言語情報と音響情報の統合的利用による感情音声コーパスの大規模化

Automatic emotional labeling using linguistic and acoustic information towards large-scale emotional dialog speech corpus

2041003

研究代表者（助成金受領者）　国立研究開発法人　理化学研究所　客員研究員　有本泰子
共同研究者　宇都宮大学　工学部　准教授　森大毅

［研究の目的］

大規模音声コーパスを必要とする音声認識や感情認識などの人間と機械との対話の円滑化を念頭にした研究では，複数のコーパスを併用することが求められる。実際の会話コーパスの開発では，収録の対象が特定の状況でのインタラクションに限定されることが多く，またコーパスの規模も比較的小さいことが多い。これらのコーパスを統合し，大規模なコーパスとして扱うことができれば，多様な感情表出をより高精度にモデリング出来ると考えられる。しかし，実際には感情の記述はコーパスごとにそれぞれ独自の方法でなされており，互換性を持たない。これは，感情という現象がまだ十分に理解されておらず，その記述方法が確立していないためである。

感情の記述には，大きく分けて，次元に基づく方法と，感情カテゴリに基づく方法がある。これらはそれぞれ，次元説と基本感情説という，感情心理学において長く対立してきた二つの理論を背景に持つ。前者は感情状態を２次元，または３次元空間上のベクトルと考え，感情の類似性をベクトルの類似性として表す。一方，後者は少数の感情カテゴリを仮定する。感情カテゴリラベルが付与されたコーパスには，Berlin Database of Emotional Speech（Emo-DB）[1]，Parameterized & Annotated CMU Let's Go（LEGO）Database [2]，Surrey Audio-Visual Expressed Emotion（SAVEE）Database [3]，FAU Aibo Emotion Corpus [4]，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：Online Gaming Voice Chat Corpus with Emotional Label）[5] がある。一方，感情次元ラベルが付与されたコーパスには Vera am Mittag（VAM）German Audio-Visual Spontaneous Speech Database [6]，宇都宮大学パラ言語情報研究向け音声対話データベース（UUDB：Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies）[7] があり，さらに，両方のラベルが付与された The Interactive Emotional Dyadic Motion Capture（IEMOCAP）Database [8] が存在する。

このような感情の記述方法が異なる複数のコーパスを統合し，大規模なコーパスとして扱うことを可能にするためには，感情ラベルの共通化が必要である。すなわち，コーパスＡにはコーパスＢと同じ枠組で，コーパスＢにはコーパスＡと同じ枠組で感情ラベルを付与することになる。しかし，この作業を全て人手により実施することは極めて高コストである。そこで，音声からの感情認識の技術を利用し，異種コーパスへの感情ラベルを自動的に付与する

ことでコーパスを共通化することが考えられる。

本研究では，現在人手により付与されている感情ラベリングを自動化することで，コーパス間で共通した感情ラベルを自動付与し，感情ラベルにおけるコーパス間共通化基盤の確立を目指す。これにより，研究の資料として利用できる音声資源が爆発的に増加し，人間と機械との対話における感情研究の発展が見込まれる。

［研究の内容，成果］

コーパス共通化のための感情推定は，一般的な感情認識とは異なり，適用先のコーパスに既に何らかの感情ラベルが存在する。つまり，感情推定器への入力に適用先のコーパスが持つ感情ラベルが利用でき，音響特徴量や言語情報のみを入力とするより感情推定精度が向上することが期待できる。本研究では，音響特徴量と感情ラベルを感情推定器への入力とすることを提案し，この手法が音響特徴量のみを入力とした一般的な感情認識手法より高い精度での推定を行えることを示す。

本論文では，感情音声コーパス共通化のモデルケースとして，UUDB [7] と OGVC [5] の感情ラベル共通化を想定する。UUDB は感情が次元で評価され，OGVC は感情がカテゴリで評価されている。UUDB では以下の感情 6 次元が付与されている。

（1）快－不快（pleasant－unpleasant）
（2）覚醒－睡眠（aroused－sleepy）
（3）支配－服従（dominant－submissive）
（4）信頼－不信（credible－doubtful）
（5）関心－無関心（interested－indifferent）
（6）肯定的－否定的（positive－negative）

OGVC は喜び（JOY），受容（ACC），恐れ（FEA），驚き（SUR），悲しみ（SAD），嫌悪（DIS），怒り（ANG），期待（ANT）に加えて感情が表出していない平静（NEU），さらに，これらのどの感情にも分類できないその他（OTH）の 10 種類の感情カテゴリが付与されている。

