frontiers in Computer Science

ORIGINAL RESEARCH
published: 09 February 2022
doi: 10.3389/fcomp.2022.837269

Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children

Manuel Milling[1], Alice Baird[1], Katrin D. Bartl-Pokorny[1,2,3], Shuo Liu[1], Alyssa M. Alcorn[4], Jie Shen[5], Teresa Tavassoli[6], Eloise Ainger[4], Elizabeth Pellicano[7], Maja Pantic[5], Nicholas Cummins[8] and Björn W. Schuller[1,5]*

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [2] Research Unit iDN – Interdisciplinary Developmental Neuroscience, Division of Phoniatrics, Medical University of Graz, Graz, Austria, [3] Division of Physiology, Otto Loewi Research Center, Medical University of Graz, Graz, Austria, [4] Centre for Research in Autism and Education, UCL Institute of Education, London, United Kingdom, [5] Department of Computing, Imperial College London, London, United Kingdom, [6] School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom, [7] School of Education, Macquarie University, Sydney, NSW, Australia, [8] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London, London, United Kingdom

OPEN ACCESS

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Chuangao Tang,
Southeast University, China
Hongli Chang,
Southeast University, China

***Correspondence:**
Manuel Milling
manuel.milling@informatik.uni-augsburg.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *16 December 2021*
Accepted: *03 January 2022*
Published: *09 February 2022*

Citation:
Milling M, Baird A, Bartl-Pokorny KD, Liu S, Alcorn AM, Shen J, Tavassoli T, Ainger E, Pellicano E, Pantic M, Cummins N and Schuller BW (2022) Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children. Front. Comput. Sci. 4:837269. doi: 10.3389/fcomp.2022.837269

Individuals with autism are known to face challenges with emotion regulation, and express their affective states in a variety of ways. With this in mind, an increasing amount of research on automatic affect recognition from speech and other modalities has recently been presented to assist and provide support, as well as to improve understanding of autistic individuals' behaviours. As well as the emotion expressed from the voice, for autistic children the dynamics of verbal speech can be inconsistent and vary greatly amongst individuals. The current contribution outlines a voice activity detection (VAD) system specifically adapted to autistic children's vocalisations. The presented VAD system is a recurrent neural network (RNN) with long short-term memory (LSTM) cells. It is trained on 130 acoustic Low-Level Descriptors (LLDs) extracted from more than 17 h of audio recordings, which were richly annotated by experts in terms of perceived emotion as well as occurrence and type of vocalisations. The data consist of 25 English-speaking autistic children undertaking a structured, partly robot-assisted emotion-training activity and was collected as part of the DE-ENIGMA project. The VAD system is further utilised as a preprocessing step for a continuous speech emotion recognition (SER) task aiming to minimise the effects of potential confounding information, such as noise, silence, or non-child vocalisation. Its impact on the SER performance is compared to the impact of other VAD systems, including a general VAD system trained from the same data set, an out-of-the-box Web Real-Time Communication (WebRTC) VAD system, as well as the expert annotations. Our experiments show that the child VAD system achieves a lower performance than our general VAD system, trained under identical conditions, as we obtain receiver operating characteristic area under the curve (ROC-AUC) metrics of 0.662 and 0.850, respectively. The SER results show varying performances across valence and arousal depending on the utilised VAD system with a maximum concordance correlation coefficient (CCC) of 0.263 and a minimum

root mean square error (RMSE) of 0.107. Although the performance of the SER models is generally low, the child VAD system can lead to slightly improved results compared to other VAD systems and in particular the VAD-less baseline, supporting the hypothesised importance of child VAD systems in the discussed context.

Keywords: affective computing, voice activity detection, deep learning, speech emotion recognition, children with autism, robot human interaction

1. INTRODUCTION

Speech emotion recognition (SER) is a prominent subfield of Affective Computing as the complexity of the human speech apparatus together with the communicative importance of emotions in speech make a good understanding of the problem both difficult and desirable, which becomes apparent from the long history of emotion recognition challenges (Valstar et al., 2013; Ringeval et al., 2019; Stappen et al., 2021). The subjective nature of emotions leads to a variety of emotion recognition tasks, which make the possibility for a one-fits-all solution not the optimal approach to capture the subtle variation in emotion expression. As most models are only focused on a single corpus, which can range from acted emotions (Busso et al., 2008) via emotions induced by a trigger (Koelstra et al., 2012) to spontaneous emotions (Stappen et al., 2020), and is often recorded for adult individuals, the application of SER models needs to be chosen with care and in general adapted to the specific scenario.

Continuous SER tasks, especially in interactive scenarios, such as robot-assisted child-robot interactions, can be prone to auditory artefacts, and limited instances of speech, creating the need to discriminate between background noise and information-rich instances. Voice activity detection (VAD) systems are therefore commonly used in SER tasks to remove unvoiced segments of the audio signal, for instance displayed in Harár et al. (2017), Alghifari et al. (2019) and Akçay and Oğuz (2020). In a scenario with more than one speaker however, VAD alone might not be enough to filter out all non-relevant information about a specific speaker's affective state.

