Advances in Artificial Intelligence Research (AAIR) **Vol. 1 (No. 1), pp. 6-12, 2021**

www.dergipark.com/aair/ **Published online: Jan 15, 2021**

Speech recognition based on Convolutional neural networks and MFCC algorithm

Arzo Mahmood [1], Utku Köse [2*]

[1] Computer engineering department, Suleyman Demirel University; [0000-0002-9652-6415]
[2] Computer engineering department, Suleyman Demirel University; [0000-0001-8101-4970]

Abstract

In this paper, an automatic speech recognition system based on Convolutional neural networks and MFCC has been proposed, we have been investigated some deep models' architecture with various hyperparameters options such as Dropout rate and Learning rate. The dataset used in this paper collected from Kaggle TensorFlow Speech Recognition Challenge. Each audio file in the dataset contain one word with one second length the total words in the dataset is 30 categories with one category for background noise. The dataset contains 64,721 files has been separated into 51,088 for the training set, 6,798 for the validation set and 6,835 for the testing set. We have evaluated 3 models with different hyperparameters configuration in order to choose the best model with higher accuracy. The highest accuracy achieved is 88.21%.

Keywords: *Convolutional neural network; FFT; MFCC; Speech recognition; Features extraction.*

1. Introduction

The automatic speech recognition is the process of recognizing the spoken words by human to a readable machine format like text, or command. This technology lets users to control their digital devices using their voice instead of using another input tool such as keyboard or mouse. The field of the speech recognition has been developed during the past decades because of its wide range of applications in many fields of life [1], the main applications of the technology are the call centers, dictation solutions and assistive applications and mobile and embedded devices.

Most of automatic speech recognition systems extract the features from the acoustic signal instead of using the full speech signal due to the large variations in the speech these variations include the pitch and speed of voice background noise, emotions and expressions. Interact with computer speech, keyboard, mouse, touchpad, etc. is helpful for people who have difficulty dealing with normal interface like. Speech recognition [2], is the process of converting a speech signal into words or phonemes. The main purpose of ASR is to overcome all difficulties encountered in the field of speech recognition such as different speech patterns, fuzzy environmental noise and the like [3, 4]. Modern speech recognition systems use deep learning techniques [5,6]. They are used to represent features and to model the language [7,8]. Better results are achieved by the recently popular convolution neural networks [9,10]. New frameworks are being created for the fastest open-source deep learning speech recognition framework [11]. Work is underway to improve their efficiency compared to existing ones such as ESPNet, Kaldi, and OpenSeq2Seq. There are also solutions in which the architecture of the Recurrent Neural Network (RNN) is used to obtain lightweight and high accuracy models that can run locally [12]. This will allow the use in real time. There are works on algorithms implemented in the frequency domain that allow speech analysis by identifying the intended fundamental frequency of the human voice, even in the presence of subharmonics [13]. The popular algorithm for features extraction used is the Mel-Frequency Cepstral Coefficients (MFCCs) [14] which is inspired from human ear and working on the Mel-space frequencies.

In this paper, we present a speech recognition system based on Convolutional neural network and Mel-frequency cepstral coefficients. The MFCC algorithm used to extract the unique features of each speech signal and then we used these features to train the CNN algorithm which also do features extraction [15]. The Purpose of using MFCC algorithm is to reduce the complexity of the model and achieve higher recognition accuracy. The dataset used in this study is the Kaggle TensorFlow Speech Recognition Challenge [16], the dataset contains a command words and non-command words spoken by various subjects each speech signal is one second length. The total dataset that we used contains 64,721 audio files separated into 51,088 training set, 6,798 validation set and 6,835 testing set. The total command words files are 23,682 and non-command words are 41,039. We evaluated three deep models with various hyperparameters options.

*Corresponding author
E-mail address: utkukose@sdu.edu.tr, janamohammed22@gmail.com

Received: 12/July/2020; Received in revised form: 25/Nov/2020; Accepted: 25/Nov/2020.

