KINETICS OF THE DARK REACTION $NO+NO_2+H_2O \rightleftarrows 2HNO_2$ IN THE GAS PHASE

Hamza K. Asar*

Research Institute
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia

الخلاصـــة :

تمت دراسة التفاعل الذي يحدث بين ثاني أكسيد النتروجين وأول أكسيد النتريك والماء لانتاج حمض النتروز في الظلام حيث ان هذا التفاعل يلعب دورا هاما في تفاعلات الضباب الدخاني لانجاز التفاعل ثم وضع مفاعل من (التفلون) في داخل غرفة اختبار بيئية محكمة . وكان تركيز ثاني أكسيد النتروجين يتبع كدالة من الوقت - باستخدام الاساليب الكيميائية المتوفرة وغير المكلفة . وقد أجريت هذه التفاعلات في درجات حرارة تتراوح بين -٢ درجة الى ١٥ درجة مئوية وتحت ظروف الضغط الجوي العادي . وتمت المحافظة على التركيزات الابتدائية للمزيج المتفاعل طوال فترة الدراسة التجريبية بمعدل ١٣٢ جزء من المليون لكل من أول أكسيد النتريك وثاني أكسيد النتروجين إضافة الى رطوبة نسبية بمعدل ٥٠٪ وما تبقى من الغاز هو نتروجين .

تمكنا من الحصول على تعبير Arrhenius الذي يتحدد تجريبيا والذي تحدد للمعدل الثابت للتفاعلات شبه الثنائية بالمعادلة الآتية .

$$k_2 = 9.9\times10^2 \exp(-1250/RT)\ \mathrm{l\,mol^{-1}\,s^{-1}},$$

والآلية المحتملة للتفاعل ستكون على النحو التالي :

(التوازن السريع) : $NO+NO_2 \underset{k_1}{\overset{k_1}{\rightleftarrows}} N_2O_3$

(محدد معدل التفاعل) : $N_2O_3+H_2O \underset{k_2}{\overset{k_2}{\rightleftarrows}} 2HNO_2$

*Address for correspondence :
KFUPM Box No. 1819
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia

ABSTRACT

The dark reaction between nitrogen dioxide (NO_2), nitric oxide (NO), and water to produce nitrous acid (HNO_2), which plays an important role in smog reactions, was investigated. A teflon bag reactor was placed inside a controlled environmental test chamber. The concentration of NO_2 was followed as a function of time using wet-chemical methods which are readily available and inexpensive.

The dark reactions were conducted in the temperature range of $-2°$ to $+15°C$ at a pressure of one atmosphere. Initial concentrations of reactant mixtures were maintained throughout the experimental study at 132 ppm for both NO and NO_2, and 50% relative humidity. The remaining gas was nitrogen.

The experimentally determined Arrhenius expression for the pseudo second-order rate constant of the rate-determining reaction step is given by:

$$k_2 = 9.9 \times 10^2 \exp(-1250/RT)\ \mathrm{l\,mol^{-1}\,s^{-1}},$$

and a tentative mechanism for the reaction is:

$$NO + NO_2 \underset{k_1}{\overset{k_1}{\rightleftharpoons}} N_2O_3 \qquad \text{(fast equilibrium)}$$

$$N_2O_3 + H_2O \underset{k_2}{\overset{k_2}{\rightleftharpoons}} 2HNO_2 \qquad \text{(rate-determining)}$$

Hamza K. Asar

KINETICS OF THE DARK REACTION $NO+NO_2+H_2O \rightleftarrows 2HNO_2$ IN THE GAS PHASE

1. INTRODUCTION

One of the important dark reactions of the oxides of nitrogen in the atmosphere is the formation of nitrous acid (HNO_2) by the overall scheme:

$$NO_2 + NO + H_2O \rightleftarrows 2HNO_2.$$

During the night, small amounts of HNO_2 are presumed to be formed, even when the concentration of nitrogen oxides are a few tenths of a ppm. When the sun rises, the HNO_2 is irradiated [1] and reacts according to the overall scheme:

$$HNO_2 + h\nu(\text{below } 4000\text{ Å}) \rightleftarrows NO + OH\cdot .$$

The OH· radical produced is extremely reactive and is considered to be one of the key intermediates in photochemical smog reactions.

Even though the HNO_2 formation reaction is considered to be important in smog formation, only a few other studies have been reported [2–8]. None of these investigators studied the overall reaction in the dark; and furthermore, none determined an activation energy. Some researchers, for example [2], used a large excess of nitric oxide NO and varied the concentration of NO_2 and water vapor between 0.5 and 2.4 mole percent. The reaction was conducted at 25°C and at 1.0 atmosphere (101 325 Pa). They found a value for the equilibrium constant of $1.65\ atm^{-1}$ and a rate constant for the disappearance of NO_2 of $4.3\times10^{7} l^2 mol^{-2} s^{-1}$. It is this value of the rate constant which most air pollution researchers have been relying upon to estimate the rate of HNO_2 formation when modeling smog reactions, even though it is only known at one temperature. Unfortunately, such an approach is limited, because it does not include the contribution of the dark reaction which may be significant in determining the initial concentration of HNO_2 in the photochemical formation of OH· radicals [9]. Thus there is a pressing need to investigate the contribution of these dark reactions.

In the present research, the dark reaction was studied in order to overcome the above limitation. The effect of temperature on the rate constant was studied, and the Arrhenius rate law was established. From this, the activation energy was calculated. Throughout the course of this study, the concentration of water vapor was kept essentially constant (at 50% relative humidity based on 23°C). The initial concentrations of NO and NO_2 were kept constant at 132 ppm each.

Wet-chemical methods invented by Saltzman [10, 11] for the analysis of NO_2 were used. The Saltzman reagent, which is specific for NO_2 (color change), is a mixture of sulfanilic acid, acetic acid *N*-(1-naphthyl)-ethylenediamine dihydrochloride, and distilled water. When NO_2 is absorbed in it, the reagent changes color from colorless to pink. The intensity of the color is directly proportional to the amount of NO_2 absorbed.

Wall effects [5], which are important in reaction kinetics, were minimized by using a teflon bag reactor. The teflon bag had walls thick enough to minimize diffusional exchange with the surroundings.

2. EXPERIMENTAL APPARATUS

2.1. Environmental Test Chamber

An environmental test chamber was used to house the teflon reactor bag. The two cubic foot chamber was equipped with heating and refrigeration assemblies to control the temperature in the range −5°C to 150°C to within ±0.5°C.

The test chamber was also equipped with a large vacuum pump and an altimeter. The chamber could be evacuated to a pressure corresponding to an altitude of 200 000 feet. A fan was located inside the test chamber near the large and trim heaters to maintain a uniform temperature inside the test chamber.

2.2. Teflon Bag Reactor

A teflon bag, 45.5 cm long × 45.4 cm wide with a wall thickness of 0.005 cm, was used to conduct the constant volume (and essentially constant pressure) batch reaction. Teflon was chosen for its inertness, so that wall effects are minimized. Transport of gases through the walls of the bag depends on the type of gas, its concentration, the temperature, and the type of plastic material used [12, 13].

Following a procedure suggested in the literature [12], the bag was first conditioned to reduce diffusional losses. Adsorption of NO_2 on the bag

walls helps in blocking the movement of NO_2 and NO through the polymer structure. Mass diffusivity of NO through teflon had been estimated to be $2.72 \times 10^{-7} cm^2 s^{-1}$. Based on this extremely low value, diffusional losses were neglected.

Since the reaction between NO, NO_2, and H_2O is exothermic, a temperature rise as a result of the reaction would be expected. This adiabatic temperature rise, ΔT, was calculated for a concentration of NO and NO_2 of 50 parts per million (ppm) each. The increase was calculated to be very small, $\Delta T = 0.0006°C$, and essentially undetectable. It is reasonable, therefore, to assume that the temperature of the reaction gas mixture inside the teflon bag was the same as the temperature inside the test chamber.

2.3. Analytical Equipment

A colorimeter was used to determine the concentration of nitrogen dioxide from the colored Saltzman reagent. The colorimeter had an operating range of wavelengths of 325 to 625 nm. Special scratch-free test tubes were used to hold the color-developed reagents during the absorbance measurements.

Temperature inside the test chamber was monitored using a thermistor placed inside the test chamber near the air-circulating fan. A recorder was attached to the thermistor, to record the temperature continuously.

3. EXPERIMENTAL PROCEDURE

3.1. Colorimeter Calibration

It has been determined [10, 11] that 0.72 moles per unit volume of sodium nitrite ($NaNO_2$) in aqueous solution give the same color intensity as that of 1.0 mole per unit volume of NO_2 gas. This ratio is called the "Saltzman Factor". After selecting the proper absorbing wavelength on the colorimeter, 0.0023 g of $NaNO_2$ was dissolved in 100 ml of distilled water (standard $NaNO_2$ solution). Graduated amounts of 0.1, 0.2,... up to and including 1.0 ml of the standard solution were added to ten separate 25 ml flasks. The flasks were then filled with measured aliquots of the absorbing Saltzman reagent. After 15 minutes, the colors were completely developed and the absorbances were measured using the colorimeter. Figure 1 shows NO_2 concentration *versus* absorbance.

