International Journal of GEOMATE, Sept., 2017, Vol. 13, Issue 37, pp. 9-15
Special Issue on Science, Engineering & Environment, ISSN: 2186-2990, Japan
DOI: http://dx.doi.org/10.21660/2017.37.2817

ENSEMBLE MLP NETWORKS FOR VOICES COMMAND CLASSIFICATION TO CONTROL MODEL CAR VIA PIFACE INTERFACE OF RASPBERRY PI

*Narissara Eiamkanitchat[1, 2], Nontapat Kuntekul[1] and Phasit Panyaphruek[1]

[1]Department of Computer Engineering, Faculty of Engineering;
[2] Social Research Institute;
Chiang Mai University, Thailand

*Corresponding Author, Received: 29 June 2016, Revised: 10 Aug. 2016, Accepted: 30 Nov. 2016

ABSTRACT: This research, exploration displays the aftereffects of utilizing the blend of the multi-layer perceptron network system to classify Thai speech. The parameters of the training process are used in the mobile application to using Thai voice commands to control the model car. The PiFace interface of the Raspberry Pi is attached to the model car for receiving the command from mobile and control the model car. The 1,000 Thai voice commands of both men and ladies are used as the training set in the experiment. The preliminary experiments have been done to find the best possible structure of the classification model, and the appropriate proportion of classes in the training set. From the experiment results using 1 network for one voice command, the average accuracy of the classification results in the environment without noise is higher than 80%, which considered favorable in the speech recognition field of study.

Keywords: Thai voices command, MLP neural network, Speech recognition, Classification.

1. INTRODUCTION

There are many researches using voice command to control electronic devices [1]-[5]. One of the useful application, is use voice command to control wheelchair for some handicap person who not able to use hand to control the wheelchair. Usually, user needs to ware headphone and speak their command in the headphone. In this research, we try to utilize the mobile phone those most people already have one, to use as the input device. Anyway, most voice commands are using the English language, and still have problem for Thai people who cannot speak English fluently. The Thai voice command is used in this research, in order to facilities Thai handicap people.

Table 1 Thai word use to control the model car.

Thai word	Pron.	English meaning	Activity of model car
ซ้าย	Ŝāy	Left	Turn left 90 degrees then move forward
ขวา	K̄hwā	Right	Turn right 90 degrees then move forward
หน้า	H̄n̂ā	Front	Move forward
หลัง	H̄lạng	Back	Turn right 180 degrees then move forward
หยุด	H̄yud	Stop	Stop moving

The Mel-Frequency Cepstral Coefficients (MFCC) for speech recognition in many researches [6]-[8] is used as the features extraction in this research. Short 5 Thai words are used as the input to control the model car. The commands are displayed in table 1.

Each word have 200 voices of both men and women, with various pronunciations. The details of the preparatory analysis are described in the following section including and the normalization method and subtle elements of the MLP structure. The consequences of the preparatory tests are lead to the general basic configuration of the framework, describe in overall structure section. Section experimental results demonstrate the Thai voice command classification results in various circumstances. The final section is the conclusion of the propose structure and suggestion for future work.

2. THE PRELIMINARY EXPERIMENT

The feature extraction process in speech recognition is techniques for converting an analog signal to the digital signal. In processing for converting MFCC algorithm has a quality point, with a particular deciding objective to work properly to the human sound. The MFCC is a procedure to convert an ordinary frequency to the

Mel frequency. The frequency of human discourse has a lower frequency by utilizing the Mel scale can increase the low frequency scale range meanwhile decrease the range of high frequency. The processing of MFCC is shown in Fig. 1. The process begins from the sampling signal process for converting an analog signal to the computerized signal. Next stride the data is partitioned to form a frame where every frame is changed over by Fast Fourier Transform (FFT) to transform data to time domain, frequency spectrum. The information from FFT is sampled by using triangular overlapping windows converted from hertz unit to Mel-scale unit. The Mel-scale value calculated the power and take logarithm. Finally, data is reconverted by the Discrete Cosine Transform (DCT) and results the MFCC feature vector, in this progression the quantity of DCT can be assigned.

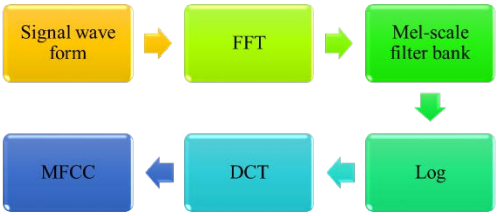

Fig. 1 The feature extraction process using MFCC.

