Modulation Components and Genetic Algorithm for Speaker Recognition System

Tariq A. Hassan
Computer Science Department
College of Education
Baghdad, Iraq

Rihab I. Ajel
Computer Science Department
College of Science
Baghdad, Iraq

Eman K. Ibrahim
Computer Science Department
College of Education
Baghdad, Iraq

***Abstract*—In this paper, the aim is to investigate weather or not that changing the filter-bank components (of the speaker recognition system) could improve the system performance in identifying the speaker. The filter is composed of 30 Gamatone filter channels. First, the channels are mel distributed of the frequency line. Then the componentsv́alues (center frequencies and bandwidths) changes with each run. Genetic algorithm (GA) is adopted to improve the filter component values that, in a result, improve the system performance. At each GA run, a new set of filter components will be generated that aimed to improve the performance comparing with the previous run. This will continue until the system reach to the maximum accuracy or the GA reach to its limits. Results show that the system will be improved at each run, however, different words might response differently to the system filter changing. Also, in terms of additive noise, the results show that although the digits affected differently by the noise, the system still get improving with reach GA run.**

***Keywords*—Computer Forensics; Digital Signal Processing**

I. Introduction

The speaker recognition system is, in general, the practical application of the speech-print idea presented by Kersta [1]. Basically, this idea open the door to the researchers to pay more attention the speech signal and find out the main characteristic that characterize one person from another. During the last 40 years, many models are suggested to parameterize the speech signal in a form that make it easy to extract features compatible (or strongly connected) with the problem in hand and ignore the others. Normally, the idea of speech-print can carry two major parts; these are, speaker recognition and speech recognition. Speech recognition is the way of understanding the work said by any speaker who try to give order or talk to the system. Speaker recognition, on the other hand, is the technique to identify the person based on his/her sound. No other biometric features should be used in the recognition process. The technique, however, is divided into two essential tasks. These are; speaker identification and speaker verification [2]. The first task is to identify who is talking to the system by assigning one utterance of speech to the already stored speakers in the system database.On the other hand, the second task is the case of the system to make sure that the incoming speech to the system is provided by the real person and not the fake one [3]. Speaker recognition, however, divided into two task depending on the style of using data. Open-set data speech is to use the same words (utterance) in both training and testing stages; while closed- set is to used one set of utterance in training stage and other set in testing stage.

Regardless of the job in hand, dealing the speech signal always encounter a wide measure of difficulties ranging from the out side noise that could, in some extent, distort the signal to the changing mood of speaker itself. So, the need for he robust system is quite challenging. One of the major key role of the system robustness can played by the speech parameterization method. Parameterization is the way of converting the speech into the set of parameters that are highly related to the problem in hand and ignoring any other features carried by the speech signal.

In this paper, a modified strategy used for speech signal parameterization is presented. The proposed strategy is to use the genetic algorithm along with the AM-FM parameter model in order to extract a set of parameters that are use for speaker identification system. The system is, basically, try to improve the performance of AM-FM model by adopting the genetic algorithm that help in selection the proper set of filter-bank channels values. So, the idea is to make the AM-FM model to be more flexible (not constrain by pre-fixed filter channels values) in estimation the modulation parameters from the speech signal.

The paper will organized as follows: First we present the method of representing the modulation components presented in speech. Then we talk about how to use the genetic algorithm in the proposed system. Our system explanation comes next with some details about how the system works. Experimental results with figures show the system performance will cone later. Conclusion will come at the end.

II. AM-FM Modulation Feature

As explained in [4] and [5], the speech signal can not be restricted with just a model presented 40 years ago; that is a source-filter model. Although this model presented some brilliant results regarding speech or speaker recognition techniques [6], [7], [8], [9]. However, its well known that some phenomena can not be captured by this model [10]. The speech instability and turbulence and other fluctuated and nonlinear open and close cycles in larynx all these phenomena can not be estimated well be the traditional source-filer model. So, the need for different model that able in some extent to estimate these and other instantaneous phenomena presented in speech signal to make the system more robust and much accurate to hold useful information in speech.