1. クロスコーパス感情ラベリング

コーパス共通化のための感情推定は，それぞれのコーパスが持たない種類の感情ラベルを推定することが目的であるが，推定された感情ラベルの評価には正解ラベルが必要となる。そこで，UUDB と OGVC に対して次元とカテゴリの両方の枠組みで人手による感情ラベリングを行った。

まず，ラベル評価者のスクリーニングテストを行った。両コーパスからそれぞれ 54 発話ずつを用いて，コーパスの対話者と同年代の 10 名（女性 6 名，男性 4 名）に，UUDB の次元と OGVC のカテゴリの両方の感情ラベルを評価させた。この実験により，3 名のラベル評価者を選定した。そして，本評価では，UUDB の 4840 発話と OGVC の 6578 発話の計 11418 発話に対してスクリーニングテストと同様に両コーパスの両方の感情ラベルを評価させた。

本研究における実験ではこの本評価で得られた評価を使用し，コーパスに元々付与されている感情ラベルは使用しない。感情次元評価値は，単純に評価者 3 名の平均値を使用する。UUDB の 6 感情次元評価の結果のうち，横軸に快－不快と縦軸に覚醒－睡眠をとった分布図と横軸に快－不快と縦軸に支配－服従をとった分布図を図 1 に示す。どちらのコーパスも似た分布となったが，OGVC では不快で覚醒の発話が UUDB より多い傾向であった。また，UUDB は不快で服従の発話が OGVC よりも多い傾向を示した。

感情カテゴリは評価者 3 名のうち 2 名以上の評価が一致した感情カテゴリ（以下，過半数一致ラベル）を使用する。3 名全員の評価が異なった発話は以降の実験データから除いた。表 1 に過半数一致ラベルのカテゴリごとの数を示す。過半数一致ラベルはコーパスごとに異なる分布であった。特に，受容や恐れ，嫌悪の数がコーパスごとに顕著に異なる傾向があった。

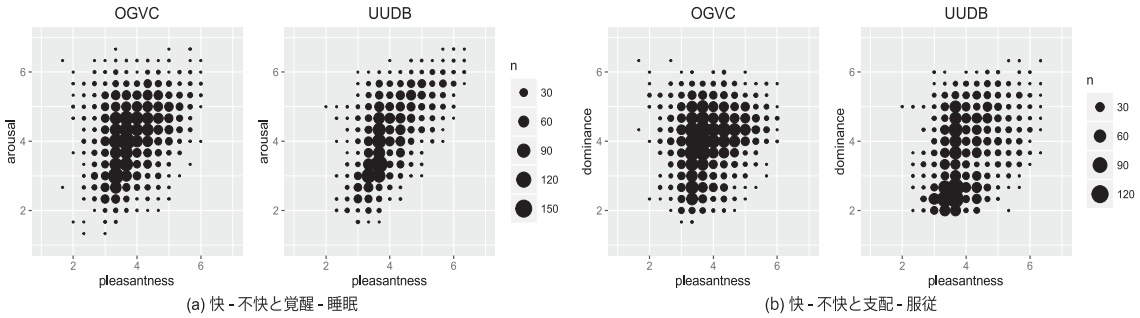

図 1　感情評価値の分布図

表 1　感情カテゴリごとの発話数

Emotion	OGVC	UUDB	Total
JOY	438	259	697
ACC	623	1030	1653
FEA	282	94	376
SUR	313	120	433
SAD	488	331	819
DIS	970	406	1376
ANG	128	39	167
ANT	186	59	245
OTH	0	0	0
NEU	18	13	31
Total	3446	2351	5797

2. クロスコーパス感情ラベル推定

UUDB と OGVC に既に付与されている種類の感情ラベルを併用することで，音響特徴量のみを用いる感情推定より高い精度での推定ができると期待できる。ここでは，感情ラベルを併用する感情推定手法を説明し，音響特徴量のみを入力とする場合，感情ラベルのみを入力とする場合，およびその両方を入力とする場合の比較を行う。

