6Y-04

字句情報，音響情報，表情から推定した話者の感情の食い違い状況の分析と食い違い自動検出手法の提案

上村譲史† 目良和也‡ 黒澤義明‡ 竹澤寿幸‡

広島市立大学情報科学部†

広島市立大学大学院情報科学研究科‡

１．はじめに

これまで人間の感情を推定するためのさまざまな手法が提案されており，その情報源として主に，発話文字列，発話音声の音響的特徴，表情などが挙げられる．しかしほとんどの手法は単独の情報源だけから感情推定を行っている．マルチモーダルな情報を用いた感情推定手法[1]も提案されているが，それらも複数の情報を組み合わせて最終的に一つの感情を推定している．しかし，実際の人間の心理状態としては，皮肉やツンデレのように，各情報源から推定される感情に食い違いがあることによって，複雑な心理状態が表現されていることもある．そこで本研究では，実際の対話から収集したデータについて，発話文字列，音響的特徴，表情それぞれから推定される感情について質問紙調査を行い，それらの一致度および不一致時の状況について分析を行う．さらに，各情報源から感情を推定する手法を用いることで複雑な心理状態の自動推定が可能かについても検討を行う．

２．人間による各情報源からの感情推定

本研究では，感情表出を伴う発話シーンを収集するため，被験者（互いに知り合いである男子大学生 9 名）に教示を与えない状態で雑談をしてもらい，その様子をビデオカメラで撮影した．音声はヘッドセットマイクで同時録音した．そして 1 発話ごとにトリミングした動画 400 件を収集した．1 発話の平均時間は約 2.9 秒である．

収集した 400 件の発話データに対して，前述の雑談者とは異なる男子大学生 5 名が感情のアノテーションを行った．感情は positive(喜)，neutral(無)，anger(怒)，sad(悲)の 4 クラスで，5 人中 3 人以上の回答が一致した感情を正解タグとした.意見が分かれたものは“正解無し”として，分析及び実験には用いなかった．

Analysis of Inconsistency among Emotions Estimated from Linguistic, Acoustic, and Facial Expression Features and A Proposal of the Inconsistency Detecting Method

Joji UEMURA, Kazuya MERA, Yoshiaki KUROSAWA, and Toshiyuki TAKEZAWA, Hiroshima City University, Japan.

本研究では各情報源からの感情推定傾向の一致度や食い違いについて分析するため，アノテータは一つの発話データに対して以下の 4 パターンの情報提示方式でアノテーションを行った．

- 映像を見ずに音声を聞く（音声）
- 音声がない状態で映像を見る（表情）
- 話者の発言したテキストを読む（言語）
- 音声付きの映像を見る（音声+表情+言語）

各情報源からの感情推定結果を表 1 に，『“音声”と“音声＋表情＋言語”』，『“表情”と“音声＋表情＋言語”』，『“言語”と“音声＋表情＋言語”』の感情推定結果の比較を表 2，3，4 に示す．

表 1： 各情報源からの感情推定結果

	音声	表情	言語	音声+表情＋言語
positive	97	156	14	103
neutral	109	190	197	131
anger	74	16	91	45
sad	83	8	55	69
正解無し	37	30	43	52

表 2：“音声”と“音声＋表情＋言語”の感情推定結果の比較

		音声			
		positive	neutral	anger	sad
音声＋表情＋言語	positive	87	3	5	0
	neutral	3	89	12	17
	anger	1	4	36	2
	sad	1	5	4	54

表 3：“表情”と“音声＋表情＋言語”の感情推定結果の比較

		表情			
		positive	neutral	anger	sad
音声＋表情＋言語	positive	101	0	0	0
	neutral	21	94	3	3
	anger	8	27	9	0
	sad	6	45	3	5

表 4：“言語”と“音声＋表情＋言語”の感情推定結果の比較

		言語			
		positive	neutral	anger	sad
音声＋表情＋言語	positive	8	54	19	6
	neutral	6	95	13	10
	anger	0	6	33	2
	sad	0	22	10	28

表１より”音声”では各感情の出現傾向に大きな偏りはみられないが，”表情”は anger と sad の感情があまり出現しないことが確認された.また“言語”は positive の感情が少なかった.表 2，表 3，表 4 より“音声＋表情＋言語”と”音声”の感情の一致率は 0.70 と最も高く,”言語”との組み合わせの一致率は 0.43 と最も低かった.

