·Article·

Frustration recognition from speech during game interaction using wide residual networks

Meishu SONG[1*], Adria MALLOL-RAGOLTA[1], Emilia PARADA-CABALEIRO[1], Zijiang YANG[1], Shuo LIU[1], Zhao REN[1], Ziping ZHAO[3], Björn W. SCHULLER[1,2]

1. *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*

2. *GLAM-Group on Language, Audio, & Music, Imperial College London, UK*

3. *Department of Computer Science, Tianjin Normal University, Tianjin* 300389*, China*

*** Corresponding author,** meishu.song@informatik.uni-augsburg.de

Received: 28 August 2020 **Accepted:** 18 October 2020

Supported by the European Union's Horizon 2020 Programmes Under Grant Agreement (826506, sustAGE).

Citation: Meishu SONG, Adria MALLOL-RAGOLTA, Emilia PARADA-CABALEIRO, Zijiang YANG, Shuo LIU, Zhao REN, Ziping ZHAO, Björn W. SCHULLER. Frustration recognition from speech during game interaction using wide residual networks. Virtual Reality & Intelligent Hardware, 2021, 3(1): 76—86
DOI: 10.1016/j.vrih.2020.10.004

Abstract **Background** Although frustration is a common emotional reaction while playing games, an excessive level of frustration can negatively impact a user's experience, discouraging them from further game interactions. The automatic detection of frustration can enable the development of adaptive systems that can adapt a game to a user's specific needs through real-time difficulty adjustment, thereby optimizing the player's experience and guaranteeing game success. To this end, we present a speech-based approach for the automatic detection of frustration during game interactions, a specific task that remains under-explored in research. **Method** The experiments were performed on the Multimodal Game Frustration Database (MGFD), an audiovisual dataset—collected within the Wizard-of-Oz framework—that is specially tailored to investigate verbal and facial expressions of frustration during game interactions. We explored the performance of a variety of acoustic feature sets, including Mel-Spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and the low-dimensional knowledge-based acoustic feature set eGeMAPS. Because of the continual improvements in speech recognition tasks achieved by the use of convolutional neural networks (CNNs), unlike the MGFD baseline, which is based on the Long Short-Term Memory (LSTM) architecture and Support Vector Machine (SVM) classifier—in the present work, we consider typical CNNs, including ResNet, VGG, and AlexNet. Furthermore, given the unresolved debate on the suitability of shallow and deep networks, we also examine the performance of two of the latest deep CNNs: WideResNet and EfficientNet. **Results** Our best result, achieved with WideResNet and Mel-Spectrogram features, increases the system performance from 58.8% unweighted average recall (UAR) to 93.1% UAR for speech-based automatic frustration recognition.

Keywords Frustration recognition; WideResNets; Machine learning

1 Introduction

Emotion-related theories define frustration as "the occurrence of an obstacle that prevents the satisfaction of a need"[1]. Frustration is an emotional state that has been studied since the early years of the 20th century[1] that is of special interest in the field of human behavior analysis. Because frustration is a common reaction triggered while playing games[2], it has also been studied within the human-game interaction paradigm. In this context, frustration is a negative emotional state that occurs in goal-oriented games when a feeling of dissatisfaction arises from a player's unfulfilled needs[3]. Experiencing frustration while interacting with a game can also trigger a variety of negative emotional responses, such as acute stress, sadness, or rage. These emotions impact a user's opinion of the playing experience, which usually negatively influences their evaluation of the game and therefore reduces their acceptance of it[4]. In this regard, recent developments in computer game analysis have served as the main avenue to improve user experience (UX)[5] through the elicitation, detection, and modelling of players' emotions while gaming. Indeed, frustration has been identified as a common emotional reaction during game interactions[6] that negatively impacts the UX[5]. Hence, the development of technology capable of effectively recognizing and reducing users' frustration while playing games is expected to be highly beneficial for improving the UX.

Although affective computing[7] technologies have already been applied in the field of gaming research[5], the automatic analysis of frustration during game play is still an underdeveloped area of investigation. Indeed, due to the difficulties linked to the collection of suitable and realistic databases[8], existing datasets that are adequate for the study of frustration during game interactions are rare. One exception is the Multimodal Game Frustration Database (MGFD)[9], which includes spontaneous interactions by players playing "Crazy Trophy," a voice-controlled game in which frustration was elicited by creating usability problems. The users' spoken and facial expressions were both recorded, and the interactions were labeled in terms of the presence or absence of frustration. An initial work on the MGFD presented a baseline using a Support Vector Machine (SVM) and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) and applying Mel-Frequency Cepstral Coefficient (MFCC) acoustic features for the automatic detection of frustration/non-frustration states from users' facial expressions and speech[9].

