VOICE RECOGNITION USING ARTIFICIAL NEURAL NETWORKS AND GAUSSIAN MIXTURE MODELS

AARON NICHIE
Computer Engineering Department, University of Ghana,
P. O. Box LG 25, Legon, Accra, Ghana
aaronichie2000@gmail.com

GODFREY A. MILLS
Computer Engineering Department, University of Ghana,
P. O. Box LG 25, Legon, Accra, Ghana
gmills@ug.edu.gh

Abstract:

The ability of recognition systems to correctly recognize speakers based on their speech waveform distribution depends largely on how the recognition system can train the model parameters so as to provide the best class of discrimination. This paper presents the results of an effort to recognize the voice of individual speakers based on their continuous speech waveform distribution using the combined frameworks of artificial neural networks (ANN) and statistical Gaussian mixture models (GMM). A feed-forward multilayer ANN architecture with 30 hidden neurons was implemented for discriminative classification and training and the statistical GMM model computed scores that were transferred to best match the speech features. The decision system determines the recognized speakers using correlation coefficient analysis to measure the goodness of match of speech feature frames of the detected speaker from the ANN and GMM frameworks. To validate performance of the system, experiments were conducted using speech utterances from 30 different speakers (20 males and 10 females). System performance showed average recognition rates of 77% for 5-word utterances and 43% when the lengths of the utterances were increased to 20-word utterances for cases of trained speech utterances. With unknown utterances, recognition rate of 18% achieved for 20-word utterances.

Keywords*:* voice recognition, artificial neural networks, Gaussian mixture model, cepstral coefficients.

1. Introduction

Voice recognition of speakers by systems is the problem of converting the information content of the speech waveform of speakers into identifiable sets of features that carry all the possible discriminative information necessary for recognition of the speakers. The ability of a recognition system to adequately recognize the voice of speakers essentially depends on the adequate capture of the time frequency and energy of the speech waveform and how well the recognition model parameters are trained to produce the best sets of discriminations so as to achieve accurate recognition. With the advent of technology, the idea of using the voice signals for the purpose of identification has found many useful applications in platforms such as access control of information, access to banking services, secured database access system, remote access to telephone services, avionics, and automobile systems, etc., [1-3]. Although many accomplishments have been demonstrated especially for isolated words, recognition based on continuous speech signals still remains an area that has gained considerable attention due to the fact that continuous speech signals depict natural flow of words [4-5]. Hence, recognition of speakers based on continuous speech signals may be useful for applications that require detection of speakers in natural conversation. Unlike isolated word recognition system where the words of the utterances are characterized by pauses, continuous speech signals do not have such pauses and this compels the recognition task to predict where each of the words in the utterances ends and the others begin so as to produce the correct utterance. In this regard, possible errors which may arise as a result of the length of utterances may broaden the variance of the class distribution of the speaker which may lead to increased classification error and subsequently, affect the recognition accuracy of continuous speech signals.

Many algorithms and approaches have been used over the past years for the recognition of speech patterns [7-21] and the most commonly used algorithm is the hidden Markov model (HMM) which has been shown to have high recognition performance. Ramesh et al [9] for example, demonstrated recognition rates of 92-94.5% using HMM for isolated words (numbers) and in the work of Katagiri et al [10], recognition rates of 67-83% were reported for isolated-word recognition. Despite its widespread use in speech recognition technology, the standard HMM algorithm has been shown to exhibit poor discriminative learning due to the training algorithm and to make up for this deficiency, various hybrid solutions have been proposed to increase the discriminative classification power [12]. The viability of machine learning approaches such as artificial neural networks (ANN) has also been explored as a useful technology to assist in statistical speech recognition [13-17] due to its discriminative and adaptive learning properties. The capability of ANN has been demonstrated in many aspects such as isolated-word recognition, phoneme classifier, and as probability estimator for speech recognizers [13, 14]. Much as ANN offers high discriminative training especially for short-time speech signals as in isolated-words, it also has issues with adequately modeling of temporal variations of long-time speech signals. Hybrid solutions that draw on the strengths of the ANN and HMM frameworks have also been demonstrated in speech recognition technology. In the work of Trentin et al. [20], a hybrid ANN/HMM architecture was used to achieve speech recognition rate of 54.9-89.27% with corresponding SNR of 5-20dB for isolated utterances. Reynolds et al., have also employed the statistical Gaussian mixture models (GMM) framework, which is a variant of HMM, to achieve speech recognition rate of 80.8-96.8% [21, 22] in a speaker-independent system using isolated utterances. Though there are considerable research activities in continuous speech recognition, most of the activities are concentrated on either correct detection or recognition of words or their positions in the utterances for example to detect when a speaker has used a word that is not in the vocabulary of a continuous speech. The goal of this work however, is to identify speakers using their continuous speech voice waveform distribution of utterances. This may be particularly useful for application systems that require detection of speakers in natural conversation environments such as forensic and security activities.

In this work we combine the frameworks of ANN and the GMM to implement voice recognition of speakers. The combined paradigm explores the discriminative and adaptive learning capabilities of the ANN and the capabilities of the GMM to model underlying properties and offer high classification accuracy while still being robust to corruptions in the speech signals for the recognition task. The performance of the recognition system is evaluated using variable lengths of speech utterances that are known and unknown to the system.

