Human Voice Recognition Using Artificial Neural Networks

Heri Nurdiyanto[a], Hendra Kurniawan[b], and Sri Karnila[c]
[a]
Informatic Enggenering Departement, STMIK Dharma Wacana, Indonesia
[b]Information System Departement IBI Darmajaya, Indonesia
[c]Information System Departement IBI Darmajaya, Indonesia

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Sound is one of the unique and distinguishable parts of the human body. Voice recognition technology is one of biometric technology that does not require much cost and does not require specialized equipment. One of the techniques of speech recognition is with artificial neural networks, where this method uses a working principle similar to the workings of the human brain. This thesis aims to apply artificial neural networks to voice recognition and create programs that simulate this method using Matlab 7.1 software.

The data used in the form of sound recordings are converted into numerical values with the Linear Predictive Coding process. The steps taken in Linear Predictive Coding include Pre-emphasis process, frame blocking, windowing Autocorrelation Analysis, Linear Predictive Coding Analysis, and change the Linear Predictive Coding parameter to the cepstral coefficient. This cepstral coefficient is a series of observations used as inputs on artificial neural networks, and will also be used for the training and testing process. In this research, artificial neural network architecture used is Learning Vector Quantization. In the process of Learning Vector Quantization neural network training using data as many as 35 votes, with learning rate 0.01, max depth 100, dec alpha 0.02 and min alpha 0,00001. Validation test results for 15 votes, obtained the conclusion that 73.34% of all validation votes successfully recognized..

1. Introduction

The sound is one means to recognize one's character. Humans can verify a person's condition by hearing his voice, for example, the gender, speaker's identity, accent, speech, emotion, and health conditions of the speaker.(Andics et al., 2010) Along with technological developments arises the phenomenon of computing model needs for speech recognition which is not only beneficial to science but also the for practical applications, for example on voice-based security systems.(Perrachione, Del Tufo, & Gabrieli, 2011) Voice-based security systems can be applied in various areas of life for both industry and household needs such as employee attendance checking systems, home-based security systems, voice-aligned passwords and many other examples.(Van Lancker & Kreiman, 1987) Technology using the computers is multiplying, almost every individual in the world needs a computer as a tool to solve the problem. Nearly all analog systems are replaced with computerized systems. The advantage is that automated systems are easier to control. In this case, for example, managing in recognizing an object. (Kisilevsky et al., 2003)Computers are endeavored to be able to work closer to the workings of the human brain.(Khashei, Zeinal Hamadani, & Bijari, 2012)

Artificial Neural Network is an alternative new computational system in the form of biological nervous system modeling so that the new computing system is capable of operating and has properties like the original neural network.(Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005) (Intelligence, 2010) Artificial Neural Network applications have been very diverse both in technology and other fields such as trade, defense, medical, insurance, and banking. One use in the field of technology is in the area of further pattern recognition related to the field of Computer Vision (Machine / Computer Vision) that tries to imitate the human eye and brain system capabilities to form and interpret the image.(Candini et al., 2014) (Ding, Li, Su, Yu, & Jin, 2013) (Johnson, 2011) Because of the complex nature of the sound type pattern, the process of identifying voice signals is aided by computational methods capable of extracting unique features of voice signals. (Ghiassi & Saidane, 2005) To meet this requirement is used artificial neural network method that has proven reliable in the process identification of image patterns, including picture sound signals in graphical form.(Bänziger, Grandjean, & Scherer, 2009) (Benítez, Castro, & Requena, 1997) This study uses a network of the artificial neural network to recognize the sound signal pattern. Synthetic neural network techniques have been widely utilized in various critical areas of the pattern recognition system: image, sound, time series prediction, and others. In this study made a system using artificial neural network back propagation method (back propagation) for voice recognition(Johnson, 2011) (Ekici & Aksoy, 2009) This system is expected to be utilized in giving computer commands, voice dialing, and others. The backpropagation neural network model is used here along with Linear Predictive Coding (LPC) and Fast Fourier Transform (FFT) methods used as initial processors. Here are the variations of structure and network parameters (number of hidden layer nodes, size step size, and momentum) to get optimum network performance.(Goh, 2005) The search for these structures and parameters aims to allow the network to learn and recognize sound with the smallest possible errors quickly.(Drapeau, Gosselin, Gagnon, Peretz, & Lorrain, 2009) Signals can be defined as physical quantities that vary over time or other independent variables that store information. Examples of messages are a human voice, morse code, the voltage across the telephone cord, the variation in the intensity of light on an optical fiber used on a telephone or computer network, etc. Signals can be classified into several types: continuous time signals, discrete time signals, signal value constant signal separate value, random signal, and

