ASEAN Engineering Journal

Full Paper

Spoken-Digit Classification using Artificial Neural Network

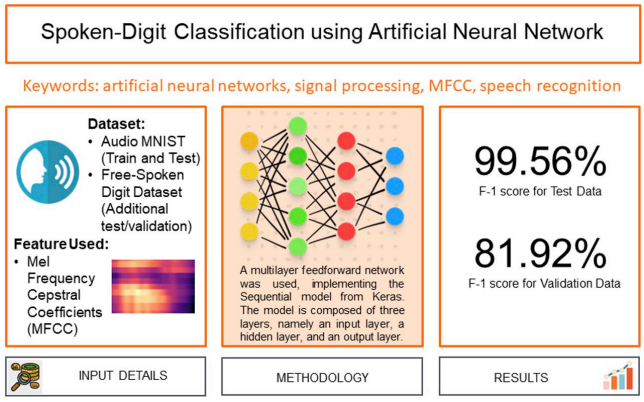
Aunhel John M. Adoptante[a], Arnie M. Baes[a], John Carlo A. Catilo[a], Patrick Kendrex L. Lucero[a], Anton Louise P. De Ocampo[a], Alvin S. Alon[a]*, Rhowel M. Dellosa[b]

[a]Batangas State University, Batangas City, Batangas, Philippines
[b]University of Northern Philippines, Vigan City, Ilocos Sur, Philippines

Article history
Received
03 March 2023
Received in revised form
21 July 2022
Accepted
09 October 2022
Published online
28 February 2023

*Corresponding author
alvin.alon@g.batstate-u.edu.ph

Graphical abstract



Abstract

Audio classification has been one of the most popular applications of Artificial Neural Networks. This process is at the center of modern AI technology, such as virtual assistants, automatic speech recognition, and text-to-speech applications. There have been studies about spoken digit classification and its applications. However, to the best of the author's knowledge, very few works focusing on English spoken digit recognition that implemented ANN classification have been done. In this study, the authors utilized the Mel-Frequency Cepstral Coefficients (MFCC) features of the audio recording and Artificial Neural Network (ANN) as the classifier to recognize the spoken digit by the speaker. The Audio MNIST dataset was used as training and test data while the Free-Spoken Digit Dataset was used as additional validation data. The model showed an F-1 score of 99.56% accuracy for the test data and an F1 score of 81.92% accuracy for the validation data.

Keywords: artificial neural networks, signal processing, MFCC, speech recognition

1.0 INTRODUCTION

Artificial Neural Networks (ANN) is a modeling technique inspired by the human nervous system that allows learning by example from representative data that describes a physical phenomenon or a decision process [1]. Globally, interest in artificial neural networks is rising. Their research has been the focus of numerous books and academic publications, and their use is expanding constantly [19]. In many aspects of daily life, including document classification, medical diagnosis, and the recognition of sounds and images, ANN models have been employed extensively. One of the applications of ANN is implementing them for audio classification.

Audio classification is the process of listening to and analyzing audio recordings. It is at the center of modern AI technology, such as virtual assistants, automatic speech recognition, and text-to-speech applications [2]. There are various phases in a voice recognition system. The extraction of speech features is the initial step. The extraction of speech features is followed by a phase of pattern recognition. Artificial neural networks (ANN) and genetic algorithms (GA) are two AI-based methods for pattern recognition in speech recognition systems [20].

A study [3] implemented an Artificial Neural Network (ANN) to classify and recognize bird species by the sound it makes. The power spectral density of the recordings per bird sound was used as input and fed to the ANN.

In another study [4], sounds of the human heart were used to determine if there were any symptoms of heart disease The Mel-frequency Cepstral Coefficients (MFCC) were extracted from the heart sounds and fed to an Artificial Neural Network (ANN) for classification.

Another study [5] focused on determining the emotion of the speaker using ANN and MFCC features. The model was used to classify audio recordings into eight emotions: happy, sad, angry, surprise, disgust, calm, and neutral.

Wahyuni [6] conducted and published a study addressing a challenging issue in recognizing spoken Arabic letters. Three letters, namely sa, sya, and tsa, have identical pronunciation when spoken by Indonesian speakers, however, they have different makhraj in Arabic. The researcher used Mel-Frequency Cepstral Coefficients (MFCC) and implemented an Artificial Neural Network.

