一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057-9 (11/20)

表情による感情推定と音声による感情推定手法の検討

Examination of voice-based sentiment estimation method using facial expression-based sentiment estimation

西田健次[1*]　山田亨[2]　糸山克寿[1]　中臺一博[1,3]
Kenji Nishida[1], Toru Yamada[2], Katsutoshi Itoyama[1], Kazuhiro Nakadai[1,3]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] 国立研究開発法人産業技術総合研究所
[2] National Institute of Advanced Industrial Science and Technology
[3] ホンダ・リサーチ・インスティチュート・ジャパン
[3] Honda Research Institute Japan

Abstract: 表情，声などから人間の感情推定を行う需要は年々高まってきており，特に脳卒中以後のリハビリテーションや認知症進行抑制のための認知活性化療法での介入効果推定手法に取り入れられ，その効果が確認されつつある．表情からの感情推定に関して，単純な線形識別器による表情識別器により，個人内での感情推定が有効なことを示してきた．音声情報による感情推定も需要が高まってきているが，リハビリテーションや認知活性化療法の効果推定手法としての有効性は確認されていない．本稿では，音声情報に関しても，単純な線形識別器による感情推定手法を適用し，その有効性を確認した．また，音声付き動画に表情識別による感情推定を適用し，表情検出によって音声情報への感情のアノテーションの可能性を探った．音声情報での感情推定に関しては，24 人の被験者に対して 24-way（23 人のデータで学習し，残る 1 人のデータで検証する）交差検証を行った結果，75%から 85%の汎化性能を達成することができた．また，この動画に対して，異なるデータセットで学習した表情識別器を適用したところ，喜び（笑顔）に対する検出性能が特異的に高く，笑顔検出による喜びの感情を表す音声情報へのアノテーションの有効性が示された．

1　はじめに

人間の感情推定は，昨今の人工知能技術において重要な課題であり，多くの分野でその適用が提案されている．特に，脳卒中後のリハビリテーションなどの，脳機能傷害に対するリハビリテーションにおいては，身体的なリハビリテーションに比べ，その有効性を評価する指標を得ることが難しかったが，笑顔検出による「快」の感情を推定することによる客観的な評価手法が提案され，その有効性が示されてきている [1]．また，認知症患者に対する心理療法の一つである回想法においても，従来より心理療法士の観察によって効果の評価が行われてきたが [2]，客観的な評価手法の確立が求められており，笑顔度による介入効果の評価への期待が持たれている．音声情報から感情推定を行う手法も提案されており，表情検出手法と同様にリハビリテーションへの適用が期待されている．

*連絡先：東京工業大学
152-8552 東京都目黒区大岡山 2–12–1 W8-18
E-mail: nishida@sc.e.titech.ac.jp

脳機能障害患者の感情推定手法を確立する際の大きな課題は，個人情報保護などのために実際の脳機能障害患者から収集された公開データセットの入手が困難であること，また，入手できたとしても充分なデータ数とは言えないことが挙げられる．そのため，公開されたデータセットによる感情推定器（感情識別器）を構成し，脳機能障害患者の感情推定に援用する手法が提案されている．西田らは，表情が乏しくなることが多い脳機能障害患者 [3, 4] に対して，健常者のデータセットで学習した線形識別器の推定値を用いることで乏しい表情変化に対応する手法を提案した [5]．この手法では，一人の被験者に対する感情推定器の最大値と最小値によって感情推定値を正規化することで，個人内での感情の変化を捉えることに成功している．そして，正規化された感情推定値を用いることで感情の変化を個人間でも比較できる可能性を示した．音声情報による感情推定においても，脳機能障害患者から収集され，感情に対するアノテーションの行われたデータは乏しく，表情検出による手法と同様に健常者のデー

タセットで学習した結果を援用する必要があると考えられる．また，感情推定の精度を向上するためには，より多くの脳機能障害患者からデータを収集する必要があり，データ収集を容易にするためにはアノテーションの自動化が必要となってくる．

本稿では，Ekman の基本感情 [6] にアノテーションされた音声付き動画データセットを用いて，音声情報による感情推定器を構成した．音声情報による感情推定では，24-way Cross-Validation（24 人分のデータセットに対して 23 人分のデータで学習し，残る 1 人のデータで評価を行う）において，77%から 85%の精度を得ることができ，十分な汎化性能があることが示された．動画部分に対して，[5] で述べた学習済み表情識別器による感情推定を適用し，音声情報に対して表情識別器を用いた感情のアノテーションの可否を検証した．その結果，喜びの表情（笑顔）検出精度 (precision) が充分に高く，笑顔（喜び）表情識別を用いた音声情報へのアノテーションが有効なことが示された．

