A Data-Base System for Speaker Identification

Bassam Ali Mustafa
University of Mosul
Mosul, Iraq.
bassamali2004@yahoo.com

Basil Y. Thanoon
University of Mosul
profbasilyt@yahoo.com

Saad Daoud Al-Shamaa
Al-Mothana District
saaddaoud2003@yahoo.com

Abstract

During the last few years many attempts were accomplished in the field of voice and speech processing aiming to build speakers identification systems. The basic views of these systems were different, but the accuracy of the final computer process result for the identification depended on varieties of factors.

Briefly, an attempt was made in this research to build and implement speaker's identification database system. This database is used for storing speaker's voices as well as all information needed for the identification purpose. Some processes are then performed on voice signals to extract some of the needed features to be stored in the database and to be used for later processing.

Many techniques in both, the time domain and the frequency domain are used for the purpose of voice feature extractions. The famous linear prediction coding (LPC) is applied to parameterize voices, and to be used later in voiced / unvoiced different feature extraction processes methods. Also the direct classical techniques for voice feature extraction purposes are used.

Some techniques dealing with LPC and based on statistical concepts are suggested for voice feature extraction. From the practical test, the performance of these techniques seems to be well as compared with other standard techniques.

A test for the identification of speaker's voice is performed by recording a new voice from one of the database members with no more information. Then the computer processes the information comparing the new recorded test voice with the information on the database and then gives the final result for the speaker's voices database system.

الملخص

خلال السنوات الأخيرة أنجزت عدة محاولات في مجال الصوت ومعالجة الكلام بهدف بناء انظمه لتشخيص المتكلمين. إن الأطر الأساسية لهذه النظم مختلفة إلا إن دقة نتائج ألمعالجه النهائية لغرض التشخيص تعتمد على عوامل مختلفة.

باختصار فقد أجريت محاوله في هذا البحث لبناء واستخدام نظام قاعدة بيانات لغرض تشخيص المتكلمين. واستخدمت قاعدة البيانات هذه لخزن أصوات مجموعة من المتكلمين إضافة إلى جميع المعلومات المطلوبة لغرض التشخيص. وقد أنجزت بعض المعالجات على الإشارات الصوتية لغرض استخلاص بعض السمات المطلوبة ثم لتخزن في قاعدة البيانات لكي تستخدم في المعالجات ألاحقة.

وتستخدم عدة تقنيات، في كل من المنطلق الزمني والمنطلق الترددي، لغرض استخلاص السمات الصوتية، وتطبق طريقة ترميز التنبؤ الخطي لمعالجة الأصوات إضافة إلى استخدامه في استخلاص السمات الموجودة في حالة الكلام. وتستخدم أيضا بعض التقنيات الكلاسيكية المباشرة لغرض استخلاص السمات الصوتية. وتقترح بعض التقنيات التي تتعامل مع طريقة ترميز التنبؤ الخطي والتي أساسها مفاهيم إحصائية لغرض استخلاص السمات الصوتية. ويبدو من خلال ألاختبار العملي لهذه الأساليب إن إنجازها جيد مقارنة مع الأساليب القياسية.

وينجز اختبار لتشخيص المتكلمين وذلك بتسجيل أصوات جديدة لأعضاء قاعدة البيانات ومن دون إعطاء أي معلومات إضافية عنهم. ويقوم الحاسوب بمعالجة المعلومات ومقارنة الصوت الجديد تحت الاختبار مع المعلومات المتواجدة في قاعدة البيانات ثم يقوم بعد ذلك بإعطاء النتيجة النهائية والمتضمنة اسم الشخص المتكلم.

1. Introduction

Voice and speech, which conveys different kinds of information between human beings, is concerned with computerized voice and speech processing research concepts. Computerized speech processing depends on a

wide spectrum of disciplines and their techniques. For a successful progress in its application areas a wide knowledge as possible of these disciplines is needed. Speech analysis, synthesis, recognition, language processing for speech understanding, phonetics, psycholinguistics, computational linguistics, signal processing, telephone systems, speech coding, data compression, voice mail, workstations, personal computers, and networks are areas for speech processing applications.