Autism is a neurodevelopmental condition that is associated with difficulties in social communication and restricted, repetitive patterns of behaviour, interests, or activities (American Psychiatric Association, 2013). The clinical picture of autism is heterogeneous, including diversity in autistic characteristics and spoken language skills, and frequently occurring comorbidities, such as anxiety disorder, attention-deficit hyperactivity disorder, developmental coordination disorder, or depressive disorders (Kopp et al., 2010; Lord et al., 2018; Zaboski and Storch, 2018; Hudson et al., 2019). Difficulties in socio-communicative skills and recognition and expression of emotion in autistic children can make interactions with their family, peers, and professionals challenging.

However, only few research projects have investigated how recent technology including Artificial Intelligence can help to better understand the needs and improve the conditions of children with autism: the ASC-inclusion project developed a platform aiming to playfully support children in understanding and expressing emotions through a comprehensive virtual world (Schuller, 2013), for instance through serious games (Marchi et al., 2018). The DE-ENIGMA project[1] focused on a better understanding of behaviour and needs of autistic children in a researcher-led robot-human-interaction (RCI) scenario, contributing to insights about robot predictability in RCI scenarios with children with autism (Schadenberg et al., 2021), as well as prediction of the severity of traits related to autism (Baird et al., 2017) and detection of echolalic vocalisations (Amiriparian et al., 2018), i.e., word or phrase repetitions of autistic children based on spoken utterances of their conversational partners. Schuller et al. introduced a task for the speech-based diagnosis of children with autism and other pervasive developmental disorders (Schuller et al., 2013). Particularly in the field of SER for individuals with autism, data appears quite sparse (Schuller, 2018), presumably caused in part due to the considerable time-expense needed to gather such data from autistic children. Rudovic et al. developed a personalised multi-modal approach based on deep learning for affect and engagement recognition in autistic children, achieving up to 60 % agreement with human annotators, aiming to enable affect-sensitive child-robot interaction in therapeutic scenarios (Rudovic et al., 2018). From this overview of related works, there have been limited works, which model emotions of autistic children with continuous labelling strategies. To the best of our knowledge, no research as of yet has explored how VAD can improve such modelling.

In this manuscript, we investigate a subset of data collected in the DE-ENIGMA project (Shen et al., 2018). The presented data consist of about 17 h of audio recordings and rich annotations including continuously perceived affective state, and manually performed speaker diarisation. The data poses numerous challenges commonly associated with in-the-wild data including noise (for instance from robot or furniture movements) or varying distances to microphones. Additionally, a particular challenge in the current dataset results from the sparsity of child vocalisations in the interaction between child, robot, and researcher, as several children who took part in the study had limited-to-no spoken communication. In contrast to common continuous emotion recognition tasks, we hypothesised that a model focusing on the child vocalisations alone would be able to outperform other models, as we expect the child vocalisations to contain the most information about the children's affective states. For this reason, in the current work, we implement a VAD system specifically trained for vocalisations of autistic

[1] https://de-enigma.eu/

children on the dataset and evaluate its performance against a trained general VAD system - trained on all vocalisations of our dataset - as well as an implementation of the Web Real-Time Communication (WebRTC) VAD (Google, 2021) and the manual speaker diarisation annotations, for the SER task at hand. The WebRTC VAD is based on Gaussian mixture models (GMMs) and log energies of six frequency bands.

The remainder of this manuscript is organised as follows. In section 2, we provide a detailed overview of the investigated dataset. Furthermore, we introduce the deep learning-based methodology for both the VAD and the SER task in section 3. Subsequently, we present experimental results for the isolated VAD experiments, as well as the SER task with a combined VAD-SER system in section 4. Finally, we discuss the results and the limitations of our approaches in section 5 before we conclude our work in section 6.

2. DATASET

The Experiments in this manuscript are based on a subset of data gathered in the DE-ENIGMA Horizon 2020 project, which were collected in a school-based setting in the United Kingdom and Serbia. In this work, we solely focus on audio data from the British study arm of the project, for which all relevant data streams and annotations are available. Here, autistic children undertook emotion-recognition training activities based on the Teaching Children with Autism to Mind-Read programme (Howlin et al., 1999), under guidance of a researcher. Ethical approval was granted for this study by the Research Ethics Committee at the UCL Institute of Education and the University College London (REC 796). Children were randomly assigned to researcher-only sessions, or to sessions, which were supported by the humanoid robot Zeno-R2. Zeno is capable of performing different emotion-related facial expressions, and which was controlled by the researcher via an external interface. The sessions were recorded with multiple cameras and microphones covering different angles of the room.

