EXPLORING EMOTIONAL SIGNALS: A STUDY ON AUDIO-BASED EMOTION CLASSIFICATION

Joylin Priya Pinto [1],Balasubramani R [2], Jason Elroy Martis [2]

[1] Department of CSE, NITTE, Bangalore, India
[2] Department of ISE,NITTE, Bangalore, India

ABSTRACT

Intersection of artificial intelligence and audio processing has witnessed a surge of interest in recent years, with various applications emerging, ranging from speech synthesis to music generation. This paper presents a comprehensive overview of various approaches to audio synthesis, utilising neural networks which classify the different emotions expressed in speech. It surveys a range of key areas such as speech classification, enhancement along with speaker dependent and speaker independent emotion analysis. By using an extensive dataset of emotional speech samples, neural network architecture effectively learns the sensitive nuances of emotional expression embedded within audio. The study delves into the utilization of various deep learning models such as long short-term memory (LSTM) networks, Auto-Encoders and convolutional neural networks (CNNs) which are capable of successfully capturing the nuanced emotional features within the input data, enabling the synthesis of emotionally enriched audio outputs. In this study, the model's capability to classify audio segments enriched with specific emotional content, such as happiness, sadness, anger, excitement etc is evidenced by the effectiveness of various approaches.

KEYWORDS

Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), Audio synthesis, deep learning, Mel-Frequency Cepstral Coefficients (MFCCs)

1. INTRODUCTION

In recent years, there has been a surge in research focused on advancing speech processing technologies, particularly in the domain of emotional voice classification and audio synthesis. Emotion classification from audio signals involves the extraction of meaningful features from speech data and the utilisation of various algorithms to analyse the underlying emotional states conveyed within the audio.

It is recognized that people's emotional states trigger physiological changes in their body. Variables such as pulse, blood pressure, facial expressions, body movements, brain waves and acoustic properties fluctuate based on these emotions. To detect changes of these variables portable medical devices are required. But facial expressions and voice signals can be observed directly without any additional equipment. Consequently, much of the research in this field concentrates on automating the recognition of emotions using visual and auditory cues. The current state-of-the-art in Speech Emotion Recognition (SER) is structured around a classification framework employing conventional machine learning techniques. This involves employing various standard algorithms to construct robust models for SER. The traditional SER pipeline involves

DOI:10.5121/cseij.2025.15132

several stages, starting with the extraction of frame-wise Low-Level Descriptors (LLDs) like fundamental frequency (F0), Mel-Frequency Cepstral Coefficients (MFCCs), and Teager Energy Operator (TEO). Subsequently, these descriptors undergo feature selection, aggregation via statistical functional, and dimensionality reduction. However, this extensive pre-processing not only constrains the accuracy of the machine learning model but also impedes its adaptability to real-world scenarios, particularly in real-time settings.

There's now a growing imperative to transition away from this standard pipeline towards an end-to-end machine learning approach. Such a methodology would possess the capability to autonomously discern low-level features from raw data and subsequently derive high-level features from these low-level ones in a hierarchical manner. This shift aims to mitigate the reliance of current models on predefined features and numerous pre-processing steps. Deep Learning emerges as a promising avenue to fulfil this objective [7].

Numerous inventive strategies harnessing deep learning methodologies and generative models have emerged to tackle diverse challenges within this domain. A deep neural network has become capable of learning the mapping between the spectral features of a speaker's voice and the desired emotional state. A trained neural network thereby has the ability to accurately convert between different emotional states.

Researchers have employed various techniques to identify the challenges of sentiment classification from audio signals. These approaches range from traditional feature extraction methods, such as Mel-frequency cepstral coefficients (MFCC) and prosodic features, to more advanced deep learning architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Additionally, the integration of multidimensional data, such as facial expressions and physiological signals, have enhanced the accuracy of emotion classification systems.

