Effectiveness of streaming ASR for real-time laughter and screaming detection

Mizuki Kurasawa* and Yoshiko Arimoto†
* Chiba Institute of Technology, Chiba, Japan
E-mail: kurasawa@mac-lab.org
† Chiba Institute of Technology, Chiba, Japan
E-mail: ar@mac-lab.org

***Abstract*—This study investigated the effectiveness of real-time laughter and screaming detection using streaming and non-streaming automatic speech recognition (ASR) models. The streaming and non-streaming ASR models using Transformer architecture were trained and evaluated using spontaneous speech corpora. The two models were compared in terms of detection accuracies for laughter and screaming, speech recognition accuracy, and real-time performance. The results showed that the real-time factor of the streaming ASR model was approximately 0.5 points lower, and the latency was about 0.3 s shorter than that of the non-streaming ASR model. The streaming ASR model processed the speech faster and yielded 7 and 6 points higher detection accuracies for laughter and screaming detection than the non-streaming ASR model, respectively, while yielding the same speech recognition accuracy as the non-streaming ASR model. These results indicate that the use of the streaming ASR model is effective in detecting laughter and screaming.**

I. Introduction

Nonverbal vocalizations such as laughter, screaming, and filled pauses often occur in human-to-human communication. Laughter is one of the social signals that indicates empathy and agreement with another [1]. In addition, laughter has other functions, such as spontaneous laughter that occurs when speakers experience positive emotions and social laughter that maintains the conversation with another [2]. This laughter is unconsciously controlled by speakers and expresses their true feelings [3]. Similarly, screaming is one of the nonverbal vocalizations that includes the speaker's emotion. Screaming includes not only negative emotions such as anger and fear but also positive emotions such as joy [4]. Thus, laughter and screaming are crucial signals for conveying emotion in human-to-human communications. Therefore, automatic detection of laughter and screaming is necessary for human-to-computer interaction to establish more humanlike communication.

Previous studies aimed to recognize such nonverbal vocalizations and speech using a batch process automatic speech recognition (ASR) model (non-streaming) [5]–[7]. Recognition of nonverbal vocalization, i.e., filled pauses and hesitation, and speech, contributed to the improvement of its recognition accuracy [5]. The studies have also been conducted on the detection of laughter and screaming [8]–[14]. Laughter and screaming detection using a non-streaming ASR model achieved detection accuracy of 84.85% in laughter and of 77.67% in screaming [8]. Therefore, nonverbal vocalizations such as laughter and screaming can be automatically recognized or detected using ASR. However, the detection time of the non-streaming ASR model is problematic because it always takes longer processing time than the duration of the input speech. In human-to-human communication, the listener instantly reacts to the speaker's laughing or screaming by laughing along or expressing surprise. On the other hand, machines cannot instantly react as humans do because it takes more time to detect laughter or screaming. The long latency to respond to speakers caused by speech recognition is one of the reasons for the dissimilarity between human-to-computer communication and human-to-human communication. For more humanlike communication, a real-time detection approach for social signals such as laughter and screaming is indispensable.

One possible solution to this problem is the application of a streaming ASR model for detecting laughter and screaming. The streaming ASR model outputs the sentences with sequential processing of fine-grained frames of an input utterance without waiting for finishing the input of the whole utterance. The streaming ASR model achieved a similar character error rate (CER) to the non-streaming ASR model and processed faster than the non-streaming ASR model [15]. For applying to a practical application, end-to-end ASR models were proposed using the streaming technique [15]–[19]. However, it hasn't been tested whether a streaming ASR model can be applied to laughter and screaming detection. One of the potential problems for applying the streaming ASR model to laughter and screaming detection is less information in one input than for the non-streaming ASR model. Less information of input for the streaming ASR model may cause worse performance in laughter and screaming detection than the non-streaming ASR model. Therefore, it should be demonstrated whether a streaming ASR model is effective when applied to laughter and screaming detection.

This study aimed at laughter and screaming detection using the streaming ASR model towards a real-time laughter and screaming detection. The streaming ASR model and the non-streaming ASR model using Transformer provided by ESPnet to compare each model's performance [20]. Two spontaneous speech corpora were used for model development. Each model was compared in terms of detection accuracies for laughter and screaming, speech recognition accuracy, and real-time performance using other corpora which were not used for

979-8-3315-7206-8/25/$31.00

training steps to demonstrate the effectiveness of a streaming ASR model for laughter and screaming detection.

