対話音声の感情認識における音響的特徴と言語的特徴の効果*

☆横山雅季, 森大毅 (宇都宮大), 有本泰子 (芝浦工大)

1 はじめに

音声が伝達する感情の研究には, 感情ラベルが付与された音声対話コーパスが必要になる。様々な状況での発話を得るためにはコーパスの併用が有効であるが, そのためには複数のコーパスで共通な感情ラベルを付与する必要がある。感情音声コーパスの共通化のため, 音響的特徴から感情ラベル推定を行う研究があるが, その推定精度は十分ではない [1]。

ところで, 既存のコーパスの感情ラベル推定には, コーパスが持つ転記テキストが利用できる。言語メッセージの選択には話者の感情が反映される [2] ため, 転記テキストに含まれる言語情報の利用は感情ラベル推定に有効だと考えられる [3, 4]。

本研究では, 既存の自然な感情音声コーパスに対する感情ラベル推定において, 音響的特徴に加えて言語的特徴を用いることの有効性を検討する。

2 音声資料と特徴量

感情ラベル推定のための学習・評価用のコーパスとして, 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)[5] を用いた。OGVC の感情ラベルが付与されている 6578 発話のうち, 異なる 3 名の評価者のうち 2 名以上が同一のラベルを付与した 3845 発話を使用した。含まれているラベルは, 恐れ (FEA), 驚き (SUR), 悲しみ (SAD), 嫌悪 (DIS), 怒り (ANG), 期待 (ANT), 喜び (JOY), 受容 (ACC), 平静 (NEU), その他 (OTH) の 10 種類である。使用した発話の感情ラベルの内訳を Table 1 に示した。

各発話の転記テキストに対して形態素解析を行い, 各単語に対応する位置の値が 1 になるようなベクトル (Bag-of-Words: BoW) を作成し, これを言語的特徴とした。予備実験の性能比較の結果を踏まえ, ストップワード (極めて高頻度に出現する, 情報の少ない語) は除外していない。ただし, 今回は名詞を使用しなかった。本実験では 473 次元の特徴量ベクトルを用いた。

音響的特徴には, Interspeech 2010 パラ言語チャレンジのベースライン特徴量 (1581 次元)[6] を用いた。

Table 1: 感情ラベルの内訳 [%]

FEA	SUR	SAD	DIS	ANG	ANT	JOY	ACC	NEU	OTH
3.69	14.7	6.32	8.71	6.16	11.1	15.5	7.88	20.8	5.20

Table 2: 感情ラベル推定精度

	WAR(%)	UAR(%)
言語的特徴のみ	33.3	23.6
音響的特徴のみ	38.6	31.3
言語的特徴+音響的特徴	40.4	34.5

3 感情ラベル推定実験

3.1 推定精度の比較

OGVC の 2 名以上一致ラベルを目的変数とした分類器 (SVM) を構築し, その性能を評価した。10 分割交差検証で評価した推定精度を Table 2 に示す。WAR (Weighted Average Recall) は単純な正解率, UAR (Unweighted Average Recall) は各感情ごとの正解率の平均である。言語的特徴または音響的特徴を単独で用いた場合に比べ, それらを組み合わせた場合の推定精度が向上していることがわかる。

3.2 音響的特徴の効果

Fig. 1 に, 音響的特徴を単独で用いた場合の混同行列を示す。正解ラベルが FEA である発話 (1 行目) に注目すると, SUR, SAD, NEU, JOY に誤る発話が多いことが分かる。これは, 恐れ (FEA) と驚き (SUR), 驚き (SUR) と悲しみ (SAD) が, プルチックの感情の円環において隣接しているため, もともと近い感情だということが関係しており, 人間の混同傾向とも類似している [5]。

正解ラベルが FEA である発話のうち, 分類結果が FEA, JOY または SAD となった 66 発話を対象として, 音響的特徴を目的変数, 分類結果を説明変数として t 検定による多重比較を行った (有意水準 5%, Bonferroni 補正)。その結果, 平均値の差が有意であった音響的特徴は 1581 次元中 501 次元であった。これらのうち, ラウドネスの 1%点 (音響的特徴 1) および第 1 線スペクトル周波数の差分の最小位置 (音響的特徴 2) の分布を Fig. 2 に示す。

Fig. 2(a) より, 誤って SAD または JOY に分類された発話はラウドネスの 1%点の値が小さい傾向があることがわかる。すなわち, これらの発話は, 無音区間を含む傾向がある。

一方, Fig. 2(b) より, 誤って JOY に分類された発話は, 第 1 線スペクトル周波数の差分の最小位置の値

*Effects of acoustic and linguistic features on the accuracy of emotion recognition from dialogue speech. by YOKOYAMA, Masaki, MORI, Hiroki (Utsunomiya University), ARIMOTO, Yoshiko (Shibaura Institute of Technology)