Understanding and categorising emotions expressed through speech is crucial for various applications, including sentiment analysis in customer service interactions, mental health monitoring, and improving human-computer interfaces. By accurately identifying emotions from audio signals, it became possible to develop systems that can adapt their responses or behaviours based on the user's emotional state, leading to more personalised and effective interactions.

2. LITERTURE SURVEY

The predominant architectures for speech-emotion recognition employing neural networks primarily consist of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) with Long-Short Term Memory (LSTM), or a combination of both [3, 4, 5, 8]. The integration of CNNs and RNNs proves effective in uncovering crucial patterns within audio files during feature extraction and entity classification [3, 6]. A primary objective of emotional speech recognition is to identify the significant features required to train a model addressed through various methodologies. Trigeorgis et al. [3] adopted a unique approach by utilising raw audio file as input for the model. They employed CNNs for pre-processing the audio samples to reduce noise and to concentrate on specific portions of the audio file. Notably, their architecture performed well over the existing models at that time.

Lim et al. [6] transformed the raw audio file taken from the EMO-DB dataset into a two-dimensional representation using the short-time Fourier transform. Subsequently, the resulting signal was fed into the models, with the initial layer being a CNN. Among their proposed models, the time distributed CNN demonstrated the most promising outcomes. Badshah et al. [5] employed a similar approach, utilising spectrograms-visual representations of audio samples generated from raw audio sample sourced from EMO-DB dataset. These spectrograms were then fed into a CNN model. While the results demonstrated satisfactory performance with an accuracy of 52% [2, 9] for all emotions.

In alignment with Badshah et al. [5], Zhao et al. [10] employed log-mel spectrograms as an input for a two dimensional CNN-LSTM network. Their findings showcased remarkable achievements, achieving an accuracy of 95.33% for speaker-dependent classification and 95.89% for speaker-independent classification of samples from the EMO-DB dataset. As far as current knowledge goes, this represents the most advanced state-of-the-art solution for addressing the challenge of speech emotion recognition within this dataset.

Chatziagapi et al. [11] introduced a novel approach to improve the performance of speech emotion recognition models by utilizing data augmentation based on Generative Adversarial Networks (GANs). They employed GANs to generate synthetic spectrograms, enhancing the dataset. Applying this technique to samples from the IEMOCAP database, authors observed a significant 10% relative performance gain.

Demircan and Kahramanli [13] introduced an innovative approach by using Mel Frequency Cepstral coefficients (MFCCs) derived from 520 samples of EMO-DB dataset. Their architecture incorporated dimensionality reduction through fuzzy C-means clustering. Subsequently, authors used multiple classifiers, including Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-Nearest Neighbor(KNN). Significantly, these classifiers achieved classification accuracies of 90%, 92.86% and 92.86%, respectively, on the test set.

Mustafa A. Qamhan et. al.[12] studied a deep learning-based speech emotion recognition approach that utilises the King Saud University Emotions' Arabic dataset. The primary system of the approach was designing a combination of convolutional neural network (CNN) and long short-term memory (LSTM), resulting in a convolutional recurrent neural network (CRNN). The use of linearly spaced spectrograms as inputs to the CRNN system is also explored. The performance of the CRNN system is compared with human perceptual evaluation, which served as the baseline system for the experiment. They analysed that CRNN system achieved high accuracies of 84.55% and 77.51% for file and segment levels which are closely comparable to human emotion perception scores.

Yoon et al. [15] adopted a multimodal approach for emotional speech recognition, leveraging both audio and text data from the IEMOCAP dataset. They utilised MFCCs and text tokens as input features for the model, resulting in an accuracy of 71.8%.

Unlike previous approaches, Huang et al. [16] developed a model which is hybrid in nature merging deep learning and traditional machine learning methods to classify the 800 entries within the EMO-DB dataset. The model is termed as semi-CNN, integrates CNN input layers reserved for learning affect-salient features, along with an SVM classifier in the final layer for categorization. Similar to earlier studies, the authors utilized spectrograms as the input for their semi-CNN. The experimental outcomes showcased an accuracy of 88.3% for speaker-dependent classification and 85.2% for speaker-independent classification on the test set.

