INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Detection of laughter and screaming using the attention and CTC models

Takuto Matsuda[1], *Yoshiko Arimoto*[1]

[1]Chiba Institute of Technology

matsuda@mac-lab.org, ar@mac-lab.org

Abstract

This study aimed to detect social signals, such as laughter and screams, in real environments. Social signals influence human-to-human communication. To effectively apply these signals in various systems, computer systems must appropriately detect social signals. In this study, social signal detection (SSD) experiments were conducted to demonstrate which of three feature sets, i.e., a spectral feature set, prosodic feature set, and spectral and prosodic feature set, was best for detecting laughter and screaming. The results showed that using both the spectral and prosodic feature sets yielded the best performance, with 81.83% accuracy for laughter and 81.68% accuracy for screams. Moreover, the detection model comparison results revealed that the bidirectional long short-term memory (BiLSTM)-connectionist temporal classification (CTC) yielded the best laughter detection performance, while attention-CTC was best for scream detection. These results suggest that CTC is effective for SSD.

Index Terms: laughter, screams, attention, CTC, detection experiment

1. Introduction

Social signals [1, 2, 3] are attitudes and postures that represent mental states that occur during human-to-human interactions. One of the communication modalities for producing social signals is the use of vocal expressions, such as laughter and screaming, which indirectly convey the speaker's emotional state to listeners. While laughter can indirectly represent positive emotions such as happiness and enjoyment, it can also express negative emotions such as fake smiles in social situations [4], taunting laughter, or schadenfreude [5, 6]. Screams can be positive screams representing extreme excitement or negative screams indicating fear or a need for help [7, 8]. The influence of laughter and screaming on human-to-human interactions is very powerful. Therefore, computer systems should be able to detect a speaker's laughter and screams to judge their emotional state, which can improve the performance of human-to-machine interactions by generating appropriate responses to the speaker.

Several studies have been conducted on the automatic detection of laughter and screams [9, 10, 11, 12, 13, 14, 15, 16, 17]. These studies aimed to detect and discriminate laughter from filler sounds and speech and screams from noise and speech, targeting the detection of sound events that occur in the same environment. Although laughter and screaming tend to have similar acoustic properties, such as higher fundamental frequency (f_o) values than typical speech [18, 19, 20], no study has investigated the detection of both laughter and screaming. The present study focused on laughter and screams due to their remarkable acoustical characteristics representing stronger emotional states in public environments such as amusement parks and sports stadiums, where people are excited during communication. Accurately detecting these two acoustically similar but emotionally different vocal phenomena is a high-priority and critical task in social signal detection (SSD). A previous study [21] that focused on discriminating between laughter and scream segments yielded high discrimination accuracies of 93.52% for the deep neural network (DNN) model and 95.54% for the support vector machine (SVM) model using 6,373 acoustic features. However, their experiments included only three classification categories for the given segmented vocal events including laughter and screams as opposed to the detection of laughter and screams occurring in time-series speech data; thus, a study on the detection of laughter and scream is needed for applications in real-world environments.

For SSD, it is necessary to train an end-to-end model that directly infers social signals on an event-by-event basis rather than on a frame-by-frame basis, considering its application to real-world environments [12]. In automatic speech recognition (ASR), the connectionist temporal classification (CTC) model [22] and attention model [23, 24] have been proposed as end-to-end models, achieving highly accurate results. In SSD, laughter and filler sounds were used in the bidirectional long short-term memory (BiLSTM)-CTC model to achieve a detection accuracy that exceeded that of the DNN-hidden Markov model (HMM), indicating that the CTC model can determine the location of social signals in time-series data [12]. In addition, a scream detection system consisting of an autoencoder module and a classifier module using an attention mechanism was constructed, achieving a detection accuracy of 71.05% [17]. However, it is not clear how effective the attention mechanism and CTC model are for detecting laughter and screaming. The development of an effective end-to-end model for SSD would be helpful in the construction of complex models for SSD and ASR.

In this study, three different end-to-end models, the CTC model, attention model, and attention-CTC model, are constructed to directly detect spontaneous laughter and screams. The models are trained with three acoustic feature sets as input and evaluated using three spontaneous dialogue speech corpora. By performing SSD using the framework of ASR, it is expected to improve the performance of both ASR and SSD. The present study focused on the detection and interrelationships of laughter and screaming, thus, speech recognition was not performed.

