン中における足，頭，腕に関して全姿勢変化数中の F, M, E それぞれの発生割合，頻度 1 位，2 位の姿勢をまとめたものを示す．また，表3には，聞き手ターン中の姿勢変化に関して表 2と同様にまとめたものを示す．

表 2 話者ターン中における足，頭，腕の姿勢変化

		ドイツ					日本				
	区間	割合	頻度1位		頻度2位		割合	頻度1位		頻度2位	
足	F	39%	LRL	25%	WRL	25%	28%	WRL	43%	SLSF	21%
	M	22%	WRL	50%	LRL	29%	26%	WRL	42%	LRL	23%
	E	34%	LRL	33%	WRL	28%	31%	WRL	32%	LRL	24%
頭	F	36%	SHd	37%	THdAP	30%	31%	SHd	43%	THdAP	33%
	M	29%	SHd	33%	THdAP	30%	23%	SHd	49%	THdAP	20%
	E	28%	SHd	36%	DsHd	19%	25%	THdAP	49%	SHd	33%
腕	F	33%	PHIPt	61%	FAs	11%	27%	PHFe	30%	PHB	25%
	M	30%	PHIPt	52%	FAs	15%	23%	JHs	26%	PHFe	23%
	E	32%	PHIPt	40%	FAs	19%	32%	JHs	40%	PHWr	20%

表 3 聞き手ターン中における足，頭，腕の姿勢変化

		ドイツ					日本				
	区間	割合	頻度1位		頻度2位		割合	頻度1位		頻度2位	
足	F	14%	LRL	50%	WRL	50%	21%	WRL	31%	SLSF	31%
	M	29%	WRL	50%	LRL	25%	33%	LRL	29%	WRL	24%
	E	57%	LRL	56%	LSF	19%	46%	WRL	24%	LRL	24%
頭	F	16%	THdAP	22%	LHdAP	22%	20%	THdAP	38%	SHd	33%
	M	39%	THdAP	34%	DsHd	16%	30%	SHd	48%	THdAP	39%
	E	45%	THdAP	44%	LHdAP	14%	50%	THdAP	54%	SHd	22%
腕	F	24%	PHIPt	33%	FAs	27%	18%	JHs	29%	PHFe	21%
	M	31%	PHIPt	38%	FAs	21%	34%	JHs	35%	PHFe	27%
	E	44%	PHIPt	44%	FAs	16%	48%	PHFe	30%	JHs	22%

3.1 各区間における姿勢変化の割合

話者ターン中の姿勢変化についてまとめた表 2より，ドイツにおける非言語行動の発生割合は，足，頭，腕，全てにおいて，区間 F での発生頻度が最も高く，特に，頭部姿勢についてはその傾向が顕著である．それに対し，日本は，頭部については区間 F での発生頻度が高いが，足，腕部に関しては，区間 E において最も発生頻度が高かった．

聞き手ターン中の姿勢変化についてまとめた表 3 より，ドイツと日本共に共通して，非言語行動の発生割合は，区間 E での発生頻度が高いことが顕著である．

A) 足部の姿勢

表 2と表 3 より，ドイツと日本に共通して頻度 1 位，2 位を占めている(日本人データでの，足をまっすぐの姿勢に戻す姿勢 (SLSF)とドイツ人データでの，足を傾ける姿勢 (LSF) を除く) ものは，右足，あるいは左足に体重をかける姿勢 (WRL，LRL) である．よって，どちらかの足に体重を乗せる姿勢は，両国に共通して頻出する姿勢だと考えられる．また，話し手ターンと聞き手ターンにおいて，発生する姿勢は，共通であることがわかる．

B) 頭部の姿勢

表 2と表 3 より，ドイツと日本に共通して頻度 1 位，2 位を占めているものが， 話し相手の方向から顔を背ける姿勢変化 (THdAP)，THdAP の姿勢から話し相手の方向へと顔を向きなおす姿勢変化 (SHd) である．発話開始時に相手から視線を背けるのは，典型的なターン開始信号であるといわれており [5]，今回，ドイツ人，日本人においても同様の非言語行動が用いられることが明らかになった．

また，話し手ターン，聞き手ターンにおいて比較した場合，ドイツにおいて，話し手であるときは，相手の方向へ顔を向ける姿勢(SHd)が多いのに対し，聞き手ターン中では，話し相手の方向から顔を背ける姿勢を行うことが顕著であった．

