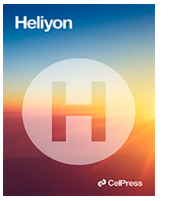
Heliyon 9 (2023) e21625

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Application of dynamic time warping optimization algorithm in speech recognition of machine translation

Shaohua Jiang [a,b], Zheng Chen [c,*]

[a] School of Humanities, Fujian University of Technology, Fuzhou 350118, China
[b] Krirk University, Bangkok 10220, Thailand
[c] Concord University College, Fujian Normal University, China

ARTICLE INFO

Keywords:
Speech recognition
Dynamic time warping algorithm
Feature extraction
Deep learning algorithm

ABSTRACT

Speech recognition is the foundation of human-computer interaction technology and an important aspect of speech signal processing, with broad application prospects. Therefore, it is very necessary to recognize speech. At present, speech recognition has problems such as low recognition rate, slow recognition speed, and severe interference from other factors. This paper studied speech recognition based on dynamic time warping (DTW) algorithm. By introducing speech recognition, the specific steps of speech recognition were understood. Before performing speech recognition, the speech that needs to be recognized needs to be converted into a speech sequence using an acoustic model. Then, the DTW algorithm was used to preprocess speech recognition, mainly by sampling and windowing the speech. After preprocessing, speech feature extraction was carried out. After feature extraction was completed, speech recognition was carried out. Through experiments, it can be found that the recognition rate of speech recognition on the basis of DTW algorithm was very high. In a quiet environment, the recognition rate was above 93.85 %, and the average recognition rate of the 10 selected testers was 95.8 %. In a noisy environment, the recognition rate was above 91.4 %, and the average recognition rate of the 10 selected testers was 93 %. In addition to high recognition rate, DTW based speech recognition also had a very fast speed for vocabulary recognition. Based on the DTW algorithm, speech recognition not only has a high recognition rate, but also has a faster recognition speed.

1. Introduction

Sound is the most widely used, natural, and fundamental information carrier in human communication. In today's highly informationized society, the application of speech recognition technology has become an indispensable part of social development, with very broad development prospects. Speech recognition essentially involves intelligent machines using certain speech recognition algorithms to convert valuable content from human speech into appropriate digital signals, which are then analyzed and processed to determine the semantic content of these speech signals. It is one of the important steps in achieving human-computer interaction. When using traditional methods for speech recognition, there are often reasons that may lead to poor performance in speech recognition. The DTW algorithm is a nonlinear technology that combines time calculation and distance measurement. Its application in speech recognition can effectively improve recognition rate, enhance speech recognition effect, and meet real-time requirements,

* Corresponding author.
E-mail address: sophia_FP@126.com (Z. Chen).

https://doi.org/10.1016/j.heliyon.2023.e21625
Received 14 June 2023; Received in revised form 7 October 2023; Accepted 25 October 2023
Available online 27 October 2023

which greatly solves the problems existing in traditional speech recognition.

Speech sound, as the most natural medium of communication in communication systems, plays an important role at all times. Natural speech recognition has become a research hotspot and has been widely applied. Ravanelli Mirco used a recurrent neural network to establish a speech recognition model that utilized temporal context to improve the efficiency of speech signal processing and the accuracy of speech recognition [1]. Zhang Yu summarized the potential impact of using giant automatic speech recognition models, which were pre trained using large-scale audio and diverse unlabeled datasets and found to greatly improve the efficiency of speech data processing [2]. Zhang Zixing believed that deep neural network-based methods had become potential substitutes for traditional unsupervised methods, and after sufficient training, they could alleviate the shortcomings of supervision in various real-world acoustic environments, bringing single channel and multi-channel technologies to the front-end and back-end development of speech recognition systems [3]. Kumar Tribhuwan introduced a new speech and emotion recognition technology based on recurrent neural networks, and used this technology to study feature extraction, improvement, and segmentation of speech emotion recognition. The experimental results showed that his proposed technology effectively improved the accuracy of speech recognition algorithms [4]. Ma Pingchuan believed that advances in deep learning and the availability of large audiovisual datasets led to the development of more accurate and robust visual speech recognition models based on predictive auxiliary tasks, and emphasized the importance of hyper-parameter optimization and appropriate data augmentation [5]. The purpose of Afouras Triantafyllos' research was to identify the spoken phrases and sentences by the speaking face, regardless of whether there is audio. Unlike previous works that focused on identifying a limited number of words or phrases, his study used lip reading as an open-world problem-unrestricted natural language sentences and addresses in videos [6]. However, these scholars' research on speech recognition is not comprehensive, and there are interference factors such as noise, speed changes, and speech differences, as well as poor robustness. They mainly rely on speech rules and language models, and have poor recognition performance and delay for standard pronunciation, different accents, dialects, etc., which is not suitable for real-time application scenarios. Moreover, speech recognition research based on DTW algorithm can have a good effect. However, DTW algorithm has good robustness to nonlinear changes and timing differences, which can improve the accuracy of speech recognition system. At the same time, because the DTW algorithm has a good tolerance for the influence of different speakers, accents and environmental noise in speech, applying it to speech recognition can improve the robustness of the recognition system to diversified speech signals, and make it more suitable for practical application scenarios.

