Supervised Contrastive Learning for Game-Play Frustration Detection from Speech

Meishu Song[1(✉)], Emilia Parada-Cabaleiro[2], Shuo Liu[1], Manuel Milling[1], Alice Baird[1], Zijiang Yang[1], and Björn W. Schuller[1,3]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
meishu.song@informatik.uni-augsburg.de
[2] Institute of Computational Perception, Johannes Kepler University Linz, Linz, Austria
[3] GLAM – Group on Language, Audio, & Music, Imperial College London, London, UK

Abstract. Frustration is a common response during game interactions, typically decreasing a user's engagement and leading to game failure. Artificially intelligent methods capable to automatically detect a user's level of frustration at an early stage are hence of great interest for game designers, since this would enable optimisation of a player's experience in real-time. Nevertheless, research in this context is still in its infancy, mainly relying on the use of pre-trained models and fine-tuning tailored to a specific dataset. Furthermore, this lack in research is due to the limited data available and to the ambiguous labelling of frustration, which leads to outcomes which are not generalisable in the real-world. Meanwhile, contrastive loss has been considered instead of the traditional cross-entropy loss in a variety of machine learning applications, showing to be more robust for system stability alternative in self-supervised learning. Following this trend, we hypothesise that using a supervised contrastive loss might overcome the limitations of the cross-entropy loss yielded by the labels' ambiguity. In fact, our experiments demonstrate that using the supervised contrastive method as a loss function, results improve for the automatic recognition (binary frustration vs no-frustration) of game-induced frustration from speech with an Unweighted Average Recall increase from 86.4 % to 89.9 %.

Keywords: Frustration recognition · Supervised contrastive learning · Speech recognition

1 Introduction

Automatically detecting frustration from users' audio-visual cues is becoming a more important part in ubiquitous sensing technology, used in a variety of applications, such as in-car-driver monitoring [44], or e-learning scenarios [13]. Although game-based applications—both entertainment and education

oriented—would benefit from such a technology to achieve user evaluation and study feedback in real-time, the automatic analysis of frustration during game play is still an underdeveloped area of research. On the one side, the lack in research is due to the still limited amount of suitable data [37]. On the other side, due to the labels' ambiguity typical of emotional data, for which instead of an objective 'ground truth', a subjective 'gold standard' [33], i.e., an agreed-upon human annotation label, is typically available. Although methods for automatic recognition of frustration from audio-visual cues have been presented, most of these are based on end-to-end models applying cross-entropy loss [9], which despite their excellent convergence speed, present a sub-optimal performance when dealing with ambiguous labels [45]— note that frustration is usually annotated according to a gold standard. Furthermore, frustration can also be confused with other highly aroused emotional states, e. g., anger, or, irritation.

In recent years, the use of contrastive learning, able to compensate for the disadvantage of cross-entropy loss, i.e., the sub-optimal performance with ambiguous labels, has led to major advances in self-supervised learning, especially showing its potential on traditional visual classification tasks [5]. The main idea behind contrastive learning is to pull together an anchor, i.e., 'positive' sample of the class intended to be recognised, into an embedding space; then, push apart the 'negative' samples, by measuring their similarity w.r.t. the anchor [21]. Since in self-supervised tasks there is no label information, in order to adapt contrastive learning to a fully supervised setting, the Supervised Contrastive (SupCon) loss has been presented, which provides label information to the system by enabling multiple positives per anchor [23].

Inspired by this idea and by the successful performance shown by the use of contrastive learning and SupCon loss in previous works [23], we aim to overcome the shortcomings of using cross-entropy loss by applying supervised contrastive learning with Residual Convolutional Neural Networks (ResNet) to the detection of frustration from speech. By using this approach on the Multimodal Game Frustration Database (MGFD) [37], we aim to mitigate the influence of ambiguous labels—typical of emotional induced datasets [31]—in the frustration recognition task. The rest of the manuscript is laid out as follows: In Sect. 2, the related literature on frustration recognition and contrastive learning is described; in Sect. 3, the methodology is presented; in Sect. 4, the considered deep learning approaches are evaluated; in Sect. 5, the experimental results are discussed; finally, in Sect. 6, the conclusions and future work are given.

