

Predicting the classification of high vowel sound by using artificial neural network: a study in forensic linguistics

Susanto Susanto^{1,2}, Deri Sis Nanda²

¹Center for Studies in Linguistics, Universitas Bandar Lampung, Bandar Lampung, Indonesia

²Department of English Education, Faculty of Teacher Training and Education, Universitas Bandar Lampung, Bandar Lampung, Indonesia

Article Info

Article history:

Received Apr 17, 2023

Revised May 7, 2023

Accepted May 23, 2023

Keywords:

Artificial neural network

Forensic linguistics

Formant frequency

Normalization method

Vowel sound

ABSTRACT

One of the tasks in forensic linguistics, especially forensic phonetics, is evaluating the speech sounds in the recordings. The speech evaluation aims at identifying and verifying speakers to predict if the sound were spoken by the suspect or not. The common problem in the task is determining which acoustic features of the speech sounds are reliable for the speaker identification and verification. The purpose of this research is studying formant frequencies to predict high vowel sounds /i/, and /u/ by using artificial neural network (ANN). Using three various normalization methods (i.e., softmax, z-score and sigmoid), we utilized multilayer perceptron on backpropagation ANN with the architectural models of 4-5-2, 4-10-2 and 4-20-2. The results show that the z-score normalization method provides higher accuracy than the other two in all formations and the 4-10-2 formation has shown the highest accuracy (92.26%).

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Susanto Susanto

Centre for Studies in Linguistics, Universitas Bandar Lampung

Jalan Zainal Abidin Pagar Alam No.26 Labuhan Ratu, Kedaton, Bandar Lampung, Indonesia

Email: susanto@ubl.ac.id

1. INTRODUCTION

Forensic linguistics is a scientific study of language applied in legal discourse. The results of forensic linguistic studies can be utilized to provide linguistic evidence that can be used as evidence in court or as an additional source of information for criminal investigations. One of the tasks in forensic linguistics, especially forensic phonetics or forensic speech science, is evaluating the speech sounds in the recordings as legal evidence [1]–[3]. In forensic phonetics, there is the application of phonetic knowledge for legal purposes, especially for the identification or verification of speakers involved in crimes or legal cases. It involves the collection and analysis of sound data, including voice recordings, analysis of sound waves, spectrograms, and phonetic parameters such as intonation, pitch, and tempo. It also requires speech analysis techniques for acoustic modeling and speakers' sound profiling.

The speech evaluation in forensic phonetics aims at identifying and verifying speakers to predict if the sounds in the legal evidence were spoken by the suspect or not. The common problem in the task is determining which acoustic features of the speech sounds are reliable for the speaker identification and verification [4], [5]. Natural variations in pronunciation can affect the acoustic features of a speaker's voice, thereby making identification and verification difficult. In addition, the type of sound, both vowel and consonant, can affect the resulting acoustic features. So, it is important to combine acoustic analysis with linguistic analysis and other forensic contexts to ensure the reliability of identification or verification results.

The purpose of this research is studying formant frequencies as the acoustic features in classifying high vowel sounds /i/ and /u/ by using artificial neural network (ANN). Each vowel sound has different acoustic characteristics, so they can be distinguished from one another. However, vowel classification requires complex pattern recognition and machine learning to ensure the accuracy and precision of the classification results. By using formant frequencies as the acoustic characteristics, a classification system can be created to identify and differentiate vowel sounds. Formant frequency refers to as the acoustic resonance of the human vocal tract which is the spectral peak of the spectrum [6], [7]. For an example, the formant frequency of vowel sound /i/ is the concentration of acoustic energy around a certain frequency in its speech sound waves as shown in Figure 1. It has several formants, each at a different frequency and each formant corresponds to a resonance in the vocal tract [6], [8]–[10].