Another kind of applications is speaker recognition, identification, and / or verification. The voice prints and forensic science of voice identification applications, which was first introduced in the mid 1960's are used in law enforcement.

History of speech sound analysis goes back more than one hundred years to Alexander Melville Bell who developed a visual representation of the spoken word.

This system developed by Bell is the phonetic alphabet which he called "visible speech". The code produced a visual representation of speech which could convey to the eye the subtle differences in which words were spoken. This system was used by both Bell and his son in helping deaf people learn to speak.

In the early 1940's a new method of speech sound analysis was developed by Potter, Kopp & Green, working for Bell Laboratories. In 1947, they published their work in a book entitled "Visible Speech". Their work is a comprehensive study of speech spectrograms designed to linguistically interpret visible speech sound patterns.

Research in the area of speaker identification slowed dramatically with the end of World War II. It was not until the late 1950's and early 1960's that the research began again [`1`][`2`][`5`].

The main aim is to work with computerized speaker identification, where it covers the individual main required information, plus his/her recorded voice, speech parameters involved, and essential variables needed for computerized speaker identification database system. This is considered as an important research topic, for its wide range of applications.

Although huge amount of work, various algorithms, and literature are encountered, but none has provided a robust system that can be adopted.

Therefore the aim of this paper is to build such a research application, by using today's most useful computer language, which can provide efficient programming tools for this matter.

2. Speech Signal and Linear Prediction

Within a wide variety of speech processing applications, the ability to present speech waveform by a small number of low information rank parameters is of fundamental importance. Fourier analysis procedures play an important role in developing efficient parametric representation of the speech signal. Based on the Fourier analysis, spectral analysis methods in speech analysis were introduced at M.I.T and Bell Laboratories.

Another important speech analysis technique, known as linear prediction (LP), was then introduced to speech analysis. The basic formulations of LP can be traced back, partly to the works of Yule, and Prong [`9`]. One of the basic formulations of LP is the Autocorrelation which was first introduced in speech analysis by Itakura and Saito in Japan.

The works of Atal and Schroeder in predictive coding led to the development of what has become to be known as the covariance method of linear prediction. As a result, linear prediction has become very popular as one of the most powerful approaches of the time domain analysis of speech.

A major use of linear prediction, besides spectral estimation, is for speech coding, where the method is called linear prediction coding (LPC).It has been found that, LPC is the most useful method of speech analysis because of their accuracy and speed of computation.

The basic idea behind linear predictive coding is that each speech sample can be represented as the input of the linear combination of its past values and the current.

Such a representation leads to a simple analysis procedure in the time domain. The different methods of determining linear prediction parameters for speech signals will be discussed in this chapter [`1`][`2`][`3`][`9`][`11`].

The content of a human speech, which is the result of a very complex and not completely understood process, is basically a concept, which is formed in the brain and somehow is converted to neural signals that travel to the muscles of the human speech production mechanism.

These components then produce an acoustic waveform that is radiated out in air from the head as a speech signal [`10`].

The speech production can be thought of as being composed of elements of analogue

information. Sampling process then takes place, where a discrete piece of waveform data is represented by a single numerical value. The fundamental of speech production and its modeling, enable a better understanding of the digital processing for a speech environment. This is available nowadays in a number of references.

Many studies were devoted to the subject of statistical behavior and characteristic of a speech waveform as illustrated in. Fig (1).

The properties of the waveform change with time, and hence speech signal, is usually considered as a non-stationary time series with significant inter-segment differences both in terms of the amplitude level and spectral (frequency) content. It will formally be termed short-time or short-term descriptions, while long-time or long-term description will refer to waveform durations, where its scope is considered several orders of greater magnitude [`9`]. The simplest description of any waveform is the plotting of the amplitude versus time which is known as the time plot.