Each child attended between one and five daily sessions (3.4 on average), yielding a total of 84 sessions with an average length of 12.4 min from 25 children (19 males, 6 females), 13 participating in researcher-only sessions and 12 participating in robot-assisted sessions, with an average age of 8.2 yrs (standard deviation: 2.5 yrs), led by three different researchers (only one researcher per child). We divided the data in a speaker-independent manner with respect to the children. As there were overall three researchers in the data set, each child only interacting with one researcher, we group our data splits based on the researchers. We do so to avoid overfitting of our machine learning models on person-specific speech characteristics of the researchers, who largely contribute to the vocalisations. An overview of the partitions is given in **Table 1**; the partitioning is being used for both types of experiments.

The sessions were richly annotated in terms of both audio and video data, following a pre-defined annotation protocol, including instructions for speaker diarisation, vocalisation type, occurrences of echolalia, type of non-verbal vocalisations, as well as emotion in terms of valence and arousal. For our study, we exploit the speaker diarisation annotations, the origin of the labels for voice activity detection, as well as valence and arousal annotations as labels for the SER system.

2.1. Speaker Diarisation Annotation

The Speaker Diarisation (in the British study arm) was performed by fluent English speakers utilising the ELAN annotation tool[2]. The task was to highlight any vocalisation of any speaker present within the session, i.e., the child, the researcher, any additionally present person (generally a teacher), or the robot Zeno. The annotators were able to base their decisions on a combination of the available video streams together with one of the video cameras' native audio recordings, as well as the according depiction of the raw audio wave form. The annotation tool further allowed annotators to skip to arbitrary points of the recording. Overall, each session was assessed by one annotator.

2.2. Emotion Annotation

The emotion annotations in the database aim to capture the emotional dimensions valence and arousal, i.e., continuous representations of how positive or negative (valence) and how sleepy or aroused (arousal) an emotional state seems. Emotional dimensions are a commonly used alternative to categorical emotions, like happy, angry, etc., when assessing people's emotional states. Five expert raters, all either native or near native English speakers, annotated their perception of the valence and arousal values expressed by the children in each session under consideration of the same video and audio data as in the speaker diarisation task. For the annotation process, raters were given a joystick (model *Logitech Extreme 3D Pro*) in order to annotate valence and arousal separately. While annotators were watching the recordings of the sessions, they changed the position of the joystick, which was continuously sampled with a sampling rate of 50 Hz and indicated degree and sign of the estimated valence or arousal values (positive in an *up* position, negative in a *down* position). The annotations of the different annotators for each session are summarised in a single gold standard sequence utilising the evaluator weighted estimator (EWE) (Schuller, 2013) gold standard. The EWE gold standard is commonly used in emotion recognition tasks (Ringeval et al., 2017, 2019) and considers annotator-specific weights depending on the pairwise correlation of the annotations. For our experiments, we use only one emotion label per second by calculating a second-wise average over the gold standard annotations.

3. METHODOLOGY

To explore the task of VAD-based SER, we employ two separate models based on feature extraction and recurrent neural networks (RNNs) with long short-term memory (LSTM) cells. The first component is a VAD component and the second is a SER component. The VAD model is presented with 1 s long audio chunks, and aims to label segments of the audio signal with a vocalisation present. The SER model is then trained on

[2]https://archive.mpi.nl/tla/elan

TABLE 1 | Overview of the three partitions of the data set: train, development (dev.), and test.

Partition	# children	# sessions	# researchers	child vocalisations	total vocalisations	total duration
Train	12	41	1	1:26:39	6:42:15	9:43:34
Dev.	4	15	1	0:18:27	1:24:37	3:14:35
Test	9	28	1	0:32:42	2:35:21	4:22:03
Overall	25	84	3	2:17:49	10:42:14	17:20:13

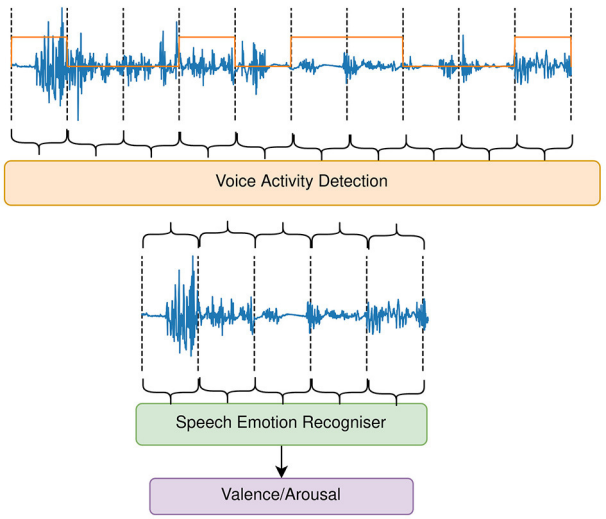

FIGURE 1 | Sequential employment of the (child) voice activity detection (VAD) and the speech emotion recognition (SER) system. The VAD system removes 1 s chunks of the audio signal, where no voice is detected. Any remaining 1 s chunks are fed into the SER system to predict continuous values for valence and arousal.

audio segments presumably containing speech, with the aim of predicting the affective dimensions valence and arousal in a continuous manner. An illustration of the combined system is depicted in **Figure 1**.