The DTW algorithm has been widely utilized in many fields to solve the matching problem of time series, which has significantly improved the recognition efficiency in this field. The use of DTW algorithm can effectively extract features, obtain feature sequences, and match the sequences with the established templates, effectively improving recognition rate [7]. Hu Y proposed a DTW algorithm for feature classification. He used the algorithm to classify the inherent physiological features of the human body, and carried out experiments on the feature classification results. He found that the recognition rate of the method for human gait recognition was up to 90 %, with good robustness [8]. TAKAYAMA Natsuki described a weak supervised learning method for continuous sign language word recognition, which included forced alignment based on dynamic time warping. Through experiments, it can be found that the average recognition rates of the initial model under the conditions of open signature and open trial were 74.82 % and 91.14 %, respectively. In addition, the average recognition performance of the adjusted model under all conditions exceeded 65.00 % [9]. Zheng Jianbin proposed a real-time motion pattern recognition algorithm based on Gaussian Mixture Model (GMM) features and DTW. The spatial distribution of sensor data was represented by GMM features. In addition, he expanded his research on curved paths and further validated the rationality of DTW algorithm for feature matching to improve recognition rate [10]. Overall, there is a considerable amount of research on DTW in improving recognition rates. However, the use of DTW algorithm for speech recognition is relatively limited. In order to improve the research related to speech recognition, the DTW algorithm is integrated to improve the recognition rate of speech recognition.

The most important purpose of speech recognition is to enable machines to understand human language, equip machines with the ability to understand human speech content, and enable machines to convert speech signals into appropriate text or commands through recognition and understanding [11]. Traditional speech recognition methods generally directly utilize acoustic features for matching, which may be affected by the speaker's speech and the environment in which they are located, resulting in low recognition rates and low recognition efficiency. For this reason, this paper used DTW algorithm to study speech recognition and further improve the acoustic model of speech recognition, and converted acoustic features into a posterior probability distribution, so as to improve the recognition rate and speech recognition efficiency.

The uniqueness of this study was based on the use of the DTW optimization algorithm in speech recognition to enhance the functionality of the system and address real-world issues by increasing recognition accuracy, compensating for shortcomings in conventional algorithms, and integrating with other technologies. The originality of these studies resides in the DTW algorithm's use in speech recognition, which analyzes its benefits in terms of recognition precision and adaptability and offers fresh approaches and ideas for the advancement of speech recognition technology.

The first chapter of this paper is an introduction, which mainly summarizes the research background and importance of DTW algorithm applied in the field of speech recognition. It also introduces the research of others on speech recognition and DTW algorithm, and focuses on the contribution and novelty of the research in this paper. The second chapter is to introduce the factors affecting speech recognition, mainly including acoustic features, acoustic models and language models. Chapter 3 introduces the DTW algorithm and explains why the DTW algorithm is chosen to study the speech devices. The fourth chapter is the research of speech recognition based on DTW algorithm, which mainly introduces the specific of DTW algorithm applies to speech recognition. In the experimental analysis part of chapter 5, the DTW algorithm is compared with the deep learning and artificial swarm algorithm.Finally, there is the conclusion.

2. Features affecting the speech recognition effect

All speech recognition systems require a suitable data sample library to store all items that can be recognized by the system [12,13]. According to the number of sample entries, they can be divided into three different categories: small vocabulary, medium vocabulary, and large vocabulary. For a speech recognition data sample library, the more samples there are, the more vocabulary can be recognized, and of course, the more difficult it is to design. When the vocabulary is small, the number of recognition samples is small, and the difficulty of design is also relatively small. The main function of speech recognition is to convert speech signals into appropriate text information, and the components of this system are very rich [14].

Usually, language recognition systems with a very large vocabulary are mostly based on the DTW algorithm for vocabulary recognition training. After extensive training, the feature sequence $P_I^K = (P_1, P_2, P_3 \cdots, P_k)$ of speech is given, and then combined with acoustic and language models to generate the feature sequence. Mathematically, this expression is displayed in the following formula (1):

$$M_i = arg\ \underset{M}{max} Q\left(M \middle| P_I^K\right) = arg\ \underset{M}{max} \frac{Q\left(P_I^K | M|\right)}{Q\left(P_I^K\right)} \tag{1}$$

Among them, $Q(M|P_I^K)$ is the acoustic model, and $Q(P_I^K)$ is the probability of acoustic feature P_I^K.

2.1. Acoustic features

Acoustic features can represent acoustic signals in some ways, and the quality of acoustic features can have a significant impact on the performance of speech recognition systems [15]. Therefore, the extraction of acoustic features is very important. It is necessary to extract the most recognizable features from the collected speech data, and the model trained with this feature data can perform well. Due to the many differences in the producers of sound, there are also differences in the sound produced. How to remove some unique features of the sound generator from the acoustic features and preserve the same parts of the speech content of the sound generator is very important for improving recognition performance. Linear prediction means that the speech signal at a certain time can be expressed linearly by the signal at a certain time. The processing formula is as follows (2):

$$CEP'(x) = DTE^{-1}(In|DTE(Frame(x))|) \tag{2}$$

Among them, $Frame(x)$ represents the speech signal at frame x.

2.2. Acoustic model

Acoustic models play a crucial role in speech recognition systems, as they describe the process of converting acoustic primitives to generate feature sets [16,17]. Given an acoustic feature vector, its probability value for each primitive is calculated based on the acoustic model. By using the maximum likelihood criterion, the state order corresponding to the feature order is obtained, which helps to improve system performance. Speech recognition uses some elements as universal primitives.

The distribution of speech signal features is difficult to explain using simple feature functions. In practical applications, the DTW function is often used to adjust the speech signal, and the output probability is mainly represented by the DTW function, namely (3):

$$f_{ij}(P) = Q(P|i,j) = \frac{1}{(2\pi)^{q/2}\left|\sum_{ij}\right|^{1/2}} exp\left\{ -\frac{1}{2}\left(P - \gamma_{ij}\right)\sum\nolimits_{ij}^{-1}\left(P - \gamma_{ij}\right) \right\} \tag{3}$$

Among them, P represents the output status. $Q(P|i,j)$ is the universal primitive in the speech acoustic model, and γ_{ij} is the sound input signal.

