Sensors & Transducers, Vol. 25, Special Issue, December 2013, pp. 91-97

www.sensorsportal.com

Sensors & Transducers

http://www.sensorsportal.com

Research on the Voice Control and its Audio Signal Processing in Flexible Manufacturing Cell

[1] Li Jing, [2] Xu Ting and [3] Shen Nanyan

[1] Shanghai Key Laboratory of Mechanical Automation and Robotics, School of Mechatronics Engineering and Automation, Shanghai University, 200072, Shanghai, China

[1] Tel.: 021-56337945, fax: 021-56331775

[1] E-mail: ianbest@shu.edu.cn

Received: 16 September 2013 /Accepted: 15 October 2013 /Published: 23 December 2013

Abstract: On the voice development platform of Microsoft Speech SDK, speech recognition and speech synthesis modules based on command control mode is built in this paper. Ethernet-based remote voice control system of intelligent flexible manufacturing cell is developed for machine tools and industrial robots. This paper designs an intelligent voice control system based on LabVIEW development environment, which realizes the human-machine voice interaction of flexible manufacturing cell and remote voice control. Audio signal processing is also designed in this paper based on MATLAB platform which largely reduces the effects caused by environmental noises. The mixing programming of LABVIEW and MATLAB can realize audio signal acquisition and processing easily and finally achieve reliable speech recognition system. Experimental studies have shown that intelligent voice control system has high speech recognition rate and system reliability both in quiet and adverse environments. *Copyright © 2013 IFSA.*

Keywords: Speech recognition, Speech SDK, Flexible manufacturing cell, Voice control, Speech signal preprocessing.

1. Introduction

Speech recognition is a high technology to make machines convert voice signals into appropriate texts or commands through recognition and understanding of the process. At present, Microsoft provides mature voice technology worldwide and has speech engine and corresponding development kit. Microsoft Speech API is a component object model (COM) and offers object programming interface, so it can be applied to any language that supports COM technology. SAPI (speech application programming interface) provides researchers with a large number of interfaces for speech recognition and speech synthesis secondary development, which can save system development time. As a result, Microsoft Speech SDK (software development kit) has broad applications [1]. In addition, Microsoft has also developed system speech recognition based on .net class library, which has the main achievement of Windows desktop operating speech recognition. It can also be able to realize simple voice secondary development based on the Windows API of WIN7 system. For the unsatisfactory results of speech recognition system based on Microsoft Speech SDK in noisy environments, many researchers have developed a robust customized system, including audio signal preprocessing, feature extraction and building speech recognition algorithm models, etc. The system can effectively improve the recognition rate in noisy environments, but it is not easy to achieve expected results considering the complexity

of development process and higher requirements on the developers. Meanwhile, in the real environment, it is impossible to capture pure speech signals which makes audio signal preprocessing essential. And the endpoint detection is relatively important which directly influences speech recognition results. So the algorithm of endpoint detection should be selected prudently.

Flexible manufacturing cell (FMC) is extensive development of flexible manufacturing equipment in recent years [2]. For its flexibility of adapting to process multi-species products and facility of voice recognition study, voice control of FMC can be achieved, then a further intelligent and humane flexible manufacturing cell can be realized.

The secondary development based on Microsoft Speech SDK is studied in this paper. Then a voice interaction module is built to apply speech recognition and speech synthesis to flexible manufacturing unit. Then Audio signal preprocessing is discussed in this paper and applied to Microsoft Speech SDK which can finally improve the recognition rate and system robustness in the adverse environment. And a voice remote control system for intelligent flexible manufacturing cell is designed and implemented.

2. Design of FMC Voice Control System

2.1. Hardware Structure of FMC

The flexible manufacturing cell in this paper consists of an EMCO lathe, an HTC50100 CNC machining center and a FANUC industrial robot. CNC machine tools and the robot are connected by input and output to realize connection of CNC and robot [3]. Meanwhile, these devices are able to communicate with the host computer via an industrial router. Then real-time status monitoring and information transmission can be achieved, so that appropriate voice commands can be sent by the host computer. The hardware structure of FMC is shown in Fig. 1.