This is the first step towards a universal social signal detection study encompassing laughter, screams, sighs, fillers, coughs, cries, and more. The main contributions of this paper can be summarized as follows.

- The effectiveness of acoustic features for four-class detection, including laughter and screams, was elucidated.
- The competence and robustness of end-to-end models for SSD was verified through comparisons with three models.

10.21437/Interspeech.2023-1412

2. Speech materials

2.1. Corpora

In these detection experiments, the Action Game Spoken Communication Corpus (AGSC) [7] was used as training data for the model. In the AGSC, the dialogue speech of 24 participants (12 males and 12 females) was recorded for an average of 60.7 minutes for each speaker. In the collection of the AGSC, unconscious laughter and screams were induced by having the participants, who were grouped in pairs, play action-oriented games against each other in a soundproof room.

The Online Gaming Voice Chat Corpus with Emotional Labels (OGVC) [25] was used for evaluation experiments. The OGVC is a publicly available spontaneous dialogue corpus that includes dialogue speech from 13 speakers (9 males and 4 females). The participants cooperatively played an online game in pairs or groups of three, conversing through voice chat, and an average of one hour of audio was recorded for each participant. In this detection experiment, the dialogue speech of 7 participants (5 males and 2 females) with laughter and screaming labels was used.

The multimodal corpus of spontaneous affective interaction during gameplay (MSAI) [26], which contains more screaming sounds than the OGVC, was used as the test data for the evaluation experiments. The MSAI was recorded with the aim of examining whether game events can be presented in response to spontaneous laughter to encourage cognitive and emotional attraction by having the participants played the game. There were 58 participants played the game for 30 minutes twice. For the convenience of the game experiment, the participants were placed next to each other to play the game, so some speech of different speakers was recorded. The dialogue speech of 12 participants was used in this detection experiment as evaluation data.

2.2. Definition of acoustic events

In this study, laughter is defined as a series of laughter (laughter episode) composed of exhalations (bouts) and inhalations [27, 28]. A laughter episode consists of bouts and inhalations with laughter. Although the spontaneous dialogue speech corpora contain speech laughs, i.e., laughing speech, they are defined as speech in this study, not as laughter episodes. Each corpus includes 2,376 (1.43s/segment, AGSC), 1,208 (0.70s/segment, OGVC), and 720 (1.30s/segment, MSAI) laughter segments.

Moreover, screaming is defined as an emotionally expressed interjection that is unconsciously uttered by the speaker due to an unexpected event that has a unique prosody or voice quality. In some of the recordings, speech and screaming cooccur; however, this is not defined as screaming in this study and is instead defined as speech. The definition of screaming varies between studies, including sustained loud voices [8, 14] and spontaneously produced prosodic-specific vocal events [7]. Since the acoustic properties of acted speech and spontaneous speech differ [25], screaming was defined on the basis of [7]. Furthermore, the context and situation should be considered to determine whether the signal is screaming or not. Since the definition in previous studies did not include that detail, it was added to the definition in our study. Each corpus includes 1,320 (0.57s/segment, AGSC), 85 (0.45s/segment, OGVC), and 284 (0.46s/segment, MSAI) scream segments.

Speech is defined as an interpausal unit based on 400 ms signal segments. Speech segments that did not include laughter, screams, or silence within 400 ms (pause) were excluded. Each corpus includes 19,867 (0.95s/segment, AGSC), 5,696 (0.88s/segment, OGVC), and 10,461 (0.91s/segment, MSAI) speech segments and 7,781 (0.15s/segment, AGSC), 1,106 (0.18s/segment, OGVC), and 1,313 (0.33s/segment, MSAI) pause segments. In this study, the single sequences in our experiments were less than 10 seconds, and laughter, screaming, speech, and pauses were labelled in segments in each sequence for each corpus. After a label was assigned, the event was treated as a single event, regardless of the duration of the utterance, until another defined event occurred. The inter-rator reliability scores among annotators and other details are described in [28, 29, 30].

3. Acoustic feature set comparison

3.1. Acoustic feature sets

The detection experiments were performed to evaluate the accuracy of detecting laughter and screaming with three corpora: a spectral feature set, a prosodic feature set, and a feature set using both spectral and prosodic features. The results clarify which feature sets are most effective for SSD.