2.3. Language model

Usually, language models describe human language, with a main focus on expressing the internal relationships that may exist between words in language. Therefore, a good language model can help improve the speed of decoding speech, while ensuring speed and improving recognition accuracy. Language models can generally be divided into two types, namely rule models and statistical models. Probability can be used to represent the likelihood of word order appearing in a language environment. For word order $M_1^t = \{M_1, M_2, M_3, \cdots, M_t\}$, its probability value can be expressed as (4):

$$Q\left(M_1^t\right) = Q(M_1)Q(M_2|M_1)Q(M_3|M_1M_2)\cdots Q(M_t|M_1M_2\cdots M_{t-1}) \tag{4}$$

Among them, $Q(M_1)$ represents the probability of M_1 occurring, and $Q(M_2|M_1)$ represents the probability of M_2 when M_1 is known.

3. Introduction of dynamic time warping optimization algorithm

DTW algorithm is a special method used to calculate the difference between two time series [18,19]. The application of DTW algorithm in speech recognition is very popular, and it has been well applied and developed in the field of speech recognition. The DTW algorithm is mainly utilized in speech recognition to calculate the distance between speech templates. The reason why it is necessary to solve the problem of distance between templates is because during the process of speech recognition, training the extracted speech can generate reference templates. After preprocessing, feature parameter templates of the same type are obtained for these voices. Finally, the feature parameter template and reference template are compared to obtain the final recognition result. The smaller the distance between these two templates, the better the recognition effect. Assuming that the test template has n frame feature parameter vectors and the reference template has X frame feature parameter vectors, denoted as W and Y, respectively, and that function X(n) satisfies the following formula (5):

$$T_n = \min_{X(n)} \sum_{n=1}^{N} p[W(n), Y(x(n))] \tag{5}$$

Among them, $p[W(n), Y(x(n))]$ is the feature parameter vector of the template to be tested in frame n, and the feature parameters of the reference template in frames $W(n)$ to $x(n)$.

Generally speaking, when the dimensions and conditions between two templates are the same, the similarity between these two templates can be expressed using Euclidean distance. However, when there are differences between templates, it is necessary to use the DTW algorithm to expand or reduce to the same ordinal number, and then calculate the distance [20]. DTW can be used as a comparison module for data matching in speech recognition systems, such as identifying whether two words represent the same word. In addition, DTW can also be used for gesture recognition, image processing, time series, and data scenes. In addition, it is also widely used in the fields of mining and information retrieval.

The basic idea of the DTW algorithm is as follows (6):

$$\begin{cases} (x_i, y_j) = (x_{i-1} + 1, y_{j-1} + 2) \\ (x_i, y_j) = (x_{i-1} + 1, y_{j-1} + 1) \\ (x_i, y_j) = (x_{i-1} + 1, y_{j-1}) \end{cases} \tag{6}$$

Among them, x_i and y_j are the number of test speech frames and the number of training speech frames, respectively.

By using the DTW algorithm for speech recognition, it can be found that its sensitivity to endpoint detection in speech recognition is very high, which can effectively compensate for the shortcomings of traditional speech recognition in terms of poor detection performance and accuracy when the background noise is too large, and avoid unexpected errors in endpoint detection results.

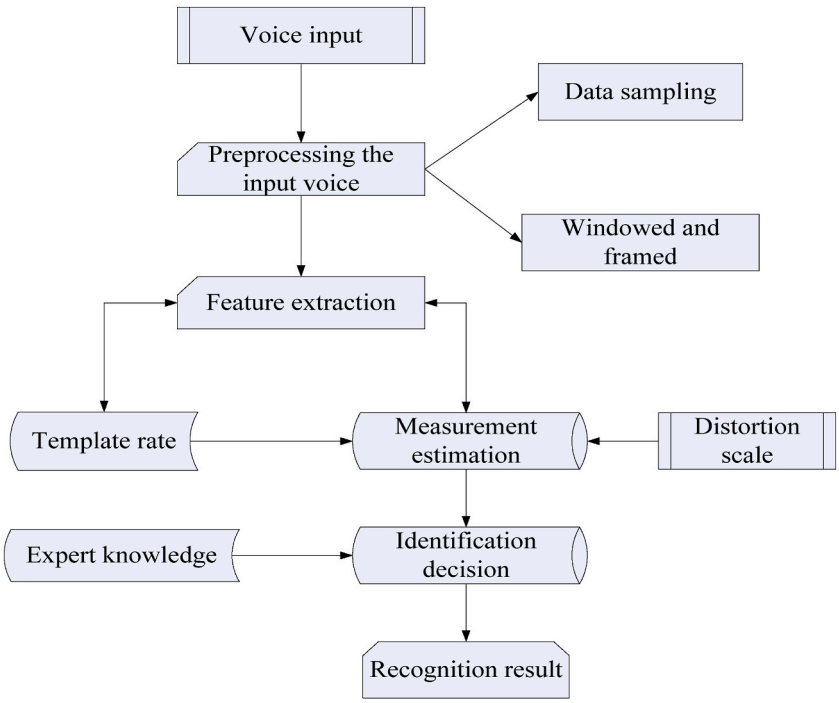


Fig. 1. Flow chart of speech recognition framework based on DTW.