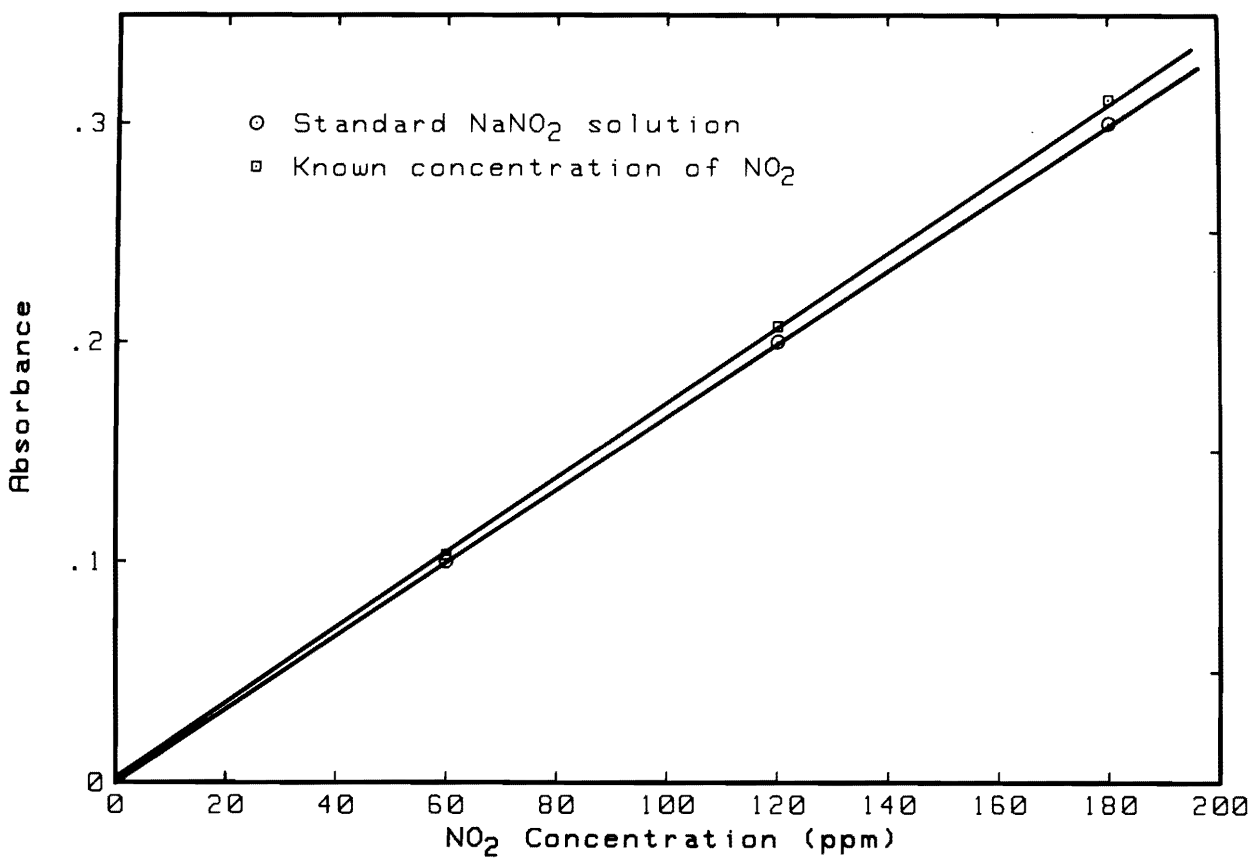

Figure 1. Determination of NO_2 Concentration From Absorbance.

3.2. Known NO_2 Concentration Calibration

A high-pressure gas cylinder containing a known concentration of 1100 ppm NO_2 in N_2 was used to prepare known concentration of NO_2 by dilution with N_2 in a small teflon bag. The results are also shown in Figure 1. The values obtained here are in very good agreement with the results obtained using $NaNO_2$ solutions.

3.3. Gas Mixture and Water Entry

The $NO-N_2$ and NO_2-N_2 gas mixtures from high pressure cylinders were directly metered into the reactor bag using flow meters. Water was introduced into the bag by saturating a portion of the diluent nitrogen. The salturation of nitrogen was obtained at 23°C by sparging it at low flow rates through a two-liter long-neck flask filled with distilled water. In all experiments, the initial concentrations of water, NO, and NO_2 were 1.63×10^4 ppm (corresponding to 50% relative humidity at 23°C), 132 ppm, and 132 ppm, respectively. These initial concentrations were obtained by metering into the reactor bag, 5 liters of water-saturated nitrogen, 3 liters of dry nitrogen, 1.1 liters of 980 ppm $NO-N_2$ gas mixture, and 1.0 liters of NO_2-N_2 gas mixture. The total volume of all the gases was 10.1 liters for each experimental run.

3.4. Sample Analyses

Each gas sample withdrawn into a 200 ml flask from the reactor was 10 ml in volume. To prevent oxidation of NO to NO_2, air was purged from each flask with pure nitrogen for a period of three to four minutes. Prior to stopping the nitrogen flow, 10 ml of Saltzman reagent was added. A gas sample was then withdrawn from the reactor and quickly injected into the 200 ml flask. The flask was shaken gently and placed in the dark to prevent any reaction of either NO_2 or HNO_2 with light *via* the reactions:

$$NO_2 + h\nu \longrightarrow NO + O\cdot$$

$$HNO_2 + h\nu \longrightarrow NO + OH\cdot\ .$$

After 15 minutes, with occasional shaking, the absorbance of the color-developed pink solution (corrected for the blank) was measured at a wavelength of 540 nm using the colorimeter. Triplicate runs were made for each analysis.

4. EXPERIMENTAL RESULTS

4.1. Conversion Rates

The experimental results of the analysis of the NO_2 concentration *versus* time for temperatures of −2, 4, 10, and 15 °C are plotted in Figure 2. It is important to note that the rate of NO_2 consumption increased with decrease in temperature.

4.2. The Integral Equation

For the overall reaction,

$$NO + NO_2 + H_2O \rightleftarrows 2HNO_2\ ,$$

the order is pseudo second-order, since water was always in large excess. The reaction was assumed to be first-order with respect to both NO and NO_2 and, hence, pseudo second-order overall. This assumption was tested using the integral method. It was found that the experimental data did fit overall pseudo second-order kinetics very well. If NO_2 is designated by the letter A, NO by the letter B, and HNO_2 by the letter C, the reaction scheme becomes,

$$A + B \underset{k_r}{\overset{H_2O,\,k_f}{\rightleftarrows}} 2C\ .$$

The integral equation in terms of the fractional conversion X_A, the equilibrium fractional conversion X_{Ae}, and initial concentration C_{Ao} is given by:

$$\ln \frac{X_{Ae}-(2X_{Ae}-1)\,X_A}{X_{Ae}-X_A} = 2k_f\left(\frac{1}{X_{Ae}}-1\right)C_{Ao}t, \tag{1}$$

A plot of $\ln \dfrac{X_{Ae}-(2X_{Ae}-1)\,X_A}{X_{Ae}-X_A}$ *versus* t should give a straight line if the actual reaction is indeed second-order. A plot of this type for the experimental data at 15°C is shown in Figure 3. The slope is given by:

$$2k_f(1/X_{Ae}-1)\,C_{Ao}\ . \tag{2}$$

4.3. The Overall Measured Rate Constant

Once the slopes have been obtained from experimental values at different temperatures, then Equation (2) allows calculation of k_f, the overall forward rate,

$$k_f = C_{H_2O}\,k,\ (C_{H_2O} = 1.63\times10^4\text{ppm})\ , \tag{3}$$

Using Equation 3, Table 1 gives values of k_f and k at various experimental temperatures.

4.4. N_2O_3 Equilibrium and the True Rate Constant

As can be deduced from Table 1, the reaction between NO, NO_2, and H_2O has a negative temper-

Table 1. Calculated Constants from Equation (3).

$T°C$	$k_f \times 10^5$ ($ppm^{-1}\,min^{-1}$)	$K \times 10^9$ ($ppm^{-2}\,min^{-1}$)
−2.0	11.8	7.24
4.0	9.0	5.52
10.0	5.8	3.56
15.0	4.8	3.00

ature coefficient; *i.e.*, the specific rate of reaction decreases with increasing temperature, and gives a negative energy of activation on an Arrhenius plot. To explain this behavior, a fast, exothermic equilibrium between NO and NO_2 is assumed, [2, 14] according to the reaction:

$$NO + NO_2 \underset{k_1}{\overset{k_1}{\rightleftharpoons}} N_2O_3 \qquad \text{(fast)} \quad (4)$$

followed by the reaction,

$$H_2O + N_2O_3 \underset{k_2}{\overset{k_2}{\rightleftharpoons}} 2HNO_2 \quad \text{(rate-determining)}. \quad (5)$$

The measured rate constant k, is then the product of the equilibrium constant for the fast step, Equation (4), and the rate constant k_2 for the slow step, Equation (5); *i.e.*:

$$k_f = C_{H_2O}\, k$$

where,

$$k = K_{N_2O_3}\, k_2 \qquad (6)$$

so that,

$$k_f = C_{H_2O}\, K_{N_2O_3}\, k_2 \qquad (7)$$

or

$$k_2 = k_f / C_{H_2O}\, K_{N_2O_3} \qquad (8)$$

Jolly [15] gives values for $1/K_{N_2O_3}$ at 25, 35, and 45 °C; these values are extrapolated to −2°C. Table 2 gives the results of the extrapolation along with values for the true rate constant, k_2. An Arrhenius plot for k_2 is given in Figure 4.

