

Article

# A Neural Network Architecture for Children’s Audio–Visual Emotion Recognition

Anton Matveev , Yuri Matveev \*, Olga Frolova, Aleksandr Nikolaev and Elena Lyakso 

Child Speech Research Group, Department of Higher Nervous Activity and Psychophysiology, St. Petersburg University, St. Petersburg 199034, Russia; aush.tx@gmail.com (A.M.); olchel@yandex.ru (O.F.); al.nikolajew@gmail.com (A.N.); lyakso@gmail.com (E.L.)

\* Correspondence: yunmatveev@gmail.com

**Abstract:** Detecting and understanding emotions are critical for our daily activities. As emotion recognition (ER) systems develop, we start looking at more difficult cases than just acted adult audio–visual speech. In this work, we investigate the automatic classification of the audio–visual emotional speech of children, which presents several challenges including the lack of publicly available annotated datasets and the low performance of the state-of-the-art audio–visual ER systems. In this paper, we present a new corpus of children’s audio–visual emotional speech that we collected. Then, we propose a neural network solution that improves the utilization of the temporal relationships between audio and video modalities in the cross-modal fusion for children’s audio–visual emotion recognition. We select a state-of-the-art neural network architecture as a baseline and present several modifications focused on a deeper learning of the cross-modal temporal relationships using attention. By conducting experiments with our proposed approach and the selected baseline model, we observe a relative improvement in performance by 2%. Finally, we conclude that focusing more on the cross-modal temporal relationships may be beneficial for building ER systems for child–machine communications and environments where qualified professionals work with children.

**Keywords:** audio–visual speech; emotion recognition; children

**MSC:** 68T10



**Citation:** Matveev, A.; Matveev, Y.; Frolova, O.; Nikolaev, A.; Lyakso, E. A Neural Network Architecture for Children’s Audio–Visual Emotion Recognition. *Mathematics* **2023**, *11*, 4573. <https://doi.org/10.3390/math11224573>

Academic Editors: Ryumin Dmitry and Ivanko Denis

Received: 9 October 2023

Revised: 28 October 2023

Accepted: 6 November 2023

Published: 7 November 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Emotions play an important role in a person’s life from its very beginning to the end. Understanding emotions becomes indispensable for people’s daily activities, in organizing adaptive behavior and determining the functional state of the organism, in human–computer interaction (HCI), etc. In order to provide natural and user-adaptable interaction, HCI systems need to recognize a person’s emotions automatically. In the last ten to twenty years, improving speech emotion recognition has been seen as a key factor in improving the performance of HCI systems. While most research has focused on emotion recognition in adult speech [1,2], significantly less research has focused on emotion recognition in children’s speech [3,4]. That is because large corpora of children’s speech, especially audio–visual speech, are still not publicly available, and this forces researchers to focus on emotion recognition in adult speech. Nevertheless, children are potentially the largest class of users of most HCI applications, especially in education and entertainment (edutainment) [5]. Therefore, it is important to understand how emotions are expressed by children and whether they can be automatically recognized.

Creating automatic emotion recognition systems in a person’s speech is not trivial, especially considering the differences in acoustic features for different genders [6], age groups [7], languages [6,8], cultures [9], and developmental [10] features. For example, in [11], it is reported that the accuracies of speech emotion recognition are “93.3%, 89.4%, and 83.3% for male, female and child utterances respectively”. The lower accuracy of

emotion recognition in children's speech may be due to the fact that children interact with the computer differently than adults, as they are still in the process of learning social and conversational interaction linguistic rules. It is highlighted in [12] that the main aim of emotion recognition in conversation (ERC) systems is to correctly identify the emotions in the speakers' utterances during the conversation. ERC helps to understand the emotions and intentions of users and to develop engaging, interactive, and empathetic HCI systems. The input data for a multimodal ERC is information from different modalities for each utterance, such as audio–visual speech and facial expressions, and the model leverages these data to generate accurate predictions of emotions for each utterance. In [13], it was found that in the case of audio–visual recognition of emotions in voice, speech (text), and facial expressions, the facial modality provides recognition of 55% of emotional content, the voice modality provides 38%, and the textual modality provides the remaining 7%. The last is the motivation to use audio–visual speech emotion recognition.

There are few studies on multimodal emotion recognition in children, and even fewer studies have been performed on automatic children's audio–visual emotion recognition. Due to the small size of the available datasets, the main approach was to use traditional machine learning (ML) techniques. The authors of [14] mentioned the following most popular ML-based classifiers: Support Vector Machine, Gaussian Mixture Model, Random Forest, K-Nearest Neighbors, and Artificial Neural Network, with the Support Vector Machine (SVM) classifier being employed in the majority of ML-based affective computing tasks. Recently, there has been a growing focus on automatic methods of emotion recognition in audio–visual speech. This is primarily driven by advancements in machine learning and Deep Learning [15], due to the presence of publicly available datasets of emotional audio–visual speech, and the availability of powerful computing resources [16].

Motivated by these developments, in this study, we have developed a neural network architecture for children's audio–visual emotional recognition. We conducted extensive experiments with our architecture on our proprietary dataset of the children's audio–visual speech.

This study offers the following main contributions:

1. An extended description of the dataset with children's audio–visual emotional speech we collected and a methodology for collecting such datasets is presented.
2. A neural network solution for audio–visual emotion recognition in children is proposed that improves the utilization of temporal relationships between audio and video modalities in cross-modal fusion implemented through attention.
3. The results of experiments on emotion recognition based on the proposed neural network architecture and the proprietary children's audio–visual emotional dataset are presented.

The subsequent sections of this paper are organized as follows. We analyze common datasets and algorithms for multimodal children's emotion recognition in Section 2. In Section 3, we present a description of the dataset we collected specifically for our purposes. We demonstrate the algorithms and the model we propose for solving the problem in Section 4. In Section 5, we describe the experiments with our data and algorithms; and in Section 6, we present the results of the experiments. Lastly, Section 7 summarizes the contributions of this article and formulates the directions for future research on multimodal children's emotion recognition.

## 2. Related Work

Several Audio–Visual Emotion Recognition (AVER) systems for adults have been discussed in the literature over the past decade. There is plenty of literature and critical analysis available on four key topics in traditional AVER: databases [17], features, classifiers, and data fusion strategies [18–21]. The majority of the common deep learning methods used in AVER are reviewed in [22]. The authors of [15] experimented with various combinations of CNN, LSTM, and SVM models in application with diverse audio–visual data for

emotion recognition, and found that the audio modality can significantly contribute to the performance of a model.

The authors of [23] introduced an AVER solution which employs training a CNN for facial expression recognition and a VGG-based feature extractor for speech emotion recognition (for audio represented in mel-spectrogram images). They then suggest to match the dimensions of the output feature tensors of both video and audio feature extractors via  $1 \times 1$  convolutions and to concatenate the results. These joint representations are then fed into an LSTM network. This approach demonstrates high performance in learning efficient representations of facial expressions.

However, there are few articles on children’s emotion recognition using single speech or facial modalities, and even fewer on children’s audio–visual emotion recognition, which is due to the lack of available children’s audio–visual emotional datasets.

Next, we consider the most relevant modern research on AVER, with a focus on children’s audio–visual speech. We pay special attention to children’s audio–visual speech emotion corpora and those approaches that use state-of-the-art machine learning and deep machine learning methods.

