ORIGINAL ARTICLE e-ISSN: 2349-0659 p-ISSN; 2350-0964

Artificial Intelligence Based Automatic Speech Emotion and Drunkenness Detection Using Convolutional Neural Network Voice Segment Dropout Optimizer

T. Geetha*, M. Annamalai, R. J. Vigneshwaran, A. Yuvaraj

Abstract

Speech recognition is far from an easy task, because of the unique nature of the human emotional text and a speech recognition model must strike a balance between being accurate. It should be accurate and sufficient enough to cause a broad group of datasets to separate and prevent errors. The models of the states created using artificial intelligence: Positive, sadly angry, and likely to fall are the easy ways to find the drunkenness or non-drunkenness. By building on the previous methods of emotional detection, drunkenness data set experiences can accurately identify these states. The solution development goal is to evaluate in real time because of the different voice recognition from consumers. It is reported to give the user an idea of what emotion they are experiencing or whether they are stumbling, if an assistant detects a drunk person, it will alert him or her about or making large online data extracted drunk information. If the model proves reliable, it can be used in voice-activated, the artificial intelligence (AI) is characterized by a lack of social or situational awareness of the person with it interacts to active on first stage. The second stage to direct on support Artificial Intelligence and Convolutional Neural Network Voice Segment Dropout Optimizer (CNN-VSDO) algorithm is a system that maintains the issues mentioned above that minimize the risk of overlapping. The convolutional neural network for classification employs filters to features from the spectrogram. This approach is used in personal assistant systems to provide the highest and most suitable of AI-human interaction and more appropriate state-based interactions between AI and humans.

Keywords: Artificial intelligence, CNN, Drunkenness recognition, Speech emotion, Voiced segment dropout optimizer

Asian Pac. J. Health Sci., (2022); DOI: 10.21276/apjhs.2022.9.4S.20

Introduction

Computerized reasoning in Personal Computer (PC) vision is utilized to examine, reproduce and learn various approaches to comprehend three-dimensional pictures from its two-dimensional views. The scenes decide the association of the meeting constructions to the genuine explicit video. It regularly has the strategies to comprehend, study and interact with computerized pictures to acquire. Video preparing is vital in local meetings, country borders, banks, jungle gyms, workplaces, air terminals, and shopping centers. The issue of human location, observing, and operational acknowledgment has acquired conspicuousness in the field of PC initially.

Identification and mobility of objects and video surveillance system are the latest developments. Functional authentication monitoring of these bodies is an important task, on recent times; these various artificially intelligent video managing and monitoring systems have been used. Applications include surveillance video, using the patient control unit, game video, and traffic management. These are used primarily to locate event and speech recognition according to the operation of the recorded video set and giving a well-organized and effective approach to security. The level of activities is likely to occur very frequently, due to the level of activities that may show differences due to the low video quality in the background. The change background, overlapping situation, different human visual points, and harassment may develop in the background.

It uses their characteristic portrayal of the character, how individuals are in their direction. The feeling of a person is utilized to interface with human robots. It gives far and wide information about the utilization of PC mechanical technology and the cooperation among people and numerous other human practices. Recognizing the kinds of various exercises is a framework that is required.

Organizations are becoming better acquainted with their customers. Only customers know as much data as possible gives the best customer experience. These data are exceptionally difficult and conventional techniques must be used. In any case, with man-made cognitive power, it is possible.

The client conducting examinations with AI capacities can save a tremendous measure of time contrasted with human representatives. Every one of the bugs that individuals can do will be taken out. However, that does not mean information examiners that are getting pointless.

The errand of this examination plan is to build up an answer dependent on the various sorts of propensities for portable application clients to recognize the kind of AI to conduct research

Department of Computer Science and Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Vinayaka Missions Research Foundation (Deemed to be University), Salem, India.