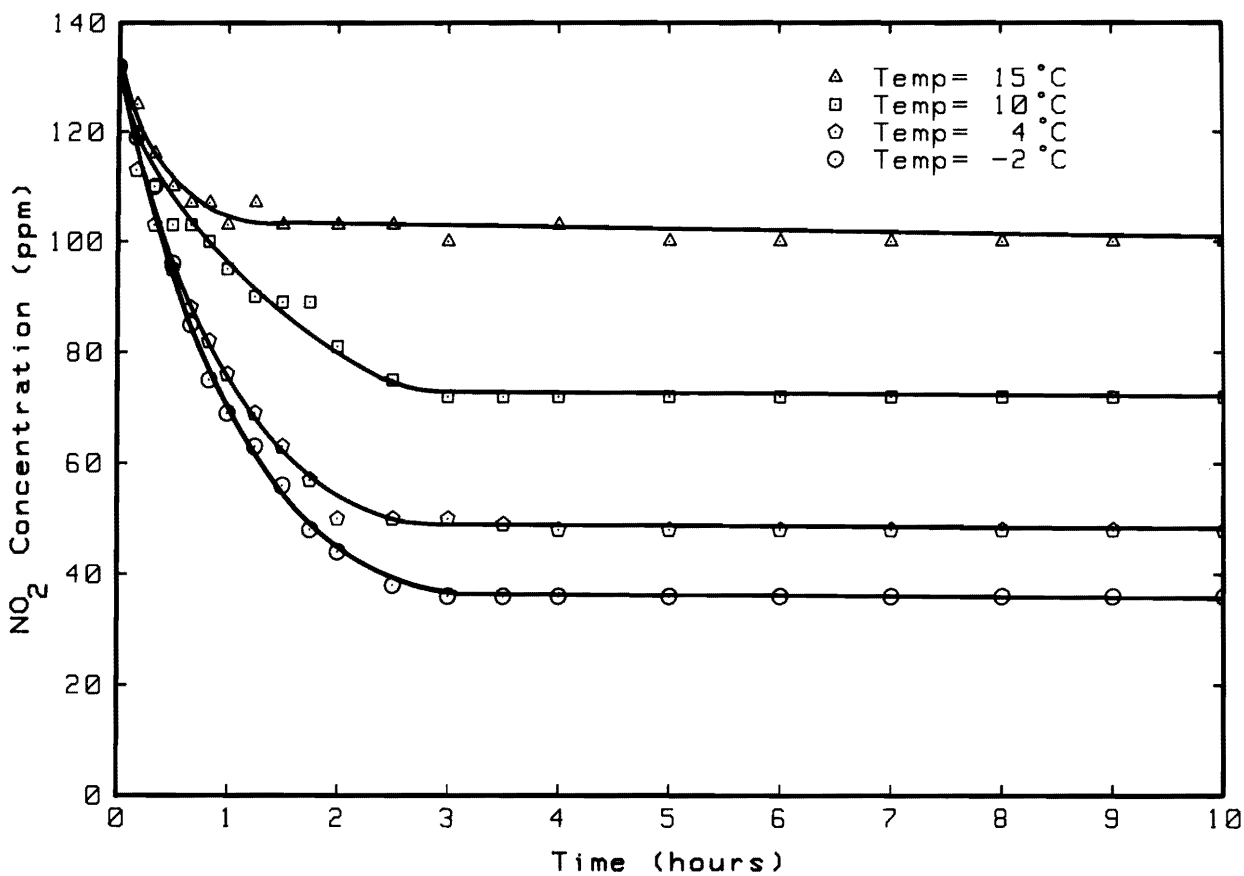

Figure 2. The Concentration of NO_2 as a Function of Time at Various Temperatures.

4.5. The True Activation Energy

From Figure 4, the slope of the curve, representing $\log k_2$ *versus* $10^3 T^{-1}$, was found to be -625 K, from which the activation energy is calculated by:

$$\text{Slope} = -E/R \,, \qquad (9)$$

giving $E = +1250$ cal, where $R = 1.987$ cal $K^{-1} mol^{-1}$.

4.6. The True Arrhenius Expression

The frequency factor k_0 for the true rate constant was calculated to be $k_0 = 9.9 \times 10^2$ (l $mol^{-1} s^{-1}$), so that the final expression relating the specific rate of reaction to temperature is:

$$k_2 = 9.9 \times 10^2 \exp(-1250/RT) \,. \qquad (10)$$

5. DISCUSSION AND CONCLUSIONS

The concentration of NO_2 as a function of time, with temperature, as the controlled parameter, is shown in Figure 2. As expected of all third order and pseudo second-order reactions, the negative temperature dependence is exhibited by the increasing consumption of NO_2 with decreasing temperature. The reaction attained equilibrium faster for higher temperatures. Equilibrium was reached after two hours at 15°C, and after three hours at −2°C. To ensure equilibrium conditions, at all temperatures tested, the reactor system was left undisturbed for ten hours, and concentrations were determined periodically.

Table 2. Calculated Constants from Equation (8).

T°C	$K_{N_2O_3}$ (l mol^{-1})	k_2 (l $mol^{-1} s^{-1}$)
−2.0	83.0	100.5
4.0	52.4	105.0
10.0	33.1	107.0
15.0	25.0	115.0

The only reaction product of NO and NO_2 is N_2O_3 [2, 14]. Equilibrium considerations exclude the formation of higher valent oxides and oxyacids of nitrogen. The N_2O_3 will react with water vapor to

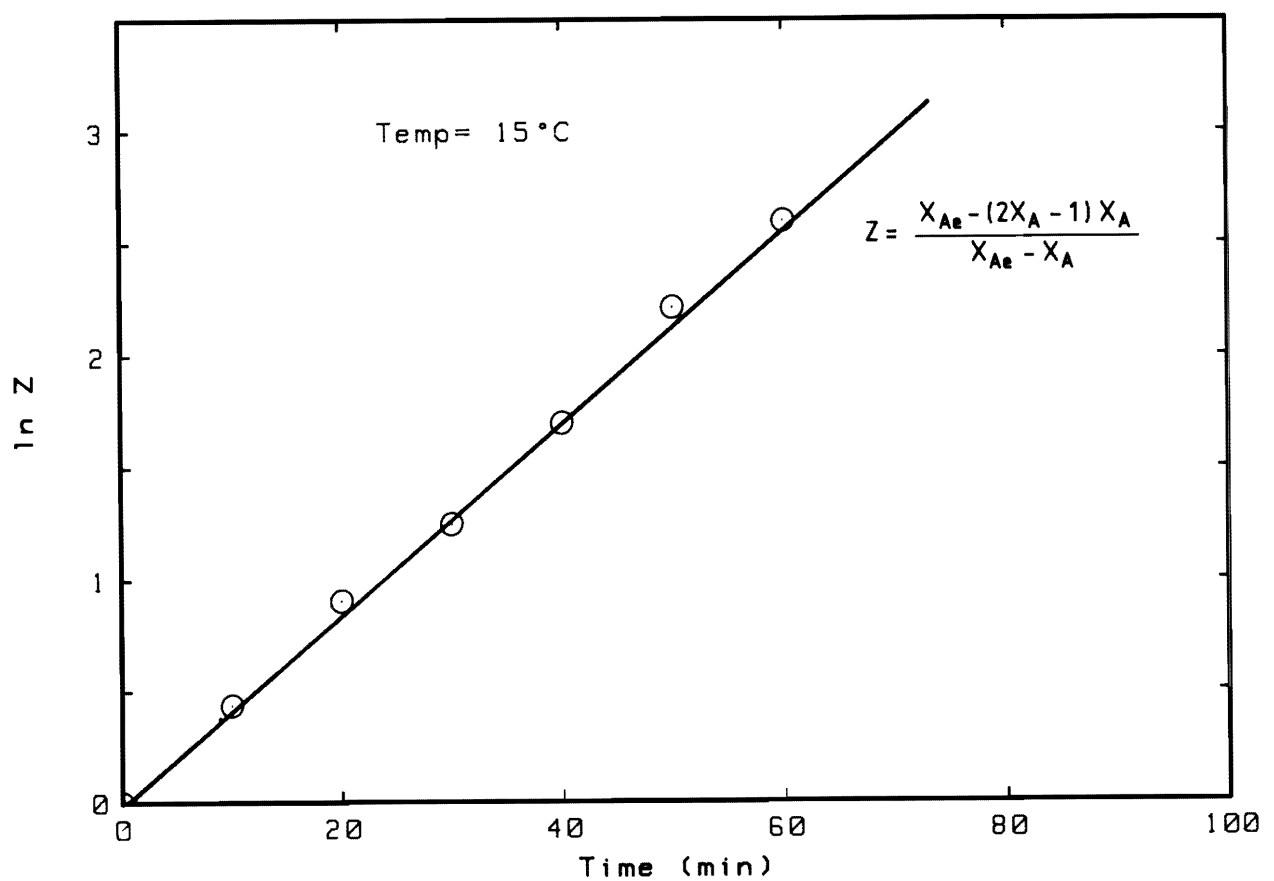

Figure 3. Determination of the Overall Rate Constant, k_f, at 15°C.

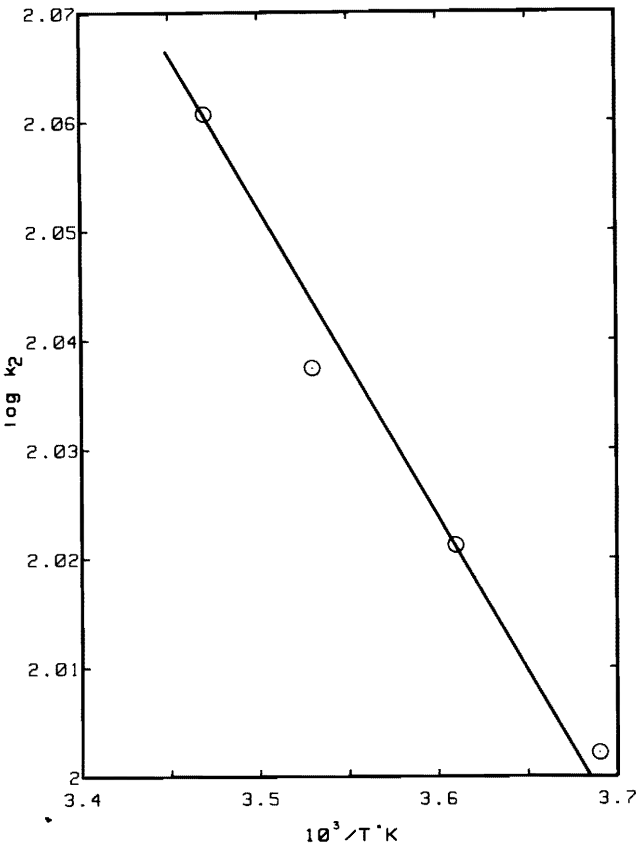

Figure 4. Arrhenius Plot for the True Rate Constant.

produce both isomers of HNO_2, namely, *cis*-HNO_2 and *trans*-HNO_2.