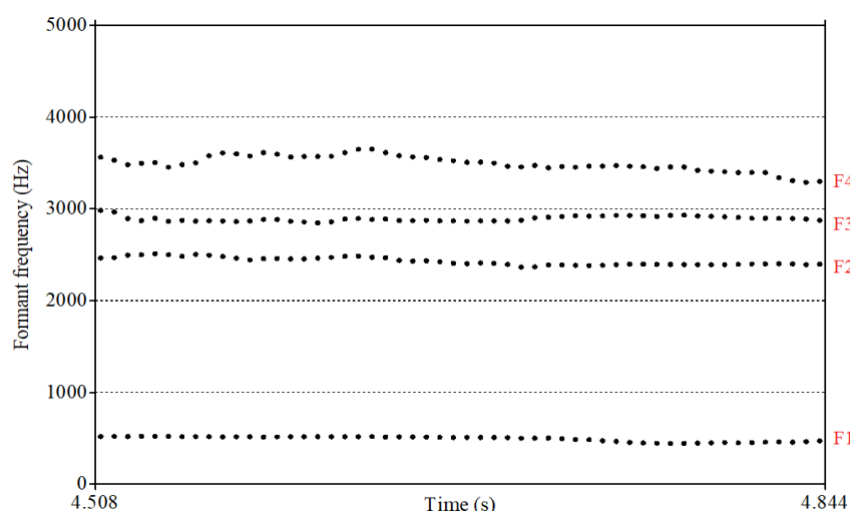


Figure 1. F1-F4 of the vowel sound /i/ (male speaker 12)

We use ANN to predict the classification of high vowel sounds /i/ and /u/. ANN is one method that can be used to predict the class of a data. One of the advantages of ANN is its ability to adapt and be able to learn from the input data so that it can map the relationship between input and output [11], [12]. In addition, ANN is able to predict the output based on the previously trained inputs. ANN has many network structures, including multilayer perceptron [13]–[15]. In this research, we utilized multilayer perceptron on backpropagation ANN with various normalization methods [16]–[19]. We analyzed the data classification using softmax [20]–[22], z-score [23]–[25] and sigmoid [26]–[28] as the normalization methods to obtain optimal classification results in predicting vowel sounds. The prediction is conducted with the formant frequencies F1, F2, F3 and F4 as the input data and the high vowel sounds /i/ and /u/ as the output data as shown in Figure 2.

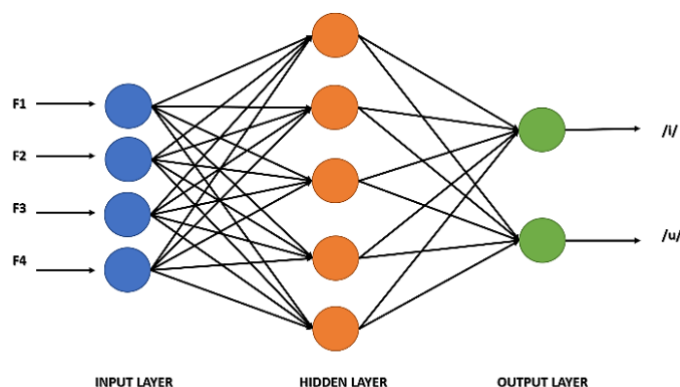


Figure 2. Multilayer perceptron ANN structure (4-5-2)

The normalization method in ANN can simplify the network optimization process and maximize the possibility of obtaining good results. Normalization can help avoid overfitting the ANN model. Overfitting occurs when the model is too complex and too specific for the training data, resulting in decreased model performance when tested on data that has never been seen before. Normalization can help reduce overfitting by normalizing input and output values, resulting in a more generalized, more generalized model. In addition, normalization can help avoid the problem of gradients that exceed the limit, either vanishing gradients or exploding gradients. Gradient exceeding the limit can cause problems in model training and cause slower convergence or even stop the training process. By using normalization, input and output values are converted into a normal distribution so that it is more stable and controllable, so that gradient problems can be avoided. Then, normalization can speed up the ANN model training process because it helps normalize input and output values. This makes it possible to use a higher learning rate, so that the training process can be carried out more quickly. And also, normalization can improve the accuracy of ANN models by reducing errors generated by abnormal input and output values. With normalization, input and output values will be converted into a normal distribution that is more controllable, so that the ANN model will be more accurate in predicting the desired output value.