Beyond SVMs and LSTM-RNNs, which have been successfully employed in speech analysis tasks[10], other neural network techniques such as Convolutional Neural Networks (CNNs) have also shown promising results for such tasks[11]. Specifically, the research in [12] obtained superior performance with CNNs as compared to RNNs. Previous work has provided some evidence that CNNs are more suitable than LSTM-RNNs for the classification of long sentences in a natural language processing task[13]. Similarly, attention-based CNNs have also achieved better results than attention-based LSTM-RNNs for answer selection using an open domain question-answer selection dataset[14]. These results may be due to the fact that while RNNs[15] compute cyclic connections for the input features, CNNs use the invariance of deep convolutions to overcome the diversity of the speech signals and multi-layer enriched features[16]. Considering this, we anticipate that CNNs will perform well for the recognition of users' frustration on the MFGD. Research on CNNs has led to a series of breakthroughs in machine learning tasks, prompting the exploration of 'very deep' models, which have become the current state-of-the-art research[17]. Nevertheless, a 'degradation' problem may arise when deep networks start to converge, diminishing the accuracy[17]. For instance, residual networks (ResNets[18]) have been shown to efficiently scale up to hundreds of layers while maintaining improved performance[19]. However, each fraction of a percentage in improved accuracy usually requires the number of layers to be nearly doubled. Likewise, training very deep residual networks tends to lead to the problem of diminishing feature reuse. Further, the information shared among all blocks is limited and does not provide a sufficient contribution[19]. To address this problem, WideResNet, which randomly disables the residual blocks during training, was designed. Wide residual blocks can outperform

their thin and deep counterparts[19].

In this work, our goal is to overcome the drawbacks of a limited feature set and of the machine learning models considered in the MGFD baseline. We thus extract a larger variety of audio feature sets than those considered in the baseline paper and apply state-of-the-art deep CNNs for the automatic identification of user frustration during game interactions. We provide a comparison of a variety of CNNs, including AlexNet, VGG, EfficientNet, ResNets, and WideResNets, and consider different feature sets: MFCCs, Mel-Spectrograms, and eGeMAPS. Moreover, we determine the most efficient configuration for frustration recognition on the MGFD.

2 Related work

Although emotional databases with a focus on frustration are rare[9], the interest in modeling this emotion computationally is evident in the variety of databases introduced in the literature that contain frustration to some extent. With regard to audio content, the UTDrive database contains recordings collected in real conditions from people driving in urban areas[20,21]. In the ChIMP-Children's Interactive Multimedia Project-database, which was collected during children-computer interactions, verbal expressions of frustration were evaluated[22]. Datasets annotated in terms of frustration are relatively common in the literature, especially for datasets containing other emotional states. For instance, DEAP, the Dataset for Emotion Analysis using Physiological Signals[23], which was recorded while subjects were watching musical videos, provides electroencephalogram (EEG) and peripheral physiological signals. Similarly, FEEDB, the Facial Expressions and Emotions Database[24], contains synchronized facial colour videos and depth maps, both of which contain annotations of frustration and other emotions. Finally, although the validity of acted emotions has been criticized[25], audiovisual datasets containing acted expressions of frustration have also been presented; examples include IEMOCAP, the Interactive EMOtional dyadic motion CAPture database[26] and the Chen-Huang database[27]. Overall, despite the interest in frustration implied by existing emotional databases, the lack of a corpus focusing on this emotion has limited the development of systems capable of automatically recognising user frustration.

Recent advances in neural network-based research have shown that properly widening the residual blocks, instead of increasing their depth, can lead to much more effective residual networks with improved performance[28]. Recent research on image-based food recognition[29] has shown that wide residual blocks with a sliced convolution can successfully capture specific information, thereby yielding better performance than existing approaches. On the ILSVRC12 image classification task, that is, the ImageNet Large Scale Visual Recognition Challenge, wide residual networks with smaller depths showed comparable accuracy to that achieved by narrower and deeper ResNets[30]. Such 'WideResNets' have also been used recently in the medical domain for lung cancer classification[31], achieving state-of-the-art accuracy results in predicting the majority of referral and non-referral nodules. Similarly, in another medical application, WideResNets have been considered for mitosis detection in breast histology images[32], an approach that was ranked 2nd in the MICCAI TUPAC 2016 competition for mitosis detection. Aditi presented the use of 3D WideResNets for disease diagnosis, specifically to obtain better quality denoised brain magnetic resonance images than the state-of-the-art approach obtained[33].