### 2.1. Children’s Audio–Visual Speech Emotion Corpora

Despite the difficulties in obtaining emotions data, there are corpora of children’s emotional speech in different languages [24–27] and emotional facial expressions [28] of children. Research is being conducted on the automatic recognition of emotions from children’s speech [29,30] and their facial expressions [31]. The accuracy of emotion recognition can be higher when using several modalities [32], for example audio and video, which requires the collection of appropriate audio–visual corpora.

A brief description of the available datasets of children’s audio–visual emotion speech is presented in Table 1 and in more detail below.

**Table 1.** Characteristics of multimodal corpora of children’s audio–visual emotions.

| Corpus                                   | Modality | Volume                      | Language           | Subjects          | Age Groups, Years |
|--|----------|-----------------------------|--------------------|-------------------|-------------------|
| AusKidTalk [33]                          | AV       | 600 h                       | Australian English | 700 TD;<br>25 ASD | 3–12              |
| AFEW-VA [34]                             | AV       | 600 clips                   | English            | 240 TD            | 8–70              |
| CHIMP [35]                               | AV       | 8 video files for<br>10 min | English            | 50 TD             | 4–6               |
| EmoReact [36]                            | AV       | 1102 clips                  | English            | 63 TD             | 4–14              |
| CHEA VD [37]                             | AV       | 8 h                         | Chinese            | 8 TD              | 5–16              |
| Dataset featuring children with ASD [32] | AVTPh    | 18 h                        | Irish              | 12 HFASD          | 8–12              |

Note: AV—audio–visual; AVTPh—audio, video, text, physiological signals (heart rate measure); TD—typical development; ASD—autism spectrum disorder; HFASD—high-functioning autism.

AusKidTalk (Australian children’s speech corpus) [33]—audio and video recordings of game exercises for 750 children aged three to twelve who speak Australian English. The study participants were 700 children with typical development and 50 children with speech disorders—25 children aged 6–12 years have a diagnosis of autism spectrum disorder. For each child, there are records that are made in a structured session 90–120 min. Speech is collected in a variety of children’s activities designed to reflect the diversity of use in children’s communications and different levels of speech skills. Video recording of the entire session is used to support manual annotation of children’s speech.

AFEW-VA database [34] of 600 real-world videos with accurate annotations of valence and arousal on 30,000 frames.

CHIMP (Little Children’s Interactive Multimedia Project) [35] dataset collected by the Signal Analysis and Interpretation Laboratory at the University of Southern California in

2005. The dataset contains recordings of 50 children aged four to six interacting with a fictional controlled character in conversational interactive games.

EmoReact (Multimodal Dataset for Recognizing Emotional Responses in Children) [36] is a dataset with 1102 audio–visual recordings of 63 children aged four to fourteen, and annotated into 17 emotion categories. To date, this is one of the few datasets containing both verbal and visual expressions of emotions by children [38].

CHEA VD (Chinese Natural Emotional Audio–Visual Database) [37] is a large-scale Chinese audio–visual corpus of natural emotions, with different age groups and emotional states, including eight hours of recordings of children aged five to sixteen.

In [32], the authors presented their plans for creating a multimodal emotion recognition system for children with high-functioning autism, where the goal is to develop human–machine interfaces specialized for children with autism. This study involved 12 children aged 8 to 12 years. Recording is carried out in three sessions of 30 min each. The total time is 18 h. The proposed system is intended to work with video, audio, text and physiological (heart rate) modalities.

It should be noted that most of the described datasets are for the English, with some for the Chinese and Irish languages but none for the Russian.

## 2.2. Audio–Visual Emotion Recognition

Emotion recognition can be formulated as a problem where some source produces several streams of data (features) of various modalities (e.g., audio and video), each with its own distribution, and the goal is to estimate the distributions and map them onto the source. That, naturally, poses several questions that ought to be answered when building an emotion recognition system: which modalities are selected and represented, how the modalities are mapped on each other, and how the joint representations are mapped onto the sources of the distributions. We will review research that answers those questions and then propose our solution.

It has been shown that regardless of the model and representations, multimodal approaches virtually always outperform unimodal ones [39], i.e., adding another modality can only benefit the performance. While this may seem obvious, the notion actually relies on the fact that, in the worst-case scenario, a model is able to learn an identity mapping for the driving modality and disregard the other one. However, as has been shown in practice, it is rarely the case that additional modalities carry no valuable information. As for the selection of modalities, the most common ones in the literature are images (or sequences of images, i.e., video), audio, and text. Since our research is focused on children, including pre-school children and children with developmental disorders, the contribution of textual modality, as already noted in Section 1, is insignificant. Therefore, we pick video and audio as our modalities of choice.

Representation is one of the key concepts in machine learning [40]. While the task of machine learning imposes a number of limitations on the representations of data, such as smoothness, temporal and spatial coherence, over the years, a bevy of various representations have been used to solve various machine learning problems, and while some are more common than the other, there is no clear rule for choosing the best representation. Traditional machine learning algorithms rely on the representation of the input being a feature and learn a classifier on top of that [41]. Meanwhile, the most agile modern models attempt to learn not only the representations but also the architecture and the hyperparameters of the model [42]. Both extremes, however, have several issues. The traditional approach lacks the capability to discover deep, latent features and is mostly unable to achieve high efficiency associated with learning hierarchical and spatial-temporal relationships within feature sets, and since there is no space to learn cross-modal relationships, multimodal models either rely on some sort of decision-level fusion or expert heuristics for joint representations. The end-to-end approach, on the other hand, has a high computational cost and requires a precise, structured approach to training [43]. With those limitations, most of the modern models take reasonably preprocessed input data, then attempt to learn their

efficient representations, including joint representations, and finally learn to classify those representations.

There are several ways to present audio data to a model. The most common include [22]:

- Waveform/raw audio, seldom used outside of end-to-end models, is simply raw data, meaning the model has to learn efficient representations from scratch;
- Acoustic features such as energy, pitch, loudness, zero-crossing rate, often utilized in traditional models, while allowing for simple and compact models, are mostly independent by design and prevent a model from learning additional latent features;
- A spectrogram or a mel-spectrogram, which shares some similar issues with raw audio, has found its way into many models due to extensive research into convolutional neural networks, since, being presented as an image, it enables learning efficient representations as shown in various practical applications;
- Mel-Frequency Cepstral Coefficients, which represent the short-term power spectrum of a sound—very commonly used as they provide a compact but informative representation.

In [44], a relatively recent example of representation learning was proposed—a large-scale self-supervised pre-trained WavLM model for speech processing. This model, which is a transformer encoder, efficiently encodes audio features for classification and is trained on a large dataset. The frozen encoder can then be utilized as a feature extractor for general purpose speech processing.

