Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190508

The Automated Method of Collecting and Labeling Data for Speech Emotion Recognition based on Face Emotion Recognition

Aisultan Shoiynbek[1], *Darkhan Kuanyshbay*[2], *Paulo Menezes*[3], *Gustavo Assunção*[3], *Bakhtiyor Meraliyev*[2,*], *Assylbek Mukhametzhanov*[2], *Temirlan Shoiynbek*[1], *and Sergey Sklyar*[4]

[1]School of Digital Technology, Narxoz Univeristy, Almaty, 050035, Kazakhstan
[2]Faculty of Engineering and Natural Sciences, SDU University, Kaskelen, 040900, Kazakhstan
[3]Institute of Systems and Robotics, University of Coimbra, Coimbra, 3030-788, Portugal
[4]General and Applied Psychology Department, Faculty of Philosophy and Political Science, Al-Farabi KazNU, Almaty, 050040, Kazakhstan

Received: 2 May 2025, Revised: 18 Jun. 2025, Accepted: 20 Jul. 2025
Published online: 1 Sep. 2025

Abstract: Speech Emotion Recognition (SER) is vital for enabling natural and effective human–machine interactions, yet its advancement is constrained by the scarcity of richly annotated emotional speech corpora, the laborious nature of manual labeling, and the difficulty of eliciting genuine expressions. We propose an automated data-collection and labeling pipeline that synchronizes video-based facial emotion recognition (FER) with audio capture to annotate speech recordings according to speakers' natural facial expressions. Applying this method, we processed 1 243 YouTube videos (1 058 hours of raw footage) and extracted 218 359 candidate utterances, which—after FER-guided filtering—yielded a high-quality corpus of 45 459 recordings (33 h 15 min of audio) across seven basic emotions in Kazakh (15 076 utterances) and Russian (30 383 utterances). We trained a deep neural network on the combined dataset and achieved 86.84% overall test accuracy, with per-language accuracies of 89.00% (Kazakh) and 85.20% (Russian) for seven-way emotion classification; a support vector machine reached 82.47% under the same conditions. By reducing manual annotation effort by over 80% while maintaining consistent labels, our approach delivers a scalable, language-agnostic solution for generating authentic emotional speech datasets, substantially cutting down on human labor and paving the way for more robust, real-world SER systems.

Keywords: Face emotion recognition, labeling, machine learning, speech emotion recognition

1 Introduction

Emotions play a significant role in human interpersonal interactions and communication. In today's context of rapid developments in robotics and artificial intelligence, one of the most promising research directions is enabling machines to recognize and reproduce human emotions. The primary goal of Emotion Recognition (ER) is to enhance human–machine interaction by enabling systems to understand emotional states.

One of the most active subfields of ER is Speech Emotion Recognition (SER), which involves classifying emotions from speech signals. In recent years, various machine learning algorithms have been applied to this task, including Support Vector Machines (SVM) [1]–[5], Recurrent Neural Networks (RNN) [6]–[8], Hidden Markov Models (HMM) [9], standard Neural Networks (NN) [10], and Gaussian Mixture Models (GMM) [11]. Researchers have also explored hybrid approaches, such as the modified Brain Emotional Learning (BEL) model [12], which integrates Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Multilayer Perceptrons (MLP). Other novel approaches include Gaussian Process (GP) classifiers using multiple kernels [7], and the Voiced Segment Selection (VSS) algorithm [13], which processes voiced signal segments as texture images.

Despite algorithmic advancements, most SER studies rely on emotional speech datasets collected under controlled laboratory conditions. These conventional datasets are often recorded by actors simulating emotions

* Corresponding author e-mail: bakhtiyor.meraliyev@sdu.edu.kz

in front of a microphone [14], which introduces several challenges:

1. The emotions are artificial, meaning the models learn to recognize exaggerated or fake expressions rather than authentic ones.
2. The datasets are limited in size and variety, lacking the quantity and diversity needed for robust real-world applications.
3. Language dependency exists in SER; models trained on one language often perform poorly on another [15], a phenomenon also observed in our prior work on Russian and Kazakh languages [16].
4. Researchers must repeatedly search for or create emotional databases in the target language, which is time-consuming and inefficient.

To overcome these challenges, we propose a novel approach that focuses on collecting audio and video data from natural, unscripted sources such as interviews, news segments, and public events. These sources provide spontaneous emotional expressions. However, manually processing and labeling such data is extremely labor-intensive and impractical at scale. Therefore, there is a strong need for an automated pipeline that can perform data collection, segmentation, and emotion labeling with minimal human intervention.

The goals of this study are twofold:

1. To develop a fully automated pipeline for collecting and labeling speech emotion data using facial emotion recognition (FER) as a supervisory signal.
2. To achieve high classification accuracy (80%) in multiple classes of emotions and languages.

