INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Gaming Corpus for Studying Social Screams

Hiroki Mori, Yuki Kikuchi

Graduate School of Engineering, Utsunomiya University
hiroki@speech-lab.org

Abstract

Screams in everyday conversation, rather than in emergencies, are considered as nonverbal behavior that makes our speech communication rich and expressive. This paper focuses on screams in conversation. Identification of screams in existing spoken dialog corpora revealed that these corpora contained only a small number of screams, so not adequate for the investigation of the screams. In order to obtain more screams that naturally occur in conversation, we recorded dialogs while playing highly action-oriented games. Following to our criteria to identify screams, 1437 screams were detected from the whole recordings. The screams in our corpus were 12 times more frequent than the existing gaming corpus. As for the number of screams per minute, a strong positive correlation was observed between two speakers of the same pair, suggesting that the interlocutors produced screams not purely spontaneously, but tried to get the screaming behavior closer to the other person. Results of the acoustic analysis showed that the typical scream is produced in 60–140 mel higher and 8 dB louder voice than typical normal speech.

Index Terms: screams, speech communication, affect bursts

1. Introduction

Emotions such as excitement and surprise may exhibit as screams. Screams can be produced not only in a crisis situation. For example, when we are watching a sports game with close friends, we may scream. By screaming, we can express our excitement and share the experience with others. Screaming in everyday conversation, therefore, is considered as one of the nonverbal behaviors, such as laughter, that makes our speech communication rich and expressive.

Most of the studies on screaming to date have been aimed at security. Some of them used sentences that were read in the style of acted screaming, rather than natural screams, as a corpus [1, 2, 3, 4]. Others used sound effects, movie soundtracks, or movie clips on the internet [5, 6, 7]. Apparently, screaming by acting is not the ideal exemplar of a natural scream. Role-playing [8] may be an option to obtain more natural screams, but it is still problematic in terms of ecological validity. Thus, there are few studies on screams that naturally occur in our everyday communication situations.

The aim of this study is to clarify the acoustic characteristics of screaming in speech communication, and the role it plays in social interaction.

2. Screams in speech communication

2.1. Definition

In this study, we define *screams* as vocal *affect bursts* [9, 10, 11] that occur spontaneously (not deliberately), and not quite conventionalized like dictionary words, therefore relatively hard to transcribe. The concept of screams in this paper covers from "raw" affect bursts at the far left in the spectrum shown in Fig. 1 to a part of expressive interjections in midway. According to this definition, some screams can be transcribed, but not others.

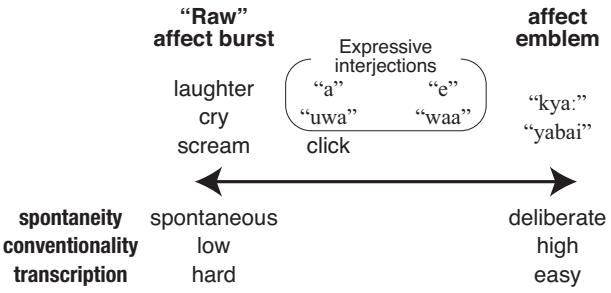

Figure 1: *Spectrum of affect bursts. [11]*

Table 1: *Number of screams in existing corpora.*

	UUDB	OGVC	Chiba3Party
# utterance	4840	8662	7599
# screams	4	53	4
# screams / min.	0.0333	0.0803	0.0111
# screams / # utt.	0.0008	0.0061	0.0005

In Japanese conversation, expressive interjections, that act as a signal to indicate the speaker's cognitive/affective state changes, include "e" in doubt, and "a" in noticing something. Each of these is easy to transcribe and conventionalized as a dictionary word, therefore not regarded as screams. However, some of the sounds expressed in response to unexpected and sudden events are so peculiar both in prosody and voice quality that they can no longer be interpreted as linguistic sounds, even though they sound close to "a," which we may regard them as screams.

2.2. Identification of screams in existing corpora

Based on the above definition, screams were identified and annotated for Japanese spoken dialog corpora. The target corpora include the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) [12], where participants were engaged in the "four-frame cartoon sorting" task, the Online Gaming Voice chat Corpus (OGVC) [13], where the voice chat during online games was recorded, and the Chiba Three-party Conversation Corpus (Chiba3Party) [14], which contains casual conversations among three participants of the same gender.

