Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.7, No.1, 2016

Performance Evaluation of Voice Classifier Algorithms for Voice Recognition Using Hidden Markov Model

O.O Adeosun and A.O Folowosele

Department of Computer Science and Engineering, Ladoke Akintola University of Technology , Ogbomoso, Oyo State, Nigeria

ABSTRACT

This paper provides performance evaluation of K mean and Gaussian mixture algorithms which are voice classifier algorithms for voice recognition using the differences in their recognition , training and testing time as parameter for the evaluation.
The performance evaluation results has shown classification efficiency of K – means & Gaussian Mixture algorithms. In the results, comparing the Average Training time for Kmeans algorithm: (Standard database = 435.6854s, Local Database = 411.4578s) while for Gaussian mixture algorithm : (Standard Database = 454.5678s, Local Database = 424.5673s). Moreover, in the considering the Average Testing time, Kmeans algorithm: (Standard database = 23.7178s, Local Database = 23.7178s) while for Gaussian mixture algorithm : (Standard Database = 25.1271s, Local Database = 20.1271s). For the Average Recognition time, Kmeans algorithm: (Standard database = 0.3388s, Local Database = 0.3388s) while for Gaussian mixture algorithm : (Standard Database = 0.4345s, Local Database = 0.4345s).
Therefore, conclusions could be made that K-mean algorithm is a better classifier for voices in a voice recognition system because it has minimum training, testing and recognition time compared to Gaussian mixture algorithms.
Key Words: Evaluation, Classification, Efficiency, Algorithm, K- means algorithm, Gaussian mixture algorithm, Training, Speaker, Recognition

I INTRODUCTION

Speaker recognition is the identification of the person who is speaking by characteristics of their voices (voice biometrics), this is also known as voice recognition (Jean-Francois Frederic, Loius – Jean, Joseph, Douglas and Ivan, 2003).The term voice recognition refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process. Voice recognition implies only that the computer can take dictation, not that it understands what is being said. (Reynolds and Rose,1995).

Many research work have been done in voice recognition system using Hidden Markov Model (HMM) but no attention has been paid to best algorithms used to classify voices. Therefore in this research work, two algorithms were considered namely, Gaussian Mixture and K – means algorithm. The performance evaluation of both algorithms using the differences in recognition , training and testing time were done.

II REVIEW OF RELATED WORKS

Voice recognition has a history dating back some four decades and uses the acoustic features of voices that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioural patterns (e.g., voice pitch, speaking style). Voice verification has earned voice recognition its classification as a "behavioural biometric". (Pollack, Pickett and Sumby, 1974).

There are two major applications of voice recognition technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity. In a sense voice verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice

print" or "voice model") whereas speaker identification is a 1:N match where the voice is compared against N templates. Voice identification systems can also be implemented covertly without the user's knowledge to identify talkers in a discussion, alert automated systems of speaker changes, check if a user is already enrolled in a system, etc. In forensic applications, it is common to first perform a voice identification process to create a list of "best matches" and then perform a series of verification processes to determine a conclusive match. (Kinnunen, Tomi and Haizhou, 2010).

Each voice recognition system has two phases: Enrolment and verification. During enrolment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print. For identification systems, the utterance is compared against multiple voice prints in order to determine the best match(es) while verification systems compare an utterance against a single voice print. Because of the process involved, verification is faster than identification (Lisa, 2004).

In the long history of voice recognition, both shallow form and deep form (e.g. recurrent nets) of artificial neural networks had been explored for many years during 80's, 90's and a few years into 2000 (Morgan, Bourlard, Renals, Cohen, and Franco, 1993) , (Robinson, 1992), (Waibel, Hanazawa, Hinton, Shikano and Lang 1989). But these methods never won over the non-uniform internal-handcrafting Gaussian mixture model/Hidden Markov model (GMM-HMM) technology based on generative models of speech trained discriminatively. (Baker, Deng, Glass, Khudanpur, Lee, Morgan, and Shaughnessy, 2009). A number of key difficulties had been methodologically analyzed in 1990's, including gradient diminishing and weak temporal correlation structure in the neural predictive models (Deng, Hassanein and Elmasry, 1994). All these difficulties were in addition to the lack of big training data and big computing power in these early days. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning starting around 2009-2010 that had overcome all these difficulties. Hinton et al. and Deng et al. reviewed part of this recent history about how their collaboration with each other and then with colleagues across four groups (University of Toronto, Microsoft, Google, and IBM) ignited the renaissance of neural networks and initiated deep learning research and applications in speech recognition.(Hinton , Deng, Dahl , Mohamed , Jaitly ., Senior , Vanhoucke ., Nguyen , Sainath , and Kingsbury, 2012).