2　感情推定手法

感情推定は，人工知能技術において重要な課題であり，これまで多くの研究がなされてきている．コンピュータビジョンを用いた感情推定は，基本感情を代表する表情を識別し，その強度の推定値を感情の推定値として用いている．音声情報を用いて感情推定を行うためには，基本感情を代表する音声情報から，感情の推定値をエう必要がある．本稿では，[5] と同様に，基本感情にアノテーションされたデータ（音声付き動画）を用いて線形識別器の学習を行い，閾値処理を行う前の識別器出力を感情の推定値としても用いることとした．

2.1　線形識別器による感情推定手法

Ekman は，基本感情を,「怒り (anger)」,「嫌悪 (disgust)」,「恐怖 (fear)」,「喜び (happiness)」,「悲しみ (sadness)」,「驚き (surprise)」の 6 種に類型化したが，其々の基本感情は有る無しの二値ではなく，その感情を持たない状態から，その感情を強く抱く状態までの連続値として考える方が妥当である．また，6 種の基本感情は，必ずしも排他的なものではなく，複数の基本感情が組み合わさった状態もあると考えられる．したがって，感情推定は，6 クラス（あるいは，感情的に中立を含めて 7 クラス）の多クラス識別器を構成するより，それぞれの感情強度を推定する識別器を構成する方が妥当である．その一方で，感情に対するアノテーションを行う際に，その強度まで指定するのは難しく，高い精度は期待できない．そこで，[1, 5] と同様に，学習データは感情の有無を示す 2 値のラベルを付け線形識別器の学習を行い，閾値処理前の識別器出力を感情強度の推定値として扱うこととした．表情識別による感情推定を例にとると，ある個人の感情（例えば,「喜び」）を代表する表情が検出できるとし，さらに，無表情からその感情（「喜び」）を抱く表情への変化が単調であると仮定するなら，変化の度合いをその表情の推定強度と考えることができる．そして，単調な変化を線形近似すると考えると顔画像の表情の強度は，式（1）で表すことができる．本手法は，この表情の強度を，感情推定値として用いるものである．

$$y = \boldsymbol{w}^T \boldsymbol{x} - h \tag{1}$$

ここで，$\boldsymbol{y}$ は表情の強度（スコア），$\boldsymbol{x}$ は顔画像から抽出された特徴量，$\boldsymbol{w}$ は係数ベクトル，$\boldsymbol{h}$ はバイアス値を示す．ある個人の表情の変化は $\boldsymbol{y}$ の変化によって示すことができるが，個人間での表情強度は直接比較することができないため，何らかの方法で正規化する必要がある．この正規化手法については，後述する．

2.2　感情推定のための学習・評価用データセット

音声情報による感情推定器の学習用データセットとして，The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[7] を使用した．RAVDESS は，24 人の話者 (12 人が男性，12 人がによる怒り，嫌悪，恐怖，喜び（笑顔），悲しみ，驚きの 6 種の感情に，中立 (neutral) と平静 (calm) の 2 種を加えた 8 種の感情にアノテーションされた音声付き動画に分類されており，2 種の文章（”Kids are talking by the door”と”Dogs are sitting by the door”）の読み上げを 2 回繰り返し，中立以外の感情では 2 種の強度（中立は強度は 1 種のみ）の一人当たり 60 本の動画が含まれ，総計 1440 本の動画が含まれている（図 1)．これに加え，同様の構成で 2 種の文章を，読み上げでなく歌ったデータセットも含まれているが，本稿では読み上げの部分のみを使用した．

表情識別器の学習用のデータセットとして The FaceGrabber Database and Software [8]（図 2） を使用した．FaceGrabber データベースは，40 人の怒り，嫌悪，恐怖，喜び（笑顔），悲しみ，驚きの 6 つの表情とニュートラルとされる表情に分類されており，一つの表情あたり 30 枚（ニュートラルに関しては 90 枚）の計 10800 枚の顔画像が含まれている．左右反転画像まで含めた計 21600 枚の画像を，1 表情分 2400 枚とそれ以外の表情全て（ニュートラル含む）19200 枚の 2 クラスに分け，2 クラス識別器による表情識別器種（怒り，嫌悪，恐怖，喜び，悲しみ，驚き）の訓練を行った．この表情