2 Related Work

2.1 Frustration Recognition

Frustration is a negative emotional state typically triggered by someone's inability to achieve the own goals [37]. In the realm of user experience research, traditional methods to measure frustration include the *Focus Group User Study* [28], the *User Questionnaire* [2], the *Expert Evaluation* [2], or the *Diary-Based*

Study [19]. Nowadays, with the advent of artificial intelligence, affective computing has opened new horizons in the automatic detection and measurement of frustration through machine learning [37]. In this regard, a variety of datasets aimed to develop frustration recognition systems has been presented in the literature. These include corpora containing spontaneous frustration in the context of driving, such as the UTDrive database [16], amongst others [26]. Similarly, educational scenarios have also been considered to record frustration, as shown by: the computer-mediated human tutoring corpus, which contains facial expression of frustration [11]; the ChIMP-Children's Interactive Multimedia Project database [3], collected during children-computer interactions, and containing verbal expressions of frustration; or the Microsoft Kinect sensors posture dataset, collected in a game-based learning environment for emergency medical training [18]. Another example is the AlloSat corpus [25], a speech-based dataset composed of real-life call centre conversations in French language recorded by Allo-Media. Finally, studies on frustration recognition from speech data have also been presented [10]. As every emotional reaction, frustration presents a variety of symptoms, which can be measured from different modalities, including physiological signals and audio-visual cues [34]. Furthermore, since frustration might arise in different situations, its recognition through such modalities has been applied in many contexts, mostly influenced by the existing corpora. For instance, driving in daily traffic scenarios has shown to be a prominent source of frustration, reason why this context has been considered to detect drivers' frustration from different modalities, such as finger temperature [44], or bimodal cues, i.e., heart rate and visual facial features [43]. Frustration is also a typical emotion that impairs a successful learning process [39].

Due to the importance of monitoring [36] and understanding students' emotions [30], systems such as ULearn, aimed to detect a student's level of frustration through computer vision techniques and natural language processing [13], as well as sensor-based methods which assess frustration through gesture and posture detection [18], have been presented. Indeed, the systems aimed to predict students' frustration and learning outcomes [12], by this enhancing student engagement [40], are always increasing. Another context very close to the educational, is the design of serious games, where motivational feedback is applied to counterbalance users' frustration, a method that can improve students' outcomes [7]. Nevertheless, despite the promising outcomes of applying automatic detection of frustration in serious games, the automatic recognition in game-interaction remains still underdeveloped [37].

2.2 Contrastive Learning

The origins of contrastive learning dates to the 90ies, and its development has spanned across many fields [24], such as, computer vision and natural language processing [22]. The core idea of this approach is learning by comparing between separate, although related, data points, without considering any supervised information like labels [1]. In 2005, Chopra et al. proposed a kernel able to map training data into a target area; thus, creating the foundation of the contrastive

learning framework [6] [15]. The training stage minimises a discriminative loss function that drives the similarity metric for the samples of the same class [6]; by this, the contrastive pair loss learns a good representation of the data. This idea was tested on the Purdue face database, which includes a very high degree of variability [6]. Further improvements were subsequently presented, for instance, Oh Song et al. improved the efficiency of comparisons in an iteration by lifting the vector of pair-wise distances within the batch [29].

In 2010, Gutmann and Hyvärinen [14] introduced the Noise Contrastive Estimation (NCE), i.e., a simple conceptual strategy for estimating an unnormalised statistical model by contrasting the data w.r.t. an auxiliary noise. Nevertheless, how to select the auxiliary noise distribution remains still an open research question. To this end, a variety of solutions has been presented in the literature. Ceylan and Gutmann [4] proposed to formulate density estimation as a supervised learning problem that unlike NCE, leverages the observed data when generating noisy samples. Also based on NCE, Mnih and Teh [27] trained powerful natural language processing models to learn word embeddings on the Microsoft Research Sentence Completion Challenge dataset [46]. In 2018, Hjelm et al. [20] investigated data representations by maximising mutual information between an input and the output as well as minimising a contrastive loss.