The main contribution of this study is to find out that real-time process does not affect the detection accuracies of laughter and screaming in spontaneous speech, by showing a 7.21% higher F-measure for laughter and a 6.48% higher F-measure for screaming. In summary, the streaming ASR model is effective for real-time laughter and screaming detection.

II. Related works

There are many studies on laughter and screaming detection [8]–[14]. The study [8] proposed a model implementing both laughter and screaming detection and ASR using a multi-task learning method and a self-supervised learning method, named wav2vec 2.0 [21]. As a result, the proposed model yields 84.13% and 62.44% detection accuracies for laughter and screaming, respectively. Another study [14] had two types of experiments on feature sets and models for laughter and screaming detection. As a result of the first experiment on feature sets, the spectral and prosodic feature sets showed the best laughter and screaming detection accuracies at 79.01% for laughter and 65.26% for screaming compared to a spectral feature set or a prosodic feature set. The second experiment on the detection model revealed that the Attention-CTC model was the best by showing the detection accuracies of 78.72% and 67.34% for laughter and screaming, respectively. Those results indicated that the non-streaming ASR model can detect laughter and screaming.

Many studies indicated that recognizing nonverbal vocalization using non-streaming ASR models was also beneficial for recognizing linguistic information [5]–[7]. The study [5] aimed at the recognition of both linguistic information and disfluent acoustic phenomena, i.e., filler and hesitation, by training an end-to-end ASR model. Moreover, the study [6] proposed an ASR model that recognized 9 types of verbal/nonverbal phenomena, e.g., laughter and filler, and linguistic information. As a result, the proposed model showed 2.57% lower CER than the joint CTC-attention Transformer model. Their results provide evidence that recognizing speech and nonverbal vocalizations improves speech recognition accuracy of ASR.

Processing time is another issue for the streaming ASR [15]–[19]. The study [15] proposed a streaming ASR model implemented using Transformer architecture. As a result, the response time of the proposed model was faster than that of the batch model by showing 0.22 seconds faster response time, while the proposed streaming ASR model showed a similar word error rate (WER) at 4.36% as the batch model. Therefore, the streaming ASR model not only processes speech faster but also maintains the same accuracy as the batch model.

III. Corpora

Four spontaneous speech corpora were prepared for our analysis: Action Game Speech Communication corpus (AGSC) [22], Online Gaming Voice chat Corpus with emotional labels (OGVC) [23], Multimodal corpus of Spontaneous Affective Interaction during gameplay (MSAI) [24], and Corpus of Spontaneous Japanese (CSJ) [25]. AGSC aimed at collecting spontaneously produced screaming during spontaneous dialogs between a pair of 24 game players (12 males and 12 females) while playing action games. OGVC includes naturalistic emotional speech during spontaneous Japanese dialogs between pairs or a group of three of 13 game players (9 males and 4 females) while playing a massively multiplayer online role-playing game. MSAI was created to study whether presenting a game event to laughing players would attract them to the virtual world. This corpus recorded spontaneous dialogs from 58 game players (38 males and 20 females), but the dialogs of 12 players were used for our analysis. CSJ consists of spontaneous speech from monologues during academic presentation and its pseudo presentation, and dialogs during interviews. A highly spontaneous dialog dataset (CSJdia) and a lowly spontaneous speech dataset (CSJEval1, CSJEval2, and CSJEval3) were used for our analysis.

Laughter and screaming were defined for our analysis. Laughter is defined as a series of laughter consisting of inhalation and bout [26]. Screaming is defined as an emotionally expressed vocalization that is unconsciously uttered by the speaker due to an unexpected event. Screaming has peculiar characteristics of prosody or voice quality and is hard to be transcribed [27]. AGSC includes 2,511 laughter and 1,315 screaming. OGVC includes 1,345 laughter and 84 screaming, and MSAI includes 726 laughter and 285 screaming.