For image processing, the traditional approaches are extremely computationally expensive. For example, when a raw image is processed through a fully connected neural network, the network has to treat each pixel as an individual input and learn to extract relevant features from all locations within the image. In contrast, a convolutional neural network (CNN) [45] can learn to recognize patterns in an image regardless of where they are located, using shared weights across the entire image and reducing the number of parameters required. By design, CNNs learn hierarchical representations of the raw input data and, due to the shown efficiency of this approach, this is the most common approach for the representation of visual data. However, while a static image is a common input for a variety of computer vision problems, there is also a large field of problems concerned with sequences of images, i.e., video. Since, for most of the practical tasks, there are strong relationships between consecutive frames of the input video. It is natural that efficient representations of those relationships are key for achieving high performance. For example, optical flow is a technique used in computer vision that enables one to recognize and track movement patterns in video footage [46]. Another option to employ an implementation of a recurrent neural network (RNN), for example a long short-term memory (LSTM) network or a convolutional RNN, in which case a network is able to collect global context and produce representations enhanced with those shared latent features [47]. Another relatively recent approach is to implement a 3D CNN [48], where the temporal dimension is added to both the input tensor and the filters. While the idea of considering a sequence of images as just another dimension of the input tensor is relatively natural, the significant increase in the number of weights presents the need for a large amount of training video data and incurs a high computational cost. However, as the CNN architectures for image processing became highly optimized and somewhat larger video datasets have become available, this approach became legitimately viable.

The key concept for multimodal classification is the fusion of modalities. Though earlier models relied on unimodal classification and consecutive ensemble learning for decision-level fusion such as averaging, voting, and weighted sum, it was quickly discovered that both the redundancy of features between modalities and latent cross-modal relationships can be utilized to achieve higher performance [18,20]. Another traditional approach is to implement an early fusion. While some of the works propose the fusion of modalities at the input data level [49], the most common approach is to combine modalities upon feature extraction, relying on some sort of heuristics [18,20]. In modern research,

fusion is applied somewhere between the feature extraction and the decision level with the goal of learning efficient joint representations to both eliminate the redundancy in order to reduce the computational cost, and to align modalities to take advantage of cross-modal relationships.

There are several strategies for this kind of intermediate fusion, but the most common technique is to implement fusion via an attention mechanism [16]. This is a method to focus on the most relevant information from each modality, to determine which parts of each modality's input should be given greater focus when making a prediction, and selecting the most important features from each modality and combining them in a meaningful way. In a more general sense, the attention technique can be understood from the distinction between soft and hard attention. To emulate human perception and reduce computations, ideally, a model should be able to ignore the clutter in the input data and attend only to the meaningful parts [50] sequentially and aggregate information over time—this approach would implement so-called hard attention. However, to achieve that, it would require the model to make choices where to look at and they are difficult to represent as differentiable functions which would be required for the most conventional techniques for training. Requiring a model to be differentiable means that the model is simply able to associate more importance with certain parts of the input data—this approach is called soft attention.

Another informative way to designate attention techniques is to focus on the dimensions across which they are applied. Though some terminology may be used interchangeably in the literature, the more common ones include:

- Channel attention—as channels of feature maps are often considered feature detectors, it attempts to select more relevant features for the task [51];
- Spatial attention—in the cases with multidimensional input data such as images, it attends to inter-spatial relationship of features [52];
- Temporal attention—though the temporal dimension can sometimes be considered simply as another dimension of input data, in practice it might be beneficial to view it separately and apply different logic to it, depending on the task [52];
- Cross-attention—mostly utilized in the cases with multiple modalities to learn relationships between modalities; since different modalities often have different dimensions, the modalities cannot be viewed as just another dimension of the input tensor, thus requiring a different approach from simply increasing the dimension of the attention maps; can be used to combine information from different modalities, in which case it is said to implement the fusion of modalities [53].

The authors of [54] suggested that applying attention along the input dimensions separately achieves lower computational and parameter overhead compared to computing attention maps with the same dimensions as the input. The authors of [55] proposed the “Squeeze-and-Excitation” block, an architectural unit that explicitly models interdependencies between channels and recalibrates feature maps channel-wise. The authors of [56] presented a self-attention mechanism for CNN to capture long-range interactions between features, which, in modern research, is mostly applied to sequence modeling and generative modeling tasks, they show that they can improve the performance of a model by increasing the number of feature maps by concatenating the feature maps with multihead attention maps. The authors of [57] implemented cross-attention for multimodal emotion recognition from audio and text modalities where the features from the audio encoder attend to the features from the text encoder and vice versa to highlight the most relevant features for emotion recognition. Though the features from those two modalities are eventually concatenated before passing them to the classifier, the attention block does not explicitly implement a fusion of modalities and is rather an example of late fusion. The authors of [58] proposed a universal split-attention block for the fusion of modalities where the attention block explicitly fuses features from different modalities and can be both placed at an arbitrary stage of a network and repeated multiple times across the network. In this paper, we consider that model as a baseline for comparison with the proposed alternative cross-modal fusion block.

After the feature maps are generated by a network, the final step is to classify the sample into one of the target categories. The most common approach is to map the feature maps onto scalar values (flatten the feature maps) and present the output as a scalar vector so that it can be presented to a fully connected network which is trained to classify the input into one of the target categories, usually by a SoftMax layer with the number of neurons equal to the number of target classes [41]. Even though this approach is utilized in most of the modern models, flattening of the feature maps effectively discards the spatial and temporal relationships. To investigate some of those issues, the authors of [59] suggested generating so-called “class activation maps”, where the class activation map points to the segments of the input image which the network considers discriminative to detect the target class. Since the outcome of this procedure can encapsulate the spatial and temporal relationships between the input and the feature maps, this information can also be employed for classification. In this paper, we demonstrate one such approach.

### 3. Corpus Description

To study children’s audio–visual emotion recognition, an audio–visual emotional corpus was collected. The corpus contains video files with emotional speech and facial expressions of Russian-speaking children.

#### 3.1. Place and Equipment for Audio–Visual Speech Recording

The recording sessions were held in a laboratory environment without soundproofing and with regular noise levels. A PMD660 digital recorder (Marantz Professional, inMusic, Inc., Sagamihara, Japan) with a SENNHEIZER e835S external microphone was used to capture a 48 kHz mono audio signal, and a SONY HDR-CX560 video camera (Sennheiser electronic GmbH & Co. KG, Beijing, China) was used to record a child’s face from a distance of one meter in 1080p resolution at 50 frames per second. During testing, the child sat at the table opposite the experimenter. The light level was constant throughout the recording session.

#### 3.2. The Audio–Visual Speech Recording Procedure

Recording of speech and facial expressions of children was carried out when testing children according to the Child’s Emotional Development Method [60], which includes two blocks. Block 1 contains information about the child’s development received from parents/legal representatives. Block 2 includes tests and tasks the purpose of which is the evaluation of expression of the emotions in the child’s behavior, speech, and facial expressions, and ability of the child to perceive the emotional states in others. Each session lasted between 60 and 90 min.

Participants in this study were 30 children aged 5–11 years.

The criteria to include children in this study were:

1. The consent of the parent/legal representative and the child to participate in this study.
2. The selected age range.
3. The absence of clinically pronounced mental health problems, according to the medical conclusion.
4. The absence of verified severe visual and hearing impairments.

The parents were consulted about the aim and the procedure of this study before signing the Informed Consent. Also, the parents were asked to describe in writing the current and the overall emotional development of their child.

The experimental study began with a short conversation with the children in order to introduce the experimenter to the child. The child then completed the following tasks: playing with a standard set of toys, co-op play, “acting play” when the child is asked to show (depict) the emotions “joy, sadness, neutral (calm state) anger, fear”; should pronounce the speech material, manifesting the emotional state in voice; video tests—for emotions recognition, standard pictures containing certain plots.

