2R-08

性格特性を考慮した音声感情認識

石田絹史朗　長名優子

東京工科大学 コンピュータサイエンス学部

1 はじめに

音声感情認識は、音声から話者の感情を認識する技術であり、現在は深層学習に基づく手法が主流となっている [1]。しかしながら、感情の表現の仕方は話者によって異なるため、話者によっては認識精度が低くなることがあるといった問題がある。それに対し、性格特性を考慮した音声感情認識 [2] が提案されている。これは、音声、テキスト情報、話者の性格特性の 3 つを考慮した感情認識手法であり、性格特性を表現するのに LIWC (Linguistic Inquiry and Word Count)[3] を利用している。LIWC はテキストに含まれる語彙をカテゴリ別にカウントするものであり、この手法では話者の使う語彙の傾向などが性格特性に影響すると考え、LIWC の出力を用いて性格特性を表現している。しかしながら、LIWC は言語に依存するため、日本語の感情認識にはそのまま利用することはできない。

また、一方で HuBERT (Hidden Unit Bidirectional Encoder Representations from Transformers)[4] という畳み込みニューラルネットワーク [5] と Transformer[6] に基づいた音声処理のための事前学習モデルが提案されている。これは、ファインチューニング/転移学習を行うことで音声信号を入力としする様々なタスクに適用することができる。

本研究では、HuBERT と LIWC を用いた性格特性を考慮した音声感情認識を提案する。

2 HuBERT

提案手法で用いる HuBERT[4] は音声処理のための事前学習モデルとして提案されたもので、音声データに対応する埋め込み表現を生成することができる。音響特徴量である MFCC (Mel-Frequency Cepstrum Coefficient)[7] を K-means 法 [8] を用いてクラスタリングした結果を疑似ラベルとして自己教師あり学習を行い、音声波形の埋め込み表現の一部をマスクした状態でも正しく出力できるように学習を行う。モデルの名前にある Hidden Unit はこの疑似ラベルに相当するものをさす。

HnBERT は CNN エンコーダと Transformer エンコーダから構成されている。入力された音声信号は、CNN エンコーダを通すことでフレームごとの音声信号の埋め込み表現に変換される。さらに、Transformer エンコーダを通して、文脈 (前後関係) を考慮した埋め込み表現が生成される。Transformer から出力される文脈埋め込み表現が疑似ラベルに対応したものになっている。

実現したいタスクに合わせた出力層を追加し、ファインチューニングや転移学習を行うことで、音声信号を入力とする様々なタスクに適用できる。たとえば、音声認識に用いる場合には、ソフトマックス層を追加し、音声信号に対して対応するテキストが出力できるように学習を行う。

3 LIWC (Linguistic Inquiry and Word Count)

LIWC[3] は心理学の領域におけるデファクト・スタンダードの感情辞書であり、日常よく使用される単語に対して、専門家によってあらかじめ分類されたカテゴリを割り当てたものである。カテゴリは、大きく言語学的なカテゴリと心的プロセスに関するカテゴリが含まれる。言語学的なカテゴリには、助動詞、接続詞などの文法的機能を示す語である機能語、名詞、動詞などの実質的な内容をもつ語である内容語とが含まれる。心的プロセスに関するカテゴリには、感情的/認知的/社会的プロセスなどに対応するカテゴリが含まれる。英語版 (1997, 2001, 2007, 2015) 以外にも様々な言語版が作成され、日本語版 (J-LIWC2015)[9] も存在する。LIWC は文に含まれる語彙がどのカテゴリに含まれるかを調べてカウントすることができ、その結果を利用することで、書き手/話し手の心理傾向の推測、心的態度の推定、感情推定などに利用できることが知られている。

4 性格特性を考慮した音声感情認識

提案する性格特性を考慮した音声感情認識では、音声と性格特性ベクトルを用いて日本語を対応とした音

Speech Emotion Recognition considering Personality
Kenshiro Ishida and Osana Yuko(Tokyo University of Technology, osana@stf.teu.ac.jp)

