ΧΑΛΙΑ ΟΛΟ

Enhancing Speech Emotion Recognition Using Deep Convolutional Neural Networks

M M Manjurul Islam[*]
m.islam@ulster.ac.uk
Intelligent System Research Centre,
Ulster University
U.K.

Md Alamgir Kabir
md.alamgir.kabir@mdu.se
Artificial Intelligence and Intelligent
Systems Research Group, Mälardalen
University
Västerås, Sweden

Alamin Sheikh
18-39230-3@student.aiub.edu
American International
University-Bangladesh
Dhaka, Bangladesh

Muhammad Saiduzzaman
18-38829-3@student.aiub.edu
American International
University-Bangladesh
Dhaka, Bangladesh

Abdelakram Hafid
abdelakram.hafid@mdu.se
School of Innovation, Design and
Engineering, Mälardalen University
Västerås, Sweden

Saad Abdullah
saad.abdullah@mdu.se
School of Innovation, Design and
Engineering, Mälardalen University
Västerås, Sweden

ABSTRACT

Speech emotion recognition (SER) is considered a pivotal area of research that holds significant importance in a variety of real-time applications, such as assessing human behavior and analyzing the emotional states of speakers in emergency situations. This paper assesses the capabilities of deep convolutional neural networks (CNNs) in this context. Both CNNs and Long Short-Term Memory (LSTM) based deep neural networks are evaluated for voice emotion identification. In our empirical evaluation, we utilize the Toronto Emotional Speech Set (TESS) database, which comprises speech samples from both young and old individuals, encompassing seven distinct emotions: anger, happiness, sadness, fear, surprise, disgust, and neutrality. To augment the dataset, variations in voice are introduced along with the addition of white noise. The empirical findings indicate that the CNN model outperforms existing studies on SER using the TESS corpus, yielding a noteworthy 21% improvement in average recognition accuracy. This work underscores SER's significance and highlights the transformative potential of deep CNNs for enhancing its effectiveness in real-time applications, particularly in high-stakes emergency situations.

CCS CONCEPTS

• **Networks** → **Session protocols.**

KEYWORDS

Speech corpus, Human speech emotion recognition, Convolutional neural network applications, Long short-term memory neural networks

ACM Reference Format:
M M Manjurul Islam, Md Alamgir Kabir, Alamin Sheikh, Muhammad Saiduzzaman, Abdelakram Hafid, and Saad Abdullah. 2024. Enhancing Speech Emotion Recognition Using Deep Convolutional Neural Networks . In *2024 9th International Conference on Machine Learning Technologies (ICMLT) (ICMLT 2024), May 24–26, 2024, Oslo, Norway*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3674029.3674045

1 INTRODUCTION

With the rapid proliferation of intelligent technologies, the demand for emotion recognition is growing rapidly, drawing increased focus in both theoretical science and engineering, owing to its significant influence on social communication and decision-making [13]. In the realm of intelligent machines, emotion recognition is essential, as these machines might behave and make judgments in ways comparable to humans. Furthermore, it can enable seamless interactions with individuals, fostering smoother conversations. Ideally, intelligent machines should accurately comprehend human emotions, given their pivotal role in social interactions.

This need for accurate emotion recognition ties into the understanding of the human voice. The human vocal folds are utilized for producing sounds during acoustic activities such as talking, singing, laughing, shouting, and more. These vocal folds, also known as cords, act as the primary source of sounds generated within the human voice frequency, forming a crucial part of the human sound production system [23]. From this same general area of the body, various other sounds can also be produced, including unvoiced consonants, clicks, whistling, and whispering. The system responsible for human voice production comprises three main components: the lungs, laryngeal vocal folds, and articulators [23]. Pitch, loudness (sound pressure), timbre, and tone are key characteristics of the human voice and its associated speech patterns [2, 26].

Various approaches to speech emotion recognition (SER) have been explored in recent research. The studies explore diverse methods for enhancing speech emotion recognition (SER). Deep learning integration in GMM and DNN classifiers achieves a 92.3% recognition rate with multiple acoustic features [13]. Approaches like CNN, RNN, dilated CNN, and BiLSTM exhibit success in emotion identification [15, 21, 27]. Strategies such as majority voting, transfer

[*] Corresponding author.

