Design and Application of Neural Network-Based bp Algorithm in Speech Translation Robot

Yuhan Jie*

Department of Intelligent Sciences, Artificial intelligence academy, Nanchang Jiaotong Institute, Nanchang, China
E-mail: 15180180173@163.com, YuhanJie202305@hotmail.com, *lsl851219@163.com
*Corresponding author

Keywords: speech translation, robots, technology, science, and medical care

Abstract

The process of turning spoken words from one language into another's spoken words is called speech translation. It entails analyzing the spoken words and producing an accurate translation in real-time utilizing cutting-edge algorithms and machine learning approaches. The usage of speech translation technology is widespread across several sectors, including travel, business, and healthcare. For instance, a doctor who speaks English may utilize a voice translation system to converse with a patient who speaks other languages, and a corporate executive would do the same while speaking with associates or customers abroad. Nowadays, speech translations are used in robots that are designed to translate speech from one language to another in real-time. in this paper, we implemented speech translation in the domain of intelligent science of medical care and technology. To assist English-speaking individuals in describing their symptoms to other language physicians or nurses, we suggested a neural network-based back propagation technique. Unlike laptops or tablets, a humanoid robot may be extended to reach out to individuals in need first and may eventually replace human labor. Finally, the neural-based control technique that was developed proved to be an efficient system for controlling human-robot voice translation, as judged both quantitatively and qualitatively. Results from a controlled trial demonstrating the translation's accuracy and success rate.

Povzetek: Opisana je nova metoda govornega prevajanja, tj. v realnem času iz govora v enem jeziku v drugega, za potrebe zdravstva s pomočjo globokih nevronskih mrež.

1 Introduction

The most popular techniques for voice translation systems are still to handle speech recognition and machine translation as independent modules. Although this method might cause error propagation and a drop in total efficiency, it is nonetheless popular since it enables the individual tuning of each component, improving performance as a whole. As neural models are better able to grasp the intricate structure of language and provide more natural-sounding translations, recent developments in neural machine translation have resulted in considerable advances in the quality of voice translation [1]. Techniques like regularizing, breaking sentences, smoothing, and anticipating punctuation may be included in the post-processing module. To lessen the influence of recognition mistakes on the translation, it could additionally incorporate error correction. By allowing real-time processing and more effective adaptation to various accents and speaking styles, artificial intelligence technology has the potential to substantially enhance voice translation systems [2]. Speech translation robots can be a useful tool for facilitating communication across language barriers. However, the effectiveness of a particular speech translation robot will depend on its performance across these different criteria, as well as the specific needs and context of its intended users. Illustrates the structure of speech translation, which is shown in Figure 1

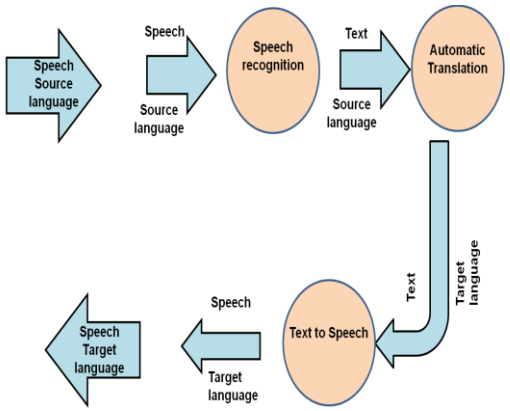

Figure 1: Implementation of speech translation

The user experience might be greatly enhanced by a machine synchronous translation system that integrates voice recognition, machine translation, and cutting-edge AI methods, which would also help the online translation business grow [3]. Due to their physical limitations, the persons deal with a variety of issues in their day-to-day lives, such as difficulty carrying out routine duties that physically fit people take for granted. Robotics and other assistive technology may be very useful in assisting people with disabilities to enjoy more independent and happy lives. The creation of prosthetic arms and hands that can be operated by voice recognition or other non-intrusive techniques is one potential use of robotics in this context [4]. The equipment enables people with limb loss to carry out a variety of activities that would be challenging or impossible without it, such as handling things, using a keyboard, or even playing musical instruments. Overall, robots and speech processing have immense potential to enhance the independence, mobility, and capacity for social interaction of people with disabilities, thereby improving their quality of life. We can anticipate seeing more creative applications that assist disabled individuals overcome their obstacles as these technologies continue to advance and become more accessible [5]. It's a difficult challenge to create voice communication capabilities for robots that seem natural and nice, and it calls for complex verbal and non-verbal interaction skills. To accomplish successful human-robot interaction, speech recognition/synthesis, conversation management, and motion recognition/generation are all essential elements that must be combined and enhanced [6]. The development of high-quality synthetic speech that sounds natural and intuitive is of great relevance in the context of service robots that are intended for voice communication. However, current Text-to-Speech (TTS) systems are often designed more for text reading than for conversation, which may result in synthetic speech that sounds repetitive, artificial, or unwelcoming [7]. Additionally, since the majority of TTS systems are built on corpora that are mostly composed of monologues, synthetic speech may lack the liveliness of actual conversations. Researchers are investigating novel approaches for enhancing TTS systems for communication, such as gathering conversational corpora and creating more realistic intonation patterns, to solve these difficulties. They anticipate seeing increasingly sophisticated and successful voice communication systems in service robots and other interactive gadgets as these technologies continue to improve [8].

