Speaker Dependent Speech Recognition in Computer Game Control

Yusra Faisal Al-Irahyim, PhD
College of Computer Science and Mathematics,
University of Mosul
Mosul Iraq

Lujain Younis Abdulkadir
College of Computer Science and Mathematics,
University of Mosul
Mosul Iraq

ABSTRACT

Playing has become an integral part of people's lives since the beginning of time, and education games have become an important part of the education process in childhood, for school students and even for university students. Insertion of the voice commands in education games considered a big challenge especially regarding the speech accuracy and rapid response, to achieve this goal an educational game was designed aimed to teach students of Computer Science the fundamental concepts of " logic ", and to enable the game to allow speech input, the game should include the speech recognition system, to build that system, in this study three algorithms for feature extraction are used (MFCC , PLP and Rasta-PLP) with three VQ Code Book generation algorithms (LBG, LBG-PSO and LBG-PSOGA) were studied and applied, and was tested on 864 sound files for different peoples (4 male, 5 female), their ages between (16-30) year, through the results it was noted that when MFCC technique with LBG-PSOGA algorithm was used higher speech accuracy up to 98.5 % was obtained compared to other algorithms and techniques.

General Terms

Game Control, Speech Recognition, Vector Quantization, Algorithms.

Keywords

Pre-processing, Classification Algorithms, LBG-PSO Algorithm, Genetic Algorithm.

1. INTRODUCTION

Speech recognition is considered to be an interesting area through its applications and communication with computer through natural language[1].

There are many examples which used speech as input in controlling games, the benefit of this interaction is found in this example. In South Africa the interaction with a computer through speech recognition and playing game helps people to learn how to use computer in normal way[2].

Speech recognition and speech synthesis two ways to communicate with a computer using natural language. The problems of speech synthesis can be solved because it's more flexible. But the problem of a speech recognition is more difficult [1].

We suggest an educational game, which introduces speech dialogue into an education environment. The game is designed with a goal to teach university students the fundamental concepts of logic.

2. RELATED WORK

Harada et al. in 2011 designed a system "the voice game controller" playing the games by voice only. The results show that the input of silent much faster than the input of speech 50%[3].

Kumar et al. in 2012 designed two mobile games based on speech input used to help children to study English[4].

Hamalainen et al. in 2013 for the purpose of teaching children between 3-10 years the basic skills of music and mathematics design an educational game based on speech recognition [5].

Booth in 2014 design an educational game for the students of financial management to understand the basics of "Time value of money,". The researcher proves the objectives of using games in education, and the game "FinMan" proves the objectives[2].

3. GAME DESIGN

The game in this study was designed using Visual Studio 2012. The aim of this game is to teach students of Computer Science the fundamental concepts of the logic, since they have no previous background regarding this topic. To control this game, the student is free to choose between mouse and keyboard input or voice input, in addition, this game consists of 2 stages: 1) learning stage and 2) quiz stage.

3.1 Learning Stage

At this stage the student learns the fundamental concepts of the logic such as:

1. How to convert truth table to karnaugh map.
2. How to draw truth table according to a number of variables.
3. Learning the main rules in simplifying karnaugh map.
4. Learning the shape and truth table of each logic gate.

3.2 Quiz Stage

This stage consists of 15 different questions the answer is only true or false. The goal of this stage is to test the knowledge that students acquired in learning to stage.

4. SPEECH RECOGNITION

Speech processing and speech recognition are two fields of signal processing. We can define speech recognition as a set of words spoken by a person word after word [7].

There are three methods in speech recognition:

A. Acoustic Phonetic Method.

B. Pattern Recognition Method.

C. Artificial Intelligence Method.

5. COMPONENTS OF SPEECH RECOGNITION SYSTEM

The basic components of speech recognition system as shown in figure: 1.