nonrandom signal.(Golan, Baron-Cohen, Hill, & Rutherford, 2007) (Karaboga & Akay, 2007) The signal analysis is the activity of extraction of information contained in a message. Linear Predictive Coding (LPC) is one of the most potent conversation analysis techniques available and provides excellent quality and efficient feature extraction for use in calculations.(Beauchemin et al., 2006) LPC was first used in 1978 to create a synthesizer of the speech signal. LPC analyzes by estimating formants, separating formants from signals, called inverse filtering processes, and assessing the intensity and frequency of the remaining conversation signal, called a residue (Kriegstein et al., 2005). Because conversation signals vary over time, these estimates are performed for every small piece of the message, called a frame this study, the scope of the problem in the stages of voice signal pattern recognition, the creation of program listing for network pattern recognition, and testing the accuracy of the network.(Jarng, 2011) The purpose of this research is to design a training program and testing of the artificial neural network to recognize the sound signal pattern and to test the accuracy of the system.(Cui & Xue, 2009)(Li & Ma, 2010).

2. Method

The system designed is a system that can recognize the voice input with various types and variations of voice input by the formulation of the problem. The system can only remember the voice of a person who has been trained, so that if the input noise is not stored in the database, then it cannot be identified. But an expected variation of voice signal input will not affect the success rate of the system so that the results obtained are valid.(Obin, Roebel, & Bachman, 2014) The designed system will recognize a person's voice after going through several processes. This can be analogized if the first time a person met a voice recognizable, the condition that is the definition of voice acquisition. Then there is a difference in the sound of a new person known to the music of others; this process is called feature extraction.(Carlini et al., 2016) In everyday life, if you meet someone sometimes forget who the person in question, therefore need a means of remembering, this method is called classification with Artificial Neural Network(Singhal & Swarup, 2011). The voice input used in this study was obtained from the sampling of the voice of adult women, adult men and children with the pronunciation of "HALO" for 5 data train each person. (Dalton & Deshmane, 1991)

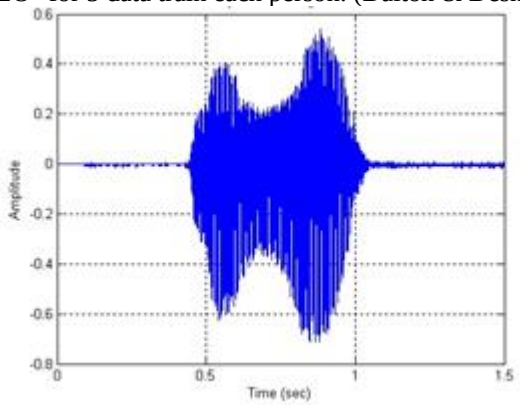


Fig. 1. Sound Signal Cut

The recorded sound is sound in WAV format In the sound signal section of the recording that is considered unnecessary cutting is required. This process is used to uniform the input format so that the signal obtained with a maximum of recording time is determined. (Obin et al., 2014) The truncated sound signal will then go through the initial processing stage:

Filtering, The input sound signal will be filtered using Band Pass Filter (BPF). The purpose of the input sound filtering is to pass the sound signal portion of the passband only, eliminating noise-noise at the sound signal inputs in the frequency regions <fco1 and> fco2, and limiting the amplitude of the input signal, thus the magnitude of the audio signal exceeds the expected.

Sampling, The filtered sound signal will be sampled with the Nyquist requirement to obtain a discrete voice signal that will facilitate the capture of its distinguishing characteristics. The sampling process aims to achieve distinct signals to determine the fundamental components of the message and assist the decimation process (reduction in the number of sampling).