2. Methods

2.1. MFCCs (Mel frequency cepstral coefficients)

MFCCs is the popular and most used Feature extraction algorithm in the field of automatic speech recognition. The algorithm proposed in the 1980's by Davis and Mermelstein[17]. Before the introduction of the MFCC algorithm, the Linear Prediction Coefficients (LPC) and the Linear Prediction Cepstral Coefficients (LPCCs)[18] where the most used methods in that time and were used together with Hidden Markov models[19] in speech recognition system on that time. The Calculation of the MFCC features is done by following steps:

1. The Algorithm works by framing the signal into short frames, and calculate the power spectrum periodogram estimate for each frame of the signal.
2. Applying the Mel-space filter banks on the power spectra and sum the filters energies.
3. Calculate the Logarithm of the filter banks energies and calculate the DCT (Discrete cosine transform) of the logarithms.
4. The DCT coefficients must be between 2-13, so the algorithm will discard the others.
5. Sometimes the frame energy, Delta and Delta-Delta features is appended to each feature vector.

Fig. 1 shows the Block diagram of the MFCC algorithm.

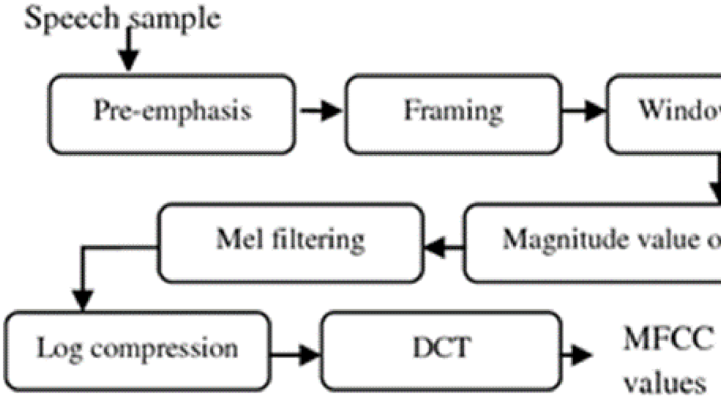

Figure 1: *Block diagram of the MFCC algorithm.*

As shown in the block diagram the pre-emphasis purpose is to amplify the higher frequencies in the input speech signal and increase the magnitude within the spectrum of the frequencies because of the high frequencies tend to have a small magnitude compared to the lower ones. The pre-emphasis filter can be represented mathematically as following:

$$y = x(t) - \alpha x(t-1)$$

Where the y is the output speech signal, x is the input signal and α is the coefficient which is typically between 0.95 – 0.97. The Pre-emphasis is useful in the FFT (Fast Fourier transform) process because of the issues with the acoustic signal values.
As described above the framing is the following step after the pre-emphasis which is divide the acoustic signal into sub-frames that have a short interval(20-40ms) and it used because of the changes in the frequencies happens in short times (milliseconds) and for other reason it’s not logical to apply the FFT on the whole speech signal.

After dividing the speech signal into sub-frames, the hamming window can be represented mathematically as following:

$$w_n = 0.54 - 0.46cos(\frac{2\pi n}{N-1})$$

where N is the window length. Fig. 2 shows the hamming window.

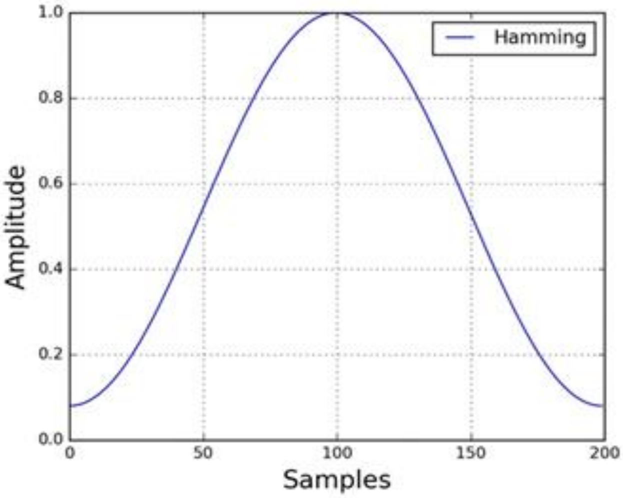

Figure 2: *Hamming window applied to speech signal.*

After applying the hamming window, the Fast Fourier transform is applied to calculate the frequency spectrum on each frame then the power spectrum is calculated as following:

