DETECTING INCONGRUITY IN THE EXPRESSION OF EMOTIONS IN SHORT VIDEOS BASED ON A MULTIMODAL APPROACH

Anastasia Laushkina
National Center for Cognitive Research
ITMO University
Russia
nastasjalausckina@mail.ru

Ivan Smirnov
National Center for Cognitive Research
ITMO University
Russia
1vany1@mail.ru

Anatoly Medvedev, Andrey Laptev, Mikhail Sinko
National Center for Cognitive Research
ITMO University
Russia
anatolmdvdv@gmail.com; nickname.avast@gmail.com; michael.v.sinko@gmail.com

Article history:
Received 27.09.2022, Accepted 17.12.2022

Index terms

Incongruence detection; Facial expression recognition; Multimodal analysis; Speech-recognition; Speech Emotion Recognition; Text analysis.

Abstract

Every day people face uncertainty, which is already an integral part of their lives. Uncertainty creates risks for various kinds of companies, in particular, the financial sector may incur losses due to various kinds of human errors. People turn to the opinion of experts who have special knowledge to eliminate this uncertainty. It is established that the expert shows insolvency if he uses incongruent manipulation techniques. In this article we propose a method that allows solving the problem of congruence estimation. The hypothesis that a person with a prepared speech and a person with a spontaneous speech will have a different level of congruence is also put forward and tested in this work. The similarity of emotional states of verbal and nonverbal channels is evaluated in our solution for determining congruence. Convolutional neural networks (CNN) were used to assess a person's emotional state from video and audio, speeth-to-text to extract the text of the speaker's speech, and a pre-trained BERT model for subsequent analysis of emotional color. Tests have shown that with the help of this development it is possible not only to distinguish the incongruence of a person, but also to point out the unnatural nature of his origin (to distinguish a simply incongruent person from a deepfake).

1 Introduction

The task of emotions recognition has recently become one of the most popular, partly because its solution is applicable in various spheres of life. The most interesting studies include augmented and virtual reality, driver monitoring systems, human-computer interaction and security systems. It is known that both verbal and nonverbal signals serve for communication, and the ability to emotions recognition is one of the most important nonverbal means that can provide this communication. Emotions recognition can be carried out by perceiving a variety of signals, such as speech, facial expressions, gestures and body language [Amer et al. (2014); Dong et al. (2022)]. Social psychologists claim that more than 65% of the information exchanged during a personal conversation occurs in the nonverbal range [Knapp (1978); Morris (1979)].

There are studies that are aimed at detecting emotions through a text channel. For example, the authors of the article [Badugu and Suhasini (2017)] used a rules-based approach and the Russell circular model to detect emotions using tweets. The [Wikarsa and Thahir (2015)] article describes an approach using a Naive Bayesian Classifier to determine the emotions of Twitter users, the accuracy of which was 83%.

The problem of emotion recognition from video was taken up by the authors of [Shan et al. (2017)], who used a deep convolutional neural (CNN) network and the k-nearest neighbor (KNN) method to extract facial expressions. In a dataset of 7 classes of JAFFE and CK+, the accuracy was 76.7442% and 80.303%, respectively. The authors of [Zhanget et al. (2019)] used a hybrid model based on spatial CNN for processing static images and temporal CNN for image flow. Using the Deep Believe Network model made it possible to combine collected objects from spatial and temporal branches at the segment level. This approach made it possible to achieve an accuracy of 75.39%.

Currently, there are solutions that can draw conclusions about the psycho-emotional state of a person and the truthfulness of his statements. For example, one of the approaches to establishing the truth of a person's statements is the analysis of the audio stream. For example, the authors of [Zhanget et al. (2019); Amiriparian et al. (2016)] analyzed nonverbal signs of speech that characterize emotional color – the timbre of the voice, volume and tempo of speech. In addition, they used some lexical, grammatical and syntactic features. In the work [Savchenko and Vasil'ev (2014)], the intensity/strength of emotions, valence/tone, and emotional regulation were used. In the work [Kirchhubel et al. (2013)], the authors studied the properties of speech and found when a person lies the pace of speech increases, the response start time and the duration of oscillations decrease.

