Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.7, No.4, 2016
www.iiste.org

Speaker Gender Recognition Using Hidden Markov Model

Dr. Yusra Al-Irhayim* Abeer abdulkafor
College of Computer Science and Mathematics, University of Mosul, Mosul Iraq

Abstract

Gender is an important demographic attribute of people. With the evolution in modern technologies in various fields of life and entering the computer systems in all applications, this led to the use of transactions instead of these technologies and human speech processing, and speaker recognition technology race.
In this research we build a system to distinguish the gender of the speaker, and through the audio information that has been obtained from the speech signal, passes the system in four phases, namely the phase of initial processing, and phase of features extraction, we use (MFCC) (Mel Frequency Cepstral Coefficients) technique, then comes the phase of training the EM algorithm was used to achieve the greatest expected limit, and finally the testing phase, which has been applied hidden Markov models in it. All algorithms and programs have been written using the language of Matlab.

Keywords: Gender Recognition, Hidden Markov Model, Mel Frequency Cepstral Coefficients, Speech Recognition

1. Introduction

Gender Classification has become area of extensive research due to it's increasingly powerful applications. Automated gender classification has attracted much attention over the last ten years since augmenting this ability with applications specific to a particular gender can provide a more user-friendly environment and human-like interaction.

Till date, the work has been emphasized on gender recognition through visual observation, but now, it has to be emphasized to computer, to perform this task. It is observable that our behavior and social interaction are greatly influenced by genders of people whom we intend to interact with. Hence a successful gender recognition system could have great impact in improving human computer interaction systems in such a way as to make them be more user-friendly and acting more human-like. Over the past decades, there have been significant advances in facial image processing, especially, in a face detection area where a number of fast and robust algorithms have been proposed for practical applications. As a result, a number of research areas attempting to extend the works have been emerging, face recognition, facial expression recognition and gender recognition(Gupta2015).

Recently there has been a growing interest to improve human-computer interaction. It is well-known that, to achieve effective Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should mimic human-human interactions. HCII is becoming really relevant in applications such as smart home, smart office and virtual reality, and it may acquire importance in all aspects of future peoples life.

Speech age & gender recognition aims at recognizing the underlying state of the speaker from his or her speech signal. This is mainly motivated by intelligent Human Machine Interaction required for different kinds of applications (Jawale2015).

In this paper, a system is presented for gender recognition using speech signal. The aim of the paper is to improve the recognition rate by finding out good feature parameters based on Mel Frequency Cepstral Coefficients. For classification we have used Hidden Markov model.

2. Previous Work

This paper provides a survey of human gender recognition in computer vision. A review of approaches exploiting information from face and whole body (either from a still image or gait sequence) is presented. They highlight the challenges faced and survey the representative methods of these approaches. Based on the results, good performance have been achieved for datasets captured under controlled environments, but there is still much work that can be done to improve the robustness of gender recognition under real-life environments(Ng2012).

This paper have proposed a new method for gender classification method which considers three features. The new method uses fuzzy logic and neural network to identify the gender of the speaker. To train fuzzy logic and

neural network, training dataset is generated by using the above three features. Then mean value is calculated for the obtained result from fuzzy logic and neural network. By using this threshold value, the proposed method identifies the speaker belongs to which gender. The implementation result shows the performance of the proposed technique in gender classification(Meena2013).

In this paper, multiple levels hierarchical techniques based on 3 Sigma control limits on Neural Network is applied for gender recognition to get the desired objectives. In order to achieve this, the proposed algorithm considers the Artificial Neural Network as basic classifier. Here, in Initial Level Hierarchy, facial features are given as input to the Neural Network. Then, the output represents the gender classification from the Neural Network is extracted. The next level of classification can be done in Core Hierarchical Decision. Findings:

This paper provides an effective approach that classifies human gender in computer vision applications. In the proposed research, a Feed Forward Neural Network works at the primary level, based on the outcome of the primary level, the further classification is done in the next higher level hierarchically. In this research, there are 1000 gray-scale with 256 gray levels facial images used for experiment. Each image size is normalized to 64×64. Among the 1000 experimental images, 800 images are used as training data, and the remaining are used as test images. Prediction of the gender is more accurate and effectively achieved the success rate of 95 percent(Dileep2015).

