表情と口調に基づく本心で誉めているか否かの自動判別手法

Automatic Detection of Insincere Utterances Based on Facial Expression and Tone of Voice

見尾　和哉[*1]　目良　和也[*1]　黒澤　義明[*1]　石野　亜耶[*2]　竹澤　寿幸[*1]
MIO Kazuya　MERA Kazuya　KUROSAWA Yoshiaki　ISHINO Aya　TAKEZAWA Toshiyuki

[*1] 広島市立大学 Hiroshima City University
[*2] 広島経済大学 Hiroshima University of Economics

We propose a method to detect insincere utterances based on facial expression and tone of voice. This paper develops our previous method by changing features. In order to decrease the bad influence due to the different mouth-movement at free sentence task, the proposed method removes the features of coordinates for mouth. Furthermore, the type of acoustic feature were significantly increased from 3 to 1,027. The experimental result represented that the recall of insincere utterance praised with the participants' own expression was improved 15 point and the accuracy of the utterance by a fixed sentence was improved 5 point over the previous method.

1. はじめに

近年，実社会において人間とコミュニケーションを行う対話システムが実用化されつつある．これらの対話システムがユーザとより円滑なコミュニケーションを行うためには，ユーザの感情を理解することが重要である．しかし，ユーザが表出している感情が常に本心であるとは限らないため，発話が本心か否かというような隠された感情まで推定する技術が求められている．

そこで本研究では，褒め発話が本心であるか否かを発話者の口調と表情から判別する手法を提案する．提案手法では，機械学習には LSTM(Long short-term memory)を使用し，時系列の変化を把握したうえで判別を行う．入力特徴量としては発話中の表情と口調の情報を利用する．

2. 本心でない発話の自動検出手法

2.1 処理全体の流れ

我々はこれまで，褒め発話時の表情特徴量と音声特徴量を算出し，機械学習に LSTM を使用することで，発話が本心であるか否かを推定する手法を提案している[見尾 19]．本手法の処理の流れを図 1 に示す．

表情特徴量は，OKAO Vision [オムロン]を用いて，顔部位の特徴点（左目頭，左目尻，右目頭，右目尻，鼻左，鼻右，口上，口元左，口元右）の座標を時系列に取得している．また音声特徴量は，OpenSMILE[Eyben 10]を用いて，“音圧”，“基本周波数（F0）”，“声らしさ”を算出し，時系列に沿って取得している．表情特徴量と音響的特徴量を LSTM に適用する方法については 2.2 節で説明する．

実験に使用する発話データは，実験参加者（大学生 10 人）に本心を誘発するであろう画像と誘発しないであろう画像各 20 枚を 1 枚ずつ見せ，例え本意でなくても必ず褒めてもらうことで収集した．その際，音声をマイクで録音，表情をビデオカメラで撮影し，特徴量の抽出に利用した．これを，褒め台詞を指定する“台詞固定”と，自由な言葉で褒めてもらう“台詞自由”の 2 パターンでデータを収集した．そして褒め発話の直後に本心であるか否かを回答させ，発話に付与する感情ラベルとした[Uemura 17]．