The DTW algorithm is chosen because it has a wide range of applications, including motion and handwriting recognition in addition to speech recognition. The potential and impact of DTW algorithms in various fields can be extended and applied by studying DTW-based speech recognition applications. The DTW algorithm can handle voice samples with low data quality and is robust against noise and signal distortion. In real applications, the recognition system frequently has to deal with issues like a noisy environment or distorted communication, and in these situations, the DTW algorithm has some advantages. In light of the aforementioned objectives and justifications, the implementation of DTM optimization algorithm in speech recognition intends to increase the system's accuracy, resilience, and responsiveness in order to handle the variety of speech signals and real-world application requirements.

4. Speech recognition based on dynamic time warping optimization algorithm

The process of speech recognition on the basis of DTW is basically the process of recognizing speech signal patterns [21,22]. Due to the fact that speech signals are typically non-stationary signals mixed with various sounds, they cannot be directly used for feature parameter extraction and require preprocessing. After preprocessing the extracted data, the next step of feature extraction can be carried out. The feature parameters extracted from speech data can be processed to a certain extent, which is beneficial for producing equivalent templates, and these targets can be stored in the template database. After the extraction is completed, the recognition phase can begin. Generally, the speech parameters are obtained first, and then the obtained speech parameters are matched with the reference template. The reference template with the highest matching score is used as the final recognition result. The flowchart of the speech recognition framework on the basis of DTW is shown in Fig. 1.

4.1. Speech preprocessing

4.1.1. Sampling

The speech data for speech recognition needs to be processed first. Most early audio data undergoes constant changes in signal amplitude or generation time. This requires that before conducting speech recognition, the original analog audio signal needs to be sampled and quantized. The purpose of this is to convert the data signals required in the experiment into digital audio signals that the system can process. Applying the DTW algorithm to speech recognition can better assist the speech recognition system in collecting speech data, and its speed for data collection is much higher than other methods of speech recognition systems [23,24]. The extracted speech waveform is shown in Fig. 2.

Each frame of the current preprocessed speech signal is transformed, and the signal is converted from the time domain to the frequency domain. The formula is as follows (7):

$$Q^t(m) = \sum_{k=0}^{K-1} q(k) e^{-j2\pi km/K}, 0 \le m \le K-1 \tag{7}$$

Among them, $q(k)$ is the preprocessed speech frame signal. $Q^t(m)$ is the frequency domain of the time domain Q^t by m transformation.

4.1.2. Adding windows and framing

Speech is considered an unstable signal, but through experiments, it can be found that speech can remain stable for a certain period of time and has quasi stationary characteristics. Before conducting speech recognition, appropriate processing and analysis can be carried out on this stable period, and the time range of this stable period is usually set to 10–30 ms. By windowing and framing, the input data signal can be divided into several short periods, and the speech signal can be segmented through the window function to output the final audio signal frame by frame. The calculation method is as follows (8):

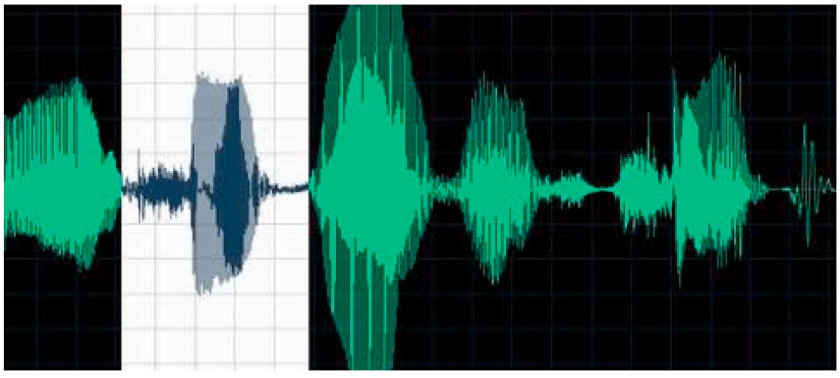


Fig. 2. A speech wave.

$$x_i(y) = \frac{1}{2}\sum_{m=0}^{M-1}\left|sgn[t_m(y)] - sgn[t_m(y-1)]\right| \tag{8}$$

Among them, $sgn[]$ is a symbolic function. t_m is the signal cycle; $x_i(y)$ is the audio signal, and y is the cycle range.

4.2. Speech feature information extraction

There are many signals in human generated speech, and these signals also include many feature parameters. Different speakers produce different speech, so the included speech feature parameters are also different, and these different feature parameters often represent different acoustic features. The feature parameters of speech signals can be regarded as simplified speech signals, so these parameters can be compressed. Of course, the prerequisite for compression is that it does not affect the recognition results. DTW based speech recognition can better extract speech feature parameters, which can help reduce irrelevant information in speech signals and enhance the recognition rate of speech recognition.

The feature extraction in speech recognition is also known as front-end processing. One of the important factors affecting recognition effectiveness is the selection of features, and good feature selection is conducive to presenting better recognition results. When selecting phonetic features, it is important to reflect the differences between words with different phonemes as much as possible, while the differences between words with the same phoneme should be small. At the same time, it is also necessary to consider the calculation frequency of feature parameters, and minimize the feature dimension as much as possible to reduce storage requirements while maintaining a high recognition rate.