The AM-FM model is, basically, try to extract the instantaneous components of speech by estimating the instantaneous frequency (phase) and the instantaneous amplitude (envelope) from the speech signal. The modulation components of speech are then used as speech-print for the speech trained by the system.

The modulation parameters are obtained using the front-end system presented in Figure 1. The speech signal is divided into fix length frame of 20 to 25 ms in length, then the low-energy frames are ignored and let only to those with high or moderate energy to contribute in the feature extraction processing. The frames are then pass through a set of filter-bank channels of gammatone filter using the following formula;

$$x_c = x_N * gm \tag{1}$$

where, $*$ is the convolution operator, x_c is single-valued signal of filter channel c, x_N frame number N of the speech signal, and g_c is the impulse response of gammatone filter.

$$gm(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi) \tag{2}$$

where f_c is the central frequency of the filter, and ϕ is the phase, the constant a controls the gain of the filter, and n is the order of the filter, and b is the decay factor which is related to f_c and is given by [11]:

After we obtain single-component frame (around one particular filter-bank center frequency) the analytic signal is calculated using

$$Ax_c = x_c + j.\hat{x_c} \tag{3}$$

where, the $\hat{x_c}$ is the Hilbert transform of speech signal frame x_c, and Ax_c in the analytic complex single-valued signal. For this complex signal, the instantaneous frequency is computed as;

$$IF_c = \frac{1}{2\pi}.\frac{d}{dt}[\arctan\left(\frac{Ax_i}{Ax_r}\right)]...... \tag{4}$$

where, Ax_i, Ax_r are the imaginary and real parts of the signal Ax_c respectively.

The instantaneous amplitude is computed as:

$$a\hat{m}p = \sqrt{Ax_r^2 + Ax_i^2}..... \tag{5}$$

These step are usually adopted in many AM-FM modulation system model for speech and speaker recognition. The trick here is the filter bank center frequencies and bandwidths values that almost match the human auditory system. As experiment done by [4], the experimental results show different identification results of different filter-bank component values. This ensure that the fixed-valued filter components (Whether it mel or linearly distributed) are not the best choice for signal feature extraction. Therefor, the proposed system try to avoid this problem by adopting different strategy that allow as to change the filter component values with each run until the system get the best filter values that give us the best description of the speaker. Next section will explain the main steps of the genetic algorithm used in filter components best value selection.

III. Genetic Algorithm Selection Process

Genetic algorithm is adopted to make our proposed system more flexible in selection the best set of filter-bank parameters (center frequencies and bandwidth). At the beginning, the system start with the definition of a filterbank of Gaussian-shape filters with Mel spaced center frequencies and bandwidths. After the first run, the system will test the results. In the case of accepted recognition accuracy, the system will adopt the current filter-bank components values. Otherwise, the genetic algorithm will take the filter-bank values and generate a new set of filter components and do the genetic algorithm step on both sets of filter-bank components. The main step that are normally adopted by the genetic algorithm are;

1) Initial population: set a number of elements (30 number) that represent an initial set of filter-bank component values. In the genetic algorithm world, each filter value represent one individual DNA in the chromosome, and each chromosome represent one suggested solution of the filter component values.
2) Evaluation: After each run, the system will evaluate the values of each produced chromosomes and give a degree that represent an objective mark for each chromosome produced in initialization step.
3) Elitism: It is an important approach in genetic algorithm system. The idea is to let some of the best solution of one generation to keep its values for the nest generation. In this step, the system will guaranteed that some of the highly mark solution will not be lost.
4) Selection: normally, this step play an important role in the genetic algorithm system since it will decide which of the chromosomes will be nominated to be mate in the next crossover step.
5) Crossover: Two strategies are usually adopted in crossover step; first, by uniformly cutting some parts of each chromosomes and do values exchange between them. Second, use a selection mask that identify the locations where exchange will be happen. In our system, we used the uniform cutting crossover.
6) Mutation: when some values some where in the chromosome changed randomly. The new value called as the mutation value. Normally, the mutation value happen within a limited probability, 10% or less is the mutation rate that are usually used.

