音声チャットを利用したオンラインゲーム感情音声コーパス*

◎有本泰子[1], △河津宏美[2] (東京工科大)

1 はじめに

これまでの感情音声研究は演技経験者などが感情を込めた発話を発声する手法により収集した音声を使って多くの成果を挙げてきた．こうした感情音声の多くは表現した感情が聞き手に知覚可能である必要があるため，感情表現が大げさとなり，日常生活で人間が表出している感情表現とは異なると言われている．このため，自発的な感情表出から話者の感情を認識することを目的とした研究には演技音声ではなく，対話中に自然に表出した感情音声の収集が必要となる．しかしながら，過去の自発音声を用いた感情音声研究では，収録した対話に含まれる感情音声の絶対数が少ない [1]．筆者らは，オンラインゲーム中のプレーヤーに音声チャットを利用させ，自然に感情が表出した音声を含む約 1 万発話を収集した．さらに，これと言語内容がほぼ一致する演技音声を 2656 発話収録し，自発音声と演技音声とを比較可能な感情音声コーパスを構築した．

本稿では，本コーパスの概要を説明するとともに，本コーパスの自発対話音声と既存の自発対話音声コーパスとを定量的に比較する．さらに，これまで実証されてこなかった演技による感情音声と自発的な感情音声との音響的な差について，本コーパスを用いて検証を行う．

本コーパスに含まれる音声・転記テキスト・感情ラベルの一部は，独立行政法人情報処理推進機構（IPA）2007 年度第 II 期未踏ソフトウェア創造事業（未踏ユース）「オンラインゲームにおける匿名性を有した音声チャットの開発」の支援により収録・作成されたものである．構築したコーパスは，オンラインゲーム感情音声コーパスとして，情報・システム研究機構国立情報学研究所音声資源コンソーシアム（NII-SRC）より無償で公開する予定である．

2 自発対話音声[3]

2.1 音声収録

対話中に自然に表出した感情音声を効率的に収集するために，オンラインゲームおよび音声チャットを用いた音声の収録を行った．感情が表出しやすい環境を作るため，日常生活とほぼ変わらないゲーム環境において，親近性のある話者間で対話収録を行った．感情誘発手法としてオンラインゲームを用い，誘発した感情を音声に表出させるために音声チャットをプレーヤー間のコミュニケーション手段として導入した．

Fig. 1 に収録環境図を示す．音声提供者は 2 名あるいは 3 名で参加し，学内に配置したの各収録サイトにてオンラインゲームに参加させた．ゲーム中のコミュニケーション手段となる音声チャットアプリケーションには Skype を用い，音声提供者間の音声を Skype

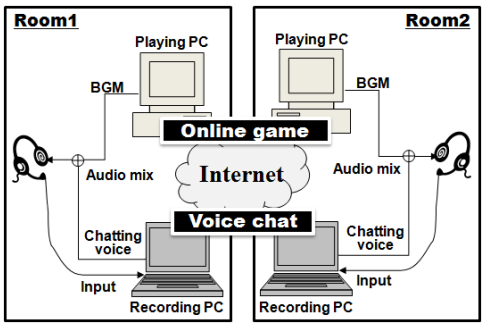

Fig. 1 Recording environment [2].

用の録音ツールである tapur を用いて録音した．音声提供者にはヘッドセットマイクを装着させ，口とマイクの位置を一定に保った．音声対話の種類は二者対話が 5 組，三者対話が 1 組である．

音声提供者はオンラインゲームの経験がある大学生 13 名（男性 9 名，女性 4 名）である．出身地は，東京 6 名，長野 2 名，神奈川／静岡／山梨／青森各 1 名，未回答 1 名である．オンラインゲーム歴は平均 38ヶ月（12〜61ヶ月），1ヶ月のプレイ時間は平均 33 時間／月（0〜100 時間）である．収録に利用したオンラインゲームは音声提供者が普段プレイしているゲーム（ラグナロクオンライン，モンスターハンターフロンティア，レッドストーン）である．

