Speech Recognition using Convolution Deep Neural Networks

Speech Recognition using Convolution Deep Neural Networks

Ayad Alsobhani[1]*, Hanaa M A ALabboodi2[2,] Haider Mahdi 3[3]

[1] Faculty of Engineering, Department of Electricity, University of Babylon, Iraq.
[2]Assint prof Dr, in Faculty of Engineering, Department of Electricity, University of Babylon, Iraq.
[3]Assint prof Dr, in Faculty of Engineering, Department of Electricity, University of Babylon, Iraq.

*Corresponding author's e-mail: ayadsobhan1992@gmail.com

Abstract. The use of a speech recognition model has become extremely important. Speech control has become an important type; Our project worked on designing a word-tracking model by applying speech recognition features with deep convolutional neuro-learning. Six control words are used (start, stop, forward, backward, right, left). Words from people of different ages. Two equal parts, men and women, contribute to our speech dataset which is used to train and test proposed deep neural networks. Collect data in different places in the street, park, laboratory and market. Words ranged in length from 1 to 1.30 seconds for thirty people. Convolutional Neural Network (CNN) is applied as advanced deep neural networks to classify each word from our pooled data set as a multi-class classification task. The proposed deep neural network returned 97.06% as word classification accuracy with a completely unknown speech sample. CNN is used to train and test our data. Our work has been distinguished from many other papers that often use ready-made and fairly consistent data of the isolated word type. While our data are collected in different noisy environments under different conditions and from two types of speech, isolated word and continuous word.

1. Introduction

Automatic speech recognition is the method of translating a speech signal into a series of words using a computer program and its algorithms. The main goal of speech recognition is to allow machines to recognize sounds and act on them. The ability of a computer to identify "receive and interpret" speech and translate it into readable form or text is known as automatic speech recognition. Automatic speech recognition is the ability of a computer to understand speech as well as execute an action based on the human's instructions [1]. The phoneme has three parts of processing, identifying the spoken words and knowing the speaker and the third one is the emotional recognition. The words are displayed either in writing or by devices. They are read as a specific command as what have been done in our work, the ability to recognize the speech signal can be divided into three categories based on the type of speech signal itself and its length [2] as shown in figure (1)

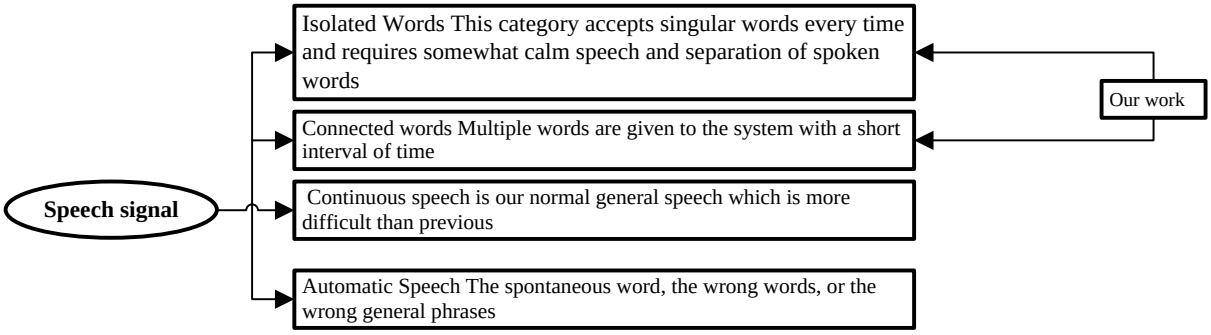

Figure 1. The speech types.

Sound recognition has a significant role in developing an automobile control circuit or robotic applications in household protection devices. Sound waves can be defined as a longitudinal wave that travels through an adiabatic compression and decompression phase in most cases. Longitudinal waves vibrate in the same direction as they fly. In sound processing, spectrograms are two-dimensional patterns that show frequency on the vertical axis and time on the horizontal axis, representing signal energy. In general, the movement of air in the vocal tract produces the consonant, which gives rise to the various sounds. The general path of completion of speech recognition is explained in figure (2).