Systematic errors resulting from experimental measurements were assessed. The major source of error is in the measurement of NO_2 concentrations. Errors in these measurements were ±1%. Incorporating these errors in subsequent calculations, it is estimated that the activation energy is within ±90 cal, which is about ±7%.

At the same temperature, the value of the reaction rate constant obtained in this study compares well with values reported in the literature by several authors [5, 6, 16]. In fact, the value reported here lies within the cited values. In the case of Chan [5], where they used a tubular stainless steel reactor with a surface to volume ratio of 0.052 cm^{-1}, the comparison is rather good, even though the surface to volume ratio used in this study was 0.21 cm^{-1}, with a teflon bag reactor. This is probably due to the fact that teflon is one of the least surface-active materials known to date. Table 3 summarizes these values along with other values reported in the literature. Considering the scatter of data in the reported literature, this work provides the first measurement of the $NO + NO_2 + H_2O \rightleftarrows 2HNO_2$ reaction as a function of temperature.

Table 3. Summary of the Reaction Rate Constants.

Investigator	Rate constant cm^6 $molecules^{-1}$	Temperature K
Wayne and Yost	1.2×10^{-34}	298
Chan *et al.*	6.0×10^{-38}	296
Graham and Tyler	3.3×10^{-38}	298
Kaiser and Wu	4.4×10^{-40}	300
This Work	3.9×10^{-39}	298

The degree to which the reaction occurs by homogeneous rather than heterogeneous pathways is not exactly known. It is likely that the reaction followed homogeneous kinetics. Kaiser [6] reported an upper limit value to the homogeneous reaction rate constant. This upper limit postulates that the value at 300 K must be less than or equal to 4.4×10^{-40} cm^6 $molecules^{-1}$. Comparing all the different values reported for the reaction rate constant, it is observed that the values obtained in this study represent the closest values to the criterion posted by Kaiser [6].

Previous investigators [5, 14] did not measure an activation energy for this reaction; however, they predicted a very low value. The actual value for the activation energy for this particular reaction is 1.25 kcal, which is rather low. The experimentally determined value of the Arrhenius preexponential factor is also low. This could be due to the formation of a transition state in which two HNO_2 molecules would form a five-member ring.

Finally, using the experimentally determined rate constant and the assumed mechanism for the dark formation of HNO_2, it is of interest to determine what level of HNO_2 concentration one might expect to be formed overnight in a typical smog basin, such as the Los Angeles basin [17]. Using typical concentrations of NO, NO_2, and H_2O at 0.3 ppm, 0.3 ppm, and 16 000 ppm, respectively, at 50°F, the amount of HNO_2 in the early morning atmosphere after 12 h of reaction is then estimated to be about 0.064 ppm. Since HNO_2 is a precursor of OH· radicals, this estimated concentration of HNO_2 is not insignificant.

ACKNOWLEDGEMENT

The author would like to extend his thanks and gratitude to Dr. R. G. Rinker for his help and to the

King Fahd University of Petroleum & Minerals for support.

REFERENCES

[1] P. A. Leighton, *Photochemistry of Air Pollution.* New York: Academic Press, 1961, p. 66.

[2] L. G. Wayne and D. M. Yost, "Gas-Phase Reaction Between NO, NO_2, and Water", *Journal of Chemical Physics,* **19** (1951), p. 41.

[3] D. M. Woldorlf and A. L. Bagg, "Vapor Phase Equilibrium of NO, NO_2, H_2O, and HNO_2", *Journal of Chemical Physics,* **39** (1963), p. 432.

[4] C. England and W. H. Corcoran, "Kinetics and Mechanism of the Gas-phase Reaction of Water Vapor and Nitrogen Dioxide", *Industrial and Engineering Chemistry Fundamentals,* **13** (1974), p. 373.

[5] W. H. Chan *et al.*, "Kinetics Study of HONO Formation and Decay Reaction in Gaseous Mixtures of HONO, NO and NO_2", *Environmental Science and Technology,* **10** (1976), p. 674.

[6] E. W. Kaiser and C. H. Wu, "A Kinetics Study of the Gas Phase Formation and Decomposition Reactions of Nitrous Acid", *Journal of Physical Chemistry,* **81** (1977), p. 1701.

[7] J. N. Pitts, Jr. *et al.*, "Photooxidation of Aliphatic Amines Under Atmospheric Conditions: Formation of Nitrosamines, Nitramines, Amides and Photochemical Oxidant", *Environmental Science and Technology,* **12** (1978), p. 946.

[8] Y. N. Lee and S. E. Schwarts, "Reaction Kinetics of Nitrogen Dioxide with Liquid Water at Low Partial Pressure", *Journal of Physical Chemistry,* **85** (1981), p. 840.

[9] J. G. Calvert and W. R. Stockwell, "Acid Generation in the Atmosphere by Gas-phase Chemistry", *Environmental Science and Technology,* **17** (1983), p. 428A.

[10] B. E. Saltzman, "Colorimetric Microdetermination of Nitrogen Dioxide in the Atmosphere", *Analytical Chemistry,* **26** (1954), p. 1949.

[11] B. E. Saltzman, "Modified NO_2 Reagent for Recording Air Analyzers", *Analytical Chemistry,* **32** (1951), p. 135.

[12] A. P. Altshuller *et al.*, "Storage of Vapor and Gases in Plastic Bags", *International Journal of Air and Water Pollution,* **6** (1961), p. 75.

[13] A. E. O'Keefe and G. C. Ortman, "Primary Standard for Trace Gas Analysis", *Analytical Chemistry,* **38** (1966), p. 761.

[14] C. England and W. H. Corcoran, "The Rate and Mechanism of Air Oxidation of Parts-Per-Million Concentrations of Nitric Oxide in the Presence of Water Vapor", *Industrial and Engineering Chemistry Fundamentals,* **14** (1975), p. 55.

[15] W. L. Jolly, *The Inorganic Chemistry of Nitrogen.* New York: Benjamin, 1964, p. 69.

[16] R. F. Graham and B. J. Tyler, "Formation of Nitrous Acid in Gas-Phase Stirred Flow Reactor", *Journal of the Chemical Society Faraday Transactions,* **14** (1972), p. 683.

[17] Shell Briefing Services, *Air Pollution: An Oil Industry Perspective.* Rotterdam: Shell Netherlands, BV, 1987, p. 4.

Paper Received 13 April 1987; Revised 12 January 1988.

IMPLEMENTATION OF AN ARABIC DIGIT RECOGNITION SYSTEM

M. S. Ahmed*

Department of Systems Engineering
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia

and

E. M. Hagos

Department of Electrical Engineering
Carlton University, Ottawa

الخلاصـــة :

تم تنفيذ نظام لتمييز الأعداد العربية لايعتمد على المتحدث بل يستخدم طريقة مطابقة النماذج . ولقد بُنِيَ النظام على بارامترات LPC وعلى مقياس (كوش) للمسافة بين إطارٍ وإطار وعلى طريقة ديناميكية لمعالجة الانبعاج لضبط بدايتي المنطوق المرجع مع المنطوق المختبر .

وحتى يكون النظام مستقلا عن المتحدث تُستعمل عدة نماذج لكل كلمة وقاعدة أقرب مجاور كمعيار للاختيار . ولقد كُـوّنت قاعدة للمعلومات بها ثمانون تكراراً لكل رقم . وجمعت هذه التكرارات عن طريق تلفظ عشرين متحدثٍ بكل رقم أربع مرات . وأمكن الحصول على النماذج المرجعية باستخدام طريقة احصائية على مكونات قاعدة المعلومات . ولقد أجريت التجارب كما يلي :

١ - تغيير عدد المتحدثين في مجموعة التدريب وعدد النماذج لكل رقم .
٢ - دراسة تأثير حجم الإطار الثابت والمتغير .

ولقد لوحظ أنَّ عدد النماذج لكل كلمة يُـؤثر بشكل كبير على الاداء . أما فيما يتعلق بثبات وتغيير حجم الإطار فإنَّ الاطار متغير الحجم يقلل زمن التشغيل والتخزين الا أن الأداء يكون سيئاً في حالة الأرقام ذات الصوت المشابه .

*Address for correspondence :
KFUPM Box No. 134
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia

ABSTRACT

A speaker-independent Arabic digit recognition system is implemented that uses template matching. The system is based upon the LPC parameters as features, the cosh measure as frame-to-frame distance, and the procedure of dynamic time warping for time alignment between the test and reference utterances. To accommodate speaker independency it uses multiple templates per word and the k-NN rule as the decision criterion. Four utterances of each digit from each of twenty speakers are collected to form a data base of eighty replications for every word. The reference templates are obtained from a statistical clustering analysis of this data base. Experiments are conducted (1) by varying the number of speakers in the training data set and the number of templates per word and (2) considering fixed and variable size framing. It is observed that a significant factor to improve the speaker independent performance is the number of templates per word. A comparison of fixed and variable size framing showed that although the latter is capable of reducing the processing time and storage, it yields poor performance in acoustically similar words.

IMPLEMENTATION OF AN ARABIC DIGIT RECOGNITION SYSTEM

1. INTRODUCTION

The problem of isolated word recognition can be viewed as a particular case of the classical pattern recognition problem which has two phases, the development or training phase and the implementation or recognition phase.

In the former patterns from known classes are used to train the system. In its simplest form, training consists of measuring, processing, clustering, and storing relevant features of the known patterns. Since in their original form, the patterns are usually very high dimensional, they are processed in order to extract features having discriminatory power. In an attempt to remove the effect of variability among the different occurrences of the same class, statistical techniques are used to cluster the samples during the training phase.

In the second phase, when the system is in use, patterns of unknown classes are received which are compared with all the stored templates and scores are obtained from the comparison. The unknown pattern is classified as belonging to a class to which it is closest. Thus this phase involves the similarity measurements and decision rules.

