Speech Recognition using Wavelets and Improved SVM

Emad Ahmed Hussien — Lecturer

Mohannad Abid Shehab Ahmed — Lecturer

Haithem Abd Al-Raheem Taha — Assistant Lecturer

Electrical Department, Engineering College, Al-Mustansirya University

Abstract:

Speaker recognition (identification/verification) is the computing task of validating a user's claimed identity using speaker specific information included in speech waves: that is, it enables access control of various services by voice.

Discrete Wavelet Transform (DWT) based systems for speaker recognition have shown robust results for several years and are widely used in speaker recognition applications.

This paper is based on text independent speaker recognition system that makes use of Discrete Wavelet Transform (DWT) as a feature extraction and kernel Support Vector Machine (SVM) approach as a classification tool for taking the decision through applying simplified-Class Support Vector Machine approach.

The proposed SVM approach can convert local Euclidean distances between frame vectors to angles by projecting these d-dimensional vectors together, and get the minimum global distance from the non-linear aligned speech path in order to address audio classification, and hence, sound recognition.

The DWT for each frame of the spoken word are taken as a tool for extracting the main feature as a data code vectors, next these data is normalized utilizing the normalized power algorithm that is used to reduce the number of feature vector coefficients then these data is scaled and tested with those stored of the training spoken words to achieve the speaker identification tasks, also the DWT gives fixed amount of data that can be utilized modesty by SVM.

Finally, the proposed method is tested and trained upon a very large data base with results limited to ten speakers only (5 males and 5 females) with words of maximally 17 phenomena and its performance gives an accurate and stable results which rises the algorithm efficiency and reduce the execution time with 97% overall accuracy.

Keyword: Discrete Wavelet Transform (DWT), Dynamic Time Warping (DTW), classifier, One-Class SVM, Speech Recognition.

الخلاصـــــــــة:

يعتبر تمييز المتكلم (تحديد الهوية والتحقق) من المهام الحسابية للتحقق من صحة المتكلم باستخدام معلومات موجات الكلام لنمكن الوصول للخدمات المختلفة بالصوت.

تحويل المويجات المتقطعة (DWT) تم استعمالها لاستخلاص مويجات الصوت وقد اثبتت كفائتها لسنوات عدة على نطاق واسع في تطبيقات التعرف على المتكلم.

في هذا العمل استعمل تحويل المويجات المتقطعة (DWT) مع نصوص الصوت المستقلة للمتكلم لاستخراج مميزات الصوت ثم استعمال ال (SVM) كمصنف ,حيث تم تطبيق تصنيف سهل (SVM) لقياس المسافة بين متجهات الصوت وتحويلها لزوايا ليسهل استخراج الاصغر بينهم رغم تباين مقايس الاداء للصوت.

يقوم ال (DWT) باستخراج ميزات كل الكلمة المنطوقة كمتجهات مصفوفة للبيانات ثم عمل تسوية (normalize) بالنسبة للقدرة لتقليل عدد معاملات البيانات التي ستدخل للمصنف (SVM) كأداة لاتخاذ القرار من خلال عمل التدريب ومعرفة الكلمات المطلوبة ، ومن سمات ال (DWT) هو استخراجه لكمية ثابتة من البيانات والتي يستفاد منها ال (SVM) على اعتبار سهولة عمل المصنفات مع البيانات الثابتة.

وأخيراً، تم اختبار الأسلوب المقترح بناء على قاعدة بيانات كبيرة جدا مع اخذ النتائج من عشرة متكلمين فقط (5 من الذكور و 5 من الإناث) مع كلمات أقصى حد لها 17 صوت للكلمة وترتفع كفاءة الخوارزمية ويقل وقت التنفيذ بنسبة 97%.

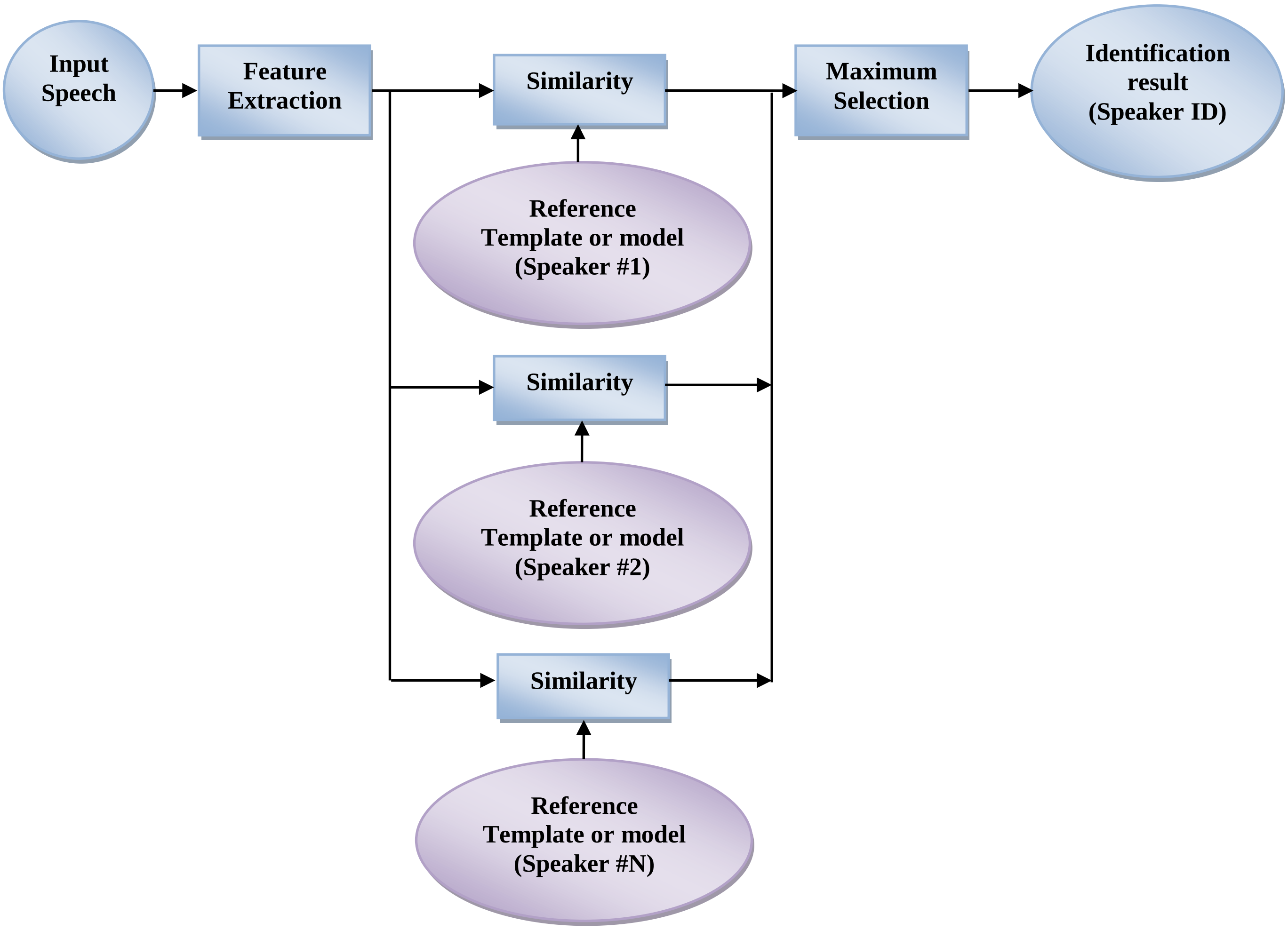
الكلمات الدليلية : تحويل المويجات المتقطعة ، ديناميكية التفاف الوقت (DTW) ، المصنف SVMمن الصنف الاول ، تمييز المتكلم

Introduction:

