社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (24)
2008－SLP－70 (24)
2008／2／9

感情音声のコーパス構築と音響的特徴の分析

― MMORPGにおける音声チャットを利用した対話中に表れた感情の識別 ―

有本泰子† 河津宏美† 大野澄雄†† 飯田仁†††

†東京工科大学大学院バイオ・情報メディア研究科
††東京工科大学コンピュータサイエンス学部
†††東京工科大学メディア学部

〒192-0982 東京都八王子市片倉町 1404-1
E-mail: ar@mf.teu.ac.jp, kawatsu@so.cs.teu.ac.jp, ohno@cc.teu.ac.jp, iida@media.teu.ac.jp

あらまし 音声に含まれる感情情報を自動認識することを目的に、オンラインゲーム中の自然な対話を収録し、感情音声のコーパスを構築した。感情の種類の分類としてプルチックの提案した感情立体モデルのうち、その一次的感情を取り上げ、収録した音声に付与した。また、収録した音声の高さ、長さ、強さ、声質に関わる 11 種の音響的特徴を抽出し、音響的特徴ごとに分散分析を行ない感情間の有意差を検証した。さらに、分散分析の結果に基づき、特定の感情と他の感情とを判別するための判別分析を行なった。その結果、驚きで 79.12%、悲しみで 70.11%と高い判別率が得られ、他の感情においてもほぼ 60%以上の判別率となった。

Designing emotional speech corpus and its acoustic analysis

― discrimination of one emotion appeared during the dialog over voice chat system for MMORPG ―

Yoshiko Arimoto, Hiromi Kawatsu, Sumio Ohno, and Hitoshi Iida

†Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
††School of Computer Science, Tokyo University of Technology
†††School of Media Science, Tokyo University of Technology

1404-1 Katakura, Hachioji, Tokyo 192-0982, Japan
E-mail: ar@mf.teu.ac.jp, kawatsu@so.cs.teu.ac.jp, ohno@cc.teu.ac.jp, iida@media.teu.ac.jp

Abstract For a purpose of automatic emotion recognition by acoustic information, we recorded natural dialogues made by two or three online game players to construct an emotional speech corpus. Two evaluators categorized the recorded utterances in a certain emotion, which were defined with referenced to the eight primary emotion of Plutchik's three-dimensional circumplex model. Moreover, 11 acoustic features were extracted from the categorized utterances and analysis of variance(ANOVA) was conducted to verify significant differences between emotions. Based on the result of ANOVA, we conducted discriminant analysis to discriminate one emotion from the others. As a result, high correctness, 79.12% for surprise and 70.11% for sadness, were obtained and over 60% correctness were obtained for every emotions.

1. はじめに

オンラインシミュレーションゲームでのコミュニケーションの手段は、文字チャットが一般的であるため、タイピングが遅いという理由だけでこのようなゲームを十分に楽しめないユーザも多い。かねてから、オンラインゲームにおいて、音声を利用したチャットシステムの機能が望まれている。一方で、音声チャットが実装されたとしても、音声には個人を特定するような情報が含まれているため、不特定多数のユーザを相手にするオンラインゲームでは利用しないとするユーザもいる。このような状況を踏まえ、我々は、音声に含まれる個人を特定するような情報を削除し、匿名性を保ちながらも、音声に含まれる感情情報を損なうことなく、コミュニケーション可能な音声チャットの開発を行なっている。この音声チャットは、ユーザの発した音声を認識することにより、言語的な情報をテキスト化するとともに、音声を手がかりとしてユーザの感情理解を行なう。そして、得られた言語的情報と感情情報をもとに自然な感情音声を合成するというものである。

この開発の一環として、音声に含まれる感情情報を自動認識することを目的に、感情音声のコーパスを構築した。このような感情認識に向けたコーパスを作成するためには、読み上げや演技などによる音声ではなく、対話中に自然に表出した感情を含んだ大量の音声を収録する必要がある。そこで、MMORPG（Massively Multiplayer Online Role-Playing Game）と呼ばれるオンラインゲーム中のプレーヤー同士に、音声チャットによるコミュニケーションを行なわせ、自然な対話で約 1 万発話を収録した。さらに、収録した音声の音響的特徴の分析を行ない、感情の種類ごとにその傾向について検討を行なった。