3 Experimental set-up

In the following, the MGFD database is described, details of the considered feature sets are provided, and the data partition and equipment setup are indicated.

3.1 MGFD

Level 6 interface of "Crazy Trophy": although the user has collected only 10 trophies, due to the

(intentional) usability problem, the panel on the right indicates a doubled count, i.e., 20 trophies; this makes it impossible to win the game.

The MGFD[1] is a database that contains 5 hours of audiovisual recordings from 67 healthy individuals (27 female and 40 male, with a mean age of 15 years) experiencing different levels of spontaneous frustration elicited by a variety of (intentional) usability issues. The MGFD was collected through from user interactions with the game "Crazy Trophy," a Wizard-of-Oz (WoZ)[34] voice-controlled game especially designed to induce frustration in the participants. During their interactions with "CrazyTrophy," users perceive that the game avatar is controlled by their voice and they are able to use the spoken commands 'left,' 'right,' 'up,' and 'down' to move the avatar. The goal of the game is to collect a specific number of trophies—indicated to the user by a counter (cf. the right panel in Figure 1) —and subsequently deliver them to a bear (cf. the right corner in the top of Figure 1). The target (i.e., the number of trophies to be collected) varies for each of the six game levels, and the participants are only given one attempt to complete each level; they must finish the task within a specific time in order to win. During levels 1-4, there were no usability problems, and the participants generally exhibited neutral/positive emotions. In contrast, in levels 5 and 6, the introduction of usability problems, such as intentional alteration of the counter to prevent users from achieving the goal, led participants to show different levels of frustration. The baseline provided with the database presents the results of a binary classification task, that is, of discriminating between frustration and lack of frustration. With regard to the experimental settings: the authors extracted MFCC acoustic features and applied SVM and LSTM to the audio channels. The best baseline result was 58.8% of the unweighted average recall (UAR) for the considered speech channel.

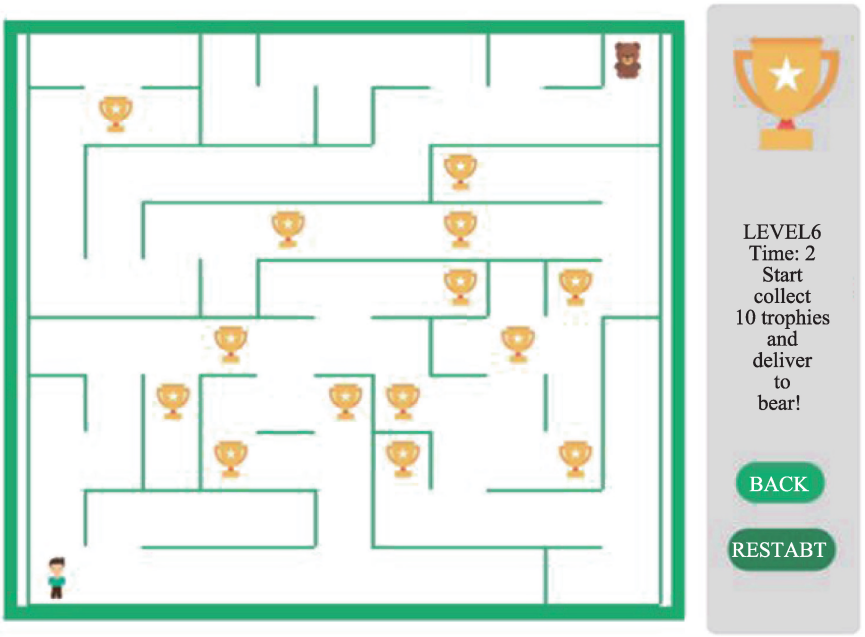

Figure 1 Level 6 interface of "Crazy Trophy": although the user has collected only 10 trophies, due to the (intentional) usability problem, the panel on the right indicates a doubled count, i.e., 20 trophies; this makes it impossible to win the game.

3.2 Feature sets

A wealth of different feature sets has been produced to perform analysis and recognition of speech. In

[1] The MGFD dataset is freely accessible in:https://zenodo.org/record/3957238#.X4A0v3X7TmF.

supervised machine learning, performance is often compared across different benchmark feature sets, which enables feature comparison for specific tasks. With this in mind, along with the MFCC features already considered for the baseline, we also evaluated the performance of the Mel-Spectrograms and eGeMAPS feature sets since they have been successfully used in the recognition of emotional speech in previous research[35,36].

