Voice Interaction Using Gaussian Mixture Models for Augmented Reality Applications

Mahfoud Hamidia[1,2], Nadia Zenati[1], Hayet Belghit[1], Kamila Guetiteni[2] and Nouara Achour[2]
[1]Centre de Développement des Technologies Avancées, CDTA, B.P. 17, 16303, Baba-Hassen, Algiers, Algeria.
[2]Faculty of Electronics and Computer Science, USTHB, B.P. 32, 16111, Bab-Ezzouar, Algiers, Algeria.
{mhamidia, nzenati, hbelghit}@cdta.dz, {k.guetiteni, noura.achour}@gmail.com, mahfoud_h@yahoo.fr

***Abstract*—This paper addresses the human computer interaction techniques for Augmented Reality (AR) applications. In fact, AR aims at inserting 2D or 3D virtual object generated by the computer in a real video filmed by a camera. On the other hand, the interaction in AR allows the user to take an action and control the virtual objects. In this work, Automatic Speech Recognition (ASR) system based on Gaussian Mixture Models (GMM) is investigated for voice interaction in AR.**
Experimental results show that good performance of the developed system. Also, the voice interaction provides an intuitive and a natural workspace for interacting with the augmented environment.

Keywords—Augmented Reality (AR); voice interaction; Automatique Speech Recognition (ASR); ARToolKit; Gaussian Mixture Models (GMM).

I. INTRODUCTION

Augmented Reality (AR) consists of adding virtual objects to the real world for enhancing the visual perception and the understanding to the user. It is a new paradigm of human machine interaction which combines real world and virtual objects in a real environment. Recently, AR has known various applications in several fields thanks to the technological evolutions such as: smart-phone, tablet, head mounted display.

AR systems are defined by Azuma [1] as having three characteristics: they combine real and virtual elements, they are interacting in a real time and they are registered in 3D. Fig. 1 shows a global structure of an AR system.

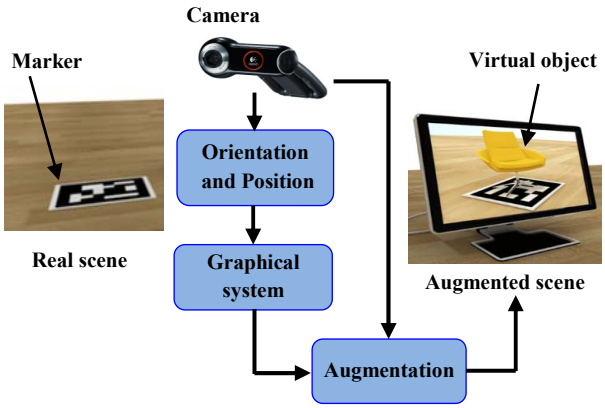

Fig. 1. Block diagram of Augmented Reality system.

The real scene is captured by the camera. Then, in order to make a virtual object in the right position into the real world, orientation and position techniques are used. Finally, a graphical system generates a virtual object when the augmentation scene is visualized by the display device.

Two approaches of AR are distinguished: marker AR and markerless AR [2]. In the first approach, artificial markers have been designed. ARToolKit [3] is the most popular marker used for AR. In markerless AR, natural features as color, shape, texture and interest point are extracted from a real scene using image processing techniques to calculate a camera's pose [4].

Human Computer Interaction (HCI) is the communication language, it consists of information exchange between the human and the computer. Several human computer interfaces have been developed to facilitate the user interaction with computer that touch and tangible interfaces. Moreover, Natural User Interfaces (NUI) take an interest part of studies in the last decade, which offer a natural interaction without intermediary. The control is performed using the human body where the user employs his voice, hands, eyes, the movement of his body, and thoughts for interacting with the computer. NUI avoids the use of visible control elements to the greatest possible extent in order to ensure a more natural control.

In Augmented Reality, the interaction becomes important, which offers the control of the virtual objects. It consists of: (1) navigation of the user, (2) selection of the virtual object, (3) manipulation such as: change of color, shape, position, and (4) application control [5].

