The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

音声からの感情推定における転移学習を用いた多言語補填

Multilingual Imputation Using Transfer Learning for Estimating Emotion from Speech

坂口巧一 *1　　加藤昇平 *1*2
Koichi Sakaguchi　　Shohei Kato

*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, Nagoya Institute of Technology

Recently, vocal communication robots attract people thanks to development of AI and robot engineering. The technology of estimating emotion from speech is important to realize a smooth dialog between human and robots. This technology needs a large number of emotional speech data, but it is difficult to collect such data a lot. We investigated the effectiveness of multilingual imputation by transfer learning using 1D convolutional bidirectional LSTM. In this paper, we report the result. The result is suggested that increasing the number of languages of emotional speech learned may exceed the performance of the model learned insufficient emotional speech in single language.

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，そのような情報からも感情を推定できることが望まれる．音声から感情を推定するには大量の感情音声サンプルが必要である．しかし，感情音声を大量に集めることは難しい．そのため，サンプル不足を補う手法が必要と考えられる．

音声から感情を推定する研究は以前から行われている．かつては何らかの特徴抽出アルゴリズムによって複数の音声特徴量を抽出し，Support Vector Machine（SVM）などに学習させて判別する手法が多かった [有本 08]．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るようになってきた [Dario 16] [George 16]．しかし，複数の言語の音声の感情について推定を行ったり，異言語間の感情の共通性について考察している研究は少ない．

Ekman[Ekman 75] が「基本な顔表情 は文化によらず普遍的であること」を示したことから，音声にも文化によらず共通の特徴が存在することが期待できる．共通の特徴が存在するならば，サンプル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，ディープラーニングにおける多言語補完の妥当性調査を試みた．

2. 提案手法

2.1 音声データの前処理

音声データは，スペクトログラムに変換して 1 次元畳込み双方向 Long Short Term Memory（LSTM）モデルに入力される．スペクトログラムとは音声データに短時間フーリエ変換（STFT）を行い，各周波数成分強度の時間変化を表す 2 次元データである．STFT のサンプル数を 882，フレーム周期を 441 とし，時間長は 200 までとした．スペクトログラムのサイズは N×200 とした．周波数方向を N としたのは，次のセクションで示す実験で周波数帯域 5kHz 上限と全領域 (22.05kHz) の場合で比較を行うためである．5kHz 上限の場合は N=101，全領域の場合は N=442 となる．これに z-score 正規化を施したものを，1 タイムステップ毎に分割して 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，韓国語，アメリカ英語の感情音声の 5 感情「怒り」「悲しみ」「喜び」「嫌悪」「驚き」を使用した．

2.2 1 次元畳込み双方向 LSTM

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言語の感情音声のみを学習するモデル（単一言語モデル）には 1 次元畳込み双方向 LSTM を用いた．

1 次元畳込み双方向 LSTM は，1 次元畳込み部と双方向 LSTM，全結合層の 3 つからなる．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なる部分であり，特徴の鋭敏化と次元圧縮を行う．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．具体的なパラメータは表 1 のようになる．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)[Nesterov 83] を用い，学習率を 0.01 とし，epoch ごとに学習率を $1e^{-6}$ ずつ減衰し，momentum 項のパラメータを 0.9 とした．

2.3 転移学習モデル

ある言語の感情音声データを別の言語の感情音声データで補填する手法として転移学習を利用した．このモデルは，特徴抽出部と判別部の 2 つからなる．

特徴抽出部には，1 次元畳み込み双方向 LSTM を用いる．各言語の感情音声で学習を行い，学習済モデル (日)，(韓)，(英) を作成する．データを 8:2 の割合で学習データとテストデータに分け，学習データのうち 2 割を検証用データとする学習を

連絡先: 坂口巧一，名古屋工業大学，〒 466-8555 名古屋市昭和区御器所町，sakaguchi@katolab.nitech.ac.jp

表 1: 1 次元畳込み双方向 LSTM 詳細

パラメータ設定	
入力層	入力サイズ:N×200 タイムステップ 上限 22.05kHz:N=442，5kHz:N=101
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:(5,1)

表 2: 判別部パラメータ詳細

パラメータ設定	
入力層	入力サイズ:100 × N
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:(5,1)

行った．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，入力データを 100 次元の特徴を抽出する特徴抽出器 (日)，(韓)，および (英) を作成した．

