

PAPER • OPEN ACCESS

Improved algorithm of DTW in speech recognition

To cite this article: HU Zhi-Qiang *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **563** 052072

View the [article online](#) for updates and enhancements.

Improved algorithm of DTW in speech recognition

HU Zhi-Qiang 1, ZHEN Jia-Qi 1, WANG Xin 1, LIU Zi-Wei 1, LIU Yong 1*

¹College of Electronic Engineering, Heilongjiang University, Harbin 150080)

Corresponding author: Liu Yong (1970-), male (Han), Heilongjiang University associate professor, tutor, research direction: the direction things were artificial intelligence,

E-mail: liuyong@hlju.edu.cn, Telpone number: 13009721364.

Abstract: This paper improves the Dynamic Time Warping (DTW) algorithm. In order to change the problem of the traditional search range of DTW speech recognition algorithm is too large, an improved DTW algorithm is proposed to limit the search path. Firstly, the traditional algorithm is analyzed to find its speech recognition search path, distortion and recognition efficiency. Secondly, the improved algorithm is introduced, the search path is limited, and simulation is carried out in MATLAB. Compare the improvement results with the improved recognition results before comparing the distortion and recognition efficiency, and the improved efficiency is calculated. Experimental results show that the improved new speech recognition algorithm is superior to the traditional algorithm in overall performance.

1. Introduction

There are many algorithms for speech recognition, and the dynamic time warping algorithm [1-4] (DTW) is a simple and efficient algorithm. Hidden Markov Model [5] (HMM) speech recognition algorithm, the basic principle is to perceive feature values through a random process. In the case of the same speech recognition environment, the DTW and HMM algorithms have similar recognition effects, but DTW is simpler than HMM. Because a large amount of voice data needs to be provided for HMM in training and double counting, DTW requires less data. Therefore, DTW is used more widely, test templates match all reference templates at once, then propose the closest template as the recognition result.

Real-time speech recognition is the object of research, and isolated words are the main object [2,6]. Better extraction of feature values is the basis for rapid identification. At present, the DTW algorithm can be used as the most proficient and simple speech recognition algorithm. The system cost is low and the recognition speed is fast. Therefore, it is a very effective recognition algorithm in a small number of word speech recognition control systems. However, the limitation of the traditional algorithm search path has a great influence on the recognition efficiency. This paper improves the recognition path and simulates its validity.

2. DTW speech recognition system works

The working principle of the speech recognition system is shown in Figure 1.



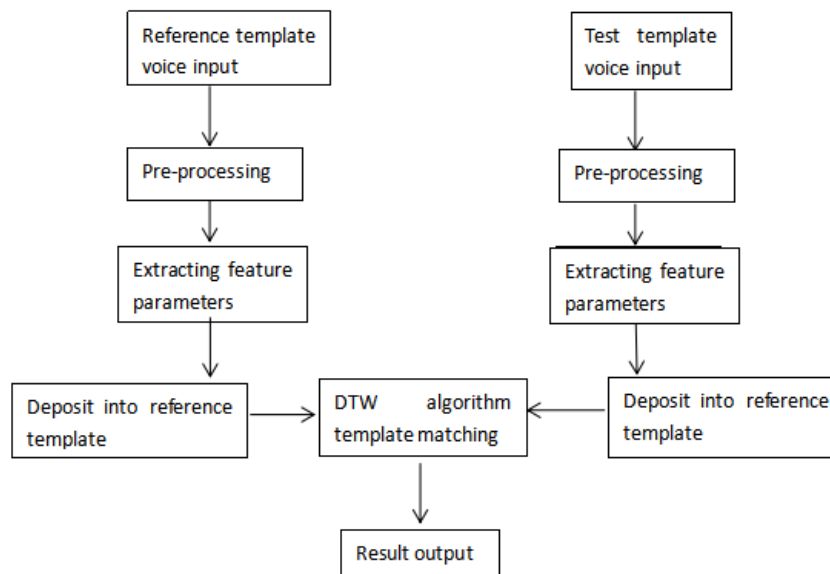


Figure 1 Schematic diagram of the speech recognition system.

The first is the establishment of a reference template, that is, the reference speech signal is input to the pre-processing unit, leaving a stationary signal after filtering out the redundant information. The left signal is extracted from the feature vector sequence [7-8] and stored in the template library. Secondly, the speech signal to be tested is similarly subjected to the above steps, and the feature vector sequence is also obtained, which is present in the test template. The DTW recognition algorithm compares the template to be tested with the reference template, and extracts the template of the maximum acquaintance [9] as the recognition result.

3. DTW recognition principle

The first is the reference template, which is the sequence of features that have been saved to the feature template library. We represent it as $R=\{R(1), R(2), R(3), \dots, R(m), \dots, R(M)\}$. m is a time series label, $m=1$ is the initial recognition frame, $m=M$ is the final identification frame, M refers to the total number of speech frames in the reference template, and $R(m)$ is the feature vector of the m frame.

