ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 134

2nd International Conference on Control, Automation, and Artificial Intelligence (CAAI 2017)

A Training System for Speech Disordered Children Based on the Intel RealSense Technology

Boyu Si[1], Zhaoming Huang[2] and Baodan Bai[1]

[1]School of Medical Instrument, Shanghai University of Medicine & Health Sciences, China

[2]Key Laboratory of Speech and Hearing Sciences Ministry of Education, East China Normal University, China

***Abstract*—A training system for speech disordered children is presented in this research. The core technology includes face tracking and speech recognition, which are supplied by Intel RealSense SDK and its relative hardware, such as 3D camera F200. The system consists of the pronouncing learning module and the speech disorder training module. The former can help children patient with almost no speech ability learn the pronouncing motion and method in Mandarin initials and finals with real-time 3D image playback. The training module which is based on the cartoon games can intervene patients' abnormal voice onset, speech volume, speech tone and vowel confusion. With the help of the speech recognition function from Intel RealSense, it can make an immersed environment which is benefit for training.**

Keywords-speech rehabilitation; Intel realsense; human-machine interaction; speech recognition

I. INTRODUCTION

The RealSense technology was developed by Intel in 2012. It is designed for natural human-machine interaction, which is a technical innovation to add human-like sensory organs for computing device based on deep analysis of the process from human sense to the realization of interaction [1]. The parts of human body which participants the interaction include eyes, ears, hands and mouth. The technical core of Intel RealSense is to supply the computing device with human-like sense hardware and software library, such as visual and auditory sense. As a result, computing device can give response and have interactions to the surrounding environment like human beings. Furthermore, we can develop more intelligent software and hardware systems by means of Intel RealSense technology to realize the recognition of users' facial expression, body gesture, speech, languages and the present scene. It can make a more immersive environment for users which is quite suitable for teaching, medical care, games etc.

Speech disorder is common in preschool children. In China only, the incidence of this disease has been over 4% in the 4-6 years old children. Speech disorder can severely affect the growth of children's language ability. Further, this situation can lead a rather harmful influence on children's communicating ability, and even the learning ability. However, due to the limitation of children's cognitive ability, the general interventions on speech disorder cannot attract young patients' attention. The rehabilitation job will not gain a satisfied effect though it is time-consuming, laborious. Some previous research announced the methods of speech disorder training based on computer games which can play a better role in attracting children patients' attention. These methods make full use of 2D or 3D cartoon frames and speech signal processing technology, and let the children patients complete the game tasks by controlling certain characteristics of pronunciation, such as the volume, tone and time length. The effect of these methods is great to the patients who have gained speech ability in some degree. But unfortunately, they are not suitable for the ones who have almost no speech ability. To them, the first important step of rehabilitation is to observe and simulate the correct pronouncing motion, get to know their own defects in speaking.

Therefore, we present a training system for speech disordered children from the view of both the learning pronunciation and the speech training. Based on the face tracking and speech recognition modules in the Intel RealSense SDK, children patients are encouraged to learn the essential ways of pronouncing first. The system can capture and play back the facial gestures of the patient by the special camera, and make a contrast with the correct pronunciation frame to help the patient be aware of his own problems. Then, after the essential ability of pronunciation is gained, the children patients can practice their speech strategies in a further step by means of completing a series of games which are developed on the speech recognition technology to achieve the goal of speech disorder intervention.

II. KEY TECHNOLOGY OF INTEL REALSENSE

A. Intel RealSense SDK and Hardware

The SDK structure is consist of multilayer components in Figure 1. The core function of the RealSense SDK is the I/O (input / output) module and the algorithm module. The former module is designed to achieve the signals from the input device or transmit the output information to the relative equipment. The algorithm module includes all kinds of pattern detection and recognition algorithms, which are the important ways of enhancing the innovated experience of human-machine interaction, such as face recognition, gestures recognition, speech recognition, text to speech, etc. The SDK has made a standardization of both the modules above in order to make the application programs realize its designed functions without dealing with the bottom system structure.

