mathematics

MDPI

Article

Automatic Speech Emotion Recognition of Younger School Age Children

Yuri Matveev [1,*], Anton Matveev [1], Olga Frolova [1], Elena Lyakso [1] and Nersisson Ruban [2]

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

[2] School of Electrical Engineering, Vellore Institute of Technology, Vellore 632014, India; ruban.ice@gmail.com

* Correspondence: yunmatveev@gmail.com

Abstract: This paper introduces the extended description of a database that contains emotional speech in the Russian language of younger school age (8–12-year-old) children and describes the results of validation of the database based on classical machine learning algorithms, such as Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP). The validation is performed using standard procedures and scenarios of the validation similar to other well-known databases of children's emotional acting speech. Performance evaluation of automatic multiclass recognition on four emotion classes "Neutral (Calm)—Joy—Sadness—Anger" shows the superiority of SVM performance and also MLP performance over the results of perceptual tests. Moreover, the results of automatic recognition on the test dataset which was used in the perceptual test are even better. These results prove that emotions in the database can be reliably recognized both by experts and automatically using classical machine learning algorithms such as SVM and MLP, which can be used as baselines for comparing emotion recognition systems based on more sophisticated modern machine learning methods and deep neural networks. The results also confirm that this database can be a valuable resource for researchers studying affective reactions in speech communication during child-computer interactions in the Russian language and can be used to develop various edutainment, health care, etc. applications.

Keywords: speech emotion recognition; child speech; younger school age

MSC: 68T10

check for updates

Citation: Matveev, Y.; Matveev, A.; Frolova, O.; Lyakso, E.; Ruban, N. Automatic Speech Emotion Recognition of Younger School Age Children. *Mathematics* **2022**, *10*, 2373. https://doi.org/10.3390/math10142373

Academic Editors: Adrian Sergiu Darabant, Diana-Laura Borza and Jüri Majak

Received: 21 May 2022
Accepted: 4 July 2022
Published: 6 July 2022

1. Introduction

Speech Emotion Recognition (SER) systems are currently being intensively used in a broad range of applications, including security, healthcare, videogaming, mobile communications, etc. [1]. The development of SER systems has been a hot topic of research in the field of human-computer interaction for the past two decades. Most of these studies are focused on emotion recognition from adult speech [2–5], and only a few are from children's speech [6–9]. This follows from the fact that sizable datasets of children's speech with annotated emotion labels are still not publicly available in the research community, which leads most researchers to focus on SER for adults.

The reason is that it is difficult to get a wide range of emotional expressions in children's speech due to the following. Children, especially those of primary school age, find it difficult to pronounce and read long texts, and it is more difficult for children than adults to express emotions in acting speech if a child has not experienced these emotions before. School-age children commonly demonstrate a limited range of emotions when interacting with unfamiliar adults due to differences in social status and with regard to training. Generally, children show vivid emotions when interacting with peers, but this situation is usually accompanied by strong noise and overlapping voices. Due to ethical standards, the provocation of vivid negative emotions, such as fear, sadness and anger should be allowed by an informed consent to be signed by the parents. However, not all of them are ready to provide such consent.

However, children are potentially the largest class of users of educational and entertainment applications with speech-based interfaces.

The characteristics of a child's voice are significantly different from those of an adult's voice. The child's voice's clarity and consistency of speech are low compared to an adult [10], while acoustic and linguistic variability in children's speech is high, which creates challenges for emotion recognition. Another severe problem is the paucity of publicly available transcribed linguistic resources for children's speech. Moreover, research on SER faces the problem that most available speech corpora differ from each other in important ways, such as methods of annotation and scenarios of interaction [11–13]. Inconsistencies in these methods and scenarios make it difficult to build SER systems.

All of the above-mentioned problems make the task of developing automatic emotion recognition in children's speech non-trivial, especially taking into account variations of acoustic features within genders, age groups, languages, and developmental characteristics [14]. For example, the authors of [15] report that the average accuracy of emotional classification is 93.3%, 89.4% and 83.3% for male, female, and child utterances, respectively.

Our analysis of databases of children's emotional speech created over the past 10 years has shown that they cover mainly monolingual emotional speech in such languages as English, Spanish, Mexican Spanish, German, Chinese, and Tamil. The only database of children's emotional speech in Russian is EmoChildRu [16], which was used to study manifestations of the emotional states of three- to seven-year-old children in verbal and non-verbal behavior [17].

Children of three to seven years of age were selected for their ability to maintain a communication with an adult while expressing more pure natural emotions than older children. The socialization of children in accordance with cultural norms and traditions mostly begins with schooling which, in the Russian Federation, starts at six to seven years old. This period characterizes the end of the preschool age.

Nevertheless, other databases we have analyzed consist mostly of children's speech of the younger school age group. The particular qualities of the primary school children are, on the one hand, the partial mastery of accepted cultural norms of emotional manifestations; while on the other hand, they represent the presence of natural spontaneous emotional manifestations characteristic of children of this age.

It is highlighted in [18] that the expansion of existing corpora of child-machine communications is necessary for building more intricate emotional models for advanced speech interactive systems designed for children. These new corpora can include other languages and age groups, more subjects, etc. Therefore, our recent works have included the collecting of children's emotional speech of younger school age (8–12-year-olds) in the Russian language, as studies of children's emotion recognition by voice and speech in Russian are sparse, especially for the younger school age group.

There are several key features of our corpus. Firstly, this is the first large-scale corpus dealing with Russian children's speech emotions, which includes around 30 h of speech. Secondly, there are 95 speakers in this corpus. Such a large number of speakers makes it possible to research techniques for speaker-independent emotion recognition and analysis. To validate the dataset [19,20], a subset of 33 recordings was evaluated by 10 Russian experts [21] who conducted manual tests to understand how well humans identify emotional states in the collected dataset of emotional speech of Russian speaking children. Results of manual evaluation validate the database reliability compared with other databases and the ability of native-speaking experts to recognize emotional states in children's speech in the Russian language with above chance performance.

In this article we report the results of experiments on SER conducted using our proprietary database and state-of-the-art machine learning (ML) algorithms. Such experiments are usually conducted to validate the ability to recognize emotional states in children's speech on specific databases in automatic mode with above chance performance. The article presents the results of the experiments and compares them both with results of SER in manual mode and results of SER in automatic mode from other authors for the same age group and the same ML algorithms.

It is noted in [22] that neural networks and Support Vector Machine (SVM) behave differently, and each have advantages and disadvantages for building a SER; however, both are relevant for the task. Multi-Layer Perceptron (MLP) is the "basic" example of a neural network. Therefore, to get a general impression of our dataset, we conducted baseline speech emotion recognition experiments using two of the state-of-the-art machine learning techniques most popular in emotion recognition, these being SVM and MLP [23,24]. These classifiers have already been used to validate several databases of emotional speech [25] and have shown superior classification accuracy on small-to-medium size databases.

The main contributions of this study are as follows:

1. An extended description of the Russian Younger School Children Emotional Speech (SCES-1) database is provided.
2. Validation of the SCES-1 dataset in automatic mode based on SVM and MLP algorithms to recognize above chance emotional states in the speech of Russian children of the younger school age group (8–12-year-old) is undertaken.

The remainder of this paper is structured as follows. In Section 2, the various existing corpora and both features and classifiers used in the literature for SER are discussed. Section 3 provides a detailed description of our proprietary dataset. In Section 4 we define the experimental setup describing our choice of tools to extract feature sets, tools to implement classifiers and evaluate the results of classification and procedures for training and testing. In Section 5, we provide the experimental results. Finally, Section 6 presents the discussion and conclusions, and topics for future research are discussed.

2. Related Work

There are a lot of studies on emotion manifestations in the voice and other biometric modalities. Recently, attention to research on automatic methods of emotion recognition in speech signals has increased due to the development of new efficient methods of machine learning, the availability of open access corpora of emotional speech and high-performance computing resources.

In the last decade, many SER systems have been proposed in the literature. There are numerous publications and reviews on three traditional SER issues: databases, features, and classifiers. Swain et al. [11] reviewed studies on SER systems in the period from 2000 to 2017 with a strong emphasis on databases and feature extraction. However, only traditional machine learning methods were considered as a classification tool, and the authors missed neural networks and deep learning approaches. The authors of [26] covered all the major deep learning techniques used in SER, from DNNs to LSTMs and attention mechanisms.