2 Literature Review

With the growing interest in SER, early efforts focused on building large-scale corpora to mitigate data scarcity. For example, [17] compiled a 187-hour dataset, but all participants were instructed to reenact emotions—anger, two levels of happiness, neutral, and sadness—resulting in non-authentic affective expressions. Researchers assembled an extensive emotional speech corpus by recording 2 965 volunteers using their own equipment. The collection encompasses both neutral utterances and five acted emotional states (anger, happy-low-arousal, happy-high-arousal, neutral, and sadness), which introduces substantial variability in room acoustics and microphone quality. Although the dataset's scale is impressive, prompts asking participants to draw on past experiences led to artificial emotion portrayal, casting doubt on ecological validity and downstream model performance. Consequently, while setting an important precedent for dataset size, its reliance on simulated affect highlights the need for annotation methods grounded in genuine emotional responses.

In the study [18], the authors trained and tested their speech emotion classifier using the Berlin [19] and Spanish [20] datasets. Both corpora consist of acted emotional expressions, which limits ecological validity. Despite this, the models reached an average accuracy of 81.1% on the Berlin set and 90.94% on the Spanish set under tenfold cross-validation. The high performance underscores the potential of current modeling techniques but also highlights the need for evaluation on naturalistic data to better gauge real-world applicability.

In [21], the authors construct three Chinese emotional-speech corpora—an acted set of twelve professional performers (six males, six females), a second collection of 51 speakers gathered under unspecified conditions, and a hybrid merging both—totaling 29 000 utterances. Although this scale and balanced gender representation are valuable, the fully manual annotation process demanded substantial time and resources, and the predominance of performed emotions risks over-stylization at the expense of ecological validity. Furthermore, the lack of transparency around the second corpus's recording protocol impedes reproducibility and comparative analysis. Collectively, these limitations highlight the need for more automated, cost-effective methods capable of capturing genuinely spontaneous emotional speech at scale.

In [22], the researchers first recorded 560 emotional speech samples from 16 professional actors—eight males and eight females—aged between twenty-five and sixty-four. They then augmented this acted corpus with the Polish Emotional Speech Database (PESD) [23], a collection of 240 recordings compiled by the Medical Electronics Division at the Łódź University of Technology, and the Polish Spontaneous Speech Database (PSSD) [24], which comprises 748 naturally expressed utterances extracted from unscripted television content such as live discussions, reality shows, and talk programs. While the PSSD's genuine emotional expressions significantly enhance the ecological validity of the combined dataset, its relatively modest size limits the robustness and generalizability of models trained exclusively on spontaneous data. Nevertheless, when an SER model was trained on this blended dataset of acted and spontaneous speech, it achieved an accuracy of 86.14%, underscoring the value of integrating both controlled and naturalistic samples despite the substantial effort required for manual collection and annotation.

In their seminal work [25], researchers introduce the first Chinese-language dataset of spontaneous emotional speech, comprising two hours of annotated segments from 219 speakers. These segments were sourced from 25 films, one television series, and 22 talk shows—media in which emotions are typically acted rather than naturally occurring—yet the authors maintain that the resulting corpus reflects authentic emotional expressions. To support this claim, every emotional segment was manually segmented and labeled by expert annotators, ensuring high fidelity in the categorization of emotional

states. This dataset thus represents a critical resource for advancing naturalistic emotion recognition in Chinese, despite the inherent challenges of deriving spontaneity from performed content.

Two years after the release of their spontaneous-emotion corpus [25], the same research team extended this work with a multimodal emotion recognition framework that integrates facial emotion recognition (FER) and automatic speech recognition (ASR) [26]. They repartitioned the original dataset into training (1,981 clips), validation (243 clips) and test (628 clips) subsets drawn from the same pool of film, television and talk-show segments. For the visual channel, convolutional neural networks were trained on the Static Facial Expression in the Wild (SFEW) [27] and FER2013 [28] benchmarks to model static facial expressions, while the audio channel used ASR-transcribed speech for textual emotion prediction. In their final evaluation, the multimodal system achieved a video-based classification accuracy of 36.56% and a text-based accuracy of 33.84%, demonstrating the viability — but also the challenges — of combining visual and verbal cues for naturalistic emotion recognition.

3 Materials and Methods

This section describes the architecture of the proposed automated method for collecting and labeling speech emotion data. The overall pipeline consists of the following sequential objectives:

- –Search and download video data from interviews, news, etc.
- –Detect speech in the video.
- –Segment the video into speech-containing parts.
- –Convert the segmented video to an audio file.
- –Recognize emotion using facial expressions from the video.
- –Recognize emotion using speech from the audio.
- –Compare recognized emotions from face and speech.
- –If the emotions match, assign the label and save the audio file in the database.

Each step in the proposed method is designed to be automated. The researcher needs only to configure basic parameters, such as language-specific models for speech detection. Figure 1 presents an overview of the system architecture.