In this paper, they propose a novel gender recognition framework based on a Fuzzy Inference System (FIS). Their main objective is to study the gain brought by FIS in presence of various visual sensors (e.g., hair, mustache, inner face). They use inner and outer facial features to extract input variables. First, They define the fuzzy statements and then they generate a knowledge base composed of a set of rules over the linguistic variables including hair volume, mustache and a vision-sensor. Hair volume and mustache information are obtained from Part Labels subset of Labeled Faces in the Wild (LFW) database and vision-sensor is obtained from a pixel-intensity based SVM+RBF classifier trained on different databases including Feret, Groups and GENKI-4K. Cross-database test experiments on LFW database showed that the proposed method provides better accuracy than optimized SVM+RBF only classification. They also showed that FIS increases the inter-class variability by decreasing False Negatives (FN) and False Positives (FP) using expert knowledge. Their experimental results yield an average accuracy of 93.35% using Groups/LFW test, while the SVM performance baseline yields 91.25% accuracy(Danisman2014).

In this paper the proposed system allows recognizing a person's emotional state from audio signals. The proposed solution is aimed at improving the interaction among humans and computers, thus allowing effective human-computer intelligent interaction. The system is able to recognize six emotions (anger, boredom, disgust, fear, happiness and sadness). This set of emotional states is widely used for emotion recognition purposes. It also distinguishes a single emotion versus all the other possible ones, as proven in the proposed numerical results. The system is composed of two subsystems namely Emotion recognition (ER)&Gender recognition (GR). For this two support vector machines (SVM'S)are used for the male and female speaker's emotion recognition. The experimental analysis shows the performance in terms of accuracy of the proposed ER system. The results highlight that the a priori knowledge of the speaker's gender allows a performance increase. The obtained results also show that the features selection adoption assures a satisfying recognition rate and allows reducing the employed features(Kumar2015).

3. Speech Process

In this section, we will describe the procedure for pre-processes. The analog voice signals are recorded thru microphone. It should be digitalized and quantified. Each signal should be segmented into several short frames of speech which contain a time series signal. The features of each frame are extracted for further processes. The procedure of such pre-process is shown in Figure(1).

The analog voice signals are recorded thru microphone. It should be digitalized and quantified. The digital signal process can be described as follows:

$$x_p(t) = x_a(t)p(t) \quad (1)$$

where $x_p(t)$ and $x_a(t)$ denote the processed and analog signal. $p(t)$ is the impulse signal.

The purpose of pre-emphasis is to increase, the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio (SNR) by minimizing the adverse effects of such phenomena as attenuation distortion.

$$y[0] = x[0]$$

y[n] = x[n]-0.95x[n] , 1<= n <= N (2)

where N is the sampling size.

While analyzing audio signals, we usually adopt the method of short-term analysis because most audio signals are relatively stable within a short period of time. Usually, the signal will be segmented into time frame, say 15 ~ 30 ms. There are always overlap between neighboring frames to capture subtle change in the audio signals. The overlapping size may be 1/3~1/2 of frame (Huang2011).

4. Feature Extraction Techniques

Feature extraction is the first step in gender recognition system. It aims to extract features from the speech waveform that are compact and efficient to represent the speech signal. The most famous features extraction techniques are:(Elkourd2014)

- Linear Predictive Coding (LPC).
- Formants
- Perceptual Linear Prediction (PLP).
- Mel-Frequency Cepstral Coefficient (MFCC).

4.1 Mel-Frequency Cepstral Coefficient (MFCC)

Mel-frequency Cepstral coefficient is one of the most prevalent and popular method used in the field of voice feature extraction. The difference between the MFC and cepstral analysis is that the MFC maps frequency components using a Mel scale modeled based on the human ear perception of sound instead of a linear scale. The Mel-frequency cepstrum represents the short-term power spectrum of a sound using a linear cosine transform of the log power spectrum of a Mel scale. The formula for the Mel scale is:

$$M = 2595log_{10}\left(\frac{f}{700}+1\right) \qquad (3)$$

MFCC as frequency domain parameters are much more consistent and accurate than time domain features. the steps leading to extraction of MFCCs: Fast Fourier Transform, filtering and cosine transform of the log energy vector. MFCCs can be obtained by the mapping of an acoustic frequency to a perceptual frequency scale called the Mel scale. MFCCs are computed by taking the windowed frame of the speech signal, putting it through a Fast Fourier Transform (FFT) to obtain certain parameters and finally undergoing Mel-scale warping to retrieve feature vectors that represents useful logarithmically compressed amplitude and simplified frequency information.by applying discrete cosine transform to the log of the Mel-filter bank. The results are features that describe the spectral shape of the signal(Makhijani2011).

5. Classification Schemes

Following are the few main classifiers found mostly used in speech Recognition system.