2.1　特徴量

OGVC の感情次元推定と UUDB の感情カテゴリ推定のどちらも音響特徴量抽出には openSMILE [9] を用いた。本研究では，事前検討において最も精度の高かった Interspeech 2010 Paralinguistic Challenge ベースラインシステム [10] と同様の 1582 次元の音響特徴量を使用した。音響モデルの学習や適用時に使用する特徴量はベクトルである必要があるため，それぞれの感情ラベルは，以下の方法でベクトル化を行なった。感情次元ラベルは，UUDB に付与されている 6 感情次元に対応する 6 次元ベクトルに変換した。感情カテゴリラベルは，OTH（その他）を除く 9 感情カテゴリに対応する 9 次元ベクトル（正解のカテゴリに対応する要素のみ 1，それ以外が 0 のベクトル）に変換した。

2.2　モデル学習

回帰モデルと分類モデルはそれぞれサポートベクター回帰とサポートベクター分類によって学習した。学習にはフリーソフト R の kernlab パッケージ [11] の ksvm 関数を用いた。カーネル関数にはガウシアンカーネルを用い，そのパラメータは kernlab により自動推定した。感情次元推定におけるハイパーパラメータは $C=1, \varepsilon=0.1$，感情カテゴリ推定におけるハイパーパラメータは，感情ラベルのみを入力とする場合では $C=8$，それ以外では $C=5$ とした。

本研究では，実験条件の数が増えて結果の解釈が煩雑になるのを避けるため，話者正規化やクラスごとのサンプル数の不均衡の補正などを行わない最も単純な条件での実験結果のみを記す。

2.3　実験結果

感情次元の推定精度は，推定されたラベルと正解ラベルの相関係数 R，および平均二乗誤差（RMSE）によって評価する。表 2 に，音響特徴量のみを使用した場合，感情ラベルのみを使用した場合，音響特徴量と感情ラベルを併用した場合の感情次元推定精度を示す。表中，太

表２　感情次元推定精度

		Pleasantness	Arousal	Dominance	Credibility	Interest	Positivity
Acoustic features only	*R*	0.312	0.764	0.703	0.339	0.613	0.207
	RMSE	0.677	0.602	0.630	0.650	0.544	0.638
Emotion labels only	*R*	**0.607**	0.426	0.346	**0.510**	0.419	**0.545**
	RMSE	**0.579**	0.886	0.892	**0.593**	0.634	**0.548**
Acoustic features ＋ Emotion labels	*R*	0.478	**0.774**	**0.709**	0.458	**0.648**	0.364
	RMSE	0.619	**0.591**	**0.626**	0.606	**0.525**	0.591

表３　感情カテゴリ推定精度（%）

	WAR	UAR	ACC	FEA	SUR	SAD	DIS	ANG	ANT	JOY	NEU
Acoustic features only	34.0	23.1	31.6	6.4	36.7	58.3	40.4	**2.6**	**8.5**	23.6	**0.0**
Emotion labels only	**50.4**	**29.3**	**58.7**	4.3	35.0	40.2	**56.7**	**2.6**	1.7	**64.9**	**0.0**
Acoustic features ＋ Emotion labels	40.2	26.6	40.0	**11.7**	**40.8**	**59.5**	46.1	**2.6**	6.8	32.0	**0.0**

字はこれら３条件での最も良い精度を示している。覚醒－睡眠，支配－服従，関心－無関心の各次元では，音響特徴量と感情ラベルを併用した場合の相関係数が最大となり，それぞれ 0.774, 0.709, 0.648 となった。快－不快，信頼－不信，肯定的－否定的の各次元でも，音響特徴量と感情ラベルを併用する方が音響特徴量のみを使用するよりも相関係数をそれぞれ 0.166, 0.119, 0.157 ポイント大きくすることができたが，感情ラベルのみを使用した場合には及ばなかった。平均二乗誤差についても，音響特徴量のみを使用する場合に比べ，音響特徴量と感情ラベルを併用する方が，より小さい誤差が得られている。

