Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 832 – 836

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

# Speech Recognition Based on Efficient DTW Algorithm and Its DSP Implementation

Jing XinXing [a], Shi Xu[a]*,

[a]*GuiLin university of electronic technology, GuangXi GuiLin, PostCode:541000,China*

**Abstract**

The present voice recognition has achieved high recognition accuracy in theory and the laboratory condition, but many speech recognition algorithms are realize on PC. A system platform with DSP core can realize real-time speech processing algorithms, and in cost, power consumption and volume has the advantages of PC did not own, and has a good application prospect. This paper emphasis on how to realize speech recognition in DSP and its whole hardware and software technological process, and analyzed the program with CCS's self-contained process analysis tools Profile. In order to improve the recognition rate, a high performance DTW algorithm which differ from the traditional DTW algorithm is used. So real-time control is achieved and obtain a high recognition rate.



*Keywords:* speech recognition, high performance DTW, DSP

## 1. Introduction

Speech recognition is an important part of pattern recognition, owns a very considerable development prospect and practical value. Since the voice signal is very random, even for the same person in the same way but in different time pronounced the same word, they can't be exactly same,because the pronunciation's sustained time is random, so directly compare each word's feature vector sequence's effect is not ideal. We must adopt the Dynamic Time Warping ( DTW ) algorithm [1]. DTW is a classical algorithm, is easy to perform and has no strict requirements on hardware resources, so it widely be used

* Corresponding author. Tel.: +8615296803696; .
*E-mail address*: shixu009@163.com.


doi:10.1016/j.proeng.2012.01.050

in speech recognition.But the DTW algorithm uses the point matching method to computate the matching distance, the matching distances between test voice and each reference speech template must be computated, when the reference template and test the voice volume increase, recognition time will increase significantly. So how to improve the recognition speed without reducing the recognition rate is the key technology of DTW recognition system.

## 2. Speech recognition system

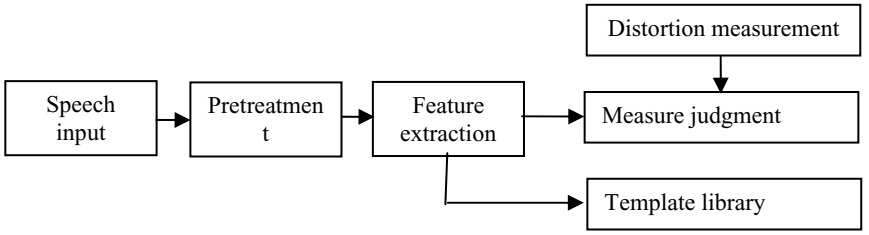


Figure 1. speech recognition block diagram

Figure 1 is the block diagram of the speech recognition system, from this figure we can see that speech recognition system is essentially a pattern recognition system, it includes a feature extraction, pattern matching, template library and so on. Because of the voice's nonstationarity and the presence of environmental noise, current interference and other factors, we cannot directly do feature extraction and must do the voice signal's pretreatment. The speech signal pre-processing includes filtering, sampling, quantization, endpoint detection and so on, the speech signal can be used to extract characteristic parameters after pretreatment. The characteristics parameters of the speech signal are a lot, such as short time energy, short-term zero-crossing rate, short-time autocorrelation coefficient and so on. Two commonly used feature parameter is LPCC ( linear predictive cepstrum coefficient ) and MFCC ( Mel frequency cepstrum coefficient ). LPCC has the advantage of extracting the speech parameters accurately, and the calculation speed is relatively fast, easy to implement in hardware. But LPCC in anti-noise performance, robustness and the recognition rate and other aspects are relatively low, so in practical applications ,MFCC which owns high noise immunity and the robustness is generally used. In the training stage, each input voice is used to extract parameters and the parameters are coexisted as the reference template, all the reference templates form template library. In the recognition stage, recognition speech's feature parameters are obtained in the same way and used to match with reference template library using the DTW algorithm , the maximum similarity reference template in the template library will be as a result of recognition output. DTW is a very successful recognition matching algorithm.

## 3. DTW algorithm

### *3.1 Triditional DTW algorithm*

Suppose the reference templates is extracted from a speech signal contained M frames, represented as {R ( 1), R ( 2), ... R ( m ), ... , R ( M ) }, where R ( I ) ( I = 1, 2, ... M ) is speech signal's characteristics vector's I frame, the test template is extracted from a speech signal contained N frames,,represented as {T ( 1), T ( 2), ... T ( n ), ... , T ( N ) }, where T ( I ) ( I = 1, 2, ... N ) is speech signal's characteristics vector's

I frame. In order to compare the similarity between them, can calculate the distance between them D [T, R], the smaller the distance is greater the similarity. In order to calculate the distortion distance,must caculate each corresponding frame's distance from T and R . Let n and m were arbitrary frame number of T and R , d [ T ( n ), R ( m ) ] is the two feature vector's distance . Distance function depends on the actual distance metric, the DTW algorithm commonly used Euclidean distance. If N = M then it can be calculated directly, otherwise must consider to align T ( n ) and R ( m ) using dynamic programming ( DP ) method.

