2H-6

会話に付随する非言語情報の分類と評価

善本 淳
情報通信研究機構

1. はじめに

二者間の対話にて各種表出されるマルチモーダルな情報を元に,対話状態を計算機等に認識・判断させる技術は今後も必要とされる技術の１つであると考えられる.

例えば自然言語処理技術にパラ言語的な抑揚やピッチの経時変化情報等を加えて発話者の意図や意味を特定する事や,また音声情報処理技術に加え話者の発話中の首振り動作に着眼したオプティカルフロー技術等を用いて発話者の肯定/否定の意図を理解する研究は現在も行われている.

ここで上述手法とは少し異なる着眼点から,今まで放置されがちであった相手発話を促す意味の扱いを含めた非言語動作を簡易な方法で自動的に認識・分離する手法を作成する事は,現在行われている他の処理を補う目的に利用可能なため有用であると考えられる.作成途中ではあるがここにその結果を報告する.

2.方針

本報告では二者間の対話状態の認識手法を作成するために,対話中の動作と音声をそれぞれ動作の有無,発話の有無を基準にして区切り,各有動作区間,有発話区間の特徴量を元に非言語動作の分類を行った.特に音声処理では話者間での有発話区間の相対的な関係性を重視した.

この手法は発話による話題等の内容に触れていないために,対象言語を選ばない手法だと考えられる.

また一般的な他手法の処理に比べ相対的に必要とされる計算機上の記憶領域や演算量も少なくなるように, 同時に加減算はともかく除積算をなるべく減らすように留意して作成した.これは将来的に廉価な装置で演算を実行させたいという意図が報告者にあるためである.

紙面の都合上,特に backchannel 性の高い相槌のグループに関して議論する.

3.対話収録と処理

3.1.対話実験方法とその収録

マイク,ヘッドホン,ビデオカメラ,モニタが設置され,互いに相手の顔を見ながら着座してビデオチャットを行える個室を２室用意し,被験者２名をそれぞれの個室に誘導した.その個室にて被験者はビデオチャットを行いながら報告者が予め準備した１つの問題を共同で解き,その回答を合同で選択してもらった.書籍[1]から引用された問題の内容は,同じ書店で働く２人の人物写真を見て,どちらが経営者なのかを推測して当てるという２択問題(正答率: 64.6%[1])であった.問題の解き始めから回答が決定するまでの期間中, 被験者の上半身側面映像と音声は DVCAM 形式（NTSC, 29.97fps, 48KHz）により音声付動画として記録された.

3.2.動画処理

上述音声付動画情報を計算機に移動後,動画の各フレーム静止画間において輝度の差分絶対値和を移動量として算出させた.移動量が一定閾値未満の場合は静止状態として処理し,時間軸上で静止状態に挟まれた一定閾値以上の移動量が存在する部分を非言語動作が発生しているとみなして自動的に分割した.以下この分割された一連の動作１つをここでは動作チャンク κ と呼ぶ.

3.2.1.閉動作と開動作

ある動作チャンクの動作開始時刻,動作終了時刻において,それぞれの時刻における被験者画像の輝度の差分絶対値和がある一定閾値未満の動作チャンクを**閉動作**と呼び,反対にある一定閾値以上の動作チャンクを**開動作**と呼ぶ.

3.3.音声処理

被験者毎に記録された音声の処理は以下のように行われた.まず一般的な発話帯域以外の高周波及び低周波情報を除去し,次に動画フレームに合わせフレーム毎に声量の総和が算出された.後は動画処理と同様に音声量が一定閾値未満の場合は無発話状態として処理し,時間軸上で無発話状態に挟まれた一定閾値以上の音声量が存在する部分を発話しているとみなし自動的に分割した.以下この分割された一連の発話をここでは発話チャンク λ と呼ぶ.

3.3.1.発話チャンクの相対発話比の定義

被験者 A のある発話チャンク(例えば被験者 A にとって i 番目の発話チャンク $\lambda_{A,i}$)の発話開始時刻,発話終了時刻をそれぞれ $\lambda_{A,i}^{start}$, $\lambda_{A,i}^{end}$ とし,それぞれ前後に一定時間の幅 t を持たせた時,その〔 $\lambda_{A,i}^{start}-t$ 〕～〔 $\lambda_{A,i}^{end}+t$ 〕の期間に, 被験者 B が被験者 B にとって j 番目から k 番目の発話チャンク($\lambda_{B,j}$ ～ $\lambda_{B,k}$)を発生させた場合,発話チャンク $\lambda_{A,i}$ の相対発話比 $R_{A,i}^{\lambda}$ は以下の式で定義される.

Classification and evaluation of nonverbal behaviors that accompany an utterance
Jun Yoshimoto
National Institute of Communications Technology

$$R^{\lambda}_{A,i} = (\lambda^{end}_{B,j} + \cdots + \lambda^{end}_{B,k} - \lambda^{start}_{B,j} - \cdots - \lambda^{start}_{B,k}) / (\lambda^{end}_{A,i} - \lambda^{start}_{A,i})$$

上述期間中に被験者 B が無発話ならば $R^{\lambda}_{A,i}=0$ となるが,被験者 B の発話長の増加に応じて $R^{\lambda}_{A,i}$ は増加する.なお,本報告では主観的ではあるが発話の大きな区切れを想定し $t=60$[フレーム]として処理した.