As we proposed, the BP approach based on neural networks might be utilized to create a speech translation robot. The applications offer a platform that may aid in the development of spoken language, making speech translation simpler and more interactive.

The remaining portion of the manuscript is organized as follows: part 2 explains the prior study about the goals or objectives of the research and highlights any shortcomings or differences from it. Part 3 gives suggestions for further study based on the results and outlines the research technique and methods utilized to gather and assess data. We go through the Discussion and Findings in Part 4 before presenting the research findings succinctly and methodically, evaluating and describing them in light of the study goals or objectives. The Study's primary components are summarized in part 5, along with its relevance and contributions, possible implications for practice or policy, and prospective future research fields.

2 Application of speech translation

The Research preferred applications based on the use of voice and visual technologies that are quickly evolving. Effective examination, swift and efficient system transfer, and appropriate medication prescriptions are a few examples. The article committed to providing a thorough analysis focusing on the requirements for the standpoint addressed, which quickly became available [9]. The case study showed Deep learning and intelligence are combined in the areas of gesture detection, voice translation, emotion identification, and intelligent robot navigation. A wide variety of recognition strategies have been proposed and put to the test in trials in related subject topics [10]. The researchers investigated the potential for developing the method, which was put into action and evaluated by the humanoid robot Pepper. A statistical study of the ratings offered by human volunteers who identified as belonging to diverse cultures is addressed, along with the preliminary results [11]. The Research evaluated that Text-to-Speech systems are designed for text reading rather than conversation; the robot utterances often sound repetitive, artificial, and unwelcoming. Here, we provide a robot voice synthesis system that is non-monologue [12]. The article conducted 24 people who participated in studies using the proposed methodology, which was tested in three situations: text reading, conversation, and domestic service robot (DSR) scenarios. The experimental findings demonstrated that, in the text-reading situation, the performance of our suggested technique was on par with the baseline method and surpassed it in the DSR scenario. The proposed system is a cloud-based voice synthesis service, making it free to use [13]. The Research determined electronics for the human-robot interfaces (HRI), which is necessary to develop thermal acoustic sound emission technology and triboelectric acoustic sensor system which may be employed as electrodes, thermal acoustic sources, or triboelectric materials [14]. The overview objective of the article is evaluated. By meticulously adjusting the structural parameters, the GHRI, which can recognize achieves significant sensitivity and operational durability [15].The research found that the artificial intelligence communication model which is also based on identified speech elements. The outcome demonstrated promising

futures for the advancement of robotic intelligence [16]. The case study demonstrated Speech is a natural approach to communicating with social robots: Speak-language conversations may aid users in naturally and adaptably expressing their intentions. Speech recognition and natural language processing have made significant strides in artificial intelligence-related spoken conversation technologies over the last several years [17]. The Research presented the incorporation of open-source conversation management, voice recognition, and natural language processing components into a robot software platform, as well as a report on a pilot test of the combined system with actual users [18]. The case study extracted the dialogue management component of the robot utterance and combined the spoken content with the robot's gestures, which are also crucial in human-robot interaction. We picked mealtime conversations about food and recipes as the discourse domain since speaking with a companion robot in these situations is seen as natural and helpful [19]. The research extracted from the second edition of Robo-Identity will provide the chance to broaden the conversation around synthetic identity after the previous event's success. We have recently been concentrating on how words and voice may be used to portray emotions. Robotic sounds that mimic expressive human voices are getting more difficult to tell apart [20]. The case study established the possibility and limitation of using emotional speech to communicate a human-like identity that can deceive others and raise ethical concerns. Should robots keep a machine-like posture, such as via robotic speech, and should more human-like emotional displays be considered as design possibilities [21]? The research invited viewpoints on difficulties and possibilities from a range of sectors since they address mutually exclusive issues and the need for interdisciplinary research. Speech, emotion, and manufactured identity will be the particular focus of this year's event [22].The research may be gathered to deploy the companion robot after the input text is transformed directly via the deep neural network (DNN). Additionally, compared to other commercial humanoid companion robots, the development of voice translation robotic applications is discussed in the article [23]. The article showed that the two-layer fuzzy multiple random forests (RF) are capable of reliably identifying emotions which improved the efficiency of speech translation [24]. The case study was calculated by combining logistic regression (LR)-optimized weightings with speech, object, and motion confidence. After that, the paper combined the measurement with gaze tracking and conducted studies with real-world human-robot interaction. According to experimental findings, the suggested technique for robot-directed speech recognition performs well, with average recall and accuracy rates of 94% and 96%, respectively [25].