In this section, we summarize some of the most relevant recent research on SER, focusing on children's speech. In doing so, we pay special attention to those approaches that use the same feature extraction (FE) and machine learning (ML) methods used in this article.

2.1. Children's Emotion Speech Corpora

Over the past three decades, many emotional datasets and corpora have been created in audio or audiovisual modalities [12,26–28]. Some of them are related to natural emotions (spontaneous emotions from real-life situations), while others are related to acted (simulated, for example, by actors) or elicited (induced/stimulated through emotional movies, stories, and games) emotions. Spontaneous emotions are preferable, but they are difficult to collect. Therefore, most of the available emotional speech corpora contain acted emotional speech. Moreover, the emotional corpora are mainly related to adult speech.

Corpora related to children's speech have also been created over the past three decades [29], but they are not so numerous, and only a few of them are related to children's emotional speech. Most of the corpora are in the English language, but some of them are in the German, French, Spanish, Mandarin, Sesotho, Filipino, and Russian languages. Moreover, these corpora vary not only by language but also by children's age range, different specific groups like specific language impairment [30], autism spectrum disorder [31], etc.

We briefly describe the eight most famous corpora with children's emotional speech in Table 1.

- MESD (Mexican Emotional Speech Database) [15,32], presented in 2021. A part of the MESD with children's speech has been uttered by six non-professional actors with mean age of 9.83 years and a standard deviation of 1.17. The database contains 288 recordings of children's speech with six different emotional categories: Anger, Disgust, Fear, Happiness, Sadness, and Neutral. The results of the evaluation of the database reliability based on machine learning analysis showed the accuracy of 83.3% of emotion recognition on children's voices. An SVM was used as a classifier together with a feature set with prosodic, voice quality and spectral features. MESD is considered a valuable resource for healthcare, as it can be used to improve diagnosis and disease characterization.
- IESC-Child (Interactive Emotional Children's Speech Corpus) [18], presented in 2020. A Wizard of Oz (WoZ) setting was used to induce different emotional reactions in children during speech-based interactions with two Lego Mindstorm robots behaving either collaboratively or non-collaboratively. The IESC-Child corpus consists of recordings of the speech spoken in Mexican Spanish by 174 children (80 girls and 94 boys) between six and 11 years of age (8.62 mean, 1.73 standard deviation). The recordings included 34.88 h of induced speech. In total, eight emotional categories are labeled: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral, and None of the above. The research on building acoustic paralinguistic models and speech communication between a child and a machine in Spanish utilized the IESC-Child dataset with prominent results.
- EmoReact (Multimodal Dataset for Recognizing Emotional Responses in Children) [33], presented in 2016. A multimodal spontaneous emotion dataset of 63 (31 males, 32 females) children ranging between four and 14 years old. It was collected by downloading videos of children who react to different subjects such as food, technology, YouTube videos and gaming devices. The dataset contains 1102 audio-visual clips annotated for 17 different emotional categories: six basic emotions, neutral, valence and nine complex emotions.
- CHEAVD (Chinese Natural Emotional Audio–Visual Database) [34], presented in 2016. A large-scale Chinese natural emotional audio–visual corpus with emotional segments extracted from films, TV plays and talk shows. The corpus contains 238 (125 males, 113 females) speakers from six age groups: child (<13), adolescent/mutation (13–16), youth (16–24), young (25–44), quinquagenarian (45–59), and elder ($\geq$60). Over 141 h of spontaneous speech was recorded. In total, 26 non-prototypical emotional states, including the basic six, are labeled by four native speakers.
- EmoChildRu (Child emotional speech corpus in Russian) [16,17], presented in 2015. This contains audio materials from 100 children ranging between three and seven years old. Recordings were organized in three model situations by creating different emotional states for children: playing with a standard set of toys; repetition of words from a toy-parrot in a game store setting; watching a cartoon and retelling the story, respectively. Over 30 h of spontaneous speech were recorded. The utterances were annotated for four emotional categories: Sadness, Anger, Fear, Happiness (Joy). The data presented in the database are important for assessing the emotional development of children with typical maturation and as controls for studying emotions of children with disabilities.
- ASD-DB (Autism Spectrum Disorder Tamil speech emotion database) [35], presented in 2014. This consists of spontaneous speech samples from 25 (13 males, 12 females) children ranging between five and 12 years old with autism spectrum disorder. The

children's voices were recorded in a special school (open space) using a laptop with a video camera. Recordings were taken in wav format at a sampling rate 16,000 Hz and a quantization of 16 bits. The emotion categories included Anger, Neutral, Fear, Happiness, and Sadness. The database was validated using MFCC features and an SVM classifier.

- EmoWisconsin (Emotional Wisconsin Card Sorting Test) [36], presented in 2011. This contains spontaneous, induced and natural Spanish emotional speech of 28 (17 males, 11 females) children ranging between 7 and 13 years old. The collection was recorded in a small room with little noise using two computers, a desktop microphone and a Sigmatel STAC 9200 sound card. Recordings are mono channel, with 16 bit sample size, 44,100 kHz sample rate and stored in WAV Windows PCM format. We recorded 11.38 h of speech in 56 sessions, two sessions per child over seven days. The total number of utterances was 2040, annotated for seven emotional categories: Annoyed, Confident, Doubtful, Motivated, Nervous, Neutral, and Undetermined. For the database validation based on categorical classification we used an SVM algorithm with 10-fold cross validation. The results of the validation showed a performance above chance level comparable to other publicly available databases of affective speech such as VAM and IEMOCAP.
- FAU-AIBO (Friedrich-Alexander-Universität corpus of spontaneous, emotionally colored speech of children interacting with Sony's pet robot Aibo) [37,38], presented in 2006. The corpus was collected from the recordings of children interacting with Sony's pet robot Aibo. It consists of spontaneous, emotionally colored German/English speech from 51 children (21 males, 30 females) ranging between 10 and 13 years old. The total number of utterances is 4525 with a total duration of 8.90 h. The audio was recorded by using a DAT recorder (16-bit, 16 kHz). Five annotators labeled each word individually as Neutral (default) or using one of the other ten emotions: Angry, Bored, Joyful, Surprised, Emphatic, Helpless, Touchy (=Irritated), Motherese, Reprimanding, Rest (non-Neutral, but not belonging to the other categories). The final label was defined using the majority voting procedure. For the database validation based on classification we used a neural network. The results of validation showed that the performance of speech emotion classification on the FAU-AIBO corpora is above chance level.

Table 1. Corpora of children's emotional speech—main features.

Corpus	Subjects	Language	Age Groups
MESD [15,32]	8	Spanish	8–11
IESC-Child [18]	174	Spanish	6–11
EmoReact [13]	63	English	4–14
CHEAVD [34]	3; 5	Chinese	<13; 13–16
EmoChildRu [16,17]	100	Russian	3–7
ASD-DB [35]	25	Tamil	5–12
EmoWisconsin [36]	28	Spanish	7–13
FAU Aibo [37,38]	51	German	10–13

The results of the analysis of the available children's emotion speech corpora show that they are all monolingual. The corpora contain mainly emotional speech of the school age group, with the exception of EmoChildRu (preschool age group) and CHEAVD (adolescence and adult age groups). All of the corpora were validated using manual perceptual tests and/or automatic SVM/MLP classifiers with different feature sets. The results of validation showed that performance of speech emotion classification on all the corpora is well above chance level.

2.2. Speech Emotion Recognition: Features and Classifiers

The classic pattern recognition task can be defined as the classification of patterns (objects of interest) based on information (features) extracted from patterns and/or their representation [39]. As noted in [40], a selection of suitable feature sets, design of proper classification methods and preparation of appropriate datasets are the key concerns of SER systems.

In the feature-based approach to emotion recognition, we assume that there is a set of objectively measurable parameters in speech that reflect the emotional state of a person. The emotion classifier identifies emotion states by identifying correlations between emotions and features. For SER, many feature extraction and machine learning techniques have been extensively developed in recent times.