Additionally, Wu et al. [14] employed traditional machine learning methods exclusively to classify samples from the EMO-DB dataset, containing 800 entries. They introduced a novel set of sound features known as modulation spectral features (MSFs). By integrating MSFs with prosodic features and applying them as input to a multi-class linear discriminant analysis (LDA) classifier, the author attained a notable 85.8% accuracy for speaker-independent classification on the test set.
Wang et al. [17] introduced a novel set of sound features named Fourier Parameter (FP) features,

derived through Fourier analysis. They focused solely on 6 out of the 7 emotion classes from the EMO-DB dataset, excluding the "disgust" class. Authors extracted FP and MFCC features from the dataset and employed them as input for an SVM classifier, achieving an average accuracy of 73.3%.

Shegokar et al.[18] employed SVM for classifying male speech samples from the RAVDESS dataset. They utilised Continuous Wavelet Transform (CWT) for feature selection and fed the chosen features into various SVM classifiers. The highest accuracy achieved was 60.1% using Quadratic SVM with a 5-fold cross-validation technique.

Zhang et al. [19] introduced a different strategy known as multi-task learning, combining features extracted from songs and speech samples from the RAVDESS dataset. They suggested that classifiers utilising the relationship between songs and speech can achieve superior accuracy. Focusing on 4 out of 8 emotion classes (angry, happy, neutral, and sad), they attained a 57.14% accuracy using the group multi-task feature selection (GMTFS) model. Similarly, Zeng et al. [20] applied the concept of multi-task learning using a deep neural network (DNN). They utilised spectrograms generated from songs and speech utterances from the RAVDESS dataset as input to their multi-task gated Residual Networks (GResNets), achieving an accuracy of 65.97%. They noted that their model outperformed task-specific ones trained separately for songs and speech.

In contrast, Popova et al. [21] opted for a fine-tuned DNN approach to classify mel spectrograms obtained from speech samples of the RAVDESS dataset. Utilising Convolutional Neural Network VGG-16 as a classifier, they achieved an accuracy of 71%.

Mustaqeem[28] introduced a novel approach to enhance speech emotion recognition (SER) accuracy while simultaneously reducing computational complexity. Method revolves around the development of an artificial intelligence-assisted deep stride convolutional neural network (DSCNN) architecture, employing the plain nets strategy. By focusing on learning salient and discriminative features from spectrograms of speech signals, authors aimed to optimise performance through a series of carefully crafted steps. Rather than conventional pooling layers, this architecture utilised specialised strides within the convolutional layers to down-sample feature maps, preserving essential local hidden patterns. In the subsequent fully connected layers, our model synthesised global discriminative features crucial for accurate emotion classification, culminating in the application of a SoftMax classifier. To validate their approach, authors evaluated its performance on two prominent datasets: IEMOCAP and RAVDESS. Results demonstrated significant improvements in accuracy, by achieving 81.75% on an IEMOCAP and 79.5% on RAVDESS datasets respectively. Impressively, these advancements are accompanied by a notable reduction in model size, slashing it by 34.5 MB comparatively Alex Net[25], Vgg16[30], ResNet50[31].

3. DATASETS USED

Researchers used three prominent audio datasets, namely RAVDESS[18,19,20,21], EMO-DB[5,10,13,14,17], and IEMOCAP [11,15,28] which are widely used in the domain of emotion recognition.

RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset comprises audio and visual recordings featuring 12 male and 12 female actors articulating English sentences across eight distinct emotional expressions. Speech samples are utilised from the database, encompassing eight emotion classes: sad, happy, angry, calm, fearful, surprised, neutral, and disgust.

EMO-DB

EMO-DB dataset consists of 535 audio utterances in German, categorised into 7 distinct emotion classes: anger, sadness, fear/anxiety, neutral, happiness, disgust, and boredom.