$$P = \frac{|FFT(x_i)|^2}{N}$$

where N is the number of FFT as typically 256 or 512 and xi is represent the ith sub-frame of signal x. The Final step in the MFCC is filters bank which is computed by applying 40 triangular filter on the mel-scale. Each filter has a triangular response which has a value of one and decrease linearly to zero. Fig. 3 shows the Mel-scale filters bank response.

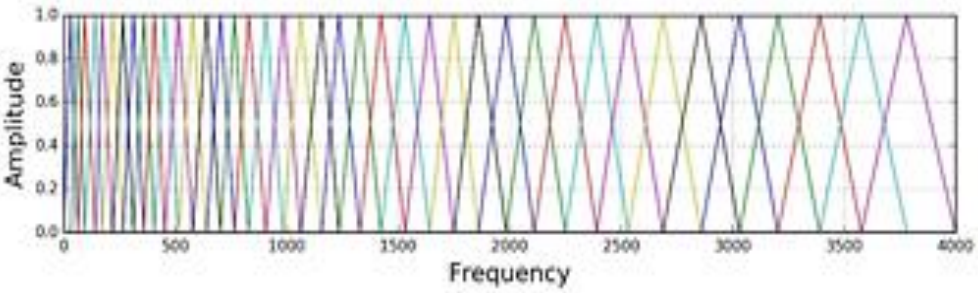

Figure 3: *The Response of the triangular mel-scale filter bank.*

The Discrete Cosine Transform (DCT) applied to the output of the mel-scale filter banks because their output is highly correlated which is an issue to the classification and machine learning algorithm to deal with. The DCT decorrelate the output and compress the filters representations.

2.2. DCNN (Deep Convolutional Neural Networks)

The DCNN are type of multi-layer perceptron. It inspired by the visual cortex of animals. The main application of the CNNs is the image processing but it has been used in many other fields like voice and video processing. The algorithm was proposed [20]. the Architecture of the CNN consists many types of layers, the main layer is the convolution layer. The convolution layer purpose is to extract features from the input data, which preserves the spatial relationship between data points by learning features using small squares of input data [21]. The other important layer is the Pooling layer which is used to reduce the dimension of the data by combining the output of the features map into single value in the next layer. Typically, the pooling size used is 2x2[22] [23]. The Last layer in each CNN network is the fully connected layer which its purpose connects each neuron in the layer to every neuron the following layer. The principle of the fully connected layer is similar to the multi-layer perceptron.

In this study, we designed three architecture of CNN network in order to classify speech signal based on its MFCC features. Table 1, Table 2 and Table 3 shows the architecture of each model.

Table 1. *Model 1 Architecture.*

Layer	Input Shape	Output Shape	Weights	Activation
Input layer	79x12	79x12	0	None
ZeroPadding1D	79x12	80x 12	0	None
Conv1D	80x 12	15x 50	6050	ReLU
Dropout	15x 50	15x 50	0	None
Conv1D	15x 50	2x100	50100	ReLU
Dropout	2x100	2x100	0	None
Average Pooling	2x100	1x100	0	None
Fully Connected Layer	1x100	1, 11	1111	Softmax

Table 1. *Model 2 Architecture.*

Layer	Input Shape	Output Shape	Weights	Activation
Input layer	79x12	79x12	0	None
Conv1D	79x12	39x20	740	ReLU
Dropout	39x20	39x20	0	None
Conv1D	39x20	19x20	1220	ReLU
Dropout	19x20	19x20	0	None
Conv1D	19x20	9x20	1220	ReLU
Dropout	9x20	9x20	0	None
Conv1D	9x20	4x20	1220	ReLU
Dropout	4x20	4x20	0	None
Conv1D	4x20	1x20	1220	ReLU
Dropout	1x20	1x20	0	None
Fully Connected Layer	1x20	1x11	231	Softmax

