Journal of Electrical and Electronic Engineering
2015; 2(6): 89-93
Published online February 6, 2015 (http://www.sciencepublishinggroup.com/j/jeee)
doi: 10.11648/j.jeee.20140206.12
ISSN: 2329-1613 (Print); ISSN: 2329-1605 (Online)

Embedded system for speech recognition and image processing

Zhengxi Wei, Jinming Liang

School of Computer Science, Sichuan University of Science & Engineering, Zigong Sichuan 643000, PR China

Email address:
413789256@qq.com (Zhengxi Wei), ljm@suse.edu.cn (Jinming Liang)

To cite this article:
Zhengxi Wei, Jinming Liang. Embedded System for Speech Recognition and Image Processing. *Journal of Electrical and Electronic Engineering*. Vol. 2, No. 6, 2014, pp. 89-93. doi: 10.11648/j.jeee.20140206.12

Abstract: In recent years, the products of voice terminal and image retrieval show the intelligentized trend, but the mature commodities are rare in the market. This paper presents an embedded design method of intelligent voice terminal based on pattern recognition. The design adopts Samsung S3C2410 ARM as target board, Philips Uda1341TS as audio codec, embedded Linux OS as software platform, and speech recognition is implemented through small-vocabulary voice training. To improve the recognized effect, we use the image retrieval technology as an auxiliary tool, which helps speech recognition module create or more accurately find a personal voice-training library. By means of image recognition, the experimental results prove that the effect of speech recognition achieves an average increase of 10 percentages.

Keywords: Speech Recognition, Embedded Development, Image Retrieval, DTW Algorithm, ARM Development

1. Introduction

In recent years, multimedia-terminal [1] products to meet people's personalized requirements show the trend of intelligent, such as support for the conversion of text and voice, speech recognition on the basis of voice communication, and the recognition of digital image coming from the camera photography. These intelligent functions are implemented under the new development environment and background based on the theory of pattern recognition, VLSI hardware platform, and customized software operation system.

Operating system platform of intelligent terminal usually provides API functions, Software Development Kit, as well as integrated graphical development and testing tools. The developer not needing to understand the complex underlying hardware structure can develop the new business by its existing experience in this domain. As a result, the technical threshold of business development is greatly reduced.

At present, multimedia retrieval [2] is one of hot technologies in small intelligent terminal. It can be divided into three categories: (1) single cross-media integration index. In this method, when a medium is able to reflect the multimedia scene very well, the mark on this medium is also used to other media in the similar scene. (2) Multi-media integration. In this form, no one medium can well reflect the contents of the multimedia scene, we have to choose two or more medium based on multimedia content and give the judgments. As a result, these judgments are integrated together to form an interpretation to a multimedia scene. (3) dual-media feature integration. Various media feature may be combined together in accordance with the multimedia timing relationships, and we can use the feature combination to analyze the various multimedia scene.

Based on the research and the analysis of speech recognition [3] and image retrieval, this paper presents an overall design method on intelligent voice terminal. Our design mainly adopts Samsung S3C2410 ARM as the core in target development board, Philips Uda1341TS as audio codecs, embedded Linux OS as software platform, and speech recognition is implemented based on small-vocabulary voice training. To improve the effect of speech recognition, we use image recognition and retrieval technology as an auxiliary tool, which helps speech recognition module create or more accurately find out a personal voice-training library. Through capturing users' image information with a camera, we can design and achieve the intelligent voice terminal with image retrieval function.

2. Speech Recognition

The speech recognition is an advanced technology to make the machine convert the voice signal into the corresponding text or command through the identification and understanding

of the process. Speech recognition has been widely used in the field of scientific research, and as to the daily life, it is more broad space for development.

2.1. Basic Principle

The speech recognition process can be attributed to pattern recognition and matching. Speech features can be extracted from the original speech signal, which should have been pre-processed and analysis-calculation, and finally speech recognition template is constructed. During the voice recognition, voice template stored in the system is to be compared to the characteristics of the input voice signal, according to certain algorithms and strategies, to identify the optimal template for matching the inputting voice, and finally to output recognition results.

In short, the speech recognition process [4] generally involves the following several key modules: signal pre-processing, speech feature extraction, matching training-library template, and outputting the matching results, as shown in figure 1.