The number of DCT utilizing as a part of this research is 21, which mean 21 of MFCC feature vectors. Subsequent to dissecting the estimation of all features, 6 features need to normalize the quality. The normalization by decimal scaling is applied in these features. The new value is figured by

$$f' = \frac{f}{10^k} \quad (1),$$

where f is the original value of feature,

k is the minimal value that makes $|v'| < 1$.

The MLP is popular algorithm to use as the classification model for the voice command [9]-[12] and other applications [13-15]. The first experiment is to locate the appropriate structure of the MLP. The principle objective is to find the superior structure with less complexity, so the number of hidden layers is not more than 2 layers. The Thai voice command to use in the experiment is the word "Ĥñā", which mean front. The first MLP structure has 3 layers including, input layer number of nodes equivalent to the number of input features. The hidden layer has 3 nodes and the output layer has 1 node. The second MLP structure has 4 layers including, input layer, 2 hidden layers and output layer. The nodes of input layer equivalent to number of input features. The following layers has 3 nodes, 2 nodes and 1 node respectively. Three experiments have been done, each experiment using 10 fold cross validation to check the precision. The experimental results appear in figure 2.

Clearly seen from every experiment displayed in figure 2 that, the MLP with 2 hidden layers have favored execution over MLP with 1 hidden layer. Notification of the parameters results of every experiments the weight estimation of input feature 1 is converted to small value almost equivalent 0. The experiment that eliminates the first input feature is done and the outcomes is approximately the same, so the 1 feature is considered not imperative for the classification and deleted.

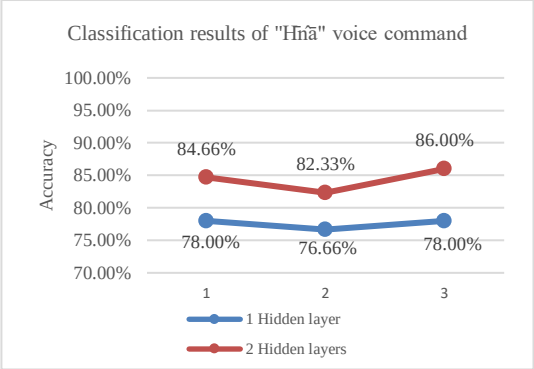

Fig. 2 The comparison of the classification results of 2 MLP structures.

Since the objective of this research is to apply the voice command classification to use with model car via the PiFace interface of Raspberry Pi, the programming language must be considered. The typical procedure of classification, including learning phase and testing phase. The learning phase using a training dataset to modify the parameters. The testing phase is receiving the value of parameters from learning process and use to classify unseen data. The consequences of using programming language as a part of learning and testing process are shown in table 2.

The results in table 2 demonstrated that using Matlab as a part of both phase is the most astounding classification result. Nonetheless, in experimental 2, clearly seen that using training parameters from Matlab cannot classify voice

command when implement by Python. This research is utilizes the Python in both learning and testing phase. The reason is a Raspberry Pi can use Python [16], [17] to implement and using Python results the same high accuracy as using Java, as found in experiment 3 and 4.

Table 2 The results of using different programming language in learning and testing process.

Expt. No.	Learning	Testing	Accuracy
1	Matlab	Matlab	85.33%
2	Matlab	Python	27.80%
3	Python	Python	84.67%
4	Java	Java	84.67%

The objective of the next preliminary experiment is to locate the portion of classes in the training data set. The preparation of the training data set in this research is for the specific purpose, to classify 5 Thai voice command. Normally perception, the desire class need to have the highest portion. The ratio using in the experiments is 80:20 of the desire class versus other classes, and 20:80 of the desire class versus other classes. Totally 3 experiments are done to each voice. The average of 5 fold cross validation is used to verify the classification accuracy. The 80:20 ratio results are displayed in figure 3, and results of 20:80 are displayed in figure 4 separately.

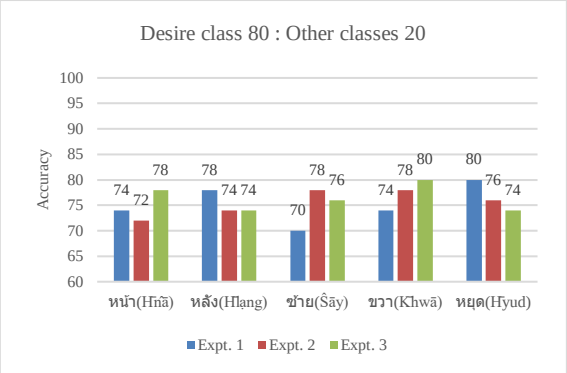

Fig. 3 The training dataset ratio 80: 20.