Instead of learning from individual data samples one at a time, contrastive learning is based on the comparison amongst data pairs, i.e., it learns a representation by maximising the distance between samples organised into similar and dissimilar pairs [42]. As in other self-supervised learning tasks, such a similarity can be defined from the data itself, thus overcoming the limitation typical of supervised learning settings, where only a finite number of label pairs are available from the data. Furthermore, while some self-supervised methods need to modify the model architecture during learning, one of the main advantages of contrastive methods is that no modification to the model architecture is needed between training and fine-tuning [24]. Indeed, contrastive learning has recently achieved state of the art performance in the field of self-supervised representation learning [38].

3 Methodology

The experiments on frustration recognition were carried out on the Multimodal Game Frustration Database (MGFD) [37], an audio-visual dataset collected within the Wizard-of-Oz framework, aimed to investigate users' audio-visual expressions of frustration during game interaction on the CrazyTrophy game (cf. Fig. 1). MGFD contains 5 h of recordings from 67 healthy individuals (27 female, 40 male, with a mean age of 15 years old) experiencing different levels of spontaneous frustration elicited by a variety of (intentional) 'inconsistency'-based usability problems. Although MGFD is suitable to assess users' frustration from both audiovisual modalities [37], i.e., audio and video, it has been shown that the recognition of the emotion of frustration from speech shows a higher performance than from facial expressions [37]. This is mostly due to the fact

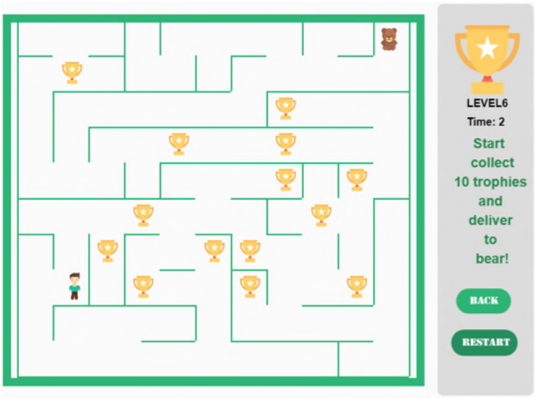
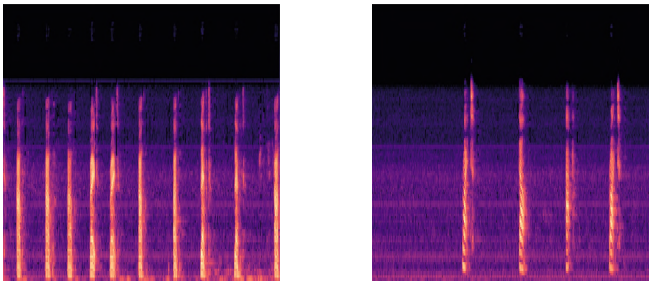


Fig. 1. Interface of the CrazyTrophy game on the sixth level. Through spoken direction words (left, right, up and down), the user controls the avatar movements in order to achieve the game's goal, i.e., to deliver the trophies to the bear (top-right of the interface). Due to a purposefully designed in-consistency usability problem, the user will receive 2 points for each collected trophy, which makes the game impossible to be won, by this intentionally eliciting frustration in the player.



(a) No frustration (b) Frustration

Fig. 2. Mel-spectrograms extracted from utterances of non-frustration and frustration (left and right, respectively). For non-frustration, the male user produces the words 'up, up, up, right, right, up, up, left, left'; for frustration, 'right, up, down, right'. Comparing the two Mel-spectrograms, we can observe that the user speaks considerably faster when not frustrated. The slower speaking during frustration might be due to the increased cognitive load: the player attempts to both win the game and to understand the usability problem [37].

that the participants of MGFD are Chinese, a culture in which it is particularly encouraged to conceal the own emotions [41]. In this regard, since facial expressions of emotion are generally accepted across cultures [8], we assume these are easier to control than speech; thus, frustration might be more difficult to hide in vocal than in visual cues. Due to this, in the present study, we will take only

into account the audio signals for the prediction of users' frustration, since it is considered a more reliable modality in the MGFD corpus.