The acoustic features included log-mel filter bank features (FBANK) for spectral information and a created ComParE feature set [31] for prosodic information, which were extracted for 25 ms frames with a frame shift of 10 ms. The FBANK features have been used in previous ASR and SSD studies [11]. This feature set consists of a 40-channel log-mel filter bank, as well as the delta and acceleration coefficients, for a total of 120 dimensions. The ComParE feature set was provided by openSMILE [32] from the Interspeech 2013 ComParE Challenge [31]. This feature set consists of 33-dimensional features, including the log energy (+ the delta and acceleration coefficients), voice probability, f_o, zero crossing rate (ZCR), and the harmonics-to-noise ratio (HNR) (+ the delta coefficient), for a total of 11 features, as well as the mean in the frame neighbourhood and the standard deviation. The ComParE features have previously been used in SSD experiments and are detected with high accuracy [11, 13]. The MFCCs are excluded from this feature set because the MFCCs were less accurate than the FBANK features for detecting social signals according to [11]. The third feature set consists of the aforementioned FBANK feature set and the ComParE feature set (FBANK+ComParE) for a total of 153 dimensions.

3.2. Experimental setup

The AGSC data were used as the training, validation and test data. To balance the amount of data for the four classes of acoustic events, the number of laughter segments was adjusted to match the number of scream segments, namely, 1320 data points. However, since the number of speech and pause segments was very large, the number of labelled speech and pause segments was minimized based on the number of single sequences of vocal events containing social signals. The created AGSC dataset was randomly divided into training, validation and test sets at a ratio of 7:2:3, resulting in a total of 1,232, 352, and 528 audio sequences in each set. Since a previous study on ASR [33, 34] successfully improved model accuracy through data augmentation methods, our original data were augmented with the following five methods, increasing the amount of training data sixfold to 7,392 sequences: (a) white noise superimposed at 20 dB, (b) time stretching to increase the duration by a factor of 1.1, (c) time stretching to decrease the duration by a factor of 0.9, (d) pitch shift to increase the pitch by a factor of

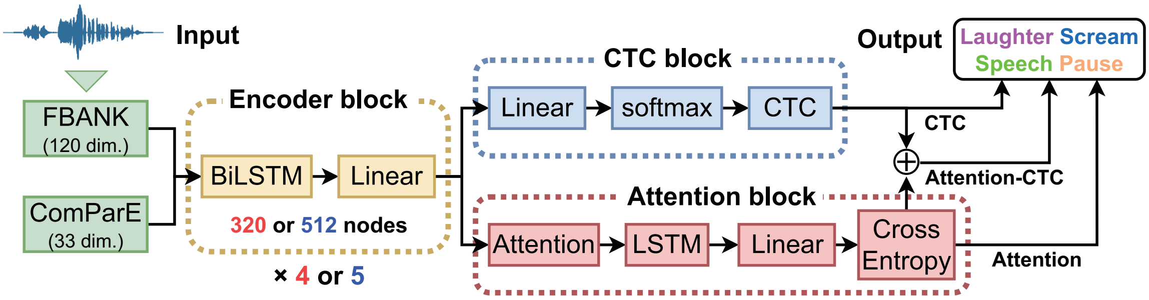

Figure 1: *The architectures of CTC, attention, and attention-CTC models in this experiments.*

1.1, and (e) pitch shift to decrease the pitch by a factor of 0.9. The numbers of sequences of OGVC and MSAI test data were also adjusted based on the amount of AGSC test data, namely, 528 sequences. The OGVC and MSAI test data were adjusted based on the number of sequences in the AGSC test data, but the class balance in each corpus was maintained.

The model used in this detection experiment is the BiLSTM-CTC model, which was also used in [12]. Figure 1 shows the architecture of CTC model. The network consists of 320 or 512 nodes with 4 or 5 connected layers (BiLSTM and linear); in addition, another linear layer and the softmax function were added at the end of the model to calculate the CTC loss function. The optimization function is Adadelta, the batch size is 10 audio sequences, and the learning rate is 1.0. To improve the learning efficiency, early stopping was used, where the learning rate was halved if the error value of the validation data did not improve, and training was stopped if the error value did not improve during three consecutive epochs. The maximum number of training epochs was 60. For fast processing, two subsampling layers were used to reduce the number of frames by half after the second and third connected layers.

The F-measure was used as an evaluation metric to assess the accuracy of detecting laughter and screaming [35]. For each condition, two parameter combinations (320 nodes with 4 connected layers and 512 nodes with 5 connected layers) were trained three times each, for a total of six results, and the average F-measure was calculated.