3.1. Voice Activity Detection

As the target of the VAD system is to remove as much information-shallow data from the audio data as possible, we compare several approaches here: at a first level, we try to filter for all vocalisations with general VAD systems, one specifically trained on our data set, the other one being an implementation of the WebRTC VAD system[3] (Google, 2021), commonly used as a comparison for other VAD systems, e.g., (Salishev et al., 2016; Nahar and Kai, 2020). The aggressiveness score of the WebRTC VAD is set equal to one. Additionally, we use the ground truth annotations for all vocalisations as a gold standard for a general VAD system. At a second level, we try to filter out only child vocalisations, which presumably contain the most information about the children's affective state. For this, we train a child VAD system on the data set mentioned above and use the ground truth annotations for child vocalisations for further comparison. Evaluations of the different impacts of general VADs and the child VAD are of further interest, as some information about the children's affective state could be retrieved from the interaction between the child and the researcher. Besides, a worse performance of the child VAD system compared to more robust general VAD systems could lead to detections of ambient noise and therefore potentially have a negative impact on the SER task.

Given the potentially short duration of vocalisations, we extract 130 ComParE2016 LLDs with a frame size of 10 ms and a hop size of 10 ms from the raw audio signal utilising the openSMILE toolkit (Eyben et al., 2010). The audio features are then fed into a two-layer bi-directional RNN with LSTM cells and a hidden layer size of 128 units, followed by a dense layer with a single output neuron indicating the confidence in the voice detection. The neural network architecture is similar to Hagerer et al. (2017), but has been adjusted based on preliminary experiments. We utilise a fixed sequence length of 100 samples during training time, i.e., the audio stream is cut into samples of 1 s length. During training of this regression problem, each frame is assigned the label 1 if speech is present or the label 0 if it is not.

The VAD models are trained for 8 epochs with a batch size of 256 utilising the Adam optimiser with a learning rate of 0.01 and mean square error (MSE) loss. We choose the rather small number of epochs based on the large amount of samples. Given that each second provides 100 sequence elements to the LSTM, the training includes around 2 000 optimisation steps. For the evaluation of the VAD system, we compute a receiver operating characteristic (ROC) curve, i.e., we vary the confidence threshold of the system, for which a frame is recognised as a detection in order to depict the relationship between true positive rate (TPR) and false positive rate (FPR).

For inference, we choose a confidence threshold, which corresponds to the equal-error-rate (EER), i.e., equal values of FPR and $1 - \text{TPR}$, visualised by the intersection of the ROC curve and the bisectional line $\text{TPR} + \text{FPR} = 1$. The VAD system is then used as a preprocessing step for the SER task, such that each second of audio is classified as containing voice activity if at least 25 % of the frames contained in 1 s are above the EER confidence threshold.

3.2. Speech Emotion Recognition

For the SER task we use 1 s chunks of audio extracted with the VAD system in order to predict a single continuous-valued valence (and arousal, respectively) value per audio chunk. The applied VAD system therefore impacts the SER task by the selection of audio chunks guided by the hypothesis that audio

[3]https://github.com/wiseman/py-webrtcvad

with child vocalisations contains the most information about the perceived affective states of the children and therefore leads to higher performance in the SER task.

Subsequently, we extract 88 functional features for each 1 s audio chunk according to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), a comprehensive expert-based audio feature selection (Eyben et al., 2015). The resulting sequence of features from one session is then used as an input to our deep learning model consisting of two RNN layers with LSTM cells and a hidden layer size of 128 units, followed by a dense layer with 128 neurons, a rectified linear unit (ReLU) activation and a dropout rate of 0.3. A final dense layer with a single neuron outputs the valence or arousal prediction for our task. The identical network architecture is trained independently for valence and arousal, respectively. With our methodology, we follow (Stappen et al., 2021), with an adjusted model architecture based on preliminary experiments.

The SER models are trained for 180 epochs with full batch optimisation – each session producing one sequence – utilising the Adam optimiser with a learning rate of 0.0001 and MSE loss. The much larger number of epochs compared to the VAD experiments is chosen based on the full batch optimisation, i.e., only one optimisation step is performed each epoch.

4. EXPERIMENTS

All experiments are implemented in Python 3 (Van Rossum and Drake, 2009), as well as TensorFlow 2 (Abadi et al., 2015) for deep learning models and training. The code is publicly available under[4].

4.1. Voice Activity Detection

For our VAD experiments, we train the architecture as described in section 3.1 with two different targets: (i) to recognise only child vocalisations, including overlap with others vocalisations and (ii) to recognise any vocalisation, including overlapping vocalisations. The two approaches are evaluated on the respective tasks. We thereby aim to evaluate the feasibility of training a general VAD system for the specifics and limitations of our dataset and to further investigate the presumably more challenging task of training a specialised VAD system for children with autism. Besides the evaluation of the VAD systems based on their raw performance, we further assess their impact on the SER task in the following section.

