INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Hybrid Dataset for Speech Emotion Recognition in Russian Language

Vladimir Kondratenko[1], *Artem Sokolov*[1,3], *Nikolay Karpov*[2,*], *Oleg Kutuzov*[1,4], *Nikita Savushkin*[1], *Fyodor Minkin*[1]

[1]SaluteDevices, Russia, [2]NVIDIA, Armenia,
[3]Laboratory of Algorithms and Technologies for Network Analysis, HSE University, Russia
[4]Department of Mathematics and Mechanics, Moscow State University, Russia

kondrat1997@yandex.ru, artsokol87@gmail.com, karpnv@gmail.com,
oleg.kutuz2018@yandex.com, savushkinnm@yandex.ru, minkin.fyodor@gmail.com

Abstract

We present a new data set for speech emotion recognition (SER) tasks called Dusha. The corpus contains approximately 350 hours of data, more than 300 000 audio recordings of Russian speech, and their transcripts. Therefore it is the biggest open bi-modal data collection with an open license for SER tasks nowadays. This data set is the first speech emotion corpus in Russian, including both crowd-sourced acted and real-life emotions from podcasts, with multiple speakers and scalable data set size. Acted subset has a more balanced class distribution than the unbalanced real-life part consisting of audio podcasts. So the first one is suitable for model pre-training, and the second is elaborated for fine-tuning purposes, model approbation, and validation. This paper describes in detail our collecting procedure, pre-processing routine, annotation, and experiment with a baseline model to demonstrate some actual metrics which could be obtained with the Dusha data set.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

There are many recent studies in human behavior analysis and automatic speech emotion recognition (SER) [1, 2, 3]. Many use various inputs such as speech, video, and transcript as multimodal data. The popular approach of such research is to invent a new neural network architecture and train it on the available data sets and benchmarks [4]. However, some aspects have a negative impact on the process of model training and evaluation. For instance, the small size of the open dataset frequently becomes a bottleneck for research. One more possible shortcoming is biasing between label annotation of the data set and user emotions in the real world [5, 6]. It is highly desirable for a data set to involve as many label evaluators as possible, but practically, it is complicated enough to implement [7]. Another issue is the lack of speaker diversity, leading to the model underperforming when it faces a new speaker in a training set or a real-time speech [8].

These issues with the existing big open data sets that cannot be solved by speech enhancement [9, 10] motivated us to develop a new extensive database with Russian speech [11]. We call it Dusha, which means Soul in Slavonic languages. It is designed to reveal such concepts as peace, openness, and the vast nature of the Eastern-European soul. We believe that our corpus can help to improve results in other languages using cross-corpus study [12] or transfer learning techniques on speech emotion recognition [13, 3]. The data set contains speech recordings and their transcripts, which is why we call it bi-modal.

Two sources of speech are used: acted crowd-sourced records and real-life podcasts in the Russian language. We consider that such a combination of domains is expected in a real-life scenario when a model developer has less data from a target domain and more from another crowd-sourced one. We select the emotions that appear in the dialogue with a virtual assistant most frequently: Anger, Happiness, Neutral emotion, and Sadness.

Each item has been labeled by several annotators using four emotional classes so that markup could be aggregated into one confident label or multi-labeled. Along with the data, we share an aggregation mechanism so that any data scientist can access them to conduct research.

This paper delivered an advanced speech emotion recognition data set with transcription to the open source. Also, it describes approaches and methods for data set collection and markup. All data and processing scripts are released on a GitHub repository[1].

2. Related work

To highlight our contribution, we analyzed existing Speech Emotional databases and compared our benchmarks with those including corpora with the Russian language.

2.1. Emotional speech datasets

The interactive emotional dyadic motion capture database (IEMOCAP) [14] is a widely used multimodal data set that is de facto preferable for modern research comparison in emotion recognition and sentiment analysis. It contains visual data, audio tracks of dialogues, and transcribed text. Besides, this database includes motion data for faces and hands only. Five male and five female semi-professional actors recorded their voices for this data set. IEMOCAP exhibits the balanced distribution of emotions from the following list: happiness, anger, sadness, frustration, and neutral emotion. This material includes about 12 hours of an audio split in 5 dyadic sessions. Although the data set is balanced, its disadvantage is that it is not very extensive and has few speakers involved. Mostly, the benchmark is applicable for model comparing, yet it can cause an issue with precision during evaluation in live speech. It is a common research practice to take a subset of IEMOCAP with four classes of emotions: happiness, sadness, anger, and neutral emotion (where the excitement is combined with happiness) [12]. This set is referred to as IEMOCAP4.