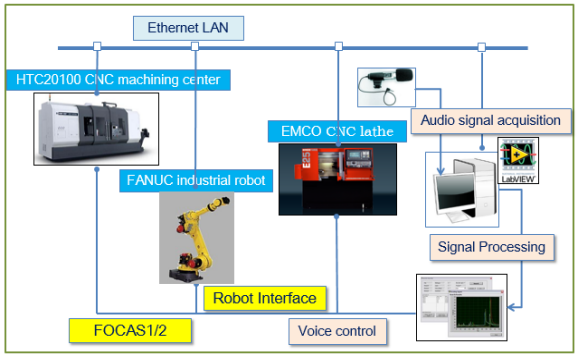

Fig. 1. The hardware Structure of FMC.

2.2. Design of Voice Control System

The intelligent flexible manufacturing cell is remotely monitored via voice commands to make the robot and CNC machines run the setup operations. For the requirements of remote control commands and signals transmission, the design of FMC system based on voice control is established. The system architecture is shown in Fig. 2.

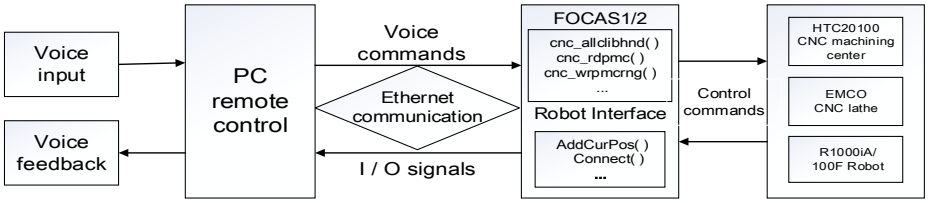

Fig. 2. The system architecture of intelligent FMC based on voice control.

The first part of the voice control system is PC remote monitoring. The most important work is to build a voice interaction module, including speech recognition module and speech synthesis module [4]. The current signal states of the robot and CNC machine tools are received at first, then PC gets input voice commands according to machines feedback signal states, and sends control commands. At the same time, the host computer obtains voice feedback signals and gives an expression of human-computer interaction [5]. The second part of the system is information exchange processing, which realizes the information transmission and remote control between host computer and FMC via Ethernet. The last part is appropriate operation of FMC according to the received commands. By calling FANUC FOCAS1/2 interface packet and robot interface packet, I/O signal status not only can be read and written ,but also can be sent to the host computer. Then the appropriate actions are performed according to the received voice commands.

3. Implementation of Audio Signal Processing of Speech Recognition System

3.1. Overview of Audio Signal Processing

The reason why speech recognition technology is difficult to apply from the laboratory to the real application environment is highly subject to noise.

The noise in the real environment can be divided into several kinds. One is periodic noise, such as the roar of the machine; other one is broadband noise, such as Gaussian noise or white noise; it also includes speech interference signal obtained from other unrelated speaker and so on. Thus, the voice signal processing is essential to speech recognition system which can largely influences the speech recognition rate and affects the stability and applicability of speech recognition system.

The audio signal preprocessing usually includes the following aspects: voice sampling, pre-emphasis, adding window and dividing frame and the endpoint detection and so on.

3.2. Implementation of Audio Signal Processing Software of Speech Recognition System

In this paper, speech signal preprocessing is carried out based on MATLAB development platform, project-oriented and scientific computing software. MATLAB itself provides certain audio processing capabilities of enabling the voice signal acquisition and playback. And it is very powerful and practical software with data analysis and processing functions [6]. Its signal processing and analysis toolbox provides a very rich set of features for speech signal analysis functions which can complete the speech signal processing and analysis quickly and realize the visualization of signals and convenient human-machine interactions.

