electronics

Article

Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets

Marta Zielonka †, Artur Piastowski †, Andrzej Czyżewski *, Paweł Nadachowski, Maksymilian Operlejn and Kamil Kaczor

Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
* Correspondence: ac@pg.edu.pl
† These authors contributed equally to this work.

Abstract: Artificial Neural Network (ANN) models, specifically Convolutional Neural Networks (CNN), were applied to extract emotions based on spectrograms and mel-spectrograms. This study uses spectrograms and mel-spectrograms to investigate which feature extraction method better represents emotions and how big the differences in efficiency are in this context. The conducted studies demonstrated that mel-spectrograms are a better-suited data type for training CNN-based speech emotion recognition (SER). The research experiments employed five popular datasets: Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Surrey Audio-Visual Expressed Emotion (SAVEE), Toronto Emotional Speech Set (TESS), and The Interactive Emotional Dyadic Motion Capture (IEMOCAP). Six different classes of emotions were used: happiness, anger, sadness, fear, disgust, and neutral. However, some experiments were prepared to recognize just four emotions due to the characteristics of the IEMOCAP dataset. A comparison of classification efficiency on different datasets and an attempt to develop a universal model trained using all datasets were also performed. This approach brought an accuracy of 55.89% when recognizing four emotions. The most accurate model for six emotion recognition was trained and achieved 57.42% accuracy on a combination of four datasets (CREMA-D, RAVDESS, SAVEE, TESS). What is more, another study was developed that demonstrated that improper data division for training and test sets significantly influences the test accuracy of CNNs. Therefore, the problem of inappropriate data division between the training and test sets, which affected the results of studies known from the literature, was addressed extensively. The performed experiments employed the popular ResNet18 architecture to demonstrate the reliability of the research results and to show that these problems are not unique to the custom CNN architecture proposed in experiments. Subsequently, the label correctness of the CREMA-D dataset was studied through the employment of a prepared questionnaire.

Keywords: speech emotion recognition; SER; machine learning; artificial intelligence; classification; convolutional neural networks

Citation: 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. https://doi.org/10.3390/electronics11223831

Academic Editors: Daniel Hládek, Matúš Pleva, Piotr Szczuko and Andrej Zgank

Received: 9 August 2022
Accepted: 15 November 2022
Published: 21 November 2022

1. Introduction

The recognition of emotions is a relatively difficult and complex task [1], even for humans. Many people could say that they can perform this task efficiently; however, they often have the opportunity to recognize emotions based on a few different aspects, such as body language, facial expression, and voice timbre or prosody. Meanwhile, speech emotion recognition (SER) is a potentially significant step toward the future as it presents a huge variety of use cases.

SER considers recognizing emotions using only one modality, voice recordings, which makes it more complex. Thus, it uses one additional medium—a microphone that may also capture some noise [2]. Achieving decent results on this type of problem could lead to the

development of machines being more humanized, as there is nothing more human-like than emotions. Enabling machines to understand human beings' moods and intentions could be used in fields such as security [3], medicine, emergency call centers [4], telemarketing, and daily work in social institutions. Previous examples present that such a tool could help understand emotions not only by machines but also by people with a less acute ability to distinguish the emotions of their interlocutor or people with perception disabilities.

A lot of research results can be found in this field of study [5–8], but they are usually obtained for only one dataset. Therefore, this approach does not promise the same results after deploying such models. This is because some of the available datasets do not employ many actors, which is not beneficial for neural networks in particular, which require vast amounts of data to perform well in every environment. Moreover, using multiple datasets can reduce the model's tendency to learn the characteristics of recordings since datasets tend to have diversified sound characteristics due to the use of a variety of recording equipment.

Having more data does not necessarily mean better results, as it first has to be prepared appropriately. The process of data preparation for this type of study can be time-consuming, and there is no typical approach that would promise the best possible results. There are a few different ways in which the data can be prepared. One of the most common ones is spectrograms and mel-spectrograms. This research aims to compare which of those brings a better performance of CNNs when used for training. Since such comparisons are nowhere to be found in the literature, exploring them could save time for many researchers.

