Vol.2019-SLP-128 No.4
2019/7/19

音声からの感情推定における多言語補填の効果検証

坂口　巧一[1,a)]　加藤昇平[1,2]

概要：近年，Aritificial Intelligence （AI） の発展に伴い人と音声対話するロボットが注目を集めている．音声から感情を推定する技術はロボットが人と円滑な対話を実現するために重要である. 機械学習を用いて音声から感情を推定するためには大量のラベル付き感情音声データが必要となるが，大量の感情音声にラベルを付けることは難しいため，データ数に限りがある．そこで我々は，転移学習を用いることで分類したい言語とは別の言語の感情音声でデータ数を補う方法を提案している．本稿は多言語補填の効果について報告する．

1. はじめに

近年，ロボティクス技術と AI の発展に伴い，音声によって人と対話するロボットが注目を集めている．人は音声対話するときに言語情報だけでなく，声の抑揚などの非言語情報も考慮しながら対話相手の感情を推定する．そのため，ロボットが人と同様に音声で対話するには，非言語情報からも感情を推定できることが望まれる．機械学習を用いて音声から感情を推定するためには大量の感情音声データが必要となる．しかし，感情音声コーパスの作成には発話を収集し，その発話にあった感情をラベルを付与する必要があるため，コーパス作成コストが高い．よって，コーパス 1 つあたりの発話データは少数であることが多い．このような背景から感情音声データの不足を補う手法が必要と考える．

音声から感情を推定する研究は以前から行われている．有本ら [1] は手動で特徴を抽出し，判別分析による判別を行った．有本のように特徴を手動で抽出し，判別を行う研究が多く見られたが．近年では，ディープラーニングの台頭により，ニューラルネットワークに自発的に音声特徴を学習させて判別する研究も盛んに行われている [2] [3]．しかし，これらの研究は 1 つの感情音声コーパスのみを使用しているものが多い．

Ekman ら [4] は 「顔表情に表出される基本感情 は文化によらず普遍的であること」を示した．このことから，基本感情 (怒り，喜び，嫌悪，驚き，悲しみ，恐怖) は生得的であり，文化によらず普遍的な感情であると考えられる．そのため，この 6 感情については同じ感情表出の手段である音声にも普遍的な特徴が存在すると期待できる．共通の特徴が存在するならば，データ数の少ないある言語の感情音声データを，別言語のデータで補填できるのではないかと考えた．本研究では，音声からの感情推定における多言語補填の妥当性調査を試みた．

[1] 名古屋工業大学 大学院工学研究科 情報工学専攻，
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology
[2] 名古屋工業大学情報科学フロンティア研究院，
Frontier Research Institute for Information Science,Nitech
[a)] sakaguchi@katolab.nitech.ac.jp

2. 提案手法

2.1 音声データの前処理

音声データは，メルスペクトログラムに変換して後述する 1 次元畳込み双方向 Long Short Term Memory（LSTM）モデルに入力される．メルスペクトログラムとは，スペクトログラムの周波数軸をメル尺度に変換することで人の音の高さの知覚感覚に近づけたものである．スペクトログラムとは，音声データを短時間フーリエ変換（STFT）により各周波数成分強度の時間変化を表す 2 次元データに変換したものである．本稿では，STFT のサンプル数を 512，フレーム周期を 256 とし，時間長は 200 までとした．このとき，メルスペクトログラムの次元数を 80 としたため，データサイズは 80×200 である．これに z-score 正規化を施したものを，時間方向に 1 ずつ分割して 1 次元畳込み双方向 LSTM モデルに入力する．なお，感情音声データは日本語，英語の感情音声データを使用した．

2.2 1 次元畳込み双方向 LSTM

1 次元畳込み双方向 LSTM は，1 次元畳込み部と双方向 LSTM および全結合層からなるネットワークである．1 次元畳込み部は，1 次元畳込み層とプーリング層の組み合わせから構築されており，特徴の鮮鋭化と次元圧縮を行

表 1　1 次元畳込み双方向 LSTM パラメータ設定

入力層	入力サイズ:80×200 タイムステップ
畳み込み層 1	フィルタ:(4,1) × 16 活性化関数:ReLu，バッチ正規化あり
畳み込み層 2	畳み込み層 1 と同様
最大プーリング層 1	プールサイズ:2，ストライド 2 ドロップアウト率:0.25
畳み込み層 3	畳み込み層 1 と同様
最大プーリング層 2	プールサイズ:2，ストライド 2
平滑化層	タイムステップごとに平滑化
双方向 LSTM 層	出力次元:512×2 隠れ層のドロップアウト:0.5，活性化関数:tanh
全結合層 1	出力ユニット数:100 活性化関数:ReLU，ドロップアウト率:0.25
全結合層 2	活性化関数:softmax l1l2 正則化:(0.01,0.01)
出力層	出力サイズ:(N,1)

