Vol.2018-SLP-125 No.6
2018/12/10

日韓混合感情音声からの1次元畳み込み双方向LSTMを用いた感情推定

坂口　巧一[1,a)]　加藤昇平[1,2]

概要：近年，ロボティクス技術と 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．また, グローバル社会の到来により異文化間交流の機会が増加しており, 対人ロボットの感情推定技術でも異なる言語間における音声感情表現の差異を吸収する必要がある. しかし, 音声から感情を推定する既存研究においてこの点を考慮しているものは少ない．そこで本研究では音声に含まれる非言語情報に注目する．1 次元畳み込み双方向 LSTM を用いて日本語と韓国語の感情音声に対する判別性能の違いを調査した．

1D-Convolutional Bidirectional LSTM for Emotional Estimation from Japanese and Korean Voice

Koichi Sakaguchi[1,a)]　Shohei Kato[1,2]

Abstract:
Recently, voice interaction robots attract people due to development of AI and robot engineering. The technology of emotional estimation from voice is important to realize a smooth dialog between human and robots. The opportunities of intercultural interactions are increasing by arrival of the globalizing society. So, the technology of emotional estimation for interaction robots need to absorb the differences in voice emotional expressions of different languages. But, the number of existing researches about the emotional estimation from voice that consider about this point is small. In this paper, we pay attention to the non-language information in voice. We then provide a comparison of performance about emotion recognition from Japanese and Korean voice.

1. はじめに

近年，ロボティクス技術と AI の発展に伴い，人と音声により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．

音声から感情を推定する研究は以前から行われている．かつてはあらかじめ決められた複数の音声特徴量を抽出して Support Vector Machine（SVM）などで学習させて判別する手法が多かった [1][2][3]．しかし，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 ら [4] は，感情音声をスペクトログラムに変換し，畳み込みニューラルネットワーク (CNN) で感情推定を行い，SVM よりも高い識別率を示した．George ら [5] は，音声データから手動で特徴抽出をした場合と畳み込み層で特徴抽出した場合の感情識別率を比較しており，後者のほうが高い識別率となったことを示している．

同じ日本語圏内であっても関東の人が関西の人の発話音声を抑揚などの非言語情報から感情を誤判別してしまうことがあるように，文化圏が異なれば音声に表れる感情表現も異なる．話す言語が異なる異言語間では尚更異なる．

[1] 名古屋工業大学,
Nagoya Institute of Technology(Nitech)
[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

また，近年のグローバル社会の到来により，異文化間交流の機会が増加している．近い将来，異文化の人々が同じ卓で会話する場面が日常的になることが予想される．対話ロボットの役割として，会話する人々の間に入って会話をサポートすることも考えられる．異なる言語文化の人々の会話をサポートするには，感情表現の違いを吸収する必要がある．そのための感情分類器を作る方法として，言語を識別して，その言語用の感情分類器で感情を分類することが考えられる．しかし，グローバル社会では相手に歩み寄るために，慣れない相手の言語を使って会話する必要に迫られる．慣れない言語を用いると，言語に表れる感情表現は本来のものと大きく異なる可能性が高い．この場合，言語を正しく分類できたとしても感情分類を正しくすることは難しい．この問題は異なる言語間に共通する感情特徴を学習することで解決できると考えられる．

しかし，音声から感情を推定する従来研究においてそのようなことを考慮しているものは少ない．そこで，本研究では異なる言語間に共通する感情特徴を学習することを試みる．その第一歩として本稿では CNN と双方向 Long Short Term Memory （LSTM）を組み合わせたモデルを用いて，日本語と韓国語の感情音声に対する判別性能の違いを調査した．

2. 提案手法

2.1 音声データの前処理

図 1 は前処理全体の流れを表したものである．音声データは，スペクトログラムに変換された後，提案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（FFT）を行うことで作られる，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．今回は，FFT のサンプル数を 512，オーバーラップ長を 120 とした． 時間長は最小のものに合わせて各データをカットし，スペクトログラムのサイズは 257 次元 ×55 タイムステップとした．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提案モデルに入力する．

2.2 提案モデル

提案モデルを図 2 に示す．このモデルは 1 次元畳み込み層+プーリング層と双方向 LSTM の 2 つの段階からなる．以下，各段階について説明する．なお，最適化関数には確率的勾配降下法を用い，学習率は 0.01，モメンタムは 0.9 とした．