There have been studies about Isarn digit speech recognition using the Hidden Markov Model (HMM) classifier [7], Pashto spoken digits recognition implementing Support Vector Machine (SVM) [8], Arabic spoken digits recognition using Deep Learning [9], and even English spoken digit recognition implementing Convolutional Neural Networks (CNN) [10]. However, to the best of the authors' knowledge, very few works focusing on English spoken digit recognition that implemented ANN classification have been done.

In this study, the authors utilized the Mel-Frequency Cepstral Coefficients (MFCC) features of the audio recording and Artificial Neural Network (ANN) as the classifier to recognize the spoken digit by the speaker.

2.0 METHODOLOGY

The dataset used in this study was described first followed by the feature extraction method, the Artificial Neural Network architecture, and finally the evaluation. Figure 1 shows the block diagram for the process of the study.

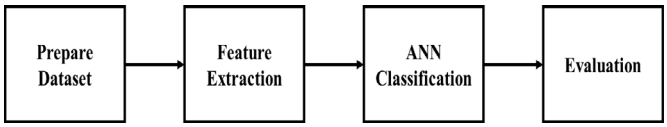


Figure 1. Block diagram of study

A. Dataset

The Audio MNIST dataset [11] was used in this work for the training and testing of the ANN model. The dataset contains 30 000 audio samples of spoken digits from 0 to 9 by 60 different speakers of different ages, gender, accent, and origin. These recordings were originally sorted into folders, each folder allocated per speaker, containing 50 samples per digit or 500 recordings per speaker. The proponent then sorted all the samples into folders with each folder containing the recording of each digit, regardless of speaker. The final result was 10 folders, one folder for each digit, containing 3000 samples.

To further validate the results of the model, a new data set was used as validation data. The dataset included the Free-Spoken Digit Dataset (FSSD) [12]. The FSSD contains a total of 3000 recordings, 50 samples per digit from 6 speakers.

B. Feature Extraction Process

Extracting the MFCC features is a widely-used method in different speech recognition applications. Figure 2 shows the block diagram of the MFCC feature extraction process.

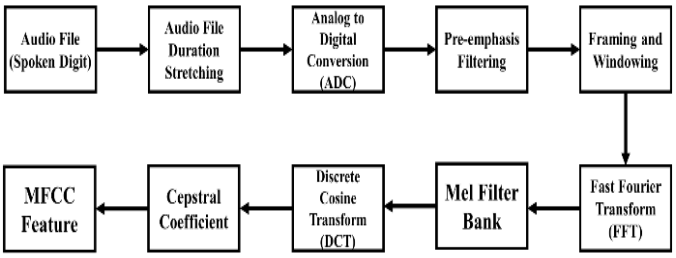


Figure 2 MFCC Feature Extraction Process Block Diagram

- Audio File Duration Stretching

  The duration of audio files is stretched to one second. This is to ensure that all the audio files have the same duration, thus making resulting in the same number of outputs later on.

- Analog to Digital Conversion (ADC)

  In ADC, the audio clips are sampled and digitized, converting the analog signal into discrete space. A sampling frequency of either 8 kHz or 16 kHz is often used in the process. In this study, a sampling rate of 8 kHz was used. The process is seen in Figure 3.

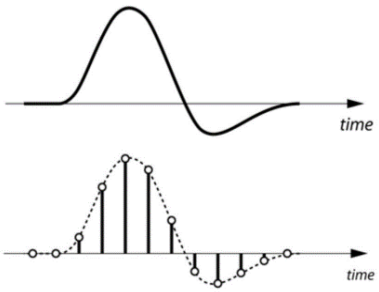


Figure 3. Analog to Digital Conversion

- Pre-emphasis Filtering

  To increase the high-frequency energy while decreasing the low-frequency energy, pre-emphasis filtering was applied at this stage, simply shown in equation (1):

$$y_t = \alpha x_t + (1-\alpha)x_{t-1} \quad (1)$$

- Framing and Windowing

  Speech signals are very dynamic. However, the signal can be viewed as stationary at certain time ranges or frames, or windows. In this process, the speech signal is divided into several windows or frames for processing. The said process can be visualized in Figure 4 [13].