Assuming that the input of the model is $X(y)$; the output is $Z(y)$, and the DTW function of the model is $F_t(m)$, the model parameters can be solved using the transfer function method (9):

$$F_t(m) = Q\frac{1+\sum_{i=1}^{n} x_I y^{-i}}{1-\sum_{i=1=1}^{q} z_I y^{-i}} = Q\frac{X(y)}{Z(y)} \tag{9}$$

5. Speech recognition experiment based on DTW

The speech signal undergoes two stages of playback and A/D conversion before entering the experiment, and then enters the MATLAB software system for data processing and analysis. The compact fully directional back pole type elect body acoustic and electrical transducer was used for the speech system's acoustic and electrical conversion. Table 1 is the descriptions of the microphone's specifications. The 2V power supply was used by the capacitive electret acoustic transducer. The values indicate that the speaker was 1 M from the microphone. The speech signal acquisition conversion was performed using the USB-6218 multifunctional DAQ data acquisition device. The acquisition card features a USB interface and a multi-function DAQ module, and it has high precision, high sample rates, and isolation from the bus power supply.

In this paper, the speech recognition experiment is based on DTW algorithm, and the experimental indicators used are the accuracy of recognition in different environments, the speed of different number of speech word recognition and the recognition rate of different speech types.

As mentioned earlier, the application of DTW algorithm in speech recognition can effectively improve the accuracy and recognition rate of speech recognition by applying DTW algorithm to speech recognition systems. For this purpose, 10 testers were randomly selected and 50 voice commands were randomly selected to have the 10 testers tested. These 10 testers were tested in quiet and noisy environments respectively, using DTW based speech recognition to recognize the speech of the testers in these two environments. After a period of calculation and matching with the reference template, the total recognition rate of each student in different environments was ultimately obtained. In order to better demonstrate the superior recognition rate of speech recognition systems based on DTW, experimental results were obtained through experiments. The experimental results were compared with those of Nassif Ali Bou [25] using deep learning (DL) for speech recognition, in order to enrich the comparison. Then, the experimental results were compared with the experimental results of Wang Min [26] using the artificial bee colony (ABC) algorithm to identify WNN (wavelet neural network) speech. The specific comparison results are shown in Fig. 3.

Fig. 3A: Quiet environment, Fig. 3B: Noisy environment.

As shown in Fig. 3, on the basis of the DTW algorithm, the speech recognition rate of the randomly selected 10 participants in both quiet and noisy environments was much higher than that of the other two types of speech recognition. As shown in Fig. 3A, in a quiet

Table 1
Performance parameters of the sound signal converter.

Sensitivity	RL = 2.2 kΩ , -44 ± dB(0 dB = 1V/Pa , 1 kHz
Frequency response range	20-16 kHz
Directivity	Full directivity
Standard operating voltage	3V
Signal-to-noise ratio	>60 dB

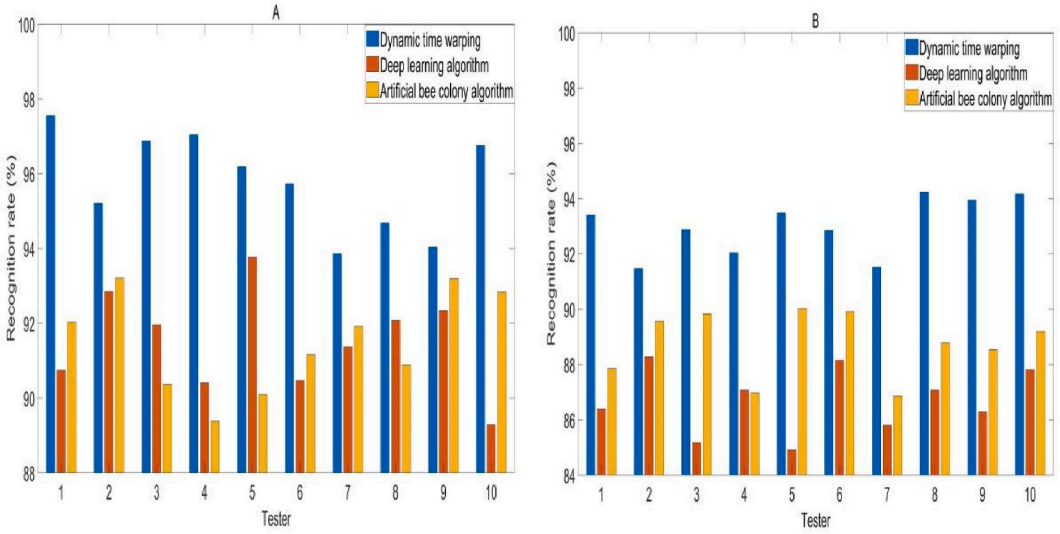


Fig. 3. Speech recognition rates under three different algorithms in different environments.

environment, the speech recognition recognition rate on the basis of the DTW algorithm was above 93.85 %, and the average recognition rate of the 10 selected testers was 95.8 %. The speech recognition based on deep learning algorithm and artificial bee colony algorithm was below 93.8 % and 93.3 %, respectively, and the average recognition rate of 10 testers extracted from these two speech recognition algorithms was 4.27 % and 4.29 % lower than that of speech recognition on the basis of DTW, respectively. Among them, speech recognition on the basis of DTW algorithm had the lowest recognition rate among the 7th tester, only 93.86 %, but it was 2.49 % and 1.94 % higher than speech recognition based on DL algorithm and ABC algorithm, respectively. The speech recognition on the basis of DL algorithm had the highest recognition rate among the 5th tester, with 93.77 %, but it was 2.42 % lower than the speech recognition on the basis of DTW algorithm. The speech recognition on the basis of the ABC algorithm had the highest recognition rate in the second tester, at 93.21 %, but it was 2 % lower than the speech recognition on the basis of the DTW algorithm.

