ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE)
Vol. 10, Issue 5, May 2024

Neural Network-Based AI for Speech Recognition: A Review

[1]Hija Happy, [2]Kamma Pusapuri Madhavi, [3]K.Safeena , [4]K.G.Shreya
[1]hijahappy2004@gmail.com,[2] kpmadhavi703@gmail.com, [3]ksafeena04@gmail.com,
[4]kgshreya2003@gmail.com

ABSTRACT

In artificial intelligence, speech recognition is extremely important as a human-machine interface. Traditional speech recognition techniques have limitations and a low learning structure. The Convolutional Neural Networks or CNNs used in this article enable speech recognition. The ability of this particular kind of neural network to represent spectral correlations in data and lower spectrum variance sets it apart from others. In addition, the paper uses backpropagation to train the neural network. Throughout the entire experiment, the neural network is evaluated using the remaining voice recordings, and five sets of recordings that we personally produced serve as training data. Throughout the entire experiment, the neural network is evaluated using the remaining voice recordings, and a set of recordings that we personally produced serve as training data.

INTRODUCTION

The signal may get noisier according to the surroundings. Speakers can occasionally add noise on their own [4]. A telephone or microphone captured acoustic signal is transformed into speech during the speech recognition process. Probably the most effective and intuitive form ofcommunication is speech. Humans have been dependent on spoken communication since they were small toddlers and have developed all the necessary abilities on their own. A similar effortless, natural, and effective means of communication between humans and robots is something that both seek. They therefore prefer voice as a tool over other heavy-duty devices like keyboards and mouse. Speaking presents several difficulties, nevertheless, as human articulators and voices are biologic organs that function apart fromour conscious minds. Speech is greatly affected bya variety of factors, including backgrounds, echoes, pitch, volume, speaking, word pronunciation, imperfections, emotional condition, gender, andspeed [1] . Speech recognition is sometimes known as automatic speech recognition, or ASR.

PATTERN IDENTIFICATION

Satellite navigation, vision in computers, machine intelligence, biology, psychology, and medicine are just a few of the scientific and engineering fields that depend on automatic detection, description, categorization, and grouping patterns. fingerprint image, the voice signal, a human face, or handwritten text in cursive can all be used as templates. The next two tasks can involve identifying and categorizing the pattern.

a) The process of supervised categorization, where patterns of inputs are assigned to pre-defined classes, is referred to as discriminant analysis.

b) Clustering without knowledge of the template is known as classification without supervision. With the classes either specified in a supervised categorization by the software architect or based on learning, the problem at hand can be identified as one of classification or problems with classification.

I. NEURAL NETWORKS

In neural networks, a multitude of nodes combine to generate a special kind of account. The term "node" refers to the standard unit of account; whether the contract can function in parallel depends on how these nodes interact with some of the academics and with each other. Neural networks can be defined as mathematical models comprising relatively simple components that mimic certain aspects of biological systems, allowing for concurrent information processing.

They are a fundamental type of algorithm, often represented in graphs or charts. These algorithms are organized into various schemes and are designed to tackle a wide range of challenging problems.

Neural network activity is demonstrated through the coding and categorization process.

A) The ability to withstand and handle noise effectively.

B) Adaptability to process distorted images.

C) High resistance to categorize broken or partially decomposed images.

D) Utilizing parallel processes with multiple operating units that are interconnected and share information.

E) A high capacity to adapt and adjust internal mechanisms using logarithms and powers, allowing for lasting changes.

F) The ability to work with nonlinear operations, including the ability to map and interpret noisy data, making them valuable for ratings and attributions.

II. NEURAL NETWORK TYPES

A. Process Operations

The method involves iteratively selecting the score that deviates the most from the mean. If this score exceeds a predetermined threshold, it is removed, and recalculations of the mean and standard deviation estimates are made. This approach leads to a notable improvement in estimating mean and variance when the process is applied to a small number of utterances. Investigations that are text dependent and independent have both used a telephone multisession database. This study introduces a speech model and demonstrates how algorithmic research improvements are influenced by the speaker's experience. Mini-problems are utilized throughout the system design process to showcase the relationship between the two factors.

The present study specifically examines voice recognition tasks using artificial neural networks. It is found that the performance in identifying phonemes within words is affected by the size of the neural network used. Additionally, the study explores the parameterization of speech signals.