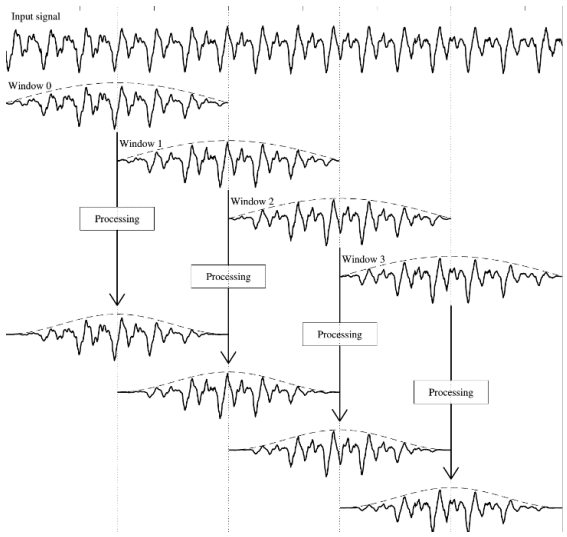


Figure 4. Illustration of the Framing Process

- Fast Fourier Transform (FFT)

Discrete Fourier Transform (DFT) plays an important role in many applications of digital signal processing. At this stage, Fast Fourier Transform (FFT), an efficient computational algorithm of DFT, was implemented. Each frame was converted from the time domain to the frequency domain [14]. Figure 5 gives us a visual representation of how FFT is implemented.

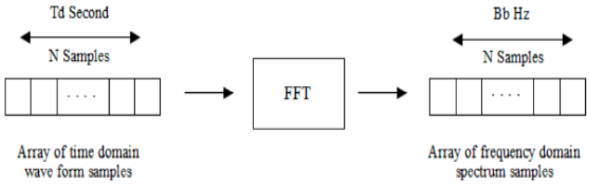


Figure 5. FFT Representation

The DFT decomposes a signal into a series of the following form:

$$X_k = \sum_{m=0}^{M-1} x_m e^{-\frac{i(2\pi km)}{M}} \rightarrow k = 0, \dots, M-1 \quad (2)$$

where x_m is a point in the signal being evaluated and the X_k is a specific 'mode' or frequency component.

The complex exponential can be rewritten as sine and cosine functions using the Euler formula:

$$e^{iy} = cos(y) + isin(y) \quad (3)$$

Such that our series contains sinusoidal waves:

$$X_k = \sum_{m=0}^{M-1} x_m \left\{\cos\left(\frac{2\pi km}{M}\right) - isin\left(\frac{2\pi km}{M}\right)\right\} \quad (4)$$
$$k = 0, \dots, M-1$$

- Mel Filter Bank

To perform linear predictions at this stage, bank-filter analysis was implemented. Computations on the Mel Scale were used, as shown in equation (5).

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (5)$$

Figure 6 shows a sample diagram of the filter bank in the Mel scale [13].

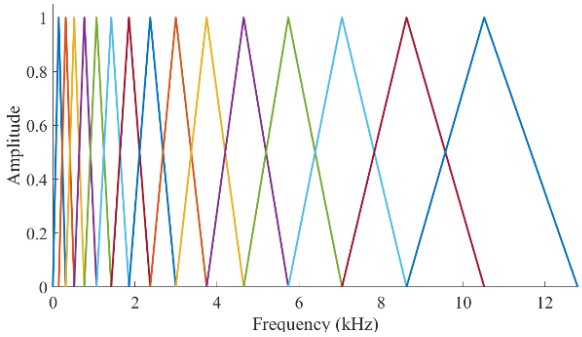


Figure 6. Mel scale filter bank

- Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT) utilizes a mixture of different cosine functions with varying frequencies to constitute a finite sequence of data. The general equation of a one-dimensional (N data points) DCT is defined in equation (6) [15].

$$F(u) = \alpha(u) \sum_{i=0}^{N-1} \cos\left[\frac{\pi u}{2N}(2i+1)\right] f(i) \quad (6)$$

where f(0, 1,…N-1) represents the discrete data sequence of signal f, N represents the number of samples, F(0, 1,…N-1) denotes cosine transform coefficients, and,

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}}, & u = 0 \\ \sqrt{\frac{2}{N}}, & otherwise \end{cases} \quad (7)$$

- Cepstral Coefficient

The final result will be 13 cepstral coefficients per window. The input to the model will be 13 MFCCs per window multiplied by several windows.