2.2.1. Features

Selecting the best feature set powerful enough to distinguish between different emotions is a major problem when building SER systems. The power of the selected feature set has to stay stable in the presence of various languages, speaking styles, speaker's characteristics, etc. Many models for predicting the emotional patterns of speech are trained on the basis of three broad categories of speech features: prosody, voice quality, and spectral features [5].

In emotion recognition the following types of features are found to be important:

- Acoustic features of prosodic: pitch, energy, duration.
- Spectral features: MFCC, GFCC, LPCC, PLP, formants.
- Voice quality features: jitter, shimmer, harmonics to noise ratio, normalized amplitude quotient, quasi-open quotient.

Most of the well-known feature sets include pitch, energy and their statistics, which are widely used in expert practice [21].

For automatic emotion recognition, we have used three publicly available feature sets:

- INTERSPEECH Emotion Challenge set (IS09) [41] with 384 features. It was found that the IS09 feature set provides good performance of children's speech emotion recognition. Nevertheless, we conducted some experiments with other feature sets.
- Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). They are popular in the field of affective computing [42]. The GeMAPS family features include not only prosody but also spectral and other features. The total GeMAPS v2.0.0 feature set contains 62 features and the eGeMAPSv02 feature set contains 26 extra parameters, for a total of 88 parameters.
- DisVoice feature set. This is a popular feature set that has shown good performance in such tasks as recognition of emotions and communication capabilities of people with speech disorders and issues such as depression based on speech patterns [43]. The DisVoice feature set includes glottal, phonation, articulation, prosody, phonological, and features representation learning strategies using autoencoders [43]. It is well known that prosodic and temporal characteristics have often been used previously to identify emotions [5]. Therefore, our next experiments were with the DisVoice prosody feature set.

2.2.2. Classifiers

Many different classification methods are used to recognize emotions in speech. There are a lot of publications on using different classifiers in SER [23,25]. The most popular are classical machine learning (ML) algorithms, such as Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Neural Networks (NN), and Multi-Layer Perceptron (MLP) as the "basic" example of NN. Currently, we see the massive application of deep learning (DL) in different fields, including SER. DL techniques that are used to improve SER performance may include Deep Neural Networks

(DNN) of various architectures [26], Generative Adversarial Networks (GAN) [44,45], autoencoders [46,47], Extreme Learning Machines (ELM) [48], multitask learning [49], transfer learning [50], attention mechanisms [26], etc.

The choice of a classification algorithm depends mainly on the properties of the data (e.g., number of classes), as well as the nature and number of features. Taking into account the specifics of children's speech, we must choose a classification algorithm that could be effective in multidimensional spaces on a relatively small-to-medium database size and with low sensitivity to outlier values.

The analysis of the literature shows that the best candidates are SVM and MLP classifiers, where SVM is a deterministic and MLP is a non-deterministic supervised learning algorithm for classification. SVM and MLP are baseline classifiers in numerous studies on emotion speech recognition.

First of all, traditional ML algorithms could provide reliable results in the case of small-to-medium size of training data [51–53]. Second, SVM and MLP classifiers often demonstrate better performance than others [22,23,54]. Some experiments have shown a supercity of SVM and MLP for emotion recognition over classical Random Forest, K-NN, etc. classifiers [54] and also over different deep neural network classifiers [55]. The failure of neural network classifiers in emotion recognition can be explained by the small size of available emotional databases, which are insufficient for training deep neural networks.

We chose to work with an SVM model to evaluate the validity of an emotional database, as it is capable of producing meaningful results with small-to-medium datasets (unlike algorithms such as deep neural networks) and its ability for handling high-dimensional spaces [1]. In [55], it was shown that the SVM classifier outperforms Back Propagation Neural Networks, Extreme Learning Machine and Probabilistic Neural Networks by 11 to 15 percent, reaching 92.4% overall accuracy. In [56] it was also shown that SVM works systematically better than Logistic Regression and LSTM classifiers on the IS09 Emotion dataset.

MLP is the "basic" example of NN. It is stated in [22] that MLP is the most effective speech emotion classifier, with accuracies higher than 90% for single-language approaches, followed closely by SVM. The results show that MLP outperforms SVM in overall emotion classification performance, and even though SVM training is faster compared to MLP, the ultimate accuracy of MLP is higher than that of SVM [57]. SVM has a lower error rate than other algorithms, but it is inefficient if the dataset has some noise [57].

The use of SVM and MLP classifiers allows us to compare the performance of emotion recognition on our database with the performance of the same classifiers on other known databases of children's emotional speech.

Moreover, the results of classification by SVM and MLP can be used as a baseline point for comparison with other models, such as deep neural networks (CNN, RNN, etc.), to see if they provide better performance for SER.

3. Corpus Description

To study emotional speech recognition of Russian typically developing children aged 8–12 years by humans and machines, a language-specific corpus of emotional speech was collected. The corpus contains records of spontaneous speech and acting speech of 95 children with the following information about children: age, place and date of birth, data on hearing thresholds (obtained using automatic tonal audiometry), phonemic hearing and videos with facial expressions and the behavior of children.

3.1. Place and Equipment for Speech Recording

The place for speech recording was the room without special soundproofing.

Recordings of children's speech were made using a Handheld Digital Audio Recorder PMD660 (Marantz Professional, inMusic, Inc., Sagamihara, Japan) with an external handheld cardioid dynamic microphone Sennheiser e835S (Sennheiser electronic GmbH & Co. KG, Beijing, China). In parallel with the speech recording, the behavior and facial expressions of children were recorded using a video camera Sony HDR-CX560E (SONY Corporation, Tokyo, Japan). Video recording was conducted as part of studies on the manifestation of emotions in children's speech [16] and facial expressions [58].

The distance from the child's face to the microphone was in the range of 30–50 cm. All children's speech and video recordings were included in the corpus. All speech files were saved in .wav format, 48,000 Hz, 16 bits per sample. For each child, the recording time was in the range of 30–40 min.

3.2. Speech Recording Procedure

Two types of speech were recorded—natural (spontaneous) speech and acting speech.

Dialogue between the children and the experimenters were used to obtain recordings of the children's natural speech. We hypothesized that semantically different questions could induce different emotional states in children [26]. A standard set of experimenter's questions addressed to the child was used. The experimenter began the dialogue with a request for the child to tell his/her name and age. Further questions included:

- What lessons at school do you like and why?
- Who are your friends at school?
- What do you play at breaks between lessons?
- What are your hobbies?
- What movies and cartoons do you like to watch?
- Can you remember are funny stories?
- Which teachers you don't like and why?
- Do you often fight with other children? Are you swearing?
- Do you often get angry? What makes you angry?
- Do you often get sad? What makes you sad?

Children's acting speech was recorded after dialogue recording sessions. The children had to pronounce speech material—sets of words, words and phrases, as well as meaningless texts demonstrating different emotional states "Neutral (Calm)—Joy—Sadness—Anger". Children were trained to pronounce sets of words, words and phrases, as well as meaningless texts. Before recording the acting speech, each child was asked to pronounce the proposed speech material two to three times depending on the age of the child. Children eight to nine years of age had difficulty pronouncing meaningless texts, so they were instructed to practice to fluently read the whole meaningless text out, rather than articulating individual words.

Two experts estimated how well the children could express different emotional states in their vocal expressions during the training and speech recording sessions. The experts asked a child to repeat the text if the child was distracted, showed an emotion that did not correspond to the given task, or made mistakes in the pronunciation of words.

The selection of this list of emotions was based on the assumption [40] that there is no generally acceptable list of basic emotions, and even though the lists of basic emotions differ, Happiness (Joy, Enjoyment), Sadness, Anger, and Fear appear in most of them. In [59] it was also noted that categories such as Happiness, Sadness, Anger, and Fear are generally recognized with accuracy rates well above chance. In [34] it is noted that regarding the emotion category, there is no unified basic emotion set. Most emotional databases address prototypical emotion categories, such as Happiness, Anger and Sadness.

A number of emotion recognition applications use three categories of speech: emotional speech with positive and negative valence, and neutral speech. In practice, collecting and labeling a dataset with such three speech categories is quite easy, since it can be done by naive speakers and listeners rather than professional actors and experts. As an extension of this approach, some SER [60–63] use four categories of emotional speech, positive (Joy or Happiness), neutral, and negative (Anger and Sadness). In this case it is also quite easy to collect a dataset, but to label the dataset we need experts.