The study may have explored and highlighted various aspects of speech translation robotics that can be improved through the use of an NN-BP approach. Some specific passages in the study that could support the idea of improving the deployment of speech translation robotics could include:

1. This study aims to enhance the effectiveness of speech translation by utilizing the NN-BP approach and assessing its impact.

2. To evaluate the effectiveness of an approach, conducting experiments and comparing their performance and satisfaction is a common method used in many fields, such as machine learning, user experience, and product design. The related works summary is presented in Table 1.

3 Proposed methodology

Speech translation robots utilize speech recognition technology to transform spoken words into text. They then apply natural language processing algorithms to examine the text and produce a translation in the desired language. Certain speech translation robots utilize speech synthesis technology to produce an audio rendition of the translated speech. In this section, we described speech translation of robot techniques in medical care, and our suggested method of neural networks based on a backpropagation algorithm is evaluated with some performance measures.

Table 1: Summary of related works

Reference	Description	Limitation
[10]	The HCI method has developed from traditional print media to intelligent media. A never-ending stream of VR, AR, and AI interactive devices have appeared, bringing seismic changes to people's lives through gesture control, voice control, dialogue robots, and other means.	HCI is just the beginning; More data from various sensors will be combined in the future.
[11]	Researchers examined the possibility of developing the procedure that Pepper, a humanoid robot, tested and evaluated.	It's very hard to predict
[13]	They proposed a cloud-based voice synthesis service, which was tested in three situations: text reading, conversation, and domestic service robot (DSR) scenarios	They focus on speech-to-sign translation without offering a vice-versa option
[19]	They extracted the robot's utterance's dialogue management component and combined it with the robot's motions.	It is only based on sounds, which means that the meanings of words may lose.
[22]	They discussed the possible advantages of including a socially supportive robot in speech therapy interventions.	A few children's temporary dissatisfaction in a small number of training sessions
[23]	Emotion analysis uses the convolutional neural network to develop the robotic voice applications for translation	facial recognition technology not used
[24]	The two-layer fuzzy multiple random forests (RF) have increased the effectiveness of voice translation by being able to accurately recognize emotions.	The TLFMRF can use an intelligent optimization algorithm like the Genetic Algorithm (GA) in the future.
[25]	The study was derived by integrating logistic regression (LR)-optimized weightings with confidence in voice, object, and motion.	It should be noted that the method's core concept

3.1 Design of the system

The software used CMU Sphinx-4, a flexible and manageable open-source Java recognition of voices toolkit, to implement speech recognition. Based on the Java Speech Grammar Format, we developed a grammar-type speech framework for the voice recognizer to analyze. After accurate identification, the text is sent to a translator for processing. A sentence is then broken down into its parts by the translation algorithm. These parts include the subject at hand, verbs, things, and accommodate phrases. It then applies Chinese syntactic rules to the parsed elements and rearranges them. Finally, the DARwIn-OP will say the Korean version of the translated statement. Pre-recorded MP3 files were utilized since there is currently no suitable Korean TTS software that can be used for this application. The hash table is used to find the corresponding Chinese voice recording for each word. Due to sensor performance limitations in the DARwIn-OP and the need for system stability while simultaneously running translating and voice recognition programs, this concept was executed on a laptop computer. So, the humanoid machine and the computing device can talk to one another thanks to TCP/IP sockets connections.

3.2 CMU organize-4 speech recognizer

CMU organize-4 is capable of voice recognition since a language model was created to identify proper phrases. Languages and statistical language models are both available via the CMU Sphinx-4 API for describing language. A statistical language model, the second kind, works well when there is a lot of data to work with. For now, however, we've decided to use the JSGF grammar approach as testing with a small number of example English sentences is sufficient. Sun Technologies' JSGF is a written representation of grammar that may be used to construct grammar for either complex phrases or basic instructions.

The JSGF grammar files specify the grammar of legitimate speech as follows:

Subject + predicate + object

The voice translation algorithm fails if the data entered does not conform to the specified syntax. As a consequence, the voice recognition component selects only entire phrases for translation.
We took for granted the typical scenario of a patient seeing a doctor for routine care. The recognizer's vocabulary of target words, phrases, and sentences was developed with this use case in mind. We spoke with the Dongguk University student healthcare facility to gain an overview of the most often reported ailments and the information that physicians need to prescribe effectively. Language files for stomachache, headaches, and the flu were defined based on this interview. Words and phrases from these dictionaries may be used to describe the individual's medical symptoms, such as the duration and severity of the individual's discomfort.