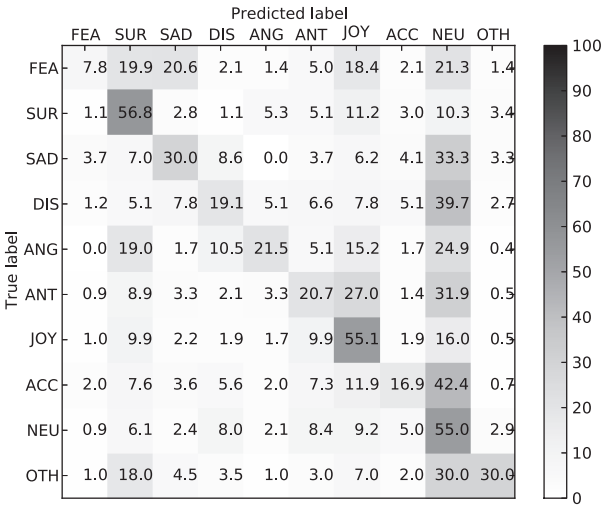

Fig. 1: 音響的特徴のみによる感情ラベル推定

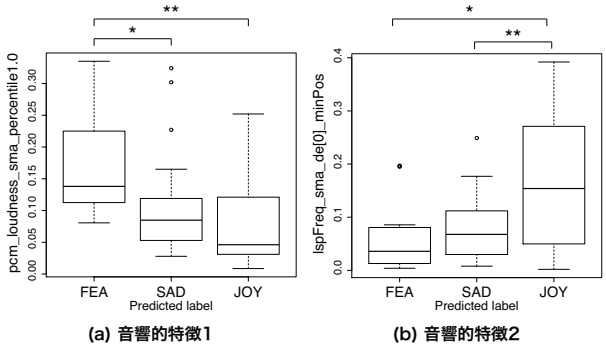

Fig. 2: FEA の推定に関連する特徴量

が大きい傾向がある。この特徴量は, スペクトルの変化が最小となる時刻を意味していると考えられ, 発話末の音節が引き伸ばされた場合に値が大きい傾向がある。JOY に分類された発話には, 「あー」, 「待ってー」, 「ひゃー」のように発話末の音節が引き伸ばされているものが多く存在していた。

3.3 言語的特徴の効果

Fig. 3 に, 音響的特徴と言語的特徴を組み合わせた場合の混同行列を示す。言語的特徴を加える事で, 正解ラベルが FEA である発話が正しく分類された割合は 7.8%から 19.1%に向上した。そこでどのような単語が FEA の推定に貢献したかを, 実際の発話に含まれる単語を確認して調べたところ, 「やばい」, 「きゃあ」, 「死ぬ」などがあった。この単語を 1 つずつ削除して結果がどう変わるか調べた。その結果を Table 3 に示す。どの単語も削除することで FEA と推定された発話の割合が低下し, SAD や JOY と判定された発話の割合が増加した。したがって, これらの単語は FEA の推定に寄与していると考えられる。

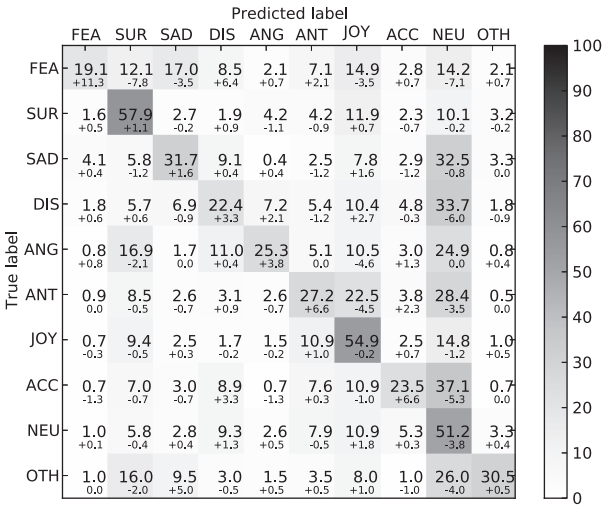

Fig. 3: 言語的特徴+音響的特徴による感情ラベル推定

Table 3: 恐れ (FEA) の推定精度に影響する単語

取り除いた単語	FEA(%)	SAD(%)	JOY(%)
なし (baseline)	19.1	17.0	14.9
「やばい」	17.0	19.9	17.0
「きゃあ」	17.0	20.6	17.0
「死ぬ」	17.0	20.6	17.7

4 おわりに

本稿では, 音響的特徴に加えて言語的特徴を用いることで感情ラベル推定の精度が向上することを示した。また, 正解ラベルが FEA の発話に対して, 推定に寄与する音響的特徴と言語的特徴を調べた。今後の課題として, 単語の埋め込み表現を利用した言語的特徴による推定精度の向上が挙げられる。

参考文献

[1] 永岡 他, 音講論 (春), 415–416, 2015.
[2] 東中 他, 人工知能, 31(5), 664–670, 2016.
[3] 有本 他, 自然言語処理, 14(3), 147–163, 2007.
[4] Nomoto *et al.*, in Proc. Interspeech 2011, 1545–1548, 2011.
[5] Arimoto *et al.*, Acoust. Sci. Tech., 32(1), 26–29, 2011.
[6] Schuller *et al.*, in Proc. Interspeech 2010, 2794–2797, 2010.