2.2 転記テキスト

収録した音声データに対して，400ms 以上のポーズによってはさまれた音声の範囲を発話単位として，発話のセグメンテーションを行なった．発話内容は漢字かな交じりの正書法で記述している．オンラインゲームに特有の用語（BOT や STRAGE）や数字・数詞はその読みをカタカナで表記した．400ms よりも短いポーズには読点（、）を，言語的な文末を表す箇所は句点（。）を付与した．文末基準は日本語話し言葉コーパスの絶対境界までを対象としている．笑い声や，咳など発話以外の言語音に対してはタグ付けを行っている．以下に付与したタグを示す．

- { 笑 }, { 咳 }
  笑い声（笑いながらの発話は除外）と咳．
- (?), (?(コメント))
  よく聞き取れない音声，自信がない場合．(コメント）の個所にコメントを記述．
- [comment=(コメント)]
  作業者のコメントがある場合．イコール以下の（コメント）の個所にコメントを記述．

Table 1 に話者ごとの発話数を，Table 2 に対話例を示す．Table 2 における句読点・エクスクラメーショ

---

*Online game emotional speech corpus using voice chat. by ARIMOTO, Yoshiko[1], KAWATSU, Hiromi[2] (Tokyo University of Technology).

[1] Current Affiliation: JST, ERATO, Okanoya Emotional Information Project and RIKEN, BSI

[2] Current Affiliation: IBM Japan, Ltd.

[3] 本節の内容は [2, 3] から抜粋し，転載したものである．

Table 1　Number of utterances for each speaker [2].

Speaker	Utterances	Speaker	Utterances
01_MMK	816	04_MNN	934
01_MAD	740	04_MSY	938
02_MTN	884	05_MYH	464
02_MEM	736	05_MKK	539
02_MFM	557	06_FTY	712
03_FMA	561	06_FWA	781
03_FTY	452		
		Total	9114

Table 2　Dialog examples [2].

Dialog 1
A: えーっとショートカットキー。じゃあ3。 うおっと何何何？何？ちょっちょちょちょっこ、 これ付いてこうぜ。これ付いてこうぜ
B: うん？
B: 何？何があった？
A: { 笑 } ねえ、あ、あれ右右右、 すごいすごいすごいすごい
A: { 笑 }
B: あらー
A: これ低レベルパーティだよね？
B: すげえ。なんか懐かしいな
Dialog 2
B: おーっと危ね
B: これは
B: あやばい死ぬ { 笑 }。死ぬ { 笑 }
A: 死ぬ { 笑 }
A: { 笑 }
B: 危ね危ね
B: オッケー
A: 生命の危機を感じたよ

ンマークは読みやすさのために付与したものである．

2.3　感情種別ラベルと感情強度ラベル

収録した 9114 発話のうち，収録音声の振幅レベルが小さく，感情の評定に使用できないと判断した 2 名の 1009 発話，および音響的分析に影響を及ぼす転記用タグが付与されている 1527 発話を除外した 6578 発話に対して，感情種別のラベル付けを行なった．各発話に対して異なる 3 名の評価者が評価を行っている．感情種別ラベルには，Pluchik の立体構造モデル [4] の基本 8 感情と「平静」および「その他」の 10 個のラベルを用いた．評価者に提示した感情種別ラベルの概要を Table 3 に示す．評価の結果，2 名以上の評価者が一致した発話数とその比率（Partial），および 3 名の評価者が一致した発話数とその比率（Complete）を Table 4 に示す．Partial および Complete の一致率はそれぞれ 58.5% and 11.0%となり，ともにチャンスレベル（28%および 1%）を上回っている．

2 名以上の評価者の評定値が一致した 3845 発話から平静の発話（NEU，798 発話）と感情種別が付与できなかった発話（OTH，200 発話）を除いた 2847 発話に対して，感情の強度ラベルを付与した．評価者は 18 名の男女（男性 13 名，女性 5 名）である．感情強度の評価は 1（弱）–3（中）–5（強）の 5 段階からひとつを選択させている．発話ごとに感情強度の平均を求め，発話の感情強度とした．各発話の感情強度の分布を感情ごとに Fig. 2 に示す．

Table 4　Emotional labeling results [2].