To put the isolated word recognition system in the above framework three major considerations are required (*i*) the preprocessing and feature selection (*ii*) a distance measure for similarity measurement, and (*iii*) a decision strategy. The analog acoustic signal is low pass filtered to remove the high frequency components which are not significant for recognition. The resulting signal is then digitized at a rate greater than twice the cut-off frequency of the low pass filter. Automatic algorithms are used to detect the end points and usually at this stage, a pre-emphasis filtering is applied for spectral smoothing. Various features that have been used for speech recognition include zero crossing rates, energy measurements in different frequency bands [1], Discrete Fourier Transform (DFT) [2], cepstral coefficients, and the Linear Prediction Coefficients (LPC) [3]. Among them the LPC modeling, where short intervals of speech comparable to the pitch period is assumed as an output of a linear time invariant system, has been extensively used because of its conceptual and computational simplicity.

The distance measures used for speech recognition are the spectral distances [4, 5], weighted spectral distances [6], Euclidean distances [1], and the covariance weighted distances [7]. In order to compensate for intra- and inter-speaker variations and the errors in automatic detection of the end points, it is necessary to time align the reference and test utterances. One frequently used scheme for time alignment is the dynamic time warping algorithm which uses dynamic programming optimization to select the best alignment [8]. Once a suitable distance measure is defined, clustering algorithms may be used in the training stage to identfy the representative prototype patterns to be stored for later use. This helps in making the recognition process independent of the speakers [9].

The last aspect of pattern recognition involves the decision strategy. Typically, the input pattern is identified as belonging to the class to which it is nearest. In the case of multiple template representation, some form of nearest neighbor rule may also be employed [10]. A rejection threshold may also be used in this stage to reject any pattern as being not belonging to any of the stored template class.

Other approaches that differ from the above template matching procedures include feature-based recognition [11] and event-based recognition [12]. In feature-based recognition, the idea is to use various features, distributed across time and frequency, about various phonetic events and the manner they change. In event-based recognition, a time domain segmentation algorithm first labels the speech waveform into different groups of sound such as stops, vowels, fricatives, *etc.* In the second stage, the speech segments are compared with reference templates of same group. The decision strategy is based upon the accumulated distances.

Relatively little research has been done on Arabic speech recognition; the only published results available are due to the IBM Cairo Scientific Center, and to the Electronics and communication department at Cairo University. Hashis and others [13] considered the recognition of the ten Arabic digits by using a number of features such as normalized autocorrelation, linear prediction coefficients, zero crossing rates and formants. Their limited experiment with two male and two female speakers has shown that high accuracy can be attained by the use of any of these features when complemented by appropriate heuristics. Emmam and Hashis [14] applied the

Hidden Markov Models (HMM) to recognition of speaker dependent Arabic words. The feature used was the energy values in twenty critical frequency bands on the Mel scale. It has been shown that with adequate training high recognition can be achieved using this approach.

A successful speaker independent isolated word recognition system should be capable of recognizing words from a variety of speakers with varied dialects and accents. In this paper, the implementation of a speaker independent isolated word recognition system for the Arabic digits zero to nine is presented. A total of twenty-six speakers with different accents and dialects have been used to train and test the system. The system is based on LPC parameters as features [3], the cosh measure as the distance function [5], the dynamic time warping for time normalization [8] and k-NN rule for decision criterion [10]. Speaker independency of the system is incorporated by generating multiple reference templates from statistical clustering analysis. A number of experiments have been conducted on the system to study the effects of (1) speaker size in the training phase; (2) number of reference templates; and (3) the fixed and variable size framing [17].

The organization of the paper is as follows. Section 2 describes the implementation of the system including data collection, preprocessing, feature extraction, template generation, and recognition. Section 3 reports the recognition experiments and results. Section 4 includes the conclusion.

2. EXPERIMENTAL IMPLEMENTATION

This section is devoted to the description of the experimental implementation of the Arabic digit recognition system.

The ten Arabic digits considered for recognition are given as follows:

Sifr	(zero)	Khemsa	(five)
Wahd	(one)	Sita	(six)
Ithnen	(two)	Sebea	(seven)
Thelatha	(three)	Themania	(eight)
Arbea	(four)	Tisea	(nine)

2.1. Analog Data Collection

Four utterances of every word from each of twenty male speakers were recorded in a normal laboratory environment using a Sony Stereo Cassette Recorder on metal cassettes of frequency response in the range of 60–13 000 Hz. The samples were taken from speakers of varied accents and nationalities in order to include various accents and dialects. These 800 utterances were used to train the system.

In order to test the system three sets of utterances were used: (1) the training utterances not selected for the reference templates; (2) 120 additional utterances created from twelve of the speakers used in the training set; and (3) another 130 utterances created from six additional speakers of different accents and dialects who were not used in the training stage.

2.2. Digitization

The analog utterances were band-pass filtered between 80–3500 Hz. The filtered signal was then digitized at 8 kHz using a Tecmar PC-Master A/D converter installed on an IBM PC-XT at 12 bits/sample. A 2.5 s portion of each utterance imbedded in silence was first sampled to obtain 20 000 samples. The sampled speech signals were played back using a D/A converter and a speaker. A sliding window of length 8192 samples (1.025 s) was then used to zoom in on the required speech utterance. This semi-automatic procedure was a first step in the end point determination of the utterances.

2.3. Data Transfer

Since the recognition system was implemented on an IBM 3033, the data was transferred from IBM PC to the mainframe using the IRMA interface card. A program was written to handle this data transfer as the IRMA data transfer software was found to be inadequate for binary data transfer both in speed and in maintaining the data integrity.

2.4. Preprocessing and Feature Extraction

The preprocessing steps required for the system were end point detection, pre-emphasis, framing, and windowing. Figure 1 depicts the sequence.

2.4.1. End Point Detection

The end-point detection was based on an algorithm proposed by Rabiner and Sambur [14] that uses energy and zero crossing computed for every 10 ms of speech. It obtains statistical information about the background acoustic environment from the first part of speech to establish the threshold values. The algorithm initially uses the energy threshold to establish the tentative end points and then uses the

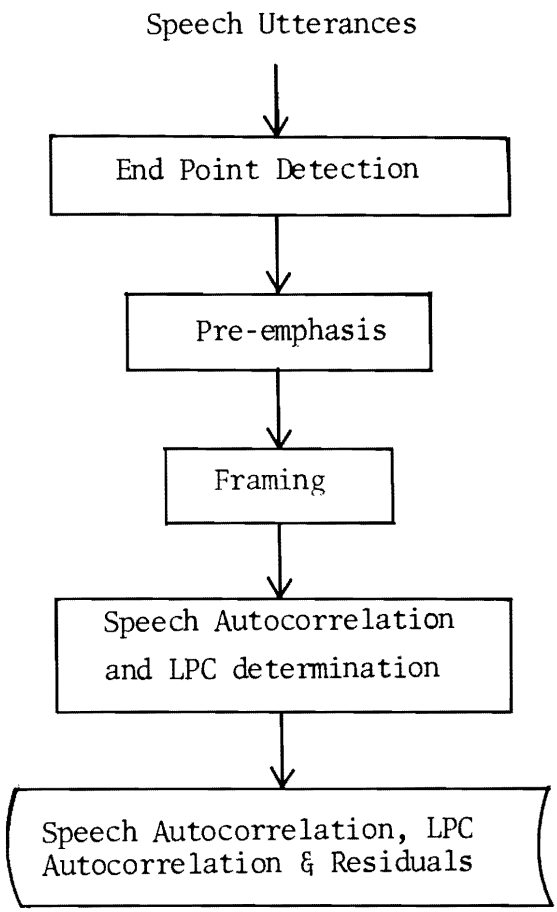

Figure 1. Block Diagram of the Feature Extraction in FFS.

zero crossing threshold to refine the end points. The refinement accounts for the presence of fricatives at the ends.

2.4.2. Pre-Emphasis

All utterances were pre-emphasized using a filter of the following type.

$$p(z) = 1 - 0.95\, z^{-1}, \qquad (1)$$

where z is a unit advance operator. The use of the pre-emphasis filter enhances the high frequency components of the speech wave. The high frequency enhancement of speech partially cancels the glottal shaping function and emphasizes the vocal tract characteristics.

2.4.3. Framing

In order to extract features, the discrete signal sequences were required to divide into small units, called frames. Frame sizes of 10–45 ms have been used in the implementation of various Isolated Word Recognition Systems (IWRS). In our study overlapping frames of length 32 ms and increment of 16 ms were used.

An obvious shortcoming of blocking the speech signal at a fixed rate is that no advantage is taken of the similarities that may exist between neighboring frames. In variable length framing the speech signal is framed at fixed length to start with and then neighboring frames are compared with one another and merged to give longer frames if they are found to be sufficiently similar. The idea has been implemented for variable speech coding in speech transmission [17] and speech recognition [17–20]. In the present study both fixed and variable length framing were implemented. In the variable length coding the average number of frames per utterance was observed to be about half of that in fixed length framing.

2.4.4. Windowing

The framing operation introduces an implicit rectangular windowing whose effect in the time domain is to create discontinuities at the boundary points. In the autocorrelation method of linear prediction, where the signal is assumed to extend from $-\infty$ to $+\infty$, the abrupt changes at the boundary points create spectral distortion. A windowing operation to taper the data at the end points and thus cause smooth transition is required. The window employed was the Hamming window [21].

2.5. Feature Extraction

Various features that have been used in IWRS can be broadly classified as either time-domain or frequency domain features. Although the time-domain features such as zero crossing rates and fundamental frequency of voiced sounds can be derived easily, the parameters more frequently used are the frequency domain features, as they provide better insight into the relationship between the speech signal and the manner of its articulation by the vocal organs. Consequently, short-time spectral representations such as the LPC, filter bank outputs, DFT features, and cepstral coefficients have found wide use. Some of the overriding considerations in the selection of features are the goodness of representation, efficiency, minimality, and ease of extraction.