The second author was engaged in the annotation work. Table 1 shows the number of utterances of each corpus and the number of screams it contains. The result suggests that screams are few in UUDB, OGVC, and Chiba3Party. Among them,

http://dx.doi.org/10.21437/Interspeech.2020-2553

screams are most frequent in OGVC. This can be interpreted that OGVC is a recording of voice chat in gaming, and gaming stimulated unexpected events more frequently. However, even the OGVC, containing only 53 screams, is insufficient for a quantitative study of screams.

3. Corpus construction for studying screams in speech communication

3.1. Recording

It turned out that the existing corpora do not contain a lot of screams. In order to get more screams that naturally occur in conversation, we newly recorded dyadic dialogs while playing highly action-oriented games. As a result of preliminary trials, two game titles were selected for the recording: *Overwatch* is a team-based multiplayer first-person shooter (FPS); *FIFA 16* is a real-time soccer simulation game.

Participants were twelve pairs of two Japanese-speaking college students of the same gender with a high degree of familiarity. Six of the pairs were female and the rest were male. All the participants except two females in a pair had at least ten years of game experience. Each pair participated in a game title they preferred, except one pair who switched the game in the middle due to VR sickness. Each participant was seated in a separate soundproof room, wore a headset (Sennheiser HMD25-1) to communicate with the partner, and played a game. The chat voice was directly digitized by a PC in the format of 16 bit 48 kHz PCM through a Roland OCTA-CAPTURE USB-DAC. In total, twelve sessions of 728.4 min. was recorded. The average recording time for each pair was 60.7 min. (SD $= 8.3$).

3.2. Screams and utterances

Following the segmentation and transcription of utterances, the second author identified the participants' screams in the same way as described in Sect. 2.2. As the total result, the number of utterances, the number of screams, the number of screams per minute, and the number of screams per utterance is 14111, 1437, 0.986, and 0.102, respectively. Comparing these figures to Table 1, it turns out that screams in our recordings are 12 times more frequent, and 27 times more numerous than OGVC.

Figure 2 shows the frequency distribution of screams and utterances. The frequency of utterances as well as screams varies greatly from speaker to speaker. Nevertheless, their correlation is weak ($R = 0.298$). This means that whether or not a speaker screams a lot cannot be explained solely by whether or not the speaker talks a lot.

Figure 2 also suggests the frequency of screams of speakers belonging to the same pair tends to be similar. To make this point clearer, Fig. 3 shows a direct comparison of two speakers who belong to the same pair. The correlation coefficient for all pairs is 0.657. The frequency of screams is surprisingly similar except for three pairs; when these pairs are excluded, the correlation coefficient becomes 0.991. This is a piece of evidence for our claim that *screaming is social*. The interlocutors produced screams not purely spontaneously. Rather, they seem to have unconsciously adapted the screaming behavior to each other to become more similar. This might be regarded as a non-linguistic manifestation of the phenomenon known as entrainment or alignment, which affects interlocutors' linguistic style [15], prosody [16], and nonverbal behaviors such as laughter [17], postures and facial expressions [18].

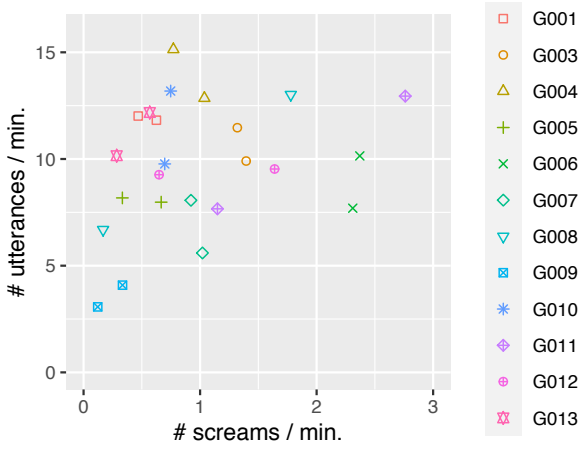

Figure 2: *Frequency distribution of screams and utterances. Each point corresponds to a speaker. Points of the same shape belong to the same pair.*