C) 腕部の姿勢

表 2より，話し手ターンにおける腕の姿勢では，足，頭の姿勢とは異なって，両文化で共通しているものがない．ドイツ人は腕全体を用いる姿勢，例えば腕を組む姿勢 (FAs) や肘をもつ姿勢 (PHEw) を行っているのに対し，日本人は，手を顔に付ける姿勢 (PHFe) や両手を組む姿勢 (JHs)，手首を持つ姿勢 (PHWr) など，主に手を使った姿勢が用いられているため，ドイツ人より，動作が小さいという印象を受ける．また，ドイツ人ではポケットに手を入れる姿勢 (PHIPt) が最も多いのに対して，日本人では，この姿勢は全く観察されなかった．

また，表 2と表 3 より，聞き手ターンにおける腕の姿勢についても，同様の傾向が見られる．

3.2 各姿勢変化の継続長とその頻度

足，頭，腕部位における各姿勢変化の継続長とその頻度の違いについて，表 4～6 にまとめる．継続長は各姿勢変化の平均継続長，頻度は会話ごとの平均値を示す．

A) 足部姿勢の特徴

次の表 4に，足の姿勢の継続長，頻度についてまとめたものを示す．

表 4　足部姿勢の継続長と頻度

	ドイツ						日本					
	話し手ターン			聞き手ターン			話し手ターン			聞き手ターン		
	姿勢	継続長	頻度	姿勢	継続長	頻度	姿勢	継続長	頻度	姿勢	継続長	頻度
頻度1位	WRL	14.34	1.88	LRL	31.49	1.63	WRL	29.28	2.7	WRL	35.62	1.6
頻度2位	LRL	26.53	1.63	WRL	15.68	1	LRL	13.40	1.5	LRL	16.75	1.6

継続長と頻度に関しては，1 対話中における足部の姿勢変化は，発生しにくく，発生した場合には 10 秒以上その姿勢変化を行うことがわかる．また，ドイツと日本において，聞き手ターンで発生した姿勢変化の継続長が，話し手ターンで発生した姿勢変化の継続長よりも長いことがわかる．

B)　頭部姿勢の特徴

次の表 5に，頭部姿勢の継続長と頻度についてまとめたものを示す．

表 5　頭部姿勢の継続長と頻度

	ドイツ						日本					
	話し手ターン			聞き手ターン			話し手ターン			聞き手ターン		
	姿勢	継続長	頻度	姿勢	継続長	頻度	姿勢	継続長	頻度	姿勢	継続長	頻度
頻度1位	SHd	1.20	7.3	THdAP	2.35	4.6	SHd	0.71	6.4	THdAP	2.10	4.9
頻度2位	THdAP	1.83	5.4	DsHd	4.47	1.6	THdAP	1.98	4.8	SHd	0.68	3.4

表 5より，継続長に関して，全体的に約 1～4(sec)となっていて，短い継続時間であることがわかる．また，頻度に関しては，ドイツと日本に共通して，聞き手ターンと比べて話し手ターン中に多く姿勢変化が発生していることがわかる．

C)　腕部姿勢の特徴

以下の表 6に，腕部姿勢の継続長，頻度についてまとめたものを示す．

表 6　腕部姿勢の継続長と頻度

	ドイツ						日本					
	ターンA			ターンB			ターンA			ターンB		
	姿勢	継続長	頻度	姿勢	継続長	頻度	姿勢	継続長	頻度	姿勢	継続長	頻度
頻度1位	PHIPt	8.67	9.5	PHIPt	5.67	6.1	JHs	16.72	3.5	JHs	14.95	2.1
頻度2位	FAs	2.55	3	FAs	5.01	3.1	PHFe	3.65	1.9	PHFe	2.62	2.1

表 6より，継続長と頻度に関して，ドイツと日本の間で，ドイツは継続長が短く頻度が多いという特徴があるのに対し，日本は継続長が長く頻度が少ないという特徴があることがわかる．また，足，頭部位と同様に，話し手ターンに偏って姿勢変化が発生しているという特徴も得られた．

以上，ドイツ人と日本人の対話データを分析した結果，ドイツ人のデータでは，姿勢変化は区間 F に偏って発生するのに対し，日本人では，区間 F, E に偏って発生している傾向が見られ，非言語行動の発生タイミングが異なることがわかった．

さらに，区間 F, M, E での発生頻度が高い姿勢変化について詳細に調べたところ，腕の姿勢に関して，ドイツと日本の間で姿勢変化の種類，姿勢継続長，頻度において大きく異なっていることがわかった．足と頭の姿勢に関しては，発生頻度が高かった姿勢が共通していたことから文化の違いに依存しない共通した姿勢であると考えられる．

話し手ターンと聞き手ターンにおいて姿勢変化の発生頻度が異なっていたことから，話し手であるときに姿勢変化が発生しやすく，また，頭部姿勢に関しては，発生する姿勢変化にも違いが見られた．