Numerous interaction techniques have been developed for AR applications. Gesture interaction is investigated. Moreover, human body based interaction is widely exploited in AR games [6]. An avatar is produced with display screen and the user can track the movement of this avatar, where the Kinect device is generally used in this technique. Also, the interaction based on hand gesture is considered the most natural way of interaction, where the user can interact with virtual objects by hand.

Gaze interaction is one of the most natural interactions, where eye tracker device is used in this technique [7]. Other techniques are investigated in different applications of AR such as: leap movement based interaction, brain based interaction, facial expression based interaction, and voice interaction.

Speech is the most natural form of communication. Automatic speech recognition (ASR) is the core challenge

978-1-4673-6673-1/15/$31.00 ©2015 IEEE

towards the natural human-to-computer communication technology [8].

In this paper, we develop a voice interface using isolated word recognition system based on Gaussian Mixture Models (GMM). The main goal is to produce voice commands for controlling the virtual object in the AR application by speech understanding using speech processing techniques.

The rest of this paper is organized as follows: Section II explains the ASR principle; then ASR system based on GMM is described in Section III. Experimental results are shown in Section IV. Finally, Section V concludes the paper.

II. Automatic Speech Recognition

In recent years, speech recognition technology has increasingly become a hotspot of research, due to the development of multimedia technology.

Speech recognition technology is a process of extracting the speech characteristic information from people's voice, and then being operated through the computer and to recognize the content of the speech.

Automatic Speech Recognition (ASR) system is used to convert spoken words into text. It has very important applications such as command recognition, dictation, foreign language translation, and security control.

Fig. 2 shows a basic structure of an ASR system, where a statistical pattern recognition method mainly used [9]. ASR system is consisting of three processes: (1) feature extraction, (2) training and (3) matching.

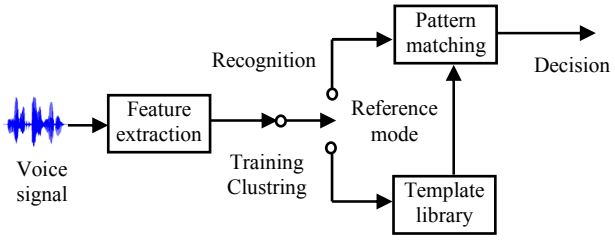

Fig. 2. Basic structure of automatic speech recognition system.

Mel-Frequency Cepstral Coefficients (MFCCs) are probably the most commonly used technique to represent the speech spectrum in ASR systems [10]. They are generated from the result of a filter bank analysis whose goal is to mimic the pitch perception of the human ear.

The MFCC process is subdivided into five phases. In the frame blocking section, the speech waveform is more or less divided into frames of approximately 30 ms. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The Fast Fourier Transform (FFT) block converts each frame from the time domain to the frequency domain.

A number B of triangular shaped filter bin functions equally spaced on the Mel scale is taken from the magnitude spectrum, it is defined as:

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f_{\text{Hz}}}{700}\right) \quad (1)$$

where f_{mel} is the frequency in Mel, and f_{Hz} in Hz .

Usually around 24 filter-banks are used to represent the spectrum. The log-spectral filter-bank outputs could be used for speech recognition. The Discrete Cosine Transform (DCT) is applied to the logarithm of the Mel-scale filtered frequencies. The first N coefficients (usually 13) are selected as a feature vector representing the selected frame

In the training process, the extracted features are trained using a Model. We have adopted a GMM modeling in this work, detailed in the next Section. Expectation Maximization (EM) algorithm is used to train the extracted features of the human voice in the system and then finally used to store in database.

III. Automatic Speech Recognition Based on GMM

After extracting features, we need to create the isolate word model using some statistical model like GMM. In fact, the Hidden Markov Models (HMMs) are widely used in speech recognition system. It covers from isolated speech recognition to very large vocabulary unconstrained continuous speech recognition.