What is more, it is common to randomly split data into training and test sets when dealing with machine learning algorithms. This is usually a good approach, but not always, and if performed incautiously, it can lead to false results. Sometimes the data can be interdependent. For example, multiple instances might be available that represent the same object repeatedly but in a different environment, or a group of data may share the same characteristics, like in the case of audio data containing multiple recordings per emotion prepared by the same actor. In this case, it is necessary to divide the data in such a way that these instances are not repeated in the training and test sets. Although unfortunately, there are still solutions provided by researchers in which no attention was paid to the above, probably by oversight [5,9,10], in order to draw attention to the problems that this may entail, it was decided to conduct research in this area.

Many cases of expressed emotions are perceivable, but they can be quite hard to label. People who prepare the datasets have different approaches to solving this problem. Sometimes, labels are retrieved based on the opinion of a speaker, of psychologists, or of a group of people to evaluate the speaker's emotions. Until now, the quality of datasets has not been verified by other researchers, only by the authors of the datasets. That is why the decision has been made to perform an additional study on the data quality. A specific study was conducted on label correctness to verify how reliable the datasets can be.

This research aims to address all these challenges by presenting a description of the data preparation and usage of the five different datasets. The approach may also be advantageous because, most of the time, datasets are prepared in different environments with various resources exploited.

2. Related Works

This chapter presents various algorithms used in speech emotion recognition introduced by other researchers. Discussing algorithms and machine learning methods is impossible if the data format is not discussed first. When working with audio data, there is a variety of approaches to choose from. Audio data can be presented in a raw waveform or in a 2D form, such as spectrograms, mel-spectrograms, mel-frequency cepstral coefficients (MFCCs), and many more.

If the input data are in raw audio format, one possible approach, besides the classical recurrent neural networks and their variations, is to use the WaveNet architecture [11]. Researchers created a solution [12] that is able to classify emotions from speech based on raw audio data and employed this architecture as a backbone.

For the 2D audio data format, Long Short-Term Memory (LSTMs) artificial neural networks are frequently used in conjunction with CNNs [13–16]. LSTMs can remember long-term relationships in the input signal, which is beneficial when dealing with sequential input data such as audio signals, and CNNs can learn features from high-dimensional input data [17]. Many variants of LSTMs have already been introduced by researchers, one of which is the Dual-Sequence LSTM Architecture [13]. Wang et al. include two LSTM architectures in their work; the first is a basic LSTM that processes MFCCs extracted from a speech sample, and the second is a dual-sequence LSTM fed simultaneously by two mel-spectrograms with different time-frequency resolutions. The final classification is based on the average calculated from the outputs of the standard LSTM and the dual-sequence LSTM. Zang et al., in their paper called "A Study on Speech Emotion Recognition Model Based on Mel-Spectrogram and CapsNet", used mel-spectrograms to classify emotions from voices and compared the performance of the Support Vector Machine (SVM), CNN, LSTM, and CapsNet algorithms. The results presented in the paper show dominance of a capsule network over the other studied algorithms. The literature also includes studies on datasets in different languages [18].

Another approach without which this subsection would not be complete is the usage of transformers [19]. For example, Tan and Soleymani, in their work published in 2022, used a pre-trained Audio-Visual Transformer for emotion recognition and achieved promising results. In addition, they used spectrograms, mel-spectrograms, and features extracted by TRILL [20] for auditory modality.

Another essential aspect may be the speed of inference, and this may be important if the goal is to implement a system that works in real time. In this case, lightweight convolutional neural networks can be an appropriate choice [12]. CNNs are proven to be an efficient method that can be optimized and reduced in size without significant performance losses and are commonly used when the input data are represented in 2D format [16,21–23].

