COMPARISON OF DIFFERENT SPEECH ENHANCEMENT METHODS ON RECOGNITION OF NOISY SPEECH

M. S. Ahmed* and A. M. Al-Marzoug

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

الخلاصـــة :

تـتم دراسة أداء مقاييس المسافة المختلفة مع طرق تـقليل الضوضاء في حالة التعرف على التَّكـلُّـم المصحوب بضوضاء والـمُستـقِل عن المتحدث . ولقد استُخدمت ضوضاء بيضاء وملونة تتبع المنحنى الطبيعي بعدة مستويات لإفساد الأصوات الأصلية . وكانت قاعدة المعلومات المستخدمة للدراسة المقارنة هي الأعداد العربية من صفر إلى عشرة . وأما المعلومات التى استُخدمت في التدريب فكانت من ثمانمائـة نطق تكلم بها عشرون متحدثا . أما في الاختبار فكانت (١٣٠) نطقاً تكلم بها ستـة متحدثون ، واستُخدمت طرق اكتشاف النهاية التلقائية والمبنية على حدود الطاقة سواء في نطق التدريب أو الاختبار . واستُخدمت البرمجة الديناميكية لنقطة النهاية الثابتـة في مرحلتي إيجاد المجموعات والتعرف . وأثبتت الدراسة أنَّ تحسناً ملحوظاً في التعرف على الكلام المصحوب بضوضاء يمكن أن ينجم باستخدام مقياس مسافة وطريقة تخفيض الضوضاء المناسبين .

*Address for correspondence:
KFUPM Box 134
King Fahd University of Petroleum & Minerals
Dhahran 31261
Saudi Arabia

ABSTRACT

The performance of different distance measures, along with different noise reduction techniques, on the LPC based, speaker independent, recognition of noisy speech is investigated. White as well as colored normally distributed additive noise of different levels was used to corrupt the test utterances. The data base used to carry out the comparative study consists of the ten Arabic digits 0−9. The training data were obtained from 800 utterances spoken by 20 different speakers, while for the test, 130 utterances were collected from six additional speakers. Automatic end point detection based on energy threshold is used for all training and test utterances. A fixed end point dynamic programming is used both in the clustering and recognition stages. Results of the study show that significant improvement in recognition of noisy speech is possible by the appropriate choice of distance measure and noise reduction technique.

COMPARISON OF DIFFERENT SPEECH ENHANCEMENT METHODS ON RECOGNITION OF NOISY SPEECH

1. INTRODUCTION

There are two quite separate directions in research on automatic speech recognition. One concentrates on exploring operational methods for applying higher level information to the decoding of the acoustic ambiguities encountered when recognizing larger vocabularies and continuous speech. The other explores the practical aspects of speech recognizers. This involves concentrating on studying the effect of field conditions, such as noise and room reverberation, on recognizers designed in the laboratory and on how to attain high recognition rates in the field. This paper concerns the latter direction, where we have considered the noise aspects of speech recognition.

In the design of a practical speech recognition system, consideration of the effect of noise on the test speech is one of the most important factors. Although, in most systems, recognition of noise-free utterances can attain a high accuracy rate, the presence of noise in the utterances is observed to severely reduce the recognition accuracy. The problem has attracted many researchers [1–5]. The approaches taken to improve the recognition accuracy may be divided into three classes: (1) designing the recognizer to function in a specific noise environment; (2) pre-processing the signal to remove the noise; and (3) use of noise-robust features and distance measures.

Approach (1) tunes either the features or the distance measures for a specific noise environment. The tuning rule is required to be determined for every new situation (*i.e.* the system requires calibration). Approach (2) uses a speech enhancement algorithm as a pre-processor. The enhanced speech is then subjected to a recognizer designed for clean speech recognition. Approach (3) uses features and distance measures which are relatively insensitive to the additive noise. The latter two approaches are therefore more flexible. In order to exploit their full power, a comprehensive study of a combination of these techniques is needed.