2. Overview of the recognition system

The architectural diagram of a typical voice and speaker recognition system is shown in Figure 1. The system is trained to recognize the voice of individual speakers with each speaker providing specific sets of utterances through a microphone terminal or telephone. The captured analog voice waveform has three components: speech segment, silence or non-voiced segment, and background noise signals. To extract the relevant speech signals, the voice waveform is digitized and signal processing is carried out to remove the noise signals and the silence or non-voiced components. Any relevant information that is ignored during this processing is completely considered as lost and conversely, any irrelevant information such as fundamental frequency of the speaker and the characteristics of the microphone that is allowed to pass is treated as useful with implications on the speech feature classification performance. The extracted speech signals are then converted into streams of template feature vectors of the voice pattern for classification and training. In the event that irrelevant information is allowed, then the speech features that may be generated from the corrupted speech signals may no longer be similar to the class distributions that are learned from the training data. The system recognizes the voice of individual speakers by comparing the extracted speech features of their utterances with the respective template features invoked from the training systems. The GMM recognizer computes scores that are used for the matching of the most distinctive speech features of speakers. The decision criteria for the voice recognition of speakers were based on correlation analysis of the speech features of speakers from the ANN and GMM.

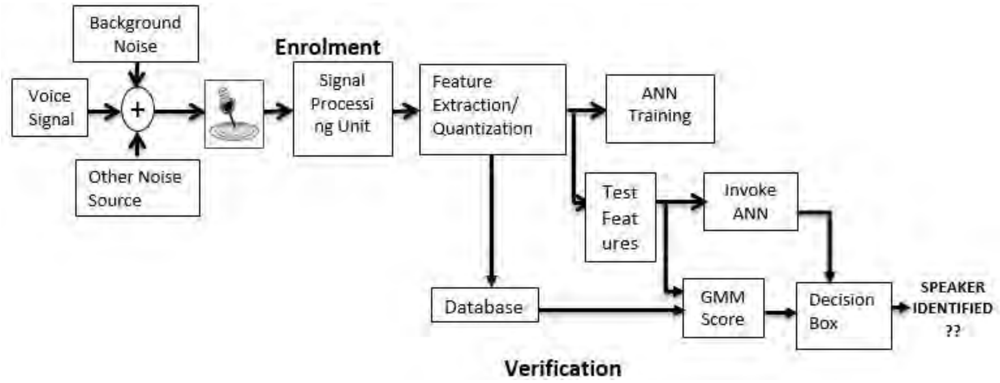


Figure 1 Architectural diagram of the voice recognition system

2.1. *Voice signal processing and feature extraction*

We denote the analog speech signals of a spoken utterance at the microphone terminal as $g(t)$ and the N-sample discretized speech signal sampled at a rate of 16 kHz and 16 bits quantization as described in equation (1). As a consequence of the digitization operation, we obtain a signal with 16,001 sample points and a digital data stream of 256 kbits/s for processing by the recognition system. Figure 2 shows the voice signal of the utterance "*all graduating engineering students must be seated in the great hall of the university by 11:30am for the chairman's address*" obtained from two speakers (male and female).

$$g[n] = g_0, g_1,,,,,,,,,,,,,,,,,,,,, g_{N-1} \qquad (1)$$

The captured speech signals of the speaker contain background noise signals which come from different sources such as ambient noise, microphone terminal, and the communication channel. To remove the low background noise signals, the discretized speech signal of the spoken utterance is passed through a second-order Butterworth IIR (infinite impulse response) highpass digital filter for the filtering (*g[n]*h[n]*). The choice for this IIR filter is based on its flexibility to easily manage signals that are heavily nonlinear in nature due to its nonlinear phase characteristics and its flexibility of requirements for meeting of constraints such as arbitrary response. Since most of the energy content of the voice signal is concentrated within the low frequency range, the filter cut-off frequency is set at 0.40 kHz. The transfer function of the filter $h[n]$ is described by equation (2).

$$H(z) = \frac{1 - 2z^{-1} + z^{-2}}{1 - 0.5979z^{-1} + 0.2355z^{-2}}\,. \qquad (2)$$