Information extracted from one modality is usually one-sided and it can lead to low accuracy [Thang et al. (2019)]. It is a reason why the issue of a multimodal approach to identifying emotions remains unexplored to identify critical situations, such as threats, lie, mental state, problems with awareness and consistency of feelings and emotions experienced, etc. In situations when there is a discrepancy between experience, awareness and the expression of feelings, emotions, incongruence manifests itself. The person's response is unclear, ambiguous in such cases. If there are inconsistencies between awareness and the message, it is more like falsehood and falseness. Determining the consistency of information (congruence) transmitted simultaneously through several channels is an important task, since it will allow us to advance in solving problems related to the definition of sarcasm, lies, human insecurity, etc. A person who has congruence will demonstrate complete sincerity and integrity in the process of speech/communication, what means he will transmit information that does not contradict social norms.

Thus, in our study we focused on identifying the relationship between the degree of manifestation of the level of congruence in the conditions of readiness of speech and spontaneous response to given topics.

2 Methods

In this part our work we will describe the approaches for working with every channel (video, audio, text). Then we will describe the method to combine extracted emotions and transfer it to congruence level evaluation.

There are different benefits that can be obtained with the ability to evaluate incongruence. For example, the ability to detect context incongruity can be the way to solve sarcasm detection problem. In the paper [Joshi et al. (2016)] to obtain context it is proposed to use different types of word embeddings: LSA, GloVE, Dependency Weights, Word2Vec.

Working with video recording allows you to get more information that describes the context of speech more comprehensively. In this case, additional channels of information should be considered for a more complete assessment of incongruities. In our work we evaluate differences between levels of emotions obtained from three channels: video, audio, text to get the speaker's level of incongruence. Figure 1 shows a sequence diagram of video processing to determine the level of congruence.

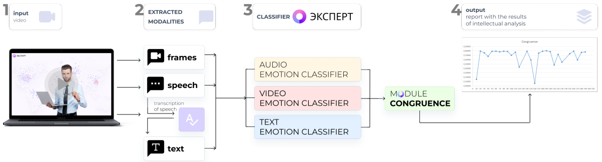


Figure 1. Video processing sequence diagram for determining the level of congruence

2.1 Video emotions

2.1.1 Related works Video modality contains important non-verbal features that open up possibilities for analyzing person's emotional state. Facial expression recognition (FER) is a method that uses machine learning methods to automatically recognize human emotional manifestations: anger, disgust, happiness, surprise, fear and sadness, which are considered to be the basic human emotions [Ekman (1992)]. A distinctive feature of facial expression recognition from the classical multi-class classification is the presence of common features in various emotional experiences. So, for example, fear, surprise and anger are characterized by wide-open eyes, raised eyebrows and a parted mouth, while the sparkle in the eyes and the openness of the face combine happy and surprised emotions [Ekman (1993)].

The authors of the method [Zhengyao et al. (2021)] propose to use the Feature Clustering Network (FCN) to optimize losses both in intra-class and inter-class predictions. The authors combine this network with the Multi-head cross Attention Network (MAN) and Attention Fusion Network (AFN). MAN allows authors to train multiple attention heads to classify facial expressions across multiple regions of the face, while AFN aims to combine attention maps obtained from multiple regions. The model was trained on a large-scale AffectNet [Mollahosseini et al. (2019)] database, which includes 287 401 images, including, in addition to 6 basic emotions, classes of contempt and neutral, which identifies the normal state of a person without vivid manifestations of emotions.