We report ROC-curves for both the child VAD system and the general VAD system on the respective tasks in **Figure 2**, as well as the EER and area-under-the-curve (AUC) in **Table 2**.

4.2. Speech Emotion Recognition

As described in section 3, we utilise our child VAD system and the general VAD system trained in the previous section in order to extract 1 s chunks from the session recordings if 25 out of the 100 frames within one second have a prediction confidence above the EER threshold. In a similar way 1 s chunks are extracted if the WebRTC VAD predicts a voice activity for at least 0.25 s of the audio. In the same manner, we use the ground truth annotations of child vocalisations, as well as ground truth annotations of all speakers to mimic a perfect child VAD and a perfect general VAD system. As a baseline, we use the audio without any VAD-based preprocessing (All Audio). **Figure 3** shows the distribution of valence and arousal values across partitions, as well as the test partition's adjusted distribution after filtering via the VAD systems and vocalisation annotations.

For evaluation, we use the root mean squared error (RMSE), as well as the concordance correlation coefficient (CCC) according to Lin (1989), which is defined between two distributions $\boldsymbol{x}$ and $\boldsymbol{y}$ as

$$\mathrm{CCC}(\boldsymbol{x},\boldsymbol{y}) = \frac{\rho(\boldsymbol{x},\boldsymbol{y})\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}}{\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{y}}^2 + (\mu_{\boldsymbol{x}} - \mu_{\boldsymbol{y}})^2}, \tag{1}$$

with the correlation coefficient ρ, as well as the mean μ and the standard deviation σ of the respective distribution. As the CCC is designed as a metric for sequences and has an inherent weakness for short sequences and sequences with little variation, we combine all predictions and labels from one data partition to two respective sequences when calculating the CCC. The results for valence and arousal are summarised in **Table 3**.

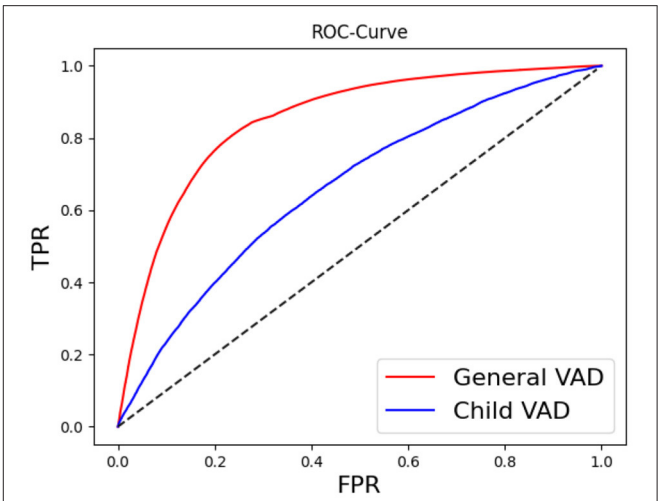

FIGURE 2 | Receiver operating characteristic (ROC)-curve of voice activity detection trained for child vocalisations specifically and for all vocalisations.

TABLE 2 | Equal-error-rates (EERs) and area-under-the-curve (AUC) for the child voice activity detection system and the general voice activity detection system evaluated on the respective task.