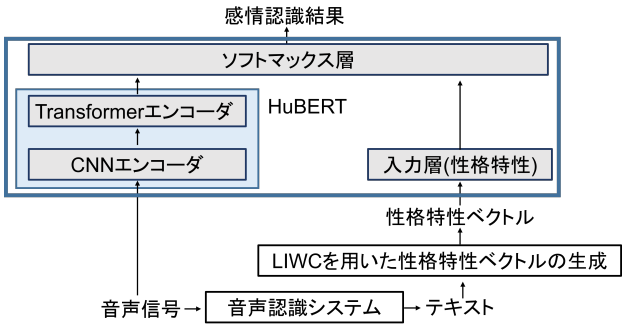

図 1: 提案システムの構成

声感情認識を実現する。感情としては、普通、喜び、悲しみ、怒り、恐れ、嫌悪、驚き、焦り、落ち着きの9つを扱う。事前学習モデルとして日本語の音声に特化したHuBERT[10]を使用し、ソフトマックス層を追加し、HuBERTから出力される音声信号に対応する文脈埋め込み表現と性格特徴ベクトルから感情認識が行えるように転移学習を行う。性格特性ベクトルは日本語版のLIWC (J-LIWC2015)[9]を用いて生成する。

4.1 構成

図1に提案システムの構成を示す。音声信号は、HuBERTに入力されると同時にLIWCを用いた性格特性ベクトルの生成にも用いられ、性格特性ベクトルとあわせて、ソフトマックス層で感情認識が行えるように学習を行うことになる。LIWCの入力はテキストデータである必要があるため、音声信号を音声認識システムを通してテキストに変換してものをLIWCへ与え，その出力を利用して性格特性ベクトルを尾止めることになる。

4.2 性格特性ベクトル

性格特性ベクトルはLIWC(J-LIWC2015)の出力を用いて生成する。LIWCの出力は話者の使用する語彙の傾向を反映したベクトルであり、ここでは、LIWCベクトルと呼ぶものとする。性格特性ベクトルは、背景話者 (訓練データに用いる話者) の特性との類似度で表現する。

話者 i の性格特性ベクトル $\boldsymbol{c}^{(i)}$ は

$$\boldsymbol{c}^{(i)} = \boldsymbol{G}\boldsymbol{g}^{(i)} \tag{1}$$

で与えられる。ここで、$\boldsymbol{g}^{(i)}$ はノルムが1になるように正規化した話者 i のLIWCベクトル、N は背景話者の数である。また、$\boldsymbol{G}$ は背景話者のLIWCベクトルをまとめた行列である。

4.3 学習

HuBERTにソフトマックス層を追加し、音声データと音声データから生成した話者ごとの性格特性ベクトルを入力として与え，音声データに対応する感情を出力するように転移学習を行う。

5 計算機実験

計算機実験を行い、性格特性を考慮することで音声感情認識の精度が上がる可能性があることを確認した。

参考文献

[1] 安藤厚志：“音声感情認識の技術動向— 深層学習に基づく手法とその最新研究 —,” 日本音響学会誌, Vol.79, No.1. pp.72–79, 2023.

[2] J. L. Li and C. C. Lee：“Attentive to individual: a multimodal emotion recognition network with personalized attention profile,” The 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2019.

[3] J. W. Pennebaker, R. L. Boyd, K. Jordan and K. Blackburn：“The development and psychometric properties of liwc2015,” https://doi.org/10.15781/T29G6Z, 2015 (2024/01/06 参照).

[4] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed：“HuBERT: self-supervised speech representation learning by masked prediction of hidden units,” https://arxiv.org/pdf/2106.07447.pdf, 2016 (2024/01/06 参照).

[5] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner：“Gradient-based learning applied to document recognition,” Proceedings of the IEEE, Vol.86, No.11, pp.2278–2324, 1998.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin：“Attention is all you need,” https://arxiv.org/pdf/1706.03762.pdf, (2024/01/06 参照).

[7] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner：“Gradient-based learning applied to document recognition,” Proceedings of the IEEE, Vol.86, No.11, pp.2278–2324, 1998.

[8] J. B. MacQueen：“Some methods for classification and analysis of multivariate observations,” Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, No.1, pp.281-297, 1967.

[9] T. Igarashi, S. Okuda,and K. Sasahara：“Development of the Japanese ersion of the linguistic inquiry and word count dictionary 2015,” Frontiers in Psychology, 13:841534. (https://doi.org/10.3389/ fpsyg.2022.841534), 2022 (2024/01/06 参照).

[10] 日本語の音声に特化した事前学習モデル HuBERT, https://rinna.co.jp/news/2023/04/20230428.html, 2023 (2024/01/06 参照).