音声言語情報処理 6－6
（1995．5．25）

マルチモーダル翻訳対話における発話とジェスチャの関連性

水梨豪　ローケン-キム キュンホ　パク ヨンドク　友清睦子　森元逞

ATR音声翻訳通信研究所
〒619-02　京都府相楽郡精華町光台2-2

マルチモーダル入力の統合的な解析の枠組みを構築するための指針を得ることを目的として、音声とジェスチャを入力として受け付けるマルチモーダル翻訳対話環境における対話データを収集し、ジェスチャ入力の特徴、発話とジェスチャの意味的関連性および時間的関連性を分析した。

その結果、ジェスチャの種類、ジェスチャが保持する情報、ジェスチャが情報を提供する発話中の要素、発話とジェスチャの意味統合の仕組、発話とジェスチャの時間的前後関係といった基本的事項に関する知見が得られた。さらに、これらの結果をもとに、発話とジェスチャの統合解析への指針も抽出した。

Relation between Speech and Gesture in Multimodal Interpreting Dialogues

Suguru Mizunashi　Kyung-ho Loken-Kim　Young-Duk Park　Mutsuko Tomokiyo　Tsuyoshi Morimoto

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02

In order to develop a framework for integrated analysis of multimodal input, we collected dialogue data in a multimodal interpreting dialogue environment in which speech and gesture are accepted as input, and analyzed characteristics of gesture input, semantics and the timing relation between speech and gesture.

As a result, we obtained basic knowledge concerning kinds of gestures, information which gestures carry, elements in speech about which gestures give information, mechanisms of integrating the meaning of both speech and gesture, and the timing relation between speech and gesture. These results represent a guide toward integrated analysis of speech and gesture.

1 はじめに

近年、計算機とその周辺機器の進歩にともない、人間と計算機の対話様式における多様性が増大している。人間が計算機に情報を入力する方法に関しても、従来のキーボードやマウスに加えて、タッチパネルを使った指による指示入力、データグローブによる動作の入力など、人間にとってより自然で効率的な方法がマルチモーダル対話システムなどに盛んに実装されるようになった。[1]-[4]

しかしながら、このようなさまざまな入力方法を用いたマルチモーダル・システムに関する従来の研究・開発においては、ユーザがマルチモーダル情報を入力する際の特性はほとんど考慮されていない。したがって、マルチモーダル入力の解析に関しても、各モダリティの特性をふまえた上での統合的な解析が行われている例はきわめて少ない。

そこで、今回は、マルチモーダル翻訳対話システムにおける、マルチモーダル入力解析の枠組みを研究するための指針を得ることを目的として、音声とジェスチャを入力として受け付けるマルチモーダル翻訳対話環境における対話データを分析し、ジェスチャ入力および発話の特性と、両者の関連性に関する知見を整理した。

以下、2章ではまず、分析対象となる対話データの収集方法を述べる。3章でジェスチャの方法による分類を示し、4章ではジェスチャが持ちうる情報を整理して、発話とジェスチャとの関連性を考察する。5章で発話とジェスチャの時間的関連性を述べ、最後に6章では、発話とジェスチャの統合解析において重要と考えられる事項を整理する。

2 データの収集

2.1 データ収集環境

実験環境としては、ATR音声翻訳通信研究所が開発した EMMI (ATR Environment for MultiModal Interactions) [5]を用いた。この環境では、通訳者を介した、2か国語間のマルチメディア／マルチモーダル・コミュニケーションのシミュレーションが可能である。

発話とジェスチャのデータを収集するため、日本語を母国語とする者と英語を母国語とする者を被験者として、「人間の通訳者を介したマルチモーダル音声翻訳実験」（以下「通訳者実験」と呼ぶ）と「機械翻訳を介したマルチモーダル音声翻訳実験」（以下「WOZ実験」と呼ぶ）の2つの実験を行った。後者の実験では、機械翻訳対話を模擬するために Wizard of Oz 方式 [6] を採用し、機械を装った発話をする通訳者2人（日本語発話通訳者と英語発話通訳者）を Wizard として配した。各 Wizard の声は、被験者に「機械が通訳している」という意識を持たせるために、ヴォイス・イフェクタを通して合成音声のように聞こえるように変換した。