The test template's each frame number n= 1-N is marked on the horizontal axis of a two-dimensional Cartesian coordinate system,and the reference template's each frame number m = 1-M is marked on vertical axis , by this integer number which stand for the frame number can draw some vertical and horizontal lines and form a network, the network's each cross point ( n, m ) stand for the intersection of T and R,and the distance between the two frames must be caculated.The DP algorithm can be come down to find a path through this network's several cross point. The path was not chosen at random,because any kind of voice pronunciation speed may have changed, but the various parts of the sequence is impossible to change, therefore the path must be start from the lower left angle, end in the upper right corner . In order to make the path not unduly skewed, can restrain the slope at ( 0.5, 2) range, as shown in figure 2.

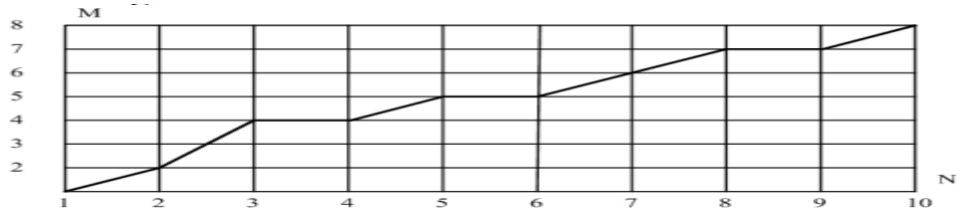


Figure 2. Traditional DTW algorithm's search path

*3.2High performance DTW*

Because the matching process defines the bending slope at range of ( 0.5, 2) ,so a lot of lattice can actually not arrive, the search path is limited in the diamond-shaped area as shown in Figure 4, outside the rhombic's lattice do not needed to calculate the corresponding distance. from the figure can see that the

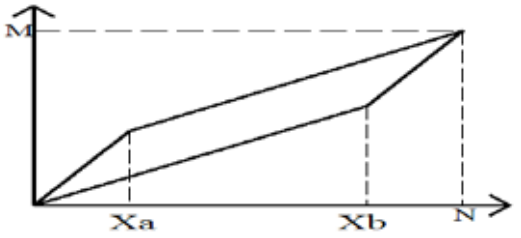


Fig.3 Efficient DTW algorithm's search path ;

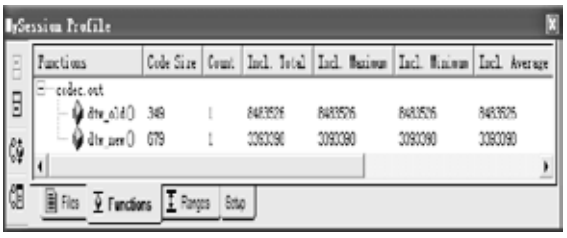


Fig.4 The analysis results between two algorithm

actual dynamic warping is splited into 3 sections, ( 1, Xa ), ( Xa+1, Xb ) and ( Xb+1, N ), where Xa = 1 / 3 * ( 2M-N ), Xb = 1 / 3 * ( 2N-M ),from which can obtain a speech frame limit: 2M-N > = 3, 2N-M > = 2. When the above conditions do not meet, think that the difference between the two voice is too big, cannot undertake dynamic bending matching. Make full use of this feature can reduce the amount of

computation. As shown in Figure 5, using CCS's self_contained program analysis tools Profile to analyze the traditional DTW algorithm ( dtw_old function ) and highly efficient DTW algorithm ( dtw_new function ), can see that a DTW operation of traditional algorithm takes 8483526 clock cycles, and efficient algorithm only needs 3393398 clock cycles, operation quantity reduced by one half.

## 4. system hardware circuit design

The system hardware circuit is shown in Figure 3, the system is mainly composed of a core chip DSP TMS320VC5402, speech collection and output module of TLV320AIC23, extended program memory FLASH, expanded data memory SRAM, decoder GAL, and JTAG download interface and a power supply module. The core module of the system is the DSP chip 5402, the chip is TI's 16 bit microprocessors with fast processing speed, when running at full speed can perform 100000000 single cycle instruction per second ( 100MIPS ), speech recognition algorithms, such as pre-emphasis, endpoint detection, feature extraction, DTW algorithm is realized by it.

