音声に明確に表出しない感情の認識のための感情音声データベースの構築*

○齋藤晶太，目良和也，△黒澤義明，竹澤寿幸（広島市立大），
瀧上順也，鈴木芳典（NTT ドコモ先進技術研究所）

1　はじめに

日常生活において，私達は相手の感情を考えながら行動している．相手が喜ぶことや嫌がることを，反応を伺い本音を推測しながら働きかけを行う．このように相手の感情を正しく認識して対応しようとする行為は，私達が日常生活でごく自然に行っていることである．感情認識を必要とするのは，もはや人間に限った話ではない．世の中では人の感情を理解するロボットやエージェントも登場し始めており，そう遠くない将来，彼らが人の感情を汲み取って，気配りの利いたサービスを提供する社会が訪れることが期待されている．

このような状況を受けて，現在，コンピュータで感情を推定する研究が盛んに行われている．感情推定に用いるモダリティとしては，表情，発話文字列，声の音響的特徴など様々なものがあるが，いずれの研究においても判別がつきやすい“明確に感情が表出された”事例を学習や評価対象に用いている．

しかしながら私達は，日常生活で生じる様々な感情を全て外側に表出しているわけではなく，時には感情をしまいこみ，表出しないようにコントロールする場合がある．そのため，相手の感情を汲み取って気配りのある対応をするためには，このような明確に表出しない感情も認識できる必要がある．

そこで本研究では，音声から人の感情を認識する音声感情認識に注目し，音声には明確に表出していない感情を音声から認識することをねらいとする．本研究では，“音声に明確に表出しない感情”を『前後の文脈を含むマルチモーダルな話者情報から総合的に話者が特定の感情を生起していることが確認できるが，発話音声単独からは感情推定が困難な感情表出状態』と定義する．そこで収集した発話音声およびその前後を含む映像を評価者に提示し，それぞれからの話者感情推定結果を“音声単独から推定した話者感情”，“総合的に推定した話者の本心の感情”としてラベル付けしている．

本論文では，データベースの構築手法および機械学習による感情推定実験結果を示す．

2　感情音声データベースの構築

2.1 音声に明確に表出しない感情の定義

“音声に明確に表出しない感情”を研究対象として収集するためには，各発話音声データに対して，**“本心の感情”**と**“音声単独から推定される感情”**の二種類の感情情報が必要である．そして，“本心の感情”が“音声単独からの感情”と一致している場合，当該音声データはその感情を音声に明確に表出していると定義する．逆に，“本心の感情”が“音声単独からの感情”と異なっている場合，当該音声データはその感情を音声に明確に表出していないと定義する．

“音声単独から推定される感情”は，文字通り音声のみ（**Voice-only**）から評価者が感情評定した結果に基づいて算出している（2.3 節参照）．一方，“本心の感情”は，前後の文脈を含むマルチモーダルな話者情報（**Multimodal**）から総合的に評定している（2.4 節参照）．なお，統計的な信頼性を担保するため，本研究では 20 名以上の評価者により感情評定を行っている．

本定義に基づく音声に明確に表出しない Happiness の例を Fig. 1 に示す．対象は「やばい！」という音声データである．Voice-only の評価では“恐れ”と評価されるが，Multimodal の評価では，喜んでいるような表情や花火を観ている状況を総合的に踏まえて，“喜び”と評価される．この時，話者の本心

---

* Designing emotional speech corpus to recognize “hidden” emotion from acoustic features, by SAITO, Shota, MERA, Kazuya, KUROSAWA, Yoshiaki, TAKEZAWA, Toshiyuki (Hiroshima City University), TAKIUE, Junya, and SUZUKI, Yoshinori (Research Laboratories, NTT DOCOMO. INC.).