Figure 2. The general path to completion of speech recognition.

The general path to completion of speech recognition theory is exhibiting, ADC (analog to digital converter) (spectral shaping) Picking the sound wave and converting it into a digital form then pre-emphasis filtering following by feature extraction stage. The extracted features will send to classification as a targeted stage. Some techniques involve parameter transformation, which entails transforming obtained features into signal parameters using a separation and concatenation procedure. Conversion of parameters in signal observation vectors is a part of statistical modeling [3]. The sound recognition plays a significant role in access control and security systems [4]. A sound wave is a characteristic sinusoidal vibration of loud tones that vibrate quickly and with a higher frequency than low tones The vibrating sound energy is converted into electrical energy by the microphone. The general shape of the sound wave gives an impression of the amount of energy with the amplitude of the signal. The nature of the sound wave is variable frequencies in its being adherent to each other, the frequency components that make up the sound wave are analyzed by the FFT, which highlights it with a diagram called the spectrum diagram. [5]. Sinusoidal waves in the air are caused by spoken voices. Higher pitches vibrate at a higher frequency than lower pitches, so they vibrate faster. A microphone can sense these sounds, which are then converted from acoustic energy carried in the sound wave to electrical energy and captured as an audio signal. The amplitude of a speech signal indicates how much acoustic energy is present in the sound and therefore how noisy it is. Around the same time, our voice is made up of a variety of frequencies. The final signal is the product of adding all of those frequencies together. The component frequencies were used as features to better interpret the signal. To decompose the signal into these components, Fourier transform has been applied. For this task, FFT algorithm (Fast Fourier Transform) is generally available. The sound transforms in to a spectrogram by using this splitting technique. signal is divide in to time frames to generate a spectrogram. Then, by using FFT, each frame is splitting each frame into frequency components. A vector of amplitudes at each frequency has been used to describe each time frame. The spectrogram can be time aligned with the original audio signal to get a graphic representation of the sound components [5].

With regard to our work, six words from thirty people in different places have been collected. Our contributed speech samples have different ages and genders. The length of recorded words was very different, which made us face many problems. Several programs have been used to deal with these differences (Audacity and Adobe Audition). The speech features have been extracted and trained by using Convolution Neural Network (CNN). CNN as a most advanced deep learning method has been applied and it performance for speech signal recognition as a multiclass classification process has been investigated. Different types of CNN learnable parameters have been tested and updated to find the best CNN structure that able to solve our multiclass classification problem. The motivation of using deep learning model is to find the best model that suitable for our collected data conditions to complete the process of controlling. As mentioned before, the recording data has many issues in terms of phonation, place of pronouncement of words, and the amount of noise in it as well as the background and recording devices noises. In general, the data have been collected in such circumstances due to our work that aims to control the movement of the machine through speech signal order for real time applications.

What distinguishes our work from other works in this field is the data that we collected ourselves in different recording areas with high, medium and little noise, for example in the market, home, garden, and laboratory. As for the other research, it mostly uses ready data, simple and clean, recorded in noise-free areas and secondly we used two types of words isolated words such as (stop and start) And connected words (figure 1) such as (backward) and other research have used separate words only for the most part. The third advantage of our work was that we were able to obtain good efficiency compared to other research by working the algorithm that deals with various data and filtering them in a way that enables the classifier to distinguish words with high accuracy [6,7,8].

2. Related work

In the fifties of the last century, research began on speech recognition from Carnegie University for digitally isolated recognition systems for ten Bell Labels, and the evolving action took place in the 1980s. [9]. In 2017, Vishal Passricha et, introduced a composite method with a non-homogeneous classification CNN and SVM where a layer was substituted softmax by SVM [10]. Y. Yorozu, M. Hirano, forward a model of very deep convolutional neural networks devoid of connected layers and showed that VDCNN worked better than CNN, when same has been experimented with MGB-3 [8]. In 2020, Yang Xuebin et, speech recognition system designed for a group of words and used three methods of classification, including CNN, and obtained results of about 92.88% [11].