作成した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．なお，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． 例として，図 1 に日本語の感情音声を転移学習モデル (英韓) で学習する場合を示す．特徴抽出器 (英)，(韓) に日本語感情音声を入力して 200 次元の特徴に変換し，その特徴を入力として学習をする．

3. 実験データ

実験に使用したデータ数一覧は図 3 のようになる．日本語，韓国語，北アメリカ英語の 3ヵ国の感情音声発話を用意し，5 感情「怒り」「悲しみ」「喜び」「嫌悪」「驚き」を実験データとして使用した．なお，サンプリング周波数は日本語と韓国語の感情音声データは CD 規格の 44.1kHz であるが，北アメリカ英語のみ 48kHz である．そのため，北アメリカ英語のデータにサンプリング周波数変換を施し，44.1kHz に変換して使用した．

3.1 日本語感情音声データ

日本語の感情音声データとして，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)[有本 13] を使用した．これ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 9 感情 (受容，怒り，期待，嫌悪，恐れ，喜び，悲しみ，驚き，平静) で演じた音声コーパスである．

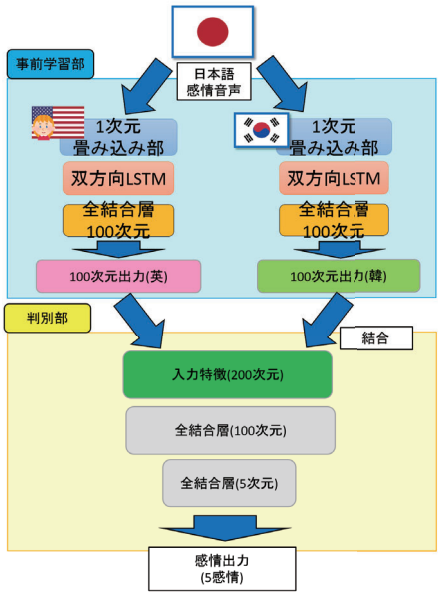

図 1: 転移学習モデル (英韓) 概要

表 3: 実験データ

	日本語	韓国語	北米英語
怒り	240	100	192
悲しみ	252	100	192
喜び	252	100	192
嫌悪	240	100	192
驚き	288	100	192
合計	1272	500	960

3.2 韓国語感情音声データ

韓国語の感情音声データとして Cho ら [Cho 09] が用いた感情音声データを使用した．これは，韓国の TV ドラマから韓国人の俳優が発した感情音声フレーズを抽出し，聴取者 2 名により 5 感情 (怒り，悲しみ，喜び，驚き，嫌悪) に分類したデータである．

3.3 北アメリカ感情音声データ

北アメリカ英語の感情音声コーパスとして，RAVEDESS[RAVEDESS 18] を使用した．これは 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．今回は発話データのみを用いた．

4. 周波数帯域による分類性能比較実験

別言語の学習済モデルを転移学習することによるデータ不足の補完の予備実験として，使用する周波数帯域による分類性能の比較実験を行った．サンプリング周波数に CD 規格の 44.1kHz を採用しため，入力として 0〜22.05kHz までの周波数のスペクトログラムを用いることができる．しかし，人の可聴範囲は一般的に 20〜20kHz といわれており，実際に感情の特徴となる周波数成分は更に狭い範囲に分布すると予想される．そのため，入力として周波数の上限を適切に定めることが必要であるのではないかと考えられる．

表 4: 韓国語の分類結果 (F 値)

	全帯域	5kHz 上限
喜び	0.346	**0.347**
悲しみ	**0.505**	0.498
怒り	**0.581**	0.571
嫌悪	0.508	**0.524**
驚き	0.367	**0.392**
平均	**0.469**	0.468

表 5: アメリカ英語の分類結果 (F 値)

	全帯域	5kHz 上限
喜び	0.483	**0.501**
悲しみ	0.564	**0.569**
怒り	0.501	**0.598**
嫌悪	0.507	**0.523**
驚き	0.523	**0.586**
平均	0.517	**0.558**

4.1 実験方法

本実験では，入力データであるスペクトログラムの周波数帯域を 5kHz 上限にした場合と 22.05kHz(全帯域) の場合で各言語ごとに判別性能を比較する．本実験では 2.2 の 1 次元畳込み双方向 LSTM モデルを使用した．5 分割交差検証を行い，各感情ごとに F 値を算出する．このとき，訓練データの 2 割を検証データとする．なお，学習エポック数は 100 回とした．