以下、2 章において音声収録の方法について述べ、3 章では、収録した音声に対して付与した感情情報について述べる。4 章では、収録した音声に対し音響分析を行ない、3 章で付与した感情の種類ごとの音響的傾向について検討した結果を述べる。5 章では、考察として音声に含まれる感情情報の自動認識の可能性について述べ、6 章で本論文をまとめる。

2. 音声チャットを利用した対話収録

1 章で述べたとおり、感情認識に向けたコーパスを作成するためには、読み上げや演技などによる音声ではなく、対話中に自然に表出した感情を含んだ音声を収録する必要がある。本研究では、オンラインゲーム中に音声チャットを使用して、プレーヤー同士で意思の疎通を図った対話音声を収録した。本章では収録の手法および収録音声の概要について述べる。

2.1. 収録環境

音声に含まれる感情情報を自動認識することを目的とした感情音声のコーパス構築に向けて、オンラインゲーム中の自然な対話を収録した。2 名あるいは 3 名のプレーヤーにオンラインゲームをプレイさせ、コミュニケーションの手段として音声チャットskype[2]を利用させた。その際の音声をskype用録音ツールであるtapur[3]を用いて録音した。録音環境として遠隔地にある部屋を利用し、ヘッドセットマイクを利用することにより、常に口とマイクとの距離は一定になるようにした。さらに、収録前にプレーヤー同士によるフリーカンバセーションを実施し、マイクレベル・マイク位置が適切となるよう調整を行った。なお、シナリオは用意せずにすべて自由発話とした。図 1に音声収録の構成図を示す。

収録に利用したオンラインゲームはラグナロクオンライン[4]、モンスターハンターフロンティア[5]、レッドストーン[6]の 3 つである。また、収録への利用回数はラグナロクオンラインが 3 回、モンスターハンターフロンティアとレッドストーンがそれぞれ 1 回である。オンラインゲームは音声提供者がアカウントをすでに取得しているサービスを利用した。ここで、ラグナロクオンラインのアカウントを取得している音声提供者が多く存在したため、ラグナロクオンラインの利用回数が多くなった。

ペアを組んだ音声提供者同士はオンラインゲームの中でもパーティを組んで行動させた。通

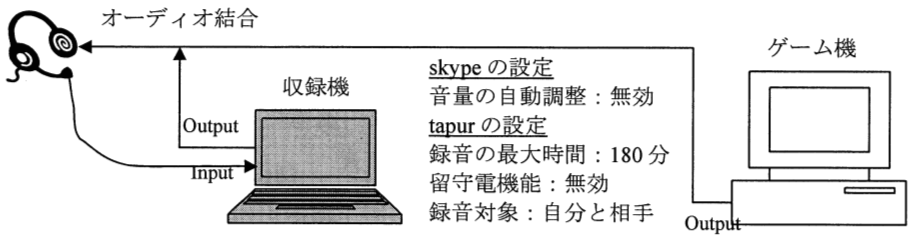

図 1　音声収録構成図

表 1　収録対話例

例 1	例 2
A: 何これ？ジョ、ジョンダ職員。わあイケメン A: 何これ？何これ？株式会社ジョンダイベントだって B: {笑} A: えーえー。た、ただ B: あれだ。あの、カプラ A: だだのカプラじゃねえか B: カプラの B: あの、なんか競争相手、的な A: へー	B: おーっと危ね B: これは A: 死ぬ{笑} A: {笑} B: あやばい死ぬ{笑}。死ぬ{笑} B: 危ね危ね A: 生命の危機を感じたよ B: オッケー

信手段はテキストチャットを禁止し、skype のみでお互いにコミュニケーションをとらせた。また、ゲーム中に不特定多数のプレーヤーから話しかけられた場合はテキストチャットで対応するよう指示を与えた。

