道案内対話におけるマルチモーダルインタラクションの収集と分析
—メタバースアバタのためのジェスチャ自動決定にむけて—

Collection and Analysis of Multimodal Interaction in Direction Giving Dialogues: Towards an Automatic Gesture Selection Mechanism for Metaverse Avatars

塚本剛生 [1]　室谷優実 [2]　岡本雅史 [2]　中野有紀子 [2]

Takeo Tsukamoto[1], Yumi Muroya[2], Masashi Okamoto[2], and Yukiko Nakano[2]

[1] 成蹊大学大学院理工学研究科理工学専攻
[1]Graduate School of Science and Technology, Seikei University
[2] 成蹊大学理工学部情報科学科
[2] Faculty of Science and Technology, Seikei University

Abstract: With the goal of building a spatial gesture generation mechanism in Metaverse avatars, this paper reports an empirical study for multimodal direction giving dialogues. First, we conducted an experiment where a direction follower asked the way to some places in a university campus, and the direction giver gave a direction to there. Then, using a machine learning technique, we annotated direction giver's right hand gestures automatically, and analyzed the distribution of the direction of their gestures. As the result, we found that the distribution of gesture directions is different depending on the proxemics of the conversational participants as well as the place of the landmarks which they are talking about. In future work, we plan to establish a method for spatial gesture generation, and implement it into a Metaverse application.

1. はじめに

セカンドライフ(SL)に代表されるインターネット上の仮想世界であるメタバースアプリケーションが，世界中で幅広く利用されつつある．このような仮想世界上で相手とコミュニケーションをする主な方法は，アバタと呼ばれる自分自身の分身を用いてテキスト入力によるチャットを行わせることであるが，現状では，吹き出しにチャットテキストが表示される方式が主流であり，対面場面での音声や身振りに近い表現は，非常に限定的にしか使うことができない．

一方，対面コミュニケーションの研究では，ジェスチャ等の非言語行動は言語行動を補う重要な情報であることがわかっている．特に，道案内を行う対話では言語表現を補足するために，方向や建物の位置関係等，空間的なジェスチャが豊富に用いられることが特徴的である．

上記の議論から，仮想世界の 2 体のアバタ間での道案内の会話において，適切な空間的ジェスチャを自動的に付与する機構が実現すれば，チャットテキストのみの表現よりも，案内が格段にわかりやすくなることが期待される．

そこで本稿では，仮想空間でジェスチャを用いながら道案内を行うアバタの実現を目指し，基礎的な検討として，人間同士の道案内対話の収集実験の実施，音声，ジェスチャデータの収集について報告する．さらに，ジェスチャの形態を規定する要因として，会話参加者の立ち位置や案内する対象との位置関係について分析・考察した結果を報告する．

2. 関連研究

ジェスチャの大部分は発話と共起して現れ，発話中の重要な概念を強調したり，発話のリズムを形成するために用いられる．McNeil[1]は，このような発話中に起こるジェスチャをその機能の観点から iconic, metaphoric, beat 等に分類している．この分類を採用したジェスチャ自動付与システムとして，[2]

は，テキストが入力されると，エージェントのアニメーションスケジュールと，合成音声を用いて発話音声を作成するシステムを実装している．[3]は，SL内で英語の文章を入力すると，ジェスチャーアニメーションを実行するシステムを開発しており，一般的な場合では beat ジェスチャー(腕や手の上下の運動)が実行される．しかし，iconic や metaphoric ジェスチャについては，その動きや形に意味のあるため，ジェスチャの形態決定も必要であるが，これらの研究ではジェスチャの形態決定については考慮されていない．その理由として，ジェスチャの形態は個人差が大きく体系的な分類が非常に難しいことがあげられる．しかし，ジェスチャを自動付与するシステムを構築するには，ジェスチャの機能上のタイプの決定に加え，ジェスチャの形態の決定方式が必要不可欠である．