5.1 Hidden markov model

In the Markov model each state corresponds to one observable event. But this model is too restrictive, for a large number of observations the size of the model explodes, and the case where the range of observations is continuous is not covered at all. The Hidden Markov concept extends the model by decoupling the observation sequence and the state sequence. For each state a probability distribution is defined that specifies how likely every observation symbol is to be generated in that particular state. As each state can now in principle generate each observation symbol it is no longer possible to see which state sequence generated a observation sequence as was the case for Markov models, the states are now hidden, hence the name of the model.

A Hidden Markov model can be defined by the following parameters:

- The number of distinct observation symbols M.
- An output alphabet ={ } 1 2 M V v ,v ,...v
- The number of states N.
- A state space Q = {1, 2,...N}

States will usually be indicated by i, j a state that ‘the model is in’ at a particular point in time t will be indicated by qt. Thus, qt = i means that the model is in state i at time t. A probability distribution of transitions between states { }ij(Jawale 2015).

A = a , where

$$a_{ij} = P(q_{t+1} = j | q_t = i) \quad 1 \le i, j \le N \qquad (4)$$

5.2 Gaussian mixture models

Gaussian mixture model is a probabilistic model for density estimation using a convex combination of multi-variant normal densities. It can be considered as a special continuous HMM which contains only one state. GMMs are very efficient in modeling multi-modal distributions and their training and testing requirements are much less than the requirements of a general continuous HMM. Therefore, GMMs are more appropriate for speech emotion recognition System phases(Ladde2013).

6. The proposed algorithm for recognition

Here we propose a new gender recognition approach based on Hidden Markov Model (HMM). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender of unseen test speakers.

7. Proposed Work

In this paper, we design a new HMM-based approach for speaker, gender estimation, which improves the accuracy of the state-of-the-art with statistical significance. In the following steps we explain the algorithm used in the training and classification phases:

Step one: The database used in this system consist of 160 samples(120 samples used in the training phase, 40 samples used in the recognition phase) recorded in a clean environment(20 male speakers*3 repetition + 20 female speakers*3 repetition =120) .

Step two: Divide incoming signal to overlapping frames of equal size, each frame consists of (N = 512) of the samples and the amount of overlap (M = 256) samples, from any interference by 50% ,the next step after creating frames.

Step three: Applying the window function, on each frame of the speech signal.

Step four: Extract features using MFCC method, features vectors represent sequence of observations(O), we generate the primary matrices for the model: $\lambda^i = (A^i, B^i, \pi^i)$, the primary values of the parameters are chosen randomly. In general the primary values of the matrices π and A equal 1⁄(N,) but the primary values of the matrix B equal 1⁄(M.)

N: represent number of states for HMM.

M: represent number of observations.

Step five: we apply Baum Welch algorithm for training, so we re estimate the parameters for the model λ=(A,B,π), The goal is to maximize the probability of the sequence of observations.

Step six: we repeat step three until the tolerance equal 10^{-4},, or the number of iterations reach to a specific number, so we get the last parameters for the specific model for each file which is

$\lambda^N = (A^N, B^N, \pi^N)$ $N = (1,2, \dots \dots \dots .32)$

Figure(2) explain training algorithm.

Step six: In classification phase we find the probability for the vectors with each model $P(O/\lambda^w)$ using the forward backward algorithm, figure(3) explain classification algorithm.

After an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Finally, Supervised HMM is applied to detect the gender group of unseen test speakers.

8. Results

In this research in the data collection stage we recorded 40 different speakers (20 male & 20 female) in a clean environment, 4 times for each speaker (3 of them are used in training and the remaining are used in the test phase).

The percentage rate = (The number of correct recognition)/(The total number of audio files) × 100%

The percentage rate for male recognition in training phase = 58/60×100%=96.6%

The percentage rate for female recognition in training phase = 53/60×100%=88.3%

The percentage rate for all speakers recognition in training phase = 111/120×100%=92.5%

The percentage rate for male recognition in test phase = 16/20×100%=80%

The percentage rate for female recognition in test phase = 17/20×100%=85%

The percentage rate for all speakers recognition in test phase = 33/40×100%=82.5%

9. Conclusion

In this paper the main goal was to develop a gender recognition system using speech signal. The feature selection is one of the most important factors in designing a gender recognition system. From the study of different previous research works it was observed that among the different features the MFCC. Thus the MFCC has been selected as the feature for recognition. Among the different technique, statistical analysis produce very good results. This is why this method was selected in recognition process to obtain improved performance. The average recognition accuracy is 82%. This approach was implemented in the working platform of MATLAB for testing. During testing if a speech signal is given as input it will identify the gender of the speaker to which speaker belongs.