In the research, we utilize only three normalization methods. One of the methods is sigmoid normalization method which converts the input value into a range between 0 and 1 with a sigmoid function. Another method is softmax which converts the value into a range between 0 and 1 using the sigmoid function with utilizing the mean and standard deviation. The last is z-score method which uses the average and standard deviation to normalize each input. In this method, each input is reduced with the mean value and its result is divided by the standard deviation value. The formulas of sigmoid, softmax and z-score are presented in (1), (2), and (3) respectively, where s is the input value, s' is the normalized input value, μ is the mean value and σ is the standard deviation value [11].

$$s' = \frac{1}{1+e^{-s}} \quad (1)$$

$$s' = \frac{1}{1+e^{-\left(\frac{s-\mu}{\sigma}\right)}} \quad (2)$$

$$s' = \frac{s-\mu}{\sigma} \quad (3)$$

2. METHOD

We used a dataset of vowel sounds recorded at the Center for Studies in Linguistics, Universitas Bandar Lampung. The dataset contains the formant frequencies F1, F2, F3 and F4 for the high vowel sounds /i/ and /u/. The number of data is 120,685 with F1 – F4 distribution of the vowel sounds for male speakers (N=46) and female speakers (N=44) shown in Table 1. Data preprocessing is done by normalizing the data into 0 and 1 using softmax, z-score and sigmoid.

Table 1. F1-F4 og high vowels /i/ and /u/ in the dataset

	Male Speakers (N: 46)				Female Speakers (N: 44)			
	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
/i/								
Min	264	1884	2693	3243	252	1854	2899	4220
Max	482	2452	3019	4174	787	2670	3182	4695
SD	63	172	192	206	94	183	210	232
/u/								
Min	335	719	2290	3182	240	641	2196	3496
Max	585	1368	3142	4023	601	1232	3215	4511
SD	65	180	196	202	63	124	182	244

The concept of the backpropagation algorithm is to adjust the network weight by propagation of the error from output to input. During training, the network minimizes errors by estimating weights and stops at minimum squared error (MSE) 0.05 or a maximum iteration of 1,000 epochs. The activation function is used with a learning rate of 0.01. The minimization procedure was carried out with gradient descent backpropagation with adaptive gain and sigmoid activation function. The ANN architecture is one input layer, one hidden layer, and one output layer. In the input layer, the neuron is the formant frequency of vowel sound with four variables, namely F1, F2, F3 and F4. In the output layer, there are two neurons, namely the results of classifying the high

vowel sounds /i/ and /u/. For the hidden layer, there are 5, 10 and 20 neurons for the architectural models of 4-5-2, 4-10-2 and 4-20-2 respectively. The stages of the research are shown in Figure 3.

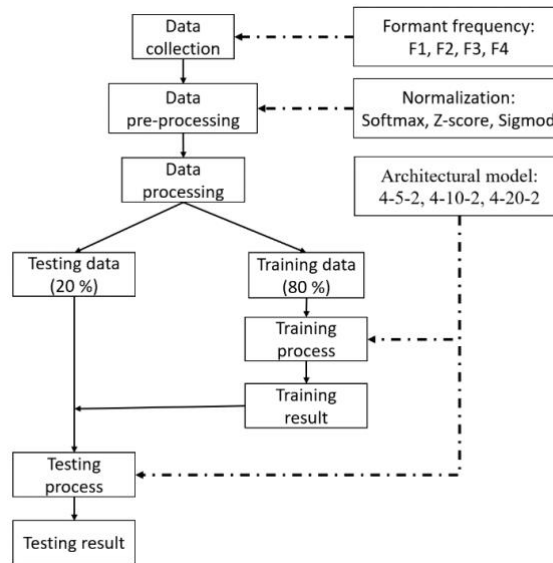


Figure 3. Research stages

3. RESULTS AND DISCUSSION

This experiment was carried out in two processes, namely the training process and the testing process. Derived from 120,685 data records, the training process uses 80% of the data, by randomizing the data from male and female voice. While the remaining 20% is for the testing process. In each variation of the experiment, one hundred repetitions were carried out. Each repetition in the training process is given an initial random weight value and the number of iterations is obtained to achieve convergence. The experiment stops at the minimum squared error (MSE) 0.05 or at the maximum iteration 1,000 epochs and is assumed to have reached convergence and produces a weight that will be used for testing. Each training weight that has converged is used for testing. The classification data from the testing results are compared with the actual classification data so that the amount of data that is predicted to be correct and those that are predicted to be incorrect is obtained. The evaluation of the experiment was carried out by taking the average epoch and accuracy of one hundred tests. The test results in this study can be seen in Table 2 for the accuracy level in each formation and normalization method and Figure 4 for each linear epoch distribution.