C. ANN Classification

In this study, a multilayer feedforward network was used, implementing the Sequential model from Keras. With a neural network as the main classifier, datasets are trained, implemented, and tested establishing the optimized parameters to be used [16]. See Figure 7 for the diagram of a multilayer feedforward network

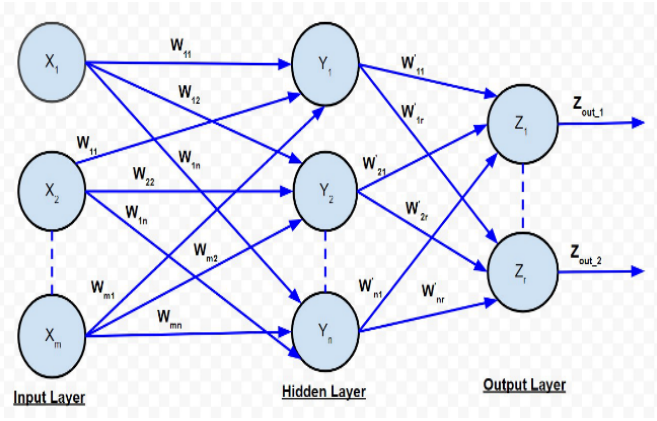


Figure 7. Diagram of a multilayer feedforward network.

The model is composed of three layers, namely an input layer, a hidden layer, and an output layer. Not including the input layers, each node is a neuron implementing a nonlinear activation function. See Table 1 for the details of the model.

Table 1. Details of each layer

Layers	Number of Nodes	Activation Function
Input	13 MFCC Features x No. of Windows	--
Hidden	1024	ReLU
Output	10	Softmax

For the hidden layer, Rectified Linear Activation Unit (ReLU) was used, as shown in equation (8) [17]. Softmax was used in the output layer, shown in equation (9) [18].

$$z = b + \sum_i x_i w_i \; ; \; y = \begin{cases} z, & z > 0 \\ 0, & otherwise \end{cases} \tag{8}$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{9}$$

$$for\ i = 1, \dots, K\ and\ z = (z_1, \dots, z_K) \in \mathbb{R}^K$$

Table 2 shows the important details regarding the training of the model.

Table 2. Details of the Model

Optimizer	Nadam
Loss	Sparse Categorical Crossentrophy
Metrics	Accuracy
Number of Epochs	50

D. Evaluation

- Accuracy

To determine how successful the classification is, we divide the number of correct classifications over the total number of classifications performed to determine the accuracy, as shown in equation (10).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}\ x\ 100\% \tag{10}$$

In which:

TP = True Positive

FP = False Positive

FN = False Negative

TN = True Negative

Using the confusion matrix shown in Figure 8, TP, TN, FP, and FN values can be obtained.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 8. Confusion Matrix

- Precision

Out of all the positive predicted, we determine the percentage of truly positive.

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

- Recall

Out of the total positive, we determine the percentage of predicted positive.

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

- F1-Score

The F1-score is the harmonic mean of precision and recall, taking both false positives and false negatives into consideration.

$$F1 - score = \frac{2*(Precision*Recall)}{Precision+Recall} \tag{13}$$

3.0 RESULTS AND DISCUSSION

In the initial stage of preparing the dataset, all audio files were stretched so that their duration is 1 second, regardless of their original duration (Figure 9). After stretching the audio files, all audio files were sampled with a sampling rate of 8 kHz.