States	Partial		Complete	
	Utterances	Percent.	Utterances	Percent.
ACC	303	4.6	27	0.4
ANG	237	3.6	60	0.9
ANT	427	6.5	69	1.0
DIS	335	5.1	45	0.7
FEA	142	2.2	33	0.5
JOY	595	9.0	174	2.6
SAD	243	3.7	49	0.7
SUR	565	8.6	177	2.7
NEU	798	12.1	116	1.8
OTH	200	3.0	30	0.5
Total	3845	58.5	780	11.0

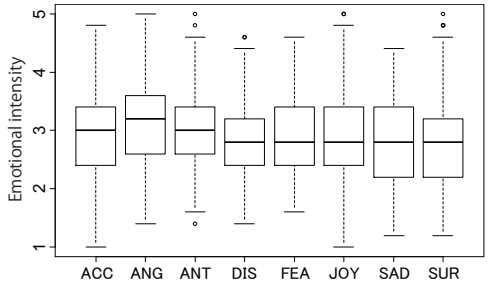

Fig. 2　Distribution of emotional intensities of utterances for each emotional state [2].

2.4　既存コーパスとの比較

本コーパスの自発対話音声と先行研究で用いられている既存の 3 つの感情音声コーパスとの比較を行なう．比較対象としたコーパスは，それぞれ収録方法が異なり，ゲーム中の音声を収録した TNO-Gaming Corpus（TNO）[5]，四コマ漫画並べ替え課題により音声を収録した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）[6]，DARPA communicator project のもとで開発された機械との疑似対話音声に感情ラベルを付けた Ang らのデータ [7] である．Table 5 に各コーパスの特徴を示す．

各コーパスの発話数を比較するために 1 分間あたりの平均発話数を求めた．Fig. 3 (a) にその結果を示す．Ang らのデータは収録時間数が不明なため，単位時間当たりの発話数が求められなかった．そのため，本比較から除外した．単位時間当たりの発話数が最も多いコーパスは UUDB（37.23 発話/分）となった．次いで本コーパスが多く（25.32 発話/分），発話数が最も少ないコーパスは TNO（13.34 発話/分）となった．UUDB と本コーパスは非対面対話であり，音声以外のコミュニケーションチャネルがないため，発話数が多くなったと考えられる．TNO は対面対話であるため，コミュニケーションチャネルが音声以外にも存在し，情動の表出が音声に集中せず，本コーパスと同じ感情誘発手法を利用しているにもかかわらず，非対面対話である本コーパスよりも発話数が少なくなったと考えられる．

コーパスに含まれる音声から聞き手が知覚する感情の割合をコーパス間で比較するために，評価対象発話に対して付与された総ラベル数に含まれる感情ラベルの割合を求めた．Fig. 3 (b) にその結果を示す．TNO はラベル基準およびその手法が本コーパスを含めた他の 3 つのコーパスと異なるため，本比較から除外している．感情種別ラベルが最も多く付与されたコーパスは本コーパス（73.0%）となった．次いで UUDB

Table 3　Abbreviations and definitions of emotional states [2].