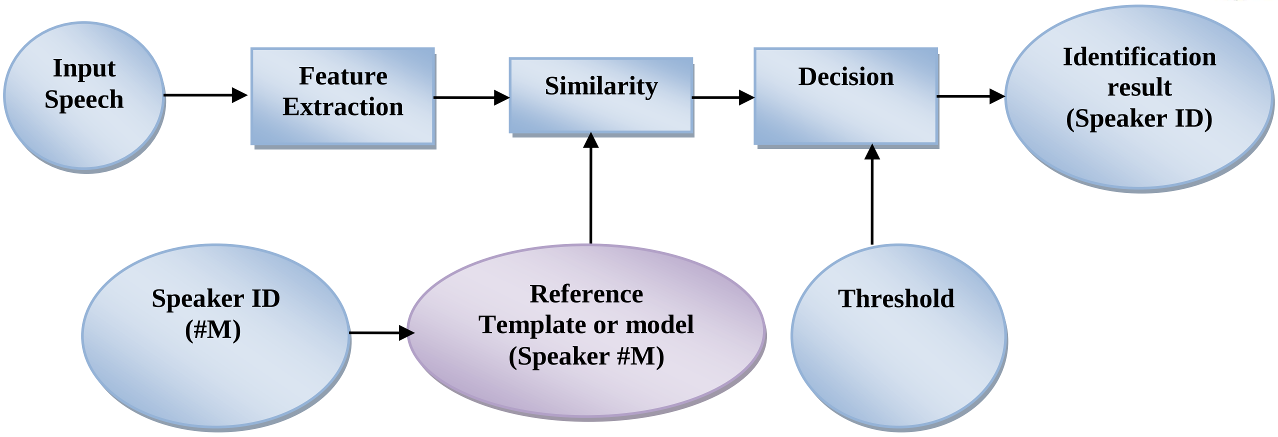
Voice recognition systems are, in general, very useful in many tasks, among those very important applications in our everyday life are secure telephony, voice-based login, and voice locks, they are also used as a security key, it can use the voiceprint of every human being, so voice recognition (both speech and speaker) plays its significant role in the field of human electronics and its wide applications [1].

Speaker recognition can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Figure (1) shows the basic structures of speaker identification and verification systems.

All speaker recognition systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.


Input Speech
Feature Extraction
Similarity
Maximum Selection
Identification result (Speaker ID)
Reference Template or model (Speaker #1)
Similarity
Reference Template or model (Speaker #2)
Similarity
Reference Template or model (Speaker #N)

(A) Speaker Identification


Input Speech
Feature Extraction
Similarity
Decision
Identification result (Speaker ID)
Speaker ID (#M)
Reference Template or model (Speaker #M)
Threshold

(B) Speaker Verification

Figure (1): basic structure of Speaker recognition systems

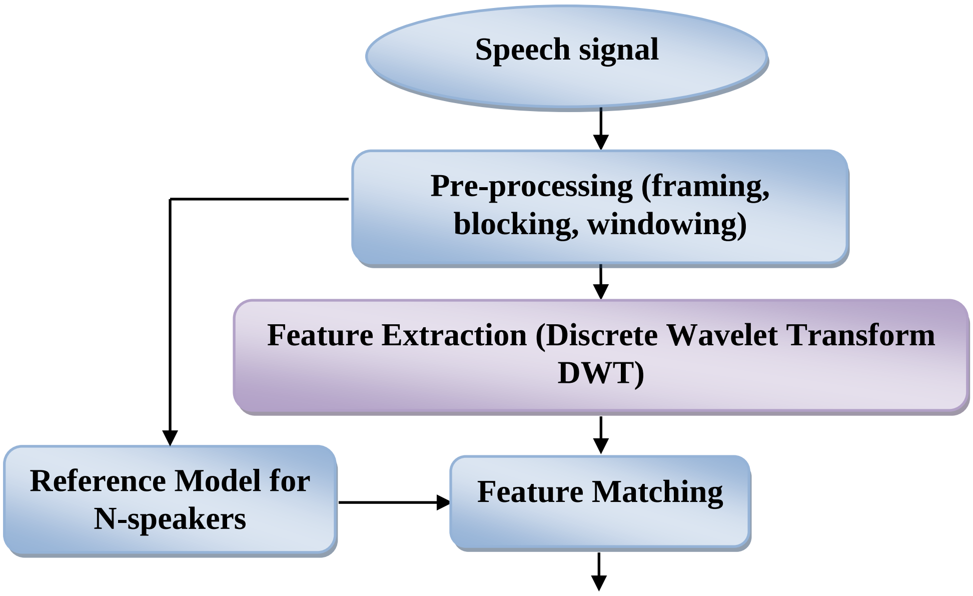
Speaker Recognition System:

Conclusive basic model of our speaker/voice identification system is as shown in figure (2).

In the pre-processing subsystem, the speech signal is first digitized by suitable sampling process, and the resulting vector is segmented into many frames to be ready for the feature extraction.

In the feature extraction subsystem, a set of essential characteristics (e.g., pitch, formant frequency, and energy profile) that can identify and represent the whole speech signal is measured. Here, the Discrete Wavelet Transform (DWT) is achieved to extract data vectors that are normalized by calculating the normalized power vectors of the specific speaker.

The classification subsystem involves the actual procedure to identify the spoken word and then takes the related true decision.


Speech signal
Pre-processing (framing, blocking, windowing)
Feature Extraction (Discrete Wavelet Transform DWT)
Reference Model for N-speakers
Feature Matching

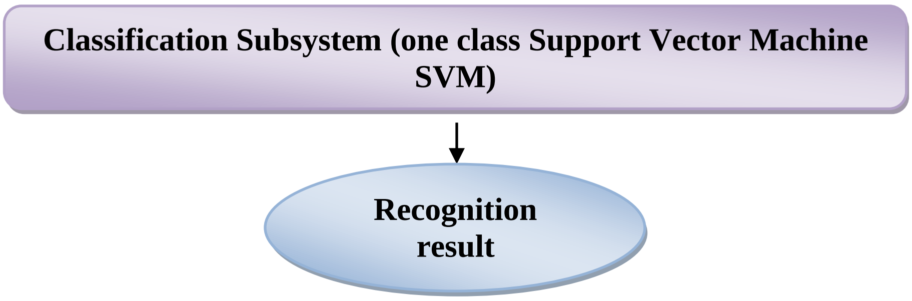

Classification Subsystem (one class Support Vector Machine SVM)
Recognition result

Figure (2): basic model of our speaker/voice identification system

Preprocessing:

The objective in the preprocessing is to modify the speech signal, so that it will be more suitable for the feature extraction analysis. The preprocessing consists of De-noising, Silence Removal, Pre-Emphasis, Framing 50% overlapped with 23ms frame time, and Windowing using Hamming window, the first three above steps are very useful for noisy speech.

Dynamic Time Warping:

Speech is a time-dependent process. Hence the utterances of the same word will have different durations, and utterances of the same word with the same duration will differ in the middle, due to different parts of the words being spoken at different rates. To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed. The best matching template is the one for which there is the lowest distance path aligning the input pattern to the template. A simple global distance score for a path is simply the sum of local distances that go make up the path.

The reference voiceprint with the lowest distance measure from the input pattern is the recognized word. The best match (lowest distance measure) is based upon Dynamic Time Warping (DTW).

A common task with continuous data is comparing one series with another. In the case where the time series have the same component shapes but do not match it must "warp" the time axis of one or both series to achieve better alignment. In our project case, the Time Warping is rarely used because the speech data spoken are saved in appropriate manner.

Discrete Wavelet Transform (DWT) [2,3]

Wavelet transform can be viewed as the projection of the signal into a set of basic functions named wavelets. Such basis functions offer localization in the frequency domain. Compare to STFT (Short Time Fourier Transform) which has equally spaced time-frequency localization, wavelet transform provides high frequency resolution at low frequency and high time resolution at high frequency.

The Discrete Wavelet Transform (DWT) of the signal x[n] is defined based on basic functions: scaling function (approximation coefficients) Φ(x) and wavelet function (detail coefficients) Ψ(x) as shown below:

$$\Phi[j,k] = \frac{1}{\sqrt{M}}\sum_{n} x[n]\Phi_{j_o,k}[n] \quad \ldots(1)$$

$$\Psi[j,k] = \frac{1}{\sqrt{M}}\sum_{n} x[n]\Psi_{j_o,k}[n] \quad \ldots(2)$$

Where:

n=0,1,2,…,M-1, j=0,1,2,..,J-1, k=0,1,2,…,2j-1 and M denotes the number of sampled to be transformed.

In the wavelet transform, it needs to decomposition process to get the detail filters coefficients as:

$$\mathrm{c}_{k,n/2} = \sum_{j=-l+n}^{n} a_{j-n}\, c_{k+1,j} \quad \ldots(3)$$

$$d_{k,n/2} = \sum_{j=-l+n}^{n} b_{j-n}\, c_{k+1,j} \quad \ldots(4)$$

Where:

n : is even and $-l_1 \le n \le l_2$.

a_k, b_k: the filter coefficients .

$c_{N,n}$: the signal .

$c_{k,n/2}$, $d_{k,n/2}$: the approximation and detail (with down sampling) at depth k .