Table 1. Implemented distribution of speakers: m(ale), f(emale); and number (#) of instances: frustration, non frustration; for each set: Train(ing), Dev(elopment), and Test. Sums across sets (Σ) are also given.

#	Train	Dev	Test	Σ
Speakers	43	12	12	67
Gender (m:f)	28:15	6:6	6:6	40:27
Frustration	456	118	118	692
Non Frustration	3798	978	978	5754

Since in supervised learning, a model's performance might be influenced by the use of specific features, Mel-spectrograms (cf. Fig. 2), widely utilised in speech emotion recognition [35], were considered appropriate for the purposes of the present study. As a standard procedure in the field, the experiments were carried out in a speaker-independent manner, considering 43 speakers for training, 12 speakers for development (i.e., validation), and 12 speakers for test. In Table 1, descriptive statistics on the considered partitioning are given.

4 Deep Learning Approach

In this study, we utilise Residual Networks (ResNet) with two loss settings: Binary Cross-Entropy Loss (BCELoss) with Logits and Supervised Contrastive (SupCon) Loss. The machine learning models are implemented through the deep-learning framework `PyTorch` [32]. In the following, the proposed architectures for frustration recognition from speech are presented.

4.1 Binary Cross-Entropy with Logits Loss

Since we perform a binary classification task: *frustration* vs *non-frustration*, we take into account the common implementation of the BCELoss with Logits as a baseline in our experiments. The BCELoss with Logits for one sample is defined as

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{2} y_i log(\hat{y}_i) = -y_1 log(\hat{y}_1) - (1 - y_1) log(1 - \hat{y}_1), \tag{1}$$

where the network assigns the probability $\hat{y}_i$ that the given sample belongs to class i. The target probability y_i of the sample belonging to class i is given by the label. Note that the right side of (1) results from the fact that only two classes are considered in binary cross-entropy, and that probabilities are normalised,

which is ensured for the network's prediction by a softmax layer. As we consider a problem for which the target probabilities are either 0 or 1, only one term on the right side of (1) contributes to the loss. Nevertheless, changes to one of the two predictions $\hat{y}_i$ affects the other one via normalisation.

4.2 Supervised Contrastive Loss

The performance of a deep learning system is directly influenced by the choice and quality of the data representation [24]. For instance, labelled datasets, especially in affective computing, are often too small, something that might impair a learning system's performance. In such a scenario, focusing explicitly on learning representation, i.e., the process of learning a parametric mapping from the raw input data to a feature vector or tensor, might be beneficial [24]. Since contrastive loss can achieve a proper data representation for distinguishing different classes, it can be applied in this context. In constrastive learning, augmented samples are generated from samples of the anchor's class; then, the network extracts a strong inductive bias from both, the anchor and the augmented samples, by this attracting positive samples, i.e., those similar to the anchor, while repealing negative ones, i.e., those dissimilar. Unlike self-supervised contrastive learning [23], supervised contrastive learning draws positive samples not only from the augmentation of the sample, but also from the augmentation of other samples belonging to the same class as the anchor.

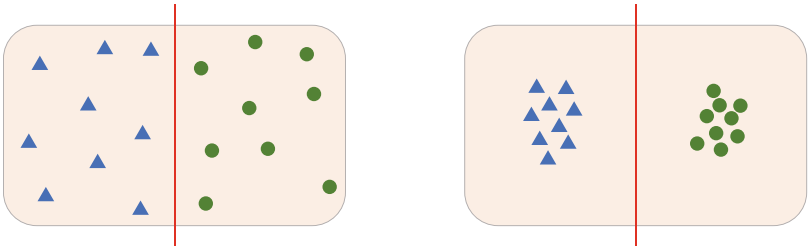
For each of N labelled samples in a given minibatch, we generate two augmentations, leading to $2N$ augmented samples in training. Our network computes the representations $\{z_l\}$ of the augmented samples in a 128-dimensional projection space. In this space, the supervised contrastive loss is defined as

$$\mathcal{L}_{\mathrm{SC}} = \sum_{i=1}^{2N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum\limits_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{2}$$

with the scalar product '$\cdot$' and the temperature hyperparameter τ. For a given augmented sample i, considered the anchor of the loss, the set $A(i)$ contains all $2N$ indices except the ones for i and $P(i)$, i.e., the set of all positive samples. In other words, $P(i)$ contains the indices of $A(i)$ with the same label as i; where $|P(i)|$ is the cardinality of $P(i)$.