Corresponding Author: T. Geetha, Department of Computer Science and Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Vinayaka Missions Research Foundation (Deemed to be University), Salem, India. E-mail: geetha@vmkvec.edu.in

How to cite this article: Geetha T, Annamalai M, Vigneshwaran RJ, Yuvaraj A. Artificial Intelligence Based Automatic Speech Emotion and Drunkness Detection Using Convolutional Neural Network Voice Segment Dropout Optimizer. Asian Pac. J. Health Sci., 2022;9(4S):103-108

Source of support: Nil

Conflicts of interest: None.

Received: 3/3/2022 **Revised:** 19/04/2022 **Accepted:** 20/05/2022

and give information to the client to propose valuable substance. They have the possibility of the benefits that singular calculations for medication and these sorts of administrations can give (e.g., human-machine communication, efficient, and canny data decrease). At present, they have theoretical evidence.

Figure 1 shows that AI is still widely accepted by companies. Where AI certainly shows the use of a standard item life cycle or arrangements.

Indeed, even in the dead business, human-made brainpower item supervisor cannot lead an AI item advancement life cycle without an item executive's range of abilities at the center, so AI arrangements cannot be refreshed. It is noticed that the correct business explicit space information shows AI arrangement understanding just as appropriate cooperation. It becomes more acquainted with the organization and item vision with clients, stays lined up with them, gets them to perceive their problem areas, sets up business associations and contacts, gets deals, and promotes help and criticism (market opportunity) on making stories. Speech Expansion bargains many speakers for counter-vibrations in dynamic circumstances.

Literature Survey

Li *et al*., (2014) authors described that customer-driven applications, for example, voice search and cell phones and home theater setups convey by voice, require robotized acknowledgment innovation to stay strong because of the full scope of certifiable clamor and other sound bending conditions talked about above. Issues mature key bits of knowledge from complete outline in this field and investigate a couple of old issues, which are still exceptionally important today.[1] Specifically, they have considered and arranged a wide scope of commotion fortifying methodologies.

Zhang *et al*., (2016), the authors discussed that the creators have portrayed over the previous decade and have reached an experienced sufficient stage that may be utilized in some genuine situations. Nonetheless, these situations require a practically peaceful climate that doesn't bargain the elements of the framework. Intellectual innovation from this way arises steadily, and grows its relevance to genuine conditions on the face. Conditions such as ecological conglomeration and convolutional disorders should be tested.[2]

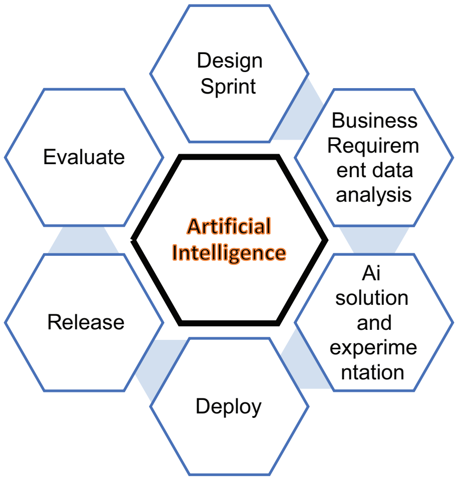
Design Sprint
Business Requirem ent data analysis
Ai solution and experime ntation
Deploy
Release
Evaluate
Artificial Intelligence

Figure 1: General architecture of artificial intelligence flow diagram

Gao *et al*., (2016), authors described that double acknowledgment of both short code and transmission capacity dependent on Deep Neural Networks (DNNs) encaged joint demonstrating methodologies for enormous scope blended band preparing infection. They can exploit customary base testing and up examining plans to go among short code and transmission capacity information. The investigation of the DNN-based Speech Bandwidth Extension (BWE) organization is done in distinguishing some solid highlights from the Wi-Fi data transfer short wave.[3]

Squartini *et al*., (2012), the authors discussed that highlight coefficients in the space are perhaps the most automatic ways to deal with speaker acknowledgment for hearty robotized decrease and clamor sound presentations: Included coefficients are standardized utilizing fitting straight or various changes to coordinate with the highlight see point framework.[4]

