SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Kartikeya Srinivas Chintalapudi
Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Guntur, India
chkartikeya02@gmail.com

Irfan Ali Khan Patan
Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Guntur, India
irfanalikhanp9@gmail.com

Harsha Vardhan Sontineni
Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Guntur, India
shvardhan08@gmail.com

Venkata Saroj Kushwanth Muvvala
Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Guntur, India
mkushwanth8082@gmail.com

SuryaKanth V Gangashetty
Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Guntur, India
svg@kluniversity.in

Akhilesh Kumar Dubey
Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Guntur, India
dubey18oct@kluniversity.in

2023 International Conference on Computer Communication and Informatics (ICCCI) | 979-8-3503-4821-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCI56745.2023.10128612

Abstract— **Speech emotion recognition is the task of automatically detecting the emotional state of a speaker from their spoken words. It is a growing area of research that has applications in various fields such as human computer interaction, education, and psychology. There are several approaches to speech emotion recognition, including the use of machine learning algorithms, which can be trained on large datasets of annotated speech samples to recognize patterns associated with different emotions. Other approaches include the use of linguistic features, prosodic features, and physiological signals such as facial expressions and heart rate. One challenge in speech emotion recognition is the variability in the expression of emotions across individuals and cultural groups. Another challenge is the need to accurately identify the underlying emotion, as opposed to simply recognizing the presence or absence of an emotion. Overall, speech emotion recognition has the potential to improve communication between humans and machines, and to provide insights into the emotional states of individuals.**

Keywords- Deep Learning, Emotion Recognition, CNN

I. INTRODUCTION

The process of utilizing a neural network to understand speech is known as speech recognition in deep learning. Accusing numerous layers of neural networks, deep learning is the process of creating artificial intelligence. It is a use of machine learning, which tries to create computer systems that function much like people. Deep learning may assist machines with speech recognition, picture perception, and text reading. Deep learning's voice recognition technology has demonstrated appreciable advancements over existing speech recognition techniques, performing human-like. In order to do sequential tasks, it combines deep neural networks with lengthy short-term memory units. The first model to attain this level of quality on this kind of work is this one, which can give real-time transcription without any pre-processing. Software for speech recognition is always improving. People are beginning to utilize it on a daily basis. It is no longer unique. One of the most well-known uses of deep learning AI, which many people might not be aware of the voice recognition (which has shown an ability to learn tasks by analyzing data without any human guidance). Companies now struggle to locate enough candidates who can talk exactly and clearly into a microphone. Speech Recognition in Deep Learning will be able to address this issue since the procedure will be automated. From greater productivity to reduced stress levels, this automation may provide several advantages to both companies and employees.

Word error rate (WER) and general word error rate are two benchmarks where deep learning has been able to outperform human accuracy (GE). Deep learning models cost a lot to compute, though. Deep learning voice recognition methods now need to be more effective.

This paper is organized as follows: Section II shows a brief of the literary works on reviewing different articles. Section III describes the observation of the existing methodologies. Section IV consists of the conclusion and future content. Section V is about the references we have gone through for this review.

II. LITERATURE WORK

We conducted an survey on speech recognition based on various challenges and platforms. The research data has been collected from well-known platforms, i.e., IEEE Xplore, Science Direct, Springer, etc. We collected more than 10 journal papers.

Had collected the data from google warehouse which is Google audio Set. They had taken the dogs and cats sound and done some data modelling by deleting some unrelated part of the audio. They used Mel-frequency cepstral coefficient (MFCC) method to convert the audio signals into the vector. The had used deep learning, TensorFlow and Keres methods for the prediction. And they had visualized the data in box plot and find the accuracy of the prediction

979-8-3503-4821-7/23/$31.00 ©2023 IEEE

using F1-score formula. The result shows the highest score of F1- score was 73% and lowest score was 60% and on average. The system has 66% of F1-score.[1] In this paper they authors had provided the information about the main applications of the speech signal processing and provide the uses cases of the speech signal processing application. The three main applications are Speech recognition, Speech synthesis, Speech compression.

A. Speech Recognition

Its nothing but converting human voice into the computer understandable format. It can read any person's speech irrespective of the voice tone and voice base.