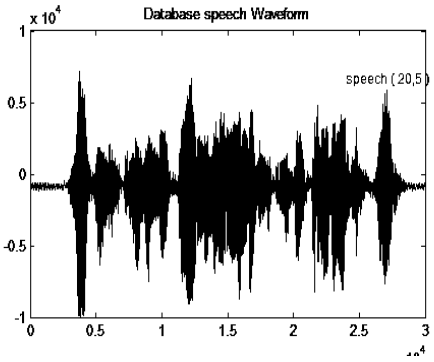

Fig (1) Speech Waveform

2.1 Digital Representation of Speech Signals

Generally, the representation can be classified into four categories as illustrated in Fig (2). [`10`]:

Linear prediction is a data model, and it has been known by different names and used in different fields during the last forty years [`9`]. A major usage of linear prediction is for speech coding, where the method is called linear prediction coding (LPC) [`1`].

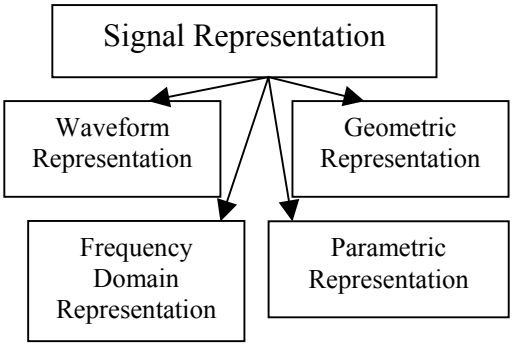

Fig (2) Categories of the Signal Representation

Once the coefficients of LPC are transmitted over a communication channel, the speech can be reconstructed (synthesized) by using the LPC coefficients. This technique can produce a very high–quality speech. Similar applications are used in some speaking toys and talking calculators, as well as in speech applications software for personal computers [`1`].

The fundamental linear prediction problem is concerned with the task of predicting the value assumed by the generally complex-valued signal elements “sample of a recorded speech”
S (n), with this prediction being based on a linear combination of the *p* most recently observed signal elements (the past *p* samples, plus the current nth speech sample) [`2`]. i.e.

$$S(n) = -\sum_{k=1}^{p} a_k \, S(n-k) + G\,u(n) \qquad (1)$$

Where $\{a_k\}$ are parameters representing the weights of the past samples on the current speech sample, and called the predictor coefficient of the signal, *u(n)* represent the nth sample of the excitation (noise), *G* is a scale factor, and *p* is the order (or the embedding dimension) of the model that represents the length of the memory of the model. For a process which fluctuates in some form in time, and it can produce a graph from plotting its values against “time” as a variable, are examples of a general type called “***time–series***”. Of course, the scale of the time–axis may vary considerably.

For speech signals we would usually measure it in seconds, or fractions of a second. If the excitation *u(n)* is unknown, then the best estimate of the speech sample based on the past speech samples is given by :

$$\hat{S}(n) = -\sum_{k=1}^{p} a_k \, S(n-k), \qquad (2)$$

Where $S^{\wedge}(n)$ denotes the predicted value of $S(n)$. The coefficients $\{a_k\}$ here are assumed to be known and called the predicator coefficients of the speech signal. Thus each output speech sample is a sum of its predicted value and the corresponding sample of the excitation.

The prediction error $e(n)$ is defined to be as the difference between the speech sample $S(n)$ and its predicted value $S^{\wedge}(n)$ as follows.

$$e(n) = S(n) - \hat{S}(n). \quad (3)$$

Substituting from (2) into (3) we get :

$$e(n) = S(n) + \sum_{k=1}^{p} a_k S(n-k), \quad (4)$$

The idea behind this is to determine the predictor coefficients $\{a_k\}$ directly from the speech signal or the recorded speech.

