Journal of Cases on Information Technology
Volume 24 • Issue 5

Influencing Neutrosophic Factors of Speech Recognition Technology in English Collection

Xizhi Chu, College of Foreign Languages, Xi'an Aeronautical University, Shaanxi, China*
Yuchen Liu, School of Automation, Xi'an Aeronautical University, Shaanxi, China

ABSTRACT

Based on the research and analysis of speech recognition system and neural network principle, combined with English related decision tree, this paper completes the construction and design of English speech recognition based on hybrid frame and series frame neural system network. Combined with relevant information, this paper analyzes the influencing factors of language recognition technology in English collection under neural network. This paper proposes a method of English Corpus collection. Through the speech recognition technology under the neural network, experiments are carried out by using the neural network algorithm, K-means clustering algorithm, and HMM/ANN cascade system to analyze the influencing factors of speech recognition technology based on the neural network in English collection. Finally, from the English accent, the authors make a detailed analysis of the proportion of speech fuzziness, speech speed, and environmental interference, so as to draw a conclusion.

KEYWORDS

Analysis of Influencing Factors, English Collection, Neural Network, Speech Recognition Technology

1. INTRODUCTION

In the past 10 years, the research of neural networks has made great progress and successfully solved many modern problems. He shows excellent intelligence characteristics for practical problems that computers are difficult to solve. With the improvement of speech recognition performance, the popularization and application of speech recognition technology in mobile Internet terminal equipment has been promoted. Due to more application requirements, the development of speech recognition technology based on deep neural networks has accelerated development. In recent years, the application of speech recognition technology has become more and more extensive, and how to improve the robustness of the speech recognition system has attracted the attention of researchers. In real life, factors such as speech sounds and emotions have a great impact on the robustness of the speech recognition system, and the neutrosophic performance of the natural speech recognition system becomes very unstable. Deep neural network has achieved excellent learning ability and high stability in the fields of speech synthesis, information classification, etc. It is the current research hotspot of speech recognition system based on neural network. This article analyzes the factors affecting the application of deep neural networks in speech recognition. Speech recognition technology can improve the accuracy of speech recognition and the efficiency of speech recognition under the application of deep neural network. Speech recognition based on deep neural network is the only way for speech

DOI: 10.4018/JCIT.295859 *Corresponding Author

recognition. Based on the research and analysis of speech recognition system and neural network principles, this paper combines English-related decision trees to complete the design and construction of an English speech recognition system based on hybrid framework and series-frame neural network.

With the development of the times, more and more people are studying the impact of speech recognition technology on people's lives. Some people have done a lot of research in English. Li P's research is biased towards the influence of external interference factors on the accuracy of recognition in English intelligent recognition. He used image recognition technology, combined with the actual elements of English recognition to set corresponding influencing factors, and proposed a reliable multi-body coding recognition method (Li, et al. 2020). His research is very helpful for English collection, but it is mainly based on image recognition technology. With the advancement of science and technology, artificial intelligence as an emerging tool has penetrated into our lives. The application of modern smart devices is becoming more and more extensive, and it is also inseparable from our lives. Jiang J produced a mobile smart terminal represented by a smart phone, which is a powerful intelligent computing and networking device with stronger perception and interaction capabilities. The experimental results show that the method he proposed is more efficient and intelligent (Jiang, et al. 2021). The above studies have explained the power of the language recognition system, some of which can be used for reference, but the effect is not very large. The method of estimating the accuracy of speech recognition without using manual transcription reference is beneficial to the research and development of automatic speech recognition technology. Ogawa A proposed an estimation method of recognition accuracy based on error type classification (ETC). ETC is an extension of confidence estimation. By adding these CSID probabilities separately, it is possible to estimate the number of CSIDs of speech data and two standard recognition accuracy measures, namely the percentage of correctness and word accuracy (WAcc) (Ogawa, et al. 2016) without using reference transcription. His research has a great effect on identification, but it basically lies in the ETC direction. Recovering sentence boundaries from speech and its transcripts is critical for readability and downstream speech and language processing tasks. Xu C suggested using a deep recurrent neural network to detect sentence boundaries in broadcast news by modeling the rich prosody and vocabulary features extracted at the position of each word (Xu, et al. 2018). XuC's research uses neural network modeling to extract text, which is of great help to our neural network-based research, but it lacks speech recognition. Speech recognition is the process of understanding human or natural language speech through a computer. In this regard, a syllable-centric speech recognition system recognizes the syllable boundaries in the input speech and converts them into corresponding written scripts or text units. The proper neutrosophic segmentation of acoustic speech signals into syllable units is an important task for the development of high-precision speech recognition systems. Panda SP proposed an automatic syllable-based segmentation technology for segmenting continuous speech signals in Indian language at syllable boundaries (Panda, et al. 2016). This segmentation technique is very practical, if applied in English, it will be very helpful to our research. With the rapid development of Internet technology, network-assisted teaching systems have developed rapidly. Li Y combines deep neural networks and game theory to reduce the dimensionality of the model through singular value decomposition and reconstruction to reduce the amount of data and improve the accuracy of the experiment (Li, 2021). The above studies all have very detailed applications and explanations for language recognition, but the practicality is not very high. All of the above references are mostly related to neural networks and speech recognition technology. Although some are not relevant, they still have a certain reference value.