IV. The Genetic AM-FM Modulation System

In order to generate one speaker feature vector, which represent the modulation components of one specific speaker presented in speech, a speech signal must be divided in to fix-length frames (25ms in our system). Short length frames would help us to analyse the speech signal in the level of phonemes (a level of one pronounce letter) rather than a level of utterance (one spoken word). Pre-processing is the next stage which include discarding some useful parts of the speech and do the pre-emphasis and windowing process. Next comes the step of breaking down the speech fames into its basic components. In other words, dividing the speech into single-valued waves that represent one band signal around the center frequency of one specific channel of the filter-bank. Multiband filtering scheme with gammatone filter-bank of 30 mel-frequncy distributed channels is the technique used in our proposed system. The filter bandwidth is computed using the following equation;

$$Bw(k) = 25 + 75\left[1 + 1.4(f_c(k)/1000)^2\right]^{0.69}. \tag{6}$$

where f_c is the centre frequency of the filterbank. The filter bandwidth is relying totaly on the center frequency. So, when the center frequencies are mel scaled so do the bandwidths. The analytic signal for each filter channels output wave is calculated using Hilbert transform. The analytic signal (complex form of the real speech signal) will help us to estimate the phase and envelope component of the speech since both components are depending in some how on the imaginary part of the signal, as well as the real part. Using equations 4, 5 to compute the instantaneous frequency and instantaneous amplitude respectively. Both values are normally combined in one entity that represent the mean amplitude-weighted instantaneous frequency (phase). The weighted-phase is computed using the following equation;

$$F_w = \frac{\int_{t_0}^{t_0+\tau}[f_n(t) \cdot \hat{a}_n^2(t)]dt}{\int_{t_0}^{t_0+\tau}[\hat{a}_n^2(t)]dt} \tag{7}$$

where τ represents the duration of the speech frame.

Using this scenario, each signal frame will be represented by just 30 modulation components, which represent the number of filter-bank channels. The modulation components of all frames in the speech signal are then collected together in one two dimensional ($Ch \times K$), where Ch represent the number of filter-channels and K represent the number of the signal frames.

At the training stage, the system will take some the speech samples of all speakers contributed in the system to build up database. In the testing stage, the system will adopt the same filter parameter values used in the training stage. Then examine the result using GMM (Gausian mixture model) with 16 (in our system) mixer component as a subsystem classifier. If the obtained result were nice and give us high accurate recognition, then the system is fine and no more action will be taken. Otherwise, if the result is not accurate, the system will produce a new set of filter-parameters values and do a new cycle of training and testing stages. This wull continue until the system reach the required accuracy level or the number of epoch set in advance.

Figure 1 shows the main steps of our proposed system. Theses steps will apply to all speech signals in the speech corpora to generate a reference database for all trained speakers. After the first run, the system will examine the recognition results; if they were fine and accepted, then the system will stop. Otherwise, the GA will generate a new set of filter components and re-run the steps of Figure 1. The system will stop until it get to the best results or it reach to the GA epoch limits.

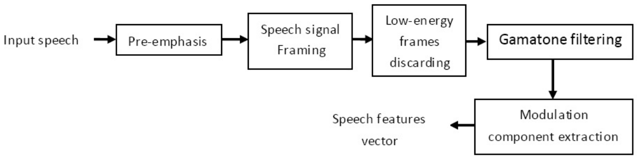


Fig. 1. Step of our proposed method of speech signal modulation component extraction

V. Experiment and Results

The training set that we adopt to evaluate our proposed system consist of 60 native English speakers saying three digits *zero*,*one*, and *nought*. Each speaker contribute in five recoding sessions with five repetitions each. each contains The first two sessions (10 repetitions) are used in the training stage and the speech from the other sessions are used in testing stage.