ICMLT 2024, May 24–26, 2024, Oslo, Norway
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1637-9/24/05
https://doi.org/10.1145/3674029.3674045

learning [20, 27], and novel architectures like ADRNN and DeepResLFLB enhance precision and recall [11, 22]. Innovations like DiverseCatAugment and adaptation address context and emotion specific challenges [18, 19]. These advancements collectively contribute to improved SER accuracy.

Furthermore, Togootogtokh et al. [24] achieve impressive accuracy using DeepEMO and transfer learning, but the potential overfitting and generalization to other datasets remain to be assessed. Chen et al. [5] introduce P-TAPT, outperforming TAPT, yet the approach's effectiveness on broader emotion categories needs validation. Few-shot learning by Feng et al. [8] shows promise, but scalability to more complex emotional nuances requires further investigation. Dossou et al. [6] demonstrate FSER's success with mel-spectrograms, but robustness across diverse languages and emotional expressions remains unexplored. Han et al. [9] present prediction-based learning, though its adaptability to real-world applications beyond specific datasets should be examined. Han et al. [10] propose RE-based learning, but the framework's complexity and computational demands warrant consideration. Aftab et al. [1] offer a compact FCNN model, but its performance on less-controlled environments requires examination. Mustaqeem and Kwon [16] propose DSCNN architecture with size reduction, but the model's response to extreme noise conditions needs evaluation. Jing et al. [12] introduce a distinct neural architecture, yet its scalability to more intricate emotion categories remains uncertain. Chatterjee et al. [4] utilize 1-D CNN for emotion classification, but the model's adaptability to unseen contexts must be investigated. Krishnan et al. [14] present entropy-based feature extraction, but its generalization to noisy environments and varied speech types needs exploration. Mustaqeem and Kwon [17] propose ConvLSTM with sequence learning, but the model's robustness to different language patterns requires further scrutiny. Anvarjon et al. [3] introduce CNN with modified pooling, but its performance on larger datasets should be assessed. Trinh Van et al. [25] compare various deep neural network models, yet the methods' reliability across diverse demographic groups needs verification.

The connection between emotion recognition and speech becomes evident as speech provides a wealth of prosodic and auditory cues that can be extracted. A variety of emotional attributes can be utilized for emotion detection; however, this might extend the detection time. Hence, meticulous feature selection becomes essential. In our research, we employed a dataset encompassing diverse conversations exhibiting a spectrum of emotions. Employing various algorithms, we successfully pinpointed all emotions with accuracy [13]. These findings highlight the intertwining significance of understanding emotion and voice within the realm of intelligent technologies and human interaction.

In the realm of emotion recognition, where human expression and technological advancement converge, this work presents a comprehensive exploration with a focus on the precise classification of the seven primary human emotions. Leveraging the power of Deep Learning and Machine Learning, our study undertakes a rigorous analysis aimed at unraveling the intricate tapestry of human emotional states.

Our primary objectives are two-fold. First, we use advanced algorithms on our dataset to train different models that can accurately predict emotions. We focus on Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) as our key tools for achieving highly accurate emotion prediction. Next, we carefully study the different aspects of speech that show how each emotion is unique. We compare these aspects in great detail, aiming to find the small but important differences that define each emotion. We confirm the success of our endeavors by using a confusion matrix, which adds practical evidence to the accuracy we achieve.

The subsequent sections of this paper are structured as follows: Section 2 outlines our empirical methodology, Section 3 presents the results and engages in a detailed discussion of our findings, while Section 4 provides a summary and the conclusion of our work.

2 METHODOLOGY

2.1 Dataset description and preprocessing

Kate Dupuis and M. Kathleen Pichora-Fuller from Toronto University's Department of Psychology created the Toronto Emotional Speech Set (TESS) dataset for research [7]. This freely available dataset, hosted on Kaggle, comprises 2800 ".wav" audio files recorded by two native voice actresses aged 26 and 64, each at a bitrate of 390kbps. The audio files follow the carrier phrase "speak the word-" and encompass a list of 200 target single words. TESS encompasses seven emotion categories: disgust, anger, happiness, fear, neutrality, sadness, and pleasing surprise. Fig.1 depicts the class details of the TESS datasets. In details, the dataset is structured into 14 folders, with each actress having a folder for each emotion. Filenames begin with "OAF" for older actresses and "YAF" for younger ones. Each audio file label consists of three sections: actress code, target term, and expressed emotion. There are 400 audio files per emotion category.