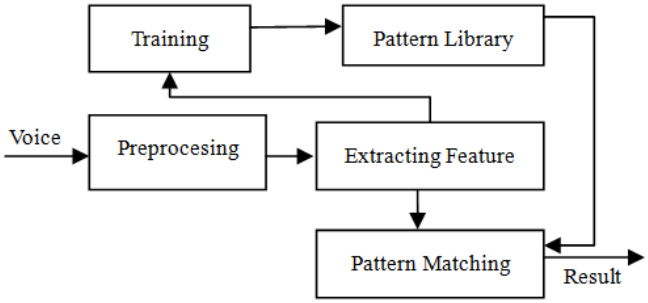


Figure 1. *Speech recognition model*

Signal pre-processing module includes sampling voice signal, removing differences of individual noise, excluding noise impact caused by the equipment and the environment, and involves the selection and endpoint detection of speech recognition unit. Speech feature-extraction module is used to extract the acoustic parameters that reflect the essential characteristics of voice, such as voice frequency, amplitude, and so on. The matching module is the core of the entire speech recognition system. It calculates the similarity (e.g. the voice speed and the likelihood probability) between the input characteristics and inventory models according to certain criteria such as word formation rules, grammar rules, semantic rules, and determines the semantic information of the inputting voice. Whereas outputting results module returns the final recognition results for its caller.

2.2. Speech Recognition Algorithm

Speech recognition algorithm is vital to the recognition effect. Good algorithms enable the signal processing to bring stable and excellent performance in the practice. DTW (Dynamic Time Warping) algorithm [5] applies the time sequence of the voice feature vector compare to every template in the reference template library. The most similar template will be acted as the recognition result. Embedded Linux OS can run in embedded ARM board and Linux-based applications can get the external voice signal. DTW speech recognition program running the embedded platform determines the syntax and semantics of the reading content. The underlying hardware of ARM board is controlled by the Linux start-up and by device drivers to complete.

DTW algorithm is based on the dynamic planning idea, departs a complex global optimization problem into many local optimization problems to deal with, and tries to automatically find a path between two feature vectors whose total distortion amount can be minimum as possible, thereby avoiding introducing time-length error. Its mathematical principle is shown in figure 2.

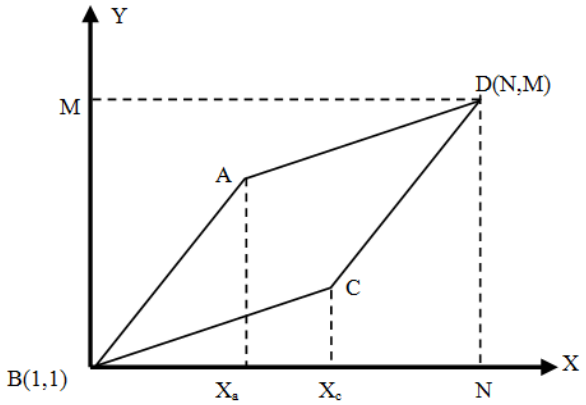


Figure 2. *DTW Matching-path constraint graph*

Assume there are M frames vector in the reference template and N frames vector in the test voice template, the dynamic-time-plan idea is to create a time revised function, as follows (1).

$$m = \omega(n) \tag{1}$$

The timeline will have test-vector n nonlinearly mapped to the reference-template-timeline m, making the cumulative-distance amount minimal between the test vector and the template vector of each frame, as the same time, the distance of matching-path between two vectors minimum, thus ensuring between the test template and reference templates with the maximum acoustic similar characteristics. Typically, the revised equation (1) is restricted to a parallelogram (assume ABCD) within the grid coordinates of the starting point is (1,1), the end point coordinates (N, M). The adjacent sides slope are 2 and 1/2. That means, simply needing calculate the matching distance of every frame that corresponds to various points in the parallelogram ABCD.

3. Image Processing

Image recognition theory contributes to image processing including image retrieval technology. It is the theoretical basis of the latter.

3.1. Image Recognition

An image recognition system [6] can be divided into three

main parts: (1) image preprocessing; (2) image segmentation and extraction feature; (3) the judgment or classification. The block diagram is seen in figure 3.