Work [Schoneveld et al. (2021)] uses a multimodal approach to classify facial expressions across 8 categories of the AffectNet dataset, proposing the use of convolutional neural networks for parallel feature extraction from audio and video modalities and a recurrent neural network for aggregation and final classification. The authors achieved an accuracy of 61.60% in recognition of expressions through the use of LSTM at the stage of directed study of features obtained from independent modalities.

2.1.2 Method We propose to use Convolutional Neural Network (CNN) with parallel convolutional layers to classify facial expressions. Combining the outputs of parallel blocks allows the use of features obtained from different regions of the image to reduce the influence of interclass differences.

We use the BlazeFace [Bazarevsky et al. (2019)] network to detect a human face in 128x128 pixel frames from a video stream, using a lightweight architecture and achieving a face position prediction accuracy of 98.61%. The encoder in our model for extracting important visual features from the image of a detected human face, which are further processed by parallel blocks of convolutional layers, is ResNet-18 [He et al. (2016)]. The combined features are sent as input to a fully connected classifier for the final emotion prediction.

The model was trained on a sample from the AffectNet dataset containing 8 emotions, but for inference it uses predictions for only three classes: anger, calmness, happiness. We condition this choice by the redundancy of a multi-class classification of emotions to determine the incongruence of a person, leaving only extreme emotional states: the area of a person's negative state (anger), normal (neutral) and the area of a person's positive state (happiness/joy), which allows us to reduce the error in classifying individual classes and group the model outputs according to semantically close manifestations of emotions and various modalities.

2.2 Audio emotions

2.2.1 Related works Audio analysis of speech allows it to work with features that provide more context to the speaker's intent. In the paper [Rana and Jha (2022)] usage of audio emotions decreases the number of false toxic speech detected utterances compared with an approach that considers only transcribed text. To extract useful features from the audio signal, signal in waveform should be preprocessed. One of the popular approaches is transformation to Mel-Frequency cepstral coefficients (MFCC). That transformation presents the audio signal in the form of coefficients based on human signal perception of frequencies [Likitha et al. (2017)].

Authors of the paper [Kozhakhmet et al. (2020)] used audio records transformed to MFCCs to solve the task of speech emotion recognition. In this paper authors focused on classification of only three base states (positive, neutral, negative), while existing datasets include 7 marks such as: anger, boredom, anxiety, joy, sadness, disgust, lack of emotions [Burkhardtm et al. (2005); Luna-Jiménez et al. (2022); Gournay et al. (2018)]. Considering a bigger number of classes provides an opportunity to make the transition from extremely negative to extremely positive emotions smoother (since sadness, which can be defined as negative, differs from anger, which belongs to the same class) [Sinko, M. et al. (2022)].

Popular method to solve the audio classification task is usage of convolutional neural networks (CNN). In the way of representing the signal in the form of MFCCs it is possible to use 2D convolutional layers [Verbitskiy et al. (2022)], this technique allows to train the model on the patterns of speech in terms of time, frequency and dynamic.

2.2.2 Method A lot of datasets are specialized for concrete countries or nationality, for this reason models trained on them can have poor performance on real records and on data from other languages [Kozhakhmet et al. (2020)]. For this reason our solution was decided to combine datasets for different languages such as: Emo-DB, RAVDESS, CaFE, SAVEE [Morris et al. (1979); Badugu and Suhasini (2017); Wikarsa and Thahir (2015)]. Thus, there was no retraining in the language of one of the datasets. These solutions also allow an increased number of records.

evaluate the level of emotions, CNN model (5 convolutional layers with batch normalization, dropout, and fully connected layer) on MFCC representations of the records was trained. For congruence were used only logits of neutral, angry, and happy statements.

2.3 Text emotions

2.3.1 Related works The deep learning approach named BERT-CNN for the text emotion classification was introduces in the article [Smetanin (2020)]. The main idea of the solution is the use utterance embeddings as input for CNN model, which has a role of classifier.