VAD System	EER	AUC
Child VAD	0.381	0.662
General VAD	0.215	0.850

[4]https://github.com/EIHW/VAD_SER_pipeline_ASC

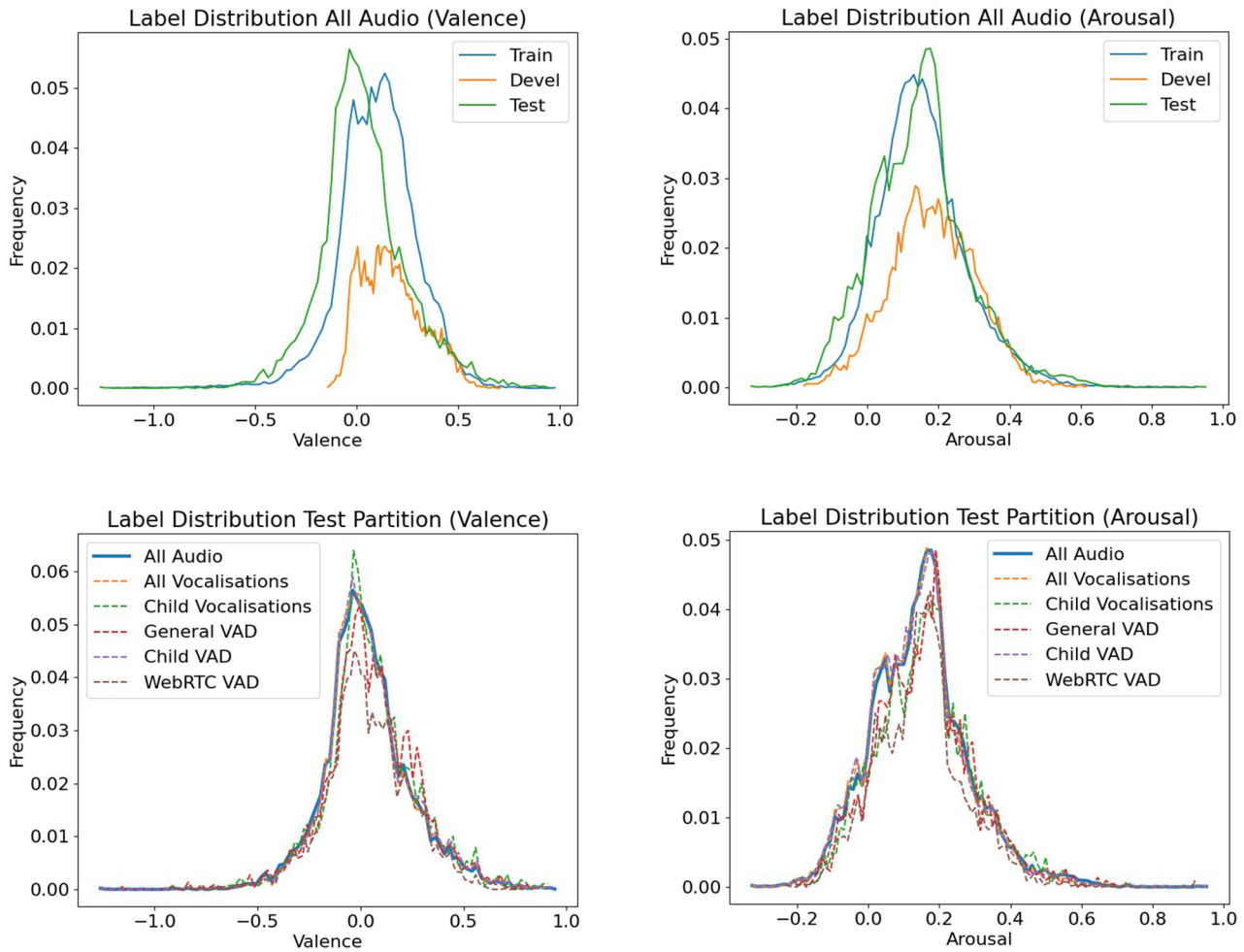

FIGURE 3 | Distributions of valence labels (left) and arousal labels (across) considering all audio data without VAD preprocessing across different partitions (top), as well the adjusted distributions of the test partition after preprocessing via the different VAD systems and vocalisation annotations (bottom).

TABLE 3 | Results of the speech emotion recognition (SER) task.

VAD System	# samples detected	Valence (CCC/RMSE)		Arousal (CCC/RMSE)	
		Dev	Test	Dev	Test
Child VAD	17,944	**0.200**/0.201	0.021/0.245	0.201/0.121	**0.168**/**0.138**
General VAD	40,013	0.012/**0.160**	0.117/0.260	0.100/0.120	0.154/0.142
WebRTC VAD	29,918	0.140/0.183	0.063/**0.224**	**0.263**/**0.107**	0.098/0.152
GT child vocalisations	10,961	0.153/0.169	0.085/0.277	0.182/0.115	0.145/0.143
GT all vocalisations	47,184	-0.032/**0.160**	**0.120**/0.231	0.166/0.114	0.105/ 0.156
All Audio	62,370	0.133/0.162	0.024/0.231	0.093/0.122	0.049/0.152

We report concordance correlation coefficient (CCC) and root mean squared error (RMSE) for valence and arousal with respect to the voice activity detection (VAD) system and ground truth (GT) annotations utilised for preprocessing of the data, as well as the baseline without a VAD preprocessing step (All Audio). Bold values indicate the best performance in each column.

5. DISCUSSION

Figure 2 and **Table 2** show that both a general voice activity system, as well as a child-specific voice activity system with a performance above-chance level can be trained from the data at hand. However, the general VAD system shows a clearly superior performance compared to the child-specific one. One apparent reason for this results from the dataset itself. **Table 1** highlights

that the dataset offers more than four times as many annotations for the general VAD compared to child VAD system, leading to a more unbalanced child VAD task. Moreover, the task of training a VAD system specialised and focused solely on autistic children appears to be generally more challenging, as the model not only needs to detect speech-typical characteristics, but also has to differentiate between speech characteristics of the speakers, i.e., the model has to find common patterns in vocalisations of children with different language levels and distinguish those from patterns in the researchers' voices. The different language levels of the children involved in the study, as well their unique ways of expression most likely made it difficult to uncover common characteristics.