Φ(x), Ψ(x) : scaling function, wavelet mother function .

l_1, l_2 : beginning and ending signal .

The form of DWT used in this work is the Daubechies wavelet transform, the filter bank structure is often used, the approximation coefficients at a higher level are passed through a high pass and low pass filter followed by down sampling by 2 to compute both the detail and approximation coefficients at low level, this tree structure is repeated for multi-level decomposition.

The decomposition process can be iterated with successive approximations being decomposed in turn, so that one signal is broken down into many lower resolution components. This is called the wavelet decomposition tree.

The wavelet decomposition of the signal 'S' analyzed at level 'j' has the following structure [cA_j, cD_j,..., cD_1]. Looking at a signals wavelet decomposition tree can reveal valuable information. The diagram of figure (3) shows the wavelet decomposition to level 3 of a sample signal 'S'.

In reality, the decomposition can only proceed until the vector consists of a single sample. Normally, however there is little or no advantage gained in decomposing a signal beyond a certain level. The selection of the optimal decomposition level in the hierarchy depends on the nature of the signal being analyzed like text, image, speech …etc or some other suitable criterion, such as the low pass filter cut off.

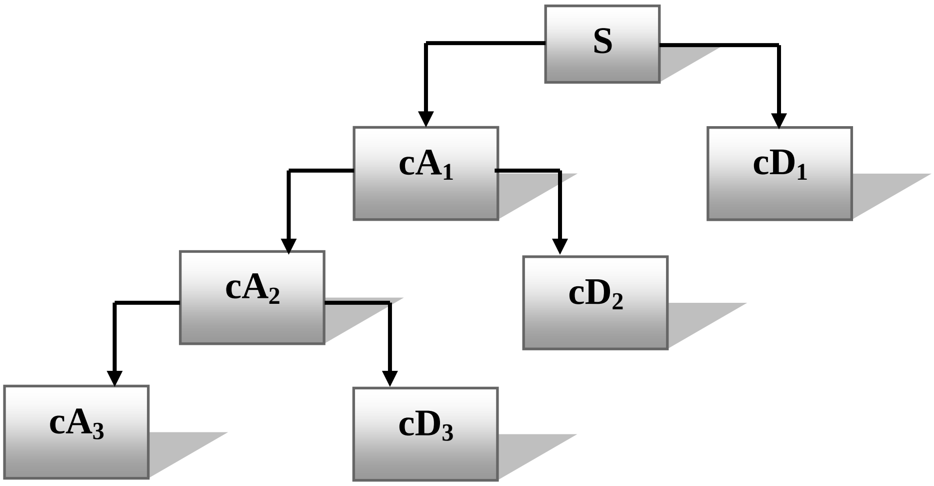


Figure (3): level 3 of the DWT

System Overview

The following paragraphs give the overall view to our project theory and calculations:

Preprocessing:

The preprocessing includes sampling, segmentation, framing, and windowing. The speech signal is sampled with sampling frequency of 11.25KHz, then the speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M<N), the second frame being M samples after the first frame, and overlaps it by N-M samples. Similarity, the third frame beings 2M samples after the first frame or (M samples after the second frame) and overlaps by N-2M samples. Typical values for N=256 (equivalent to 23 msec) and M=128.

Finally, the signal must be limited to finite range of time, one way to do this is by multiplying the signal by some function such function called window, here the Hamming window is used, which has the form:

$$w(n) = 0.54 - 0.46\ \cos\left(\frac{2n}{N-1}\right)$$

$$0 \leq \mathrm{n} \leq \mathrm{N} - 1 \qquad \ldots (5)$$

Dynamic Time Warping (DTW):

Frame distances between the processed speech frames and those of the reference templates are summed to provide an overall distance measure of similarity. But, instead of taking frames that correspond exactly in time, you would do a time “warp” on the utterance (and scale its length), so that similar frames in the utterance line up better against the reference frames. A dynamic programming procedure finds a warp that minimizes the sum of frame distances in the template comparison. The distance pro- before time warp amplitude produced by this warp is chosen as the similarity measure. In the illustration here, the speech frames that make up the test and reference templates are shown as scalar amplitude values plotted on graphs of figure (4) with time as the x-axis. In practice, they are multidimensional vectors, and the distance between them is usually taken as the Euclidean distance. The graphs show how warping one of the templates improves the match between them [4] .

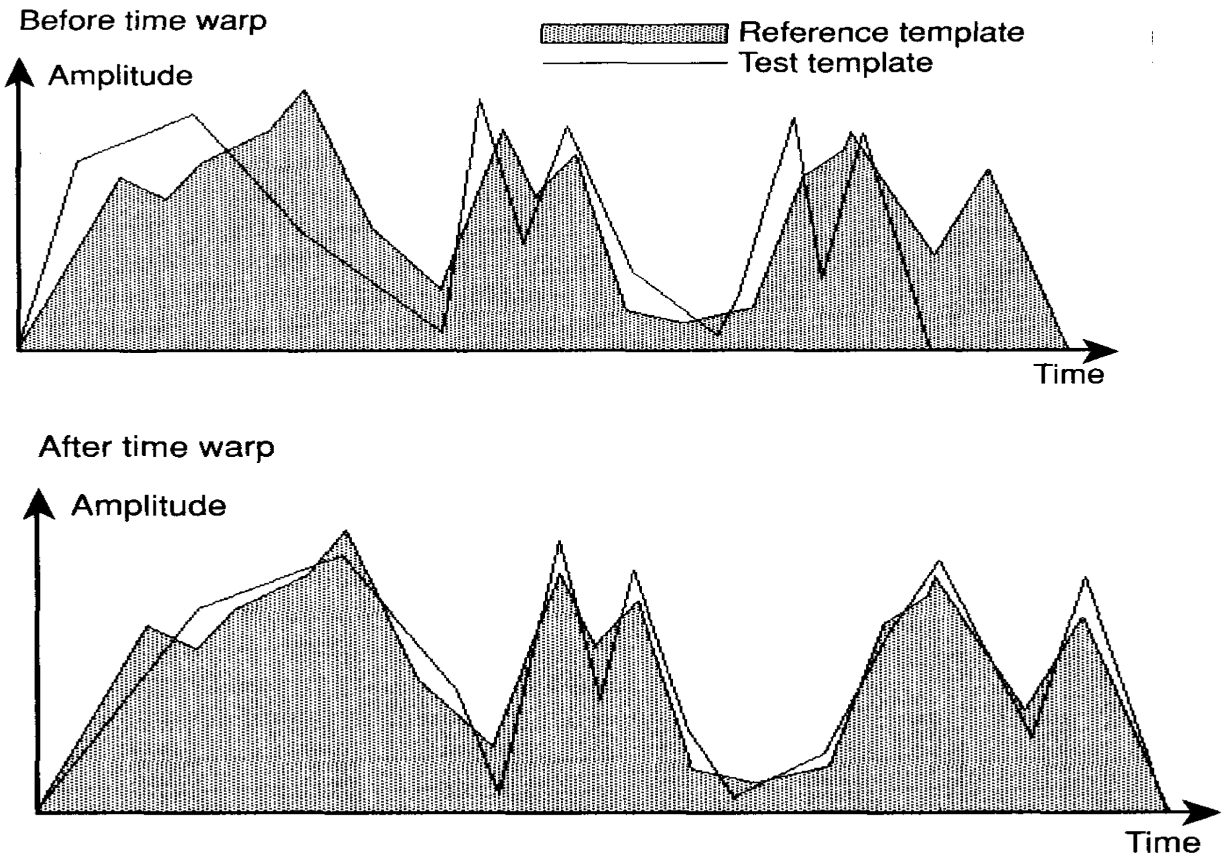


Figure (4): time dynamic warping

Feature Extraction:

The feature extraction of the speech signal is based on the DWT through getting the decomposition (low pass only and down sampling). The ability of DWT to extract features from the signal is dependent on the choice of the mother wavelet function and the level of the decomposition; here the mother function used is the Daubechies with 8-decompostion levels.

The steps of feature extraction are as follows:

Each frame of the spoken words is now expanded using the DWT up to 8 levels of decomposing, and the overall coefficients in each branch of decomposing is calculated according to equation (6) [5]:

$$floor\left(\frac{n-1}{2}\right) + N \qquad \ldots (6)$$

Where:

n: the number of input vector to the low pass filter .

N: the order of the wavelet type .