被験者数は、通訳者実験では日本語話者、英語話者各9人、WOZ実験では各10人であった。実験では、日本語話者を国際会議の事務局員の役（以下「エージェント」と呼ぶ）、英語話者を国際会議の参加のために京都を訪れた客の役（以下「クライアント」と呼ぶ）に設定して、両者の間で京都駅から国際交流センターへの道案内とホテル予約に関する対話を行ってもらった。

実験を通じて、エージェント、クライアント、通訳者の音声をDAT (Digital Audio Tape) に録音し、通訳者とクライアントのディスプレイのイメージを、ダウン・コンバータを通してビデオ・テープに録画した。なお、通訳者のディスプレイでは、エージェントの描画の様子がわかるようになっている。

2.2 道案内タスクで使用した地図

道案内のタスクでは、京都駅構内地図、京都市内の鉄道路線図、国際交流センター付近の地図を用いた。図1に京都駅構内の地図を示す。いずれの地図も、構造を持たないひとつのグラフィック・オブジェクトとしてディスプレイ上の21cm×16cmの領域に表示される。

地図上には、「オブジェクト」と「名前」が描かれている。「オブジェクト」とは、円や矩型で表わされる建造物、道路、鉄道、場所など、「名前」とは、オブジェクトの名前を表わす文字列（約8mm×6mm）である。

被験者のディスプレイにはタッチパネルが組込まれており、被験者は指を使って地図上に自由に線などを描くことができる。図1中の上半分あたりに描かれた線は、その一例である。

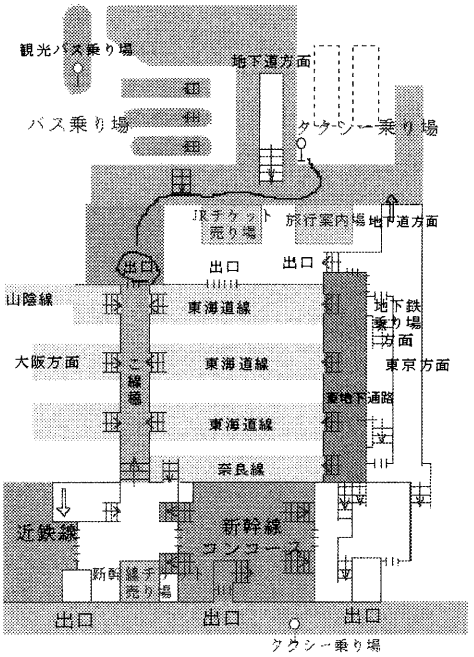

図1　京都駅構内の地図

3　ジェスチャの分類

本稿でいう「ジェスチャ」とは、タッチパネルを通して被験者がディスプレイの地図上に線などを描く行為を指す。

実験の結果、ジェスチャ全体の90%以上が、道案内のタスクに集中して用いられていることがわかった。これは、もうひとつのタスクであるホテル予約では予約シートに名前や日付などのデータをキーボードを用いて書き込む作業が中心で、地図がほとんど使用されなかったことによる。よって、今後ジェスチャの使用頻度などの統計情報を扱う際には、データの適正さを考えて、道案内タスクにおける対話のみを考慮することにする。