図 1: RAVDESS DB 顔画像の例

図 2: FaceGrabber DB 顔画像の例

識別器を RAVDESS データセットの動画に適用し，表情識別による感情の推定値とした．

2.3 音声情報による感情推定器

RAVDESS の動画に含まれる約 2 秒の読み上げ音声に対して短時間フーリエ変換を行ったスペクトログラムを特徴量として用いた．RAVDESS 動画の音声は，48,000Hz でサンプリングされていたため，8,000Hz のローパスフィルターを通した後，16,000Hz でリサンプリングを行った．リサンプリングされたデータに対して，周波数ビン 513，時間方向ビン 766 のパワースペクトログラムを生成し，特徴量とした（図 3）．このスペクトログラムを元に，怒り，嫌悪，恐怖，喜び，悲しみ，驚きの 6 種の 2 クラス線形サポートベクトルマシン（SVM）の訓練を行ったが，汎化性能を向上するため，23 人のデータで学習し，人分を未学習データとして扱うことで，24-way Cross Validation を行いソフトマージンに対するコスト係数を決定した．

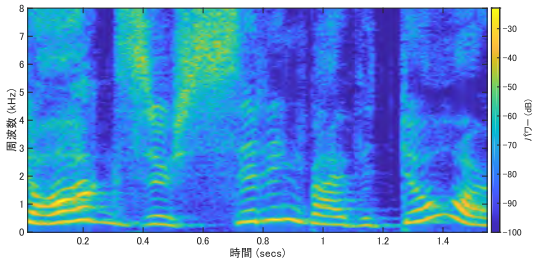

図 3: RAVDESS 動画音声のパワースペクトログラムの例

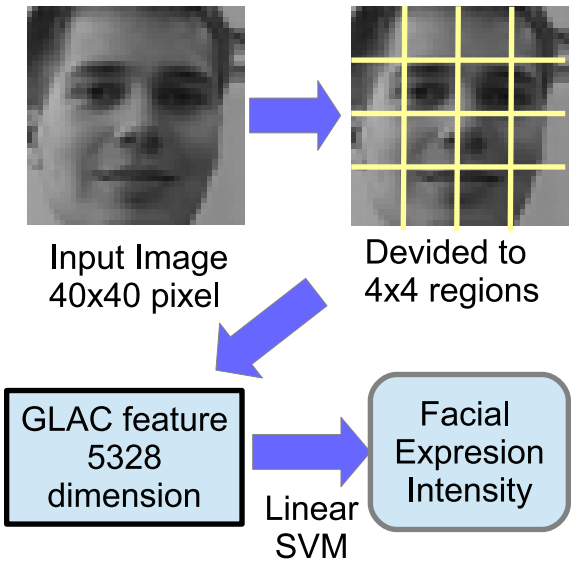

図 4: 表情識別器の構成

2.4 表情識別器による感情推定器

表情識別器は，特徴量として位置ずれや照明条件に頑健な GLAC (Gradient Local Auto-Correlation)[9] 特徴を採用した．検出された顔画像は 40×40 のグレースケール画像に変換され，4×4，計 16 個の領域に分割される．各領域について 333 次元の GLAC 特徴が抽出され，顔画像 1 枚につき 5328 次元の特徴量が $\boldsymbol{x}$ として抽出される（図 4 ）．係数ベクトル w，バイアス h は，特定の表情に対する 2 クラス線形サポート・ベクトル・マシン (SVM) を学習することによって得られる．