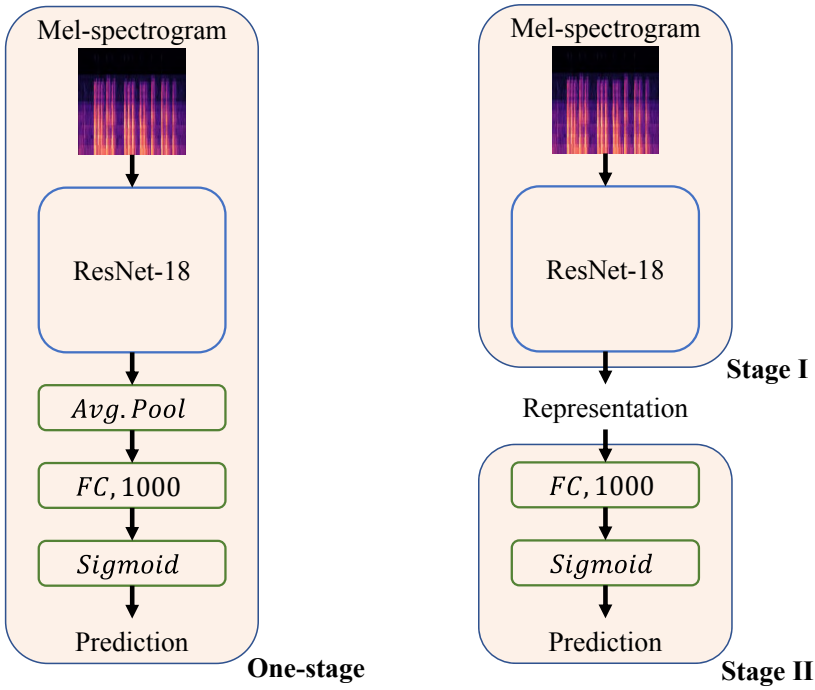
In order to train a classifier with the help of SupCon Loss, two subsequent steps are performed. First, the model learns the representations of the data, which are well separable utilising the SupCon Loss. In a second step, a classifier based on the BCELoss is trained on the learnt representations. Note that, during the training of the classifier backbone of the model is frozen, i. e., learning is disabled for any layer being involved in the calculation of the representations.

In Fig. 3, a visual representation of the procedure followed by the BCELoss and SupCon Loss for data representation is displayed. For the BCELoss training and optimisation, i.e., looking for a good data representation and for the optimal decision boundary, are performed simultaneously. Differently, when considering



(a) BCELoss with Logits (b) SupCon Loss

Fig. 3. Representation of the Binary Cross Entropy Loss (BCELoss) with Logits and the supervised Contrastive (SupCon) Loss, associated with a linear classifier.



(a) ResNet-18 (BCELoss with Logits) (b) ResNet-18 (SupCon Loss)

Fig. 4. ResNet-18 architectures: (a) with Binary Cross Entropy with Logits Loss (BCELoss); (b) with Supervised Contrastive Loss (SupCon Loss). For both architectures, in the last fully connected layer, a Sigmoid activation function is considered. For the model with SupCon Loss, we apply the ResNet model to extract a data representation; then, this is considered as input of a Multilayer Perceptron.

the supervised contrastive loss, finding a good representation and classification are two separate procedures, as at this moment, there is no sophisticated model for explicit training of a classifier using the supervised contrastive loss. It has

previously been shown that this two step-training can facilitate the task of the linear classifier and therefore improve its performance [24].