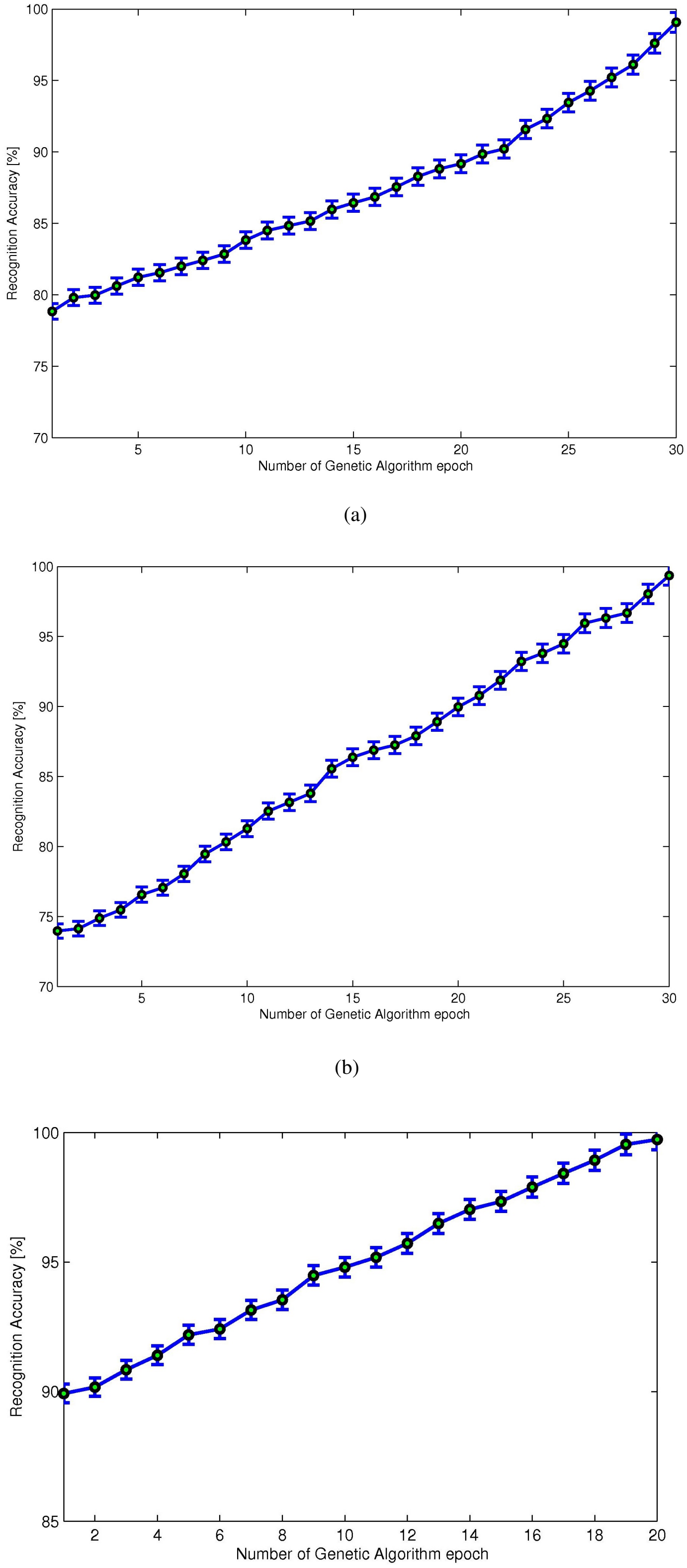
The strategy is to train the system with the 60 speakers saying one specific word (saying for example the digit *zero*) and then use the same word but in different session (recorded some time later after the first two sessions). This is the strategy of text-depending speaker identification. Also, we try to divers the accuracy examining of our proposed system by add some noise to the speech data and repeat the testing process.