Hamdy K .Elminir, Mohamed Abu Elsoud ,and L.M Abou El – maged (2012) worked on the evaluation of different feature extraction techniques for continuous speech recognition. The main feature extraction techniques they used are Mel – Frequency Cepstral Coeficients (MFCC), Linear Predictive Coding (LPC), Zero Crossing with Peak Amplitudes (ZCPA)..principal Component Analysis (PCA) was used to obtain their results and it was discovered that ZCPA is best techniques as a comparison to other techniques but with their results there are still some problems especially with continuous speech. Also, this was done in non – noisy environment.

Teenu Therese Paul, Shiju George (2013) worked on voice recognition based secure android model for inputting Smear Test Results. In their work, the voice recognition technology was applied into a laboratory information system for identifying each technician's voice. By using the user's voice sample, a secure authentication system was developed where the unique features of the user's voice were extracted and stored at the time of registration. Afterwards during the login stage, unique features of the user's new voice sample are extracted. It was then used to compare the features with all the stored features rather than the just previous one. For this, a unique username is set to all the users. The comparison operation was performed with all voice samples under that particular user name. the voice feature comparison process was done by using Fast Fourier Transform Techniques. After a successful login the user can enter the results of the smear test through his voice rather than typing into the system. The system that was developed consists of two parts , a client system and a server system. The client system was developed using Android and the server system was implemented in Java. A major weakness of this system is that they fail to take into consideration voice changes due to illness , mood e.t.c.

III RESEARCH METHODOLOGY

Each person has biologically distinct features such as fingerprint, face, hand shape, palm, iris and voices. The voice of each person is unique which can therefore be used as biometrics for security purposes. The voice recognition system analyses the voices patterns of an individual to verify the individual's identity. But most biological features (for example fingerprint, hand and face) vary with age. For reliable fault tolerance voice

recognition, it is necessary to develop an efficient method for pre-processing and classifying the different voices. The basic stages that would be involved in this system are:

A Voice Acquisition

Voices was used for the development of this systems will be the voice samples that was collected from people. A total of 140 voices was collected and saved into the database. 50 voices was used for training and 20 voices will be used for testing.

B Feature Extraction

Mel-Frequency Cepstral Coefficient (MFCC) Computation was used to extract the voices. In a broad sense, feature extraction aims for data reduction by converting the input signal into a compact set of parameters while preserving spectral and/or temporal characteristics of the speech signal information.

C Training the Voices

Viterbi algorithm was used in training the voices. Viterbi algorithm will be used for the recursive procedure of how the voices will be maximised. It has the following steps

D Classification Methods

Once you have produced the feature vectors, the next task is classification, that is to build a unique model for each speaker in the database. The speech produced by the speaker whose identity is to be recognized, will be compared with all speaker's models in the database. Then, the speaker identity will be determined according to a specific algorithm. In these research work two algorithms are going to be used to classify our voices, the reason is because these are the two major algorithms that are been used for classifying voices but they have not been compared before. The two algorithms includes:

1. Voice Classification Using The K – Means Algorithm

K-means algorithm partitions the T feature vectors into M centroids. The algorithm first randomly chooses M cluster-centroids among the T feature vectors. Then each feature vector is assigned to the nearest centroid, and the new centroids are calculated for the new clustres. This procedure is continued until a stopping criterion is met, that is the mean square error between the feature vectors and the cluster-centroids is below a certain threshold or there is no more change in the cluster-center assignment.

2. Voice Classification Using The Gaussian Mixture Algorithm

The pattern matching is probabilistic (evaluating probabilities) and results in a measure of the likelihood, or conditional probability, of the observation given the model. Here, a certain type of distribution is fitted to the training data by searching the parameters of the distribution that maximize some criterion.