3.1 Video Parser

The first component of the pipeline is responsible for acquiring videos containing natural emotional expressions. YouTube was selected as the primary video source due to its vast content availability. An existing open-source tool from GitHub [29] was used to download

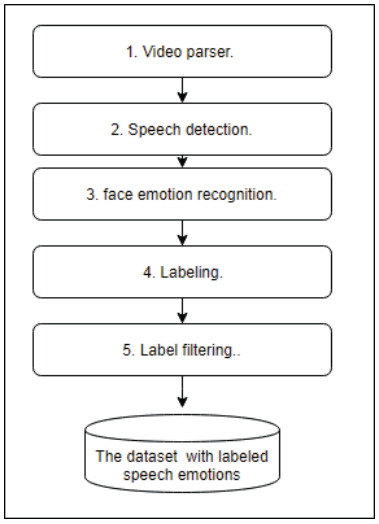


Fig. 1: Architecture of the proposed method for collecting and labeling speech emotion data.

videos. The tool supports multiple modes for retrieving video links:

1. Search for playlists using a single keyword.
2. Search for playlists using multiple keywords.
3. Extract video links from a specific target playlist.
4. Search for individual video links using multiple keywords.
5. Search for individual video links using a single keyword.

The user needs to provide a list of relevant keywords or a target playlist URL to download videos that likely contain natural emotional expressions. Figure 2 illustrates the available parsing modes.

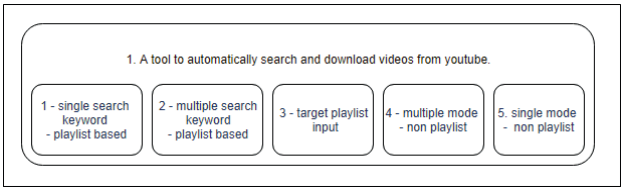


Fig. 2: Modes of operation in the video parser module.

3.2 Speech Detection

The speech detection module extracts segments containing human speech from video files. The process consists of six sequential steps:

1. **Audio Extraction:** The audio track is extracted from a video file in `.wav` format using the `moviepy` Python library.

2. **Audio Conversion:** The extracted `.wav` file is converted into a unified format using the FFmpeg codec. The target format specifications are: `pcm_s16le` (16-bit), mono channel, and a sample rate of 16 kHz.
3. **Audio Splitting:** The audio file is split into 1-second chunks. For instance, an audio file of 2 seconds and 36 milliseconds is divided into three parts: 0–1s, 1–2s, and 2–2.36s.
4. **Feature Extraction:** Mel Frequency Cepstral Coefficients (MFCCs) are extracted from each chunk to serve as input features for speech classification.
5. **Speech Prediction:** Each chunk is passed through a pre-trained speech detection model [30], which identifies whether the chunk contains speech.
6. **Result Storage:** The seconds with detected speech are stored as a list of serial numbers to be used in subsequent steps.

3.3 Face Emotion Recognition and Labeling

Following speech detection, the next module focuses on identifying emotions from facial expressions in the video frames corresponding to detected speech. The full process is illustrated in Figure 3.

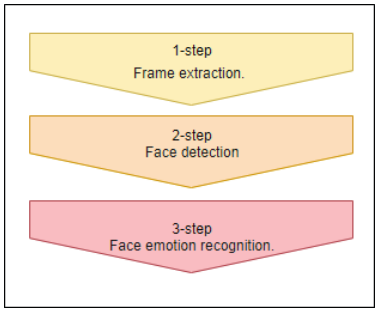


Fig. 3: The face emotion recognition (FER) process.

3.3.1 Frame Extraction

Video is a sequence of images shown in rapid succession. Standard video plays at 25 frames per second (fps), but for this application, only 10 fps were extracted to optimize processing. The `moviepy.editor` Python library was used to extract frames at 100ms intervals from video segments with detected speech.

3.3.2 Face Detection

Each extracted frame was analyzed using the Viola-Jones method based on Haar cascades [31]. The `OpenCV` library was employed to detect faces, ensuring that only a single face per frame is processed to avoid ambiguity. Frames with multiple faces (e.g., Figure 4) were excluded. Detected faces were cropped and resized to 64x64 pixels for the emotion recognition model (Figure 5).

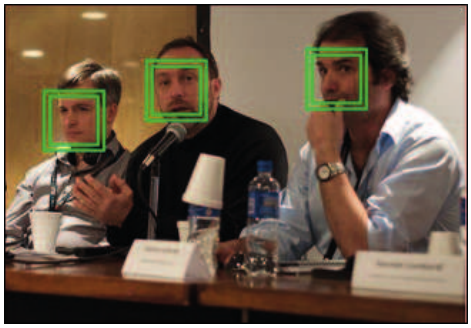


Fig. 4: Example of a frame with multiple detected faces.

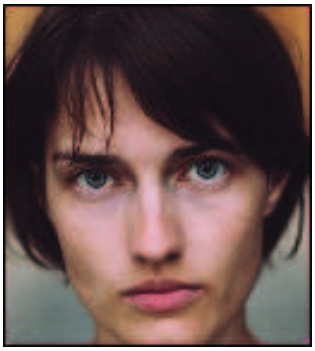


Fig. 5: Cropped rectangle containing a single detected face.