Table 2. The results of epoch average and accuracy

Formation	Normalization	Epoch	Accuracy
4-5-2	Softmax	84	86.45%
	Z-score	198	91.14%
	Sigmoid	406	89.48%
4-10-2	Softmax	196	83.24%
	Z-score	307	92.26%
	Sigmoid	643	85.81%
4-20-2	Softmax	179	88.51%
	Z-score	484	91.87%
	Sigmoid	672	89.73%

For the comparison of the average epochs of various normalization methods shown in Table 2, it can be seen that, in all formations, softmax has the lowest average epoch, i.e., 84 epochs in 4-5-2 formation, 196 epochs in 4-10-2 formation, and 179 epochs in 4-20-2 formation. That means the softmax method has the shortest time to achieve convergence. While the comparison of the average accuracy, z-score has the highest accuracy in all formations, i.e., 91.14% in 4-5-2 formation, 92.26% in 4-10-2 formation, and 91.87% in 4-20-2 formation. Based on the test results seen in Table 2, it can be considered the z-score normalization

method is the best method in normalizing the input data of formant frequencies F1-F4 to predict the high vowel sounds /i/ and /u/. And it can also be considered that the 4-10-2 formation is the best architectural model used in this research.

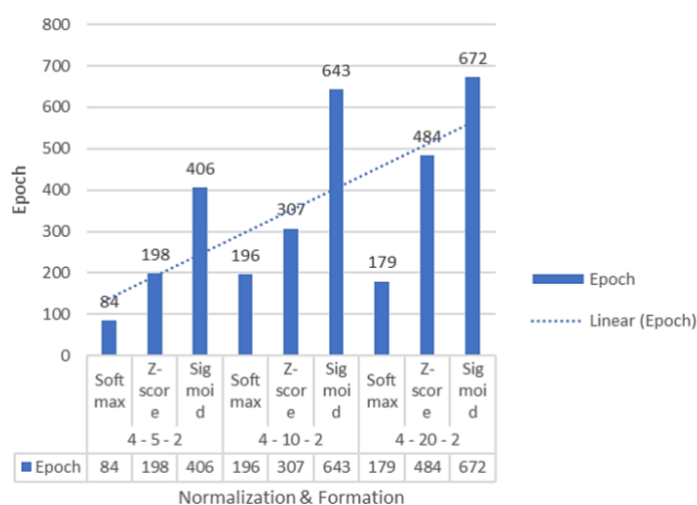


Figure 4. Linear epoch distribution

4. CONCLUSION

In predicting the classification of high vowel sounds /i/ and /u/ by using the four input variables of formant frequencies F1-F4 in this study, the results showed that the prediction can obtain an accuracy of 92.26% by using the backpropagation artificial neural network. We utilized the normalization methods of softmax, z-score and sigmoid and the architectural models of 4-5-2, 4-10-2 and 4-20-2. The highest level of accuracy can be obtained with the architectural model of 4-10-2 and the normalization method of z-score. It is also concluded from the research that the softmax normalization method has the shortest time to achieve convergence although it did not achieve the best accuracy comparing with the z-score and sigmoid normalization methods. For further research, normalization methods and other architectural models can be used to compare with the results obtained in this study. In addition, it can be done with different input variables or different classifications of vowel sounds. It is hoped that this will be a contribution in forensic linguistics, especially forensic phonetics in identifying or verifying sound data as legal evidence.

ACKNOWLEDGEMENTS

The authors would like to thank the Indonesian Ministry of Education, Culture, Research, and Technology for supporting the research [grant number: 1424/SP2H/LT/LL2/2021, and grant number: 1360/LL2/PG/2022]. We thank our research assistants at the Center for Studies in Linguistics, Universitas Bandar Lampung for helping us in analyzing the data by using the artificial neural networks.