2.5.1. Linear Prediction Modeling

In the present study, LPC features were used. In LPC modeling, speech is assumed as an output of an all-pole digital filter excited by an impulse train for voiced speech and by white noise for unvoiced speech. The all-pole filter may be expressed as:

$$x_k = -\sum_{i=1}^{p} a_i x_{k-i} + e_k$$

or $$x_k = [1/A(z^{-1})]\, e_k$$

with $$A(z^{-1}) = \sum_{i=0}^{p} a_i z^{-i} \text{ and } a_0 = 1 \ ;$$

where a_i, $i = 1, 2, \ldots, p$ are the LPC parameters or simply LPC's, e_k is the excitation function and x_k is the speech sequence. A particularly interesting aspect of this modeling approach is that the speech spectral estimation is reduced to the evaluation of the LPC which using the autocorrelation method reduces to solving a set of simultaneous equations for which an efficient algorithm exists [22].

There exist efficient distance measures for comparing two speech sounds that are represented by their LPC parameters. It will be discussed in the next section that one measure depends on the error residual of the linear prediction method which in the autocorrelation method is obtained as a byproduct of the Durbin's [22] algorithm which substantially decreases the computational load of evaluating the distance between two frames.

An appealing characteristics of the LPC's is that the entire analysis is performed in the time-domain. In this study a 12 order LPC modeling was conducted. The relevant features stored for each frame were the autocorrelation coefficients of the data, autocorrelation coefficients of the LPC's and the least sum squared residual of LPC modeling [4].

2.6. Similarity Measurement

The matching of utterances to establish their closeness is a fundamental aspect of speech recognition. In fact, a major consideration in the selection of features is the requirement that the feature space allows for a definition of a distance measure. In the case of IWRS, the features are usually a time-sequence of vectors derived over short-periods of time. Since the two utterances to be compared are generally of unequal length, there is the added requirement of defining a matching procedure for this time vectors. Such a need, in the context of speech, arises due to the variability inherent in speech utterances. Hence, similarity measure becomes a three-step procedure, firstly the definition of a distance measure between two frames of speech, secondly having such a measure the establishment of a procedure to match two utterances that are of different time duration, and thirdly application of a decision rule to classify the unknown pattern. In addition to these, a clustering step may be employed during the training stage to accommodate the intra-speaker variability.

2.6.1. Distance Definition

It is possible to define distance functions both in the temporal and spectral domains. Different distance measures used in pattern recognition application include Euclidean distance [1], weighted covariance or Mahalanabish distance [7], and various spectral distances [5, 6]. The spectral distances are found to be most useful in speech recognition. Another distance measure suitable for LPC-based speech recognition is known as the log likelihood ratio originally proposed by Itakura [4] in which two utterances are compared by computing the ratios of linear prediction residual energy of the frames when they are passed through the inverse LPC filter derived from one of them. What follows that given two speech frame sequences $[x_0, x_1, \ldots, x_{N-1}]$ and $[y_0, y_1, \ldots, y_{N-1}]$ and the corresponding LPC models $1/A(z^{-1})$ and $1/B(z^{-1})$ respectively, the four possible prediction residual energies E_a^x, E_a^y, E_b^x, and E_b^y may be computed as shown in Figure 2.

Itakura [4] defined two likelihood ratios as

$$d = E_b^x / E_a^x \text{ and } d' = E_a^y / E_b^y \qquad (3)$$

which can be utilized as dissimilarity measures between frame x and frame y. The individual measures however are not symmetric. In order to obtain a symmetric measure their geometric mean may be taken, which can be shown to be related to the cosh of their spectral distance [15]. In order to obtain the distance in the decibel scale, the following distance was considered

$$d(A, B) = \log(dd')^{1/2} = (\log d + \log d')/2 \ . \qquad (4)$$

The computation of the distance however does not require explict filtering. An efficient computation of the likelihood distances is available due to Itakura [4] and is given as:

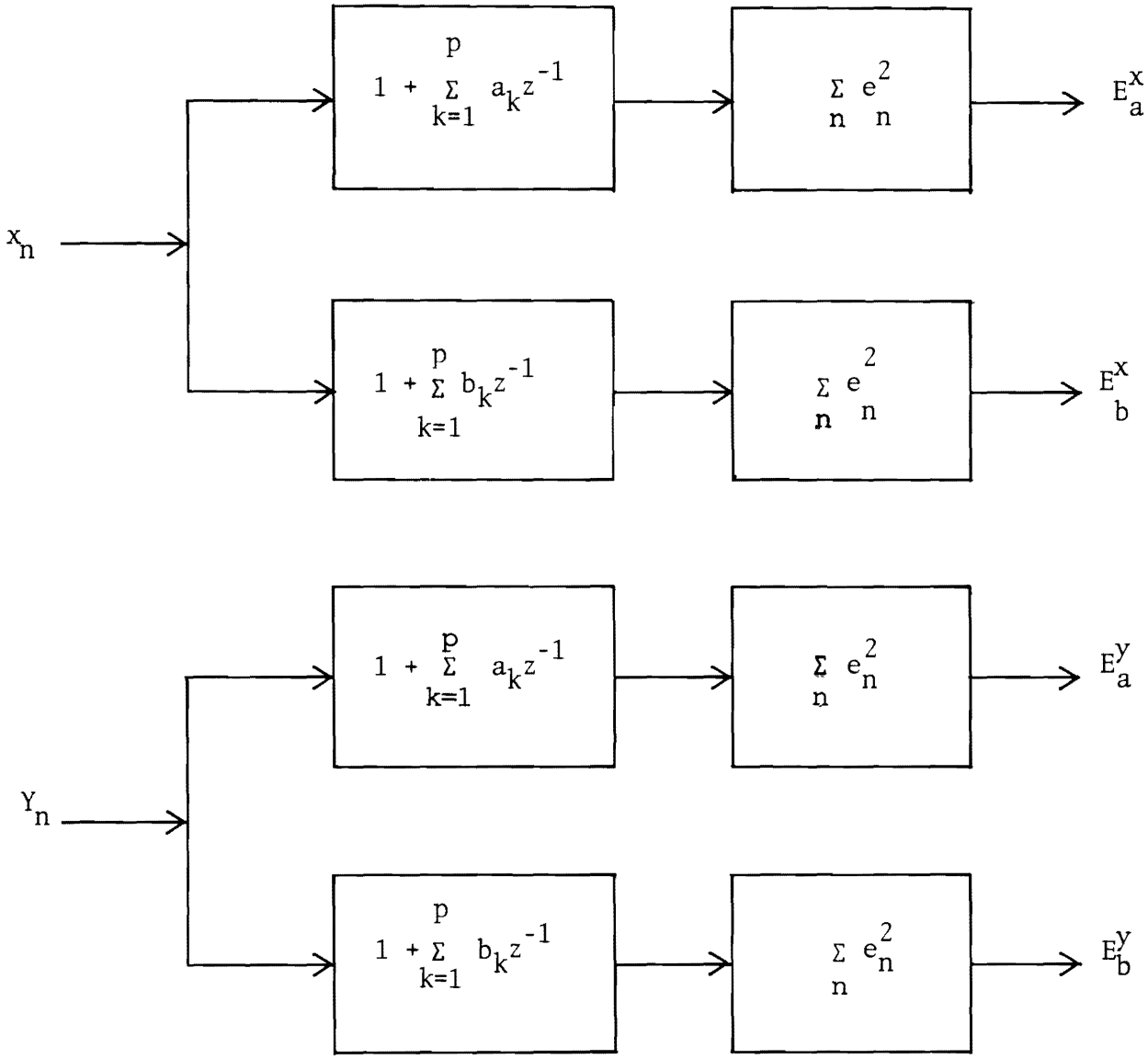

Figure 2. Four Possible Residuals From Two Frames.

$$E_a^x = \sum_{i=-p}^{p} r_a(i)\, r_x(i)$$

$$E_b^x = \sum_{i=-p}^{p} r_b(i)\, r_x(i)$$

$$E_a^y = \sum_{i=-p}^{p} r_a(i)\, r_y(i)$$

and

$$E_b^y = \sum_{i=-p}^{p} r_b(i)\, r_y(i) \qquad (5)$$

where $r_a(i)$ and $r_b(i)$ are the autocorrelation sequence of the coefficients of $A(z^{-1})$ and $B(z^{-1})$, $r_x(i)$, and $r_y(i)$ are the corresponding data autocorrelation coefficients and p is the LPC order. For details the reader is referred to references [4, 23].

2.6.2. *Time Alignment*

Generally, speaking rates vary considerably not only among different speakers but also among different utterances of the same word spoken by one speaker. As a result, different utterances of the same word may assume unequal lengths. This creates some difficulty in speech recognition. The problem becomes worse due to errors in end point detection.

The time alignment step is used to accommodate this problem which selects an optimal correspondence between the frames of the training and recognition utterances.

There are various approaches to achieve the time normalization which include the linear time alignment [1], alignment based on event matching [12], and dynamic time warping [8]. Among them the dynamic time warping method that employs the dynamic programming to match two utterances by optimizing the overall accumulated distance in a constrained path is found to be most successful in speech recognition. The optimization is carried on subject to three kinds of constraints.

A boundary constraint is enforced to align the end points, a monotonicity constraint in the alignment function is imposed since the speech samples are strictly ordered in time due to its physical nature, and a slope constraint in the alignment function is introduced which accounts for the local variations of speech wave form but disallow those distortions which are not physically possible. A widely and successfully used constraint assumes that one frame of an utterance cannot be assigned to more than two consecutive frames of the other utterance which is equivalent to constrain the slope of the alignment function between ½ and 2 [24]. A further consideration in the distance calculation is the selection of a weighting function. The usual approach is to divide the accumulated distance by the number of frames of the reference utterance which reflects the average interframe distance between two utterances. In our implementation a fixed end point dynamic time alignment algorithm was used.