3 実験結果

3.1 音声情報による感情推定結果

24-way Cross Validation による平均精度を，表 1 に示す．男性，女性双方が含まれるデータセットにおい

表 1: 音声スペクトログラムによるクロスバリデーション結果

感情種別	平均精度
怒り	85.3%
嫌悪	77.5%
恐怖	81.3%
喜び	77.1%
悲しみ	74.3%
驚き	77.5%

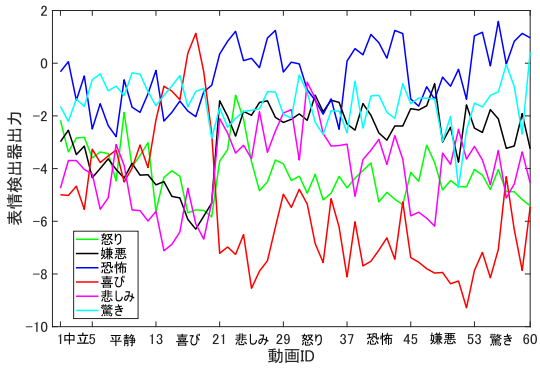

図 5: RAVDESS 話者 1 に対する表情識別器出力

て，一人分の未学習データに対して高い識別精度を達成している．これは，音声情報による感情推定器は，高い汎化性能が持つことが示唆されている．

3.2 表情識別器による感情推定結果

FaceGrabber データセットで学習した表情識別器を RAVDESS データセットに適用した結果を，図 5 に示す．動画 ID の 1 から 4 が中立（neutral），5 から 12 が平静（calm），13 から 20 が喜び，21 から 28 が悲しみ，29 から 36 が怒り，37 から 44 が恐怖，45 から 52 が嫌悪，53 から 60 が驚きにアノテーションされたもので，それぞれの線は対応する表情識別器の出力を示す．喜び識別器の出力は，喜びにアノテーションされた動画に対して高い推定値を出力しており，喜び以外の感情にアノテーションされた動画に対しては低い推定値を出力している．喜び以外の識別器は，対応する感情にアノテーションされた動画だけでなく，他の感情にアノテーションされた動画に対しても高い推定値を出力しているものが多く，必ずしも一つの感情に対する推定器とはなりえていないと考えられる．

次に，表情識別器による感情への自動アノテーションの可能性を検討した．中立と平静は基本感情に含まれていないため，怒り，嫌悪，恐怖，喜び，悲しみ，驚きの 6 種にアノテーションされた動画について検討を行った．

代表的な表情識別器，喜び識別器と怒り識別器の出力を，話者間で比較してみる．図 6 に話者 6 人分の喜び識別器の出力を示す．実際には 24 人分のデータに対して処理を行ったが，表示が煩雑になるため，6 人分だけを図示する事とした．喜び識別器出力は，喜びにアノテーションされた動画に対して高い数値を示し，それ以外の感情にアノテーションされた動画に対しては相対的に低い値を出力していることがわかる．しかし，しかし，喜び表情検出器の出力は，話者間での絶対値には差があるため，単純な閾値では喜びの感情推定とすることはできない．そこで，一人の話者に対する一つの表情識別器出力の最大値と最小値により，表情識別器の出力を正規化する．正規化は，式（2）にしたがって行った．

$$\tilde{y} = \frac{y - \min(y_e)}{\max(y_e) - \min(y_e)} \tag{2}$$

$\tilde{y}$ は識別器出力 y の正規化値，y_e は表情 $\{e|e = anger, disgust, fear, happy, neutral, sad, surprise\}$ での識別器出力を示す．

図 7 は，それぞれの喜び表情識別器の出力を最大最小値で正規化し，話者 6 人分を重ねてプロットしたものである．全ての話者で喜びの動画に対する出力が高くなっているが，その他の感情の動画に対する出力との差は，話者によって異なっている．そこで，適合率 (Precision) 最大となる閾値を求めることとした．適合率 Pr は式 (3) に従って計算される．

$$Pr = \frac{TP}{TP + FP} \tag{3}$$

TP は True Positive，FP は False Positive となったデータ数を示す．

また，再現率 (Recall) も同時に求めることとした．再現率 $Recall$ の定義を，式 (4) に示す．

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

FN は，False Negative となったデータ数を示す．

表 2 に，他の表情識別器に対しても，喜び識別器と同様の正規化処理と適合率最大の閾値を行った結果の適合率，再現率，閾値を示す．喜び表情識別器の喜び動画に対する適合率が 0.94 と，他の表情識別器に比べて高い値を示している．再現率は 0.44 と決して高くはないが，自動アノテーションを行える可能性の高い性能である．嫌悪表情識別器は，適合率は 0.79 と決して低い値ではないが，再現率が 0.10 と低い．その他の表情識別器は，適合率最大の閾値を求めたにもかかわら