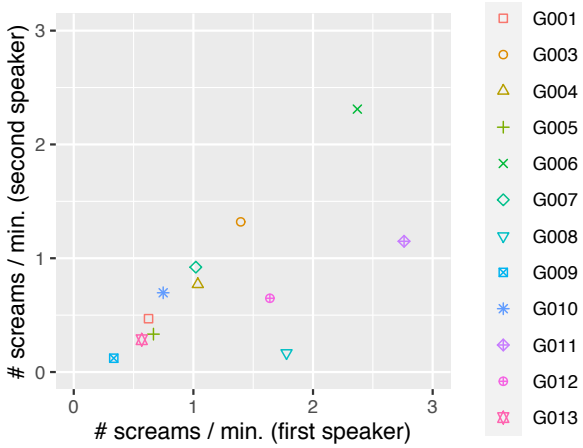

Figure 3: *Comparison of the frequency of screams between speakers of the same pair. Horizontal axis indicates the number of screams of the speaker who screamed more.*

3.3. Agreement

Annotation of screams in conversation is challenging. Despite the definition of screams given in Sect. 2.1, it is not straightforward to test whether a given vocal expression is a scream or not. To check the reproducibility of the scream annotation described in Sect. 3.2, the two authors separately annotated the screams from an excerpt of recordings consisting of four 10-minute soundtracks of two females and two males, then the agreement was calculated. For each utterance in the 40-minute set, their annotations were regarded as agreeing either (1) if no segment was annotated as a scream, or (2) if more than 50% of the segments that were annotated as a scream overlapped.

Calculated Cohen's κ was 0.681. This may be interpreted that the agreement of identified screams of the two annotators is substantial [19]. When the annotation of screams by the second author is assumed to be the ground truth, the precision and recall of the annotation of screams by the first author are 0.643 and 0.800, respectively. While the values of precision/recall are not quite satisfactory, the annotated screams in Sect. 2.1 can be said