Zhang *et al*., (2014), the authors discussed that the two-way long memory utilizing without hands voice control gadgets for vibration is repeatedly tended to by neural organizations. Such organizations can abuse a self-trained degree of mainstream setting to utilize memory modules in secret units. The principal reason for this system is to diminish the confound between inaccessible (reverberation/misshaped) and close. To accomplish this, the campestral perspective is prepared by space planning from the closest channel, the identical far off channel based on keen relapse to the organization law.[5]

Han *et al*., (2014), the authors described that direct opposite contorts human disuse, particularly in hearing-weakened audience members, and ordinarily, disuse can have mentally unfortunate results. It causes execution corruption in programmed acknowledgment and speaker acknowledgment frame works.[6] The proportion issue should be dealt with in the everyday listening climate along these lines.

Zhang *et al*., (2016), the authors discussed that technologies of Recuperation Neural Organizations present direct private associations between memory cells in progressive layers of Distributed Long Short-Term Memory (DL-STM). These immediate connections get interstate associations, empower continuous data progression across various layers, and accordingly take out gradient blurring when constructing profound LSTMs.[7]

Rotili *et al*., (2013), the authors discussed that commotion Speech Expansion bargains many speakers for counter-vibrations in dynamic circumstances. The creators proposed room vibration in far-off signals for programmed ongoing handling. They proposed the front-end to decrease twists within sight of foundation commotion by consequently accomplishing a huge improvement in every speaker's quality.[8]

Omar *et al*., (2010), the authors explored that the principal non-local speakers and general enormous information frameworks have programmed recognition in their language. They have presently led a few investigations to show that framework Fisher Information and Foreign Accented English (FAE) data beat both non-local speakers in their language separately.[9]

Compernolle *et al*., (2017), authors discussed that diminishing word mistake rate is an individual test to build up a Automatic Speech Recognition (ASR) framework. The exhibition of this framework is a long way from great, sound models and language models are the rudiments for building a solid R motor. This review presents the difficulty or practice of creating sound models on the size of sound and word. Mel Frequency Cepstral Coefficients (MFCC) rating of 35% of the edges are on top of one another in every 25 ms of the sound sign. This survey analyzes and shows Kannada language jargon word and phonetic level sound example shows. The framework elements are recorded for various jargon levels, and the Word Error Rate (WER) is determined for phonetic and word sound models.[10]

Zoughi *et al*., (2015), authors discussed that technologies of Deep Neural Network (DNN) have become an extremely well-known way to deal with acknowledgment, which has had a great deal of progress. Because of the DNN preparing is to adjust and the size of the data set, its various activities are troublesome.[11]

Vipperla *et al*., (2010), authors discussed that human voices will, for the most part, be portrayed by expanded roughness and changes in the phonetic structure, including numerous changes. In this examination, the impact is analyzed and afterward discovered and looked at the basal recurrence, quake, glare, harmonics, and spectral top significance in grown-ups and a few voice factors, including more seasoned men. A considerable lot of these boundaries show the critical factual distinction between the two gatherings. Nonetheless, incorrectly amplifying vibration, data loss, does not affect brightness. Artificially, the execution is marginally decreased to the base recurrence level, yet the exhibition can be defeated by somewhat utilizing this drop Voice Track Length Standardization (VTLS).[12]

Potamianos *et al*., (2003), authors discussed that social changes present the signal delivered by kids into ghastly age-subordinate and common varieties. N present tests in such varieties for solid programmed acknowledgment of kids'. Regarding programmed acknowledgment, youngsters exhibit the recurrence impacts of otherworldly envelope boundaries like age-related sound properties from the analyzer. Acknowledgment tests plainly show age-related execution weakening in the offspring of various ages by utilizing grown-up to preparing. Word mistake rates are multiple times more awful in kids than in grown-up. Different methodologies for improving youngsters' conversational R capacities are recommended.[13]