データ分析により得られた非言語行動の発生タイミングの違いと，区間 F, M, E での頻出姿勢の違い，そして個々の姿勢変化の継続長・頻度の違いを用いて，次章では，姿勢の自動生成を行うシステムを提案する．

4.　動作決定システム

提案するシステムは，エージェント動作決定システムとして，チャットシステムや外国語教材として応用することを考えている．

4.1 システム構成

システム構成は，図 3の通りである．以下，図中の番号順に，システムの処理フローを説明する．

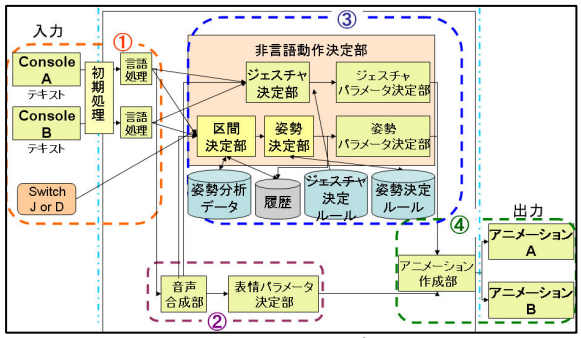

図 3　システム構成図

① テキストが入力されると，最初に入力文の初期処理，言語処理を行う．その後，入力テキストを履歴管理部，音声合成部，ジェスチャ決定部，区間決定部に出力する．

② 音声合成部では，入力テキストを受け取ると，合成音の音素タイミングを計算し，それを区間決定部に，また，作成された音声を表情パラメータ決定部に出力する．表情パラメータ決定部では，音声とモデルとのリップシンクを行い，表情パラメータを作成し，そのデータをアニメーション作成部に出力する．

③ 入力された音素タイミングデータ，入力テキスト，文化選択スイッチの値が区間決定部に入力されると，まず，履歴データと個々の姿勢変化分析データを比較し，部位ごとに (1) 姿勢変化を行うかどうか，(2) 行う場合はその区間を決定（詳しくは次項目で述べる）する．その後，姿勢パラメータ決定部にて，姿勢決定部で選択された姿勢の具体的な表現に関する詳細なパラメータを決定し，そのデータをアニメーション作成部に出力する．また，ジェスチャ決定部においても，ジェスチャ決定ルールに従って，入力のテキストと音声のタイミングデータから，どのタイミングでどのジェスチャを行うかを決定する．その後，ジェスチャパラメータ決定部にてジェスチャの強さ，スピードなどのパラメータが決定され，これらの情報がアニメーション作成部に出力される．

④ アニメーション作成部では，種類の入力データ（表情パラメータを含む音声データ，姿勢パラメータ，ジェスチャパラメータ）から，アニメーションを生成し，音声と同期したアニメーションを出力する．

4.2 非言語動作決定部

前節の③にあたる，非言語動作決定部中の姿勢決定方式について，さらに詳細に説明する．非言語動作決定部における処理のフローチャートを図 4に示す．

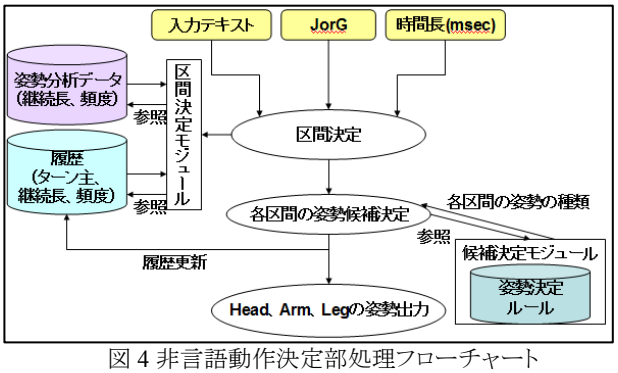

図 4 非言語動作決定部処理フローチャート

例えば，テキスト「おはようございます，今日もいい天気ですね．」が入力され，文化選択は，J（エージェントに日本人の姿勢を行わせる，ドイツ人の姿勢を行わせたい場合には，G）が選択され，音声合成器から出力された合成音声の長さが 2500msec となるとき，腕の姿勢を決定する際には，以下のような処理が行われる．