In this study, the performance of different LPC distances, along with different noise reduction techniques, on the recognition of speaker-independent noisy isolated utterance is investigated. The LPC distances considered are LPC cepstrum [6], two forms of asymmetrical log likelihood ratios [7], symmetrical log likelihood ratio [7], and symmetrical and asymmetrical weighted likelihood ratios [8]. The noise reduction techniques considered are simple spectral subtraction [9–11], spectral over-subtraction with use of a spectral floor [12], spectral subtraction with residual noise removal [13], and time and frequency domain adaptive minimum mean square error (MMSE) filtering [14].

In reference [1], the performance of the above distance measures and enhancement algorithms are studied in terms of the spectral perturbation. It is further shown that the use of an enhancement algorithm can improve the recognition of noisy phonemes. In reference [3], WLR is shown to be a better distance measure compared to CEP and LLR in the recognition of noisy speech when no enhancement is applied. Further, references [1] and [20] pointed out the general superiority of MMSE enhancement algorithms compared to spectral subtraction methods in matching the spectral properties with the clean speech.

However, a distance measure that is superior for noisy speech recognition may not retain its superiority in enhanced speech recognition. Even though one may conjecture that a combination of distance measure and enhancement algorithm that results in a smaller distance between the clean and enhanced frames of similar sound is a better candidate for noisy speech recognition, it may be also argued that the enhancement, being a filtering operation, may also decrease the distance measure between frames of dissimilar sounds by spectral smoothing. The recognition accuracy may further be affected by error in end point detection of noisy speech.

In order to arrive at a sound conclusion on the effectiveness of distance measures and enhancement algorithms for a noisy isolated-word recognition system, and to pick up the "best" combination of them, a comparative study is needed directly on an isolated speech recognition system. The present study aimed precisely in that direction.

The data base used to train the speaker-independent recognition experiment consists of four replications of the ten Arabic digits zero to nine from each of twenty speakers, containing a total of eighty training utterances per digit. The reference templates are

obtained by applying clustering [15] corresponding to each LPC distance function. The data base used for the recognition consists of thirteen utterances of each digit from an additional six speakers. White, as well as colored, normally-distributed noise at different levels was added to these test utterances.

Automatic end point detection based on energy threshold is used for all training and test utterances. Fixed end point dynamic programming [16] and the appropriate distance function are used both in the clustering and in the recognition stages.

2. LPC DISTANCE MEASURES

One of the advantages of LPC modeling is that it allows for computationally efficient frame-to-frame distance measures. They are extremely sensitive to change in the spectral shape of the speech segment. They have been widely used in speech and speaker recognition and speech coding. Gray [6] considered several properties which the distance measure should satisfy. They are: (1) non-negativity; (2) physically meaningful interpretation in the frequency domain; and (3) computational efficiency.

The spectrum matching measures considered in this study are the LPC cepstrum (CEP), two forms of the asymmetrical log likelihood ratios (LLR-1, LLR-2), symmetrical log likelihood ratio (SLLR), symmetrical weighted likelihood ratio (SWLR), and the asymmetrical weighted likelihood ratio (UWLR). CEP is the mean square log spectral difference between the LPC power spectrums of two speech frames. LLR-1 and LLR-2 are the logarithm of the ratios of prediction residuals that results when one speech frame is filtered through the inverse LPC filter of another frame. This measure also may be given a frequency domain interpretation [6], as related to the LPC power spectrum ratio integral. Since these distances are not symmetric, a symmetrical LLR (SLLR) can be obtained by taking their average.