う．LSTM は，適切な過去の入力を保存することで，時間依存性の強いデータに対して効果を発揮する．一般的な LSTM は過去から未来への一方向の流れのみを考慮するが，双方向 LSTM は未来から過去への方向も考慮する．表 1 に本稿における 1 次元畳み込み双方向 LSTM のパラメータ設定を示す．出力サイズを N としたのは，評価するデータによってラベル数が異なるからである．誤差関数は categorical crossentropy を用いた．最適化アルゴリズムには Nesterov accelerated gradient（NAG）[5] を用い，学習率を 0.01，momentum 項のパラメータを 0.9 とし，1epoch ごとに学習率を 1.0e -6 ずつ減衰させた．

2.3　提案モデル

サンプル数の不足を複数のコーパスで補填する手法として転移学習を利用した．提案モデルは，評価するコーパスとは別種のコーパスを用意し，コーパスごとに別々のモデルを事前学習し，学習された複数のモデルから抽出した特徴を入力とするモデルを学習し，このモデルの出力を最終出力とする．コーパスが違うと話者，発話環境，言語等が異なるが，同一の感情識別というタスクについて学習しているため，関連した特徴が得られると考えられる．そのため，このように複数のモデルで学習した特徴から総合的に判断することで汎化性能の向上が期待できる．ここでは提案モデルにコーパス A の訓練データ，コーパス B の全データを学習させて，コーパス A のテストデータに対する性能を評価する場合 (転移学習モデル (B)) を例に説明する．図 1 に転移学習モデル (B) における感情推定時の流れを示す．

事前学習部には，1 次元畳み込み双方向 LSTM を用いる．各データ (コーパス A の訓練データ，コーパス B の全データ) でそれぞれ学習を行い，2 種類の学習済モデルを作成する．このとき，コーパス B のデータはコーパス A に共通で存在する感情のデータのみ使用した．なお，学習エポック数は 100 回とした．学習終了後，学習済モデルの最終層を取り除き，入力データから 100 次元の特徴を抽出する特徴抽出器とした．

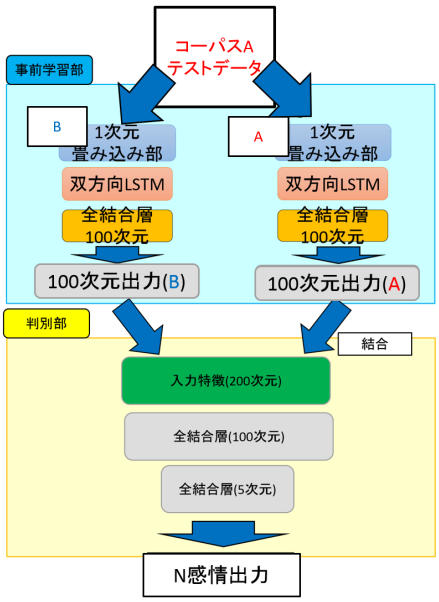

図 1　転移学習モデル (B) 概要

表 2　判別部パラメータ設定

入力層	入力サイズ:100 × S(特徴抽出器の数)
全結合層 1	出力ユニット数:100 活性化関数:ReLU
全結合層 2	活性化関数:softmax l1l2 正則化:(0.01,0.01)
出力層	出力サイズ:(N,1)

作成した複数の特徴抽出器の出力を入力として，全結合層 2 層からなる判別部を学習する．このとき，判別部の学習には，学習済みモデル A の学習に用いたものと同様のデータを用いる．誤差関数と最適化関数については 1 次元畳込み双方向 LSTM モデルと同様である．

3.　実験データ

図 3 に実験に使用したデータ数一覧を示す．日本語の感情音声データには感情評定値付きオンラインゲーム音声チャットコーパス (OGVC)[6] の演技発話コーパスと自発話コーパスを，英語の感情音声データとしては RAVEDESS[7] と IEMOCAP[8] を使用した．これらのコーパスのうち，OGVC の演技発話コーパスと RAVEDESS は補填用のデータとして用いた．

3.1　OGVC の自発話コーパス

OGVC の自発話コーパス (以下 OGnat) はオンラインゲームの音声チャット対話に基づく対話により感情を喚起させた自然発話コーパスである．このコーパスは 11 名の大学生 (男性 8 名，女性 3 名) の計 6578 発話に 3 名の評価者が 10 種類の感情評価 (喜び，受容，恐れ，驚き，悲しみ，嫌悪，怒り，期待，平静，その他) を付与してある．使用した発話は，感情評価が 3 名中 2 名以上一致した 5 感情 (怒