When the predictor coefficients $\{a_k\}$ are unknown, then they can be estimated by using standard statistical methods. Let us denote these estimates by $\{a^{\wedge}_k\}$, then the estimated prediction error, known as the prediction residuals, denoted by $e^{\wedge}(n)$ and can be obtained as follows :

$$e^{\wedge}(n) = S(n) + \sum_{k=1}^{p} a^{\wedge}_k S(n-k). \quad (5)$$

The LPC parameters refer to a variety of methods for finding estimators to the problem of modeling speech waveform.
The difference between these methods is the way of viewing the problem [`1`]. One of these methods is the autocorrelation function estimate.

The sampling rate at which a waveform audio driver performs audio-to-digital or digital-to-audio conversion is 44.1 kHz.

But Possible Values could be 8000, 11025, 22050, 32000, 44100, or 48000 kHz. The default value is 44100 kHz., [`7`].

The chosen estimate function is the autocorrelation estimates method for LPC to be calculated and stored in the individual speech database for further processing use.

With typical values suggested for the recorded voices, the sampling rate frequency should be 10 kHz, and LPC order value of p=14, which is different from the above Windows Xp operating system default, or suggestion [`9`].

3. Feature Extraction for Sensing

It is important to notice first that the sampled parameters that represent a single segment, whose duration is the sampling interval and constitute effectively the minimum segmental unit in the recognition system, is called a *frame*.

Some systems calculate a number of additional acoustic characteristics (or features) associated with each frame which is known to be relevant to segmentation and labeling, or to make comparisons between the signal and the stored references more effective. Such acoustic features, generally involve [`1`] [`2`] [`11`] :

A- The overall amplitude function.

B- Fundamental frequency:
This is usually used for segmentation into voiced and unvoiced portions of the signal and for differentiating between voiced consonants and their unvoiced counterpart, and detecting stress and boundaries.

C- Intensity:
This is the energy of speech signal.

D- Formant frequency:
This is for vowel identification and for identification of the place of articulation of stop consonants.

E- Zero crossing rates:
This is for detecting burst and fricatives.

F- Time spectrum derivative:
This is the spectral distance between two successive sounds.

3.1 Parametric LPC Features Extraction

It is known that many features can be extracted from a speech signal which can be considered as censer detections that can help in the identification, or the verification of one speech for a speaker person, from another.

The parameters of LPC for a speech signal, which are based on statistical calculations, are constructed using both of the following classifications:

1) A LPC multi-feature voiced/unvoiced (V / UV) speech signal classifier.
2) A LPC voiced (V) speech signal classifier.

To implement such a scope for identifying procedures, different approaches for feature extraction in the time domain, and in the frequency domain have been considered in this research [`1`] [`2`] [`4`].

In the following paragraphs, we give some approaches based on parametric LPC, which can be used for feature extraction.

The first approach is in the time domain and is suggested by the author. In this approach, three techniques based on statistical concepts are suggested for feature extraction. The second approach is in the frequency domain which is usually used by many users.

3.2 Statistical Properties of frames

Assume that we have a speech signal with length *N* and we are fitting a LP model of order 14 to it. Let $\upsilon = \{\upsilon_{ij}\}$ denotes the LPC matrix, where *i = 1, 2, ... , N* and *j = 1, 2, ... , 15*. The number 15 here is the number elements of each speech frame. This number corresponds to the order of LP model plus 1 (the coefficient of the constant term in the model). Let υ_i denotes the row vector that corresponds to the j-th frame. Note that the first element of this vector is always 1. Let

$$\eta_i = \upsilon_i \upsilon_i^{l} \quad ; i = 1,2,\ldots,N \qquad (6)$$

Where " ^{l} " denotes the transpose. In fact, η_i is an indicator of the energy of the i-th frame. According to our practical experience, we found that, $\eta_1, \eta_2, \ldots, \eta_N$ possess vital information which can be used for feature extraction. Hence, we suggest using the following statistics for this goal.

I. Mean LPC energy feature extraction :

This feature extraction is denoted by μ and calculated in the usual way as:

$$\mu = Mean\ (\eta 1, \eta 2, \ldots, \eta_N) = N^{-1} \sum_{i=1}^{N} \eta_i \ .$$

This feature extraction is a measure of the central tendency of LPC energy.