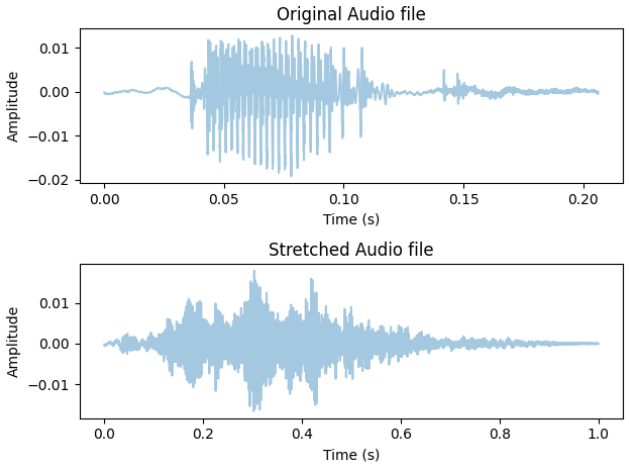


Figure 9 Sample of audio file duration stretching

The signal is then pre-emphasized, as shown in Figure 10, then framed and windowed. Each frame is 0.25 seconds long with a window step of 0.2 seconds, visually represented in Figure 11.

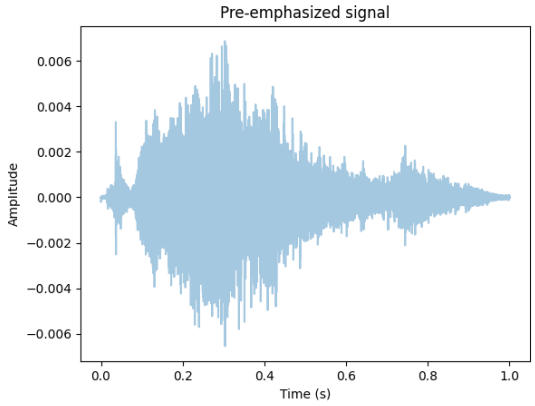


Figure 10 Pre-emphasized Signal

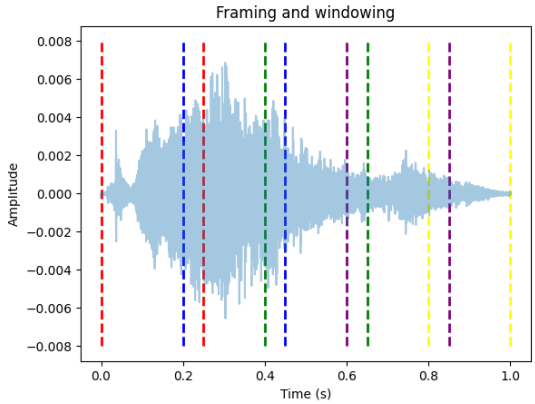


Figure 11 Framing and windowing of pre-emphasized signal

For every window, the signal is converted from the time domain to the frequency domain using FFT and then Mel-filter banks were extracted. Shown in Figure 12 is a sample of the extracted Mel spectrogram.

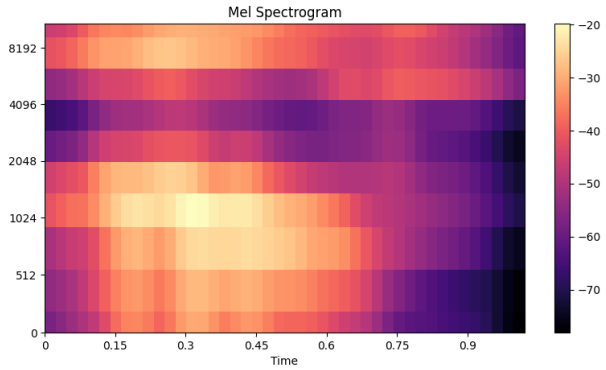


Figure 12. Extracted Mel Spectrogram

From the Mel spectrogram, the cepstral coefficients were extracted (Figure 13). The resulting MFCCs were saved in a JSON file and used as features in the artificial neural network.

The performance of the model per digit using the test data can be better visualized in the table below (Table 3), followed by its corresponding confusion matrix (Figure 14). The model had the highest precision in recognizing the number 2 at 99.78% and the lowest precision in recognizing the number 8 at 99.42%. The highest recall was recorded in recognizing the number 7 at a perfect 100% while the lowest recall was observed in recognizing the number 1 at 99.22%. Recognizing the number 7 exhibited the highest F1 score at 99.78% while the lowest F1 score was observed in recognizing the number 1 at 99.33%. It can also be observed that there wasn't any significance in the performance per digit recognition, indicating that the model has performed quite well in recognizing all the digits.