As shown in Fig. 3B, in a noisy environment, the speech recognition recognition rate on the basis of the DTW algorithm was above 91.4 %, and the average recognition rate of the 10 selected testers was 93 %. However, speech recognition on the basis of deep learning algorithms and artificial bee colony algorithms were below 89 % and 90.1 %, respectively, and the average recognition rates of 10

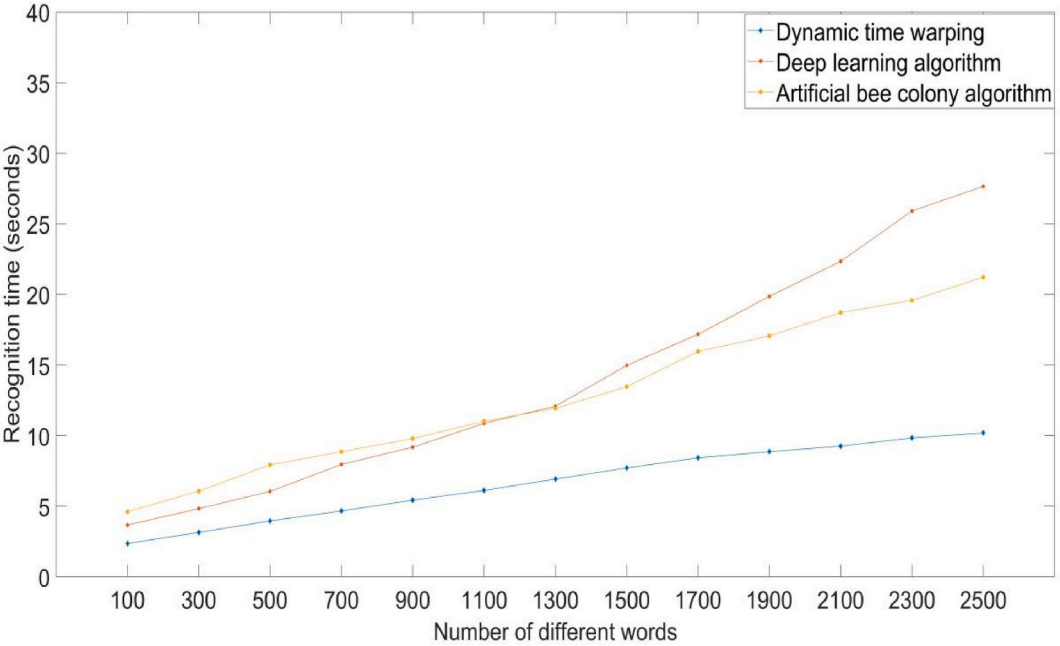


Fig. 4. Comparison of three different speech recognition methods for different numbers of vocabulary recognition times.

testers extracted from these two speech recognition algorithms were 6.3 % and 4.24 % lower than those based on DTW speech recognition, respectively. By comparing Fig. 3A and B, it can be found that in a quiet environment, the average recognition rates of the three speech recognition methods were higher than those in noisy environments. The quieter the environment, the less interference and noise it receives, so the accuracy of recognition is also higher.

The time required for speech recognition varies for different vocabulary sizes. From the speech database, different vocabulary words were randomly extracted and recognized to see the time required for DTW based speech recognition to recognize these words. In order to better demonstrate the superiority of DTW based speech recognition in recognition speed, a comparative study was conducted with speech recognition systems based on deep learning algorithms and artificial bee colony algorithms. The specific comparison outcome is illustrated in Fig. 4.

As shown in Fig. 4, DTW based speech recognition can recognize different amounts of vocabulary speech at a much faster speed than the other two types of speech recognition. Moreover, as the number of recognized vocabulary continues to increase, the growth rate of the required time gradually decreases, and the fluctuations in the required time are also smaller and more stable. When the vocabulary was 100, the time required for speech recognition based on DTW was 2.36 s, which was 1.31 s and 2.25 s less than the time required for speech recognition based on DL algorithm and ABC algorithm. When the vocabulary was 2500, the time required for DTW based speech recognition was 10.19 s, which was 17.47 s and 11.02 s less than the time required for DL algorithm and ABC algorithm based speech recognition. Moreover, based on the data in Fig. 4, it can be found that when the vocabulary was below 1100, the DL algorithm based speech recognition required less time than the ABC algorithm based speech recognition, but more time than the DTW algorithm. However, when the vocabulary was above 1300, speech recognition based on DL algorithm required more time for recognition than both ABC algorithm and DTW based speech recognition. In summary, regardless of the recognized vocabulary, speech recognition based on DTW takes less time and recognition speed is faster.

By studying the DTW algorithm, it can be found that its application in speech recognition can help speech recognition outperform speech recognition based on other algorithms in terms of recognition rate and speed. In order to further demonstrate the benefits of speech recognition on the basis of DTW algorithm, speech recognition is carried out for isolated words, continuous words, specific people, non-specific people, small vocabulary, medium vocabulary, and large vocabulary. This can reflect the recognition rate based on DTW algorithm when facing different numbers of recognition or continuous vocabulary. For small vocabulary, medium vocabulary, and large vocabulary, 50 words, 2000 words, and 6000 words were selected, respectively. For continuous word speech recognition, the tester was asked to record three consecutive sentences. Speech recognition based on DTW algorithm was used to recognize the seven different types mentioned above, and the final recognition rate was compared with the recognition rate obtained by speech recognition systems based on deep learning algorithm and artificial bee colony algorithm. The specific comparison outcome is illustrated in Fig. 5.