Figure 3 shown the average accuracy from 5 fold cross validation of training dataset ratio 80:20 of desire class and other classes. The most noteworthy accuracy form 3 experiments of ratio 80:20 of each voice are 74.60%, 75.30%, 74.60%, 77.30% and 76.60%. The average accuracy from all voices of this type of ratio is 75.68%.

Fig. 4 The training dataset ratio 80: 20.

Figure 4 shown the average accuracy from 5 fold cross validation of training dataset ratio 20:80 of desire class and other classes. The highest accuracy form 3 experiments of ratio 20:80 of each voice are 83.30%, 82.60%, 80.60%, 86.60% and 84.60%.

The average accuracy from all voices of this type of ratio is 83.54%. The results unmistakably demonstrated that the ratio of desire class 20: other classes 80, results better accuracy in every experiment. The reason is the portion 80:20, to create a training data set for voices “H̄nā”, the 200 voices of “H̄nā” are added in the training data set. The voices of other classes, including, H̄lạng, Ŝāy, K̄hwā, and H̄yud are added in the training dataset totally 50 voices. The neural network has just a few sample of other classes, result to higher miss classification. On the other hand, the portion 20:80 with the same example the 50 voices of “H̄nā” are added in the training data set. The voices of other classes, including, H̄lạng, Ŝāy, K̄hwā, and H̄yud are added in the training dataset totally 200 voices. The neural network has enough sample of other classes to learn, result to higher accuracy of classification.

3. THE OVERALL STRUCTURE

As the prior mansions, the objective of the Thai voice command development is aimed to be able to apply to utilizing mobile and the PiFace interface of Raspberry Pi. The overall structure is appeared in figure 5.

The structure in figure 5 comprises of 4 main processes. The first process is to discover the parameters to classify the Thai voice command using MLP. The second process is to implement the voice command classification models using the parameters from the training process to install in

Raspberry Pi. The third process is the mobile application implementation and the car model execution. The last process is the moving activity process, the voice command from mobile is send to classify the order by Raspberry Pi and control the movement of the car by PiFace.

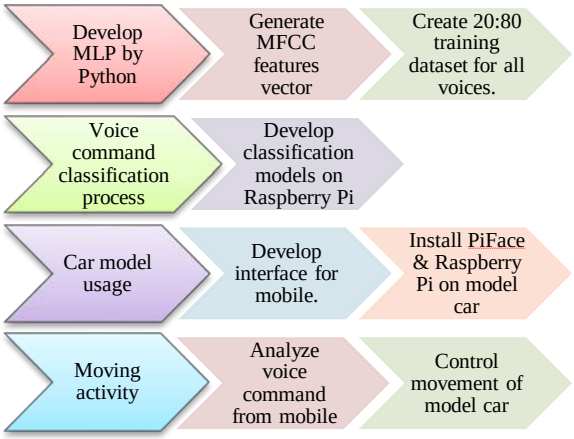

Fig. 5 The overall structure of the Thai voice command application system.

The multi-layer perceptron neural network is used as the classification model in this research. Each network can have numerous layers, and each layer can have multiple perceptron or neural. The structure of 1 perceptron is consists of the input values, weights between perceptron, bias of each perceptron and 2 computation functions. The first function will consolidate the value of all inputs those already adjusted by weights. The second function is called activation function or transfer function that will transfer value from the first function to the output value. The sigmoid is utilized as a transfer function in this research, the output results from sigmoid is displayed in figure 6.

Firstly, all input features to each node are adjust by multiply to weight it connected, and summarize including with the bias of each node as

$$z = \sum_{i=1}^{n} w_i x_i + bias \qquad (2),$$

x_i represent the value of input feature, and

w_i is the weight connected to each input feature.

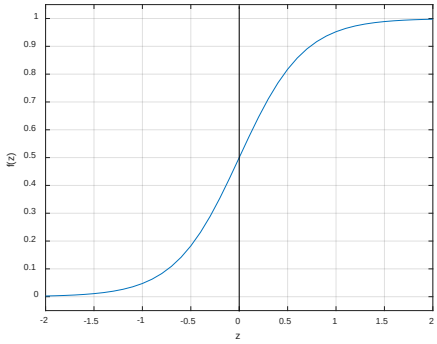

Fig. 6 The output from sigmoid function.