Of the entire sensory hardware device, the visual sensor can gain the largest amount of input information. But the ordinary camera can only make the real world image into a projection of 2D surface without the depth information, which is the

character of the real view. The depth information is quite important in completing the deep level visual tasks, such as object recognition, motion estimation, etc. In the Intel RealSense projects, 3D camera is employed to capture the depth information based on the infrared imaging technology [2]. It can easily capture the real 3D view into the digital device with the accurate coordinates of pixels which can be used in rebuild the real scene on screen. There are two types of Intel RealSense camera: the user facing camera F200 and the world facing one R200. In this paper, F200 is used mainly in our system, as shown in Figure 2.Its facial tracking range is 0.3 – 1.2 m. The valid gesture range is 0.2 – 0.6 m. It supports the maximum solution of 1080p with 30 fps of RGB image and the solution of 640 x 480 with 60 fps of the depth image capture.

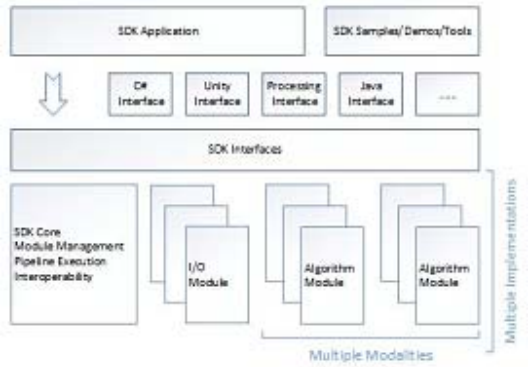

FIGURE I. OVERVIEW OF INTEL REALSENSE SDK

FIGURE II. INTEL 3D CAMERA (MODEL:F200)

B. Face Tracking

The face tracking module in the SDK supplies several kinds of facial algorithm: face detection, facial landmark detection, pose detection, facial expression detection and face recognition.

- Face detection. The system can detect the face from one frame of image or the video steam and locate the face with a rectangular box.

- Facial landmark detection. To a fixed given rectangular box, system can recognize the feature points further, such as eyes, mouth, etc. The position of eyes is rather important, because most application programs ensure the location which the users are looking at on the screen according to the position of eyes [3].

- Pose detection. When the user looks to a certain block of the screen, this function can estimate the orientation of user's face.

- Facial expression detection. The system can found some facial expressions, for example, closing eyes, frown and so on.

- Face recognition. The system can judge the user's identity by comparing the characters of the current face and the reference images in the database.

In this system, we employed the *Enable Face* function of the Intel Real Sense SDK to activate the face detection algorithm in the *Sense Manager* pipeline [4]. Then the callback function *Sense Manager* is used. When there is data to be process, SDK will response and start the callback function *On Module Processed Frame*automatically.

1) If the face detection algorithm is enabled during the initialized period, the C# code is:

Face Configuration. Detection. Is Enabled = true;

The function *Query Detection* is used to gain the location data of the current face.

2) If the facial landmark detection is enabled, the C# code is:

Face Configuration. Land marks. Is Enabled = true;

The function *Query Land marks* is used to gain any one of the landmarks detected by SDK, which can support the quantity of landmarks dynamically. Therefore, the system can access all the landmarks by function *Query Points*, or access them in groups by function *Query Points By Group*, or achieve a designated single landmark by function *Query Point*.

3) If the pose detection is enabled, the C# code is:

Face Configuration. Pose. Is Enabled = true;

The function is employed to gain any data of the detected poses, including *Head Position*, *Euler Angles*, *Position Quaternion* and *Rotation Matrix.*

C. Speech Recognition

The SDK support speech recognition, which means it can either recognize and execute the speech orders or transfer the input speech signal into text.

1) The interface *Audio Source* can be used to enumerate and choose one audio input device. The audio source can also be extracted from an audio file.