3. Data set

In our project. six control words (start, left, right, backward, forward and stop) for thirty people are collected. Each person uttered these six words once and in English languages. The people were half men and the other half were women and their ages started from the age of 16 years and over. The six words were recorded in different locations. In the laboratory, on the street, in the garden, in the market, in places free from noise and in other noisy areas. The recorded words are different in length of rely on the word itself. Also, some words differ in length from one person to another depend on the speaker himself and his pronounce ways. These conditions made the work more complicate especially for the training and classification process. Audacity and Adobe Audition were used to clean and pre-process the input data to prepare them for classification stage. The input words have length in ranges from (1s to 1.35 s). The data have been classified into seven classes according to the words that required to conduct to the control circuit. The recorded words classes are forward class, backward class, start class, stop class, left class, right class and unknown class. The unknown class includes the words intended to complete the training process and these words are (yes, no, are, is, friend, hello, he and she). Figure (3) show three audios with spectrograms of them from our data

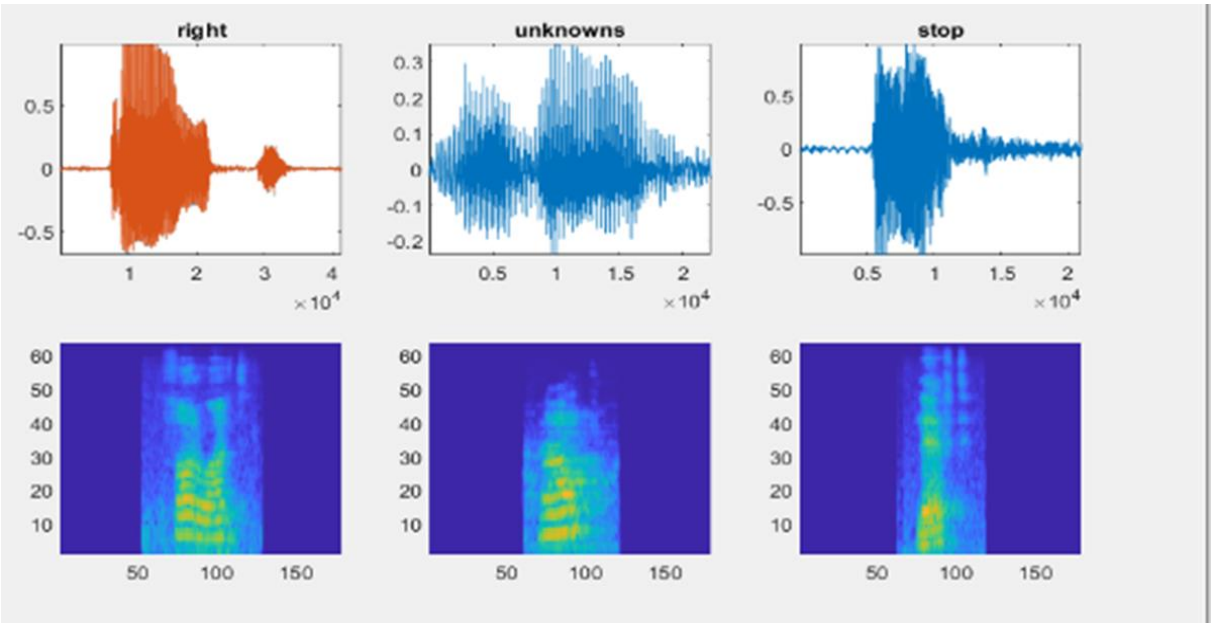

Figure 3. Three acoustics with their spectral diagrams from our data.

4. Convolutional neural network (CNN)

Convolutional networks were the beginnings Hubel and Wiesel who found that a single network architecture could reduce complexity in the feedback neural network when studying neurons used for local sensitivity and orientation selection in the cerebral cortex of cats.

CNN is often used with image processing that requires a two-dimensional matrix containing features and may be three-dimensional, the pixel values are in the horizontal and vertical coordinate indicators. CNN is a neural network model. Its architecture has three main ideas, as explained in figure (4). Each one of them has the susceptibility to improve speech recognition performance [12]. figure (5) explain the CNN layers and how CNN work.