All procedures were approved by the Health and Human Research Ethics Committee (HHS, IRB 00003875, St. Petersburg State University) and written informed consent was obtained from parents of the child participant.

### 3.3. Audio–Visual Speech Data Annotation

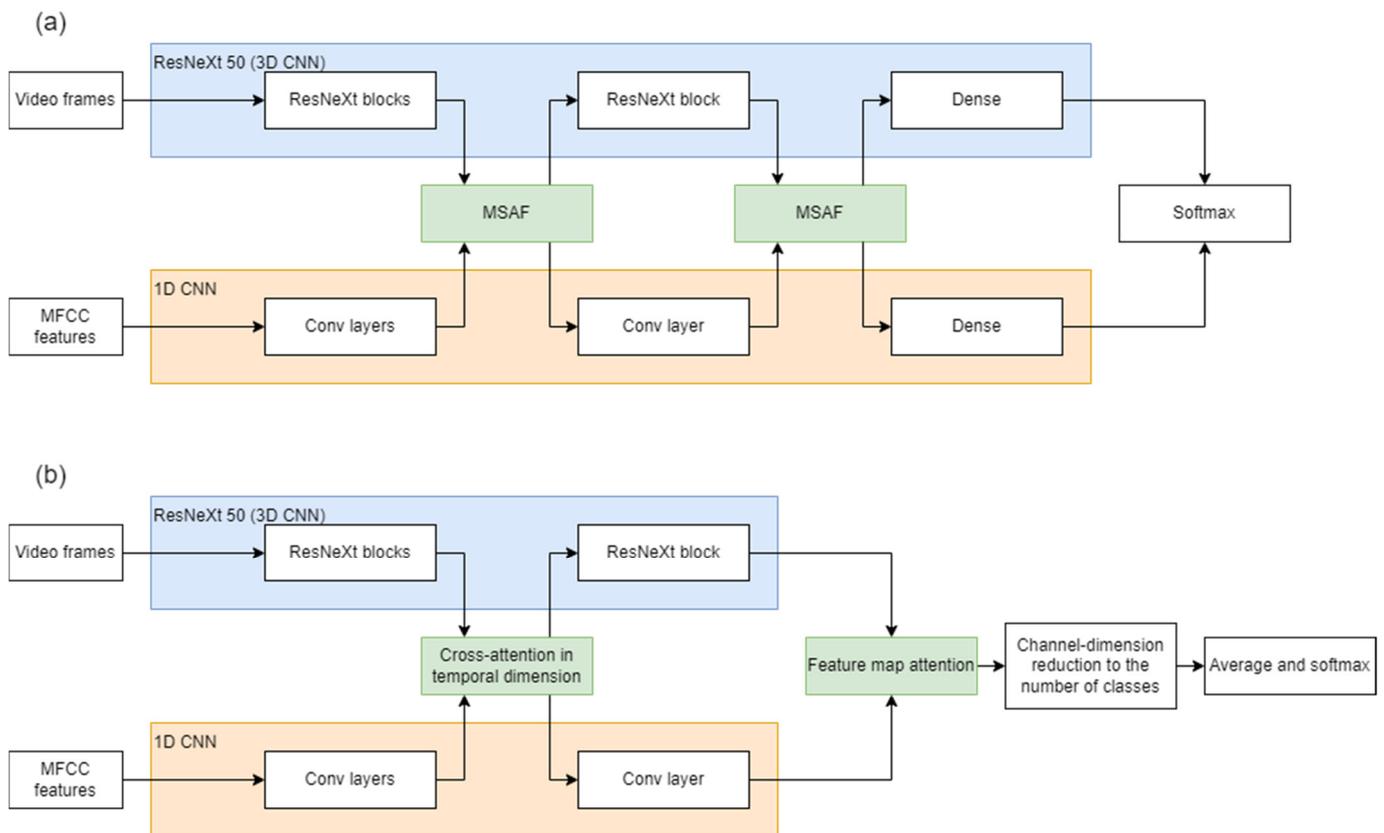
Specifically, for training the neural network based on our approach with 3D CNN, we have prepared an annotated dataset that contains relatively short video segments with audio. First, we performed facial landmark detection across the whole video dataset and automatically selected the segments with continuous streams of video frames with fully visible faces (as per the data collection procedure, most of the frames with fully visible faces belong to a child being recorded). Further, we applied speaker diarization and selected the segments in which continuous streams of video frames with fully visible faces overlap with continuous speech. Next, a group of 3 experts reviewed the obtained video segments to either annotate them with emotions expressed by a child, or to annotate the segment with additional timestamps when across the video segment a child expresses different emotions at different times. If the face or speech of a non-target person appears in the recording, experts should reject the segment. A segment receives a label only if all experts agree with the expressed emotion, otherwise the segment is rejected. Once the annotation process was complete, the annotations were used to filter the dataset and further categorize the video segments by expressed emotion where appropriate. Finally, we randomly split the segments into subsegments of 30 frames in length, which were then used to train the neural network.

## 4. A Neural Network Architecture Description

To classify children’s emotions, we propose a neural network based on 3D CNN for video processing and 1D CNN for audio processing. To demonstrate the performance of our solution, we took as the baseline the architecture from [58], as that solution has shown a state-of-the-art performance for the target problem. Note, however, that in [58], the authors propose a modality fusion block while utilizing existing approaches for video and audio processing to demonstrate the performance of their proposed solution for several machine learning problems, including emotion detection. Similarly, in this manuscript we do not discuss in detail the underlying models and refer the reader to the original article [58]. Our goal here is to demonstrate that, by optimizing the attention component of the model to the particularities of the source data, we can improve the performance of the emotion classification for children’s speech.

Per the research on children’s speech, some of which is reviewed in Section 1, the temporal alignment of video and audio modalities is highly informative for detecting emotions in children’s speech. Furthermore, research seems to indicate that this temporal alignment may depend not only on the psychophysiological aspects of children in general, but may also differ for typically and atypically developing children, and, moreover, for different types of atypical development. This naturally provides for an assumption that by increasing the focus and granularity of modeling the inter-modal temporal relationships may result in an improved performance of a model. To address this problem, we propose a modification of the cross-attention fusion module introduced in [58], followed by a classifier inspired by [59], based on the application of “Squeeze-and-Excitation”-like attention [55] to the feature maps of the final layer for a classification. This preserves more spatial relationships than the traditional approach of flattening the feature maps and attaching a fully connected network.

For a comparison between the baseline and the suggested in this paper architectures, see Figure 1.