ここでは、はじめにジェスチャをともなうターンの数を集計してジェスチャが用いられる傾向を示し、その後、ジェスチャの使用形態の分類を示す。

3.1　ジェスチャをともなうターン

3.1.1　通訳者実験と WOZ 実験

表1、表2はそれぞれ、通訳者実験、WOZ実験におけるエージェントとクライアントについて、ひとつ以上のジェスチャをともなうターンとジェスチャをともなわないターンの数を集計したものである。全ターン数（エージェントとクライアントのターンの合計）を比較すると、WOZ実験が通訳者実験の約65%であるにもかかわらず、ジェスチャをともなうターン数に関しては逆に前者は後者の1.4倍となっている。この原因としては、WOZ実験においては、エージェントとクライアントが、「機械が通訳している」という認識を持っているので、人間の通訳者に対するよりもわかりやすく効率的に情報を伝えなければならないという意識が働き、結果として、言語情報（音声）よりも直接的でわかりやすい視覚情報（ジェスチャ）を多用したということが考えられる。

表1　通訳者実験におけるターンの数

	エージェント	クライアント	計
ジェスチャ有	37	10	47
ジェスチャ無	260	288	548
計	297	298	595

表2　WOZ実験におけるターンの数

	エージェント	クライアント	計
ジェスチャ有	46	20	66
ジェスチャ無	151	167	318
計	197	187	384

3.1.2　エージェントとクライアント

表3は、ひとつ以上のジェスチャをともなうターン数の、全ターン数に対する割合を整理したものである。エージェントは、クライアントと比較すると、通訳者実験では4倍弱、WOZ実験では2倍強のジェスチャを使用していることがわかる。この原因は、対話における両者の役割にあると考えられる。すなわち、エージェントは地図上のオブジェクトの位置や進路のとり方などの説明を行うため

にジェスチャを多用するが、質問して情報を要求することが主な役割であるクライアントは、相対的にジェスチャの使用量が少なくなるということである。

表3　ジェスチャをともなうターンの割合

	エージェント	クライアント	全体
通訳者実験	12.1%	3.4%	7.9%
WOZ実験	23.4%	10.7%	20.8%

3.2　ジェスチャの使用形態

3.2.1　ジェスチャの方法

被験者が実験で用いたジェスチャを、その方法に応じて、サークリング、ドラッギング、ポインティングおよびマーキングに分類した。なお、各ジェスチャの例で使用する '<' と '>' の表記は、それぞれジェスチャの始まりと終わりを示し、'>' の右にある 'C' はそのジェスチャが「サークリング」であることを表わす。同様に、Dは「ドラッギング」、Pは「ポインティング」、Mは「マーキング」を表わす。

○サークリング

画面上に丸形の線を描くジェスチャ。ジェスチャの始点と終点が隣接している。使用例を発話とともに示す。

・<蹴上はここです>C 。
・<この場所がタクシー乗り場です>C 。
・<あなたは今ここにいます>C 。

囲まれるものは、地図上のオブジェクト（ホテルや駅）、オブジェクトの名前などであり、サークリング・ジェスチャの90%以上において、オブジェクトやその名前の外側に丸が描かれた。

○ドラッギング

直線やそれに準ずる線を描くジェスチャ。サークリングに比べて、ジェスチャの始点と終点が離れている。使用例を発話とともに示す。

・<ここで乗り換えて、そして、一つ、二つ目の駅で降りてください>D 。
・はい、タクシーでしたら、<近鉄線から>C <真っ直ぐ行っていただいて、ＪＲのチケット売場がここに>D。
・このように<出口を出られまして>D 。

あるオブジェクトとオブジェクトの間に線を引いたり、道などにそって線を引いたりする例が95%以上を占めた。

○ポインティング、マーキング

点を打ったり（ポインティング）、マーク（∨）をつけたり（マーキング）するジェスチャ。使用例を発話とともに示す。

・OK, (one) one is here?<>P

オブジェクトやその名前の上に点やマークをつける例が大半であった。

3.2.2　ジェスチャの使用頻度

表4は、前項で分類したそれぞれのジェスチャの、実験における使用頻度を整理したものである。

各実験とも、使用された全ジェスチャの90%近くがサークリングとドラッギングで占められていることがわかる。

表4　ジェスチャの使用頻度

	通訳者実験	WOZ実験
サークリング	17 (34%)	44 (36%)
ドラッギング	30 (60%)	64 (52%)
ポイント・マーク	3 (6%)	15 (12%)
計	50	123