IEMOCAP

This dataset contains audio, video, and face motion capture samples collected from five pairs of male and female actors. These samples are distributed across five sessions, each featuring data from a specific pair of actors. The audio files in the dataset are categorised into ten emotion classes: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other.[28]

4. Features Extraction Techniques [2]

Audio or speech feature extraction is a critical process in speech recognition. Numerous studies have explored the identification of emotions from auditory data. This involves extracting features from a collection of emotional speech, followed by the classification of emotions based on these features. The accuracy of emotion classification hinges significantly on the effectiveness of feature extraction. The paper explores various feature extraction methods that are employed to extract relevant information from audio signals, enabling machines to understand and analyse audio data.

A. Time-domain Features

Time-domain feature extraction is a fundamental technique in audio emotion classification, involving the analysis of raw audio signals directly in the time domain. Several key features can be extracted from the time-domain waveform to capture various aspects of the signal related to emotion such as Root Mean Square (RMS) Energy, Zero-Crossing Rate (ZCR), Temporal Centroid, and Signal Variance.

B. Frequency-domain Features

Frequency-domain features analyse the spectral characteristics of audio signals. Fourier Transform is often used to convert audio signals from the time domain to the frequency domain. Spectral features include spectral centroid, spectral bandwidth, and spectral roll-off which helps in identifying the vocal characteristics.

C. Mel-frequency Cepstral Coefficients (MFCC)

MFCCs are widely used in speech and audio processing. They represent the short-term power spectrum of audio signals after passing through a Mel filter bank. MFCCs capture both spectral and temporal information, making them effective for speech and music analysis. They transform fundamental features such as pitch, power in decibels, and phase characteristics into a set of 12 MFCC features. [22] The effectiveness of the Least Squared SVM Bound (LSBOUND) algorithm underscores the superiority of MFCC features over linear prediction coefficient (LPC) features in minimising Cross Validation (CV) errors for multi-class classifiers.

D. Chroma Features

Chroma features focus on the pitch content of audio signals. Pitch is recognized as the fundamental frequency, a highly responsive element tied to the auditory perception.[23] It denotes the periodicity of a wave pulse produced by air compression through the glottis originating from the lungs. Chroma features represent the distribution of pitch classes, disregarding octave variations. Chroma features are particularly useful for music-related tasks such as chord recognition and genre classification.

E. Rhythm Features

Rhythm features capture temporal patterns and rhythmic structures in audio signals. Examples include tempo, beat histogram, and rhythmic complexity measures. These features are essential for tasks such as music genre classification and tempo estimation.

F. Wavelet Transform

Wavelet Transform decomposes audio signals into different frequency components at multiple resolutions. It provides a time-frequency representation of audio signals, allowing for efficient feature extraction.

5. Analysis of Previous Work on Different Datasets

To achieve reliable and robust performance in Speech Emotion Recognition (SER) systems, the selection of appropriate datasets plays a crucial role. In this study, we investigated and compared the performance of Speech Emotion Recognition models trained on two widely used datasets: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Berlin Emotional Speech Database (EMO-Db).