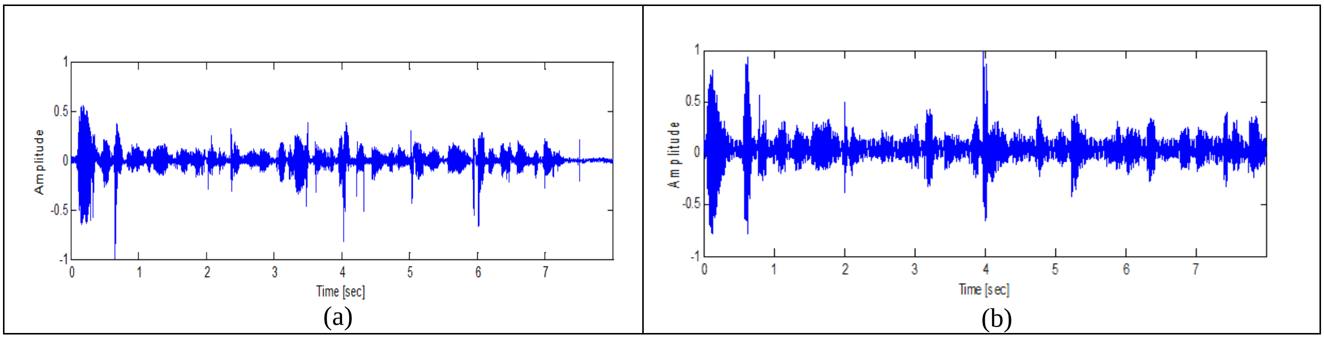
The digital filter is designed using the Matlab filter design and analysis (FDA) toolbox and the following design specifications: sampling rate = 16 kHz, passband frequency = 3.5 kHz, stopband frequency = 0.3 kHz, stopband attenuation = 60 dB, and passband ripple = 0.015 dB. The amplitude of the filtered voice signal is normalized to unity in accordance with equation (3) to better gauge the variations in the signal.

$$x[n] = \frac{g'[n]}{\max|g[j]'|} \qquad 0 \le j \le N-1, \qquad 0 \le n \le N-1, \qquad (3)$$

To remove the silence or non-voiced components in the spoken utterance, the processed continuous speech signal is divided into overlapping frames so as to enable signal distribution in each frame to be independently analyzed and represented appropriately by a feature vector. The number of frames N_S is estimated as:

$$N_S = floor\left(\frac{r - N + M}{M}\right) + 1, \qquad (4)$$

where, r is the length of the speech signal, N is the number of samples per frame, and M is the separation between consecutive frames where $M < N$. Thus, if the first frame consists of the first N sample, then the second frame begins M samples after the first frame and the overlaps it by N-M samples. For this work, a frame size of 30ms (480 samples) was obtained every 10ms (160 samples) which represents 34% of the frame size. The signal distribution in each frame is also assumed to have stationary characteristics for easy analysis and representation.



(a) (b)

Figure 2 Speech signal of the utterance "all graduating engineering students must be seated in the great hall of the university by 11:30am for the chairman's address" from (a) male speaker (b) female speaker.

To minimize signal leakages from one frame to the other, each framed voice signal $x[m,n]$ is passed through a Hamming window function as described in equation (5) to taper the edges of the signal in the frame smoothly to zero. Figure 3 shows a plot of the clean speech signal component of the voice waveform of the male speaker shown in Figure 2 (a) with all the background noise signals and the silence components removed.

$$y[m,p] = x[m,p] * \left\{ 0.54 - 0.46\cos\left(2\pi \frac{p}{N-1}\right)\right\}, 0 \le p \le N-1, \quad 0 \le m \le N_S - 1. \qquad (5)$$

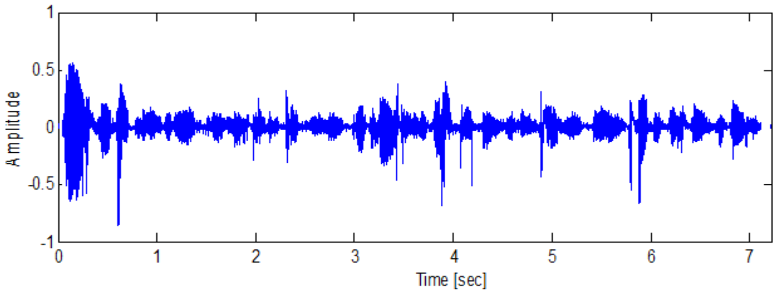


Figure 3 Processed utterance of a male speaker with suppressed noise and silence components.

Following suppression of all the irrelevant information in the voice signals, the clean speech is then converted into streams of feature vectors coefficients containing only the information about the spoken utterance that is required for its recognition. The extraction of the feature vectors of the speech signal may be carried out using either time domain analysis or spectral domain analysis. Whilst the time domain is more straightforward and takes less calculation time, the spectral domain approach is very effective but requires more computational time since it involves frequency transformation. We however used the well-established spectral analysis technique *of mel* frequency cepstral coefficient (MFCC) algorithm [24, 25] for the feature vectors extraction due to its robustness. Figure 4 shows the MFCC flow process for the feature vectors coefficients extraction of the speech signal. The windowed speech signals are first converted to the frequency domain using the discrete cosine transform (DCT) to allow for the energy compaction of the spectral data in the lower order coefficients. The magnitudes of the spectral distribution are then passed through a *mel* filter bank to obtain the Mel-domain spectral distribution. The power spectral distribution of each frame at the output of the filter are then smoothened using a logarithm function followed by computation of inverse DCT operation to convert the *mel*-domain spectral distribution into the time domain *mel* coefficients *C(i,j)* as:

$$C(i,j) = \frac{1}{m}\sum_{k=0}^{m-1} \log\{E(i,k)\}\cos\left[(k-0.5)\frac{j\pi}{m}\right], \qquad 0 \le j \le m-1, \qquad 0 \le i \le N_S - 1, \quad (6)$$

where *m* is the number of filters in the *mel* filter bank, and *E*(.) is the power spectrum. The *mel* filter is computed using the following parameters: sampling rate = 16 kHz, minimum frequency = 10 Hz, and maximum frequency = 8 kHz, and number of filters *m* = 30. The *melcepst* of the Voicebox Matlab toolbox was used for the feature extraction computation [26]. A feature vector of 12 dimensional MFCC was extracted from each frame of the speech utterance. Since the speech feature space was quite large for the computations this was reduced to a small set of representative feature vectors (codebook) sufficient to adequately describe the extracted speech features of the utterance. The vector quantization (VQ) algorithm [27] was used for the feature space reduction.