The above measures do not consider human hearing perception. The weighted likelihood ratios (WLR) relate the distance measures to human auditory characteristics by weighting the dissimilarities in the frequency domain at the spectral peak (or formants) to which the human hearing system is most sensitive. The symmetrical WLR (SWLR) weighs the distance at the peak of both the reference and the test frames while the asymmetrical WLR (UWLR) weighs the distance only at the peak of the reference frames. Since the references are created from noise free utterances, this approach puts more weight at the clean speech spectral peaks and therefore is expected to be noise robust. The distance measures used in this study along with the equations used for their computation are listed in Table 1. The unprimed variables indicate the template parameters obtained from the training utterances, while the primed variables indicate the corresponding parameters of the test utterances.

3. SPEECH ENHANCEMENT ALGORITHMS

The enhancement algorithms considered in this study are spectral subtraction [9–11], spectral over-subtraction with use of a spectral floor [12], spectral subtraction with residual noise removal [13], time-domain minimum mean square error (T-MMSE) algorithm [14], and frequency-domain minimum mean square error (F-MMSE) algorithm [14]. Other enhancement algorithms such as comb filtering [17] and adaptive noise canceling [18] are not considered as they required the computationally demanding and error prone step of pitch determination from noisy speech.

The algorithms considered are very briefly described below. In addition, Table 2 shows the parameter values used for different enhancement algorithms.

3.1. Spectral Subtraction Method

A spectral subtraction method to enhance degraded speech is considered by Weiss *et al.* [9–10] and Lim [11]. The algorithm divides the speech into frames and subtracts the short time spectral magnitude of the noise from that of the noisy speech. The time-domain speech is reconstructed from the estimated spectral magnitude and the unprocessed phase. More specifically, the spectral subtraction method computes the speech spectrum as

$$\hat{S}(\omega) = [X(\omega)^{\theta} - V(\omega)^{\theta}]^{1/\theta} \qquad (1)$$

where $\hat{S}(\omega)$, $X(\omega)$, and $V(\omega)$ are the DFT of the enhanced speech, noisy speech, and the noise respectively. $\hat{S}(\omega)$ is set to zero when Equation (1) is negative and θ is a parameter introduced for generality, which affords a degree of flexibility. It can be shown that the choice $\theta = 2$ is equivalent to the correlation subtraction method.

3.2. Spectral Over-Subtraction and Use of a Spectral Floor

The spectral subtraction method, although capable of reducing the wide band noise, also introduces a

Table 1. LPC Distance Measures.

Abbr.	Distance	Reference	Equation
CEP	LPC Cepstrum	[10]	$\sum_{i=1}^{q} (c_i - c'_i)^2$
LLR1	Log Likelihood Ratio 1	[11]	$\log \frac{\sum_{i=1}^{p} A_i r'_i}{\sum_{i=1}^{p} A'_i r'_i}$
LLR2	Log Likelihood Ratio 2	[11]	$\log \frac{\sum_{i=1}^{p} A'_i r_i}{\sum_{i=1}^{p} A_i r_i}$
SLLR	Symmetrical LLR	[11]	½(LLR1+LLR2)
SWLR	Symmetrical Weighted Likelihood Ratio	[12]	$\sum_{i=1}^{q} (c_i - c'_i)(r_i - r'_i)$
UWLR	Asymmetrical Weighted Likelihood Ratio	[12]	$(c_0 - c'_0) r'_0 + \sum_{i=1}^{q} (c_i - c'_i) r_i$

$c_i \triangleq$ LPC cepstrum coefficient; $r_i \triangleq$ speech autocorrelation coefficient;
$A_i \triangleq$ LPC autocorrelation coefficient; $p \triangleq$ LPC model order = 12; q = 18.

Table 2. Parameters in the Enhancement Algorithms.