4　ジェスチャが保持する情報

ジェスチャは、自身が持つ情報を発話に提供し、その結果として、発話とジェスチャの統合的な意味が定まると考えられる。本章では、ジェスチャが発話に提供する情報とはどのようなものか、ジェスチャの情報が発話にどのようにわたされるのかを考察する。

ジェスチャの情報が発話にわたされる形態には、大別して2通りあることがわかった。発話中に指示詞をおいて、ジェスチャと直接的に対応をとる方法とそうでないものである。ジェスチャをともなう発話を解析する枠組みを考える際にも、この2つの形態の差は大きいと考えられる。なぜなら、前者は、「タクシー乗り場は<ここです>C」のよ

うに、発話中に明示的に「ここ」などの指示詞が含まれているので、ジェスチャからの情報がなければ「ここ」が何を指しているのか不明で意味がとれないが、後者は、「一番近いホテルは、蹴上駅の駅前の<都ホテル>Cになりますが」のように、指示詞が含まれていないので、ジェスチャからの情報がなくても、発話文だけで意味がとれるという差があるからである。

ジェスチャをともなう発話文を、指示詞が含まれるものと含まれないものに分類すると、表5のように、ジェスチャをともなう全発話文のうち、60%以上が指示詞を含んでいることがわかった。

表5　ジェスチャをともなう発話における
指示詞を含む文／含まない文の数

	エージェント	クライアント	計
指示詞を含む文	54	23	77
指示詞を含まない文	36	7	43

そこで、ここでは、ジェスチャをともなう発話文を、指示詞を含むものと含まないものに分け、それぞれの場合における、各ジェスチャが発話に対して提供する情報とその提供の仕方を、実験結果をもとに考察する。なお、ポインティングとマーキングに関しては、使用率が低く、さらに、機能に関してはサークリングに準ずるので割愛した。

4.1　指示詞を含む発話文の場合

○サークリングが持つ情報

丸が描かれた「位置」

丸を描くことによって、丸が描かれた「位置」という情報が発話に提供される。たとえば、広い新幹線コンコースの一部を丸で囲みながら、「<あなたは今ここにいます>C」と発話する場合がこれにあたる。この場合、地図上の丸が描かれた位置 (x,y)（たとえば丸の重心）が、発話中の指示詞「ここ」に対して提供され、「あなた」が「地図上の位置 (x,y)」にいるという解釈がなされる。

実験の結果、ジェスチャからの「位置」の情報を受け取る発話中の要素としては、指示詞「ここ」、「こちら」、指示詞を含む名詞句「この場所」、「このあたり」があることがわかった。

丸で囲んだ「オブジェクト」

あるオブジェクトやその名前を丸で囲むことによって、その「オブジェクト」という情報が発話に提供される。たとえば、「<ここで降りて>C 下さい」と発話しながら三条駅を丸で囲む場合がこれにあたる。この場合、丸で囲まれた「三条駅」というオブジェクトが、発話中の指示詞「ここ」に対して提供され、「『三条駅』で降りて下さい」と解釈される。

この情報を受け取る発話中の要素としては、指示詞「ここ」、「こちら」があることがわかった。

○ドラッギングが持つ情報

線の始点、終点の「位置」

線の始点あるいは終点の「位置」の情報が発話に提供される。たとえば、「<ここからここまで行って下さい>D」と発話しながら、ある場所からある場所へ線を引く場合がそうである。この場合、線の始点の位置 (xs,ys) と終点の位置 (xe,ye) がそれぞれ最初の「ここ」と二番目の「ここ」に対して提供され、「地図上の位置 (xs,ys) から (xe,ye) まで行って下さい」と解釈される。