4.2 実験結果

結果を表 4，表 5，表 6 に示す．比較して高い方を太字で表記した．

まず，韓国語の感情音声について分類した結果を比較すると (表 4),「悲しみ」「怒り」については 0〜22.05kHz の周波数帯域のデータを入力とした場合のが F 値が良い結果になった．しかし,「喜び」「嫌悪」「驚き」については 5kHz 上限で区切ったスペクトログラムを入力した場合の方が良い結果となった．

次に，アメリカ英語の感情音声について分類した結果を比較すると (表 5),「喜び」「悲しみ」「怒り」「嫌悪」「驚き」全てにおいて 5kHz 上限で区切った場合の方が良くなった．特に「怒り」については 0〜22.05kHz の周波数帯域のデータを入力した場合に比べて F 値が約 0.10 改善された．

最後に，日本語の感情音声について分類した結果を比較すると (表 6),「喜び」「悲しみ」については 0〜22.05kHz の周波数帯域のデータを入力とした場合の方が良い結果となった．しかし,「怒り」「嫌悪」「驚き」については 5kHz 上限で区切った場合の方が良い結果となった．特に「怒り」については F 値が約 0.05 改善された．

4.3 考察

全体的に，5kHz 上限でスペクトログラムを区切った場合の方が良い結果になった．この要因として 5kHz 上限で入力を区切ったことにより，入力次元が 4 分の 1 以下まで削減されたことが大きいと思われる．入力データのサイズが小さいほど情報が減ってしまうのは確かであるが，その分モデルの規模が小さくなり，求めるべきパラメータが減少する．これにより学習に必要となる学習データの規模が小さくなる．特に本実験のようにデータ数が少ない場合は次元削減によるメリットが大きかったと考えられる．また，感情分類に深くかかわる周波数帯域を削除してしまうと逆に性能が悪化することが考えられる．しかし，本実験の結果を見てみると全体的には性能の向上が見られたことから，感情分類に必要となる周波数の大部分は 5kHz 以内に分布すると推測される．

表 6: 日本語の分類結果 (F 値)

	全帯域	5kHz 上限
喜び	**0.444**	0.423
悲しみ	**0.426**	0.391
怒り	0.279	**0.327**
嫌悪	0.432	**0.472**
驚き	0.560	**0.595**
平均	0.429	**0.443**

5. 転移学習による学習データ補填実験

5.1 実験方法

本実験では，単一言語モデルと転移学習モデルの性能を各言語ごとに比較する．ある言語 A の感情を推定する場合 (A 以外の言語を B，C とする) は以下の 4 つのモデルを比較する．なお，スペクトログラムは 5kHz 上限のサイズ 101×200 のものを使用する．

- 単一言語モデル
- 転移学習モデル (言語 B)
- 転移学習モデル (言語 C)
- 転移学習モデル (言語 BC)

5 分割交差検証を行い，各感情ごとに F 値を算出する．このとき，訓練データの 2 割を検証データとする．なお，学習エポック数は 100 回とした．

5.2 実験結果

実験結果は表 7，表 8，表 9 のようになった．全体で最も結果が良いものを太字，転移学習モデルの中で最も結果が良いものを赤字で示した．

まず，韓国語の感情音声について分類した結果を比較すると (表 7)，全体的に韓国語の単一言語モデルが最も良い結果となった．転移学習モデル 3 つについて比較すると,「悲しみ」「怒り」「驚き」については，転移学習モデル (英日) が最も良い結果が得られた．また，F 値の平均を比較と，転移学習モデル (英日) が最も良い結果であった．

次にアメリカ英語の感情音声について分類した結果を比較すると (表 8)，全体的にアメリカ英語の単一言語モデルが最も良い結果となった．特に「喜び」「悲しみ」「怒り」については単一言語モデルと転移学習モデルの間に 0.20 以上の F 値の差が見られた．転移学習モデル 3 つを比較すると,「喜び」「悲しみ」「嫌悪」については転移学習モデル (日韓) が最も良い結果が得られた．また，F 値の平均を比較すると，転移学習モデル (日韓) が最も良い結果であった．

最後に日本語の感情音声について分類した結果を比較すると (表 9)，全体的には日本語の単一言語モデルが最も良い結果となった．特に「怒り」については顕著であり，単一言語モデルが転移学習モデルよりも約 0.20 高い F 値となった.「喜び」については転移学習モデル (英韓) が最も良い結果になった．転移学習モデル 3 つについて比較すると,「喜び」「怒り」「嫌悪」