2.2. 音声提供者

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。音声提供者の出身地は東京 6 名、長野 2 名、神奈川／静岡／山梨／青森各 1 名、未回答 1 名であった。音声提供者の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）であった。

音声提供者には同性同士でペアを組んでもらった。対話は 2 者対話と 3 者対話の 2 種類とし、2 者対話で 5 組、3 者対話で 1 組の音声を収録した。表 1に収録された対話例を示す。

2.3. 発話内容の転記

収録したオンラインゲーム中のプレーヤーの対話音声に対し、発話単位基準を 400ms 以上のポーズによってはさまれた音声の範囲とし、発話内容の書き起こしを行なった。さらに、発話内容の書き起こしに加えて、以下の 3 つの転記用タグを付与した。

表 2　音声提供者ごとの発話数

音声提供者	*発話数*	*音声提供者*	*発話数*
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

- {笑}、{咳}タグ
  笑い声、咳の場合（笑いながらの発話については記述なし）。
- (?)、(? （コメント）)タグ
  よく聞き取れない音声、自信がない場合。スペース以下の（コメント）の個所にコメントを記述。
- [comment:（コメント）]タグ
  作業者のコメントがある場合。コロン以下の（コメント）の個所にコメントを記述。

上記発話単位基準による転記によって、収録発話数は 9114 発話となった。

表 2に音声提供者ごとの発話数を示す。音声提供者の表記は、「二桁の対話番号_性別を表わす記号（MまたはF）イニシャル（姓名）」とした。

表 3　感情の分類と定義

分類	*説明*
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
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
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
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
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
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
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
期待	望ましい事態の実現、好機の到来を心から待つ感情
平静	まったく感情が表れていない
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

表 4　判定が一致した発話数と一致率

感情	*一致数*	*一致率*
喜び	287	38.2%
受容	145	18.9%
恐れ	82	23.5%
驚き	345	43.6%
悲しみ	112	27.8%
嫌悪	172	24.9%
怒り	116	31.7%
期待	236	30.0%
平静	334	27.5%
合計/平均	1829	29.6%

3. 感情情報のタグ付け

音声に付与する感情の種類の選定およびその定義付けを行い、収録した音声に表出した感情の種類を付与した。本章では、収録した各発話に感情情報を付与した手法について述べる。

3.1. 感情の種類の選定と定義

聴取した 9114 発話の音声に感情情報のタグ付けを行うため、複数の評定者に評価する発話を分担して判定を行う。そのため、評定者間に感情の定義のずれが生じうる。その感情の定義のずれを極力小さくするため、プルチックの提案した感情立体モデル[1]の 8 種の一次的感情に加え、まったく感情が表現されていない「平静」と上記 8 種の感情に分類不能な「その他」の計 10 種を判定対象の分類とし、[7]などの辞書を利用して、各感情の定義を行った。感情の種類とその定義を表 3に示す。

3.2. 各発話への感情の種類の付与

収録した 9114 発話のうち、収録音声の振幅レベルが小さく評定に使用できないと判断した 2 名（03_FMA、02_MFM）の音声提供者の 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（2.3参照）を除外した 6578 発話に対して、感情の種類の判定を行なった。

判定対象の発話を 8 つのセットに分け、各セットに含まれる発話に対して、評定者が感情の種類を判定した。判定では、聴取した音声に、表 3で示した感情のうち、どの感情が表れていると感じたかを判定し、その感情にマークさせた。この際、言語的な内容にとらわれず、音声から感じた感情の種類を判定させた。

以後の分析では、すべての発話について評定が完了している 2 名分の評定値が一致した 1829 発話を対象とする。表 4に一致した発話数とその一致率を感情ごとに示した。

4. 音響分析

感情の種類の識別を行なう音響的特徴量の抽出を行い、相関分析により各感情の識別に有効な特徴量を検討する。

4.1. 音響的特徴量の抽出

収録した全発話を対象に音響的特徴の分析を行なった。音響的特徴量は文献[8]を参考に、声の高さ、声の長さ、声の強さ、声質に関する計 11 種の特徴量を抽出した。無声音のみの発話や短すぎる発話など、8 発話においてこれらの特徴量を抽出できなかった。そのため、特徴