When we use text models, the description of the text occurs due to the information obtained at the training stage. In [Kwok and Wang (2013)] the Bag of Words approach is used to detect aggressive statements. This approach was compared with recurrent neural networks and networks with attention mechanisms in [Smetanin (2020)] in the problem of detecting toxic comments, where the BERT model [Devlin et al. (2019)] demonstrated the best accuracy.

2.3.2 Method For working with the text channel it was necessary to solve the problem of transcription (decoding information from audio or video to text form) speech. To solve the problem, a Vosk toolkit [Shmyrev

(2020)] was used. Vosk toolkit is based on a common Deep Learning NN and Hidden Markov Model (DNN-HMM architecture) [Lv et al. (2021)]. The toolkit provided for transcription supports more than 20 languages and is trained on a large datasets.

The text obtained after transcription of speech is necessary for further emotional classification. Our method uses a pre-trained distilled version of the RoBERTa model. The model has 6 layers, 768 dimensions and 12 heads, totalizing 82M parameters. We use it to classify the text according to 6 basic emotions and the class of calmness, which reflects the absence of bright emotional manifestations in the text. Next, we extract extreme emotional states (anger, calmness, joy) and add them to the overall congruence analysis. A separate study of the text channel allows us to form a conclusion about the congruence of a person not only based on the intonation of his voice, but also the color of the words, the use of special expressions in relation to the object under discussion.

2.4 Congruence

In our study congruence is understood as a characteristic that allows us to assess the level of consistency of information transmitted by a person simultaneously verbally and nonverbally (via audio, video and text channels).

To do this, we divide a person's speech into 10-second intervals, for which we calculate 3 extreme emotions (anger, neutral, joy) on 3 channels separately. The output is a 3 by 3 matrix for each interval. Then the standard deviation for individual emotions and individual channels is calculated. The minimum difference between the average value and all indicates a small spread of values (mismatch of emotions and channels), and the high one indicates incongruence (for example, joy is highlighted by audio, and anger is determined on the face).

We took the value 0.5 as a threshold, where values above the threshold value are considered incongruent (high spread of emotions).

2.5 Experiment

The study involved 22 women and 13 men of various professional activities, the age range was from 20 to 28 years. The restrictions on participation were the following: the age limit of 18 years, diagnosed psychological diseases. All participants signed an informed consent to conduct the experiment and process personal data. Each of the participants was tested and recorded in comfortable conditions for him to eliminate the possible influence of third-party factors.

Based on the analysis of the spheres of activity and the position held, the distribution was shown in Figure 2, including:

1. work with plant organisms, animal organisms, microorganisms;
2. work with technical objects (machines, mechanisms), materials, types of energy;
3. work with conventional signs, numbers, codes, natural or artificial languages;
4. work with phenomena, facts of artistic representation of reality;
5. work related to communication with people.

After the initial selection, 27 out of 35 subjects remained. The videos were divided into two categories:

1) a video in which participants answered questions without first preparing for the answers (27 videos). The total duration of the video was x minutes;

2) a video in which respondents told a prepared story from an area in which they are well versed (27 videos). The total duration of the video was 156.34 minutes.

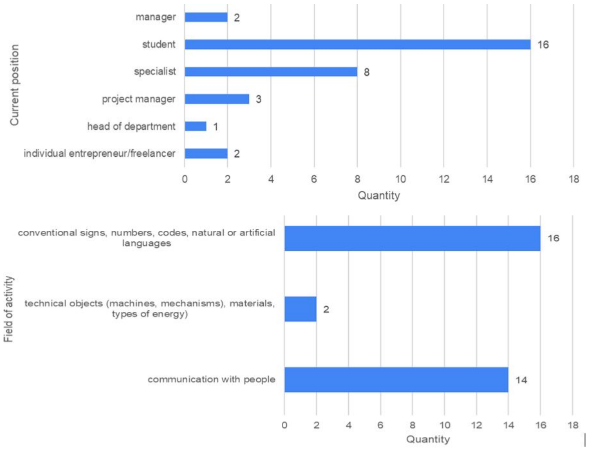
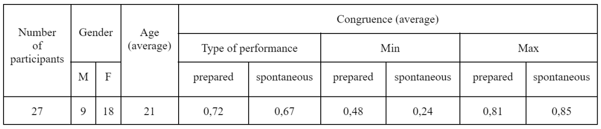