2.3.2.発話チャンクの単独発話比の定義

被験者 A のある発話チャンク(例えば被験者 A にとって i 番目の発話チャンク $\lambda_{A,i}$)にて被験者 A が発話した総時間を ${}^{\lambda^{end}_{A,i}}_{\lambda^{start}_{A,i}}S_A$,同様に被験者 B が発話した総時間を ${}^{\lambda^{end}_{A,i}}_{\lambda^{start}_{A,i}}S_B$ とする時,発話チャンク $\lambda_{A,i}$ の単独発話比 $I^{\lambda}_{A,i}$ は以下の式で定義される.

$$I^{\lambda}_{A,i} = 1 - {}^{\lambda^{end}_{A,i}}_{\lambda^{start}_{A,i}}S_B \,/\, {}^{\lambda^{end}_{A,i}}_{\lambda^{start}_{A,i}}S_A$$

上述期間中に被験者 B が無発話ならば $I^{\lambda}_{A,i}=1$ となるが,被験者 A が発話しているにもかかわらず被験者 B が発話し続けた場合は $I^{\lambda}_{A,i}=0$ となる. なお,本報告では ${}^{\lambda^{end}_{A,i}}_{\lambda^{start}_{A,i}}S_A \leftarrow \lambda^{end}_{A,i} - \lambda^{start}_{A,i}$ として処理した.

4.分析

ある被験者の 4 分 11 秒の記録から発話チャンクと動作チャンクの作成を行った.孤立した 5 フレーム未満のチャンクや,同一チャンク中で最大移動量が一定閾値を超えないチャンクは排除した.このような長さと移動量の足切処理により,動作チャンクでは瞬きや,口の開閉のみ等の微少動作が排除された.このようにして 100 個の発話チャンクと 145 個の動作チャンクがそれぞれ分離された.相手被験者の発話チャンクも同様に分離を行い,二者間での相対発話比,単独発話比の算出を行った.その後 145 個の動作チャンク中,閉動作である 89 個のみを選択し,そのそれぞれに対して動作持続時間,動作期間中の平均自己相関発話比,平均他者相関発話比,平均自己単独発話比,平均他者単独発話比,動作中に発生した音量総和の 6 種の属性値を求めた.その後各属性は標準化され,その各属性値を用いてクラスター分析(UPGMA 法[2])を行った.その結果を図１に樹形図として示した.

5.結果と考察

図１で示されたグループの中から特徴的なグループ(図１最右端)の解説を以下に行う.動作チャンク#6，9，106，107，108，109，118，136 が属するグループは特別な傾向を有していると考えられる.#9 を除けばこの動作チャンクは概ね頭部において頷きの動作を発生させ,また付随して発生した対象被験者の発話は典型的な backchannel である「んー(尻下がり)」であった(表１.)

本報告において対象とした被験者は閉動作が多いという特徴があったため分類に成功したが,被験者によっては閉動作が少ない場合やあまり動作を伴わない backchannel を行う場合があり,実際には各個人に応じて各種閾値や手法を変更する必要があると思われた.例えば対象被験者の相手被験者を対象被験者と同じ閾値で処理すると動作チャンク数は 108 個であるがその内閉動作は 27 個とサンプル数が少ないため分析後の評価が困難であった.

表 1．対象被験者の動作と付随した発話内容

K #	動作長 [フレーム]	直前の相手被験者の発話	対象被験者の発話
6	20	何の根拠もないけど	んー
9	24	んー	えー
106	12	なんか人物１が	んー
107	24	店の経営者で	んー
108	23	なんか話してそれに対して	んー
109	32	人物２が	んー
118	22	経営者は結構裏，裏で	んー
136	37	今１の方は	んー

参考文献

[1] 工藤力, 市村英次, “ボディ・ランゲージ解読法”, 誠信書房, pp.204-206, (1988).

[2] 西田英郎, 佐藤嗣二, “実例クラスター分析”, 内田老鶴圃, (1992).

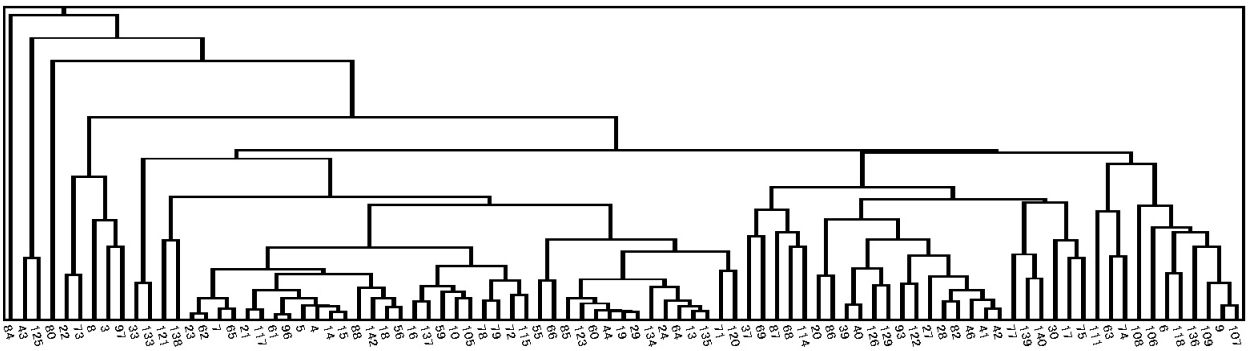

図 1．非言語動作樹形図