The set of words and set of words and phrases were selected according to the lexical meaning of words /joy, beautiful, cool, super, nothing, normal, sad, hard, scream, break, crush/ and phrases /I love when everything is beautiful, Sad time, I love to beat and break everything/. The meaningless sentences were pseudo-sentences (semantically meaningless sentences resembling real sentences), used e.g., [64] to reduce the influence of linguistic meaning. A fragment of "Jabberwocky", the poem by Lewis Carroll [65] and the meaningless text (sentence) by L.V. Shcherba "Glokaya kuzdra" [66] were used. Children had to utter speech material imitating different emotional states "Joy, Neutral (Calm), Sadness, Anger".

After the completion of the sessions recording the children's speech, two procedures were carried out: audiometry, to assess the hearing threshold, and a phonemic hearing test (repetition of pairs and triple syllables by the child following the speech therapist).

All procedures were approved by the Health and Human Research Ethics Committee of Saint Petersburg State University (Russia), and written informed consent was obtained from parents of the participating children.

Speech recording was conducted according to a standardized protocol. For automatic analysis, only the acting speech of children was taken.

3.3. Process of Speech Data Annotation

After the raw data selection, all speech recordings are split into segments which contain emotions and are manually checked to ensure audio quality. The requirements for the segments are as follows:

a. Segments should not have high background noise or voice overlaps.
b. Each segment should contain only one speaker's speech.
c. Each speech segment should contain a complete utterance. Any segment without an utterance was removed.

The acting speech from the database was annotated according the emotional states manifested by children when speaking: "neutral—joy—sadness—anger". The annotation was made by two experts based on the recording protocol and video recordings. The speech sample is assigned to the determined emotional state if the accordance between two experts is 100%.

2505 samples of acting speech (words, phrases, and meaningless texts) were selected as a result of annotation: 628 for the neutral state, 592 for joy, 646 for sadness, and 639 for anger. In contrast to other existing databases, the distribution of the emotions in our database is approximately uniform.

4. Experimental Setup

4.1. Features and Classifiers

To generate the features vector (eGeMAPS feature set) we have used the openSMILE toolkit v2.0.0 (audEERING GmbH, Gilching, Germany) [67], which is an open-source C++ feature extractor. The code is freely available and well documented.

To implement SVM and MLP classifiers, we have used the Scikit-learn machine learning library. A multiclass version of the SVM model [68] with a linear kernel and default values of all other parameters was chosen. We also used an MLP model that optimizes the log-loss function using LBFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) or stochastic gradient descent algorithms. As it is noted in [69], the stochastic gradient descent solvers work rather well on datasets with thousands of training samples or more in terms of both training time and validation score. However, in our case, we have a small dataset, so we used LBFGS (an optimizer in the family of quasi-Newton methods) because it converges faster and works better for small datasets. To do this we set the parameter solver = 'lbfgs'. After some experiments, we set the parameters alpha = 1×10^{-5}, hidden_layer_sizes = (64,16), max_iter = 500, and the other parameter values are by default.

4.2. Training and Testing Setup

A total of 2505 Russian emotional speech samples were used. We separated the data into 80% (2004 samples) for training and 20% (501 samples) for testing our classification models.

For cross-validation we have used the Stratified K-Folds cross-validator [70]. It provides a greater balance of classes, especially when there are only a few speakers. Due to the small size of the dataset, we used Stratified K-Folds cross-validation with K = 6 (5:1).

4.3. Evaluation Setup

As a tool for evaluation of classification models, we have used a multi-class confusion matrix library PyCM written in Python [71].

In our experiments, we used such evaluation metrics as the per class Accuracy, Precision, Recall, and F1-score. Due to the unequal number of samples in each test class (unequal priors), we have analyzed the results using Unweighted Average Recall (UAR) for multiclass classifiers, closely related to the accuracy as a good or even better metric to optimize when the sample class ratio is imbalanced [72]. UAR is defined as the average across the diagonal of the confusion matrix.

5. Experimental Results

First, we analyzed the results of automatic emotion recognition in children's speech with two state-of-the-art ML classification algorithms, SVM and MLP. Then, following [73], we compared results of manual and automatic speech emotion recognition to understand how well the emotional state in emotional speech is recognized by humans and machines.

5.1. Results of Automatic Emotion Recognition on Extended Feature Set

From Figure 1 and Tables 2 and 3 we can see that performance of SVM and MLP classifiers is approximately the same. For all individual emotions, both classifiers perform noticeably above chance. This is consistent with the results of monolingual emotion recognition in the Russian language [74].

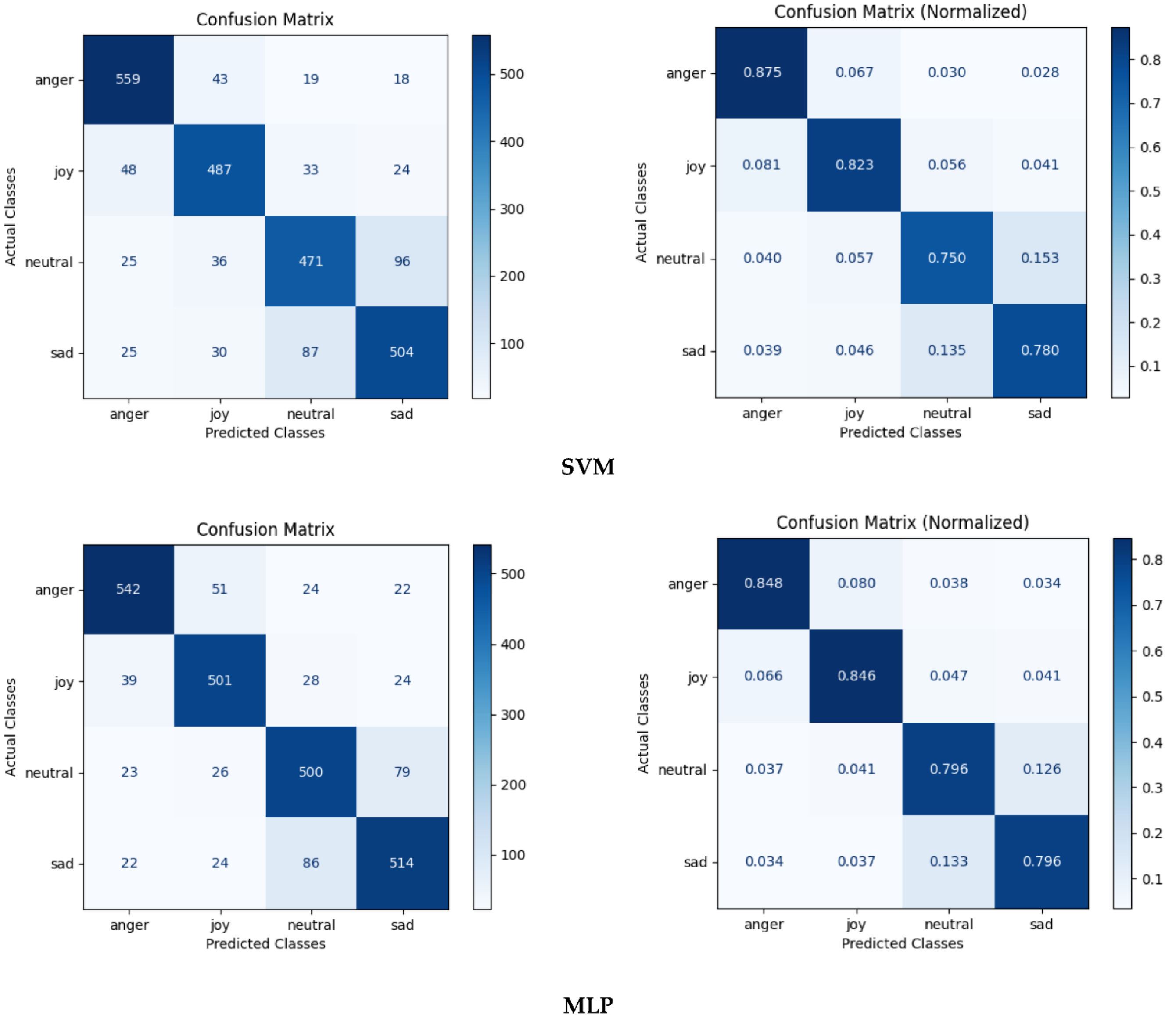
These results are comparable to evaluations in other languages. For example, Sowmya and Rajeswari [75] reported that they achieved an overall accuracy of 0.85 for automatic children's speech emotion recognition in the Tamil language with an SVM classifier on prosodic (energy) and spectral (MFCC) features. Rajan et al. [13] reported that they achieved an Average Recall of 0.61 and Average Precision of 0.60 in the Tamil language using a DNN framework, also on prosodic and spectral features.