IV. Experimental setup

A. Model Development

This study adopted streaming and non-streaming ASR models using the Transformer provided by ESPnet [20]. Fig. 1 shows the architecture of the streaming mechanism using Transformer. For the real-time processing in the streaming ASR model, the encoder and decoder operate synchronously. The model processes an input utterance (h_n) in blockwise, while the decoder immediately receives the encoder's output and sequentially outputs the text. These approach enable a real-time process. In contrast, the non-streaming ASR model processes an entire utterance ($h_0, ..., h_n, ..., h_N$) at once, with the encoder and decoder functioning asynchronously. The decoder only receives the encoder's output after the encoder completes to process the input, resulting in the batch process of the text output. This approach prevents real-time processing. Although the streaming ASR model facilitates real-time speech recognition, the short length of frame-wise input would be less informative to improve recognition accuracy. Therefore, the recognition accuracy of the streaming ASR model might be worse than that of the non-streaming ASR model in laughter and screaming detection. For model development, an NVIDIA GeForce RTX 4070 was used. The maximum number of training epochs was set to 300, with early stopping at 9 epochs. The number of batch bins was set to 650,000. All other parameters were set in accordance with [15].

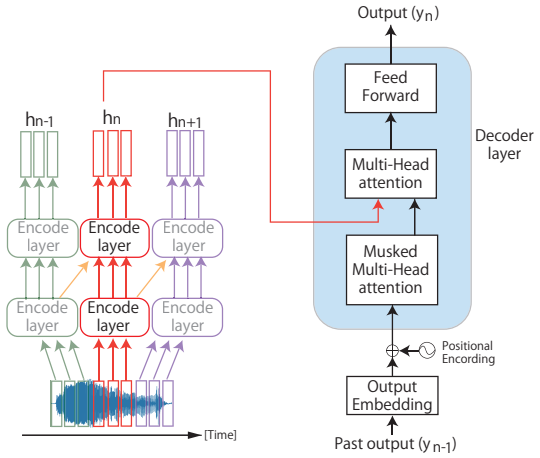


Fig. 1: Architecture of the streaming ASR model

B. Experimental Condition

Our objective is to examine whether the streaming ASR model can detect laughter and screaming using short-duration frame-wise features. This study compared two models, streaming and non-streaming, to demonstrate whether laughter and screaming can be detected within the framework of streaming speech recognition, which requires less input information than the non-streaming ASR model to achieve real-time processing. To compare the effectiveness of two models, this study used AGSC and CSJdia for training, validation, and evaluation (corpus-closed) while the OGVC, MSAI, and CSJEval (CSJEval1, CSJEval2, and CSJEval3) were used only for evaluation (corpus-open). The number of laughter and screaming is few and limited in each corpus. Therefore, the data was augmented to increase the number of these sounds. Data augmentation methods employed pitch shift (by factors of 0.9 and 1.1), time stretch (by factors of 0.9 and 1.1), and noise superimposition at 20 dB. As a result, laughter and screaming increase six-fold compared to the amount of training data. The training dataset includes 49,281 utterances, 9,894 laughter, and 5,244 screaming. The validation dataset has 1,200 utterances (from AGSC and CSJdia each with 600 utterances), 255 laughter and 92 screaming. The evaluation dataset has 1,400 utterances in each corpus (AGSC, CSJdia, OGVC, and MSAI). The AGSC includes 560 laughter and 286 screaming, the OGVC includes 716 laughter and 84 screaming, and the MSAI includes 540 laughter and 271 screaming. The CSJEval1, CSJEval2, and CSJEval3 have each 2,918, 3,067, and 2,484 utterances. To obtain a reliable result, each model was trained once and its predicted value of each sample was estimated three times.

C. Evaluation Indicator

This study calculated F-measure and CER as metrics for the detection accuracy of laughter and screaming, and speech recognition accuracy. Real-time Factor (RTF) and Latency were also calculated to evaluate the effect of real-time processing on laughter and screaming detection. RTF is the rate of time required for processing a speech against the duration of the speech when the duration of the speech is 1.0. RTF was calculated by

$$RTF = \frac{ProcessTime[s]}{Duration[s]}. \quad (1)$$

RTF less than 1 indicates fast processing speed. Latency is the time to process a speech. Latency was calculated by

$$Latency = \frac{1}{N}\sum_{i=1}^{N} ProcessTime_i. \quad (2)$$

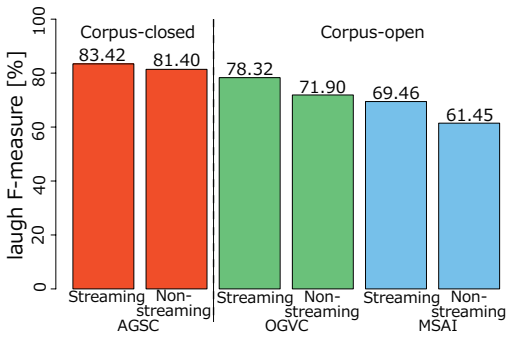
The higher F-measure represents the better detection accuracy. The lower CER represents the better speech recognition accuracy. The lower RTF and Latency show the faster process speed. All the evaluations were averaged across the three trials.