Voice signal is one-dimensional analog signal with continuous changing time and amplitude. So voice signals should be converted into digital signals for computer analysis, and the analog voice signal is divided into two steps: sampling and quantization. Sampling frequency is selected as 8 kHz in this paper. For the convenient processing, amplitude normalized processing of signals is carried out by the command $x = x / \max(abs(x))$ which may finally realize the sampling and quantization [7].

Generally, pre-emphasis is used before speech signal analysis and the aim is to enhance the high frequency section and make the signal spectrum flat which can realize the right spectrum analysis. The transfer function of digital pre-emphasis filter is presented in equation (1).

$$r(n) = s(n) - es(n-1), \tag{1}$$

where (n) is the original voice signal sequences, r(n) is the sequence after pre-emphasis, e is a pre-emphasis coefficient which is a constant close to 1 and the number from 0.9 to 1.0 is always proper. Then select the command "open the door" as an input and the original wave and the wave with pre-emphasis are shown in Fig. 3.

Speech signal changes with time which is a non-stationary stochastic process. Meanwhile, the characteristics of speech signal is considered as the same in the interval from 10 ms to 30 ms. In order to reduce the voice frame truncation effects and avoid dramatic changes at both ends of voice frame, voice frame is needed to multiply a window function.

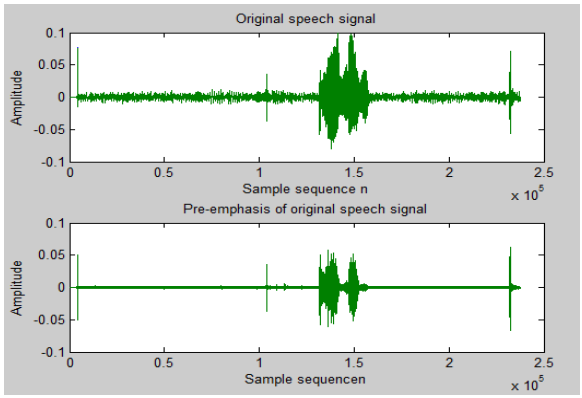

Fig. 3. Pre-emphasis of original speech signal.

The Hamming window is a common window function which is presented in equation (2).

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)] & 0 \le n \le N \\ 0 & n < 0 or n > N \end{cases} \tag{2}$$

Endpoint detection is to identify the start point and the end point of various passages (such as phonemes, syllables, morphemes, words, etc.) by digital voice signal processing techniques, which is a very important step for speech signals processing. The accuracy of endpoint detection directly affects the speed and result of speech signals processing which makes the research of endpoint detection algorithm a hotspot of speech signals processing, especially in adverse environments. In general, there are no completely pure speech signals in real environments and the speech signals are always with noises or other interferences. So the start point and the end point are needed to be identified from the input signal since the objective of the speech recognition system is the effective voice signal excluded the pure noise passages.

Many studies have been carried out based on the endpoint detection algorithm. One research is focused on cepstral distance measurement method which is to identify the noise frame and the speech frame by cepstral distance measurement [8]. This method has good robustness in adverse environments while selection of the threshold has great influences on system performance and it is hard to distinguish the processing speech from non-stationary noise when the door suddenly makes a loud sound or the phone rings, or the cepstral distance is too small. One other novel endpoint detection method based on time domain characteristic parameters has been studied by focusing on the changing rate of short-term energy. This method also has certain robustness, but failures

during the low SNR (Signal to Noise Ratio) still exist. Another study of the endpoint detection is carried out by comparing the results of autocorrelation method, HMM model and artificial endpoint detection which emphasizes on the relevance of voices. The feasibility of this new method has been proved although the processing algorithm is complex and needs large computation.

In conclusion, there are two general trends of most researches by using time-domain characteristic parameters for endpoint detection. One is the combination of multiple time domain characteristic parameters of voice signal. The other is the improvement of the existing endpoint detection method. Therefore, the double threshold endpoint detection algorithm is discussed in this paper which combines short-term energy with short-time zero-crossing rate. Speech signal is including silent voice signal, unvoiced segment and voiced segment. And the short-term energy is used to detect voiced sound while the short-time zero-crossing rate is used to detect the unvoiced sound. The combination can realize reliable endpoint detection.