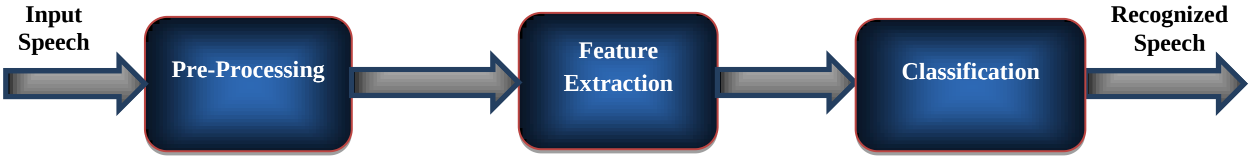

Figure 1: Basic components of speech recognition system

5.1 Preprocessing

A speech analysis is performed after taking an input through microphone from a user.

5.1.1 Silence Removal

Since processing the silent frames spend a time with no benefit, so the silence must be removed before processing from speech signal which contain silence at different positions such as beginning of the signal, in between the words, end of the signal etc. [8]. This can be done by:

a) We start extracting the general shape of the speech signal by making an envelope around it as follows:

1. Taking the absolute value of the speech signal.
2. Compute the mean value of the signal.
3. Use a function called imdilate() which essentially "dilates" a signal to a specified degree, allowing the creation of an amplitude envelope of sorts when applied to a speech signal.

b) The signal envelope is then compared to a threshold, the value of threshold that used in this study is 0.05.

c) Then erased from the signal any value that less than the threshold.

5.1.2 Pre – emphasis

In this stage we pass the speech signal through a filter which “emphasizes higher frequencies. This procedure will increase the energy of the signal at higher frequency”.

$$Y[n] = X[n] - a * X[n-1] \qquad (1)$$

Where

- $X[n]$ Input speech signal acquired after silence removal.

We assume a = 0.97, so that 97% of the samples are produced from the previous samples [9].

5.1.3 Normalization

To be ensured that the volume of the speech signal of a speaker during recording does not influence the analysis, we normalize it as follows [10]:

$$x = \frac{\left(\left(x - Mean(x)\right)\right)}{MAX\left(ABS\left(\left(x - Mean(x)\right)\right)\right)} \qquad (2)$$

Where

- x input speech signal acquired after pre-emphasis.

5.1.4 Framing

In this step the speech signal is divided into frames, the length of each frame equal 50 msec. Each frame contains N samples, Adjacent frames are being separated by M ($M < N$). The used values here are $M = 128$ and $N = 256$ [9].

5.1.5 Windowing

Hamming window is used as window shape. The Hamming window equation is given as[9]:

$$Y(n) = X(n) * W(n) \; 0 \le n \le N-1 \qquad (3)$$

Where

- $Y(n)$ Output signal.
- $X(n)$ Is an input speech signal acquired after framing.
- $W(n)$ Is a hamming window given by Eq. (4).

$$W(n) = 0.54 - 0.46\, cos\left[\frac{2\pi n}{N-1}\right] 0 \le n \le N-1 \qquad (4)$$

Where

- N number of samples in each frame.

5.2 Feature Extraction

There are many algorithms used in speech recognition for features extraction. Below some of them [7]:

A. Mel-Frequency Cepstrum Coefficients (MFCC).
B. Perceptual Linear Prediction (PLP).
C. Relative Spectral (RASTA-PLP).
D. Linear Predictive Coding (LPC).
E. Linear Discriminant Analysis (LDA) .
F. Discrete Wavelet Transform (DWT).
G. Linear Prediction Cepstral Coefficients (LPCC).
H. Principal Component analysis (PCA).

In this research the following techniques are used:

5.2.1 Mel-Frequency Cepstrum Coefficients (MFCC)

We use MFCC method for feature extraction, because it was considered as a standard method, also it was sensitive to the noise. The number of coefficients used in speech processing is about 20[11]. The steps of MFCC are shown in figure 2.