3.3.3 Emotion Recognition

For FER, a CNN model [32] pre-trained on the FER2013 dataset [28] was used. The model outputs probabilities for seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Accuracy on the test set was 66%, which is acceptable due to majority-voting across 10 frames per second.

3.4 Labeling Process

3.4.1 Video Segmentation

Video segments are determined using a logical operator that separates frames when the temporal gap between

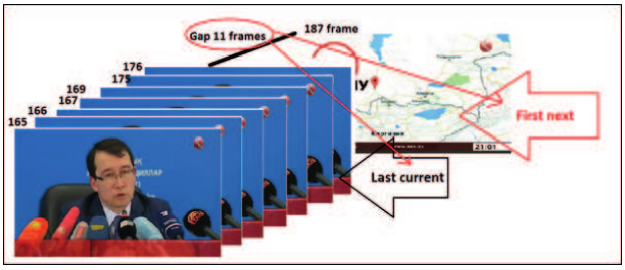


Fig. 6: Detection of first and last frames in a segment.

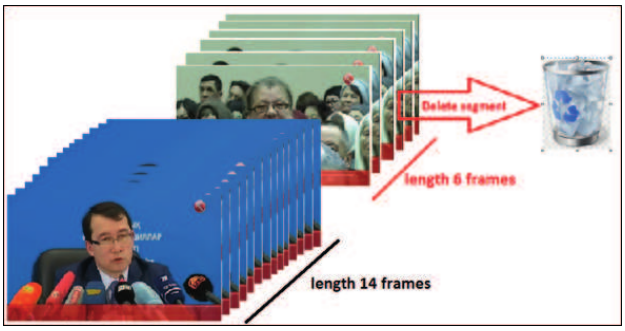


Fig. 7: Segment splitting and majority-vote labeling.

adjacent frames exceeds 10 frames (Figure **??**). Segments shorter than 11 frames are discarded to reduce error propagation.

3.4.2 Emotion Classification of Segments

If a segment exceeds 3 seconds, it is divided into smaller sub-segments to account for emotion variability. Each frame in the segment is labeled individually, and the final emotion label for the segment is determined via majority voting.

3.4.3 Saving Labels and Data

Labeled segments are exported as audio files using the `pydub` library and saved in folders corresponding to their emotion classes.

3.5 Label Filtering

Two major sources of label confusion are mismatched face and voice emotions, and off-screen voiceovers. To address these, an auxiliary SER model [33] with 85.6% accuracy was applied to revalidate each labeled audio file. Only samples with consistent emotion labels across modalities were retained.

The SER model was trained on six benchmark datasets: SAVEE [34], EMOVO [35], RML [36], EMODB [19], ELRA-S0329 [20], and RAVDESS [37]. It combines VGGVox [38] features with an SVM classifier [39].

Figure 8 illustrates the complete pipeline of our automated method for collecting and labeling speech emotion data.

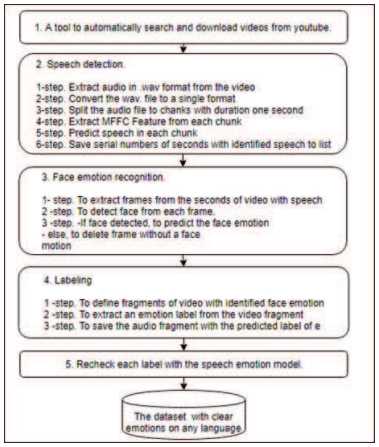


Fig. 8: Full pipeline of the automated method for collecting and labeling speech emotion data.

3.6 Experimental Setup

3.6.1 Data Collection using Video Parser

To build a large-scale emotional speech dataset, a total of 1,243 videos were collected from YouTube using the custom-built Video Parser. Of these, 343 videos were in the Kazakh language and 900 in Russian. The sources included TV shows, news interviews, and real-world recordings.

The Russian-language sources consisted of:

- **DNA** – A program about paternity verification.
- **Wait for Me** – Reunification of missing family members.
- **House-2** – Reality TV centered on relationship-building.
- **Windows** – Scandal-driven reality show.
- **Seems Groped** – A game show involving tactile guessing.

Kazakh-language sources included:

- **Sorry Me** – A show built around gossip and intrigue.
- **Literal Truth** – Investigative journalism on controversial topics.
- **What Is It** – Kazakh adaptation of ”Seems Groped.”

Additionally, clips from traffic stops with emotional escalation and interviews with celebrities were used. The total video duration was 452 hours in Kazakh and 753 hours in Russian.

Criteria for Show Selection:

1. Scenarios should reflect real-life emotional interactions.
2. Content must be unscripted and genuine.
3. Rich emotional expressions should be present.

