Kinect を用いた会話と会話時の振る舞いの解析

An Analysis of Conversation and Gesticulation Using Kinect

馮建美 [1] *　岨野太一 [1]　今井倫太 [1]
Jianmei Feng[1]　Taichi Sono　Michita Imai[1]

[1] 慶応義塾大学
[1] Keio Univeristy

Abstract: In conversation analysis, both verbal information and non-verbal information are used to create a transcript. In this paper, we use Microsoft Kinect to record audio and video, to recognize gestures people make in group discussion. By mapping each audio to the narrator, we also recognize the interval when people are not talking or silence.

1 はじめに

対面コミュニケーションにおける会話解析では, 言語表現だけに限らず, ジェスチャを含む手の動作,「間」(silence gaps) などの非言語表現が情報伝達に重要な役割を果たしている. 会話中のジェスチャや「間」は, 話者が何かを伝えようとするために用いられたり, 発話内容の説明に役に立つ情報が含まれる. 本研究の目的は, 会話で発生するジェスチャ,「間」の分析を行うこととする.

このように会話解析をするために, 手の動作のラベリング, 音声データの記録,「間」の識別が必要となる.

マルチモーダル情報を言語情報で表す研究は盛んに行われている. 小林ら [3] は, 視覚情報として取得される人と物体の振る舞いを表す時系列データと動作を表す自然言語の説明文との対応を対数線形モデルを用いて学習し, 動作の意味を表す中間表現を判別する方法を提案している. 三吉ら [4] は光学式モーションキャプチャシステムを導入し, 赤外線カメラで被験者の身体につけられたマーカを光学的に捉え, 身体を動かす時のマーカの位置の 3 次元数値データを時間情報とともに自動的に得る会話データ収集システムを構築した. 岡田ら [5] は会話中に観測されるハンドジェスチャの機能を認識するために有効な特長量を抽出, 分析し, ハンドジェスチャの機能認識モデルの構築, 評価を行った.[5] の研究でハンドジェスチャの機能認識にジェスチャフェーズと会話参加者のマルチモーダル非言語特徴量の 2 つの特徴量を提案した.

しかしながら, 従来研究では「間」に関する解析が行われていない. 会話で発生する「間」は以下の 3 つのタイプに分類される [1].

1. 人の話を聞く時に発生する受話の休止
2. 人が発話する時に文または文の間で発生する長い発言の中止
3. 人が発話する時に語または音節間で発生する短い発言の中止

つまり,「間」であることを識別するためには, 語間と文間の区切りを見つける必要がある.

本研究では音声データ, 手の動作,「間」を Microsoft 社の Kinect を用いて認識し, トランスクリプトにとり, 会話解析を行う. 本研究は [5] の提案を利用し, ジェスチャフェーズと非言語特徴量の 2 つの特徴量で音声とジェスチャを時系列データとして捉える.

2 会話解析

2.1 会話解析の手法

会話分析という研究領域では, 会話の秩序を解明することを目的とし, 実際に発生した会話を録音・録画し, それをできるだけそのままの形で書き起こした後, 詳細な分析するが行われている. ここで, トランスクリプトという技術を用いられている. トランスクリプトとは, 現実の会話を一定のシンボルにしたがって転記した記録のことである. 十分に完全なトランスクリプトは, 現実に起きたと考えられる会話を, 会話者が少なくとも潜在的に意味あるものとして捉えられた形のままに書き表したものである [7].

2.2 会話分析における非言語情報の扱い

対面コミュニケーションを 3 つの層として捉えることが可能である [8].

*連絡先：慶応義塾大学理工学部情報工学科
横浜市港北区日吉 4－1－1
E-mail: hyokenmi@keio.jp

言語層 文法規則に沿うように並べられた単語による情報の中核部分を形成するものであり, これにより他者に情報を伝達・共有することが可能となる. 容易に言語化可能である.

パラ言語層 発話行為により, 言語層に付随して露見する層である. 発話によって生じる情報から言語層を除いた部分のことであり, 声の高さとその抑揚, 声の大きさ,「間」などが挙げられる.

非言語層 対面会話を行っている時の視覚的な情報のことである.