記号	感情	説明
ACC	受容	心がひきつけられ，積極的に受け入れよう，接し続けようとする感情
ANG	怒り	許しがたい事柄に接し，不快感を抑えきれず，いらだった状態の感情
ANT	期待	望ましい事態の実現，好機の到来を心から待つ感情
DIS	嫌悪	その状態・行為をすんなりと受け入れることができず，避けようとする感情
FEA	恐れ	危害が及ぶことを心配してびくびくし，その人やその物と接することを避けたがる感情
JOY	喜び	良いことに出会って非常に満足し，うれしい，ありがたいと思う感情
SAD	悲しみ	不幸なことに出会った時など，取り返しのつかない事を思い続けて泣きたくなる感情
SUR	驚き	意外なことを見聞きして心が強く動揺し，平静を失う，どう判断すべきか戸惑う感情
NEU	平静	まったく感情が表れていない
OTH	その他	ノイズが大きい場合など 8 種の感情に分類不能のもの

Table 5　Characteristics of each corpora.

	Emotion induction technique	
Environment	Game	Task-oriented
Non-Face-to-Face	This material	UUDB Ang et al.
Face-to-Face	TNO	

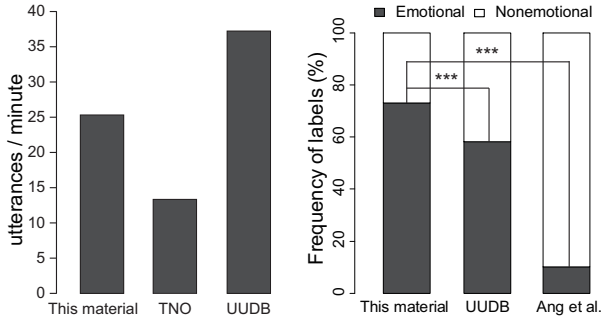

Fig. 3　Comparison with the existing corpora [3, 2]

が多く（58.2%），感情種別ラベルが最も少ないコーパスは Ang らのデータ（9.9%）となった．比率の差の検定の結果，3 つのコーパスの感情種別ラベルの割合は有意に異なることを示した ($\chi^2(2) = 27,659.87$, $p < 0.001$)．さらに，本コーパスが他の二つのコーパスよりも有意に感情種別ラベルの割合が多いことも示した ($p < 0.001$, Fig. 3 (b) に '***' として示す)．本コーパスは，非対面対話である点は他の二つのコーパスと同様であるが，感情誘発手法にオンラインゲームを用いている点が他の二つのコーパスと異なる．感情誘発手法にオンラインゲームを用い，さらに音声チャットを利用させることで誘発した感情を音声に表出させる本手法が，感情を知覚させる自発的な音声の効率的な収集に有効であったことが示唆される．

3　演技音声

3.1　発話テキスト

自然に感情が表出した音声と比較可能な演技音声を収録するため，自発対話音声の言語内容とほぼ一致する音声の収録を行った．

自発対話音声の転記テキストから 17 個の対話シーケンスを抜粋し，読み上げの対象とした．対話シーケンスは 10 前後の発話テキストで構成されている．全 664 発話には発話ごとに演じる感情が指定されている．Fig. 6 に発話テキストの例を示す．Fig. 6 には左から順に対話シーケンス内における発話の通し番号（Uid），話者記号（Sid），演じる感情（Emotion），発話内容（Utterance）が示されている．

Table 6　Example of acted material.

Uid	Sid	Emotion	Utterance
01	A:	ANT	属性は？
02	B:	DIS	属性はね
03	B:	ANG	水
04	A:	FEA	水、水武器を持ってけばいいの？
05	B:	ACC	そうだね
06	A:	FEA	持ってたっけな
07	B:	ANT	水武器じゃないや。風武器
08	A:	SUR	風武器？
09	B:	DIS	ああ
10	A:	SAD	風ダマあったかな？
11	A:	ANG	ちょっと待って。 今狩ってる場合じゃないんだ
12	A:	ANG	準備時間なんだよ

Table 7　Number of utternaces for each emotion.