2.6.3. *Clustering*

The purpose of clustering is to choose a small set of templates that can be used to represent a large number of replications of each word in the vocabulary. It is an essential part of all pattern recognition systems that are based on the statistical behavior of patterns. In particular, when the set of patterns of a given object can be partitioned into a number of close clusters, it is possible to represent these patterns by one pattern representative of the set.

Some of the popular algorithms for clustering are the k-means and ISODATA algorithms [10]. However, a more suitable algorithm for use in speaker independent speech recognition is designed by Rabiner and Wilpon [9]. The algorithm proceeds sequentially at each step by identifying a cluster center, removing the identified cluster center from further considerations and repeating the procedure until all the patterns in the original set have been associated with some cluster centers. This algorithm was used in our implementation.

2.6.4. *Decision Strategy*

This refers to the set of rules applied to classify an unknown utterance. Simplest of these is the nearest neighbor (NN) rule, where minimum of the distance is used to classify the utterance. This rule is effective when single template representation is used. For the multiple template representation a more sophisticated algorithm is the k-NN rule where for each class the average distance of k-nearest templates to the input is evaluated and the minimum of these averages is used to classify the input utterance.

In order to avoid system errors with utterances not belonging to the system, a rejection threshold may be used such that whenever the computed minimum distance exceeds the threshold, the system rejects the utterance. During the distance evaluation, computation can be abandoned when the accumulated distance exceeds a value corresponding to the threshold. This procedure reduces the computational load.

2.7. Operational Modes

The system was designed for three operational modes. Figure 3 illustrates the software consideration. Three modes consisting the training,

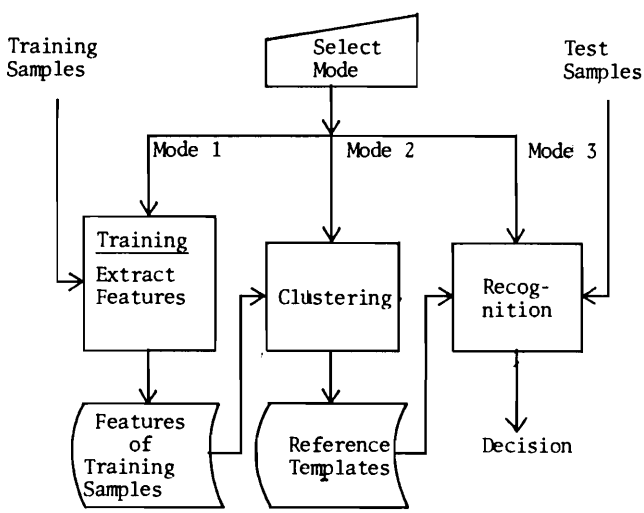

Figure 3. Block Diagram of the Software.

clustering, and recognition were incorporated. In the training mode, preprocessing and feature extraction steps were included. For every training utterance two files were stored. The first one contained the LPC autocorrelation coefficients and LPC least sum squared residual, and the second one contained the data autocorrelation coefficients.

In the clustering mode, a clustering algorithm was run to obtain a reasonable number of cluster centers as representative of each class. The clustering algorithm proposed by Rabiner and Wilpon [9] was used for this purpose. The distance between the training utterances required for the algorithm were the average frame to frame cosh measure computed through a fixed end point dynamic time alignment. The clustering algorithm rejected any cluster center that contain only one pattern.

In the recognition mode, the unknown test utterances were dynamically time warped [8] against all stored templates. The distance function used was the cosh measure. The decision strategy used was k-nearest neighbor (k-NN) rule. Details of each of these stages are given in Hagos [23].

3. RESULTS

This section describes the results which include various observations during the training, as well as the recognition stages.

3.1. Clustering

The user-supplied inputs to the clustering algorithm were a distance threshold T depicting the maximum intra class distance and a maximum iteration count K_m to limit computation. Suitable values of T were found as 0.6 and 0.5 respectively for the Fixed Frame Size (FFS) and Variable Frame Size (VFS) implementations. A suitable value for K_m was found as 5. The clustering algorithm rejected any cluster center that included only itself in the cluster set (called outliers).

Table 1 shows the results of clustering algorithm. In general inter utterance distances for VFS were smaller than those of FFS method. This was due to the relatively smaller number of frames in the former which necessitated the use of a smaller threshold parameter T in it. It can be also observed that, in general, the number of cluster centers in VFS is large compared to that of FFS (except for /Tisea/). This exceptional behavior of /Tisea/ was due to the presence of large number of outliers in the case of VFS. In order to obtain meaningful comparison between the two methods, the number of clusters in the VFS method were made at most equal to those of the FFS method by discarding the cluster centers with the smallest number of patterns in the cluster set.

Table 1. Results of Clustering Algorithm

Digit	FFS Number of Clusters	VFS Number of Clusters
Sifr	11	13
Wahd	11	15
Ithnen	12	13
Thelatha	11	12
Arbea	11	12
Khemsa	10	12
Sita	9	11
Sebea	9	11
Themania	11	15
Tisea	14	11

3.2. Recognition

The system was tested using three test sources.

1. From the 800 training utterances, 109 for the FFS and 105 for the VFS were selected as reference templates. The remaining 691 for the FFS and 695 for the VFS formed the first test group.
2. A second batch of test utterances was created from 12 of the speakers used in the training phase. 120 new utterances, one utterance per word from each speaker was collected. These utterances were not included into the training data.
3. A third group of test utterances was formed from 6 additional male speakers of varying accents and dialects not used in the training stage. Two utterances for each word from five speakers and three utterances for each word from the sixth speaker gave a total of 130 test utterances in this batch. It is to be noted that only this source provides any real test of a speaker-independent system. The earlier two sources were used to test the training.

Each of the test utterances were dynamically time wrapped against all the stored reference templates. The decision strategy adopted was the k-nn rule where k was set to 3. A rejection threshold for each digit was also used and was set equal to the mean of the distances between utterances of the same digit. Table 2 depicts these threshold values.

Table 2. Rejection Thresholds for the Two Implementations

Word	FFS	VFS
Sifr	0.892	0.783
Wahd	0.899	0.754
Ithnen	0.946	0.856
Thelatha	0.826	0.703
Arbea	0.841	0.744
Khemsa	0.809	0.675
Sita	0.757	0.691
Sebea	0.749	0.747
Themania	0.869	0.753
Tisea	0.894	0.748

3.2.1. Experiments With Different Test Sources

The recognition results for different test sources are shown in Table 3 while the summary (the overall recognition result) is displayed in Table 4. Tables 5 and 6 depict the confusion matrices for the misrecognized words in FFS and VFS implementations respectively.

A comparison of the two methods FFS and VFS in Table 4 shows that the former method is, in general, superior when the overall performance is considered. More careful examination of Table 3 however shows that recognition improves in the VFS method in some cases while there is marked degradation in

Table 3. Recognition Results (%)

Digit	Test Type 1 (Training Patterns)		Test Type 2 (12 Utterances/digit)		Test Type 3 (13 Utterances/digit)	
	FFS	VFS	FFS	VFS	FFS	VFS
Sifr	97.1	97.0	83.3	75.0	76.9	76.9
Wahd	94.2	97.0	91.7	91.7	100.0	92.3
Ithnen	92.7	91.0	91.7	100.0	100.0	92.3
Thelatha	93.3	95.6	91.7	100.0	100.0	100.0
Arbea	94.9	94.1	91.7	75.0	100.0	92.3
Khemsa	97.1	92.7	91.7	91.7	100.0	100.0
Sita	93.0	94.2	100.0	100.0	84.6	100.0
Sebea	93.0	92.8	100.0	83.3	92.3	92.3
Themania	93.0	95.4	91.7	100.0	92.3	100.0
Tisa	100.0	95.5	100.0	100.0	100.0	100.0

Table 4. Overall Recognition Results (%)

Test Type	FFS			VFS		
	Recognize	Reject	Misrecognize	Recognize	Reject	Misrecognize
1	94.8	1.2	4.0	93.5	1.9	4.6
2	93.4	0.8	5.8	91.6	1.7	6.7
3	94.6	0.0	5.4	94.6	0.8	4.6

Table 5. Confusion Matrix of Misrecognized Utterances for FFS

	Misrecognized	0	1	2	3	4	5	6	7	8	9
Sifr	7				2	1			1		3
Wahd	2								2		
Ithnen	6							1	2	1	2
Thelatha	3	1							1		1
Arbea	4								4		
Khemsa	2					1					1
Sita	5	1									4
Sebea	6					5					1
Themania	5	4							1		
Tisea	0										
Total	40	6	0	0	2	7	0	1	11	1	12

Table 6. Confusion Matrix of Misrecognized Utterances for VFS

	Misrecognized	0	1	2	3	4	5	6	7	8	9
Sifr	7			1	1		1	1	2	1	
Wahd	1								1		
Ithnen	7	1			2					4	
Thelatha	1									1	
Arbea	7								7		
Khemsa	1								1		
Sita	2			1							1
Sebea	7				2	4					1
Themania	3			1	1				1		
Tisea	10	1					1	8			
Total	46	2	0	3	6	4	2	9	12	6	2

other cases. While FFS resulted in a perfect score for the word /Tisea/, VFS yielded the poorest score, perhaps due to loss of valuable information. On the other hand, merging frames in /Wahd/, /Thelatha/, and /Themania/ by VFS removes redundant information improving the recognition rates.

From Tables 5 and 6 the confusion between the acoustically similar words /Arbea/ and /Sebea/, and /Sita/ and /Tisea/ is evident. Figure 4 shows the distance distribution from the reference templates of /Sita/ to the training utterances of /Sita/ and /Tisea/ for the FFS and VFS methods. The Figures demonstrate the overlap between the two words especially in the VFS explaining frequent mis-recognition of /Tisea/ as /Sita/. Another reason of the bad performance of /Tisea/ in the case of VFS was the rejection of large number of outliers from the cluster centers. Which was a direct consequence of lose of valuable information in VFS coding. Perhaps this confusion can be reduced by using weight in the distance function to different phonemes of the word. Such an approach is suggested in Lee [12].