Use case:

- It was used in Health care at the medical documentation process.
- It was used in Military air crafts to operate the radio frequencies, weapon release.

B. Speech Synthesis

Its reverse of Speech recognition, that it was mainly used to convert the test into speech, in which speech recognition convert the speech into text. There are some certain software's to convert the speech into text.

Use Case:

- It was used in some accessibilities like headphone, AI assistance, etc.
- It was used to narrate some dialogues based on the user's specification.

C. Speech Compression

Its nothing but store the speech audio files into the databases after compressing the data. In simple we can say as Digital voice storage.[2] The identification of potential outliers in a database can be viewed as controlling the quality of recordings in a voice database. Generally speaking, based on the volume of additional details offered for the recordings, a speech. Many various applications, including speaker identification, emotion classification, voice disorder diagnosis, speaker identification, language recognition, and ambient sniffing, can be performed using databases. As a result, outliers in a particular voice database can vary between applications. But in this paper, our primary focus is on locating outliers that emerge during the remote data collection procedure.[3] Speech quality evaluation is required due to the rapidly growing use of speech processing algorithms in multi-media and telecommunications applications. Thus, accurate and trustworthy speech quality assessment is becoming essential for the end-user or customer satisfaction of the implemented speech processing systems. A mathematical comparison of the unprocessed and processed speech waveforms constitutes objective evaluation. By measuring the numerical "distance" between the original and processed signals, objective measurements can estimate quality.[4] According to the research we have described in this post, neural networks are the most frequently employed solution. Numerous research views can be taken into consideration by what we have offered in this post. In order to create an intelligent interface based on computer vision and work to receive user voice, we first want to use neural networks in our approach to automatic speech processing. This will enable intelligent interaction with users and establish natural and simple communication between the machine and human. Additionally, we intend to enhance our interface by combining speech with other human senses. Finally, we want to design a user interface that enables real-time decision-making.[5] The subjective quality ratings were derived using the ITU-TP.835 method. The relationships between a number of objective measurements and these three subjective rating systems are investigated in this study. Several innovative composite objective measures are also proposed, and the different objective measures are merged utilizing nonparametric and parametric regression analysis techniques. Most of the basic objective measurements can predict signal distortion and overall quality with accuracy, but not background distortion.[6] Using CNN As several neural networks, this data is sent into the initial layer of each network. Some discernible characteristics are sent to the second layer. Considered as an example is a signal that is composed of a two-dimensional array of pixels. It is a checkerboard with either a light or dark color in each square. CNN determines if a signal has a change in frequency or amplitude based on the pattern it observes.[7]

III. METHODOLOGIES(APPROACH)

The study of voice signals and signal processing techniques is known as speech processing. Speech processing can be viewed as a unique application of digital signal processing since voice signals are frequently processed in a digital representation. Speech processing involves the capture, alteration, storing, moving, and output of speech signals. There are two major types of techniques supervised and unsupervised. The supervised classification technique selects samples (training data) from the voice and visually classifies them in order to provide statistical measures that may be applied to the entire audio. Figure 1 below shows how our model's suggested methods might work. As shown in the flow chart, we imagine the data, do a binary classification in which we also receive the accuracy, and then construct evaluation matrices from that.

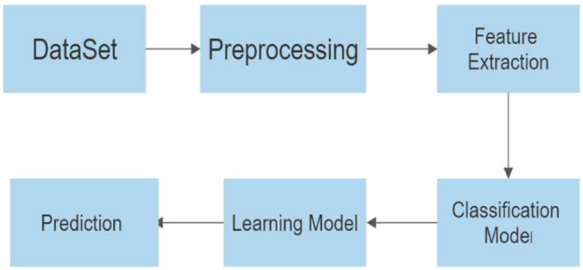

Figure 1. Methodologies

A. Gathering Dataset

As we observe from the above figure-1 we can know that the data set that we use comes from a variety of sources, and we assume that the best collection of data should be used rather than the most extensive one. After completing the whole literature study, we must now develop a technique to solve all the shortcomings in the preceding approaches. The methodologies we utilized in this study and CNN, ASR For deep learning will both produce accurate results. After

getting the findings, we used the Speech Recognition dataset that is accessible on Kaggle [8].

