第 5 回 東北地区音響学研究会（2022 年 11 月 30 日）資料番号 5－12

言語特徴と音響特徴の後期融合による音声感情認識の検討

○佐藤清秀(山形大学/ビットアート)，岸恵汰，小坂哲夫 (山形大学)

1. はじめに

人間対人間の対話と同様に感情などの非言語情報は，円滑な対話を実現するために大きな役割を果たすことになり，感情を推定する感情認識技術の重要性は高まっている．

感情認識の手法としては従来音響情報のみが使われる場合が多かったが，[1]では音声認識により言語情報を取得し，音響および言語情報の両方を使用して感情認識する方法が提案されている．この場合音声認識の精度が感情認識の精度に影響を与える可能性があるが，書き起こしによる正解テキストを利用し，音声認識の精度が 100%を仮定した実験と比較して，遜色のない結果が得られている．

[1]では BERT[2]による言語特徴，統計量および LLD による音響特徴の 3 種類の認識出力を重み付き加算することにより後期融合し感情認識を行った．しかしタスクごとに重みを計算し融合するのは実用上問題がある．本研究では後期融合についても深層学習モデルを用いることにより認識実験を行った結果について述べる．

2. 後期融合による感情認識手法の概要

音響特徴による感情認識では，時系列特徴である LLD からの感情認識および発話全体の統計量からの感情認識を併用する．時系列特徴からの認識では識別器として双方向 LSTM（BiLSTM）・双方向 GRU（BiGRU） を用い，統計量を用いた認識では識別器として DNN を用いる．

言語特徴による感情認識では，まず音響モデルと言語モデルを用いて音声認識を行い，得られた認識結果を利用して BERT により感情の推定を行う．

最終的に音響特徴および言語特徴の認識結果を DNN による融合法で感情認識を行い，「怒り(ang)」「喜び(joy)」「悲しみ(sad)」「平静(neu)」の 4 感情の出力を得る．感情認識システムの構成図を図 1 に示す．

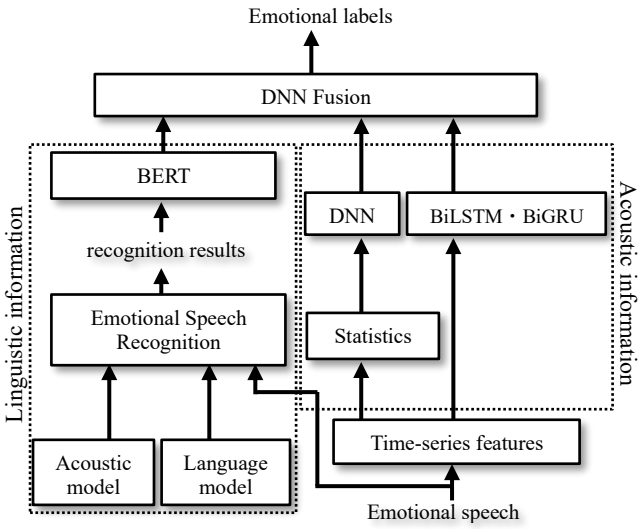

Fig. 1 Block diagram of the proposed system

3. 言語特徴･音響特徴を用いた感情認識

3.1. 言語特徴を用いた感情認識

言語特徴による感情認識のため，まず感情音声認識を行う．認識には[3]で示す音声認識手法を用いる．この認識システムでは，感情音声に対して，音響モデル（acoustic model:AM）と言語モデル（language model:LM）の適応を行った．AM 適応では，感情音声データを適応データとして用い，バックプロパゲーション法により適応した．LM 適応には，twitter 上より収集したツイートデータを適応データとして用いた．この音声認識結果を利用して言語特徴による感情認識を行う．本研究では BERT を利用し，予め大量のラベルなしデータを学習しておき，目的のタスクのため少量のラベル付きデータを転移学習させる手法を用いる．まず日本語の事前学習済みモデルに感情ラベル付きデータで転移学習することで，4 感情の感情識別器を学習する．事前学習済みモデルには日本語 Wikipedia のテキストを用いて学習したモデル[4]を用い，転移学習デ

---

* A Study of speech emotion recognition by late fusion of linguistic and acoustic features by SATO, Kiyohide (Yamagata Univ. / Bit Art), KISHI, Keita and KOSAKA, Tetsuo (Yamagata Univ.)

ータは JTES 1963 文を使用し，このうち 1 割を検証データとして使用した. AM と LM 適応後の認識精度は 82.2 %となった。

3.2. 音響特徴を用いた感情認識

感情認識のための音響特徴量としては，感情特徴を表す時系列として LLD を抽出し, そこから発話全体の特徴を表すさまざまな統計量を算出する．まず入力された音声を窓関数によってフレーム毎に分割し，基本周波数などの特徴量（LLD）を抽出する. ここでは LLD として The large openSMILE emotion feature set[5]を用い，1 フレーム(10msec)あたり 168 次元の特徴ベクトルを用いる. 次に 168 次元の LLD から 39 種類の統計量を算出し, 39×168 の計 6552 次元の特徴量を統計的特徴用の DNN に入力した．DNN の構造は 2048 個のユニットの 3 つの隠れ層からなる.