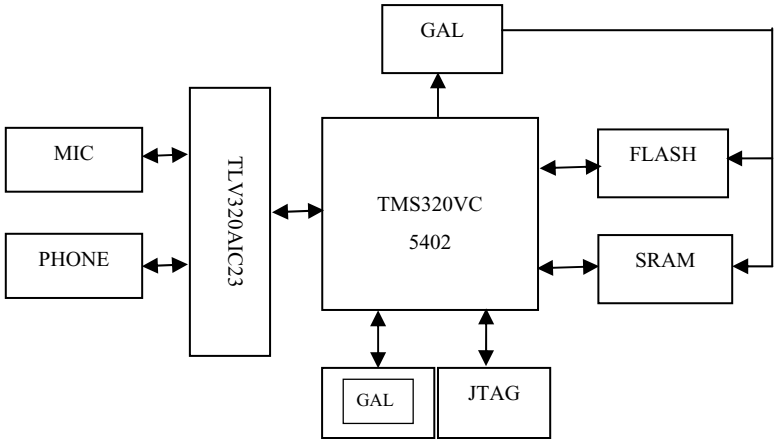


Fig.5 System block diagram of hardware system

## 5. DSP software programming and system application

Software design include speech acquisition program, pretreatment, endpoint detection procedures, MFCC feature extraction procedures, procedures for DTW algorithm. The software includes two phases:training and recognition.

In the training stage, the serial port controls AIC23 to do voice signal collection, because DTW algorithm is very sensitive for speech endpoint, so endpoint detection must be done after the speech has been collected. In order to enhance the high frequency component and restrain the power frequency interference,the pre-emphasis must be done after endpoint detection,put the speech signal into a one order low pass filter. After preprocessing the speech signal can be used to extract the characteristic parameters of MFCC, the speech signal is divided into frames, each frame contains 256 points,frame shift is 128 points. Extracte characteristic parameters for each frame, extracted feature parameters are stored in the off-chip's non-volatile FLASH as template library.

In the training stage, the serial port controls AIC23 to do voice signal collection, because DTW algorithm is very sensitive for speech endpoint, so endpoint detection must be done after the speech has been collected. In order to enhance the high frequency component and restrain the power frequency interference,the pre-emphasis must be done after endpoint detection,put the speech signal into a one order low pass filter. After preprocessing the speech signal can be used to extract the characteristic parameters of MFCC, the speech signal is divided into frames, each frame contains 256 points,frame shift is 128 points. Extracte characteristic parameters for each frame, extracted feature parameters are stored in the off-chip's non-volatile FLASH as template library.

In the recognition stage, the same was used to collect speech, do pretreatment and frame parameter extraction. Take the test speech's feature parameters as test template and calculated matching distance with the reference template in FLASH by DTW algorithm, the minimum distance's speech is taken as the output results.

## 6. Conclusion

The experiment finally used the CCS's built-in program analysis tools Profile to analyse the entire program, as shown in Figure 5, it can be seen from the figure that in a recognition program ,do endpoint detection 1 times, feature extraction 22 times ( 22 voice signal frames ), DTW algorithm 5 times ( for a total of 5 reference template ). The program runs industry takes 39350562 clock cycles, through the software configuration make DSP running on the 100M clock condition, that 39350562 clock cycle took 393ms, it fully achieve the real time control.

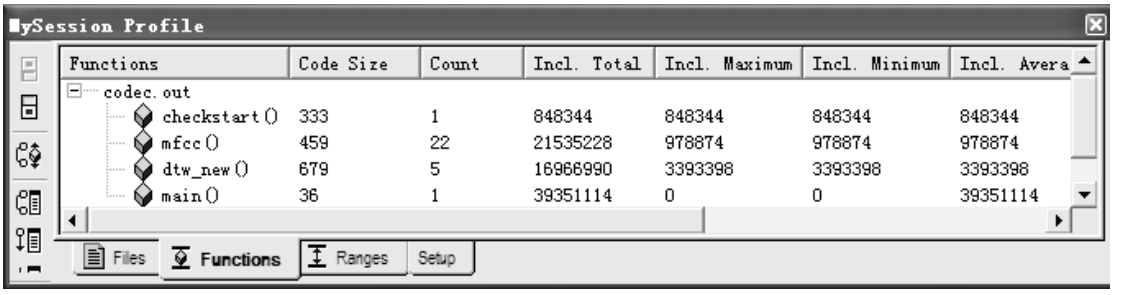


Fig. 6 functions operation cycle diagram obtained by Profile

## References

[1]Zhang ling, Zhang Min. Speech Recognition System Based Improved DTW Algorithm， 2010 Interational Conference on Computer, Mechatronics, Control and Electronic Engineering. 2010;3(5): 468-472。

[2] Ma B,Guo LL,speech recongation rearch and design base on DSP, mrocomputer information,2008;24(8-2)

[3] Wang K M,improved recongation method for online sign based on DTW,technology paper,2010;28(13):101-103