References

Danisman, Bilasco, &Martinet,(2014),” Boosting Gender Recognition performance with a Fuzzy Inference System”,Vol 42, Issue 5, pages 2772-2784.

Dileep, &Danti,(2015),” Multiple Hierarchical Technique to Predict the Gender of a Person based on 3 Sigma Control Limits on Neural Network”, Indian Journal of Science and Technology, Vol 8(14).

Elkourd,(2014),”Arabic Isolated Word Speaker Dependent Recognition System”,Msc Thesis, Islamic University, Gaza, Palestine, Deanery of Higher Studies, Faculty of Engineering, Computer Engineering Department.

Gupta,(2015),”Gender Detection using Machine Learning Techniques and Delaunay Triangulation”, International Journal of Computer Applications (0975– 8887) Vol.124 No.6.

Huang,(2011),” An Effective Approach for Chinese Speech Recognition on Small size of Vocabulary”, Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.2.

Jawale, patil, &Agrawal,(2015),”Identification of Age And Gender Using HMM”,

(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) ,1643-1647.

Kumar, RangaBabu,(2015),” Emotion and Gender Recognition of Speech Signals Using SVM”, International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 4, Issue 3.

Ladde,(2013),” Hybrid Classifiers for Gender Driven Emotion Recognition”, International Journal of Science and Research (IJSR),ISSN (Online): 2319-7064.

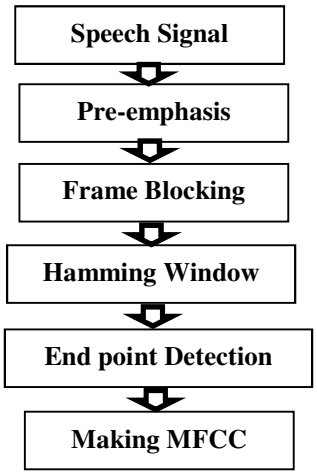

Figure 1. Pre-processes of Speech Signal

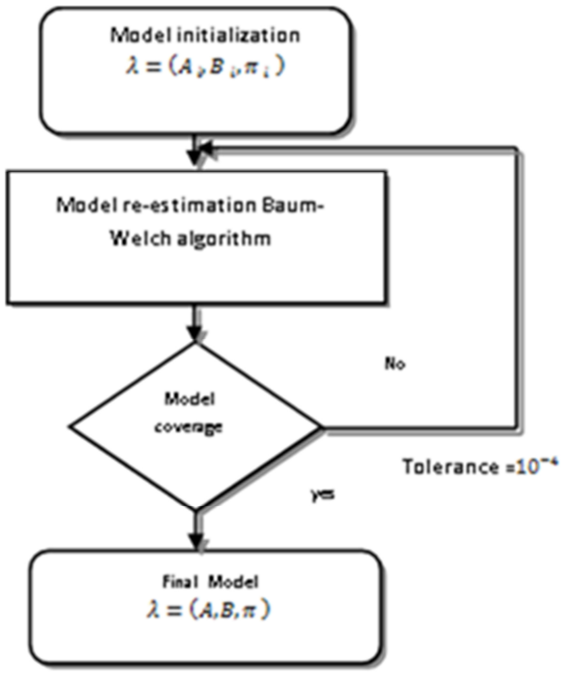

Figure 2. Flowchart of training algorithm

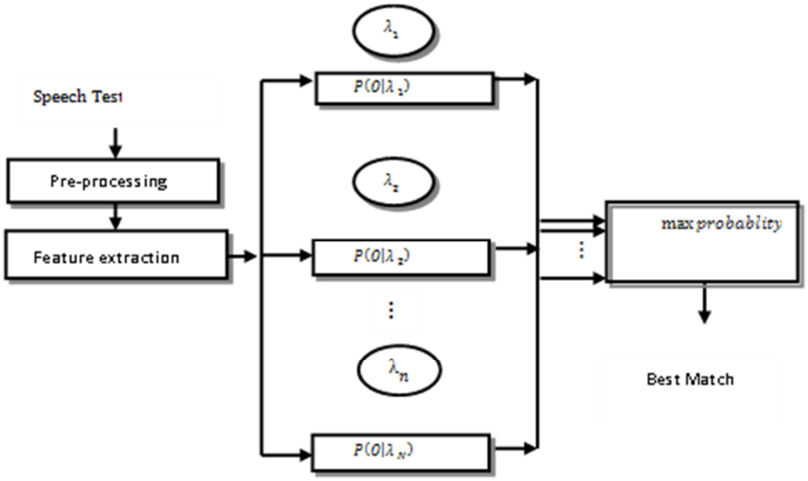

Figure 3. Flowchart of classification algorithm