applied sciences

MDPI

Article

Polish Speech and Text Emotion Recognition in a Multimodal Emotion Analysis System

Kamil Skowroński [1,*], Adam Gałuszka [1,*] and Eryka Probierz [1,2]

[1] Department of Automatic Control and Robotics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland; erykaprobierz@gmail.com

[2] Lukasiewicz Research Network, Institute of Innovative Technologies, EMAG, 40-189 Katowice, Poland

* Correspondence: kamil.skowronski@polsl.pl (K.S.); adam.galuszka@polsl.pl (A.G.)

Abstract: Emotion recognition by social robots is a serious challenge because sometimes people also do not cope with it. It is important to use information about emotions from all possible sources: facial expression, speech, or reactions occurring in the body. Therefore, a multimodal emotion recognition system was introduced, which includes the indicated sources of information and deep learning algorithms for emotion recognition. An important part of this system includes the speech analysis module, which was decided to be divided into two tracks: speech and text. An additional condition is the target language of communication, Polish, for which the number of datasets and methods is very limited. The work shows that emotion recognition using a single source—text or speech—can lead to low accuracy of the recognized emotion. It was therefore decided to compare English and Polish datasets and the latest deep learning methods in speech emotion recognition using Mel spectrograms. The most accurate LSTM models were evaluated on the English set and the Polish nEMO set, demonstrating high efficiency of emotion recognition in the case of Polish data. The conducted research is a key element in the development of a decision-making algorithm for several emotion recognition modules in a multimodal system.

Keywords: speech emotion recognition; mel spectrogram; polish text emotion analysis; multimodal emotion recogntion; social robots

check for updates

Citation: Skowroński, K.; Gałuszka, A.; Probierz, E. Polish Speech and Text Emotion Recognition in a Multimodal Emotion Analysis System. *Appl. Sci.* **2024**, *14*, 10284. https://doi.org/10.3390/app142210284

Academic Editors: J. Ernesto Solanes and Luis Gracia

Received: 15 October 2024
Revised: 4 November 2024
Accepted: 6 November 2024
Published: 8 November 2024

1. Introduction

The diversity of the human individual has an impact on many aspects of our lives. Biological, cultural, or national diversity, especially in the process of interpersonal communication, creates problems in mutual understanding and expressing emotions. Starting with the mundane problems of everyday life in relationships between partners or between children and parents, we often make mistakes in understanding the expressiveness of emotions and the behavior of the other person. Situations become more serious when, in the face of experiencing a tragedy or falling into various traumas, we lose the ability to express and, at the same time, understand human emotions. We even talk about a disease called alexithymia [1]. It is obvious that in such situations we seek help from specialists, but as we enter the digital world more and more, we also expect to understand contacts with artificial intelligence software or futuristic robots, which are slowly becoming our everyday life.

Social robotics, which aims to create robots capable of interacting naturally with humans, is becoming increasingly advanced thanks to modern artificial intelligence technologies. One of the key elements of these interactions is emotion recognition, which allows robots to better understand the needs and reactions of users. Recent advances in artificial intelligence, especially in the field of deep learning and neural networks, have significantly improved the ability to recognize emotions through multimodal approaches. Such approaches integrate different data sources, such as images, sounds, and text, which allows for a more comprehensive understanding of emotions, taking into account facial expressions, tone of voice, biomedical data, and analysis of the content of utterances [2]. Despite these advances, the challenges associated with recognizing emotions in Polish, both in speech and text, remain significant [3]. AI models are often trained on English-language resources [4,5], which leads to limited accuracy in the case of languages such as Polish, where grammatical complexity and specific phonetic features can make analysis difficult. A particular deficiency in the Polish context is the dearth of large, diverse datasets that could be used to effectively train emotion recognition models. In the case of speech, these challenges also include variability in accent, intonation, and dynamics of utterance, while in texts, it is particularly difficult to analyze emotions hidden in the subtleties of syntax and context.

The main objective of this work is to evaluate the latest and most precise deep learning models for the problem of speech emotion analysis using Mel spectrograms on a dataset based on the Polish language. The primary goal of this work is to show the differences in emotion recognition in text and speech, and propose and implement a dual-track audio module based on the indicated algorithms in a multimodal system for analyzing the emotions of social robots. The system itself can become a specific tool for collecting data in Polish for the problem of recognizing emotions in speech. Thanks to the feedback, in which the user of this system assesses the correctness of the recognized emotion, it is possible to create a rich set of data and these data for which emotions are not artificially induced.

2. Related Works

2.1. Social Robotics and Human-Robot Interaction

Social robotics has been around since the 1990s, when one of the first social robots was created at MIT's AI Lab [6]. Kismet was designed to engage in face-to-face social interactions inspired by the exchanges between an infant and a caregiver. As social robotics developed, the aforementioned concept of HRI (human–robot interaction) emerged. HRI is a field concerned with the design, understanding, and evaluation of systems that include robots and with which humans interact [7]. This field fits perfectly into the area of social robots, which, thanks to a better understanding of human emotions, can constantly improve and improve communication with people.

In recent years, with the development of artificial intelligence algorithms, the slow stabilization of industrial robotics, and the demand for the use of robots in everyday life, there has been a rapid development of social robotics. Social robotics has applications in many areas, including healthcare [8], education [9], daily life management, and support for the elderly [10]. These areas illustrate how social robots can positively impact various aspects of people's lives and support them in everyday tasks.

2.2. Speech and Text Emotion Recognition

The concept of basic emotions has evolved over time, with significant contributions from psychologists such as Paul Ekman, who identified six basic emotions in the 1970s: happiness, sadness, fear, anger, surprise, and disgust. This framework was influenced by earlier theories, including William James and Carl Lange's ideas about the physiological basis of emotion and Charles Darwin's work on emotional expressions as evolutionary adaptations [11]. Over the years, researchers have expanded on these ideas, recognizing the complexity of emotions and the role of cultural factors while also recognizing the importance of these basic emotions as a basis for understanding human affective experiences.

Emotion and its recognition has become one of the challenges and goals of social robotics. Emotion recognition refers to the process of detecting and interpreting human emotions using cues such as facial expressions, tone of voice, body movements, or physiological signals. Speech emotion recognition analyzes features such as pitch, tone, and speech patterns to identify emotions, while text emotion recognition uses NLP (natural language processing) to assess emotional content based on word choice, structure, and meaning in written text [12]. These methods aim to capture emotional states during real-time interactions, increasing the effectiveness of human–robot communication.

2.3. Mel Spectrograms

The complexity of the human body in expressing emotions means that detailed and sophisticated methods are necessary to extract most of the information. In the case of analyzing speech emotions, the acoustic features of the utterance are of great importance. Mel spectrograms are spectral representations of sounds that use the Mel scale, a nonlinear frequency scale that better corresponds to the perception of sounds by the human ear. Creating a Mel spectrogram involves transforming the audio signal into a frequency spectrum using a Fourier transform, and then converting this spectrum to the Mel scale, which allows for a better representation of the acoustic experience. Mel spectrograms represent the intensity of sound at different frequencies over time, which allows for the identification of subtle differences in tone, pitch, and timbre of the voice that may indicate different emotions. Images prepared in this way allow for the use of advanced machine learning algorithms, especially CNNs (convolutional neural networks), in search of appropriate patterns. Additionally, they allow for preliminary analysis of the input data and noise reduction. Thanks to this, spectrograms are often used in many studies for the analysis of emotions [13], acoustic scene classification [14], automatic speech recognition [15] or even in medical problems for Alzheimer's dementia recognition [16]. Examples of Mel spectrograms are presented in Figure 1.