この問題を解決しようとした研究として，[4]は道案内の対話に焦点を当て，道案内場面で用いられるジェスチャの形態の決定方式を提案している．彼らの方式では，ランドマークとなる建物等の形の特徴をいくつかの次元における程度として表現している．概念を特徴づける多次元の属性と値のセットを単語に登録することにより，発話の意味表現から，発話の言語表現とジェスチャから構成されるマルチモーダルな表現を生成する際，単語に付与された属性値を参照することによりジェスチャの形態を決定することができる．これにより，単語エントリとジェスチャ形態との直接的な組み合わせをジェスチャ辞書としてあらかじめ用意することなく，より汎用性が高く，柔軟なジェスチャ選定が可能になるとしている．さらに，[5]では，ベイジアンネットワークを使用して，個人間のジェスチャ形態の差異を表現している．モーションキャプチャを使用して被験者のジェスチャの手のひらの方向，指の方向，移動の方向等を取得し，それらをベイジアンネットワークの各ノードとすることにより，個人間のジェスチャの差異を予測するジェスチャ生成システムを提案している．

そこで本研究においても，道案内場面に限定し，ジェスチャの形態決定方式の確立を目指す．また，[4][5]では，ランドマークの形状に着目した方式が提案されているが，本研究では，会話参加者の立ち位置や説明対象との空間的な位置関係に着目して研究を進める．

3. 道案内におけるマルチモーダルインタラクションデータの収集

仮想空間でジェスチャを用いながら道案内を行うアバタを実現するためには，道案内の説明中に行われるジェスチャの形態を規定する要因を解明する必要がある．本節では，その基礎となるデータ収集を目的とした実験について報告する．

3.1 実験手順

本実験では，成蹊大学に初めて訪れる訪問者と道案内者役の成蹊大学生がペアになり，大型のスクリーンに映し出された仮想空間内の学内の画像 (図 1) を見ながら (図 4-b)，訪問者役の被験者が大学内の特定の建物の場所を案内役に尋ね，その建物まで行く道順を案内役が説明する会話を収録した．

(1) 実験手続き：訪問者役には，案内者役との会話を通して，目的の場所までの道順を十分理解するように指示した．一方，案内者役の被験者には，訪問者役が道順を正しく理解していることを確認する様に指示した．

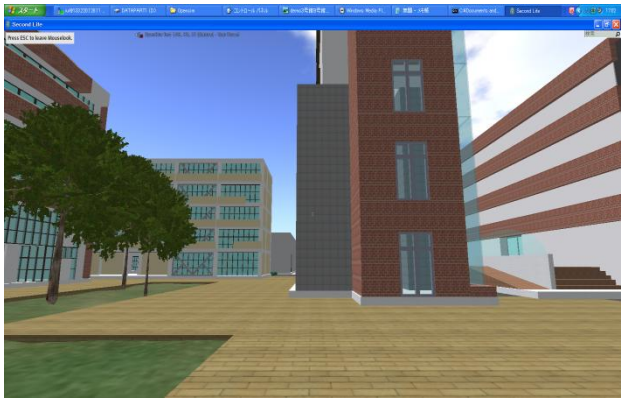


図 1　案内する際に使用する画像例

また案内者役には，目的地までの道のりを説明する際に，必ず言及しなければならないポイントを各目的地につき 2 つずつ設定した．それ以外の説明については案内者役の自由に任せた．訪問者役が案内者役の説明を正しく理解していることを確認するために，一通り説明した後，目的地までの道のりを訪問者役が案内者役に説明するフェーズを設定した．その際，訪問者役の説明が間違っていると案内者役が判定した場合は，案内者役が再度説明を行った．

(2) 実験材料：目的地となる大学内の場所として，全部で 6 か所設定し，目的地毎に図 1 に示すような，学内の仮想空間モデルから切り出した画像を用意した．1 か所の目的地の説明を 1 セッションとし，各ペアは 6 セッション行い，全ての目的地についての会話を 1 回ずつ収録した．

(3) 実験条件：実験の条件として，セッション開始時の訪問者役の初期位置を以下の 3 種類設けた（図 2）.

