Hindawi
Security and Communication Networks
Volume 2022, Article ID 9853158, 1 page
https://doi.org/10.1155/2022/9853158

Retraction

Retracted: Hybrid Algorithm for English Translation Speech Recognition Based on Deep Learning Model and Clustering

Security and Communication Networks

Received 10 November 2022; Accepted 10 November 2022; Published 17 November 2022

Security and Communication Networks has retracted the article titled "Hybrid Algorithm for English Translation Speech Recognition Based on Deep Learning Model and Clustering" [1] due to concerns that the peer review process has been compromised.

Following an investigation conducted by the Hindawi Research Integrity team [2], significant concerns were identified with the peer reviewers assigned to this article; the investigation has concluded that the peer review process was compromised. We therefore can no longer trust the peer review process, and the article is being retracted with the agreement of the Chief Editor.

References

[1] B. Zhang, "Hybrid Algorithm for English Translation Speech Recognition Based on Deep Learning Model and Clustering," *Security and Communication Networks*, vol. 2022, Article ID 9308188, 11 pages, 2022.

[2] L. Ferguson, "Advancing Research Integrity Collaboratively and with Vigour," 2022, https://www.hindawi.com/post/advancing-research-integrity-collaboratively-and-vigour/.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 9308188, 11 pages
https://doi.org/10.1155/2022/9308188

Research Article

Hybrid Algorithm for English Translation Speech Recognition Based on Deep Learning Model and Clustering

Baicheng Zhang

School of Foreign Languages, Wuhan Polytechnic University, Wuhan 430023, Hubei, China

Correspondence should be addressed to Baicheng Zhang; zhangbaicheng2021@163.com

Received 29 March 2022; Revised 15 April 2022; Accepted 30 April 2022; Published 21 May 2022

Academic Editor: Mohammad Ayoub Khan

Speech recognition is the most important research direction in human-computer interaction. It is the key to the connection between human beings and machines and the expression of intelligence and automation in the information society. Taking English as the research object, using the related knowledge of speech recognition, it is based on the hidden Markov model technology of deep learning and clustering analysis algorithm and evaluated according to the cross-language English phonemic recognition system of sparse autoencoder (SA) method. By studying the speech recognition algorithm of the English translation, the influence of the speech recognition environment on the accuracy of speech recognition is confirmed. This provides a direction for humans to study speech recognition at a deeper level. Based on the language model of Transformer and the language model based on Seq2Seq, it sets different vocabularies, and the data are collected in the laboratory and outdoors, respectively, and the posttest template library is formed after collection. In the task of restoring phonetic symbols to English characters when phonemes are modeling units, the error rate is the lowest. The error rate on the test set reached 9.54%, which was 6.97 percentage points higher than that of the syllable modeling unit.

1. Introduction

The research of speech recognition technology has a history of nearly 100 years since the initial prototype of speech recognition [1, 2]. Speech recognition can not only make the computer receive and understand the information expressed by human beings more directly but also help people-to-people communication and human-computer communication through automated equipment and intelligent operation. It is an important bridge between people. The research of speech recognition occupies an important position in the field of scientific and technological development. HMM recognition technology has become the technology of modern speech recognition. The vast majority of existing person-neutral, large-vocabulary, continuous speech recognition systems are based on HMM models.

With the interest of the modern artificial intelligence industry and the development of deep learning theory and technology, computer scientists apply various computing methods to study speech recognition. Therefore, speech recognition technology has made great achievements in both theory and application. Speech recognition has moved from theoretical knowledge and laboratory to people's daily life, providing great convenience and expectation for daily life.

By describing the relevant theories of speech recognition technology, it covers the principles of speech recognition and the theoretical basis of deep learning. According to hidden Markov model technology and cluster analysis algorithm, according to the difference of syllables and phonemes, based on deep learning and modeling under different units, the acoustic model and language model of English speech recognition are designed. The Transformer-based language model has a slightly lower error rate than the Seq2Seq-based language model in the task of restoring phonemes and syllables to English characters.

2. Related Work

Speech recognition technology is widely used in the commercial market and has high commercial value. How to use this technology in a low-cost and reliable way in daily life is its future development direction. Zhu et al. analyzed remote sensing data through deep learning, reviewed recent advances, and provided resources that make deep learning in remote sensing seem ridiculously easy. He encouraged remote sensing scientists to bring their expertise to deep learning. And he uses it as an implicit universal model to address unprecedented, large-scale, and influential challenges such as climate change and urbanization [3]. Montazeri Ghahjaverestan et al. proposed a method for detecting apnea and bradycardia in premature infants based on a coupled hidden semi-Markov model (CHSMM). For simulated data, the proposed algorithm was able to detect the desired dynamics with 96.67% sensitivity and 98.98% specificity. The results show that the CHSMM-based algorithm is a robust tool for monitoring apnea remission in preterm infants [4]. Zhang started with the influence of cultural context on Chinese-English translation and discusses the context in Chinese-English translation combined with practical work experience as well as the understanding and practice of translation activities from the perspective of cultural translation. Between the two languages, due to the profound influence of culture, translators gradually form their own unique and personalized cultural understanding and translation concepts in translation practice [5]. Bharathi and Selvarani analyzed the occurrence, propagation, and transition of errors from start to finish execution cycle through the hidden Markov model (HMM) technique. Attempts at the design level can help design engineers to improve the quality of their systems in a cost-effective manner [6]. Zhihao adapted the TMS320DM365 series multimedia processors based on the current development and application direction of DaVinci digital image processing technology. The hardware circuit design of the system mainly includes a power management module, a serial port fault diagnosis module, and an Ethernet communication module. Finally, this paper studies and discusses the accuracy of trade English [7]. Jang and Hitchcock applied model-based cluster analysis to data on types of democracies, creating tools for typology [8]. Pakoci et al. used the largest existing Serbian speech database and the best n-gram-based language model made for general purpose, changing the parameters of the system to achieve the best word error rate (WER) and character error rate (CER). In addition to tuning the neural network itself, its layers, complexity, layer concatenation, etc. have explored other language-specific optimizations [9]. These studies are instructive to a certain extent, but the studies are too single and can be further improved.

3. Theories Related to Speech Recognition

In the process of speech recognition, the original signal of speech data is mainly collected by the machine [10–12]. When the collected speech raw data samples are detected by the speech recognition system, more than half of the recognition errors are due to endpoints. Therefore, the noisiness of the real environment in which the original signal is collected directly affects the difficulty of accurate endpoint detection of the speech signal [13]. Therefore, during speech recognition, it is necessary to cut off the silence at the beginning and end of the collected speech samples so as not to affect the later detection of speech recognition.