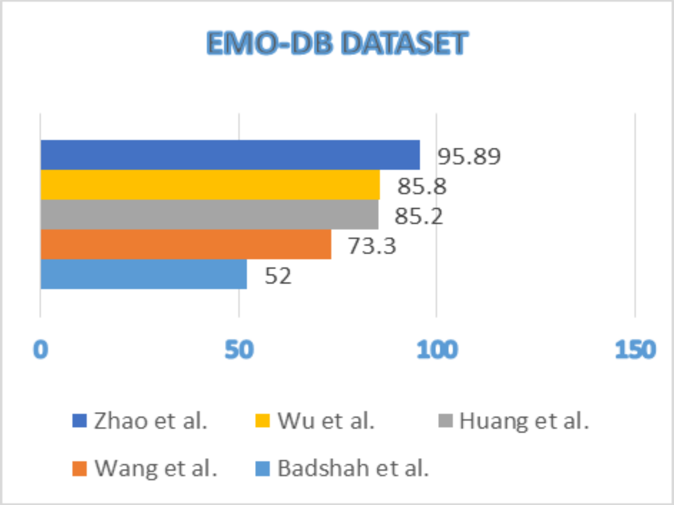


FIGURE 1. EMO-DB dataset Analysis

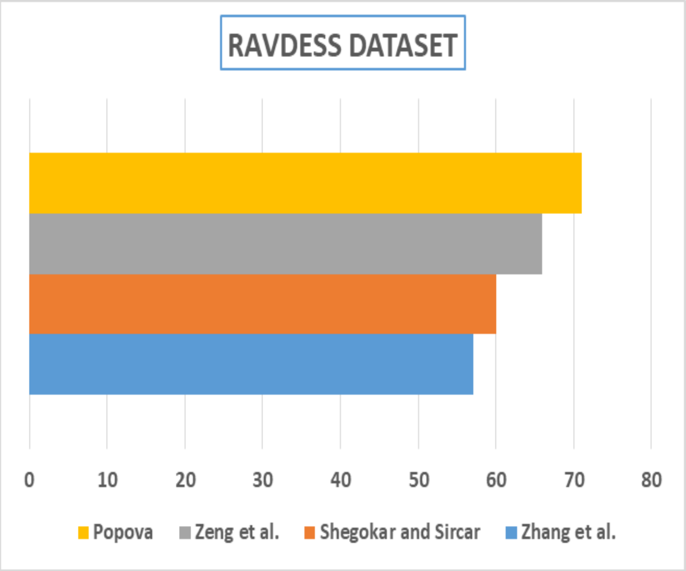
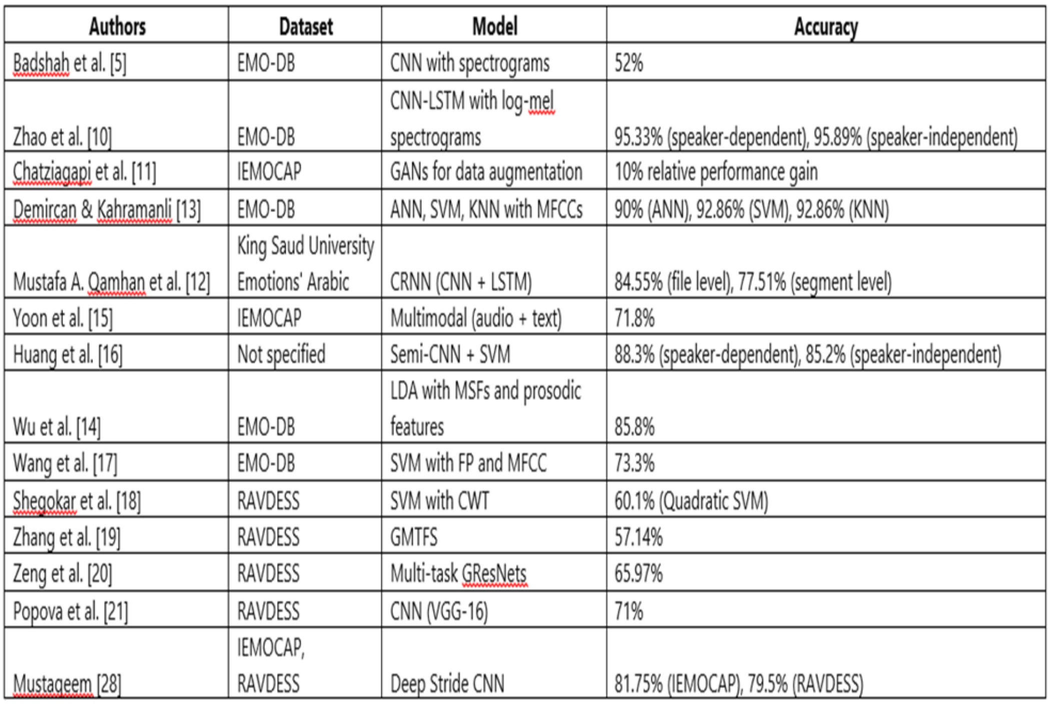