GMM model is a parametric probability density function represented as a weighted sum of Gaussian component densities as proposed in [11]. The probability density functions of many random processes, such as speech, are non-Gaussian. A non-Gaussian probability density function may be approximated by a weighted sum (i.e. a mixture) of a number of Gaussian densities of appropriate mean vectors and covariance matrices. GMM is a special case of an HMM, by assuming independent and identically distributed consecutive frames. GMM is very competitive when compared to other pattern recognition techniques. It is more simple and faster than HMM with very small or no performance degradation.

GMM parameters are estimated from training data using the iterative EM algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model.

GMM is a weighted sum of M component Gaussian densities. It is denoted by its three parameters; mean, covariance matrix and weights by the notation:

$$\{\mu_{\text{i}}, \sum_{\text{i}}, p_{\text{i}}\}_{\text{i}=1}^{M} \quad \text{with the constraint} \quad \sum_{\text{i}=1}^{M} p_{\text{i}} = 1$$

For a feature vector x the mixture density can be stated as:

$$p(x/\lambda) = \sum_{i=1}^{M} p_{\text{i}}\, b_{\text{i}}(x) \quad (2)$$

where

$$b_{\text{i}}(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_{\text{i}}|^{\frac{1}{2}}} e^{\left\{-\frac{1}{2}(x-\mu_{\text{i}})^{t}\, \Sigma_{i}^{-1} [(x-\mu_{\text{i}})]\right\}} \quad (3)$$

where D is the dimension of the feature space, λ is the target isolated word model.

In matching process, the system authenticates the registered person by matching the real time voice sample to the stored voice samples in the database.

IV. Experimental Results and Evaluations

In this Section, we describe the developed system of isolated word recognition applied to an application of augmented reality. After that, we evaluate the performance of this system using multiple tests with discussion.

A. Isolated word recognition system

The main goal of this work is to develop a voice interface using MATLAB environment for controlling the virtual object in AR application. For this reason, we have realized an isolated word recognition system based on GMM. In the first step, sixteen voice commands are defined to correspond the AR application. These commands contain four application commands; each one includes four commands for manipulating the virtual object as shown in Tab I.

TABLE I. Voice commands of the developed system.

Command	Content
Rotation	Roll, Pitch, Yaw
Move	Horizontal, Oblique, Vertical
Zoom	Small, Average, Big
Color	Red, Blue, Yellow

In the second step, we produce a database which consists of a 320 set of the sixteen commands in English language. These commands are spoken by four speakers (two males and two females) with five repetitions for each command. It was recorded by Algerian speakers in a quiet environment, in a 16 bits WAV file format, with a sampling frequency of 8 kHz. The record duration of each command is one second.

MFCCs speech features are extracted from 30 ms frame duration and 20 ms overlapping with the previous frame of the command speech signal. We used the number $N = 12$ of MFCCs in this work. These coefficients are trained into GMM including: clustering, expectation, and maximization.

The Maximum Likelihood (ML) method is now the most popular parametric estimation method for GMM. The purpose of maximum likelihood estimates is to get model parameters of training data. Initializing model parameters is necessary. It includes the number M of Gaussian component and parameters of each Gaussian component.

For the test, MFCCs features are extracted from a recorded command speech signal by a microphone and compared with the GMM model of each command. The developed system based on GMM recognizes the command on the basis of log probability. It recalculates the log probability of voice vector and compares it to previously stored value.

B. Augmented reality system

For the AR application, we use visual studio 2010 C# environment with the popular open source library ARToolKit [12]. The computer processor is Intel ® core TM, i3, M380, 2.53 GHz. The Logitech Quick Cam Pro 9000 webcam is used with the resolution 320×240.

In augmentation task, AR systems need to know the relationship between the real scenes (3D world) and the corresponding images (2D) of this one. So transformation matrix must be calculated using the least squares method to obtain the matrix which relates 3D points with their projections. To insert the virtual object into a real scene, we use the obtained reference point from the ARToolKit maker to be recognized by ARtoolKit.

