6X-06

会話のタイミングを検討するためのトランスクリプトの自動生成法の検討

馮建美† 岨野太一‡ 今井倫太†

慶應義塾大学理工学部† 慶應義塾大学大学院理工学研究科‡

1.初めに

会話解析において，会話時に発生する沈黙である「間」，視線，ジェスチャなどの非言語情報は，言語情報と共にトランスクリプト(図１)を用いて手動で書き記すのが一般的である．トランスクリプトに含まれる内容は，「間」，発話内容，視線，ジェスチャなどがある．

そこで，センサー・認識技術の発達に伴い，自然会話で発生するジェスチャや音声を自動で取得する研究が行われてきた．しかし，会話中の「間」の取得，可視化する研究は行われていない．

本研究は，会話時の「間」に注目し，対面コミュニケーションにおけるジェスチャ，「間」及び音声データを自動で取得するシステムの構築について述べる．会話時に発生する「間」に情報が含まれると主張する社会学的研究[4]があるため，会話時の「間」を捉え，可視化をすることには意味がある．

1A：そうですね，11時までが，作業で* そ] のあと15分間休憩
2F： *はい]
3A：あって*：：] そのあと45分づつ，３こうたい*で：：，いち] 時半まで
4F： *うん]
5K： *あっなるほど]
6A：°みんなおべんとう*たべるの，ね°]
7F： *ひるやすみを] 順番でとるわけですね
8A：そうしたらあとは：：(0.5) ２時40分までまた作業で，
9A：そのあと15分間また休憩で，あと４時半まで，しごと（す*る），

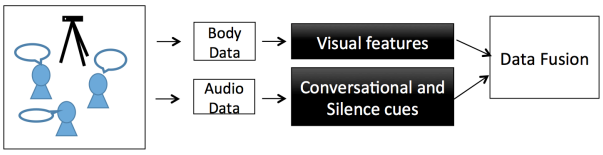
図.１トランスクリプト例[５]

2.関連研究

Yasir ら[1]の研究では音声・画像特徴で二人間の会話からソシオメトリーを取得する．彼らのシステムは，Kinect から RGB と Depth data で視覚的情報を，マイクロホンから音声情報を取得する．非言語情報を特徴として，システムに機械学習させることより１０個の社会的指標，例えば，同情，困惑，礼儀正しさを自動的に取得している．

D.S.Shete[2]らは Zero Crossing rate と音声信号のエネルギーの両方で音声の有声と無声を分離した．Zero Crossing rate というのは，信号の符号の変化率（正から負へ，またはその逆）のことである．有声部分では zero crossing rate が低く，エネルギーが高い一方で，無声部分では zero crossing rate が低く，エネルギーが低いという結果を得た．

M. Cristani[3]らは，生成構造(generagtive structure)に基づいた，会話シナリオを自動で分析するシステムを提案した．このシステムでは Gaussian mixture model と影響モデル(observed influence model)で構成される生成モデルに，低レベル聴覚のソーシャルシグナル(low level auditory social signal)を用いている．彼らの研究では，連続した沈黙・発話をスロットで表現しており，話者交代を考慮している．



図． 2 システム概要

3.自動会話解析システム

3.1 本システムの課題

本システム（図 2）では，発話と，発話時に発生するジェスチャのイベントを時系列データとして記録し，最終的に可視化を行う．本システムの課題として，以下の３つが挙げられる．

1. 発話の有無の判断
2. 発話者の特定
3. ジェスチャの認識

ここで，話者交代などによる無声区間を「間」として認識する．

A study of automatical transcript generation for speech timing analysis
† Jianmei FENG, Michiata IMAI
Faculty of Science and Technology, Keio University
‡ Taichi SONO
Graduate School of Science and Technology, Keio University

3.2 Kinect で取得するデータ

Microsoft 社の Kinect を使用し，３人で行われる日常会話の解析を行う．Kinect で Body data と Audio data を取得する．Body data で人の特定と，ジェスチャの判定を行う．Audio data で音声信号のエネルギーを計算する．また，会話全体の音声信号を wav ファイルに記録する．

3.3 ジェスチャ認識

ジェスチャ認識は Kinect SDK の Visual Gesture Builder を使用する．会話時に観測できるジェスチャを事前にビデオに撮り，アノテーションし，機械学習させることにより，ジェスチャを認識し，ラベルをつける．

3.4 発話者の識別

発話者の識別は Audio Data に付随した TrackingID と Body Data の TrackingID を照合することによって行う．同一 TrackingID を使用することによって，イベント列に記録した人の ID に一貫性を持たせることができる．

4.「間」の認識

間の認識は，音声の有無から判断する．本システムでは，10msec ベースで音声信号のエネルギーと zero crossing rate を計算する．この２つの値に閾値を設けることより，発話の有無を判定する．10msec 以上の無音区間を「間」としてみなす．

5. 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.「間」の可視化は,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）をデータベースに格納し,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.

6.まとめ

本システムは自然会話で発生する「間」，ジェスチャを解析する．音声信号のエネルギー，zero crossing rate から発話の有無を判定し，無声区間を「間」として認識する．自然会話で発生する言語情報，非言語情報を時系列データとして記録し，イベント列の可視化を行う．

参考文献

[1]Yasir Tahir, Debsubhra Chakraborty, Tomasz Maszczyk, et al., ”Real-Time Sociometrics from Audio-Visual Features for Two-Person Dialog”, IEEE 2015

[2]D.S.Shete, Prof.S.B. Patil. “Zero crossing rate and Energy of the Speech Signal of Devanagari Script”, IOSR Journal of VLSI and Signal Processing(IOSR-JVSP),=vol4, Issue 1, Ver.I (Jan.2014),PP01-05s

[3]M.Cristani, A.Pesarin, C.Drioli, A.Tavno,et al.,”Auditory Dialog Analysis and Understanding by Generative Modelling of Interactional Dynamics”,IEEE 2009

[4]Michal Ephrat,”The functions of silence”,Journal of Pragmatics 40(2008) 1909-1938,2008

[5] 樫村志郎, 会話分析の課題と方法, The Japanese Journal of Experimental, social Psychology. 1996, Vol. 36, No. 1, 148-15