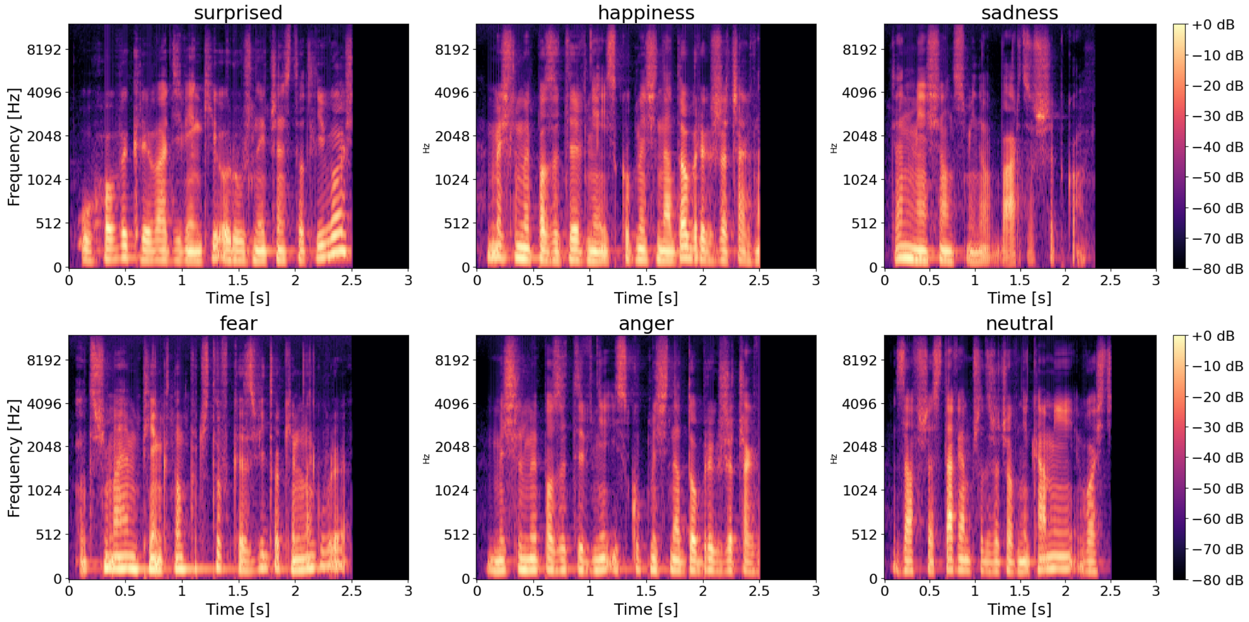

Figure 1. Mel spectrograms created for each category from nEMO dataset.

3. Multimodal Emotion Analysis System

3.1. Architecture

The architecture of the multimodal system fits perfectly into the HRI theory: a separate human and a robotic unit that interact with each other. Additionally, due to the possibility of collecting new data and retraining machine learning models, it is an excellent tool for teaching emotion recognition in narrower groups of people (emotion recognition is a complex problem and is not a universal process), e.g., among people with autism. The hardware architecture of the system consists of data acquisition devices. In the case of audio/video data, these are a regular microphone and camera, while in the case of biomedical data, the Aidmed sensor (Gdańsk, Poland) was used. The Aidmed device is a wearable medical tool designed for remote monitoring of sleep disorders, particularly sleep apnea, by tracking vital signs like heart rate, respiratory rate, and oxygen saturation. The Aidmed device can be adapted for emotion recognition by monitoring physiological signals. Through its real-time data collection and analysis, it could potentially help detect emotional changes based on variations in these biometric indicators. On the other hand, for better visual effects, a robotic head was used. In this case, it is a more advanced Furhat head (Stockholm, Sweden) already equipped with devices for basic acquisition and basic artificial intelligence algorithms. Furhat is a social robot with an expressive, customizable face designed to interact with humans in natural conversations, combining speech recognition, facial expressions, and AI to enhance social interactions. However, in the case of less demanding aspects, simpler solutions can be used, such as the OhBot robotic head (London, UK). As for software, three artificial intelligence algorithms for analyzing vision, sound, and biomedical signals are based on neural networks written in Python (v.3), and will be launched in the cloud or on physical machines communicating via the Internet, which will make the system more flexible and portable. Decision algorithms and self-correction are the final stage of preparing the system and will be developed after combining all three data processing algorithms. The system architecture is shown in Figure 2.

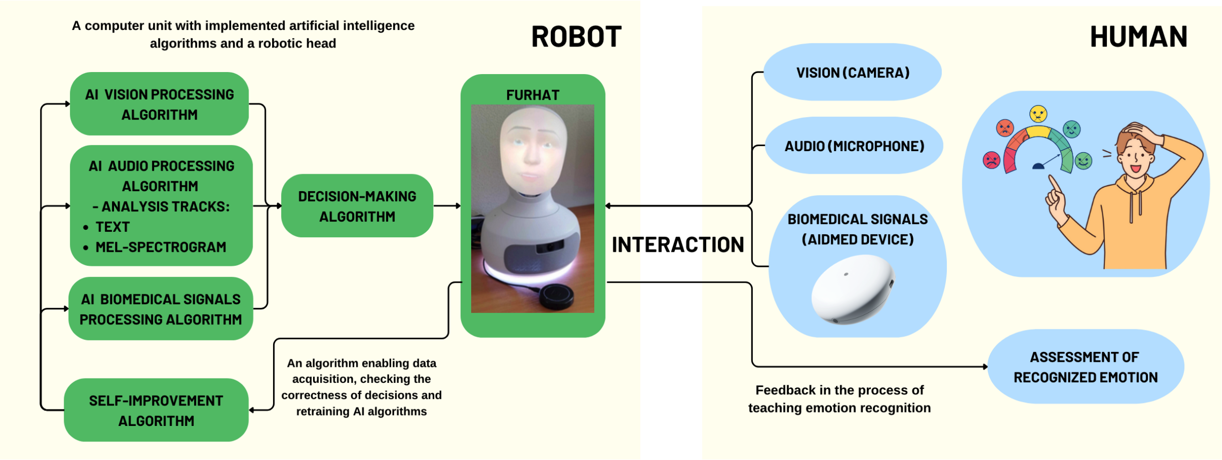
A computer unit with implemented artificial intelligence algorithms and a robotic head
ROBOT
AI VISION PROCESSING ALGORITHM
AI AUDIO PROCESSING ALGORITHM
- ANALYSIS TRACKS:
• TEXT
• MEL-SPECTROGRAM
AI BIOMEDICAL SIGNALS PROCESSING ALGORITHM
SELF-IMPROVEMENT ALGORITHM
DECISION-MAKING ALGORITHM
FURHAT
An algorithm enabling data acquisition, checking the correctness of decisions and retraining AI algorithms
INTERACTION
HUMAN
VISION (CAMERA)
AUDIO (MICROPHONE)
BIOMEDICAL SIGNALS (AIDMED DEVICE)
ASSESSMENT OF RECOGNIZED EMOTION
Feedback in the process of teaching emotion recognition

Figure 2. Architecture of multimodal emotion analysis system.

3.2. Dual-Track Speech Emotion Recognition Module

The channel for recognizing emotions in speech is the main research area of this work. Although one device (microphone) is used to collect speech data, the complexity of this channel is more complicated. We can distinguish two data paths: text and phonetics of speech. Algorithms processing this data in a certain area may return similar results, but there are procedures such as irony of speech, which will give a different result for the text processing algorithm than in the case of examining the tone of speech. Despite these differences, both algorithms must return a common decision (emotional category) based on the assumptions and results of the research carried out in this work. Focusing on the audio processing algorithms themselves, which are based on neural networks, each of them contains preprocessing methods. In the case of text analysis, it is necessary to use speech-to-text methods, which, for a given language (in our case, Polish), will allow for transcription. These tools are popular among the largest computer companies because they significantly facilitate communication on mobile devices. It is also possible to integrate these tools into your own solutions. The most popular speech recognition engines are Google Speech Recognition, Google Cloud Speech API, Houndify API, IBM Speech to Text, Vosk API, and OpenAI Whisper. In the case of phonetics and analysis of emotions from the tone of speech, it is necessary to translate the audio track into a frequency representation in the form of Mel spectrograms. In the case of the Python programming language, the natural choice is the librosa package (v.0.10.2) for audio analysis.

3.3. Other Emotion Recognition Channels

Work on the remaining two channels in the multimodal emotion recognition system has already been done in earlier stages. The most popular algorithm is emotion recognition based on facial images, where the linguistic context does not have such an impact as in the case of analyzing emotions in speech. The human face is also one of the sources from which it is easiest to read the expressed emotion. Many solutions for analyzing human facial expressions are based on datasets categorized, as in this case, according to basic emotions defined by Paul Ekman. The datasets worth mentioning are the FER2013 dataset (Facial Expression Recognition 2013 Dataset) [17] and its extension FER+ [18], in which the photos were captured spontaneously. The third dataset is the CK+ (extended Cohn–Kanade dataset) [19], which is a video set of simulated transition behavior from a neutral state to expressing the target emotion. Based on these datasets, many solutions for image classification algorithms were created, mainly based on ResNet (residual network)-type neural networks. ResNet is a deep learning model designed to solve the problem of vanishing gradients in very deep neural networks by using the so-called "skip connections" or residual blocks, which allows training much deeper networks with better accuracy. Example solutions from recent years indicate the accuracy of models at the level of 80–90% [20–22], which in the case of recognizing a subjective thing such as emotion is a very good result. Additionally, in the case of the vision system, work was also carried out on algorithms for detecting an active speaker, which in the case of a multimodal system will allow for the separation of people whose emotional state is analyzed from potential interlocutors or people appearing in the background [23].

The third information channel through which emotions are transmitted is biomedical data-physiological signals of the human body. We can distinguish the following biomedical signals: EDA (electrodermal activity), BVP (blood volume pulse), EEG (electroencephalography), EMG (electromyography), RSP (respiration), SKT (skin temperature), and HRV (heart rate variability). In their work, datasets concerning these signals were compared, but under two conditions, which are conditioned by the multimodal system using the Aidmed One device and excluding equipment and signal studies that require clinical/medical conditions (the system is to be user-friendly). On this basis, the YAAD (Young Adult's Affective Data) [24] dataset was selected and classic classification models such as decision trees, random forest, support vector machine, k-nearest neighbor, extreme gradient boosting, linear regression, and voting classifier were tested. The accuracy of these models oscillated around 60%, which gave satisfactory results for sets that did not have a very large sample of data (most of the datasets are signals collected from a group of several dozen people). In the future, it is planned to use neural networks for this algorithm, but the priority will be to create a dataset thanks to the multimodal emotion analysis system.

3.4. Future Applications

Multimodal emotion analysis systems can find various applications, such as customer service to assess satisfaction using voice and facial expressions, and mental health monitoring to track patient emotions. They can also improve HRI, increase engagement in education by assessing student responses, and provide insights during job interviews. Additionally, by targeting such a system to a specific group of people, such as children with autism, it can support the process of learning to recognize emotions. However, the key application of this system is for research purposes, such as collecting new data for retraining emotion recognition models and assessing the usefulness of the system by humans.

4. Speech Emotion Recognition Datasets

In accordance with Paul Ekman's theory on the basic palette of emotions, datasets described by six basic emotions or variants containing a similar amount of emotions were selected for further work. Due to the global nature of the English language, many current solutions focus on this language. While the results of emotion recognition algorithms in speech should not differ depending on the language of the utterance (however, there are differences in the phonetic expression of emotions between nations), language is of great importance in the case of text-based algorithms. Based on these assumptions, it was decided to select datasets in the form of audio tracks for English and in the form of audio tracks and text for Polish.

The most important information about the datasets are collected in Table 1.

Table 1. Comparison of technical information of datasets for the problem of emotion recognition in Polish speech.

Dataset	Language	Emotions	Audio Track/ Transcription	Num. of Samples	Technical Info
RAVDESS	English	angry, calm, disgust, fearful, happy, sad, surprise, and neutral	✓/✗	1440	4 s, 48 kHz, 16 bit/sample
TESS	English	anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral	✓/✓	2800	lack of informations
CREMA-D	English	anger, disgust, fear, happy, sad, and neutral	✓/✗	7442	lack of informations
IEMOCAP	English	angry, disappointed, excited, fear, frustrated, happy, sad, surprise, and neutral	✓/✓	1150	lack of informations
SAVEE	English	anger, disgust, fear, happiness, sadness, surprise, and neutral	✓/✗	480	44.1 kHz
Chatbot	Polish	anger, disgust, fear, happiness, sadness, surprise, and neutral	✗/✓	8458	-
nEMO	Polish	anger, fear, happiness, sadness, surprise, and neutral	✓/✓	4481	2.5 s, 24 kHz, 16 bit/sample

4.1. English Language Datasets

4.1.1. RAVDESS

RAVDESS (Ryerson Audio–Visual Database of Emotional Speech and Song) [25], is an advanced dataset designed for research on emotion recognition and vocal expression. It contains recordings of both speech and singing that are intended to reflect emotions such as joy, sadness, fear, anger, surprise, disgust, calm, and neutral. The collection consists of recordings from 24 professional actors, evenly divided between women and men, who perform the same phrases in different emotional states and perform singing fragments expressing the same emotions. The main goal of the database is to support research on emotion perception through the analysis of audio and visual signals. The RAVDESS dataset consists of 1440 recordings of both audio and video. The speech recordings last about 4 s, while singing lasts about 12 s. Various formats are available: audio-only or audio–visual, allowing for the analysis of both vocal and visual expressions of emotion. An important feature of RAVDESS is that emotions are expressed in different degrees of intensity–each of them occurs in both low- and high-intensity versions, which allows for detailed studies of the nuances of emotional expression.

4.1.2. TESS

TESS (the Toronto Emotional Speech Set) [26] dataset was developed to support research on emotion recognition in speech. It consists of recordings in which two actresses utter the same phrases, expressing different emotions. Each actress recorded 200 different sentences, which were selected to evoke a variety of emotional states. The TESS dataset contains recordings reflecting six basic emotions: anger, disgust, fear, happiness, pleasant surprise, and sadness, as well as a neutral state. The audio files are of high quality, which allows them to be used in research on the acoustic aspects of emotional speech. Thanks to carefully selected recordings and the participation of actresses of different ages, TESS is a valuable tool in emotion recognition research, especially in the context of age differences in the perception and expression of emotions.

4.1.3. CREMA-D

CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset) [27] is a publicly available dataset used in research on emotion recognition from face, voice, and multimodal signals. It contains video, audio, and image recordings of actors playing different emotions by saying specific phrases. The dataset consists of material from 91 actors, including 48 men and 43 women, representing different ethnicities and accents. The dataset includes six main emotions: anger, disgust, fear, happy, sad, and neutral. CREMA-D is particularly valuable in multimodal analyses.