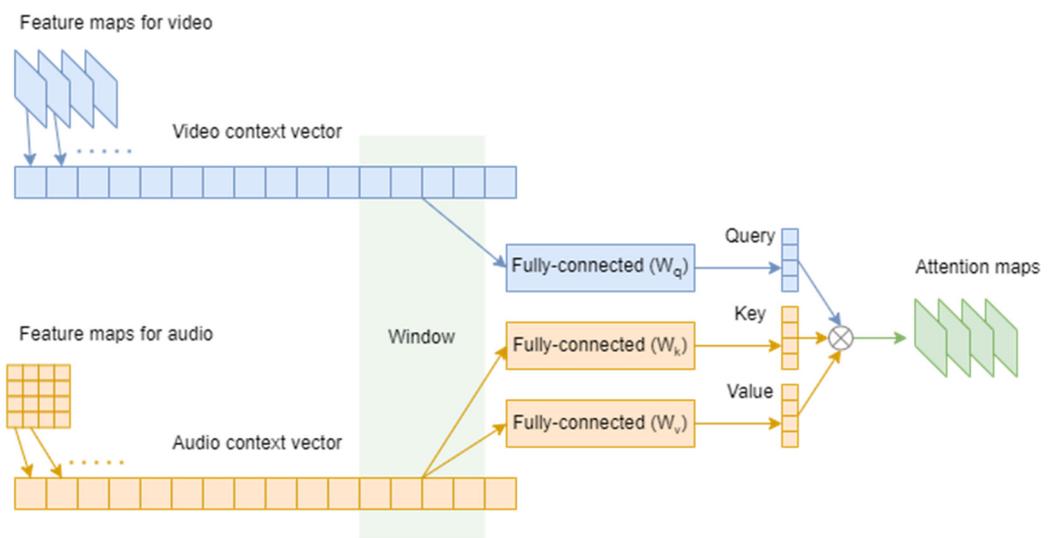
**Figure 1.** An overview of the baseline (a) architecture [58], where MSAF refers to the suggested Multimodal Split Attention Fusion and the suggested architecture (b). The blocks highlighted with green signify the implementations of the multimodal fusion over the base models for video and audio processing.

Let us underscore a couple of differences between the proposed and the baseline models. First, in this paper, we present a different implementation for the fusion block, in which the fusion is performed in a window and using the query-key-value approach to calculate attention. Second, in the baseline model, the fusion block is placed at two locations, while in our model, we found that a single block is sufficient. However, it is important to highlight that neither we nor the authors of the baseline model require a specific placement of the fusion block. Both consider the fusion block as a black box or, in a sense, a layer that can be placed at arbitrary positions and an arbitrary number of times, depending on various circumstances such as a choice of the baseline models for video and audio processing. Third, in our work, we propose a different approach to classification. Instead of the traditional flattening of feature maps with the dense layer, we deploy an attention layer to transform feature maps into class maps matching the number of target classes.

#### 4.1. An Algorithm for Multimodal Attention Fusion

Following [58], we do not assume a specific placement of the attention block in the architecture; essentially, we only consider the attention block in the context of a neural network architecture as a black box with feature maps in—feature maps out. Briefly (for a more detailed explanation we direct the reader to [58]), the cross-attention fusion module for video and audio modalities takes feature maps  $F = \{F_v, F_a\}$ , where  $F_v$  are feature maps for the video modality and  $F_a$  are feature maps for the audio modality and, as an input and produces modified feature maps  $F' = \{F'_v, F'_a\}$  with the goal of enhancing the quality of representations of features of each modality by attending to them according to the information learned from another modality. As a side note, here we do not make an explicit

distinction between the sets of feature maps and the blocks of sets of feature maps where the notion of blocks appears from the concept of cardinality in the ResNeXt architecture, which refers to an additional dimension to the data passing through a network. Both our approach and the approach in [58] are essentially agnostic to this distinction in the sense that both simply operate on vectors containing feature maps. To calculate the modified feature maps, first each modality must be mapped to a space with only a temporal dimension, which for our task simply means that the spatial dimensions of the video modality are collapsed into a scalar by global average pooling. After obtaining the channel descriptors, a commonly called global context descriptor has to be formed as the source of the information about cross-modal relationships. Here, we propose the following approach: to capture the more immediate relationships between the modalities, we calculate the query, key, and value [61] in a window of length  $S$  for the context vectors  $F_v^c$  and  $F_a^c$  for video and audio modalities, respectively (see Figure 2).



**Figure 2.** Schema to calculate the query, key, and value in a window of length  $S$  for the context vectors  $F_v^c$  and  $F_a^c$  for video and audio modalities, respectively.

Since this approach originally appeared in the context of natural language processing and is often explained in terms of that field, here we want to provide a brief intuition for applying this approach in more general terms. In the case of one modality, the goal is to find relationships between different sections of a feature map of that modality. For additional clarity, when we consider a video, i.e., a modality with both spatial and temporal dimensions, we can consider either the self-attention within a single image, where sections are represented as regions of pixels in the image, and the self-attention within the temporal dimension obtained by collapsing the spatial dimensions of a series of images. The “query, key, and value” approach is agnostic to whichever one we choose.

In this article, we are always talking about the attention in the temporal dimension. To achieve that, each section is mapped to three different vectors: “query”—functioning as a request, “value”—functioning as a response, and “key”—a map between queries and values. Nevertheless, it is important to understand that attributing a function or role to those vectors serves mostly for the purposes of human understanding, while from a purely technical standpoint, the procedure is implemented simply through tripling a fully connected layer and then another layer joining the outputs together.

Let us call the learnable transformations for the “query”, “key”, and “value” ( $\bar{q}$ ,  $\bar{k}$ , and  $\bar{v}$ ) vectors  $T_Q$ ,  $T_K$ , and  $T_V$ , respectively. Then, for the context vectors  $F_v^c$  and  $F_a^c$  for video and audio modalities, and for their windowed segments  $F_v^{c, S_i}$  and  $F_a^{c, S_i}$ , we calculate:

$$\bar{q} = T_Q(F_v^{c, S_i}), \quad \bar{k} = T_K(F_a^{c, S_i}), \quad \bar{v} = T_V(F_a^{c, S_i}) \tag{1}$$

While the dimensions of the value vectors are not required to match the dimensions of the query and key vectors, unless there is a specific reason to choose otherwise, most commonly the dimensions do match, for simplicity. We follow this approach, so  $\bar{q}, \bar{k}, \bar{v} \in \mathbb{R}^D$ . Strictly speaking, the key vectors do not provide a one-to-one mapping between queries and values, instead, they encapsulate the likelihood or the strength of the relationship between each query and each value. Also, since we consider each segment of each windowed context vector to be independent, we are only interested in the relative likelihood, which we, following the common approach, implement using *softmax*.

So, for each query  $\bar{q}_l$ , we calculate:

$$\text{softmax}\left(\left[\bar{q}_l, \bar{k}_m\right]\right) \text{ for each key } \bar{k}_m, \quad l, m \in 1, \dots, M \tag{2}$$

or, in matrix form:

$$\text{softmax}\left([\bar{q}_1, \dots, \bar{q}_M][\bar{k}_1, \dots, \bar{k}_M]^T\right). \tag{3}$$

This result, in some sense, is a heatmap, showing the strength of the relationships between queries and values.

Now, at this point, we still have to construct a function that would take this heatmap and the values, and produce a new set of feature maps, and while in principle this function can also be learned. It has been shown that a simple weighted average provides a good balance between the performance and the computational resources required, since it can, again, be calculated as a straightforward matrix multiplication.

Summarizing the algorithm, we can present the equation for joining the outputs (the attention) as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where  $\sqrt{d_k}$  is a simple scaler.