Scripted movies and series were intentionally excluded as they often involve acted (not natural) emotional expressions. Table 1 provides a breakdown of the collected Kazakh and Russian speech emotion datasets before filtering.

Table 1: Collected datasets in Kazakh and Russian (pre-filtering)

Emotion	Kazakh	Russian
Anger	5,165	14,873
Disgust	3,947	982
Happiness	14,343	15,740
Neutral	40,065	81,072
Sadness	3,965	18,107
Fear	4,335	13,613
Surprised	336	1,816
Total	**72,156**	**146,203**

3.6.2 Label Extraction

The label extraction phase was computationally intensive, requiring approximately six days on a high-performance computer with the following specifications:

– Intel(R) Core(TM) i7-7700 @ 3.60GHz, 8 cores
– 47 GB RAM
– NVIDIA GeForce GTX1060 (6GB)

Notably, emotions such as *Surprise* and *Disgust* were rarely detected. This is consistent with the challenges of capturing authentic surprise in real-life scenarios.

4 Results and Discussions

4.1 Label Filtering and Final Dataset Statistics

After the automatic filtering process using the auxiliary SER model, the final number of labeled samples was significantly reduced to ensure label consistency. Table 2 presents the cleaned datasets for Kazakh and Russian.

Table 2: Filtered emotional speech datasets

Emotion	Kazakh	Russian
Anger	548	1,131
Disgust	269	123
Happiness	4,066	3,598
Neutral	9,019	20,553
Sadness	286	2,005
Fear	872	2,900
Surprised	16	73
Total	**15,076**	**30,383**
Duration	11h 10m	22h 5m

Based on these cleaned datasets, four main datasets were derived:

– **Dataset 1:** Kazakh-only (Table 2)
– **Dataset 2:** Russian-only (Table 2)
– **Dataset 3:** Merged Kazakh and Russian (with all emotion classes) (Table 3)
– **Dataset 4:** Same as Dataset 3, excluding *Disgust* and *Surprised* (Table 4)

Table 3: Dataset 3: Combined Kazakh + Russian dataset (all classes)

Emotion	Utterances
Anger	1,679
Disgust	392
Happiness	7,664
Neutral	29,572
Sadness	2,291
Fear	3,772
Surprised	89
Total	**45,459**
Duration	33h 15m

Table 4: Dataset 4: Combined dataset (excluding Disgust and Surprised)

Emotion	Utterances
Anger	1,679
Happiness	7,664
Neutral	29,572
Sadness	2,291
Fear	3,772
Total	**44,978**
Duration	32h 51m

All datasets were split into training, development, and test sets using an 80/10/10 proportion.

4.2 Feature Extraction and DNN Model

For feature extraction, the VGGVox model [38] was used to convert each audio file into a 1024-dimensional vector.

Multiple classifiers were trained, including SVM, Logistic Regression, Random Forest, K-means, Decision Tree, and a Deep Neural Network (DNN). The DNN achieved the highest accuracy.

DNN Architecture: figure 9 illustrates our deep neural network architecture, which begins with batch normalization applied to the input and proceeds through eight fully connected hidden layers of decreasing size—1024, 512, 256, 128, 64, 32, 16 and 8 neurons—each using ReLU activation. A dropout layer with rate 0.5 is inserted between the fourth and fifth hidden layers to mitigate overfitting. The network terminates in a two-neuron output layer with softmax activation, and all weights are initialized according to the Glorot uniform scheme [40].

Training Configuration:

–Optimizer: Stochastic Gradient Descent (SGD)
–Learning rate: 0.11
–Loss function: Binary Cross-Entropy (Logloss) [41]
–Momentum: 0.1
–Batch size: 256
–Epochs: 100

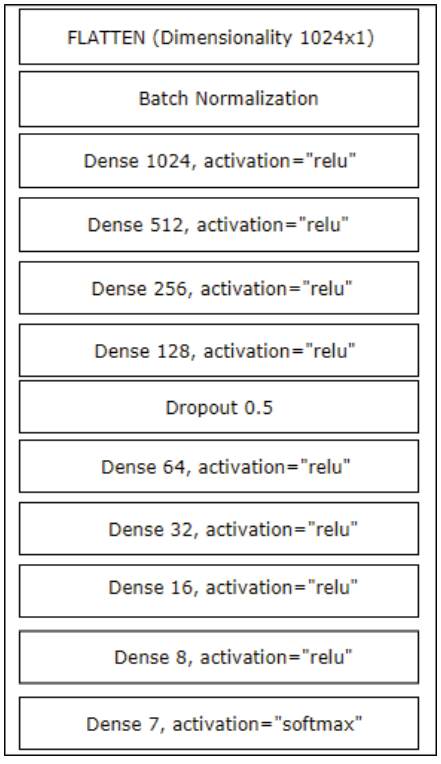


Fig. 9: Architecture of the proposed DNN model.