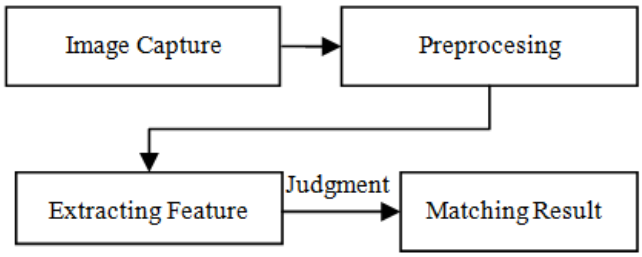


Figure 3. *Image recognition model*

Any kind of image recognition method first through a variety of sensors converts a variety of physical variables to values or set of symbols that the computer can receive. Traditionally, the space of this value or symbol is called the pattern space. In order to extract effective identified information from these numbers or symbols, it must be the following processing, removing noise, excluding irrelevant signals, calculating feature (such as the shape of the object, perimeter, area, etc.) as well as the necessary transformation (such as Fourier transformation).

Then by feature selection and extraction, the pattern feature-space is established. The subsequent pattern classification or pattern matching is based on the feature space. Finally, the system will output the object type or a model number that means this object in model database is the most correlative to the object to be searched.

3.2. Image Retrieval

Images can be manually labeled character information based on contents in the text annotations. Web Crawler can collect pictures from the web environment or extract some image marked similar text information in an HTML page, and establish originally keywords. Then it performs a pre-processes to these image, which includes de-nosing, setting standard size, and so on. The image is stored to the memory in development board and its feature-index will be further perfected after processing of relevant algorithms. Moreover, this index can later be retrieved and compared to the search keywords. In this way, it can determine whether they are the retrieved objects.

The image information will be abstracted to a generic string main through caliphemir algorithm library. Such as color histogram and other information can be extracted through the adjustment of parameters. Open source tools package (Java caliphemir) extracts the features such as color histogram and layout, convert them to the corresponding string from the image. The correspondence between extracted strings and images is established through the inverted algorithm used to file-search, co-exist in the index file. Different picture information can be stored in different fields, together constitute one document for the query. The feature string of image acts as a search keyword, and a picture of the maximum likelihood is found by querying the index file. Finally, a group of image in the picture library are found and extracted from their path information so that these retrieved pictures could be displayed on terminal screen to users by the web-explorer.

4. Embedded Development Platform

4.1. Hardware Platform

The development board with Samsung S3C2410 microprocessor is selected as the hardware platform. CPU frequency is up to 203MHz in the board. Start-up codes, OS kernel and users' application programs are together stored in a FLASH whose capacity is 64MB. Application programs run in 64MB SDRAM, which can also be used as the room of various data and the stack. A camera capturing videos is connected to a USB interface in the board. The captured video is processed according to the image-matching rules. Subsequently, the result will be transmitted to speech recognition module. ARM Board of the system is shown in figure 4.

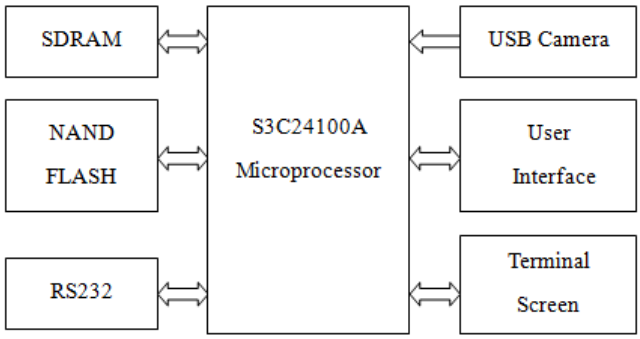


Figure 4. *ARM board*

The system uses UDA1341TS audio codec made by PHILIPS Company, which can achieve mutual conversion between stereo analog signal and digital signal. For digital signal, the chip also provides a DSP (digital signal processing) function. Input and output are composed of a microphone, a speaker and LCD (liquid crystal display). Voice analog signal inputting from microphone is first preprocessed, including A/D converter, AGC etc. A/D sampling frequency is set at 6 kHz, which is right the sampling frequency of voice signal, and Flash chip is used as the storage.