3.3. Results and discussion

Table 1 shows the results of the laughter, scream, and speech feature sets for each corpus. In the corpus-closed evaluation experiment using the AGSC, laughter was detected with an accuracy of 81.83% and screaming was detected with an accuracy of 81.68% using the FBANK+ComParE feature set. Laughter and screaming were detected with an F-measure of approximately 80%, indicating that laughter and screaming can be distinguished. In the corpus-open evaluation experiment using the OGVC and MSAI, laughter was detected with accuracies of 79.38% (OGVC) and 78.63% (MSAI), and screaming was detected with accuracies of 58.89% (OGVC) and 71.63% (MSAI), with the best results achieved with the FBANK+ComParE feature set. Laughter and screaming were detected over 70% in the two-speaker mixed vocal audio samples in the MSAI dataset indicates the robustness of the model in real-world environments.

Comparing the F-measures of laughter and screams for all feature sets, the FBANK+ComParE feature set was superior to the other feature sets. This result indicates that spectral and prosodic features are essential for SSD. Although the ComParE feature set has only 33 dimensions, it shows a difference in F-measure of ±1.84–6.29% from the FBANK feature set, which is nearly equivalent. This result suggests that sudden prosodic changes that are unique to affect bursts can be distinguished from speech. Thus, prosodic features may be essential for SSD.

Table 1: *Results of acoustic feature set comparison among different corpora for detection F-measures [%] of laughter, screaming, and speech.*

Event	Feature set	**AGSC**	**OGVC**	**MSAI**
Laughter	FBANK	80.19	71.31	76.29
	ComParE	77.31	74.01	70.00
	FBANK+ComParE	**81.83**	**79.38**	**78.63**
Screaming	FBANK	80.84	52.52	64.18
	ComParE	78.57	50.68	68.47
	FBANK+ComParE	**81.68**	**58.89**	**71.63**
Speech	FBANK	78.46	**72.35**	64.69
	ComParE	74.83	63.36	63.41
	FBANK+ComParE	**78.96**	71.06	**69.40**

4. Detection model comparison

4.1. End-to-end models

In these detection experiments, the FBANK+ComParE feature set was applied to three end-to-end models, and the detection accuracy of laughter and screaming was evaluated based on three corpora. Then, an effective model for SSD was developed. The following three types of end-to-end models were tested: the CTC (baseline) model, attention model, and attention-CTC model. Figure 1 shows the architecture of attention and attention-CTC models.

The attention model [23, 24] is an encoder-decoder model that extracts latent representations by focusing on input features that are important according to the token output. In this work, the attention model consists of 320 or 512 nodes with 4 or 5 connected BiLSTM and linear layers in the encoder. Then, the attention weights are computed based on the output of the encoder. In the decoder, an LSTM network is established using the attention weights, the estimated previous output, and prior information as inputs. Then, the social signal is output through a linear layer. The loss function is a cross-entropy function. Teacher forcing is conducted to ensure stable learning.

The attention-CTC model is an attention model that adds the CTC loss function to the cross-entropy loss function used by the attention model [36]. This suppresses unnatural alignment estimation caused by the attention mechanism and improves the ASR accuracy by stabilizing attention learning. In the attention-CTC model used in this experiment, the encoder and decoder used the same network as in the attention model. In addition, the CTC loss value is calculated based on the encoder output in

Table 2: *Results of end-to-end model comparison among different corpora for detection F-measures [%] of laughter, screaming, and speech.*

Event	Model	**AGSC**	**OGVC**	**MSAI**
Laughter	CTC (*baseline*)	**81.83**	79.38	**78.63**
	Attention	80.71	79.75	76.59
	Attention-CTC	81.40	**80.54**	76.90
Screaming	CTC (*baseline*)	81.68	58.89	71.63
	Attention	82.92	53.75	69.95
	Attention-CTC	**83.12**	**61.68**	**73.00**
Speech	CTC (*baseline*)	**78.96**	**71.06**	69.40
	Attention	73.42	64.62	62.06
	Attention-CTC	76.66	69.30	**69.86**

addition to the cross-entropy error, and the weight of the CTC loss function is set to 0.2.

Other parameter values, optimization functions, and methods for early stopping and subsampling are the same as those used in Section 3.