2.2.1 1 次元畳み込み層+プーリング層

1 次元畳み込み層+プーリング層は 1 次元畳み込み層と 1 次元プーリング層の組み合わせからなる．畳み込み層は特徴の鋭敏化の役割を担い，プーリング層は次元圧縮の役割を担う．提案モデルでは畳み込み層の出力に対してバッチ正規化を，プーリング層ではドロップアウトを行っている．

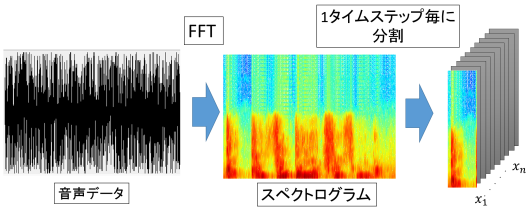

図 1 前処理の流れ
Fig. 1 Flow of preprocessing

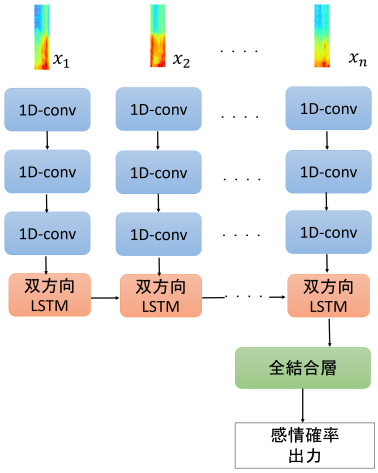

図 2 提案モデル概要
Fig. 2 Overview of our proposed model

表 1 提案モデル詳細
Table 1 Detail of our proposed model

パラメータ設定	
入力層	入力サイズ:257 次元 ×55 タイムステップ
畳み込み層 1	フィルタ:(3,1) × 32 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 ドロップアウト率:0.25
畳み込み層 3	フィルタ:(3,1) × 64 バッチ正規化あり
最大プーリング層 2	プールサイズ:2 ドロップアウト率:0.50
双方向 LSTM 層	ユニット数:512×2
全結合層	活性化関数:softmax l1l2 正則化:(0.01,0.001)
出力層	出力サイズ:(5,1)

2.2.2 双方向 LSTM

LSTM は主に時間依存性の強いデータに対して効果を発揮する再帰型ニューラルネットワークである．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ことで分類性能の向上が期待できる．

3. 実験方法

本実験では，以下に示すデータで学習した分類器の性能を比較した（発話数は学習サンプルの数）．

表 2　日本語データの f1-score
Table 2　Japanese emotional voice estimation results

f1-score			
	モデル韓	モデル日	モデル混
怒り	0.45	**0.48**	**0.48**
悲しみ	0.34	0.36	**0.42**
嫌悪	0.45	0.44	**0.48**
驚き	0.17	**0.27**	0.18
喜び	**0.23**	0.20	0.12
平均	0.33	**0.35**	0.34

表 3　韓国語データの f1-score
Table 3　Korean emotional voice estimation results

f1-score			
	モデル韓	モデル日	モデル混
怒り	**0.60**	0.57	0.55
悲しみ	**0.55**	0.42	0.42
嫌悪	**0.74**	0.58	0.65
驚き	0.25	**0.45**	0.36
喜び	**0.37**	0.17	0.15
平均	**0.50**	0.44	0.43

- (モデル韓) 韓国語の感情音声データ 400 発話
- (モデル日) 日本語の感情音声データ 400 発話
- (モデル混) 日韓混合の感情音声データ 800 発話

学習とテストには cho らが [3] で用いた感情音声データを用いる．このデータは日本と韓国の TV ドラマや映画から俳優が感情を込めて発話したフレーズを抽出し，聴取実験により感情が適切に表現されていると判断された音声資料を「怒り」「悲しみ」「嫌悪」「驚き」「喜び」の 5 感情に分類したデータである．聴取実験は，日本語と韓国語をそれぞれ母国語とする 2 名の聴取者により行った．この音声データは日本語のデータと韓国語のデータがそれぞれ 500 発話 (1 感情あたり 100 発話) ずつ収録されている．このデータの 2 割（日本語データ 100 発話，韓国語データ 100 発話）をテスト用のデータとし，残りを学習用データとするホールドアウト法により実験を行った．