V. Result

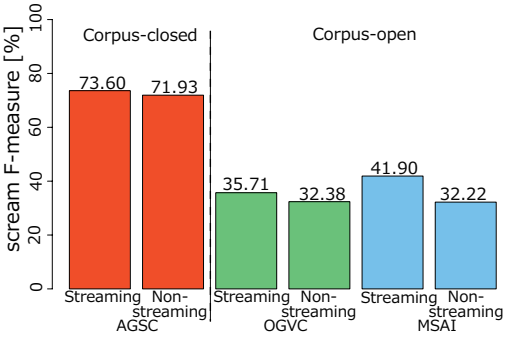
Fig. 2 shows the F-measure for laughter and screaming detection for the two models. Fig. 2(a) shows the F-measure for laughter detection, and Fig. 2(b) shows for screaming detection performance. The vertical axis of Fig. 2 represents the F-measure, and the horizontal axis represents each corpus used for decoding. Each corpus has two values: the left value shows the performance of the streaming ASR model and the right shows that of the non-streaming ASR model. The red bars in Fig. 2 show the result for laughter and screaming detection to use AGSC (corpus-closed condition) and green and blue show the result for laughter and screaming detection to use OGVC and MSAI (corpus-open condition), respectively. According to Fig. 2(a), the mean F-measure of the streaming ASR model for laughter detection across the corpora was 73.89% (OGVC: 78.32%, MSAI: 69.46%), while the non-streaming ASR model was 66.68% (OGVC: 71.90%, MSAI: 61.45%). Similarly, according to Fig. 2(b), the mean F-measure of the streaming ASR model for screaming detection across the corpora was 38.81% (OGVC: 35.71%, MSAI: 41.90%), while the non-streaming ASR model was 32.33% (OGVC: 32.43%, MSAI: 32.22%).

Fig. 3 shows the mean CER across the three trials. The vertical axis represents the CER, and the horizontal axis indicates the corpus used as the evaluation data. Similar to Fig. 2, there are two bar plots for each corpus; the left one represents the results of the streaming ASR model and the right one represents the results of the non-streaming ASR model. Red and orange bars in Fig. 3 indicate the CERs for the AGSC and CSJdia (corpus-closed), while green, blue, purple, pink, and violet represent the CERs for OGVC, MSAI, CSJEval1, CSJEval2, and CSJEval3 (corpus-open), respectively. The mean CER of the streaming ASR model across the corpora was 30.93% (OGVC: 42.74%, MSAI: 41.08%, CSJEval1: 27.51%, CSJEval2: 24.59%, CSJEval3: 18.75%). On the other hand, the mean CER of the non-streaming ASR model across the corpora was 30.35% (OGVC: 42.71%, MSAI: 39.24%, CSJEval1: 27.27%, CSJEval2: 23.91%, CSJEval3: 18.63%).

Fig. 4 shows the mean RTF across the three trials. The vertical axis shows the RTF, and the horizontal line at 1.0



(a) Result for laughter detection (F-measure).



(b) Result for screaming detection (F-measure).

Fig. 2: The F-measures for laughter and screaming detection using the streaming and non-streaming ASR models (each color represents each corpus). The left bar shows the streaming ASR model, and the right bar shows the non-streaming ASR model for each corpus.

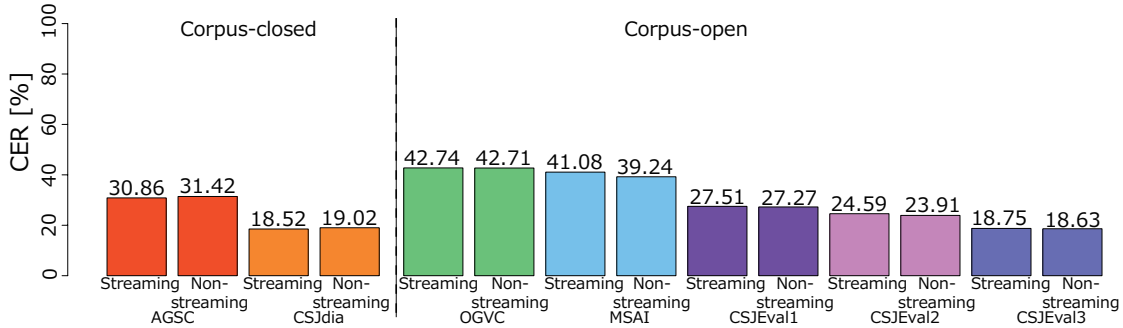


Fig. 3: The CERs for the streaming and non-streaming ASR model (each color represents each corpus). The left bar shows the streaming ASR model, and the right bar shows the non-streaming ASR model for each corpus.