3. Selected Approach

The method studied in this paper recognizes emotions based on the retrieved spectrograms [7] and mel-spectrograms [24] from short voice recordings. It can be seen in the literature that when using CNNs, the primary approach to input data is to use spectrograms, mel-spectrograms, or raw speech signals [25]. In this study, it was decided to combine spectrograms and mel-spectrograms and investigate which feature extraction method performs better. After a systematic literature review, it is ambiguous which data type is better suited for speech emotion classification problems, and a comparison of their performance is nowhere to be found. These types of feature extraction were chosen because previous studies have shown that it is more suitable than, for example, raw signal combined with CNN [26]. This is why it is important to measure which data type to choose and support the decision with the results of experiments. Previous studies introduced by many researchers in the past [16,17,21–23,27] have proven that convolutional neural network (CNN) architecture is suitable for the problem, so this type of architecture is used in all experiments in this work since it is relatively simple. The goal is not to develop the best possible model but to highlight the researchers' challenges. This type of architecture is applied in all experiments in this work. However, using only a single dataset is common [5,6], which does not necessarily show whether the trained models can be utilized in a real environment. In this article, five different datasets are employed, namely CREMA-D, RAVDESS, SAVEE, IEMOCAP, and TESS; the exploitation of such amounts of data has not been usedin studies on SER before.

4. Datasets

The chosen datasets have been used in a few different configurations, individual and mixed with others. They differ in the number of emotions and the distribution of actors by gender. Information on datasets is included in the following subsections.

4.1. CREMA-D

CREMA-D stands for Crowd-sourced Emotional Multimodal Actors Dataset [28]. It presents multimodal data, meaning data corresponding to multiple modalities, visual and audio data. The dataset consists of 7443 clips prepared with the participation of 91 actors. Categorical emotion labels were obtained using crowdsourcing from 2443 raters, making this dataset reliable. The emotions presented in CREMA-D are neutral, happiness, anger, disgust, fear, and sadness. Such emotional examples from many different actors make it possible to train neural networks exclusively on this dataset.

4.2. RAVDESS

RAVDESS represents the Ryerson Audio-Visual Database of Emotional Speech and Song [29]. This dataset delivers two types of data: speech and song. The dataset consists of 7356 files from 24 professional actors—12 males and 12 females—from which only 1440 are speech audio-only files. The emotions represented in RAVDESS are: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. In this article, not all of the emotions from the RAVDESS dataset were used. Emotions of calm and surprise were dropped because they rarely occur in different datasets, and the RAVDESS dataset is too small to be used solely for deep neural network (DNN) training. Therefore, the number of examples of each emotion is a little uneven, as with the "neutral" emotion, for which there are only 48 files compared to 96 for others.

4.3. SAVEE

The name SAVEE is short for Surrey Audio-Visual Expressed Emotion [30]. The dataset consists of audio, visual, and audio-visual modalities. The authors of the database have collected the recordings of four English male actors expressing seven emotions. These emotions are anger, disgust, fear, happiness, sadness, surprise, and neutrality. This dataset presents an unbalanced distribution of classes as it consists of 90 examples of "neutral" emotion, which is two times more frequent than the others. It is also a relatively small dataset in terms of artificial neural network (ANN) training, so no research was conducted on this dataset alone. However, it was a perfect fit to combine it with the RAVDESS dataset, which presents a smaller number of "neutral" emotions but is also not big enough.

4.4. TESS

Toronto emotional speech set—TESS [31] represents seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. This dataset contains recordings of two females that are aged 26 and 64 years old. The TESS consists of 2800 audio files. Even though the number of audio files is large, preparing the CNN model solely on this dataset would be almost impossible as it would require having half of the dataset reserved for the test because only two actors were taking part in the study. That is why it is used only in combination with datasets: CREMA-D, RAVDESS, SAVEE, and IEMOCAP.

4.5. IEMOCAP

The Interactive Emotional Dyadic Motion Capture [32] is a commonly used database for emotion classification [8,33]. It serves multiple modalities: Motion Capture Face Information, Speech, Videos, Head Movement and Head Angle Information, Dialog Transcriptions, and Word-level, Syllable-level, and Phoneme-level alignment. The IEMOCAP dataset presents a highly unbalanced distribution of classes which is why it was only used for additional studies with only four emotions. Since the "Happy" emotion is considered crucial to be recognized and given that it has a small number of occurrences in the dataset,

it was merged with the emotion "Excitement," which is a common technique presented in other articles [8,34]. Another common approach presented in the same articles is to use this dataset only to recognize four emotions because of an insufficient number of examples of other emotions.