言語学において, 非言語動作を分離・分類する試みは Birdwhistell[2] の研究がある. しかし, 実際は言語学で行われるような分類手法のみによって, 全ての非言語行動を分類することが難しいと言われている [8]. 非言語動作が発生した会話の背景抜きでは, 非言語動作を分類できないと主張されている.Ekman と Friesen は非言語動作を構造的に分類するのではなく, その動作の目的, 意味, 意図によっていくつかの基本的な分類を行うという動作学に関しての外部変数的なアプローチが提案された. この方針が研究者の間で広く受けられている [8].

2.3 会話の「間」を分析システム構築の課題

会話の「間」を識別するために, 発話の有無の認識が必要である. 発話と発話の間の時間が「間」として捉えることができる. これを実現するために, 話者の識別と発話中と認識することが必要である. 本研究では, 会話解析を行うために, 発話音声を録音することでこの 2 つの問題を解決する.

2.3.1 時系列で 1 文語ごとの録音

自然会話の定量的分析に適するように考案された基本的な文字化の原則 (Basic Transcription System for Japanese:BTSJ) では「発話文」を基本的分析の単位としている [6]. 発話文とは, 実際の会話中で発話された文である [6]. 日本語で, スピーチレベルの解析など,「文」単位でコーティングをする必要があると思われている. 発話文の定義は, 会話という相互作用の中における文とする. 基本的に, 一人の話者による「文」を成していると捉えられるものを「1 発話文」とする. しかし, 自然言語では, 最後まで続かないいわゆる「中途終了型発話」のような構造的に「文」が完結していない発話もある. このため,「話者交替」や「間」という 2 つの要素が重要になる.

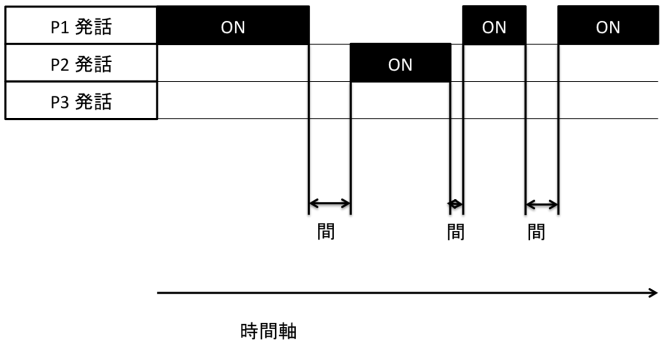

図 1: 間の認識

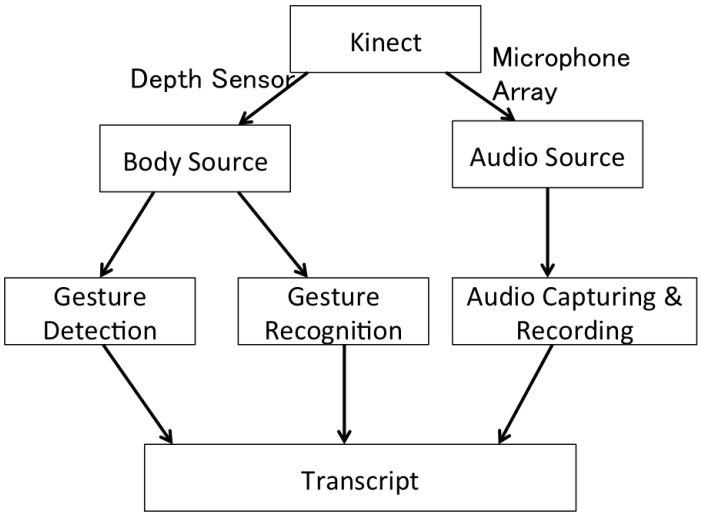

図 2: システム構成

3 「間」の認識

3.1 システムの概要とシステム構成

本実験は Microsoft 社の Kinect v2 を用いて,3 人のグループ会話を記録, 解析する. 音声, ジェスチャを時系列データとして記録する.

「間」は, 時間軸上の音声データを分析することで識別する. 図 1 で示すように, 発話者 P 1 と P 2 との話者交替で発生する時間間隔, P1 の発言の間で空いている時間間隔を「間」として捉える.