4.1.4. IEMOCAP

IEMOCAP (Interactive Emotional Dyadic Motion Capture) [28] is a dataset designed for research on emotion recognition from speech, facial expressions, and gestures. It consists of recordings of dialogues between pairs of actors who act out scenes both improvised and according to prepared scripts. It contains audio and video data, as well as information about body movement obtained through motion capture technology, which makes. Five sessions were recorded in IEMOCAP with 10 actors (5 males and 5 females), and each session was manually annotated for emotions such as angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral. The recordings were divided into smaller segments corresponding to individual utterances, with assignment to specific emotions, subjectively rated by several annotators. IEMOCAP is widely used in research on emotions in speech, dialogue systems, and analysis of nonverbal communication, providing key information in both linguistic and emotional aspects.

4.1.5. SAVEE

SAVEE (Surrey Audio–Visual Expressed Emotion) [29] is a dataset designed for research on emotion recognition in speech, including both audio and visual components. The data comes from recordings made by four male native speakers of English. During the recordings, the actors expressed seven emotions: anger, disgust, fear, happiness, sadness, surprise, and a neutral emotional state, allowing for the study of different forms of emotional expression. The recordings were made in controlled conditions, ensuring high quality of both audio and visual signals. The SAVEE dataset consists of 480 files, including audio recordings and audio–visual materials. The sentences spoken by the actors were carefully selected to correspond to the emotions expressed, allowing for precise analysis of both acoustic and facial expressions. Due to its structure and diverse emotions, the dataset is suitable for the development of artificial intelligence technologies, speech recognition systems, and applications that analyze emotions in real time.

4.2. Polish Language Datasets

4.2.1. Chatbot Dataset

This regards the author's text dataset for the problem of recognizing emotions in text [30]. It was created based on the popular chatbot ChatGPT3.5 by OpenAI using the model "text-davinci-003". The main idea was to create prompts to generate Polish text related to one of six emotions and a neutral state. Thanks to this, a dataset consisting of a training and test set was created. The training set was fully generated thanks to artificial intelligence and contained about 6800 sentences described in emotional categories. The test set of about 1700 sentences was described in emotional categories by a human. Thanks to this, it was possible to examine the usefulness of such a set and show the shortcomings of the OpenAI tool. This set is the basis for emotion recognition algorithms for a multimodal system. Due to the way the dataset was created based on artificial text generation, the dataset does not contain audio tracks of people expressing emotions. The dataset is mainly used for the problem of analyzing emotions in text.

4.2.2. Nemo Dataset

A modern dataset for analyzing speech emotions in Polish [31]. The authors indicate the following directions for using this set: audio classification, automatic speech recognition, and text-to-speech. The set contains about 4500 simulated recordings expressing five basic emotions (disgust emotion is not included in the set) and a neutral state. Additionally, the above sentences were transcribed, which allows the use of the dataset for both problems: analysis of voice and text emotion. From the weighted technical information, the average length of recordings is about 2.5 s. All audio samples were transformed so that their highest volume was 0 dB, with a sampling frequency of up to 24 kHz, using 16 bits for each sample.

5. Machine Learning Algorithms

5.1. Speech Emotion Recognition Models

The problem of recognizing emotions in speech using Mel spectrograms is based on classification models of artificial intelligence. Below are presented popular neural networks for image processing, which were trained based on the sets presented earlier.

5.1.1. Convolutional Neural Networks

CNNs are deep learning architectures suited to grid-structured data such as images, using convolutional layers to automatically capture spatial patterns and feature hierarchies. They are highly effective at tasks such as image classification, object detection, and face recognition, due to their ability to efficiently identify patterns and features in images. Based on four datasets RAVDESS, TESS, SAVEE, and CREMA-D, a convolutional network architecture was developed [32]. Using CNN treats Mel spectrograms as images, leveraging CNN's strength in feature extraction from 2D data. The model architecture includes several convolutional layers (Conv2D), each followed by pooling and batch normalization to reduce complexity and maintain stability. After multiple convolutional layers, a GAP (global average pooling) layer is used to reduce the number of parameters, followed by a dropout layer to prevent overfitting, and finally a dense layer with a softmax activation function to classify emotions. This approach allows for efficient feature learning from audio data without the need for extensive preprocessing.

5.1.2. Convolutional Neural Networks: ResNet

ResNet is a neural network architecture designed to solve the vanishing gradient issue in deep networks by introducing residual connections that skip layers, allowing the model to learn identity functions more effectively. These shortcut connections enable the training of very deep models with improved performance, as seen in tasks like image recognition and object detection. The number in ResNet refers to the total number of layers in the network, which include convolutional layers, pooling layers, fully connected layers, and residual blocks. The higher the number, the deeper the network, with more layers used to extract features and process information, allowing the network to capture more complex patterns. The comparison of neural networks for the problem of emotion recognition in speech was made not only based on ResNets, but also on classical convolutional networks [33]. The datasets used in the comparison are CREMA-D, IEMOCAP, RAVDESS, SAVEE, and TESS. The ResNet 50, 34, and 18 networks in different configurations with categories were used in the comparison.

5.1.3. Recurrent Neural Networks: LSTM

LSTM (long short-term memory) is a widely used deep learning architecture that excels at processing sequential data due to its capability to retain information over extended time steps. This makes LSTMs particularly effective for tasks involving time-sensitive or sequential data, such as speech recognition, language translation, and time series forecasting, where maintaining context over long sequences is crucial. The authors proposed four architectures [34] based on the RAVDESS dataset:

- Stacked Time Distributed 2D CNN-LSTM integrates several 2D convolutional layers with an LSTM layer, enabling the effective learning of spatial features from sequences of images while also capturing temporal relationships.
- Stacked Time Distributed 2D CNN-Bidirectional LSTM with Attention improves upon this by adding a bidirectional LSTM and an attention mechanism, which enhances the model's ability to consider context from both preceding and subsequent frames in the input data.
- Parallel 2D CNN-Bidirectional LSTM with Attention employs parallel 2D CNNs to simultaneously extract spatial features from multiple inputs and channels them into a bidirectional LSTM with Attention, enhancing the model's focus on critical information throughout the input sequence.
- Parallel 2D CNN-Transformer Encoder leverages parallel 2D CNNs for feature extraction, followed by processing with a Transformer encoder that utilizes self-attention mechanisms, thereby facilitating a deeper understanding of intricate patterns in sequential data.

The comparison of the presented classification models for emotion recognition along with their most important parameter, accuracy, is presented in Table 2.

Table 2. Comparison of neural network models for the emotion recognition problem using Mel spectrograms.

Classification Model	Training Datasets	Num. of Categories	Accuracy
Convolutional Neural Network	CREMA-D, RAVDESS, SAVEE, TESS	7	71.00%
ResNet 18	CREMA-D, IEMOCAP, RAVDESS, SAVEE, TESS	6	67.62%
ResNet 34		6	68.15%
ResNet 50		4	75.76%
ResNet 50		5	68.86%
ResNet 50		6	62.18%
Stacked Time Distributed 2D CNN-LSTM	RAVDESS	8	90.00%
Stacked Time Distributed 2D CNN-Bidirectional LSTM with Attention		8	94.02%
Parallel 2D CNN-Bidirectional LSTM with Attention		8	96.55%
Parallel 2D CNN-Transformer Encoder		8	96.78%

5.2. Text Emotion Recognition Model

The model for analyzing emotions in Polish texts was developed using an artificially generated Chatbot dataset [30]. Its neural network architecture relied on LSTM recurrent networks, with the input layer incorporating embeddings derived from distribution models specifically designed for the Polish language. In NLP tasks, LSTMs are frequently combined with word embeddings, which enhances the model's capability to grasp and analyze the contextual and semantic relationships between words. Word embeddings serve as numerical representations that capture the semantic and syntactic associations between words in a dense vector space, enabling algorithms to better interpret word meaning and context. Common techniques for generating these embeddings, such as Word2Vec or GloVe, are widely used in tasks like sentiment analysis, classification, and translation. For this model, the distribution models were generated using the Word2Vec technique and trained on data from the NJKP (National Corpus of Polish) and Polish Wikipedia. Multiple models were applied, varying in type, architecture, and vector size, to optimize performance for the specific linguistic requirements of the task. The model for recognizing emotions in text is characterized by high efficiency. The model tested on generic data has an accuracy of about 96%, while in the case of data labeled by a human, the model's accuracy is 87%.