4.2. Experiments setup

Our detection experiments use only the FBANK+ComParE feature set. The details of the dataset and the CTC model are the same as those in Section 3. The evaluation metrics used in this experiment are the F-measure [35].

4.3. Results and discussion

Table 2 shows the detection results of the three end-to-end models based on the F-measures of laughter, screaming, and speech. For laughter, the CTC model was the most accurate based on the AGSC (81.83%) and MSAI (78.63%) datasets, while the attention-CTC model was the most accurate based on the OGVC (80.54%) dataset. For screaming, the attention-CTC model was the most accurate based on all corpora, with accuracies of 83.12% (AGSC), 61.68% (OGVC), and 73.00% (MSAI). The models with the best performance (CTC and attention-CTC) were both constructed with CTC. Thus, the models that include CTC may be effective for SSD.

For laughter detection, the CTC model is superior to the attention-CTC model. Laughter has different sequential acoustic characteristics because it is composed of a time series of voiced exhalation, voiced inhalation, unvoiced exhalation, and unvoiced inhalation components. The CTC model, which learns the sequential characteristics of speech signals, might learn complicated sequence characteristics of laughter better than the other models and thus show better laughter detection performance. The attention model yielded the worst performance among the three models for all event detection experiments. The attention model often produced errors such as "speech -> pause -> speech -> pause -> speech -> ..." in the detection results. Attention mechanisms require a large amount of training data, and the attention mechanism may not have been trained well in this experiment because of the insufficient training data. Additional detection experiments were conducted for the attention-CTC model using smaller training datasets, i.e., the AGSC dataset without data augmentation. As a result, similar errors were observed. Therefore, the insufficient training dataset may cause the lower accuracy of the attention model. Since collecting more spontaneous laughter and screaming data would be costly, training data cannot be easily increased. Thus, it is important to explore data augmentation methods specialized for SSD.

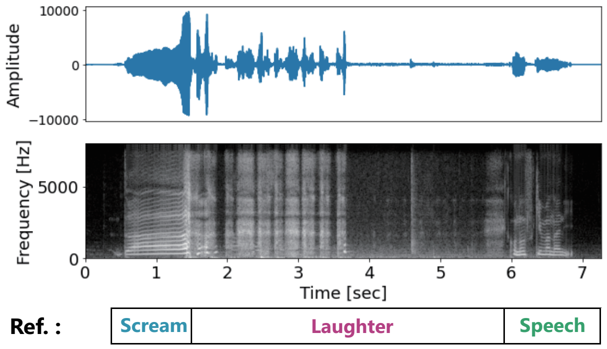

Figure 2: *Audio waveforms and spectrograms of audio sequences with all events correctly distinguished.*

In this experiment, misdetected sequence data were investigated to determine how laughter and screaming were confused. More accurate detection results were obtained with audio sequences consisting of fewer events than those consisting of many labelled events. Figure 2 shows the audio waveforms and spectrograms of the audio sequences containing laughter, screaming, and speech that were correctly distinguished by all models. Laughter was often confused with pauses. This may be due to the acoustic similarity between pauses and the unvoiced inhalation in the latter part of the laughter signal, as shown in Fig. 2 (between 4 and 6 seconds). Screaming is often confused with speech. The spectrogram of the scream shows clear harmonics. This acoustic tendency is similar to the harmonics of the speech, indicating that misdetection may have occurred.

5. Conclusion

In this study, laughter and screaming detection experiments were conducted to facilitate human-machine interactions. First, four-class SSD experiments including laughter and screaming were conducted using three feature sets. The results showed that in the closed corpus experiments, 81.83% of laughter and 81.68% of screaming were accurately detected using the FBANK+ComParE feature set, indicating that laughter and screaming were detectable. In the open corpus experiments, 79.38% of laughter and 71.63% of screaming were accurately detected using the FBANK+ComParE feature set, indicating that the model was robust. Then, detection experiments were conducted with three end-to-end models, and the results showed that the CTC and attention-CTC models achieved the most accurate laughter and screaming detection results, respectively, indicating that CTC is effective for SSD. Consequently, the attention-CTC model was effective for SSD when using large amounts of data. Future research should be developed a more appropriate SSD method to overcome the influence of a lack of laughter and screaming speech material on the results, and an architecture simultaneously performs SSD and ASR to accurately distinguish various social signals from speech.

6. Acknowledgements

This study was partially supported by JSPS KAKENHI Grant Numbers 22K18477 and 22K12107 and by the 30th Anniversary Research Grant of Hayao Nakayama Foundation for Science, Technology and Culture.