Speech recognition technology is a type of pattern recognition. The basic principle is that the machine processes, analyzes, recognizes, and understands the speech signal and converts it into text. Speech recognition technology involves many fields. In addition to basic applications, it can be combined with other natural language processing technologies such as spoken language recognition technology, speech synthesis technology, and machine translation technology to build more complex and intelligent applications. The first step in speech recognition is speech signal preprocessing. Speech signal preprocessing is the premise and foundation of speech recognition, and it is also a very critical step in the feature extraction of speech signals. Only when the characteristic parameters that can represent the essence of the speech are extracted in the preprocessing stage of the speech signal, the best similarity effect language can be obtained by comparing the compared speech with the standard speech.

3.1. Relevant Principles of Speech Recognition. The speech recognition system mainly includes two models: acoustic model and language model. Acoustic models are mainly classified according to the acoustic characteristics of speech signals. The language model performs semantic-level scoring on the feature discrimination results of the acoustic model. The two models are the key to the performance of the entire speech recognition system [14].

In the development of speech recognition technology, although different researchers have proposed many different solutions, the basic principles are the same. In the processing of the speech signal, any speech recognition system can use Figure 1 to represent its general recognition principle. The most important modules of the speech recognition system are speech feature extraction and speech pattern matching [15].

Speech recognition is to process the collected raw speech data samples and calculate according to the signal data and the acoustic parameters in the computer speech database. Then, the relevant characteristic parameters of the original speech data samples are inferred. Whether it is the recognition link or the training link in the speech recognition system, it is necessary to analyze and extract the characteristic parameters of the speech based on the preprocessing of the speech signal [16]. Then, it is compared with the training template library according to the installation rules of the audio signal function. It obtains the recognition result through the recognition algorithm. The quality of speech recognition results is directly related to the parameters and parameter selection in the template library. Its basic structure is shown in Figure 2.

The preprocessing in the principle block diagram of the speech recognition system is mainly to prehighlight and segment the collected speech signal and remove other noises or interference signals. It detects the front and back ends of

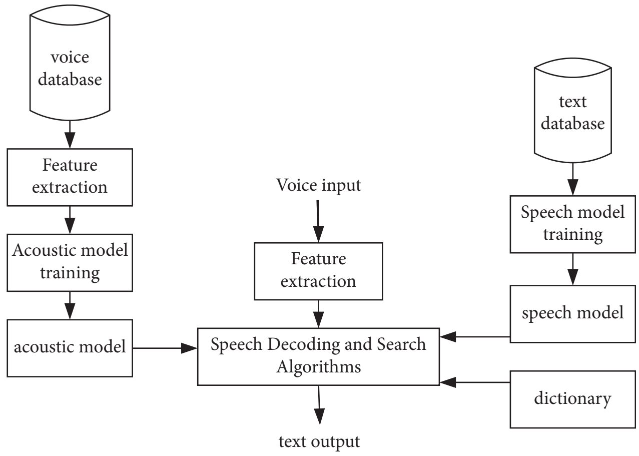


Figure 1: Block diagram of speech recognition system.

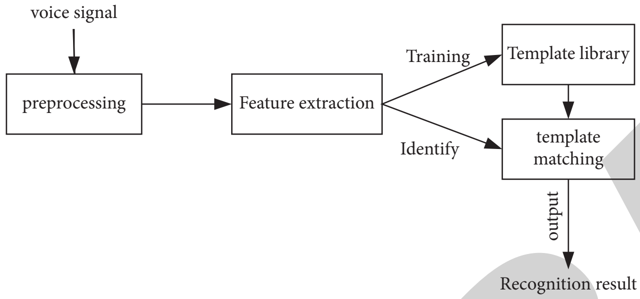


Figure 2: Block diagram of speech recognition system.

speech and retains valid speech fragments [17]. Similarity matching is performed on the extracted information according to the extracted speech segment to reduce the dimension and reduce the calculation amount of subsequent processing. It performs computational matching on amplitude, zero-crossing velocity, time-based energy, and frequency-based linear prediction coefficient *t*.

3.2. Dynamic Time Warping (DTW) Technology. Dynamic time warping is to make the reference template and test template different in time through the principle of dynamic programming to achieve the best match. It is to bend two speech sequences with different times on the time axis so that the two speeches can be better matched. In the speech evaluation system, the similarity between the user speech and the reference speech can be calculated by comparing the difference between the two characteristic parameters. However, since the speech to be evaluated and the reference speech have obvious differences in speech length and pronunciation speed, if the two are directly compared, the result is bound to be inaccurate. DTW is a very classic algorithm in speech recognition. The idea of the algorithm is to stretch or shorten the unknown until it is the same length as the reference template.

In speech recognition, the most frequently used discriminant algorithm is the dynamic time warping (DTW) algorithm [18]. DTW obtains a dynamic form that combines dynamic and time-coordinated methods to obtain distances between vectors. This is a classic course in speech reporting [19]. This workaround is to define a reference sample for each file and further calculate the corresponding path to get corrupted by the smallest sign in the vector distance. The common distance of the dashboard is minimal when the vectors are parallel. The meaning of differences between markers and specimen features addresses the stochastic problem of speech signals. The parameters involved in the dynamic time warping DTW technology mainly include speech feature vector, frame distortion and frame matching distance, and other related parameters.

Assuming that the frame vector parameter in the template is M, $\{R(1), R(2), L, R(m), L, R(M)\}, R(m)$ is the speech feature vector of the mth frame, the test template has N frame vectors, and $\{T(1), T(2), L, T(n), L, T(N)\}, T(n)$ is the speech feature vector of the nth frame. A match comparison point is denoted by (n, m) at each intersection where the parameter template and the test pattern intersect. The frame distortion at this intersection is $D[T(n), R(m)]$. The main purpose of the DTW algorithm is to make the measured feature template nonlinearly map to the reference template by determining an optimal time warping function $\phi(i_n)$. This minimizes the cumulative distortion D, which satisfies the following formula:

$$D = \min_{\phi(i_n)} \sum_{i_n=1}^{n} d\left(T(n), tRn\left(\phi\left(i_n\right)\right)\right). \tag{1}$$

The DTW algorithm to find the minimum distortion is shown in Figure 3.