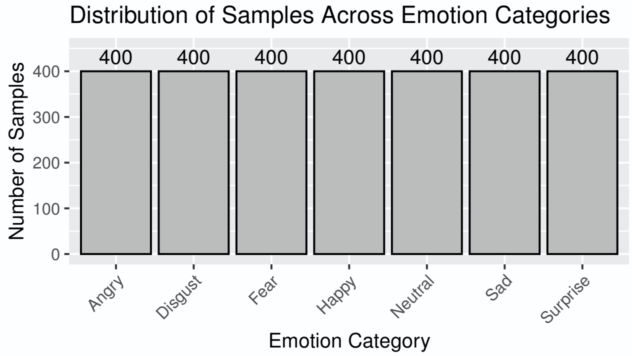

Figure 1: Class details of TESS dataset.

2.2 Methods

In this section, we elaborate on the methodology employed to preprocess and utilize the data for model training, testing, and validation. The subsequent figure provides an overview of the research methodology employed. The audio dataset is initially transformed into 2D images as part of a fundamental preprocessing step. These images are subsequently flattened to facilitate their application across various algorithms. The resulting flattened image vectors are then inputted into the models for the training process. Fig.2 illustrates the working process of emotion detection through speech using deep learning. The feature extraction sub-model utilizes either a CNN or LSTM deep CNN. This network is trained exclusively using the images present within dataset. Our model was trained over 20 iterations, affording you the freedom to experiment and meticulously adjust other parameters to your preference. In this context, we present one of the outcomes derived from training our network.

2.3 Audio signal analysis

The TESS audio files have undergone preprocessing to ease the challenges of working with raw recordings. This preprocessing minimizes the need for additional steps by researchers, reducing the complexity of handling the data. Working directly with original audio files is impractical due to their large size. Unprocessed audio presents challenges for machine learning models to effectively learn feature distributions, potentially resulting in poor overall performance. Providing substantial data directly to a model is computationally intensive and unfeasible for practical applications. As a solution, the audio files have been transformed into image format using spectrograms, significantly reducing data size and computational demands. This conversion enhances manageability and facilitates more efficient model training and analysis. Fig. 3 illustrates the process of audio signal analysis.

2.4 Tools

In order to address the initial problem effectively, we have leveraged a range of tools and libraries to facilitate our solutions. The implementation section encompasses diverse tools and libraries, each serving distinct purposes tailored to specific use cases. These resources play a pivotal role in supporting deep learning research, particularly within the sub-field of Natural Language Processing. Additionally, they contribute significantly to the formulation of various methods and enhance overall functionality. Notably, all tools and libraries utilized in this thesis are written in Python, aligning with the language used for the thesis implementation.

Convolutional neural networks have been employed in our experiment, utilizing functions such as conv2d, maxpool2d, and dropout to construct the model and ensure its regularization. Additional functions have played a role in model validation and interpreting outcomes. To partition the dataset into training and testing sets, we have predominantly utilized scikit-learn. The libraries from scikit-learn have also been instrumental in computing the confusion matrix for quantitative comparison. Further, when dealing with extensive audio data on a large scale for various purposes, such as voice identification or extracting personal traits from audio recordings, this experiment primarily relies on Librosa. Leveraging a range of signal processing techniques, including Chroma Energy Normalized, Mel-Frequency Cepstral Coefficients, and Zero Crossing Rate, Librosa serves a dual role. It aids in visualizing audio signals and facilitates the extraction of crucial features from them.

2.5 Performance metrics

The effectiveness of a classification model is assessed through a confusion matrix, which is an N x N matrix representing all target classes, encompassing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It compares the model's predicted values with the actual values, providing insights into the strengths and weaknesses of the classification model. We consider accuracy, precision, and recall for evaluating the model's performance. Recall measures the quantity of instances from a specific class that are correctly recognized. It signifies the ratio of correctly classified instances of a true class. Precision means the ratio of TP to the sum of FP and TP and recall measures the ratio of of TP to the sum of TP and FN. Here, TP denotes the positive class that is predicted correctly. TN represents the correctly predicted negative class. FP indicates the incorrectly predicted negative class. FN refers to the positive class that was predicted incorrectly.