Decimation, The decimation is the reduction of the sample on the sampling result. This decimation process aims to reduce the number of examples of voice signals are too much, but the results of the decimation obtained can still represent patterns are removed. Parameter n is a parameter that determines how many times or how many samples you want.

The preemphasis process is performed on the input signal to minimize the signal change area. The sound signal is passed to a digital filter that serves to flatten the spectral signal. In the processing of speech signal preemphasis filter is required after the sampling process. (Drapeau et al., 2009) The purpose of this filtering is to obtain a smoother spectral form of speech signal frequency. Where the spectral shape is relatively high value for low areas and tends to fall sharply to the high-frequency region. The Reshapes process is used to resize the matrix

according to the expected input. (Amato et al., 2013) In Windowing step is done weighting function on each frame that has been formed in the previous step. In the Autocorrelation Analysis phase, each windowed frame is autocorrelated with the highest autocorrelation value, i.e., the order of the LPC analysis. Fast Fourier Transform is done to improve the performance of the system, because with the Fast Fourier Transform then the difference between the sound signal pattern with other sound signal pattern more visible.(Schmidhuber, 2015) Where the data produced by Fast Fourier Transform is an input for Artificial Neural Networks In the Training and Identification with Artificial Neural Network conducted training and sound identification using Backpropagation Neural Network. The output of the system is that the sound detected is either recognized as a target or not recognized. The training process is done until the network gets the target error value as small as possible. Once these conditions are met, we apply the weight of each segment on the system. The masses will be used in the testing process. Backpropagation Artificial Neural Network training through 3 phases: advanced phase, reverse phase, and weight change. In the concept of training is also used to make the practice run smoothly and adequately to get the output as desired.

In this testing process used some test sound for the sound already trained and to test a voice that has not been trained at all. Each vote is tested to see if the sound can be recognized correctly in the original classification, recognized as someone else, or not recognized at all. By the system so that the system is expected to recognize well in the input data, although previously never done learning about the pattern of input data. From the system can be formulated as Grade Success System is Percentage of system success. With Grade Success System can be known the level of system success that has been made, the result can be formulated as follows:

$$\Gamma\Sigma\Sigma = \frac{RTD+RRD}{2} \times 100 \quad (1)$$

Result Training Data is the result of a previously trained data test, whether the system can recognize correctly, does not recognize correctly, or does not recognize it at all. From the system can be formulated as follows:

$$\mathrm{PT}\Delta = \frac{The\ amount\ of\ data\ is\ successful}{Amount\ of\ data} \quad (2)$$

By the system so that the system is expected to recognize well in the input data, although previously never done learning about the pattern of input data. From the system can be formulated as follows:

$$\mathrm{PP}\Delta = \frac{The\ amount\ of\ data\ is\ successful}{Amount\ of\ data} \quad (3)$$

Where the data is successful is the number of test data that is recognized, and the amount of data is the overall test data in the test.

3. Results and Discussion

Trainer voice signals are obtained from 5 respondents, each of which consists of 5 voice signals. The test voice signal is obtained from 5 respondents which each respondent consists of 5 sound signal patterns. Voice signals are stored in Waveform Audio (WAV) format. In the next process Implementation of the system by utilizing Linear Predictive Coding toolbox (LPC) and Artificial Neural Network (ANN) Where success is the number of input data that is successfully recognized, and the amount of data is the overall data in the test. Result Random Data is the result of a completely unrecognized data test in Matlab programming version 7.1. Pattern recognition with backpropagation is a process in which the network can recognize input sounds correctly. However, before recognizing the pattern of input, training must be done with the correct parameter values so that the network can recognize the sound well. The training process in this final project uses bipolar sigmoid activation function. This study was conducted using parameters set on fixed values are; Error target = 0, Learning rate (Alpha) = 0.05, mu (μ) = 10-3.