The gap between high and low signal is artificially increased since the short-term energy does square operation to the signal. So use the short-term average to indicate the change of energy. The calculation formulation of short-term energy is shown in equation (4).

$$E(i)=\sum_{n=1}^{N}\left|x_i(n)\right| \tag{4}$$

The calculation formulation of short-time zero-crossing rate is presented in equation (5).

$$ZCR(i)=\sum_{n=1}^{N}\left|x_i(n)-x_i(n+1)\right|, \tag{5}$$

The waveforms of short-term energy and short-time zero-crossing rate are presented in Fig. 4.

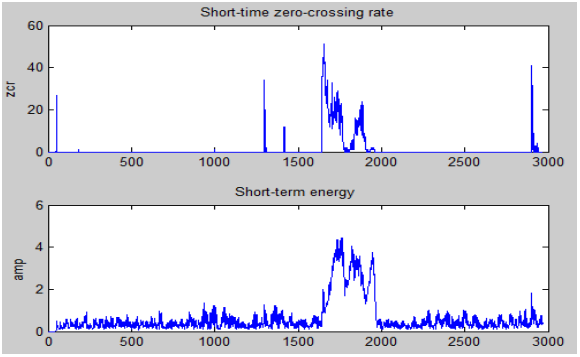

Fig. 4. Short-term energy and short-time zero-crossing rate of speech signal.

The waveform of endpoint detection of speech signal is shown in Fig. 5.

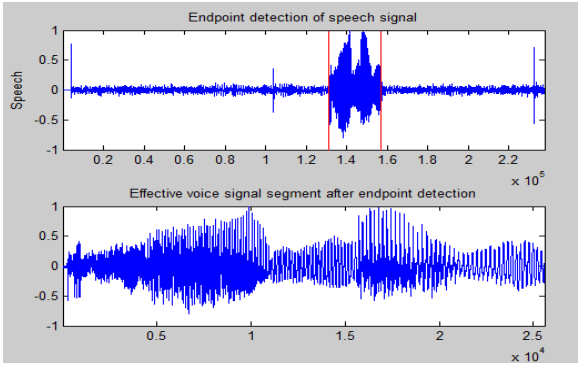

Fig. 5. Endpoint detection of speech signal.

4. Implementation of Speech Recognition System Based on Speech SDK

4.1. Overview of Microsoft Speech SDK

Microsoft Speech SDK is a COM-based development tools package on Windows operating system, which contains SAPI and Microsoft speech synthesis engine (TTS) and so on [9]. The underlying protocol is completely independent of the application layer as a COM component, which certainly shields the user from complex voice technology. As a result, users only need focusing on their applications, instead of spending much time in the complex work of building their own acoustic models. The development kit provides Chinese speech packets so that speech recognition and speech synthesis applications in Chinese can be easily developed. Fig. 6 shows the structure of Speech SDK.

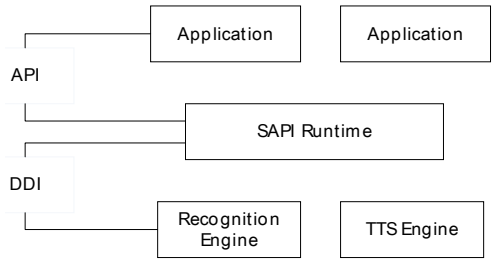

Fig. 6. The structure of Speech SDK.

According to the structure of speech SDK, the application communicates with SAPI by API, and recognition engine interacts with SAPI by the device driver interface (DDI). A series of work related to voice is done by COM component. Speech recognition is controlled by recognition engine and speech synthesis is managed by TTS engine. The main-used COM interfaces are shown in Table 1.

