自発対話音声に対する叫び声アノテーション

白鳥 恵大 *, 大久保 港, 松田 匠翔, 有本 泰子（千葉工業大学情報科学部）

Scream and shout annotation for spontaneous dialog speech

Keita Shiratori, Minato Okubo, Takuto Matsuda, Yoshiko Arimoto
(Faculty of Information and Computer Science, Chiba Institute of Technology)

要旨

叫び声は突発的な感情表現を示す音声現象の一つである．先行研究では，自発対話音声に含まれる叫び声を感情表出系感動詞と区別して定義していた．しかし，先行研究の定義を基に叫び声と感情表出系感動詞のアノテーションを行っても，言語表現が似ている音声現象を音響的特性のみで区別する必要があるため，この二つの音声現象を区別することは困難であった．そこで，叫び声と感情表出系感動詞を区別するために改めて叫び声 (scream) の定義を行った．また，発話の特徴と叫び声の特徴を併せ持った音声を発話と叫びの共起（shout）として区別した．これらの定義を基に自発対話音声に含まれる音声を収録した音声資料に対して叫び声アノテーションを行った．複数人でアノテーションした際の一致率算出を行って新たな定義と先行研究の定義との比較を行う．さらに，叫び声の事例をいくつか示し，自発的な叫び声がどのような音声言語現象として発せられているかについて考察する．

1. はじめに

叫び声は突発的な感情表出を示す音声現象であり，なにかに襲われたときに危機的状況を表したり，スポーツ観戦などで場を盛り上げたりと様々な場面で発声される．叫び声に関する研究はまだ少ないものの，自発的な叫び声を収集する基礎的な研究 (Mori and Kikuchi 2020) から，叫び声検出 (Laffitte et al. 2016) や叫び声合成 (土井敦也・有本泰子 2022) といった工学的応用研究まである．例えば，叫び声を検出し異常事態の発声を検知する監視システムの構築を行った研究がある (Laffitte et al. 2016)．監視カメラの死角からでも叫び声を検出することで映像では確認できない異常事態を知ることができる．他にも，叫び声を音声合成する研究 (土井敦也・有本泰子 2022) が実現可能となれば，スポーツ観戦時に機械が叫び声を発することで場を盛り上げるなど，機械と人間とのコミュニケーションをより豊かなものにすることができる．このように，叫び声の工学的応用研究は様々な場面で我々の社会生活で役に立つ．しかし，これらの研究を達成するためには，研究の資料となる自発的に発声された叫び声の音声資源が必要となる．叫び声は日常生活の中で頻出する音声現象ではなく，その資料を効率的に収集するのは非常に困難である．

* shiratori@mac-lab.org

Mori and Kikuchi (2020) は，この課題に果敢に挑戦し，叫び声を多く含んだ自発対話コーパスを作成した．このコーパスでは，叫び声は非言語音の中でも急激な感情表出を示す社会的シグナルの一つであり，感情表出系感動詞の中でも，様式化の程度が低く，話者の制御下にある度合いが低い（思わず発せられた）ものとしていた．しかし，Mori and Kikuchi (2020) の叫び声の定義では，音声の音響的特性のみで区別する必要があり，同じような言語情報を持つ感情表出系感動詞と叫び声を区別することは困難であった．また，叫び声には「うわー」や「きゃー」といった言語的な意味を伴わない叫び声と「助けて」や「やったー」といった言語的内容を伴った発話と叫びが共起した音声が存在しているが，この二つを明確に区別して定義していない．

本研究では，音声対話中に出現する叫び声を改めて定義する．提案する定義と Mori and Kikuchi (2020) での定義とでアノテーションを実施し，一致率の比較を行った．また，提案した定義を基に，自発対話音声中に出現する叫び声はどんな音声かを調査するため，叫び声の分節音の分析，発話と叫びが共起した際の言語分析を行った．これらの分析によって，自発的な叫び声がどのような音声言語現象として発せられているかについて考察する．