*Work performed while at Salute Devices, Russia.

[1]https://github.com/salute-developers/golos/tree/master/dusha

10.21437/Interspeech.2023-311

The authors of the paper [15] propose an approach to effectively build an extensive, naturalistic emotional database based on podcasts. The majority of the sentences in podcasts are emotionally neutral, so they balance the content that has been collected. The authors implemented a non-trivial procedure to provide accurate pre-processing and confident data labeling. Crowd-sourced platform workers obtained the annotation. They yielded the corpora with over 27 hours of emotional data recorded by 83 speakers and called it MSP-PODCAST. Even though the data set is almost two times larger than IEMOCAP, it still has a size limitation to be employed for modern deep learning-based solutions.

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) database [16] is another human multimodal language benchmark. The data set is the next generation of CMU-MOSI [17] and involves YouTube video recordings with 1000 distinct English speakers, text transcription of audio, and emotion annotation for each utterance. In addition to the size of CMU-MOSEI, one of its strong points is that emotions are not acted. However, the emotion annotation of this benchmark was conducted by only three crowdsourced persons. Potentially, such a few number annotators could lead to a gap in accuracy for the evaluation and include some bias compared to real data, even if they pass special training.

Among widely-spoken languages, Chinese (Mandarin) and Spanish are also covered by numerous data sets. German domain is widely represented in emotion databases too [18], [19]. The most famous one is EmoDB [20].

An attempt to create an enormous repository by joining several various languages was described in [21]. The authors presented a united database that included subsets with English, German, Chinese, Turkish and other languages.

2.2. Datasets in Russian domain

Currently, very few data collections for emotional speech recognition are available in the Russian language.

One of the first attempts to organize a Russian emotional data set is described in [22]. This set of audio utterances and their transcriptions is called the Russian Language Affective speech database (RUSLANA). Students of various Russian universities, participating as speakers, dictated 6.400 utterances with the corresponding emotions.

Russian Multimodal Corpus of Dyadic Interaction for Studying Emotion Recognition (RAMAS) [23] is another Russian language data set that is widely used in research of emotion recognition [8, 24]. Similar to IEMOCAP, it includes acted recordings with 7 hours of emotional speech. The corpus provides video and audio modality, transcripts, motion, and physiology data. It annotated the following emotions: Anger, Sadness, Disgust, Happiness, Fear, and Surprise. Ten actors participated in the recording of the video clips for this benchmark.

One more Russian database which could be employed for SER is Multimodal Russian Corpus (MURCO) [25], which is a part of the Russian National Corpus (RNC). It stores clips from Russian cinematography, TV and radio programs, recordings of usual conversations, etc. Although MURCO has millions of recordings, it has obsolete and unfriendly interfaces for automatic data retrieval. The complete list of emotion classes is not defined.

We consider the problem of large-scale data sets for SER tasks. Moreover, there is a substantial gap in real data in the Russian domain. When faced with natural emotions, the data set would become a framework to conduct research and establish a connection between obtained results in the laboratory and system behavior. In addition, MLS [26] and Golos [11] datasets play a significant part in the automatic speech recognition (ASR) task [27, 13, 28]. Therefore, we decided to collect and share a large multimodal (audio and text) [29] data set in the Russian domain, involving both acted and real-life data. The first part of corpora is suitable for model pre-training, and the second is elaborated for fine-tuning purposes, model approbation, and validation. Most of the sentences in real life are emotionally neutral. Relying on this fact, we do not balance our emotional data, but we provide a mechanism of aggregation to do it.

3. Data acquisition

The Dusha data set consists of two logical parts obtained differently. The first one is collected with the assistance of non-professional actors on a popular crowd-sourcing platform. Yandex Tolloka[2]. Further in the text, we call it "*Crowd domain*" or "*Crowd*". The second part consists of a speech from various emotional podcasts. We call it "*Podcast domain*" or "*Podcast*".