Figure 2. Distribution by position and field of activity

Each of the participants had to pass a test that described their communication style and patterns of social behavior. On average respondents noted that 52% of the time during the day they have to communicate with people, the maximum value was noted by one respondent (100% of the time), the minimum value was noted by 4 respondents (10% of the time). Also, 87% of respondents noted that they had to convince other people of what they disagreed with, and on average such cases accounted for 24.7% of the total time of communication with other people, with 2 respondents noting 70%. The prevailing number of respondents adhere to their interests in the conversation (72.7%), 15.2% of participants noted ignoring their interests, 9.1% always agree with the opinion of the interlocutor and one respondent noted a lack of interest in the topic of the interlocutor. At the same time, in 54.5% of cases, respondents have a feeling of depression when they have to talk about unknown topics, 45.5% remain in the same state.

3 Results

The results of the answer to the question "When you communicate with an interlocutor, you try" were not included in the final analysis due to the peculiarities of the sample (most of the respondents were students) and the imbalance of data (Figure 3).



Number of participants	Gender		Age (average)	Congruence (average)					
				Type of performance		Min		Max	
	M	F		prepared	spontaneous	prepared	spontaneous	prepared	spontaneous
27	9	18	21	0,72	0,67	0,48	0,24	0,81	0,85

Figure 3. A descriptive statistic for prepared and spontaneous speeches.

All data were averaged by the type of performances and by the user. Some subjects were removed as outliers, as they had a significantly greater difference in congruence between different types of performances than other subjects. The distribution of residuals did not significantly differ from normal (Shapiro-Wilco - p-value = 0.67). It is acceptable to use parametric comparison methods. The Student's T-test for paired samples showed significantly differences between the types of performances (p-value = 0.0178). The average congruence for prepared speeches was 0.72, for spontaneous speeches 0.67. The effect size was calculated using Cohen's d coefficient and was 0.49, which corresponds to the average effect size. This means that the existing model (congruence depends on the type of speech) describes the variance in congruence at a sufficient level (Fig. 4).

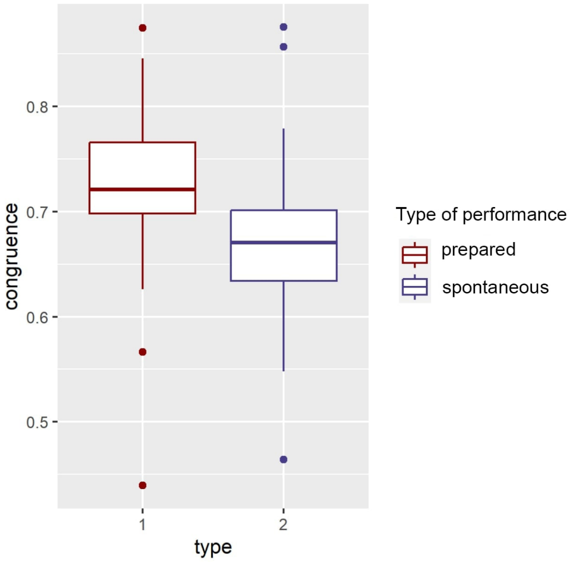


Figure 4. Scope diagram for prepared and spontaneous performances

In addition it is necessary to note the statistical tendency that men are more congruent than women p-value = 0.06 (Mann-Whitney U-test). However, when adjusted for multiple Hill comparisons, the significance drops to 0.18. In this case, it is necessary to consider this issue in the future on a larger sample and a more balanced dataset. In the future, it is necessary to investigate this issue on a larger sample, as well as to investigate the issue related to determining the difference in the level of congruence in various situations related to professional activity, including in the banking sector.