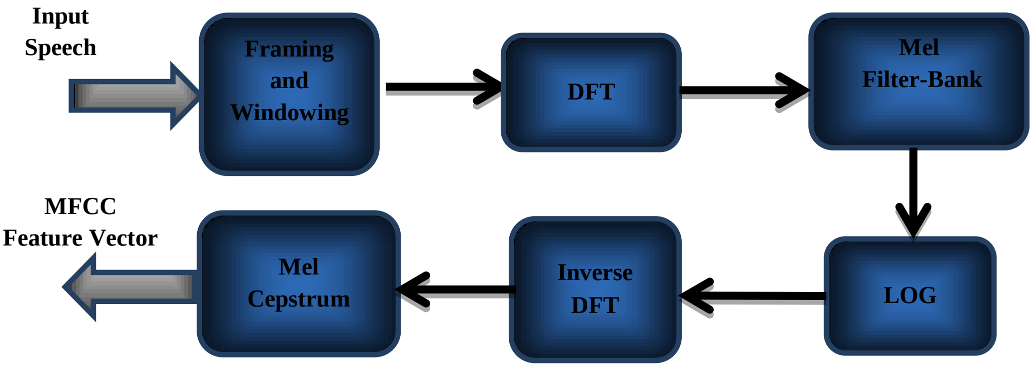

Figure 2: Block diagram of MFCC

5.2.1 Perceptual Linear Prediction (PLP)

The PLP and LPC are two methods for feature extraction, they are based on the short term spectrum of speech, but PLP method describes the psychophysics of human hearing more accurately [11]. The steps of PLP are shown in figure3.

5.2.3 Relative Spectral(Rasta-PLP)

PLP and Rasta two different methods. PLP was used to make the differences between speakers low during the processing of speech information. But Rasta was used to apply in each frequency of the energy a band-pass filter to smooth variations of noise and to remove any constant offset [11].

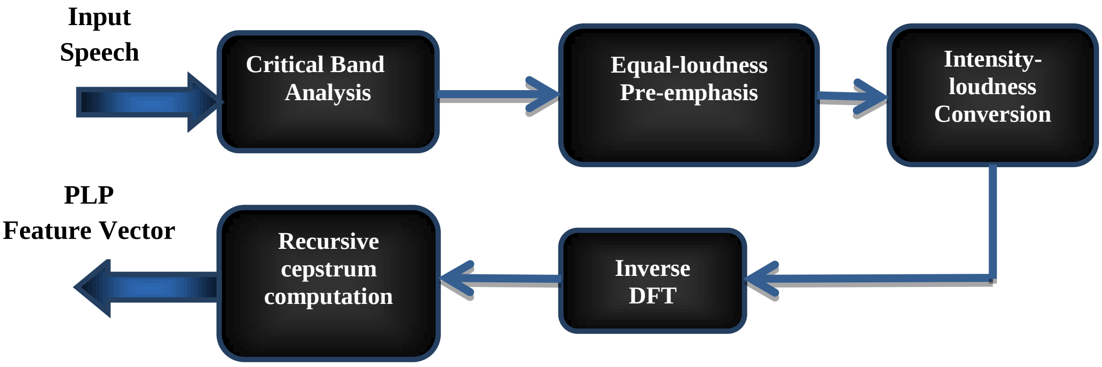

Figure 3: Block diagram of PLP

5.3 Classification Algorithm

5.3.1 Vector Quantization (VQ)

Mathematically VQ is a mapping function that maps k dimensional vector space to a finite set $CB = \{C_1, C_2, \dots \dots, C_N\}$, the set CB is called Code Book consisting of N number of code vectors and (also called code words) and each code vector $C_i = \{C_{i1}, C_{i2}, \dots \dots, C_{ik}\}$ Is of dimension k. The speech input signal is divided into set of training vector $X_i = \{x_{i1}, x_{i2}, \dots \dots, x_{ik}\}$. We search the Code Book to find the nearest code word C_{min} Using Eq. (5) Which represent the square of Euclidean distance with vector X_i With all the code words of the codebook CB[1 [].

$$d(x_i, C_j) = \left\{\sum_{i=1}^{dw} (x_{ik} - C_{jk})\right\}^{1/2} \qquad (5)$$

Where

- $d(x_i, c_j)$ The distance between vectors x_i And centers c_j.
- dw number of features in data vector.
- $(x_{ik} - c_{jk})$ Euclidean distance between a vector x_{ik} And codeword c_{jk}.