Moreover, the study utilizes the method of selecting the score that deviates the most from the mean iteratively. If this score exceeds a predefined threshold, it is removed and the estimates of the standard deviation and mean are calculated once again. This method proves to be highly beneficial in estimating mean and variance when working with a limited number of utterances.

Both text-dependent and text-independent investigations have utilized a multisession telephone database for their experiments.

Finally, the study introduces a speech model and highlights the relationship between algorithmic research and system improvements in the context of speaker experience. This relationship is illustrated through the use of mini-problems during the system design process.

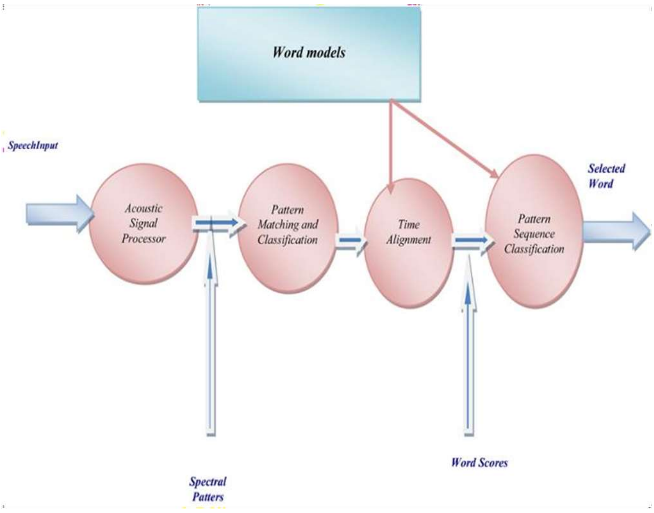
Word models
SpeechInput
Acoustic Signal Processor
Pattern Matching and Classification
Time Alignment
Pattern Sequence Classification
Selected Word
Word Scores
Spectral Patters

Figure 1 : Voice transmission system schematic illustration.

B. The Identification Process Algorithm for Recognition

Employing a computer-based neural network to assess the proximity between specific audio characteristics; assigning characteristics to the sound range; choosing word boundaries according on the sound input; comparison using the norms of dictionaries.

Once the audio data has undergone processing, a neural network is utilized to receive the processed audio signal as an input and produce an array of signal segments as an output. Each signal segment represents a range of values that describe the amplitude spectrum of a signal, which are then used to generate the neural network's signal output through calculation. Table 2 displays thenumerical values obtained during this process, with "I" denoting the total values for a set, and each row representing a set of values for each frame. Abbreviations and Acronyms may be included in the table to provide a shortened version of commonly used terms.

In order to ensure clarity and understanding, it is important to provide definitions for abbreviations and acronyms the first time they are used in the text, even if they have already been defined in the abstract. However, commonly known abbreviations such as SI, ac, and dc do not need to be defined. Abbreviations that include periods should not have spaces between them, for example, "C.N.R.S." instead of "C. N. R. S." Additionally, it is advisable to avoid using abbreviations in the article title unless they are necessary and cannot be avoided.

C. Formulae

The following sequential stages must be finished in order to determine the neural network's output [9]:
Step 1: Establish the undetected layer of context for each neuron;
Step 2: Program the synthetic neural network using the initial set of integers. Determine the hidden layer's output.

$$y_j=f(\Sigma^1{}_{i=1}\,\omega_{ij}\,X_{1i}+\beta_i+\omega_j X_j)$$

The non-linear activation function, or F(x)

$$y_j = \frac{1}{1+e^{-\alpha S}}$$

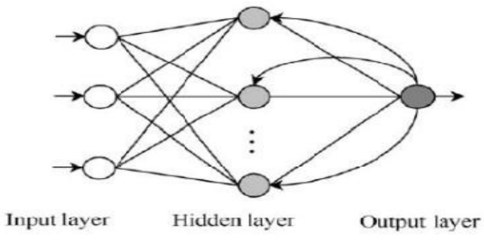

Figure 2: A neural network's architecture with a feed table

The artificial neural network can be programmed using the initial set of integers. Discover what the buried layer's output is

$$E=1/2N\,\Sigma^N{}_{i=1}(y_{ki}-d_i)^2$$

A neuron's prototype is found in the biology of nerve cells. A neuron's soma, or cell body, is made up of two kinds of exterior, wood-like branches: dendrites and axons. The plasma, which includes chemicals required for the production and transfer of the components required to form a neuron, and the core, which houses genetic material, are located inside the structure of the cell. Via its dendrites, a neuron can connect with other neurons. Moreover, an axon, which ends at synapses and forks into a fiber, can carry signals produced by body cells to another neuron. A mathematical model that depicts the neuron's democratic ratio