The innovation of this paper is to design a simple and effective English related acoustic model based on the combination of hybrid framework and English related decision tree, and an English related acoustic model based on the single-level optimal network structure of tandem framework and the combination of two-layer cascaded mlan and English related decision tree, which improves the average recognition rate of English test sets. The absolute improvement over the baseline English related acoustic model was 3.12% and 2.22%, respectively. It is of great help to English acquisition.

2. BASED ON THE NEUTROSOPHIC METHOD OF SPEECH RECOGNITION TECHNOLOGY ON NEURAL NETWORK IN ENGLISH COLLECTION

2.1 Speech Recognition Technology

Speech recognition technology is an important development direction of computer technology. The arrival of the multimedia era urgently needs to solve the automatic speech recognition technology. Speech recognition technology is an important technology popularized by hundreds of millions of people. It is an epoch-making technology in the information industry and will undoubtedly become an important function of computers in the future. It can be seen from the linear prediction coefficients of speech recognition technology based on neural network that neural network plays a very important role in speech recognition technology.

As a result, the routing protocol frequently performs calculations, and this change must achieve the uniformity of the entire network, that is, it takes a certain time to reach the route convergence. During this period, the route selection is inaccurate, which affects the network transmission. If this kind of change is frequent and causes a long-term instability of the entire network, it is called network fluctuation. Voice recognition technology is very permeable, slowly entering our lives, changing our lifestyle anywhere. Figure 1 shows the background and theme of the application (Wangi, et al. 2019).

Speech recognition hides the possibility of wide application. With the rapid development of the information industry and the Internet, the demand for speech recognition is becoming more and more urgent (Gábor, et al. 2016). For example, in voice services, it is very difficult to use neutrosophic traditional voice technology to query stock transactions, flight status, station ticket status, etc. in real time. Voice technology can solve these huge information dynamic queries. In addition, people hope to make general information equipment through embedded systems and send Internet information to thousands of families through the phone. Hope that in the near future, the combination of sound, image and animation will be even more exciting. Network technology has produced a new generation of multimedia in virtual reality. Voice technology has greatly lowered the threshold for people to enter the information age and opened up many new information services and applications. That is now forming a new industry. Therefore, countries in the world not only include sound technology in high-tech research plans, but also play an important competitive market role in the information industry of this century.

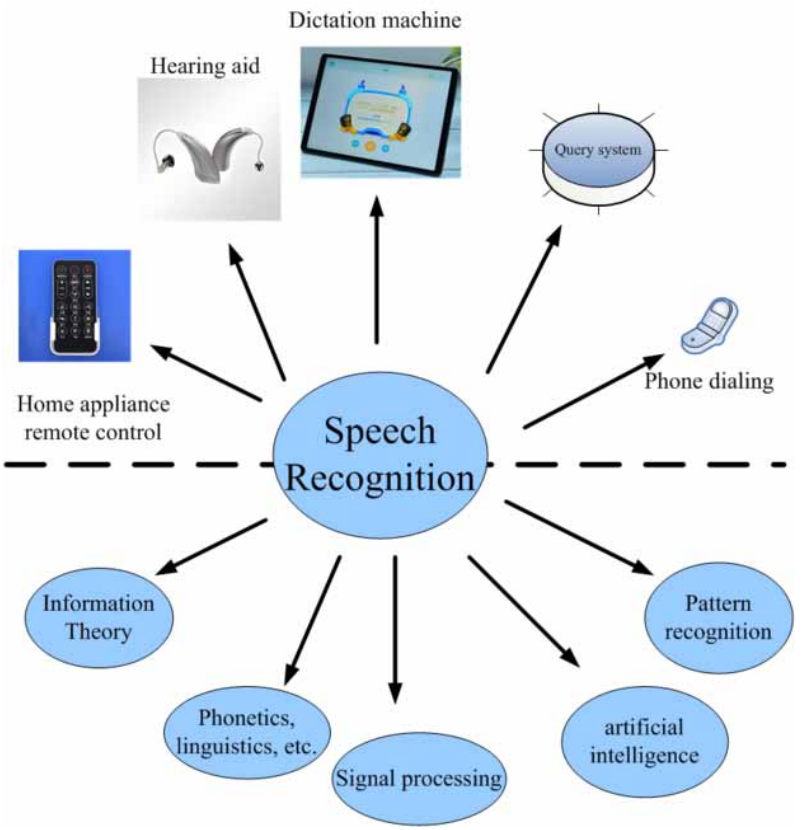
(1) Linear Prediction Neutrosophic Coefficient (Lpc)