FIGURE 2. RAVDESS dataset Analysis

The strategies employed and the size of the dataset utilized determine how accurate the machine learning and deep learning models are. Table 1. Captures the key aspects such as datasets used, models implemented, and the achieved accuracies for the discussed studies in speech-emotion recognition. Here is the comparison table based on the provided literature review:

Table 1. Deep Learning and Machine Learning Model Comparison Table



Authors	Dataset	Model	Accuracy
Badshah et al. [5]	EMO-DB	CNN with spectrograms	52%
Zhao et al. [10]	EMO-DB	CNN-LSTM with log-mel spectrograms	95.33% (speaker-dependent), 95.89% (speaker-independent)
Chatziagapi et al. [11]	IEMOCAP	GANs for data augmentation	10% relative performance gain
Demircan & Kahramanli [13]	EMO-DB	ANN, SVM, KNN with MFCCs	90% (ANN), 92.86% (SVM), 92.86% (KNN)
Mustafa A. Qamhan et al. [12]	King Saud University Emotions' Arabic	CRNN (CNN + LSTM)	84.55% (file level), 77.51% (segment level)
Yoon et al. [15]	IEMOCAP	Multimodal (audio + text)	71.8%
Huang et al. [16]	Not specified	Semi-CNN + SVM	88.3% (speaker-dependent), 85.2% (speaker-independent)
Wu et al. [14]	EMO-DB	LDA with MSFs and prosodic features	85.8%
Wang et al. [17]	EMO-DB	SVM with FP and MFCC	73.3%
Shegokar et al. [18]	RAVDESS	SVM with CWT	60.1% (Quadratic SVM)
Zhang et al. [19]	RAVDESS	GMTFS	57.14%
Zeng et al. [20]	RAVDESS	Multi-task GResNets	65.97%
Popova et al. [21]	RAVDESS	CNN (VGG-16)	71%
Mustaqeem [28]	IEMOCAP, RAVDESS	Deep Stride CNN	81.75% (IEMOCAP), 79.5% (RAVDESS)

6. Discussion and Conclusion

Conducting a survey on Speech Emotion Recognition (SER) using deep learning models offers valuable insights into the state-of-the-art techniques, challenges, and potential future directions in this field. Through the exploration of various deep learning architectures, datasets and feature extraction techniques, we have gained valuable insights into the advancements and challenges within this domain.

Deep learning models, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants such as Long Short-Term Memory (LSTM) networks, have demonstrated remarkable success in automatically extracting discriminative features from speech signals for emotion recognition tasks. These models have shown promising results across different datasets and modalities, showcasing their potential for real-world applications. However, despite the progress made, several challenges remain to be addressed. These include the need for larger and more diverse datasets to capture the nuances of human emotion, the development of robust feature extraction methods that can effectively handle noisy and variable input data. In conclusion, Speech Emotion Recognition using deep learning models holds great promise for understanding and responding to human emotions in various contexts.

References

[1] Hadhami Aouani and Yassine Ben Ayed, Speech Emotion Recognition with deep learning, 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science 176 (2020) 251–260, 1877-0509 © 2020 published by Elsevier B.V, 10.1016/j.procs.2020.08.027

[2] M.N.Hasrul, M.Hariharan, Sazali Yaacob, Human Affective (Emotion) Behaviour Analysis using Speech Signals: A Review, 2012 International Conference on Biomedical Engineering (ICoBE),27-28 February 2012,Penang, 978-1-4577-1991-2/12/$26.00 ©2011 IEEE

[3] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: 2016 IEEE International Conference on Acoustics, Speech and Signa l Processing (ICASSP), IEEE, 2016, pp. 5200–5204

[4] W. Lim, D. Jang, T. Lee, Speech emotion recognition using convolutional and recurrent neural networks, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–4.