B. Data Pre-Processing

In this crucial phase, the unnecessary data are removed from the data collection, giving us the precise results, we need to predict the favorable outcomes. Even if we do have the biggest files to get the greatest data, categorization in our data collection is still made easier because of this. Data pre-processing generally transforms the data into a format that can be handled more easily and effectively in data mining, machine learning, and other data science tasks. These scans may be used to detect noise, one of several aberrations that audio might have. These artefacts can be removed using audio filtering methods. A geometric mean filter is applied to the input Audio to reduce the amount of noise.

C. UnderStanding and Visualizing the data

The user's analysis of the data set's outcomes is the next task on the agenda after the work has been finished through data pre-processing. Due to the extensive data set must be easily understood by everyone, so we will present the information so that a person can see it with just their eyes. Despite the fact that the brain processes 80 percent of information visually, every individual learns differently. While some people learn best through movement, others learn best through listening. But a significant portion of people—more specifically, 65 percent—learn best visually. Data visualization and online data visualization tools enable quick understanding of the provided information. Look at figure 2. Since the invention of spreadsheets, modern technology has transformed data into aesthetically pleasing, easily readable charts and graphs. A technique that has been employed for visually presenting data and that uses online data visualizations.

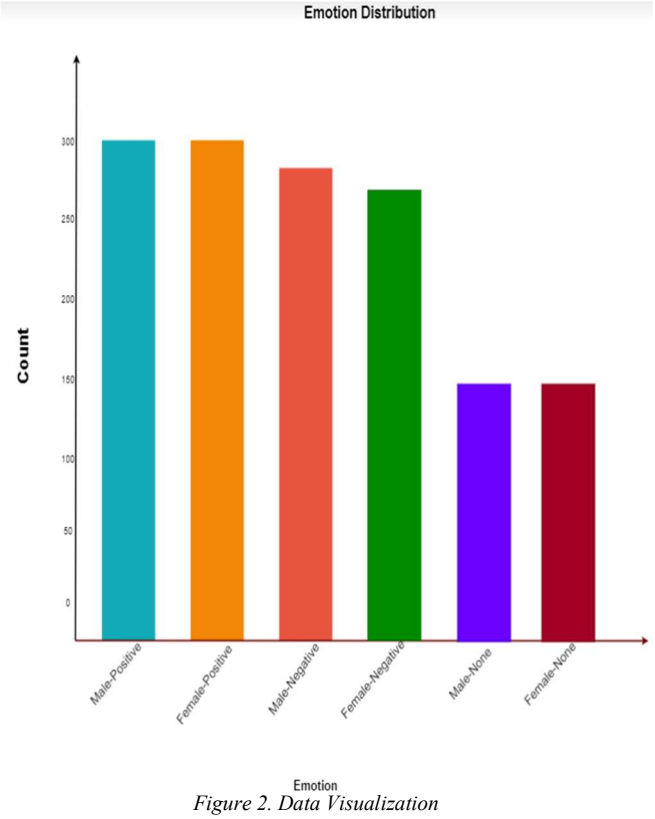

Figure 2. Data Visualization

D. Classification

ASR involves converting a particular auditory speech into text. We require a text transcript of an audio utterance in order to train and test an ASR system. No further processing was required to produce training or validation data for ASR since the collected data already contains this alignment by design. ASR is becoming used in multinational organization customer care divisions. Other organizations and some governmental institutions also use it. Simple ASR systems can recognize single-word submissions like yes-or-no questions and spoken numbers. When someone speaks with a strong accent or dialect, an ASR system may not always be able to recognize their input accurately. It also has significant issues when someone combines words from two different languages out of habit.

The Convolutional Neural Network (CNN) algorithm, which uses various modules for emotion detection and classifiers to distinguish between emotions including happiness, surprise, anger, neutrality, and sadness, is the basis for spoken emotion recognition. In order to categorize each word from our aggregated data set as a multi-class classification problem, Convolutional Neural Network (CNN) is used as an advanced deep neural network. With a completely unidentified speech sample, the suggested deep neural network produced a word classification accuracy score of 46.85%. Our data is trained and tested using CNN.