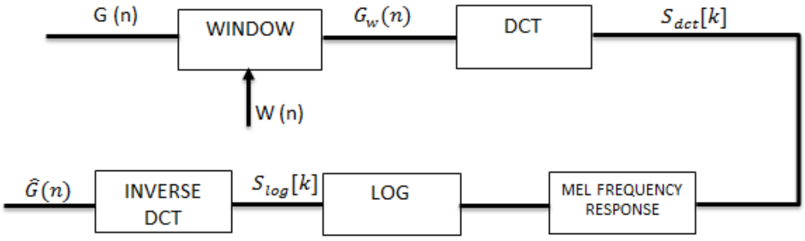


Figure 4 Block diagram of the MFCC operation for the voice feature data

2.2. *Neural network architecture and training*

The multilayer feed-forward ANN architecture with supervised training of the extracted MFCC feature vectors of the utterances was implemented using the Matlab platform. The ANN is trained using the back propagation algorithm, which has been demonstrated [13, 14] as effective in achieving minimization of recognition error rates. The ANN architecture plays a critical role in achieving a good and desirable result and this depends largely on the number of neurons at the input and hidden layers and the number of layers for the task. To achieve the required ANN architecture, we started with the basic structure of single input and hidden layer neurons in a three layer system and varied the number of neurons in each layer through training of the network till the optimal number that gave the best training was achieved. A variety of utterances from five speakers were used for testing the ANN architecture. The extracted feature vectors of each of the utterances were presented at the input layer and the most probable speech and speaker identified at the output layer. Since the back propagation method usually requires provision of a target value that is used during training, which for speech recognition, is usually not available, we therefore set the target output to “1” for the correct speech signals (utterance) and “0” for others. This somehow reinforced the correct output and weakened the wrong output. The network was trained by adjusting the values of the connection weights between the network elements in accordance with equation (7) [14, 18]:

$$w_{ij}(k+1) = w_{ij}(k) + \alpha\left\lfloor -\partial E / \partial w_{ij} \right\rfloor + \beta\left\lfloor w_{ij}(k) - w_{ij}(k-1)\right\rfloor, \qquad (7)$$

where $w_{ij}(.)$ is the weight matrix between the layers, E is the mean-square-error function for the training, β is the momentum factor, and α is the learning rate which represents the speed of learning process. An important consideration in the ANN design is the choice of learning rate, momentum factor, and number of epochs for the network training. These parameters were varied and their effects were observed by repeating testing utterances for each variation. For example, when a smaller learning rate was used, the algorithm took significant time to converge due to the gradual learning process. Conversely, when a larger learning rate was used, the algorithm diverged due to the acuteness in the learning process. The learning rate was therefore tuned for stable learning by updating the learning rate after each training phase in accordance with the expression below and an update value φ (from 1% – 20%):

$$\alpha(k) = \alpha(k-1) * (1 - \varphi / k)\ . \qquad (8)$$

Following a series of experiments on the variation of neurons in the layers and the different sets of utterances, ANN architecture with 20 neurons at the input layer and a hidden layer with 30 neurons was found as suitable for the task. Table 1 presents the summary of results of variation of hidden layer neurons and the average recognition rates and times for different utterances (1-word to 5-word utterances). This ANN architecture was adopted for the voice recognition system.

Table 1 Results of data sets for the optimal ANN architecture design.

Test utterances	Hidden layer neurons	Recognition rate (%)	Recognition time (s)

transform;
aaron nichie;
signal processing;
voice recognition system;
move volume control up and down;
	10	40.91	0.60

	15	45.45	1.00
	20	59.00	2.00
	25	72.73	2.00
	30	77.27	4.00
	35	77.27	6.00

2.3. *GMM recognition model development*

The GMM for speaker recognition is formulated to compute the GMM parameters of extracted feature vectors of spoken utterances that best match the speech feature templates in the system database. There are several techniques that may be used to estimate the parameters of the GMM (mixture weights, mean vector, covariance matrix) that describe the component distribution of the extracted speech feature vectors. So far, the most popular and established method is the iterative expectation maximization (EM) algorithm [21, 28, 29] which is used to obtain the maximum likelihood (ML) estimates. In the development of the GMM recognizer model, we used a one-state HMM with multiple Gaussian distributions describing the single state so as to conform to the ANN model and also to enable the model capture more variations in the speech signal. It is also possible to model the GMM as a multi-state HMM with a single Gaussian distribution describing each state. The governing equation for the GMM recognizer model of a speaker is expressed as [21]:

$$P(x) = \sum_{k=1}^{M} \beta_k\, f(x, \mu_k, C_k)\,, \qquad (9)$$

where M is the number of Gaussian components, x is the N-dimensional MFCC speech feature vector (x_1, x_2, ,,,,, x_N) β_k (k = 1, ..., M) is the weights of the mixture component k, and $f(.)$ is the multivariate Gaussian with mean vector μ_k and covariance matrix C_k, where each Gaussian is given by:

$$f(x_n, \mu_k, C_k) = \frac{\exp[-0.5(x_n - \mu_k)^T C_k^{-1}(x_n - \mu_k)]}{(2\pi)^{N/2} \left|C_k\right|^{1/2}}, \qquad 1 \le n \le N\ . \qquad (10)$$