感情カテゴリの推定精度は，Weighted Average Recall（WAR）と Unweighted Average Recall（UAR）によって評価する。WAR はカテゴリごとのテストデータ数の偏りを考慮しない再現率であり，UAR はカテゴリごとのテストデータ数の偏りを平均化した再現率である。N カテゴリの分類問題では，カテゴリごとのテストデータ数を $c_1, c_2, \cdots, c_N$，カテゴリごとの正解数を $t_1, t_2, \cdots, t_N$ とすると，WAR と UAR は以下の式となる。

$$\mathrm{WAR} = \frac{\sum_i^N t_i}{\sum_i^N c_i}$$

$$\mathrm{UAR} = \frac{1}{N}\sum_i^N \frac{t_i}{c_i}$$

表３の二重線で区切った左側に，音響特徴量のみを使用した場合，感情ラベルのみを使用した場合，音響特徴量と感情ラベルを併用した場合の感情カテゴリ推定精度を示す。音響特徴量のみを使用した場合に比べ，感情ラベルを併用することで，WAR では 34.0% から 40.2% と 6.2 ポイント改善し，UAR では 23.1% から 26.6% と 3.5 ポイント改善した。しかしながら，この性能は音響特徴量を使用せず感情ラベルのみを使用した場合に比べ低いものとなった。感情ラベル併用の効果をより詳細に検討するため，感情カテゴリごとの再現率を表３の二重線で区切った右側に示す。受容，恐れ，驚き，悲しみ，嫌悪，喜びの各感情では，音響特徴量のみを使用した場合に比べ，音響特徴量と感情ラベルを併用した場合により高い再現率が得られていることがわかる。しかしながら，受容，嫌悪，喜びの各感情では，むしろ音響特徴量を使用しないほうが再現率は高くなっている。

［今後の研究の方向，課題］

本報告では，プロジェクト期間内に検証を行った実験の一部について報告した。報告した内容以外にも実験を行い，検証してきた。成果

の発表として挙げた[1]では，学習用のコーパスが本来持っていない種類の感情ラベルを，音響特徴量から推定した感情ラベルで代用する手法を提案しているので，参照されたい。また，現在は音響特徴量のみを用いたモデルを検証したが，今後は言語情報も利用してモデルを作成し，さらなる精度向上を目指す予定である。

[参考文献]

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proceedings of Interspeech 2005*, 2005, pp. 3–6.

[2] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system," *Proc. Lr. 2012*, pp. 3369–3373, 2012.

[3] S. Haq and P. J. B. Jackson, "Multimodal Emotion Recognition," in *Machine Audition : Principles, Algorithms and Systems*, IGI Global, 2010, pp. 398–423.

[4] A. Batliner, C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russell, and M. Wong, "' You stupid tin box' — children interacting with the AIBO robot : A cross-linguistic emotional speech corpus.," in *Proceedings of LREC 2004*, 2004, pp. 171–174.

[5] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoust. Sci. Technol.*, vol. 33, no. 6, pp. 359–369, 2012.

[6] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *2008 IEEE International Conference on Multimedia and Expo*, 2008, pp. 865–868.

[7] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Commun.*, vol. 53, no. 1, pp. 36–50, Aug. 2011.

[8] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP : Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[9] F. Eyben, M. Woellmer, and B. Schuller, *openSMILE the Munich open Speech and Music Interpretation by Large Space Extraction toolkit*. 2010.

[10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proceedings of Interspeech 2010*, 2010, no. September, pp. 2794–2797.

[11] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab — An S4 Package for Kernel Methods in R," *J. Stat. Softw.*, vol. 11, no. 9, pp. 1–20, 2004.

[成果の発表，論文等]

[1] 永岡 篤，森 大毅，有本泰子，"感情音声コーパスへの異種感情ラベル自動付与における既存感情ラベル併用の効果，" 日本音響学会誌，2016.（投稿予定）

[2] H. Mori and Y. Arimoto, "Accuracy of Automatic Cross-Corpus Emotion Labeling for Conversational Speech Corpus Commonization," in *Proceedings of LREC2016*, 2016, pp. 4019–4023.

[3] 有本泰子，森 大毅，"クロスコーパス感情ラベリングによる対話音声の比較，" 日本音響学会 2016 年春季研究発表会講演論文集，pp. 359–360, 2016.

[4] 【招待講演】有本泰子，"コミュニケーション場面におけるリアルな感情表出の分析，" 日本音響学会 2015 年秋季研究発表会講演論文集，pp. 1317–1320, 2015.

[5] 永岡 篤，森 大毅，有本泰子，"複数の対話音声コーパスにおける感情ラベルの相互推定，" 日本音響学会 2015 年春季研究発表会講演論文集，pp. 415–416, 2015.

[6] 永岡 篤，森 大毅，"サポートベクター回帰による自発音声の感情次元推定，" 日本音響学会 2014 年秋季研究発表会講演論文集，pp. 389–390, 2014.