4 Discussion

The proposed method analyzes the congruence of a person in three channels, evaluating the manifestations and dynamics of changes in extreme emotional states of a person during speech. The increase in the number of fake video performances on the Internet, including videos created using machine learning methods, motivated us to test our method on artificially generated videos.

We used speeches by real people on various topics and generated videos (deepfakes). We concluded that the assessment of the congruence of verbal and non-verbal markers of a person's performance allows not only to determine his emotional state, but also to identify videos that replace a human personality (face, voice) or completely generate a model of a person. Thus, ”synthetic” people show low emotional consistency from the first fragments of the video, and during the performance the difference increases between modalities. In turn, the assessment of the person's speech video modality with a replaced face is generally characterized by low dynamics of changes in emotions and a small dispersion between the states of a person in different parts of the video.

We found that the assessment of the human congruence and the analysis of individual modalities opens up opportunities for analyzing the unnatural origin of the video distinguishing an incongruent person from a deepfake.

5 Conclusion

The article presents an approach to the problem of determining the consistency of information simultaneously transmitted by a person through verbal and nonverbal communication channels (congruence). Our approach is based on several machine learning models for sequence analysis. In the course of the work a data set was collected that allowed us to test the hypothesis that a person will have a different level of congruence depending on the degree of preparedness for the performance. The developed approach makes it possible to analyze congruence/incongruence of a person through all three channels, assessing the manifestations and dynamics of changes in extreme emotional states during a speech:

– Anger (negative attitude);
– Calmness (neutral attitude);
– Joy (positive attitude).

In addition, during the development of this approach it was found that the assessment of congruence by the general consistency of human emotions and the analysis of particular indicators allows not only to identify incongruence, but also to indicate the unnatural nature of its origin (to distinguish a simply incongruent person from a deepfake). One of the promising areas of application of the technology is the identification of unreliable or unverified multimedia content on the open Internet, verification and training of HR candidates and contractors for business.

This research is financially supported by The Russian Science Foundation, Agreement №17-71-30029 with co-financing of Bank Saint Petersburg.

References

Abas, A. R., Elhenawy, I., Zidan, M. and Othman, M. (2022). Bert-cnn: a deep learning model for detecting emotions from text. *Computers, Materials & Continua*, **71** (2), pp. 2943–2961.

Amer, M. R., Siddiquie, B., Richey, C. and Divakaran, A. (2014). Emotion detection in speech using deep networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3724-3728.

Amiriparian, S., Pohjalainen, J., Marchi, E., Pugachevskiy, S. and Schuller, B. (2016). Is deception emotional? *An emotion-driven predictive approach. Proceedings of INTERSPEECH-2016*, pp. 2011–2015.

Badugu, S. and Suhasini, M. (2017). Emotion detection on twitter data using knowledge base approach. *International Journal of Computer Applications*, **162** (10), pp. 28-33.

Bazarevsky, V., Kartynnik, Yu, Vakunov, A., Raveendran, K. and Grundmann, M. (2019). BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. *Proceedings of the Computer Vision and Pattern Recognition Workshop on Computer Vision for Augmented and Virtual Reality*.

Burkhardtm F. et al. (2005). A Database of German Emotional Speech. *Interspeech*, pp. 1517–1520.

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, pp. 4171-4186.

Dong, K. J., Guan, Z., Rosenthal, E., Fu, J., Rafailovich, S. and Polak, M. (2022). Detection of (Hidden) Emotions from Videos using Muscles Movements and Face Manifold Embedding.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, **6** (3-4), pp. 169-200.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, **48** (4), pp. 384-392.