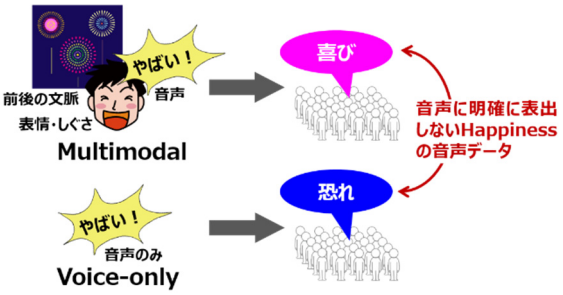


Fig. 1 Example of hidden Happiness voice data

の感情は“喜び”であると推定されるが，音声にはその感情が明確に表出していないため，音声だけを聞いても正解を導くことが難しいと考えられる．このようにして，音声に明確に表出しない感情を特定する．

2.2 収集した音声データの仕様

本研究で構築するデータベースは，インターネット上から収集した 1,825 件の自発音声に評価者 20 名または 25 名が感情ラベルを付与するものであり，各発話音声データに対して，音声のみを参照して行う感情ラベル付け（Voice-only）と，発話者の表情やしぐさ，発話の前後の文脈が分かる映像全体を参照して行う感情ラベル付け（Multimodal）を行う．そしてこれらの感情ラベル付けの結果を比較することにより，音声には明確に表出しない感情の音声データを特定する．

本データベースに収録している音声は，以下の条件を満たす 1,825 件である．

- 話者の性別や年代に偏りがないこと
- 演技でない音声であること．ニュース，ドラマ，舞台など決められたセリフの発話や特定のキャラクターを演じている音声などは含めない
- 対象とする話者以外の音声（効果音，BGM，背景雑音，対象話者以外の相槌や笑い声など）が含まれていないこと
- 一発話内で感情の種類が変化しないこと
- 音声の品質が悪くないこと

また，対象の発話音声だけでなく，発話の前後数十秒を含む映像も併せて収集した．

2.3 発話音声単体からの感情評定

本データベース中の音声データは，感情の種類ごとに 4 段階で生起強度を評定された．評価者は発話者と異なる 20 歳から 47 歳の男女 20 名（男性 10 名，女性 10 名）または 25 名（男性 12 名，女性 13 名）であり，性別や年齢に偏りがないように評価者を選定した．感情の種類は Ekman の基本 6 感情[1]（怒り，嫌悪，恐れ，喜び，悲しみ，驚き）を採用した．各感情の強度レベルは，以下の 4 段階の言語表現を被験者に提示して評定された．なお評定の際には，X に{怒っている，喜んでいる，悲しんでいる，恐れている，驚いている，嫌っている／嫌がっている}をそれぞれ入れて被験者に提示した．

0：なし（＝知覚できない）
1：やや X
2：X
3：非常に X

被験者は各音声データについて 6 感情それぞれの強度レベルを回答する．そのため，話者にいずれの感情も生起していないと感じた場合は6種類の感情全て強度0と回答される．

2.4 マルチモーダルな情報からの感情評定

本データベースでは，対象となる音声データだけでなく，音声データの前後数十秒を含む映像についても同様に話者の生起感情を評定する．先に，2.3 節の手順に基づき評価対象となる音声データから話者感情を評定する（Voice-only）．このとき，評価者は音声から感じられる話者の感情を感じたまま評価し，発話の内容は参考程度とするよう教示した（例：怒ったエピソードでも喜んで話していると感じたら“喜び”と評定する）．その後，評価対象となる音声が含まれる映像をその前後を含めて視聴し，対象箇所で話者がどのような感情であるかを評定し，ラベル付けを行う（Multimodal）．その際，評価者は発話の音響的特徴だけでなく，話している内容や前後の文脈，シチュエーション，表情やしぐさを踏まえて総合的に感情を判断する．評価対象となる音声データ（Voice-only）と映像で提示される発話前後を含むシーン（Multimodal）の関係を Fig. 2 に示す．

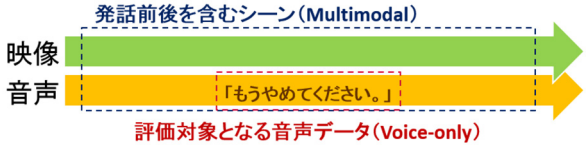


Fig. 2 Voice-only and Multimodal data

2.5 感情ラベル決定方法

本節では, 2.3 節および 2.4 節の結果から音声データの感情ラベルを決定する方法について述べる．各音声データに付与するラベルは下記の 2 種類とした.

(A) 基本 6 感情それぞれの生起の有無
- ・怒り　{Angry or Not Angry}
- ・嫌悪　{Dislike or Not Dislike}
- ・恐れ　{Fear or Not Fear}
- ・喜び　{Happiness or Not Happiness}
- ・悲しみ　{Sadness or Not Sadness}
- ・驚き　{Surprise or Not Surprise}

(B) 各感情の有無が音声に明確に表出している（Explicit）か明確に表出していない（Hidden）か

(A)のラベルを付与するために, まず評定者からの 4 段階の回答について「レベル 0：対応する感情なし（Not X)」,「レベル 1〜3：対応する感情あり(X)」の 2 クラスに分類する．そして，統計的に評価者の過半数に支持された場合のみ，各感情の有無（X or not X）をラベル付けした. 具体的には評価者 20 名の場合 14 名以上, 評価者 25 名の場合 17 名以上支持された場合のみ感情ラベル付けを行い，所定の人数に満たない場合, 感情は不定（neither）として扱った.

(A)のラベル付けを Voice-only と Multimodal 両方の回答について行った．Voice-only および Multimodal の評定における各ラベルの付与データ数をそれぞれ Table 1 と Table 2 に示す.

次に, Explicit／Hidden のラベル付け方法について述べる．本研究では，各感情について Multimodal の感情ラベルが Voice-only の感情ラベルと一致している場合，当該音声データはその感情を音声に明確に表出する（**Explicit**）音声データと定義した．逆に，Multimodal の感情ラベルが Voice-only の感情ラベルと異なっている場合，当該音声データはその感情を音声に明確に表出しない（**Hidden**）音声データと定義した．X／not X ラベルに基づく Explicit／Hidden 感情の定義を Table 3 に示す.

この定義に基づき，収集した 1,825 件の音声データに対して Explicit／Hidden のラベル付けを行った. データ数の内訳を Table 4 に示す.

Table 1 Number of labeled data (Voice-only)

ラベル		データ数	計
Angry	X	458	1,825
	not X	969	
	neither	398	
Dislike	X	485	1,825
	not X	796	
	neither	544	
Fear	X	175	1,825
	not X	1,209	
	neither	441	
Happiness	X	278	1,825
	not X	1,211	
	neither	336	
Sadness	X	264	1,825
	not X	1,146	
	neither	415	
Surprise	X	329	1,825
	not X	1,031	
	neither	465	

Table 2 Number of labeled data (Multimodal)

ラベル		データ数	計
Angry	X	336	1,825
	not X	1,255	
	neither	234	
Dislike	X	475	1,825
	not X	920	
	neither	430	
Fear	X	178	1,825
	not X	1,296	
	neither	351	
Happiness	X	444	1,825
	not X	987	
	neither	394	
Sadness	X	278	1,825
	not X	1,236	
	neither	311	
Surprise	X	276	1,825
	not X	1,123	
	neither	426	

Table 3 Definition of Explicit/Hidden emotion

		Multimodal		
		X	neither	not X
Voice-only	X	Explicit X	none of them	Hidden not X
	neither	Hidden X		
	not X			Explicit not X

Table 4 Number of labeled data (Explicit/Hidden)

		Explicit X	Explicit not X	Hidden X	Hidden not X
X	Angry	291	925	45	330
	Dislike	315	682	160	238
	Fear	104	1,069	74	227
	Happiness	225	883	219	104
	Sadness	181	1,029	97	207
	Surprise	216	880	60	243