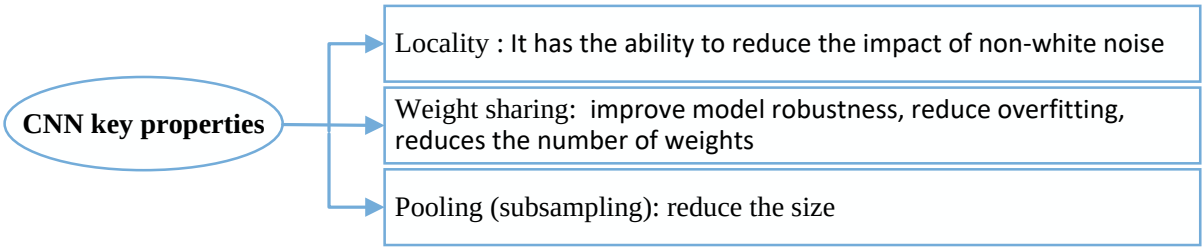

Figure 4. Architecture of CNN properties

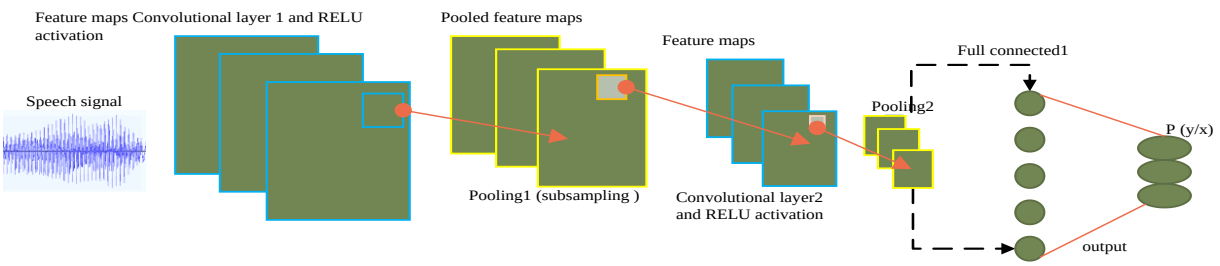

Figure 5. Architecture of CNN layers

CNN has a filter that shifts over the image to produced feature map at convolution layers, through this window or filter, the weights of the network can identify the different features of the incoming image. The activation function decides if a particular feature is present at a particular location in the image. Usually uses a lot of filters over the image to find the necessary features [13]. CNN is often called the local network because the individual units computed in a specific location of the window depend on the local area that the window is currently looking at. Convolutional architecture is coordinated by three main layers arranged in the forward feed structure. The convolutional layer for feature extraction, sub-sampling layers, the aggregation(pooling) layer, to reduce the dimensions of the input data and the output which a fully-connected layer for final classes prediction [14]. linear filter and a nonlinear activation function, One of the most important elements [15]. In a convolutional layer, each plane is connected to one or more feature maps of the preceding layer [16] .an activation function is applied on to the result obtain the plane's output. The plane output is a 2-D matrix called a feature map; this name arises because each convolution output indicates the presence of a visual feature at a given pixel location [16]. A convolution layer produces one or more feature maps. Each feature map is then connected to exactly one plane in the next sub-sampling(pooling) layer [15].
Sharing of weights and location are essential to the properties of the pooling, feature values computed at different locations are grouped together and represented by a single value in order to minimize differences in the extracted features along the frequency dimension when the input patterns are shifted. This is important when dealing with the small frequency shifts common in speech resulting from different path lengths vocal. CNN used activation faction [16] as figure (6) and table (1) explain some properties of CNN layers

After converting the audio into a Spectrograms as an image, let's say we have an image of dimensions N = 6 * 6, and the filter was F = 3 * 3 kernel filter, in example below and after making the convolution (*) the result is 4 * 4 according to the formula Out =N-F+1.
N=6, F=3 with padding P=0 and Strided convolution S=1, the out =4=n-f+1