Table 3 further shows that all considered VAD systems have a largely varying sensitivity. By the term sensitivity we mean in this context the total number of voice detection events independent of the correctness of the detections. The sensitivity of the child VAD system, aiming to detect the ground truth child vocalisations, can be considered too high with almost twice as many detected events compared to the number of human target annotations. The two remaining VAD systems naturally seem to be much more sensitive than the child VAD, as they do not aim at filtering out the child vocalisations only. However, both the out-of-the-box WebRTC VAD, as well as our trained general VAD both seem to show a lower sensitivity than the ground truth annotations of all speaker vocalisations with the WebRTC's deviation of detection events being considerably higher.

The top part of **Figure 3** shows that there are no large difference between the label distributions in the train and test partition for the SER task. The development partition however deviates substantially. The bottom part of **Figure 3** indicates the difference in emotion label distributions in the test set caused by the preprocessing via the different VAD approaches. Even though the choice of the VAD system has only little impact on the label distribution and therefore should not give any considerable, label-related advantage to any of the resulting SER experiments, it still affects the comparability of the results as it alters the test data.

According to **Table 3**, the best test results for arousal in our SER experiments are obtained with the child VAD preprocessing, even outperforming the preprocessing based on ground truth annotations. These results seem in-line with the hypothesis that considering only child vocalisations could improve the performance of SER systems for autistic children and they further suggest a reasonable system performance of the child VAD. However, this analysis only holds to a certain amount for the arousal development set and even less for the valence experiments, which tend to achieve lower performance in acoustic SER tasks compared to arousal experiments. Nevertheless, the VAD-based systems outperform the VAD-less system in most experiments, suggesting a clear advantage of VAD-based systems for the task at hand. Limitations to the expressiveness of the results discussed here have to be taken into account, as small improvements together with a low overall performance of the SER models are not always consistent across the investigated evaluation metrics.

Future work shall further investigate the impact of a child-specific VAD system in a multi-modal emotion recognition approach. Given the complex scenarios resulting from sessions with autistic children, it is inevitable that not all modalities are available at all times, as children for instance move out of the focus of the cameras or are silent for an extended period of time. The detection and consideration of those missing modalities, for instance in form of a VAD system contributing to a weighted feature fusion, might therefore have a substantial influence on model behaviour and even help with explaining the decisions of applied approaches.

6. CONCLUSION

With this contribution, we discussed the feasibility and utility of a VAD system, specifically trained on autistic child vocalisations, for SER tasks in robot-assisted intervention sessions for autistic children in order to improve programme success for children with autism. Given the size as well as the noise-heavy quality of the dataset, we showed that the voice activity component could be trained with reasonable performance, while being inferior to an identically trained general VAD system. Our results further suggest that the use of VAD systems, and in particular child VAD systems, could lead to slight improvements of continuous SER for autistic children, even though an overall low performance across SER models, most likely caused by the challenges of the task at hand, weaken the expressiveness of the results. Further research based on this work will examine the use of child VAD systems as a basis for missing data strategies in multi-modal SER tasks.

DATA AVAILABILITY STATEMENT

The data set presented in this article cannot be made publicly available in its current form due to ethical reasons.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee at UCL Institute of Education, University College London (REC 796). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MM: literature search, experimental design, experiment implementation, computational analysis, manuscript preparation, and drafting manuscript. AB: experimental design, feature extraction, literature search, annotation guideline development and implementation, and manuscript preparation. KB-P: literature search, manuscript preparation, consultancy on experimental task and study-related aspects of child development. SL: experimental design and advise on implementation. AA, JS, and EA: study design and implementation, annotation guideline development and implementation. TT: study design and implementation. EP:

study design and manuscript preparation. NC: study design, annotation guideline development and implementation, and manuscript preparation. MP: study design. BS: study design and manuscript editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the EU's Horizon 2020 Programme under grant agreement No. 688835 (RIA DE-ENIGMA), the European Commission's Erasmus+ project – 'EMBOA, Affective loop in Socially Assistive Robotics as an intervention tool for children with autism' under contract No. 2019-1-PL01-KA203-065096, and the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

ACKNOWLEDGMENTS

We are enormously grateful to all of the children, parents, teachers, researchers, and schools who so generously took part in our study. We would also like to thank the DE-ENIGMA student volunteers at the UCL Institute of Education for their contributions to school studies and data annotation.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems.* Available online at: https://www.tensorflow.org/ (accessed December 13, 2021).

Akçay, M. B., and Oğuz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* 116, 56–76. doi: 10.1016/j.specom.2019.12.001

Alghifari, M. F., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., Janin, Z., et al. (2019). On the use of voice activity detection in speech emotion recognition. *Bull. Elect. Eng. Inf.* 8, 1324–1332. doi: 10.11591/eei.v8i4.1646

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, Arlington, VA: APA.