The recognition engine application can create two types of engine by using IspRecognizer interface: inproc (InProc Recognizer) engine and shared (shared Recognizer) engine. The object of inproc

engine can only be used by created applications, while that of shared engine can be used by multiple applications. The activated grammar rules loaded by applications also have two types: dictation (Dictation Grammar) syntax and command and control (Command and Control Grammar) syntax. Dictation grammar is used for continuous speech recognition, which can identify large vocabulary of engine dictionary. Command syntax is used to identify some specific command words and sentences that the user customized in a syntax file. These grammar rules are written in XML (Extensive Markup Language) file format, then activated and loaded by using IspRecoGrammar interface.

Table 1. The main COM interfaces for speech recognition and speech synthesis.

Main COM interfaces	Main functions
IspRecognizer Interface	To create an instance of speech recognition engine
IspRecoContext Interface	1. To receive and send an event message associated with voice recognition message 2. To load and unload recognition grammar resources
IspRecoGrammar Interface	To load activated grammar rules
IspPhrase Interface	To obtain recognition results, including texts, grammar rules and so on
IspVoice Interface	To get access to the TTS engine to realize text-to-speech conversion, then get voice feedback

4.2. Architecture and Implementation of Speech Recognition Software

Voice interaction module and FMC operation module in this paper are built and implemented based on LabVIEW development platform. And the speech signal processing developed on MATLAB platform can also be implemented on LabVIEW which makes the mixing programming possible and the speech recognition system be realized. LabVIEW has graphic programming environment and use data streams instead of the traditional text programming methods. This programming approach emphasizes the actual process of signal processing, which is conducive to simplify programming and reduce development time. Compared to other development languages, LabVIEW has facility of development, interactive graphic front panel, and full compatibility with the Microsoft Speech SDK. So this paper combines LabVIEW and MATLAB development platform with Microsoft Speech SDK to establish a voice remote control system of intelligent flexible manufacturing cell.

In this paper, voice command control mode is used to build speech recognition system of small vocabulary, isolated words and speaker-independence [10]. The recognition rate can reach 92 % or more, but grammar rules need to be created to match voice input. Therefore, the implementation of speech recognition program based on SAPI includes two major steps: the creation of grammar rules and the establishment of recognition engine.

XML files should be written and saved at first according to the mode specified by SAPI, which define necessary words, including control commands of the industrial robot, CNC machines and system control commands, such as "robot, load the workpiece", "lathe, clamp the chuck", "Exit" and so on.

The function CoInitialize must be called to initialize the COM library before using COM interfaces of SAPI5.4, and be unloaded by using function CoUninitialize. Then create a voice recognition engine (Recognizer) object and a speech recognition context object (RecoContext).After that, set user mechanism and message events in the application, connect RecoContext object with the message processing by function SetNotifyWindowMessage, and use function SetInterest to determine the messages that applications concern. Finally, create recognition grammar object (RecoGrammar), then call function LoadCmdFromResouce to load syntax rules from the XML file. Speech synthesis is achieved by ISpvoice interface and its member function Speak which realizes the text-to-speech conversion [11].

The robot or CNC machines will perform the appropriate action in real time if FMC operation commands are successfully recognized [12]. Some basic operations and interaction of machine tools and the robot can be voice controlled on LabVIEW platform. The user's operation commands and feedback commands are responded by TTS. The process of speech recognition and speech synthesis has the implementation of human-computer interaction. Design of the program is shown in Fig. 7.

5. Experimental Results and Analysis

According to the real-time status information got by the host computer in this experiment, the user gives out the corresponding voice commands. After recognition, the voice commands are transmitted to the robot and CNC machines via Ethernet to complete its real-time control effectively. The voice commands are given out by five people, including three boys and two girls. Two experiments are conducted in this paper. One is designed in both noisy and quiet laboratory environment to study the noises influences on speech recognition rate. The other experiment is designed to identify the effects of audio signal preprocessing by comparing the speech recognition system with audio signal processing and the system without any preprocessing under the adverse environment. The effects of using audio signal processing can be seen from these two cases.