Computing the power in each segment of each level of the decomposition according to equation (7), the obtained feature vectors describe the power distribution over the time-frequency plane. This scale power density along every segment describes the power variation in each scale.

$$P_{norm} = \frac{\sum_{i=1}^{n} S_i{}^2}{n} \qquad \dots (7)$$

Where:

S_i : the element of the coefficient set .

n : the number of elements .

The variance of the segments overall power for each of the 8 levels is computed according to equation (8):

$$power\_var = var(P_{norm}) \qquad \dots (8)$$

Where:

var: the square of the standard deviation of the normalized power vector.

SVM Classification for Decision:

In many cases, SVMs outperform most state of the art classifiers. In recent years, many pattern recognition problems have been tackled using SVMs, ranging from computer vision to text classification with better performance than more traditional techniques [6] .

The one-class SVM approach (1-SVM approach) has been successfully applied to various learning problems that can halve the error rate of conventional phone-based approaches [7][8]. It consists of learning the minimum volume contour that encloses most of the data in a dataset.

General maximum margin classifiers are of increasing interest in ASR (Automatic Speech Recognition) due to their robustness against both sparse data and rapid transient changes in acoustic sequences, SVMs explicitly minimize a hypothesized upper bound on the expected classification error by orienting a hyper-plane between classes such that the norm of its orthogonal vector maximizes the margin between the nearest data. Let us assume training set of m examples $T_{train} = \{(x_i, y_i)\}_{i=1}^{m}$ where $x_i \in R^d$ are d-dimension input vector and $y_i \in \{-1, 1\}$ is the target class. Looking for parameters $(w \in R^d, b \in R)$, the simplest binary classifier one can think of is the linear classifier is

$y_i(x) = sign(w. x_i + b)$ …(9)

When the training set is said to be linearly separable, there is potentially an infinite number of solutions $(w \in R^d, b \in R)$ that satisfy equation (9). Hence, the SVM approach looks for the one that maximizes the margin between the two classes, where the margin can be defined as the sum of the smallest distances between the separating hyper-plane and points of each class.

This can be expressed by the following optimization problem:

$min_{w,b} \ \frac{1}{2} \|w\|^2$ …(10)

While it is difficult to solve equation (10), the following formulation is computationally more efficient:

$max_{\alpha} \{\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i x_j\}$ …(11)

Where most α_i are zero except those corresponding to examples in the margin or misclassified, often called support vectors (hence the name of SVMs).

One problem with this formulation in speech classifier is that if it not linearly separable, there might be no solution to it. Hence one can relax the constraints by allowing errors with an additional hyper-parameter C that controls the trade-off between maximizing the margin and minimizing the number of training errors, as follows:

$min_{w,b} \ \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$ …(12)

where C and ξ_i are called slack variables.

In order to look for nonlinear solutions, one can easily replace x by some nonlinear function $\phi(x)$, and hence generate kernel function $k(x_i, x_j)$ as

$k(x_i, x_j) = \phi(x_i). \phi(x_j)$ …(13)

We hope to use a soft margin SVM here and extend the process to k −class discrimination by training $k(k-1)/2$ binary classifiers, each delineating two class regions.

In speech recognition SVMs depend on kernel functions due to its non linearity data to describe the distance between two points of data.

The first common and popular kernel is a symmetric Radial Basis Function (RBF) k_{RBF} that generalizes to non-linear decision boundaries using the following function:

$$k_{RBF}(\mathrm{x_i},\mathrm{x_j}) = \exp(-\frac{(\mathrm{x_i}-\mathrm{x_j})^2}{2\sigma^2}) \qquad \ldots\ (14)$$

given vectors $\mathrm{x_i},\mathrm{x_j}$ and width parameter σ [9].

In this project we use another kernel k_{DTW} that is a sequence kernel which be generalized to arbitrary sequences u and v having non-equal lengths, this kernel exploits the notion of distance between sequences inherent in Dynamic Time Warping (DTW), and converts it to a form amenable for use in SVMs i.e. it can converting local Euclidean distances between frame vectors to angles by projecting these d-dimensional vectors onto a unit hyper-sphere H centered α units from their origin in the many dimension. Namely, every vector u_i is converted to the unit vector $\dot{u}_\imath$ sharing an origin with H by:

$$\hat{u}_i = \frac{1}{\sqrt{{u_i}^2+\alpha^2}}\begin{bmatrix}u_i\\ \alpha\end{bmatrix} \qquad \ldots(15)$$

Given two unit vectors, $\hat{u}_i$ and $\hat{v}_i$ that define points on the surface of H, the angle between them is by definition by:

$$d_s(\hat{u}_i\,,\hat{v}_i)= \theta_{\hat{u}_i\,,\hat{v}_i} = \cos^{-1}(\hat{u}_i\,,\hat{v}_i) \qquad \ldots\ (16)$$

Now, given these local distances, apply symmetric DTW on whole sequences u and v and get the minimum global distance from the non-linear aligned Viterbi path (speech path) Γ with

$$D_{global}(u,v) = min_{\Gamma}\,\frac{1}{\|\Gamma\|}\sum_{p=1}^{\|\Gamma\|} d_s(\hat{u}_i\,,\hat{v}_i) \qquad \ldots(17)$$

This distance is then converted to the kernel

$$k_{DTW}(u,v) = \cos(\,D_{global}(u,v)) \qquad \ldots(18)$$

which is symmetric if the symmetric version of DTW is used.

After the completion of programming the proposed SVM approach in Matlab software, a comparison with SVM package software called “Spider” which depends on kernel quadratic programming is made, leading us that this approach of kernel is very good for ASR data and convenient to small data and hence small storage.

The flow chart of the proposed speaker identification system is presented in figure (5):