4.2. Embedded Operating System

The embedded Linux2.6.12 is a kind of miniature operating system, which is designed for the demand of the embedded OS. It has some advantages, such as small code amount, fast running speed, strong stability, and so on. This OS cuts out the normal Linux and becomes much smaller in size. It can even be solidified in a memory chip with a few KB or MB. The kernel of Linux2.6.12 can be customized by development engineers in terms of the actual demand. So it is regarded as the ideal software platform to develop embedded application programs. Speech recognition application adopts the DTW algorithm to implement speech processing and

matching.

The implementation details are as follows. DTW algorithm is firstly to compare time series of speech feature vector with the reference template library to find the highest similarity template as the recognition result. Its core code based on DTW in the embedded Linux platform is encoded by C language as follows. In the program, f denotes the amount of test template, dist[i] represents the shortest distance to a feature vector.

```
for(f=0;f<2;f++){ //read parameters of test template
  wave_read(t,filename_test[f],f);
  for(d=1;d<=M;d++){
    // read parameters of reference template
  wave_read(r,filename_ref[d],d);
    for (i=0;i<n;i++) {
      for (j=0;j<n;j++) {
        float sum1=0•0;
          for(g=0;g<M;g++)
            sum1+=sqrt( t[i][g]- r[j][g]));
      D1[i][j]=sum1; //machting distance matrix D1[n][n] } }
      D2[0][0]=D1[0][0];
    for(i=1;i<n;i++)
      for(j=0;j<n;j++){
        D3=D2[i-1][j];
        if(j>0) D4=D2[i-1][j-1];
        else D4=REALMAX;
        if(j>1) D5=D2[i-1][j-2];
        elseD5=REALMAX;
        D2[i][j]=D1[i][j]+zmin(D3,D4,D5);
      }// zmin( ) calculate the minimum among D3,D4,D5
    strdata[d].data=D2[i-1][j-1];
    aa=strdata[1].data; g=0;
    for(k=1;k<=10;k++){
    if(aa>strdata[k].data){
    aa=strdata[k].data; g=k; }
    }strdata[g].data=aa;
    printf(" %s and %s is min and value""dist[%d] is %e\n\n",
          filename_tdata[g-1].chstr[g-1], g, strdata[g].data);}
}
```

In summary, the development platform includes the target board with the S3C24100A microprocessor and the embedded Linux2.6.12. The former constitutes hardware system architecture and development environment. In addition, the latter, as embedded OS provides powerful support for the development of search retrieval software.

4.3. Image Retrieval Module

Image Retrieval module is main to create image index and search users' picture. The image information will be abstracted to a generic string main through Caliph & Emir algorithm [7] library. Such as color histogram and other information can be extracted through the adjustment of parameters. Open source tools package (Java Caliphemir) extracts the features such as color and layout, converts them to the corresponding string from the image.

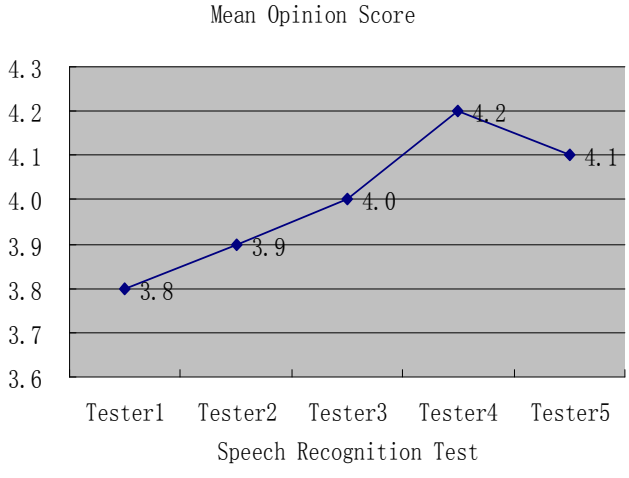
The correspondence between extracted strings and images is established through the inverted algorithm used to file-search, co-exist in the index file. Different picture information can be written in different fields, together constitutes one document for the query. The feature string of images acts as a search keyword, and a picture of the maximum likelihood is found by querying the index file. The most similar image in the picture library is obtained from its path information.

Finally, Image Retrieval module will establish a connection with the speech library. The information of matching-image or a new-user will be sent to the speech recognition module in order to accurately find or create a personal voice-training library, and complete speech recognition.