3.2.1 Mel-Spectrogram

Mel-Spectrograms are generated by applying a Mel filter bank to spectrograms, which are extracted from audio signals via the short-term Fourier transform (STFT)[37]. The window length was 2000. We used a hop length of 800. The N-FFT value was 2000. The Mel filter bank converts spectrograms into the Mel scale, which was considered because it emphasizes low frequencies over higher ones, mirroring the perceptual capability of human ears. To compute the Mel-Spectrograms, we used the librosa Python package in our experiment. In Figure 2, a sample frustration Mel-Spectrogram is shown.

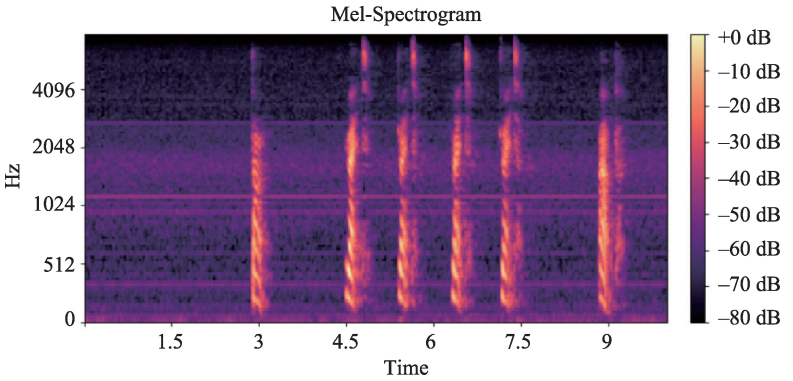

Figure 2 Sample Mel-spectrogram extracted from a clip of female frustration speech (seconds).

3.2.2 MFCCs

MFCCs are a representation that is derived by computing the cepstrum of melodic frequencies[38]. Owing to their high performance[39], MFCCs are one of the most commonly used filter bank-based feature types for speech processing applications, such as speech recognition[39], speaker verification/identification[40], and language identification[38]. Furthermore, MFCCs offer the advantages of low dimensionality and the independence of the position of the partial narrow-band corruption across feature dimensions[41]. In this work, we extracted a total of 39 dimensional MFCC features, including 13 MFCC coefficients and the first and second delta regression coefficients (both the first and second deltas having 13 dimensions).

3.2.3 eGeMAPS

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)[42,43] is a small (low dimensional) knowledge-based acoustic feature set designed for the purpose of achieving a high level of robustness when capturing emotion from speech[42]. It consists of 42 low-level descriptors (LLDs) and 2 functionals, that is, the arithmetic mean and the coefficient of variation[42].The overall dimension of the eGeMAPS is 88.

3.3 Data partitioning

As in the baseline experimental setting, we applied the independent Leave-Three-Speakers-Out Cross Validation method, meaning that the 67 participants (40 males, 27 females) were divided into two subsets. One subset was employed as a fixed test set (10 males, 12 females), while the other (30 males, 15 females)

was considered as training and development set for cross validation and then subsequently partitioned into 15 subsets, each including 2 males and 1 female. During the training process, one subset was chosen as the development set, while the other 14 subsets were used as the training set. The fixed test set was used to evaluate the model. The distributions of speakers and instances in the subsets are given in Table 1.

Table 1 Partitioning the dataset by participants and instances

Models	Train & Dev	Test	Σ
Speakers	45	22	67
Gender (M:F)	30:15	10:12	40:27
Frustration	483	209	692
Non-Frustration	3763	1991	5754

3.4 Equipment set-up

High-performance computing systems increasingly incorporate the computational power provided by accelerators, especially GPUs. Thus, to implement the experiments, we applied the NVidia GeForce GTX Titan X as our GPU in order to increase the computational performance. Our deep learning models were programmed using PyTorch with a MacBook[2].

4 Deep learning approaches

Current state-of-the-art deep learning methods have explored the use of residual networks (ResNets) in applications such as earthquake signal detection[44], infant crying recognition[45], and pediatric pneumonia diagnosis[46]. These methods are characterized by the use of particularly deep networks. Model training is especially time consuming when the depth of the ResNets increases, a problem that may be alleviated by the use of WideResNets[19], as outlined above. One of the properties that WideResNets share with ResNets is the inclusion of shortcut connections[31], that is, connections that skip one or more layers and which help to reduce the vanishing gradient problem. The main difference between the ResNets and WideResNets is the width of the networks.