時系列特徴である LLD は, 経時変化に適したモデルの使用が考えられる．よって，本研究では識別器として LSTM などを利用した手法を用いた．前述の LLD を用いて双方向 LSTM(BiLSTM)への入力とした．ネットワーク構造は，入力側から BiLSTM 1 層，BiGRU 1 層，全結合層(fc1～fc3) 3 層の構成とした. BiLSTM と BiGRU の隠れ層はそれぞれ 100 個，全結合層 fc1 は 512 個, 全結合層 fc2 は 256 個, 全結合層 fc3 は 64 個のユニット，学習率は 0.0005，最適化手法として adam，損失関数は cross entropy を使用した.

4. 後期融合

4.1. 重み付き融合法

まず[1]の手法を用いて, LLD と統計量を次の式にて重み付き融合法による後期融合を実施した.

$$Ae = \alpha Se + (1-\alpha)Te \quad (1)$$

最終的な LLD と統計量を用いた音響特徴による認識感情スコアを Ae とする．Se は統計量の尤度であり, Te は LLD の尤度である. α は重み付き係数であり 0 から 1 の間 0.1 刻みで変化させた.

次に言語特徴と音響特徴を下記の式により重み付き融合法による後期融合を実施した.

$$emo = \beta Le + (1-\beta)Ae \quad (2)$$

最終的な認識感情スコアを emo とする. Le は言語特徴の感情尤度であり, Ae は音響特徴の感情尤度である．β は重み付き係数であり 0 から 1 の間 0.1 刻みで変化させた.

4.2. DNN による融合法

DNN による融合法による後期融合を実施した. DNN のネットワーク構造は，全結合層 fc1 に 128 個，全結合層 fc2 に 32 個，全結合層 fc3 に 32 個のユニットの構成とした．学習率は 0.00004，最適化手法として adam，損失関数は cross entropy を使用した.

5. 感情音声コーパス

本研究ではコーパスとして感情音声データベース JTES(Japanese Twitter-based emotional speech)を用いた[6]．このコーパスは，Twitter のつぶやきの中から感情表現語を含む口語的な文章を，音韻や韻律のバランスを考慮し選出されている. 話者は 100 名(男女各 50 名)，感情は「怒り」「喜び」「悲しみ」「平静」の 4 感情で各感情 50 文，計 20000 発話が用意されている．発話者には「自分が意図する感情をロボット(機械) に伝える」ように発話するよう指示されている．また，怒りには「hot anger(激しい怒り)」と「cold anger(押し殺した怒り)」があるが，韻律的特徴の有効性が既に示されている「hot anger」を対象とし，収録の際には「hot anger」を意識するよう指示されている.

6. 感情認識実験

6.1. 実験条件

学習データには JTES 14400 発話(40 文×4 感情×(男性 45 話者+女性 45 話者))，評価データには学習データを含まない JTES 400 発話(10 文×4 感情×(男性 5 話者+女性 5 話者))を使用した．重み付き融合法と DNN による融合法の後期融合では，音響特徴と言語特徴ともに同じ感情認識結果から得た LLD 4 次元，統計量 4 次元，BERT 4 次元の特徴ベクトルを入力データとして用いた．言語特徴には音声認識(automatic speech recognition:ASR)の認識結果と，比較のため音声認識率 100%

と仮定した正解テキストを利用する場合の 2 種類を試みた．以降前者を言語特徴(ASR)，後者を言語特徴(正解)と記載する．重み付き融合法の LLD と統計量の後期融合では，式(1)の重み α は実験的に 0.35 に設定し，次に言語特徴と音響特徴の後期融合では，式(2)の重み β は実験的に正解テキストとの融合では 0.35 に，ASR との融合では 0.25 に設定した．

6. 2. 実験結果

音響特徴おける感情認識については LLD のみの認識率が 75.75%となり，統計量のみの認識率が 69.25%となった．重み付き融合法の LLD と統計量による後期融合の結果は 77.25%となった．重み付き融合法による音響特徴と言語特徴(正解)，または言語特徴(ASR)との後期融合の結果を図 2 に示す．