6. Results

In order to create a dual-track audio analysis module for a multimodal emotion recognition system and to analyze the relationship between emotion recognition in speech and text, based on the presented concepts and comparisons of models and datasets, the following research was conducted according to speech emotion recognition:

- The Chatbot model was used to analyze emotions in text by testing it on the nEMO dataset to justify the need to recognize emotions in speech.
- The LSTM models for speech emotion recognition that performed best in the comparison were re-trained on the original datasets to prove their accuracy and suitability for the given problem.
- The nEMO dataset containing data for the Polish language was preprocessed, and this set was adapted for training on previous models to check the model's accuracy on new Polish-language data.
- The models were cross-compared by testing them on opposite test datasets.

6.1. Testing nEMO Dataset on Chatbot Text Emotion Recognition Model

The first step in the research on the dual-track module for recognizing emotions in speech and text was to prove that text or speech is not always the carrier of information about the expressed emotion. For this purpose, a model for recognizing emotions in text was used, which was previously trained on generic Chatbot data based on ChatGPT [30]. One of the few sets in Polish was used for this purpose-nEMO, which has both an audio track and raw text. By nature, the sentences spoken in the set are neutral, and only the manner and tone of the statement give them an emotional overtone. In this situation, an attempt was made to evaluate the model on this dataset. The prediction results are presented in Table 3.

Table 3. Evaluation of a text emotion recognition model on the nEMO dataset with neutral texts.

Emotion	Number of Model Predictions
Surprise	96
Disgust	0
Anger	0
Fear	6
Sadness	6
Happiness	0
Neutral	432

Based on the above results, it can be concluded that the model behaved correctly. Assuming from the outset that the dataset contains only emotionally neutral texts, the model correctly predicted 80% of the entire set. Several cases were predicted as fear and sadness, and several dozen as a surprise. This may indicate a slight imperfection of the model, which can be improved in future work by adding data to the training set. However, the predictions regarding the model's behavior were confirmed and it can be safely stated that the speech emotion recognition module cannot rely only on the text itself. In this case, it is necessary to also recognize the tone and timbre of the utterance.

6.2. Speech LSTM Models Based on RAVDESS Dataset

The first stage of research on speech emotion recognition models was to select a group of neural network models that gave the best results in comparison. At first glance, the group of LSTM models stood out: Stacked Time Distributed 2D CNN-LSTM, Stacked Time Distributed 2D CNN-Bidirectional LSTM with Attention, Parallel 2D CNN-Bidirectional LSTM with Attention, and Parallel 2D CNN-Transformer Encoder. The accuracy of these models, according to the authors, oscillated around 90–97% accuracy, which in comparison to other models gave a gain in accuracy of around 20%. This result allowed us to state that this is an ideal solution. Therefore, it was decided to confirm the given results by re-training the proposed LSTM model architectures on the RAVDESS database. These data allow for the classification of the following emotions: angry, calm, disgust, fearful, happy, sad, surprise, and neutral. According to the assumptions, the set was divided into training, validation, and test sets in the proportions of 80:10:10, then Mel spectrograms containing the appropriate time window of the spoken text were prepared and subjected to appropriate preprocessing methods. Models were trained on the prepared data and tested, taking into account the most important metrics for classification algorithms. Accuracy reflects the proportion of correct predictions out of all predictions made by a model, whereas precision focuses on the ratio of true positives to all positive predictions. Recall measures the model's ability to identify all actual positive cases, and the F1 score combines precision and recall into a single metric, giving a balanced measure of performance when dealing with imbalanced data. The results of the mentioned metrics for each category and the average value are summarized in Table 4. For each emotion and the average value, the best results are bolded.

At first glance, the difference between the evaluation of the authors of the LSTM models and the retraining is visible. The most accurate model is the Parallel 2D CNN-Bidirectional LSTM with Attention with an accuracy of 65%, which gives a result worse by several percent than other models proposed in comparison with a similar number of categories. The metrics for individual emotions are mostly the best for the aforementioned model. However, two other models cannot be rejected-Parallel 2D CNN-Transformer Encoder and Stacked Time Distributed 2D CNN-Bidirectional LSTM, whose accuracy is several percent lower. For a better analysis of individual emotions, it was decided to present a confusion matrix for each model. A confusion matrix is a summary table used to evaluate the performance of a classification model by displaying the counts of true positives, true negatives, false positives, and false negatives. It offers insights into the model's ability to correctly predict each class, helping identify areas where the model may be biased or making errors. The confusion matrices for the evaluated models are presented in Figure 3.

Confusion matrices are clearly diagonal, which allows us to claim that the predicted values match the true emotions. From the evaluation of metrics, it was possible to notice a poor accuracy of the models in the case of the neutral state, which is confirmed in the above matrices. This results from the small sample of the neutral category in the set. For the most accurate model, the number of well-recognized categories for each emotion is distributed evenly, which indicates that the model best balances the prediction results despite the low accuracy.

Table 4. Models evaluation for emotion speech recognition on Mel spectrograms based on RAVDESS dataset.

Models	**Emotions**								
	Average	**Surprise**	**Neutral**	**Calm**	**Happy**	**Sad**	**Angry**	**Fear**	**Disgust**
Stacked Time Distributed 2D CNN -LSTM									
Accuracy [%]	53.00	**65.00**	40.00	65.00	40.00	10.00	65.00	55.00	80.00
Precision [%]	54.00	**76.47**	66.67	48.15	61.54	20.00	65.00	39.29	55.17
Recall [%]	52.50	**65.00**	40.00	65.00	40.00	10.00	65.00	55.00	80.00
F1-Score [%]	51.70	**70.27**	50.00	55.32	48.48	13.33	65.00	45.83	65.31
Stacked Time Distributed 2D CNN -Bidirectional LSTM with Attention									
Accuracy [%]	62.00	45.00	**50.00**	75.00	**65.00**	45.00	70.00	**70.00**	70.00
Precision [%]	63.20	56.25	**71.43**	75.00	48.15	47.37	66.67	66.67	73.68
Recall [%]	61.20	45.00	**50.00**	**75.00**	**65.00**	45.00	70.00	**70.00**	70.00
F1-Score [%]	61.70	50.00	**58.82**	75.00	55.32	46.15	68.29	**68.29**	71.79
Parallel 2D CNN -Transformer Encoder									
Accuracy [%]	63.00	45.00	30.00	70.00	**65.00**	40.00	**95.00**	55.00	**90.00**
Precision [%]	61.90	42.86	37.50	**70.00**	**86.67**	**50.00**	**67.86**	73.33	66.67
Recall [%]	61.30	45.00	30.00	70.00	**65.00**	40.00	**95.00**	55.00	**90.00**
F1-Score [%]	60.60	43.90	33.33	70.00	**74.29**	44.44	**79.17**	62.86	76.60

Table 4. *Cont.*

Models	Emotions								
	Average	Surprise	Neutral	Calm	Happy	Sad	Angry	Fear	Disgust
Parallel 2D CNN -Bidirectional LSTM with Attention									
Accuracy [%]	**65.00**	**65.00**	30.00	**90.00**	55.00	**60.00**	70.00	60.00	70.00
Precision [%]	**64.20**	**76.47**	33.33	64.29	73.33	**50.00**	58.33	**70.59**	**87.50**
Recall [%]	**62.50**	**65.00**	30.00	**90.00**	55.00	**60.00**	70.00	60.00	70.00
F1-Score [%]	**62.60**	**70.27**	31.58	**75.00**	62.86	**54.55**	63.64	64.86	**77.78**