4.3 Network Learning and Accuracy Results

Each of the four datasets described earlier was used to train and evaluate multiple classifiers, including a Deep Neural Network (DNN). Table 5 summarizes the classification accuracy obtained on the test set for all classifiers across the datasets.

Detailed evaluation metrics such as confusion matrices, precision, recall, and F1 scores for the DNN classifier on each dataset are presented in the following tables (Figure 10).

4.4 Discussions and Future Work

The evaluation results, including F1 scores and confusion matrices, clearly demonstrate the influence of dataset size on classification performance. Specifically, emotions such as *Neutral* and *Happiness*, which had the largest number of samples, achieved the highest recognition rates across all datasets—most notably in Dataset 3.

This trend underlines a crucial insight: the success of Speech Emotion Recognition (SER) models is strongly dependent on the amount of available labeled data. Emotions with limited training examples, such as *Surprised* and *Disgust*, consistently showed lower recognition accuracy and were ultimately excluded in Dataset 4 to reduce class imbalance and noise.

The automated pipeline proposed in this work offers a scalable way to generate large emotional speech datasets, particularly for underrepresented languages like Kazakh. This automation allows researchers to selectively target and expand underrepresented emotion categories in future iterations.

Future Work:

–Enhance the video parser with advanced filtering for context-specific emotional scenes.
–Fine-tune the face and speech emotion recognition models with domain-specific data.
–Introduce multilingual alignment strategies for improved emotion consistency across languages.
–Explore transformer-based architectures for emotion classification.
–Apply the pipeline to collect data from social media and real-time streams for generalization.

5 Conclusion

Speech Emotion Recognition (SER) remains a critical task within the broader field of Artificial Intelligence, particularly in the context of human-computer interaction. This paper addressed three major challenges in SER: the scarcity of labeled emotional speech data across languages, the time-intensive nature of manual labeling, and the difficulty of capturing authentic emotional expressions.

Table 5: Accuracy on the test set for different classifiers

Classifier	Dataset 1 (%)	Dataset 2 (%)	Dataset 3 (%)	Dataset 4 (%)
SVM	86.95	80.00	82.47	84.38
Logistic Regression	59.88	61.28	60.37	61.09
Random Forest	65.36	67.58	64.23	67.96
K-means	58.35	55.43	56.27	56.98
Decision Tree	62.93	61.23	59.88	61.97
DNN	**89.00**	**85.20**	**86.84**	**88.56**

(a) Dataset 1

	anger	happiness	neutral	sadness	fear	Recall	Precision	F1 score
anger	32	18	3	0	1	0.5818	0.6956	0.6339
happiness	8	354	32	2	9	0.8697	0.8697	0.8696
neutral	2	15	873	5	6	0.9678	0.9407	0.954
sadness	0	3	8	11	4	0.3793	0.4782	0.4229
fear	1	15	6	4	62	0.7126	0.7126	0.7125
total						0.5675	0.6183	0.5894

(b) Dataset 2

	anger	happiness	neutral	sadness	fear	Recall	Precision	F1 score
anger	64	15	3	0	31	0.5663	0.8205	0.67
happiness	7	230	36	2	85	0.6371	0.7098	0.6715
neutral	1	37	1975	18	25	0.9606	0.9440	0.9522
sadness	0	1	38	141	21	0.7014	0.8057	0.7502
fear	5	37	35	10	181	0.6753	0.5186	0.5869
total						0.5058	0.5426	0.5186

(c) Dataset 3

	anger	happiness	neutral	sadness	fear	Recall	Precision	F1 score
anger	102	36	12	0	16	0.6071	0.7286	0.6623
happiness	18	608	63	5	70	0.7927	0.7845	0.7885
neutral	8	34	2772	26	22	0.9682	0.9374	0.9526
sadness	0	9	49	153	16	0.6522	0.75	0.6978
fear	10	80	45	13	227	0.6005	0.6486	0.6236
total						0.5565	0.648	0.5882

(d) Dataset 4

	anger	happiness	neutral	sadness	fear	Recall	Precision	F1 score
anger	127	29	4	0	8	0.756	0.7938	0.7744
happiness	20	598	59	8	82	0.7797	0.8147	0.7968
neutral	6	33	2762	35	27	0.9647	0.9518	0.9582
sadness	0	7	34	163	26	0.7087	0.7581	0.7329
fear	7	67	43	9	252	0.6666	0.638	0.652
total						0.7751	0.7913	0.7828

Fig. 10: Confusion matrices and performance metrics of the models on 4 datasets.

To tackle these challenges, we proposed an automated method for collecting and labeling emotional speech data using Face Emotion Recognition (FER) as a supervisory mechanism. The method comprises five modular components and enables scalable, multilingual dataset generation.

Using this pipeline, we collected a comprehensive dataset of emotional speech in Kazakh and Russian. The source material included highly emotional video content such as interviews, news reports, and real-life dramatic moments. In total, 1,243 video files from YouTube—amounting to 1,058 hours of raw footage—were processed. The method yielded 218,359 labeled utterances, which were filtered down to 45,459 high-quality samples comprising 33 hours and 15 minutes of audio.