横条件：訪問者役の横に説明対象物 (スクリーン)がある

前方条件：訪問者役の前方に説明対象物(スクリーン)がある

後方条件：訪問者役の後方に説明対象物(スクリーン)がある

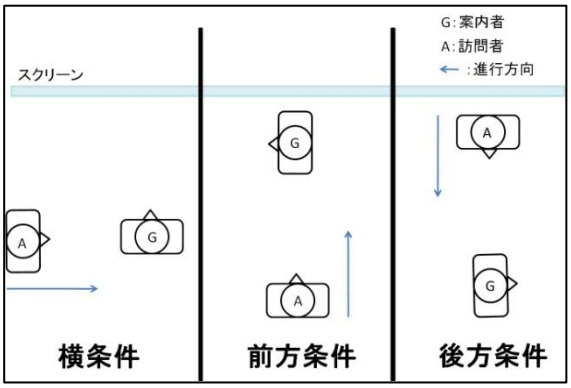


図 2　実験の条件

実験のスタートの合図で，訪問者役が図 2 に示す条件の初期位置から案内者役に近づいて声をかけることで，道案内対話が始まる． 6 種類の目的地に対し，3 種類の初期位置条件が 2 回ずつ適用された．また，目的地と初期位置条件の組み合わせは偏りがないよう被験者ごとに変更した．

一方, 案内者役の初期位置も条件によって異なる．横条件では，案内役は図 3 の位置(A)，前方条件では図 3 の位置(B), 後方条件では図 3 の位置(C)に立つ．従って，どの条件においても，訪問者役から見ると案内者役が横を向いている状態にある時に，歩いて近づいてゆき, 声をかけるという設定となっている．

さらに, 案内者役には, 可動範囲の制限を設けた．案内者役は最初の立ち位置を示すマットからどちらかの片足が入っている範囲でなければ動くことが出来ない．訪問者役には可動範囲の制限は設けていない．これは，将来的に SL を使用したシステムの実装する際，案内者が移動するのではなく，訪問者を誘導することにより，両者の位置関係が適切になるような方式を実現するためである．

(4) 被験者：訪問者役は成蹊大学関係者ではない男性 7 名，女性 7 名．案内者役は成蹊大学関係者の男性 14 名であり，全て大学生，大学院生である．これらの訪問者役と案内者役のペア，計 14 組 28 名が実験に参加した．

3.2 実験環境

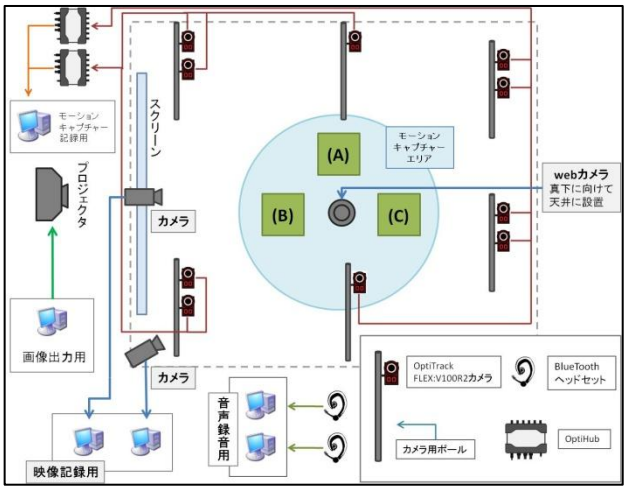


図 3　実験環境図

実験環境の概要を図 3 に示す．被験者 2 人には音声データを収録するための Bluetooth ヘッドセットマイクとモーションキャプチャデータを取得するためのマーカが装着されたカーディガンを着衣してもらった．

2 者のインタラクションの様子は，スクリーンの上部と横に設置してあるビデオカメラ（図 4-a），天井に固定してある Web カメラ（図 4-c）で収録し，Bluetooth ヘッドセットマイクで被験者 2 人の音声データを収録した．また，3m×3m の正方形の周りにポールを 6 か所設置し，そこにモーションキャプチ