システムの構成は図 2 で示す. Kinect センサーを用いて, Body source からジェスチャの検出と認識を行う. Audio source から, 音声の有無を判定し, 録音を行う. ジェスチャの識別結果と録音した音声をトランスクリプトに出力する.

3.2 会話で発生するジェスチャの認識

本研究では, 手を振る動作, 指さし動作のデータベースを作り,Visual Gesutre Builder[9] で機械学習させる. このデータベースで判定できない動作, 例えば単純な手の移動などは, 各関節の位置座標を使用し, 手の動きの記録を取る.

3.3 音声の切り取り

会話を一文単位で切り取るために, 以下のステップを経る.

1. 発話者の識別 Kinect v2 の Depth Sensor から取得する body Source, Microphone アレイから取得する Audio source を元に, 発話者の ID を特定できる.

2. 発話区間 Mircrophone アレイから音源の推定とその信頼値 (Confidence) が与えられている. 信頼値から発話の開始と終了の判定を行う.

3. 録音 発話区間においてマイクの録音機能を使用し, 録音を行う.

3.4 「間」の可視化

認識できた「間」を GUI 上で可視化を行う. 時間軸における「間」を色つけることで表現する.

4 使用例

本システムは会話解析のために使われることを想定している. 人の手を借りずに会話の内容を記録したい場合の使用が可能となる. 可視化された「間」は, 会話解析における「間」の意味の研究に役に立つ. 前後のコンテキストと「間」の情報（例えば,「間」の長さ, 意味）をデータベースに納め, 機械学習させることにより,「間」の意味を推測することが可能となる.

5 まとめと今後の課題

本研究の目的は, 音声, ジェスチャ,「間」から会話解析を行う.「間」を解析するためには,「間」の認識と後からの意味付けが必要である. 本研究では,「間」の認識は, 音声を時間軸での分析から行う.

今後の課題として, 以下の 2 つが挙げられる.

1. 同時発話

   本システムでは, 複数の人の同時発話, つまり発話の重なりを考えていないが, 自然会話においては, 同時発話は少なくない. 今後,2 台以上の Kinect を使用し, 複数の人が話すときのおのおのの録音を考えるのが課題となる.

2. 「間」の意味の解析

   「間」の認識自体に意味がないが,「間」の意味解析が会話分析のために役に立つ. 前後のコンテキストから,「間」の意味を考えるのが課題となる.

参考文献

[1] Harold P. Stern, Samy A. Mahmoud, Kin-Kwok Wong: Modeling the On-Off Patterns in Conversational Speech, Including Short Silence Gaps and teh Effects of Interaction Between Speaking Parties, *Vehicular Technology Conference*, pp.1296-1300 vol.2, Jun 1994

[2] Ray L. Birdwhistell: Introduction to Kinescis: An annotation system for analysis of body motion and gesture, *Washington, Dpet. of State, Foreign Service Institute*, 1952

[3] 小林瑞季, 小林一郎, 麻生秀樹: Kinect により観測された人の動作を説明する確率的言語生成への取り組み, *ARG WI2*, No.3, (2013)

[4] 三吉秀夫, 関進, 綿貫啓子: マルチモーダルインタフェース, *www.sharp.co.jp/corporate/rd/journal-77/pdf/77-09.pdf*

[5] 岡田将吾, 坊農真弓, 高梨克也, 角康之, 新田克巳: 非言語マルチモーダル情報を利用したグループ対話におけるジェスチャの機能認識, 電子情報通信学会論文誌 A Vol. J98-A No.1 p63-75

[6] 宇佐美まゆみ: 改訂版:基本的な文字化の原則 (Basic Transcription System for Japanese: BTSJ) 2007 年 3 月 31 日改訂版

[7] 樫村志郎：会話分析の課題と方法 *The Japanese Jornal of Experimental Social Psychology*, Vol. 36, No. 1, 148-159, (1996)

[8] 善本淳, 水上悦雄, 山下耕二, 矢野博之: 非言語に着目した対話時のインタラクション解析ヒューマンコミュニケーション特集, 情報通信研究機構季報 Vol.53 No.3, (2007)

[9] Visual Gesture Builder (VGB), https://msdn.microsoft.com/en-us/library/dn785304.aspx