engineering proceedings

MDPI

Proceeding Paper

Feature Fusion for Emotion Recognition †

Awais Mahmood

College of Applied Computer Science, King Saud University, Riyadh 11451, Saudi Arabia; mawais@ksu.edu.sa; Tel.: +966-545798315

† Presented at the 8th International Electrical Engineering Conference, Karachi, Pakistan, 25–26 August 2023.

Abstract: The information that needs to be communicated to the partner depends heavily on how emotions are expressed in human communication. There are many different ways that people can communicate their feelings. Body language, facial expressions, eye contact, humor, and tone of speech are all potential indicators. The people of the world speak a variety of languages, but even when communicating without understanding the other partner's language, one can nearly always infer what they are trying to say from their emotional expressions. The combined features were provided to CNN. The RAVDESS dataset was used in this work and all the emotions were used for accuracy. The feature parameters used for recognition include the Mel spectral coefficients, chroma_STFT, zero crossing, RMS, pitch frequency, and Mel spectrogram. The results have shown that our proposed technique achieved 87.17% accuracy, which outperforms the existing literature.

Keywords: speech emotion recognition; CNN; MFCC; RMS

1. Introduction

Emotions are defined by the Oxford dictionary as "a strong feeling such as love, fear or anger; the part of a person's character that consists of feelings". There are many types of emotions and each one refers to a subtle tone or movement. People may not care for others' emotions because either they are not aware, or they do not know how to interpret them; most probably, it is the latter. With the power of technologies, we have today; new novel ways have been introduced to interpret emotions. The speech emotion recognition system is useful in psychiatric diagnosis, lie detection, call center conversations, customer voice review, and voice messages [1].

Emotion recognition identifies and detects human feelings, which have an important impact on our daily lives. A computer is used to make it easier to interpret emotions and it is applied in many aspects. The target is to create a system that can classify human emotions into the different categories already established in human studies. These systems can be used: speech recognition, image recognition and brainwaves, and so on [2–4]. These systems are being created to improve humanity's social psychology and interpretation. Many models can be used to benefit emotion recognition, which is mostly based on machine learning and deep learning. Building a model to serve emotion recognition, like building any other model, starts by collecting the training and testing data. There are a lot of available datasets that contain audio samples such as RAVDESS and TESS.

Different types of feature extraction/selection techniques were used to classify the emotions automatically. These features were used basically for speech recognition specifically such as MFCC, GFCC, RASTA, etc. [4–7]. These can be used for any speech-based technology. After extracting the needed features, those features are fed to a classifier that will detect emotions and assign them to their correct classes. However, before selecting the model, it is important to keep a lookout for computational complexity and how the algorithm scales with different features, patterns, and categories. Common techniques for classification are SVM, KNN, CNN, and MLP [8–14].

check for updates

Citation: Mahmood, A. Feature Fusion for Emotion Recognition. *Eng. Proc.* **2023**, *46*, 20. https://doi.org/10.3390/engproc2023046020

Academic Editors: Abdul Ghani Abro and Saad Ahmed Qazi

Published: 21 September 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Finally, start training the model using training data and learn the parameters of the model; it is possible to experiment with supervised learning methods and unsupervised ones, which will lead us into the evaluation phase, to see the accuracy of the model after the training has been done and check if the model suffers from problems such as overfitting, underfitting, or being biased.

There have been many research works published during the last decade where different types of feature extraction techniques and modeling techniques were used to classify emotions. In [5], the authors extracted MFCC features and trained the model using GMMs and HMMs to classify emotions. The dataset used was RAVDESS. Their findings showed that SVM provided a classification rate of 86.6% higher than GMM and HMM. The authors in [7] used feature extraction techniques MFCC and STFT with a deep learning approach CNN (convolutional neural networks). They used the RAVDESS dataset and scored an 85% accuracy rate.

Mustageem in [8] aimed to improve the accuracy without increasing the complexity of the model with deep stride convolutional neural network (DSCNN) and learning by salient and discriminative features from the spectrogram of speech signals and using the SoftMax classifier and testing the model on a different dataset. The author came up with a new idea to test the model using cross-dataset experiments RAVDESS and IEMOCAP with four emotions, decreasing the model accuracy from 81% to 56%, which proves that different languages and cultures have a huge impact.

Gender can affect the accuracy according to [10]. The authors in this research used feature extraction techniques such as MFCC, CHROMA-GRAM, and Mel Spectrogram and fed into three different classifiers, SVM, DNN, and KNN, to see if gender would affect the accuracy of the model. Their results showed an increase in all of the classifiers such that DNN scored 55.89% without gender and then scored 57.52% with it. Adding intensity to the equation increased the accuracy to 77.92%.