Sinha *et al*., (2016), author clarified the work introduced in Adult Children's Speech on Language Automatic Speech Recognition. In the unique situation, Pitch prepares versatile campestral highlights good examples. Huge commotion among preparing and test information have been noted in cases, for example, acknowledgment rates, which are extraordinarily diminished. Early investigations have shown that on account of sound Mel-frequency Campestral Coefficient (MFCC), there is help for pitch synchronization and absence of reliable youngster speakers. Roused by that, this work lessens the affectability of the pitch versatile campestral highlights to add up to pitch varieties.[14]

Steidl *et al*., (2010), authors that programmed acknowledgment of it is reported that leaving children cannot be a test, so diminishing disuse analysis is accepted to assume a part in affecting exhibition. In this commitment, examine the two occasions together, far-reaching trials against unbiased have been done with preposterous assurance of 1K jargon, Continuous Speech acknowledgment, and irate kids'. Tests address the topic of what explicit feelings mean for word precision. In a first situation, "enthusiastic" recognizers are equivalent to a recognizer in unbiased preparing. In contrast with this, preparation information is utilized in equivalent measure for each sincerely related state.[15]

Proposed Method

The first is the justification for using a CNN to accomplish objectives, which is particularly valid in all these drunken speech detections. Emotion identification is substantially less dense than research. Because of this knowledge gap, using the neural loop is the most effective way to improve drunken speech that a wide-band spectrum analyzer over a narrow-band spectrogram for pre-processing is chosen. Wide-band spectrograms have a lower time resolution, allowing for displaying the independent laryngeal pulse rather than precise voice modulation. When the air moves through the vocal folds, the laryngeal lid is accountable for the voice pitch when a brief burst of energy happens. The pulse frequency determines for speech energy changes, another alternative is to amplify speech data and add white noise as a pre-processing step. This is because there are fixes that allow it to work with a smaller database. It is applicable when using over-CNN models to train speech data and it works best when the sample output is negatively impacted when evaluating new data. The "power" of the interference model is enhanced by CNN sensitivity to minor changes in input. It allows any restrictions on the output-input to be ignored.

This section describes him for implementation the pattern is depicted in Figure 2 show the overall process and. AI awareness, extract the line segments are described a number of aspects to separate the signals.

Data Source

The data used in the training of emotional states (happy, sad, angry, and neutral) were taken from the Ryerson Audio-Visual Database of RAVDS. This information consisted of mono audio recordings of 24 speech recognition (about three to few seconds long) - 12 male and 12 female. Two forms of emotional frequencies (normal and strong) were used in the files, along with two situations ("children talking through the door" and "dogs sitting by the door") and two sentence fragments. It produced a total of 672 audio files based on this data.

Self-raw drunken speech models are chosen because the corpus of anesthetized speech is relatively small and open. For lessons, when the speech stigma started, drag voice recordings of encounters or shows impaired to extend set. Furthermore, when network needs solely English preparation, this corpus includes models exclusively. If several databases are used, most studies isolate their samples by language and the possible outcomes are language-specific.

Sense recognition variations exist with English systems having a higher recognition rate and higher accuracy rates, according to native speakers. It chose to stick to only plain English details because English showed high accuracy, and their findings showed minor variations in all languages. The used 96 samples per color

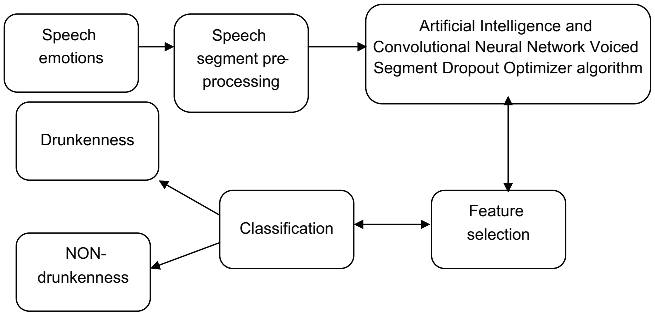
Speech emotions
Speech segment pre-processing
Artificial Intelligence and Convolutional Neural Network Voiced Segment Dropout Optimizer algorithm
Drunkenness
Classification
Feature selection
NON-drunkenness

Figure 2: Proposed flow diagram

impression balance the size after consuming as much information as possible.