表 3　各コーパスにおける感情音声データの内訳

	OGVC 自然発話コーパス	OGVC 演技発話コーパス	IEMOCAP	RAVEDESS
怒り	237	240	1103	192
悲しみ	243	252	1084	192
喜び	595	252	595	192
嫌悪	335	240	-	192
驚き	565	288	-	192
合計	1975	1272	2782	960

り，悲しみ，喜び，嫌悪，驚き) のラベルが付いた発話である．

3.2 OGVC の演技発話コーパス

OGVC の演技発話コーパス (以下 OGact) は，オンラインゲームの音声チャットの感情発話を，4 名の俳優 (男性 2 人，女性 2 人) が対話形式で 8 感情 (受容，怒り，期待，嫌悪，恐れ，喜び，悲しみ，驚き) で演じた音声コーパスである．それぞれの感情音声には感情とその強度 (0-1-2-3) のラベルがつけられており，強度 0 は平静状態を表す．このうち 5 感情 (怒り，悲しみ，喜び，嫌悪，驚き) のラベルが付いた発話を使用した．

3.3 IEMOCAP

IEMOCAP は 10 名の俳優 (男性 5 名，女性 5 名) が対話形式でシナリオにそって感情を表現したコーパスであり，3 名の評価者が 10 種類の感情評価 (怒り，嫌悪，興奮，恐怖，不満，喜び，平静，悲しみ，驚き，その他) を付与している．感情評価が 3 名中 2 名以上一致したものをその発話の感情ラベルとし，このうち 3 感情 (怒り，喜び，悲しみ) のラベルが付いた発話を使用した．他のコーパスと違い 3 感情のみを使用したのは，嫌悪と驚きのデータ数が 2，107 と極端に少なかったためである．

3.4 RAVEDESS

RAVEDESS(以下 RAV) は 24 人の役者 (男性 12 人，女性 12 人) が発話と歌で 8 感情 (平静，落ち着いた，喜び，悲しみ，怒り，恐怖，嫌悪，驚き) を演じたコーパスである．今回は 5 感情 (怒り，悲しみ，喜び，嫌悪，驚き) のラベルが付いた発話を使用した．

4. 複数言語コーパスによる学習データ補填実験

4.1 実験方法

提案モデルである転移学習モデルの事前学習，および比較手法である単一コーパスの感情音声のみを学習するモデル（単一モデル）には 1 次元畳込み双方向 LSTM を用いた．本実験では，単一学習モデルと転移学習モデルの性能を OGnat と IEMOCAP でそれぞれ比較する．IEMOCAP の感情を推定する場合は以下の 4 つのモデルを比較する．

- 単一学習モデル
- 転移学習モデル (RAV)
- 転移学習モデル (OGact)
- 転移学習モデル (OGactRAV)

5 分割交差検証を行い，各感情ごとに再現率，適合率を算出する．再現率はあるラベルが振られたデータのうち，正しくそのラベルと分類できた割合を表す．適合率はあるラベルと予測したデータのうち，実際にそのラベルが正解である割合を表す．このとき，訓練データの 2 割を検証データとする．学習エポック数は 100 回とした．OGnat の場合は 5 感情分類，IEMOCAP の場合は 3 感情分類を行う．

4.2 実験結果

表 4 に OGnat の推定結果を，表 5 に IEMOCAP の推定結果を示す．各感情ごとの適合率，再現率それぞれにおいて，4 モデルのうち最も結果が良いものを太字で表記した．

まず，OGnat の分類結果について比較する．「嫌悪」と平均に関しては再現率，適合率の両方において，単一モデルよりも転移学習モデル全体の方が高い結果となった．「喜び」においては，転移学習を行うことでわずかに再現率が低下したモデルがあるものの，適合率は全てのモデルで上昇がみられた．「怒り」に関しては転移学習モデル (RAV) において適合率の減少がみられたものの，再現率については全ての転移学習モデルで上昇がみられた．「怒り」はデータ数が少なく，単一モデルでは再現率が低い感情である．そのため，提案モデルを用いて少ない「怒り」のデータを補填することにより，「怒り」に関する特徴をより捉える学習を行うことができたと考えられる．また，別言語による補填効果の検証として，単一モデルと転移学習モデル (RAV) を比較すると，転移学習モデル全体で改善された項目に加え，「悲しみ」について再現率と適合率が上昇した．このことから，別言語の感情音声から学習した特徴も，推定性能の向上に寄与できることが示唆された．一方，転移学習モデル内で比較すると，転移学習モデル (OGactRAV) は複数の学習済みモデルを用いたモデルであるのに関わらず，「驚き」以外は再現率と適合率の両方で他のモデルより向上したものはなかった．