2. 音声資料

本研究では，アクションゲーム音声コミュニケーションコーパス（AGSC）(Mori and Kikuchi 2020) を使用した．AGSC は 2 人 1 組でゲームをプレイしている時の音声を収録している．会話中の自発的な叫び声を得るため，アクション性の高い 2 種のゲーム（First-person shooter(FPS) ゲームである OverWatch（Blizzard Entertainment, Inc.），サッカーゲームである FIFA16(Electronic Arts Inc.)）を使用し収録された．コーパスには，24 名（男性 12 名，女性 12 名）の音声が量子化ビット数 16bit，サンプリング周波数 48kHz で収録されている．一人当たり平均 60.7 分，全体で 728.4 分の音声コーパスである．AGSC は praat によって，収録した音声に対して発話番号や発話内容，先行研究の定義による叫び声・感情表出系感動詞がアノテーションされている．言語音・非言語音が 400ms 以上の途切れがなく連続している区間を発話のセグメントとし，発話のセグメンテーションと発話の書き起こしは大学生 6 名によって行われ，最終的に著者によって修正されている．

この音声資料の 24 名分全ての音声を使用して，叫び声のアノテーションを行った．

3. 叫び声アノテーション

3.1 叫び声の定義

叫び声の定義は，Mori and Kikuchi (2020) の定義（以下，先行研究の定義）を参考に改めて定義した．感動詞は，フィラー，感情表出系感動詞，応答表現，挨拶表現，呼び掛け・掛け声の 5 つに分類される．そのうち，感情表出系感動詞は驚いた時や落胆したときなどに発する表現である．Mori and Kikuchi (2020) は，叫び声を次のように定義していた．

- 叫び声は非言語音の中でも急激な感情表出を示す社会的シグナルの一つであり，感情表出系感動詞の中でも，言語的な様式化の程度が低く，話者の制御下にある度合いが低い

（思わず発せられた）もの

叫び声は急激な感情表出を示す社会的シグナルであり危機的状況下などの突発的なイベントに対して発声される．そのため，文脈や状況を考慮してアノテーションを行う必要があるが，先行研究の定義では考慮されていない．そこで，

- 叫び声は予想外の出来事によって話者が無意識に発したもので，韻律あるいは声質が特異なもの

とし，叫びのうち聞き手が単語として意味を理解できないもの (scream) と定義した．

また，叫びには様式化の程度が低いものに加えて，「助けて」や「やったー」といった発話と叫びの共起によって言語内容を伴った叫び声が存在するが，先行研究の定義ではこの現象を説明できていない．笑いの研究では，発話と笑いの共起である音声 (speech laugh) を発話や発話を伴わない真の笑い声と区別して定義されている (Nwokah et al. 1999)．そこで，叫びのうち聞き手が単語として意味を理解できるもの (shout) を定義に加え，単語としての理解の可否で scream と区別した．

3.2 アノテーション

アノテーションは praat を用いて話者の音声のみを使用して行った．AGSC の TextGrid には 3 つの層があり，そのうち utterance 層は発話番号，word 層は発話内容，affectburst 層は叫び声（scream，shout），感情表出系感動詞が記述されている．word 層は，発話内容以外にも，叫び声には {shout} というラベルが付与されている．

この affectburst 層に追加・変更し，提案した定義による叫び声および感情表出系感動詞を記述した．提案した叫び声の定義をもとに，sceam と shout および感情表出系感動詞を区別するためのラベルを定義し，アノテーションに使用した．提案した定義でのアノテーションで使用したラベルとその定義を以下に示す．

- {s}：予想外の出来事によって話者が無意識に発したもので，韻律あるいは声質が特異なもののうち，聞き手が単語として理解できないもの (scream)
- {shout}：予想外の出来事によって話者が無意識に発したもので，韻律あるいは声質が特異なもののうち，聞き手が単語として理解できるもの (shout)
- {a}：感動詞のうち感情表出系感動詞 (scream と shout は含まない)
- {o xxx}：{s} や {shout}，{a} のどれかにあたるが，音量が小さいなどの定義に合わないもの (xxx に定義に合わない理由を記述)

提案した叫び声の定義の妥当性を検証するため，まずは G011_L(男性話者，58 分) のデータのみを使用してアノテーションを実施した．先行研究の定義でのアノテーションで使用したラベルとその定義を以下に示す