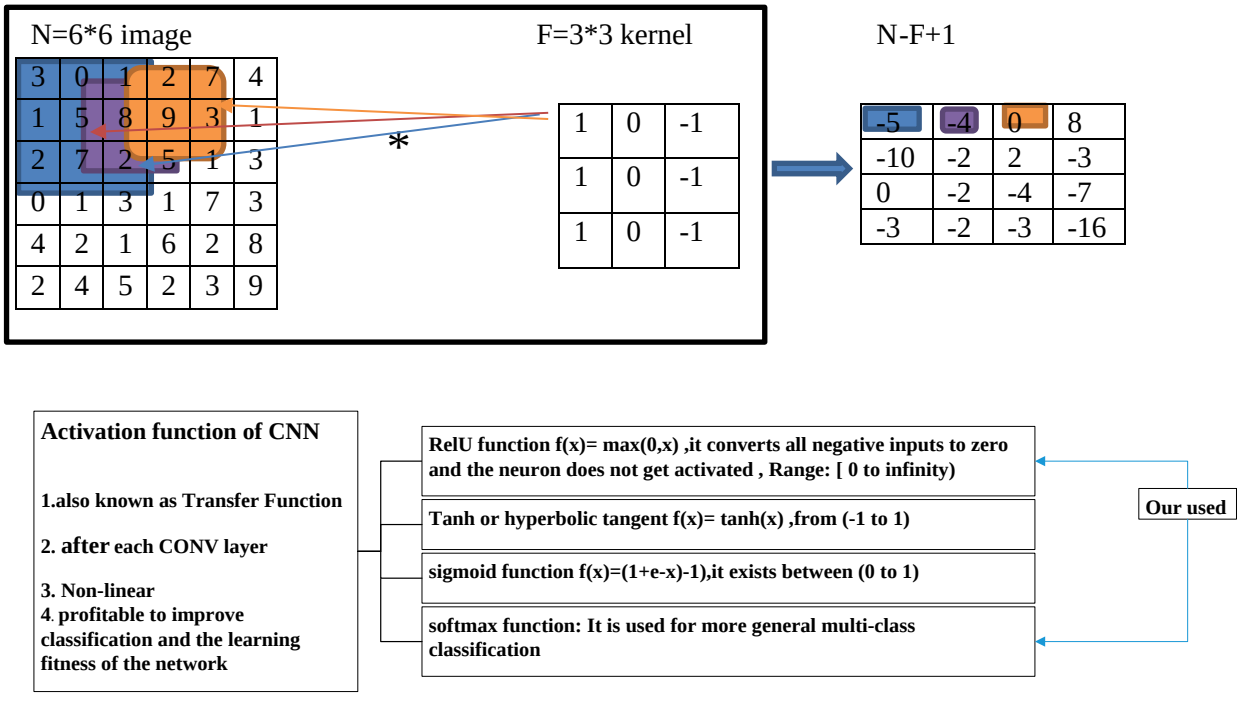

Figure 6. Activation function and it work [16]

Table 1. Illustrates the properties of CNN layers [13,14,15].

Convolution layer	**Pooling layer**	**Full connected layers**
Filters are included to find features of an image	Reduce dimensionality	Aggregate information from final feature
The filter consists of small kernels (number of kernels)	Maximum or average area is extracted	General final classification
One bias per filters		
For every value of feature map must apply activation function	Sliding window approach	Parameters full connected, (number of nodes, activation function; usually changes depending on role of layers. RELU used for aggregating information, and SOFTMAX for producing final multi-classification)
parameters of CONV layers ,(size of kernels ,activation function ,stride, padding and regularization type and value)	Parameters of pooling, (stride and size of window). [16]	