次に，IEMOCAP の判別結果について考察する．単一モデルと転移学習モデル全体を比較すると，「怒り」において再現率と適合率の両方で単一モデル以上の性能が得られた．IEMOCAP の単一モデルの判別結果とデータ数を見ると (表 3，表 5)，「喜び」は全体に占める発話数が少なく他の 2 感情に誤分類されやすいため，再現率が低下した．「喜び」の再現率を単一モデルと転移学習モデルについて比較すると，全ての転移学習モデルにおいて単一モデルよりも再現率が減少した．このことから「喜び」を「怒り」「悲しみ」と誤推定する傾向が強くなったと考えられる．また，複数の学習済みモデルを用いた転移学習モデル (OGactRAV)

表 4　OGnat における単一モデルと転移学習モデルの比較

	再現率				適合率			
	単一	転移学習モデル			単一	転移学習モデル		
	モデル	OGact	RAV	OGact RAV	モデル	OGact	RAV	OGact RAV
嫌悪	0.400	**0.412**	0.406	0.409	0.445	0.455	**0.466**	0.452
喜び	0.659	0.657	**0.662**	0.657	0.613	0.622	**0.623**	0.621
悲しみ	0.461	0.453	**0.469**	0.457	0.459	0.460	0.465	**0.468**
怒り	0.342	**0.376**	0.359	0.363	0.365	**0.379**	0.359	0.368
驚き	0.676	0.673	0.674	**0.680**	**0.673**	0.668	0.670	**0.673**
平均	0.508	**0.514**	**0.514**	0.513	0.511	**0.517**	**0.517**	0.516

表 5　IEMOCAP における単一モデルと転移学習モデルの比較

	再現率				適合率			
	単一	転移学習モデル			単一	転移学習モデル		
	モデル	OGact	RAV	OGact RAV	モデル	OGact	RAV	OGact RAV
怒り	0.845	0.851	**0.853**	0.850	0.795	**0.800**	0.797	0.795
喜び	**0.410**	0.380	0.382	0.375	0.510	0.528	**0.530**	0.517
悲しみ	0.817	**0.841**	0.839	0.836	**0.783**	0.773	0.775	0.773
平均	**0.691**	**0.691**	**0.691**	0.687	0.696	0.700	**0.701**	0.695

は他のモデル以上の結果となる項目がなかった．

5. まとめ

本研究では，音声には言語文化によらない共通の特徴があり，ある言語の感情音声のデータの不足は別言語の感情音声で補填できるという考えのもと，感情推定を行う言語とは別の言語について学習したネットワークを転移学習で利用することを提案した．

別言語の感情音声による補填効果の検証方法として，単一モデル，転移学習モデル (同言語)，転移学習モデル (別言語)，転移学習モデル (同言語と別言語) の 4 つのモデルの性能を，IEMOCAP と OGnat でそれぞれ比較した．その結果，提案手法により，IEMOCAP では「怒り」の再現率と適合率について，また OGact では「怒り」の再現率と「嫌悪」「喜び」「悲しみ」の再現率と適合率について上昇がみられた．このことから，たとえ別言語の感情音声について学習したモデルであっても，分類器の性能の向上に寄与できる可能性が示唆された．一方，どちらの実験においても，学習済みモデルを複数用いた転移学習モデルよりも，1 つの学習済みモデルを用いた転移学習モデルの方が良い項目が多かった．そのため，今後は複数の学習済みモデルを利用する方法を今後改善していく必要があると考えられる．また，OGnat ではデータの少ない「怒り」の再現率が提案モデルで改善が見られたが，IEMOCAP においてはデータ数の少ない「喜び」の再現率が減少した．今後，割合の少ないデータをアップサンプリング等で補うことも検討する予定である．また，複数のコーパスの学習済みモデルを用いて学習を行う本モデルは，1 つのコーパスを用いる場合に比べて多くの話者の感情について学習している．そのため，未知の話者に対して頑健なモデルとなることが期待できる．今後は未知話者の感情の識別性能についても検証していく．

参考文献

[1] 有本泰子ら," 感情音声のコーパス構築と音響的特徴の分析" 情報処理学会研究報告音楽情報科学（MUS）,pp.133-138,2008
[2] Dario Bertero et al,"Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems" in Proceedings of the 2016 Conference on Empirical Meth-Ods in Natural Language Processing“, pp. 1042-1047
[3] George Trigeorgis et al,“Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network" in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5200-5204.
[4] Ekman, P. and Friesen, W. V." Unmasking the Face,Prentice-Hall" ,1975
[5] Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). Doklady ANSSSR (translated as Soviet.Math.Docl.), vol. 269, pp. 543-547.
[6] 有本泰子, 河津宏美, “音声チャットを利用したオンラインゲーム感情音声コーパス”，日本音響学会 2013 年秋季研究発表会講演論文集, 1-P-46a, pp. 385-388, 2013.
[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 in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.
[8] C. Busso et al, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.