In this work, the architectures of ResNets and WideRestNets are compared with three other widely used CNNs: AlexNet, VGG, and EfficientNet. The AlexNet[47] model contains five convolutional layers, two fully connected layers, and a softmax output layer. In contrast to previous shallow networks, rectified linear units (ReLUs), overlap pooling, and dropout are used in this model[47]. The VGG-11 network consists of eleven weight layers, eight convolutional layers, and three fully connected layers. It is classified by a softmax classifier layer for its output. The promising EfficientNet[48] has been presented as a benchmark for balancing network depth, width, and resolution. Indeed, convolutional neural networks have been widely scaled up to improve network accuracy, one simple composite scaling method is based on a fixed set of scaling coefficients, thus uniformly scaling the depth, width, and resolution parameters. This scaling method can be used to efficiently avoid tedious fine-tuning processing[48].

An overview of the architectures of ResNets and WideResNets used in this work is illustrated in Figure 3. One of the critical parameters in WideResNets is the widening factor k. It is a coefficient that is multiplied by the width of the residual blocks for transformation from ResNets to WideResNets. However, before multiplication with the k factor, the best value cannot be deterministically defined. For this reason, we decided to compare the performance of this architecture using k=2, k=3, and k=4. In Figure 3, the k factor is set up in the residual blocks. Batch normalization[49] is used at the output of the convolutional layer, and ReLU is employed as the activation function in all the residual blocks. Our network was trained

[2] For reproducibility, the code to re-implement the experiments is freely accessible in : https://github.com/Meishu619/frustrationrecognition-fromspeech.

with a batch size of 32 samples. It uses weighted cross-entropy as the loss to be optimised with the Adam optimizer. The learning rate in our experiments was 0.0001. In our architecture, the input sizes for the networks are 1000×39 for MFCCs, 1000×900 for Mel-Spectrograms, and 1000×88 for eGeMAPS.

5 Experimental results

The classification performance is measured using the unweighted average recall (UAR), an evaluation metric already used in previous work on the same dataset[9] and broadly used in the field as it is well suited for the commonly encountered class imbalance. Table 1 presents the results obtained using the frustration recognition models that were trained using different acoustic feature sets. From the evaluation of the results, we observe that the WideResNets50-2 architecture achieved the best performance using Mel-Spectrogram acoustic features (UAR=93.1%; cf. test in Table 2); that is, our best results outperformed the baseline by 34.3%. Better results for MFCCs were also achieved with the WideResNets50-2 architecture (UAR=92.9%; cf. test in Table 2), and when using eGeMAPS as acoustic features, WideResNets50-3 outperformed the other architectures (UAR=85.7%; cf. test in Table 2). We found that the best k value for Mel-Spectrograms and MFCCs is 2, whereas for eGeMAPS, the optimal k is 3, which indicates that the selection of k is input-dependent and that k therefore needs to be fine-tuned for different input features. Note that we did not test other values because we observed that higher k values yielded lower performance.

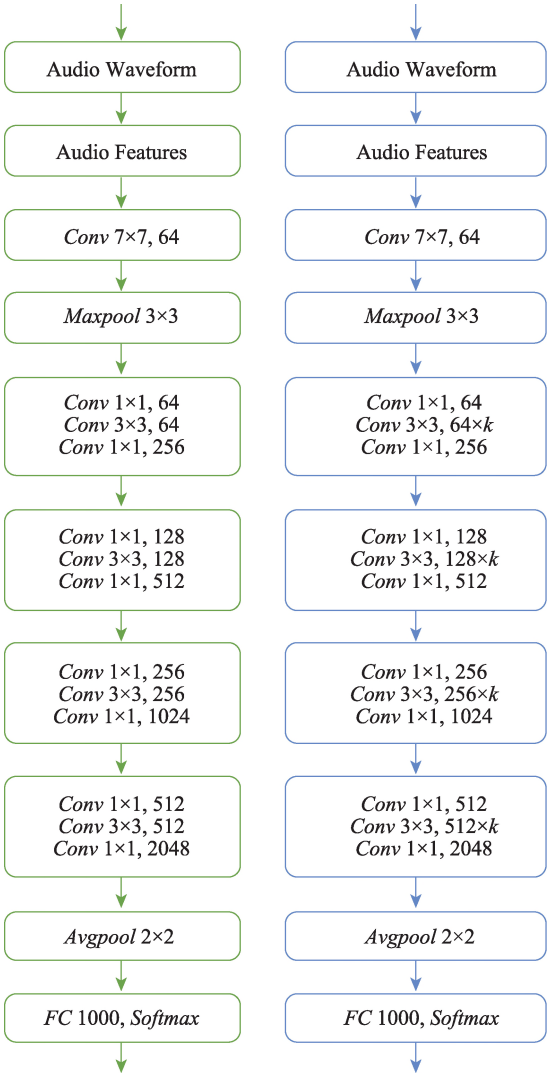

Figure 3 Illustration of the WideResNets (blue) and ResNets (green) model architectures.