- {s}：感情表出系感動詞の中でも，言語的な様式化の程度が低く，話者の制御下にある度合いが低いもの
- {a}：感動詞のうち感情表出系感動詞

大学生 2 名に，先行研究での {a} と {s} ラベル，提案した定義での {a} と {s} ラベル，それぞれに対してアノテーションを行わせ，アノテータ間の一致率を求めて比較した．提案した定

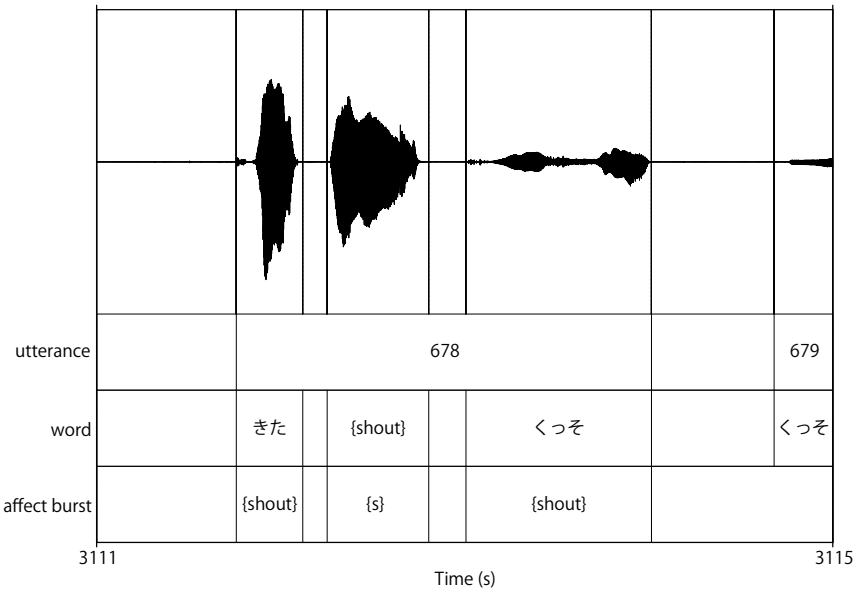

図 1　praat アノテーション画面の例

義により一致率が向上したことを確認したのち，叫び声の研究に従事している大学生 2 名および大学院生 1 名の合計 3 名によって AGSC のすべてのデータに対してアノテーションを実施した．一人当たり 8 名分のデータに対してアノテーションさせている．判断が難しい音声に対しては，アノテーションを行った 3 名でディスカッションを行い，最も適切だと思われるラベルへ振り分けた．

AGSC の praat のアノテーション画面の例を図 1 に示す．1 段目は音声波形，2 段目は TextGrid による書き起こしの内容を示している．

3.3　結果

提案した叫び声の定義の妥当性を検証するため，評価の指標として，正解率と F 値を算出した．正解率と F 値を求める際は，ラベルが付与された各区間の開始時刻と終了時刻の差が 0.2s 以内であれば一致しているとした．先行研究の定義の {s} と {a} ラベル，提案した定義の {s} と {a} ラベルのアノテーションの一致率を表 1 に示す．先行研究の定義の {a} ラベルの正解率は 30%，{s} ラベルの正解率は 29%，提案した定義の {a} ラベルの正解率は 55%，{s} ラベルの正解率は 58% であった．先行研究の定義の {a} ラベルの F 値は 46%，{s} ラベルの F 値は 45%，提案した定義の {a} ラベルの F 値は 71%，{s} ラベルの F 値は 74% であった．

AGSC の全データに対してアノテーションをした結果として，各ラベルが付与された区間の数を話者ごとに表 2 に示す．AGSC には，男性話者による scream が 534 個，shout が 278 個，女性話者による scream が 856 個，shout が 364 個，合計して scream が 1390 個，shout が 642 個含まれていた．