E. Finding Total Parameters

Figure 3 below shows the results of the summary () technique. Since each row represents a layer, we can refer to these layers by their row names alone without any further explanation confusion. The layers that were added to the model in the last code sample are all visible in the figure below.

```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_18 (Conv1D)          (None, 40, 64)            384

 activation_24 (Activation)  (None, 40, 64)            0

 max_pooling1d_12 (MaxPoolin  (None, 10, 64)           0
 g1D)

 conv1d_19 (Conv1D)          (None, 10, 128)           41088

 activation_25 (Activation)  (None, 10, 128)           0

 max_pooling1d_13 (MaxPoolin  (None, 2, 128)           0
 g1D)

 conv1d_20 (Conv1D)          (None, 2, 256)            164096

 activation_26 (Activation)  (None, 2, 256)            0

 dropout_11 (Dropout)        (None, 2, 256)            0

 flatten_6 (Flatten)         (None, 512)               0

 dense_6 (Dense)             (None, 8)                 4104

 activation_27 (Activation)  (None, 8)                 0

=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
```

Figure 3. Finding Total Params

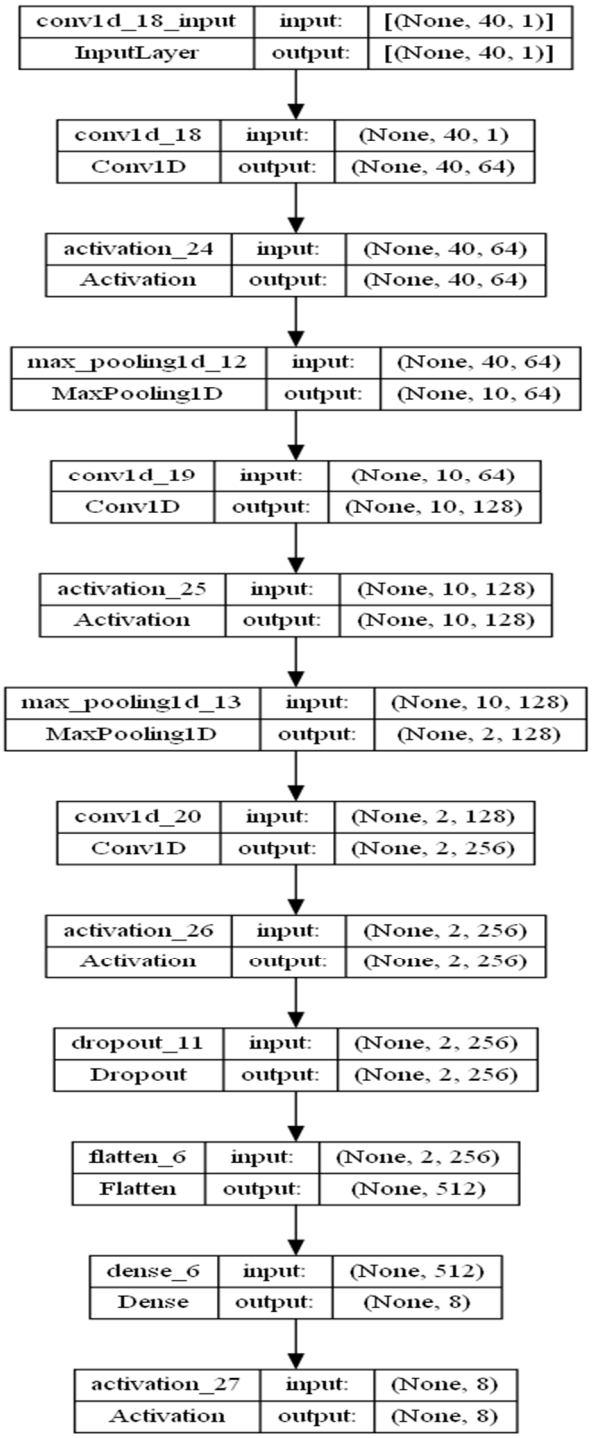

Figure 4. Finding Total Parameters by using plot

Parameters are often weights that are trained weights. They are weight matrices that are altered during the back-propagation procedure, increasing the model's ability to predict the future. Depending on the training method you employ, particularly the optimization approach, they change their values.