E Feature Matching

The feature matching will be done using vector quantization. In the recognition phase an unknown speaker, represented by a sequence of feature vectors {x1,x2,….,xT}, will be compared with the codebooks in the database. For each codebook a distortion measure is computed, and the speaker with the lowest distortion is chosen,

$$C_{best} = \underset{1 \le i \le N}{argmin} \{ s(X, C_i) \} \quad \text{.....} \qquad \text{equation 1}$$

One way to define the distortion measure, which is the sum of squared distances between vector and its representative (centroid), is to use the average of the Euclidean distances:

$$s(X, C_i) = \frac{1}{T} \sum_{t=1}^{T} d(\boldsymbol{x}_t, \boldsymbol{c}_{min}^{i,t}) \qquad \text{equation 2}$$

The well known distance measures are Euclidean, city distance, weighted Euclidean and Mahalanobis. Euclidean distance will be used in this work.

Where **c**min denotes the nearest codeword**x**t in the codebook *Ci* and d(.) is the Euclidean distance. Thus, each feature vector in the sequence *X* is compared with all the codebooks, and the codebook with the minimized average distance is chosen to be the best.

The Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn),

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$ equation 3

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

F HMM Generator of Observations

Given appropriate values of *N*, *M*, **A**, **B**,π and an alphabet the HMM can be used as a generator to produce an observation sequence $\mathbf{O} = o_1 o_2 o_T \in V^*$ by performing the following steps:

I. Choose an initial state $q_1 = i$ according to the initial state distribution π. Set $t = 1$.
II. Choose $o_t = v_k$ according to $b_j(k)$.
III. $t = t + 1$, transit to a new state $q_{t+1} = j$according to a_{ij}.
IV. Return to step 3 if $t<T$.

Recognition of voices with a Hidden Markov model works will work this way; An observation sequence from some process we want that we want to study, a speech signal out of the database represented by a sequence of vectors, and having a number of models that represent the things what to recognize, for example the words that can be spoken. Then the next thing is to know the probability that the observation sequence was produced by the model λ.

Now, having an observation sequence, it is not clear how the model generates that sequence, because the underlying state sequence is hidden. Since practically any state sequence could generate each observation sequence there is no correct state sequence to be found here. All that will done is try to solve this problem as best as possible and seek the most likely state sequence given the observation sequence and the model.

With this having a model available, an important question is of course how to obtain such a model. This is called the training problem, since the model parameters of a HMM can be obtained from a set of example data.

Decoding: Decoding is to find the single best state sequence, $Q = (q_1, q_2,\ldots,q_T)$, for the given observation sequence $O = (o_1, o_2, \ldots,o_T)$. Consider $\delta_t(i)$ defined as

$_t(i)$ max $P[q_1,q_2.....q_t \quad i,o_1,o_2.....o_t \mid]$
$(q1,q2.....qT \quad 1)$

that is $\delta_t(i)$ is the best score along single path at time *t*, which accounts for the *t* observations and ends in state *i*. by induction,

$_i \ _1(j)$ [max $_t(i)a_{ij}]b_j(o_t \ _1)$

Recognition and Training Time Assessment

The first step before assessing the two algorithms for classification is to build a speaker-database Cdatabase = { C1 , C2 , …. , CN} consisting of N codebooks, one for each speaker in the database. This is done by first converting the raw input signal into a sequence of feature vectors X={x1, x2, …. ,xT}. These feature vectors are clustered into a set of M codewords C={c1, c2, …. , cM}. The set of codewords is called a codebook .

This system was tested with a total of 140 sample audio signals with seven samples for each class. A Class consists of seven voices belonging to a subject. The system works 20 Classes with voices per class.

For the recognition and training time assessment to be done, it will follow the various steps

1. Load the database created.
2. Train the database.
3. Test the database.

After the three steps have been achieved, now calculating the training and testing time for Gaussian Mixture Algorithm and K – Means Algorithm, thereby comparing to know which one has a better performance in terms which training and testing time is relatively low.