3.3 Algorithm of translation

Studying three of the 7 patterns of English sentence building (Subject + Predicate + Complement + Object) was the focus of this investigation. S+P comes before C or O in English. In contrast, a standard Chinese phrase consists of S + [C | O] + P, with P placed at the final. This variation in word form was addressed throughout the alteration process.

This system's English-Chinese software for translation is made up of 5 major modules: tokenization, part-of-speech tags, word grouping, Chinese language usage, and sentence-by-sentence interpretation.

A.Tokenization:

Spacing and articles ("a," "an," and "the"), which have no significance in Chinese, are used by the tokenization when possible module to divide the string into tokens once the words recognizer has finished its transcription. This transforms every token into a real word.

B. Part of speech tagging:

Parsing the phrases into their parts (subject, predicate, and object) requires first looking at the part of speech of each word. Each word is assigned a part of speech by looking it up in a database that categorizes words according to their meaning. Given the small sample size of input phrases, we can only focus on one sense of the term at a time. Words like "feel" are consistently identified as verbs in this system. Each word is now labeled as a noun, verb, or adjective based on this process.

C. Structure sentence grouping:

Each token is then assigned a part of the speech label before being sorted into sentences. Each clump of nouns and adverbs, for instance, produces a noun phrase. Time adverb phrases are categorized in the dictionary by using prepositions like "since," "for," and "from." It is conventional to treat the first noun word in a line as the participant and the second as a complement or object. An entirely verbal sentence functions as a predicate.

D. Chinese grammar application:

This stage entails rearranging parts in accordance with syntactic norms used in Korean. Prepositional phrases undergo a morphological and syntactic shift toward the Chinese norm. For example, prepositions like "for" and "since" are placed at the conclusion of a sentence in Chinese. A single Chinese character may represent many English phrases; thus, a manifestation same "runny nose" or "sore throat" becomes one token.

E. Word-by-word translation:

Each English sentence is then matched to a keyword in Chinese after the tokens have been organized in a Chinese sentence sequence. Since we do not use a TTS application and instead use matching pre-recorded audio files for each word, this process was streamlined by merging it with the following stage of Korean speech creation.

3.4 Generation of Chinese speech

At this early point, a TTS system was not used in the creation of Korean speech. DARwIn-OP searches a hash table for each token after receiving them in their new order through TCP/IP from the client of the translation application. The DARwIn-OP uses a hash table where the "keys" are English words and the "values" are system links to Chinese voice recording files. Once the search is complete, an in-built DARwIn-OP library method is called to play a folder's worth of MP3s in order. Therefore, DARwIn-OP plays an MP3 file for each Chinese word, resulting in a whole Chinese phrase.

3.5 Design of neural network

This study uses a 3-layer neural network, with an input layer, a hidden layer, and an output layer, as its basis in the neural network framework. The average grain size Z output (MD) is calculated from the input value W input"(GR, DEN, CNL, AC)," which is the responsiveness measure of the average rock grain size. The following is the format of its data:

$$W_{input} = \{w_{11}, w_{12}, \dots, w_{kn} \ \vdots \ \ddots \ \vdots \\ w_{m1}, w_{m2,} \dots, w_{mn} \ \}^S = \{HQ_1, CFM_1, DMK_1, BD_1, \vdots \vdots \vdots \vdots \\ HQ_m, CFM_m, DMK_m, BD_m, \}^S \quad (1)$$

$$Z_{output} = (z_1, z_2, \dots z_m)^{s=} \ (NC_1, NC_2, \dots NC_m)^S \ (2)$$

The neural network has four inputs and one output, therefore, m is the number of training observations, and n is the input sensitivity variable for each sample. The network's nonlinear mapping may be written as $Q^m \to Q^1$. Each concealed layer node's input is if the hidden layer is o, and else it is,

$$T_i = \sum_{j=1}^{n} \quad x_{ji} w_j - \theta_i (i = 1,2, \dots, o), \ (3)$$

Where x_{ji} is the weight of the link between the input and hidden layers. Node θ_i in the hidden layer has a threshold value of θ_i.

For the chosen activation processes, the corresponding ReLU and tanh calculation formulae and derivative forms are as follows:

$$Relu(w) = \{w, (w > 0), 0, (w \leq 0), \ (4)$$

$$Relu(w) = \{w, (w > 0), 0, (w \leq 0), \ (5)$$

$$Tanh(w) = \frac{f^w - f^{-w}}{f^w + f^{-w}}, \quad (6)$$

$$Tanh(w) = 1 - \frac{(f^w - f^{-w})^2}{(f^w + f^{-w})^2} \quad (7)$$

ReLUis used as the usable mechanism in input and hidden layer networks because of the features of its data selection, which eliminate the issue of slow progressive descent in the learning procedure of neural networks and cause gradient descent to converge much quicker than other traditional activation functions. However, because of its unilateral inhibition, a ReLU neuron will no longer be activated on any data when a very big gradient passes through it and the parameters have been updated. As a result, the gradient of the neuron will then always be 0. The majority of the network's neurons will probably be inactive if the learning rate is too high, and the tanh activation function will only operate when the features are noticeably changed.