C. Performance Evaluation

To evaluate the developed isolated word recognition system performance, we use the Correct Recognition Rate (CRR%) criterion measure, which is defined as follow:

$$\mathrm{CRR}(\%) = \frac{\text{the recognized words number}}{\text{the total number of tested words}} \times 100 \qquad (4)$$

CRR is calculated for each command of the developed system; four speakers record their voice commands in the test. A total number of tests are 15 for each speaker, and the average of CRR is calculated with $M = 64$ number of Gaussian components. The obtained results for each command are shown in Tab II.

TABLE II. CCR evaluation of the different voice commands.

Command	CRR(%)	Command	CRR(%)
Rotation	100	Oblique	93,33
Move	86,67	Vertical	100
Zoom	86,67	Small	93.33
Color	93,33	Average	100
Roll	73,33	Big	86.67
Pitch	93,33	Red	93,33
Yaw	86.67	Blue	100
Horizontal	100	Yellow	100

These results demonstrate the good performance of the developed isolated word recognition system based on GMM in terms of high CRR for the most commands. A little difference of CRR between some commands is due to a pronunciation similarity.

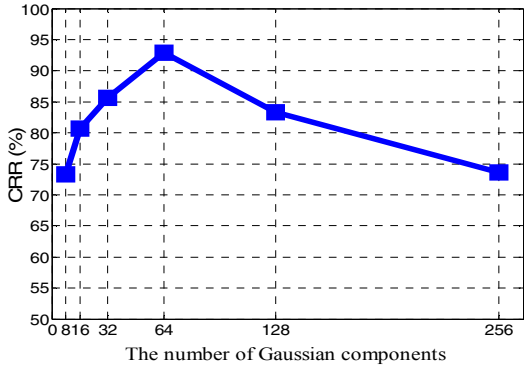

Fig. 3. CRR evaluation for different number of Gaussian components.

Fig. 3 shows the influence of the number of Gaussian components on the performance of the developed system. The experiment is carried out repeatedly 15 times by four speakers; we get the average as the result. The obtained results show that the CRR is maximum in 64 numbers of Gaussian components. This number is sufficient to give a good performance of the developed system.

We have applied the developed system of isolated word recognition to the AR application as shown in Fig. 4, which represents one of AR applications, an example of an advertisement for a car uses AR technology. We add a voice control option in this application to control the virtual 3D car.

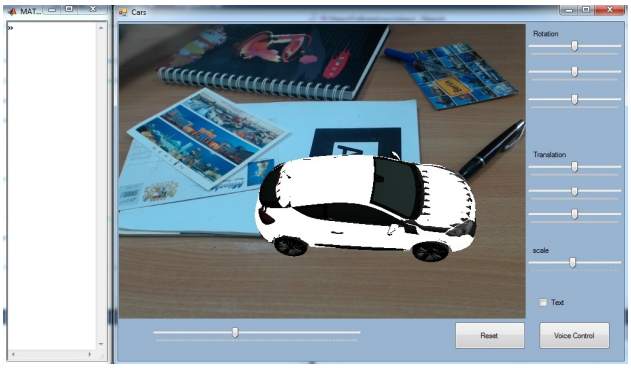

Fig. 4. Virtual car augmentation in the initail state.

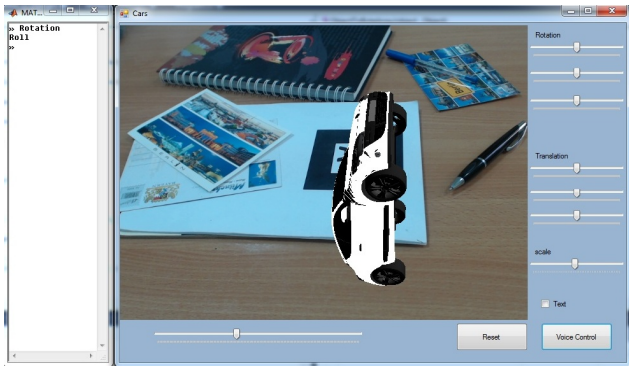

Fig. 5. Roll rotation of the virtual car.