After the DTW algorithm, the calculated MFCC parameters of the test speech and the standard speech are respectively used as the actual parameters of *t*, *r* in the function DTW(*t*, *r*), that is, as the parameter input value of the function DTW(*t*, *r*). By calculating the similarity between the two, the most similar voice to the test voice can be found in the standard voice library. Finally, the output of the program can achieve the purpose of speech recognition. DTW algorithms cannot fully exploit the temporal and dynamic properties of speech signals, making them suitable for relatively simple speech recognition systems such as isolated words and small vocabularies. DTW technology optimizes the calculation results as a whole without considering the local optimization problem, which is easy to use [19].

3.3. Hidden Markov Model Technology for Deep Learning. Deep Learning (DL) originates from the deep understanding of knowledge and is a new direction in the field of machine learning [20]. Deep learning is often applied to various supervised mode tasks such as speech recognition, natural language models, and image recognition. The traditional HMM model is currently the most widely used model, mainly based on statistical signals. It is mainly used in the modeling of speech recognition systems. HMM is widely used in various fields of speech processing, such as endpoint detection, speech compression, speech enhancement, and speech recognition. This method is now the mainstream of speech recognition technology.

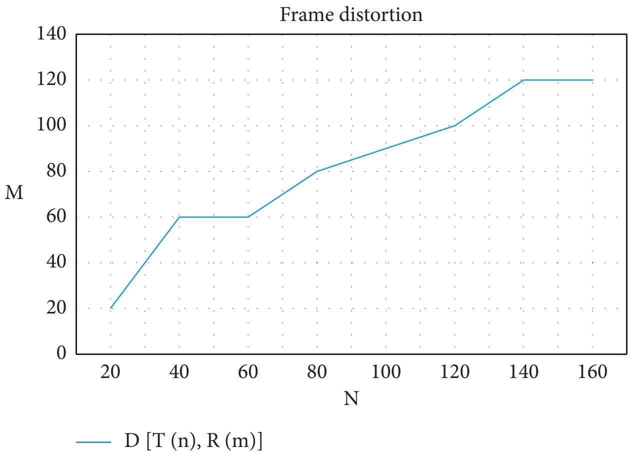


Figure 3: DTW algorithm to find minimum distortion.

(1) With the given observation value sequence $O = \{o_1, o_2, \ldots, o_r\}$, r is used to represent the frame number and model parameter $\lambda = \{\pi, A, B\}$ of the speech signal. By computing estimates, it effectively observes the probability of a sequence of values, finding the best model that matches the sequence of observations with multiple model parameters.

(2) If the observation sequence $O = \{o_1, o_2, \ldots, o_r\}$ and the model $\lambda = \{\pi, A, B\}$ are known, how to choose the corresponding optimal state sequence $Q = \{q_1, q_2, \ldots, q_t\}$? The problem is mainly to find the paths the model can take to generate this sequence of observations and take the path with the highest chance. In practical applications of this identification problem, optimization criteria are usually chosen to solve this problem.

(3) How to adjust the model parameter $\lambda = \{\pi, A, B\}$ to make $P(O \mid \lambda)$ the largest? This is a training process. It is used to train the HMM model parameters so that the recognition performance under the model parameters is the best.

3.4. Transformer Model. The fully connected neural network is fixed in the input and output layers, and the length of the input and output sequences of the recurrent neural network must be the same. In order to better realize the tasks where the length of input sequence and output sequence is not equal, such as machine translation and speech recognition, the Seq2Seq model is proposed, which consists of an encoder and a decoder [21]. Early Seq2Seq models used LSTMs or RNNs to map one sequence as input to another output sequence. Transformer is also structurally a Seq2Seq model, originally used in the field of machine translation. Different from the design of RNN and CNN, before the sequence of Transformer is input to the encoder and decoder, positional encoding is required to obtain the timing information of the sequence. The multihead self-annotation mechanism combines the context with the remote words sequentially, while processing all the words in parallel. The entire model framework completely adopts the multihead attention-grabbing mechanism and neural network for tracking. Its speed and training effect are better than the Seq2Seq model of RNN.

Although RNN-based speech recognition systems have ample room for development, they still have shortcomings such as poor dynamic deformation, long training time, and difficulty in implementation. The recognition rate is not necessarily better than cognitive-based speech recognition. Therefore, in the statistical model, this algorithm is only in the experimental research stage.

4. English Translation Speech Recognition System

A speech recognition system is essentially a model recognition system [22]. Speech recognition is a process of matching models and similar similarity measurement rules based on data in a database. It matches the acquired speech with the existing data model. The successful application of HMM technology in speech recognition lies in its powerful modeling ability for time series structure. But it still has certain limitations. In the process of speech matching, the English translation speech recognition system needs to calculate a large amount of speech data, and the HMM model algorithm needs to involve many parameters, so this makes the HMM model training time-consuming. However, the HMM model has high recognition accuracy and can meet the needs of real-time speech recognition in the daily time of the real society.

4.1. Speech Recognition System Based on HMM Model. The successful application of HMM in speech recognition has completely changed the history of speech recognition and has far-reaching effects [23]. As a statistical model, HMM was introduced into speech recognition in the 1970s, and in recent years, it has successfully realized the modeling of complex problems such as speech recognition and biological sequence analysis. The emergence of HMM has made a substantial breakthrough in speech recognition systems. To understand the HMM model, it must first introduce the concept of Markov chains. The Markov chain describes the changes of N states in a finite-state machine within time T. Let S represent the finite state set, $S\{S_1, S_2, \ldots, S_N\}$, then the state X_i of the state machine at a certain time t can only be equal to one of the states s_i in the finite state set S, where $t = 1, 2, \ldots, T, i = 1, 2 \ldots, N$. The state X of the state machine in time T constitutes a state chain $X = X_1, X_2, \ldots, X_T$ in chronological order, and its probability satisfies the following formula:

$$P = (X_1, X_2, \ldots, X_T) = \sum_{i=1}^{T} P(X_1, X_2, ..., X_{i=1}). \qquad (2)$$

A state chain X that satisfies the formula is called a Markov chain. Further, if the state chain X satisfies the "Markov assumption," the probability that the situation X_t of the state chain X at a certain time t belongs to the finite set S is only associated with its previous situation $X_{i=1}$. And it does not matter any time after time $t-1$. Then, the state chain $X = X_1, X_2, \ldots, X_T$ satisfies the following probability formula:

$$P = (X_1, X_2, \ldots, X_T) = P(X_1)\prod_{t=1}^{T} P(X_t \mid X_{t-1}). \quad (3)$$

At this time, the Markov chain composed of the state sequence X is called "homogeneous Markov chain."