Table 2. Per-class scores in multi-class classification, eGeMAPS feature set.

Classifier	**SVM**				**MLP**			
Emotion	**Anger**	**Joy**	**Neutral**	**Sad**	**Anger**	**Joy**	**Neutral**	**Sad**
Accuracy	0.929	0.915	0.882	0.888	0.928	0.923	0.894	0.897
Recall	0.875	0.823	0.750	0.780	0.848	0.846	0.796	0.796
Precision	0.851	0.817	0.772	0.785	0.866	0.832	0.784	0.804
F1-score	0.863	0.820	0.761	0.783	0.857	0.839	0.790	0.800

Table 3. Average scores in multi-class classifications, eGeMAPS feature set.

Classifier	**SVM**	**MLP**
Overall Accuracy	0.903	0.911
Unweighted Average Recall	0.807	0.822
Unweighted Average Precision	0.806	0.822



SVM

MLP

Figure 1. Confusion matrices for SVM and MLP classifiers, eGeMAPS feature set, a dataset of Russian children's emotional speech with 2505 samples for training and testing.

5.2. Results of the Subjective Evaluation of Emotional Speech Recognition

In our first experiment on the subjective evaluation of emotional speech recognition [21], 10 native Russian experts with a mean age of 37.8 years (standard deviation ± 15.4 years) and lengthy (mean ± SD = 14.2 ± 10.3 years) professional experience as experts in the field of speech science manually recognized Happiness/Joy, Anger, Sadness, and Neutral emotional states in children's speech from the dataset consisting of 16 samples of meaningless text from "Jabberwocky" [65] and "Glokaya kuzdra" [66]). There was no preliminary training of experts. The experts were given the task of recognizing emotional states (selected from a list of four emotions) of children while listening to the test sequence. Different utterances were separated by a pause of 7 s, and each utterance was repeated once. This subjective evaluation was conducted to understand how well humans identify emotional states in the acting emotional speech of Russian children. The experts used perceptual evaluation and spectrographic analysis [21] of four samples for each emotional state, with 16 samples in total. The results of this subjective evaluation are shown in Figure 2 and Tables 4 and 5.

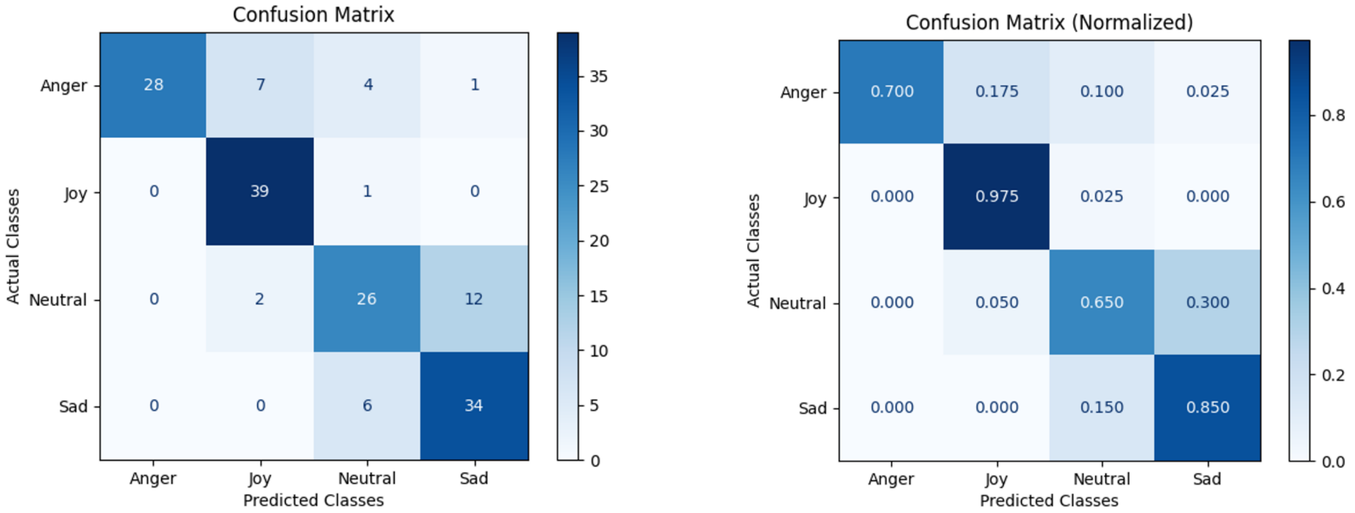


Figure 2. Confusion matrix of subjective evaluation of emotions in acting children's speech in Russian language by Russian experts.

Table 4. Per-class scores in multi-class classification, expert's feature set with 16 samples in total.

Metrics	**Emotion**			
	Anger	**Joy**	**Neutral**	**Sad**
Accuracy	0.925	0.940	0.845	0.880
Recall	0.700	0.980	0.650	0.850
Precision	1.000	0.817	0.707	0.720
F1-score	0.824	0.891	0.677	0.780

Table 5. Overall scores in multi-class classification, expert's feature set with 16 samples in total.

Metrics	**Accuracy**
Overall Accuracy	0.898
Unweighted Average Recall	0.795
Unweighted Average Precision	0.811

From Figure 2 and Table 4 we can see that experts recognize all emotions noticeably above chance (0.25 for 4-class classification). This is consistent with the results of perception tests for adult emotion speech. For example, Rajan et al. [13] reported comparable results of subjective evaluation using a perception test with an overall accuracy of 0.846 in recognizing emotional states—happiness, anger, sadness, anxiety, and neutral—from acted emotional speech in the Tamil language.

In this article, we present the results of an extended experiment with the dataset that consists of 16 samples of meaningless text supplemented with 17 samples of words and phrases, for 33 samples in total. The results of this subjective evaluation are shown in Figure 3 and Tables 6 and 7.

Table 6. Per-class scores in multi-class classification, extended expert's feature set with 33 samples in total.

Metrics	**Emotion**			
	Anger	**Joy**	**Neutral**	**Sad**
Accuracy	0.925	0.938	0.844	0.881
Recall	0.700	0.975	0.650	0.850
Precision	1.000	0.813	0.703	0.723
F1-score	0.824	0.886	0.675	0.782

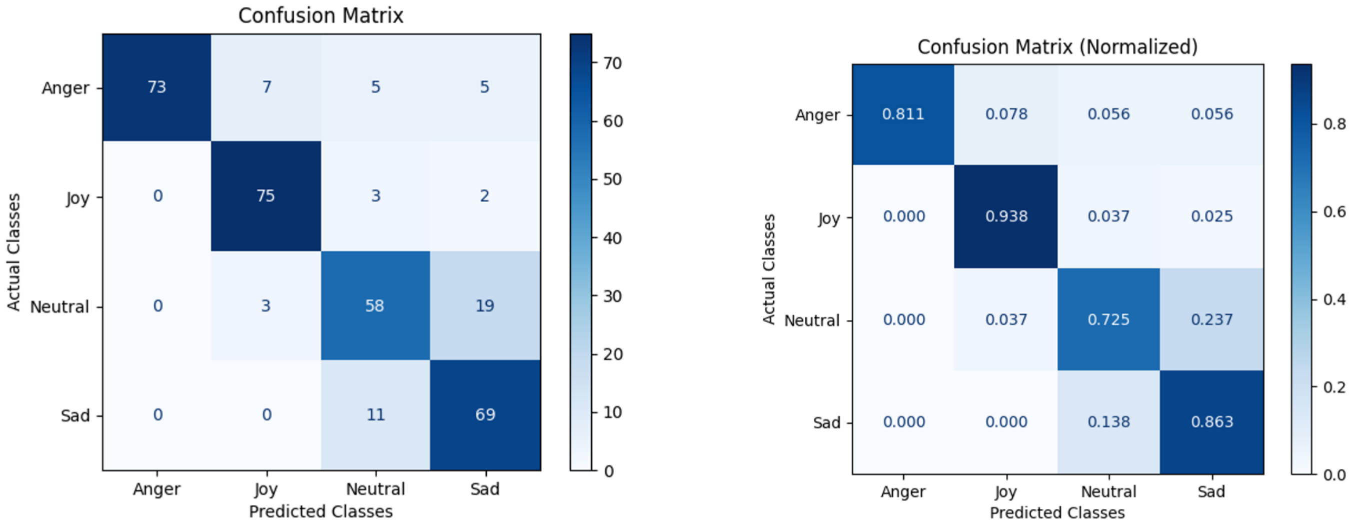


Figure 3. Confusion matrix of subjective evaluation of emotions in acting children's speech in the Russian language by Russian experts on the extended dataset.