To estimate the GMM parameters for a given set of the *N*-dimensional MFCC feature vectors an utterance, we first organized the data into a number of cluster centroids (example 256) using K-means clustering technique and the cluster centroids are further grouped into sets of 32 which are then passed to each component of the GMM. The EM algorithm is then employed to obtain the ML distribution parameter estimates. The iterative procedure first computes estimation of the current iteration values of the *k-th* Gaussian component for the next iteration using equation (11) followed by a maximization operation where the predicted values are then maximized to obtain the real values for the next iteration based on equation (12).

$$y(k,t) = \frac{\beta_k^i f(x, \mu_k^i, C_k^i)}{\sum_{j=1}^{M} \beta_j^i f(x, \mu_j^i, C_j^i)}. \qquad (11)$$

$$\mu_k^{i+1} = \frac{\sum_{t=1}^{T} y(k,t)x}{\sum_{t=1}^{T} y(k,t)}; \qquad \beta_k^{i+1} = \frac{1}{T}\sum_{t=1}^{T} y(k,t); \qquad C^{i+1}(k,j) = \frac{\sum_{t=1}^{T} y(k,t)\left(x_j - \mu_{k,j}^{i+1}\right)^2}{\sum_{t=1}^{T} y(k,t)}, \qquad (12)$$

To establish the number of Gaussian distributions useful for the GMM speaker model for the recognition task, we performed experiments using varying number of Gaussian distributions and MFCC feature vectors from the utterance "*increase volume upwards*" for two speakers. Table 2 shows the results obtained of recognition rates with varying number of Gaussians. Based on the experimental results, 20 Gaussians was found adequate for the GMM speaker models.

Table 2 Number of Gaussians and recognition rates

No. of Gaussians	Recognition Rate
3	21%
5	43%
12	77%
20	79%
25	79%
30	79%

Following estimation of the GMM for the utterance of each speaker, recognition is achieved by comparing the sets of extracted MFCC feature vectors of testing utterances against the GMM speaker models to measure a matching score that best maximizes the log-likelihood value. To compute this score, we used the log-likelihood (LL_{GM}) measure defined in equation (13), where $P_{GM}(x)$ is the trained GMM speaker model for trained feature vectors $\{x\}$ and $\{x_t\}$ is the observation or testing feature vectors of utterance to the trained GMM speaker model for evaluation on how well the GMM recognizer model fits the given observation data.

$$LL_{GM} = \frac{1}{T}\sum_{t=1}^{T} \log\left[P_{GM}(x_t)\right]. \qquad (13)$$

Following the computations, the GMM model that results in the highest LL_{GM} value is determined as the recognized speaker because that model gives the best probability of producing the same speech features of the spoken utterance of the speaker. Once the speaker is identified, the features of the selected speaker is extracted from the database and submitted to the decision for verification.

2.4. *Recognition decision system*

The decision task for the recognition of speakers from the combined frameworks of ANN and the GMM is based on finding the correlation coefficient measure as described in equation (14). This evaluates the goodness of match by comparing the recognized feature of the speech frames of the detected speaker from the ANN and GMM to measure the degree of similarity. This scheme takes advantage of the differences in the training abilities of the ANN and the GMM to find the maximum probability of recognition of speakers. If the MFCC feature vectors of a testing utterance of a speaker is X_n (n = 1, 2, .., N) and the GMM results in a speech feature vectors X_{GMM} for the detected speaker and that of the ANN identifies a speech feature vector X_{ANN}, for the detected speaker, then the correlation measure ρ is estimated as:

$$\rho = \frac{\sum_{k=1}^{N}\left(X_{GMM,k} - E[X_{GMM}]\right)\left(X_{ANN,k} - E[X_{ANN}]\right)}{\sqrt{\sum_{k=1}^{N}\left(X_{GMM,k} - E[X_{GMM}]\right)^2}\;\sqrt{\sum_{k=1}^{N}\left(X_{ANN,k} - E[X_{ANN}]\right)^2}}, \qquad (14)$$