The data vectors was putted in clusters. Using Eq. (6) calculates cluster.

$$C_i = \frac{1}{n}\sum_{x \in s} x_j \qquad (6)$$

Where

- c_i Are the centers or code word i in the vector space.
- n number of data vectors in subset s.
- x_j The input data vector.

In this section we discuss VQ codebook generation algorithms: LBG, LBG-PSO and LBG-PSOGA used in the research.

$$x_i^{d+1} = x_i^d + V_i^{d+1} \qquad (9)$$

5.3.1.1 LBG Algorithm

Linde-Buzo-Gray(LBG), It's used a mapping function to partition training vectors in N clusters. The mapping function is defined as: $Rk \to CB$. Let $X = (x_1, x_2, \dots \dots, x_k)$ be a training vector and $d(X; Y)$ be the Euclidean distance between any two vectors. Below the steps of LBG for a codebook generation [13] :

1. The initial codebook CB_0 generated randomly.
2. $i = 0$.
3. Perform the following process for each training vector, compute the Euclidean distances between the training vector and the code words in CB_i, the Euclidean distance by Eq.(5) search the nearest code word among CB_i.
4. Partition the Code Book into N cells.
5. Compute the centroid of each cell to obtain the new Code Book CB_{i+1}.
6. Compute the average distortion for CB_{i+1}. If it is changed by a small enough amount since the last iteration, the Code Book may converge and the procedure stops.
7. Otherwise $i = i + 1$ and go to Step 3.

5.3.1.2 LBG-PSO Algorithm

A new algorithm is particle swarm optimization(PSO), it's a branch of evolutionary computation technique. On the design of the Code Book we use the fitness function which assign fitness values to the potential solution, it's based on Eq.(7)[14].

$$fitness = 1 \Big/ \sum_{j=1}^{k} \sum_{x \in c} d(x_i, c_j) \qquad (7)$$

- x_i^{d+1} Is the position of the next iteration.
- x_i^d Is current position.
- V_i^{d+1} Is velocity of next iteration.

The advantage of LBG algorithm, it can converge faster, but at the local minimum solution is terminated. The advantage of PSO algorithm, it can search for the global best solution, but the LBG algorithm is faster than it. The algorithm of LBG-PSO put the results of LBG algorithms into the initial global best particle. The steps of this proposed algorithm are as follows[14]:

1) Execute the LBG algorithm for one time.
2) We put the result of a LBG algorithm into one particle and initialize locations of rest particles by applying k-mean algorithm for one iteration and we randomly link velocity of all particles.
3) Using Eq.(7) we compute the fitness value for each particle.
4) We compare the fitness value to each particle's with the previous best value. If the value is better, we update p-best and we take current location as the particle's best location.
5) We find the highest fitness value of all particles. If the fitness value is better than g-best, we change g-best with this fitness value, and we take the global best location.
6) We update the velocity and position of each particle using Eqs. (1) and (2) respectively.
7) The sum of distance between each data vector and its cluster center is computed using Eq. (3) in K-means algorithm.
8) We assign each data vector to the closest cluster based on the new location of each particle, and then we calculate the cluster center using the Eq. (4).
9) Repeat the steps from 3 to 8 until the maximum iteration exceeds.

Where

- $d(x_i, c_j)$ The distance between vectors x_i and centers or codewords c_j.

Each particle has two locations. One location is named global best (gbest), the other is called personal best (pbest)location, the population of particles is flying in the search space and every particle changes his location according the gbest and the pbest with Eqs. (8) and (9) respectively[14].

$$V_i^{d+1} = wV_i^d + c_1 * r_1 * \left(pbest_i^d - x_i^d\right) + c_2 * r_2 \left(gbest^d - x_i^d\right) \qquad (8)$$

Where

- r_1, r_2 Is a uniformly distributed random variable that can take any value between 0 and 1.
- V_i^{d+1} Is velocity of next iteration.
- V_i^d Is current velocity.
- x_i^d Is current position.
- $pbest_i^d$ Is the location of the particle that experiences the best fitness.
- $gbest^d$ Is the location of the particle that experiences a global best fitness value.
- c_1 and c_2 are two positive acceleration constants responsible for degree of informed consideration of personal and swarm memory respectively.
- w represents ithe inertia weight, which is usually linearly decreasing during the iterations.