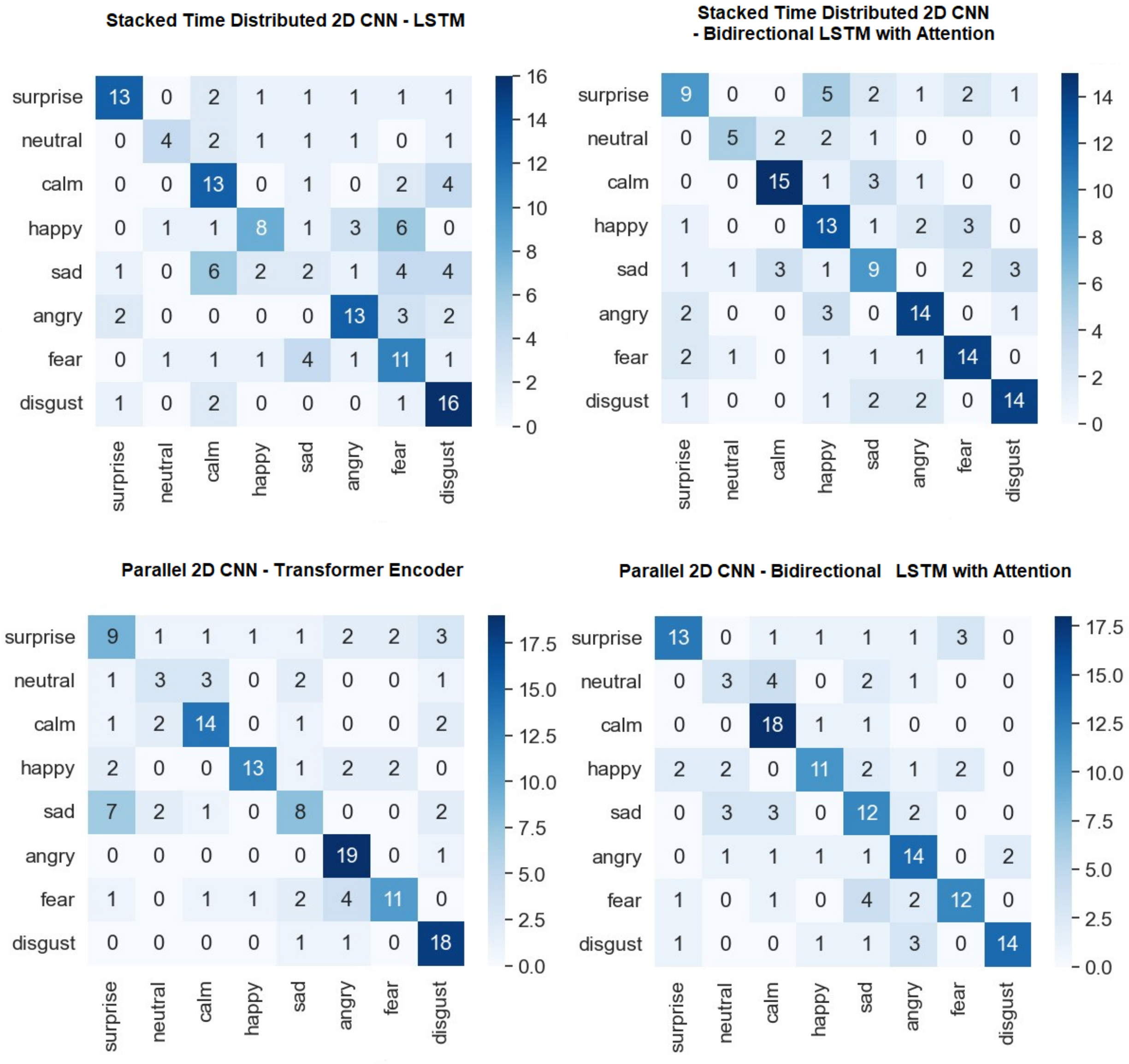

Figure 3. Confusion matrices for speech emotion analysis models based on the English-language RAVDESS dataset.

6.3. Speech LSTM Models Based on nEMO Dataset

Despite the lack of the assumed accuracy of LSTM models on the RAVDESS dataset, it was decided to test them on the Polish-language nEMO dataset. The process of preparing Mel spectrograms, selecting the appropriate utterance window, and their preprocessing methods were carried out in exactly the same way as in the previous study. This dataset has a smaller number of categories (anger, fear, happiness, sadness, surprise, and neutral), which were also taken into account. The evaluation of the Accuracy, Precision, Recall, and

F1-score metrics for individual metrics and four LSTM models is presented in Table 5. The best results are bolded in the table.

Table 5. Models evaluation for emotion speech recognition on Mel spectrograms based on polish nEMO dataset.

Models	**Emotions**						
	Average	**Anger**	**Fear**	**Happiness**	**Sadness**	**Surprised**	**Neutral**
Stacked Time Distributed 2D CNN -LSTM							
Accuracy [%]	72.20	86.67	87.84	30.67	77.92	61.19	**88.89**
Precision [%]	77.70	63.73	63.11	**100.00**	92.31	77.36	69.90
Recall [%]	72.20	86.67	87.84	30.67	77.92	61.19	**88.89**
F1-Score [%]	70.80	73.45	73.45	46.94	84.51	68.33	78.26
Stacked Time Distributed 2D CNN -Bidirectional LSTM with Attention							
Accuracy [%]	71.30	**90.67**	86.49	36.00	90.91	44.78	79.01
Precision [%]	75.90	75.56	58.18	93.10	81.40	81.08	65.98
Recall [%]	71.30	**90.67**	86.49	36.00	90.91	44.78	79.01
F1-Score [%]	69.90	82.42	69.57	51.92	85.89	57.69	71.91
Parallel 2D CNN -Transformer Encoder							
Accuracy [%]	**84.60**	86.67	**91.89**	**84.00**	**98.70**	**62.69**	83.95
Precision [%]	**85.00**	**89.04**	83.95	79.75	84.44	**82.35**	**90.67**
Recall [%]	**84.60**	86.67	**91.89**	**84.00**	**98.70**	**62.69**	83.95
F1-Score [%]	**84.50**	**87.84**	**87.74**	**81.82**	91.02	**71.19**	**87.18**
Parallel 2D CNN -Bidirectional LSTM with Attention							
Accuracy [%]	79.40	84.00	86.49	69.33	92.21	**62.69**	81.48
Precision [%]	80.00	75.00	**88.89**	81.25	**94.67**	54.55	85.71
Recall [%]	79.40	84.00	86.49	69.33	92.21	**62.69**	81.48
F1-Score [%]	79.50	79.25	87.67	74.82	**93.42**	58.33	83.54

In the case of the Polish-language dataset, it can be seen that LSTM architectures perform much better than in the previous case, ensuring at least 10% higher efficiency. The Parallel 2D CNN-Transformer Encoder model was trained best on this data, providing an accuracy of 84.60%, which is about 10% higher efficiency than for the ResNet 50 models from the theoretical comparison. Again, the Parallel 2D CNN-Bidirectional LSTM with Attention model deserves special mention, with an efficiency of 79.40%. Considering the metrics for individual emotions, the first model shows greater efficiency, but it is also worth mentioning that the lowest efficiency occurs for the emotion of surprise and does not drop below 62%. For a better insight into the distribution of model predictions, it was decided to present confusion matrices again, which can be analyzed in Figure 4.