REFERENCES

- [1] M. Jessen, "Forensic phonetics," *Linguistics and Language Compass*, vol. 2, no. 4. Wiley, pp. 671–711, May 2008, doi: 10.1111/j.1749-818X.2008.00066.x.
- [2] F. Nolan, "Forensic phonetics," *Journal of Linguistics*, vol. 27, no. 2, pp. 483–493, Sep. 1991, doi: 10.1017/S0022226700012755.
- [3] H. Hollien, *Forensic phonetics*. London: Pinter, 2013.
- [4] S. Susanto, W. Zhenhua, W. Yingli, and D. S. Nanda, "Forensic linguistic inquiry into the validity of F0 as discriminatory potential in the system of forensic speaker verification," *Journal of Forensic Sciences & Criminal Investigation*, vol. 5, no. 3, Sep. 2017, doi: 10.19080/jfsci.2017.05.555664.
- [5] H. Hollien and J. H. Bradford, *The acoustics of crime: the new science of forensic phonetics*, vol. 90, no. 3. Springer Verlag, 1991.
- [6] W. J. Hardcastle, J. Laver, and F. E. Gibbon, *The handbook of phonetic sciences*. Chichester, 2007.
- [7] P. Rao and A. Das Barman, "Speech formant frequency estimation: evaluating a nonstationary analysis method," *Signal Processing*, vol. 80, no. 8, pp. 1655–1667, Aug. 2000, doi: 10.1016/S0165-1684(00)00099-2.
- [8] G. Fant, *Speech acoustics and phonetics*, vol. 24. Dordrecht: Springer Netherlands, 2005.
- [9] A. C. Cohn, J. Clark, and C. Yallop, *An introduction to phonetics and phonology*, vol. 68, no. 1. Oxford: Wiley-Blackwell, 1992.
- [10] W. H. Chapman, E. Olsen, I. Lowe, and G. Andersson, *Introduction to practical phonetics*. High Wycombe: Summer Institute of Linguistics, 1989.