7. References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing : Survey of an emerging domain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] I. Poggi and F. D'Errico, "Social signals: a framework in terms of goals and beliefs," *Cognitive Processing*, vol. 13, no. 2, pp. 427–445, 2012.

[3] P. Brunet and R. Cowie, "Towards a conceptual framework of research on social signal processing," *Journal on Multimodal User Interfaces*, vol. 6, pp. 101–115, 2012.

[4] J. Vettin and D. Todt, "Laughter in conversation: features of occurrence and acoustic structure," *Journal of Nonverbal Behavior*, vol. 28, pp. 93–115, 2004.

[5] D. P. Szameitat, K. Alter, A. J. Szameitat, C. J. Darwin, D. Wildgruber, S. Dietrich, and A. Sterr, "Differentiation of emotions in laughter at the behavioral level," *Emotion*, vol. 9, pp. 397–405, 2009.

[6] D. Szameitat, A. Szameitat, and D. Wildgruber, "Vocal expression of affective states in spontaneous laughter reveals the bright and the dark side of laughter," *Scientific Reports*, vol. 12:5613, 2022.

[7] H. Mori and K. Yuki, "Gaming corpus for studying social screams," in *Proc. Interspeech 2020*, 2020, pp. 520–523.

[8] R. V. D. Handa, "Distress screaming vs joyful screaming: An experimental analysis on both the high pitch acoustic signals to trace differences and similarities," *Indo - Taiwan 2nd International Conference on Computing, Analytics and Networks, Indo-Taiwan ICAN 2020 - Proceedings*, pp. 190–193, 2020.

[9] H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 4282–4287.

[10] A. Batliner, S. Steidl, F. Eyben, and B. Schuller, "On laughter and speech-laugh, based on observations of child-robot interaction," *Arxiv*, 2019.

[11] G. Gosztolya, T. Grósz, and L. Tóth, "Social signal detection by probabilistic sampling dnn training," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 164–177, 2020.

[12] H. Inaguma, K. Inoue, M. Mimura, and T. Kawahara, "Social signal detection in spontaneous dialogue using bidirectional lstm-ctc," in *Proc. INTERSPEECH 2017*, 2017.

[13] H. Kaya, A. Ercetin, A. Salah, and S. Gurgen, "Random forests for laughter detection," in *Proc. WASSS 2013*, 2013, pp. 1–5.

[14] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[15] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, p. 161–165.

[16] T. Fukumori, "Deep Spectral-Cepstral Fusion for Shouted and Normal Speech Classification," in *Proc. Interspeech 2021*, 2021, pp. 4174–4178.

[17] S. Baghel, M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, "Automatic detection of shouted speech segments in indian news debates," in *Proc. Interspeech 2021*, 2021, pp. 4179–4183.

[18] E. Nwokah, H.-C. Hsu, and A. Fogel, "The integration of laughter and speech in vocal communication : A dynamic systems perspective," *J. Speech Lang Hear Res*, no. 42, 1999.

[19] J. A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *The Acoustical Society of America*, vol. 110, no. 3, p. 1581–1597, 2001.

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

[21] T. Matsuda and Y. Arimoto, "Acoustic discriminability of unconscious laughter and scream during game-play," in *Proc. Speech Prosody 2022*, 2022, p. 575–579.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML 2006*, 2006, p. 369–376.

[23] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949.

[24] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[25] 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.

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

[27] J. Trouvain, "Segmenting phonetic units in laughter," *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003.

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

[29] Y. Arimoto, R. Imanishi, and H. Mori, "Laughter components estimation using emotional information towards natural and expressive laughter synthesis," *Information Processing Society of Japan*, vol. 63, no. 4, pp. 1159–1169, 2022, (in Japanese).

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

[31] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. Interspeech 2013*, 2013, pp. 148–152.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, p. 1459–1462.

[33] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario," in *in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7554–7558.

[34] K. Lounnas, M. Lichouri, and M. Abbas, "Analysis of the effect of audio data augmentation techniques on phone digit recognition for algerian arabic dialect," in *2022 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, 2022, pp. 1–5.

[35] G. Gosztolya, "On evaluation metrics for social signal detection," in *Proc. Interspeech 2015*, 2015, pp. 2504–2508.

[36] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2017, p. 4835–4839.