Audio Processing

To make network's training easier, perform various before measures on database, first, the audio files were designed at 44.1 kHz from 16.1 kHz to minimize preparation time by manipulating the remote sensing data talk features below, which had a very similar correlation matrix. Except that, each audio file is fitted with a specific dataset aided at 15 beats per minute Signal Noise Ratio –SNR. The output signal initially grows 20% of the original duration. It is discovered that accuracy results in amplification were better after studying network preparation section is very thorough.

Network Architecture

A convolutional neural network was used to build models in each of the states that trust them, detecting circulation, funding, and completely linked all layers at the same time. In this analysis, CNN could have used two convolutional layers (each following a max-accumulation layer) and one completely integrated layer with 1024 secret neurons. The first vibration layer has a kernel size of 10 with 8 kernels, and the second convolutional layer 5 has a kernel size of 16 grain. The kernel size for affected entire 2. 2 is set to represent a CNN architecture.

The random variable of classes and a neutral class, a five-way smooth max unit class, was estimated using emotions. Completely incorporated into the updated positive units and input variables in the convolution operation were used to implement the model predictability feature. Even though the Adam Optimizer algorithm was used, inter was used to quantify the loss, and the confusion matrix on the mini package was reduced (between 16 and 256). The number of improvements in the test differed, but the preparation iteration's actual value reached 100.

Classification Using CNN

As a set of cognition of expressions, texture parameters of exposure level are given. Used a recently advanced classifier named VSDO because the size of the training claims is limited. When used in small groups, the VSDO training kit has been shown to produce pledge outcomes. The VSDO is typically a single hidden layer multilayer perceptron that needs more hidden units than traditional neural networks (CNNs) to achieve substantial classification accuracy, VSDO's preparation approach is straightforward. Unlike traditional NNs, which require the redistribution algorithm to change their weights, the VSDO input is stable after the estimated assignment of weights between the layer and the input. Inside the analysis process, the weights between the hidden and output layers are calculated by a simple constructive interference and CNN reverse movement is expressed as possible.

One of the several forms of classification based on audio monitoring is recognition that identifies speech, emotion, and alcohol intake is an analysis of show Figure 3 Convolutional Neural Network-CNN. Machine learning feature selection algorithms are used to produce the best outcomes, since then, neural network and framework have used the chosen features to implement various artificial intelligence methods. Finally, like simulations, these various approaches are compared to one another. In any case, they achieve the best performance and outcomes.

The audio dataset will have nine sections at first, Figure 4 depicts data preprocessing. Some of the nine columns contain noise, non-noise, parameters, and so on to provide a broad range of values for these variables to be triggered. By design, +1 is applied to the first column. On the speech and alcohol database, dimensionality reduction audio has been used in comparison to the frequency.

Voiced Segment Dropout Optimizer Algorithm -VSDOA

The proposed method of Voiced Segment Dropout Optimizer Algorithm has a Speech emotion and drunkenness recognition test level of data analysis to explain details below, and system support.

```
Start: Speech emotion and drunkenness recognition data
testing level
LADS→Load Audio Data Set
SRT→Speech Recognition Test
BrAC-Breath Alcohol Concentration
    HC→Human Condition to SRT
        HiLD→High Level Drunkenness
    Nd→Non- drunkenness
        Non-→Normal
DRE→Drug Recognition Expert
        DRE -→SRT
        Drug Recognition Testing:-→Human Morning Time
        (HMT) and Empty
        Stomach (ES)
If  DRE > value (VSDOA)
    DRE -→ Drug Recognition Testing VSDOA
                If else DRE >= BrAC
        DRE to Active VSDOA
    To find the drunkenness
    DRE <=BrAC
    Health is good level
    Non-drunkenness
End
```

While the defendant appears to be inebriated, the blood alcohol concentration (BAC) is below the legal limit. After that, the officer would summon a drug recognition expert (DRE) officer.