量が求まらなかった 8 発話を以後の分析の対象から除外した。表 5に抽出した特徴量の記号とその説明を示す。

4.2. 音響的特徴量間の相関

1821 発話に対して抽出した 11 種の音声の特徴量間で相関分析を行なった。その結果、“Fmini と Fmean”、“Fmaxi と Fmean”でそれぞれ高い相関がみられた。一方で“Fmini と Fmaxi”には、高い相関がみられなかったため、Fmean を除く 10 種の音声の特徴量を利用し、以後の分析を行なった。

5. 各感情の識別実験

音声の特徴量ごとに分散分析を行ない、識別に有効な特徴量を絞り込んだ。さらに、分散分析の結果をもとに、判別分析により識別実験を行った。

5.1. 各特徴量の識別精度の検証と識別実験

音声の特徴量ごとに分散分析を行ない、平静を含む 9 種類の感情について平均値の差の検定を行った。

受容、恐れ、嫌悪を除く 6 種類の感情については、抽出した音声の特徴量を利用することで、他の感情と判別できることが分かった。また、悲しみ－嫌悪、驚き－怒り、受容－恐れ－嫌悪－期待のそれぞれをグループにすることで、いくつかの音声の特徴量に、他の感情との有意な差を認めることができた。この際、多くの感情間に有意差を認めたのは Pmaxi であり、Pmaxi が識別に有効な特徴量であることが分かった。

分散分析において、特定の感情が他の感情と有意に区別できるとされた音声の特徴量を利用し、特定の感情と他の感情とを判別するための判別分析を行なった。その結果、驚きでは 79.12%、悲しみでは 70.11%と高い判別率が得られ、他の感情においてもほぼ 60%以上の判別率となった。分散分析および判別分析の結果を表 6に示す。表中の各感情は*が付与されている特徴量で判別が可能（有意水準 5%）であることを示している。

5.2. グループ内の感情識別実験

5.1の分散分析の結果により、複数の感情をグループ化することで他の感情と有意差を認めることができた感情について、各グループ内の

表 5　音響特徴量一覧

記号	*説明*
Fmini	男女差正規化 F_0 値の発話内最低値
Fmaxi	男女差正規化 F_0 値の発話内最高値
Fmean	男女差正規化 F_0 値の発話内平均
Fstdv	F_0 値の発話内標準偏差
Dmora	読点区切りの平均モーラ数
Drate	平均発話速度（mora/s）
Pmaxi	短時間平均パワーの発話内最大値
Pstdv	短時間平均パワーの発話内標準偏差
Pmagn	短時間平均パワーの発話内変動量
Cmean	第 1 次ケプストラム係数の発話内平均
Cstdv	第 1 次ケプストラム係数の発話内標準偏差

表 6　分散分析により有意な差を認めた特徴量とその特徴量を使用した判別分析の結果（基本 8 感情＋平静）

感情	*Fmini*	*Fmaxi*	*Fstdv*	*Dmora*	*Drate*	*Pmaxi*	*Pstdv*	*Pmagn*	*Cmean*	*Cstdv*	***全体***	***感情***	***その他***
喜び						*					51.87%	59.15%	50.52%
受容											-	-	-
恐れ											-	-	-
驚き	*	*		*							79.12%	71.76%	80.81%
悲しみ						*					70.11%	64.29%	70.49%
嫌悪											-	-	-
怒り		*								*	63.52%	65.52%	63.38%
期待				*	*						67.03%	71.19%	66.41%
平静						*					61.98%	53.01%	63.98%
悲しみ-嫌悪	*	*									66.59%	70.42%	65.89%
驚き-怒り			*			*					63.30%	66.67%	62.17%
受容-恐れ-嫌悪-期待						*					53.41%	47.80%	56.42%

表 7　分散分析により有意な差を認めた特徴量とその特徴量を使用した判別分析の結果（悲しみ-嫌悪）

感情	Fmini	Fmaxi	Fstdv	Dmora	Drate	Pmaxi	Pstdv	Pmagn	Cmean	Cstdv	全体	感情	その他
悲しみ		*				*					67.14%	76.19%	53.57%
嫌悪		*				*					67.14%	53.57%	76.19%