To further observe the performance of the model, the results of the recognition using the validation data are shown in the table below (Table 4), followed by its corresponding confusion matrix, shown in Figure 15. The model had the highest precision in recognizing the number 9 at 99.58% and the lowest precision in recognizing the number 5 at 64.95%. The highest recall was recorded in recognizing the number 8 at 99.00% while the lowest recall was observed in recognizing the number 6 at 50.33%. Recognizing the number 0 exhibited the highest F1 score at 89.44% while the lowest F1 score was observed in recognizing the number 6 at 65.09%. It can be observed that the performance of the model had a significant change.

Table 5 compares the model's average performance using the test data versus the validation data. The support column indicates the total number of samples used in the classification process. There is an observable difference in the performance of the model when using a different dataset. The model exhibits an F1 score of 99.56% using the test data as inputs compared to the F1 score of 81.92% when using the validation data.

MFCC Features:
[[-9.0319574424 -10.4534552748 19.3054834901 -14.8452151713 -50.1194860610 5.9582132035 11.3758384460 2.3031549447 22.6241006926 -11.1476340677
-18.4193970715 -1.2404508018 -18.1876075928],
[-6.8814607886 -10.8886258443 16.3103757042 -22.4133909407 -53.0798941929 7.2923459229 6.2489552545 -2.2808533667 2.8830692019 -18.5149882264
-19.9167218210 -7.4848033523 -14.5646974322],
[-6.0735867171 -6.1785359303 18.6321166589 -24.6979969885 -58.2258486303 1.2497979729 0.2635419561 -6.5458015254 3.5952628515 -18.1179847611
-20.1854889515 -11.9373618309 -27.5320834135],
[-5.5182454753 -6.7402167190 18.2998245761 -22.2860046621 -53.5063616182 4.9933587831 1.5638258890 1.3600512726 7.8674688797 -10.2507245396
-15.2935870865 -8.5435749646 -19.2381331965],
[-5.6638389393 -3.6780152865 25.8278061743 -23.6243148678 -56.4218832520 -2.5023080219 -21.3386221787 -5.0677261490 -1.6508632218 -7.4391235367
-14.9916535081 -14.0930393728 -26.4297197884],
[-6.0214502003 -2.8413640639 30.6541347781 -24.6783010941 -57.3557607840 -6.3040959220 -16.6934893650 -6.0622614749 -3.3112631517 -8.0332222442
-12.2232174699 -8.8481940767 -23.1782229440],
[-6.9681977017 -2.8402611479 36.3857839487 -20.2225952317 -48.5842693961 -13.4189429725 -19.0906881673 -2.1553927697 3.7930760328 3.0381451624
-3.1272189044 -2.5747005186 -13.2111378557],
[-7.7981168803 -3.8863024185 32.4961598896 -10.9840760685 -45.2697681928 -20.2205176978 -17.2244623083 -10.3781239482 -0.5409544096 -1.5597357999
-5.8936472977 -8.3544613381 -16.3932791359],
[-8.4851126668 -8.0281153451 20.8094907918 -5.3018076184 -31.8478130787 -35.7420507343 -4.1603785134 -20.9495820204 -6.5210211307 11.0968401244
-7.7514807899 -8.4467491230 2.3440018284],
[-8.8084868451 -16.2830006494 7.2655047697 -7.5832523586 -20.0571716550 -23.2247570124 8.2856759096 -18.7664114105 -10.0307044774 3.8857600439
-11.9276353692 -7.2591881239 6.9793374005],
[-8.1827570187 -25.9577802332 5.7485517812 -10.3123116544 -18.5312064355 -20.0186747669 28.6026727833 -7.7318642000 -3.8008769650 15.2379024087
4.7759994695 -12.1902346816 16.6477535896],
[-9.3815069395 -26.3398333132 3.9205405164 -0.7020526122 -15.2789056997 -20.5944827303 26.2934977636 5.9123605460 -2.3527580053 16.7667676356
15.0888383939 -19.1138142974 14.5426777416],
[-10.5445164457 -25.5571485244 2.7856323026 2.1792435707 -9.8285517841 -22.6699645643 29.9664472527 6.5177732406 -7.5623001874 12.8674805612
16.7338182150 -21.2344782971 3.7877969291],
[-13.0407845097 -26.1850829167 1.7479608015 2.9584083865 -14.4740223516 -20.1628393746 26.9928833681 5.5798399147 -10.7580177451 13.2114083433
11.4082929538 -23.4519380667 -0.9475801255]]