Five people respectively provide five commands in each experiment, including door, chuck, tailstock, spindle and program, which have each of 10 repetitive trials. As a result, each set is tested 50 times, that is each 250 times in noisy environment and relatively quiet environment. The result is shown in Table 2. The other experiment has each 250 times for the speech recognition system with audio signal processing and the system without any preprocessing. Thus, the second experiment is carried out by comparing two cases. Case 1 is using the speech recognition engine to receive the input audio directly without any signal processing. Case 2 is combining speech signal processing algorithm with speech recognition engine. The results of two cases are presented in Table 3.

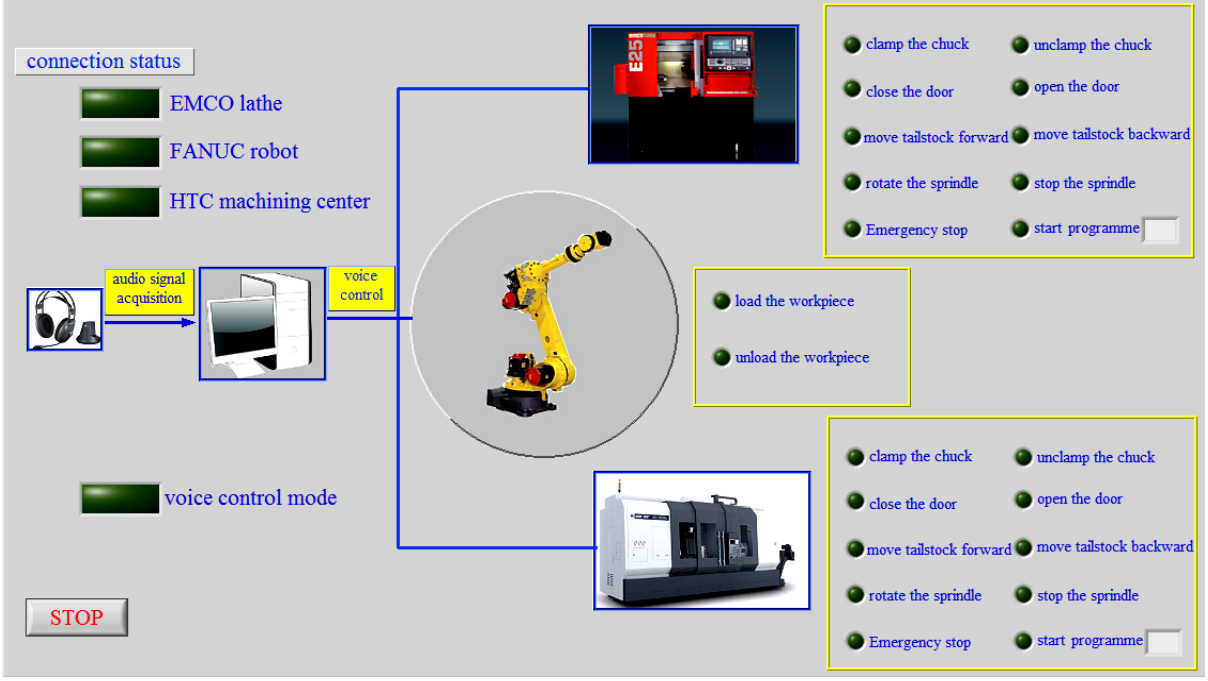

Fig. 7. The voice control platform of intelligent FMC.