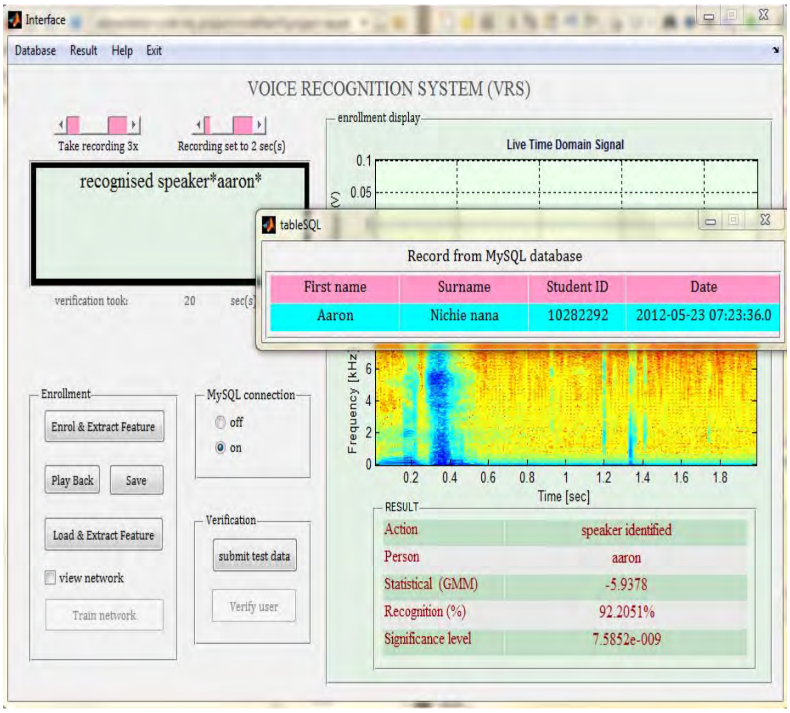
where *E*[.] denotes the mean value. Once there is a strong degree of correlation value between the outputs of the two frameworks within a good significance level of less than 5%, the decision system considers the speaker as recognized and the name of the recognized speaker is extracted from the enrollment database and displayed on the graphic user interface as in Figure 5 below.

3. Experiment and results

We first created two databases; a local database and an external MySQL database. The local database was used to store user features, user name, codebook, and the trained neural network. The external database implemented in MySQL was used to store other information about the speakers so that additional information about a speaker could be retrieved from the database after recognition. We captured and enrolled the utterances of 40 speakers (students of the university) comprising 25 male speakers and 15 female. The utterances of the speakers were captured through a standard headset microphone (Enet rated at 808 mV) with 32 Ω impedance and 105 dB sensitivity in a laboratory environment with relatively moderate noise (not perfect as desired). Figure 5 shows the user graphic interface system in Matlab platform for the speech data streaming and processing for recognition. For the purposes of uniformity, a fixed recording time of 30s (adequate for slow speakers) was set for capturing of the utterances. Each of the 40 speakers was first made to utter (read out) the same four different utterances (English words) comprising two of 5-word utterances such as "*move volume control to left*" and two of 20-word utterances. The choice for the 5-word utterance was based on the fact speech can generally be fast and speakers on the average can speak at 200 words per minute. Each speaker repeated the spoken utterance four times at different instances of time from February to April 2012. These utterances were used for training of the system. Following the training, testing was done in real-time environment to recognize the voice of 30 speakers (20 males and 10 females) whose speech waveform patterns have been trained. Each of the 30 speakers was made to provide two utterances comprising one 5-word utterance and 20-word utterance that are known to the system, followed by another two utterances comprising one 5-word utterance and 20-word utterance that are not known to the system at different times from April to May 2012. Each speaker repeated the utterance four times and the average recognition rate was estimated using:

$$RR = \frac{NCR}{NTU} \quad x \quad 100\ , \qquad (15)$$

where *NCR* represents the total number of correct recognition of a speaker from the system and *NTU* is the total number of spoken utterances presented to the system for recognition. Tables 3 and 4 below show the summary results for 10 speakers for the case of 5-word utterances. It can be observed that speaker 1 was successfully recognized by the system as Aaron in all the 4 attempts for a specific spoken utterance with an average correlation coefficient of 98.7%. In the case of speaker 2, the first attempt resulted in false rejection but the speaker was subsequently recognized as Yonny in the remaining with an average correlation coefficient of 92.3%. With speaker 3, both first and second attempts resulted in false rejection but the speaker was recognized in the third attempt whilst the fourth attempt resulted in false acceptance.



recognised speaker*aaron*

verification took: 20 sec(s)

MySQL connection: ○ off ● on

☐ view network

Record from MySQL database

First name	Surname	Student ID	Date
Aaron	Nichie nana	10282292	2012-05-23 07:23:36.0

RESULT

Action	speaker identified
Person	aaron
Statistical (GMM)	-5.9378
Recognition (%)	92.2051%
Significance level	7.5852e-009

Figure 5 User interface for the voice recognition system

Table 3 Test results for voice recognition of speakers using 5-word utterances

Speakers	**1st Attempt**	**2nd Attempt**	**3rd Attempt**	**4th Attempt**
Aaron	Recognition	Recognition	Recognition	Recognition
Yonny	False Rejection	Recognition	Recognition	Recognition
Adjoa	False Rejection	False Rejection	Recognition	False Acceptance
Barbara	Recognition	Recognition	False Acceptance	Recognition
Louis Mark	Recognition	Recognition	Recognition	Recognition
Naa Kai	Recognition	Recognition	Recognition	Recognition
Rockson	Recognition	Recognition	Recognition	Recognition
Sarpong	False Acceptance	False Acceptance	False Rejection	False Acceptance
Fiawoo	Recognition	Recognition	Recognition	Recognition
Nelson	Recognition	Recognition	Recognition	False Acceptance