Since there is no time $t = 0$, the state X_1 of the state machine at $t = 1$ is determined by matrix $\pi = [\pi_1, \pi_2, \ldots, \pi_n]$. The matrix π is the initial state probability distribution matrix. The components $\pi_i, i = 1, 2, \ldots, N$ of π, respectively, represent the probability that the initial state X_1 in the homogeneous Markov chain is equal to the ith state s_i in the finite state set, namely,

$$\pi_i = P(X_1 = S_i), \quad i = 1, 2, \ldots, N. \quad (4)$$

In addition to the probability distribution matrix π of the initial state, a square matrix $A = \{a_{ij}\}$ of order N is defined. The value of the square element a_{ij} represents the probability of one step transition from s_i to s_j, so the square matrix A represents the state transition matrix. The formula for calculating the element a_{ij} is as follows:

$$a_{ij} = P\big(X_{i+1} = S_j \mid X_t = S_i\big), \quad 1 \le i, j \le N. \quad (5)$$

To sum up, a first-order Markov chain λ can be represented as $\lambda = \{\pi, A\}$ by the initial state probability distribution matrix π and the state transition matrix A. HMM uses the Markov chain to simulate the change process of the signal and then indirectly describes the change through the sequence of observations. Therefore, it is a double random process, which can well describe the overall nonstationarity and short-term stationarity of speech signals.

4.2. Hmm Topology and Classification. HMM uses the states in the Markov chain to represent the pronunciation process of speech. During word generation, the system transitions from one state to another. An output is generated in each state until the word is output. According to the different state transition methods, HMM models have different topological structures. According to the structure, the common types are the first type, the HMM that experiences various states, as shown in Figure 4.

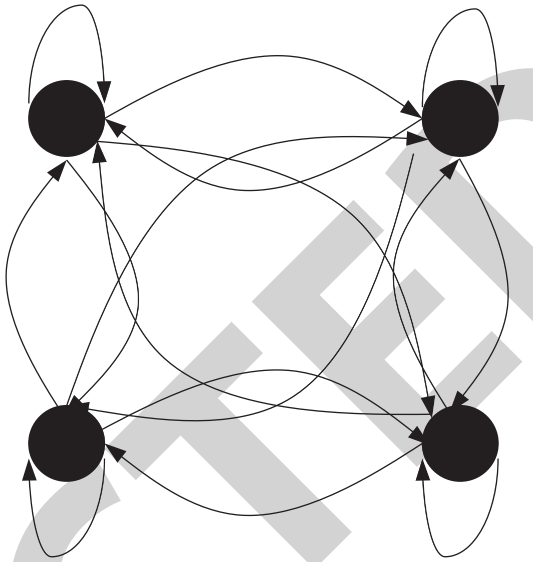


Figure 4: Topological diagram of HMM undergoing various states.

The second, the two-transfer HMM, is shown in Figure 5:

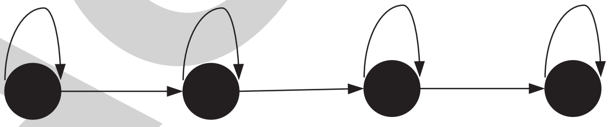


Figure 5: Topological diagram of the two-transfer HMM.

The third, three-transfer HMM, is shown in Figure 6:

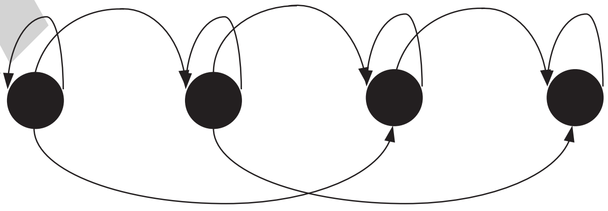


Figure 6: Topological diagram of the three-transfer HMM.

As can be seen from the figure, the states in the latter two topologies can only reside in the original state or perform state transitions from left to right. Therefore, this topology is also called the left-right model, that is, the transition must start from the first state [24]. This left-to-right HMM model is commonly used in speech recognition because this left-to-right HMM model is the main manifestation of the temporal structure.

The forward-backward algorithm is exactly the method used to solve the first problem of HMM. For a given HMM model $\lambda = \{\pi, A, B\}$ and observation sequence $O = \{o_1, o_2, \ldots, o_r\}$, according to the general idea, the method of calculating the output probability $P(O \mid \lambda)$ is as follows: if the given state sequence is $Q = \{q_1, q_2, \ldots, q_t\}$, there are

$$P(O \mid Q, \lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda) = b_{q1}(o_1)\, b_{q2}(o_2) \ldots b_{qr}(o_T). \quad (6)$$

The formula is in the q_t state, and the output is the product of the probabilities of o_t. And because for a given HMM model, the conditional probability of generating a state sequence $Q = \{q_1, q_2, \ldots, q_r\}$ is $P(Q \mid \lambda)$, and its calculation formula is

$$P(Q \mid \lambda) = \pi_{q1} a_{q1q2} \cdots a_{qr-1qr}. \quad (7)$$

Then under the conditions of the given HMM model, the probability of outputting the observation sequence $O = \{o_1, o_2, \ldots, o_T\}$ is

$$P(O \mid \lambda) = \sum_{\text{allQ}} P(O \mid Q, \lambda)\, P(Q \mid \lambda). \quad (8)$$

To perform $(2T-1)N^T$ multiplications and $N^T - 1$ additions is too computationally intensive. Therefore, it is necessary to consider a more concise and efficient algorithm, that is, a forward-backward algorithm.

4.2.1. Forward Algorithm. First, it defines the forward probability vector: $\alpha_t(i) = P(o_1, o_2, \ldots, o_t, q_t = \theta_i \mid \lambda)$ represents the probability that, given the HMM model λ, the partial observation sequence output from time 1 to time t is $\{o_1, o_2, \ldots, o_t\}$ and is in state θ_t at time t.