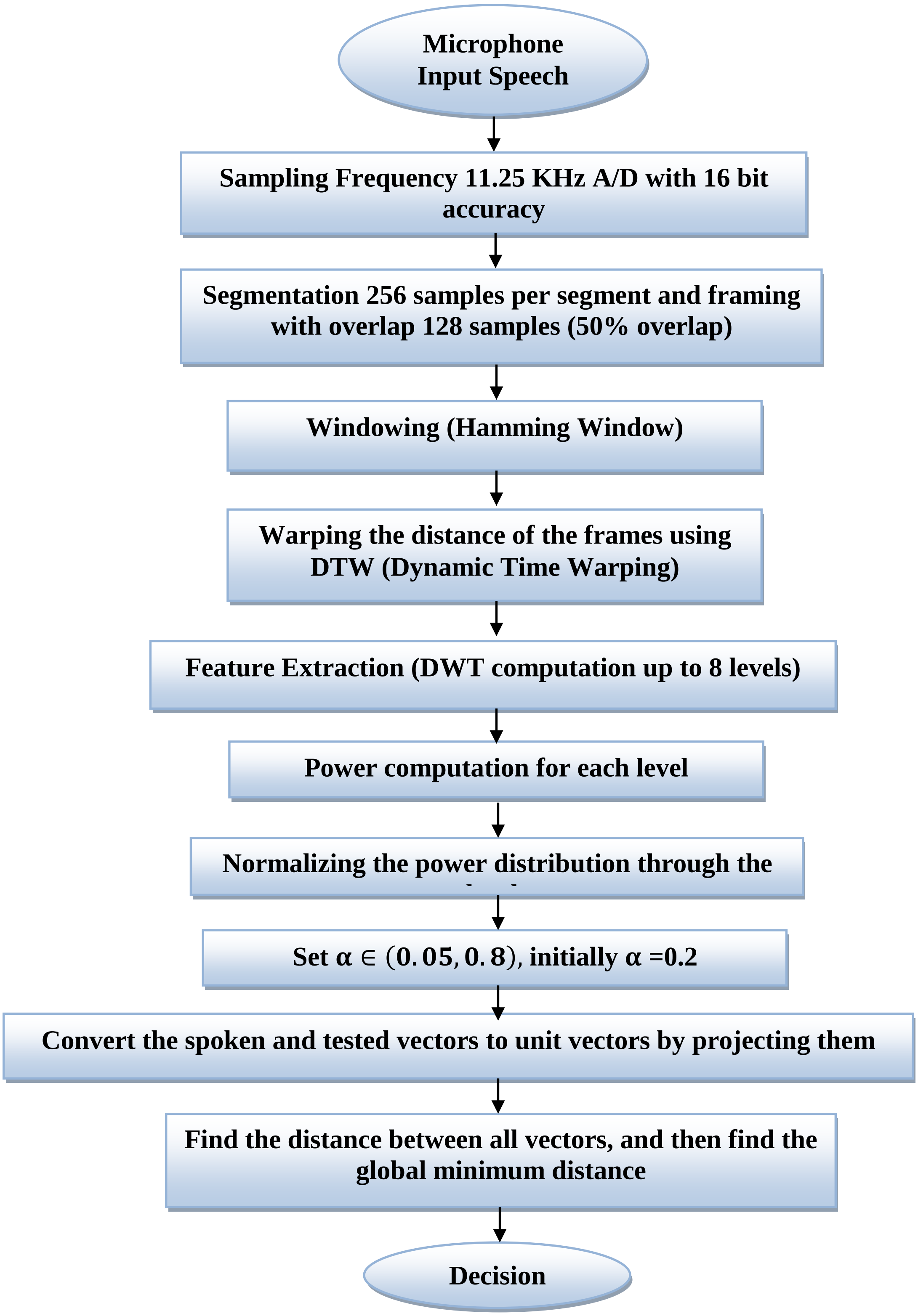

Microphone Input Speech
Sampling Frequency 11.25 KHz A/D with 16 bit accuracy
Segmentation 256 samples per segment and framing with overlap 128 samples (50% overlap)
Windowing (Hamming Window)
Warping the distance of the frames using DTW (Dynamic Time Warping)
Feature Extraction (DWT computation up to 8 levels)
Power computation for each level
Normalizing the power distribution through the
Set α ∈ (0.05, 0.8), initially α =0.2
Convert the spoken and tested vectors to unit vectors by projecting them
Find the distance between all vectors, and then find the global minimum distance
Decision

Figure (5): overall system flow chart

Simulation Result and Evaluation Test

The text spoken by each person (half number of the tested peoples is males and the second half is females) with different ages in order to check the algorithm fidelity is splitted into a tested word “يمين”.

The word “يمين” is divided into four characters namely “ي” , “م”, “ي”, “ن” and phenomena “ya”, “m”, “ee”,”n”, the word segmentation is as shown in figure (6).

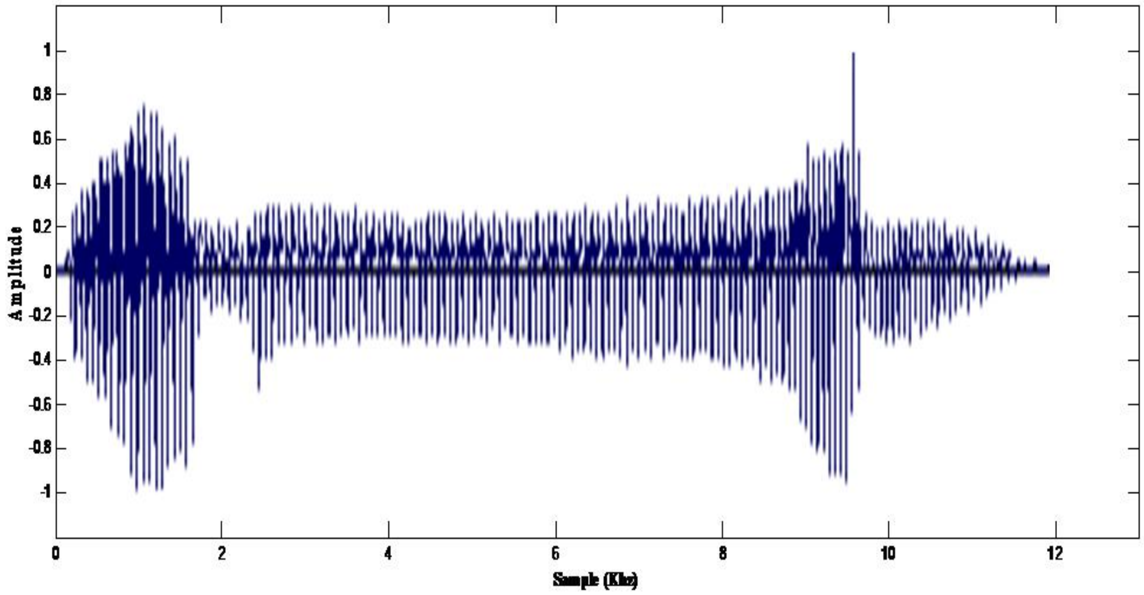


Figure (6): spectrum of word “يمين"

The normalized length of the word “يمين” for ten speakers of the noiseless speech is as shown in table (1).

Table (1): normalized length of word “يمين” for noiseless

Speaker	sex	From the tested word				From the spoken text			
		ي 1	م 2	ي 3	ن 4	ي 1	م 2	ي 3	ن 4
Sp1	m	0.65899	0.38710	1.00000	0.65899	0.68203	0.37327	1.00461	0.65438
Sp2	m	0.65899	0.38710	0.96774	0.65438	0.64516	0.39171	0.96774	0.64516
Sp3	m	0.66820	0.37788	0.96774	0.69124	0.66820	0.37327	0.95853	0.65438
Sp4	m	0.65899	0.38710	0.95853	0.66820	0.65438	0.38710	0.95853	0.66820
Sp5	m	0.66820	0.40092	1.00922	0.64516	0.64516	0.40092	1.00000	0.66820
Sp6	f	0.68203	0.40092	1.00000	0.68203	0.67281	0.39171	1.00000	0.68664

Sp7	f	0.65438	0.40092	1.00000	0.66820	0.65438	0.40092	1.00000	0.66820
Sp8	f	0.66820	0.39631	0.98618	0.65899	0.66820	0.38710	1.00922	0.65899
Sp9	f	0.67281	0.40092	1.00000	0.68203	0.68203	0.40092	1.00922	0.68203
Sp10	f	0.66820	0.40092	1.00000	0.66820	0.66820	0.40092	1.00000	0.66820

Where m & f denote to male & female

Figure (7) shows a three dimensional surfaces which represent the phoneme length relation for training and testing of the noiseless word speech.