Amiriparian, S., Baird, A., Julka, S., Alcorn, A., Ottl, S., Petrović, S., et al. (2018). "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Proceedings of the Interspeech 2018* (Hyderabad), 2334–2338.

Baird, A., Amiriparian, S., Cummins, N., Alcorn, A. M., Batliner, A., Pugachevskiy, S., et al. (2017). "Automatic Classification of autistic child vocalisations: a novel database and results," in *Proceedings of the Interspeech 2017* (Stockholm), 849–853.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *MM '10* (New York, NY: Association for Computing Machinery), 1459–1462.

Google (2021). *WebRTC.* Available online at: https://webrtc.org/. (accessed December 13, 2021).

Hagerer, G., Pandit, V., Eyben, F., and Schuller, B. (2017). "Enhancing lstm rnn-based speech overlap detection by artificially mixed data," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, Erlangen.

Harár, P., Burget, R., and Dutta, M. K. (2017). "Speech emotion recognition with deep learning," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)* (Noida: IEEE), 137–140.

Howlin, P., Baron-Cohen, S., and Hadwin, J. (1999). *Teaching Children With Autism to Mind-Read: A Practical Guide for Teachers and Parents*. Chichester: J. Wiley & Sons Chichester.

Hudson, C. C., Hall, L., and Harkness, K. L. (2019). Prevalence of depressive disorders in individuals with autism spectrum disorder: A meta-analysis. *J. Abnormal Child Psychol.* 47, 165–175. doi: 10.1007/s10802-018-0402-1

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). Deap: A database for emotion analysis ;using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Kopp, S., Beckung, E., and Gillberg, C. (2010). Developmental coordination disorder and other motor control problems in girls with autism spectrum disorder and/or attention-deficit/hyperactivity disorder. *Res. Develop. Disabil.* 31, 350–361. doi: 10.1016/j.ridd.2009.09.017

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268. doi: 10.2307/2532051

Lord, C., Elsabbagh, M., Baird, G., and Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *Lancet* 392, 508–520. doi: 10.1016/S0140-6736(18)31129-2

Marchi, E., Schuller, B., Baird, A., Baron-Cohen, S., Lassalle, A., O'Reilly, H., et al. (2018). The asc-inclusion perceptual serious gaming platform for autistic children. *IEEE Trans. Games* 11:328–339. doi: 10.1109/TG.2018.2864640

Nahar, R., and Kai, A. (2020). "Effect of data augmentation on dnn-based vad for automatic speech recognition in noisy environment," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)* Kobe, 368–372.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al. (2019). "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop AVEC '19* (New York, NY: Association for Computing Machinery), 3–12.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., et al. (2017). "Avec 2017: real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17* (New York, NY: Association for Computing Machinery), 3–9.

Rudovic, O., Lee, J., Dai, M., Schuller, B., and Picard, R. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science* 3:eaao6760. doi: 10.1126/scirobotics.aao6760

Salishev, S., Barabanov, A., Kocharov, D., Skrelin, P., and Moiseev, M. (2016). "Voice activity detector (vad) based on long-term mel frequency band features," in *Text, Speech, and Dialogue*, eds P. Sojka, A. Horák, I. Kopeček, and Pala, K. (Cham: Springer International Publishing), 352–358.

Schadenberg, B. R., Reidsma, D., Evers, V., Davison, D. P., Li, J. J., Heylen, D. K., et al. (2021). Predictable robots for autistic children-variance in robot behaviour, idiosyncrasies in autistic children's characteristics, and child–robot engagement. *ACM Trans. Comput. Human Interact.* 28, 1–42. doi: 10.1145/3468849

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association* (Lyon).

Schuller, B. W. (2013). *Intelligent Audio Analysis* (Berlin: Springer Publishing Company, Incorporated).

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 90–99. doi: 10.1145/3129340

Shen, J., Ainger, E., Alcorn, A., Dimitrijevic, S. B., Baird, A., Chevalier, P., et al. (2018). Autism data goes big: A publicly-accessible multi-modal database of child interactions for behavioural and machine learning research. In *International Society for Autism Research Annual Meeting* (Kansas City, MO).

Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., et al. (2020). "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop MuSe'20* (New York, NY), 35–44.

Stappen, L., Meßner, E.-M., Cambria, E., Zhao, G., and Schuller, B. W. (2021). "Muse 2021 challenge: Multimodal emotion, sentiment, physiological-emotion, and stress detection," in *Proceedings of the 29th ACM International Conference on Multimedia MM '21* (New York, NY), 5706–5707.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "Avec 2013 - the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge AVEC '13* (New York, NY), 3–10.

Van Rossum, G., and Drake, F. L. (2009). *Python 3 Reference Manual*. (Scotts Valley, CA: CreateSpace).

Zaboski, B. A., and Storch, E. A. (2018). Comorbid autism spectrum disorder and anxiety disorders: a brief review. *Future Neurol.* 13, 31–37. doi: 10.2217/fnl-2017-0030

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.