II. Standard deviation LPC energy feature extraction:

This feature extraction is denoted by σ and calculated as follows:

$$\sigma = Standard\ Deviation\ (\eta 1, \eta 2, \ldots, \eta_N) = \left(\sum_{i=1}^{N} \eta_i - \mu \right)^2 / N \Big)^{1/2} .$$

This feature extraction is a measure of dispersion of LPC energy.

III. CV LPC energy feature extraction:

This feature extraction is the ratio of the standard deviation with respect to the mean and it is calculated as follows:

$$CV = \sigma / \mu$$

3.3 Spectra of LPC based Speech Feature

Assume that we are given a LP model of the form given in equation (1). Then a smooth parametric estimate of the spectra can be obtained directly from the LP model as follows [`8`]:

$$h(\omega) = \frac{G^2\ \sigma u^2}{2\pi} \frac{1}{|1+a1\ exp(-iw)+a2exp(2iw)+\ldots+ap|^2} \quad , \quad -\pi \le \omega \le \pi$$

Where $\sigma^2 u$ is the variance of $\{u(n)\}$ and $i=\sqrt{-1}$.

3.4 Direct Speech Signal Feature Extraction

Many features can be extracted directly from a speech signal data, which can be considered as censer detections helping in the identification, or the verification of one speech for a speaker person, from another.

When a recorded speech is required to be identified, the differentiation between different records is needed, and the zero crossing technique is used as an identifier to detect when the input signal crosses the zero line in either the rising or falling direction, and computationally can be presented as [`4`] [`12`];

$$ZCR = \#\ (\underset{i \epsilon R}{IF}\ S(i).S(i+1) <= 0\ ) \qquad \ldots. (8)$$

Where, R refers to the entire number of records within a file, and the # represents the number of repetitions.

Adaptive zero crossing can be considered as a dynamic computational algorithm compared with the static type zero crossing computational algorithm. The adaptive type responds to incremental changes over time in response to an external threshold input variable, which represents the minimum significant values of the zero crossing to be calculated. Adaptive zero crossing is presented as;

$$AZCR = \#\ (\underset{i \epsilon R}{IF}\ S(i).S(i+1) < 0\ , and\ |S(i)| >= Threshold\ ) \ldots (9)$$

3.5 Spectra Feature Extraction

Let $\{s(n);\ n=1, 2, \ldots, N\}$ be a realization from a speech signal $\{S(n)\}$. To obtain a non-parametric estimate of the spectrum of this signal, the auto-covariance function $R_{ss}(k)$ is first calculated as follows:

$$R_{ss}(k) = (N-k-1)^{-1} \sum_{n=k+1}^{N} [\{s(n) - \hat{s}\}\ \{s(n+k) - \hat{s}\}] \ ;$$

$$k = 0, \pm 1, \pm 2, \ldots, \pm(n-1) \ (10)$$

$$\textbf{Where } \hat{s} = N^{-1} \sum_{n=1}^{N} s(n).$$

Then, a window type estimate of the spectra is given by [`8`]

$$h(\omega) = \frac{1}{2\pi} \sum_{k=-m}^{m} \lambda(k)\, R_{ss}(k)\, cos(\omega k)\ ; \quad -\pi \le \omega \le \pi \qquad (11)$$

Where m is the truncation point (e.g. $m = \sqrt{N}$) and $\lambda(k)$ is some window function satisfies the two conditions:

1. $\lambda(k) \ge 0$ for all k.
2. $\int_{k} \lambda(k)\, dk = 1$.

3.5 Pitch and Jitter, the Fundamental Frequency Feature Extraction

The pitch or the fundamental frequency feature extraction is derived from the recorded voice signal, which is stored on the database. After processing the recorded voice, the glottal closer index (GCI) is extracted [`1`]. The pitch feature is then extracted from the voice signal, using the glottal closer index (GCI).