3.1. Crowd subset collection

The text for crowd recordings was chosen from genuine requests which users fulfilled via virtual voice assistant Salute and SmartSpeech[3] - service for speech recognition. The raw data set included tens of millions of recordings and their transcriptions. It is evident that most voice requests involve an urge to do something like "Salute, turn on YouTube", "Salute, sign me up for a hairdresser", and other phrases and talks, which users send to their voice assistant with neutral emotion. To balance our data, we filtered out requests and kept recordings with conversation (chit-chat) because this subset could include more explicit emotional utterances. To do so, we employed the Salute internal intent classifier, which separates various types of voice commands and selects chatter requests when no action except response is required. The resulting subset was several million of utterances.

Next, we applied emotional pseudo-labelling of texts to establish what emotions could be acted for utterances. We employed a simple classifier on top of a BERT-large version of well-known BERT architecture [30], which was trained from scratch internally and could classify our texts for four target sentiments: anger, happiness, sadness, and neutral emotion. The investigation result demonstrates that neutral emotion dominated in many cases. To evaluate our pseudo labels, we conducted a survey on a crowd-sourcing platform where we were asked to label manually a tiny part (˜10.000) of utterances and compare them with classifier results. It shows that our pseudo labels are sufficiently accurate. We use them to sample emotional utterances and decrease the count of neutral recordings.

Next, we performed audio voicing with non-professional actors' help on a crowd-sourced platform. We took pseudo labels predicted in the previous step into account. For each phrase, we set one emotion from the label and one more with similar emotion valence or neutral sentiment. For instance, we organized emotions in pairs positive/neutral, sadness/neutral, anger/sadness, etc.

Thus, the actors had to pronounce the text with one of the emotions from the pair. Also, we described how to voice the

[2] https://toloka.yandex.ru

[3] https://github.com/salute-developers/salute-speech

emotion better.

We obtained 201 850 acted emotions with 2 068 unique speakers where neutral emotion dominates as in real-life situations; however, other classes are pretty balanced. The blue column on Figure 1 represents the time length distribution. As people use their equipment, the quality of audio files differs. Audio can contain background noises such as children, animal voices, or street sounds. The total length is about 255 hours.

3.2. Podcast subset collection

The Podcast subset was designed to diversify data in the Dusha database. Emotions in these recordings are not performed but rather sincere. Furthermore, the distribution of emotions for this data set corresponds better to their distribution in usual human speech. *Podcast domain* is not balanced, and the neutral emotion class substantially outnumbers other classes. Moreover, since acted emotions may differ slightly from spontaneous real-life emotions, we consider it reasonable to keep this subset with natural class distribution in the Dusha. The Podcast could be used for fine-tuning goals and assessing the quality of the model for the production system.

We obtained a topic diversity and included entries on politics, IT, games, relationships, etc. We do not fulfill any specific podcast choosing or filtering and try to cover various conversation topics. Recordings were sliced into 5-second segments by a voice activity detector (VAD) to simplify emotion annotation (See Figure 1 orange color). Text obtained with SmartSpeech solution pre-trained on [11]. A total of 6240 podcasts were used, of which 102 113 samples were selected. The Podcast audio is recorded with professional equipment and has better quality than the Crowd. We normalized files to 16-bit, 16 000 Hz. The total length is greater than 90 hours.

3.3. Post-processing and annotation

To avoid implicit bias in annotation on the crowd-sourcing platform, each person took the training and passed the exam. All participants with a passing score above 80%

Participants listened to the audio only and did not have access to the transcript to evaluate emotions of *Crowd* and *Podcast domain*. Annotators were given instructions to choose their labels using one of the five options:

Positive: The text is spoken with a smile, laughter, admiration, playful tone, or pronounced stresses on words emphasizing the positive.

Neutral: The voice is still and calm; there is no emotion. At the same time, even if the text is negative (for example, “how tired you are”) or positive (for example, “what a fine fellow you are”), this emotion is not expressed in voices; it is necessary to mark the emotion as neutral.

Sadness: The text is pronounced with sadness, melancholy, and a faded voice.

Anger/Irritation: if the text is spoken with anger or annoyance, the user is yelling or speaking through gritted teeth, or there are pronounced stresses on words emphasizing the negative.