While the value of system parameters to be altered is Hidden layer = 1, 2, 3, and 4, the number of neurons each hidden layer = 25, 20, 15, 10 and the amount of train data = 5. Change the values of this parameter aims to know the combination of systems that provide the most optimal training time. The training process is done three times for each combination of system parameters

Tabel 1. Network Training Results for five training data with Bipolar Sigmoid Activation Function

Parameter		***Attempt to***	***Training***		***Mean Square Error (MSE)***
Hidden Layer	***Neuron***		***Iterasi***	***Time***	
2	15,10	1	137	45,844	9,127 x 10^{-6}
		2	844	405,58	9,453 x 10^{-6}
		3	112	37,27	1,785 x 10^{-6}

From Table 1 the system parameters provide the most optimal training time with hidden layer = 2 and the neuron of each hidden layer = 15, 10. The average training time obtained to approach target error 0 is 1.7850 x

10-6 and training time is 37.266 seconds with 112 iterations. Hidden layer I neuron I = 15 and Hidden layer II neurons = 10.

Backpropagation testing process aims to measure how much network success based on parameters that have been obtained previously. Here are the results obtained from some network recognition tests Tests with sound training is the first test of the network conducted on the voice input that has been trained previously. This experiment aims to determine whether the system can recognize well or cannot recognize the sound that has been prepared. The results of tissue testing for the bipolar sigmoid activation function can be seen in Table 2.

Table 2. Introduction of Trained Coaching pattern for five training data with Sigmoid Bipolar Activation Function

No	***Input Exercise***	***Recognized as***				
		Fatin	***Abizar***	***Anggara***	***Safira***	***Aktansi***
1	Fatin 1	V				
	Fatin 2	V				
	Fatin 3	V				
	Fatin 4	V				
	Fatin 5	V				
2	Abizar 1		V			
	Abizar 2		V			
	Abizar 3		V			
	Abizar 4		V			
	Abizar 5		V			
3	Anggara 1			V		
	Anggara 2			V		
	Anggara 3			V		
	Anggara 4			V		
	Anggara 5			V		
4	Safira 1				V	
	Safira 2				V	
	Safira 3				V	
	Safira 4				V	
	Safira 5				V	
5	Aktansi 1					V
	Aktansi 2					V
	Aktansi 3					V
	Aktansi 4					V
	Aktansi 5					V

Information : MSE = 1,785 x 106
Iteration = 112
Time = 37,266
number of samples = 25
Success = 25
Failed = 0
Success rate = 100%

After the network training with parameters that have been set each network, then the system whose introduction rate above 98% then will be continued in the testing process. This test process uses five input data voice signal that has not been trained at all. This testing process aims as a network learning process.

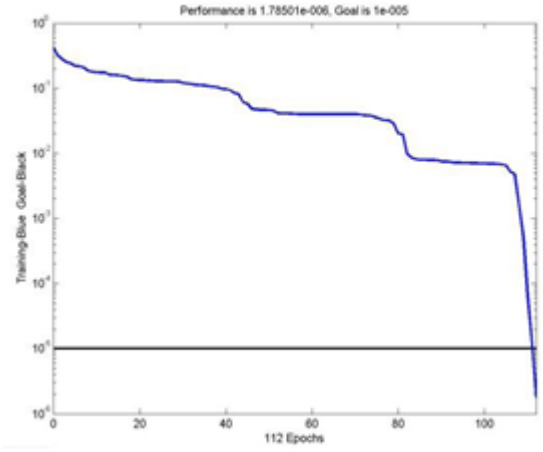
Table 3 Introduction of Trainer Patterns for five training data with the Bipolar Sigmoid Activation Function

No	***Input Exercise***	***Recognized as***				
		Fatin	***Abizar***	***Anggara***	***Safira***	***Aktansi***
1	Fatin 1	V				
	Fatin 2	V				
	Fatin 3					V
	Fatin 4	V				
	Fatin 5					V
2	Abizar 1		V			
	Abizar 2		V			
	Abizar 3					V
	Abizar 4					V
	Abizar 5			V		

3	Anggara 1				V
	Anggara 2		V		
	Anggara 3		V		
	Anggara 4				V
	Anggara 5		V		
4	Safira 1			V	
	Safira 2			0	
	Safira 3	V			
	Safira 4		V		
	Safira 5				
5	Aktansi 1		V		V
	Aktansi 2				V
	Aktansi 3				V
	Aktansi 4				0
	Aktansi 5				V

Information :number of samples = 25
Success = 12
Failed = 13

Level of Success in the process of sound training test for five training data of 100%. While the success rate on the method of testing the voice that has not been trained for five training data of 48%. The success rate is obtained from each election each tested network that has the highest percentage rate. For 5 data train With the comparison of the curve above then the network that has two hidden layers are more effective to use than networks with 1 or 3 hidden layers.