Table 4 Correlation coefficient evaluation based on 5-word utterances

Speakers	**1st Attempt**	**2nd Attempt**	**3rd Attempt**	**4th Attempt**
Aaron	99.24%	98.22%	98.66%	98.77%
Yonny	32.56%	99.49%	89.19%	88.33%
Adjoa	24.23%	33.75%	89.20%	79.75%

A summary of the average speaker recognition rates for the 30 speakers used in the testing of the system for the 5-word and 20-word utterances are shown and Figure 6 below. Figure 6(a) shows that with the 30 testing speech samples the system is able to adequately recognize the voice of the speakers at a success rate of 77% with false acceptance and false rejection rates of 9% and 14% respectively, when 5-word utterances similar to the trained data sets were used. It is possible for the recognition accuracy to be improved further if more training data is used. Since the characteristics of the microphone plays important role in the quality of the recognition accuracy, it is likely that the measured performance value could be improved. The recognition rate of 77% however, reduced to 43% when the length of the utterances was increased to 20-word utterance as depicted in Figure 6(b). The reduction for may be attributed to the increased complexities of the large-sized utterances which somehow affected the learning of relationships. Also the issue of level of mismatch associated with the training and testing of the large-sized speech patterns tend to be significant due to increased level of variabilities which affects the density distribution and subsequently the recognition rates. The results in Figure 6(c) on the other hand show recognition performance for the case of 20-word utterances that are unknown to the training system. The speaker recognition rates rapidly declined to an average of 18% with high false rejection and false acceptance rates of 55% and 27%, respectively. The extremely low recognition rates point to the fact that the power spectral density distributions of the testing speech data are not fully consistent with that of the training speech utterances. Although the recognition rate of 18% may be too low for application in speaker-independent recognition systems, it somehow also shows that the system may be adequate to identify imposters in a speaker-dependent recognition system.

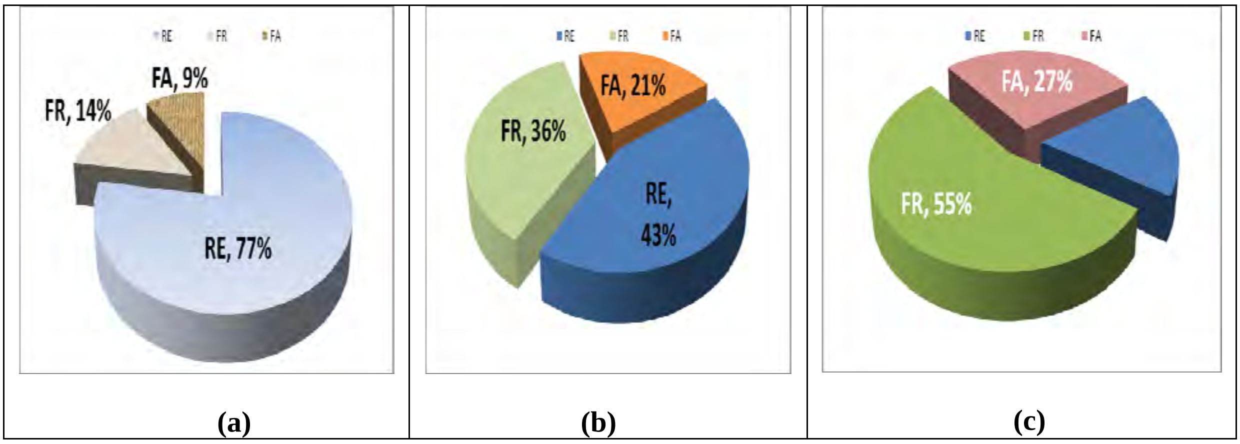


Figure 6 Recognition rates using (a) 5-word utterances known to training system (b) 20-word utterances known to the training (c) 20-word utterances unknown to the training system: RE – recognition, FR – false rejection, FA – false acceptance

4. Conclusion

In this work we have discussed the results of speaker recognition system based on the use of continuous speech utterances and the combined frameworks of ANN and the GMM. We have demonstrated through testing of speech utterances from 30 different speakers with each speaker providing four different utterances and results show the ability of the system to recognize speakers with success rate of 77% for the case of 5-word utterances and 43% for the case of 20-word utterances for situations where the utterances are known to the training system. In the case of speech utterances that are unknown to the system, a recognition rate of only 18% was possible for 20-word utterances. This low rate makes it possible for the system to detect imposters when used as a speaker-dependent system though much lower rates may be required for efficiency. The ability to adequately recognize speakers using their continuous speech waveform may find useful applications in systems that require detection of speakers in natural conversation environments. In further studies we are expanding the database of speakers to make more tests with varying lengths of utterances and extend the application to the recognition of speakers using local Ghanaian languages. We are also investigating unsupervised self organizing ANN architecture to improve on the autonomous classification and learning accuracy.

Acknowledgments

The authors wish to acknowledge the support and cooperation received from over 40 students of the University of Ghana and the Faculty of Engineering Sciences for making available the voice recordings for the training and testing of the voice recognition system.