5. Experiments results and discussion

Supervised learning of with suggested deep neural model have been used to train and test the convolution neural networks CNN. The suggested CNN structure has 13 layers. In CNN, sounds file has been entered directly to the designed network structure that contained multilevel learning procedure to achieve the model multi-classification task to classify six labels for words (start, stop, right, left, forward, backward). The size of data has been enlarged by increasing the number of training data through a process augmentation. The number of augmentations are three for every image(spectrogram) of audio. Auditory Spectrograms in frequency sampling 48K has been computed with segmentation duration equal 1.8 and frame duration 0.02Ms, FFT length 1024 and number of band 64 over melspectum have been applied. Table (2) explain all details of our cnn layers

Table 2. The layers of CNN for our proposed model

Cnn layers (our work)	Description
1st layer	Image input layer to adjust image dimensions by (number of hopes and number of bands)
2nd layer	To add filter size of pixels (padding equal 3)
3rd layer	To balance the data and put mean and standard deviation equal to zero and do smoother gradients ,faster training and better generalization accuracy [normalization] with RelU layer
4th layer	Add pooling to reduce size with 3 stride and 2 padding
5th layer	2*number of filter(numF) for padding [number of filter =10]
6th layer	Batch normalization layer with RelU layer and maxpooling2dlayer with stride 3 and padding 2
7th layer	Convolution 2D Layer (3,4*numF, ‘Padding’, ‘same'), numf=10
8th layer	Batch Normalization Layer (RelU layer)
9th layer	Max Pooling 2D Layer([timePoolSize,1])
10th layer	Dropout Layer (dropout rob), dropout, prevent overfitting
11th layer	Fully Connected Layer (Number of Classes)
12th layer	Softmax Layer , compute probability of each label
13th layer	Classification Layer, classify based on softmax, cost will be x-entropy

Figure (7) explains CNN training and validation stages and it is showing clearly the all training parameters. Our speech samples have been trained and validated by applying suggested deep neural network. The model learning process started from acquiring the corresponding learning patterns in the input speech spectrum. The learning features go through our suggested network layers and training parameters have been updated through learning process. Figure (7) shows that the words classification accuracy for both training and validation data have been enhanced by increasing the number of iterations to reach the better accuracy with 275 iterations. On the other hand, the misclassification data are decreased throughout the training and validation progress. The model consists of 13 layers, the first layers has 179 neurons that corresponded to the speech samples in the input matrix and the final layer which is the classification layer has 7 neurons that refer to our seven classes (targeted labels). All the model details explained clearly in figure (7). While figure (8) shows the confusion matrix of validation results, which shows the location of the error and only one error is seen from our validation data, where the error in the word 'start' appeared in the unknown class. Figure (9) shows the training label distribution that compared with targeted label distributions.