表 1　先行研究の定義と提案した定義のアノテーション一致率

	先行研究		提案	
	{a}	{s}	{a}	{s}
正解率	0.30	0.29	0.55	0.58
F 値	0.46	0.45	0.71	0.74

3.4　考察

先行研究の定義より，提案した定義による {a} ラベルの正解率は 25%，F 値は 25% 高く，{s} ラベルの正解率は 29%，F 値は 29% 高かった．提案した定義は先行研究の定義よりも安定したアノテーションが可能であることがわかった．先行研究の定義と提案した定義の異なる点は文脈や状況を考慮したことである．このことより，自発対話音声における叫び声は，突発的なイベントに対するリアクションであるという視点を追加することで他の音声との区別がより可能になると考えられる．また，先行研究の {s} で 2 名のアノテータが一致せず，かつ提案手法の {s} で一致した音声は，話者が無意識に発声したかどうか曖昧な音声が多かった．先行研究では対象となる叫び声の前後の文脈は考慮せず，その叫び声の音声情報のみでアノテーションを行っているため，曖昧な音声は判断がつきにくいが，提案した定義では音声の直前の会話内容から予想外の出来事の有無を確認し，ある程度無意識に発声したかどうか判断できるため，アノテータ間でより一致した判断が得られた．

男性話者よりも女性話者の scream の数は 322 個多く，shout の数は 86 個多かった．今回の参加者の一月当たりの平均ゲーム時間は，男性話者が 92.5（SD=60.47），女性話者が 29.17（SD=42.15）であり，ゲーム経験の差が叫び声の表出数に差を生み出した可能性がある．つまり，今回の参加者は，ゲームの内容になじみ深く，ある程度どんなことが起こるか想像することができるため，あまり叫び声をあげることがない参加者が男性に多く，また，ゲームの内容をあまり知らず，ゲームの展開を予想できなかったために，予想外の出来事が発生したときに，叫び声を上げた参加者が女性に多かった可能性がある．また，同じ男性話者でも G009_R は scream が 5 個，shout が 0 個に対して G011_L の話者は scream が 133 個，shout が 55 個と大きな差が見られた．そのため，叫び声の発声しやすさは性差ではなく個人差があると考えられる．

4.　叫び声の言語現象の分析

4.1　分析の目的

自発的な叫び声がどのような音声言語現象として発せられているかについて調査した．叫び声がどのような言語音や単語で発せられるかあるいは発せられやすいかを，コーパスを用いて研究された例はまだ見ず，それらを明らかにして，工学的な応用研究に役立てることが求められる．例えば，叫び声を合成する際に，どのような音韻で叫び声を構成すれば最も叫び声らしくなるかの判断の一助となる．まずは，scream がどのような音素で発声されているかを明ら

表 2　ラベルが付与された音声区間の数

話者	性別	{a}[個]	{s}[個]	{shout}[個]
G001_L	男	137	17	7
G001_R	男	106	34	15
G003_L	女	239	89	24
G003_R	女	254	70	43
G004_L	女	275	61	25
G004_R	女	410	66	72
G005_L	女	145	10	3
G005_R	女	80	35	9
G006_L	女	208	113	22
G006_R	女	123	131	37
G007_L	女	167	52	34
G007_R	女	55	55	22
G008_L	女	80	11	9
G008_R	女	364	135	64
G009_L	男	51	16	2
G009_R	男	35	5	0
G010_L	男	145	35	23
G010_R	男	209	50	19
G011_L	男	283	133	55
G011_R	男	58	62	27
G012_L	男	114	94	55
G012_R	男	79	35	26
G013_L	男	156	21	20
G013_R	男	156	21	20
男性話者	男	1627	534	278
女性話者	女	2400	856	364
合計		4027	1390	642

かにする．また，発話と叫びが共起する shout では，「助けて」など意味のある単語で発せられるのか，「が」や「を」などの助詞のような意味のない単語で発せられるのかを明らかにする．さらに，shout では，話をしている途中で叫んで shout となるのか，叫んだ (scream) 後に shout となるのか，それとも発話と叫びを同時に発声するのかを明らかにする．