2) The function *Create Impl* is used to create an *PXC Speech Recognition* instance to realize the location.

3) The function *Query Profile* is used to achieve the available configuration and set the configuration by the *Set Profile* function.

4) The function *Set Dictation* is used to set the working mode of speech recognition module: order and controlling mode or dictation mode. The default working mode of speech recognition module is the dictation mode.

In the dictation mode, the system can recognize speech information on any vocabulary. While in the other mode, the

system can only recognize the specific speech tissues in the order list of SDK and ignore any word outside the list.

5) The function *Start Rec* is used to start speech recognition, while the function *Stop Rec* is used to stop. The application programs can accept any incident relative to the recognition activities.

Besides, it is rather necessary to set a high quality audio sampling to achieve a high recognition rate. The value of amplitude peak should be between FS-3 dB and FS-12 dB, here FS is the full-scale value of the A/D converter (ADC).

III. DESIGN AND REALIZATION OF THE SYSTEM

The main functional modules of this system consist of pronouncing learning module and speech disorder training module. The designing principle of these two modules follows the common regulation of learning language for Chinese children. Especially for the children patients who have almost no speech ability, the first important task is to let them study all kinds of pronouncing motion. In this paper, we design the system content by taking the children patients learning Mandarin as objects. The speech disorder training module aims to enhance the speech skills of the patients after they have mastered the basic pronouncing motion and methods, such as speech volume, tone, voice onset and vowel contrast.

A. Ponouncing Learning Module

The syllables in Chinese Mandarin include the initials and finals. The initials are mainly unvoiced, while all finals are voiced [5]. The quantity of finals is relatively larger, includes nasal finals, compound finals and single finals [6]. The type of the initials can be distinguished by the methods or the location of organs during pronouncing. The user interface of this module is shown in figure 3 to figure 6, which correspond to the nasal finals, the compound finals, the single finals and the initials. The left part of the screen shows the target phoneme, which are designed as a button individually. The right part is divided into two parts. The simulation cartoon of standard pronouncing motions is on the top, and the real-time speaking movement of the patients captured by the camera is on the bottom.

FIGURE III. NASAL FINALS LEARNING (MANDARIN)

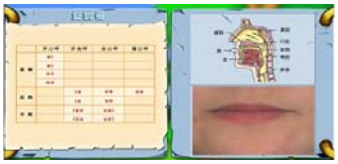

FIGURE IV. COMPOUND FINALS LEARNING (MANDARIN)

FIGURE V. SINGLE FINALS LEARNING (MANDARIN)

FIGURE VI. INITIALS LAEANING (MANDARIN)

B. Speech Disorder Training

The pathological types of children's speech disorder are mainly divided into voice onset disorder, abnormal volume, abnormal tone and vowel confusion. In this paper, we develop four training game modules for the four types separately in Figure 7 to Figure 10.

1) Voice onset: When the patient pronounces /a/ or the similar vowel, system starts the function of speech recognition. Each time when the patient manages the voice onset successfully, the gift box in the training game will give out a present as in Figure 7. The time for playing a game is limited, therapists can set a proper time according to the patients' ability on voice onset to finish the task. The time range can be set from 5 s to 10 s.

FIGURE VII. VOICE ONSET TRAINING

FIGURE VIII. SPEECH VOLUME TRAINING

2) Speech volume: Abnormal speech volume means the patients' volume is either too loud or too weak when they speak. As a result, we develop two types of training games for lower and increase the patients' speech volume. Take the lower-volume game as example shown in Figure 8. At the beginning, the patient speaks /a/ or a similar vowel with his

original speech volume. He will find that the spacecraft on the upper left of the screen cannot descend to rescue the astronaut on the lower right position. The therapist tells him to try to lower his pronunciation volume, and after a period of time for training in this way, the patient will raise his ability in controlling the volume during speaking. The theory of this training game for raise increase the speech volume is almost the same, but the training target is just opposite. The time for complete the games is limited, therapist can set it to 5-10 s. The principle is that the patient can make full use of the setting time to control his speech volume to finish the game.