この情報を受け取る発話中の要素としては、指示詞「ここ」があることがわかった。

線の始点や終点の付近にある「オブジェクト」

あるオブジェクト付近から線を引き始める、あるいはあるオブジェクト付近で線を引き終えることによって、線の始点や終点の付近にある「オブジェクト」を発話に対して提供する。たとえば、「<…ここまで行って下さい>D」と発話しながら「蹴上駅」を表わすオブジェクト付近で線を引き終える場合がこれにあたる。この場合、終点付近にある「蹴上駅」というオブジェクトが、発話中の「ここ」に対して提供され、「『蹴上駅』まで行って下さい」と解釈される。

この情報を受け取る発話中の要素としては、指示詞「ここ」があることがわかった。

線の通過点付近にある「オブジェクト」

あるオブジェクト付近に線を通過させることによって、その「オブジェクト」の情報を発話に対して提供する。たとえば、「<こちらの方から出て、…>D」と発話しながら「出口」を表わすオブジェクトの上に線を通過させる場合がこれにあたる。この場合、線の通過点付近にある「出口」と

いうオブジェクトが、発話中の「こちらの方」に対して提供され、「『出口』から出て、…」と解釈される。

この情報を受け取る発話中の要素としては、指示詞「ここ」、指示詞を含む名詞句「こちらの方」があることがわかった。

線の軌跡が示す「動作の方法」

「<このように行っていただいて…>D」と発話しながら線を引く場合がこれにあたる。この場合、線の軌跡が示す「動作の方法」という情報が、発話中の「このように」に対して提供され、「『線の軌跡にそって』行っていただく」と解釈される。

この情報を受け取る発話中の要素としては、指示詞を含む副詞句「このように」があることがわかった。

4.2 指示詞を含まない発話文の場合

○サークリングが持つ情報

丸が描かれた「位置」

オブジェクトを丸で囲むことによって、丸が描かれた「位置」という情報が発話に提供される。たとえば、京阪三条駅を丸で囲みながら、「<京阪三条に着いたら>C」と発話する場合がこれにあたる。この場合、「京阪三条駅は丸で囲んだあたりにある」という補足的な情報として、地図上の丸が描かれた位置 (x,y)（たとえば丸の重心）が、発話中の固有名詞「京阪三条」に対して提供される。

指示詞を含まない発話文にともなうサークリング・ジェスチャには、位置の情報以外に発話に対して提供する情報がないことがわかった。また、この情報を受け取る発話中の要素は、例外なく、オブジェクトの名前である固有名詞であった。

○ドラッギングが持つ情報

線の始点、終点、通過点の「位置」

「<京阪三条へ行って下さい>D」と発話しながら、京阪三条駅付近で線を引き終える場合のように、線の終点の「位置」の情報が、サークリングと同様に補足的な情報として、固有名詞「京阪三条」に提供される。これに対して、上記の「京阪三条」が普通名詞の「ホーム」や「出口」などになる場合もあり、この場合は、地図上にいくつか存在する「ホーム」や「出口」のうち、ある特定のホームや出口を示すために、ジェスチャからの位置情報は必須の情報となる。

この情報を受け取る発話中の要素は、オブジェクトの名前である名詞であったが、そのうちの約80%が固有名詞であった。

線の軌跡が示す「動作の方法」

「<東地下通路をまっすぐ行っていただいて…>D」と発話しながら線を引く場合がこれにあたる。この場合、線の軌跡が示す「動作の方法」という情報が、発話中の「まっすぐ」に対して提供される。

この情報を受け取る発話中の要素としては、副詞「まっすぐ」があることがわかった。

5 発話とジェスチャの時間的関連性

前章では、ジェスチャと発話中の対応要素の意味的な関連性を考察したが、それらの対応関係を同定する際には、時間的な前後関係も重要な手がかりとなる。そこで、ここでは、各ジェスチャごとに、ジェスチャと発話中の対応要素の時間的前後関係を調査した結果を示す。