The purpose of this study is to provide an algorithm that understands a small selection of simple audio commands. The selection of commands the algorithm should understand should be (Yes, No, Down,up, right, left, off, on, go, stop). The purpose of the work is to Objectively work to minimize the number of layers in the architecture to maintain some simplicity within the model and to minimize runtime, using convolutional layers with a wide-enough convolution window and stride to identify patterns and using pooling layers to reduce dimensionality between convolutional layers.

Each of the network uses 1-dimensional convolutional layers that slide along the time axis, convoluting MFCCs and reducing the size of the first dimension of the audio data. The idea behind each of these networks is to recognize patterns among MFCCs across time. Differentiating features among these three models were the size of the convolution windows, the overlapping stride of the windows, the number of filters, and the number of layers.

The three model architectures that introduced are (1) Large Windows/Few Layers/Many Filters - this model consists of 1 padding layer, 2 convolutional layers with window size of 10 and stride of 5 each, an average pooling layer, and a dense output layer. The first and second convolutional layers have 50 and 100 filters, respectively. Dropout layers have been added after each convolutional layer. (2) Small Windows/Fewer Filters - this model consists of 5 sets of convolutional layers with window size of 3 and strides of 2 followed by a dense output layer. Each convolutional layer has 20 filters. Dropout layers have been added after each convolutional layer. (3) Moderate Windows/Increasing Filters - this model consists of a set of 1 padding and 2 convolutional layers with window size 4 and stride 2, followed by another set of 1 padding and 3 convolutional layers with small window sizes. The number of filters increase with each additional layer until the final dense output layer. Dropout layers have been added after each convolutional layer.

Hyperparameters that changed with each model were the dropout rate (0%,25%,50%) and the learning rate (0.01, 0.05, 0.10). Therefore, the total number of models evaluated was 81.

2.3. DATASET

The algorithm trained and tested using the Speech Commands Dataset released by Google on August 3, 2017. The data contains 64,727 one-second audio clips of 30 short words. The audio files were crowdsourced by Google with the goal of collecting single-word commands (rather than words as said and used in conversation).

A group of 20 core words audio files were recorded and repeated 5 times by the most of speakers. An additional group 10 words were recorded to help distinguish unrecognized words; most speakers recorded these words once. The core words consist of (Yes, No, Down,up, right, left, off, on, go, stop) and the numbers zero through nine. Auxiliary words consist of "Bed", "Bird", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila", "Tree" and "Wow".

This algorithm will take as inputs: a file of PCM-encoded data to be decoded into a 16-bit [16000, 1] integer tensor and Output will be a [1, 11] tensor representing the prediction of the algorithm. Values will be between 0 and 1.

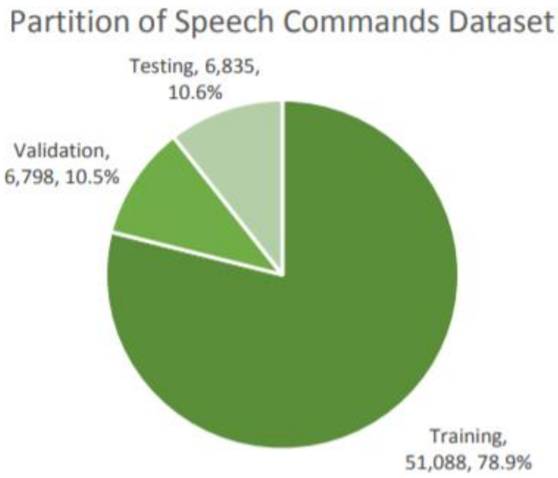

Figure 4. *The Chart show the Dataset separation into Training,validation and testing subsets.*

3. Results

The three-model architecture that presented in this paper, trained using the dataset. The program has been implemented using python and Keras deep learning library with TensorFlow backend. The Table below shows the Test results of the Models.