Linear prediction coefficients are a set of prediction coefficients based on linear prediction analysis (Zhihao, 2021), and its main theoretical basis is related to the digital model of speech signal generation. Figure 2 shows the simplified model block diagram of sound production:

Time-varying filter is an indispensable device on the pipeline of conveying medium. It is usually installed at the inlet of pressure reducing valve, pressure relief valve, fixed water level valve, Fanggong filter and other equipment. The filter consists of a cylinder, a stainless steel filter screen, a sewage part, a transmission device and an electrical control part. As long as the vocal tract model parameters can be accurately estimated (Nagyeong, 2017), it is possible to use this parameter as a feature value for speech recognition. The time-varying digital filter X(z) in the figure can be simulated by the following all-pole system:

$$\mathrm{X}(z) = \frac{1}{1 - \sum_{i=1}^{p} a_1 z^{-1}} \tag{1}$$

Figure 1. Application of speech recognition



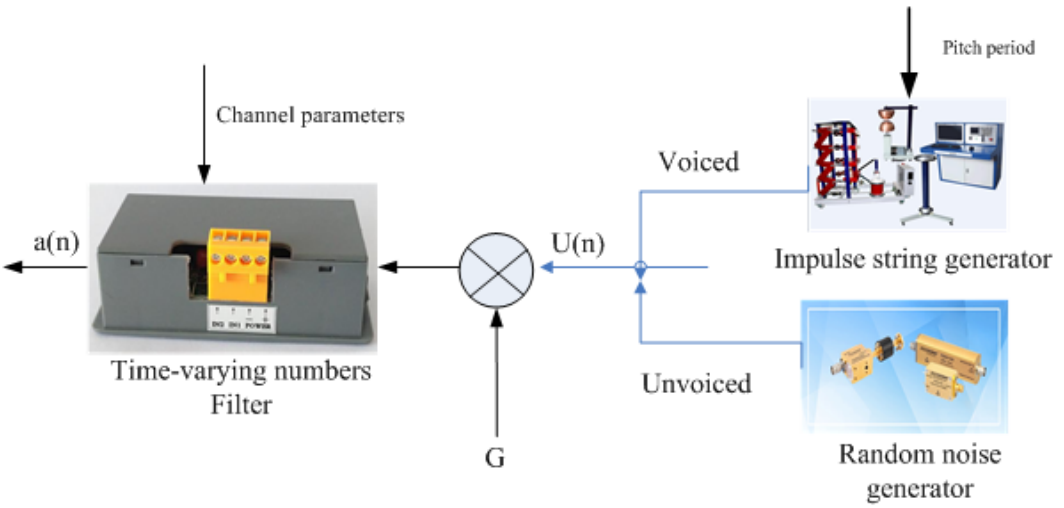
The idea of LPC model (Tu, et al. 2018) is: based on the correlation between speech signal sampling points, use the first sampling points to estimate the current or future sample values, that is, the estimated value of speech signal uses the linear combination of previous s sampling values to weigh the recent previous sample points and speculate the current or future sample values. That is, the estimated value of the sound signal is weighted and estimated by a linear combination of previous sampling values, that is:

$$\hat{\mathrm{H}}(n) = \sum_{i=1}^{s} a_i h(n-i) \tag{2}$$

Then the difference between $\mathrm{H}(n)$ and $\hat{\mathrm{H}}(\mathrm{n})$ is the error a(n) introduced by the s-order linear predictor a(n) :

$$a(n) = h(n) - \sum_{i=1}^{s} a_i h(n-i) \tag{3}$$

Figure 2. Simplified diagram of sound model



$a_i(i=1,2,\ldots,s)$ is the prediction coefficient of the s-time linear predictor (Zia, et al. 2019). The main problem of the LPC model is to determine these coefficients based on the principle of minimizing the total mean square error.

2.2 Neural System Network

Logical thinking refers to the process of reasoning according to logical rules; it first converts information into concepts and expresses them with symbols, and then logically reasoning according to the serial mode according to symbolic operations; this process can be written as serial instructions and let the computer implement. However, intuitive thinking is to integrate distributed stored information, and the result is a sudden idea or a solution to a problem.

(1) Neuron Learning Algorithm

The research of artificial neurons originated from the theory of brain neurons. At the end of the 19th century, in the fields of biology and physiology. It is recognized that the complex nervous system is composed of a large number of neurons. Hebb algorithm is the basis of neuron learning algorithm, and most other learning algorithms have evolved from this. The basic idea of the Hebb algorithm is: a neuron λ_i receives the output from another neuron λ_j as a condition for its own excitation. If both neurons are excited, then the weight ω_{ij} between them will increase. The algorithm can be described in the following form:

$$\Delta\omega_i = \mu y x_i \tag{4}$$

The formula (1) $\Delta\omega_i$ is the result of adjusting the i-th weight, μ is the learning speed coefficient, which can be set by yourself.