The diagonal is visible on all matrices, which allows us to conclude that the models cope with recognizing emotions. The performance of the Parallel 2D CNN-Transformer Encoder model over others is visible in recognizing the emotion of happiness. The accuracy of the other models in recognizing this emotion is much lower. The only imperfection that can be noticed is the model confusing the emotion of surprise with emotions such as anger, fear, and happiness. Surprise can be expressed very expressively, which also indicates that it is confused with other expressive emotions. In future work, more data could be collected for this emotion, which would allow the model to be more resistant to these mistakes.

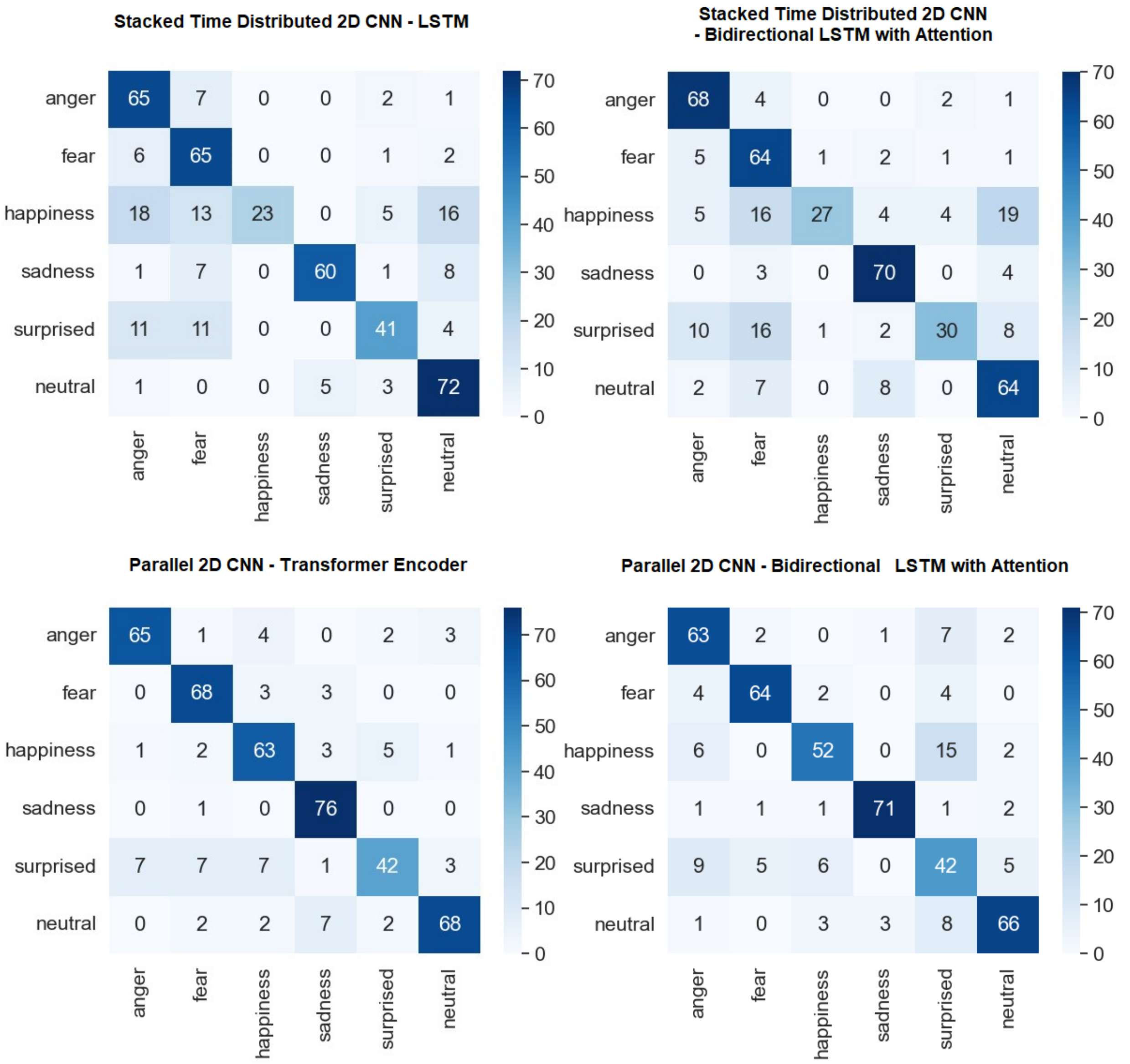

Figure 4. Confusion matrices for emotion analysis models in speech based on the Polish-language nEMO dataset.

6.4. Comparison of Speech Models

In the undertaken studies, an attempt was made to cross-test the most accurate models: Parallel 2D CNN-Bidirectional LSTM with Attention and Parallel 2D CNN-Transformer Encoder. Models trained on the RAVDESS set were tested on the nEMO test set and vice versa. However, in both cases, the results did not exceed 30% efficiency, which indicates diversity in the datasets, especially in tones and manners of speech. The manner and characteristics of speech are culturally and, to some extent, nationally conditioned. Factors that can affect accuracy during the cross-test are temporal characteristics of speech such as the length of the speech, tempo, intonation, rhythm, or pauses (punctuation). There are several ways to create a more accurate model for recognizing emotions in speech for different languages. One of them is to maintain the aforementioned features at a similar level between languages or people of different nationalities. However, this creates another problem, which is the unnaturalness of speech. A model trained on such data may have a problem with spontaneous speech in different languages. The second, more universal way is to create a new data set (or combine several existing ones). This set should represent data collected in several of the most popular languages from people with different ways of speaking. This diversity will cause the neural network to have more patterns for a given emotion, which may ultimately increase the effectiveness of such a model. For a multimodal system for recognizing emotions in Polish, such a model is sufficient. In future studies, one can focus on creating a more universal model based on datasets in different languages.

7. Discussion and Conclusions

The main goal of this project was to research deep learning algorithms for recognizing emotions in Polish speech and text as one of the modules of a multimodal emotion recognition system. Initial recognition of datasets showed a niche in datasets other than English. Recently, the nEMO dataset has appeared for the Polish language, thanks to which it was possible to train artificial intelligence algorithms. A comparison was also made of existing neural network solutions for the problem of speech recognition using Mel spectrograms, which showed the highest efficiency among LSTM models and later convolutional networks including ResNet. The next stage of work consisted of checking the Polish language dataset on an earlier model prepared for recognizing emotions in Polish texts. Due to the natural texts of the dataset, the emotional coloring of which consisted of changing the tone of the utterance, the model recognized the text as emotionally neutral in most cases. It can be stated that using only one audio analysis algorithm in a multimodal system could give poor results. The diverse nature of people means that they can express emotions for neutral-sounding texts or people may not emanate such expression when they say texts indicating specific emotions. At this point, it is necessary to use two algorithms for speech and text, which will be activated alternately, when one of them gives a neutral result. The problem arises when the algorithms return contradictory emotions, but this may be an aspect of further research. In the next stage, it was decided to evaluate 4 LSTM models: Stacked Time Distributed 2D CNN-LSTM, Stacked Time Distributed 2D CNN-Bidirectional LSTM with Attention, Parallel 2D CNN-Bidirectional LSTM with Attention, and Parallel 2D CNN-Transformer Encoder on the English dataset. The evaluation showed lower accuracy than that provided by the authors, and the most accurate model was characterized by an accuracy of 65%. The models were re-evaluated on the Polish nEMO dataset and it was possible to obtain an accuracy of 84.6%, which is very satisfactory. The Parallel 2D CNN-Transformer Encoder model copes well with all emotional categories, thanks to the balanced dataset. It can be said that this architecture on the nEMO dataset is an excellent point for implementation in a multimodal emotion analysis system. Additional research has shown that there are differences in cross-language testing, which allows us to conclude that the model on Polish or English data is not a universal model, but it is sufficient for implementation in the system.