Category	Intensity				Total
	N	W	M	S	
ACC	80	80	80	80	320
ANG	80	80	80	80	320
ANT	80	80	80	80	320
DIS	80	80	80	80	320
FEA	80	80	80	80	320
JOY	84	84	84	84	336
SAD	84	84	84	84	336
SUR	96	96	96	96	384
Total	664	664	664	664	2656

3.2　音声収録

プロの俳優の 4 名（男性 2 名，女性 2 名）に感情を演じさせて音声を収録した．同性同士の俳優にペアを組ませて，対話形式で音声の収録を行った．話者間で発話数を同一にするために，役を交代させて収録を行なった．防音室内に話者 2 名を入室させ，収録時には話者間の発話を重複させないために発話間に十分な間をあけるよう指示を出した．Table 3 に示す 8 つの感情（ACC，ANG，ANT，DIS，FEA，JOY，SAD，SUR）を，感情を含まない平静状態（N）と，弱（W）－中（M）－強（S）の 3 段階の感情強度で表現させた．話者 1 名につき 664 発話，計 2656 発話を収録している．Table 7 に感情ごとの発話数を示す．

3.3　自発音声と演技音声との韻律的特徴量の比較

自発対話音声と演技音声との違いについては，これまで頻繁に議論されてきたものの，実際に音響分析を行いその差を検証する研究はあまり例を見ず，ほとんど実証されていない．本コーパスを用いて，自発的な感情音声と演技による感情音声との韻律的特徴量の差について検証する．

検証に用いた自発対話音声は Table 1 に示した 2 名

以上の評価者の評定値が一致した 3845 発話（Partial）から NEU の発話（798 発話）と OTH の発話（200 発話）を差し引いた 2847 発話である．演技音声は収録した全発話（2656 発話）を用いた．

比較する韻律的特徴量として，感情音声で伝統的に用いられてきた発話内の平均基本周波数（F_{mean}），基本周波数の発話内標準偏差（F_{stdv}）および平均発話速度（D_{rate}）を求めた．各特徴量の計算方法は [2] とほぼ同様である．F_{mean} を求める際は，声の高さの個人差の影響を除去するため，話者ごとに発話全体の平均値を求め，その値を各発話の値から差し引くことで正規化を行った．各特徴量に対して一要因分散分析を行い，自発的な感情音声と演技による感情音声との音響的な差を感情ごとに検証した．その結果を Fig. 4 に示す．自発的な感情音声と演技による感情音声との間に有意差が認められた感情には，その感情記号にアスタリスクを付与している（$^{*}p < 0.05$, $^{**}p < 0.01$，$^{***}p < 0.001$）．

Fig. 4 の上段のパネルから，F_{mean} はいずれの感情においても自発的な感情音声と演技による感情音声とのあいだに有意な差があることが示された．感情ごとにその傾向は異なり，ACC・ANT・FEA・JOY・SAD・SUR の 6 つの感情では自発的な感情音声は演技による感情音声よりも F_{mean} が有意に低くなるが，ANG と DIS においては自発的な感情音声は演技による感情音声よりも有意に高くなっていることが分かる．中段のパネルからは，ACC・ANT・DIS・JOY・SUR の 5 つの感情において自発的な感情音声は演技による感情音声よりも F_{stdv} が有意に小さくなることが分かる．下段のパネルからは，すべての感情において自発的な感情音声は演技による感情音声よりも D_{rate} が有意に遅くなることが示された．

ほぼすべての感情で，自発的な感情音声と演技による感情音声との間には音響的な表現に差があることが示された．感情音声コーパスを利用する際にはその音響的表現の差に注意し，その目的に応じてコーパスを選定する必要がある．

4 おわりに

オンラインゲーム中のプレーヤーに音声チャットを利用させ，自然に感情が表出した発話を約 1 万発話収集し，さらに，これと言語内容がほぼ一致する演技音声を 2656 発話収録することで，自発対話音声と演技音声を比較可能な感情音声コーパスを構築した．本コーパスを用いて，本コーパスの自発対話音声と既存の自発対話音声コーパスとの定量的比較を行うとともに，これまで議論に挙がるだけで実証されてこなかった演技による感情音声と自発的な感情音声との音響的な差について検証した．