Figure 13. Extracted MFCC Features

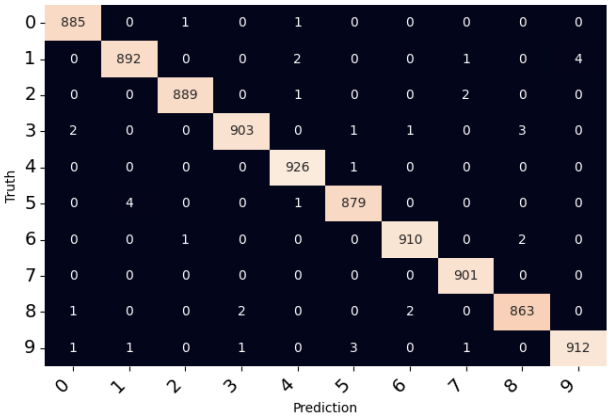


Figure 14. Confusion Matrix for Test Data

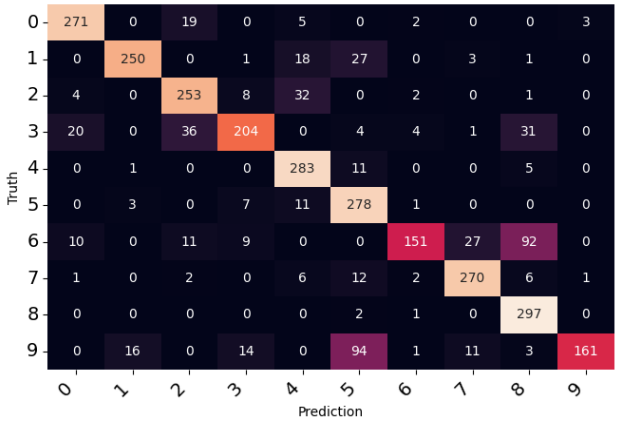


Figure 15. Confusion Matrix for Validation Data

Table 3. Performance of Model per Digit using Test Data

Digit	Precision	Recall	F1-Score	Support
0	99.55%	99.77%	99.66%	887
1	99.44%	99.22%	99.33%	899
2	99.78%	99.66%	99.72%	892
3	99.67%	99.23%	99.45%	910
4	99.46%	99.89%	99.68%	927
5	99.43%	99.43%	99.43%	884
6	99.67%	99.67%	99.67%	913
7	99.56%	100.00%	99.78%	901
8	99.42%	99.42%	99.42%	868
9	99.56%	99.24%	99.40%	919
Accuracy	-	-	99.56%	9000

Table 4. Performance of Model per Digit using Validation Data

Digit	Precision	Recall	F1-Score	Support
0	88.56%	90.33%	89.44%	300
1	92.59%	83.33%	87.72%	300
2	78.82%	84.33%	81.48%	300
3	83.95%	68.00%	75.14%	300
4	79.72%	94.33%	86.41%	300
5	64.95%	92.67%	76.37%	300
6	92.07%	50.33%	65.09%	300
7	86.54%	90.00%	88.24%	300
8	68.12%	99.00%	80.71%	300
9	97.58%	53.67%	69.25%	300
Accuracy	-	-	81.92%	3000

Table 5. Average Performance of the Model

Data	Accuracy	
	F1-score	Support
Test Data	99.56%	9000
Validation Data	81.92%	3000

4.0 CONCLUSION

The study has presented a spoken digit classification system that uses the Mel-Frequency Cepstral Coefficient of the audio for the features and an artificial neural network for the classification process. The model showed an F-1 score of 99.56% accuracy for the test data and an F1 score of 81.92% accuracy for the validation data.

This shows that the model can recognize the data well if it has a history of processing similar data in the training period since the test data is also part of the data set from which the training data was derived. However, when using separate data, the performance slightly decreases but the performance can still be considered passable.