5.3.1.3 LBG-PSOGA Algorithm

The PSO algorithm converges to a stable point, the Genetic algorithm takes more iteration for finding best cluster centers. The hybrid between the two algorithms mean the position of particle updated, by applying a crossover operation, the data fly to a new search area if it swapped between two particles. So to avoid the local maxima we apply mutation to PSO to increase the diversity of the population[15,16]. Because of the mentioned disadvantages of GA and PSO, in this study, PSO hybrid with GA is proposed to improve the performance of each of those algorithms. The steps of this proposed algorithm are as follows:

1) We run the LBG algorithm for one time.
2) The result of LBG algorithm was assigned to one particle and the locations of rest particles was initialized by applying k-mean algorithm for one

iteration and associated velocity of all particles was initialized randomly.

3) The fitness value was computed using Eq.(7) for each particle.

4) We select the two best particles using its fitness value and we add them in the next generation.

5) We select two particles from the rest particles randomly then:

   a. Apply crossover between particles and find out two offspring .

   b. Calculate the fitness value of two offspring by Eq. (7).

   c. Select best two among old particle and two offspring .

   d. Update population by best two.

   e. Repeat this step(5) until the number of particles is over.

6) After the crossover operation is done do:

   a. Select one particle randomly.

   b. Apply mutation operation on this particle and find out offspring.

   c. Calculate the fitness value of offspring using Eq. (7).

   d. Update population by best one.

7) The fitness value of each particle's was compared with the previous best value for the same particle's, If the result is better, we update p-best and we take the current location as the best location for the same particle's.

8) We find the highest fitness value of the whole particles. If the value is better than g-best, we replace g-best with this fitness value, and we take the global best position.

9) We use Eqs. (8) and (9) to update velocity and location of each particle.

10) In K-means algorithm we use Eq.(5) to find the sum of distance between each data vector and its cluster center.

11) The data vector is assigned to the closest cluster depending on the new location of each particle, and then the cluster center is calculated using Eq. (6).

12) Repeat the steps from 3 to 11 until exceed the limit of maximum iteration.

6. THE RESULTS

After training the sound and design of optimal Code Books in different algorithms, has been tested on 864 sound files and the speech accuracy for each algorithm & techniques were calculated according to the Eq. (10) :

$$Recognition\ accuracy = \frac{1}{n\sum_{k=0}^{n} RR} \qquad (10)$$

Where

- n is the total number of persons.
- RR value of rethe recognition ratioor each person according to Eq. (11) :

$$Recognition\ Ratio(RR) = \left(\frac{tw}{rw}\right) * 100 \qquad (11)$$

Where

- tw total number of words.
- rw number of words recognized correctly.

Figure 4: illustrates the speech accuracy for each technique of feature extraction techniques and for each algorithm of classification algorithms that have been implemented in this study.

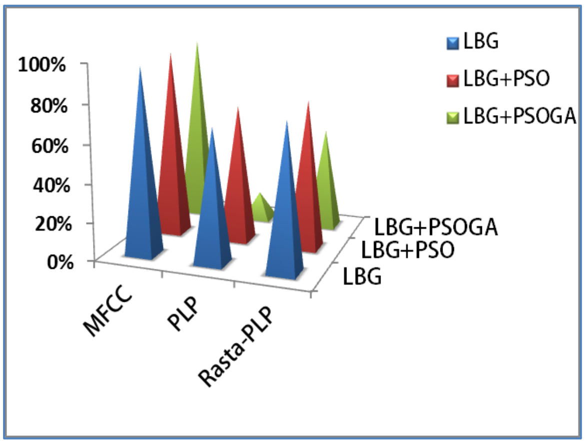

Figure 4: Speech Accuracy

7. CONCLUSION

The purpose of this study was to design a speech recognition system. 3 feature extraction techniques (MFCC, PLP and Rasta-PLP) with 3 VQ codebook generation algorithms (LBG , LBG-PSO and LBG-PSOGA) were studied and applied, and tested on 864 sound files for different people, through the results it was noticed that when MFCC technique with LBG-PSOGA algorithm was used gave the higher speech accuracy up to 98.5 % compared to other algorithms and techniques.