As for the learnable transformations of the query, key, and value for multiple modalities, in our case we obtain them via projection of the windowed segments of the context vectors  $v_s$  and  $a_s$  for video and audio modalities, respectively, with learnable parameters  $w_q, w_k$ , and  $w_v$ :

$$q = w_q v_s, \quad k = w_k a_s, \quad v = w_v a_s \tag{5}$$

After obtaining the attention maps (4), we can calculate the new feature maps:

$$F' = \{F'_V, F'_A\} = \{F_V \odot A_V, F_A \odot A_A\}. \tag{6}$$

Here, just as we do not distinguish between feature maps and sets of feature maps, we also can view our suggested windowed attention as adding another dimension to a collection of feature maps which we can simply flatten when necessary, e.g., when passing them to a classifier.

#### 4.2. An Algorithm for Feature-Map-Based Classification

Regarding the classifier, inspired by the concept of class activation maps in [59], we propose the following intuition first: with  $N$  feature maps at the final layer, our goal is to obtain  $C$  feature maps, each representing the category we are attempting to detect. To realize this transformation, we propose to apply the ‘‘Squeeze-and-Excitation’’-type attention [55]  $C$  number times each with different learnable parameters assuming that this procedure would allow to learn the relationships between the low-level feature descriptors, represented by the feature maps of the final layer, relevant to each target class separately. This way, after applying *softmax* to the globally average pooled class maps, we are expecting to obtain a probability distribution for the target classes.

Comparing to [55], we omit the initial transformation step for the feature maps, as we assume the feature maps at the final layer already represent low-level features and do not require additional transformations for spatial information. So, for each of the  $C$  class

maps, we perform global average pooling, followed by the excitation operation (see [55], Section 3.2):

$$s = \sigma(W_2 \delta(W_1 z)), \quad (7)$$

where  $\sigma$  is a sigmoid function,  $\delta$  is ReLU,  $W_{1,2}$  are learnable parameters also implementing a bottleneck with a dimensionality reduction–expansion hyperparameter, and  $s$  is the vector further used to scale the channels of the feature volume  $\hat{F}'_i = \hat{F}_i * s_i$ .

The final output of the model is then:

$$R = \text{softmax}\left(\text{GAP}\left(\left[\hat{F}^{1\dots C}\right]\right)\right). \quad (8)$$

## 5. Experimental Setup

### 5.1. Software Implementation

To achieve higher efficiency in conducting the experiments, we created a software environment based on the Docker Engine 24.0 ([www.docker.com](http://www.docker.com), accessed on 28 October 2023). The aim of this framework was to simplify running the experiments of different machines, conducting ablation studies, and experimenting with image and audio processing models. We employed the PyTorch 2.0 (<https://pytorch.org>, accessed on 28 October 2023) for implementing the components of our model and we followed the SOLID [62] approach to software development to simplify reconfiguration of the model. Then, we created docker configuration scripts which would dynamically pull the necessary source code and downloadable resources such as base models, set up an execution environment and external libraries, and run the experiments. We ran the experiments on two machines with NVIDIA GeForce RTX 3090 Ti GPUs.

### 5.2. Fine-Tuning

Similar to [58], we used the baseline models trained on the Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS) [63], and we further fine-tuned the models with samples from our proprietary dataset of children’s emotional speech.

### 5.3. Performance Measures

For evaluation of the results of the experiments, we selected several common metrics often used for similar tasks. First of all, we collected the multiclass recognition results into confusion matrices and calculated the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) metrics.

Then, we calculated the accuracy, precision, and recall as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (11)$$

respectively.

Additionally, we calculated the F1-scores per class as

$$F1(\text{class}) = \frac{2 \times \text{Precision}(\text{class}) \times \text{Recall}(\text{class})}{\text{Precision}(\text{class}) + \text{Recall}(\text{class})}. \quad (12)$$

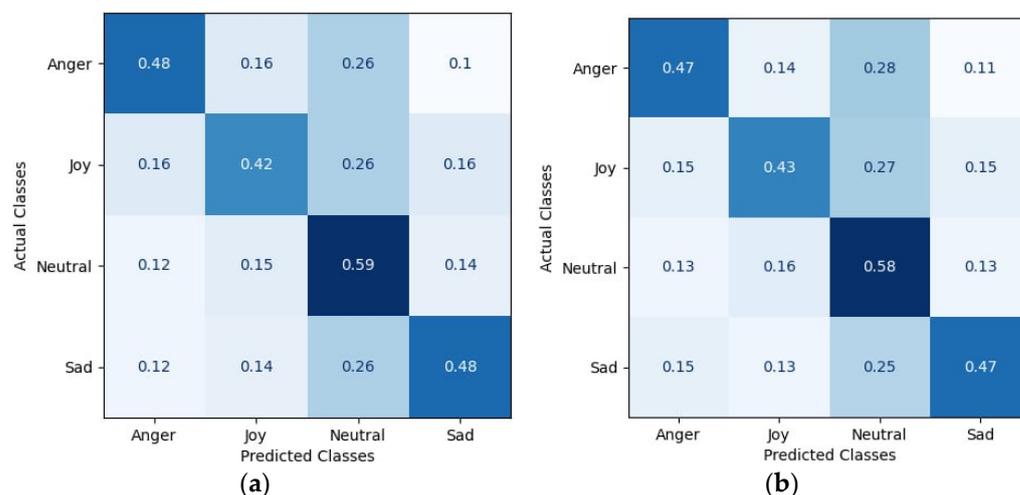
## 6. Experimental Results

From the corpus of child speech, we selected 205 recorded sessions and after processing them as described in Section 3.3 we obtained 721 video segments with variable length, annotated with an expressed emotion. Due to a relatively small volume of data, we randomly extracted 30-frame-length non-intersecting segments, ensuring the balance

between classes and repeated the process 6 times and averaged the results. For each batch, we performed a k-fold cross-validation with 80% of samples used for training and 20% for testing.

In addition, we conducted an ablation study where we tested the fusion block separately from the classifier.

The results of automatic emotion recognition are presented in Figure 3, and Tables 2 and 3.



**Figure 3.** Confusion matrices for both the fusion block and the classifier (a) and for the fusion block only (b). The color shade visualizes the cell value with a darker shade corresponding to a higher value.

**Table 2.** Per-class scores in multiclass classification.

| Per-Class Performance |       |      |         |      |
|-----------------------|-------|------|---------|------|
| Emotion               | Anger | Joy  | Neutral | Sad  |
| Accuracy              | 0.77  | 0.74 | 0.70    | 0.77 |
| Recall                | 0.48  | 0.42 | 0.59    | 0.48 |
| Precision             | 0.54  | 0.48 | 0.43    | 0.54 |
| F1-score              | 0.51  | 0.45 | 0.50    | 0.51 |

**Table 3.** Average scores in multiclass classifications.

| Classifier                | Overall Accuracy |
|---------------------------|------------------|
| Fusion block + classifier | 0.492            |
| Fusion block only         | 0.487            |

Compared with the performance of the state-of-the-art (baseline) model at 0.482, our proposed approach demonstrates a relative improvement in performance by approximately 2%.

### 7. Discussion and Conclusions

We propose a hypothesis that by focusing more on the temporal relationships between different modalities for multimodal automatic emotion recognition in children, we can achieve improvements in performance. Due to the complexity of the problem, in the modern scientific literature, one can find a wide variety of approaches and models. To test our hypothesis, we selected several common and popular approaches that demonstrate state-of-the-art performance on similar tasks and took them as a baseline. Since it is not viable to test fusion and classification modules in isolation, to make sure that the difference in performance between the proposed solution and the baseline model emerges from

the implementation of the proposed solution, it is important to minimize the differences with the baseline neural network architecture. Unfortunately, in machine learning, even repeating the same experiment with the same model and data, it is impossible to produce exactly the same results. However, we attempted our best to utilize the same models and mostly the same training data, except for our novel corpus of children's emotional speech.