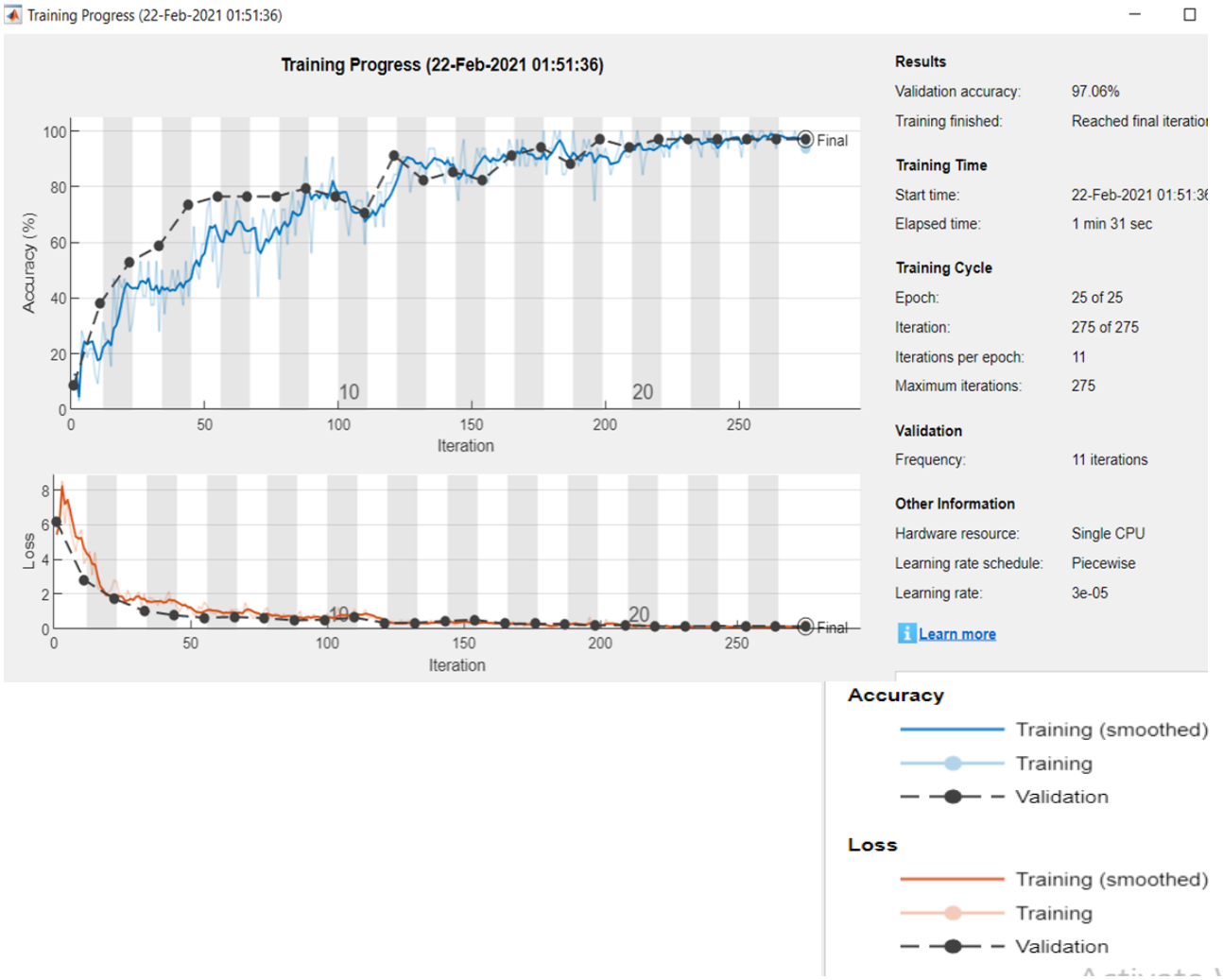

Figure 7. training and validation results of CNN method

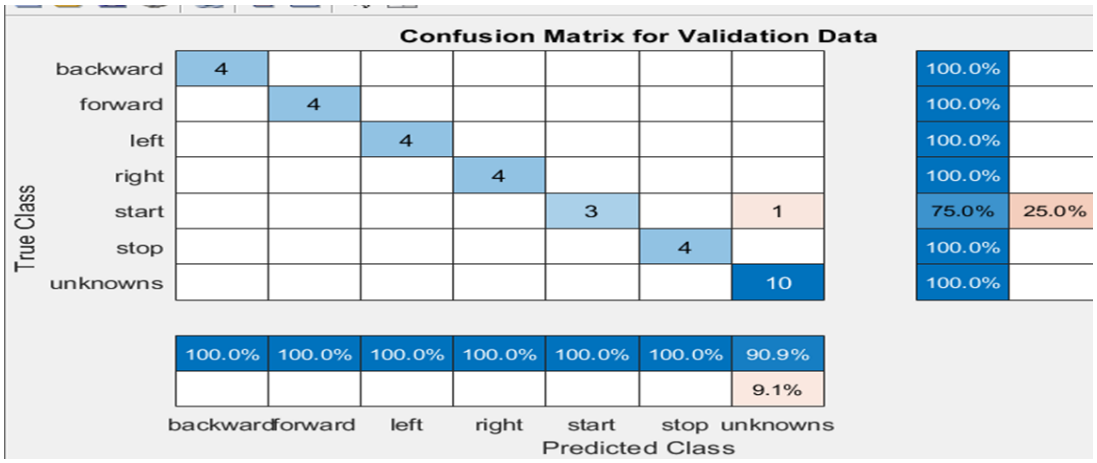

Figure 8. Confusion Matrix for Validation Data of CNN

Figure 9. Training label distribution

6. Conclusion

Our data consists of six words for 25 persons, each word represents one label from our six labels for words (start, stop, right, left, forward, backward) that applied for supervised learning of our suggested deep neural model. The suggested CNN structure has 13 layers. In CNN, sounds file has been entered directly to the designed network structure that contained multilevel learning procedure to achieve the model multi-classification task. The experiment results show that the deep neural networks have ability to solve speech recognition challenges that related with noisy and low resolution sound signals. CNN returned an acceptable performance for word recognition task using our noisy speech signal. The model performance could be higher and returned better classification accuracy when enough training data for deep learning model is available. For this purpose, the data has been increased and another class has been added to our six classes of word to become 7 classes. The seventh class is unknown class which have about 50 words, now, the total amount of data and words classes contribute to

enhance the convolution neural network in word speech recognition task. And the model classification accuracy increased to reach 97. 06%. CNN requires a large amount of data for the purpose of learning, the more data is given, CNN will return a better and more accurate classification accuracy. Our classification results were based on the percentage of data division, and the training rate was 85% and validation was 15% of input data.

7. References

[1] Fadlilah A. F., Djamal E. C. (2019) Speaker and speech recognition using hierarchy support vector machine and backpropagation. In: 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). IEEE, p. 404-409

[2] Shaikh Naziya S., Deshmukh R. R. (2016) Speech recognition system—a review. IOSR J. Comput. Eng, 8.4: 3-8. [3] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 J. Cryst. Growth 222 701

[3] Kesarkar M. P., Rao, P. (2003) Feature extraction for speech recognition. Electronic Systems, EE. Dept., IIT Bombay.