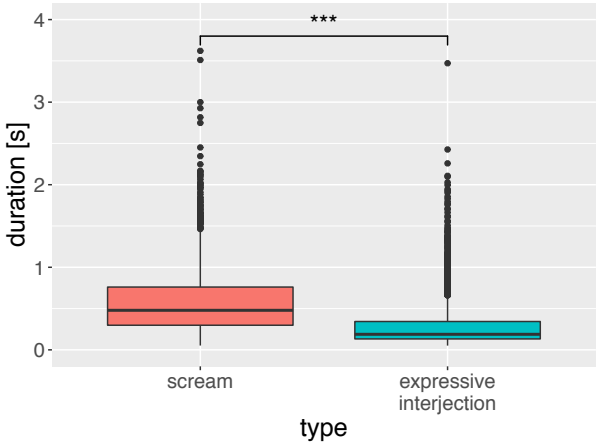

Figure 4: *Duration distribution.*

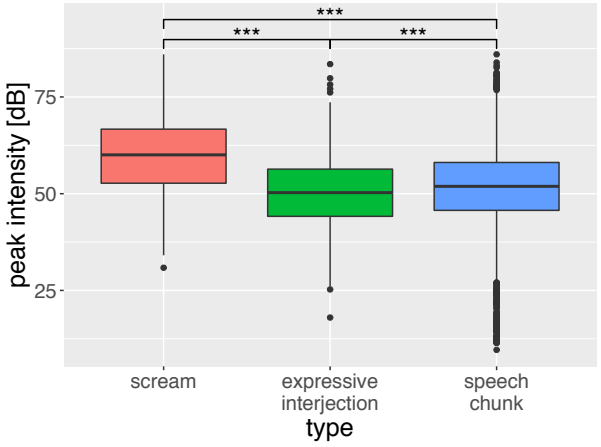

Figure 6: *Peak intensity distribution.*

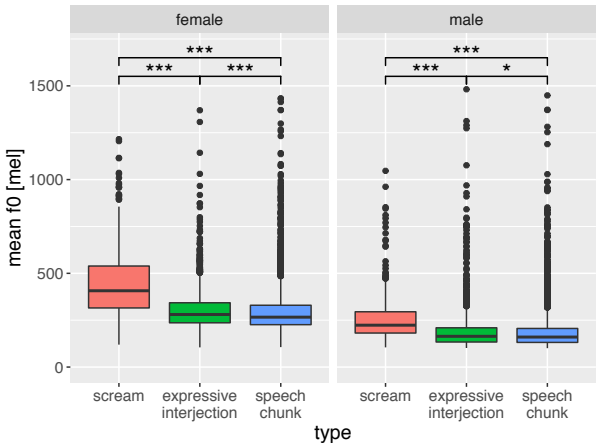

Figure 5: *Mean f_o distribution.*

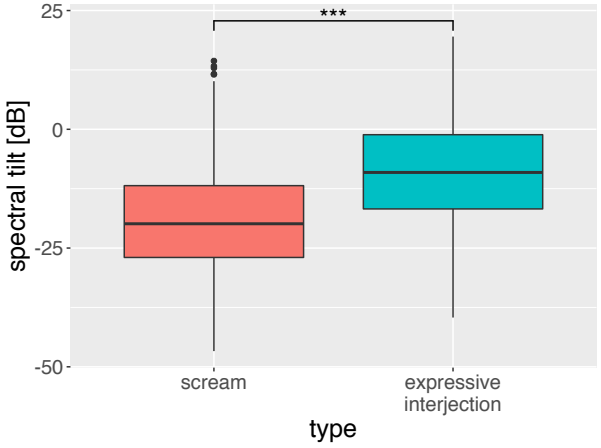

Figure 7: *Spectral tilt distribution.*

to have reasonable reliability.

4. Acoustic analysis

As a preliminary study of the acoustics of screams, an analysis of key speech features for our corpus was performed. Throughout this section, an *expressive interjection* is that defined in [11], but screams are excluded from expressive interjections because of having a dedicated category *scream*. In addition, a *speech chunk* is defined as any spoken segment that is neither a scream nor an expressive interjection.

4.1. Duration

Figure 4 shows the distribution of duration for screams and expressive interjections. The median was 0.479 sec. and 0.188 sec., respectively. As a general tendency, the scream is longer than the expressive interjection.

4.2. f_o and intensity

As prosodic features, mean f_o and peak intensity were calculated. F0 values were obtained with the *To Pitch (ac)* command of Praat, and converted to the mel scale. To avoid errors due to unstable phonations, the mean calculation was performed only for ones with five voiced frames or more. Likewise, intensity values were obtained with the *To intensity* command of Praat. Note that the recording levels were not calibrated, so it is not possible to directly compare the intensity of different speakers.

Figure 5 shows the distribution of mean f_o for screams, expressive interjections, and speech chunks. The median was 407 mel, 280 mel, and 266 mel for female speakers, and 223 mel, 164 mel, and 160 mel for male speakers, respectively. Although the pitch of expressive interjections was higher than speech, that of screams was even much higher; the median of screams was about 140 mel (for female speakers) or 60 mel (for male speakers) higher than that of normal speech. It is also worth pointing out that there was great diversity in the mean f_o of the female speakers' screams; the interquartile range was 224 mel. Female speakers seemed to actively produce screams with a wide variety of expressivity.

Figure 6 shows the distribution of peak intensity for screams, expressive interjections, and speech chunks. The median was 60.0 dB, 50.3 dB, and 51.9 dB, respectively. This shows a clear conclusion that screams are, as expected, generally loud.