The data output in the final output layer uses the tanh activation function since the feature effect will be constantly extended throughout the cycle.

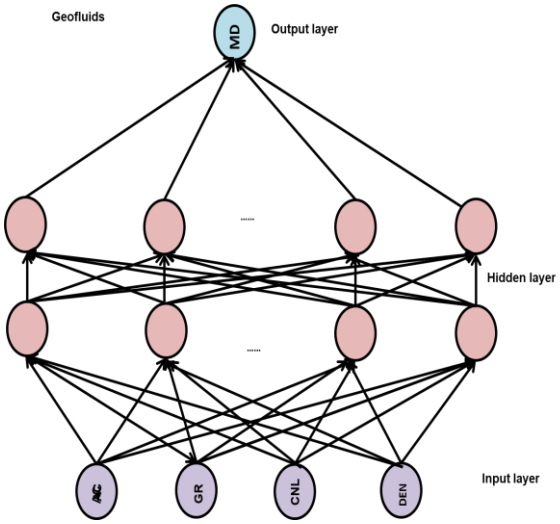

Figure 2: Structure of neural networks

According to the average size of particles and sensitivity variables of low-permeability sandstone, both superficial and DNN were constructed. Deep neural networks have two additional concealed layers compared to superficial neural networks. The mechanisms for activation of the hidden and input layers of the neural networks are ReLU; the input is the sensitive performance of the average size, and the output is the median grain size. The loss function that is partly used to solve the regression issues is the MSE loss function. As a result, two neural networks were constructed and then developed by the data input for the organize-permeability sandstone medium-size forecast model. Figure 2 depicts the general framework of the neural network.

4 Results and discussion

The experimental outcomes are determined by two metrics: the proportion of correct translations at both the initial and conclusion stages of the process.

The success levels for the three primary processes: are English speech-to-text, word reorganization according to Chinese sentence structure, and English text-to-Chinese speech displayed in Figure 3 and the values are termed in Table 2.

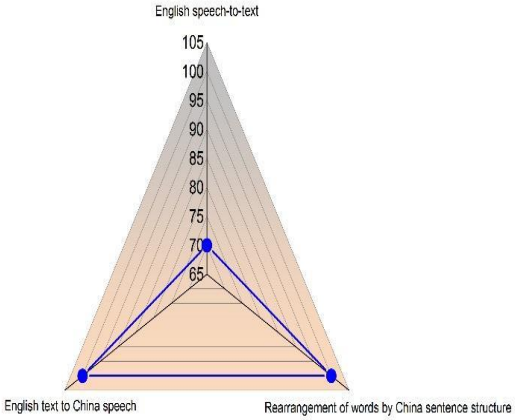

Figure 3: Success rate in breakpoint

Table 2: Success rate in each breakpoint comparison

The success rate in breakpoint	
English STT	70%
China's rearranged sentence structure uses different terms	100%
Chinese voice rendered in English text	100%

4.1 English language-to-text

The efficiency of CMU Sphinx-4 and the translation model determine the outcome of this rate. The success rate for English speakers, both native and non-native, is 70% on average. The distinction between "have" and "had" in terms of phonemes is the fundamental reason why the rate decreased. Other phrases were virtually precisely dictated.

4.2 Chinese sentence structure-based word rearrangement

If CMU Sphinx-4 is able to identify a spoken sentence as described in grammar files, it will be able to transform all of the example sentences provided in our grammatical files. We only allowed Subject + Predicate + [Complement | Object] phrases to be accepted as legitimate input, which made it possible.

4.3 English words to the Chinese language

Assuming that the word rearranging was correct, the outcome of creating English words to Chinese voice worked flawlessly, as it matched the pre-saved Chinese word files to each restructured word in English.

The percentage of accurate S2S translations is displayed in Figure 4 and Table 3. As can be seen above, this rate is affected by how well speech is recognized. CMU Sphinx4's acoustic model seems to be more accommodating to the native language of English since the success rate was 20 percent higher with the native language of English than the non-native language of English. Chinese translations that are accurate in terms of grammar and meaning can have awkward expressions. This restriction arose because the English elements were rearranged into Chinese word order and then translated word by word.