As we mentioned above, the encoding of the speech signal in a form of AM-FM parameter to generate a set of feature vectors is required fine tuning of the filter-bank components (center frequencies and bandwidths). The best tuning will be obtained by the support of the genetic algorithm process. The importance of using GA is to allow us to select the best set of filter parameters that make the system operate with high accuracy. At each GA run, a new set of filter components will be produced, at these filter components the system will be tested to see to what extent that these components will improve the performance. If the recognition accuracy is accepted then the system will stop at this point and filter components will be taken to be a filter-bank standard components. Otherwise, the system will take another round to choose a new set of filter components.

The efficiency of our system is evaluated using a speech data of text-dependent speaker recognition task. We compare the performance of the system under cleaned data speech and noisy data. The testing will include three words of the database, *zero*,*one*, and *nought*). In fact, the speech database contain more digits that can be used in our system but we just select those words since they can, in some extent, reflect the whole image of the speech database,

Figure 2 summarizes the recognition accuracy results of cleaned data speech of the frequency range (0..4)kHz using Gamatone filter bank with components are firstly mel-spaced between (100.. 3900)Hz. As shown in the figure, the results is improved with each GA run until they reach to the maximum recognition accuracy or it reach the epoch limit. The error areas represent the standard divination values of results around the mean.

Different words (digits) need different number of GA epoch. For example, digits (One, Nought) required 30 GA epoch to reach to the best accuracy, while the digit (Zero) re-



(a)

(b)

(c)

Fig. 2. The recognition accuracy results of Text-dependent speaker identification of clean speech database with mel-scaled centre frequency and bandwidth and frequency range of (0..4) kHz of three digits; (a) word “One” , (b) word “Nought” , (c) word “Zero” ,

quired only 20 GA epoch to reach to the maximum recognition accuracy. This is might depends, in some way, on the amount of voiced sound presented i speech signal, or could rely on the kind of the composed speech phonemes. The phonemes that strongly linked to the speaker rather than speech are defiantly need less GA epoch and give more accurate results.

Figure 3 shows the accuracy results of the system with noisy data speech of 30% Guassian white noise. The results clarify that different words could effected diffrently with the noise, this is clear in the recognition results.

The recognition accuracy has differently affected by the additive noise to the speech. The GA method try to get the best filter components values that manage to alleviate the noise effect and boost the system performance.

VI. Conclusion

This paper has set a different strategy that used the GA method and the modulation components presented in speech signal on order to extract and estimate the speaker features presented in speech signal. The strategy state that updating the filter-bank components at each run will improve the system performance and increase the recognition accuracy rate. This idea stems from the fact that different people have different shape of the filter, and also, that the same person could change, unintentionally, its auditory filter when listen to different sounds. Also, in terms of estimated features, the modulation components of speech are well proved to hold more informations about the speaker and less affected by the noise comparing with other speech signal models. Results show that different digits in the database (different words) need different GA epoch to reach its maximum accuracy. Also, in terms of speech signal noise, as we saw, words are affected differently by the additive noise.

References

[1] L. G. Kersta, “Voiceprint identification,” *Sceince*, vol. 196, pp. 1253–1257, 1962.

[2] D.A. Reynolds, “An overview of automatic speaker recognition technology,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP ’02).*, 2002, vol. 4, pp. IV–4072 – IV–4075.

[3] F. Bimbot, J-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A tutorial on text-independent speaker verification,” *EURASIP Journal of Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.