Results and Discussion

This tool connects with the design python tool anaconda which is simulated using a program created in the python language. The various factors and values are considered in simulation. In these parameters, the Anaconda web mining tools create a dataset that describes information.

Figure 5 describes speech emotion and drunkenness recognition performance in speech emotion using the CNN-Voiced Segment Dropout Optimizer Algorithm (CNN-VSDOA). The suggested algorithm gives the superior performance with 92% compared to existing Support Vector Machine (SVM) algorithms with 75%, and K-Nearest Neighbor (KNN) with 78% and with Random Forest (RF) with 83%.

Figure 6 describes concerns about the sensitivity of the speech emotion recognition in the drunkenness. People suffering from emotion recognition can cause excessive drunkenness leading to sensitivity. Thus, the CNN-Voiced Segment Dropout Optimizer

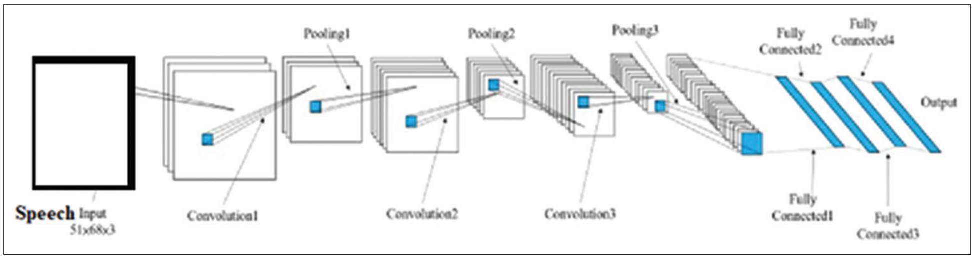

Figure 3: CNN classification based speech emotion and drunkenness recognition

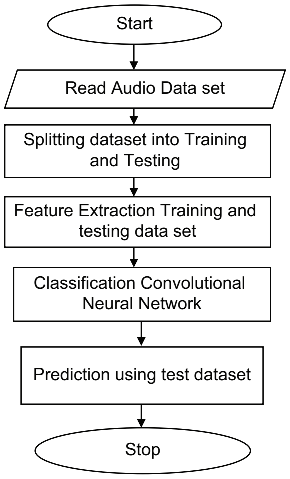
Start
Read Audio Data set
Splitting dataset into Training and Testing
Feature Extraction Training and testing data set
Classification Convolutional Neural Network
Prediction using test dataset
Stop

Figure 4: Process of CNN

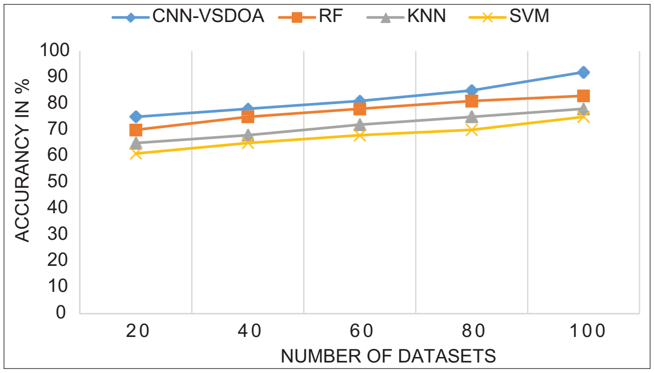

Figure 5: CNN-VSDOA algorithm classification performance

Algorithm (CNN-VSDOA), achieves the extraordinary result of 90% sensitivity compared to existing algorithms Support Vector Machine (SVM) achieves 75% sensitivity, K-Nearest Neighbor (KNN) achieves 78% sensitivity, and Random Forest achieves (RF) 80% sensitivity.