As for the implementation of our solution, we focused on the parts of the model responsible for the multimodal fusion via attention. To help the model to focus more on the temporal relationships between different modalities, we proposed to window the context vectors of the modalities, calculate the attention with the query-key-value approach, and perform modality fusion utilizing the obtained attention maps. Additionally, since this approach focuses on the temporal dimension, we also introduced an approach to classification based on the concept of class activation maps that elevates the attention to the spatial dimensions. However, it is important to highlight that our original hypothesis only related to the temporal dimension and, even though, eventually, we observed a cumulative improvement in performance. We did not explicitly test the hypothesis that the proposed approach to classification works as a universal drop-in replacement, we consider it only as an extension of the proposed fusion module.

By evaluating the results of the experiments, we confirmed that with a significant degree of certainty our solution can improve the performance of automatic children's audio-visual emotion recognition. A relatively modest result at approximately 2% performance improvement is nevertheless promising, since there is significant space for further improvement. Our goal here was to demonstrate specific optimizations to the fusion and classification components of the network without optimizations to the overall network architecture, which means that further fine-tuning of the architecture is possible. Our ongoing work on collecting a large dataset of children's audio-visual speech provides us with data to sufficiently improve the fine-tuning of the baseline models and further take advantage of the proposed solution. In addition, since this work only used a dataset with samples where all experts were in agreement with the emotions expressed, a larger dataset with more "difficult" samples with disagreements between the experts would be more helpful for our proposed solution by design. In future research, we plan to focus on collecting more data, particularly, for children with atypical development, and testing our solution on more diverse data. Also, we want to develop more practical tools and applications for people working with children with typical and atypical development to stress-test our solution in a real-time environment.

**Author Contributions:** Conceptualization, A.M., Y.M. and E.L.; methodology, Y.M.; software, A.M.; validation, O.F., E.L. and Y.M.; formal analysis, O.F. and E.L.; investigation, A.M. and Y.M.; resources, O.F., A.N. and E.L.; data curation, O.F., A.N. and E.L.; writing—original draft preparation, Y.M.; writing—review and editing, A.M., Y.M., O.F. and E.L.; visualization, A.M.; supervision, Y.M.; project administration, E.L.; funding acquisition, E.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was financially supported by the Russian Science Foundation, grant number № 22-45-02007, <https://rscf.ru/en/project/22-45-02007/> (accessed on 28 October 2023).

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions on the public dissemination of this data imposed by the informed consent signed by the parents of the minors whose audio-visual data were used in this research.

**Conflicts of Interest:** The authors declare that they have no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

- Schuller, B.W. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
- Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* **2019**, *7*, 117327–117345. [CrossRef]
- Lyakso, E.; Ruban, N.; Frolova, O.; Gorodnyi, V.; Matveev, Y. Approbation of a method for studying the reflection of emotional state in children’s speech and pilot psychophysiological experimental data. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 649–656. [CrossRef]
- Onwujekwe, D. Using Deep Learning-Based Framework for Child Speech Emotion Recognition. Ph.D. Thesis, Virginia Commonwealth University, Richmond, VA, USA, 2021. Available online: <https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=7859&context=etd> (accessed on 20 March 2023).
- Guran, A.-M.; Cojocar, G.-S.; Diosan, L.-S. The Next Generation of Edutainment Applications for Young Children—A Proposal. *Mathematics* **2022**, *10*, 645. [CrossRef]
- Costantini, G.; Parada-Cabaleiro, E.; Casali, D.; Cesarini, V. The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. *Sensors* **2022**, *22*, 2461. [CrossRef] [PubMed]
- Palo, H.K.; Mohanty, M.N.; Chandra, M. Speech Emotion Analysis of Different Age Groups Using Clustering Techniques. *Int. J. Inf. Retr. Res.* **2018**, *8*, 69–85. [CrossRef]
- Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. *Electronics* **2020**, *9*, 1725. [CrossRef]
- Lyakso, E.; Ruban, N.; Frolova, O.; Mekala, M.A. The children’s emotional speech recognition by adults: Cross-cultural study on Russian and Tamil language. *PLoS ONE* **2023**, *18*, e0272837. [CrossRef]
- Matveev, Y.; Matveev, A.; Frolova, O.; Lyakso, E. Automatic Recognition of the Psychoneurological State of Children: Autism Spectrum Disorders, Down Syndrome, Typical Development. *Lect. Notes Comput. Sci.* **2021**, *12997*, 417–425. [CrossRef]
- Duville, M.M.; Alonso-Valerdi, L.M.; Ibarra-Zarate, D.I. Mexican Emotional Speech Database Based on Semantic, Frequency, Familiarity, Concreteness, and Cultural Shaping of Affective Prosody. *Data* **2021**, *6*, 130. [CrossRef]
- Zou, S.H.; Huang, X.; Shen, X.D.; Liu, H. Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. *Knowl.-Based Syst.* **2022**, *258*, 109978. [CrossRef]
- Mehrabian, A.; Ferris, S.R. Inference of attitudes from nonverbal communication in two channels. *J. Consult. Psychol.* **1967**, *31*, 248–252. [CrossRef] [PubMed]
- Afzal, S.; Khan, H.A.; Khan, I.U.; Piran, J.; Lee, J.W. A Comprehensive Survey on Affective Computing; Challenges, Trends, Applications, and Future Directions. *arXiv* **2023**, arXiv:2305.07665v1. [CrossRef]
- Dresvyanskiy, D.; Ryumina, E.; Kaya, H.; Markitantov, M.; Karpov, A.; Minker, W. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. *Multimodal Technol. Interact.* **2022**, *6*, 11. [CrossRef]
- Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83–84*, 19–52. [CrossRef]
- Haamer, R.E.; Rusadze, E.; Lüsi, I.; Ahmed, T.; Escalera, S.; Anbarjafari, G. Review on Emotion Recognition Databases. In *Human-Robot Interaction-Theory and Application*; InTechOpen: London, UK, 2018. [CrossRef]
- Wu, C.; Lin, J.; Wei, W. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, E12. [CrossRef]
- Avots, E.; Sapiński, T.; Bachmann, M.; Kamińska, D. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **2019**, *30*, 975–985. [CrossRef]
- Karani, R.; Desai, S. Review on Multimodal Fusion Techniques for Human Emotion Recognition. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 287–296. [CrossRef]
- Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
- Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249. [CrossRef]
- Schoneveld, L.; Othmani, A.; Abdelkawy, H. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognit. Lett.* **2021**, *146*, 1–7. [CrossRef]
- Ram, C.S.; Ponnusamy, R. Recognising and classify Emotion from the speech of Autism Spectrum Disorder children for Tamil language using Support Vector Machine. *Int. J. Appl. Eng. Res.* **2014**, *9*, 25587–25602.
- Chen, N.F.; Tong, R.; Wee, D.; Lee, P.X.; Ma, B.; Li, H. SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese. In Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH), San Francisco, CA, USA, 8–12 September 2016; pp. 1545–1549. [CrossRef]
- Matin, R.; Valles, D. A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions. In Proceedings of the Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 2–3 October 2020; pp. 1–6. [CrossRef]
- Pérez-Espinosa, H.; Martínez-Miranda, J.; Espinosa-Curiel, I.; Rodríguez-Jacobo, J.; Villaseñor-Pineda, L.; Avila-George, H. IESC-Child: An Interactive Emotional Children’s Speech Corpus. *Comput. Speech Lang.* **2020**, *59*, 55–74. [CrossRef]