(2) Network Learning Algorithm

The network learning algorithm aims to meet the performance(Shukla, et al. 2018) required by the network, and thus determines the adjustment method of the weights of interconnected neurons. The network learning algorithm is divided into supervised (comparing the actual value obtained by the sample with the expected value) and unsupervised (characterizing the statistical value between the samples in the sample space between the weights of the neurons). The core is the activation function, and the following are commonly used:

1) Linear Function:

$$y(x) = k * x + z \tag{5}$$

2) Threshold Function:

$$y(x)\begin{cases}1, x \geq z\\ 0, x < z\end{cases} \tag{6}$$

3) S-Type Function:

$$y(x) = \frac{1}{1 + e^{-\alpha x}} (0 < y(x) < 1) \tag{7}$$

2.3 Influencing Factors of Speech Recognition Technology in English Collection

(1) Hmm/Ann Cascade (Tandem) System

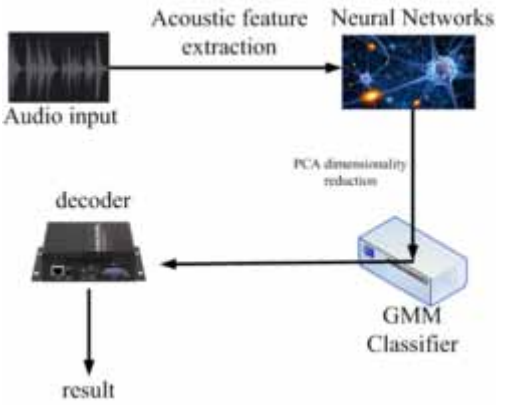
Hidden Markov Model (HMM) is the simplest dynamic Bayesian network. It is a particularly famous directed graph structure. It is mainly used for modeling time series data and is widely used in speech recognition, natural language processing and other fields. In the word segmentation algorithm, Hidden Markov is often used as an algorithm that can discover new words. Through massive data learning, it can identify people's names, place names, and new words on the Internet one by one, which has a wide range of application scenarios. In the hybrid HMM/ANN system, the ANN (Pawar, et al. 2021) is used to estimate the scaled state output probability of the HMM, and then the viterbi search algorithm is used for decoding. However, in the case of continuous speech recognition with a large vocabulary, a more complex neural network is required, and the amount of calculation will be greatly increased, which is not the optimal method. The cascade method of hybrid HMM/ANN exists in theory, so multiple experiments are needed to verify its scientificity.

To this end, the article proposes a cascading method, which is improved on the basis of hybrid HMM/ANN(Dutta, et al. 2019), combining the feature distinguishing processing ability of ANN and the representation ability of the mixed Gaussian model for complex distributions. The feature sequence of speech is used as the input of ANN, and the classification output result of ANN is regularized as the training feature of the traditional Gaussian mixture model. ANN is used as the extraction of discriminative feature(Zoughi, et al. 2019), which almost completely quotes the advantages of the HMM method. Self-adaptation can be done on the basis of HMM, which will be a better way to improve the recognition performance of the system.

In addition, according to different experimental requirements and task requirements, multiple ANN networks based on different conditions (bandwidth, accent, gender, background noise, etc.) can

be trained, combined with the original acoustic feature parameters, and then Gaussian model training can be performed. It have to expand the performance ability of characteristic parameters. Figure 3 shows the basic cascaded ANN/HMM system block diagram:

Figure 3. Block diagram of cascaded HMM/ANN system



(2) k-Means Clustering Algorithm

The K-mean clustering algorithm selects the training set that has been classified from the data, and uses the data mining classification technology on the training set to establish a classification model and classify the test data.

K-means clustering method, also known as fast clustering method, is a clustering method proposed and named by MacQueen in 1967. This method(Hamdan, et al. 2016) is mainly used for good geometric and statistical significance.