As shown in Fig. 5, it can be observed that the recognition rate of DTW based speech recognition for different types of speech was very high, far higher than the other two types of speech recognition. The recognition rate of speech recognition based on DTW algorithm for different types of speech was above 92.8 %, while the recognition rates of speech recognition on the basis of DL algorithm and ABC algorithm were below 91.5 % and 92.3 %, respectively. Among them, the speech recognition on the basis of DTW algorithm had the lowest recognition rate for large vocabulary speech recognition, only 92.81 %, which was 2.87 % and 0.93 % higher than the recognition rates based on DL algorithm and ABC algorithm, respectively. The speech recognition based on DTW algorithm had the highest recognition rate for small vocabulary speech recognition, with 95.87 %, which was 5 % and 3.61 % higher than the recognition

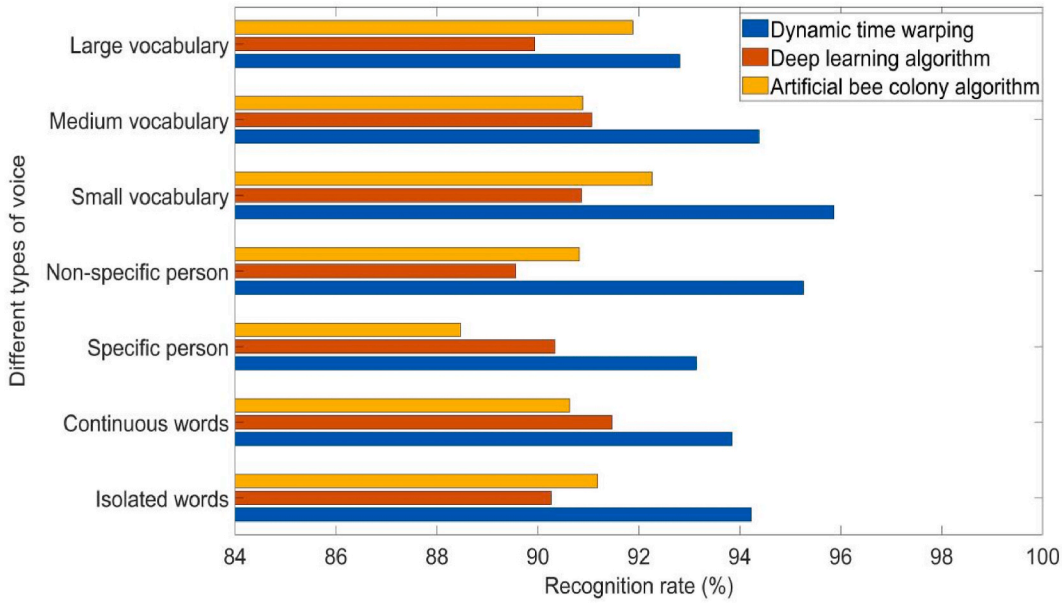


Fig. 5. Recognition rates of three different speech recognition methods for different speech types.

rates based on DL algorithm and ABC algorithm, respectively. The speech recognition based on DL algorithm had the highest recognition rate for continuous words, at 91.47 %, but it was still 2.38 % lower than the speech recognition on the basis of DTW algorithm. The speech recognition based on ABC algorithm had the lowest recognition rate for specific people, only 88.47 %, which was 4.67 % lower than the speech recognition based on DTW algorithm.

6. Conclusions

Language is the most effective and convenient way for humans to exchange ideas and transmit information. Speech recognition is based on human language as the research object. It mainly utilizes mobile terminal devices to assist intelligent tools in automatically understanding human language through speech signal processing and pattern recognition. At the same time, speech recognition can also help intelligent machines convert speech signals into appropriate text and recognize these texts. However, many speech recognition technologies currently have problems such as low recognition rate, inaccurate recognition, and slow recognition speed. Therefore, this article used the DTW algorithm to study speech recognition, effectively improving the recognition rate of speech recognition. By using the DTW algorithm, more representative speech reference templates can be obtained, which improves the accuracy of recognition. Through experiments, it can be learned that under the same conditions, the recognition rate based on DTW algorithm is much higher than that based on DL algorithm and ABC algorithm, and the recognition time is shorter. This paper mainly obtains the following results.

(1) DTW algorithm is speech recognition in quiet or noisy environment, and the speech test recognition rate is much higher than that of other speech recognition.
(2) DTW based speech recognition has a much faster recognition speed for different numbers of words and speech compared to the other two types of speech recognition. Moreover, with the increasing number of the words recognized, the growth rate of the time needed is slowly decreasing, and the ups and downs of the time needed are smaller and more stable.
(3) DTW-based speech recognition has a very high recognition rate for different types of speech recognition.

In the use of speech recognition research, the DTW optimization technique has the following benefits. Strong accuracy: The DTW algorithm can accurately align speech signals on nonlinear time scales, improving speech recognition for various speakers and speech signal adaptability to various environments. Strong adaptability to nonlinear changes: The DTW algorithm is capable of handling nonlinear changes in speech signals. It also adapts well to variations in sound speed, pitch, and other factors. However, there are significant drawbacks to using the DTW optimization technique for speech recognition. High computational complexity: The standard DTW technique requires a lot of computing power and time to calculate, especially for long speech segments. The DTW method has a high demand for large amounts of data, although its capacity requirements are lower than some statistical models. This is because the DTW algorithm is computationally demanding and requires more time and resources when dealing with large amounts of data.

Data availability

All data generated or analyzed during this study are included in this published article.