[5] A.M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 International Conference on Platform Technology and Service (PlatCon), IEEE, 2017, pp. 1–5.

[6] W. Lim, D. Jang, T. Lee, Speech emotion recognition using convolutional and recurrent neural networks, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–4

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 05/28/print 2015.

[8] Y. Niu, D. Zou, Y. Niu, Z. He, H. Tan, Improvement on speech emotion recognition based on deep convolutional neural networks, Proceedings of the 2018 International Conference on Computing and Artificial Intelligence (2018) 13–18.

[9] N. Weibkirchen, R. Bock, A. Wendemuth, Recognition of emotional speech with convolutional neural networks by means of spectral estimates, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2017, pp. 50–55.

[10] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomed. Signal Process. Control 47 (2019) 312–323.

[11] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, Data augmentation using gans for speech emotion recognition, Proc. Interspeech 2019 (2019) 171–175.

[12] Mustafa A. Qamhan, Ali H. Meftah, Sid-Ahmed Selouani, Yousef A. Alotaibi, Mohammed Zakariah, Yasser Mohammad Seddiq, Speech Emotion Recognition using Convolutional Recurrent Neural Networks and Spectrograms, IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 978-1-7281-5442-8/20 ©2020 IEEE

[13] S. Demircan, H. Kahramanli, Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech, Neural Comput. Appl. 29 (2018) 59–66.

[14] S. Wu, T.H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Commun. 53 (2011) 768–785.

[15] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118

[16] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn, Proceedings of the 22nd ACM International Conference on Multimedia (2014) 801–804

[17] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using fourier parameters, IEEE Trans. Affect. Comput. 6 (2015) 69–75.

[18] P. Shegokar, P. Sircar, Continuous wavelet transform based speech emotion recognition, in: 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), IEEE, 2016, pp. 1–8.

[19] B. Zhang, E.M. Provost, G. Essi, Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5805–5809.

[20] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimed. Tools Appl. (2017) 1–18.

[21] A.S. Popova, A.G. Rassadin, A.A. Ponomarenko, Emotion recognition in sound, in: International Conference on Neuroinformatics, Springer, 2017, pp. 117–124

[22] Coskucan Buyukyildiz, İsmail Saritas, Ali Yasar, Classification of Emotion with Audio Analysis, Yuzuncu Yil University Journal of the Institute of Natural & Applied Sciences, Volume 28, Issue 2 (August), 467-481, 2023

[23] Dias Issa, M. Fatih Demirci, Adnan Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control 59 (2020) doi: https://doi.org/10.1016/j.bspc.2020.101894 1746-8094/© 2020

[24] S.R. Livingstone, F.A. Russo, 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 13 (2018) e0196391

[25] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1097–1105.

[26] J. Abeber, ‘‘A review of deep learning based methods for acoustic scene classification,’’ Appl. Sci., vol. 10, no. 6, p. 2020, Mar. 2020, doi: 10.3390/app1006202045. A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, ‘‘A survey of the recent architectures of deep convolutional neural networks,’’ Artif. Intell. Rev., vol. 53, no. 8, pp. 5455–5516, Apr. 2020, doi: 10.1007/s10462-020-09825-6.

[27] N. Weibkirchen, R. Bock, A. Wendemuth, Recognition of emotional speech with convolutional neural networks by means of spectral estimates, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2017, pp. 50–55.

[28] Mustaqeem and Soonil Kwon, A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition, Sensors 2020, Open Access Journal, 20, 183; doi:10.3390/s20010183

[29] Dias, M.; Abad, A.; Trancoso, I. Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2057–2061.

[30] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

[31] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.

[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. Lang. Resour. Eval. 2008, 42, 335. [CrossRef]

[33] 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. [CrossRef]

[34] E. Lakomkin, C. Weber, S. Magg, S. Wermter, Reusing Neural Speech Representations for Auditory Emotion Recognition, 2018 arXiv preprint arXiv:1803.11508.

[35] M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process. Lett. 25 (2018) 1440–1444.