Table 2. Experimental results of speech recognition in different environment (%).

	door	chuck	tailstock	spindle	program
quiet	98	94	96	98	96
noisy	80	72	82	76	76

Table 3. Experimental results of speech recognition in two cases (%).

	door	chuck	tailstock	spindle	program
Case1	80	72	82	76	76
Case2	92	86	90	86	88

From the experimental results, it can been seen that recognition rate of command control words based on Microsoft Speech SDK can reach up to 94 % in the quiet laboratory environment. Meanwhile, in the noisy environment, recognition rate decreased due to noise interference which makes speech engine have low sensitivity of command control words during audio signal acquisition. Meanwhile, after the speech signal processing, the speech recognition rate has increased greatly. Recognition rate in adverse environment can also reach 86 % or more. So it can be concluded that Microsoft recognition engine can get a higher recognition rate in a quiet environment while in the case of noise, the application of speech signal processing obtains a desirable recognition rate which is not easy to realize without audio signal processing.

6. Conclusions

This paper studies speech signal processing on MATLAB and secondary development based on Microsoft Speech SDK by calling SAPI interface functions to achieve speech recognition on Labview development platform. Voice interactive system of intelligent FMC is developed, which has rich interfaces, convenient of implementation and high speech recognition rate. From the experimental results, recognition rate of command control words in the relatively quiet environment is up to 94 % or more and the rate in adverse environment can also reach 86 % or more after speech signal processing which can largely avoid unrecognized commands or

error recognized commands caused by environmental noise. The mixed programming with MATLAB and LabVIEW can completely meet the requirements of practical application of FMC voice interaction and realize the high recognition rate and reliability of the speech recognition system.

Acknowledgements

The research discussed in this paper is supported by Shanghai Science and Technology Commission project under Grant No. 13DZ1101602.

References

[1]. Microsoft Speech SDK5.4 Help, (http://www.Microsoft.com).

[2]. T. Qiu, S. B. Chen, Y. T. Wang and L. Wu, Information flow analysis and Petri net based modeling for welding flexible manufacturing cell, *The International Society for Optical Engineering*, Vol. 4192, 2000, pp. 449-456.

[3]. Irfan Ullah, Qurban Ullah, Furqan Ullah, and Seoyong Shin J., Sensor-Based Autonomous Robot Navigation with Distance Control, *Journal of Computational Intelligence and Electronic Systems*, Vol. 1, Issue 2, 2012, pp. 161-168.

[4]. Kawashima, K, Sasaki, T, Miyata, T. Field test of remote control system for construction machines using robot arm, in *Proceedings of the IEEE International Conference on Control Applications*, Vol. 2, 2004, pp. 1171-1176.

[5]. Zhou Fengyu, Li Jinhuan, Tian Guohui, Research and implementation of embedded voice interaction system based on ARM in intelligent space, *IEEE Material Science and Information Technology*, 2012, pp. 5620-5627.

[6]. Md. Ahasan Habib, Mahe Zabin, and Jia Uddin, An Approach to Wavelet Based Image Denoising, *Journal of Computational Intelligence and Electronic Systems*, Vol. 1, Issue 1, 2012, pp. 144-148.

[7]. Liu Jing, Wang Lu, Qu Jinyu, Speech recognition simulation of isolated words based on DTW algorithm, *Journal of Shandong Polytechnic University*, Vol. 27, 2013, pp. 63-66.

[8]. Zhang Zhenyu, Experimental study of speech endpoint detection based on Matlab, *Journal of Zhejiang Science and Technology University*, Vol. 19, 2007, pp. 197-201.

[9]. Dale Rogerson, Inside facts of COM technology, 1999.

[10]. Wu Ping, Hu Rui Min and Ai Haojun, Research and Implementation of Speech recognition in Train Ticket Qury System, *Computer Engineering and Applications*, Vol. 39, 2003, pp. 33-37.

[11]. G. L. Li, H. B. Zhang, Design of an English Self-study System Based on TTS and SR, *Journal of East China Jiaotong University*, Vol. 26, 2009, pp. 86-90.

[12]. N. K. Mittal, Mohd Ahmed, and Farha Naaz, Efficient Palmprint Recognition System Implementation Using 2D Gabor Filters, *Journal of Computational Intelligence and Electronic Systems,* Vol. 1, Issue 1, 2012, pp. 154-160.