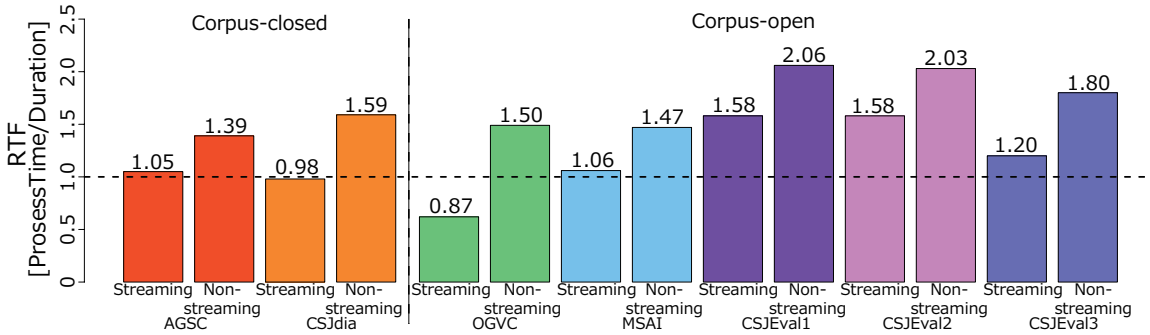


Fig. 4: The RTFs using the streaming and non-streaming ASR models (each color represents each corpus). The left bar shows the streaming ASR model, and the right bar shows the non-streaming ASR model for each corpus.

represents the duration of the target audio. Other details are consistent with Fig. 3. The mean RTFs across the corpora were 1.26 for the streaming ASR model (OGVC: 0.87, MSAI: 1.06, CSJEval1: 1.58, CSJEval2: 1.58, CSJEval3: 1.20) and 1.77 for the non-streaming (OGVC: 1.50, MSAI: 1.47, CSJEval1: 2.06, CSJEval2: 2.03, CSJEval3: 1.80).

Fig. 5 shows the mean latency across the three trials. The vertical axis shows the latency, and the other details are consistent with Fig. 3 and Fig. 4. The mean latency of the streaming ASR model across the corpora was 3.98 s (OGVC: 1.99 s, MSAI: 3.42 s, CSJEval1: 5.23 s, CSJEval2: 5.44 s, CSJEval3: 3.83 s), whereas the non-streaming ASR model recorded a latency of 4.28 s (OGVC: 2.58 s, MSAI: 3.39 s, CSJEval1: 5.54 s, CSJEval2: 5.63 s, CSJEval3: 4.28 s).

VI. Discussion

The F-measure for laughter and screaming detection (Fig. 2) show that the streaming ASR model outperforms the non-streaming ASR model by showing 7.21% higher accuracy for laughter and 6.48% higher accuracy for screaming. The streaming ASR model demonstrated superior detection performance for laughter and screaming, despite requiring less input information for sequential processing. In other words, the streaming mechanism did not affect the accuracy of detecting laughter and screaming. The false detection counts for laughter