Table 3. *Test results of each model with hyperparameters values.*

Dropout rate	*Learning Rate*	*Model 1*	*Model 2*	*Model 3*
0.00	0.01	84.61%	82.27%	83.61%
	0.05	84.96%	82.11%	82.79%
	0.10	82.93%	82.44%	83.44%
0.25	0.01	87.96%	72.28%	83.44%
	0.05	88.21%	74.48%	83.79%
	0.10	88.21%	74.92%	82.91%
0.50	0.01	84.80%	62.44%	66.99%
	0.05	85.41%	62.44%	67.17%
	0.10	85.21%	62.44%	65.85%

Model architecture 1 was vastly superior to the other two models with all scores above the average of the testing data. This suggests that a greater convolution window, stride, and number of filters may render direct improvement in classification accuracy. A dropout rate of 0.25 was better than both 0 and 0.5 which isn't surprising considering that a rate of 0 leads to overfitting of test data and a substantial dropout rate will impede training as backpropagated calculations will fail to persist. This suggests an appropriate range of dropout rate should be narrower and skewed toward lower values greater than 0. In general, a learning rate of 0.05 tested better than rates at 0 and 0.1, suggesting that models tested in a narrower range including 0.05 may train and test better.

Discussion

Many researches and studies on speech recognition have been done in the literature. M. Karakaş, Using Mel Frequency Cepstral Coefficients and Dynamic Time Bending Algorithms "OPEN", "CLOSE", "START" and "STOP" on MATLAB, an average accuracy of 88.5% and an average accuracy of 82% were obtained independently of the speaker [24] . Fezari meat. get. used RSC 364 board for speech recognition and PIC 16F876 as microprocessor. Experiments with regular and disorganized speech outside the laboratory and regular and disorganized speech resulted in average success rates of 85%, 73%, 78% and 65%, respectively [25]. P. Leechor et al. in. Visiual Basic 6 user interface used Hidden Markov Model user kit. In a noisy environment, when a remote-controlled car was controlled with voice commands, it achieved 98% accuracy, while in a very noisy room it decreased by up to 44% [26]. V. A. Petrushin used the Artificial Neural Networks algorithm for speech recognition. According to the calls made to the call center in the study, 30 people tried to test 5 different emotions and achieved 70% success [27]. used in this article the dataset contains 64,721 files has been separated into 51,088 for the training set, 6,798 for the validation set and 6,835 for the testing set. The difference of our article we have evaluated 3 models with different hyperparameters configuration in order to choose the best model with higher accuracy. The highest accuracy achieved is 88.21%.

Conclusion

This research paper has been proposed an automatic speech recognition system based on two algorithms for features extraction, firstly the unique features of the speech acoustic signal has been extracted using MFCCs and feed theses features for the CNN algorithm for further features learning and classification. The Study compared three model architectures of the CNN in order to see the best results of the model with various options of hyperparameters and layers. The study shows that the using of MFCC as a feature extraction and feed these features for the CNN model for further features extraction will improve the accuracy and reduce the complexity of the model.

References

[1] M. A. Anusuya and S. K. Katti. “Speech Recognition by Machine, A Review”. In: International Journal of Computer Science and Information Security, IJCSIS, Vol. 6, No. 3, pp. 181-205, December 2009, USA (Jan. 2010).

[2] S. Sinha, S.S. Agrawal, A. Jain. “Continuous density hidden markov model for Hindi speech recognition”, GSTF J. Comput., 3(2) (2018).

[3] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, G. Penn. “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition Acoustics”, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE (2012, March), pp. 4277-4280.

[4] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu. “Convolutional neural networks for speech recognition” IEEE/ACM Trans. Audio Speech Lang. Process., 22(10) (2014), pp. 1533-1545.