Other: The recording is too quiet, hissing, rattling, or has no voice.

To ensure markup quality, each person occasionally got a control task in which we knew the correct label. We named such control tasks ”honeypots”. If an answer to the control task were right, they would continue to mark up. 303 963 recordings were evaluated during annotation, and 1 715 301 emotion labels were accumulated.

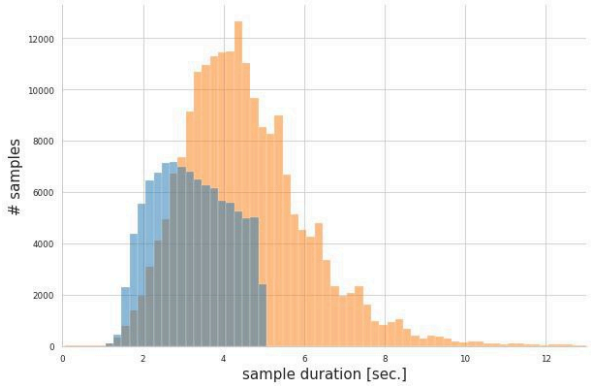

Figure 1: *Single audio length distribution in Dusha corpus: blue - Crowd domain, orange - Podcast domain.*

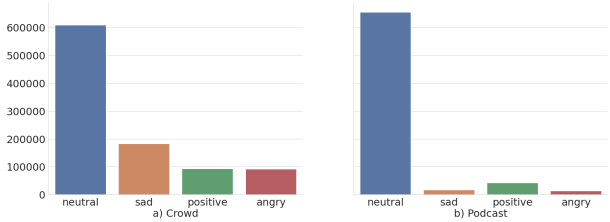

Figure 2: *Class duration [sec.] distribution in Crowd and Podcast subsets. blue - Neutral, orange - Sad, green - Positive, red - Angry.*

4. Dataset overview

4.1. Raw dataset

Our raw metadata includes at least three labels given by independent annotators per sample and several fields for pure emotional markup without aggregation. Independent annotators have an independent opinion about emotion labels. In case of disagreement, more people were involved in marking one sample.

A list of fields of raw metadata is provided below: **wav_path** - relative path to audio file; **annotator_id** - unique id of annotator; **annotator_emo** - emotion mark given by annotator; **annotator_emo** - emotion mark given by annotator; **golden_emo** - emotion mark of control tasks (honeypots); **speaker_text** - original speaker text to pronounce; **speaker_emo** - intentional emotion of the audio; **source_id** - unique id of speaker or podcast;

Metadata stores information about all applicable emotions to each recording, voting results, and other specific data. It enables researchers to explore markup consistency and try various methods to customize markup for data sampling with a particular annotation confidence level. To get data set for machine learning purposes, we have to group labels by audio files and aggregate them into single-labels or multi-labels. We call this ”aggregation mechanism” or ”aggregation”. For an aggregation of raw data, we use Dawid-Skene (DS) algorithm [31] with a confidence threshold to limit the level of agreement. We choose an empirically selected threshold of 0.9. Unlike raw corpus, the subset we get could be employed for the SER model.

The emotion distribution per domain of aggregated annotation is depicted in Figure 2 and also described in Table 1. A list of fields of this metadata is provided below: **wav_path** - relative path to audio file; **emotion** - aggregated emotion mark; **speaker_text** - original text in the audio record; **speaker_emo**

Table 1: *Emotion files distribution after aggregation mechanism using Dawid-Skene algorithm with threshold 0.9*

Domain	Pos	Sad	Ang	Neu	Oth	Total
Crowd	15446	23316	17120	106850	1655	164387
Podcast	5909	1170	2057	81104	222	90462

Table 2: *Amount of data after aggregation mechanism using Dawid-Skene algorithm with threshold 0.9*

Domain	Training files and hours		Test files and hours	
Crowd	150352	188 h. 44 min.	14035	18 h. 29 min.
Podcast	79825	71 h. 23 min.	10637	09 h. 25 min.
Total	230177	260 h. 07 min.	24672	27 h. 54 min.