The jitter feature extraction is derived from the recorded voice signal which is stored on the database. After the process to extracted the fundamental frequency feature from the voice signal, a smoothing process for glottal closer index (GCI) takes place and then the following calculation process is done for the final jitter result of the voice [`1`]:

J = standard deviation (smoothing - index – Pitch)/Pitch. (12)

4. Two-dimensional Voice Storage Database

For the purpose of storing 3 seconds or more of uttered voice for each individual, the sophisticated two-dimensional unfixed uttered voices storage database was established for the checking of speaker identification. And an interactive software toolbox system was built through the command window of the MATLAB [4]. The execution of the interactive Speaker Identification System software, gives the following step levels of information as an output. Fig (4.1) shows the questionnaire main menu information page. There exist six options in the main menu. A number should be inputted referring to the desired option needed to be executed.

A brief description will be provided as in the following:

1- Option number 1, if it is selected from the main menu, it means that you want to hear all voices and information recorded on the database concerning a specific individual.
2- Option number 2, if it is selected means that you want to hear the first voice signal recorded for each individual speaker on the whole database, one after the other until the last speaker. Together speaker's voice plotting figure as in Fig (4.2) time domain waveform, hearing the first stored recorded voice, and recorded information on the database concerning each individual as in Fig (4.3), are all displayed on screen.

This procedure helps an observer to have an idea concerning each voice signal that is being displayed on the screen, for all speakers one after the other, by pressing any key on the keyboard for displaying each voice signal
3- Option number 3, if it is selected means that you want to add a new voice to the voice-database, and the system will require some information to decide the storage place.
4- Option number 4, if it is selected means that you want to delete a voice from the Voice-Database.
5- Option number 5, if it is selected means that you want to test, whom this voice belongs to? Comparison of the unknown voice with previously recorded voices of the database should take place.
6- Option number 6, if it is selected means Exit, and end the speaker identification system.

Fig (4.1) Main Menu of Voice Identification System

5. Conclusions

The following conclusions can be drawn from the research work.

1) Human voices are so much time varying that one recorded voice signal of a short time can never convey to distinguish speaker identification almost 100%.