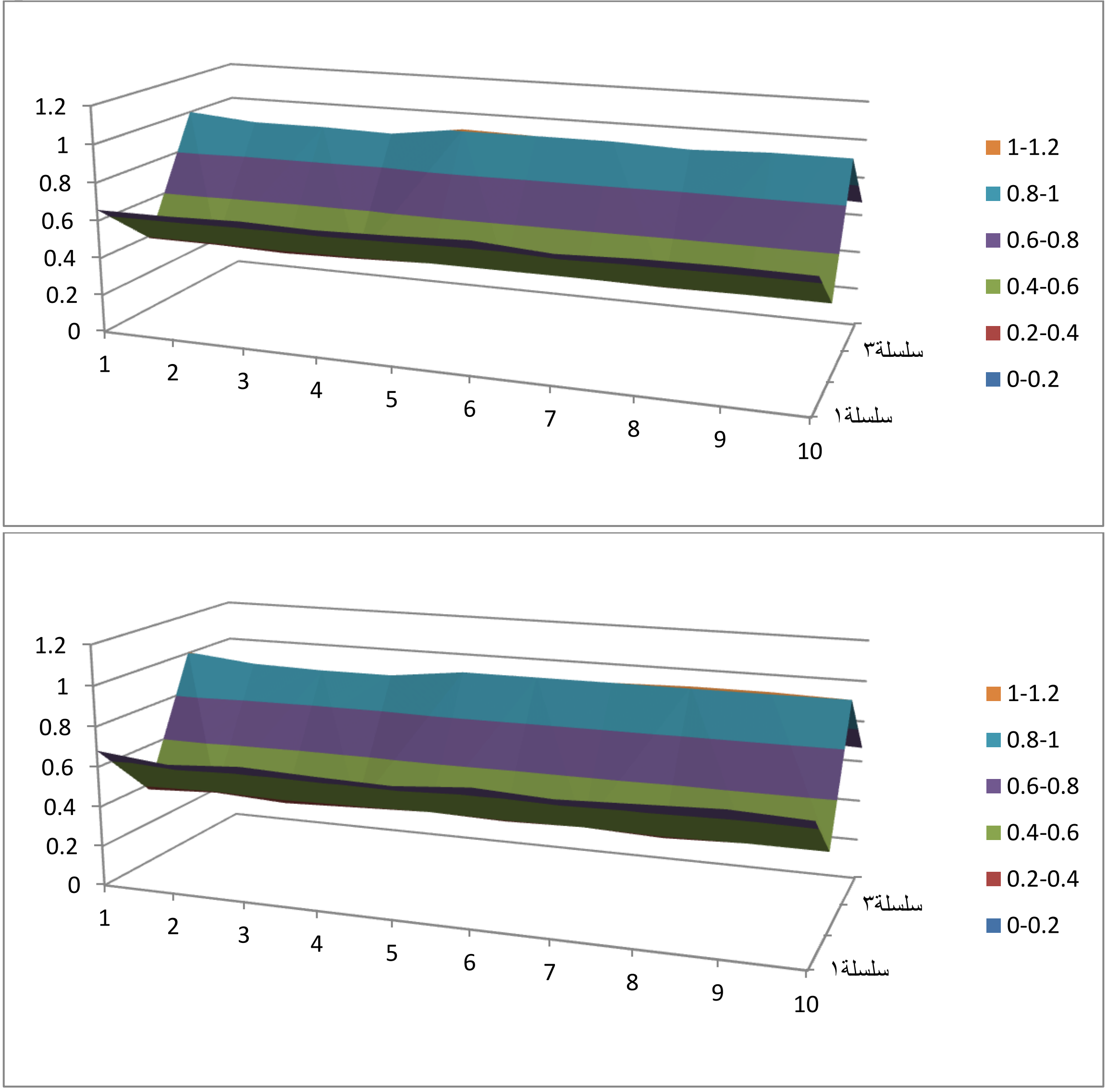


Figure (7): three dimensional surfaces for noiseless word speech

The normalized length of the word “يمين” for ten speakers of noisy speech with SNR of 30 dB as shown in table (2).
The purpose of above figures is to provide an overall looking to the match and mismatch between tested and spoken words.

Table (2): normalized length of word “يمين” for SNR=30 dB

Speaker	sex	From the tested word				From the spoken text			
		ي 1	م 2	ي 3	ن 4	ي 1	م 2	ي 3	ن 4
Sp1	m	0.59447	0.37327	0.94931	0.64516	0.61290	0.38710	0.92166	0.62673
Sp2	m	0.59447	0.37327	0.94931	0.64516	0.59908	0.37327	0.92166	0.64055
Sp3	m	0.64516	0.37788	0.94931	0.64055	0.64055	0.37327	0.94009	0.64516
Sp4	m	0.66820	0.38249	0.93548	0.64516	0.65899	0.40092	0.96774	0.65899
Sp5	m	0.56682	0.36406	0.80184	0.38710	0.58065	0.35023	0.87558	0.41935
Sp6	f	0.59908	0.37788	0.95853	0.65438	0.59908	0.37327	0.95853	0.65438
Sp7	f	0.65899	0.38249	0.94470	0.53917	0.65438	0.39171	0.98618	0.55300
Sp8	f	0.66820	0.40092	0.94931	0.60369	0.64055	0.37327	0.98157	0.59908
Sp9	f	0.66820	0.40092	1.00000	0.65438	0.66359	0.38249	0.99539	0.65899
Sp10	f	0.64516	0.40092	0.95392	0.61751	0.64516	0.38710	0.96313	0.63134

Figure (8) shows a three dimensional surfaces which represent the phoneme length relation of the SNR=30 dB word speech.

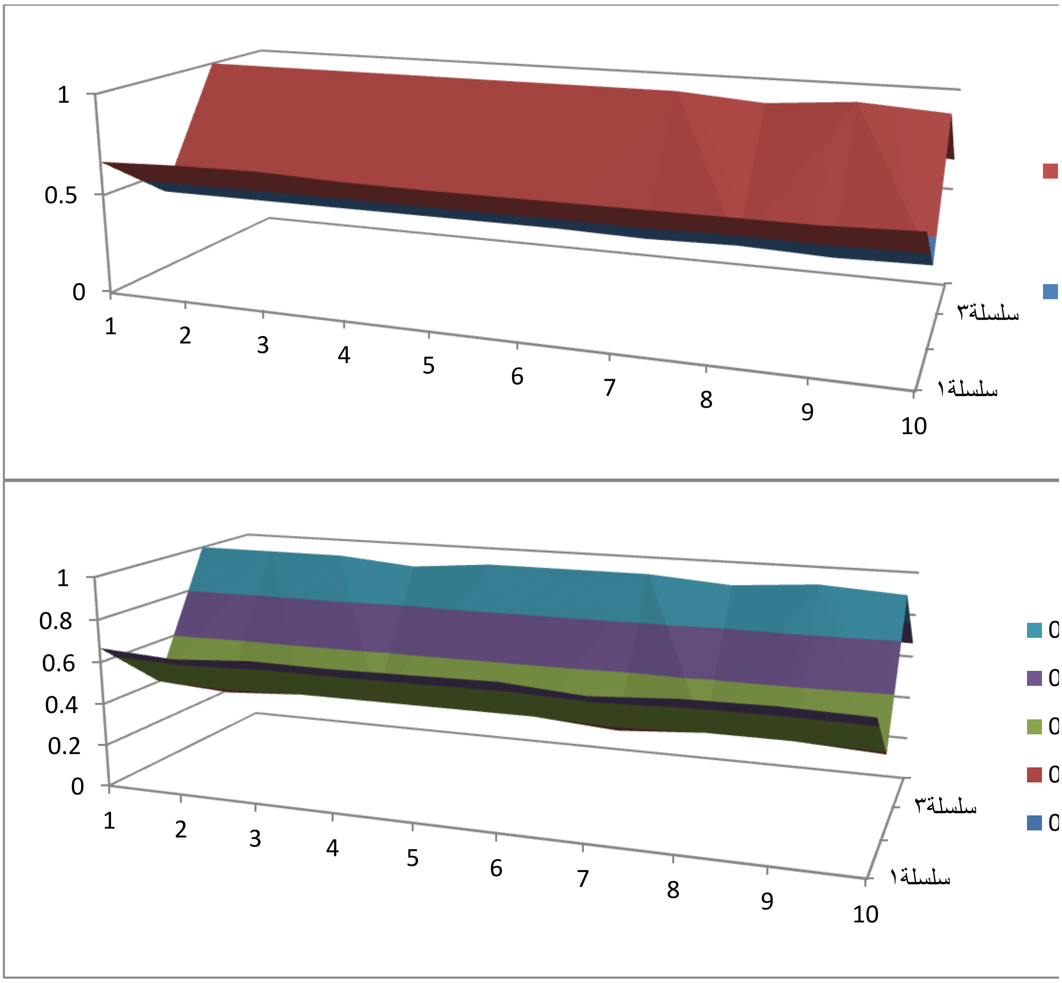


Figure (8): three dimensional surfaces for SNR=30 dB word speech

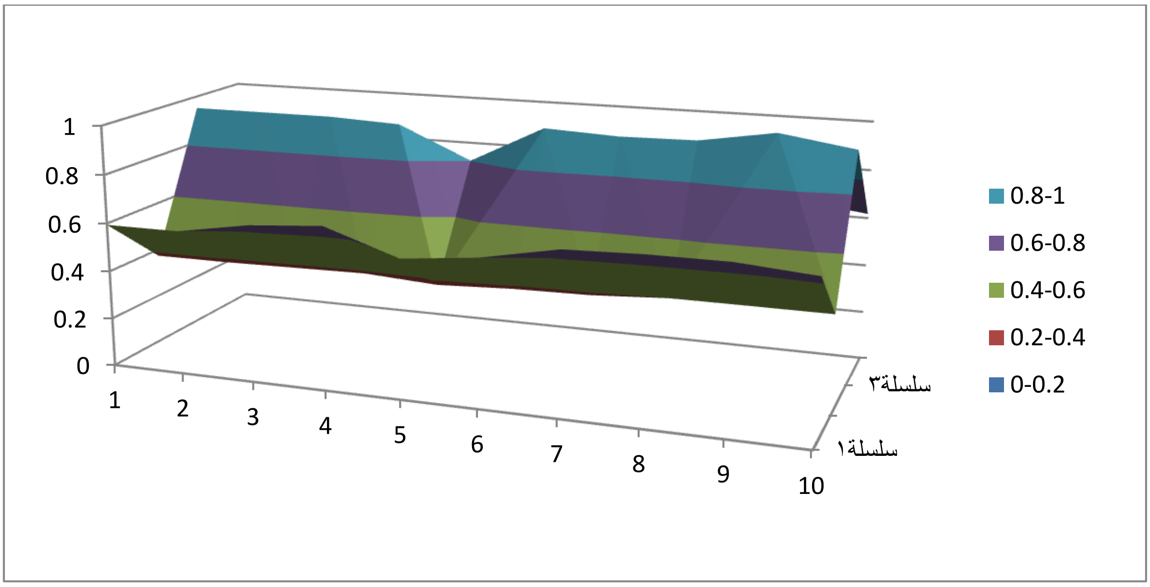
The normalized length of the word “يمين” for ten speakers of noisy speech with SNR of 20 dB as shown in table (3).