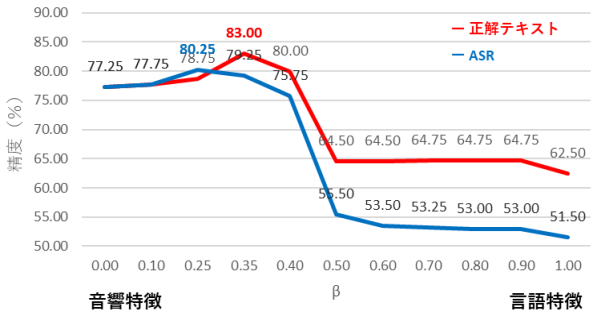

Fig. 2 重み付き融合結果

さらに，DNN による融合法による LLD および統計量の音響特徴と言語特徴(正解)，または言語特徴(ASR)との後期融合を行った．

以上の認識実験のまとめを表 1 に示す．

Table 1 融合条件毎の感情認識結果 (%)

融合法＼条件	音響特徴＋言語特徴(正解)	音響特徴＋言語特徴(ASR)
重み付き融合法	83.00	80.25
DNN による融合法	85.50	82.25

重み付き融合法と DNN による融合法の比較の結果，音響特徴と言語特徴(正解)，または言語特徴(ASR)を用いた融合において，どちらの場合も DNN による融合法の方が高い認識率を示す結果となった．一方，言語特徴(正解)と言語特徴(ASR)との比較では，言語特徴(正解)との融合の方が多少認識精度は高いものの，感情音声認識結果の優劣にはほぼ影響を受けないことが確認できた．

次に a)言語特徴(ASR)のみ，b)重み付き融合法による音響特徴のみ，c) LLD および統計量の音響特徴と言語特徴(ASR)の DNN 融合の 3 種類の認識結果について比較を行った．a)，b)それぞれの認識率は図 2 より 51.5%，77.25%であり，c)については表 1 より 82.25%となる．a)の混同行列を表 2，b)の混同行列を表 3，c)の混同行列を表 4 に示す．

Table 2 言語特徴(ASR)のみの混同行列(%)
認識率 51.5%

正解＼予測	ang	joy	neu	sad
ang	52	4	20	24
joy	12	55	6	27
neu	10	23	46	21
sad	20	25	2	53

Table 3 重み付き融合法による音響特徴のみの混同行列(%) 認識率 77.25%

正解＼予測	ang	joy	neu	sad
ang	85	14	1	0
joy	38	57	2	3
neu	1	6	85	8
sad	1	2	15	82

Table 4 音響特徴と言語特徴を DNN による融合法用いた感情認識結果の混同行列(%)
認識率 82.25%

正解＼予測	ang	joy	neu	sad
ang	86	8	6	0
joy	23	72	2	3
neu	2	3	90	5
sad	1	4	14	81

言語特徴のみの場合は neu を joy や sad に，sad を ang や joy に間違える傾向を示しているが，発話内容テキストによる言語特徴のみでは neu や sad の感情を判別しにくいことを表している．

音響特徴のみの場合では，表 3 における ang と joy の混同が多いのは，発話の強弱や基本周波数の変動が他の感情に比べ大きいといった音響的類似性が影響したと考えられる．一方，表 2 の言語特徴においては，そのような影響を受けないため比較的混同が少ない．

音響特徴と言語特徴を DNN による融合法により後期融合した結果，これらの認識誤りが改善されて感情認識の精度が向上していることが確認できる．

7. まとめ

本研究では，言語特徴と音響特徴の DNN による融合法を用いた後期融合による音声感情認識の手法を提案した．実験の結果，音響特徴と言語特徴の融合により各々の低認識率部分を補うことで高い精度の認識が可能であることが確認できた．また，重み付き融合法を上回る認識精度を得て，本システムの有効性が示された．しかし，JTES コーパスを用いたクローズドタスクでは認識性能がほぼ上限に達したと考えられるため，今後はオープンタスクを対象とした手法を検討したい．

謝辞

感情音声データベース JTES を東北大学能勢准教授にご提供頂いた．実験の一部について櫻井美咲氏(現 NEC ネッツ SI)，須々田和基氏(現アドバンスト・メディア)にご協力頂いた．本研究の一部は科研費(課題番号 22K12087)によった．以上記して感謝する．

参考文献

[1] 櫻井，須々田，小坂，音講論(春)，3-3-6，2022.

[2] Jacob Devlin, Ming-Wei, Chang Kenton, Lee Kristina Toutanova,“ BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ”, arXiv:1810.04805v2 [cs.CL] 24 May 2019.

[3] K. Saeki et al., Proc. of APSIPA ASC 2020, 371-375, 2020.

[4] 東北大乾・鈴木研, Pretrained Japanese BERT models, "https://github.com/cl-tohoku/bert-japanese".

[5] Florian Eyben et al., "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proceedings of the 18th ACM international conference on Multimedia, pp.1459-1462, 2010

[6] 武石，能勢，伊藤，音講論(春) , 1-R-47, 2015.