4.6. Comments on Datasets

The main focus was to gather as much data as possible, but some restrictions had to be established. The selected dataset had to be annotated with at least five different emotions and in the form of an audio file in the wav format as long as it represents uncompressed audio. Another restriction was to have at least two actors while preparing the dataset. It was necessary to be able to obtain samples from one actor in a training set and the other in a test set because having interdependent data in both sets could induce misleading results. This situation is discussed in chapter 6 by data analysts [35] and pertains to some papers [36,37]. The number of retrieved labeled recordings for each emotion from each dataset is shown in Table 1. Datasets with fewer actors than four are used only in combination with other datasets, and thus never alone. Finally, only audio files of recorded speech were selected from each dataset as this research considers only this modality.

Table 1. The amount of retrieved labeled recordings for each emotion from each dataset.

Dataset	Anger	Disgust	Fear	Happy	Neutral	Sad
CREMA-D	1271	1271	1271	1271	1087	1271
RAVDESS	192	192	192	192	48	192
SAVEE	60	60	60	60	120	60
TESS	400	400	400	400	400	400
IEMOCAP	1103	2	40	1636 [1]	1708	1084

[1] This number denotes merged recordings of excitement (1041 recordings) and happiness (595 recordings) emotions.

5. Architecture

Recent advancements in the field of SER demonstrated that the usage of Convolutional Neural Networks (CNNs) can produce satisfying results [7,8]. Creating the best possible model was not the goal of this research. The following sections describe studies that have been conducted and answer specific questions, addressing specific problems. Considering the above, the basic model architecture was developed and used in all experiments with some minor adjustments to make training possible. The base model consists of several convolution layers, and max-pooling layers to which a dropout has been added. The network is completed with a flattened layer and two dense layers. The architecture is shown in Figure 1.

The input data used in experiments are spectrograms and mel-spectrograms extracted from raw wav files. The sampling rate is 22,050 Hz. The sizes of both data types were $231 \times 349 \times 3$.

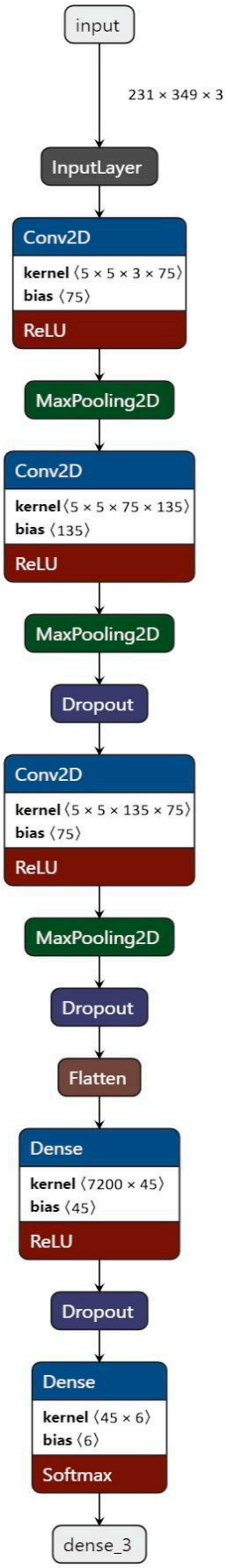

Figure 1. The artificial neural network architecture used in experiments.

6. Performance of Different Feature Extraction Methods

There are many options with regard to the input data format for emotional speech recognition. For example, one can use raw audio data in a wav format [12,38] and send it to a neural network or use Mel Frequency Cepstral Coefficients (MFCCs) [39–41], another audio representation format. However, this chapter focuses on establishing whether the basic models perform better with data in the form of a spectrogram or a mel-spectrogram.