$$y=f(s),swx_iw_iwb$$

where s is the input signal, y is the signal output neuron, f-function is active, wi is the synapse and weight (b)-offset value, and n is the number of inputs to the neuron. Neuron block diagram: The activation function is denoted by F(S); the output signal, y, is a set of weights; neural control handles the conversion result of nonlinear thresholds and carries out basic operations such weighted summation. Because of the body of its simple homogenous pieces, one of the characteristics of the neural network technique is its capacity to handle complicated interactions between entities. The functional characteristics of the network as a whole are defined by the structure of relations. Neural networks' properties are determined by the evolutionary history and functional characteristics of individual neurons.

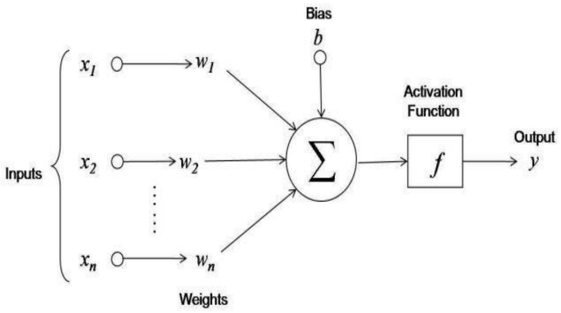

Figure 3: A representation of a neuron's technical model is shown

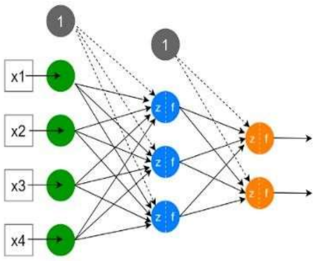

Figure 4: Two-layer neural network's structural diagram

Applying the two-layer neural network of direct action model yields the following mathematical representation.

$$y(\theta)=F_i(\Sigma^{nh}{}_{j-1}W_{ij}\ f_i\ (\Sigma^{nh}{}_{j-1}\ W_{ij}\ \Theta_{j\ +}\ W_{jo}\ ) + W_{jo}$$

where nh is the number of neurons in the buried layer and nφ φ is the size of the neural network; The vector θ represents the programmable parameters of the neural network, such as weights,and neuron-by offset (wji, Wij); the activation function of the neurons in the hidden layer is fj(x), and the activation function of the neurons in the output layer is Fi(x). The most important feature ofthe neural network approach is its capacity to handle data in parallel. When there are multiple international brain connections present, this feature allows the signal data processing process to be significantly accelerated. the potential for real-time audio signal processing. Some neural network properties are claimed to be present in artificial intelligence.

CONCLUSION

It may be possible to turn the speech transmission method for elders into an app in the future, which would make phone use more simple for them and gradually ease their transition to accepting smartphones. There is a lot of promise for thefuture in voice recognition technologies andapplications tailored to specific groups. Efficient phone use can also be facilitated by similar strategies for kids or those with mental health conditions such as autism. Besides becoming amore practical way for them to use their phones, it would definitely help us understand more about the worlds within their brains. The integration of the sections on health-related instructions and emergency procedures is one component that mainly helps determine the user's health state. "Health section," a recently introduced section.

REFERENCES

[1] S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique for digital communications channel equalization using radial basis function networks,” IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[2] J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[3] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, “Rotation, scale, and translation resilient public watermarking for images,” IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

[4] Choudhary, A. and Kshirsagar, R. (2012) Process Speech Recognition System Using Artificial Intelligence Technique. International Journal of Soft Computing and Engineering (IJSCE), 2.

[5] Ovchinnikov, P.E. (2005) Multilayer Perceptron Training without Word Segmentation for Phoneme Recognition. Optical Memory & Neural Networks (Information Optics), 14, 245-248.

[6] Guo, X.Y., Liang, X. and Li, X. (2007) A Stock Pattern Recognition Algorithm Based on Neural Networks. Third International Conference on Natural Computation, 2.

[7] Dai, W.J. and Wang, P. (2007) Application of Pattern Recognition and Artificial Neural Network to Load Forecasting in Electric Power System. Third International Conference on Natural Computation, 1.