3) Speech tone: The user interface of the speech tone training game is show in figure 9. There are three objects: a little fish, a shark and an octopus. Before training, therapist should move the shark and octopus to a proper position on the screen according to the actual ability of the patient in speech tone. When the game starts, the patient should continue pronouncing to control the little fish to move forward. The target of this game is to help the patient control his speech tone by raising or lowering it to change the little fish's route, in order to get away from the shark and octopus. If the patient fails, the little fish will be eaten by the shark or octopus. The gaming time is also limited to 5-10 s. It depends on the therapist' decision. Totally, patients should be give the reasonable time to finish the game and manage to control their speech tone.

FIGURE IX. SPEECH TONE TRAINING

4) Vowel contrast: The vowel contrast training is to improve patients' clarity and accuracy of pronunciation, avoid the happening of phoneme replacement or confusion. Before training, system needs to sample patients' best speech of target phoneme as the template for speech recognition. Hence, patients must finish the study of pronouncing learning module before starting this part. As the speech disordered patients' speech is generally instable, the system supplies a function of saving templates to control the problems. The therapists save the patients' best speech during several-time exercises as the templates for the following training. During the game process in figure 10, when the patient pronounces the target vowel, which is highly similar with the template, the boy on the screen will shoot the relative balloon. The gaming time is also limited to 5-10 s. It depends on the therapist' decision. The principle is that patients should be given the reasonable time to finish the game in the condition of maintaining the accuracy of vowel pronunciation.

FIGURE X. VOWEL CONTRAST TRAINING

IV. CONCLUSION

As the rapid development of human-machine interactive technology, computing device have been added more and more sensory hardware to realize the human-like functions, just like the eyes, mouth, ears and hands. Intel RealSense technology, as an excellent combination of software and hardware, has played rather an important role in this field. In this paper, we present a training system for children's speech disorder rehabilitation based on Intel RealSense technology. It supplies an immersed training environment for the children patients. We proposed a two-step rehabilitation plan. First, children with almost no speech ability should finish pronunciation motion and strategy study. This part is designed in the contrast pattern. Patients can tell their speech problems by observing the pronouncing motion differences between his and the standard norms. Then, the disordered users who have pronouncing skills in some degree can join the speech disorder training games. They will improve the ability on voice onset, speech volume, speech tone and vowel contrast. In this part, speech recognition technology is employed in analyzing speech signal and extracting the parameters to control the game. During the procedure, patients manage will gradually improve their speech ability.

The future work will focus on developing more immersed games for speech rehabilitation to make the training program more interesting and attractive. On the other hand, we are discovering more application to enhance the functions of the current system, such as eye-move tracking, gesture locating, to improve the effects of rehabilitation.

REFERENCES

[1] Draelos, M., Qiu, Q., Bronstein, A., and Sapiro, G. . "Intel realsense = Real low cost gaze," IEEE International Conference on Image Processing , IEEE, pp.2520-2524, 2015.

[2] Patil, Jayashree V., and P. Bailke. "Real time facial expression recognition using RealSense camera and ANN." International Conference on Inventive Computation Technologies, IEEE, 2017.

[3] O'Driscoll, G. A., et al. "Neural correlates of eye tracking deficits in first-degree relatives of schizophrenic patients: a positron emission tomography study," Archives of General Psychiatry, pp.1127, vol.12,1999.

[4] Martin, Miguel Vargas, V. Cho, and G. Aversano. "Detection of Subconscious Face Recognition Using Consumer-Grade Brain-Computer Interfaces," Acm Transactions on Applied Perception.vol.1, pp.1-20, 2016.

[5] Pai, H. F., Wang, H. C.. "A two-dimensional cepstrum approach for the recognition of mandarin syllable initials," Pattern Recognition, vol.4, pp.569-577,1993.