Overall, a comparison with the baseline results[9] shows that CNN-based architectures offer a significant improvement (from 58.8% to 93.1% UAR), which is evident for all the evaluated CNN models—even the worst result (UAR=73.5%), achieved by the VGG11 architecture using the eGeMAPS feature set, outperformed the baseline. This may be partially attributable to the data collection procedure: in the "non-frustration" clips, individuals pronounced the commands "left," "right," "up," and "down" sequentially, frequently and confidently. In contrast, in the "frustration" clips, long silences were introduced between the commands. We hypothesize that although the MGFD dataset contains time-sequential data, the numerous silences biased the performance of the LSTM architecture when modeling frustration in the MGFD dataset. Another interesting outcome is that in most cases the WideResNets architectures outperform the ResNets-based models, confirming that an increase in the width rather than the depth of the residual blocks leads to better performance on our speech-based frustration detection task. Despite the remarkable performance of EfficientNet in image classification tasks[48], this type of architecture obtained lower UAR values than the ResNets and WideResNets architectures considered here, which may be because the resolution of the

speech signals in the MGDF dataset is lower than the image dimensions in image classification. Another possible explanation is that in EfficientNet there are no skip connections.

6 Conclusion

In the present research, following our work on speech-based frustration recognition during game interaction, we showed that the use of Mel-Spectrogram acoustic features with a WideResNets architecture yields a significant improvement (34.3% greater UAR) over the baseline results. Through a comparison of several models, our results confirm that convolutional neural networks (CNNs) in general, and Wide Residual Networks (WideResNets) in particular, are suitable architectures for successfully retrieving emotional content from speech. Future work will need to re-evaluate this finding on other datasets including frustration and more general paralinguistic tasks to provide additional evidence of the value of WideResNets architectures in the field.

Table 2 Unweighted Average Recall (UAR[%]) results obtained by the evaluated CNN-based models. Three acoustic representations of the input audio signals are considered: MFCCs, eGeMAPS, and Mel-Spectrograms. The UAR results are given; the best results for each feature set are highlighted in bold. The baseline results from a previous paper are included[9]

Model	MFCCs	eGeMAPS	Mel-Spectrograms
Baseline (SVM) [9]	58.8	None	None
Baseline (LSTM) [9]	57.4	None	None
AlexNet	88.4	82.3	89.2
VGG11	80.1	73.5	80.0
EfficientNet-0	86.1	75.5	87.0
EfficientNet-1	86.2	76.7	89.3
EfficientNet-4	89.2	80.1	90.8
ResNets18	83.7	80.2	88.6
ResNets34	90.0	84.2	89.7
ResNets50	87.3	80.7	91.9
WideResNets50-2	**92.9**	83.2	**93.1**
WideResNets50-3	89.3	**85.7**	91.9
WideResNets50-4	90.8	84.7	90.4

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

1 Scheirer J, Fernandez R, Klein J, Picard R W. Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers, 2002, 14(2): 93–118
DOI:10.1016/s0953-5438(01)00059-5

2 Caroux L, Isbister K, Le B L, Vibert N. Player-video game interaction: a systematic review of current concepts. Computers in Human Behavior, 2015, 48: 366–381
DOI:10.1016/j.chb.2015.01.066

3 Craig S D, D'Mello S, Witherspoon A, Graesser A. Emote aloud during learning with AutoTutor: applying the facial action coding system to cognitive–affective states during learning. Cognition & Emotion, 2008, 22(5): 777–788
DOI:10.1080/02699930701516759

4 Picard R W, Klein J. Computers that recognise and respond to user emotion: theoretical and practical implications. Interacting with Computers, 2002, 14(2): 141–169
DOI:10.1016/s0953-5438(01)00055-8

5 Yannakakis G N, Isbister K, Paiva A, Karpouzis K. Emotion in games. IEEE Transactions on Affective Computing, 2014, 5(1): 1–2