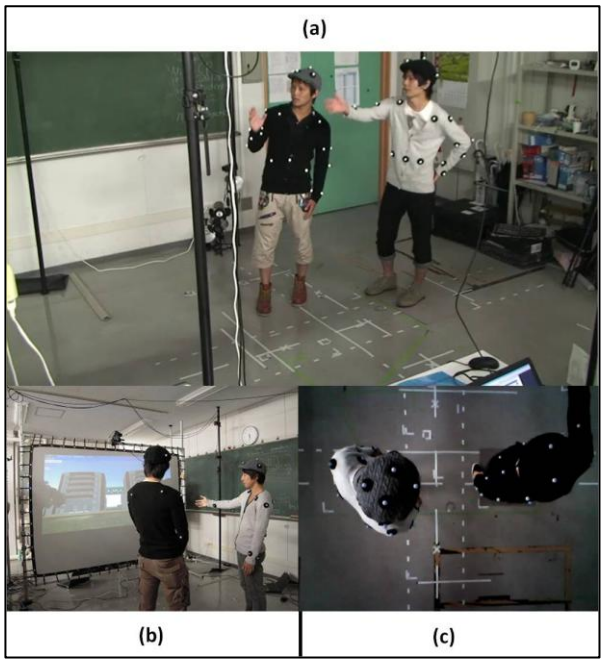


図 4　実験風景

ヤ OptiTrack 用のカメラ 10 台を設置し，上半身の動作データを取得した．実験の様子を図 4 に示す.

3.3 収集データ

本実験で収集したデータは，会話全体の様子を収録した 2 方向からのビデオデータ，被験者 2 人の立ち位置の様子を天井から収録した Web カメラデータ，会話の言語情報を収録した音声データ，およびその書きお越しテキスト，被験者 2 人の上半身の動作を取得したモーションキャプチャデータである.

実験で収集した会話データは, 6 セッション／ペア×14 ペアで, 全 84 会話である. 1 会話の平均時間は 68.5 秒であった．モーションキャプチャデータは被験者 2 人の頭部，肩，腹部，背中，右腕，左腕にマーカを付与し，モーションキャプチャソフト OptiTrack を使用し，100fps で会話データ収集時間と同時間のモーションキャプチャデータを取得した．収集した会話の一例を図 5 に示す.

発話者	発話開始時間（秒）	発話終了時間（秒）	発話内容	ジェスチャ有無	対象物方向
A	3.0253	5.9301	すいません四号館への行き方を教えていただけますか		
G	6.113	6.473	はい		
G	6.8153	7.9568	え一っと<声>		
G	8.3196	10.591	建物がいっぱいあってちょっと分かりにくいと思うんですけども	rh	center
A	10.5869	10.7703	はい		
G	10.8244	12.1624	あの一渡り廊下が	rh	center
G	12.3124	13.4045	正面にあるじゃないですか	rh	center
A	13.0544	13.3586	はい		
G	13.5765	14.5465	(W_ダレーガ|であれが)		
G	14.7526	15.1474	あの		
G	15.3469	16.6556	(W_ワタリロカ一ガ|渡り廊下が)	rh	center
G	17.113	19.3981	<声>左手が八号館で右手が	rh	left, right
G	19.7186	22.8118	七号館なんですけどまその二つが渡り廊下で繋がってるんですね	rh	right, center
A	21.9782	22.2199	はい		
G	22.8493	23.7953	でその下をとりあえず	rh	center
G	23.9948	24.9333	くぐっていただいて	rh	center
A	24.6707	24.8541	はい		
G	25.4867	26.5088	でそうすると		

図 5　データ一例

4. 分析

訪問者役と案内者役の立ち位置や，説明対象の位置によって，空間的ジェスチャの形態がどのように異なるのかを調べるために，本実験にて収集したデータのうち，今回はランダムに選んだ 5 ペアについて，3 セッションずつ 15 セッションの案内者役の右腕の動きをモーションキャプチャデータを用いて分析した．以下では，特にジェスチャの向きに着目した分析を行う.