表 7: 韓国語における単一言語モデルと転移学習モデル (英), (日), (英日) の比較 (F 値)

	単一言語モデル	転移学習モデル (英)	(日)	(英日)
喜び	**0.347**	0.163	0.138	0.144
悲しみ	**0.498**	0.344	0.284	0.380
怒り	**0.571**	0.484	0.495	0.540
嫌悪	**0.524**	0.384	0.354	0.374
驚き	**0.392**	0.350	0.374	0.384
平均	**0.468**	0.346	0.342	0.366

表 8: アメリカ英語における単一言語モデルと転移学習モデル (日), (韓), (日韓) の比較 (F 値)

	単一言語モデル	転移学習モデル (日)	(韓)	(日韓)
喜び	**0.501**	0.296	0.247	0.301
悲しみ	**0.569**	0.291	0.274	0.342
怒り	**0.598**	0.311	0.384	0.378
嫌悪	**0.523**	0.323	0.366	0.416
驚き	**0.586**	0.412	0.369	0.392
平均	**0.558**	0.330	0.332	0.368

「驚き」については，転移学習モデル (英韓) が他の転移学習モデル以上の F 値であった．また，F 値の平均を比較すると，転移学習モデル (英韓) が最も良い結果であった．

5.3 考察

単一言語モデルと転移学習モデルの結果を比較すると，全体的に転移学習モデルの方が悪い結果になった．2 言語転移学習モデルと 1 言語転移学習モデルの分類結果の比較では，どの言語も 5 感情中 3 感情以上で 2 言語転移学習モデルの方が良い結果が得られた．今後，学習する感情音声の言語数を増やすことで，データ数が不十分な単一言語の感情音声で学習した判別器以上の性能が得られるのではないかと考えられる．

6. まとめ

本研究では，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を別言語の感情音声で補完できるのではないかと考え，別の言語について学習したモデルを転移学習で利用することを提案した．

まず，スペクトログラムの周波数帯域を適切なところで制限した方が分類性能が向上するのではないかと考え，上限 5kHz で区切った場合と，サンプリング周波数の半分 22.05kHz までを入力とした場合で言語ごとに学習を行い，性能の比較実験を行った．その結果，日本語と韓国語は 5 感情中 3 感情が，アメリカ英語は 5 感情全ての感情について，区切った場合の方が高い F 値が得られた．よって，5kHz までの範囲に感情に関連する周波数が多く分布するのではないかと考えられる．

次に，ある言語の感情音声のデータ不足を別の言語の感情音声を学習したモデルを転用することで補う方法について検討を行った．結果としては，日本語の「喜び」以外は 1 から学習したモデルが分類性能において最も良い結果であった．しかし，1 言語転移学習モデルよりも 2 言語転移学習モデルの方が全体的に良い結果が得られたことから，学習する言語数を増やすことで，不十分なデータ量の単一言語モデルよりも高い性能が得られる可能性が示唆された．

表 9: 日本語における単一言語モデルと転移学習モデル (英), (韓), (英韓) の比較 (F 値)

	単一言語モデル	転移学習モデル (英)	(韓)	(英韓)
喜び	0.423	0.412	0.344	**0.437**
悲しみ	**0.391**	0.362	0.312	0.307
怒り	**0.327**	0.102	0.070	0.128
嫌悪	**0.472**	0.374	0.340	0.428
驚き	**0.595**	0.519	0.505	0.519
平均	**0.443**	0.365	0.334	0.378

これらの結果を踏まえ，今後は更に多くの言語の感情音声を入手し，本実験で示された可能性について検証していく予定である．また，言語数を増加させていくと，本手法では次元数が増加してしまう．対策として，出力次元を減らすことや，文化的背景が近い言語ごとに系統分けすることなどを検討していく予定である．

参考文献

[Dario 16] Dario Bertero et al,”Real-Time Speech Emotion and Sen-timent Recognition for Interactive Dialogue Systems ” Proceedings of the 2016 Conference on Empirical Meth-Ods in Natural Language Processing “ , pp. 1042-1047

[George 16] 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.

[Ekman 75] Ekman, P. and Friesen, W. V.” Unmasking the Face,Prentice-Hall” ,1975

[Nesterov 83] 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.

[RAVEDESS 18] 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.

[有本 08] 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

[有本 13] 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.

[Cho 09] 趙章植ら,” ベイジアンアプローチに基づく感情発話音声からの感情推定における日韓感性の比較” 日本感性工学会論文誌 Vol.8No.3 pp.913-919，2009