[4] Alías F., Socoró J. C., Sevillano X. (2016) A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Applied Sciences, 6(5), 143.

[5] Student P. G. (2016) Feature Selection and Extraction of Audio Signal. algorithms, 5(3).

[6] Abdel-Hamid O., Mohamed A. R., Jiang, H., Deng L., Penn, G., Yu D. (2014) Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

[7] Li X., Zhou Z. (2017) Speech Command Recognition with Convolutional Neural Network. CS229 Stanford education

[8] Poudel S., Anuradha, R. (2020) Speech Command Recognition using Artificial Neural Networks. JOIV: International Journal on Informatics Visualization, 4(2), 73-75.

[9] Song Z. (2020) English speech recognition based on deep learning with multiple features. Computing, 102(3), 663-682

[10] Passricha V., Aggarwal R. K. (2019) Convolutional support vector machines for speech recognition. International Journal of Speech Technology, 22(3), 601-609

[11] Yang X., Yu H., Jia L. (2020) Speech recognition of command words based on convolutional neural network. In: 2020 International Conference on Computer Information and Big Data Applications (CIBDA) (pp. 465-469).

[12] Saitoh T., Zhou Z., Zhao, G., Pietikäinen M. (2016) Concatenated frame image based cnn for visual speech recognition. In: Asian Conference on Computer Vision. p. 277-289.15- Phung, Son Lam, and Abdesselam Bouzerdoum. "Visual and Audio Signal Processing Lab University of Wollongong", 2009.

[13] Nanni L., Costa Y. M., Aguiar R. L., Mangolin, R. B., Brahnam S., Silla C. N. (2020) Ensemble of convolutional neural networks to improve animal audio classification. EURASIP Journal on Audio, Speech, and Music Processing, 1-14.

[14] Patel S. (2020) A Comprehensive Analysis of Convolutional Neural Network Models. International Journal of Advanced Science and Technology, 29(4), 771-777.

[15] Kubanek M., Bobulski J., Kulawik, J. (2019) A method of speech coding for speech recognition using a convolutional neural network. Symmetry, 11(9), 1185.

[16] Nwankpa C., Ijomah W., Gachagan, A., Marshall, S. (2018) Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378.

[17] Phung S. L., Bouzerdoum A. (2009) Visual and Audio Signal Processing Lab University of Wollongong