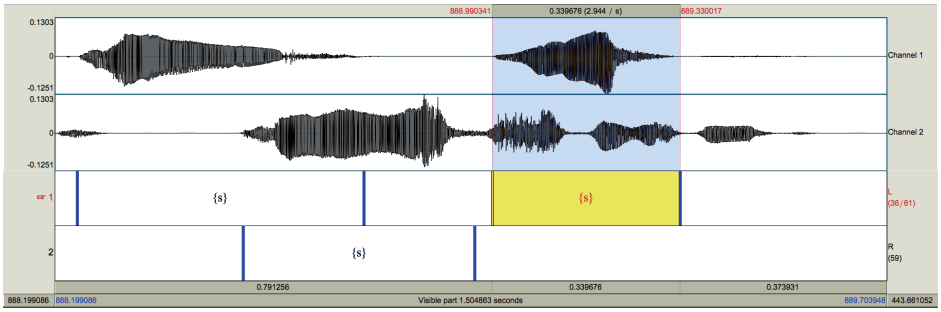

Figure 8: *Co-occuring screams.*

4.3. Spectral tilt

Vowel spectral tilt was calculated as a voice quality feature related to the glottal source. The first LPC cepstrum coefficient (c_1) was obtained with the *To LPC (autocorrelation)* command of Praat for the midpoints of the segments, then log power at 0 Hz relative to 3000 Hz was calculated using the spectral envelope estimated by c_1. Figure 7 shows the distribution of spectral tilt for screams and expressive interjections. Note that a larger positive value of the spectral tilt indicates that the higher frequency component is relatively low, as in the case of lax, breathy voice. The result indicates the spectral tilt of screams tended to move towards negative values, which implies that the screams were produced with a more tense voice quality.

5. Conclusions and future works

To elucidate the characteristics of screams in conversation, an action game-based spoken dialogue corpus was constructed. A preliminary analysis of the acoustic feature was also conducted. The results showed that the screams are produced in a higher and louder voice than the normal speech.

One of our main concerns is how the interlocutor's screams affect the other's behavior. It has been shown at Sect. 3.2 that the frequency of screams of interlocutors tends to be similar. This overall trend is an aspect of the entrainment. Another aspect might be a temporally local one, often referred to as synchrony [20]. We observed many interesting examples of how a participant's scream elicited another participant's scream. Figure 8 shows an example. It gives the impression that they enjoyed making screams. In future works, this co-occurrence of screams will be investigated quantitatively to clarify how the synchrony reflects the interlocutors' emotions and social relationships.

6. References

[1] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," in *Proc. Interspeech 2007*, 2007.

[2] H. Nanjo, H. Mikami, H. Kawano, and T. Nishiura, "A fundamental study of shouted speech for acoustic-based security system," in *Proc. Interspeech 2009*, 2009, pp. 1027–1030.

[3] M. K. Nandwana and J. H. Hansen, "Analysis and identification of human scream: Implications for speaker recognition," in *Proc. Interspeech 2014*, 2014, pp. 2253–2257.

[4] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *Proc. ICASSP 2015*, 2015, pp. 161–165.

[5] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proc. EUSIPCO 2007*, 2007, pp. 1216–1220.

[6] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. ICASSP 2009*, 2009, pp. 165–168.

[7] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *Proc. ICIEA 2010*, 2010, pp. 2115–2120.

[8] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proc. ICASSP 2016*, 2016, pp. 6460–6464.

[9] K. R. Scherer, "Affect bursts," in *Emotions: Essays on emotion theory*. Hillsdale, NJ: Lawrence Erlbaum, 1994, pp. 161–193.

[10] M. Schröder, "Experimental study of affect bursts," *Speech Communication*, vol. 40, no. 1-2, pp. 99–116, 2003.

[11] H. Mori, "Morphology of vocal affect bursts: Exploring expressive interjections in Japanese conversation," in *Proc. Interspeech 2015*, 2015, pp. 1309–1313.

[12] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, Jan. 2011.

[13] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoustical Science and Technology*, vol. 33, no. 6, pp. 359–369, 2012.

[14] Y. Den and M. Enomoto, "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation," in *Conversational informatics: An engineering approach*, T. Nishida, Ed. Hoboken, NJ: John Wiley & Sons, 1994, pp. 307–330.

[15] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[16] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech 2011*, 2011, pp. 3081–3084.

[17] K. P. Truong and J. Trouvain, "On the acoustics of overlapping laughter in conversational speech," in *Proc. Interspeech 2012*, 2012, pp. 851–854.

[18] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[20] T. Kawahara, T. Yamaguchi, M. Uesato, K. Yoshino, and K. Takanashi, "Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening," in *Proc. APSIPA ASC 2015*, 2015, pp. 392–395.