Table 7. Overall scores in multi-class classification, extended expert's feature set with 33 samples in total.

Metrics	Accuracy
Overall Accuracy	0.897
Unweighted Average Recall	0.794
Unweighted Average Precision	0.810

From the normalized confusion matrices in Figures 2 and 3 it can be seen that the diagonal values of the matrix in Figure 3 are distributed more evenly than the diagonal values of the matrix in Figure 2, but the overall scores in both cases are approximately the same; see Tables 4–7.

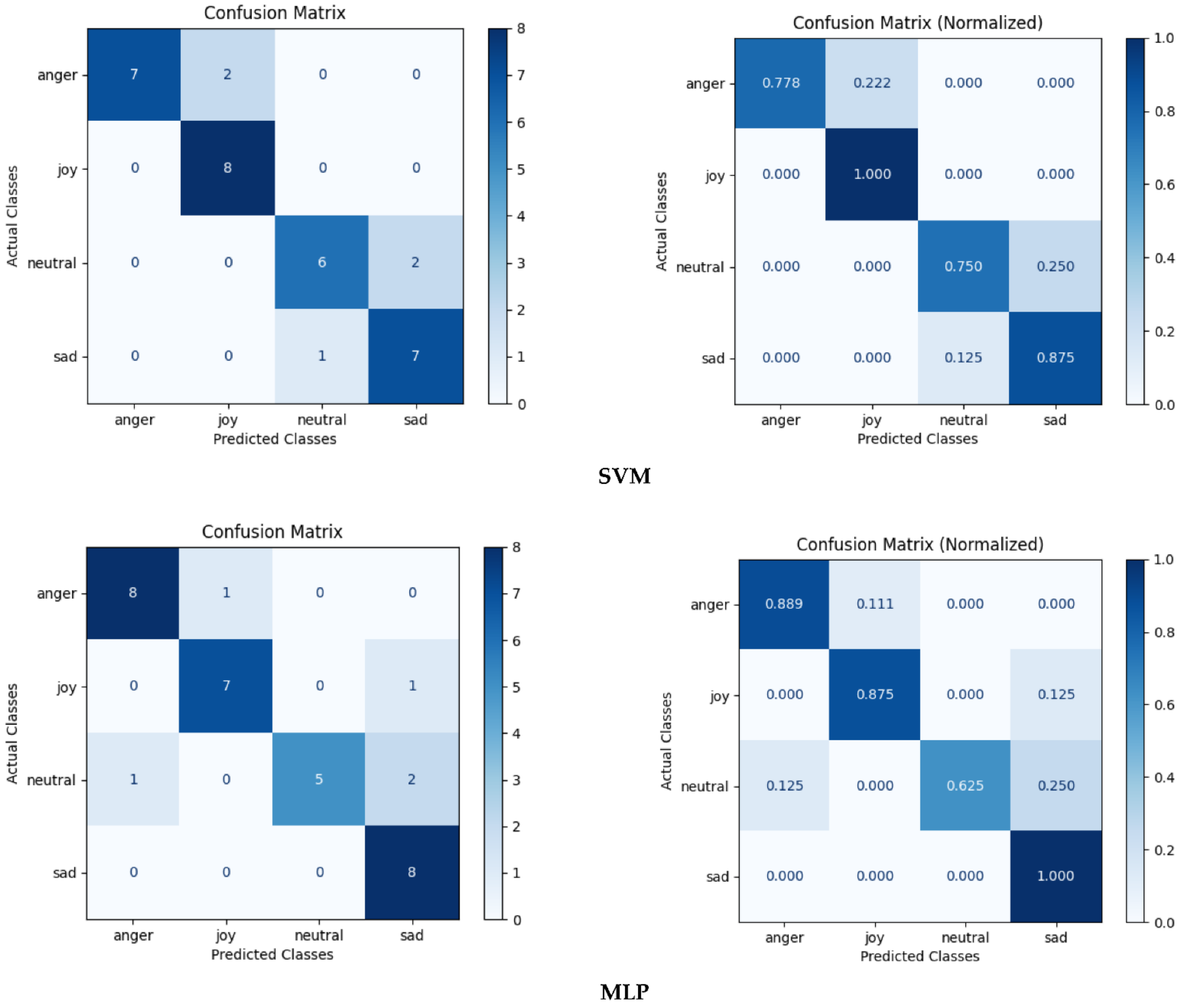
5.3. *Comparison of the Subjective Evaluation and Automatic Emotional Speech Recognition*

Figure 4 shows the confusion matrices for the SVM and MLP classifiers. The training dataset consists of 2505 samples of the acting emotional speech of Russian children. The test set for automatic evaluation consists of 33 separate samples of acting emotional speech of Russian children used in perception tests [21]. Both classifiers were trained based on the eGeMAPS feature set [72].

The results of automatic classification are shown in Tables 8 and 9. Both SVM and MLP classifiers recognize emotional states with comparable performance, which is noticeably above chance. Their overall accuracy is comparable to the overall accuracy of the subjective evaluation, achieving an overall accuracy of 0.833 in automatic emotion recognition for both SVM and MLP classifiers, and the overall accuracy of 0.898 in the subjective evaluation of Russian children's emotional speech.

Table 8. Per-class scores in automatic multi-class classification, eGeMAPS feature set, testing on 33 samples from subjective evaluations.

Classifier	SVM				MLP			
Emotion	**Anger**	**Joy**	**Neutral**	**Sad**	**Anger**	**Joy**	**Neutral**	**Sad**
Accuracy	0.939	0.939	0.909	0.909	0.939	0.939	0.909	0.909
Recall	0.778	1.000	0.750	0.875	0.889	0.875	0.625	1.000
Precision	1.000	0.800	0.857	0.778	0.889	0.875	1.000	0.727
F1-score	0.875	0.889	0.800	0.824	0.889	0.875	0.769	0.842



SVM

MLP

Figure 4. Confusion matrices for SVM and MLP classifiers, a dataset of Russian children's emotional speech with 2505 samples for training, an eGeMAPS feature set, and an additional 33 samples from subjective evaluations for testing.

Table 9. Overall scores in automaticmulti-class classification, eGeMAPS feature set, testing on 33 samples from subjective evaluations.

Metrics	**Classifier**	
	SVM	**MLP**
Overall Accuracy	0.924	0.924
Unweighted Average Recall	0.851	0.847
Unweighted Average Precision	0.859	0.873

As a result of the comparison of the subjective evaluation and automatic emo-tional speech recognition, we can conclude:

1. Humans are able to identify emotional states in Russian children's emotional speech above chance. Following [13], this ensures the emotional quality and naturalness of the emotions in the collected corpus of Russian children's emotional speech.
2. The performance of automatic emotion recognition using state-of-the-art ML algorithms/classifiers trained on the collected corpus is at least comparable to the perfor-

mance of perception tests. This also proves that the collected corpus can be used to build automatic systems for emotion recognition in Russian children's emotion speech.

6. Discussion and Conclusions

Our experiments were conducted on our proprietary dataset of children's emotional speech in the Russian language of the younger school age group [21]. To our knowledge, this is the only such database.

We have conducted a validation of our dataset based on classical ML algorithms, such as SVM and MLP. The validation was performed using standard procedures and scenarios of the validation similar to other well-known databases of children's emotional speech.