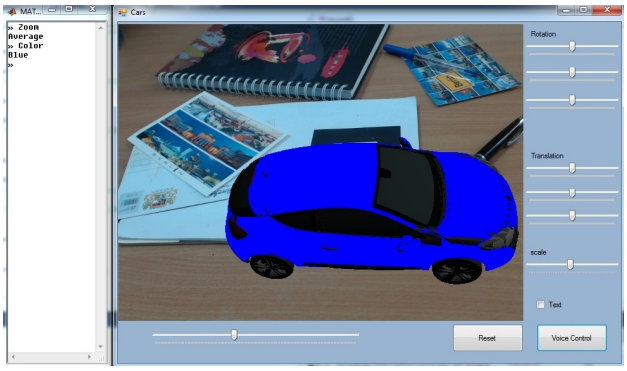

Fig. 6. Average zoom and blue color of the virtual car.

Figs. 6 and 7 show the virtual car manipulation using voice interaction; two examples are given: roll rotation and average zoom with a blue color. The decision of the speech recognition system is given in the MATLAB workspace. This technique of interaction allows the user to interact in a real time with the augmented environment by a natural way using voice commands.

V. Conclusion

In this paper, voice interaction has been investigated in AR. This latter can enhance the user visual perception by adding a virtual object. We have developed an isolated word recognition system based on GMM. This system has adopted for augmented reality application; voice commands are defined for controlling the virtual object. Experimental results demonstrate a good performance of the developed system in terms of high values of corrected recognition rate. Moreover, the voice commands offer a natural way of the user interaction with the augmented environment; manipulating of the virtual objet by a remote command with hands-free.

As future work, we plan to use voice activity detection to improve the speech recognition performance. Also a multimodal interaction technique will be investigated by voice and gesture recognition fusion.

References

[1] R. T. Azuma, “Survey of augmented reality,” Presence, Vol. 6, No. 4, pp. 355-385, 1997.

[2] M. Hamidia, N. Zenati-Henda, H. Belghit, M. Belhocine, “Markerless tracking using interest window for augmented reality applications,” In proc IEEE, International Conference on Multimedia Computing and Systems (ICMCS), pp. 20-25, 2014.

[3] H. Kato, M. Billinghurst, “Marker tracking and hmd calibration for a video-based augmented reality conferencing system,” In proc of the 2nd International Workshop on Augmented Reality, San Francisco, USA, pp. 85-94, 1999.

[4] M. Hamidia, N. Zenati-Henda, H. Belghit, A. Bellarbi, “Object recognition based on ORB descriptor for markerless augmented reality,” 9ème Conférence sur le Génie Electrique, EMP, Bordj El Bahri, Alger, 2015.

[5] D. A. Bowman, “Interaction techniques for common tasks in immersive virtual environments,” Doctoral dissertation, Georgia Institute of Technology, 1999.

[6] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, “Pfinder: Real-time tracking of the human body,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 780-785, 1997.

[7] F. Djelil, S. Otmane, S. Wu, “Apport des NUIs pour les applications de réalité virtuelle et augmentée: État de l’art,” Les 8èmes journées de l’AFRV, Laval, France, 2013.

[8] F. Fauziya, G. Nijhawan, “A Comparative study of phoneme recognition using GMM-HMM and ANN based acoustic modeling,” International Journal of Computer Applications, Vol. 98, No. 6, pp. 12-16, 2014.

[9] M. An, Z. Yu, J. Guo, S. Gao, Y. Xian, “The teaching experiment of speech recognition based on HMM,” In proc IEEE, the 26th Chinese Control and Decision Conference (CCDC), pp. 2416-2420, 2014.

[10] S. B. Davis, P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions Acoustics Speech and Signal Processing, Vol. 28, pp. 357-366, 1980.

[11] P. Zolfaghari, A. J. Robinson, “Formant analysis using mixtures of Gaussians,” In Proc SLP, pp. 904-907,1996.

[12] ARToolKit. http://www.hitl.washington.edu/artoolkit/.