3 RESULTS AND DISCUSSION

The CNN exhibited superior performance compared to other algorithms in our evaluation. Although it encountered some challenges in classifying instances based on categories, its performance was comparatively more favorable than the alternative methods. The accuracy dropped to 7% for the 7-class classification due to the algorithm's struggles. It's worth noting that training a CNN with multiple layers can be time-consuming, particularly without a powerful GPU. Additionally, the effectiveness of a ConvNet hinges on the availability of a substantial dataset for training purposes.

In the results (i.e., Table 1), our CNN model achieved an impressive 98% accuracy on the training dataset. Furthermore, precision and recall also reached 98%, underscoring the model's robust performance.

LSTM networks offer the advantage of propagating input values through multiple LSTM layers and across different time steps within a single LSTM cell. This layered approach facilitates a balanced parameter distribution, enabling comprehensive processing at each step. The reduced accuracy in the 7-class classification is attributed to the algorithm encountering difficulties. For ConvNets, a substantial dataset is essential for effective processing and training of LSTM models, particularly those with multiple layers. This requirement can lead to prolonged training times if a powerful GPU is not available.

In the table (i.e., Table 1), our CNN models achieved a 77% accuracy on the training dataset, accompanied by precision and recall values also at 77%. CNN demonstrates superior performance with an accuracy of 98%, while LSTM achieves an accuracy of 77% (Table 2). This indicates that CNN provides more accurate overall predictions for the 7-class task. For each of the seven classes (Table 3), CNN consistently achieves precision values above 0.93, with some classes reaching as high as 0.99. LSTM, on the other hand, exhibits variable precision values ranging from 0.59 to 0.95. This highlights CNN's ability to maintain high precision across multiple classes. Similar to precision, CNN consistently exhibits strong recall values across the classes, ranging from 0.95 to 1.00. In contrast, LSTM's recall values vary, with the lowest being 0.59 and the highest being 0.95. This indicates that CNN is generally better at correctly identifying instances across all classes.

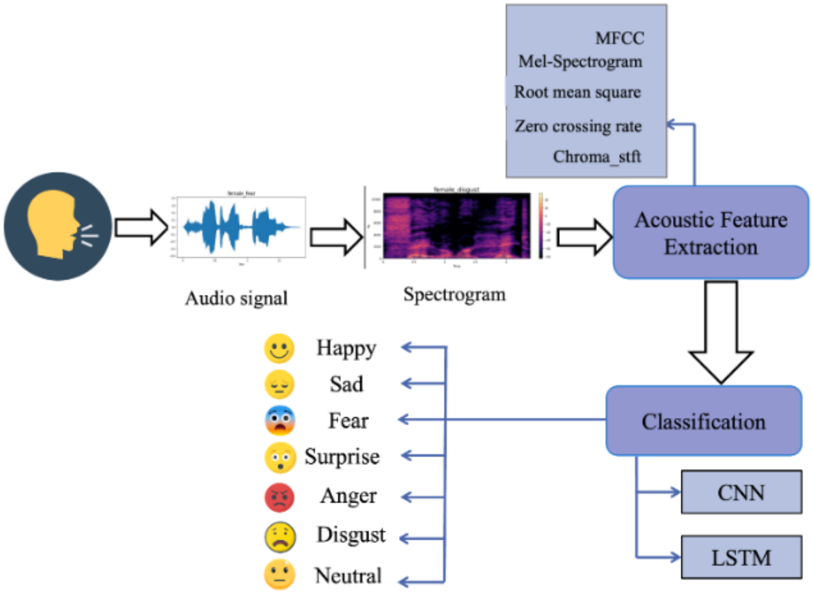

Figure 2: Working methods of Emotion Detection through Speech using Deep Learning.