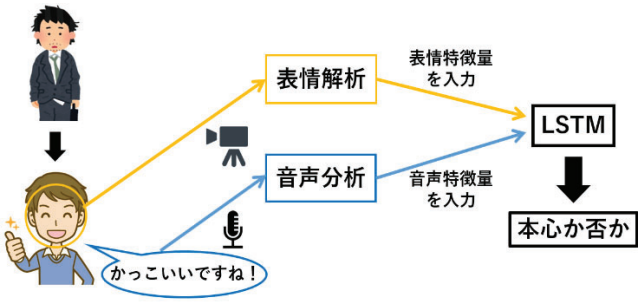

図 1　本心でない発話の自動検出手法の処理の流れ

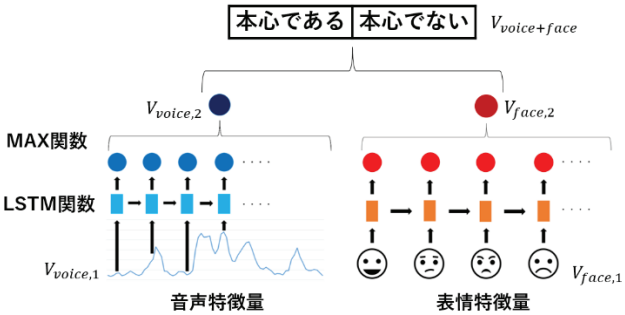

図 2　LSTM を利用した本心か否かの推定手法

2.2 LSTM 手法の処理の流れ

特徴量入力から本心か否かを推定するまでの処理の流れを図 2 に示す．図中の長方形は LSTM 関数を，丸は 100 次元の出力をそれぞれ表している．

まず，音声特徴量$v_{voice,1}$と表情特徴量$v_{face,1}$を，LSTM 関数に入力し，100 次元のベクトルに変換する．次に，MAX 関数を適用し$v_{voice,2}$，$v_{face,2}$を得る．そして連結後，線形関数を適用し，2 次元のベクトル$v_{voice+face}$へ変換する．最後に，$v_{voice+face}$の中で最も値の大きい次元に対応するラベルを予測ラベルとする．

[見尾 19]では，LSTM 手法のモデルパラメータの最適化手法に Adam，中間層の層数は 2，中間層のユニット数は 100，バッチサイズは50と設定した．実験の結果，台詞固定の正解率0.57，台詞自由の正解率 0.55 であった．本研究では，[見尾 19]の手法を改良し，より適切な手法を提案する．

連絡先：目良和也，広島市立大学大学院情報科学研究科，
mera@hiroshima-cu.ac.jp

3. 従来手法からの改良点

- 台詞自由において，表情特徴量のうち口座標（口上，口元左，口元右）を除いた特徴量を使用する．イメージ図を図 3 に示す．
- 台詞固定においては，音響分析ツール WORLD [Morise 2016]を使用し，“基本周波数 1 次元”，“スペクトル包絡 513 次元”，“非周期性指標 513 次元”の合計 1,027 個の特徴量を使用する
- 表情特徴量に関しては，OpenFace [Baltrusaitis 2016]を用いて算出できる目線や顔座標など 714 個の特徴量を使用した場合，714 個の特徴量の中から有効と考えられる特徴量のみ使用した場合で実験を行った．

4. 評価実験

[見尾 19]の手法をベースに，様々なパラメータを変更し実験を行った．評価手法は，既存手法の 5 分割交差検定を改良した評価手法を適用した．

4.1 実験条件

台詞固定において，様々なハイパーパラメータの組み合わせで予備実験を行った．ハイパーパラメータは，中間層の層数は 1 または 2，中間層のユニット数は 100，300，500，バッチサイズは 32 あるいは 50 とした．そしてこれらを組み合わせて予備実験を行った結果，最も高い正解率を得られた組み合わせは中間層=1 層，ユニット数=500，バッチサイズ=50 であった．

本研究では，台詞固定において，5 分割交差検定により生成される 5 つの学習器それぞれで loss が最小となった結果を平均した場合，5 つの学習器それぞれ異なるバリデーションデータとテストデータを使用し，バリデーションデータで正解率が最大となったエポックでテストデータをテストした結果を平均した場合，で実験を行った．それぞれの場合において，LSTM を BiLSTM（Bidirectional LSTM）に変更した場合でも実験を行った．

4.2 実験結果

(1) 台詞自由発話における口座標の影響の検証実験

台詞自由において，口座標を除いた特徴量を使用した場合(口なし)の実験結果を口座標を除かない場合（口あり）の実験結果と比較する．結果を表 1 に示す．

口座標を除くことで再現率を 15 ポイント，F 値を 7 ポイント向上させることができた．これは，自由発話による多様な口の動きが判定を困難にさせていたためであると考えられる．

(2) 台詞固定発話に対する評価実験

次に，台詞固定において以下の実験結果を比較する．

- 既存手法：[見尾 19]の実験結果．LSTM を使用
- 既存手法（Bi）：[見尾 19]の実験結果．BiLSTM を使用
- OpenFace：OpenFace で算出した特徴量を入力とする実験
- Loss：loss が最小となったエポックの学習器にテストデータを適用した結果．5 分割のため結果を平均している．LSTM を使用
- Loss（Bi）：loss が最小となったエポックの学習器にテストデータを適用した結果．5 分割のため結果を平均している．BiLSTM を使用
- Valid：バリデーションデータの正解率が最も高いエポックの学習器にテストデータを適用した結果．5 分割のため結果を平均している．LSTM を使用