4. 結果

表 2 と表 3 は各モデルを日本語及び韓国語の感情音声データについてそれぞれ感情分類を行った結果の f1-score（f 値）である．

モデル韓の f 値についてみてみると，韓国語の音声データの分類に関しては他のモデルと比べて最もよく（f 値 0.50），日本語の音声データの分類は他のモデルと比べて最もわるい（f 値 0.33）．これは韓国語の感情音声特有の特徴をよく学習したからだと思われる．

モデル日の分類結果についてみてみると，日本語の感情音声データに対する分類性能 (f 値 0.35) はモデル韓 (f 値 0.33) の場合と比較して良かった．しかし，モデル日の日本語感情音声データに対する分類性能（f 値 0.35）と韓国語感情音声データに対する分類性能（f 値 0.44）を比較すると，後者の方が良い結果だった．この結果から，日本語音声を用いた感情推定が韓国語の場合よりも困難なのではないかと推測される．

モデル混の結果についてみてみると，日本語データについては「怒り」「悲しみ」「嫌悪」の f 値が他のモデルに比べてよくなった（表 2）．しかし，韓国語のデータについては他のモデルよりもよい結果にはならなかった（表 3）．

5. 考察

特定の言語文化圏のみの感情音声データを学習した場合，その言語文化圏の音声に含まれる非言語情報のみに表れる感情特徴を学習する可能性が高い．そのため，複数の言語文化圏の感情音声データで学習することで言語文化圏の違いにとらわれない感情特徴を見つけることを期待した．モデル混の結果を見ると，韓国語データを学習データに加えることで日本語の感情分類性能の向上につながったが，日本語データを学習データに加えても韓国語の感情分類性能の向上にはつながらなかった．また，どちらの言語についても十分な特徴を見つけることができたとは言い難い．その原因としてサンプル数の不足が考えられる．各言語の音声感情データが 400 ずつであり，深層学習で用いるには学習サンプル数が不足していることが考えられる．この問題の解決手段として，ファインチューニングが考えられる [6]．この手法は対象タスクに関連するタスクで学習したパラメータを初期値として，対象タスクで更に学習を進める手法である．一から学習する場合に比べてよい初期値が得られることが期待できるため，特に学習データが少ない場合は，よりよい学習結果を得ることが期待できる．

また，複数の言語文化の感情分類の方法として，言語文化の分類器を加えることが考えられる．感情分類器は 1 つの言語について学習すればよいため，ある程度の性能の向上が期待できる．しかし，正しく分類するには言語と感情の両方について正しく分類しなければならない．そのため，異なる言語間に共通する特徴を学習する方がよい結果になるかもしれない．

以上の点を考慮して，今後は感情に関連するタスクの日本語と韓国語の学習サンプルを獲得し，それについて学習したモデルを感情分類モデルに転用するファインチューニングをし，混合感情音声データで学習した場合と，言語文化の分類器を付け加えた場合を比較することを検討していく．

6. おわりに

本稿では日本語と韓国語の感情音声について 1 次元畳み込み双方向 LSTM で感情分類を行った．韓国語のみを学習したモデルは韓国語の感情特徴を上手く学習できたが，

日本語のみを学習したモデルは日本語の感情特徴を上手く学習できなかった．また，モデル混は分類性能が十分とはいえず，異なる言語間に共通する十分な感情特徴を発見できたとはいえない．原因として学習データ数の不足と共通する感情特徴の不足が考えられる．この問題を解決するための手法として，今後は日本語と韓国語の音声認識のデータを獲得し，それについて学習したモデルを感情分類モデルに転用するファインチューニングや，どの言語文化の音声かを判別する分類器を加える手法との比較について検討していきたい．

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] 酒造正樹ら:「情動・感情判別のための自然発話音声データベースの構築」情報処理学会誌 Vol.52 No.3,p.1185-1194,2011
[3] 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
[4] Dario Bertero et al:“Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1042-1047
[5] 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.
[6] R. Girshick et al :” Rich feature hierarchies for accurate object detection and semantic segmen-tation.” In Proc. IEEE CVPR, 2014.