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Shaohua Jiang: Investigation. **Zheng Chen:** Investigation, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the funding of the following research projects: 1“Exploration on the Reform of College English Grammar Teaching by Educational Informationization” (No. JZ180077); 2“Corpus-assisted English Grammar Teaching Innovation” (No.2018CG02644) ; 3“An Innovative Model of Blended English Teaching by SPOC” (No. FJJKCGZ18-793) .

References

[1] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, Yoshua Bengio, Light gated recurrent units for speech recognition, IEEE Transactions on Emerging Topics in Computational Intelligence 2 (2) (2018) 92–102.

[2] Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, et al., Bigssl: exploring the frontier of large-scale semi-supervised learning for automatic speech recognition, IEEE Journal of Selected Topics in Signal Processing 16 (6) (2022) 1519–1532.
[3] Zixing Zhang, J. Geiger, J. Pohjalainen, A.E.D. Mousa, W. Jin, B. Schuller, Deep learning for environmentally robust speech recognition: an overview of recent developments, ACM Transactions on Intelligent Systems and Technology (TIST) 9 (5) (2018) 1–28.
[4] Tribhuwan Kumar, S.S. Rajest, K.O. Villalba-Condori, D. Arias-Chavez, K. Rajesh, M.K. Chakravarthi, An evaluation on speech recognition technology based on machine learning, Webology 19 (1) (2022) 646–663.
[5] Pingchuan Ma, Stavros Petridis, Maja Pantic, Visual speech recognition for multiple languages in the wild, Nat. Mach. Intell. 4 (11) (2022) 930–939.
[6] Triantafyllos Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2018) 8717–8727.
[7] Alireza Entezami, Hashem Shariatmadar, Structural health monitoring by a new hybrid feature extraction and dynamic time warping methods under ambient vibration and non-stationary signals, Measurement 134 (1) (2019) 548–568.
[8] Y. Hu, A method of DTW based gait recognition and gait data from kinect, International Journal of Computer Techniques 5 (1) (2018) 14–19.
[9] Natsuki Takayama, Hiroki Takahashi, Weakly-supervised learning for continuous sign language word recognition using DTW-based forced alignment and isolated word HMM adjustment, IIEEJ Transactions on Image Electronics and Visual Computing 7 (2) (2019) 88–96.
[10] Jianbin Zheng, Z. Li, L. Huang, Y. Gao, B. Wang, M. Peng, Y. Wang, A GMM-DTW-based locomotion mode recognition method in lower limb exoskeleton, IEEE Sensor. J. 22 (20) (2022) 19556–19566.
[11] Arul Valiyavalappil Haridas, Ramalatha Marimuthu, Vaazi Gangadharan Sivakumar, A critical review and analysis on techniques of speech recognition: the road ahead, Int. J. Knowl. Base. Intell. Eng. Syst. 22 (1) (2018) 39–57.
[12] Rajeev Ranjan, Abhishek Thakur, Analysis of feature extraction techniques for speech recognition system, Int. J. Innovative Technol. Explor. Eng. 8 (7C2) (2019) 197–200.
[13] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2018) 8717–8727.
[14] Saliha Benkerzaz, Youssef Elmir, Abdeslam Dennai, A study on automatic speech recognition, Journal of Information Technology Review 10 (3) (2019) 77–85.
[15] G. Ciaburro, G. Iannace, A. Trematerra, I. Lombardi, M. Abeti, The acoustic characteristics of the “dives in misericordia” church in rome, Build. Acoust. 28 (2) (2021) 197–206.
[16] J.X. Zhang, Z.H. Ling, L.J. Liu, Y. Jiang, L.R. Dai, Sequence-to-sequence acoustic modeling for voice conversion, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (3) (2019) 631–644.
[17] S.H.K. Parthasarathi, N. Sivakrishnan, P. Ladkat, N. Strom, Realizing petabyte scale acoustic modeling, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 9 (2) (2019) 422–432.
[18] Mohammad Saleem, Bence Kovari, Preprocessing approaches in DTW based online signature verification, Pollack Period. 15 (1) (2020) 148–157.
[19] Y. Yao, X. Zhao, Y. Wu, Y. Zhang, J. Rong, Clustering driver behavior using dynamic time warping and hidden Markov model, Journal of Intelligent Transportation Systems 25 (3) (2021) 249–262.
[20] Omer Gold, Micha Sharir, Dynamic time warping and geometric edit distance: breaking the quadratic barrier, ACM Trans. Algorithm 14 (4) (2018) 1–17.
[21] Ye Shuo, Chuntang Peng, He Juan Du Zhenzhen, Design of isolated word speech recognition system based on DTW [J], Journal of Yangtze University (Natural Science Edition) 15 (17) (2018) 33–37.
[22] Yao Fang, Design of oral English intelligent evaluation system based on DTW algorithm, Mobile Network. Appl. 27 (4) (2022) 1378–1385.
[23] Patmi Kasih, Voice recognition untuk sistem keamanan PC menggunakan metode MFCC dan DTW, Generation Journal, Department Of Informastics Engineering 2 (1) (2018) 57–68.
[24] Noor Fita Indri Prayoga, Yenni Astuti, Catur Budi Waluyo, Analisis speaker recognition menggunakan metode dynamic time warping (DTW) berbasis matlab, Aviation Electronics, Information Technology, Telecommunications, Electricals, Controls 1 (1) (2019) 77–85.
[25] Ali Bou Nassif, Speech recognition using deep neural networks: a systematic review, IEEE Access 7 (1) (2019) 19143–19165.
[26] Wang Min, Xu Juan, Yao Chenhong, Zhao Yuan, Research on optimizing WNN speech recognition based on ADSABC algorithm, Chinese Journal of Liquid Crystal & Displays 33 (7) (2018) 1–2.