言語，音声，表情から推定した話者感情の一致･食い違い状況の分析と食い違い自動検出手法の提案

Analysis of consistency and inconsistency among emotions estimated from linguistic, acoustic, and facial expression and a proposal of emotion inconsistency situation detecting method

上村　譲史
Joji Uemura

広島市立大学大学院　言語音声メディア工学研究室
Language and Speech Research Laboratory, Graduate School of Information Sciences, Hiroshima City University

概要　本研究では，対話中の言語，音声，表情の各情報源に対し質問紙調査を行い，感情の一致・食い違い状況の分析を行う．また，言語，音声，表情からそれぞれ感情推定を行うことにより，感情の食い違いの自動検出手法を提案する．言語は評価極性辞書と文法から感情推定する．音声は音響的特徴から感情推定を行う．表情はフレームごとに感情推定を行い，一度でも出現した感情を用いる．

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いる． そしてその情報源としては主に，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．しかしほとんどの研究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アノテーションを行い，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　人間による感情推定結果の一致・食い違い状況の分析

本研究では，発話者の感情推定を行うため，ビデオカメラとヘッドセットマイクで雑談等の動画の収集を行った．対象者は男子大学生 21 歳から 23 歳の 9 名，収集した動画は発話ごとにトリミングを行い， 400 件の発話動画データを収集した．

2.1 アノテーションによる実験

収集した 400 件のデータを~~動画~~対話収集対象者とは別の男子大学生 21 歳から 23 歳の 5 名に感情分類の 4 つ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 人以上が一致する感情をそれぞれの質問の正解とした．質問紙調査は以下の 4 パターンで行った．

1. 音声:映像無しで音声を聞く
2. 表情:音声がない状態で映像を見る
3. 言語:話者の発言したテキストを読む
4. 表情+音声+言語:音声付きの映像を見る

表 1 に「『音声』と『言語』」のアノテーション結果を示す．

表 1: 「『音声』と『言語』」のアノテーション

		言語			
		positive	neutral	anger	sad
音声	positive	8	55	13	7
	neutral	6	81	5	10
	anger	0	13	53	2
	sad	0	35	7	33

表 1 から音声が positive かつ言語で anger のデータ 13 件に対して，例えば「バカじゃねえの（笑）」のような“笑いながらバカにしている発言”は 8 件確認できた．またその 8 件に対し表情の分類結果を調査した結果，8 件中全てが positive に分類されていることが分かった．この結果から複数の情報源から感情を推定することで複雑な心理状態を検出することが可能だと考えられる．

アノテーションにより分類したデータのまとめを表 2 に示す．また，表 3 に分類したデータでの感情の一致率を示す．「『音声』と『表情+音声+言語』」，「『表情』と『表情+音声+言語』」，「『言語』と『表情+音声+言語』」の順番で相関が高いことが分かった．よって，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．そのため次節では，感情を推定する上で順位の高い『音声』と『表情』それぞれについての感情推定手法を検討する．

表 2:アノテーション結果の感情タグ付きデータ(400 件)

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

表 3:アノテーションごとの一致率

質問紙調査の組み合わせ	一致率
『音声』と『表情+音声+言語』の感情のペア	0.70
『表情』と『表情+音声+言語』の感情のペア	0.54
『言語』と『表情+音声+言語』の感情のペア	0.43

3　感情推定手法を用いた食い違い状況の自動検出

3.1 音響情報からの感情推定

音声的な処理として，音響分析ツール openSMILE[2]を用いて音響分析を行い，音圧，メル周波数ケプストラム（12 段階），ゼロ交差率，有声確率，基本周波数のそれぞれ 12 種類の素性とそれらの⊿成分，合計で 384 種類の素性を音響的特徴量として抽出する．抽出した特徴量を機械学習器 Support Vector Machine (SVM)に入力し，発話を音響的に 4 クラスの感情(positive, neutral, anger, sad)に分類する．

3.2 表情からの感情推定

表情推定ツールとしてオムロン社の OKAO Vision[3]を使い，映像を 0.1 秒刻みのフレームに分割し，そのフレームごとに 5 クラスの感情（無，喜，驚，怒，悲）の出現確率推定を行う．本研究では，各フレームにおいて一度でも最尤感情と推定されたものは全て，その発話シーンから推定される感情として用いる．

4　評価実験

4.1 機械学習による感情推定の食い違いの検出

アノテーションにより正解の感情タグを付けた表 2 の音声データ 363 件を用い感情推定の一致・食い違いの実験を行う．音声は機械学習器 SVM により感情推定を行う．また，表情は表情推定ツール OKAO Vision を用いて各フレームごとに表情推定を行い，最尤感情と推定されたものは全て，その発話シーンから推定される感情として用いる．

4.2 感情推定結果と『表情＋音声＋言語』の比較

音声，表情それぞれ感情推定し，音声のみで感情推定した結果と表情のみで感情推定した結果，音声と表情で感情推定結果が一致したデータ（AND）の 3 パターンに対し，アノテーション『表情+音声+言語』と比較した結果の精度，再現率，F 値を表 4 に示す．

表 4 より，音声のみから感情推定した結果と表情＋音声＋言語の結果 F 値が 0.54 と分類結果はあまり高くないことが分かった．表情のみの場合はフレームごとに出現したすべての感情を分類に用いるため精度が極端に低いことが分かる．

音声と表情の結果が一致(AND)が 3 パターンの間でもっとも F 値が高かったことから，音声，表情それぞれ別々に感情推定を行うよりも組み合わせたほうが感情の　一致率が高いことが分かる．

表 4:提案手法によって推定した感情と人間が推定した感情の一致

	精度	再現率	F 値
音声のみ	0.54	0.54	0.54
表情のみ	0.29	0.62	0.40
音声と表情が一致	0.65	0.61	0.63

4.3 感情推定結果と『表情＋音声＋言語』の感情の食い違い状況の検出

音声と表情の感情の食い違いについて調査を行うため，前節の手法から，音声と表情の感情推定の組み合わせとアノテーション『表情+音声+言語』の組み合わせを比較する．neutral は考慮せず，anger と sad はまとめて negative のクラスとした結果を表 5 に示す．

168 件中 23 件の感情の食い違いデータに対して再現率は 0.43 となった．

表 5:提案手法による感情の食い違い状況の検出

		人間による音声と表情からの感情推定結果			
		pos+pos	pos+neg	neg+neg	精度
提案手法による音声と表情からの感情推定結果	pos+pos	66	7	2	0.88
	pos+neg	54	10	8	0.14
	neg+neg	7	6	8	0.38
	再現率	0.52	0.43	0.44	0.50

5　おわりに

本研究では，アノテーションにより聞き手が発話者のどの情報から感情を判断しているのか調査し，学習器 SVM と表情推定ツール OKAO Vision を用いて，音声と表情から発話者の感情の推定を試みた．結果として聞き手は話者の音声>表情>言語の順に感情を推定する指標にしていると考えられる．また音声，表情の感情が一致する場合もあるが，それぞれ違う感情出現することも考えられる．今後は提案手法の表情推定の改良と，言語からの感情推定手法の考案と，複合感情による評価方法を検討し，最終的には発話から 3 種類の感情をリアルタイムで解析するシステムを構築する予定である．

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

6　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/, (2016/07/05 アクセス).