First of all, we demonstrated the following performance of automatic multiclass recognition (four emotion classes): SVM Overall Accuracy = 0.903, UAR = 0.807, and also MLP Overall Accuracy = 0.911, UAR = 0.822 are superior to the results of the perceptual test with Overall Accuracy = 0.898 and UAR = 0.795. Moreover, the results of automatic recognition on the test dataset, which consists of 33 samples used in the perceptual test, are even better: SVM Overall Accuracy = 0.924, UAR = 0.851, and also MLP Overall Accuracy = 0.924, UAR = 0.847. This can be explained by the fact that for the perceptual test we selected samples with clearly expressed emotions.

Second, a comparison with the results of validation of other databases of children's emotional speech on the same and other classifiers showed that our database contains reliable samples of children's emotional speech which can be used to develop various edutainment [76], health care, etc. applications using different types of classifiers.

The above confirms that this database is a valuable resource for researching the affective reactions in speech communication between a child and a computer in Russian.

Nevertheless, there are some aspects of this study that need to be improved.

First, the most important challenge in the applications mentioned above is how to automatically recognize a child's emotions with the highest possible performance. Thus, the accuracy of the speech emotion recognition must be improved. The most prominent way to solve the problem is use deep learning techniques. However, one of the significant issues in deep learning-based SER is the limited size of the datasets. Thus, we need larger datasets of children's speech. One of the suitable solutions we suggest is to study the creation of a fully synthetic dataset using generative techniques trained on available datasets [77]. GAN-based techniques have already shown success for similar problems and would be candidates for solving this problem as well. Moreover, we can use generative models to create noisy samples and try to design a noise-robust model for SER in real-world child communications including various gadgets for atypical children.

Secondly, other sources of information for SER should be considered, such as children's face expression, posture, and motion.

In the future, we intend to improve the accuracy of the automatic SER by

- extending our corpus of children's emotional speech to be able to train deep neural networks of different architectures;
- using multiple modalities such as facial expressions [58], head and body movements, etc.

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

Funding: The research was financially supported by the Russian Foundation for Basic Research (projects 19-57-45008–IND) and Russian Science Foundation (project 22-45-02007). Database collection and selection the feature set, powerful to distinguish between different emotions were financially supported by the RFBR project 19-57-45008; automatic emotion speech classification was financially supported by the RSF project 22-45-02007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The publicly available dataset was analyzed in this study. This dataset can be found here: https://github.com/hydra-colab/Emotion-Recognition-of-Younger-School-Age-Children, accessed on 6 May 2022.

Acknowledgments: The study was carried out as part of a joint project with Department of Science and Technology (DST), School of Electrical Engineering, Vellore Institute of Technology, Vellore, India (INTRUSRFBR382).

Conflicts of Interest: The authors declare that they have no conflicts 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

1. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
2. Lefter, I.; Rothkrantz, L.J.M.; Wiggers, P.; van Leeuwen, D.A. Emotion Recognition from Speech by Combining Databases and Fusion of Classifiers. *Lect. Notes Comput. Sci.* **2010**, *6231*, 353–360. [CrossRef]
3. Schuller, B.W. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
4. Ganapathy, A. Speech Emotion Recognition Using Deep Learning Techniques. *ABC J. Adv. Res.* **2016**, *5*, 113–122. [CrossRef]
5. 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]
6. Polzehl, T.; Sundaram, S.; Ketabdar, H.; Wagner, M.; Metze, F. Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. In Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, UK, 6–10 September 2009; pp. 340–343. [CrossRef]
7. 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. Available online: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse91912020.pdf (accessed on 20 April 2022). [CrossRef]
8. Cao, G.; Tang, Y.; Sheng, J.; Cao, W. Emotion Recognition from Children Speech Signals Using Attention Based Time Series Deep Learning. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 1296–1300. [CrossRef]
9. Onwujekwe, D. Using Deep Leaning-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 April 2022).
10. Kaur, K.; Singh, P. Punjabi Emotional Speech Database: Design, Recording and Verification. *Int. J. Intell. Syst. Appl. Eng.* **2021**, *9*, 205–208. [CrossRef]
11. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [CrossRef]
12. Akçay, M.B.; Oguz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *166*, 56–76. [CrossRef]
13. Rajan, R.; Haritha, U.G.; Sujitha, A.C.; Rejisha, T.M. Design and Development of a Multi-Lingual Speech Corpora (TaMaR-EmoDB) for Emotion Analysis. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 3267–3271. [CrossRef]
14. 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]
15. 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]
16. Lyakso, E.; Frolova, O.; Dmitrieva, E.; Grigorev, A.; Kaya, H.; Salah, A.A.; Karpov, A. EmoChildRu: Emotional child Russian speech corpus. *Lect. Notes Comput. Sci.* **2015**, *9319*, 144–152. Available online: https://www.researchgate.net/profile/Heysem-Kaya/publication/281583846_EmoChildRu_Emotional_Child_Russian_Speech_Corpus/links/55ee969208aedecb68fca34c/EmoChildRu-Emotional-Child-Russian-Speech-Corpus.pdf (accessed on 20 April 2022). [CrossRef]

17. 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]
18. 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]
19. van den Heuvel, H. The art of validation. *ELRA Newsl.* **2000**, *5*, 4–6. Available online: http://www.elra.info/media/filer_public/2013/12/04/elranews_v12-1.pdf (accessed on 20 April 2022).
20. van den Heuvel, H.; Iskra, D.; Sanders, E.; de Vriend, F. Validation of spoken language resources: An overview of basic aspects. *Lang Resour. Eval.* **2008**, *42*, 41–73. [CrossRef]
21. Lyakso, E.; Frolova, O.; Ruban, N.; Mekala, A.M. The Child's Emotional Speech Classification by Human Across Two Languages: Russian & Tamil. *Lect. Notes Comput. Sci.* **2021**, *12997*, 384–396. [CrossRef]
22. 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]
23. Javaheri, B. Speech & Song Emotion Recognition Using Multilayer Perceptron and Standard Vector Machine. *arXiv* **2021**, arXiv:2105.09406.
24. Zanaty, E.A. Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification. *Egypt. Inform. J.* **2012**, *13*, 177–183. [CrossRef]
25. Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria, Y.B. Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network. *Sensors* **2020**, *20*, 6008. [CrossRef] [PubMed]
26. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249. [CrossRef] [PubMed]
27. Emotional Databases. Available online: http://kahlan.eps.surrey.ac.uk/savee/Introduction.html (accessed on 20 April 2022).
28. Devi, S.; Kumbham, V.; Boddu, D. A Survey on Databases and Algorithms used for Speech Emotion Recognition. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 7032–7039. [CrossRef]
29. List of Children's Speech Corpora. Available online: https://en.wikipedia.org/wiki/List_of_children%27s_speech_corpora (accessed on 20 April 2022).
30. Grill, P.; Tučková, J. Speech Databases of Typical Children and Children with SLI. *PLoS ONE* **2016**, *11*, e0150365. [CrossRef] [PubMed]
31. Matin, R. Developing a Speech Emotion Recognition Solution Using Ensemble Learning for Children with Autism Spectrum Disorder to Help Identify Human Emotions. Unpublished Thesis, Texas State University, San Marcos, TX, USA, 2020. Available online: https://digital.library.txstate.edu/handle/10877/13037 (accessed on 20 April 2022).
32. Duville, M.M.; Alonso-Valerdi, L.M.; Ibarra-Zarate, D.I. Mexican Emotional Speech Database (MESD). *Mendeley Data* **2021**, *V2*, 1644–1647. [CrossRef]
33. 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. Available online: http://multicomp.cs.cmu.edu/wp-content/uploads/2017/09/2016_ICMI_Nojavanasghari_Emoreact.pdf (accessed on 20 April 2022). [CrossRef]
34. Li, Y.; Tao, J.; Chao, L.; Bao, W.; Liu, Y. CHEAVD: A Chinese natural emotional audio–visual database. *J. Ambient Intell Hum. Comput* **2017**, *8*, 913–924. Available online: http://www.speakit.cn/Group/file/2016_CHEAVD_AIHC_SCI-Ya%20Li.pdf (accessed on 20 April 2022). [CrossRef]
35. 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.
36. Pérez-Espinosa, H.; Reyes-García, C.; Villaseñor-Pineda, L. EmoWisconsin: An Emotional Children Speech Database in Mexican Spanish. In Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII), Memphis, TN, USA, 9–12 October 2011; pp. 62–71. [CrossRef]
37. Batliner, A.; Hacker, C.; Steidl, S.; Nöth, E.; D'Arcy, S.; Russell, M.; Wong, M. "You Stupid Tin Box"—Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 26–28 May 2004; pp. 171–174. Available online: http://www.lrec-conf.org/proceedings/lrec2004/pdf/317.pdf (accessed on 20 April 2022).
38. FAU Aibo Emotion Corpus. Available online: https://www5.cs.fau.de/en/our-team/steidl-stefan/fau-aibo-emotion-corpus/ (accessed on 20 April 2022).
39. Pattern Recognition. Available online: https://www.sciencedirect.com/topics/engineering/pattern-recognition (accessed on 20 April 2022).
40. Basharirad, B.; Moradhaseli, M. Speech Emotion Recognition Methods: A Literature Review. *AIP Conf. Proc.* **2017**, *1891*, 020105-1–020105-7. [CrossRef]
41. Schuller, B.; Steidl, S.; Batliner, A. The INTERSPEECH 2009 Emotion Challenge. In Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, UK, 6–10 September 2009; pp. 312–315. [CrossRef]

42. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; Andre, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
43. Disvoice's Documentation. Available online: https://disvoice.readthedocs.io/en/latest/index.html (accessed on 20 April 2022).
44. Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 171–175. [CrossRef]
45. Eskimez, S.E.; Dimitriadis, D.; Gmyr, R.; Kumanati, K. GAN-based Data Generation for Speech Emotion Recognition. In Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 3446–3450. [CrossRef]
46. Ying, Y.; Tu, Y.; Zhou, H. Unsupervised Feature Learning for Speech Emotion Recognition Based on Autoencoder. *Electronics* **2021**, *10*, 2086. [CrossRef]
47. Andleeb Siddiqui, M.; Hussain, W.; Ali, S.A.; ur-Rehman, D. Performance Evaluation of Deep Autoencoder Network for Speech Emotion Recognition. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 606–611. [CrossRef]
48. Cai, X.; Yuan, J.; Zheng, R.; Huang, L.; Church, K. Speech Emotion Recognition with Multi-Task Learning. In Proceedings of the 22th Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czechia, 30 August–3 September 2021; pp. 4508–4512. [CrossRef]
49. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore, 14–18 September 2014; pp. 223–227.
50. Padi, S.; Sadjadi, S.O.; Sriram, R.D.; Manocha, D. Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI)*; Montréal, QC, Canada, 18–22 October 2021, pp. 645–652. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=932172 (accessed on 20 April 2022). [CrossRef]
51. Rumagit, R.Y.; Alexander, G.; Saputra, I.F. Model Comparison in Speech Emotion Recognition for Indonesian Language. *Procedia Comput. Sci.* **2021**, *179*, 789–797. [CrossRef]
52. Ali Alnuaim, A.; Zakariah, M.; Kumar Shukla, P.; Alhadlaq, A.; Hatamleh, W.A.; Tarazi, H.; Sureshbabu, R.; Ratna, R. Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. *J. Healthc. Eng.* **2022**, *6005446*, 1–12. [CrossRef] [PubMed]
53. Poojary, N.N.; Shivakumar, G.S.; Akshath Kumar, B.H. Speech Emotion Recognition Using MLP Classifier. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2021**, *7*, 218–222. [CrossRef]
54. Kaur, J.; Kumar, A. Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest. *Lect. Notes Data Eng. Commun. Technol.* **2021**, *58*, 499–509. [CrossRef]
55. Yang, N.; Dey, N.; Sherratt, R.S.; Shi, F. Recognize Basic Emotional States in Speech by Machine Learning Techniques Using Mel-Frequency Cepstral Coefficient Features. *J. Intell. Fuzzy Syst.* **2020**, *39*, 1925–1936. [CrossRef]
56. Goel, S.; Beigi, H. Cross Lingual Cross Corpus Speech Emotion Recognition. *arXiv* **2020**, arXiv:2003.07996v1.
57. Chugh, V.; Kaw, S.; Soni, S.; Sablan, V.; Hande, R. Speech Emotion Recognition System Using MLP. *J. Emerg. Technol. Innov. Res.* **2021**, *8*, 222–226.
58. Lyakso, E.; Frolova, O.; Matveev, Y. Chapter 14—Facial Expression: Psychophysiological Study. In *Handbook of Research on Deep Learning-Based Image Analysis Under Constrained and Unconstrained Environments*; Raj, A., Mahesh, V., Nersisson, R., Eds.; IGI Global: Pennsylvania, PA, USA, 2021; pp. 266–289. [CrossRef]
59. Laukka, P.; Elfenbein, H.A. Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis. *Emot. Rev.* **2021**, *13*, 3–11. [CrossRef]
60. Rajoo, R.; Aun, C.C. Influences of languages in speech emotion recognition: A comparative study using malay, english and mandarin languages. In Proceedings of the IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE), Penang, Malaysia, 30–31 May 2016; pp. 35–39. [CrossRef]
61. Heracleous, P.; Yoneyama, A. A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. *PLoS ONE* **2019**, *14*, e0220386. [CrossRef]
62. Latif, S.; Qadir, J.; Bilal, M. Unsupervised Adversarial Domain Adaptation for Cross-Lingual Speech Emotion Recognition. *arXiv* **2020**, arXiv:1907.06083v4.
63. Latif, S.; Qayyum, A.; Usman, M.U.; Qadir, J. Cross lingual speech emotion recognition: Urdu vs. western languages. *arXiv* **2020**, arXiv:1812.10411.
64. Gilam, G.; Hendler, T. Deconstructing Anger in the Human Brain. *Curr Top Behav Neurosci.* **2017**, *30*, 257–273. [CrossRef]
65. Carrol, L. *Through the Looking-Glass and What Alice Found There*; Macmillan and Co.: London, UK, 1872.
66. GLOKAYA KUZDRA. Available online: http://languagehat.com/glokaya-kuzdr (accessed on 20 April 2022).
67. openSMILE Python. Available online: https://github.com/audeering/opensmile-python (accessed on 20 April 2022).
68. Support Vector Machines. Available online: https://scikit-learn.org/stable/modules/svm.html#svm (accessed on 20 April 2022).

69. Multi-Layer Perceptron Classifier. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html (accessed on 20 April 2022).
70. Stratified K-Folds Cross-Validator. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html (accessed on 20 April 2022).
71. Haghighi, S.; Jasemi, M.; Hessabi, S.; Zolanvari, A. PyCM: Multiclass confusion matrix library in Python. *J. Open Source Softw.* **2018**, *3*, 729. [CrossRef]
72. Schuller, B.W.; Weninger, F. Ten recent trends in computational paralinguistics. *Lect. Notes Comput. Sci.* **2012**, *7403*, 35–49. Available online: https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/72653/file/72653.pdf (accessed on 20 April 2022). [CrossRef]
73. Werner, S.; Petrenko, G.K. A Speech Emotion Recognition: Humans vs Machines. *Discourse* **2019**, *5*, 136–152. [CrossRef]
74. Verkholyak, O.V.; Kaya, H.; Karpov, A.A. Modeling Short-Term and Long-Term Dependencies of the Speech Signal for Paralinguistic Emotion Classification. *Tr. SPIIRAN* **2019**, *18*, 30–56. [CrossRef]
75. Sowmya, V.; Rajeswari, A. Speech emotion recognition for Tamil language speakers. *Adv. Intell. Syst. Comput.* **2020**, *1085*, 125–136. [CrossRef]
76. Guran, A.-M.; Cojocar, G.-S.; Dioşan, L.-S. The Next Generation of Edutainment Applications for Young Children—A Proposal. *Mathematics* **2022**, *10*, 645. [CrossRef]
77. Kaliyev, A.; Zeno, B.; Rybin, S.V.; Matveev, Y.N.; Lyakso, E.E. GAN acoustic model for Kazakh speech synthesis. *Int. J. Speech Technol.* **2021**, *24*, 729–735. [CrossRef]