A comparison between MFCC and GFCC is presented in reference [11]. The proposition that GFCCs are superior to MFCCs, using the RAVDESS dataset and those two models, proved that GFCC is accurate but not superior. It is only accurate with a small additional fraction of what MFCC predicts. The accuracy achieved for emotion recognition was 75.6%.

Vani, H. [12] presents the use of a decision tree and CNN as a classifier to classify emotions. The performance of CNN has been identified as the best classifier for emotion recognition. Emotions were recognized with 72% and 63% accuracy using CNN and decision tree algorithms, respectively. MFCC features were extracted from the audio signals and the model was trained, tested, and evaluated accordingly by changing the parameters. CNN outperformed the decision tree and achieved 72.2% accuracy.

Authors in reference [13] used two datasets: TESS and RAVDESS. The machine-learning technique used was DNN and the features extracted were MFCC. Their feature is Log Mel Spectrogram, and their architecture was 2D CNN with a global average pool, but they proved that if the number of classes is increased, the accuracy will start to decrease. After the literature survey, it was observed that the dataset most of the researchers used was RAVDESS due to its speech quality, the number of different emotions, and the number of samples for each emotion type. The second observed thing was that MFCC was used for baseline or sometimes fused with other techniques for the best accuracy. Due to the success of CNN during the last 5 years, it was decided to evaluate its performance using this machine-learning technique. The next section will explain the feature extraction and modeling technique, Section 3 presents the proposed technique, whereas Section 4 illustrates the experimental result and discussion. In the end, the conclusion is presented.

2. Feature Extraction and Modeling Technique

2.1. Feature Extraction Technique

There are different feature extraction techniques that can be utilized, such as MFCC, PLP, LPCC, etc. In this research work, the fusion of different types of features was used but the main feature extraction technique was Mel-frequency cepstral coefficients (MFCC) since

it is a conventional technique, reliable, easy, and fast to compute [15,16]. The second type of feature extracted was zero-crossing rate. It is mostly used to extract one feature by finding the time between each zero crossing to determine the emotion; for example, if the speaker is angry, the frequency goes up which means it takes more time for it to cross the zero threshold. The third type of feature was short-term Fourier transformation (chroma_STFT). It is widely used in music analysis to classify pitch into 12 classes. In this work, it is used for categorizing pitch sensitivity, which is greatly affected by emotions. The next feature used was Root Mean Square (RMS); it captures the average of the audio wave by dividing the peak value by the square root of two. RMS was used in this work as explained; testing the model concluded that using RMS improves the recognition rate of the model. The fifth feature castoff was pitch frequency (PF). It is extracted by calculating the difference between the peaks obtained from the autocorrelation of the speech signal, and the pitch of the speech is estimated. The last type of feature used was the Mel Spectrogram. It is a frequency of spectrogram that is converted to a Mel scale. A spectrogram shows a plot where one axis represents frequencies and the other represents time. First, the speech signal is decomposed into a number of frames. Later, each frame is multiplied by the Hamming window. FFT is applied to each windowed frame and passes through the Mel filter. Later, the signal magnitude is decomposed into its components for each window corresponding to the frequencies in the Mel-scale.

2.2. Modeling Technique

Depending on the feature extraction techniques explained previously, the best classifier is the convolutional neural network (CNN), testing with many different hidden layers, until the best amount of layers gives the highest accuracy in training and testing data. CNN was used because of its flexibility [7,8]. Research work has shown that it is more compatible with speech datasets compared to commonly used techniques for classifications, e.g., SVM, KNN, MLP, etc., since it is considered an advanced and complex deep-learning technique [12,14–20].

3. Proposed Work

In this research work, five different types of features were used: those are MFCC, chroma_STFT, zero-crossing, RMS, PF, and Mel spectrogram. Later, for the preprocessing, the data was scaled by removing the mean and scaling to unit variance. Following the feature extraction and preprocessing phases, that data was ready to be trained into the proposed machine-learning technique CNN.

The structure of CNN may contain many convolutional layers depending on its usage. For this work, MFCC coefficients were used with a 1-dimensional image of the MFCC spectrogram which was converted from its source form into a target array using Conv1D as a start layer for the model. For this work, 13 layers were used to train the data, putting it through four 1-dimensional convolutional layers (Conv1D, Conv1D_1, Conv1D_2, Conv1D_3) throughout the entire model while having MaxPooling layers in between them.