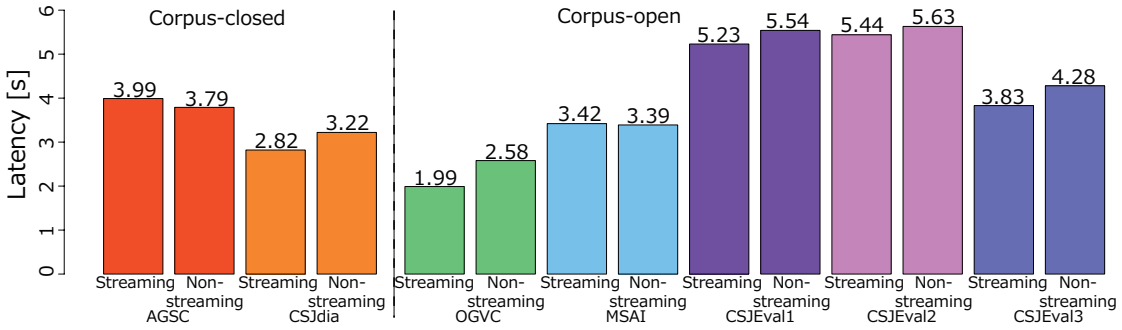


Fig. 5: The Latency using the streaming and non-streaming ASR models (each color represents each corpus). The left bar shows the streaming ASR model, and the right bar shows the non-streaming ASR model for each corpus.

and screaming in each of the three trials showed that the non-streaming ASR model produced more errors than the streaming ASR model. For laughter, the streaming ASR model had 27 false detections while the non-streaming ASR model had 44. Similarly, the false detection counts for screaming were 33 and 38 for the streaming and the non-streaming ASR models, respectively. The precise inspection of the outputs from each model shows that the non-streaming ASR model did not output explicit labels for laughter or screams. Instead, it outputs the context of phones. i.e., ”uN” for laughter and ”a:” for screaming. Since the streaming ASR model has the structure of sequential processing, the detection of laughter and screaming with short duration is only affected by the acoustic properties contained in each frame, rather than the context of the speech. Consequently, laughter and screaming with specific expressions might be detected as phones with the closest acoustic properties within the frame. Another reason why the model outputs ”a:” for screaming might be that no prosodic information was provided to the model. The training data includes a high frequency of the ”a:” phone (5,600 instances) and many screaming in the evaluation set were also uttered as ”a:”. As only speech waveforms were used to train the model, the difference in prosodic information between ”a:” in speech and ”a:” in screaming might not be trained. It has been shown that screaming have a higher fundamental frequency (F_0) than speech [28]. Moreover, the previous study of [14] exhibited a higher result than our work, using the 2013 ComPasrE feature set. Since this feature set includes prosodic features, the study [14] showed better results in discriminating screaming from speech. For more reliable detection of screaming, the model needs both the speech waveform and prosodic information.

The CER showed that the streaming and non-streaming ASR models performed similar CERs of 30.93% and 30.35%, respectively (Fig. 3). The streaming ASR model provided speech recognition results in real-time without waiting for the end of an utterance. In contrast, the non-streaming ASR model processes the speech only after the utterance has ended. Therefore, there is a difference in the amount of information that each model can process for speech recognition. Despite this difference, both models demonstrated comparable. This aligns with the findings of previous research on streaming speech recognition [15].

The RTF results (Fig. 4) show that the streaming ASR model has 0.51 lower RTF than the non-streaming ASR model. The latency results (Fig. 5) also show that the streaming ASR model was faster than the non-streaming ASR model by 0.3 s. The streaming ASR model's architecture of sequential speech processing is working effectively, and it has achieved faster processing speeds than the non-streaming ASR model.

Based on the above results, the streaming speech recognition model showed speech recognition accuracy comparable to the non-streaming ASR model, superior laughter and screaming detection accuracy, and faster processing speed. Consequently, the streaming ASR model is more practical and effective for real-time laughter and screaming detection.

VII. Conclusion

This study demonstrated the effectiveness of the streaming ASR model for detecting laughter and screaming. The two models were trained and compared; one based on sequential processing (streaming) and the other based on batch processing (non-streaming). The two models were evaluated using four indices: F-measure, CER, RTF, and latency. The results showed that the streaming ASR model had shorter latency and faster processing speed compared to the non-streaming ASR model. While the both models yielded similar speech recognition accuracies (CER), the streaming ASR model was 7.21% more accurate in detecting laughter and 6.48% more accurate in detecting screaming. In summary, our study proved that the model is effective at detecting laughter and screaming through streaming speech recognition. To enhance the accuracy of detecting these vocalizations, further improvement of the model performance will be planned by implementing multi-task learning that focuses on both speech recognition and the detection of laughter and screaming.

Acknowledgment

This study was partially supported by JSPS KAKENHI Grant Number JP22K18477 and by Kayamori Foundation of Information Science Advancement (K36 Ken-XXIX-662).

References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[2] K. Wang, C. Ishi, and R. Hayashi, "Acoustic analysis of several laughter types in conversational dialogues," in *Proc. Speech Prosody*, 2024, pp. 667–671.