Table 1 show the details CAN-VSCODE error rate performance details to compare the existing algorithm RF, KNN, and SVM table details explained.

Figure 7 explains the error rate in the speech emotion recognition data prediction. CNN-Voiced Segment Dropout Optimizer Algorithm (CNN-VSDOA) gives a low error rate compared to existing algorithms Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). The suggested proposed CNN-VSDOA algorithm achieves a 32% of low error rate, and Support Vector Machine (SVM) gains 52% error rate, K-Nearest Neighbor (KNN) gains 60% of error

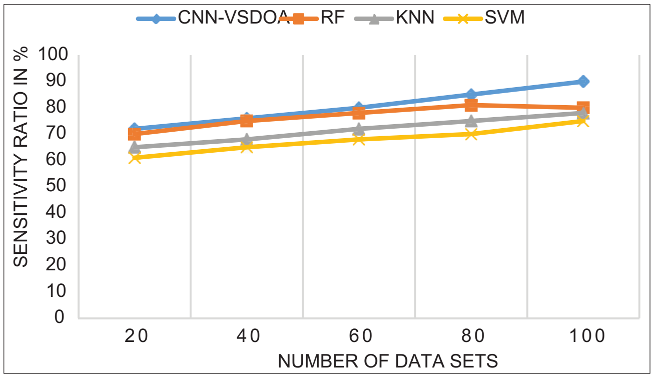

Figure 6: Sensitivity

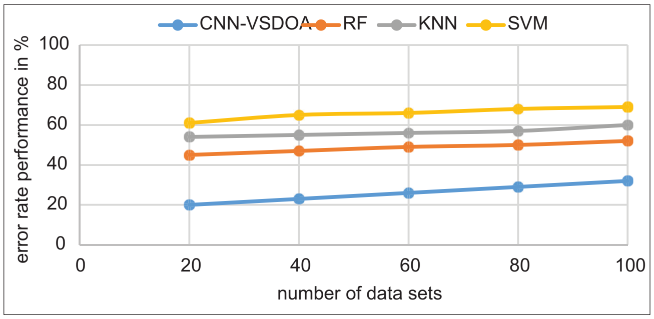

Figure 7: Speech emotion recognition error rate

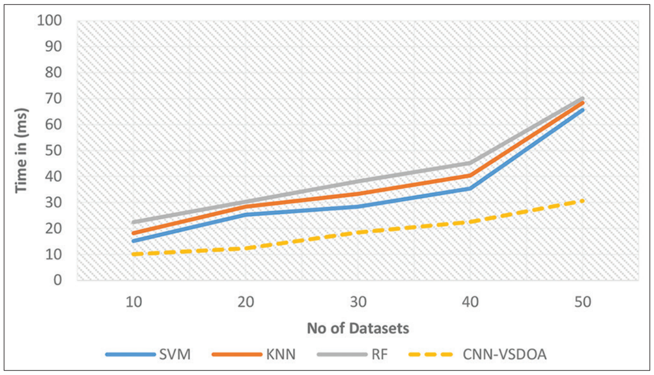

Figure 8: Time performance

rate and finally, and Random Forest (RF) achieves 69% of high error rate. Hence, compared to the existing algorithms, the proposed CNN-VSDOA algorithm gives a low error rate.

Figure 8 explains the time performance on CNN-Voiced Segment Dropout Optimizer Algorithm (CNN-VSDOA) which is the best method. This suggested algorithm achieves 30.6 ms

Table 1: Error rate performance

Number of data sets	*CNN-VSDOA*	*RF*	*KNN*	*SVM*
20	20	45	54	61
40	23	47	55	65
60	26	49	56	66
80	29	50	57	68
100	32	52	60	69

of less-time performance result than existing Support Vector Machine (SVM) algorithms with 65.6 ms, K-Nearest Neighbor (KNN) with 68.4 ms and Random Forest (RF). Compared to the current methods, the proposed CNN-VSDOA method proves to be the best.