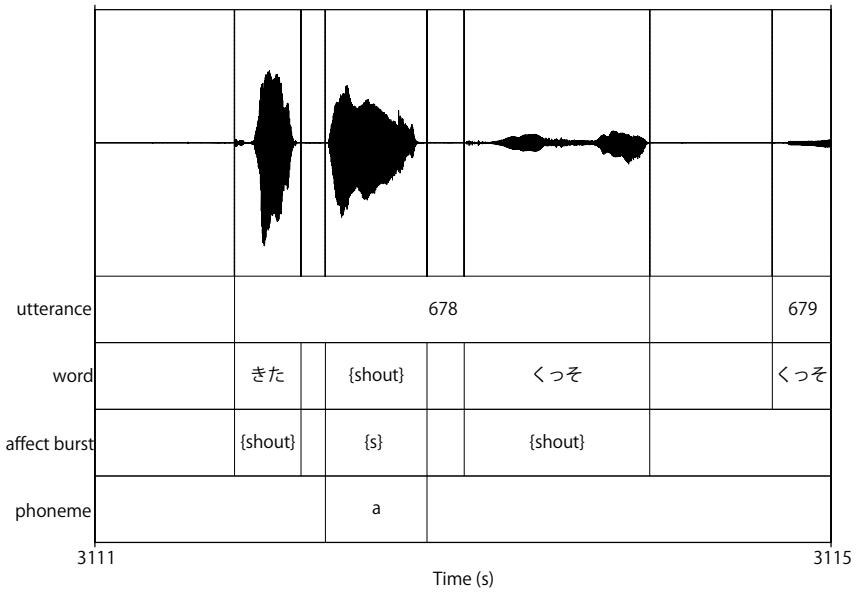

図 2　分節音アノテーションの例

4.2　分析方法

分析に使用したデータは, 男女それぞれで叫び声の数が多かった G006_R, G008_R, G011_L, G012_L の 4 名のデータを使用した．使用したデータには，{s} の音声が 493 個，{shout} の音声が 211 個含まれていた．

scream がどのような音素で発声されているかを明らかにするため，アノテーションした scream に対して分節音アノテーションを行い，各分節音の出現率を求めた．分節音アノテーションの例を図 2 に示す．

phoneme 層を新たに追加し，affectburst 層に {s} のラベルが付与されている区間に分節音ラベルを記述した．分節音を判断する際に，いずれかの母音に該当するがどの母音にするか判断できない音声については X ラベルを付与した．各分節音の出現率の統計的な差を確認するために，分節音ラベルを母音と子音に分類し，帰無仮説を「すべての母音の発生率に差はない」,「すべての子音の発生率に差はない」としてカイ二乗検定を行った．

shout を発するときに，意味のある語で叫んでいるか意味のない語で叫んでいるかを調べるため，shout の冒頭が内容語であった数と機能語であった数を集計した．アノテーションを行った {shout} ラベルの部分の発話内容について MeCab による形態素解析を行い，出力された品詞を内容語と機能語に分類した．帰無仮説を「内容語と機能語の数に差はない」として二項検定を行った．

さらに，shout の直前の音声を集計し，「これさー，やばいー」のように発話から shout が発声されるか，「わー，やばいー」のように scream から shout が発声されるか，「やばいー」のように発話と叫びが同時に開始されて（shout の直前には無音区間がある．あるいは shout の直前に shout がある）shout が発声されるかを調査した．アノテーションを行った {shout} ラベルの部分の直前 0.2s の区間を調査し，その区間の音声が発話か scream かそれ以外かで,

発話から開始した shout か，scream から開始した shout か，発話と scream が同時に発声した shout かを分類した．発話から shout が発声されるか，scream から shout が発声されるか，発話と scream が同時に開始し shout が発声されるかの帰無仮説を「shout の直前の音声に差はない」としてカイ二乗検定を行い，調査した．

4.3 結果

scream の分節音の集計結果を，母音と子音に分けてそれぞれ図 3 と図 4 に示す．図 4 の子音のグラフは，2 個以上出現した分節音のみを掲載した．母音では/a/が 392 個と最も多く，/i/は 43 個，/u/は 69 個，/e/は 76 個，/o/は 79 個，/X/は 18 個であった．子音では，/h/で 46 個，/w/で 41 個，/y/で 16 個，/r/で 4 個，/d/で 2 個，/n/で 2 個であった．母音の出現率のカイ二乗検定の結果，$\chi^2(5) = 852.82, p < 0.001$ となり，帰無仮説が棄却され「すべての母音の出現率に差がある」ことが分かった．子音の出現率のカイ二乗検定の結果，$\chi^2(6) = 117.22, p < 0.001$ となり，帰無仮説が棄却され「すべての子音の出現率に差がある」ことが分かった．