IV. RESULT AND FINAL ANALYSIS

We experimented a Dataset file to get its results by plotting the waveform.

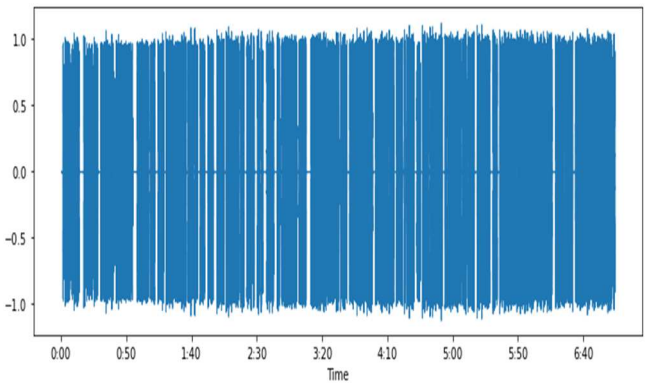

Figure 5. Audio Waveform

The training and testing loss on our dataset is depicted in the image below. According to the graph, "training and testing" errors decrease as the number of training model epochs rises.

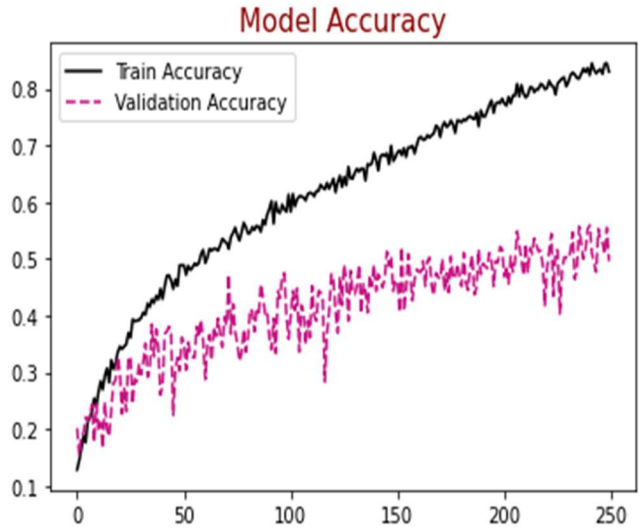

Figure 6. Model Accuracy

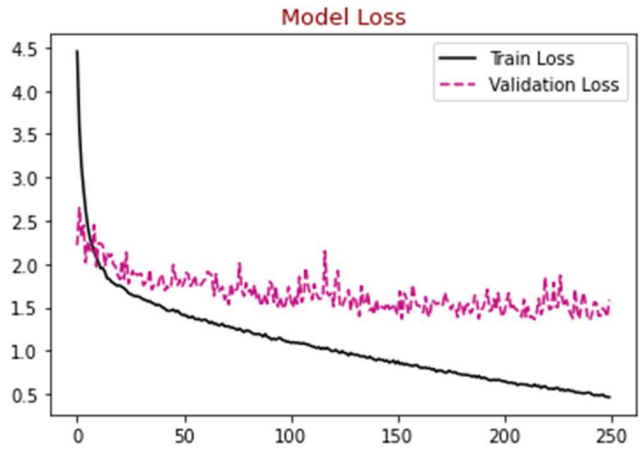

Figure 7. Model Loss

Finally, we get a conclusion on the speech emotion recognition performance assessment, performance metrics can be used to evaluate a model's performance by calculating their metrics.

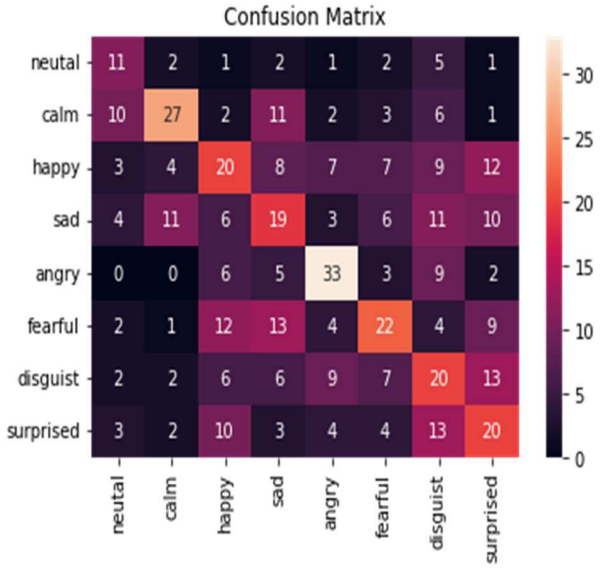

Figure 8. A Study on Various Existing State of Art Methodologies of Speech Recognition