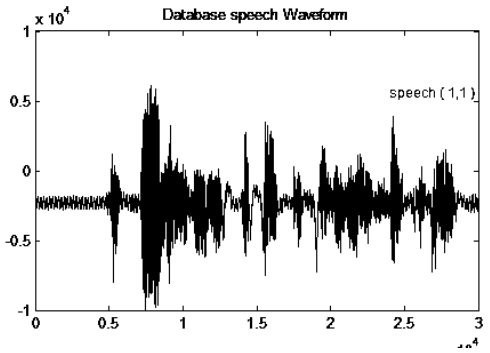

Fig (4.2) Waveform Voice Database plot

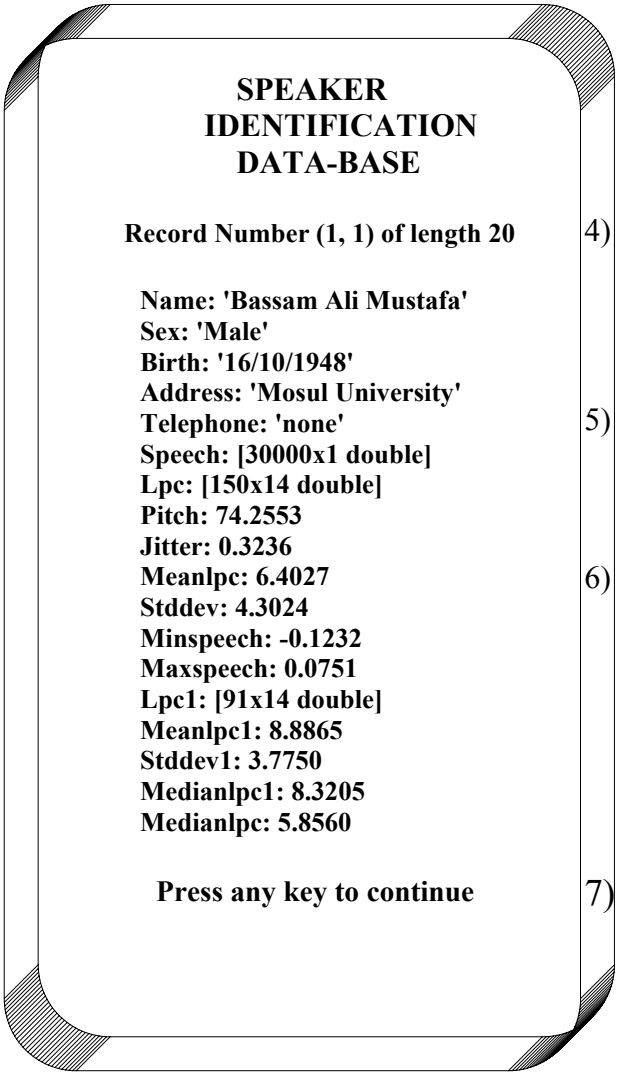

Fig (4.3) Voice Database Information Record display

2) The recorded voices on the database are most likely to be renewed every suggested period of time (6 months, 9 months, or 1 year) due to the variation of voices with time. Voices always change as long as the human being is growing in age with time.
3) On the speaker's identification voice database system, the recorded voices were for males and female persons, grown ups, and under ages. A problem was encountered when under age voices were mixed with grown up voices on the same database used for speaker identification. Therefore, for a better performance of the speaker identification database system it is wise to separate speakers on different databases: One for male grown ups, another for male under ages, third for female grown ups, and fourth for female under ages. Also this requirement coordinates with the speech recognition engine that is supported with the windows XP operating system.
4) For a better speaker identification final target, it is important to understand that the surrounding recording environment has a major effect on the decision efficiency the system provides as a final result.
5) For a better decision making, it is not the number of features extractions that counts, but the decision efficiency of the voice feature extraction that effects.
6) Some speakers have a monotonous way of speaking, and all their voice signals look very much a like on the different calculations of the speaker identification. This kind of situation no variety of sounds might ruin the final decision result. It requires varieties of sounds from each speaker persons. The more the varieties for all speakers the better the final result is.
7) It is well known that speech wave files need large storage capacity in the memory. The environment of

MATLAB seems to be quite suitable for dealing with such files.

REFERENCES

1. D. G. Childers, Speech Processing and Synthesis Toolboxes , John Wiley & Sons Inc. 2000.
2. Frank Fallside and William. A. Woods, Computer Speech Processing, Prentice Hall International 1985.
3. مياده غانم حموشي ، معالجة الصوت في الوسائط المتعددة ، تقرير ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل ٢٠٠٠.
4. MathWorks, The MathWorks Inc., MATLAB The Language of Technical Computing, Version 6.0.0.88 Release 12 “, September 22, 2000.
5. Michael C. McDermott, (mike@mcdltd.com), Tom. Owen, (owlmax@aol.com), Frank M. McDermott, Voice Identification: The Aural Spectrographic Method, Internet 1996.
6. MSDN,The Microsoft MSDN, http://msdn.microsoft.com/downloads/ 2002.
7. M. B. Priestley, Spectral Analysis and Time Series, Academic Press 1981.
8. D. R. Reddy , Speech Recognition, Academic Press 1975.
9. Majid Abdullah Sukker Shaalan, New High Synthetic coding Methods for Compressing Digital Speech Signals , M.sc Thesis, College of Science, University of Baghdad 2000.
10. Sahar Dakhel Al-Sudani, Speaker Recognition System, M.sc Thesis, College of Science, Technology University 1997.
11. مازن شاكر جاسم الزيوري، نظام حاسوبي لموائمة وتمييز الاصوات البشرية ، رسالة دكتوراه في الفيزياء، كلية العلوم، جامعة بغداد ٢٠٠٢.