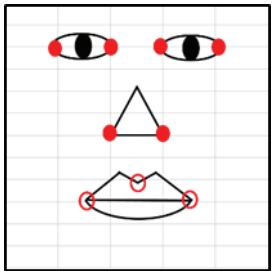

図 3：口座標（口上，口元左，口元右）を除いた特徴量

表 1：口座標あり/なし実験結果

	正解率	精度	再現率	F 値
口あり	0.55	0.51	0.38	0.44
口なし	0.54	0.50	0.53	0.51

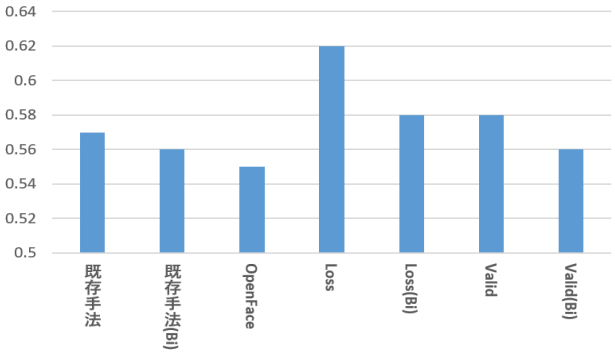

図 4：台詞固定実験の正解率

- Valid（Bi）：バリデーションデータの正解率が最も高いエポックの学習器にテストデータを適用した結果．5 分割のため結果を平均している．BiLSTM を使用

それぞれの正解率を図 4 に示す．LSTM を用いた Loss 実験において，既存手法より 5 ポイント高い正解率 0.62 を得ることができた．これは，各学習器において十分に学習できたエポックの結果に着目できたためであると考えられる．

5. まとめ

本研究では，既存研究を改良した手法を提案した．台詞自由において，表情特徴量のうち口座標（口上，口元左，口元右）を除いた特徴量を使用した．台詞固定においては，算出する音声特徴量を音響分析ツール WORLD を使用し，“基本周波数 1 次元”，“スペクトル包絡 513 次元”，“非周期性指標 513 次元”の合計 1,027 個の特徴量を使用した．実験の結果，台詞自由において，F 値を 7 ポイント向上させることができた．台詞固定においては，loss が最小となったエポックの学習器を用いる手法により，既存手法より正解率を 5 ポイント向上させることができ，本論文で提案した改良が有効であることを明らかにした．

謝辞

本研究は，国立研究開発法人科学技術振興機構（JST）の研究成果展開事業「センター・オブ・イノベーション（COI）プログラム」グラント番号 JPMJCE1311 の助成を受けたものです．また本研究を行うにあたり，オムロン（株）から openSMILE SDK をご提供いただいております．

参考文献

[Baltrusaitis 16] T. Baltrusaitis, P. Robinson, and L.-P. Morency: OpenFace: An open source facial behavior analysis toolkit, Winter Conference on Applications of Computer Vision, IEEE, 2016.

[Eyben 10] F. Eyben, M. Wöllmer, and B. Schuller: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor, Proc. of ACM Multimedia Conference – MM, pp.1459-1462, 2010.

[見尾 19] 見尾和哉, 石野亜耶, 目良和也, 竹澤寿幸:LSTM を用いた本心でない発話の自動検出, 2019 年度人工知能学会全国大会(第 33 回), 人工知能学会, 2019.

[Morise 16] M. Morise, F. Yokomori, and K. Ozawa:WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, IEICE TRANS. INF. & SYST., VOL.E99-D, NO.7, pp.1877-1884, 2016.

[オムロン] OMRON Japan, OKAO Vision | オムロン人画像センシング サイト, https://plus-sensing.omron.co.jp/technology, (2020 年 2 月 22 日アクセス).

[Uemura 17] J. Uemura, K. Mera, Y. Kurosawa, and T. Takezawa: Suppressed Negative-Emotion-Detecting Method by using Transitions in Facial Expressions and Acoustic Features, Proc. of The Second Workshop on Processing Emotions, Decisions and Opinions (EDO 2017), The 8th Language and Technology Conference (LTC), pp.122-127, 2017.