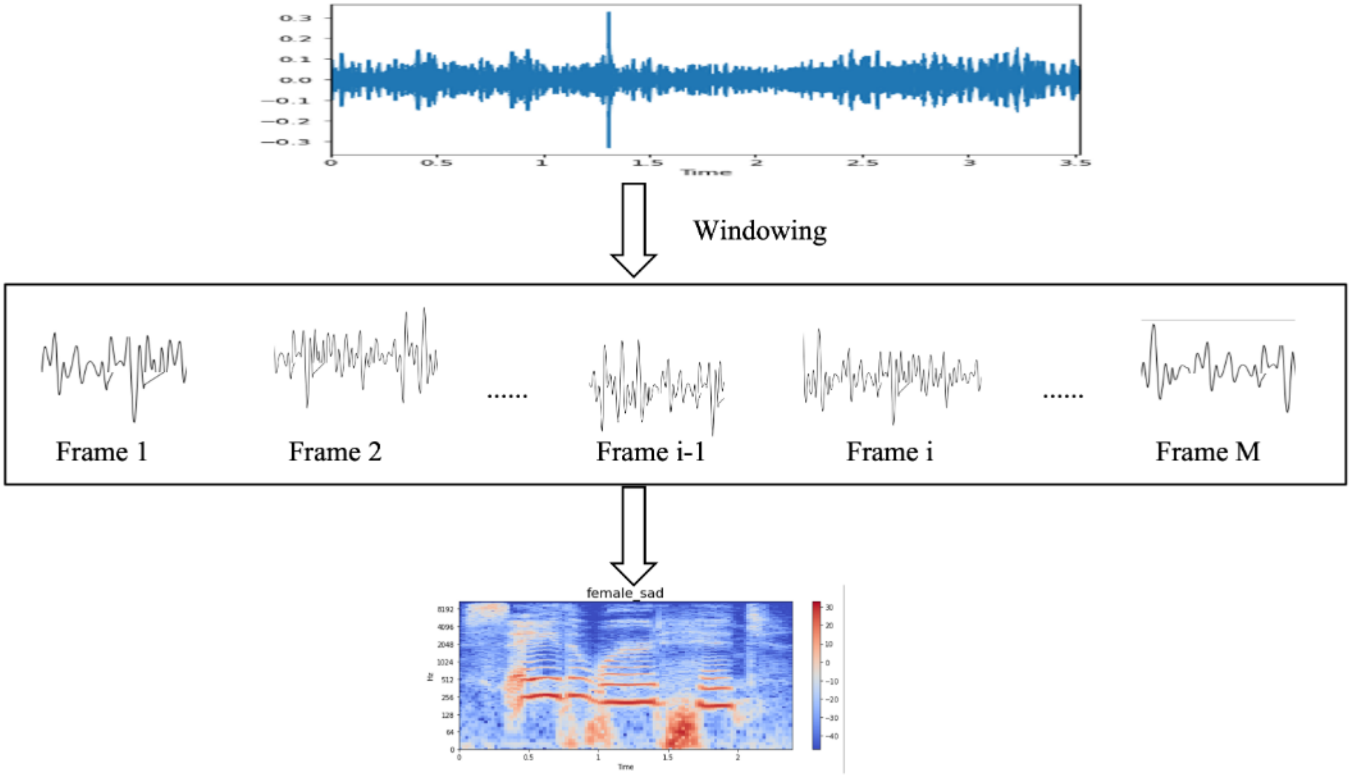

Figure 3: Audio signal analysis.

The observed differences in classifier performance suggest that the choice between CNN and LSTM depends on the specific requirements of the classification task. CNN excels in terms of overall accuracy and precision, making it a strong candidate when minimizing false positives and achieving high accuracy is crucial. On the other hand, LSTM, despite having lower overall accuracy, may be

Table 1: Accuracy, Precision and Recall Table for CNN

CNN	Classes	Accuracy	precision	Precision Average	Recall	Recall Average
7-class	Female_angry	98%	1.00	0.98	0.98	0.98
	Female_disgust		0.95		0.96	
	Female_fear		0.99		1.00	
	Female_happy		0.97		0.99	
	Female_neutral		1.00		1.00	
	Female_sad		0.98		1.00	
	Female_surprise		0.97		0.95	