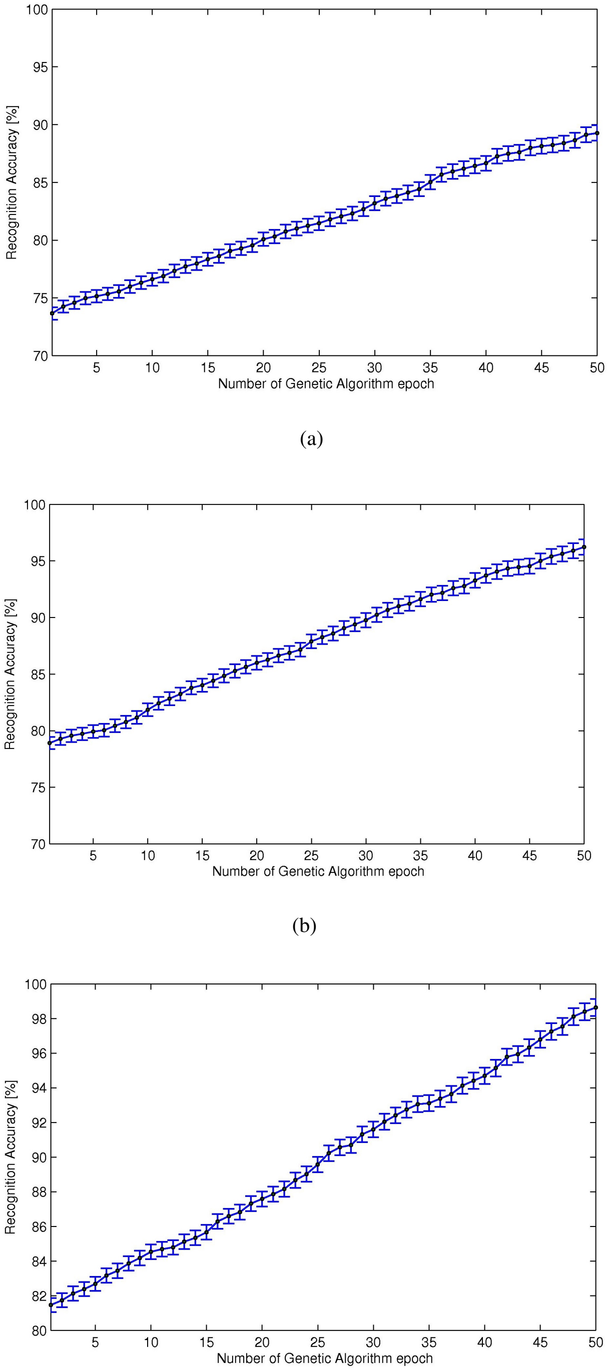
[4] M. Grimaldi and F. Cummins, “Speaker identification using instantaneous frequencies,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, Aug. 2008.

[5] Dhananjaya N Gowda, Rahim Saeidi, and Paavo Alku, “Am-fm based filter bank analysis for estimation of spectro-temporal envelopes and its application for speaker recognition in noisy reverberant environments.,” in *INTERSPEECH*. Citeseer, 2015, pp. 1166–1170.

[6] Md Sahidullah and Goutam Saha, “A novel windowing technique for efficient computation of mfcc for speaker recognition,” *IEEE signal processing letters*, vol. 20, no. 2, pp. 149–152, 2013.

[7] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman, “Mfcc-gmm based accent recognition system for telugu speech signals,” *International Journal of Speech Technology*, vol. 19, no. 1, pp. 87–93, 2016.

[8] Prashant Borde, Amarsinh Varpe, Ramesh Manza, and Pravin Yannawar, “Recognition of isolated words using zernike and mfcc features for audio visual speech recognition,” *International Journal of Speech Technology*, vol. 18, no. 2, pp. 167–175, 2015.



(a)

(b)

(c)

Fig. 3. The recognition accuracy results of Text-dependent speaker identification of noise (30% Guissian white noise) speech database with mel-scaled centre frequency and bandwidth and frequency range of (0..4) kHz of three digits; (a) word “One” , (b) word “Nought” , (c) word “Zero” ,

[9] Khan Suhail Ahmad, Anil S Thosar, Jagannath H Nirmal, and Vinay S Pande, “A unique approach in text independent speaker recognition using mfcc feature sets and probabilistic neural network,” in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*. IEEE, 2015, pp. 1–6.

[10] Mohammadi Zaki, J Nirmesh Shah, and Hemant A Patil, “Effectiveness of multiscale fractal dimension-based phonetic segmentation in speech synthesis for low resource language,” in *International Conference on Asian Language Processing (IALP), 2014*. IEEE, 2014, pp. 103–106.

[11] Hui Yin, Volker Hohmann, and Climent Nadeu, “Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency,” *Speech Communication*, vol. 53, no. 5, pp. 707 – 715, 2011.