[8] Shahrin, A.N., Omar, N., Jumari, K.F. and Khalid, M. (2007) Face Detecting Using Artificial Neural Networks Approach. First Asia International Conference on Modelling & Simulation.

[9] Lin, H., Hou, W.S., Zhen, X.L. and Peng, C.L. (2006) Recognition of ECG Patterns Using Artificial Neural Network. Sixth International Conference on Intelligent Systems Design and Applications, 2.

[10] Al Smadi, T.A. (2013) Design and Implementation of Double Base Integer Encoder of Term Metrical to Direct Binary.

[11] K.M. He, X.Y. Zhang, S.Q. Ren and J. Sun, Deep Residual Learning for Image Recognition. In Proc. of CVPR, pp. 770–778, 2015.

[12] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. Journal of Machind Learning Research, 12, 2493–2537.

[13] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. In: Proc. of EMNLP, pp. 1746–1751.

[14] Cho, K., Merrienboer, B., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proc. of EMNLP.

[15] Hunt, A. J., Black, A. W. (1996) Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. of ICASSP, pp. 373–376.

[16] H. Kawai, T. Toda, J. Ni, et al. (2004) XIMERA: A new TTS from ATR based on corpus-based technologies. In Proc. of Fifth ISCA Workshop on Speech Synthesis.

[17] Ling, Z. H., Wang, R. H. (2007) HMM-based hierarchical unit selection combiningKullback-Leibler divergence with likelihoodcriterion. In Proc. of ICASSP, pp. 1245–1248.

[18] Black, A. W., Zen, H., & Tokuda, K. (2007). Statistical parametric speech synthesis. Proc. ICASSP, 4, 1229–1232.

[19] Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical Parametric Speech Synthesis. Speech Communication, 51(11), 1039–1064.

[20] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (1999) Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Proc. of Eurospeech, pp. 2347–2350.

[21] Ling, Z. H., Kang, S. Y., Zen, H., Senior, A., Schuster, M., Qian, X. J., Meng, H., & Deng, L. (2015). Deep Learning for Acoustic Modeling in Parametric Speech Generation. Journal of IEEE Signal Processing Magazine, 32, 35–52.

[22] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural probabilistic language model. Journal of Machine Learning Research, 1137–1155.

[23] Mikolov, T., Karafiat, M., Burget, L. (2010) J. BHonza^ Cernocky and S. Khudanpur, BRecurrent neural network based language model. In Proc. of INTERSPEECH, pp. 1045–1048.

[24] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12, 2493–2537.

[25] Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. Proc. of ACL, 1, 873–882.

[26] T. Mikolov, K. Chen, G. Corrado and J. Dean (2013) Efficient estimation of word representations in vector space. In Proc. of CoRR.

[27] Kang, S., Qian, X., & Meng, H. (2013). Multi-distribution deep belief network for speech synthesis. In Proc. of ICASSP, pp.7962–7966.

[28] Ling, Z.-H., Deng, L., & Yu, D. (2013) Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis. In Proc. of ICASSP, pp. 7825–7829.

[29] Fan, Y.-C., Qian, Y., Xie, F.-L. & Soong, F. K. (2014) TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks. In Proc. of Interspeech, pp.1964–1968.

[30] Siniscalchi, S. M., Yu, D., Deng, L., & Lee, C.-H. (2013). Exploiting Deep Neural Networks for Detection-Based Speech Recognition. Neurocomputing, 106, 148–157.

[31] C.-H. Lee and S. M. Siniscalchi, BAn Information-Extraction Approach to Speech Processing: Analysis, Detection, Verification and Recognition,^ Proceedings of the IEEE, Vol. 101, No. 5, pp. 1089–1115, May 2013.

[32] Caruana, R. (1997). Multitask learning. Machine Learning Journal, 28, 41–75.

[33] Wu, Z., Valentini-Botinhao, C., Watts, O., & King, S. (2015). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In Proc. of ICASSP, pp. 4460– 4464.

[34] Tokuda, K., Kobayashi, T., & Imai, S. (1995). Speech parameter generation from HMM using dynamic features. Proc. of ICASSP, pp. 660–663.

[35] Song, E., Joo, Y.-S., & Kang, H.-G. (2015) Improved Time Frequency Trajectory Excitation Modeling for a Statistical Parametric Speech Synthesis System. In Proc. of ICASSP.