3　機械学習による感情推定実験

従来の感情推定処理では，声だけで判別可能な事例を対象として研究を行っていた．つまり，Table 3 における（Voice-only = X）vs（Voice-only = not X）の識別を行っていたことになる．しかし，**“話者の声がどう聞こえるか”**ではなく**“話者が実際にどう感じているか”**を知るためには，（Explicit X + Hidden X）vs（Hidden not X + Explicit not X）を識別すべきである．

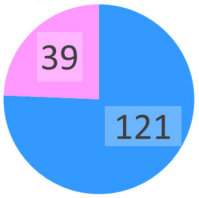
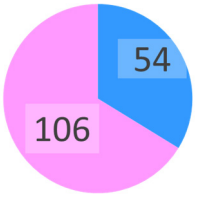
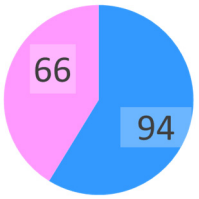
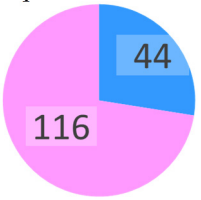
そこで本稿では本心推定手法の第一歩として，今回構築したデータベースのうち嫌悪（Dislike）のデータを対象として Support Vector Machine (SVM)を用いた判別実験を行った．

機械学習に使用する素性値は openSMILE[2]で算出される The INTERSPEECH 2009 Emotion Challenge feature set (IS09)を用いた．IS09 は，32 種類の特徴量（音のエネルギー，メル周波数ケプストラム[1]～[12] ，基本周波数，声である確率，ゼロ交差率およびこれらの一次微分）の変化曲線から算出される 12 種類の基本統計量（最大値，平均値，レンジなど）計 384 種類の特徴量から構成される．また，機械学習を行う際はデータ数に偏りが無いように，データ数が最も少ない Hidden Dislike の 160 個に合わせて他のクラスのデータをランダムにダウンサンプリングした．

Multimodal 情報からの話者感情推定結果を“本心”の感情と定義し，本心の嫌悪（Explicit Dislike + Hidden Dislike）と，本心から嫌悪でない（Hidden not Dislike + Explicit not Dislike）状態を識別する学習器を SVM により構築した．Leave-one-out 方式で実験を行った結果を Table 5 に示す．

Table 5 より，Dislike および not Dislike 推定の精度はそれぞれ 0.69, 0.68 であった．また，本心の Dislike および not Dislike の再現率はそれぞれ 0.67，0.69 であった．そして全体の正解率は 0.68 であった．データ種別では，Explicit より Hidden の正解率が低いことが確認された．これは本心としてアノテーションされた話者感情の推定において，音響的特徴から推定される感情が影響しているためと考えられる．しかし，Hidden なデータでも 6 割程度以上の正解率が得られていることから，今後，有効な特徴量の絞り込みや Hidden データの音響的特徴の分析などにより，更なる正解率向上が期待できる．

Table 5 Experimental result to detect “Dislike”

	Multimodal（本心）	
	Dislike	not Dislike
Data type	Explicit Dislike  正解率 0.76	Hidden not Dislike  正解率 0.66
	Hidden Dislike  正解率 0.59	Explicit not Dislike  正解率 0.72
再現率	0.67	0.69

4　おわりに

本研究では，音声に明確に表出しない感情を認識するための感情音声データベースを構築した．さらに，Dislike を対象として SVM を用いた音声に明確に表出しない感情の識別実験を行なった．その結果，全体の正解率は 0.68 となり，また Explicit より Hidden なデータの正解率が低いことが確認された．

今後の課題としては，特徴量の絞り込みやデータ分析などにより，音声に明確に表出しない Hidden な各感情状態の検出精度の向上を目指す．

参考文献

[1] P. Ekman and W. V. Friesen, Unmasking The Face: A Guide to Recognizing Emotions from Facial Expressions, Malor Books, 2003.

[2] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor,” In Proc. ACM Multimedia (MM), pp.835-838, 2013.