表 8　分散分析により有意な差を認めた特徴量とその特徴量を使用した判別分析の結果（驚き-怒り）

感情	Fmini	Fmaxi	Fstdv	Dmora	Drate	Pmaxi	Pstdv	Pmagn	Cmean	Cstdv	全体	感情	その他
驚き	*	*		*			*	*		*	78.95%	80.00%	75.86%
怒り	*	*		*			*	*		*	78.95%	75.86%	80.00%

表 9　分散分析により有意な差を認めた特徴量とその特徴量を使用した判別分析の結果（受容-恐れ-嫌悪-期待）

感情	Fmini	Fmaxi	Fstdv	Dmora	Drate	Pmaxi	Pstdv	Pmagn	Cmean	Cstdv	全体	感情	その他
受容							*				62.66%	58.33%	63.93%
恐れ											52.53%	55.00%	52.17%
嫌悪	*	*		*					*		63.92%	74.42%	60.00%
期待				*	*			*			63.29%	69.49%	59.60%

識別の可能性を分散分析および判別分析により検証した。その結果を表 7、表 8、表 9に示す。表 9中の嫌悪については分散分析の結果、有意な差を認める特徴量がなかったが、他の 3 つの感情の判別結果から判別率を求めた。

悲しみ－嫌悪の識別実験では全体で 67.14%の判別率であるものの、悲しみは 76.19%と高い確率で判別できることが分かる。また、驚き－怒りの識別実験では全体で 78.95%、驚きの判別で 80.00%、怒りの判別で 75.86%と高い判別率を示した。さらに、受容－恐れ－嫌悪－期待の識別実験では、いずれの感情も全体の判別率は 60%前後となったが、嫌悪は 74.42%と高い判別率を得た。

6. まとめ

本論文では、音声に含まれる感情情報を自動認識することを目的とした感情音声のコーパス構築に関して、その音声の収集方法と感情情報のタグ付け手法について述べた。

さらに、感情音声の高さ、長さ、強さ、声質に関わる 11 種の音響的特徴を抽出し、音響的特徴ごとに分散分析を行ない、感情間の有意差を検証した。分散分析の結果に基づき、特定の感情と他の感情とを判別するための判別分析を行なった。その結果、驚きで 79.12%、悲しみで 70.11%と高い判別率が得られ、他の感情においてもほぼ 60%以上の判別率となり、音響的特徴を利用した感情情報の自動識別の可能性が示唆された。

今後は感情ごとにその程度を付与する主観評価実験を進め、感情の種類および程度を推定する仕組みについて研究を進めていく予定である。

謝辞

本研究は独立行政法人情報処理推進機構（IPA）の 2007 年度第 II 期未踏ソフトウェア創造事業（未踏ユース）の支援を受け、東京大学大学院の竹内郁雄教授よりアドバイスを受けた。また、コーパス構築にご協力いただいた東京工科大学の戸上雅夫氏・上野智子氏、音声収録にご協力いただいた東京工科大学の酒井優子教授に深く感謝する。

参考文献

[1] Robert Plutchik, “Emotion - A Psycho-evolutionary Synthesis”, Harper, Row, 1980.
[2] Skype, http://www.skype.com/intl/ja/
[3] Tapur, http://www.tapur.com/jp/
[4] RAGNAROK online, http://www.ragnarokonline.jp/
[5] Monster Hunter Frontier, http://www.mh-frontier.jp/
[6] Red Stone, http://www.redsonline.jp/
[7] 山田忠雄, 柴田武, 酒井憲二, 倉持保男, 山田明雄, ”新明解 国語辞典 第六版”, 三省堂, 2005.
[8] 有本泰子, 大野澄雄, 飯田仁, “「怒り」の発話を対象とした話者の感情の程度推定法”, 言語処理学会誌「自然言語処理」, vol. 14, no. 3, pp. 131-145, 2007.