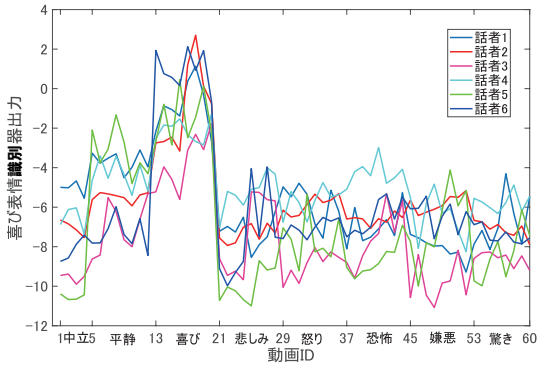

図 6: 話者 6 人に対する喜び識別器の出力

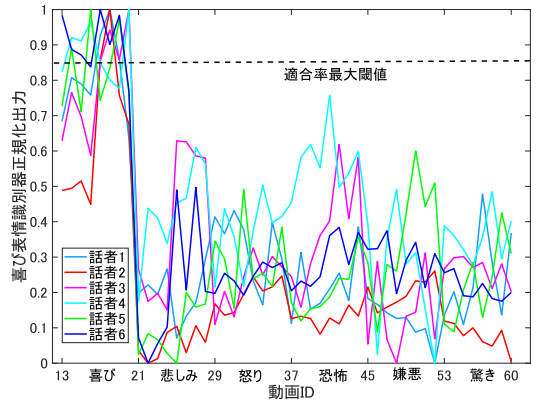

図 7: 話者 6 人の喜び表情検出器の正規化出力

ず適合率が低く，アノテーションに使用できるレベルではない．これは，[5] で確認された，喜びは他の感情に比べて特異的に判別性が高いことを再現したと考えられる．

4 結論

RAVDESS データセットを用いて，音声情報による感情推定器を構成し，汎化性の高い感情推定が可能なことを示した．また，FaceGrabber データセットで訓練した表情識別器を RAVDESS データセットの動画に適用し，表情検出による感情アノテーションの可能性を探った．その結果，喜びに関しては表情検出によるアノテーションの可能性を示すことができた．本稿での用いた音声データセットは，決められた文章を読み上げるものであり，音声の持続時間もほぼ揃っているものであり，一般的な音声情報での感情推定にはなっていない，今後は，よりバリエーションの大きな音声データセットをもちいて提案手法の有効性を検証していきたいと考えている．また，喜びの表情識別は未学習データに対しても有効なことが示されたので，これを用いて喜び表情 (笑顔) データの収集を行い，同時に音声情報に対するアノテーションの有効性を確認していきたいと考えている

表 2: 表情識別器の適合率，再現率，閾値

識別器	適合率	再現率	閾値
怒り	0.23	0.47	0.63
嫌悪	0.79	0.10	0.99
恐怖	0.58	0.18	0.94
喜び	0.94	0.44	0.85
悲しみ	0.50	0.12	0.94
驚き	0.44	0.06	0.99

謝辞

表情識別器構成手法に関して有益なご助言をいただいた産業技術総合研究所人間情報研究部門松田圭司氏，ならびに，笑顔度識別器のプロトタイプの有用性を示していただいた筑波大学人間系山中克夫准教授に感謝いたします．本研究は JSPS 科研費 20H01765 の助成を受けた．

参考文献

[1] 嶋田敬士, 山田亨, 高橋友香, 野口祥宏, 山崎郁子, 福井和広: SVM による笑顔度推定技術を用いた音楽療法効果の評価, 情報処理学会論文誌, Vol. 55, No. 12, pp. 2569–2581, (2014).

[2] 中谷淳, 山中克夫: 認知症ケアにおける回想法, 保険の科学, Vol. 48, No. 4, pp. 254–258, (2006).

[3] Borod,J. C., Koff, E., Perlman Loach, M., Nicholas, M., Welkowittz, J.: Emotional and non-emotional facial behaviour in patients with unilateral brain damages, *J. of neurology, Nuerosurgery, and Psyhiatry*, Vo. 51, pp. 826–832, (1988).

[4] Patel, S., Oishi, K., Wright, A., Sutherland-Foggio, H., Saxena, S., Shppard, S. M., Hillis, A. E.: *Frontiers in Neurology*, Vol. 9, Article 224, pp. 1–7, (2018).

[5] 西田健次, 山田亨, 糸山克寿, 中臺一博: リハビリテーション効果推定のための感情識別器の構成と

評価, 人工知能学会 AI チャレンジ研究会, SIG-Challenge-055-8, pp. 41–47, (2019).

[6] Ekman, P., Davidson, R. J. (Eds.). (1994). Series in affective science.The nature of emotion: Fundamental questions. Oxford University Press.

[7] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions, *North American English. PLoS ONE* 13(5): e0196391.

[8] D. Merget, T. Eckl, M. Schweirer, P. Tiefenbacher, and G. Rigoll, Capturing Facial Videos with Kinect 2.0: A Multithreaded Open Source Tool and Database, in *Proc. WACV*, IEEE, 2016.

[9] Kobayashi, T., Otsu, N.: Image Feathre Extraction Using Gradient Local Auto-Correlations, *European Conference on Computer Vision (ECCV)*, pp. 346–356. (2008).