6 Gilleade K M, Dix A. Using frustration in the design of adaptive videogames. In: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology-ACE '04. Singapore, New York, ACM Press, 2004
DOI:10.1145/1067343.1067372

7 Picard R W. Affective computing. MIT Press, Cambridge, MA, USA, 2000

8 Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G. Cross-corpus acoustic emotion

recognition: variances and strategies. IEEE Transactions on Affective Computing, 2010, 1(2): 119−131
DOI:10.1109/t-affc.2010.8

9 Song M S, Yang Z J, Baird A, Parada-Cabaleiro E, Zhang Z X, Zhao Z P, Schuller B. Audiovisual analysis for recognising frustration during game-play: introducing the multimodal game frustration database. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). Cambridge, United Kingdom, IEEE, 2019, 517−523
DOI:10.1109/acii.2019.8925464

10 Li C K, Wang P C, Wang S, Hou Y H, Li W Q. Skeleton-based action recognition using LSTM and CNN. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Hong Kong, China, IEEE, 2017, 585−590
DOI:10.1109/icmew.2017.8026287

11 Zhao J F, Mao X, Chen L J. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 2019, 47: 312−323
DOI:10.1016/j.bspc.2018.08.035

12 Kollias D, Zafeiriou S. A multi-component CNN-RNN approach for dimensional emotion recognition in-the-wild. 2018

13 Zainab M, Usmani A R, Mehrban S, Hussain M. FPGA based implementations of RNN and CNN: a brief analysis. In: 2019 International Conference on Innovative Computing (ICIC). Lahore, Pakistan, IEEE, 2019, 1−8
DOI:10.1109/icic48496.2019.8966676

14 Yin W P, Kann K, Yu M, Schütze H. Comparative study of CNN and RNN for natural language processing. 2017

15 Keren G, Schuller B. Convolutional RNN: an enhanced model for extracting features from sequential data. In: 2016 International Joint Conference on Neural Networks (IJCNN). Vancouver, BC, Canada, IEEE, 2016, 3412−3419
DOI:10.1109/ijcnn.2016.7727636

16 Baird A, Amiriparian S, Cummins N. Automatic classification of autistic child vocalisations: a novel database and results. In: Proceedings of Interspeech 2017. Stockholm, Sweden, International Speech Communication Association, 2017

17 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, IEEE, 2016, 770−778
DOI:10.1109/cvpr.2016.90

18 He K M, Zhang X Y, Ren S Q, Sun J. Identity mappings in deep residual networks. In: Computer Vision−ECCV 2016. Cham, Springer International Publishing, 2016, 630−645
DOI:10.1007/978-3-319-46493-0_38

19 Zagoruyko S, Komodakis N. Wide residual networks. Computer Science ArXiv, 2016

20 Angkititrakul P, Petracca M, Sathyanarayana A, Hansen J H L. UTDrive: driver behavior and speech interactive systems for in-vehicle environments. In: 2007 IEEE Intelligent Vehicles Symposium. Istanbul, Turkey, IEEE, 2007, 566−569
DOI:10.1109/ivs.2007.4290175

21 Boril H, Sadjadi S O, Kleinschmidt T. Analysis and detection of cognitive load and frustration in drivers' speech. In: Conference of the International Speech Communication Association. Makuhari, Chiba, Japan, DBLP, 2010

22 Arunachalam S, Gould D, Andersen E, Byrd D, Narayanan S. Politeness and frustration language in child-machine interactions. In: 7th European Conference on Speech Communication and Technology. 2001

23 Koelstra S, Muhl C, Soleymani M, Lee J S, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I. DEAP: a database for emotion Analysis;Using physiological signals. IEEE Transactions on Affective Computing, 2012, 3(1): 18−31
DOI:10.1109/t-affc.2011.15

24 Szwoch M. FEEDB: a multimodal database of facial expressions and emotions. In: 2013 6th International Conference on Human System Interactions (HSI). Sopot, Poland, IEEE, 2013, 524−531
DOI:10.1109/hsi.2013.6577876

25 Douglas-Cowie E, Campbell N, Cowie R, Roach P. Emotional speech: towards a new generation of databases. Speech Communication, 2003, 40(1/2): 33−60
DOI:10.1016/s0167-6393(02)00070-5