4.1 ジェスチャの自動アノテーション

案内者役が右腕でジェスチャをしているか否かを自動的に判定するために，モーションキャプチャにより取得したデータを用いて, 決定木学習を行った. 学習には　データマイニングツール Weka の J48 を用いた．用いた特徴量は，右腕の x 座標，y 座標，z 座標の移動量，右腕の x 軸回転，y 軸回転，z 軸回転，右腕と肩との相対 x 座標，相対 y 座標，相対 z 座標，右腕と肩との距離の 10 個である．判別は右腕のジェスチャの有無の 2 値である. 教師データには，案内役 2 人 6 セッション分のビデオについて，右腕のジェスチャをしている箇所のラベル付けを行った結果を用いた．ラベリング作業には，ビデオアノテーションツール Anvil 5.0 を用いた.

決定木学習の結果，サイズが 873，葉の数が 437 の決定木が生成され，10 回の交差検定における分類精度は 97.5%であり，十分な精度が得られた.

機械学習で得られた決定木を，15 セッション分のモーションキャプチャデータに適用し，会話中に案内者役が右腕のジェスチャを行っている区間を自動でアノテーションした.

4.2 ジェスチャ分布パターンの分類

得られた 15 セッション分の案内者役の右腕のジェスチャについて，その 2 次元平面上の位置をプロットした．その結果，ジェスチャの位置の分布を以下の 4 つのパターンに分類することが出来た(図 6).

(i) 右下がりの長方形に分布
(ii) 正方形あるいは円形に均等に分布
(iii) 孤あるいは波線を描くように分布
(iv) 縦型に分布

15 セッション中，(i)に分類されたものが 6 個，(ii)が 3 個，(iii)が 4 個，(iv)が 2 個であった.

4.3 立ち位置の分類

次に，天井から撮影した web カメラからの映像を観察し，各セッションの被験者ペアの立ち位置を分類した．その結果，以下の 3 種類に分類できることが分かった(図 7).

(a) ハの字型：訪問者役と案内者役の両者がほぼ同じ角度でスクリーンに対してやや斜めに立ち，2 者の相対的な立ち位置によりハの字を形成している.
(b) L 字型：訪問者役か案内者役のどちらかが，スクリーンに対して正面を向いており，もう一方がやや斜めに立っている.
(c) V 字型：訪問者役と案内者役の両者がスクリーン

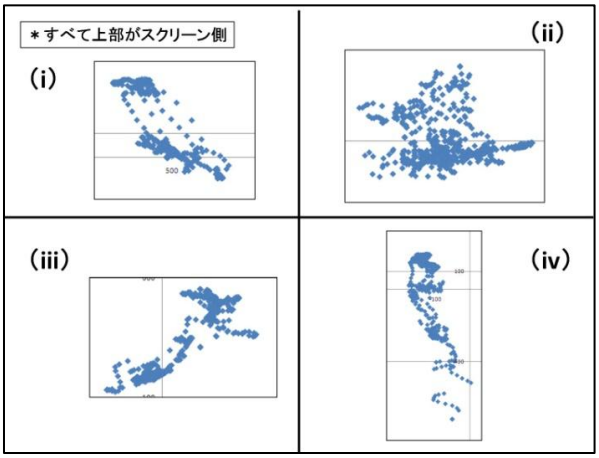


図 6　プロット分類

に対して大きく斜めに立ち，2 者が前後にずれて立っている．

4.4 立ち位置とジェスチャ分布との関係

さらに，4.2 節で示した(i)から(iv)のジェスチャ分布パターンと 4.3 節で示した立ち位置パターンとの関連を調べると，表 1 のような対応関係があることが分かった．

この結果から次のことが考察できる．ハの字型の立ち位置では，訪問者役と案内者役の前方に，均等に比較的広く空間が確保されているため，ジェスチャを提示するための共有空間が広い．そのため，ジェスチャも左右方向に幅広く広がっている．次に，L 字型の立ち位置では，スクリーンに対する角度が 2 者間で大きく異なるため，共有空間がやや歪んでいる．そのため，ハの字型に比べてジェスチャが分布する範囲が狭く，一定の角度にジェスチャが集中している．最後に，V 字型では，共有空間がほとんどない，あるいは縦に細長い空間しかない．そのため, ジェスチャの左右方向の広がりはほとんどなく，ジェスチャの分布は縦長になる．