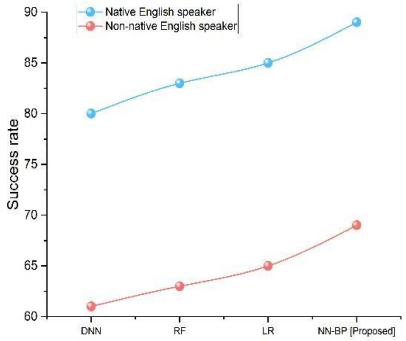

Figure 4: The success level of S2S translation

Table 3: The success level of S2S translation comparison

Methods	The success level of S2S translation	
	The native language of English	Non-native language of English
DNN	80	61
RF	83	63
LR	85	65
NN-BP [Proposed]	89	69

4.4 Accuracy

A performance indicator called accuracy assesses how accurate the system's predictions were overall. Accuracy in our BP algorithm-based platform refers to the system's capacity to accurately recognize learners' requirements and preferences and to provide them with the right kind of feedback and direction. High accuracy values obtained by our technique show that it is capable of providing accurate and useful forecasts. Our research indicates that our BP algorithm platform's accuracy is good enough to allow real-world applications in voice translation robots and that it may be further enhanced by including more representative and varied data, as well as by enhancing the rule set and feature selection.

$$Accuracy = (Truepositives + TrueNegatives)/(Truepositives + Truenegatives + Falsepositives + Falsenegatives) = (TP + TN)/(TP + TN + FP + FN) \quad (8)$$

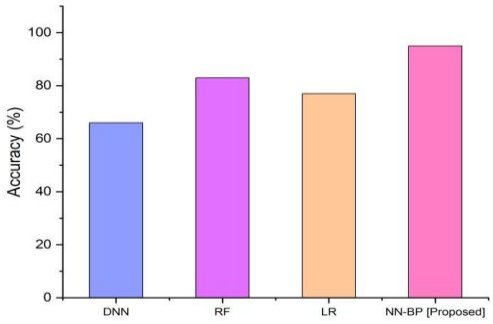

Figure 5: Accuracy of proposed and existing method.

Figure 5 displays the accuracy of the proposed and current methods. A percentage of the total is often used to represent the accuracy level. There are indicators of the possibility of erroneous forecasts in both the current methodology and the one that is being suggested. This threat is recognized by both systems. In comparison to DNN's accuracy of 66%, RF's accuracy of 83%, and LR's accuracy of 77%, the recommended technique, NN-BP, achieves an accuracy of 95%. Therefore, the strategy that is suggested has the best accuracy rate. The proposed approach accuracy is shown in Table 4

Table 4: Comparison of accuracy

Methods	Accuracy (%)
DNN	66
RF	83
LR	77
NN-BP [Proposed]	95

4.5 Precision

The algorithm's processing speed, resource efficiency, and capacity to deal with various accents, dialects, and languages. We would normally utilize metrics to assess the correctness of the system's transcription of spoken language when evaluating the precision of a neural network-based back propagation (BP) algorithm in a voice translation robot. Overall, a neural network-based BP algorithm's precision in a voice translation robot will be influenced by several variables, such as the amount and quality of training data, the model's architecture's complexity, and the efficiency of the learning and optimization methods used.

$$Precision = Truepositives/(Truepositives + Falsepositives) = TP/(TP + FP) \quad (9)$$

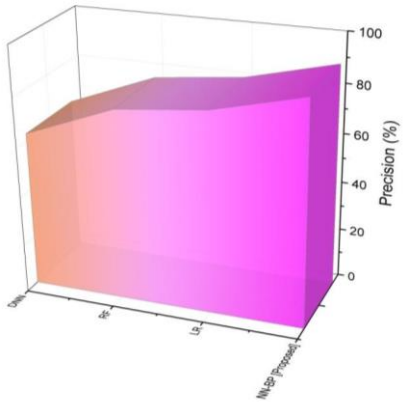

Figure 6: Accuracy of proposed and existing method.

The precision of the proposed and existing approaches is shown in Figure 6. A percentage of the total is often used to represent precision levels. Both the existing approach and the one that is being discussed show signs of the potential for inaccurate estimates. This hazard is recognized by both systems. The suggested technique, however, has a 90% precision rate, compared to 63% for DNN, 77% for RF, and 81% for LR. Consequently, the proposed technique has the greatest Precision rate. Table 5 displays the proposed procedure's precision.