IV RESULTS

Table 1 Comparison Between Gaussian Mixture And Kmeans of Training of Sampling Frequency

Sampling Frequency	Gaussian Mixture	Kmeans
20000	-0.14937	-0.231149
19000	-0.117733	-0.297271
18000	-0.0573784	-0.225666
17000	0.0628527	-0.135463
16000	0.089878	-0.166473
15000	0.200797	-0.0601331
14000	0.330616	0.228493
13000	0.506557	0.388194
12000	0.405279	0.328512
11000	0.702558	0.614919
Avg mean Value	**0.1974**	**0.0444**

It can be found that classification with K means gives a better classification results as its values tend to be the minimum than that of the Gaussian mixture model.

Table 2 Comparism Of The K Means And Gaussian Mixture In Terms Of Average Training And Testing Time.

DATABASE	TRAINING TIME (Gaussian Mixture) (Secs)	TESTING TIME (Gaussian Mixture) (Secs)
Standard Database	454.5678	25.1271
Local Database	424.5673	20.1271

DATABASE	K – MEANS ALGORITHM	GAUSSIAN MIXTURE ALGORITHM
Standard base	0.3388	0.4345
Local base	0.3388	0.4345

Table 3 Showing Average Recognition Time for K- means and Gaussian mixture Algorithms

DATABASE	TRAINING TIME (K means) (Secs)	TESTING TIME (K mean) (Secs)
Standard Database	435.6854	23.7178
Local Database	411.4578	23.7178

From the result above it can be seen that, Kmeans has less training, testing and recognition time than Gaussian Mixture model.

Therefore, it can be concluded that Kmeans is a better classifier than Gaussian mixture model, because the average training, testing and recognition time is lowered compared to Gaussian mixture model.

V CONCLUSION AND RECOMMENDATION

A Conclusion

The main purpose of the centre for the improvement the voices are been classified using several algorithms. Therefore, evaluation of K-means and Gaussian mixture model which are used as classifier for the voices was examined. It was found out that K-means is a better classifier in terms of the training, testing and recognition time which tends to be minimum compared to the Gaussian Mixture Model.

B Recommendation

After a couple of months for this research work, the following recommendations are been made both for other to work on this project and also for future purpose:-

i. It is also recommended that out of K – means and Gaussian Mixture algorithms that are used to classify voices, K – means is a better classifier than Gaussian Mixture algorithm because of it's minimum training, testing and recognition time.

VI REFERENCES

Baker, J.; Deng L.; Jim, G., Khudanpur, S.; Lee, C.H.; Morgan, N. and O'Shaughnessy D. (2009). "Research Developments and Directions in Speech Recognition and Understanding, Part 1," IEEE Signal Processing Magazine, vol. **26**(3): 75-80.

Hamdy K. E, Mohamed A. E, and Abou El-Maged L.M. , (2012)"Evaluation of predictive model with applications to speech recognition, Journal on Springer Science & Business Media". vol**. 7**(2.):331-339

Jean-Francois, B.; Frederic, B.; Louis-Jean, B.; Joseph, P. C.; Douglas, A. R.; Ivan,M.C (2003).["http://www.iscaspeech.org/archive/eurospeech_2003/e03_0033.html""http://www.iscaspeech.org/archive/archive_papers/eurospeech_2003/e03_0033.pdf" "Person Authentication by Voice: A Need for Caution"] (PDF). *Person Authentication by Voice: A Need for Caution*. 8th European Conference on Speech Communication and Technology. Geneva, Switzerland: isca-speech.org.

Kinnunen, Tomi; Li, and Haizhou (2010). "An overview of text-independent speaker recognition: From features to super vectors". Speech Communication Vol **52** (1), 12–40.

Lisa M. (2004). "An Exploration of Voice Biometrics "SAN institute, 1 - 17

Morgan, Bourlard, Renals, Cohen, and Franco (1993) "Hybrid neural network/hidden Markov model systems for continuous speech recognition. ICASSP/IJPRAI" New York Times, 18 – 25.

Pollack, Pickett, and Sumby (1974). "Experimental phonetics". MSS Information Corporation, 251–258.

Robinson T. (1992) A real-time recurrent error propagation network word recognition system, ICASSP, 20 – 25.

Teenu Therese Paul, Shiju George (2013) Voice Recognition Based Secured Model *International Journal of Engineering Sciences & Emerging Technologies* .ISSN: 2231 – 6604. Volume 6, Issue 3, pp: 344-351.

Waibel, Hanazawa, Hinton, Shikano,and Lang. (1989) "Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech and Signal Processing."