A Wireless Sensor Network-Speech Recognition Scheme Using Deployments of Multiple Kinect Microphone Array-Sensors

Ing-Jr Ding* and Shih-Kai Lin

Department of Electrical Engineering, National Formosa University, Yunlin, Taiwan.

Received 22 February 2016; received in revised form 28 March 2016; accepted 13 April 2016

Abstract

Speech recognition has successfully been utilized in lots of applications recently. With the development of the Kinect sensor device from Microsoft, speech recognition could be further promoted to be used in an ubiquitous environment where a wireless sensor network using Kinect sensors is deployed. This study develops a wireless sensor network (WSN)-speech recognition scheme using deployments of multiple Kinect microphone-array sensors. Presented speech recognition by Kinect-WSN could effectively capture the acoustic data made from the talking speaker and then perform the corresponding voice command control on certain target. In this study, different strategies to deploy multiple Kinect microphone-array sensors for constructing an ubiquitous Kinect-WSN speech recognition environment are investigated. Several different acoustic sensing data fusion methods are also explored for achieving superior performance on Kinect-WSN speech recognition. The presented method in this paper is evaluated the efficiency and effectiveness in an 5m×5m laboratory environment in which any of four test speakers is to make the voice command anywhere. Developed Kinect microphone array sensor-deployed WSN speech recognition in this work is finely utilized in various different applications in control.

Keywords: speech recognition, Kinect microphone array- sensor, wireless sensor network

1. Introduction

Speech recognition has been a matured technique for human machine-interaction (HCI) in the recent years. With the development of internet of things (IoTs) technology nowadays, the smart home scenario that most of equipment and devices in a family are connected and communicated with each other via wireless networks will be a practical integrated application. Conventional speech recognition viewed as the category of voice control interactions is that the voice command data provided by the specific user is acquired by the microphone in a very short distance from the user [1, 2]. Such voice-control speech recognition application can be widely seen in speech recognition on the smart phone platform and speech recognition on the central multimedia control panel platform in car. In the new technology of IoTs nowadays, different to conventional voice command-control speech recognition, in order to control all things connected to the internet in a home, office or other indoor environments, a new strategy for speech recognition developments, wireless sensor network (WSN)-speech recognition will be attracted much attention and a new and challengeable technique issue.

This study explores the utilization of the Kinect sensor [3, 4] for sensing and then acquiring the acoustic voice command of the user in an office environment. As many persons know, in addition to gesture recognition by Kinect [5-7], the Kinect device can also be used to perform speech recognition due to the embedded microphone array design composed of four microphones [8]. Speech recognition in this work will be performed in an acoustic sensing area that is properly deployed by multiple Kinect microphone-arrays. In the presented scheme of WSN-speech recognition by multiple Kinect microphone array-sensors in this paper, technical issues such as the (1) deployment method of the Kinect microphone array, (2) the establishment of client server-based wireless sensor network by Kinect sensors, (3) investigations of acoustic sensing data fusion methods, and (4) the possible application with practice in a real life using presented WSN-speech recognition by Kinect microphone array sensors will be considered, which will be detailed in the following section.

* Corresponding author. Email: ingjr@nfu.edu.tw

2. Speech Recognition via Wireless Sensor Network by Using Kinect Microphone Array-Sensors

The framework of WSN-speech recognition by Kinect microphone array-sensors explored in this study mainly contains Kinect sensor deployments by two Kinect microphone-arrays, client-server WSN establishments using TCP/IP protocol, acoustic sensing data fusion using a simple and computationally fast strategy. The developed framework is further performed in an application of voice sensing and remote control to the multimedia player component on a smart phone, which is depicted in Fig. 1.

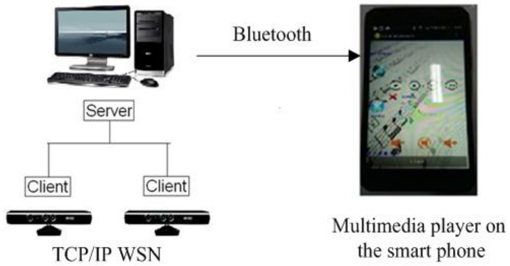

Fig. 1 WSN-speech recognition by deployed Kinect microphone array-sensors and its control application

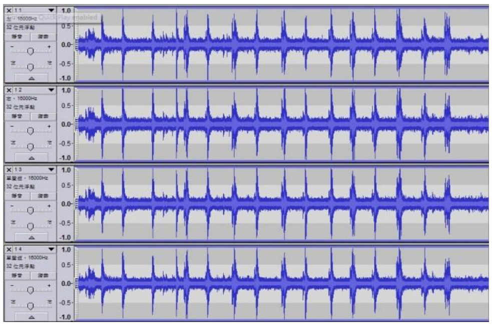

Fig. 2 Sensed data from a Kinect microphone array-sensor for data fusion calculations (four-channel voice data contained simultaneously in the unique Kinect sensor)