Given the sample set $\mathrm{A}=\left\{\mathrm{a}_1,\mathrm{a}_2,\ldots,\mathrm{a}_n\right\}$, the number of categories z can be known in advance, and the criterion function of the sum of squares of the error is selected as the desired objective function: $Y_e=\sum_{i=1}^{c}\sum_{j=1}^{n}x_{ij}\left\|a_j-w_i\right\|^2$. In:

$$x_{ij}=\begin{cases}1\blacklozenge\text{if a}_j\text{ belongs to category i}\\0\blacklozenge\text{if }a_j\text{ not in category i}\end{cases} \tag{8}$$

w_i is the cluster center of the i-th category(Tachioka, et al. 2016; Hourri, et al. 2020). $\mathrm{Y_e}$ represent the sum of errors generated when classifying samples of category c. If the cluster centers are different, the value of $\mathrm{Y_e}$ is also different. The cluster that minimizes $\mathrm{Y_e}$ is the best solution under the standard of error sum of squares. The K-means algorithm continuously adjusts the clustering centers (Mudimbe-Boyi, et al. 2016), so that the error sum of squares benchmark gets the minimum value. The specific algorithm is as follows:

1) Select the sample as the initial set point (clustering seed), or divide all samples into initial clusters, and use the center of gravity (average) of each cluster as the initial set point.

2) Classify all samples except the total points one by one, and classify each sample into the category closest to the total point. The total points of the category (Nagajyothi, et al. 2018; Kaur, et al. 2018) will be updated to the current average value of the category until all samples are classified.
3) Repeat step 2) until all samples cannot be distributed.

3. EXPERIMENTS ON THE NEUTROSOPHIC INFLUENCING FACTORS OF SPEECH RECOGNITION IN ENGLISH COLLECTION

This experiment mainly analyzes the influence of English accent factors on English collection, combined with the application of neural network (Swietojanski, et al. 2016) in language recognition, makes a correlation analysis of English collection, and explores the influencing factors.

3.1 Introduction to Experimental Data

The data involved in all the experiments in this article mainly include the collection of English corpus from various regions, mainly in the style of reading aloud(Hernandez, et al. 2020), and fully consider the balance of voice segments, and try to use some internationally-used standards in the production. Both data are sampled at 16K Hz with 16bit precision and stored in Wave format.

3.2 English Phonetics Database in Various Regions

English is the most popular language in the circulation area, but the number of mother tongues is the third in the world, second only to Chinese and Spanish. English pronunciation in different regions will be different. Our experiment is to collect recording corpus from different regions to find out the influence of English pronunciation on English collection (Abdel-Basset, et al. 2017).

(1) Recorder

There are 200 people in each region, distributed by age, gender, pressure, and education level.

(2) Recording corpus

There are two types of recording corpus: spoken language and read aloud. The specific content of each recording corpus is shown in Table 1. ν

3.3 Select Feature Parameters

The characteristic parameter reflects the characteristics of the sound, which is different from the RGB value that reflects the color of the image. It is difficult to determine a very accurate and effective method for the characteristics of speech, especially its own. At present, the widely used applications include: Mel frequency distribution coefficient method, linear prediction situation spectrum coefficient (LPCC), linear prediction coefficient (LPC), etc.

Linear prediction coefficient (LPC): the analysis of linear prediction coefficient focuses on the principle of people's oral vocalization. It is regarded as an all-pole model with a digital filter. The signal at a specific moment can be weighted by the value of the previous moment and several moments. In the form of, the estimated weighting factor is obtained from the minimum mean square error. The weighting coefficient is a linear prediction coefficient. When represented by n past sampling points, the prediction is called n predictions.

Assuming that the speech signal of a meal is $a_n(n = 0,1,2,...,n-1)$, where the first p values are known, the p+1th value can be expressed as:

Table 1. The pronunciation corpus of each speaker in the four major English accent regions

Recording item	Pronunciation method	Content description
0	Spontaneous speech	Oral narration of natural monologue: 3-5 minutes
1-15	Spontaneous speech	Answer 15 questions
16-35	read speech	20 common spoken sentences
36-150	read speech	Phonetic balanced sentences around 110 words

$$\hat{a}(n) = -\sum_{i=1}^{p} x_i a(n-i) \tag{9}$$

The error from the ideal value is called the prediction error, and $e(n)$ is calculated as:

$$e(n) = a(n) - \hat{a}(n) = \sum_{i=0}^{p} x_i a(n-i), x_0 = 1 \tag{10}$$

When the mean square error ρ reaches the minimum, the predicted result is closer to the actual value:

$$\rho = \mathrm{E}\left[e^2(n)\right] = \rho_{\min} \tag{11}$$

Assuming that the speech signal is an autoregressive signal, when calculating the prediction coefficients of each frame, only the correlation function must be calculated, the Tiptz matrix must be established, and the Revenson-German algorithm must be used to recursively solve it quickly. Linear prediction cepstrum coefficient (LPCC): Linear prediction cepstrum is to convert linear prediction coefficients into cepstrum data. In practical applications, linear prediction coefficients are almost not used, but replaced by their grammatical coefficients. This is because the cepstral coefficients can be reflected more intensively. Through analysis and research, this is an excellent characteristic parameter. The solution of LPCC is obtained through recursion after obtaining LPC. The formula is as follows:

$$\begin{cases} c(n) = a_n + \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c(n=k) 1 < n \le p \\ c(1) = a_1 \\ c(n) = \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c(n=k) n \ge p \end{cases} \tag{12}$$

Similar to other signal transmission systems, the transfer function of the sound channel also expresses the characteristics of speech, which is often expressed in low-time domain spectrum, and the order of the cepstrum is often a value between 10 and 16.