4.3 ResNet Architectures

To assess the efficiency of the two evalauted losses, i.e., the BCELoss with logits and the SupCon Loss, we consider a ResNet architecture in our experiments. This was chosen since it has been shown that shallow networks with a Residual Block usually perform better than traditional convolutional neural networks [17]. Furthermore, this architecture does not need additional parameters, such as shortcuts, by this reducing computational complexity. The experiments were carried out on three types of RestNet: ResNet-18, ResNet-34, and ResNet-50; i.e., three convolutional networks with 18, 34, and 50 layers depth, respectively. The Residual Block enables the stacked layers to fit a residual mapping, i.e., the layers in a traditional network are learning the true output whereas the layers in a residual network are learning the residual. Specifically in our study, there are two steps utilising SupCon Loss. Firstly, we apply a ResNet and SupCon Loss to otain a data representation. Afterwards, we send these representation data into a classifier. In this study, we use a Multi-Layer Perceptrum (MLP) as classifier. The normalised activations of the final pooling layer are used as the representation vector. In Fig. 4, the ResNet-18 architecture with both BCELoss with logits and SupCon Loss is shown.

For the experiments with the SupCon Loss, the initial learning rate is 0.0001, the batch size is 64, training epochs are 1 000, and the SGD optimiser with momentum equalling 0.9 is considered. After getting all the data representations, these are considered as input for the MLP classifier with one hidden layer, for which, the BCELoss with logits is chosen as activation function. For the experiments with the BCELoss, we input the data directly into the different ResNet models. For comparability, the hyperparameters are setup as for the experiments with the SupCon Loss (described above).

5 Results

The models' performance is measured using Unweighted Average Recall (UAR), an evaluation metric suitable for class-imbalanced classification tasks [37]. In Table 2, the experimental results are given. For both ResNet-18 and ResNet-34, the experiments considering SupCon Loss outperform the ones given by BCELoss with logits: 79.3 % and 81.2 % vs 75.1 % and 78.5 % for Dev and Test in ResNet-18; 88.7 % and 89.9 % vs 84.2 % and 86.4 % for Dev and Test in ResNet-34; cf. SupCon Loss vs BCELoss, respectively, in Table 2.

Differently, for ResNet-50, the BCELoss with logits performs slightly better than the SupCon Loss: 84.1 % and 83.6 % vs 85.0 % and 84.8 % for Dev and Test; cf. ResNet-50 for SupCon Loss vs BCELoss, respectively, in Table 2. Yet, these differences are very small. The best UAR was achieved by ResNet-34 with SupCon Loss on the Test set (cf. 89.9 % in Table 2).

Table 2. Evaluation Results [%] of the ResNets models: ResNets-18, ResNets-34, and ResNets-50; with Supervised Contrastive Loss (SupCon loss) and Binary Cross Entropy Loss (BCELoss) with Logits. Results are presented both for the Dev(elopment) and Test sets. Unweighted Average Recall (UAR) is used as evaluation metric; the best results are highlighted in bold.

UAR [%]	Achitecture	Dev	Test
SupCon Loss	ResNets-18	79.3	81.2
	ResNets-34	88.7	**89.9**
	ResNets-50	84.1	83.6
BCELoss	ResNets-18	75.1	78.5
	ResNets-34	84.2	**86.4**
	ResNets-50	85.0	84.8

We interpret that the superiority of using the SupCon Loss is due to the fact that—through data augmentation—this architecture enables 2 views per sample; thus, its batch size is effectively double w.r.t. the BCELoss-based architecture. Furthermore, the model with BCELoss focuses on maximising the probability of recognising the correct class, but it does not take into account the samples distance to each other. Differently, the model with SupCon Loss aims to create a good representation where the samples of each class are close to each other and distant to the opposite class. Nevertheless, training an architecture with SupCon Loss is considerably more time-consuming than using BCELoss, a downside that should be taken into account.