There have been previous experiments conducted that focused on creating a CNN model and measuring its performance. For a better comparison between experimental results, a similar architecture consisted of several convolutional and max-pooling layers followed by two dense layers. Only minor adjustments were made to achieve better results as experiments were run on various datasets that differed in size. The main focus was on basic Convolutional Neural Networks. The chosen approach was to design artificial neural networks and train them from scratch using two input variants: spectrograms and mel-spectrograms. For spectrograms, no padding was applied, which is the opposite approach to using mel-spectrograms as input, where padding was introduced. For regularization, dropout layers were added. The experiments aimed to present the differences in performance between CNNs that were similar but used different datasets. In each experiment, two separate models with different input data were developed, one with spectrograms and one with mel-spectrograms, so that the difference in the usage of these two types of spectrograms is also checked. Moreover, the experiments were performed with different datasets and combinations of multiple datasets. The division of the training and testing sets was controlled, and it was ensured that the actors from the training set did not repeat in the test set.

Results presented in Table 2 indicate a slight advantage of using mel-spectrograms instead of spectrograms, as in almost every experiment, models using mel-spectrograms were able to achieve better scores. Two tasks with different complexity can be differentiated: the classification of four emotions and the classification of six emotions. The best results for classifying four emotions were acquired by a model trained on four datasets (CREMA-D, SAVEE, RAVDESS, TESS), where the test accuracy was 55.89%. For the recognition of six emotions, the highest test accuracy was equal to 57.42%, achieved by the model trained and tested on the CREMA-D dataset. To verify the reliability of the study, ResNet18 [42] was used as a popular architecture. ResNet18 achieved better results on spectrograms than custom CNNs when working on the same combination of data. The latter allows for the comparison between the results of the ResNet18 architecture and the custom CNNs using mel-spectrograms.

Table 2. Test accuracy in each experiment with differentiation of the type of input data.

Architecture	Datasets Used	Spectrogram	Mel-Spectrogram
Custom CNN architecture	CREMA-D (6 emotions)	0.4675	0.5366
	SAVEE, RAVDESS (6 emotions)	0.3256	0.3000
	IEMOCAP (4 emotions)	0.1439	0.5326
	CREMA-D, RAVDESS, SAVEE, TESS (4 emotions)	0.5245	0.5589
	CREMA-D, RAVDESS, SAVEE, TESS (6 emotions)	0.4331	0.5742
	CREMA-D, IEMOCAP, RAVDESS, SAVEE, TESS (4 emotions)	0.5032	0.5558
ResNet18	CREMA-D, RAVDESS, SAVEE, TESS (6 emotions)	0.4970	0.5537

Based on the results described above, the mel-spectrograms showed better results, and thus, they were used for further experiments, as is presented later in the article.

7. The Importance of Data Division into Training and Test Sets

As is generally known, incorrectly split data in an SER task can lead to misleading results from deep learning algorithms. For example, suppose the recordings of the same actor and the same emotion are mixed up in the training and test data sets [36,37]. In this case, a high result may be observed during the evaluation. Still, because the model would learn some specific features of the given actors, it would not be well prepared for a completely new speaker, leading to the model's failure during deployment. The reason for such behavior would be the model's ability to distinguish particular actors and their emotions but a lack of the ability to classify the emotions of unknown speakers in a real-life environment.

However, there has been no attempt in the literature to show what impact this can have on performance. Three experiments were conducted, first on the TESS and IEMOCAP datasets, and then on all datasets combined to demonstrate the difference in performance between correctly and incorrectly split data. In the proper split, samples of the same actor are not mixed between the training and test set. Firstly, the model was trained and then evaluated on properly split data. Subsequently, the same procedure was repeated for the data split without consideration of separating actors among training and test datasets. Finally, the same comparison was prepared for all three cases.

Table 3 presents the results of the three experiments on two different datasets and a combination of all datasets. Similarly to the previous experiment, ResNet18 was used as a reference. The most striking difference can be seen in the first experiment on the TESS dataset. Nevertheless, in all cases, a random split produced better accuracy results. This is due to the model's ability to distinguish between actors and their emotions and not only emotions. There is a risk that the model would not be able to handle a completely new actor.

Table 3. Test accuracy in each experiment with differently divided data.