The value of z is fed to the sigmoid activation function to calculate the output as

$$f(z) = \frac{1}{(1+e^{-z})} \qquad (3).$$

In the machine learning literature, the adaptive learning is the most important function. The learning process of MLP, delta rule is applied in the backpropagation algorithm. The parameters of each node, including the weights and bias connected to it. The small values call delta of weight (Δw) and delta of bias ($\Delta bias$) are calculated from miss classification error and use for the parameter modification. After the feed forward process, the errors from output are calculated by

$$e_n = t_n - y_n \qquad (4).$$

The output calculated from the feed forward process is represented by y_n, the target value is represented by t_n. The error is used to calculate Δw as Eq. (5).

$$\Delta w(n) = \eta \delta(n) x(n) \qquad (5),$$

where η is learning rate , the local gradient is represented by $\delta(n)$, which can be separate into 2 cases. The gradient for output layer calculate from Eq. (6) and for hidden layer calculate from Eq. (7).

$$\delta(n) = e(n)\varphi'(z(n)) \qquad (6),$$

where $\varphi'(z(n))$ is the first derivative of the activation function. For hidden layer

$$\delta(n) = \varphi'(z(n))\sum_{i=1}^{n}\delta_i(n)w_i(n) \quad (7),$$

where $\sum_{i=1}^{n}\delta_i(n)$ is the summation gradient of the adjacent layer. The weights of the next iteration is updated by

$$w(n+1) = w(n) + \Delta w \quad (8).$$

The bias value of each neuron are updated similarly to the weights as displayed in Eq. (5) to Eq. (6). The initial parameters of learning rate is 0.1, the maximum iteration of the training process is 10,000 rounds. The sigmoid is utilized as the activation function for every node. As the experimental results shown earlier, the structure of MLP with 2 hidden layers result preferable execution over 1 hidden layer.

4. EXPERIMENTAL RESULTS

The primary experiment is to find the best neural network structure for 5 voices command classification. Three analyses have been done to compare whether using 1 model to classify 5 voices or 5 models for 5 voices. The experimental results are appeared in table 3.

Table 3. The average accuracy results of using 1 model comparing with using 5 models.

Expt. No.	5 fold cross validation of all voices command	
	1 network	5 networks
1	78.80%	84.00%
2	78.00%	85.60%
3	79.20%	83.20%

The structure of 1 network has 4 layers, including input layer with 20 nodes, first hidden layer with 15 nodes, second hidden layer with 8 nodes and lastly an output layer with 3 nodes. The other structure is the combination of 5 networks. The nodes of input layer are equivalent to input features. The first hidden layer has 3 nodes, the second input layer has 2 nodes and output layer has 1 node for 1 voice command. Agreeing from table 3, the outcomes from using 5 networks for classification are higher than using only 1 network in all experiments. Furthermore, the learning time of using 1 network, is longer than using 5 networks in 1 structure. Along these lines, the final structure of this research are as demonstrated in figure 7.

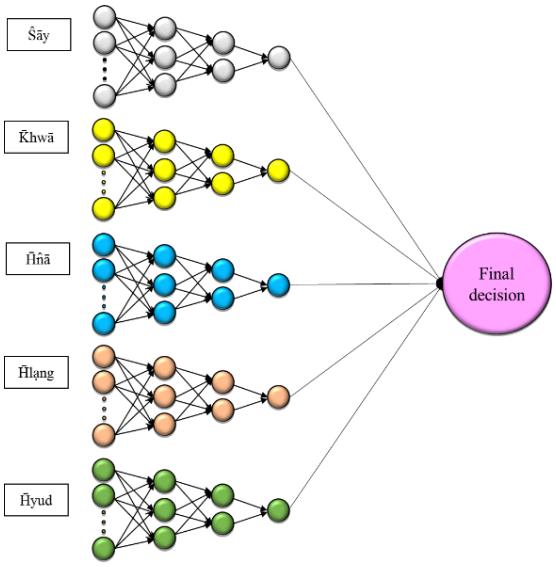

Fig.7 The overall structure of voices commands classification model.

Next experiment using the structure in figure 7 to actualize in PiFace and Raspberry Pi to control the model car and separated experiments into 4 sections. The first experiment, using men voices in both training and testing process. The second experiment, using women's voices in both training and testing process. The third and fourth experiments utilizing both men's voices and women's voices in the training process and using men's voices and women's voices in testing processing separately. The average of 5 fold cross validation, experimental results are shown in table 4.

Table 4 The experimental results based on gender.

Expt. No.	Training	Testing	Avg. Acc.
1	Male	Male	85.33%
2	Female	Female	80.67%
3	Male & Female	Male	76.00%
4	Male & Female	Female	74.67%

The results in table 4 demonstrated that using men's voice as a part of both training and testing process result the most astounding precision. Utilizing blend gender voices in the training process and test with female voices result the lowest accuracy compared to all experiments.