Acknowledgment

The authors would like to thank and acknowledge the Electronic Systems Research Center and the Digital Transformation Center of Batangas State University, The National Engineering University for their support in this research.

References

[1] R. Sadiq, M. J. Rodriguez and H. R. Mian, 2019 Encyclopedia of Environmental Health (Second Edition),.

[2] TELUS International.2021. What is audio classification?," [Online]. Available: https://www.telusinternational.com/articles/what-is-audio-classification. Accessed: Aug-2021

[3] M. M. M. Sukri, U. Fadlilah, S. Saon, A. K. Mahamad, M. M. Som and A. Sidek,2020 "Bird Sound Identification based on Artificial Neural Network," in *2020 IEEE Student Conference on Research and Development (SCOReD)*, Batu Pahat, Malaysia, 342-345.

[4] M. Rahmandani, H. A. Nugroho and N. A. Setiawan, 2018."Cardiac Sound Classification Using Mel-Frequency Cepstral Coefficients (MFCC) and Artificial Neural Network (ANN)," in *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia,

[5] H. Dolka, A. X. V. M and S. Juliet, 2021. "Speech Emotion Recognition Using ANN on MFCC Features," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, Coimbatore, India,

[6] E. S. Wahyuni, 2017. "Arabic speech recognition using MFCC feature extraction and ANN classification," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia,

[7] S. Sangjamraschaikun and P. Seresangtakul, 2017."Isarn digit speech recognition using HMM," in *2017 2nd International Conference on Information Technology (INCIT)*, Nakhonpathom, Thailand,

[8] S. Nisar, I. Shahzad, M. A. Khan and M. Tariq, 2017. "Pashto spoken digits recognition using spectral and prosodic based feature extraction," in *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, Doha, Qatar

[9] S. M. B. Wazir and J. H. Chuah, 2019."Spoken Arabic Digits Recognition Using Deep Learning," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Selangor, Malaysia.

[10] R. V. Sharan, 2020. "Spoken Digit Recognition Using Wavelet Scalogram and Convolutional Neural Networks," in *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Thiruvananthapuram, India.

[11] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller and W. Samek, 2018"Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals," [Online]. Available: https://arxiv.org/abs/1807.03418.

[12] Z. Jackson, 2018 "Free Spoken Digit Dataset (FSDD)," [Online]. Available: https://github.com/Jakobovski/free-spoken-digit-dataset. doi: 10.5281/ZENODO.1342401.

[13] T. Bäckström, "Windowing," Aalto University Wiki, Aug-2019. [Online]. Available: https://wiki.aalto.fi/display/ITSP/Windowing. [Accessed: Aug-2021].

[14] J. Hrisko, 2018."Audio Processing in Python Part I: Sampling, Nyquist, and the Fast Fourier Transform," [Online]. Available: https://makersportal.com/blog/2018/9/13/audio-processing-in-python-part-i-sampling-and-the-fast-fourier-transform. Accessed: Aug-2021

[15] D. Salomon, 2004. Data Compression: The Complete Reference, Springer Science & Business Media

[16] R. G. d. Luna, R. G. Baldovino, E. A. Cotoco, A. L. P. d. Ocampo, I. C. Valenzuela, A. B. Culaba and E. P. D. Gokongwei, 2017. "Identification of Philippine Herbal Medicine Plant Leaf Using Artificial Neural Network," in *2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, Manila, Philippines

[17] G. E. Hinton, "2012 Coursera Course Lectures," [Online]. Available: https://www.cs.toronto.edu/~hinton/coursera/lecture1/lec1.pdf Accessed: Aug-2021

[18] Goodfellow, Y. Bengio and A. Courville, 2016. Deep Learning, MIT Press,

[19] T. D. Kainova and A. A. Zhilenkov, 2022 "Artificial Neural Networks in the Geometric Paradigm," *Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 2022, pp. 321-325, doi: 10.1109/ElConRus54750.2022.9755786.

[20] R. K. C. Billones et al., 2015 "Speech-controlled human-computer interface for audio-visual breast self-examination guidance system," *2015 International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, 1-6, doi: 10.1109/HNICEM.2015.7393236.