3.2.2. *Experiments on the Number of Speakers for Training*

To see the effect of number of speakers in the training phase two experiments were performed. In the first one the number of speakers in the training phase were taken as 5, 10, 15, and 20. The average number of templates per digit were decided by a constant threshold parameter T. Table 7 shows the recognition results when FFS and test source 3 were used.

In the second experiment the same number of speakers as in the preceding experiment were used in

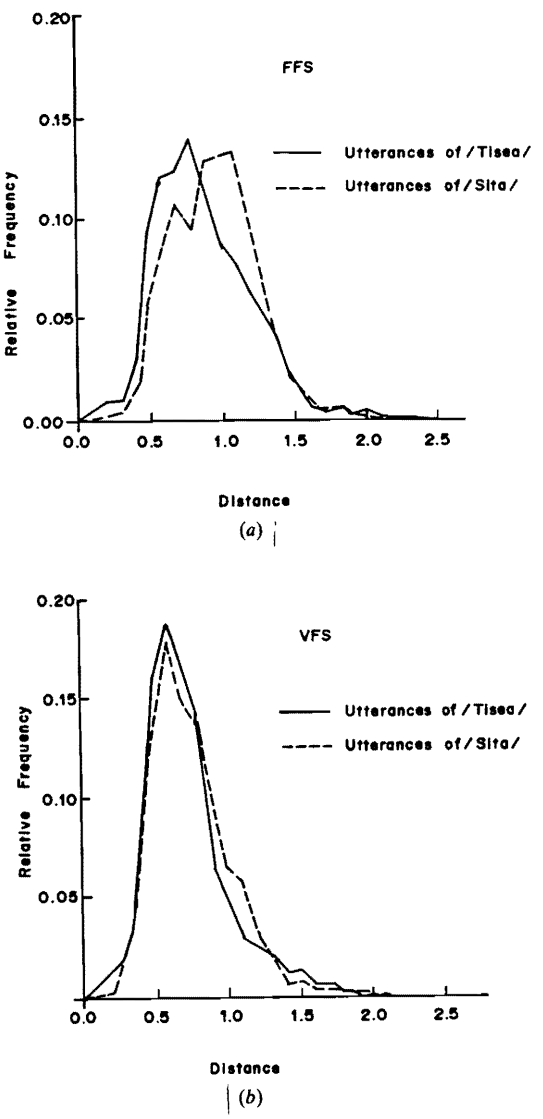

Figure 4. Distance Distribution from the Reference of /Sita/ to the Training Samples of /Sita/ and /Tisea/.

the training phase but the average number templates in each case was fixed to 10. Table 8 shows the recognition results for FFS implementation with test source 3.

Although the test sample size was small (130 utterances), the steady increase in recognition rate with the increase in speaker population of the training set is clearly evident from Table 7. However, Table 8 shows that the steady increase in recognition result is more due to the number of stored reference templates and not due to the increase in number of speakers. Although, it is expected that recognition accuracy of a speaker independent system should increase with more training speakers, the results of Table 8 perhaps indicate lack of consistency in the repetitive utterances of the training speakers. All our speakers were untrained junior undergraduate students. Therefore a significant factor to improve the performance of speaker independent recognition systems for untrained speakers can be considered as the number of multiple templates in the reference pattern.

4. CONCLUSION

A speaker-independent Arabic digit recognition system is implemented and tested. A total of twenty-six speakers of varying accents and dialects have been used in the training and test of the system. The system used LPC parameters as features, the cosh measure as frame to frame distance, a fixed end point dynamic programming for time warping and a statistical clustering analysis to select the reference templates. An overall accuracy of about 95% is achieved.

A comparison of Fixed Frame Size (FFS) and Variable Frame Size (VFS) implementations indicated no general superiority of one over the other in terms of performance. The latter is however capable of reducing the processing time and storage. Experiments on the number of templates per word revealed that an important factor in increasing the accuracy of a template-based speaker-independent recognition system is to incorporate large number of multiple templates.

The experiments reported in this paper is of an exploratory nature. Recognition rates vary across languages and test conditions. At least a hundred speakers are required to simulate realistic situations. However, a speaker population of a quarter of that size as used in our experiments adequately demonstrates the effectiveness of the implementation.

Recognition accuracy of the system may further be increased by incorporating heuristics. Another approach of improving the accuracy could be the use of other distance measures. It has been shown in [25] that the Weighted Likelihood Ratio (WLR) is more effective in vowel recognition. Since the Arabic language is dominated by vowel phonemes, WLR may be a more effective distance measure. The effectiveness of different LPC distance measures on Arabic speech recognition and their noise robustness properties are presently under study.

Table 7. Recognition Results (%) as Number of Training Speakers were Varied

Average Number of Speakers	Templates	Recognized	Rejected	Misrecognized
5	5	83.1	7.7	9.2
10	7	85.4	1.5	13.1
15	8	89.2	0.8	10.0
20	10	94.6	0.0	5.4

Table 8. Recognition Results (%) as Number of Templates were Kept Constant

Average Number of Speakers	Templates	Recognized	Rejected	Misrecognized
5	10	93.1	0.0	6.9
10	10	91.5	1.5	6.9
15	10	93.9	0.8	5.3
20	10	94.6	0.0	5.4

REFERENCES

[1] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition Theory and Selected Applications", *IEEE Transactions on Communications,* **29(5)** (1981), p. 621.

[2] D. A. Adams, "The Carnegie–Mellon University Distributed Speech Recognition System", *Speech Technology,* March/April 1986, p. 14.

[3] J. Makhoul, "Linear Prediction: a Tutorial Review", *Proceedings of the IEEE,* **63** (1975), p. 561.

[4] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Transactions on Acoustics Speech and Signal Processing,* **23(1)** (1975), p. 67.

[5] A. H. Gray and J. D. Markel, "Distance Measure for Speech Processing", *IEEE Transactions on Acoustics Speech and Signal Processing,* **24(5)** (1976), p. 380.

[6] M. Sugiyama and K. Shikano, "LPC Peak Weighted Spectral Matching Measures", *IEEE Transactions,* **J65-A(5)** (1981), p. 965.

[7] S. I. Nakagawa, "Speaker-Independent Phoneme Recognition in Continuous Speech by a Statistical Method and a Stochastic Dynamic Time Warping Method", *Technical Report, Computer Science Department, Carnegie–Mellon University,* January 1986.

[8] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEE Transactions on Acoustics Speech and Signal Processing,* **26(1)** (1978), p. 43.

[9] L. R. Rabiner and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition", *Journal of the Acoustical Society of America,* **66(3)** (1979), p. 663.

[10] J. T. Tau and R. C. Gonzalez, *Pattern Recognition Principles.* Reading, Massachusetts: Addison Wesley, 1974.

[11] R. A. Cole, R. M. Stern, and M. J. Lasry, "Performing Fine Phonetic Distinctions: Templates *versus* Features", in *Variability and Invariance in Speech Processes.* eds. J. S. Parkell and D. H. Klatt. Hillsdate, New Jersey: Lawrence Erlbaum, 1985.

[12] K. F. Lee, "Incremental Network Generation in Word Recognition", *Proceedings of International Conference on Acoustics Speech and Signal Processing,* vol. 1, 1986, p. 77.

[13] M. A. Hashis, A. T. Elkheshen, and M. R. Elghonemy, "Experiences in Isolated Arabic Word Recognition", *Workshop in Computer Processing of the Arabic Language,* Kuwait, April 14–16, 1985.

[14] O. S. Emam and M. A. Hashis, "Application of Hidden Markov Models to the Recognition of Isolated Arabic Words", *Proceedings of the 10th National Computer Conference, Jeddah,* February–March 1988.

[15] A. H. Gray and J. D. Markel, "Cosh Measure for Speech Processing", *Journal of the Acoustical Society of America,* **58** (Fall Suppl.) (1975), p. xxx.

[16] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the End Points of Isolated Utterances", *The Bell System Technical Journal,* **54(2)** (1975), p. 297.

[17] V. R. Viswanathan, J. Makhoul, R. M. Schwartz, and A. W. F. Huggins, "Variable Frame Rate Transmission: A Review of Methodology and Application to Narrow-Band LPC Speech Coding", *IEEE Transactions on Communications,* **30(4)** (1982), p. 674.

[18] R. Pieraccini and R. Billi, "Experimental Comparison Among Data Compression Techniques in Isolated Word Recognition", *Proceedings of International Conference on Acoustics Speech and Signal Processing,* **3** (1983), p. 1025.

[19] J. L. Gauyaip, J. Mariani, and J. S. Lienard, "On the Use of Time Compression for Word-Based Recognition", *Proceedings of International Conference on Acoustics Speech and Signal Processing,* **3** (1983), p. 1029.

[20] C. K. Chuang and S. W. Chan, "Speech Recognition Using Variable Frame Rate Coding", *Proceedings of International Conference on Acoustics Speech and Signal Processing,* **3** (1983), p. 1033.

[21] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", *Proceedings of the IEEE,* **66** (1978), p. 51.

[22] J. Durbin, "The Fitting of Time-Series Models", *Review of Institute of International Statistics,* **28(3)** (1960), p. 223.

[23] E. M. Hagos, "Implementation of an Isolated Word Recognition System", *M. S. Thesis, University of Petroleum and Minerals, Dhahran,* 1985.

[24] C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance Tradeoffs in Dynamic Time Wrapping Algorithm for Isolated Word Recognition", *IEEE Transactions on Acoustics Speech and Signal Processing,* **28** (1980), p. 622.

[25] K. Shikano, "Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition", *Technical Report, Department of Computer Science, Carnegie Mellon University,* February 1986.

Paper Received 29 September 1987; Revised 23 May 1988.