shout の冒頭における内容語・機能語の集計結果を図 5 に示す．内容語であった数は, 160 個, 機能語であった数は 51 個であった．内容語と機能語の数に対する二項検定の結果，$p < 0.001$ となり，帰無仮説が棄却され，内容語が機能語に対して有意に多いことが分かった．

shout の直前の音声の集計の結果を図 6 に示す．発話から開始されたものは 18 個，scream から開始されたものは 69 個，発話と叫びが同時に開始されたものは 124 個であった．shout の直前の音声に対するカイ二乗検定の結果，$\chi^2(2) = 79.92, p < 0.001$ となり帰無仮説が棄却

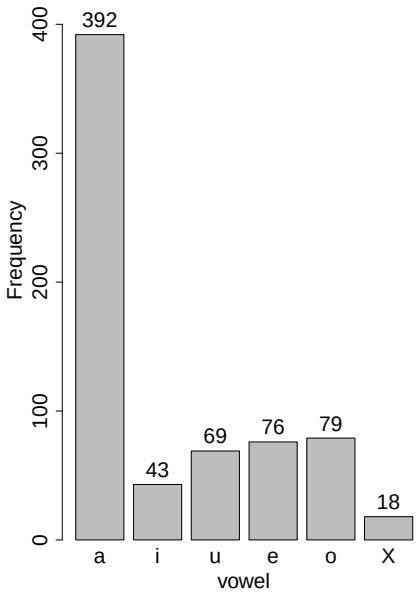

図 3　母音の集計結果

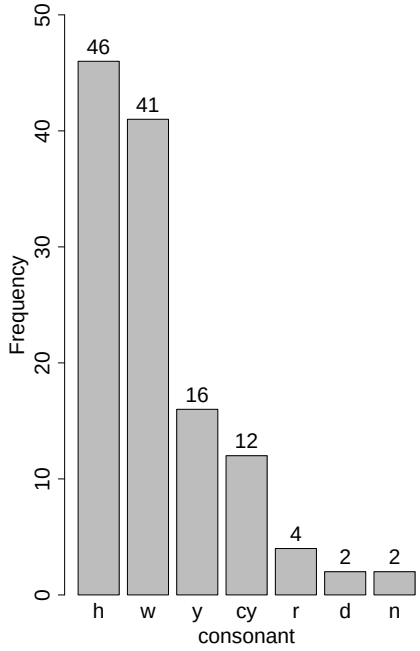

図 4　子音の集計結果

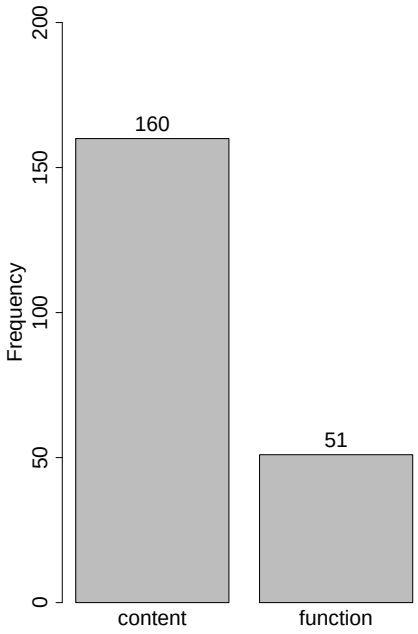

図 5　shout が内容語か機能語かの集計結果

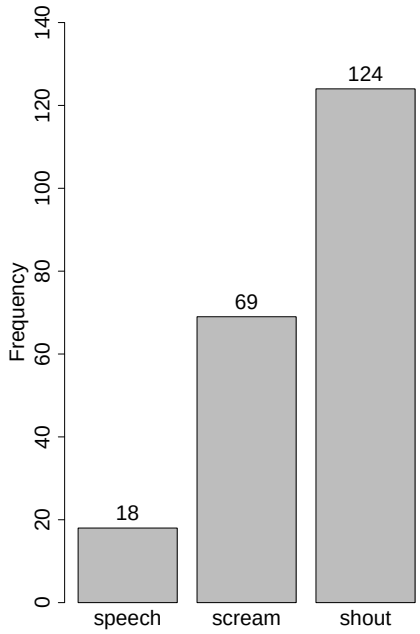

図 6　shout 直前の音声の集計結果

され「shout の直前の音声に差がある」ことが分かった．

4.4　考察

scream の音声の分節音の分析によって，叫び声は母音では/a/，子音では/h/や/w/で発声されやすいことが分かった．なかでも，分析対象となる 493 個の scream のうち母音/a/は 392 個と圧倒的に多く，ほとんどの scream が/a/で開始されていることが分かる．我々が思い浮かべる scream の典型的な音韻列は「きゃー」であるが，図 4 に示した子音の集計結果からも明らかなように/ky/で発声される scream はほとんどない．自発対話音声における scream には「きゃー」といったような叫び声はほとんど存在せず，単純に「あー」といった scream が多く発声されることが示唆される．

shout の内容語・機能語の分析では，言語内容を伴った叫び声は内容語で発声されやすいことが分かった．shout の音声は，「行け」や「取れ」といったゲーム内の自分が操作しているアバターに対して指示を出す際に思わず発することが多かった．このことより，応援や鼓舞をする強い気持ちが叫び声となって表出されると考えられる．

shout の直前の音声の分析では，発話から変化するのでもなく叫びから変化するのでもなく，発話と叫びの現象が同時に生起しやすいことが分かった．shout の音声は，「行け，行け，行け」のように短い shout の区間を何度も繰り返すことが多かった．このことより，同じ言葉を繰り返し叫ぶことで伝えたい内容を強調していると考えられる．また，予想外の出来事によって scream が発声され，その後 shout が発声されることが多かった．scream によって予想外の出来事に対するリアクションをし，shout によってその状況を踏まえて伝えたい内容を発声