Then, it uses the defined forward vector $\alpha_t(i)$ to calculate the output conditional probability $P(O \mid \lambda)$, and the steps are as follows:

The first step: initialization. When $1 \leq i \leq N$, there are

$$\alpha_t(i) = \pi_i b_i(o_1). \tag{9}$$

The second step: recursive calculation. For all $1 \leq t \leq T-1; 1 \leq j \leq N$, there are

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{i+1}). \tag{10}$$

The third step: terminate the calculation.

$$P(O \mid \lambda) \sum_{i=1}^{N} \alpha_T(i). \tag{11}$$

The schematic diagram of the forward algorithm of the HMM model is shown in Figure 7:

4.2.2. Backward Algorithm. Similar to the forward algorithm, first define the backward probability vector: $\beta_t(i) = P(o_{t+1}, o_{t+2}, \ldots, o_T, q_t = \theta_i \mid \lambda)$, and the probability of being in state θ_i at time t. Then, it uses the defined backward vector to calculate the output conditional probability $P(O \mid \lambda)$ as follows:

Step 1: initialize. When $1 \leq i \leq N$, there are

$$\beta_T(i) = 1. \tag{12}$$

The second step: recursive calculation. For all $t = T-1, T-2, \ldots, 1; \quad 1 \leq i \leq N$, there are

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j). \tag{13}$$

Step 3: terminate the calculation.

$$P(O \mid \lambda) \sum_{i=1}^{N} \beta_1(i). \tag{14}$$

The recursive process of the backward algorithm of the HMM model is shown in Figure 8:

Therefore, the parameter reestimation problem and training problem are often solved by the Baum–Welch algorithm.

According to the HMM model definition, $\varepsilon_t(i, j)$ represents the probability at time t and i, given a specific model and training sequence O, and its expression is

$$\varepsilon_t(i, j) = P(q_t = i, q_{t+1} = j \mid o, \lambda). \tag{15}$$

It is deduced that

$$\varepsilon_t(i, j) = \frac{\alpha_i(i) a_{ij} b_j \beta_{t+1}(j)}{P(O \mid \lambda)}. \tag{16}$$

Then, the probability that the Markov chain of the model is in state i at time t is

$$\begin{aligned} \varepsilon_t(i) &= P(q_t = \theta_i, O \mid \lambda), \\ &= \sum_{j=1}^{N} \varepsilon_t(i, j) = \frac{\alpha_i(i) \beta_t(i)}{P(O \mid \lambda)}. \end{aligned} \tag{17}$$

In the formula, $\alpha_t(i)$ and $\beta_t(i)$ are the forward probability and the backward probability, respectively. From this, it can be deduced that the reestimation formula of HMM parameters is

$$\overline{\pi} = \varepsilon_t(i), \tag{18}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i, j)}{\sum_{t=1}^{T-1} \varepsilon_t(j)}, \tag{19}$$

$$\overline{b_{ij}} = \frac{\sum_{t=1}^{T} \varepsilon_t(i, j)}{\sum_{t=1}^{T} \varepsilon_t(j)}. \tag{20}$$

It can be seen from the above formula that the training process of the HMM model is a process of finding the extreme value of the functional. At present, there is no analytical method for this kind of problem [25]. Because the given training sequence O is finite, there cannot be an optimal way to estimate the parameter λ. The Baum–Welch algorithm uses the idea of recursion to find the parameters that make $P(O \mid \lambda)$ local maximum, which is the result of the reevaluation optimization. At the same time, HMM can also be used in isolated word speech recognition system in speech recognition, and the recognition rate is higher than the DTW method, which has a wide range of applicability.

4.3. Density-Based Clustering Algorithm Steps. As far as the principle of clustering is concerned, both the hierarchical method and the division method are measured on the basis of the distance measurement standard [26]. The main idea of density-based clustering is to continue clustering as long as the number of objects or data points in the adjacent area exceeds a certain threshold. The accuracy of the grid cluster placement is related to the size of the unit cell. Unlike density-based methods, its classification statistics are not measured using distinct distances at all but instead classify data objects belonging to a relevant density domain according to whether they belong to that density domain [27].

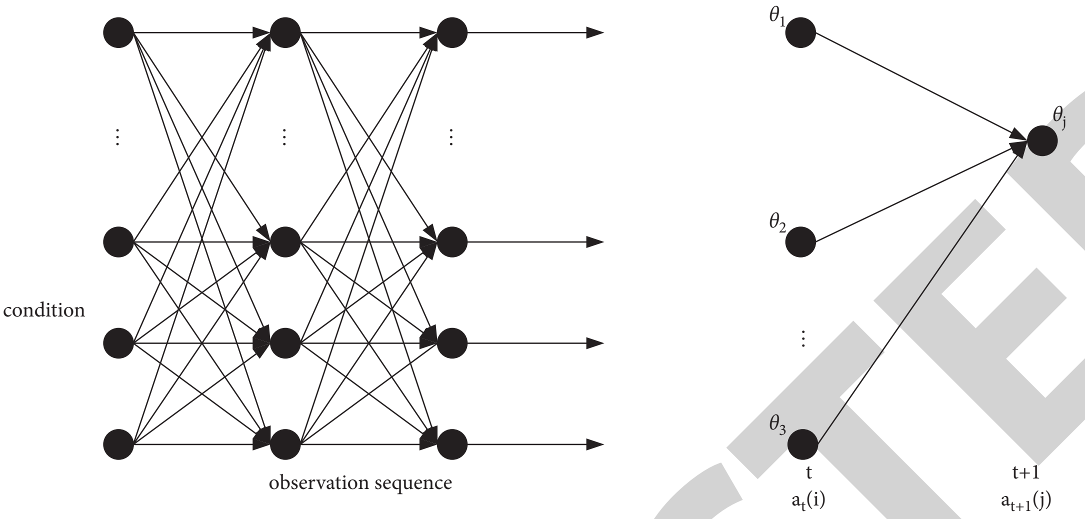


Figure 7: Schematic diagram of the forward algorithm of the HMM model.

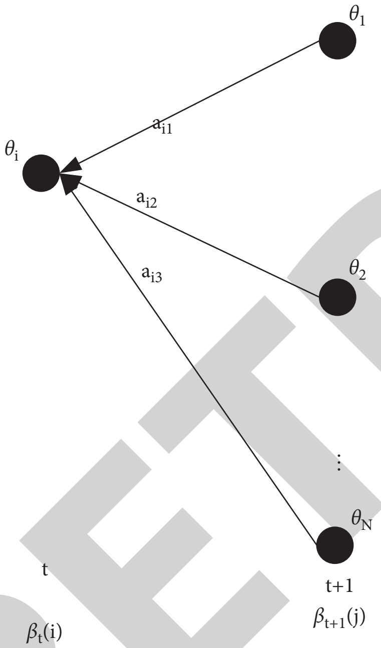


Figure 8: Schematic diagram of the recursive process of the forward algorithm of the HMM model.

(1) Input data and calculate Euclidean distance $d(x_i, x_j) = \sqrt{(x_i, x_j)^2}$.