Future research includes the following aspects: improving the quality of the algorithm for textual emotion analysis, implementing a decision module for the entire multimodal system, real-world testing of the system along with collecting further training data for all modules, developing an original dataset based on real tests, and training speech emotion analysis algorithms to work universally with different languages. Additionally, further research is planned to take into account the temporal features of speech. Developing advanced machine learning models that can accurately analyze speech patterns, including duration, rhythm, and intonation, in different cultural contexts could impact the effectiveness of emotion recognition in speech.

Author Contributions: Conceptualization, K.S. and A.G.; methodology, K.S.; software, K.S.; validation, K.S.; formal analysis, K.S.; investigation, K.S.; resources, K.S.; data curation, K.S.; writing—original draft preparation, K.S.; writing—review and editing, K.S., A.G. and E.P.; visualization, K.S.; supervision, A.G. and E.P.; project administration, A.G. and E.P.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research of Kamil Skowroński was supported by the Silesian University of Technology through the subsidy for maintaining and developing the research potential grant in 2024 for young researchers in analysis (02/060/BKM24/0062). The work of Adam Gałuszka was supported in part by the Silesian University of Technology through the subsidy for maintaining and developing the research potential grant in 2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BVP	Blood Volume Pulse
CK+	Extended Cohn-Kanade dataset
CNN	convolutional neural network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
EDA	Electrodermal activity
EEG	Electroencephalography
EMG	Electromyography
FER2013	Facial Expression Recognition 2013 Dataset
GAP	Global Average Pooling
HRI	human-robot interaction
HRV	Heart Rate Variability
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LSTM	Long Short-Term Memory
NJKP	National Corpus of Polish
NLP	natural language processing
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
ResNet	Residual Networks
RSP	Respiration
SAVEE	Surrey Audio-Visual Expressed Emotion
SKT	Skin temperature
TESS	The Toronto Emotional Speech Set
YAAD	Young Adult's Affective Data

References

1. Hogeveen, J.; Grafman, J. Alexithymia. In *Handbook of Clinical Neurology*; North -Holland Publishing Company: Amsterdam, The Netherlands, 2021; Volume 183; pp. 47–62.
2. Abdullah, S.M.S.A.; Ameen, S.Y.A.; Sadeeq, M.A.; Zeebaree, S. Multimodal emotion recognition using deep learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 73–79. [CrossRef]
3. Powroźnik, P. Polish emotional speech recognition using artifical neural network. *Adv. Sci. Technol. Res. J.* **2014**, *8*, 24–27. [CrossRef] [PubMed]
4. Satt, A.; Rozenberg, S.; Hoory, R. Efficient emotion recognition from speech using deep learning on spectrograms. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
5. Zielonka, M.; Piastowski, A.; Czyżewski, A.; Nadachowski, P.; Operlejn, M.; Kaczor, K. Recognition of emotions in speech using convolutional neural networks on different datasets. *Electronics* **2022**, *11*, 3831. [CrossRef]
6. Breazeal, C.L. Sociable Machines: Expressive Social Exchange Between Humans and Robots. Ph.D. Thesis, Massachusetts Institute of Technology, Massachusetts , CA, USA, 2000.
7. Murphy, R.R.; Nomura, T.; Billard, A.; Burke, J.L. Human–robot interaction. *IEEE Robot. Autom. Mag.* **2010**, *17*, 85–89. [CrossRef]
8. Johanson, D.L.; Ahn, H.S.; Broadbent, E. Improving interactions with healthcare robots: A review of communication behaviours in social and healthcare contexts. *Int. J. Soc. Robot.* **2021**, *13*, 1835–1850. [CrossRef]
9. Pachidis, T.; Vrochidou, E.; Kaburlasos, V.; Kostova, S.; Bonković, M.; Papić, V. Social robotics in education: State-of-the-art and directions. In Proceedings of the Advances in Service and Industrial Robotics: Proceedings of the 27th International Conference on Robotics in Alpe-Adria Danube Region (RAAD 2018), Zagreb, Croatia, 9–12 December 2019; pp. 689–700.
10. Broekens, J.; Heerink, M.; Rosendal, H. Assistive social robots in elderly care: A review. *Gerontechnology* **2009**, *8*, 94–103. [CrossRef]
11. Prinz, J. Which emotions are basic. *Emot. Evol. Ration.* **2004**, *69*, 88.

12. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118.
13. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
14. Zhang, T.; Feng, G.; Liang, J.; An, T. Acoustic scene classification based on Mel spectrogram decomposition and model merging. *Appl. Acoust.* **2021**, *182*, 108258. [CrossRef]
15. Lin, T.Q.; Lee, H.y.; Tang, H. Melhubert: A simplified hubert on mel spectrograms. In Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; pp. 1–8.
16. Meghanani, A.; Anoop, C.S.; Ramakrishnan, A. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 670–677.
17. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Republic of Korea, 3–7 November 2013; Proceedings, Part III 20, pp. 117–124.
18. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo Japan, 12–16 November 2016; pp. 279–283.
19. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
20. Pham, L.; Vu, T.H.; Tran, T.A. Facial expression recognition using residual masking network. In Proceedings of the 2020 25Th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4513–4519.
21. Li, H.; Wang, N.; Ding, X.; Yang, X.; Gao, X. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Trans. Image Process.* **2021**, *30*, 2016–2028. [CrossRef] [PubMed]
22. Le Ngwe, J.; Lim, K.M.; Lee, C.P.; Ong, T.S.; Alqahtani, A. PAtt-Lite: Lightweight patch and attention MobileNet for challenging facial expression recognition. *IEEE Access* **2024**, *12*, 79327–79341. [CrossRef]
23. Skowroński, K. Active speaker detection in a social human-robot interaction maintenance system for social robots. In Proceedings of the 2023 European Simulation and Modelling Conference, Toulouse, France, 24–26 October 2023; pp. 155–160.
24. Dar, M.N.; Rahim, A.; Akram, M.U.; Khawaja, S.G.; Rahim, A. YAAD: Young adult's affective data using wearable ECG and GSR sensors. In Proceedings of the 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), Rawalpindi, Pakistan, 24–26 May 2022; pp. 1–7.
25. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
26. Pichora-Fuller, M.K.; Dupuis, K. *Toronto Emotional Speech Set (TESS)*; University of Toronto, Psychology Department: Toronto, ON, Canada, 2020. [CrossRef]
27. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef] [PubMed]
28. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
29. Jackson, P.; Haq, S. *Surrey Audio-Visual Expressed Emotion (Savee) Database*; University of Surrey: Guildford, UK, 2014.
30. Skowroński, K. Application of data-collecting chatbot for Polish text emotion analysis. In Proceedings of the 2023 European Simulation and Modelling Conference, Toulouse, France, 24–26 October 2023; pp. 161–165.
31. Christop, I. nEMO: Dataset of Emotional Speech in Polish. *arXiv* **2024**, arXiv:2404.06292.
32. Speech Emotion Recognition (SER) Using CNNs and CRNNs Based on Mel Spectrograms and Mel Frequency Cepstral Coefficients (MFCCs). Available online: https://datascrutineer.com/speech-emotion-recognition-cnns-tensorflow/ (accessed on 10 October 2024).
33. Speech Emotion Recognition. Available online: https://github.com/KanikeSaiPrakash/Speech-Emotion-Recognition (accessed on 10 October 2024).
34. Speech Emotion Classification with PyTorch. Available online: https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch/ (accessed on 10 October 2024).

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