Table 3: *Experiment results on Dusha benchmark*

Training setup	Crowd test			Podcast test		
	UA	*WA*	*F1*	*UA*	*WA*	*F1*
Dusha	**0.83**	0.76	0.77	**0.89**	0.53	0.54

- intentional emotion of the audio; **source_id** - unique id of speaker. The number of items and duration in the aggregated training and test subsets are represented in Table 2.

4.2. Baseline implementation details

We conduct experiments using the shallow baseline model to simplify the entry threshold for researchers who will benchmark using our data set.

We use standard metrics for SER tasks: macro F1 score (*F1*), Unweighted Accuracy (*UA*), and Weighted Accuracy (*WA*). These validation metrics are calculated on Crowd and Podcast testing sets, created using the Dawid-Skene algorithm with confidence > 0.9.

We train a baseline model from scratch with both Dusha parts (**Crowd** and **Podcast**). Additionally, we train our baseline model on IEMOCAP4 to compare it with other state-of-the-art (SOTA) solutions for speech emotion recognition.

For our experiments, we employ an audio modality only. As input, we pass 64 Mel-filterbank calculated from 20ms windows with a 10ms overlap. Next, features are received during the inference [32] of a simple MobileNetV2[33] based architecture with a self-attention layer described in SAGAN[34]. Input Mel features are passed through a sequence of inverted residual blocks as done in [33] but with custom layers configuration. Then we apply a convolutional self-attention layer followed by a global average pooling. After that, we pass the resulting vector (one number for each feature map) through a fully connected layer to get classification results.

The model is implemented in Pytorch, using the Adam[35] optimizer with a learning rate 0.001, a weight decay of 10^{-6}, and without gradient clipping. We train models 100 epochs with batch size 64.

4.3. Benchmark results

The results of our experiments are presented in Table 3. For all test subsets, *UA* is higher than *WA*. The neutral emotion dominance could explain it, and the corpus includes emotion distribution as people face it. However, each researcher or engineer can filter out emotions as they want. Confusion matrices on the Firure 3 demonstrates that differences between *UA* and *WA* are connected to unbalanced aggregation and the fact of neutral

(a) *Crowd test set.* (b) *Podcast test set.*

Figure 3: *Confusion matrices for baseline experiment.*

emotion dominance.

Our baseline model trained on the IEMOCAP4 subset of IEMOCAP shows 0.59 ± 0.01 unweighted accuracy *UA*, 0.59 ± 0.01 weighted accuracy *WA*, and 0.59 ± 0.01 *macro F1* score with five sessions cross-testing. The actual SOTA results we showed with IEMOCAP were considerably better, but we didn't set the goal to obtain the best metrics. We demonstrated the abilities of the utilized architecture for the famous dataset.

A demonstration notebook was developed to showcase the accuracy of speech emotion recognition achieved by a model trained with our dataset. It is available in our GitHub repository[4], as well as other experiment setups with more accurate metrics.

5. Conclusion

This study introduces the novel speech data set for emotion recognition called "Dusha", and this is the first large dataset in Russian. The data has been taken from two different sources. The first one is 255 hours of audio with text transcriptions, and this is an acted subset obtained and labeled via a crowd-sourcing platform. The second subset is taken from various podcasts, and its size is about 90 hours. Acted part of corpora could be employed for model pre-training or general research purposes in the SER area. The natural one is elaborated for fine-tuning purposes and approbation of solutions.

The distinctive feature of Dusha is that we provide a raw emotional data set and an example of an aggregation mechanism. The Dusha's markup can be aggregated into single labels or multi-labels. The research community can use our example of a label aggregation or set-up in their experiments with customized filtering. We open-sourced a code to benchmark models using Dusha and experimented with a baseline model to demonstrate obtained metrics with default emotion distribution. We shared a demo to prove the excellent quality of emotion recognition we achieved with a solution trained on data of "Dusha". In further works, we want to conduct more experiments with the dataset and compare it deeply with other famous frameworks.

6. Acknowledgements

The publication was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.

[4] https://github.com/salute-developers/golos/tree/master/dusha

7. References

[1] P. Tzirakis, Z. Jiehao, and B. W. Schuller., "End-to-end speech emotion recognition using deep neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[2] A. Velichko, M. Markitantov, H. Kaya, A. Karpov *et al.*, "Complex paralinguistic analysis of speech: Predicting gender, emotions and deception in a hierarchical framework," *Interspeech*, pp. 4735–4739, 2022.