5.1 サークリング

サークリング・ジェスチャと発話中の対応要素の時間的前後関係には以下のような形態があった。

a 対応要素が発声された時点で、ジェスチャが開始される場合 (35%)
b 対応要素が発声された時点で、ジェスチャが終了する場合 (41%)
c 対応要素が発声された時点で、ジェスチャが継続している場合 (8%)
d 対応要素が発声された後に、ジェスチャが開始される場合 (6%)

発話中の対応要素とサークリング・ジェスチャの時間的前後関係をみると、ジェスチャの開始時刻、終了時刻、継続時間中に対応要素が発話される場合（a,b,c）が90%以上を占めることがわかった。したがって、サークリング・ジェスチャに関しては、ジェスチャが行われる時刻とその対応要素が発声される時刻はきわめて近いと判断できる。

5.2 ドラッギング

サークリング・ジェスチャがジェスチャ全体でひとまとまりの情報を持つのに対して、ドラッギング・ジェスチャは、線の各部（始点、終点、通過点など）や軌跡からさまざまな情報が生じている。したがって、ここでは、ドラッギング・ジェスチャの各部ごとに、ジェスチャと対応要素の時間的前後関係を調査した。

線の始点

ジェスチャが開始されると同時に、対応要素の約80%が発声された。

線の終点

ジェスチャの終了時刻と対応要素の発声時刻には特徴のある関係は見られなかった。

線の通過点

ジェスチャの線がオブジェクト付近を通過すると同時に、対応要素の約90%が発声された。

線の軌跡

ジェスチャの継続中に、対応要素の約80%が発声された。

このように、線の終点に関するジェスチャと対応要素の発声のタイミング以外は、きわめて高い時間的一致をみた。

6 発話とジェスチャの解析への指針

以上の分析から、発話とジェスチャの統合解析の枠組みを考える際には、以下の事項が特に重要であることがわかった。

・ジェスチャの種別の認識
・ジェスチャが持つ情報の表現方法
・ジェスチャと対応する発話中の要素の同定
・発話とジェスチャの時刻情報の獲得
・地図情報の構造の策定

7 むすび

マルチモーダル入力の統合的な解析の枠組みを研究するための指針を得る目的で、音声とジェスチャを入力として受け付けるマルチモーダル翻訳対話のデータを収集し、ジェスチャ入力の特徴、発話とジェスチャの意味的関連性および時間的関連性を分析した。

その結果、ジェスチャの種類、ジェスチャが保持する情報、ジェスチャが情報を提供する発話中の要素、発話とジェスチャの意味統合の仕組、発話とジェスチャの時間的前後関係といった基本的事項に関する知識を得ることができた。さらに、それをもとにして、発話とジェスチャの統合解析の枠組みを考える際の指針も抽出した。

今後は、今回得られた知識をもとに、音声とジェスチャを入力とするマルチモーダル対話システムの設計と実装を行う予定である。

参考文献

[1]Bolt,R.A. ""Put-That-There": Voice and Gesture at the Graphics Interface" Computer Graphics 14(3) 262-270 (1980)

[2]Stock,O. and the ALFRESCO Project Team "ALFRESCO:Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration" INTELLIGENT MULTI MEDIA INTERFACE, (ed.) Maybury,AAAI Press/MIT Press(1993)

[3]Vo,M.T. and Waibel,A. "Multimodal Human-computer Interaction" Proc. ISSD-93(1993)

[4]Allgayer,J.,Jansen-Winkeln,R.,et al. "Bidirectional use of knowledge in the multi-modal NL access system XTRA" Proc. IJCAI-89(1989)

[5]ローケン・キム、谷戸、森元 "マルチモーダル音声翻訳通信のためのシミュレータ" 情報処理学会オーディオビジュアル複合情報処理研究会 4-16(94.3.18)

[6]Norman M. Fraser and G. Nigel Gilbert, "Simulating Speech Systems", Academic Press Limited, 1991