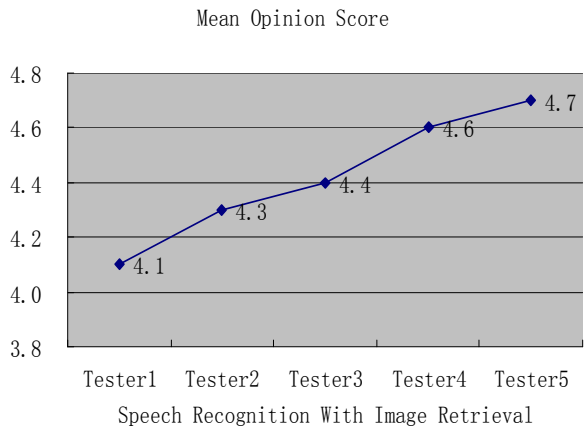
5. Experimental Results

We have firstly used a simple prototype system with sine-lifted cepstral-coefficients. With context-independent templates and no optimization, an error rate of about 2% is obtained. This setup is used in the comparative experiments on database size, and template selection. It should be noted that the error rate is readily reduced to below 1% when using context-dependent templates.

After that, we have more completed tests to our hardware and software system is as following. Five testing persons whose ages and sex are different select several sentences randomly from a 100-sentence library. Everyone says a few words in turn by a microphone. After voice signals inputting from the microphone are preprocessed, converted and encoded, recognition results are obtained by the DTW algorithm and templates, and voice signals will be output from the speaker. Listeners hearing the output voice give a score and the scoring criteria are as follows: 1-worst, 2-bad, 3-general, 4-good, and 5-excellent. The score reference is the pronunciation of an average person, and its value is set as 5. Mean opinion score (MOS [8]) of subjective perception experiment is 4.0, shown in figure 5(a).



(a)



(b)

Figure 5. *MOS result*

For further testing the auxiliary effect of image recognition, the voice terminal under the same conditions loads the image recognition module in the second test. Camera captures different people's face image to the development board, for each one to establish image index, so that the next step could be automatically identify the speaking person. Subsequently, application programs are able to create or change the voice-training file of that user and support for the speech recognition. The experimental results are shown in Figure 3(b). From the experimental results, Mean Opinion Score is 4.4. The effect of speech recognition achieves an average increase of 10 percentages when the intelligent terminal is assisted through image retrieval module.

6. Conclusions

Combined with image-recognition technology, the paper presents an embedded-design method about developing speech recognition and image retrieval for small intelligent terminal. The experimental results show that the speech recognition accuracy is improved 10%, and the design idea for developing intelligent voice terminals can be referenced.

After a few modifications, the intelligent terminal can be used for much equipment, such as used in mobile phones, automatic answering machines and other portable devices; can also be used for the intelligent building systems, instrumentation-control, smart toys and other occasions, with good application prospects.

Acknowledgements

The research was supported by Artificial Intelligence Key Laboratory of Sichuan Province (No. 2013RYY04) and the Sichuan Provincial Education Department's Key Project (No.14ZA0210).

Our work was also supported by university Key Laboratory of Sichuan Province (No. 2013WYY09) and Fund Project of Sichuan Provincial Academician (Experts) Workstation (No.2014YSGZZ02).

References

[1] Shen Y T. Portable personal multimedia terminal: U.S. Patent D689, 856[P]. 2013-9-17.

[2] Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval[C]//Proceedings of the international conference on Multimedia. ACM, 2010: 251-260.

[3] Rabiner L R, Schafer R W. Digital Speech Processing [J]. The Froehlich/Kent Encyclopedia of Telecommunications, 2011, 6: 237-258.

[4] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. Signal Processing Magazine, IEEE, 2012, 29(6): 82-97.

[5] Muscillo R, Schmid M, Conforto S, et al. Early recognition of upper limb motor tasks through accelerometers: real-time implementation of a DTW-based algorithm [J]. Computers in biology and medicine, 2011, 41(3): 164-172.

[6] Zhu B B, Yan J, Li Q, et al. Attacks and design of image recognition CAPTCHAs[C]//Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010: 187-200.

[7] Lux M, Klieber W, Granitzer M. Caliph & Emir: semantics in multimedia retrieval and annotation[C]//Proceedings of the 19th International CODATA Conference. 2004: 64-75.

[8] Viswanathan M, Viswanathan M. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale [J]. Computer Speech & Language, 2005, 19(1): 55-83.