We benchmarked several classifiers on the final dataset, with the Deep Neural Network (DNN) model achieving a test accuracy of 86.84%. To our knowledge, this is the largest and first automatically collected emotional speech corpus with natural emotions for the Kazakh and Russian languages.

The results highlight the viability of our automated approach for SER dataset creation, paving the way for further advances in emotion-aware AI systems.

Acknowledgement

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP22786670.

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

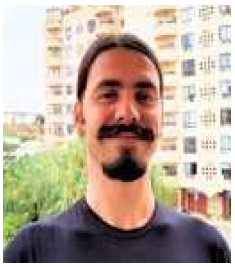
References

[1] Milton A, Roy SS, Selvi TS. SVM scheme for speech emotion recognition using MFCC feature. Int. J. Comput. Appl. 2013;69.
[2] Sree GS, Chandrasekhar P, Venkateshulu B. SVM based speech emotion recognition compared with GMM-UBM and NN. IJESC. 2016.
[3] Melki G, Kecman V, Ventura S, Cano A. OLLAWV: Online learning algorithm using worst-violators. Appl. Soft Comput. 2018;66:384–393.
[4] Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. Int. J. Smart Home. 2012;6:101–108.
[5] Peipei S, Zhou C, Xiong C. Automatic speech emotion recognition using support vector machine. IEEE. 2011;2:621–625.
[6] Alex G, Navdeep J. Towards end-to-end speech recognition with recurrent neural networks. ICML. 2014;32.
[7] Chen S, Jin Q. Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks. Brisbane, 2015.
[8] Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. Asia Pacific. 2017:1–4.
[9] Ingale AB, Chaudhari D. Speech emotion recognition using hidden Markov model and support vector machine. Int. J. Adv. Eng. Res. Stud. 2012:316–318.
[10] Sathit P. Improvement of speech emotion recognition with neural network classifier using speech spectrogram. IWSSIP. 2015:73–76.
[11] Martin V, Robert V. Recognition of emotions in German speech using Gaussian mixture models. LNAI. 2009;5398:256–263.
[12] Sara M, Saeed S, Rabiee A. Speech Emotion Recognition Based on a Modified Brain Emotional Learning Model. Biol. Inspired Cogn. Arch. 2017;19:32–38.
[13] Yu G, Eric P, Hai-Xiang L, van den HJ. Speech emotion recognition using a voiced segment selection algorithm. ECAI. 2016;285:1682–1683.
[14] Ververidis D, Kotropoulos C. A state of the art review on emotional speech databases. 1st Richmedia Conf., Lausanne, 2003:109–119.
[15] Rajoo R, Aun CC. Influences of languages in speech emotion recognition. ISCAIE. 2016.
[16] Shoiynbek A, et al. This article.
[17] Smith AJ, et al. Crowdsourcing emotional speech. ICASSP. 2018.
[18] Kerkeni L, et al. Automatic speech emotion recognition using machine learning. IntechOpen. 2019.
[19] Burkhardt F, et al. A database of German emotional speech. Interspeech. 2005;5:1517–1520.
[20] ELRA. Emotional speech synthesis database s0329. 2012. https://catalogue.elra.info/en-us/repository/browse/ELRA-S0329/
[21] Huang C, et al. Practical speech emotion recognition based on online learning. Math. Probl. Eng.
[22] Kamińska D. Emotional Speech Recognition Based on Committee of Classifiers. Stat. Mach. Learn. for Human Behaviour Analysis. 2019.
[23] Ślot K, et al. Emotion recognition with Poincare mapping. ICAISC, Zakopane. 2019:886–895.
[24] Arruti A, et al. Feature selection for SER in Spanish and Basque. PLoS ONE. 2014;9:e108975.
[25] Bao W, et al. Building a Chinese natural emotional audio-visual database. ICSP. 2014:583–587.
[26] Li Y, et al. MEC 2016: Multimodal Emotion Recognition Challenge. CCPR. 2016.
[27] Dhall A, et al. Static facial expression analysis in tough conditions. ICCV Workshops. 2011:2106–2112.
[28] FER2013 dataset. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
[29] YouTube Parser Tool. https://github.com/spidezad/Youtube-Videos-Search-and-Download
[30] Shoiynbek A, et al. This article.
[31] Lienhart R, et al. Analysis of detection cascades for object detection. PRS. 2003:297–304.
[32] FER Model. https://github.com/omar178/Emotion-recognition
[33] Assunção G, et al. Premature overspecialization in emotion recognition. AES Int. Conf. Audio Forensics. 2019.
[34] Haq S, Jackson PJB. Speaker-dependent audiovisual emotion recognition. AVSP. 2009.
[35] Costantini G, et al. EMOVO corpus: Italian emotional speech database. LREC. 2014.
[36] Xie Z, Guan L. Multimodal information fusion for emotion recognition. ICME. 2013.
[37] Livingstone SR, Russo FA. RAVDESS: North American emotional speech and song. PLoS ONE. 2018.
[38] Nagrani A, et al. VoxCeleb: Large-scale speaker ID dataset. INTERSPEECH. 2017.
[39] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM TIST. 2011;2(3):1–27.
[40] Glorot X, Bengio Y. Difficulty of training deep feedforward NNs. AISTATS. 2010;9:249–256.
[41] Zhang T. Solving large-scale linear prediction problems with SGD. ICML. 2004:116.