Table 2: Accuracy, Precision and Recall Table for LSTM

LSTM	Classes	Accuracy	Precision	Precision Average	Recall	Recall Average
7-class	Female_angry	77%	0.77	0.78	0.59	0.77
	Female_disgust		0.78		0.77	
	Female_fear		0.77		0.79	
	Female_happy		0.68		0.64	
	Female_neutral		0.89		0.91	
	Female_sad		0.86		0.95	
	Female_surprise		0.66		0.74	

Table 3: CNN and LSTM comparison for accuracy, precision and recall

Classifier for 7-class	Accuracy	Precision							Recall						
		f-a	f-d	f-f	f-h	f-n	f-sa	f-s	f-a	f-d	f-f	f-h	f-n	f-sa	f-s
CNN	98%	0.99	0.95	1.00	0.98	1.00	0.99	0.94	0.98	0.98	0.98	0.93	1.00	0.99	0.96
LSTM	77%	0.77	0.78	0.77	0.68	0.89	0.86	0.66	0.59	0.77	0.79	0.64	0.91	0.95	0.74

suitable when dealing with classes that require balanced precision and recall. However, to gain a more comprehensive understanding, a detailed comparison across different classes is necessary to determine which model excelled in specific classification challenges and exhibited superior precision or recall, or both, for particular classes.

Finding high-quality datasets is a significant challenge when working on emotion detection through speech recognition. Getting more relatable datasets could increase the accuracy and efficiency in a more remarkable way of the research. Although the TESS dataset has an ample amount of audio files, all of the files contain a single word rather than a sequence of words. Emotional features extracted from a single word were rather limited. Having a sequence of words could better train our models and increase the emotion recognition process more efficiently and accurately simultaneously, making it more robust. However, a dataset with a sequence of words would have consumed more memory and working on them directly might not be feasible with our existing hardware and online tools such as Google Colab. So, we were heavily dependent on the TESS dataset. We also found that TESS only contains female voiced audio files. It limited our study to females only. No features in this work were extracted from male speech. Moreover, in the TESS dataset, all the data has been recorded in a controlled environment by voice artists. Mostly in real life, data will have more interference, and noises and voice pitch might have up or down more than usual. As a result, the model may not receive enough data to learn about the diverse kinds of features that are actually present in human speech. So, there can be bias in the models trained using this dataset.

Since the audio files were created in a controlled environment by voice actors and only included a single word, in-depth pre-processing couldn't be performed as the audio files were already refined. However, basic pre-processing has been performed. If the audio files contained a sequence of words instead of one word & created in a regular environment, a wide range of feature selection could have been extracted as needed.

4 CONCLUSION

Our study delved into the realm of Speech Emotion Recognition (SER) with a primary focus on leveraging Deep Convolutional Neural Networks (CNNs) to enhance the accuracy and precision of emotion prediction. We meticulously explored the intricate relationship between speech and emotion, recognizing the pivotal role that speech plays in conveying human emotional states. Throughout our research, we successfully harnessed advanced algorithms, with a specific emphasis on CNNs and Long Short-Term Memory networks (LSTMs), to develop models capable of accurately discerning the seven primary human emotions. Our rigorous analysis and empirical evaluation showcased the superior performance of the CNN model, yielding a noteworthy 21% improvement in average recognition accuracy. This achievement underscores the potential of deep learning techniques, particularly CNNs, in revolutionizing the field of SER. In conclusion, our study not only advances the field of SER but also lays the groundwork for future research endeavors. By continually refining and expanding upon the insights gained in this study, we can unlock new possibilities for the accurate recognition of human emotions in speech, with far-reaching implications for human-machine interactions and beyond.

Several promising avenues for further research emerge. Firstly, the refinement and optimization of deep learning models for SER, potentially through hybrid architectures combining CNNs with Long Short-Term Memory networks, offer opportunities for even greater accuracy. Additionally, future research in this domain can leverage robust datasets to enhance training, supported by advanced hardware capable of processing real-time speeches with non-native accents. While current algorithms demonstrate promise, exploring alternative approaches such as support vector machines could yield more accurate results, and the exploration of unsupervised methods like auto-encoders holds further potential for future investigations. Additionally, all scripts for replication are available online[1].

REFERENCES

[1] Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, and Benoit Champagne. 2022. LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6912–6916.