まず，履歴を参照したとき，現在エージェントが行っている姿勢を参照すると，2000（msec）前から PHFe を行っていることが得られたとする．また，表 6 から，PHFe の平均継続長が 3500（msec）であることがわかる．履歴の継続長 2000msec と音声の時間長が 2500msec であるため，この発話において，PHFe の姿勢を継続した場合，表6から得られた PHFe の平均継続長を超える．このため，音声の時間超 2500msec のうち，最後の 1000msec に対応する区間 M もしくは E において姿勢変化を行うことを決定する．もし，発話音声が短く，同じ姿勢を継続しても，PHFe の平均継続長を超えない場合には，姿勢変化を行わない．区間 M と E のどちらで行うかは，表 2 に示した区間 M と E の割合に応じて決定する．例えば，区間 E で姿勢変化を行うことが決定された場合，次に，候補決定モジュールに「JAE」（Japanese Arm End）の3つ組のパラメータが入力され，区間 E における姿勢変化の候補が出力される．ここで，候補決定モジュールは，表 2 に示した姿勢変化の分布（JHs：40%，PHWr：20%）に応じた姿勢変化（例：JHs）を返す．

4.3 応用例 1：チャットシステム

エージェント動作決定システムをチャットシステムに組み込むことにより，図 5に示すようなキャラクターアニメーション付きチャットステムが実現できる．チャットシステムを使うように，テキストを双方が入力すると，それが音声とキャラクターアニメーションに変換されて相手側の画面に表示される．

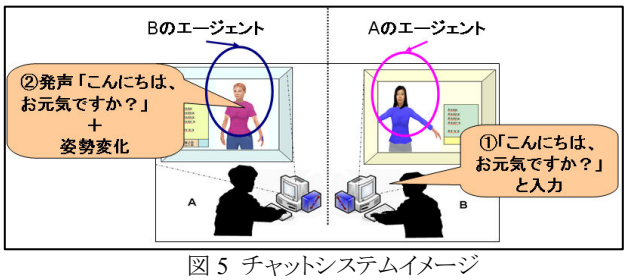

図 5 チャットシステムイメージ

例えば，利用者 B（図中：右側の人）がテキストボックスにテキスト（「こんにちは，お元気ですか？」）を打ち込むと，利用者 A の画面上（図中：左側の画面）で，B のエージェントが音声（「こんにちは，お元気ですか？」）に同期して姿勢の変更やジェスチャを行う．また，ドイツ人／日本人モードに設定することにより，文化に適応したエージェント非言語行動が自動的に生成される．

4.4 応用例 2：外国語教材

外国語教材としての使用イメージを図 6に示す．チャットシステムとは異なり，アニメーションの出力は，生徒側へのみ行われる．

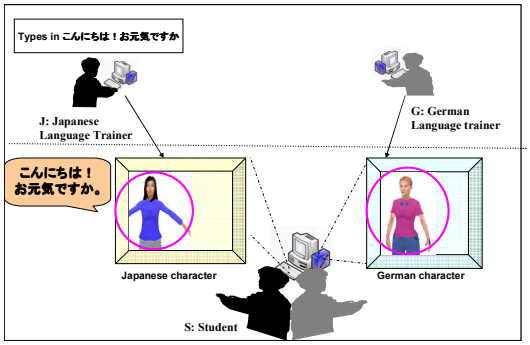

図 6 外国語教材システムイメージ

例えば，日本語の先生 J が入力テキストボックスにテキストを打ち込むと，利用者 S の画面上で，J のエージェントが音声言語に同期して姿勢の変更やジェスチャを行う．また，ドイツ人／日本人モードに設定することにより，文化に適応した動作を行う先生エージェントが自動的に生成される．このようなシステムにより，生徒は，言語のみならず，非言語的なコミュニケーションについても学ぶことができる．

5. おわりに

本稿では，会話における姿勢の特徴をドイツ人と日本人の会話データを比較することにより明らかにし，これに基づき，文化に適応した会話エージェントの動作決定機構を提案した．本研究では，発話の内容と非言語行動との関係については，未分析であるが，ジェスチャ等の非言語行動が，特に発話内容に依存している可能性は高い．今後はこの点もシステム実装において考慮するために，発話内容とジェスチャの依存関係を再度分析する必要がある．

参考文献

[1]Rehm, M.et al.: Creating a Standardized Corpus of Multimodal Interactions for Enculturating Conversational Interfaces, IUI2008 Workshop on Enculturating Conversational Interfaces by Socio-cultural Aspects of Communication, 2008.

[2] Kipp, M. Anvil - A Generic Annotation Tool for Multimodal Dialogue. In Proceedings of *the 7th European Conference on Speech Communication and Technology*, pp. 1367--1370, 2001.

[3]Pelachaud, C. Multimodal expressive embodied conversational agents. In Proceedings of ACM Multimedia, pp. 683–689, 2005.

[4] Bull, P.E., Posture and Gesture: Pergamon Press, 1987.

[5] Duncan, S., On the structure of speaker-auditor interaction during speaking turns. *Language in Society*. **3**: pp. 161-180, 1974.