28. Egger, H.L.; Pine, D.S.; Nelson, E.; Leibenluft, E.; Ernst, M.; Towbin, K.E.; Angold, A. The NIMH Child Emotional Faces Picture Set (NIMH-CHEPS): A new set of children's facial emotion stimuli. *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 145–156. [CrossRef]
29. Kaya, H.; Ali Salah, A.; Karpov, A.; Frolova, O.; Grigorev, A.; Lyakso, E. Emotion, age, and gender classification in children's speech by humans and machines. *Comput. Speech Lang.* **2017**, *46*, 268–283. [CrossRef]
30. Matveev, Y.; Matveev, A.; Frolova, O.; Lyakso, E.; Ruban, N. Automatic Speech Emotion Recognition of Younger School Age Children. *Mathematics* **2022**, *10*, 2373. [CrossRef]
31. Rathod, M.; Dalvi, C.; Kaur, K.; Patil, S.; Gite, S.; Kamat, P.; Kotecha, K.; Abraham, A.; Gabralla, L.A. Kids' Emotion Recognition Using Various Deep-Learning Models with Explainable AI. *Sensors* **2022**, *22*, 8066. [CrossRef]
32. Sousa, A.; d'Aquin, M.; Zarrouk, M.; Hollowa, J. Person-Independent Multimodal Emotion Detection for Children with High-Functioning Autism. *CEUR Workshop Proceedings*. 2020. Available online: <https://ceur-ws.org/Vol-2760/paper3.pdf> (accessed on 28 October 2023).
33. Ahmed, B.; Ballard, K.J.; Burnham, D.; Sirojan, T.; Mehmood, H.; Estival, D.; Baker, E.; Cox, F.; Arciuli, J.; Benders, T.; et al. AusKidTalk: An Auditory-Visual Corpus of 3- to 12-Year-Old Australian Children's Speech. In Proceedings of the 22th Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 3680–3684. [CrossRef]
34. Kossaiifi, J.; Tzimiropoulos, G.; Todorovic, S.; Pantic, M. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* **2017**, *65*, 23–36. [CrossRef]
35. Black, M.; Chang, J.; Narayanan, S. An Empirical Analysis of User Uncertainty in Problem-Solving Child-Machine Interactions. In Proceedings of the 1st Workshop on Child, Computer, and Interaction Chania (WOCCI), Crete, Greece, 23 October 2008; paper 01. Available online: [https://www.isca-speech.org/archive/pdfs/wocci\\_2008/black08\\_wocci.pdf](https://www.isca-speech.org/archive/pdfs/wocci_2008/black08_wocci.pdf) (accessed on 28 October 2023).
36. Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C.; Morency, L. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI), Tokyo, Japan, 12–16 November 2016; pp. 137–144. [CrossRef]
37. Li, Y.; Tao, J.; Chao, L.; Bao, W.; Liu, Y. CHEAVD: A Chinese natural emotional audio-visual database. *J. Ambient. Intell. Humaniz. Comput.* **2017**, *8*, 913–924. [CrossRef]
38. Filntisis, P.; Efthymiou, N.; Potamianos, G.; Maragos, P. An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications. In Proceedings of the 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 791–795. [CrossRef]
39. Chiara, Z.; Calabrese, B.; Cannataro, M. Emotion Mining: From Unimodal to Multimodal Approaches. *Lect. Notes Comput. Sci.* **2021**, *12339*, 143–158. [CrossRef]
40. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *8*, 1798–1828. [CrossRef] [PubMed]
41. Burkov, A. *The Hundred-Page Machine Learning Book*; Andriy Burkov: Quebec City, QC, Canada, 2019; 141p.
42. Egele, R.; Chang, T.; Sun, Y.; Vishwanath, V.; Balaprakash, P. Parallel Multi-Objective Hyperparameter Optimization with Uniform Normalization and Bounded Objectives. *arXiv* **2023**, arXiv:2309.14936. [CrossRef]
43. Glasmachers, T. Limits of End-to-End Learning. In Proceedings of the Asian Conference on Machine Learning (ACML), Seoul, Republic of Korea, 15–17 November 2017; pp. 17–32. Available online: <https://proceedings.mlr.press/v77/glasml17a/glasml17a.pdf> (accessed on 28 October 2023).
44. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
45. Alexeev, A.; Matveev, Y.; Matveev, A.; Pavlenko, D. Residual Learning for FC Kernels of Convolutional Network. *Lect. Notes Comput. Sci.* **2019**, *11728*, 361–372. [CrossRef]
46. Fischer, P.; Dosovitskiy, A.; Ilg, E.; Häusser, P.; Hazırbaş, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning Optical Flow with Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015; pp. 2758–2766. [CrossRef]
47. Patil, P.; Pawar, V.; Pawar, Y.; Pisal, S. Video Content Classification using Deep Learning. *arXiv* **2021**, arXiv:2111.13813. [CrossRef]
48. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6546–6555. [CrossRef]
49. Ordóñez, F.J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* **2016**, *16*, 115. [CrossRef]
50. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; Volume 2, pp. 2204–2212. Available online: [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf) (accessed on 28 October 2023).
51. Hafiz, A.M.; Parah, S.A.; Bhat, R.U.A. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. *arXiv* **2021**, arXiv:2106.07550. [CrossRef]
52. Bertasius, G.; Wang, H.; Torresani, L. Is Space-Time Attention All You Need for Video Understanding? *arXiv* **2021**, arXiv:2102.05095. [CrossRef]

53. Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; Wu, F. Multi-Modality Cross Attention Network for Image and Sentence Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10938–10947. [[CrossRef](#)]
54. Woo, S.; Park, J.; Lee, J.-L.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Part VII; pp. 3–19. [[CrossRef](#)]
55. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [[CrossRef](#)]
56. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3285–3294. [[CrossRef](#)]
57. Krishna, D.N.; Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 4243–4247. [[CrossRef](#)]
58. Lang, S.; Hu, C.; Li, G.; Cao, D. MSAF: Multimodal Split Attention Fusion. *arXiv* **2021**, arXiv:2012.07175. [[CrossRef](#)]
59. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929. [[CrossRef](#)]
60. Lyakso, E.; Frolova, O.; Kleshnev, E.; Ruban, N.; Mekala, A.M.; Arulalan, K.V. Approbation of the Child’s Emotional Development Method (CEDM). In Proceedings of the Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI), Bengaluru, India, 7–11 November 2022; pp. 201–210. [[CrossRef](#)]
61. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [[CrossRef](#)]
62. Martin, R.C. *Agile Software Development: Principles, Patterns, and Practices*; Alan Apt Series; Pearson Education: London, UK, 2003.
63. Livingstone, S.; Russo, F. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American English. *PLoS ONE* **2018**, *13*, e0196391. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.