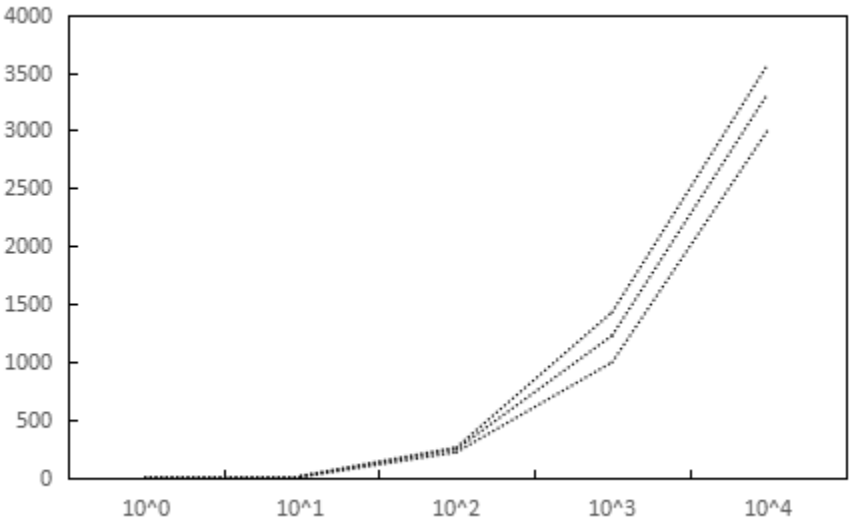
Mel frequency cepstrum coefficient (MFCC): since the human ear does not increase linearly with the increase of sound frequency, instead, it increases exponentially as shown in the figure. The

dynamic range can be reduced through transformation. Mel frequency describes this transformation relationship, and the relationship with linear frequency is shown in Figure 4:

$$f_{mel} = 25951g(1 + \frac{f}{700}) \tag{13}$$

Through the conversion, the result is linear, so that part of the redundant voice information can be removed. Under low frequency conditions, the Mel frequency cepstrum coefficient will still change linearly with time, but the magnitude of the change is not as drastic as under high frequency conditions.

Figure 4. Correspondence between frequency and Mel frequency



4. THE IMPACT OF SPEECH NEUTROSOPHIC RECOGNITION TECHNOLOGY ON ENGLISH COLLECTION

4.1 The Impact of Different Factors on Speech Recognition Technology

Language recognition refers to the process of using machines to analyze language signals, based on the characteristic parameters and grammatical rules of phonetic units such as phonemes, syllables or words, and even the regularity of textual meaning between speeches to make logical judgments to recognize the language. In different time periods, the data obtained through experiments are analyzed (Biswas, et al. 2016), and the collection results are shown in Table 2:

4.2 Proportion Analysis of Influencing Factors in English Collection

Analyzing the data in the table, we get Figure 5:

Table 2. The impact of speech recognition technology on English collection

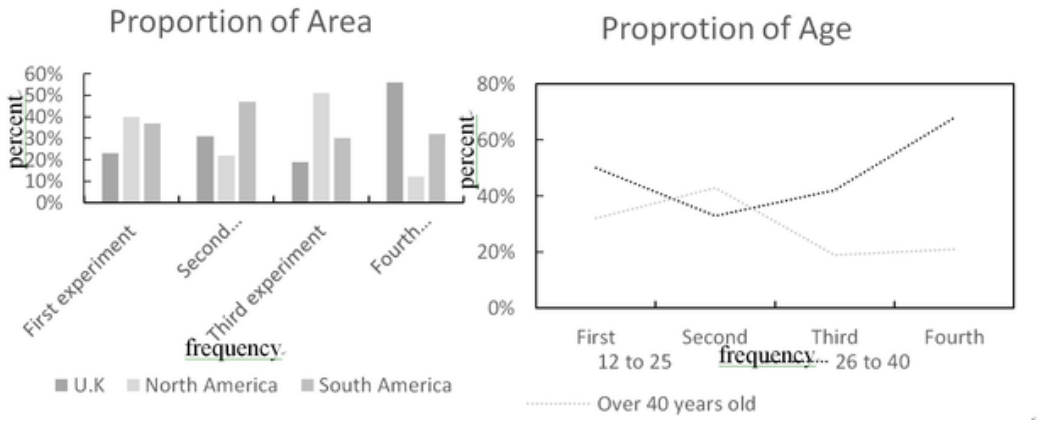
Factor	English accent	Volume of voice information	Speech ambiguity	Speaking speed and intonation	Environmental interference
Percentage	13.4%	8.7%	25.6%	16.9%	35.4%
Number of samples	2683	1745	5127	3381	7088