Next is the test template, which is the sequence of speech features to be recognized. We represent it as $T=\{T(1), T(2), T(3), \dots, T(n), \dots, T(N)\}$. n is the timing label of the test speech frame, N is the total number of speech frames in the test template, and $T(n)$ is the n th frame feature vector.

The reference template and the test template use the same feature vector, frame length, and frame shift and frame window. When template matching, the most important thing is to calculate the degree of acquaintance between the two templates. The higher the degree of acquaintance, the more the two match, so we use the distortion to express the matching relationship between the two. The degree of distortion of the degree of acquaintance is inversely proportional. The lower the distortion, the higher the degree of acquaintance.

The frame number $n=1\sim N$ of the speech template to be tested is represented on the horizontal axis N axis in the two-dimensional Cartesian coordinate system, and each frame number $m=1\sim M$ of the reference template is in the two-dimensional Cartesian coordinate system. The vertical axis is represented on the M -axis, and a network lattice matching the path is obtained. The intersections in the grid represent the intersection of a frame in the test. The reference speech signal template is used to test the frame number calculated by the speech signal template, that is, the point at which the matching route has passed. Paths can't be chosen at will, because the speed of different voices may change at

any time. However, the order of each part does not change, so the route selected at the time of matching must be from the lower left corner to the upper right corner, and the sum of the frame distortions of all the intersections on the route is minimized, as shown in picture 2:

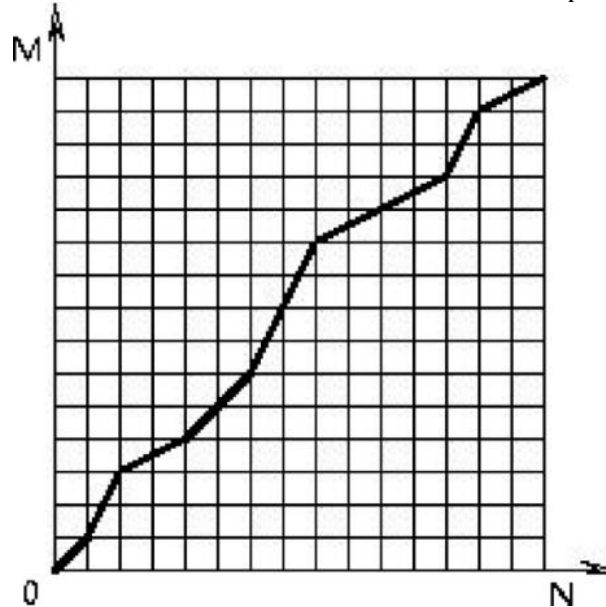


Figure 2 DTW algorithm search path

To plot the matching path, we set the path through which the path passes $(n_1, m_1), (n_i, m_i), \dots, (n_N, m_N)$, where $(n_1, m_1) = (1, 1), (n_N, m_N) = (N, M)$. Path search condition: 1, the slope of the restricted route must be between $1/2$ and 2 , so as to avoid the line transition tilt; 2, assuming that the route has passed the point (n_{i-1}, m_{i-1}) , then the point that will pass There are only three cases (n_i, m_i) : $(n_i, m_i) = (n_{i-1}+1, m_{i-1}+2), (n_i, m_i) = (n_{i-1}+1, m_{i-1}+1), (n_i, m_i) = (n_{i-1}+1, m_{i-1})$. Use β to represent constraints. The problem of finding the shortest path can be translated into finding the best path function β that satisfies the constraints. $n_{i-1} = n_i - 1, m_{i-1}$ can be determined by 2-1:

$$\begin{aligned} D[(n_i-1, m_i-1)] &= \min \{ D[(n_i-1, m_i)], \\ &D[(n_i-1, m_i-1)], D[(n_i-1, m_i-2)] \} \end{aligned} \quad (2-1)$$

4. Dynamic time algorithm improvement

The traditional dynamic time algorithm still has a loose limit on the matching process path, which causes points in many areas to be not counted in the matching process at all. In order to reduce the amount of computation in the path matching process, the matching path is now limited. Figure 3:

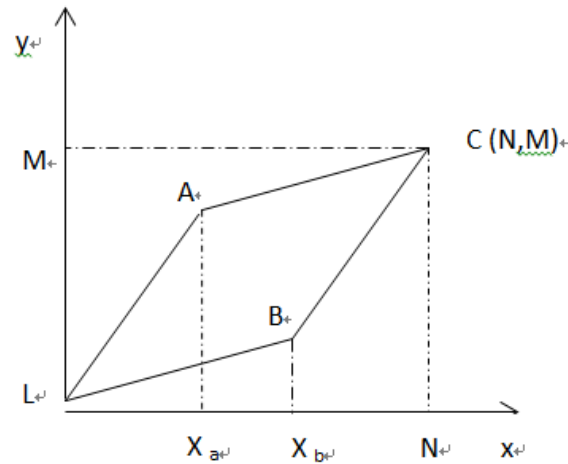


Figure 3 Improved path search area

The slope of AL and CB is $3/2$, and the slope of BL and CA is $1/2$. Areas outside the diamond do not need to be calculated. As can be seen from the figure, the actual dynamics can be divided into three parts (1, X_a), (X_a+1 , X_b) and (X_b+1 , N), where 2-2:

$$\begin{cases} X_a = M - N / 2 \\ X_b = 3N / 2 - M \end{cases} \quad (2-2)$$

At that time, the actual dynamics could be divided into two parts (1, X_a), (X_a+1 , N); at that time, the actual dynamics could be divided into three parts (1, X_b), (X_b+1 , X_a) and (X_a+1 , N).

5. Analysis of MATLAB simulation results

We use 0-9 ten digital voice files as the voice database. To further prove the correctness of the algorithm, we use a voice file together with the "two" voice file in the speech database to be identified and the "two" file in the template voice database. Take pre- and post-improvement simulation recognition results, including the recognition of distortion and the time to identify ten digits. Table 1 shows the distortion comparison before and after the improvement. Table 2 shows the results of the recognition of ten voices before and after the improvement and the time required.

Table 1 Comparison of distortion before and after improvement

Audio file \ Distortion factor	NO.0	NO.1	NO.2	NO.3	NO.4	NO.5	NO.6	NO.7	NO.8	NO.9
Before improvement	2.1869	0	1.6127	2.0934	1.7092	1.6418	1.8664	1.4171	1.1444	1.8260
After improvement	1.8442	0	1.3765	1.9874	1.3420	1.2661	1.6660	1.3402	1.1405	1.8147

Table 2 Identification results before and after improvement and identification time-consuming

Audio file Distortion factor	NO.0	NO.1	NO.2	NO.3	NO.4	NO.5	NO.6	NO.7	NO.8	NO.9	Time- consuming
Identify results after improvement	0	1	2	3	4	5	6	7	8	9	17.5s
Identify results after improvement	0	1	2	3	4	5	6	7	8	9	15.0s

In the case of correct identification in MATLAB[10] simulation. The "2" speech in Table 1 is improved after the improvement and the distortion is 0, which proves the stability of the speech feature vector extraction in the recognition system. Comparing the distortions before and after the improvement of Table 1, it can be found that the improved recognition distortion of the ten test speech files is reduced to a different extent than the distortion before the improvement (except for the "2" file).

In Table 2, in the case of identifying ten test speeches and the recognition result is complete, the pre-improvement time is 17.19 seconds, and after the improvement, correctly identifying ten test speech files takes about 15.06 seconds, saving about 2.13 seconds. It is equivalent to improving the recognition efficiency by about 12.39%.

6. Conclusion

It can be seen that after the traditional speech recognition algorithm is improved, not only the distortion is reduced, but also the recognition time is reduced. Improve the efficiency of speech recognition.

References:

- [1] A reference
Chen Liwan. Research on Improvement Technology of DTW Algorithm Based on Speech Recognition System[J]. Microcomputer, 2006(05): 267-269.
- [2] Another reference
DTW-based speech recognition system for isolated words [D]. Liao Zhendong. Yunnan University 2015.
- [3] More references
Zhu Shuqin, Zhao Wei. Research and Analysis of DTW Speech Recognition Algorithm[J]. Microcomputer Information, 2012, 28(05): 150-151+163.
- [4] Hou Ruizhen. Online speech recognition and simulation based on DTW algorithm [D]. North China University of Technology, 2014.
- [5] Jiang Yubo. Research on speech recognition technology based on HMM and ANN hybrid model [D]. University of Electronic Science and Technology, 2016.
- [6] Ye Shuo, Peng Chuntang, Du Zhenzhen, He Juan. Design of Isolated Word Speech Recognition System Based on DTW[J]. Journal of Yangtze University (Self Edition), 2018, 15 (17): 33-37 + 5.
- [7] Yang Dali, Xu Mingxing, Wu Wenhui. Study on the Method of Selecting Speech Recognition Feature Parameters[J]. Journal of Computer Research and Development, 2003(07): 963-969.
- [8] DHINGRA S D, NIJHAWAN G, PANDIT P. Isolated speech recognition using MFCC And DTW[J]. International Journal of Advanced Research in Electrical Electronics & Instrumentation Engineering, 2013, 2(8): 4085-4092.

- [9] Shu Qi. Research on small-vocabulary isolated speech recognition method [D]. Wuhan University of Technology, 2012.
- [10] He Qiang, He Ying. Matlab extended programming [M]. Beijing: Tsinghua University Press, 2002: 330 -349.