[5] Mohamed, A.R.; Dahl, G.E.; Hinton, G.E. Acoustic Modeling Using Deep Belief Networks. IEEE Trans. Audio Speech Lang. Process. 2012, 20, 14–22. [CrossRef]

[6] Yu, D.; Seide, F.; Li, G. Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.

[7] Tüske, Z.; Golik, P.; Schlüter, R. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.

[8] Arisoy, E.; Sainath, T.N.; Kingsbury, B.; Ramabhadran, B. Deep Neural Network Language Models. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT; Association for Computational Linguistics: Montréal, QC, Canada, 2012; pp. 20–28.

[9] Sainath, T.N.; Mohamed, A.R.; Kingsbury, B.; Ramabhadran, B. Deep convolutional neural networks for LVCSR. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8614–8618.

[10] Mitra, V.; Wang, W.; Franco, H.; Lei, Y.; Bartels, C.; Graciarena, M. Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In Proceedings of the Fifteenth annual conference of the international speech communication association, Singapore, 14–18 September 2014.

[11] Pratap, V.; Hannun, A.; Xu, Q.; Cai, J.; Kahn, J.; Synnaeve, G.; Liptchinsky, V.; Collobert, R. wav2letter++: The

Fastest Open-source Speech Recognition System. arXiv 2018, arXiv:1812.07625.

[12] De Andrade, D.C. Recognizing Speech Commands Using Recurrent Neural Networks with Attention. Available online: https://towardsdatascience.com/recognizing-speech-commands-using-recurrent-neuralnetworks-with-attention-c2b2ba17c837 (accessed on 27 December 2018).

[13] Andrade, D.C.; Trabasso, L.G.; Oliveira, D.S.F. RA Robust Frequency-Domain Method For Estimation Of Intended Fundamental Frequency In Voice Analysis. Int. J. Innov. Sci. Res. 2018, 7, 1257–1263.

[14] Han, Wei, et al. "An efficient MFCC extraction method in speech recognition." *2006 IEEE international symposium on circuits and systems*. IEEE, 2006.

[15] Jiang, Fei, et al. "An event recognition method for fiber distributed acoustic sensing systems based on the combination of MFCC and CNN." *2017 International Conference on Optical Instruments and Technology: Advanced Optical Sensors and Applications*. Vol. 10618. International Society for Optics and Photonics, 2018.

[16] Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

[17] S. Davis and P. Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: IEEE Transactions on Acoustics, Speech, and Signal Processing 28.4 (Aug. 1980), pp. 357–366.

[18] Harshita Gupta and Divya Gupta. “LPC and LPCC method of feature extraction in Speech Recognition System”. In: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). IEEE, Jan. 2016.

[19] D.J. Mashao, Y. Gotoh, and H.F. Silverman. “Analysis of LPC/DFT features for an HMM-based alphadigit recognizer”. In: IEEE Signal Processing Letters 3.4 (Apr. 1996), pp. 103–106.

[20] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: Proceedings of the IEEE 86.11 (1998), pp. 2278–2324.

[21] Hamed Habibi Aghdam and Elnaz Jahani Heravi. “Traffic Sign Detection and Recognition”. In: Guide to Convolutional Neural Networks. Springer International Publishing, 2017, pp. 1–14.

[22] Dan Claudiu Ciresan et al. “Convolutional Neural Network Committees for Handwritten Character Classification”. In: 2011 International Conference on Document Analysis and Recognition. IEEE, Sept. 2011.

[23] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[24] M. Karakaş, "Computer Based Control Using Voice Input", Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi, Eylül 2010.

[25] M. Fezari, M.B. Salah, "A Voice Command System for Autonomous Robots Guidance", IEEE AMC-06 0-7803-9511, 2006.

[26] P. Leechor, C. Pornpanomchai, P. Sukklay, “Operation of a Radio Controlled Car by Voice Commands”, 2010 2nd International Conference on Mechanical and Electronics Engineering (ICMEE 2010), 2010.

[27] V.A. Petrushin, "Emotion in Speech Recognition and Application to Call Centers", Andersen Consulting.