As depicted in Fig. 1, a voice command from a speaking user is sensed by two Kinect microphone array-sensors that are properly deployed in an office space. Each sensed data from each of these two Kinect sensors is sent to the server via the TCP/TP protocol, and the server end performs data fusion calculations for determining the recognition result of the sensed voice command. After the data fusion estimate, the recognized voice command is then sent to the smart phone device via the Bluetooth protocol to carry out a series of functional control on the multimedia player application program. In this work, two Kinect microphone array-sensors are designed to be appropriately localized inside to acquire almost most of all possible voice data.

Fig. 2 shows sensed data in a form of four channels. All these data come from the unique Kinect microphone array-sensor. Two Kinect microphone array-sensors deployed in this work are composed of 2 sets of the four-channel sensed voice data. These data are then considered the content of the voice command using data fusion calculations. The data fusion strategy employed in this study is a voice energy-based method. As shown in Fig. 2, each voice data from the Kinect sensor have the different values of energy. The value of the voice energy is dependent on the amplitude value of the data in certain time duration. The microphone in the Kinect microphone array has the large value of voice energy in case the voice data source (i.e. the speaking user) is located extremely near the microphone. Conversely, when the microphone in the Kinect microphone array is far away from the voice data source, the estimated voice energy to this microphone will be significantly small. Based on the above design thought-line, the primary principle of the data fusion method in this study is that the microphone with the sensed data of large-sized values will have more effects on the recognition decision of voice commands. The simplest method based on such the designed fusion principle is that the data fusion result is the recognition outcome of the microphone receiver where the sensed data has the largest values of energies.

In the part of control applications, the Bluetooth (BT) protocol is employed in this work to handle transmissions of the fused recognition command. The BT connection tunnel is firstly established in an initialization process to form a peer-to-peer connection pair between the server (the command provider) and the end-device of the smart phone (the command receiver). For speeding up command transmissions via BT, a command table containing a series of labels, each label with a text form representing a corresponding voice command, is properly devised. The multimedia player application platform in the smart phone will be finely operated by "remotely sensed voice commands made by the speaking user" under the regulation of the presented method.

3. Conclusions

In this paper, a wireless sensor network-speech recognition approach is presented by deploying the Kinect microphone array-sensors. Compared to conventional speech recognition, the presented framework considering sensor deployments, sensing data fusion, wireless communication scheme establishments, and possible extension application with practice provides an acoustic sensing way for command control in the application of internet of things. In addition, the presented approach with the use of Kinect sensors will also avoid the property of 'surveillance' and therefore can be much more acceptable by the users.

Acknowledgement

This research is partially supported by the Ministry of Science and Technology (MOST) in Taiwan under Grant MOST 104-2815-C-150-01 7-E.

References

[1] I. J. Ding, C. T. Yen and D. C. Ou, "A method to integrate GMM, SVM and DTW for speaker recognition," International Journal of Engineering and Technology Innovation, vol. 4, no. 1, pp. 38-47, 2014.

[2] I. J. Ding and Y. M. Hsu, "An HMM-like dynamic time warping scheme for automatic speech recognition," Mathematical Problems in Engineering, vol. 2014, Article ID 898729, 8 pages, 2014.

[3] I. Tashev, "Kinect development kit: a toolkit for gesture- and speech based human-machine interaction," IEEE Signal Processing Magazine, vol. 30, no. 5, pp. 129–131, 2013.

[4] Z. Zhang, "Microsoft kinect sensor and its effect," IEEE Multimedia, vol. 19, no. 2, pp. 4-10, 2012.

[5] I. J. Ding and C. W. Chang, "An eigenspace-based method with a user adaptation scheme for human gesture recognition by using Kinect 3D data," Applied Mathematical Modelling, vol. 39, no. 19, pp. 5769-5777, 2015.

[6] I. J. Ding and C. W. Chang, "Feature design scheme for Kinect-based DTW human gesture recognition," Multimedia Tools and Applications, pp. 1-16, July, 2015

[7] K. Qian, J. Niu and H. Yang, "Developing a gesture based remote human-robot interaction system using Kinect," International Journal of Smart Home, vol. 7, no. 4, pp. 203-208, 2013.

[8] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh and I. Tashev, "Microphone array processing for distant speech recognition: towards real-world deployment," Proc. Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012.