[2] Trevor R. Agus, Simon J. Thorpe, Clara Suied, and Daniel Pressnitzer. 2010. Characteristics of human voice processing. In *2010 IEEE International Symposium on Circuits and Systems (ISCAS)*. 509–512. https://doi.org/10.1109/ISCAS.2010.5537589

[3] Tursunov Anvarjon, Mustaqeem, and Soonil Kwon. 2020. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors* 20, 18 (2020), 5212.

[4] Rajdeep Chatterjee, Saptarshi Mazumdar, R Simon Sherratt, Rohit Halder, Tanmoy Maitra, and Debasis Giri. 2021. Real-time speech emotion analysis for smart home assistants. *IEEE Transactions on Consumer Electronics* 67, 1 (2021), 68–76.

[5] Li-Wei Chen and Alexander Rudnicky. 2023. Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[6] Bonaventure FP Dossou and Yeno KS Gbenou. 2021. FSER: Deep convolutional neural networks for speech emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3533–3538.

[7] K Dupuis and KP Fuller. 2010. Toronto emotional speech set (TESS) Collection.

[8] Kexin Feng and Theodora Chaspari. 2021. Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing* (2021).

[9] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. 2017. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5005–5009.

[10] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. 2017. Reconstruction-error-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2367–2371.

[11] Pengxu Jiang, Hongliang Fu, Huawei Tao, Peizhi Lei, and Li Zhao. 2019. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* 7 (2019), 90368–90377.

[12] Wei Jiang, Zheng Wang, Jesse S Jin, Xianfeng Han, and Chunguang Li. 2019. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors* 19, 12 (2019), 2730.

[13] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* 7 (2019), 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

[14] Palani Thanaraj Krishnan, Alex Noel Joseph Raj, and Vijayarajan Rajangam. 2021. Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition. *Complex & Intelligent Systems* 7 (2021), 1919–1934.

[15] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. 2019. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE access* 7 (2019), 125868–125881.

[16] Mustaqeem and Soonil Kwon. 2019. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* 20, 1 (2019), 183.

[17] Mustaqeem and Soonil Kwon. 2020. CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. *Mathematics* 8, 12 (2020), 2133.

[18] YC Pan, MX Xu, LQ Liu, and PF Jia. 2006. Emotion-detecting based model selection for emotional speech recognition. In *The Proceedings of the Multiconference on" Computational Engineering in Systems Applications"*, Vol. 2. IEEE, 2169–2172.

[19] Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Laureano Moro-Velazquez, and Najim Dehak. 2021. Beyond isolated utterances: Conversational emotion recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 39–46.

[20] K Sarker and KR Alam. 2014. Emotion recognition from human speech: Emphasizing on relevant feature selection and majority voting technique. In *Proceedings of the 3rd International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh*. 23–24.

[21] Mandeep Singh and Yuan Fang. 2020. Emotion recognition in audio and video using deep neural networks. *arXiv preprint arXiv:2006.08129* (2020).

[22] Sattaya Singkul, Thakorn Chatchaisathaporn, Boontawee Suntisrivaraporn, and Kuntpong Woraratpanya. 2020. Deep residual local feature learning for speech emotion recognition. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part I 27*. Springer, 241–252.

[23] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. 2018. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology* 21 (2018), 93–120.

[24] Enkhtogtokh Togootogtokh and Christian Klasen. 2021. DeepEMO: deep learning for speech emotion recognition. *arXiv preprint arXiv:2109.04081* (2021).

[25] Loan Trinh Van, Thuy Dao Thi Le, Thanh Le Xuan, and Eric Castelli. 2022. Emotional speech recognition using deep neural networks. *Sensors* 22, 4 (2022), 1414.

[26] Chunxi Wang, Maoshen Jia, Yanyan Zhang, and Lu Li. 2023. Multi-speaker Speech Separation under Reverberation Conditions Using Conv-Tasnet. *Journal of Advances in Information Technology* 14, 4 (2023).

[27] Sitong Zhou and Homayoon Beigi. 2020. A transfer learning method for speech emotion recognition from automatic speech recognition. *arXiv preprint arXiv:2008.02863* (2020).

[1] Source code: https://bit.ly/HumanSpeechEmotionRecogniztion