表 1：立ち位置とジェスチャ分布の関係

立ち位置	ジェスチャ分布
(a) ハの字型	(i)長方形型，(ii)正方形型
(b) L 字型	(iii)波線型
(c) V 字型	(iv)縦長型

上記の考察の妥当性を検証するために，3 種類の立ち位置について，説明対象物がスクリーンの右，左，中央にある場合にそれぞれ分けて，ジェスチャをプロットした．図 8-(a),(b),(c)にその結果を示す．ハの字型のプロットを見ると，右，左，中央と幅広

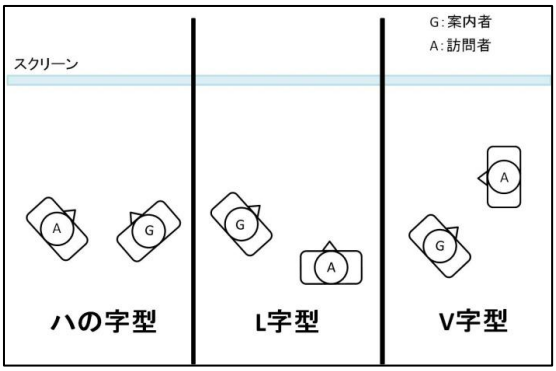


図 7　立ち位置分類

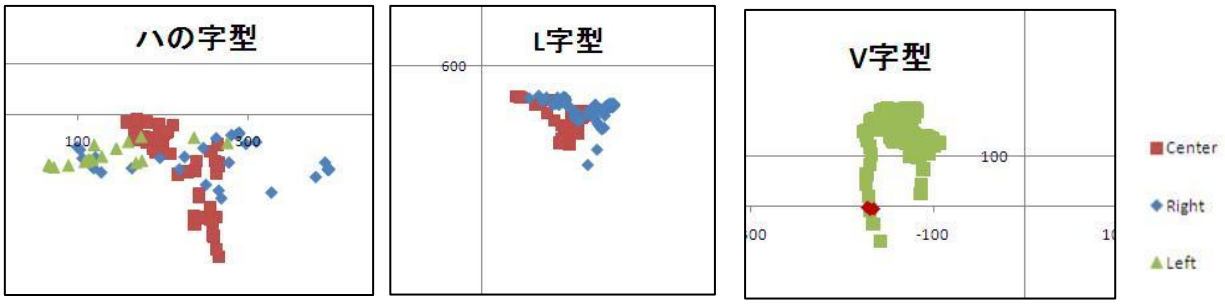
くジェスチャが分布しているのが見て取れる．L 字型では対象物が中央にある場合のジェスチャと右にある場合のジェスチャがほぼ重なっており，L 字型と V 字型の立ち位置は，ハの字型に比べて明らかにジェスチャの範囲が狭くなっているのが見て取れる．

以上の結果から，2 者の立ち位置によって，ジェスチャの分布範囲が異なること，さらに説明対象の位置によってジェスチャの向きが異なることが分かった．このことは，説明対象が同じでも立ち位置の形状が異なることにより，適切なジェスチャ方向が異なることを示す結果であり，立ち位置と説明対象の両方がジェスチャの方向を決定する要因として，重要であることを示唆している．今後は，これらの結果に基づき，仮想空間にけるアバタの位置に応じてジェスチャの方向を自動的に決定する方式の検討を進める．