6 Conclusion and Future Work

Our study confirmed that supervised contrastive learning is an appropriate method to recognise frustration from speech, showing to be robust in handling the inaccurate labels typical of emotional datasets. To the best of our knowledge, this is the first time that supervised contrastive learning has been applied in the detection of frustration from speech. Our experimental results demonstrate that the proposed method considerably outperforms the cross-entropy loss-based method, as shown by the comparison on the same model configuration. The promising outcomes show that supervised contrastive learning is a robust method to automatically detect users' frustration from speech, which suggest that the use of SupCon loss might also be a turning point in audio-visual applications for emotion recognition. In future work, one should further investigate how different percentages of noisy labels might alter the performance of a supervised contrastive learning framework in comparison to other loss functions.

Acknowlwdgement. The authors acknowledge funding from the German Research Foundation (DFG) under the Reinhart Koselleck Project grant AUDI0NOMOUS (No. 442218748). The responsibility lies with the authors.

References

1. Becker, S., Hinton, G.E.: Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature **355**(6356), 161–163 (1992)
2. Bevan, N.: What is the difference between the purpose of usability and user experience evaluation methods. In: Proceedings of the Workshop UXEM, pp. 1–4. Uppsala, Sweden (2009)
3. Byrd, D., McLaughlin, M., Khurana, S., Landes, M., Ucar, T.: Chimp: Children interacting with machines project
4. Ceylan, C., Gutmann, M.U.: Conditional noise-contrastive estimation of unnormalised models. In: Proceedings of the International Conference on Machine Learning, pp. 726–734. Vienna, Austria (2018)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning, pp. 1597–1607. Virtual (2020)
6. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 539–546. San Diego, USA (2005)
7. DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C.: Detecting and addressing frustration in a serious game for military training. Int. J. Artif. Intell. Educ. **28**(2), 152–193 (2018)
8. Ekman, P., Keltner, D.: Universal facial expressions of emotion. In: Segerstrale U, P. Molnar, P., (eds.) Nonverbal communication: Where nature meets culture vol. 54 no. 2, pp. 27–46 (1997)
9. Franz, O., Drewitz, U., Ihme, K.: Facing driver frustration: towards real-time in-vehicle frustration estimation based on video streams of the face. In: Proceedings of the International Conference on Human-Computer Interaction, pp. 349–356. Virtual (2020)
10. Goetsu, S., Sakai, T.: Different types of voice user interface failures may cause different degrees of frustration. arXiv preprint arXiv:2002.03582 (2020)
11. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically recognizing facial indicators of frustration: a learning-centric analysis. In: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 159–165. Geneva, Switzerland (2013)
12. Grafsgaard, J.F., Wiggins, J.B., Vail, A.K., Boyer, K.E., Wiebe, E.N., Lester, J.C.: The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In: Proc. International Conference on Multimodal Interaction, pp. 42–49. Istanbul, Turkey (2014)
13. Grewe, L., Hu, C.: Ulearn: understanding and reacting to student frustration using deep learning, mobile vision and nlp. In: Proceedings of the Signal Processing, Sensor/Information Fusion, and Target Recognition XXVIII. p. 110. Maryland, USA (2019)
14. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 297–304. Sardinia, Italy (2010)
15. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1735–1742. New York, USA (2006)