Aisultan Shoiynbek holds a PhD, a Master's, and a Bachelor's degree in Computational Engineering and Software. As Project Leader and Leading Research Fellow, he coordinates all management and organizational activities for the project. His research spans AI-based deception detection, speech emotion recognition for Kazakh and Russian, automatic speech recognition systems for low-resource languages using connectionist temporal classifiers, robust spectral audio feature extraction, and optimization algorithms for neural network models.

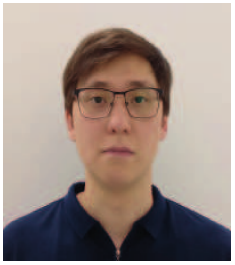
Darkhan Kuanyshbay received his PhD degree in Computer Science and also holds a Master's degree in Information Systems as well as a Bachelor's degree in Computer Systems Processing and Management. As Co-Project Leader and Chief Research Fellow, he is responsible for defining and executing the conceptual framework of the project. His research focuses on speech data collection systems, automatic speech recognition for the Kazakh language—using connectionist temporal classifiers and transfer learning—voice identification techniques, emotion recognition in speech, and optimization algorithms for deep neural network models in speech processing.

Paulo Menezes received his PhD in Electrical and Computer Engineering (Informatics) from the University of Coimbra, following an MEng in Systems and Automation and a BEng in Electrical Engineering (Informatics) at the same institution. Since July 2024, he has been an Associate Professor in the Department of Electrical and Computer Engineering at the University of Coimbra's Faculty of Sciences and Technology, teaching Computer Architecture, Computer Graphics and Augmented Reality, and Interactive Systems and Robotics. As a senior researcher at the Institute of Systems and Robotics and head of its Immersive Systems and Sensory Stimulation Laboratory, he leads work on human–robot interaction, affective computing, AR/VR applications for therapy and rehabilitation, and serious games for emotional learning; he also serves as WP lead for Robotic Social Interaction in the H2020 RISE Lifebots project, advises multiple PhD candidates, has published extensively in top journals and conferences, contributed to major European research initiatives, and acts as a peer reviewer and editorial board member for leading robotics and AI journals .



Gustavo Assunção received his M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from the University of Coimbra, Portugal, in 2019 and 2024, respectively. Since 2018, he has been a Senior Researcher at the Immersive Systems and Sensory Stimulation Lab (IS3L) of the Institute of Systems and Robotics, University of Coimbra, where he focuses on biologically-inspired deep learning, computational neuroscience, neurocognitive computing, and the development of bio-inspired mathematical foundations for artificial intelligence. He has authored numerous articles in leading journals and conferences and serves as a reviewer for several international scientific publications.

Bakhtiyor Meraliyev is a PhD candidate in Computer Science at SDU University (2024–2027) and holds a Master's degree in Computer Science. His research interests include ML, data science, NLP, predictive analytics, and intelligent data analysis. As a Senior Research Fellow, he is responsible for developing algorithms and executing the project's current tasks. He has authored publications in leading international conferences—covering topics such as HR analytics, computer vision in educational contexts, and NLP-based social media analysis



Assylbek Mukhametzhanov received his Bachelor's degree in Information Systems and is currently pursuing a Master's degree in Data Engineering at SDU University, Kaskelen, Kazakhstan. His technical expertise and research interests encompass natural language processing, data analysis, web scraping, backend development, and frontend web application design. As a Junior Research Fellow, he is responsible for data collection, execution of ongoing project tasks, and maintenance of project documentation.

Temirlan Shoiynbek received his Bachelor's degree in Mechanical Engineering Technology and his Master's degree in Information Systems. As a Research Fellow, he is responsible for creating and administering the project website and developing lie-detection algorithms. He also serves as a Senior Lecturer at Narxoz University's Faculty of Digital Technologies.

Sergey Sklyar received the Candidate of Medical Sciences degree in Psychiatry from the Republican Scientific and Practical Center for Mental Health (2007–2010). His research interests encompass suicide-prevention methodologies, psychotherapeutic and psycho-pedagogical approaches across the lifespan, data analysis and ethical oversight in mental-health interventions, digital cognitive training for children with ADHD, and the assessment of quality of life in older adults. He has published numerous research articles in reputed international journals and conference proceedings in psychiatry and psychology, and serves as a reviewer and editor for scientific journals.