Architecture	Datasets Used	Proper Split	Random Split
Custom CNN architecture	TESS	0.4416	0.9979
	IEMOCAP	0.5326	0.6913
	ALL DATASETS	0.5558	0.6596
ResNet18	CREMA-D, RAVDESS, SAVEE, TESS	0.5537	0.6429

8. Human-Based Speech Emotion Classification on CREMA-D

An additional study was conducted involving the classification of emotions in recordings from the CREMA-D dataset by humans to verify the possible results. This dataset was chosen for this study because it is the biggest and one of the most recent datasets. Regarding preparing this dataset, Cao et al. published [28] a detailed description of the data collection process along with some statistical analysis.

This dataset was selected for the study because it was labeled during crowdsourcing. The crowdsourcing process relies on the labeling of data by volunteers. In the experiment, 54 volunteers aged 22–58 of Polish nationality participated. The subjects were asked to classify emotions for the presented recordings. Thirty recordings were randomly chosen and used, five recordings per emotion. The experiment was in the form of an online questionnaire. Recordings were played as many times as requested by participants.

The confusion matrix is presented in Table 4. The most challenging emotions to classify were disgust and sadness, for which the correct choices were not the most common answers. The interviewees classified anger with the highest accuracy of 76%. Overall, the mean score achieved during the study was 14.63 (48.76% accuracy), with scores ranging from 2 to 21 where 30 is the highest possible score. Additionally, the median was equal to 15.

Table 4. Confusion matrix of human-based speech emotion recognition. Intensity of the red color corresponds to how many times the particular emotion was selected by the respondents.

Confusion Matrix of Emotion Classification						
	Anger	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**
Anger	207	62	20	18	10	6
Disgust	18	71	11	13	22	26
Fear	9	15	129	38	7	39
Happy	7	22	9	109	5	2
Neutral	26	73	53	81	194	117
Sad	3	27	48	11	32	80

To verify if the answers were selected at random, a statistical t-test was performed. The chosen null hypothesis (H_0) indicates that the mean for a sample is equal to 5, which is the accuracy at the level of 16.67%. Therefore, the null hypothesis has been rejected with a p-value lesser than 0.001. These results confirm that humans are able to distinguish emotions from the recordings.

Another statistic performed was the calculation of the confidence intervals for the mean estimation in the population. The confidence intervals acquired via this study are (0.4562848288, 0.5177891712), which means that based on the questionnaire results, one may be confident that the population mean will fall from 95% to between ~45.63% and ~51.78%.

In the article describing the CREMA-D dataset [28], it is stated that the labeling process used a crowdsourcing method, but the labels are only attached as additional information; the labels presented in the name of each file, however, are derived from what kind of emotion actors were trying to imitate. Hence the study presented above gives another set of possible labels for the 30 audio files included, and in the results, these files have three possible labels. These annotations are compared in the table, which can be found in the appendix. Comparing labels from the study with those from CREMA-D crowdsourcing resulted in six disagreements, which is significant when considering a sample size of 30. These disagreements raise concerns about the number of annotators used in CREMA-D as each audio file was annotated by only 7–11 people compared to the 54 in the study above. These conclusions also present how complex the data preparation process is as no specific guideline clearly states all the necessary steps to produce a good dataset for artificial intelligence (AI) models.

9. Conclusions

This study investigated the difference in the performance of CNN models while using spectrograms and mel-spectrograms. For this, two different architectures were exploited—the popular ResNet-18 and a custom CNN architecture similar to the classic LeNet. Most of the conducted experiments demonstrared that the exploitation of mel-spectrograms as a feature extraction method significantly improves the accuracy metric. However, there was only one experiment where models trained on spectrograms outperformed the ones trained on mel-spectrograms, and it occurred by a small margin. This leads to the conclusion that it is usually better to choose mel-spectrograms as a data processing method. Despite the effectiveness of mel-spectrograms in speech recognition—which is an outcome that one could have anticipated—it can be seen in the literature that many authors still use spectrograms. Our goal, therefore, was to visualize the differences in the effectiveness of CNN training in both cases quantitatively. The results are presented in the form of a benchmark concerned with different datasets and their combinations, and the type of input data.

Additionally, a model trained on all gathered datasets was also prepared. Despite not showing the best results in terms of accuracy, it has the most significant potential for a real-world environment. Although the data were collected from different actors and using

different microphones, the performance of the trained model should be verified in further studies in real-life scenarios.