[3] L. Savchenko and A. V Savchenko, "Speaker-aware training of speech emotion classifier with speaker recognition," in *Proceedings of the International Conference on Speech and Computer (SPECOM)*. Springer, 2021, pp. 614–625.

[4] W.-C. Lin and B. Carlos, "An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks," *Intespeech*, 2020.

[5] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, 2018.

[6] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[7] L. Devillers, V. Laurence, and L. Lori, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, 2005.

[8] A. Savchenko and L. Savchenko, "Audio-visual continuous recognition of emotional state in a multi-user system based on personalized representation of facial expressions and voice," *Pattern Recognition and Image Analysis*, vol. 32, no. 3, pp. 665–671, 2022.

[9] I. Shchekotov, P. K. Andreev, O. Ivanov, A. Alanov, and D. Vetrov, "FFC-SE: Fast Fourier Convolution for Speech Enhancement," in *Interspeech*, 2022, pp. 1188–1192.

[10] V. Kuzmin, F. Kravchenko, A. Sokolov, and J. Geng, "Real-time streaming Wave-U-Net with temporal convolutions for multi-channel speech enhancement," *arXiv preprint arXiv:2104.01923*, 2021.

[11] N. Karpov, A. Denisenko, and F. Minkin, "Golos: Russian dataset for speech research," in *Proceedings of Interspeech*, 2021, pp. 1419–1423.

[12] R. Milner, M. A. Jalal, R. W. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[13] A. Sokolov and A. V. Savchenko, "Gender domain adaptation for automatic speech recognition," in *Proceedings of IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 2021, pp. 000 413–000 418.

[14] B. Carlos, B. Murtaza, L. Chi-Chun, K. Abe, M. Emily, K. Samuel, N. C. Jeannette, L. Sungbok, and S. N. Shrikanth, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, no. 42, pp. 335–359, 2008.

[15] R. Lotfian and B. Carlos, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings." *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[16] Z. Amir, L. Paul, Pu, V. Jonathan, P. Soujanya, T. Edmund, C. Erik, C. Minghai, and L.-P. Morency., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018.

[17] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016.

[18] F. Schiel, S. Steininger, and U. Türk, "The SmartKom multimodal corpus at BAS," in *LREC*. Citeseer, 2002.

[19] B. Schuller, A. Dejan, R. Gerhard, W. Matthias, and R. Bernd, "Audiovisual behavior modeling by combined feature spaces,"in proceedings of the," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[20] F. Burkhardt, P. Astrid, R. Miriam, F. S. Walter, and W. Benjamin, "A database of German emotional speech," *Interspeech*, 2005.

[21] M. Gerczuk, A. Shahin, O. Sandra, and B. W. Schuller., "Emonet: A transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing)*, 2021.

[22] V. Makarova and V. A. Petrushin, "RUSLANA: A database of Russian emotional utterances," *Seventh international conference on spoken language processing*, 2002.

[23] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian multimodal corpus of dyadic interaction for affective computing," *Proceedings of International Conference on Speech and Computer (SPECOM)*, 2018.

[24] E. Ryumina, O. Verkholyak, and A. Karpov, "Annotation confidence vs. training sample size: Trade-off solution for partially-continuous categorical emotion recognition." in *Interspeech*, 2021, pp. 3690–3694.

[25] S. Savchuk and A. Makhova, "Multimodal Russian corpus and its use in emotional studies," *Russian Journal of Communication*, 2021.

[26] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Interspeech*, 2020, pp. 2757–2761.

[27] A. Savchenko and Y. I. Khokhlova, "About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems," *Optical Memory and Neural Networks*, vol. 23, no. 1, pp. 34–42, 2014.

[28] A. V. Savchenko, "Phonetic words decoding software in the problem of Russian speech recognition," *Automation and Remote Control*, vol. 74, pp. 1225–1232, 2013.

[29] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh *et al.*, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," *arXiv preprint arXiv:2104.10121*, 2021.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018.

[31] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[32] A. V. Savchenko, "Fast inference in convolutional neural networks based on sequential three-way decisions," *Information Sciences*, vol. 560, pp. 370–385, 2021.

[33] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," 2018.

[34] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980