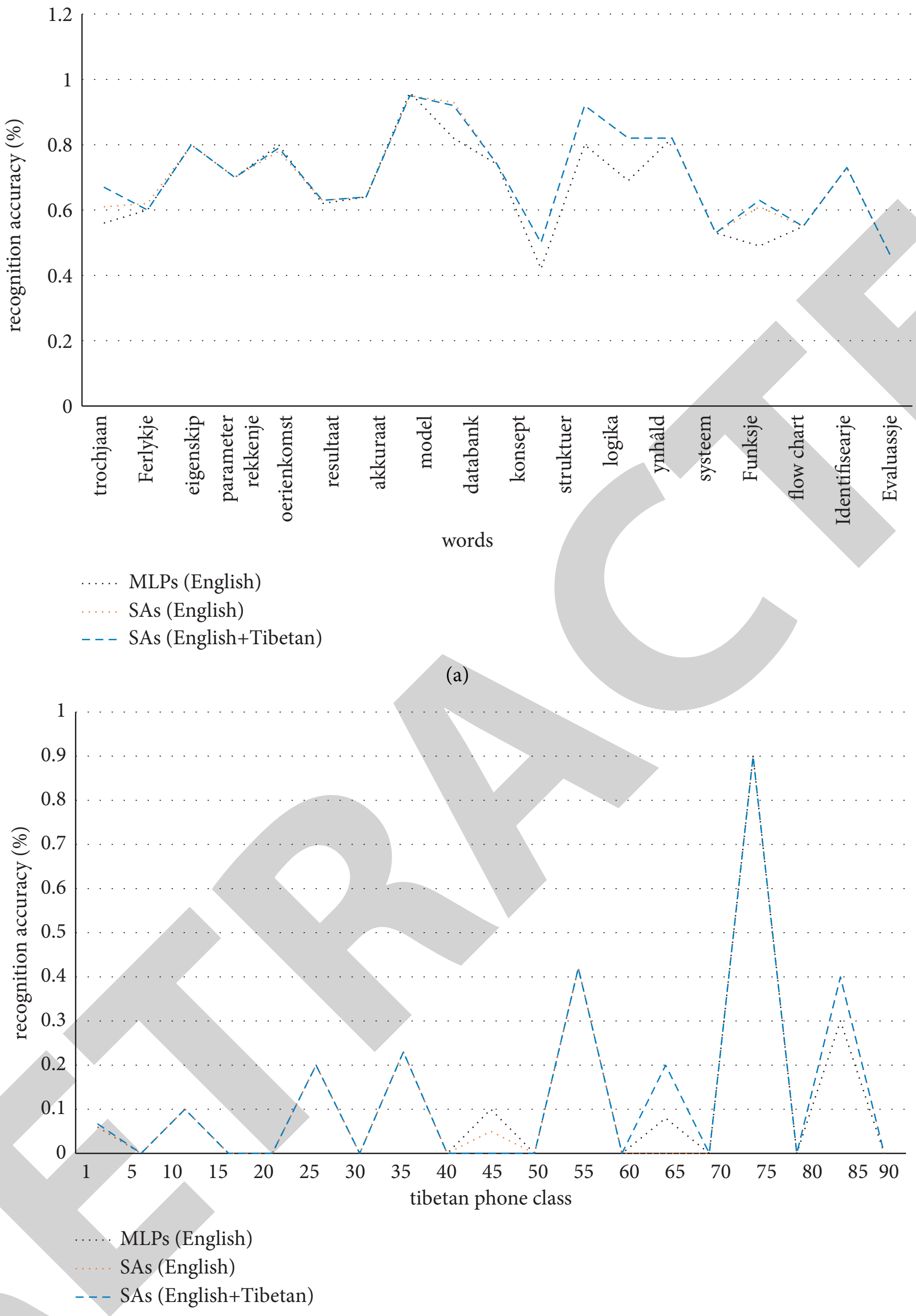
(2) The result obtained in step (1) is input, preprocessed, and given the parameter t used to calculate the cutoff distance d_c. It calculates the distance d_{ij}, and let $d_{ij} = d_{ji}$, $i/j \in S$, and determines the cutoff distance d_c.

(3) Calculate density $d_c\{\rho_i\}_{i=1}^{N}$ and generate descending subscript order $\{q_i\}_{i=1}^{N}$.

(4) Calculate $\{\delta_i\}_{i=1}^{N}$ and classification number $\{n_i\}_{i=1}^{N}$, determine the final cluster center, and output the result.

4.4. Experiments and Results. The objective methods commonly used to evaluate the performance of speech recognition models include word error rate, sentence error rate, and character error rate [28]. The experiments test the model performance by taking the word error rate, which is the error rate for phoneme recognition in English speech recognition.

The cross-language English phoneme recognition system is evaluated according to the sparse autoencoder (SA) method [29]. First, it compares the performance of SA and single-hidden-layer MLP in AF-based speech attribute detection. Then, it compares their performance in cross-lingual English phoneme recognition. The TIMIT English data set is used as the training data for the source language. 70% of the 1000 English continuous speech sentences are extracted as pretraining data, and the remaining data are used as test data [30]. These English continuous speech data sets can be downloaded from the Internet. The sampling rate of all original voice data is 8KHZ, each detection uses a 10 ms window function, and a 3 ms window is superimposed to extract a 39-dimensional MFCC specialization. The input layer of WSA and MLP has 39 nodes. From the TIM hit data set, including silence, it can get 34 English phonemes from English sentences. The 20 MLPs are all set into a structure of 18 hidden layer nodes and 2 output layer nodes, and then trained into 20 speech attribute detectors using the TIMIT data set. Likewise, 20 SA models are trained using the TIMIT and English data sets. In order to evaluate the results of the AF-based speech attribute embolizer and phoneme recognition, the evaluation criteria used were the attribute and phoneme unit-by-hour comparisons of speech [31]. For each frame, if the recognition results in the speech attribute detector and phoneme checker agree with the reference value, the score of S is incremented. The recognition accuracy (RA) is as follows: r represents the detection number of all speeches in the speech set. This evaluation criterion takes into account the phoneme recognition results and temporal information. The accuracy rate and recognition rate of English phonetic attributes using SAs and MLPs methods are shown in Figure 9. “English” and



(a)

