Embedded Speech recognition interaction system research

Qiong Luo

College of Foreign Languages, Wuhan Institute of Technology, Wuhan, China

Keywords: Speech recognition; embedded speech system; Recognition.

Abstract. With the continuous improvement and development of speech recognition technology, the numerous special purpose chips for speech recognition have been developed, thus, the practical products of speech recognition have been gradually appeared in the market. This paper will take the interpretation of the embedded speech system as the breakthrough point, combined with the analysis of the basic structure of speech recognition in the embedded system, it discusses the architecture of the speaker independent speech recognition system.

Introduction

With the development of computer technology, signal processing as well as the development of pattern recognition technology, speech recognition technology has been improved gradually, the fields of application have been more and more extensive, at the same time and a lot of speech recognition products have been appeared. Speech recognition products are widely used in voice dialing system, English and Chinese translation system, intelligent toys controlling, intelligent home controlling system, smart phones, stock trading system, banking service system, medical intelligence service, the intelligent automobile navigation, industrial control and some other fields.

The Interpretation of the Embedded Speech System

Embedded speech recognition system refers to the application of using various advanced microprocessor with the realization of speech recognition technology in board-level or chip-level with software or hardware. Embedded speech recognition system is required to achieve the optimization of algorithm under the premise of ensuring the recognition effect as much as possible, so as to adapt to the characteristics of the embedded platform with less storage resources and real-time. The large vocabulary continuous speech recognition system with high performance in the advanced level of laboratory can represent today's advanced speech recognition technology. But because of the limitation of the embedded platform in the aspects of resources and speed, the embedded implementation has still not been mature. While, because the algorithm of small vocabulary speech commands recognition system is relatively simple, whose demand for resources is small, and the rate of high recognition and robustness is rather high, which can meet the requirements of most applications, thus it has become the main focus of the embedded application.

The Basic Structure of Embedded Speech Recognition System

The Structure of Speech Recognition System. Speech recognition is a pattern recognition in essence, but its voice signal is more complicated, plus its content is quite rich, so speech recognition is much complicated than the general pattern recognition, the speech recognition system mainly includes speech signal pre-treatment, endpoint detection, feature parameter extraction, pattern matching, reference template library, and several other modules, the principle diagram of the system is shown in Figure 1.

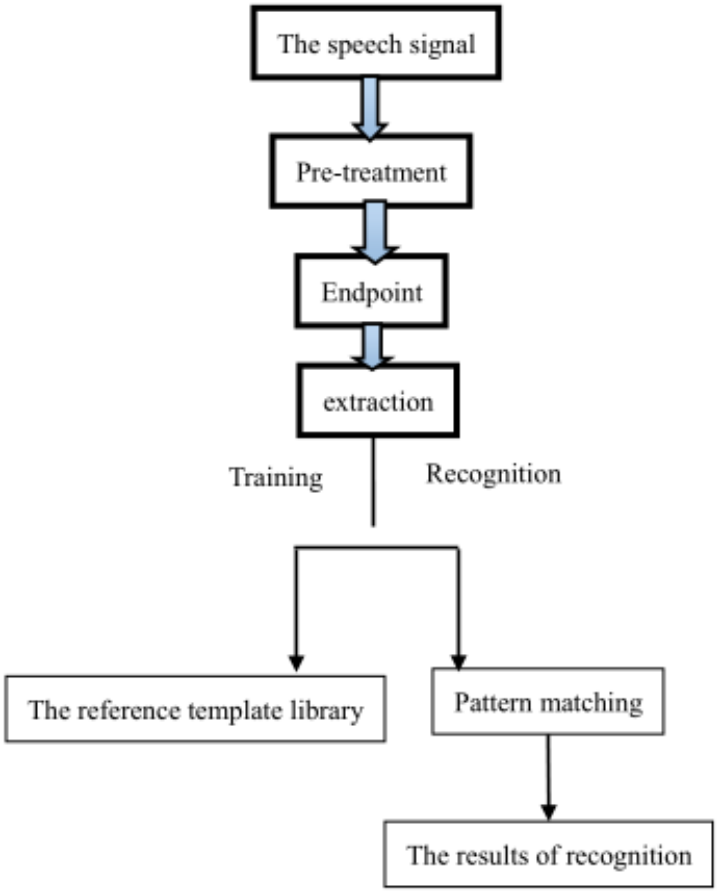

Fig.1 Block diagram of speech recognition system

The basic structure of speech recognition system includes: (1)pre-treatment includes the collection of speech signal and the operation of the pre -emphasis, window adding and framing operation and so on; (2)the endpoint detection can separate the speech signal effectively from the collected speech signals; (3)feature parameter extraction refers to extract key characteristic parameters from the signals that can reflect the characteristics of the speech signals; (4)the training phase refers to the acquired feature parameter vector after the speech signal is input by the user, the pre-treatment of the speech signal, endpoint detection and feature extraction,which takes the characteristic parameters of each speech signal as the template and the formation of the reference template library. (5) The recognition stage refers to contrasting the similarity between the feature parameter vector of the unknown speech signal and the reference template in the template library, taking the highest similarity model as the recognition results and then output.

Speech monitoring system of embedded PI transmission can realize the localization of the processing of speech information, so as to improve the performance of the server. Each device can have access to the Internet with the service function, namely, for each speech equipment, it can be regarded as an independent network terminal and thereby it can greatly improve the quality and scope of monitoring.