5. ジェスチャ自動付与に向けて

本節では，ジェスチャ自動付与機能を有するアバタの実現に向け，現在我々が開発中の SL 上のシステムを紹介する．

本システムは, SL 上のチャットウィンドウに文章を打ち込むと，それに応じたジェスチャを自動的に付与し，音声合成による音声とアニメーションを同期させて出力するものである．システム構成図を図 9 に示す. 本ジェスチャ自動付与システムは, (1)Text Receiver，(2)ジェスチャ決定部，音声合成器，(3)Action-Voice Controller (AVC)，の 3 つのモジュールから成る．まず，Text Receiver では SL 独自のスクリプト言語であるリンデンスクリプト言語 (LSL) を用いる．SL ではプリムと呼ばれるオブジェクトに LSL を組み込み，それをアバタに装着させることでアバタの制御ができる．この機能を用いて，ユーザが入力したテキストをジェスチャ決定部に送信する機能を実装した．



(a) (b) (c)

図 8　立ち位置によるジェスチャ分布範囲の違い

ジェスチャ決定部は CAST [2] を拡張することにより実装している．まず，テキストを形態素解析 JUMAN, 構文解析 KNP による言語タグ付与機構に送り，形態素情報を付与したのち，CAST に既に組み込まれているジェスチャ決定ルールが適用され，beat ジェスチャが決定される．これに加え，[4]を参考に文中の特定の単語に意味的な特徴を付与することによって，iconic ジェスチャを自動付与する機構を組み込んだ．最後に，入力テキストに対して上記の方法でジェスチャが付与されると，音声合成器から得られる音素タイミングから，リップシンクのための viseme とあわせてアニメーションのタイムスケジュールが自動生成される．AVC では，作成した日本語合成音声を再生し，タイムスケジュール通りにアバタのジェスチャを実行する．

以上の機構により，SL で日本語文章を入力すると，日本語音声合成が再生されると同時に，その文章に適したジェスチャが出力される．今後はジェスチャ決定部を拡張し，本稿で分析した空間的なジェスチャの決定方式を組み込む予定である．

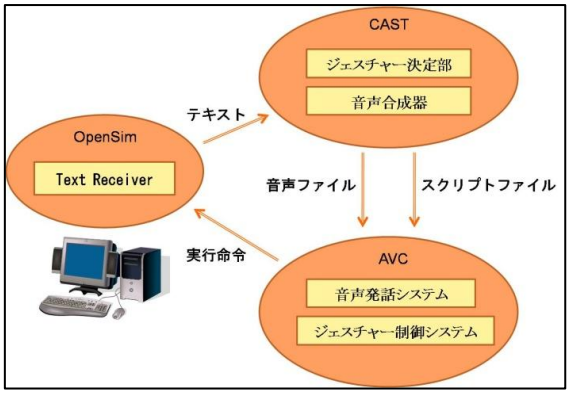


図 9　システム設計図

7. まとめ

本論文では，道案内対話収集実験について報告し，モーションキャプチャから得られる案内役の腕の動きから，ジェスチャの方向データを自動抽出し，会話参加者の立ち位置や案内する対象との位置関係によるジェスチャ方向の差異を分析・考察した．今後は，定量的な分析を進め，ジェスチャ形態自動付与システムに応用していく予定である．

謝辞

本研究は科研費基盤(S)(課題番号:19100001)の助成による．

参考文献

[1] David McNeill : Hand and Mind, University of Chicago Press (1992)

[2] Yukiko Nakano, Masashi Okamoto, Daisuke Kawahara, Qing Li, and Toyoaki Nishida : Converting Text into Agent Animations: Assigning Gestures to Text, In Proc. of HLT-NAACL 2004 Companion Volume, pp. 153–156, (2004)

[3] Werner Breitfuss, Helmut Prendinger, and Mitsuru Ishizuka : Automatic generation of gaze and gestures for dialogues between embodied conversational agents, Int'l J of Semantic Computing, Vol. 2, No. 1, pp 71-90, (2008)

[4] Paul Tepper, Stefan Kopp, and Justine Cassell : Content in Context: Generating Language and Iconic Gesture without a Gestionary, Proceedings of the Workshop on Balanced Perception and Action in ECAs at AAMAS '04, (2004)

[5] Kirsten Bergmann and Stefan Kopp : GNetIc – Using Bayesian Decision Networks for Iconic Gesture Generation, Proceedings of the 9th International Conference on Intelligent Virtual Agents, pp 76-89, (2009)