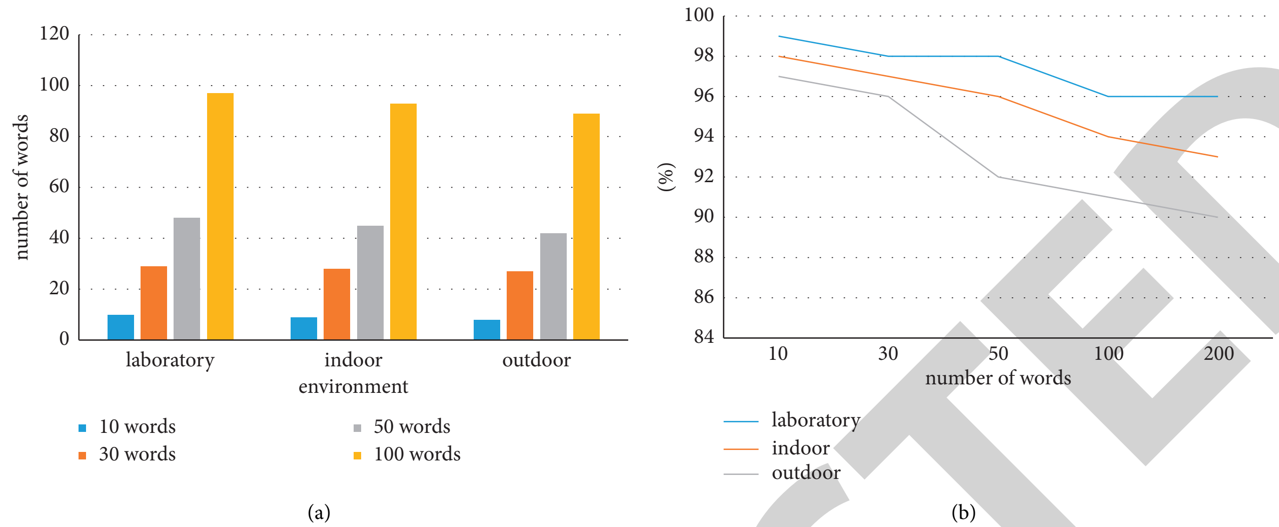
(b)

Figure 9: English speech accuracy and recognition rates for SAs and MLPs.

"English + Tibetan" in Figure 9 represent the languages used to train the model.

As can be seen from Figure 9, the SAs trained in a semisupervised way can detect AF speech attributes better than the other two methods, and it can recognize 14 Mandarin phonemes. However, the MLPs method and SAs trained in a supervised manner with English data can only recognize 10 and 12 Chinese phonemes, respectively. Furthermore, Figure 9 shows that the SAs trained on English and Mandarin data have higher phoneme recognition accuracy than the other two models. These results demonstrate that the sparse autoencoder trained in a semisupervised pretraining manner can learn shared phonetic properties between English and Mandarin. And the learned shared speech attributes can be used to effectively improve the accuracy of speech recognition.

In this test, 5 groups of different vocabulary sizes were set, which were 10, 30, 50, 100, and 200 isolated words, and 100 random tests were performed on each group. The test voices this time came from classmates, and 5 sets of data were collected in the laboratory and outdoors. After collection, this test template library is composed. The recognition rate test results are shown in Figure 10.



(a)

(b)

Figure 10: Number of correct recognitions and recognition rates in different environments.

Although the recognition rate in the outdoor environment is lower than that in the laboratory due to the influence of outdoor noise, the recognition results this time still meet the needs of practical applications. In addition, according to the test results, the larger the test vocabulary, the lower the recognition rate. However, the recognition rate can still reach more than 90%. In general, the recognition rate of the system can meet the actual use requirements.

According to the structure of the Transformer model, it belongs to the Seq2Seq model, so the Seq2Seq model is selected as the benchmark model, and the model performance on different modeling units is compared. The experimental results are shown in Table 1.

Transformer-based language models and Seq2Seq-based language models have the lowest error rates in the task of restoring phonetic symbols to English characters when phonemes are modeling units. The error rates on the test set reach 9.54% and 11.21%, respectively, which are 6.97 and 6.1 percentage points higher than the modeled units of syllables, respectively.

Under the same experimental conditions, the Transformer-based language model has a slightly lower error rate than the Seq2Seq-based language model in the task of restoring phonemes and syllables to English characters, and its test running speed is faster than that of the Seq2Seq model. This is because the Transformer model has the advantage of parallel computing.

In order to test the performance of the speech recognition system combined with the language model and the acoustic model, the acoustic model based on CNN-CTC was combined with the language model of Transfomer and compared with the speech recognition system using only the acoustic model based on CNN-CTC. Table 2 shows the test results of the speech recognition system by selecting different modeling units.

Table 1: Language model performance evaluation.

Model	Modeling unit	WER (%)	
		Validation set	Test set
Sep2Sec	Syllable	16.37	17.31
	Phoneme	10.09	11.21
Transformer	Syllable	15.23	16.51
	Phoneme	9.16	9.54

Table 2: Performance evaluation of speech recognition system.

Voice recognition system		Test set	WER (%)
Acoustic mode	Language model		
CNN-CTC	Without	Character	55.19
	Transformer	Syllable	47.21
		Phoneme	42.53

The experimental results show that the English speech recognition performance of the combination of the acoustic model and the language model is better than the speech recognition whose modeling unit is a word. In addition, in a speech recognition system in which the modeling units are syllables and phonemes, it is better to use phonemes as the recognition unit for training. Although the calculation of the language model is saved when the word is the modeling unit, the recognition speed is faster, but when the phoneme is the modeling unit, the recognition effect is better than that of the word and syllable as the modeling unit model. The error rate of its language recognition system reached 42.53, which was 12.66 percentage points higher than that of the word modeling unit and 4.68 percentage points higher than that of the syllable modeling unit. HMM technology mainly needs to make a priori assumptions about the current state sequence distribution in speech recognition. The ability to model high-level acoustic phonemes is weak, which makes acoustically similar words easy to confuse.

5. Discussion

Language is the main way for human thought and emotion to communicate. It is the behavioral performance of information and intelligence level and the crystallization of human civilization and wisdom. Although the effect of deep learning models on speech recognition is better than that of traditional models, the HMM model has a strong demand and dependence on data, and the size and quality of data directly affect the model effect. If it wants to realize the potential of the model, it needs to rely on a large amount of data for training.

The implementation details and processing difficulty of speech recognition systems vary, but the basic technical routes are similar. No matter which of the above speech recognition systems is used, appropriate technologies in modeling units, speech signal preprocessing, feature extraction, system modeling, and pattern matching must be selected. As a statistical model of speech signal, HMM can reasonably imitate human speech process. In the future, humans can realize some things more conveniently and quickly through voice interaction and enjoy more modern services.