8. REFERENCES

[1] Nagy M., 2007, " Penguin Quart-Slovak Digit Speech Recognition Game Based on HMM ", in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovation , eds. Maglogiannis I., Karpouzis K. and Bramer M., (Boston: Springer), pp. 179-186.

[2] Booth M., 2014, " Combining Games and Speech Recognition in a Multilingual Educational Environment ", MSc. dissertation in School of Information Technology North-West University, Vaal Triangle Campus.

[3] Harada S., Wobbrock J., Landay J., 2011, "Investigation into the Use of Non-Speech Voice Input for Making Computer Games More Accessible ", © IFIP International Federation for Information Processing.

[4] Kumar A., Reddy P., Tewari A., Agrawal R., Kan M., 2012, " Improving Literacy in Developing Countries Using Speech Recognition-Supported Games on Mobile Devices ".

[5] Hämäläinen A., Pinto F., Rodrigues S., Júdice1 A., Silva S., Calado A., Dias M., 2014, " A Multimodal Educational Game for 3-10-Year-Old Children:

Collecting and Automatically Recognizing European Portuguese Children’s Speech ", International Conferenceon Computational Processing of Portuguese, © OATAO, pp. 1-11 .

[6] Shrawankar U., Dr. Thakare V., 2014, " Techniques for Feature Extraction in Speech Recognition System ", Research Student, Computer Science & Engg., SGB Amravati University.

[7] Kekre H. B., Sarode T. K., Save J. K., 2012, "New Clustering Algorithm for Vector Quantization Using Walsh Sequence ", International Journal of Computer Applications (0975 – 8887), Vol. 39, No.1.

[8] Mittal M., Lamba R., 2013, " Image Compression Using Vector Quantization Algorithms ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6.

[9] Kaveh A., Rad M., 2010, " Hybrid Genetic Algorithm and Particle Swarm Optimization for the Force Method-Based Simultaneous Analysis and Design ", Iranian Journal of Science & Technology, Vol. 34, No. B1, pp. 15-34.

[10] Premalatha K., Natarajan A. M., 2010, "Hybrid PSO and GA models for Document Clustering ", Int. J. Advance Soft Comput. Appl., Vol. 2, No. 3, Copyright © ICSRS Publication.

[11] Meena Y., Shashank, Singh V., 2012, " Text Documents Clustering Using Genetic Algorithm and Discrete Differential Evolution ", International Journal of Computer Applications (0975 – 8887), Vol. 43, No.1.

[12] Strik H., Luigi Palumbo L., Wet F. D., Cucchiarini C ., 2015, "Web-based mini-games for language learning that support spoken interaction ", © ISCA Workshop on Speech and Language Technology in Education, SLaTE Group.

[13] Gamit M., Prof. Dhameliya K., Dr. Bhatt3 N., 2015, " Classification Techniques for Speech Recognition ", A Review, International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 2.

[14] Jasmine J. M., Sandhya S., Dr. Ravichandran K., Dr. Balasubramaniam D.,2016 " Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameter ", International Journal of Innovative Research in Computer and Communication Engineering, Volume 4, Issue 4.

[15] Rehmam B., Halim Z., Abbas Gh., Muhammad T., 2015, " Artificial Neural Network- Based Speech Recognition Using DWT Analysis Applied on Isolated Words From Oriental Language ", Malaysian Journal of Computer Science, Vol. 28, No. 3, pp. 242-262.

[16] Easwari N., Ponmuthuramalingam P., 2015, " A Comparative Study on Feature Extraction Technique for Isolated Word Speech Recognition ", International Journal of Engineering and Techniques.