Table I. It Provides The Relevant Classification System Efforts

Author & Year	Advantages	Disadvantages
Y. Hu,2019	Naturally discriminative, easily integrates with statistical methods & modular design.	It cannot predict the background distortion on speech sample
A.H. Poorjam, M.S. Hussain Jan 18, 2019	The simple and efficient nature of this approach significantly reduces the work required to remove noise from the noisy sample. The robust center and scatter of the data are likewise calculated using the deterministic MCD approach.	When there are samples of mixed noise rather than assuming all the samples in the database are the same, this technique might not be as accurate in all the datasets.
P. Kumawat,2019	This method is good for detecting noise signals on a large scale of data in form of a 2D CNN's and outputs the percentage of the parameters detecting the noisy speech segments.	Since it is challenging to directly evaluate speech data that has been distorted, the suggested deep learning model may occasionally produce false results without the aid of the ASR model

V. CONCULSION AND FUTURE SCOPE

In the future, speech recognition will no longer be a human assisted task. Speech recognition will become an automated process. In the future, voice recognition will likely be carried out automatically. It would not be surprising if one day we're all walking around with earpieces like C-3PO - except they'll be listening to everything we say. The scope for speech recognition in deep learning is being researched and developed. There are a number of promising advancements to be made in the realm of speech recognition. One of the most interesting developments is the use of neural networks to analyze sound patterns, which could in turn produce better artificial intelligence algorithms. Speech recognition is a subset of deep learning and has been studied for over fifty years. Nowadays, speech recognition systems are more accurate than ever. The future scope for speech recognition has never been more promising. Deep learning techniques have made it possible to train speech recognition models on larger datasets and with more computing power to help these algorithms process their data better.

VI. REFERENCES

[1] P. Lakkhanawannakun, Speech Recognition using Deep Learning, June2019
[2] Raghib, E.Sharma, T.Ahmad, F.Alam, Emotion Analysis and Speech Signal Processing, June 2018
[3] A.H. Poorjam, Quality Control in Remote Speech Data Collection, Jan 18, 2019
[4] Philipos C. Loizou Speech Quality Assessment, Vol 346
[5] S. Benkerzaz, Y. Elmir, A. Dennai, A Study on Automatic Speech Recognition, Aug2019.
[6] Y. Hu, Evaluation of Objective Quality Measures for Speech Enhancement, Feb 2008.
[7] N. Dimmita P. Siddaiah, Speech Recognition Using Convolutional Neural Network, Sep 2019
[8] https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
[9] M.S. Hossain and G. Muhammad, ''Emotion recognition using deep learning approach from audio–visual emotional big data,'' Inf. Fusion, vol. 49, pp. 69–78, Sep. 2019.
[10] M. Chen, P. Zhou, and G. Fortino, ''Emotion communication system,'' IEEE Access, vol. 5, pp. 326– 337, 2016.
[11] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, ''Speech emotion recognition,'' in Proc. Int. Conf. Adv. Electron. Comput. Commun. (ICAECC), Oct. 2014, pp. 1–4.
[12] K. R. Scherer, ''What are emotions? And how can they be measured?'' Social Sci. Inf., vol. 44, no. 4, pp. 695– 729, 2005.
[13] S. G. Koolagudi and K. S. Rao, ''Emotion recognition from speech: A review,'' Int. J. speech Technol., vol. 15, no. 2, pp. 99–117, 2012.
[14] J. Schmidhuber, ''Deep learning in neural networks: An overview,'' Neural Netw., vol. 61, pp. 85–117, Jan. 2015.
[15] S. Demircan and H. Kahramanlı, ''Feature extraction from speech data for emotion recognition,'' J. Adv. Comput. Netw., vol. 2, no. 1, pp. 28–30, 2014.