Conclusion

The need for an easily accessible CNN network to the end helps to find the highest level of emotions and drinking. Has a very limited definition of a very limited knowledge and training, the able to capture the features in these states using clever pre-processing and network design. It is able to make some changes over the methods used to study our connection weights. One of the first could train and test our model using these databases into support on CNN support and with speech emotion database. It will get similar results in terms of accuracy this is a topic to be addressed in the future. The neural network results and machine learning applications are to be continued. The CNN-Voice segment Dropout Optimizer (CNN-VSDOA) is a speech and recognition algorithm that uses speech emotion to recognition. The suggested algorithm gives the superior performance with 78.7% compared to existing Support Vector Machine (SVM) algorithms with 72.6%, and K-Nearest Neighbor (KNN) with 73.4% and with random forest (RF) with 75.4%. On comparison, it is proved that CNN-VSDOA offers best accuracy in speech recognition system.

References

1. Li J, Deng L, Gong Y, Haeb-Umbach R. An overview of noise-robust automatic speech recognition. IEEE ACM Trans Audio Speech Lang Proc 2014;22:745-77.
2. Zhang Z, Ringeval F, Han J, Deng J, Marchi E, Schuller B. Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Auto Encoder with LSTM Neural Networks. Proceedings Interspeech; 2016. p. 3593-7.
3. Gao J, Du J, Kong C, Lu H, Chen E, Lee CH. An Experimental Study on Joint Modeling of Mixed-Bandwidth Data Via Deep Neural Networks for Robust Speech Recognition. Vancouver, Canada: International Joint Conference on Neural Network; 2016. p. 588-94.
4. Squartini S, Principi E, Rotili R, Piazza F. Environmental robust speech and speaker recognition through multi-channel histogram equalization. NeuroComputing 2012;78:111-20.
5. Zhang Z, Pinto J, Plahl C, Schuller B, Willett D. Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices. IEEE Trans Consum Electron 2014;60:525-33.
6. Han K, Wang Y, Wang D, Woods WS, Merks I, Zhang T. Learning spectral mapping for speech dereverberation and denoising. IEEE Trans Audio Speech Lang Process 2015;23:982-92.
7. Zhang Y, Chen G, Yu D, Yaco K, Khudanpur S, Glass J. Highway Long Short-term Memory RNNS for Distant Speech Recognition. Proceedings International Conference on Acoustics Speech and Signal Processing (ICASSP); 2016. p. 5755-9.
8. Rotili R, Principi E, Squartini S, Schuller B. A real-time speech enhancement framework in noisy and reverberated acoustic scenarios. Cognit Comput 2013;5:504-16.
9. Omar M, Pelecanos J. A Novel Approach to Detecting Non-Native Speakers and Their Native Language. Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP); 2010. p. 4398-401.
10. Compernolle DV, Smolders J, Jaspers P, Hellemans T. Speaker Clustering for Dialectic Robustness in Speaker Independent Recognition. Proceedings 2nd Eurospeech; 1991. p. 723-6.
11. Zoughi T, Homayounpour MM. Gender Aware Deep Boltzmann Machines for Phone Recognition. Proceedings International Joint Conference on Neural Networks (IJCNN); 2015. p. 1-5.
12. Vipperla R, Renals S, Frankel J. Ageing voices: The effect of changes in voice parameters on ASR performance. EURASIP J Audio Speech Music Proc 2010;2010:525783.
13. Potamianos A, Narayanan S. Robust recognition of children's speech. IEEE Trans Speech Audio Proc 2003;11:603-16.
14. Sinha R, Shahnawazuddin S, Karthik PS. Exploring the Role of Pitch-adaptive Cepstral Features in Context of Children's Mismatched ASR. Proceedings. International Conference on Signal Processing and Communications (SPCOM); 2016. p. 1-5.
15. Steidl S, Batliner A, Seppi D, Schuller B. On the impact of children's emotional speech on acoustic and language models. EURASIP J Audio Speech Music Proc 2010;2010:783954.