16. Hansen, J.H., Busso, C., Zheng, Y., Sathyanarayana, A.: Driver modeling for detection and assessment of driver distraction: examples from the utdrive test bed. IEEE Signal Process. Mag. **34**(4), 130–142 (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the Computer Vision and Pattern Recognition, pp. 770–778. Las Vegas, USA (2016)
18. Henderson, N.L., Rowe, J.P., Mott, B.W., Brawner, K., Baker, R., Lester, J.C.: 4d affect detection: improving frustration detection in game-based learning with posture-based temporal data fusion. In: Proceedings of the Artificial Intelligence in Education, pp. 144–156. Beijing, China (2019)
19. Hertzum, M.: Frustration: a common user experience. DHRS2010 p. 11 (2010)
20. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
21. Inoue, N., Goto, K.: Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition. In: Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1641–1646. Virtual (2020)
22. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2021)
23. Khosla, P., et al.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
24. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: a framework and review. IEEE Access (2020)
25. Macary, M., Tahon, M., Estève, Y., Rousseau, A.: Allosat: a new call center French corpus for satisfaction and frustration analysis. In: Proceedings of the Language Resources and Evaluation Conference, pp. 1590–1597. Virtual (2020)
26. Malta, L., Miyajima, C., Kitaoka, N., Takeda, K.: Analysis of real-world driver's frustration. IEEE Trans. Intell. Transp. Syst. **12**(1), 109–118 (2010)
27. Mnih, A., Teh, Y.W.: A fast and simple algorithm for training neural probabilistic language models. arXiv preprint arXiv:1206.6426 (2012)
28. Oehl, M., Ihme, K., Drewitz, U., Pape, A.A., Cornelsen, S., Schramm, M.: Towards a frustration-aware assistant for increased in-vehicle UX: F-RELACS. In: Proceedings of the Automotive User Interfaces and Interactive Vehicular Applications, pp. 260–264. Utrecht, Netherlands (2019)
29. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4004–4012 (2016)
30. Parada-Cabaleiro, E., Batliner, A., Baird, A., Schuller, B.: The perception of emotional cues by children in artificial background noise. Int. J. Speech Technol. **23**(1), 169–182 (2020). https://doi.org/10.1007/s10772-020-09675-1
31. Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M., Schuller, B.W.: DEMoS: an Italian emotional speech corpus: elicitation methods, machine learning, and perception. Lang. Resour. Eval. **54**, 341–383 (2020)
32. Paszke, A., et al.: Automatic differentiation in PyTorch. In: NIPS-W (2017)
33. Schuller, B., Batliner, A.: Computational paralinguistics: emotion, affect and personality in speech and language processing. Wiley, Sussex, UK (2014)
34. Shoumy, N.J., Ang, L.M., Seng, K.P., Rahaman, D.M., Zia, T.: Multimodal big data affective analytics: a comprehensive survey using text, audio, visual and physiological signals. J. Netw. Comput. Appl. **149**, 102447 (2020)

35. Song, M., et al.: Frustration recognition from speech during game interaction using wide residual networks. Virtual Reality & Intelligent Hardware **10**, (2020)
36. Song, M., et al.: Predicting group work performance from physical handwriting features in a smart English classroom. In: Proceedings of the International Conference on Digital Signal Processing (ICDSP). Chengdu, China (2021)
37. Song, M., et al.: Audiovisual analysis for recognising frustration during gameplay: introducing the multimodal game frustration database. In: Proceedings of the Affective Computing and Intelligent Interaction, pp. 517–523. Cambridge, the UK (2019)
38. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243 (2020)
39. Tyng, C.M., Amin, H.U., Saad, M.N., Malik, A.S.: The influences of emotion on learning and memory. Front. Psychol. **8**, 1454 (2017)
40. Valdez, M.G., Hernández-Águila, A., Guervós, J.J.M., Soto, A.M.: Enhancing student engagement via reduction of frustration with programming assignments using machine learning. In: Proceedings of the International Joint Conference on Computational Intelligence, pp. 297–304. Funchal, Portugal (2017)
41. Wei, M., Su, J.C., Carrera, S., Lin, S.P., Yi, F.: Suppression and interpersonal harmony: a cross-cultural comparison between chinese and european americans. J. Couns. Psychol. **60**(4), 625 (2013)
42. Xiao, T., Wang, X., Efros, A.A., Darrell, T.: What should not be contrastive in contrastive learning. arXiv preprint arXiv:2008.05659 (2020)
43. Zepf, S., Stracke, T., Schmitt, A., van de Camp, F., Beyerer, J.: Towards real-time detection and mitigation of driver frustration using SVM. In: Proceedings of the Machine Learning and Applications, pp. 202–209. Florida, USA (2019)
44. Zhang, M., Ihme, K., Drewitz, U.: Discriminating drivers' fear and frustration through the dimension of power. In: Proceedings of the Humanist Conference, p. 98. Hague, Netherlands (2018)
45. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:1805.07836 (2018)
46. Zweig, G., Burges, C.J.: The microsoft research sentence completion challenge. Microsoft Research, Redmond, WA, USA, Technical report. MSR-TR-2011-129 (2011)