Table 5: Comparison of precision

Methods	Precision (%)
DNN	63
RF	77
LR	81
NN-BP [Proposed]	90

4.6 Recall

The recall is a metric used in machine learning and information retrieval to assess how thorough or full the findings are. It calculates the proportion of relevant things that were found among a bigger collection of objects. How many of the right words or characters are properly identified and included in the output transcription is often the metric used to assess the recall of a neural network-based back propagation (BP) algorithm in a voice translation robot.

$$Recall = \frac{FN}{FN+TP} \quad (10)$$

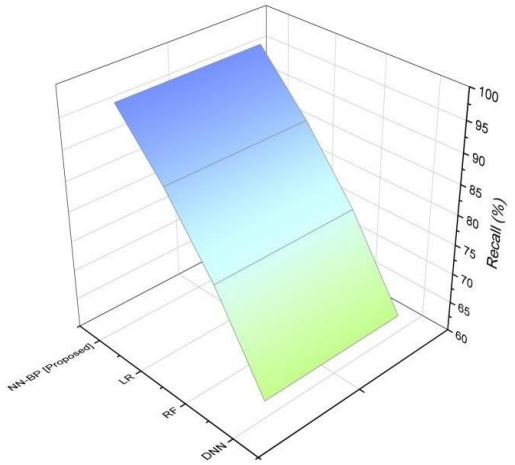

Figure 7: Recall of proposed and existing method.

Figure 7 depicts the recall of the proposed and current methods. A percentage of the total is often used to represent the recall level. There are indicators of the possibility of erroneous forecasts in both the current methodology and the one that is being suggested. This threat is recognized by both systems. A recall rate of 97% is produced by the recommended method, NN-BP, as opposed to recall rates of 63% for DNN, 77% for RF, and merely 88% for LR. Therefore, the strategy that is suggested has the best accuracy rate. The proposed procedure recall is shown in Table 6.

Table 6: Comparison of Recall

Methods	Recall (%)
DNN	63
RF	77
LR	88
NN-BP [Proposed]	97

4.7 Discussion

The research was derived by integrating logistic regression (LR)-optimized weightings with confidence in voice, object, and motion [10]. Artificial neural networks or regression algorithms can be used to estimate linear predictor coefficients as well as other critical parameters for quantized construction [11]. Based on GANs trained with converter audio and normalized key points retrieved from a chosen dataset, the system has been tested with people who identified with different cultures, and they provided favorable comments on its capacity to embed cultural features [19]. The power analysis demonstrates that the inquiry is limited by the preliminary exploratory results because of the small number of subjects [22]. The average emotion recognition accuracy is 77.82%. Additionally, the accuracy effectively increases and can reach up to 79.81% on average if the speech data is further analyzed by voice data augmentation to increase the total number of data [23]. The computation times for the TLFMRF, RF, and BPNN are 0.0579s, 0.0128s, and 0.0013s, respectively. Although the suggested technique has the longest computing time, it is still under a second, which can still ensure that real-time tracking accuracy is within acceptable bounds [24]. The approach is exceedingly efficient and fulfills a crucial need for natural and secure human-robot interaction [25]. The success rate of sentence-to-sentence translation is 89% when done by a native speaker of English but drops to 69% when done by someone whose first language is not English.

5 Conclusion

Finally, patients who speak languages other than English may benefit greatly from the use of voice translation robots in hospitals. By offering precise and immediate interpretations of spoken language, these robots may aid healthcare practitioners in communicating more effectively with their patients.In medical environments where communication is essential, such as emergency rooms, intensive care units, and clinics, speech translation robots may be very helpful. These robots may assist in enhancing the standard of care and guaranteeing that patients get the right therapy by accurately translating symptoms among patients and medical histories for healthcare professionals.We have created the first humanistic application model that can be utilized for English-Chinese S2S translation in a medical setting to meet the needs of the ever-increasing number of patients from outside of China who want services for translation.DARwIn-OP, a humanoid robot, serves as a base because of the increased mobility and versatility that comes with this kind of robot in the workplace. Furthermore, it can communicate with anybody, including people who are unfamiliar with technological devices. As a result, it is anticipated that the system suggested in this work would be able to significantly increase its capability.Speech translation software and a rule-based machine translation approach were brought together to create this system. This software first employs a narrower model of CMU Sphinx-4 for voice recognition and then uses rule-based translation techniques to convert the detected English text into Chinese. S2S translation has an 89% success rate when using a native language of English but only a 69% success rate when using a non-native language of English. Currently, the translation functions flawlessly with a fixed voice recognition domain and predefined rules in the translation techniques. Therefore, the outcome relies on speech translation identification.

References

[1] Tan, Y., (2022). Design of Intelligent Speech Translation System Based on Deep Learning. *Mobile Information Systems.*

[2] *M*athur, M., Samiulla, S., Bhat, V. and Jenitta, J., (2020), October. Design and Development of Writing Robot Using Speech Processing. *In 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC)* (pp. 1-4). IEEE.

[3] Shahin, I., Hindawi, N., Nassif, A.B., Alhudhaif, A. and Polat, K., (2022). Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Systems with Applications*, 188, p.116080.

[4] Wang, Y., (2022). The Performance of Artificial Intelligence Translation App in Japanese Language Education Guided by Deep Learning. *Computational Intelligence and Neuroscience*.