4. Experimental Work and Discussion

There are different datasets that are widely used in speech emotion recognition (SER) mainly containing audio files ordered in categories and created by several actors. Based on the literature review, RAVDESS is the most used dataset when it comes to emotion classification due to the great quality of the speakers and recordings, having a lot of actors, and different genders. RAVDESS has eight emotions but it lacks a variety of sentences; the entire dataset is based on two sentences and the audio clip count is not equal for all the emotions. All emotions have 192 recordings except “Neutral” which has 96 recordings, which might affect the model accuracy. Table 1 presents the distribution of emotion type and number of samples.

Table 1. RAVDESS data distribution.

				Emotions							
	Actors	**Genders**	**Sentences**	**Natural**	**Calm**	**Happy**	**Sad**	**Angry**	**Fear**	**Disgust**	**Surprised**
RAVDASS	24	12 Male 12 Females	2	96	192	192	192	192	192	192	192

For the experimental setup, speech utterances were used. The preliminary result achieved an accuracy of 39.81%, where MFCC and chroma features were extracted and MLP was used for the modeling. Later on, MFCC features were extracted and fed to a CNN classifier that improved the accuracy to 40.2%. The detailed evaluation is provided in Table 2 with different types of evaluation metrics.

Table 2. Preliminary result for RAVDESS.

Emotion	**Precision**	**Recall**	**F1-Score**	**Over-All Accuracy**
Angry	0.47	0.80	0.59	40.2%
Disgust	0.34	0.43	0.38	
Fear	0.35	0.44	0.39	
Happy	0.27	0.43	0.33	
Neutral	0.72	0.23	0.35	
Sad	0.36	0.08	0.13	
Surprise	0.43	0.51	0.47	

The third experiment was conducted using MFCC-, chroma-, and Mel-classified features with the MLP modeling technique, improving the accuracy to 41%. Table 3 describes the features extracted with the modeling technique and accuracy. It clearly shows that CNN is better as compared to MLP.

Table 3. Preliminary results for CNN and MLP.

Emotion	**Experiment 1**		**Experiment 2**		**Experiment 3**	
	Features	**Classifier**	**Features**	**Classifier**	**Features**	**Classifier**
	MFCC, Chroma_Mel, contrast tonnetz	MLP	MFCC	CNN	MFCC, Chroma_Mel	MLP
Accuracy	39.81%		42.9%		41%	

The experiments (Table 4) show how different feature extraction techniques and classification techniques can have an impact on accuracy. In the first experiment, MFCC, Chroma_STFT, and Mel Spectrogram are used to extract features from the eight emotions category, which makes up a total count of 179 features. Then, the model was trained using Multilayer Perceptron (MLP), which achieved an accuracy of 71.05%. This result was much better as compared to the previous results. For the second experiment, three more feature extraction techniques ZCR, PF, and RMS were added to the modeling technique MLP.

Table 4. Accuracy using different feature extraction techniques.

Experiment No.	Feature Extraction	Classification	Accuracy
1	MFCC, Chroma_STFT, Mel spectrogram	MLP	71.07%
2	MFCC, Chroma_STFT, ZCR, RMS, PF, Mel Spectrogram	MLP	84.44%
3	MFCC, Chroma_STFT, ZCR, RMS, PF, Mel Spectrogram	CNN	87.17%

This method improved the recognition rate by 13%, which yielded 84.44%. Since most of the recent work used CNN, we tried to work with CNN where the features remain the same, and the recognition rate was raised to 87.17% as illustrated in Table 4. Figure 1 presents the accuracy graph for training and testing accuracy.

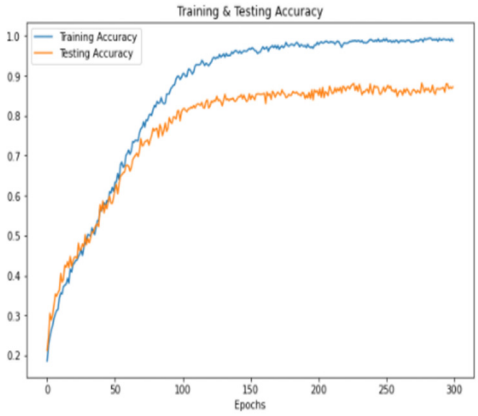


Figure 1. Accuracy graph.

Compared to the accuracy results of the literature review as shown in Table 5, the accuracy of this project has surpassed all the other references above, since most of the researchers decided to drop an emotion from RAVDESS. In this work, we also conducted an experiment and removed the "calm" emotion because it might conflict with the "neutral" emotion since they have similar features, and using multiple feature extraction techniques compared to the ones they used, that gave this project a boost in accuracy.

Table 5. Comparison of accuracy with the available literature.