5. CONCLUSION

From aggregate analyses in this research the best proportion for creating a training data set is 20% of desire class and 80% of all other classes. The structure of neural networks is using 2 hidden layers and using 1 network for classifying 1 voice command. Obviously seen from experiment that using the same gender voices as a part of training and testing process results better than using blend gender voices in the training process. The proposed research results demonstrated that the propose methodology is appropriate to apply to utilizing Thai voice command to control the model car. Furthermore, can be used as a prototype model to implement for using Thai voice commands to control a wheelchair or other electronic appliances later on.

6. REFERENCES

[1] Ansari, J.A., Sathyamurthy, A., Balasubramanyam, R., "An Open Voice Command Interface Kit", IEEE Transactions on Human-Machine Systems, Vol. 46, pp. 467- 473.

[2] Bartišiūtė, G., Ratkevičius, K., Paškauskaitė, G., "Hybrid recognition technology for isolated voice commands", Advances in Intelligent Systems and Computing, Vol.432, 2016, pp.207-216.

[3] Srijiranon, K., Eiamkanitchat, N., "Application of neuro-fuzzy approaches to recognition and classification of infant cry", IEEE Region 10 Annual International Conference, Proceedings/ TENCON, 2015.

[4] Srijiranon, K., Eiamkanitchat, N., "Thai speech recognition using Neuro-fuzzy system", ECTI-CON 2015 - 2015 12th International Conference on electrical engineering / Electronics, computer, Telecommunications and Information Technology, 2015.

[5] Xiao, X. , Zhao, S. , Ha Nguyen, D.H. , Zhong, X. , Jones, D.L. , Chng, E.S., Li, H., "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation", Eurasip Journal on Advances in Signal Processing, Vol. 2016, December 2016, pp.1-18.

[6] Mannepalli, K., Sastry, P.N., Suman, M., "MFCC-GMM based accent recognition system for Telugu speech signals", International Journal of Speech Technology, Vol.19, March 2016, pp.87-93.

[7] Ali, S.M., Karule, P.T., "MFCC, LPCC, formants and pitch proven to be best features in diagnosis of speech disorder using neural networks and SVM", Vol. 11, March 2016, pp. 897-903.

[8] Borde, P., Varpe, A., Manza, R., Yannawar, P., "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition", International Journal of Speech Technology, Vol. 18, June 2015, pp.167-175.

[9] Zarrouk, E., Ben Ayed, Y., Gargouri, F., "Hybrid continuous speech recognition systems by HMM, MLP and SVM: A comparative study", International Journal of Speech Technology, Vol. 17, September 2014, pp.223-233.

[10] Ben Nasr, M., Saoud, S., Cherif, A., Optimization of MLP using genetic algorithms applied to Arabic speech recognition, International Review on Computers and Software, Vol.8, February 2013, pp. 653-659.

[11] Park, J., Diehl, F., Gales, M.J.F., Tomalin, M., Woodland, P.C., "The efficient incorporation of MLP features into automatic speech recognition systems", Computer Speech and Language, Vol. 25, July 2011, pp. 519-534.

[12] Pujol, P., Pol, S., Nadeu, C., Hagen, A., Bourlard, H., "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system", IEEE Transactions on Speech and Audio Processing, Vol. 13, January 2005, pp. 14-22.

[13] Kazuhiro O., Minsun L., Shotaro K., "Spatial Interpolation of consolidation properties of Holocene clays at Kobe Airport using an artificial neural network", International Journal of GEOMATE, Vol. 4, April 2015, pp 423-428.

[14] Sakthivel P.B., Ravichandran A., Alagumurthi. N., "Modeling and prediction of flexural strength of hybrid mesh and fiber reinforced cement-based composites using artificial neural network (ANN)", International Journal of GEOMATE, Vol. 10, October 2015, pp 1623-1635.

[15] Hafez Dahlia H., Mahgoub A.G., Abu Kiefa. Mostafa A., "General regression neural network modeling of soil characteristics from field tests", International Journal of GEOMATE, Vol. 12, June 2016, pp 132-139.

[16] Guerra, H., Cardoso, A., Sousa, V., Leitao, J., Graveto, V., Gomes, L.M., "Demonstration of programming in Python using a remote lab with Raspberry Pi", exp.at 2015 - 3rd Experiment International Conference: Online Experimentation, 29 April 2016, pp. 101-102.

Weychan, R., Marciniak, T., Dabrowski, A., "Implementation aspects of speaker recognition using Python language and Raspberry Pi platform", Signal Processing - Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, SPA, December 2015, pp. 162-167.