Curve 2 Hidden Layer for 5 test data

Performance analysis of the system is the main parameter to know the success rate of system design. These parameters are RTD, RRD, and GSS. RTD is a successful parameter of identification with the test sound is the previously trained sound. The RTD testing process is performed on the bipolar sigmoid activation function:

$$\text{PT}\Delta = \frac{The\ amount\ of\ data\ is\ successful}{Amount\ of\ data} = \frac{25}{25} = 1{,}0 \qquad (4)$$

In addition to RTD, RRD is also a parameter that determines the level of system success. After the results of the test data that has been trained, then the system is also tested with data that has not been tested at all. It aims to determine whether the system can still recognize the random sounds, as well as the system, recognizes the sound of the train. The equation is obtained as follows:

$$\text{PP}\Delta = \frac{The\ amount\ of\ data\ is\ successful}{Amount\ of\ data} = \text{RRD} = \frac{12}{25} = 0.48 \quad (5)$$

System success is calculated from GSS based on the previous equation. Since the value of RTD and RRD has been obtained, then the success rate of the system can be calculated. Based on the calculation of RTD and RRD obtained GSS as follows.

$$\Gamma\Sigma\Sigma = \frac{RTD+RRD}{2} = \frac{1.00+0.48}{2} \times 100 = 74 \quad (6)$$

4. Conclusion

From the analysis of sound recognition test system using Linear Predictive Coding and Artificial Neural Network Backpropagation for identification of voice signal pattern, it can be concluded as follows: Network with 4 layers consisting of 1 layer input (input), 2 hidden layer (hidden layer) and 1 layer output. The number of neurons in each layer is 24, 15, 10, 5. The value of the characteristics of the network is the value of learning rate = 0.05, the value of mu (μ) = 10-3 and using the bipolar sigmoid activation function. The network success rate for voice testing for 5 trainer data trained with the bipolar sigmoid activation function is 100% and the network success rate for untrained voice testing reaches 74%. The more trained data processed in the network the higher the success rate obtained (voice signal can be recognized).

References

1. Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine, 11(2), 47–58. https://doi.org/10.2478/v10136-012-0031-x
2. Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. NeuroImage, 52(4), 1528–1540. https://doi.org/10.1016/j.neuroimage.2010.05.048
3. Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion Recognition From Expressions in Face, Voice, and Body: The Multimodal Emotion Recognition Test (MERT). Emotion, 9(5), 691–704. https://doi.org/10.1037/a0017088
4. Beauchemin, M., De Beaumont, L., Vannasing, P., Turcotte, A., Arcand, C., Belin, P., & Lassonde, M. (2006). Electrophysiological markers of voice familiarity. European Journal of Neuroscience, 23(11), 3081–3086. https://doi.org/10.1111/j.1460-9568.2006.04856.x
5. Benítez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes? IEEE Transactions on Neural Networks, 8(5), 1156–1164. https://doi.org/10.1109/72.623216
6. Candini, M., Zamagni, E., Nuzzo, A., Ruotolo, F., Iachini, T., & Frassinetti, F. (2014). Who is speaking? Implicit and explicit self and other voice recognition. Brain and Cognition, 92, 112–117. https://doi.org/10.1016/j.bandc.2014.10.001
7. Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., … Zhou, W. (2016). Hidden Voice Commands. Usenix Security, 1–18.
8. Cui, B. C. B., & Xue, T. X. T. (2009). Design and realization of an intelligent access control system based on voice recognition. 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, 1, 229–232. https://doi.org/10.1109/CCCM.2009.5270462
9. Dalton, J., & Deshmane, a. (1991). Artificial neural networks. IEEE Potentials, 10, 1–8. https://doi.org/10.1109/45.84097
10. Ding, S., Li, H., Su, C., Yu, J., & Jin, F. (2013). Evolutionary artificial neural networks: A review. Artificial Intelligence Review, Vol. 39, pp. 251–260. https://doi.org/10.1007/s10462-011-9270-6
11. Drapeau, J., Gosselin, N., Gagnon, L., Peretz, I., & Lorrain, D. (2009). Emotional recognition from face, voice, and music in dementia of the alzheimer type: Implications for

music therapy. Annals of the New York Academy of Sciences, 1169, 342–345. https://doi.org/10.1111/j.1749-6632.2009.04768.x

12. Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. Advances in Engineering Software, 40(5), 356–362. https://doi.org/10.1016/j.advengsoft.2008.05.003
13. Ghiassi, M., & Saidane, H. (2005). A dynamic architecture for artificial neural networks. Neurocomputing, 63(SPEC. ISS.), 397–413. https://doi.org/10.1016/j.neucom.2004.03.014
14. Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. Journal of Experimental Psychology: Learning Memory and Cognition, 31(1), 40–53. https://doi.org/10.1037/0278-7393.31.1.40
15. Golan, O., Baron-Cohen, S., Hill, J. J., & Rutherford, M. D. (2007). The "Reading the Mind in the Voice" test-revised: A study of complex emotion recognition in adults with and without autism spectrum conditions. Journal of Autism and Developmental Disorders, 37(6), 1096–1106. https://doi.org/10.1007/s10803-006-0252-5
16. Intelligence, A. (2010). Fundamentals of Neural Networks Artificial Intelligence Fundamentals of Neural Networks Artificial Intelligence. Fundamentals of Neural Networks : AI Course Lecture 37 – 38, Notes, Slides.
17. Jarng, S. S. (2011). HMM voice recognition algorithm coding. 2011 International Conference on Information Science and Applications, ICISA 2011. https://doi.org/10.1109/ICISA.2011.5772321
18. Johnson, B. E. (2011). The speed and accuracy of voice recognition software-assisted transcription versus the listen-and-type method: A research note. Qualitative Research, 11(1), 91–97. https://doi.org/10.1177/1468794110385966
19. Karaboga, D., & Akay, B. (2007). Artificial Bee Colony (ABC) Algorithm on Training Artificial Neural Networks. 2007 IEEE 15th Signal Processing and Communications Applications, 1–4. https://doi.org/10.1109/SIU.2007.4298679
20. Khashei, M., Zeinal Hamadani, A., & Bijari, M. (2012). A novel hybrid classification model of artificial neural networks and multiple linear regression models. Expert Systems with Applications, 39(3), 2606–2620. https://doi.org/10.1016/j.eswa.2011.08.116
21. Kisilevsky, B. S., Hains, S. M. J., Lee, K., Xie, X., Huang, H., Ye, H. H., … Wang, Z. (2003). Effects of experience on fetal voice recognition. Psychological Science, 14(3), 220–224. https://doi.org/10.1111/1467-9280.02435
22. Kriegstein, K. von, Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of Face and Voice Areas during Speaker Recognition. Journal of Cognitive Neuroscience, 17(3), 367–376. https://doi.org/10.1162/0898929053279577
23. Li, Y., & Ma, W. (2010). Applications of Artificial Neural Networks in Financial Economics: A Survey. 2010 International Symposium on Computational Intelligence and Design, 211–214. https://doi.org/10.1109/ISCID.2010.70
24. Obin, N., Roebel, A., & Bachman, G. (2014). On automatic voice casting for expressive speech: Speaker recognition vs. speech classification. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 950–954. https://doi.org/10.1109/ICASSP.2014.6853737
25. Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. Science (New York, N.Y.), 333(6042), 595. https://doi.org/10.1126/science.1207327
26. Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. Neural Networks, Vol. 61, pp. 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

27. Singhal, D., & Swarup, K. S. (2011). Electricity price forecasting using artificial neural networks. International Journal of Electrical Power & Energy Systems, 33(3), 550–555. https://doi.org/10.1016/j.ijepes.2010.12.009
28. Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. Neuropsychologia, 25(5), 829–834. https://doi.org/10.1016/0028-3932(87)90120-5