Through this experiment to choose the standard of the data set and its credibility to discuss, it shows that the experiment has a great credibility, and the data selection also meets the standard. Therefore, this paper comes to the following conclusions. The application of neural network is becoming more and more widely. We study, experiment and analyze the aspects of English Acquisition Based on the language recognition technology based on neural network, and get the conclusions of the influencing factors (Ye, et al. 2018; Huang, et al. 2019). We start from the aspects of English accent, speech information, speech fuzziness, speech speed, speech harmony and environmental interference. Through the neural network algorithm, K-means clustering algorithm and understanding the HMM / Ann cascade system, it is concluded that the environmental factors have the greatest impact on English acquisition (Deli, et al. 2017; Meng, et al. 2021). The noisy environment seriously affects the operation of speech recognition technology, and the amount of speech information has the least impact on English Acquisition, mainly because the current speech recognition technology is powerful enough. It can quickly read the speaker's voice.

5. CONCLUSION

Through the analysis of conclusions drawn from many experiments, there are great changes in the speech recognition technology under the neural network. The neural network algorithm can strengthen the speech recognition technology, and then based on the K-means clustering algorithm to get the English accent and the amount of speech information. The influence of factors such as speech ambiguity, speech speed, intonation, and environmental interference on English collection is finally analyzed, and it is concluded that the influence of speech recognition technology based on neural network under English collection accounts for the largest proportion of 35.4%. The least

Figure 5. The influence of English accent on English collection



accounted for only 8.7%, so it can be seen that the amount of voice information is not a big problem for language recognition. The conclusions drawn in this experiment also have many shortcomings. The experimental samples are not extensive enough and the experimental data are not accurate enough. Therefore, in the next experiment, we should extensively absorb the corpus of the sample, perform repeated calculations, and use perfect The algorithm is analyzed to obtain a more realistic data analysis report, so that people can understand the power of the speech recognition technology under the neural network, and how many factors affect the English collection.

REFERENCES

Abdel-Basset, M., Mohamed, M., Zhou, Y., & Hezam, I. (2017). Multi-criteria group decision making based on neutrosophic analytic hierarchy process. *Journal of Intelligent & Fuzzy Systems*, *33*(6), 4055–4066. doi:10.3233/JIFS-17981

Biswas, P. S., Pramanik, S., & Giri, B. C. (2016). TOPSIS method for multi-attribute group decision-making under singlevalued neutrosophic environment. *Neural Computing & Applications*, *27*(3), 727–737. doi:10.1007/s00521-015-1891-2

Deli, I., & Şubaş, Y. (2017). A ranking method of single valued neutrosophic numbers and its application to multi-attribute decision making problems. *International Journal of Machine Learning and Cybernetics*, *8*(4), 1309–1322. doi:10.1007/s13042-016-0505-3

Dutta, A., Ashishkumar, G., & Rao, C. (2019). Designing of Gabor filters for spectro-temporal feature extraction to improve the performance of ASR system. *International Journal of Speech Technology*, *22*(4), 1085–1097. doi:10.1007/s10772-019-09650-5

Gábor, G., & Tamás, G. (2016). Domain Adaptation of Deep Neural Networks for Automatic Speech Recognition via Wireless Sensors. *Journal of Electrical Engineering*, *67*(2), 124–130. doi:10.1515/jee-2016-0017

Hamdan, S., & Shaout, A. (2016). Hybrid Arabic Speech Recognition System Using FFT, Fuzzy Logic and Neural Network. *International Journal of Information Technology and Computer Science*, *6*(4), 2249–9555.

Hernandez, A. A., & Fajardo, A. C. (2020). Convolutional Neural Network for Automatic Speech Recognition of Filipino Language. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(1.1), 34-40.

Hourri, S., & Kharroubi, J. (2020). A deep learning approach for speaker recognition. *International Journal of Speech Technology*, *23*(2), 123–131. doi:10.1007/s10772-019-09665-y

Huang, Y., Sheng, W., Jin, P., Nie, B., Qiu, M., & Xu, G. (2019). A Node-Oriented Discrete Event Scheduling Algorithm Based on Finite Resource Model. *Journal of Organizational and End User Computing*, *31*(3), 67–82. doi:10.4018/JOEUC.2019070104

Jiang, J., & Wang, H. H. (2021). Application intelligent search and recommendation system based on speech recognition technology. *International Journal of Speech Technology*, *24*(1), 23–30. doi:10.1007/s10772-020-09703-0

Kaur, G., Srivastava, M., & Kumar, A. (2018). Genetic Algorithm for Combined Speaker and Speech Recognition using Deep Neural Networks. *Journal of Telecommunications and Information Technology*, *2*(2), 23–31. doi:10.26636/jtit.2018.119617

