IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Review on Classification of Speech Signal Using Traditional and Deep Learning Techniques

Chisnah C.J
PG Scholar
Department of Electronics and Communication Engineering
Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

Gayathri R
Assistant Professor (Sr.Gr.)
Department of Electronics and Communication Engineering
Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

Abstract- Speech classification could be a broad space of analysis that has gained abundant attention in recent years. Feeling recognition from speech signals is a vital however difficult part of Human-Computer Interaction (HCI). One amongst the largest variations between a machine and a person knows the emotions of others and behaving consequently. Within the literature of speech emotion recognition (SER), several techniques are utilized to extract emotions from signals, as well as several well-established speech analysis and classification techniques. This paper presents an overview of traditional techniques, Deep Learning techniques and discusses some recent literature where these methods are utilized for speech-based emotion recognition. The review covers databases used, emotions extracted, and contributions made toward speech emotion.

Keywords – Speech classification, Emotions, Deep Learning, speech emotion recognition, traditional techniques,

I. INTRODUCTION

Human beings categorize their feelings, opinions, views, and notions orally through speech. For humans, it's straightforward to grasp what someone is saying. Humans will comprehend each word, each character therefore simply while not a sweat that is sort of stunning. Our brain may be a powerhouse that might method each word that someone is an expression with instance accuracy. Even once the accent of sure folks is tough to grasp, humans tend to simply have to be compelled to understand the few words and also the context within which the words are being spoken to urge the concept of what someone can be expressing. However, once it involves machines, this can be wherever things get a bit sophisticated. Creating a machine perceive the language and creating a way of what has been spoken is extremely abundant tough. Even though it's potential to exhaust code the principles and let the machine apprehend all the various words and in what context they'll be spoken however that's as so much as go along with traditional programming. Such machines are solely restricted to the information that will impact them.

Speech or sound is nothing however a form of a wave. Sound is that the vibrations that travel through the air or another medium and might be detected after they reach a person's ear. So, after the person speaks one thing, the molecules on the point of the mouth start vibrating and these vibrating molecules run into the neighbouring molecules that ultimately reach the ear of another person. And this means the opposite person will hear what has been spoken. The information in the speech signal is portrayed by short term amplitude spectrum of the speech waveform. This permits us to extract features that supported the short-term amplitude spectrum from speech (phonemes). However, it tends to still cannot achieve natural interaction between humans and machines because of the present machines cannot sufficiently perceive the emotional standing of humans. This discrepancy has a crystal rectifier to a replacement analysis field: speech emotion recognition.

Speech classification may be a means that for automatic classification of audio signals. Speech classification aims to try the analysis of audio signals by providing vital information concerning the content of speech signals. Speech signals should be classified since it's supported statistical techniques and thus, it's necessary to supply appropriate training material for the task. Deep learning permits computational models that are composed of multiple processing layers to learn representations of information with multiple levels of abstraction. Deep learning algorithms are used for increasing complexity and abstraction. In the existing system of speech classification, data sets were considered for only 4 to 5 emotions. Traditional Machine learning Techniques used are Bayes Classifier and K-Nearest Neighbor Classifier. Deep Learning Algorithms used are DSCNN, CNN, MSCNN+SPU+Attention Mechanism, BLSTM, CNN+BLSTM, CNN+LSTM+Attention Mechanism, DNN, and CRNN. Speech segregation and music separation had been done using the Deep Belief network.

Deep Learning has been thought of as a rising analysis field in machine learning and has gained further attention in recent years. Deep Learning techniques for SER have several edges over traditional strategies, at the side of their capability to discover the advanced structure and options whereas not the need for manual feature extraction and tuning; tendency toward extraction of low-level options from the given information, and talent to deal with un-labeled data. Deep Learning may well be a set of Machine Learning that achieves body politic edges by learning to represent the planet as a nested hierarchy of ideas, with every conception made public about easier ideas, and additional abstract representations computed in terms of less abstract ones. In the associate elaborate approach, a deep learning technique learns categories incrementally through its hidden layer style, method low-level categories like letters initial then very little or no higher-level categories like words then higher-level categories like sentences. Every neuron or node at intervals the network represents one facet of the complete and therefore the whole they supply a full illustration of the image. Every node or hidden layer is given a weight that represents the strength of its relationship with the output and since the model develops the weights area unit adjusted.

In traditional Machine learning techniques, most of the applied options ought to be identified by a domain expert to chop back the complexes of the data and make patterns a lot visible to learning algorithms to work. The need for domain expertise and hard-core feature extraction is eliminated.

Deep learning strategies area units comprised of assorted nonlinear elements that perform computation on a parallel basis. However, these strategies ought to be structured with deeper layers of design to beat the restrictions of different techniques. Deep learning techniques like Deep Boltzmann Machine (DBM), recurrent Neural Network (RNN), recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN), and auto Encoder (AE) are some of the fundamental deep learning techniques used for SER, that considerably improves the general performance of the designed system.

II. Summary of Literature on Deep Learning Techniques for Ser

Several existing methodologies associated with the analysis are surveyed which can be delineated below.

An assortment of techniques has been given within the space of Speech feeling Recognition (SER), wherever the most focus is to acknowledge the silent discriminant and helpful options of speech signals. These options bear the method of classification to acknowledge the feeling of a speaker. In recent times, deep learning techniques have emerged as a breakthrough in speech feeling recognition to notice and classify emotions. Modified a recently developed completely different specification of convolutional neural networks, i.e., Deep Stride Convolutional Neural Networks (DSCNN) [6], by taking a smaller range of convolutional layers to extend the machine speed while still maintaining accuracy. Besides, it tends to train the state-of-art model of CNN and planned DSCNN on spectrograms generated from the SAVEE speech feeling dataset. For the analysis method, four emotions angry, happy, neutral, and sad, were thought about. The prediction accuracy of 87.8% for the planned design DSCNN, outperforms CNN with 79.4% accuracy results were showed.

In recent years, increasing attention is given to the analysis of the emotions present in speech. Numerous systems area unit developed attending to sight the emotions within the speaker's statements. One among the most important variations between a machine and somebody is, knowing the emotions of others and behaving consequently. Researchers square measure acting on bridging this gap by recognizing emotions in speech or voice. A deep learning-based technique for speech feeling recognition (SER) is proposed. The SER system is predicated on numerous techniques that use distinguished modules for feeling recognition. The model differentiates emotions like neutral state, happiness, sadness, anger, surprise, etc. The performance of the arrangement is predicated on options extracted and generated models. The options used during this embody energy, pitch, chromagram, Mel-frequency spectrum coefficients (MFCC), and Gammatone frequency spectrum coefficients (GFCC). The emotions area unit is classified employing a two-dimensional Convolutional Neural Network (CNN). The classification model achieved an Associate in Nursing overall accuracy of 92.59% on the check information that is relatively higher than the previous algorithmic program. In the future, the intention is to extend the system performance and sight a lot of emotions.

The advancements within the field of deep learning and feeling recognition are increasing within the recent past. The work presents a model framework that understands the feeling pictured on the face and from the voice. The first goal of this work remains to boost human-computer cooperation. Hood frontal face pictures and numerous voice cuts square measure provided by the model system. 90 samples of the Amrita emote database (ADB) were used for testing and 25838 samples were used for training. The speech information consists of four completely different datasets, with a complete of 20,000 examples. 3/4 of this info is employed to organize and 1/4 of the data used is for testing. Intermittent neural networks (RNNs) and ancient neural networks (CNNs) square measure nervous system-based come to that use speech and image management to regulate emotions: pleasure, sadness, anger, hatred, surprise, and fear [7].

Robust automatic speech emotional-speech recognition architectures supported hybrid convolutional neural networks (CNN) and feedforward deep neural networks are planned and named during this paper as BFN, CNA, and HBN. Bag-of-Audio-word (BoAW) and feedforward deep neural network combined to form BFN, CNA supported CNN, finally, HBN is a hybrid design between BFN and CNA. Overall accuracy is achieved by investment Mel-frequency cepstral constant options and bag-of-acoustic-words to feed the network, leading to promising classification performance. The three projected models area unit trained on eight emotional categories from the Ryerson Audio-Visual information of Emotional Speech and Song audio (RAVDESS) dataset. Our planned models achieved overall exactitude between 81.5% and 85.5% and overall accuracy between 80.6% and 84.5%, therefore outperforming progressive models victimization a similar dataset.

Emotion recognition from speech may be a difficult task. Recent advances in deep learning have crystal rectifier bi-directional repeated neural network (Bi-RNN) and a spotlight mechanism as a customary technique for speech emotion recognition, extracting and attending multi-modal options - audio and text, and so fusing them for downstream feeling classification tasks. The planned framework victimization multi-scale convolutional layers (MSCNN) to get each audio and text hidden representation. Then, an applied mathematics pooling unit (SPU) is employed to additional extract the options in every modality. Besides, the Associate in Nursing attention module is designed on high of the MSCNN-SPU (audio) and MSCNN (text) to additional improve the performance. Intensive experiments show that the planned model outperforms previous progressive strategies on the IEMOCAP dataset with four emotion classes (i.e., angry, happy, sad, and neutral) in each weighted accuracy (WA) and unweighted accuracy (UA), with Associate in the Nursing improvement of 5.0% and 5.2% severally below the ASR set.

Non-linguistic speech cues aid the expression of assorted emotions in human communication. Demonstrated the appliance of deep long remembering (LSTM) continual neural networks for frame-wise detection and classification of laughter and filler vocalizations in speech information. Further, proposed a unique approach to perform classification by incorporating cluster-info as an extra feature wherever in the clusters within the dataset square measure extracted via a k-means bunch rule. In-depth simulation results demonstrate that the planned approach achieves vital improvement over the standard LSTM-based classification strategies. Lastly, for classification of the temporally related to speech information thought-about during this work, a comparison with widespread machine learning-based techniques validates the prevalence of the planned LSTM-based theme [8].

A model trained on a traditional speaking rate is best able to acknowledge speech at a traditional pace however fails to attain similar performance once tested on slow or quick speaking rates [11]. A recent study has shown that a drop of just about ten proportion points within the classification accuracy is ascertained once a deep feed-forward network is trained on the conventional speaking rate and evaluated on slow and quick speaking rates. Convolutional neural networks (CNN) area unit accustomed see if this model will scale back the accuracy gap between completely different speaking rates. Filter bank energies (FBE) and Mel frequency cepstral coefficients area unit evaluated on multiple configurations of the CNN wherever the networks area unit trained on traditional speaking rate and evaluated on slow and quick speaking rates. The Deep neural network results obtained are compared. A breakdown of phoneme-level classification results and therefore the confusion between vowels and consonants is additionally given. The CNN design once used with FBE options performs higher on each slow and quick speaking rate are shown in the experiment. Associate in the Nursing improvement of nearly 2% in case of quick and 3% in case of slow speaking rates is ascertained.

The King Saud University Emotions' Arabic dataset is used in a speech feeling recognition technique supported by a deep learning neural network. The convolutional neural network and long remembering (LSTM) area unit accustomed style the first system of the convolutional repeated neural network (CRNN). This study additional investigates the employment of linearly spaced spectrograms as inputs to the emotional speech recognizers. The performance of the CRNN system is compared with the results obtained through the Associate in Nursing experiment evaluating the human capability to understand the feeling from speech. This human sensory activity analysis is taken into account because of the baseline system. The CRNN system achieves 84.55% and 77.51% accuracies for file and section levels, severally.

Emotions recognition from the speech is one among the foremost necessary subdomains within the field of signal process [13]. The system may be a two-stage approach, specifically feature extraction and classification engine. Firstly, two sets of options square measure investigated that are: 39 Mel-frequency Cepstral constant (MFCC) coefficients and sixty-five MFCC options extracted supported. Secondly, used the Support Vector Machine (SVM) because the main classifier engine since it's the foremost common technique within the field of speech recognition. Besides that, investigated the importance of the recent advances in machine learning together with deep kernel learning, furthermore because of the numerous sorts of auto-encoder (the basic auto-encoder and also the stacked auto-encoder). An oversized set of experiments square measure conducted on the SAVEE audio information. The experimental results show that DSVM methodology outperforms the quality SVM with a classification rate of 69.84% and 68.25% mistreatment thirty-nine MFCC, severally. To boot, the auto-encoder methodology outperforms the quality SVM, yielding a classification rate of 73.01%.

Table -1 Summary of Literature on Deep Learning Techniques for SER

References	Databases Used	Emotions	Features	Deep Learning	Accuracy
Zixuan Peng, Yu Lu, Shengfeng Pan, Yunfeng Liu, 2021 [2]	IEMOCAP dataset	angry, happy, sad and neutral	32-dimensional MFCC	MSCNN+SPU+ Attention Mechanism	weighted accuracy (WA) 80.3% and unweighted accuracy (UA)81.4%
Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Hasmah Mansor, Mira Kartiwi, Nanang Ismail, 2020 [6]	SAVEE speech emotion dataset	angry, happy, neutral, and sad	spectrograms using short-term Fourier transform (STFT)	Deep Stride Convolutional Neural Networks (DSCNN) Convolutional Neural Networks (CNN)	DSCNN with accuracy of 87.8%, CNN with 79.4% accuracy
Mustafa A. Qamhan,Ali H. Meftah,Mohammed Zakariah,Sid-Ahmed Selouani,Yasser Mohammad Seddiq,Yousef A. Alotaibi,2020	King Saud University Emotions (KSUEmotions)	neutral, sadness,happiness, surprise and anger	spectrograms using Discrete Fourier transform (DFFT)	CRNN	84.55%
Chuanzheng Wei, Xiao Sun, Fang Tian, Fuji Ren, 2019 [9]	CASIA-Chinese Emotional Speech Corpus	anger, fear, sadness, neutral, happiness and surprise	spectrogram by using STFT	CNN+LSTM+ Attention Mechanism	93.16%
Pavol Harar, Radim Burget, Malay Kishore Dutta, 2017 [15]	Berlin Database of Emotional Speech	angry, neutral, sad	downsampled to 16kHz (mono) standardized to have zero mean and unit variance	DNN	96.97%

Table 1 show the Summary of Literature on Deep Learning Techniques for SER. The database used, types of emotions chosen for classification, techniques used for classification and accuracy for each techniques are given in the table.

III.CONCLUSION

The emerging growth and development in the field of AI and machine learning have led to a new era of automation. Many advantages can be built over the existing systems if besides recognizing the words, the machines could comprehend the emotion of the speaker (user). A challenging product of creating machines with emotion is to incorporate a sarcasm detection system. Sarcasm detection is a more complex problem of emotion detection since sarcasm cannot be easily identified using only the words or tone of the speaker. Therefore, in the future, there would emerge many applications of a speech-based emotion recognition system. This paper has provided a detailed review of the deep learning techniques for SER.

REFERENCES

[1] Sumanathilaka TGDK,Viggnah Selvarai, Uddav Raj, Venkatesh P Raiu, Jay Prakash, "Emotion Detection Using Bi-directional LSTM with an Effective Text Pre-processing Method", *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, July 2021.

[2] Zixuan Peng, Yu Lu, Shengfeng Pan, Yunfeng Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2021.

[3] Shambhavi Sharma, "Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks", *11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, March 2021.

[4] Mai Ezz-Eldin, Ashraf A. M. Khalaf, Hesham F. A. Hamed, Aziza I. Hussein, "Efficient Feature-Aware Hybrid Model of Deep Learning Architectures for Speech Emotion Recognition", *IEEE Access*, 9, January 2021.

[5] Mehmet Bilal Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features", *IEEE Access*, 8, 221640–221653. doi:10.1109/access.2020.3043201, December 2020.

[6] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Hasmah Mansor, Mira Kartiwi, Nanang Ismail, "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks", *6th International Conference on Wireless and Telematics (ICWT)*, November 2020.

[7] R. Chinmayi, Narahari Sreeja, Aparna S. Nair, Megha K. Jayakumar, R. Gowri, Akshat Jaiswal, "Emotion Classification Using Deep Learning", *Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* - Tirunelveli, India, October 2020.

[8] Himanshu Joshi, Ananya Verma, Amrita Mishra, "Classification of Social Signals Using Deep LSTM-based Recurrent Neural Networks", *International Conference on Signal Processing and Communications (SPCOM)* - Bangalore, India, August 2020.

[9] Chuanzheng Wei, Xiao Sun, Fang Tian, Fuji Ren, "Speech Emotion Recognition with Hybrid Neural Network", *5th International Conference on Big Data Computing and Communications (BIGCOM)* - QingDao, China, November 2019.

[10] Sandeep Kumar Pandey, H. S. Shekhawat, S. R. M. Prasanna, "Deep Learning Techniques for Speech Emotion Recognition", *29th International Conference Radioelektronika (RADIOELEKTRONIKA)* - Pardubice, Czech Republic, June 2019.

[11] Abdolreza Sabzi Shahrebabaki, Ali Shariq Imran, Negar Olfati, Torbjorn Svendsen, "A Comparative Study of Deep Learning Techniques on Frame-Level Speech Data Classification", *Circuits, Systems, and Signal Processing, Springer,* April 2019.

[12] Tang Baolong, Li Yuanqing, Li Xuesheng, Xu Limei, Yan Yingchun, Yang Qin, "Deep CNN Framework for Environmental Sound Classification using Weighting Filters", *IEEE 2019 IEEE International Conference on Mechatronics and Automation (ICMA)* - Tianjin, China, 2019.

[13] Hadhami Aouani, Yassine Ben Ayed, "Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder", *4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* - Sousse, Tunisia, May 2018.

[14] Norman Weiskirchen, Ronald Bock, Andreas Wendemuth, "Recognition of emotional speech with convolutional neural networks by means of spectral estimates", *Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* - San Antonio, TX, USA, February 2018.

[15] Pavol Harar, Radim Burget, Malay Kishore Dutta, "Speech emotion recognition with deep learning", *4th International Conference on Signal Processing and Integrated Networks (SPIN)* - Noida, Delhi-NCR, India, September 2017.