している可能性がある．

5. おわりに

叫び声検出や叫び声合成などの工学的な応用研究に対話中に自然に表出した叫び声を利用することを目的とし，自発対話音声における叫び声の定義を提案した．また，叫び声がどんな言語表現で表出するか調査するため，基本的な言語分析を行った．その結果，提案した叫び声の定義は Mori and Kikuchi (2020) の定義よりもアノテータ間の一致率が高く，比較的安定して叫び声ラベルを付与できることが分かった．言語内容を伴っていない叫び声である scream は母音 a で発声されることが多いことが分かった．言語内容を伴った叫び声である shout は，内容語で，発話と叫びの現象が同時に引き起こされやすいことが分かった．叫び声がいつ，どのように発声されるかを明らかにすることで，機械が対話の中で自然に叫び声を発声し場を盛り上げるなど，機械と人間とのコミュニケーションを円滑にできる．今回はゲーム場面での自発音声を扱ったが，今後は他の自発対話音声について叫び声の分析をし，今回の結果との差異を比較していきたい．

謝　辞

AGSC を快くご提供くださった宇都宮大学の森大毅准教授に感謝する．本研究の一部は公益財団法人中山隼雄科学技術文化財団設立 30 周年研究助成，および JSPS 科研費 JP22K18477, JP22K12107 の助成を受けた．

文　献

Hiroki Mori, and Yuki Kikuchi (2020). “Gaming Corpus for Studying Social Screams.” *Proc. Interspeech 2020*, pp. 3132–3135.

Pierre Laffitte, David Sodoyer, Charles Tatkeu, and Laurent Girin (2016). “Deep neural networks for automatic detection of screams and shouted speech in subway trains.” *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6460–6464.

土井敦也・有本泰子 (2022).「WaveNet による叫び声合成の実現に向けたコンテキストラベルの検討」　日本音響学会 2022 年春季研究発表会講演論文集, pp. 951–952.

Eva Nwokah, Hui-Chin Hsu, P Davies, and Alan Fogel (1999). “The integration of laughter and speech in vocal communication: A dynamic systems perspective.” *Journal of speech, language, and hearing research : JSLHR*, 42, pp. 880–94.