Li, P., & Jiang, S. (2020). Analysis of the characteristics of English part of speech based on unsupervised machine learning and image recognition model. *Journal of Intelligent & Fuzzy Systems*, *39*(99), 1–11. doi:10.3233/JIFS-179960

Li, Y. (2021). Speech-assisted intelligent software architecture based on deep game neural network. *International Journal of Speech Technology*, *24*(1), 57–66. doi:10.1007/s10772-020-09722-x

Meng, F., Ji, Q., Zheng, H., Wang, H., & Chu, D. (2021). Modeling and Solution Algorithm for Optimization Integration of Express Terminal Nodes With a Joint Distribution Mode. *Journal of Organizational and End User Computing JOEUC*, *33*(4), 142–166. doi:10.4018/JOEUC.20210701.oa7

Nagajyothi, D., & Siddaiah, P. (2018). Speech Recognition Using Convolutional Neural Networks. *IACSIT International Journal of Engineering and Technology*, *7*(4), 133–137. doi:10.14419/ijet.v7i4.6.20449

Nagyeong. (2017). A Study on English Translation of the Muk'am Collection - Focusing on the elements of literary style. *The Journal of Transatlantic Studies*, *18*(1), 65–93.

Ogawa, A., Hori, T., & Nakamura, A. (2016). Estimating Speech Recognition Accuracy Based on Error Type Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(12), 1–1. doi:10.1109/TASLP.2016.2603599

Panda, S. P., & Nayak, A. K. (2016). Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technology*, *19*(1), 9–18. doi:10.1007/s10772-015-9320-6

Pawar, M. D., & Kokate, R. D. (2021). Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. *Multimedia Tools and Applications*, *80*(10), 15563–15587. doi:10.1007/s11042-020-10329-2

Shukla, S., Jain, M., & Dubey, R. K. (2018). Increasing the performance of speech recognition system by using different optimization techniques to redesign artificial neural network. *Journal of Theoretical and Applied Information Technology*, *97*(8), 2404–2415.

Swietojanski, P., & Renals, S. (2016). Differentiable Pooling for Unsupervised Acoustic Model Adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(10), 1773–1784. doi:10.1109/TASLP.2016.2584700

Tachioka, Y., & Ishii, J. (2016). Long short-term memory recurrent-neural-network-based bandwidth extension for automatic speech recognition. *Acoustical Science and Technology*, *37*(6), 319–321. doi:10.1250/ast.37.319

Tu, Y. H., Du, J., & Lee, C. H. (2018). A Speaker-Dependent Approach to Single-Channel Joint Speech Separation and Acoustic Modeling Based on Deep Neural Networks for Robust Recognition of Multi-Talker Speech. *Journal of Signal Processing Systems for Signal, Image, and Video Technology*, *90*(7), 963–973. doi:10.1007/s11265-017-1295-x

Wangi, N., & Madekhan, M. (2019). Mobile Application for Training of Foreign Students with Gamification Techniques and Speech Recognition Technology. *EDUTEC Journal of Education And Technology*, *2*(2), 38–46. doi:10.29062/edu.v2i2.32

Xu, C., Xie, L., & Xiao, X. (2018). A Bidirectional LSTM Approach with Word Embeddings for Sentence Boundary Detection. *Journal of Signal Processing Systems for Signal, Image, and Video Technology*, *90*(7), 1063–1075. doi:10.1007/s11265-017-1289-8

Ye, J. (2018). Neutrosophic number non-linear programming problems and their general solution methods under neutrosophic number environment. *Axioms*, *7*(13), 1–9.

Zhihao, J. (2021). Simulation of ocean surface temperature based on audio signal collection and accuracy of trade English translation. *Arabian Journal of Geosciences*, *14*(16), 1–15. doi:10.1007/s12517-021-07859-w

Zia, T., & Zahid, U. (2019). Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. *International Journal of Speech Technology*, *22*(1), 21–30. doi:10.1007/s10772-018-09573-7

Zoughi, T., & Homayounpour, M. M. (2019). A Gender-Aware Deep Neural Network Structure for Speech Recognition. *Iranian Journal of Science and Technology - Transactions of Electrical Engineering, 43*(3), 635-644.

Xizhi Chu was born in Xi'an, Shaanxi, P.R. China, in 1985. She received the Master's degree from Xi'an Polytechnic University, P.R. China. Now, she works in Xi'an Aeronautical University. Her research interests include English teaching, social linguistics, and data analysis. Corresponding Author E-mail: 20150413@stu.nun.edu.cn.

Yuchen Liu was born in Xi'an, Shaanxi, P.R. China, in 1988. He received the Master's degree from Hunan University, P.R. China. Now, he works in Xi'an Aeronautical University.