Reference	Feature Extraction	Classifier	Accuracy
[7]	MFCC, STFT	CNN	85%
[10]	MFCC	DNN	77.92%
[11]	GFCC, Gama tone filter back	FCNN	75.6%
Proposed Technique	MFCC, Chroma_STFT, ZCR, RMS, PF, Mel spectrogram	CNN	87.17%

5. Conclusions

In this article, we presented the results of speech emotion recognition with the RAVDESS corpus. Different types of feature extraction techniques and machine-learning techniques were evaluated. Our proposed technique outperforms others using the RAVDESS dataset for seven different emotions. Different features hold the characteristics of emotions that complement each other and that results in the best accuracy. Different researchers obtained good recognition rates but with a smaller number of emotions. When it came to the complete corpus result, our proposed technique performs best. In the future, different types of datasets will be used, and even for the cross datasets, the proposed technique will be evaluated.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is freely available on https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio (accessed on 1 July 2023).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Zheng, S.; Du, J.; Zhou, H.; Bai, X.; Lee, C.H.; Li, S. Speech Emotion Recognition Based on Acoustic Segment Model. In Proceedings of the 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 24–26 January 2021; pp. 1–5.
2. Scotti, V.; Galati, F.; Sbattella, L.; Tedesco, R. Combining Deep and Unsupervised Features for Multilingual Speech Emotion Recognition. In Proceedings of the International Conference on Pattern Recognition, Talca, Chile, 7–19 April 2022; Springer: Cham, Switzerland, 2022; pp. 114–128.
3. Kerkeni, L.; Serrestou, Y.; Mbarki, M.; Raoof, K.; Mahjoub, M.A. Speech Emotion Recognition: Methods and Cases Study. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence, Funchal, Portugal, 16–18 January 2018. [CrossRef]
4. Moine, C.L.; Obin, N.; Roebel, A. Speaker Attentive Speech Emotion Recognition. In Proceedings of the International Speech Communication Association (INTERSPEECH), Brno, Czechia, 30 August–3 September 2021.
5. Nema, B.M.; Abdul-Kareem, A.A. Preprocessing signal for Speech Emotion Recognition. *Al-Mustansiriyah J. Sci.* **2017**, *28*, 157–165. [CrossRef]
6. Tripathi, S.; Ramesh, A.; Kumar, A.; Singh, C.; Yenigalla, P. Learning Discriminative Features using Center Loss and Reconstruction as Regularizer for Speech Emotion Recognition. In Proceedings of the Workshop on Artificial Intelligence in Affective Computing, Macao, China, 10 August 2019; pp. 44–53.
7. Huang, A.; Bao, M.P. *Human Vocal Sentiment Analysis*; NYU Shanghai CS Symposium: Shanghai, China, 2019.
8. Mustaqeem; Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* **2019**, *20*, 183. [CrossRef] [PubMed]
9. Guo, L.; Wang, L.; Dang, J.; Zhang, L.; Guan, H.; Li, X. Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018. [CrossRef]
10. Johnson, A. *Emotion Detection through Speech Analysis*; National College of Ireland: Dublin, Ireland, 2019.
11. Harár, P.; Burget, R.; Dutta, M.K. Speech emotion recognition with deep learning. In Proceedings of the 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2–3 February 2017. [CrossRef]
12. Liu, G.K. Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech. *arXiv* **2018**, arXiv:1806.09010.
13. Agtap, S. Speech based Emotion Recognition using various Features and SVM Classifier. *Int. J. Res. Appl. Sci. Eng. Technol.* **2019**, *7*, 111–114. [CrossRef]
14. 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. [CrossRef]
15. Tripathi, S.; Kumar, A.; Ramesh, A.; Singh, C.; Yenigalla, P. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. *arXiv* **2019**, arXiv:1906.05681.
16. Zhang, Y.; Du, J.; Wang, Z.; Zhang, J.; Tu, Y. Attention Based Fully Convolutional Network for Speech Emotion Recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018. [CrossRef]
17. Damodar, N.; Vani, H.Y.; Anusuya, M.A. Voice Emotion Recognition using CNN and Decision Tree. *Int. J. Innov. Technol. Explor. Eng. Regul. Issue* **2019**, *8*, 4245–4249. [CrossRef]
18. Venkataramanan, K.; Rajamohan, H.R. *Emotion Recognition from Speech*; Cornell University: Ithaca, NY, USA, 2019.
19. Jiang, W.; Wang, Z.; Jin, J.S.; Han, X.; Li, C. Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network. *Sensors* **2019**, *19*, 2730. [CrossRef] [PubMed]
20. Han, K.; Yu, D.; Tashev, I. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In Proceedings of the INTERSPEECH, Singapore, 14–18 September 2014; pp. 223–227.

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.