6. Conclusion

In order to improve the performance of English speech recognition, this paper mainly studies the design of acoustic model and language model for English speech recognition. A speech recognition system with syllables and phonemes as modeling units is realized by combining the CNN-CTC-based acoustic model and the Transformer-based language model.

In this paper, there is no other preprocessing algorithm design except for the anti-interference of Markov adaptive learning noise environment. Subsequent data augmentation methods can be used to perform time warping, frequency masking, time masking, etc. on the time domain or frequency domain of the speech. In this way, meaningful speech signals can be extracted from the noise background, and the performance of the model can be improved.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] T.-Y. Kim, H. Ko, S.-H. Kim, and H.-D. Kim, "Modeling of recommendation system based on emotional information and collaborative filtering," *Sensors*, vol. 21, no. 6, p. 1997, 2021.

[2] Y. Liu and G. Fu, "Emotion recognition by deeply learned multi-channel textual and EEG features," *Future Generation Computer Systems*, vol. 119, pp. 1–6, 2021.

[3] X. X. Zhu, D. Tuia, L. Mou et al., "Deep learning in remote sensing: a comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[4] N. Montazeri Ghahjaverestan, M. B. Shamsollahi, D. Ge, A Beuchée, and A. I. Hernández, "Apnea bradycardia detection based on new coupled hidden semi Markov model," *Medical, & Biological Engineering & Computing*, vol. 59, no. 1, pp. 1–11, 2021.

[5] X. e. Zhang, "A study of cultural context in Chinese-English translation," *Region - Educational Research and Reviews*, vol. 3, no. 2, pp. 11–14, 2021.

[6] R. Bharathi and R. Selvarani, "Hidden Markov model approach for software reliability estimation with logic error," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 305–320, 2020.

[7] J. Zhihao, "Simulation of ocean surface temperature based on audio signal collection and accuracy of trade English translation," *Arabian Journal of Geosciences*, vol. 14, no. 16, pp. 1614-1615, 2021.

[8] J. Jang and D. B. Hitchcock, "Model-based cluster Analysis of democracies," *Journal of Data Science*, vol. 10, no. 2, pp. 297–319, 2021.

[9] E. Pakoci, B. Popović, and D. J. Pekar, "Improvements in Serbian speech recognition using sequence-trained deep neural networks," *SPIIRAS Proceedings*, vol. 3, no. 58, pp. 53–76, 2018.

[10] Z. Hass, M. Woodhouse, and G. Arling, "Using a semi-markov model to estimate medicaid cost savings due to Minnesota's return to community initiative," *Journal of the American Medical Directors Association*, vol. 22, no. 3, pp. 642–647, 2021.

[11] H. Wei, Y. Long, and H. Mao, "Improvements on self-adaptive voice activity detector for telephone data," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 623–630, 2016.

[12] H. Wei and N. Kehtarnavaz, "Determining Number of Speakers from Single Microphone Speech Signals by Multi-Label Convolutional Neural Network," *IEEE IECON*, 2018.

[13] T. Ochiai and T. Enomoto, "Multi-hazard evaluation using cluster Analysis-for designated evacuation centers of yokohama," *Journal of Geographic Information System*, vol. 13, no. 02, pp. 243–259, 2021.

[14] C. Li, H. J. Yang, F. Sun, J. M. Cioffi, and L. Yang, "Adaptive overhearing in two-way multi-antenna relay channels," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 117–120, 2016.

[15] H. Wang, T. Asefa, and A. Sarkar, "A novel non-homogeneous hidden Markov model for simulating and predicting monthly rainfall," *Theoretical and Applied Climatology*, vol. 143, no. 1-2, pp. 627–638, 2021.

[16] K. Leahy, C. Gallagher, P. O'Donovan, and D. T. J. O'Sullivan, "Cluster Analysis of wind turbine alarms for characterising and classifying stoppages," *IET Renewable Power Generation*, vol. 12, no. 10, pp. 1146–1154, 2018.

[17] J. Han, D. Zhang, G. Cheng, N Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[18] W. Peng, "Unreliable paratexts in intralingual translation: a case study of an excerpted English translation of san guo yan yi," *Language & Semiotic Studies*, vol. 5, no. 01, pp. 120–141, 2019.

[19] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 852–855, 2018.

[20] M. F. Dixon, N. G. Polson, V. O. Sokolov, and V. O. Sokolov, "Deep learning for spatio-temporal modeling: dynamic traffic flows and high frequency trading," *Applied Stochastic Models in Business and Industry*, vol. 35, no. 3, pp. 788–807, 2019.

[21] H. A. Haenssle, C. Fink, R. Schneiderbauer et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[22] S. S. Han, M. S. Kim, W. Lim, and G. H. I. S. E. Park, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

[23] F. Grassmann, J. Mengelkamp, C. Brandl, and S. M. E. B. A. I. M. C. B. H. F. Harsch, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, 2018.

[24] R. W. Jernigan and R. H. Baran, "Testing lumpability in Markov chains," *Statistics & Probability Letters*, vol. 64, no. 1, pp. 17–23, 2003.

[25] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *IET Biometrics*, vol. 7, no. 1, pp. 81–89, 2018.

[26] R. Ranjan, S. Sankaranarayanan, A. Bansal et al., "Deep learning for understanding faces: machines may Be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66–83, 2018.

[27] A. Andrea, D. Amir, and E. Jean-Pierre, "Large deviations theory for Markov jump models of chemical reaction networks," *Annals of Applied Probability*, vol. 28, no. 3, pp. 1821–1855, 2018.

[28] T. Um, S.-Y. Koh, and N. Chung, "Comparison of movement density and destination cluster analysis of residents and tourists using Jeju Island navigation data before and after the COVID-19," *The Journal of Internet Electronic Commerce Resarch*, vol. 21, no. 2, pp. 35–51, 2021.

[29] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: deep learning for the internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.

[30] L. Tan and J. Xie, "Research and engineering evaluation practice of medical equipment purchase decision model based on improved Markov model," *Zhongguo yi liao qi xie za zhi=Chinese journal of medical instrumentation*, vol. 45, no. 3, pp. 344–348, 2021.

[31] A. Obeidat and T. S. Binti Mahadi, "The English translation of idiomatic collocations in the noble quran: problem and solutions," *Issues in Language Studies*, vol. 9, no. 2, pp. 78–93, 2020.