[3] R. R. Provine, "Laughing, tickling, and the evolution of speech and self," *Current Directions in Psychological Science*, vol. 13, no. 6, pp. 215–218, 2004.

[4] D. Handa and R. Vig, "Distress screaming vs joyful screaming: An experimental analysis on both the high pitch acoustic signals to trace differences and similarities," in *Proc. Indo - Taiwan ICAN*, 2020, pp. 190–193.

[5] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-end spontaneous speech recognition using disfluency labeling," in *Proc. Interspeech 2022*, 2022, pp. 4108–4112.

[6] N. Shione, yukoh Wakabayashi, and N. Kitaoka, "Automatic speech recognition model simultaneously recognizes linguistic information and verbal/non-verbal phenomena," *in Proc. 2023 Autumn Meeting Acoustical Society of Japan*, 2023, (in Japanese).

[7] H. Inaguma, K. Inoue, M. Mimura, and T. Kawahara, "Social signal detection in spontaneous dialogue using bidirectional LSTM-CTC," in *Proc. Interspeech 2017*, 2017, pp. 1691–1695.

[8] T. Matsuda and Y. Arimoto, "Laughter and screaming detection utilizing automatic apeech recognition using wav2vec 2.0," *in Proc. the Speech Processing Meeting The Acoustical Society of Japan*, vol. 4, no. 1, pp. 19–24, Feb. 2024, (in Japanese).

[9] J. Gillick, W. Deng, K. Ryokai, and D. Bamman, "Robust laughter detection in noisy environments," in *Proc. Interspeech 2021*, 2021, pp. 2481–2485.

[10] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Laughter and filler detection in naturalistic audio," in *Proc. Interspeech 2015*, 2015, pp. 2509–2513.

[11] S. Petridis, M. Leveque, and M. Pantic, "Audiovisual detection of laughter in human-machine interaction," in *Proc. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 129–134.

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

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

[14] T. Matsuda and Y. Arimoto, "Detection of laughter and screaming using the Attention and CTC models," in *Proc. Interspeech 2023*, 2023, pp. 1025–1029.

[15] E. Tsunoo, Y. Kashiwagi, and S. Watanabe, "Streaming Transformer ASR with blockwise synchronous beam search," in *Proc. IEEE SLT Workshop*, 2021, pp. 22–29.

[16] R. Wang, S. Nadig, D. Kulko, *et al.*, "Double decoder: Improving latency for streaming end-to-end ASR models," in *Proc. the ICNLSP*, 2024, pp. 83–91.

[17] Y. Sudo, S. Muhammad, K. Nakadai, J. Shi, and S. Watanabe, "Streaming automatic speech recognition with re-blocking processing based on integrated voice activity detection," in *Proc. Interspeech 2022*, 2022, pp. 4641–4645.

[18] W. Kang, Z. Yao, F. Kuang, *et al.*, "Delay-penalized transducer for low-latency streaming ASR.," in *Proc. ICASSP*, IEEE, 2023, pp. 1–5.

[19] X. Song, D. Wu, Z. Wu, *et al.*, "Trimtail: Low-latency streaming ASR with simple but effective spectrogram-level length penalty," in *Proc. ICASSP 2023*, 2023, pp. 1–5.

[20] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech 2018*, 2018, pp. 2207–2211.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460.

[22] H. Mori and Y. Kikuchi, "Gaming corpus for studying social screams," in *Proc. Interspeech 2020*, 2020, pp. 3132–3135.

[23] 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.

[24] M. Fukuda and Y. Arimoto, "Effects of reactions generated by a virtual world on game players under laughing/non-laughing conditions," *in Proc. SIG-SLUD, JSAI,*, vol. 97, pp. 92–97, 2023, (in Japanese).

[25] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, paper MMO2.

[26] H. Mori, T. Nagata, and Y. Arimoto, "Conversational and social laughter synthesis with WaveNet," in *Proc. Interspeech 2019*, 2019, pp. 520–523.

[27] K. Shiratori, M. Okubo, T. Matsuda, and Y. Arimoto, "Scream and shout annotation for spontaneous dialog speech," *in Proc. Language Resources Workshop*, vol. 1, pp. 365–374, 2023, (in Japanese).

[28] J. H. L. Hansen, M. K. Nandwana, and N. Shokouhi, "Analysis of human scream and its impact on text-independent speaker verificational," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. 2957–2967, 2017.