その結果，本コーパスの自発対話音声においては既存の二つのコーパスよりも有意に感情種別ラベルの割合が多く，感情を知覚させやすい感情音声が収録できたことを示した．さらに，ほぼすべての感情において，自発的な感情音声と演技による感情音声との間には音響的な表現に差があることを明らかにした．感情音声コーパスを利用する際にはその音響的表現の差に注意し，目的に応じたコーパス選定が必要である．

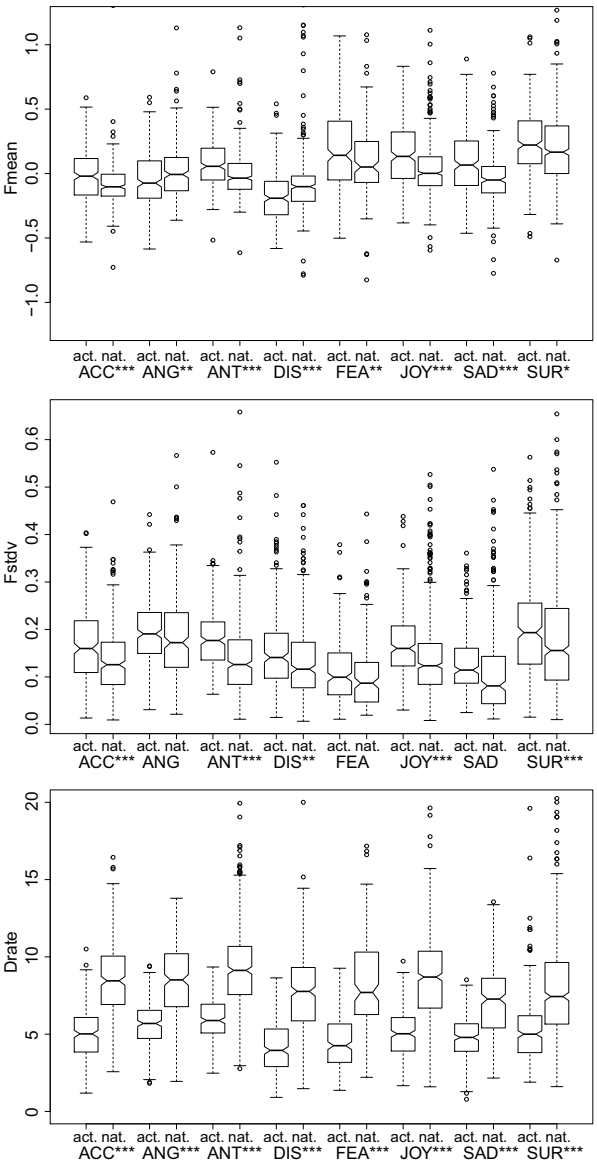

Fig. 4 Boxplots of mean F_0 (top panel), standard deviation of F_0 (middle panel) and speaking rate (bottom panel) for each emotional category.

謝辞 本発表の一部は日本学術振興会の科学研究費（25871139）の助成を受けた．

参考文献

[1] Cowie, Phil. Trans. R. Soc. B, 364 (1535), 3515–3525, 2009.
[2] Arimoto *et al.*, Acoust. Sci. Tech., 33 (6), 359–369, 2012.
[3] Arimoto-Mori, Doctoral Thesis, TUT, 2013.
[4] Plutchik, “Emotions: A psychoevolutionary synthesis.” Harper & Row, 1980.
[5] Truong *et al.*, Speech Communication, 54 (9), 1049–1063, 2012.
[6] Mori *et al.*, Speech Communication, 53 (1), 36–50, 2011.
[7] Ang *et al.*, Proc. ICSLP2002, 2037–2040, 2002.