References

[1] Rabiner, L. R., "Applications of speech recognition in the area of telecommunication", IEEE Proc., 1997, pp. 501-510.

[2] Picone, J. W., "Signal modeling techniques in speech recognition," IEEE Proc., Vol. 81, No. 9, 1993, pp. 1215-1247.

[3] Campbell, J. P. Jr., "Speech recognition: A tutorial", IEEE, Vol. 85, No. 9, 1997, pp. 1437-1462.

[4] Morgan, N., and Bourland, H., "Continuous speech recognition using multilayer perceptrons with hidden Markov models", International Conference on Acoustics, Speech and Signal Processing, Albuquerque, 1990, pp. 413-416.

[5] Bourlard, H., and Morgan, N., "Continuous speech recognition by connectionist statistical methods", IEEE Trans on Neural Networks, Vol. 4, No. 6, 1993, pp. 893-909.

[6] Nichie, A., "Voice recognition system using artificial neural network", Bachelor Thesis, Computer Engineering Department, University of Ghana, Legon, June 2012.

[7] Huang, X. D., Ariki, Y., and Jack, M., "Hidden Markov models for speech recognition", Edinburgh University Press, Edinburgh, 1990.

[8] Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", IEEE Proc., Vol. 77, 1989, pp. 257-286.

[9] Ramesh, P., and Wilpon, J. G., "Modeling state durations in hidden Markov models for automatic speech recognition", IEEE , Vol. 9, 1992, pp. 381-384.

[10] Katagiri, S., and Chin-Hui, L., “A new hybrid algorithm for speech recognition based on HMM segmentation and learning vector quantization”, IEEE Trans on Speech and Audio Processing, Vol. 1, No. 4, 1993, pp. 421-430.

[11] Frikha, M., Ben Hamida, A., and Lahyani, M., “Hidden Markov models (HMM) isolated-word recognizer with optimization of acoustical and modeling techniques”, Int. Journal of Physical Sciences, Vol. 6, No. 22, 2011, pp. 5064-5074.

[12] Johansen, F. T., “A comparison of hybrid HMM architectures using global discriminative training”, Proceedings of ICSLP, Philadelphia, 1996, pp. 498-501.

[13] Bengio, Y. “Neural network for speech and sequence recognition”, Computer Press, London, 1996.

[14] Lippman, R. P., “Review of neural networks for speech recognition”, Neural Computing, Vol. 1, 1989, pp. 1-38.

[15] Renals, S., and Bourlard, H., “Enhanced phone posterior for improving speech recognition”, IEEE Trans on Speech, Audio, Language Processing, Vol.18, No. 6, 2010, pp. 1094-1106.

[16] Yegnanarayana, B., and Kishore, S., “ANN: an alternative to GMM for pattern recognition”, Neural Networks, 2002, pp. 459-469.

[17] Biing-Hwang, J., Wu, C., and Chin-Hui, L., “Minimum classification methods for speech recognition”, IEEE Transactions on Speech and Audio Processing, Vol.5, No. 3, 1997, pp. 257-265.

[18] Haykin, S. O., “Neural networks and learning machines”, 3rd Ed., Prentice Hall, 2008.

[19] Riis, S. K., and Krogh, A., “Hidden neural networks: A framework for HMM/NN hybrids”, International Conference on Acoustics, Speech, and Signal Processing, Munich, 1997, pp. 3233-3236.

[20] Trentin, E., and Gori, M., “Robust combination of neural network and hidden Markov models for speech recognition”, IEEE Transactions on Neural Network, Vol. 14, No. 6, 2003, pp. 1519-1531.

[21] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., “Speaker verification using adapted Gaussian mixture speaker models”, Digital Signal Processing, Vol 10, No. 103, 2000, pp. 19-41.

[22] Reynolds, D. A., and Rose, R. C., “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Transactions on Speech and Audio Processing, Vol 3, No. 1, 1995, pp. 72-83.

[23] Brown, J. C., and Smaragdis, P., “Hidden Markov and Gaussian mixture models for automatic call classification”, Journal of Acoustic Society of America, Vol 125, No 6., 2009, pp. 221-224.

[24] Furui, S., “Speaker dependent feature extraction, recognition and processing techniques”, Speech Communication, Vol 10, No. 5-6, 1991, pp. 505-520.

[25] Furui, S., “Cepstral analysis technique for automatic speaker verification”, IEEE Trans on Acoustics, Speech and Signal Processing, Vol 29, No. 2, 1981, pp. 254-272.

[26] Brookes, M., “Voicebox: speech processing toolbox for Matlab” Imperial College, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, (March 3, 2012)

[27] Linde, Y., Buzo, A., and Gray, R., “An algorithm for vector quantizer design”, IEEE Transactions on Communications, Vol. 28, 1980, pp. 84-95.

[28] Xuan, G., Zhang, W., Chai, P., “EM Algorithms of Gaussian mixture model and Hidden Markov model”, IEEE, 2001, pp. 145-148.

[29] Dempster, A., Laird, N., Rubin, D., “Maximum Likelihood from incomplete data via the EM algorithm”, Journal of Royal Statistical Society, Vol 39, No. 1, 1977, pp. 1-38.