Table (3): normalized length of word “يمين” for SNR=20 dB

Speaker	sex	From the tested word				From the spoken text			
		ي 1	م 2	ي 3	ن 4	ي 1	م 2	ي 3	ن 4
Sp1	m	0.66820	0.40092	1.00000	0.66820	0.66820	0.40092	1.00000	0.66820
Sp2	m	0.68203	0.40092	1.00000	0.65438	0.64516	0.37788	1.00000	0.64977
Sp3	m	0.68664	0.40092	1.00000	0.68664	0.66820	0.40092	1.00000	0.68203
Sp4	m	0.67281	0.40092	1.00000	0.66820	0.66359	0.40092	0.96774	0.66820

Sp5	m	0.66820	0.40092	1.00000	0.67281	0.66820	0.40092	1.00000	0.66820
Sp6	f	0.66820	0.40092	1.00000	0.67281	0.67281	0.40092	1.00000	0.66820
Sp7	f	0.66820	0.39171	1.00000	0.65899	0.64516	0.37327	1.00000	0.66820
Sp8	f	0.66820	0.40092	0.96774	0.67281	0.66820	0.40092	0.97235	0.66820
Sp9	f	0.66820	0.38710	1.00000	0.67281	0.67281	0.40092	1.00000	0.66820
Sp10	f	0.68203	0.39631	0.97235	0.64977	0.66359	0.38249	0.97696	0.66359

Figure (9) shows a three dimensional surfaces which represent the phoneme length relation of the SNR=20 dB word speech.



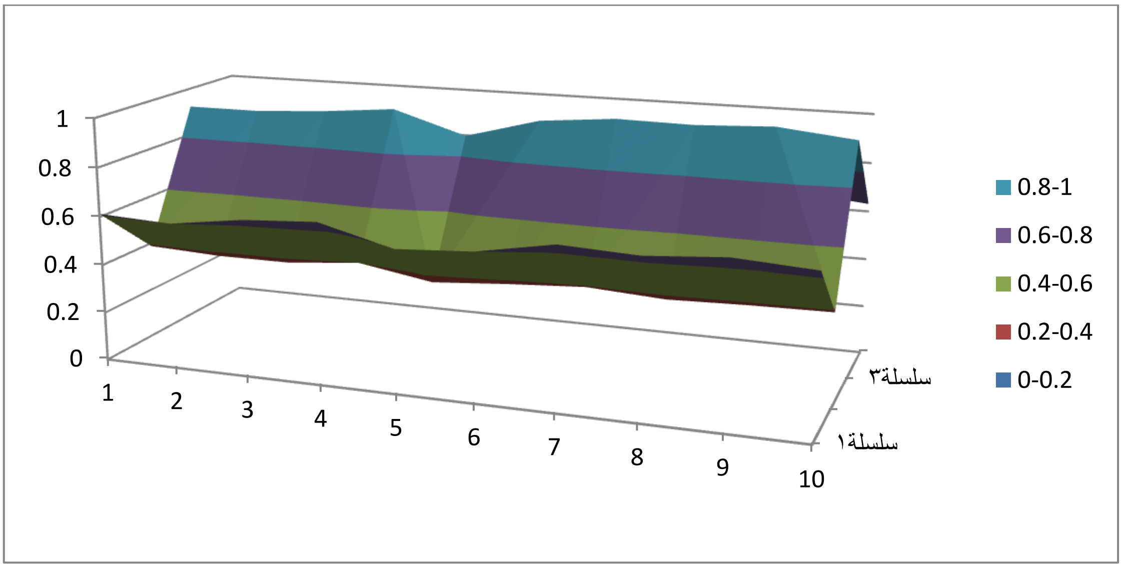


Figure (9): three dimensional surfaces for SNR=20 dB word speech

The normalized length of the word “يمين” for ten speakers of noisy speech with SNR of 10 dB as shown in table (4).

Table (4): normalized length of word “يمين” for SNR=10 dB

Speaker	sex	From the tested word				From the spoken text			
		ي 1	م 2	ي 3	ن 4	ي 1	م 2	ي 3	ن 4
Sp1	m	0.66820	0.40092	1.00000	0.75576	0.65438	0.40092	0.97235	0.75576
Sp2	m	0.67281	0.40092	1.00000	0.65438	0.64516	0.39171	0.96774	0.66820
Sp3	m	0.66820	0.40092	1.00000	0.66820	0.66820	0.40092	1.00000	0.66820
Sp4	m	0.66820	0.40092	1.00000	0.66359	0.66820	0.40092	1.00000	0.66820
Sp5	m	0.66820	0.40092	1.00922	0.67281	0.67281	0.40553	0.99539	0.66359
Sp6	f	0.67281	0.40092	1.00000	0.66820	0.66820	0.40092	1.00000	0.66820
Sp7	f	0.63134	0.36406	0.87097	0.55760	0.62673	0.82949	0.88018	0.55300
Sp8	f	0.65438	0.41014	0.96774	0.66820	0.65438	0.38710	0.96313	0.64516
Sp9	f	0.67281	0.40092	1.01382	0.68664	0.68203	0.40092	1.01382	0.69124
Sp10	f	0.67281	0.40092	1.00000	0.65438	0.64516	0.40092	1.00000	0.65438

Figure (10) shows a three dimensional surfaces which represent the phoneme length relation of the SNR=10 dB word speech.

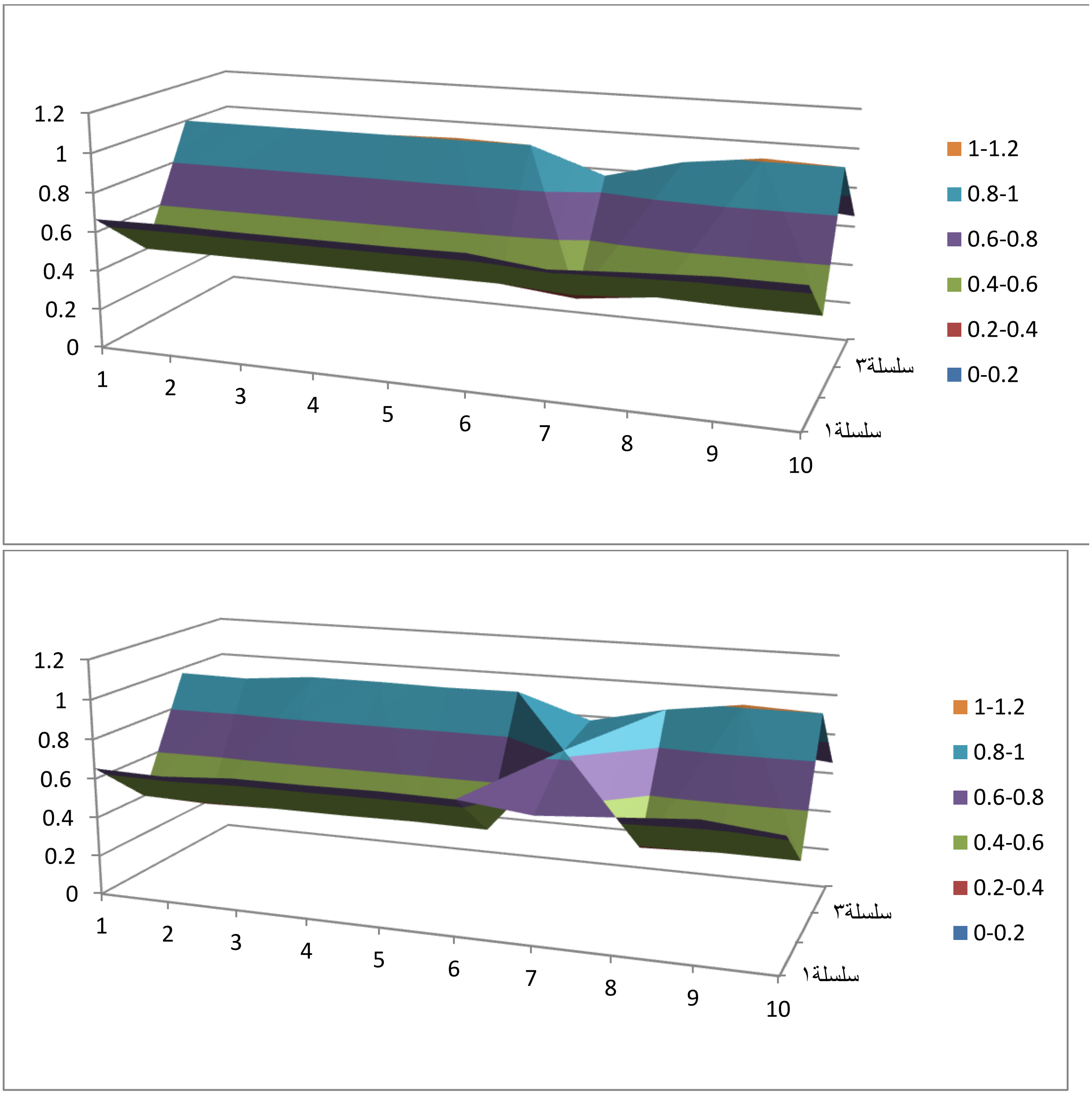


Figure (10): three dimensional surfaces for SNR=10 dB word speech

In all previous tables, numbers are of comparable values, converting them to angles of large values leads to ease in finding the global minimum distance; this is the summary of our proposed kernel.