[5] Yuvaraj, S., Badholia, A., William, P., Vengatesan, K. and Bibave, R., (2022), May. Speech Recognition Based Robotic Arm Writing. In Proceedings of International Conference on Communication and Artificial Intelligence: ICCAI 2021 (pp. 23-33). *Singapore: Springer Nature Singapore.*

[6] Shah, H.D., Sundas, A. and Sharma, S., (2021), September. Controlling Email System Using Audio with Speech Recognition and Text to Speech. *In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-7). IEEE.

[7] Kargathara, A., Vaidya, K. and Kumbharana, C.K., (2021). Analyzing desktop and mobile applications for text-to-speech conversation. *In Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS* 2020 (pp. 331-337). Springer Singapore.

[8] Walker, N.T., Ultes, S. and Lison, P., (2022). GraphWOZ: Dialogue Management with Conversational Knowledge Graphs. *arXiv preprint* arXiv:2211.12852.

[9] Yadav, S.P., Zaidi, S., Mishra, A. and Yadav, V., (2022). Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). *Archives of Computational Methods in Engineering*, 29(3), pp.1753-1770.

[10] Lv, Z., Poiesi, F., Dong, Q., Lloret, J. and Song, H., (2022). Deep Learning for Intelligent Human–Computer *Interaction. Applied Sciences*, 12(22), p.11457.

[11] Delić, V., Perić, Z., Sečujski, M., Jakovljević, N., Nikolić, J., Mišković, D., Simić, N., Suzić, S. and Delić, T., (2019). Speech technology progress based on new machine learning paradigm. *Computational intelligence and neuroscience*.

[12] Raheem, A.K.A. and Zuhair, M., (2023), March. Real-time speech recognition of Arabic language. *In AIP Conference Proceedings* (Vol. 2591, No. 1, p. 030018). *AIP Publishing LLC.*

[13] Homburg, D., Thieme, M.S., Völker, J. and Stock, R., (2019). Robotalk-prototyping a humanoid robot as a speech-to-sign language translator.

[14] Sun, H., Gao, X., Guo, L.Y., Tao, L.Q., Guo, Z.H., Shao, Y., Cui, T., Yang, Y., Pu, X. and Ren, T.L., (2023). Graphene-based dual-function acoustic transducers for machine learning-assisted human–robot interfaces. *InfoMat*, 5(2), p.e12385.

[15] Gkeka, E., Agorastou, E. and Drigas, A., (2019). Artificial Techniques for Language Disorders. Int. J. Recent Contributions Eng. Sci. IT, 7(4), pp.68-76.

[16] Chen, X., (2021). Simulation of English speech emotion recognition based on transfer learning and CNN neural network. *Journal of Intelligent & Fuzzy Systems*, *40*(2), pp.2349-2360.

[17] Fujii, A. and Kristiina, J., (2022), March. Open source system integration towards natural interaction with robots. In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 768-772). IEEE.Minato, T., Higashinaka, R., Sakai, K., Funayama, T., Nishizaki, H. and Nagai, T., (2022). Overview of dialogue robot competition 2022. arXiv preprint arXiv:2210.12863.

[18] Gjaci, A., Recchiuto, C.T. and Sgorbissa, A., (2022). Towards Culture-Aware Co-Speech Gestures for Social Robots. *International Journal of Social Robotics*, 14(6), pp.1493-1506.

[19] Laban, G., Le Maguer, S., Lee, M., Kontogiorgos, D., Reig, S., Torre, I., Tejwani, R., Dennis, M.J. and Pereira, A., (2022), March. Robo-identity: Exploring artificial identity and emotion via speech interactions. In 2022 17th ACM/IEEE International *Conference on Human-Robot Interaction* (HRI) (pp. 1265-1268). IEEE.

[20] Guljajeva, V. and Canet Sola, M., (2022), October. Dream Painter: An Interactive Art Installation Bridging Audience Interaction, Robotics, and Creative AI. *In Proceedings of the 30th ACM International Conference on Multimedia* (pp. 7235-7236).

[21] Esfandbod, A., Rokhi, Z., Meghdari, A.F., Taheri, A., Alemi, M. and Karimi, M., (2023). Utilizing Lee, M.C., Chiang, S.Y., Yeh, S.C. and Wen, T.F., (2020). Study on emotion recognition and companion Chatbot using deep neural network. *Multimedia Tools and Applications*, *79*, pp.19629-19657.

[22] Chen, L., Su, W., Feng, Y., Wu, M., She, J. and Hirota, K., (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot *interaction. Information Sciences*, 509, pp.150-163.

[23] Zuo, X., Iwahashi, N., Taguchi, R., Funakoshi, K., Nakano, M., Matsuda, S., Sugiura, K. and Oka, N., (2010), September. Detecting robot-directed speech by situated understanding in object manipulation tasks. In 19th International *Symposium in Robot and Human Interactive Communication* (pp. 608-613). IEEE.