Gournay, P., Lahaie, O. and Lefebvre, R. (2018). A canadian french emotional speech dataset. *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 399–402.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770.

Joshi, A. et al. (2016). Are Word Embedding-based Features Useful for Sarcasm Detection? *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1006–1011.

Kirchhubel, C., Stedmon, A. and Howard, D.M. (2013). Analyzing deceptive speech. *Proceedings of EPCE-2013. – Heidelberg: Springer*, **8019**, pp. 134–141.

Knapp, M.L. (1978). Nonverbal Communication in Human Interaction (2nd Edition) Holt. *Rinehart and Winston Inc.*

Kozhakhmet, K., Zhumaliyeva, R., Shoiynbek, A. and Sultanova, N. (2020). Speech Emotion Recognition For Kazakh And Russian Languages. *Applied Mathematics & Information Sciences*, **14** (1). pp. 65–68.

Kwok, I. and Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks.*Proceedings of the AAAI Conference on Artificial Intelligence*, **27** (1), pp. 1621-1622.

Likitha, M.S. et al. (2017). Speech based human emotion recognition using MFCC. *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). IEEE*, pp. 2257–2260.

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M. and Fernández-Martínez, F. (2022). A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. *Applied Sciences*, **12** (1), pp. 327.

Lv, H. et al. (2021). LET-Decoder: A WFST-Based Lazy-Evaluation Token-Group Decoder With Exact Lattice Generation. *IEEE Signal Processing Letters*, **28**, pp. 703-707.

Mollahosseini, A., Hasani, B. and Mahoor, M.H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, **10** (1), pp. 18-31..

Morris, D., Collett, P., Marsh, P. and O'Shaughnessy, M. (1979). Gestures, their Origin and Distribution. *Jonathan Cape Ltd*.

Rana, A. and Jha, S. (2022). Emotion Based Hate Speech Detection using Multimodal Learning.

Savchenko, V.V. and Vasil'ev, R.A. (2014). Analiz emotsional'nogo sostoyaniya diktora po golosu na osnove foneticheskogo detektora lzhi [The analysis of the emotional condition of the an-nouncer on the voice on the basis of the phonetic lie detector]. *Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta – Belgorod State University Scientific Bulletin*, **21** (192), pp. 186–195.

Schoneveld, L., Othmani, A. and Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-Visual emotion recognition. *Pattern Recognition Letters*, **146**, pp. 1-7.

Shan, K., Guo, J., You, W., Lu, D. and Bie, R. (2017). Automatic facial expression recognition based on a deep convolutional-neural-network structure. *IEEE*

15th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 123-128..

Shmyrev, N. (2020). Vosk Speech Recognition Toolkit: Offline speech recognition API for Android, iOS, Raspberry Pi and servers with Python, Java, C# and Node.

Sinko, M. et al. (2022). Method of constructing and identifying predictive models of human behavior based on information models of non-verbal signals. *Procedia Comput Sci.*, **212**, pp. 171–180.

Smetanin, S.I. (2020). Toxic comments detection in Russian. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*, pp. 1149–1159.

Thang, P. C. et al. (2019). An Adaptive Algorithm For Restoring Image Corrupted By Mixed Noise. *Cybernetics and Physics*, **8** (2), pp. 73–82.

Verbitskiy, S., Berikov, V. and Vyshegorodtsev, V. (2022). ERANNs: Efficient Residual Audio Neural Networks for Audio Pattern Recognition. *Pattern Recognition Letters*, **161**, pp. 38-44.

Wikarsa, L. and Thahir, S. N. (2015). A text mining application of emotion classifications of Twitter's users using Naïve Bayes method. *2015 1st International Conference on Wireless and Telematics (ICWT)*, pp. 1-6.

Zhang, S., Pan, X., Cui, Y., Zhao, X. and Liu, L. (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*, **7**, pp. 32297–32304.

Zhengyao, W., Wenzhong, L., Tao, W. (2021). Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition.