IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORK BASED ON DEEP LEARNING

[1]Pavithra P, [2]Priya N, [3]Naveenkumar E, [4]Dr.Uma
[1]PG Scholar, [2]Assistant Professor, [3]PG Scholar, [3]Professor In-Charge PG-CSE
[1-4]Hindusthan College of Engineering and Technology, Coimbatore.

Abstract

In modern days, person-computer communication systems have gradually penetrated our lives. One of the crucial technologies in person-computer communication systems, Speech Emotion Recognition (SER) technology, permits machines to correctly recognize emotions and greater understand users' intent and human-computer interlinkage. The main objective of the SER is to improve the human-machine interface.It is also used to observe a person's psychological condition by lie detectors. Speech recognition has been used in medicine and forensic medicine.Automatic Speech Emotion Recognition(SER) is vital in the person-computer interface, but SER has challenges for accurate recognition.In this work to resolve the above problem, automatic Speech enhancement shows that deep learning techniques effectively eliminate background noise.Using Deep leaning models for four states were created: happy, sad, angry, and intoxicated. Recurrent Neural Network (RNN) algorithm used to reduce the possibility of over fitting by randomly omitting neurons in the hidden layers. For accurate classification results, preprocessing filters extract voice and non-voice features from speech. Then feature selection is extracting the speech relevant features to predict accurate results.The proposed RNN method could be implemented in personal assistant systems to give better and more appropriate state-based interactions between humans. In the simulation results shows Improving accuracy, Time complexity, Error rate is also reduced to using the proposed method.

Keywords: Speech Emotion Recognition (SER), Speech emotion detection, deep leaning, Recurrent Neural Network (RNN), preprocessing, feature extraction.

1. Introduction

Speech is one of the basic and natural ways of communicating with human beings. Emotions make speeches more expressive and useful. Humans express their emotions in various ways, such as laughing, screaming, teasing and crying.Emotions play a vital role in everyday relationships. It is essential for our rational and informed decision making. It helps to integrate and understand the feelings of others by communicating our feelings and giving them feedback.

Emotion conveys a lot of information about an individual's mental state. Its primary goal is to understand and receive the desired emotions.Since speech is one of the main ways of expressing emotions, it is important that the natural human-machine interface recognizes, interprets, and responds to the emotions expressed in speech.Emotions affect both speech properties and the linguistic content of the speech. This paper aims to the phonological properties of speech to detect possible emotions.

Also, it is not easy to discover an objective fundamental truth about a particular person's current emotional state in real life, but it is a prerequisite for automatic recognition. For these reasons, emotional recognition is rarely used in business products because the authenticity accuracy of current systems is very low.

This study using deep learning technique have numerous advantages over existing algorithms, such as manually disassembling complex structures and features and identifying features without the need to adjust features of speech recognition. A deep learning-based approach extracts low-level features from specific source data and handles unlabeled data features.More precisely, it recognizes the speech characteristics of the emotion, such as pitch, noise and frequency spectrum distribution, using the proposed RNN algorithm.There is a growing interest in using in-depth learning to learn useful aspects from emotional speech data automatically.

2. Related work

The type of phoneme defines the phoneme characteristics that are not represented. Deep neural networks (DNN) are used to evaluate the probability of speech class in the speech signals. Theoretically, a unique combination of phoneme characteristics creates a phoneme identity [1].Therefore, the probabilistic rationality of speech classes can assess their integrated phonological possibilities. The new information theory framework is designed to measure the information transmitted by each phoneme attribute and to evaluate the quality of the phoneme recognition of speech development.

That study investigates an end-to-end speech recognition system for Japanese people. Speech recognition systems have difficulty recognizing their speech because it is often irregular or vague.The former was derived from non-Japanese rough data, the latter from non-Japanese rough data [2] [3].However, these methods require a large amount of training data, and it is difficult to collect sufficient data from these some patients.

Traditional speech development methods based on frame, multi-frame or segment assessment require knowledge of noise. That study describes a new approach to reducing or effectively eliminating this need [4].Using the zero-mean normalization

correlation coefficient as a comparative measure and extending the effective length of the speech segment to fit the speech with longer sentences, the results provide an accurate speech evaluation from the noise without the need for specific knowledge.

The traditional automated speech recognition systems trained by neutral speech are significantly reduced.To analyze this paradoxical training/experiment situation in-depth and develop effective whisper recognition methods, this study first analyzes the phonological characteristics [5] [6]. It identifies the problem of whisper recognition under disagreement conditions.Further analysis of septum distance, septum coefficient distribution, confusing matrix, and reverse filter tests shows that the voice of speech stimuli is a major factor in the misinterpretation of the word in random training/test situations.

The RNNLM convert to new domains is an open issue, and current approaches can be classified as feature-based or model-based. Functional-based adaptation adds sub-functionality to RNNLM inputs while fine-tuning the model-based adaptation model and introducing an adaptation layer into the network [7].In multi-type broadcast speech recognition, two types of adaptation characteristics are explored. It will review the cutting edge of these two adaptations and explore model-based adaptation technologies: linear hidden network adaptive layers and K-component adaptive RNNLM.

The majority demanding task in audiovisual automated speech recognition is to attract research interest [8]. Over the past few decades, several methods have been proposed to integrate the audio and video system to improve the performance of automated speech recognition in a clean, noise-free environment.However, some studies in the literature have compared the AV-ASR models with different fusion models [9]. Some studies compare audiovisual fusion models with large vocabulary continuous speech recognition models using DNNs.

Home appliances with microphone sets, such as car navigation devices and headsets, use gradient-based speech enhancement technology to handle noise for speech enhancement [10].However, although these techniques were originally developed for voice communication and can increase the signal-to-distortion rate, they do not always increase the accuracy of automatic speech recognition. For this reason, the parameters for pre-final speech development have been designed by human experts to suit every environment and sound model.

Recognition is a problem in recognizing speech produced by people with motor speech disorders that interfere with the body's speech production. People with pronunciation disorders have less tone control, making it harder to pronounce certain sounds, resulting in unwanted audio changes [12].The latest automated speech recognition systems designed for normal speakers are ineffective for people with dysarthria due to speech variation. Used to capture KL-HMM audio transitions.

The lack of adequately labelled voice data for practice severely restricts the widespread use of supervised learning methods in speech-emotional recognition. Therefore, considering the wide availability of unlabeled speech data, that novel recommends a semi-monitored automated encoder to improve speech sensory recognition [13].The

purpose is to derive from a combination of named and unnamed data. The proposed model carefully combines the objectives of supervised learning and extends the automated code that is not generally overseen.

Automatic Speech Recognition systems are sensitive to noise levels in real-world use. Adding visual cues to the ASR system is an attractive alternative to enhancing system strength and reflecting audiovisual and cognitive processes used during human interaction [15].The experimental evaluation of this study underscores this problem when training a sophisticated audiovisual hybrid system with a DNN and a Hidden Markov Model (HMM). That study suggests a framework for solving this problem, improving or maintaining performance when using visual features.

3. Materials and Method

This section express theSpeech Emotion Recognition (SER) based on happy, sad, angry, and neutral using the proposed Recurrent Neural Network (RNN) algorithm. The proposed method contains three process there are Feature extract, speech pitch and loudness analysis and classification.

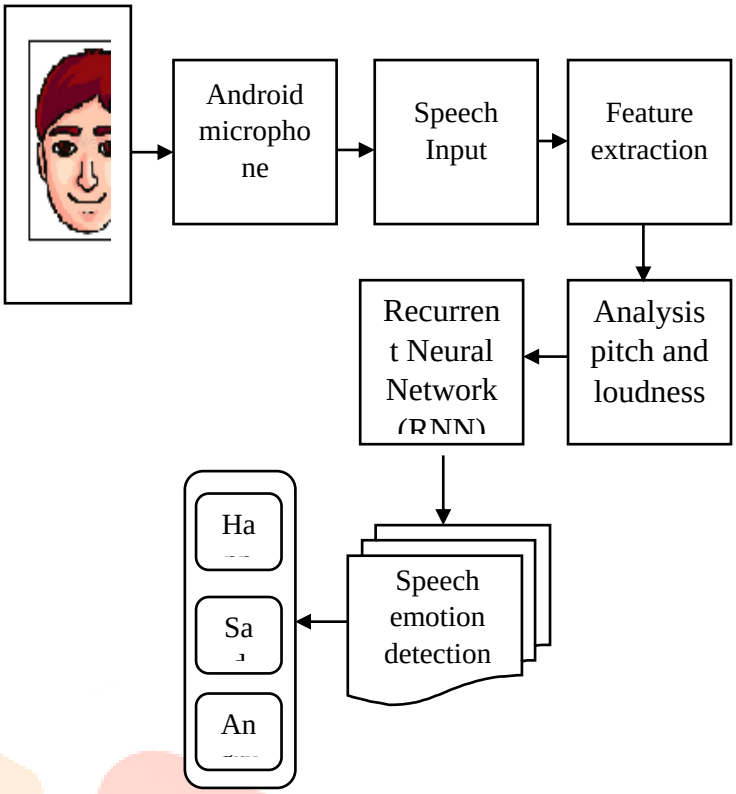

Figure 1: Proposed Diagram

Figure 1 illustrates the proposed algorithm SER classification process. Audio sample collection first. Extract the second aspect vector created by the features. Next, identify the most appropriate features to differentiate each emotion. These features have been introduced into the deep learning classifiers to identify them.

3.1 Feature extraction

The original audio signal can be converted into sound features such as continuous features, spectral features, and standard features at the feature extraction stage.Audio signals have many parameters that reflect sensory characteristics. Which features should use is one of the difficulties in recognizing emotions.The spectral envelope representation of abstract voice digital signals used in voice and signal processing is called the linear prediction model.Advanced voice analysis technology and modern methods for high-quality

voice at low bitrates provide accurate estimates of voice parameters.

Algorithm steps

Input:User voice (U_v)

Output: Features extract voice E_v

Start

Import user voice U_v

$$U_{vn} = U_{v1,}U_{v2}\ldots U_{vn}$$

Analysis the speech signal prediction$\widehat{U_{vn}}$

$$\widehat{U_{vn}} = \sum P_s(i-t)$$

Extract relative speech information E_v

$$E_v = \sum \widehat{U_{vn}}\, exp\left(\frac{2\pi}{n}\right)$$

Obtain E_v

End

Where P_s presents linear predictive signal, i refers to iteration input user voice, t refers to time index. The above algorithm steps first calculates speech signal, then extract the relevant speech informationE_v.

3.2 Analysis of pitch and Loudness

Pitch is the ratio of the frequency of a sound wave. Frequency is the number of peaks and troughs within a sound wave. These peaks and troughs represent the vibrations of the sound waves.Lower sounds have the lowest number of vibrations, and the highest number of sounds has the highest number of vibrations. The treble seems very unpleasant, but the bass is comfortable. In general, the angry one will give a loud voice.

To calculate vocal pitch and loudness (V_{pl})

$$V_{pl} = E_v + p_l^{-(E_v)}(p_l^{-1})$$

Where p_l refers to speech pitch and loudness

Loudness depends on the amplitude of the sound wave. The amplitude is the distance from peak to idle or tank to idle. If the displacement is large, the noise is high. If the amplitude is high and the displacement is small, the noise is low.Some people's emotions can be easily identified by their voices.

3.3 Recurrent Neural Networks (RNN)

RNNs are ideal for learning time series data and improving the efficiency of classification tasks.RNNs use a multilayer perceptron-like system designed to reduce processing needs.RNN layers consist of input, output and multiple convolution layers. Removing controls and increasing the effectiveness of speech recognition will make your computer more efficient.

Algorithm steps

Input: vocal pitch and loudness (V_{pl})

Output:Classified result(c_r)

Begin

Import vocal pitch and loudness (V_{pl})

$$V_{pln} = V_{pl1,}V_{pl2}\ldots V_{pln}$$

Calculate speech frequency (s_f)

$$s_f = \frac{S_{mx} - S_{mn}}{257}(Hz)$$

Evaluate the weight function F_w

$$F_w = \sum s_f * V_{pl}$$

Update the values in c_r

Obtain classified result c_r

End

Let assume Hz refers to Hertz, S_{mx} denotes speech frequency maximum values, S_{mn}refers to minimum values. The algorithm steps give efficient classification results based on users' speech, anger, sadness, happiness, and neutral.

4. Result and discussion

This section provides and describes simulation experiment test results. The simulation language is java, and the tool is the android studio on Linux Operating System, as shown in table 1.

Table 1: Parameters for Automatic SER

Constraints	Values
Tool	Android studio
Language	Java
Processor	Intel corei5
OS	Linux

From table 1 conclude the parameters constraints and Values for automatic SER. The Previous algorithms are Hidden Markov models (HMMs), and Deep Neural Networks (DNNs).

Table 2: Analysis ofSER accuracy performance

Comparison methods	Accuracy performance in %
HMM	68
DNN	74
RNN	83

From table 2 conclude for RNN algorithm has high performance compared with previous algorithmsare HMM and RNN.

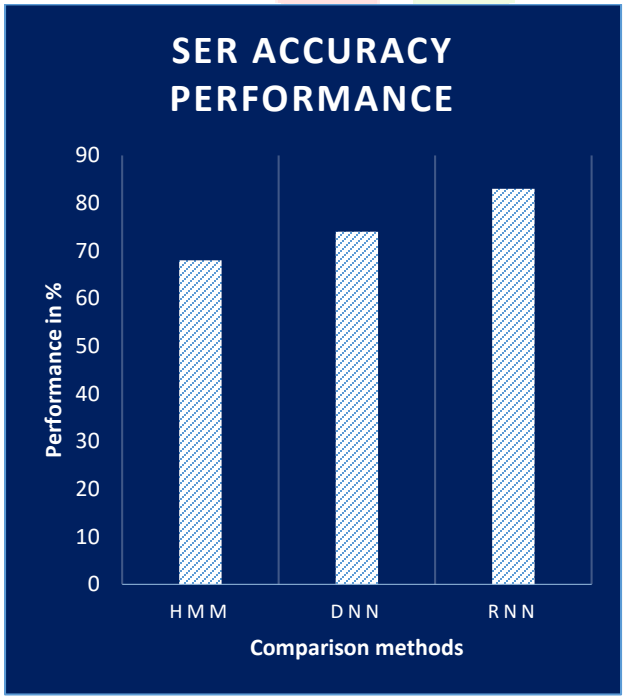

Figure 2:Analysis of SER accuracy performance

Figure 2 explores an analysis of SER accuracy performance comparison approaches present in the graph. The proposed RNN algorithm accomplish 83%, with the existing algorithm results are Hidden Markov models (HMMs) algorithm accomplish68%, and Deep Neural Networks (DNNs) algorithm accomplish74%.

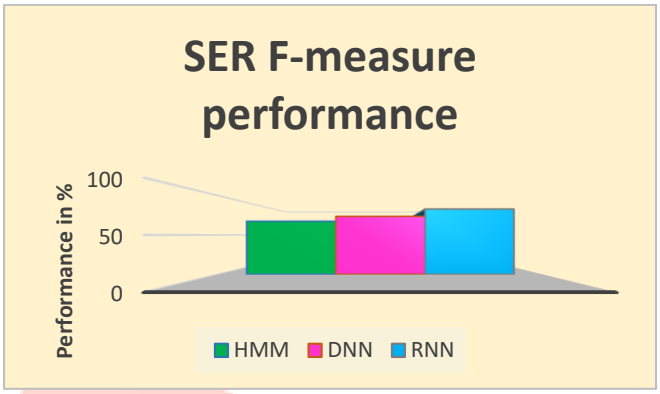

Figure 3: Analysis of SER F-measure

Figure 3 explores an analysis of SER F-measure performance comparison approaches present in the graph. The proposed RNN algorithm accomplish 82.9%, with the existing algorithm results are Hidden Markov models (HMMs) algorithm accomplish 67.4%, and Deep Neural Networks (DNNs) algorithm accomplish 73.6%.

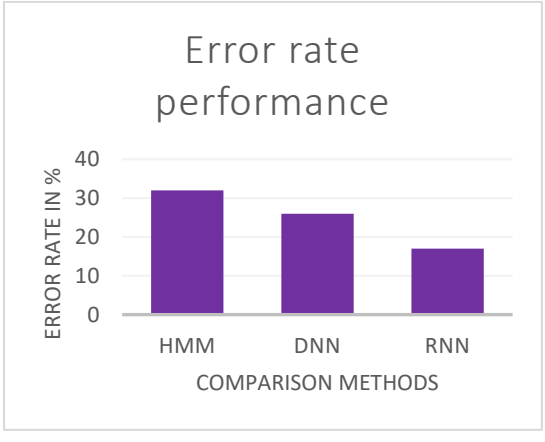

Figure 4: Analysis of error rate performance

Figure 4 explores an analysis of error rate performance comparison approaches present in the graph. The proposed RNN algorithm accomplish 17%, with the existing algorithm results are Hidden Markov models (HMMs) algorithm accomplish 32%, and Deep Neural Networks (DNNs) algorithm accomplish 26%.

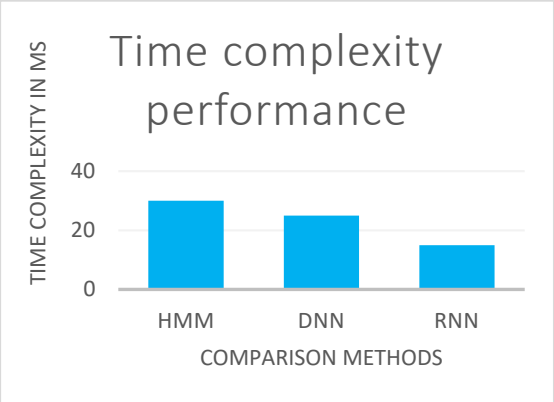

Figure 5: Time complexity performance

Figure 5 explores an analysis of time complexity performance comparison approaches present in the graph. The proposed RNN algorithm accomplish 15ms, with the existing algorithm results are Hidden Markov models (HMMs) algorithm accomplish 30ms, and Deep Neural Networks (DNNs) algorithm accomplish 25%.

5. Conclusion

The paper shows that the recurrent neural networks (RNN) algorithm is powerful for audio signal classification. The simplified model can even identify small voice signals.The performance of the SER depends on the quality of the feature extraction.Then feature selection is extracting the speech relevant features to predict accurate results. The proposed RNN method could be implemented in personal assistant systems to give better and more appropriate state-based interactions between humans. The proposed method achieves the results are accuracy with 83%, F-measure with 82.9%,Error rate with 17%, and time complexity with 15ms.In the simulation results shows Improving accuracy, Time complexity, Error rate is also reduced to using the proposed method.

References

1. A. AsaeiCernak and H. Bourlard, "Perceptual Information Loss due to Impaired Speech Production," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2433-2443, Dec. 2017, doi: 10.1109/TASLP.2017.2738445.
2. Y. Takashima, R. Takashima, T. Takiguchi and Y. Ariki, "Knowledge Transferability Between the Speech Data of Persons With Dysarthria Speaking Different Languages for Dysarthric Speech Recognition," in IEEE Access, vol. 7, pp. 164320-164326, 2019, doi: 10.1109/ACCESS.2019.2951856.
3. J. Ming and D. Crookes, "Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 531-543, March 2017, doi: 10.1109/TASLP.2017.2651406.
4. Đ. T. Grozdić and S. T. Jovičić, "Whispered Speech Recognition Using Deep DenoisingAutoencoder and Inverse Filtering," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2313-2322, Dec. 2017, doi: 10.1109/TASLP.2017.2738559.
5. S. Deena, M. Hasan, M. Doulaty, O. Saz and T. Hain, "Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast

Speech Recognition and Alignment," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 3, pp. 572-582, March 2019, doi: 10.1109/TASLP.2018.2888814.

6. H. Abdelaziz, "Comparing Fusion Models for DNN-Based Audiovisual Continuous Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 3, pp. 475-484, March 2018, doi: 10.1109/TASLP.2017.2783545.
7. T. Kawase, M. Okamoto, T. Fukutomi and Y. Takahashi, "Speech Enhancement Parameter Adjustment to Maximize Accuracy of Automatic Speech Recognition," in IEEE Transactions on Consumer Electronics, vol. 66, no. 2, pp. 125-133, May 2020, doi: 10.1109/TCE.2020.2986003.
8. M. Kim, Y. Kim, J. Yoo, J. Wang and H. Kim, "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 9, pp. 1581-1591, Sept. 2017, doi: 10.1109/TNSRE.2017.2681691.
9. J. Deng, X. Xu, Z. Zhang, S. Frühholz and B. Schuller, "SemisupervisedAutoencoders for Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 31-43, Jan. 2018, doi: 10.1109/TASLP.2017.2759338
10. C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu and Z. Wen, "Gated Recurrent Fusion With Joint Training Framework for Robust End-to-End Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 198-209, 2021, doi: 10.1109/TASLP.2020.3039600.
11. L. Chai, J. Du, Q. -F. Liu and C. -H. Lee, "A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 106-117, 2021, doi: 10.1109/TASLP.2020.3036783.
12. G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy and J. C. Kline, "Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2386-2398, Dec. 2017, doi: 10.1109/TASLP.2017.2740000.
13. B. Wu et al., "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1289-1300, Dec. 2017, doi: 10.1109/JSTSP.2017.2756439.
14. M. Sakurai and T. Kosaka, "Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results," 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), 2021, pp. 824-827, doi: 10.1109/GCCE53005.2021.9621810.
15. F. Tao and C. Busso, "Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1290-1302, July 2018, doi: 10.1109/TASLP.2018.2815268.