次に，単独の情報源からの感情推定結果と全ての情報を使っての感情推定結果の相関を表 5 に示す．『“音声”と“音声+表情+言語”』，『“表情”と“音声+表情+言語”』，『“音声”と“表情”』の順に相関が高いことが分かった.また，“言語”は他の要因とあまり相関が高くないことが分かった．これより，聞き手が話者の感情を推定するのに重要な情報は音声＞表情＞言語の順だと考えられる.

また，“音声”からの感情推定結果と“表情”からの感情推定結果の比較を表 6 に示す.“音声”と”表情”では,positive と neutral は大多数が一致しているが，anger と sad の音声は表情からだとほとんど neutral に分類されている.また positive 表情の発話における音声の一部は neutral か anger に分類されている傾向があった.

表 5:アノテーションごとの感情推定結果の相関

	音声	表情	言語	音声+表情＋言語
音声	1			
表情	0.52	1		
言語	0.35	0.12	1	
音声+表情+言語	0.81	0.57	0.37	1

表 6：人手による“音声”と“表情”の感情推定結果

		音声			
		positive	neutral	anger	sad
表情	positive	94	18	19	4
	neutral	1	76	43	59
	anger	0	4	7	5
	sad	0	2	0	6

３．感情検出手法と人間の推定結果の比較

本節では，感情推定する上で重要度の高い“音声”と“表情”それぞれについて感情推定手法を適用し，人間の感情推定結果との比較を行う．音響情報からの感情推定手法として，INTERSPEECH2009 Emotion Challenge[2]で用いられた 384 種類の特徴量を機械学習器 SVM に学習させたものを用いた．また，表情からの感情推定手法としては，オムロンの OKAO Vision[3]によって 0.1 秒ごとの静止顔画像から算出された推定感情（無，喜，驚，怒，悲）のうち驚を省いた 4 つの感情について動画全体で一度でも推定結果として出力された感情すべてを動画からの推定感情とした．なお，Leave-One-Out による音響情報からの感情推定手法の正解率は 0.60,表情からの感情推定結果と人間のアノテーション結果の一致率は 0.55 であった.

音声を用いた感情推定手法の結果と，動画中で出現回数が最大の感情を正解とする表情推定手法の結果の比較を表 7 に示す．表 7 から表情の推定結果は neutral にほとんど分類されておらず，また sad の表情は，positive 以外の音声に多く分類された.音声と表情では positive が最も一致しやすいが，表情で positive のデータは音声の neutral にも多く分類されている.

また，音声，表情それぞれの感情推定手法の結果が一致したデータと“音声+表情+言語”のアノテーション結果の比較を表 8 に示す.positive は精度，再現率ともに 0.80 以上の高い値を示した.しかし sad と neutral は音声と表情の両方の情報を合わせても判別しにくいと考えられる.

表 7：推定手法による“音声”と“表情”の感情推定結果

		音声（推定手法）			
		positive	neutral	anger	sad
表情（推定手法）	positive	71	33	15	14
	neutral	3	8	0	3
	anger	14	17	4	4
	sad	1	52	22	18

表 8：感情推定手法結果とアノテーション結果の比較

		感情推定手法結果				
		positive	neutral	anger	sad	再現率
アノテーション	positive	66	0	2	2	0.94
	neutral	8	14	4	21	0.30
	anger	5	3	11	7	0.42
	sad	3	4	1	31	0.79
	精度	0.80	0.67	0.61	0.51	0.67

謝辞

本研究は国立研究開発法人科学技術振興機構(JST)の研究成果展開事業「センター・オブ・イノベーション(COI)プログラム」及び JSPS 科研費 26330313 の助成を受けたものです．また，オムロン（株）から画像センシング技術 OKAO(R)Vision をご提供いただいております.

参考文献

[1] M. Kurisu et al.,“A Method using Linguistic and Acoustic Features to Detect Inadequate Utterances in Medical Communication,” Proc. of IWCIA2013, pp.197-200 (2013).

[2] B. Schuller et al.,“The INTERSPEECH2009 Emotion Challenge,” Proc. of INTERSPEECH2009, pp.312-315 (2009).

[3] OKAO Vision | オムロン人画像センシングサイト, http://plus-sensing.omron.co.jp/ technology/, (2016/01/03 アクセス).