- [11] J. Fulcher, *Artificial neural networks*, vol. 16, no. 3. Springer Verlag, 1994.
- [12] D. Graupe, *Principles of artificial neural networks (3rd Edition)*. New Jersey: World Scientific, Cop, 2013.
- [13] O. Rudenko, O. Bezsonov, and O. Romanyk, "Neural network time series prediction based on multilayer perceptron," *Development Management*, vol. 17, no. 1, pp. 23–34, May 2019, doi: 10.21511/dm.5(1).2019.03.
- [14] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: architecture optimization and training," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, p. 26, 2016, doi: 10.9781/ijimai.2016.415.
- [15] J. Rynkiewicz, "Efficient estimation of multidimensional regression model using multilayer perceptrons," *Neurocomputing*, vol. 69, no. 7-9 SPEC. ISS., pp. 671–678, Mar. 2006, doi: 10.1016/j.neucom.2005.12.008.
- [16] R. Acharya, J. Pal, D. Das, and S. Chaudhuri, "Long-range forecast of Indian summer monsoon rainfall using an artificial neural network model," *Meteorological Applications*, vol. 26, no. 3, pp. 347–361, Mar. 2019, doi: 10.1002/met.1766.
- [17] W. Yu, A. S. Poznyak, and X. Li, "Multilayer dynamic neural networks for non-linear system on-line identification," *International Journal of Control*, vol. 74, no. 18, pp. 1858–1864, Jan. 2001, doi: 10.1080/00207170110089816.
- [18] R. El Hamdi, M. Njah, and M. Chtourou, "Multilayer perceptron training using an evolutionary algorithm," *International Journal of Modelling, Identification and Control*, vol. 5, no. 4, pp. 305–312, 2008, doi: 10.1504/IJMIC.2008.023515.
- [19] F. Gironi and T. Poggio, "Networks and the best approximation property," *Biological Cybernetics*, vol. 63, no. 3, pp. 169–176, Jul. 1990, doi: 10.1007/BF00195855.
- [20] K. Banerjee, C. Vishak Prasad, R. R. Gupta, K. Vyas, H. Anushree, and B. Mishra, "Exploring alternatives to softmax function," *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications, DeLTA 2021*, pp. 81–86, Nov. 2021, doi: 10.5220/0010502000810086.
- [21] J. Zhou, X. Jia, L. Shen, Z. Wen, and Z. Ming, "Improved softmax loss for deep learning-based face and expression recognition," *Cognitive Computation and Systems*, vol. 1, no. 4, pp. 97–102, Nov. 2019, doi: 10.1049/ccs.2019.0010.
- [22] P. Blanchard, D. J. Higham, and N. J. Higham, "Accurately computing the log-sum-exp and softmax functions," *IMA Journal of Numerical Analysis*, vol. 41, no. 4, pp. 2311–2330, Aug. 2021, doi: 10.1093/imanum/draa038.
- [23] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using Z score transformation," *Journal of Molecular Diagnostics*, vol. 5, no. 2, pp. 73–81, May 2003, doi: 10.1016/S1525-1578(10)60455-2.
- [24] U. Dauda and B. M. Ismail, "A study of normalization approach on K-means clustering algorithm," *International Journal of Applied Mathematics and Statistics*, vol. 45, no. 15, pp. 439–446, Nov. 2013.
- [25] C. Cheadle, Y. S. Cho-Chung, K. G. Becker, and M. P. Vawter, "Application of z-score transformation to Affymetrix data," *Applied bioinformatics*, vol. 2, no. 4, pp. 209–217, 2003.
- [26] Y. V. Koteswararao and C. B. Rama Rao, "Single channel source separation using time-frequency non-negative matrix factorization and sigmoid base normalization deep neural networks," *Multidimensional Systems and Signal Processing*, vol. 33, no. 3, pp. 1023–1043, May 2022, doi: 10.1007/s11045-022-00830-2.
- [27] P. Chandra, "Sigmoidal function classes for feedforward artificial neural networks," *Neural Processing Letters*, vol. 18, no. 3, pp. 185–195, Dec. 2003, doi: 10.1023/b:nepl.0000011137.04221.96.
- [28] S. Narayan, "The generalized sigmoid activation function: competitive supervised learning," *Information Sciences*, vol. 99, no. 1–2, pp. 69–82, Jun. 1997, doi: 10.1016/S0020-0255(96)00200-9.

BIOGRAPHIES OF AUTHORS



Susanto Susanto    is a senior lecturer at English Education Study Program, Teacher Training and Education Faculty, Universitas Bandar Lampung (UBL), Indonesia. He is also the Head of Centre for Studies in Linguistics UBL. He received his BA in English Literature from Universitas Islam Sumatera Utara, Medan, Indonesia, MA in English Applied Linguistics from Universitas Negeri Medan, Medan, Indonesia, MA in English from Central Institute of English and Foreign Languages, Hyderabad, India, and PhD in Linguistics and Phonetics from English and Foreign Languages University, Hyderabad, India. In his education, he has joined workshops on Artificial Neural Networks, and Python for Data Science and Machine Learning. He conducted postdoctoral research at Shanghai Jiao Tong University, China and Massachusetts Institute of Technology, USA. Some of his major interests are linguistics, phonetics, language metafunction, discourse analysis, forensic linguistics and artificial intelligence. He can be contacted at email: susanto@ubl.ac.id.



Deri Sis Nanda    is a senior lecturer at English Education Study Program, Teacher Training and Education Faculty, Universitas Bandar Lampung (UBL), Indonesia. She is also the Head of English Education Department UBL. She received her BA in English Literature from Universitas Islam Sumatera Utara, Medan, Indonesia, MA in English Literature from Central Institute of English and Foreign Languages, Hyderabad, India, and PhD in English Literature from English and Foreign Languages University, Hyderabad, India. During her education, she has actively joined workshops on Machine Learning Fundamentals, Introduction to Artificial Intelligence, Artificial Neural Networks, and Deep Learning. She is the member of Indonesian Community for Forensic Linguistics. Her major interests include English literature, postcolonial literature, English education, cyber literature, forensic linguistics and artificial intelligence. She can be contacted at email: derisisnanda@ubl.ac.id.