26 Busso C, Bulut M, Lee C C. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4):335−359

DOI: 10.1007/s10579-008-9076-6

27 Chen L S. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. University of Illinois at Urbana-Champaign, 2000

28 Aggarwal V, Wang W L, Eriksson B, Sun Y F, Wang W Q. Wide compression: tensor ring nets. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, IEEE, 2018, 9329–9338
DOI:10.1109/cvpr.2018.00972

29 Martinel N, Foresti G L, Micheloni C. Wide-slice residual networks for food recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, NV, USA, IEEE, 2018, 567–576
DOI:10.1109/wacv.2018.00068

30 Zhangy Y, Ozayy M, Li S H, Okatani T. Truncating wide networks using binary tree architectures. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, IEEE, 2017, 2116–2124
DOI:10.1109/iccv.2017.231

31 Ferreira C A, Aresta G, Cunha A, Mendonça A M, Campilho A. Wide residual network for lung-rads™ screening referral. In: 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG). Lisbon, Portugal, IEEE, 2019, 1–4
DOI:10.1109/enbeng.2019.8692560

32 Zerhouni E, Lányi D, Viana M, Gabrani M. Wide residual networks for mitosis detection. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne, VIC, Australia, IEEE, 2017, 924–928
DOI:10.1109/isbi.2017.7950667

33 Panda A, Naskar R, Rajbans S, Pal S. A 3D wide residual network with perceptual loss for brain MRI image denoising. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Kanpur, India, IEEE, 2019, 1–7
DOI:10.1109/icccnt45670.2019.8944535

34 Dahlbäck N, Jönsson A, Ahrenberg L. Wizard of Oz studies: why and how. In: Proceedings of the 1st International Conference on Intelligent User interfaces-IUI '93. Orlando, Florida, USA, NewYork, ACMPress, 1993
DOI:10.1145/169891.169968

35 Xue W, Cucchiarini C, van Hout R, Strik H. Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech. In: SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education. ISCA, 2019
DOI:10.21437/slate.2019-9

36 Valstar M, Pantic M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, Scherer S, Stratou G, Cowie R. AVEC 2016: depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge-AVEC'16. Amsterdam, TheNetherlands, ACM Press, 2016
DOI:10.1145/2988257.2988258

37 Shen J, Pang R, Weiss R J. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. IEEE, 2018

38 Eyben F, Weninger F, Gross F, Schuller B. Recent developments in open SMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia-MM'13. Barcelona, Spain, ACM Press, 2013
DOI:10.1145/2502081.2502224

39 Schmitt M, Ringeval F, Schuller B. At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech. In: Interspeech 2016. ISCA, 2016
DOI:10.21437/interspeech.2016-1124

40 Geiger J T, Hofmann M, Schuller B, Rigoll G. Gait-based person identification by spectral, cepstral and energy-related audio features. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada, IEEE, 2013, 458–462
DOI:10.1109/icassp.2013.6637689

41 Winursito A, Hidayat R, Bejo A. Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. In: 2018 International Conference on Information and Communications Technology (ICOIACT). Yogyakarta, Indonesia, IEEE, 2018, 379–383

DOI:10.1109/icoiact.2018.8350748

42 Eyben F, Scherer K R, Schuller B W, Sundberg J, André E, Busso C, Devillers L Y, Epps J, Laukka P, Narayanan S S, Truong K P. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 2016, 7(2): 190–202
DOI:10.1109/taffc.2015.2457417

43 Cummins N, Amiriparian S, Hagerer G, Batliner A, Steidl S, Schuller B W. An image-based deep spectrum feature representation for the recognition of emotional speech. In: Proceedings of the 2017 ACM on Multimedia Conference-MM '17. Mountain View, California, USA, ACM Press, 2017
DOI:10.1145/3123266.3123371

44 Mousavi S M, Zhu W Q, Sheng Y X, Beroza G C. CRED: a deep residual network of convolutional and recurrent units for earthquake signal detection. Scientific Reports, 2019, 9: 10267
DOI:10.1038/s41598-019-45748-1

45 Xie X, Zhang L, Wang J. Application of residual network to infant crying recognition. Journal of Electronics and Information Technology, 2019, 41(1): 233–239
DOI: 10.11999/JEIT180276

46 Liang G B, Zheng L X. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Computer Methods and Programs in Biomedicine, 2020, 187: 104964
DOI:10.1016/j.cmpb.2019.06.023

47 Xiao L S, Yan Q, Deng S Y. Scene classification with improved AlexNet model. In: 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). Nanjing, China, IEEE, 2017, 1–6
DOI:10.1109/iske.2017.8258820

48 Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks. 2019

49 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015