The performance of a much bigger ResNet-18 architecture was slightly worse compared to our custom model. This shows that there is no real advantage in using bigger models, as the available data model of 800 thousand parameters (the custom CNN model) outperformed one with 11 million (Resnet-18). The smaller model would be quicker in inference after deployment and significantly lighter. One of the goals of this article is to show how important the proper way of splitting datasets is for training and testing. Researchers always strive for the best possible performance of their AI models, but sometimes the problem of interdependent data may be overlooked [36,37]. Even though some of the research claims to have produced excellent results in the study of SER, they are difficult to compare with other results unless there is no specific strategy for dataset splitting [5,9,10,35]. Some of the papers addressed the problem, but unfortunately, the results cannot be verified directly, as software is not always shared with the research community. Therefore, we decided to show the results of experimental comparisons in this regard.

Another consideration of this article is how well-prepared for SER the currently available datasets are. The study presented in Section 5 shows that the dataset might not have the highest possible quality. Based on this research and the interview participants' comments, the authors found no particular need to classify more than four emotions. For many people, emotions like disgust or sadness are hardly ever conveyed in speech, and it is even harder to imitate them by actors preparing the datasets. Suppose examined subjects classify emotions from speech with an average accuracy of just under 49% (study presented in Section 5); in that case, it is difficult to establish what should be expected from machine learning models. Namely, will 100% accuracy be possible in the future, and what would it mean for people if machines could classify emotions that cannot even be described by the speakers being tested? Since it is not yet possible to solve these problems at this stage of research development, they remain rhetorical questions.

The source code of the software developed by the authors has been shared on the GitHub platform [43].

Author Contributions: M.Z. and A.P. are the main and equal contributors. Conceptualization, M.Z. and A.P.; Methodology, M.Z. and A.P.; Software, M.Z., A.P., P.N., M.O. and K.K.; Supervision, A.C.; Validation, M.Z. and A.P.; Writing—original draft, M.Z. and A.P.; Writing—review and editing, M.Z., A.P., A.C., P.N., M.O. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Gdansk University of Technology. Internal grant No. 033014.

Data Availability Statement: The research was carried out using databases available on the Internet [28,29,31] and databases available on request [30,32], the sources of which are cited in the paper.

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

References

1. Milner, R.; Jalal, M.A.; Ng, R.W.M.; Hain, T. A Cross-Corpus Study on Speech Emotion Recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019. [CrossRef]
2. el Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
3. TTsouvalas, V.; Ozcelebi, T.; Meratnia, N. Privacy-preserving Speech Emotion Recognition through Semi-Supervised Federated Learning. In Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Pisa, Italy, 21–25 March 2022.
4. Deschamps-Berger, T.; Lamel, L.; Devillers, L. End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings. In Proceedings of the 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 28 September–1 October 2021. [CrossRef]
5. Ristea, N.-C.; Ionescu, R.T. Self-Paced Ensemble Learning for Speech and Audio Classification. *arXiv* **2021**, arXiv:2103.11988.
6. Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L.; Schmauch, B. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. *arXiv* **2018**, arXiv:1802.05630. [CrossRef]

7. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017. [CrossRef]
8. Padi, S.; Sadjadi, S.O.; Sriram, R.D.; Manocha, D. Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021. [CrossRef]
9. Lee, K.H.; Kim, D.H. Design of a Convolutional Neural Network for Speech Emotion Recognition. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 21–23 October 2020; pp. 1332–1335. [CrossRef]
10. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Mansor, H.; Kartiwi, M.; Ismail, N. Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks. In Proceedings of the 2020 6th International Conference on Wireless and Telematics (ICWT), Yogyakarta, Indonesia, 3–4 September 2020; pp. 1–6. [CrossRef]
11. Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499.
12. Pandey, S.K.; Shekhawat, H.S.; Prasanna, S.R.M. Emotion Recognition from Raw Speech using Wavenet. In Proceedings of the TENCON 2019—2019 IEEE Region 10 Conference (TENCON), Kochi, India, 17–20 October 2019; pp. 1292–1297.
13. Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; Tarokh, V. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6474–6478. [CrossRef]
14. Zhang, W.; Jia, Y. A Study on Speech Emotion Recognition Model Based on Mel-Spectrogram and CapsNet. In Proceedings of the 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 10–12 December 2021; pp. 231–235. [CrossRef]
15. Huang, C.; Narayanan, S.S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 583–588.
16. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
17. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* **2019**, *7*, 117327–117345. [CrossRef]
18. Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. *Electronics* **2020**, *9*, 1725. [CrossRef]
19. Tran, M.; Soleymani, M. A Pre-Trained Audio-Visual Transformer for Emotion Recognition. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 4698–4702. [CrossRef]
20. Shor, J.; Jansen, A.; Maor, R.; Lang, O.; Tuval, O.; Quitry, F.d.; Tagliasacchi, M.; Shavitt, I.; Emanuel, D.; Haviv, Y. Towards Learning a Universal Non-Semantic Representation of Speech. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020.
21. Zheng, W.; Yu, J.; Zou, Y. An experimental study of speech emotion recognition based on deep convolutional neural networks. In Proceedings of the 2015 International Conference on IEEE Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 827–831.
22. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882, preprint.
23. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. *arXiv* **2014**, arXiv:1404.2188.
24. 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]
25. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* **2021**, *10*, 1163. [CrossRef]
26. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
27. Stolar, M.N.; Lech, M.; Bolia, R.S.; Skinner, M. Real time speech emotion recognition using RGB image classification and transfer learning. In Proceedings of the 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), Surfers Paradise, Australia, 13–15 December 2017; pp. 1–8.
28. 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]
29. 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]
30. Jackson, P.; Haq, S.U. *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*; University Surrey: Guildford, UK, 2014.
31. Pichora-Fuller, M.K.; Dupuis, K. Toronto emotional speech set (TESS). *Sch. Portal Dataverse* **2020**. [CrossRef]

32. 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. *J. Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
33. Neumann, M.; Vu, N.T. Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7390–7394. [CrossRef]
34. Jalal, M.A.; Milner, R.; Hain, T. Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 4113–4117.
35. 'Using CNN for Speech Emotion Recognition—What Is Wrong with It?' Sopra Steria. Available online: https://www.soprasteria.se/blogg/using-cnn-for-speech-emotion-recognition (accessed on 28 September 2022).
36. Sehgal, S.; Sharma, H.; Anand, A. Smart and Context-Aware System employing Emotions Recognition. In Proceedings of the 2021 2nd International Conference for Emerging Technology (INCET), Belgaum, India, 26–28 May 2021. [CrossRef]
37. Sahoo, S.; Kumar, P.; Raman, B.; Roy, P.P. A Segment Level Approach to Speech Emotion Recognition Using Transfer Learning. In Proceedings of the Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, 26–29 November 2019; Revised Selected Papers, Part II, Auckland, New Zealand; pp. 435–448. [CrossRef]
38. Mocanu, B.; Tapu, R. Emotion Recognition from Raw Speech Signals Using 2D CNN with Deep Metric Learning. In Proceedings of the 2022 IEEE International Conference on Consumer Electronics (ICCE), Pingtung, Taiwan, 17–19 July 2022; pp. 1–5.
39. Nasrun, M.; Setianingsih, C. Human Emotion Detection with Speech Recognition Using Mel-frequency Cepstral Coefficient and Support Vector Machine. In Proceedings of the 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Delft, The Netherlands, 12–16 July 2021; pp. 1–6.
40. DMuttaqin; Suyanto, S. Speech Emotion Detection Using Mel-Frequency Cepstral Coefficient and Hidden Markov Model. In Proceedings of the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 10 December 2020; pp. 463–466.
41. Rajasekhar, A.; Hota, M.K. A Study of Speech, Speaker and Emotion Recognition Using Mel Frequency Cepstrum Coefficients and Support Vector Machines. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), Tamilnadu, India, 3–5 April 2018; pp. 114–118.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
43. GitHub repository. Available online: https://github.com/Amikirami/Speech-Emotion-Recognition (accessed on 1 October 2022).