The combined DWT and SVM approach is applied to many sentences with isolated and continuous words which lead to accuracy of 97% for isolated and of 90% for continuous words. The number of the correct identified speakers for ten persons for both noiseless and noisy speech of SNR 30, 20, 10 dB with 20 sec training time can be shown in figure (12).

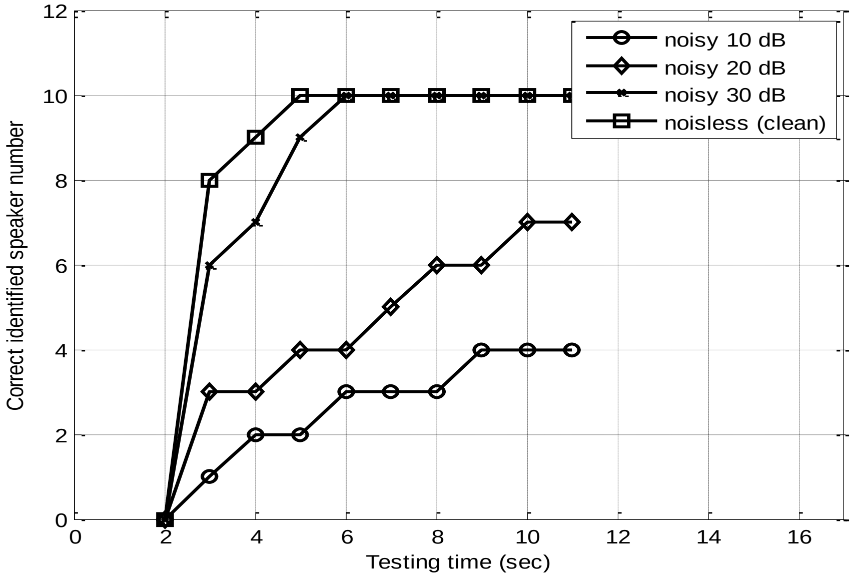


Figure (11): number of correct identifier to all types

It can be noted that almost type of noisy speech, the speaker can be identified correctly after 10 sec of testing time for 20 sec training time.

It can also compare the SVM feature classification method response to that of Gaussian Mixture Mode (GMM) as shown in figure (13), which leads to little needing to testing time for same speaker identifier or less correct identifier speaker number for same time in the optimization approach in the SVM to that of statistical approach in the GMM.

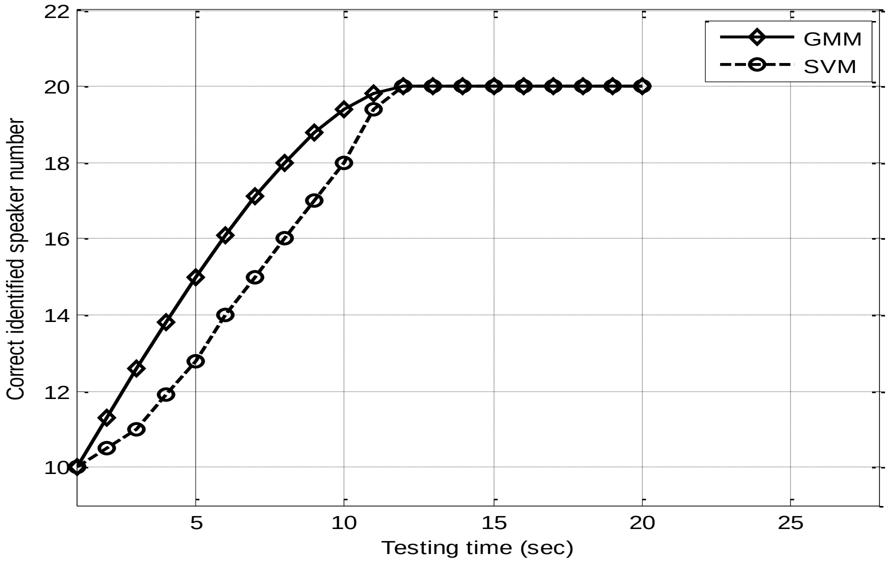


Figure (12): comparison between the SVM and GMM feature classification methods

Conclusion and future works:

- The previous results show that multi-level discrete wave transform (as a feature extraction) and SVM (as a classification) is valid and good approaches for speaker identification method system in both simulation and real world.
- The Daubechies (Db8) filter banks is chosen in the processing phase, this wavelet transform have some characteristics that are useful for speaker recognition like the simplicity and the fast processing. The way of selection of the basis function of the (Db8) is done experimentally rather than theoretically.
- The proposed kernel is suitable to data generated from DTW (Dynamic Time Warping) that really applied to the symmetric voice generated from wavelet transform to improve the ASR response.
- Since ASR needs to non linear classifier type, it is clear needing to kernel in SVM. In our approach converting the distance between two vectors (tested words and spoken words) to angle allow us to find the minimum global distance clearly and convenient to small data values.
- As a result of slower process time and the less memory consumption, the proposed method achieves the speaker identification task with lower number of calculations as compared to those methods that depends on the statistically calculations with probability like GMM (Gaussian Mixture Mode), but SVM needs to large data base.
- Using the wavelet transform yields fixed and stable extracted values and the SVM yields improving and raising the algorithm efficiency.
- The proposed method has been achieved upon a very large data base showing only that data consists of ten speakers (5 males and 5 females) with accurate and numerically stable results in the case of Text-Independent.

Future work can include:

- Combining the proposed kernel type with that of RBF to study the changing of some parameters that affect on the SVM response like σ (the Kernel width) and ν (the fraction of support vectors) to get optimum process time and efficiency.

- Mixing the obvious methods with other reinforced methods like adaptive approaches, HMM, GMM, MLLR, CMLLR, etc.
- Speech compression is the technology of removing the redundancy between neighboring samples of a speech signal and/or between the next cycles to get the speaker identification over IP (Internet Protocol).
- Also it can developing projects and applications that based on the speech real time actual works like **emotion recognition** and **lie detector** circuit and simulators.

References

1. Khalid Saeed, Member, IEEE, and Mohammad Kheir Nammous, **“A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image”**, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, VOL. 54, NO. 2, APRIL 2007.
2. Meysam Mohamad pour, Fardad Farokhi, **“An Advanced Method for Speech Recognition”**, World Academy of Science, Engineering and Technology, Vol. 25, 2009.
3. Nitin Trivedi, Sachin Ahuja, Dr. Vikesh Kumar, Raman Chadha, Saurabh Singh, **“Speech Recognition by Wavelet Analysis”**, International Journal of Computer Applications (0975 – 8887),Volume 15– No.8, February 2011.
4. Toni Giorgino, **“Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package”**, University of Pavia, 2009.
5. Siwar Rekik, Driss Guerchi, Habib Hamam & Sid-Ahmed Selouani, **“Audio Steganography Coding Using the Discrete Wavelet Transforms”**, International Journal of Computer Science and Security (IJCSS), Volume (6), 2012.
6. A. Rabaoui , M. Davy, S. Rossignol, Z. Lachiri, N. Ellouze**, “Improved One-Class SVM Classifier for Sounds Classification”**, Campus University, Lille France.
7. W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, **“Phonetic Speaker Recognition with Support Vector Machines”**, MIT Lincoln Laboratory.
8. Asma Rabaoui, Manuel Davyy, St´ephane Rossignoly, Zied Lachiri and Noureddine Ellouze, **“Using One-Class SVMs and Wavelets for Audio Surveillance Systems”**, Campus University, Lille France.

9. Joseph Keshet, and Samy Bengio, **“Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods”**, John Wiley & Sons, 2009.