Embedded IP speech equipment as well as CP are connected in the network, each speech equipment of the whole system and CP can be regarded as the network device that has a unique PI address, thus each equipment can be recognized by the address of network PI. For each speech equipment, the most basic function is to collect speech, playback, compress the code and decode, act as network interface, etc. Each speech equipment is equivalent to a speech acquisition and monitoring equipment, they are working under the remote monitoring of CP, which can complete the acquisition of data, compress the code and transmit the data. Remote CP can directly monitor the location of the speech device and store the speech information, when it is necessary, it can broadcast through the network in a single or multiple speech broadcasting equipment.

The Structure of Speaker Independent Speech Recognition System

English large vocabulary continuous speech recognition system (LVCSR) [2-6] commonly adopts the acoustic model of the context as the modeling unit, which can search the algorithm once or multiple times, with N-GRAM language model, moreover, the amount of vocabulary generally can reach tens of thousands of words, so the requirements for the operation platform of computing power and storage capacity are very high, at present, it can only keep running in the main PC. Dictation machine used to be the main application mode of LVCSR, but in practical application, because the recognition rate of speech recognition engine and the robustness can not satisfy the requirement of application, so the application of dictation machine can not be widely spread. However, the characteristics of LVCSR system and natural language interaction with speaker independent is what the speech interaction interface persistently pursued.

BASELINE System

Figure 2 presents us the frame structure of BASELINE system of speaker independent speech recognition.

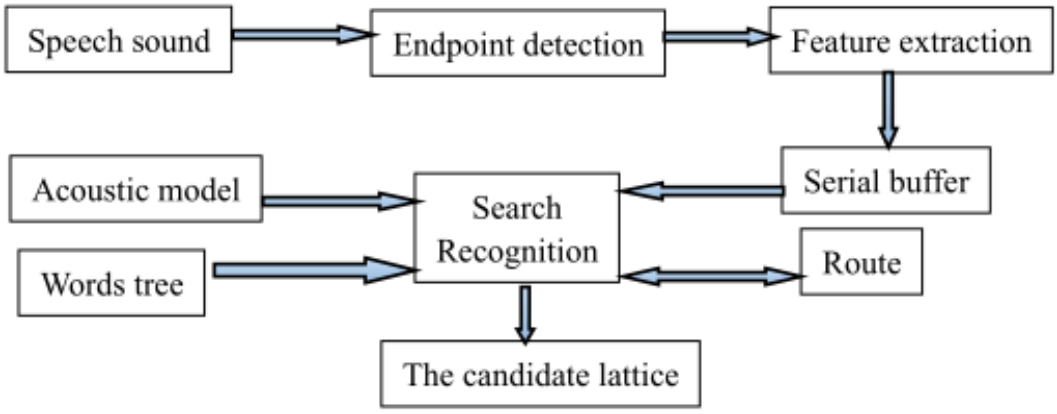

Figure 2 Framework of speech recognition system of speaker independent

The BASELINE of this system can be regarded as a simplified version of LVCSR. As for the specific simplified version it includes: ignoring the extension between words, so that the system can become speech recognition system of a command word; ignoring the language model, because without the extension between the words, speech recognition engine can be no longer sequential, thus it needs no language model; Lower the vocabulary, generally speaking, the smaller the vocabulary is, the lower the confusion degree of the word is, the higher the rate of the recognition engine is, at the same time, the smaller the data storing space, the searching space and the amount of computation is; Adopting the context-free acoustic model with voiceless sound, as for small vocabulary, context-free BASEPHONE model is much smaller than the contextual TRIPHONE models in data storage space and computation, whose recognition rate at the same time can also meet the requirements of the practical application. While, using tones can make the size of the model increase 5 times, which is sensitive to accent, therefore it is also ignored; reducing the sampling rate from 16 KHZ to 8 KHZ, which is shown by the experimental results that, as for small words in the table, the the drop of the recognition rate of the recognition engine caused by the drop of the sampling rate is less than 1%, but it can save 50% of the dynamic storage space of the front speech recognition and reduce 25% of the amount of calculation of front recognition. As for choice of the acoustic features, choosing "energy + MFCC + first order difference" with 26 dimensions totally, compared with the acoustic features of39 dimensions, it can save one third of the characteristics of buffer space.

Conclusion

Embedded speech recognition system has a broad prospect of market application. Speaker independent speech recognition system is introduced in this paper, which has many advantages compared with the speaker dependent isolated word speech recognition system, thus it can become the main focus of researching the embedded speech recognition system research as well as implementation. As for the embedded platform, researching and developing the front-end processing module of special speech recognition can make it perform more complex speech front-end signal processing algorithms.

Reference

[1]Lee K F, Hon H W. 1990, Speeker independent phone recognition using hidden markov models. IEEE Trans on ASSP, vol.37, p1641-1648.

[2]Lee T H,Wang Q C,Tan W K. 1993, A Framework for Robust Neural Network-Based Control of Nonlinear Servomechannism. Computer Speech and Language,vol.3, p190-197.

[3]Robert E.Uhrig. 1994,Application of Artificial Neural Networks in Industrial Technology. IEEE Transaction on Consumer Electronics,vol.10, p371-377.

[4]A.V.I Rosti, M.J. F Gales. 2004,Factor analysed hidden Markov models for speech recognition. Computer Speech and Language,vol.18, p181-200.

[5]Tomio Takara,Akira Hirayssu. 1997, Speech recognition using the model structure determined by the Genetic Algorithm. Nonlinear Analysis Methods and Applications,vol.30, p2969-2979.