Algorithm	Ref.	Parameter values
Spec. Sub.	[9]	$\theta = 1$, see Equation (1).
Berouti	[12]	$\alpha = \begin{cases} \alpha_0 + 5/\nu & \text{if SNR} < -5\,\text{dB} \\ \alpha_0 - \text{SNR}/\nu & \text{if } -5\,\text{dB} < \text{SNR} < 20\,\text{dB} \\ 1 & \text{if SNR} > 20\,\text{dB} \end{cases}$ $\beta = 0.2$, $\theta = 1$, $\alpha_0 = 2$, $\nu = 20$; $\alpha \equiv$ spectral overcorrection factor; $\beta \equiv$ spectral floor factor; $\theta \equiv$ a spectral parameter, see Equation (1).
Boll	[13]	$\theta = 1$, see Equation (1) MNR is computed in three successive frames; Enhanced signal is further attenuated by −30 dB if SNR < −12 dB.
TMMSE	[14]	Filter length $r = 12$, see Equation (3); Correlation functions are computed from 5 adjacent frames with the target frame being at the center.
FMMSE	[14]	$\gamma = 2$, see Equation (6); PDS are computed from 5 adjacent frames with the target frame being at the center.

SNR = Computed signal to noise ratio in a frame; see [12, 13]

new "musical" or "warbling" noise due to the presence of remaining spectral peaks. To reduce the musical noise, Berouti *et al.* [12] suggested over-subtraction of the noise spectrum and introduction of a non-zero spectral floor. The over-correction eliminates most of the spectral peaks and the use of a positive spectral floor fills the deep valley, the combined effect of which is to reduce the "musical" noise. For guidelines on the over-correction and the spectral floor parameters, the reader is referred to [12].

3.3. Spectral Subtraction with Residual Noise Removal

Boll [13] suggested a method of reducing the "musical" noise by measuring the frame-to-frame randomness of the noise. The maximum noise residual (MNR), *i.e.*, the maximum value by which the noise spectrum may exceed the average noise spectrum, is computed for each frequency bin. After the spectral subtraction, if the MNR in any of the frequency bins exceeds the spectral magnitude of three successive frames, then the center frame is processed further. For implementation details the reader is referred to [13].

3.4. Time-Domain Adaptive MMSE Filtering

Minimum mean squared Error (MMSE) was pioneered by Wiener [19]. The resulting filter which is generally non-causal, tends to suppress the noise while leaving the signal relatively unaffected. The design of such filters requires both the signal and noise to be stationary. Although speech is a non-stationary signal, the fact that its statistics change slowly with time can be used advantageously in applying the adaptive MMSE filtering.

Time-domain adaptive filtering of noisy speech exploiting local stationarity is considered by Ahmed [14]. The algorithm segments the noisy speech into overlapping frames, windows each frame, filters each frame by a time domain linear filter, and overlap adds the filtered frames. Suitable windowing ensures smooth transition of the filter coefficients. The filter coefficients are obtained as follows:

Define the speech, noise, and noisy speech by s_t, v_t, and x_t such that:

$$x_t = s_t + v_t. \tag{2}$$

The enhanced speech in each windowed frame is obtained as

$$s_{t+\tau} = \sum_{i=0}^{r} c_i x_{t-i} \tag{3}$$

with $\tau = -r/2$ and r the order of the enhancement filter. The filter coefficients c_i are estimated from

$$r_{xs}(\tau + l) = \sum_{i=0}^{r} c_i r_{xx}(l-i), \quad l = 0, 1, \ldots, r, \tag{4}$$

where $r_{xs}(i) = r_{ss}(i) = r_{xx}(i) - r_{vv}(i)$; r_{xx} and r_{vv} are the corresponding autocorrelation functions computed from the time averages assuming local stationarity. Since the system of equations in (4) is topelitz, a computationally efficient algorithm may be used to estimate c_i. Implementation details are given in [14].

3.5. Frequency-Domain MMSE Filtering

Frequency-domain adaptive MMSE filtering that is based upon the same principle, but with the filtering performed in the frequency domain, is also considered in [14]. The algorithm computes the DFT of overlapping frames, filters each frame in the frequency domain and transforms back to the time-domain through an IDFT.

In each windowed frame, the enhanced speech $\hat{S}(\omega)$ is obtained as

$$\hat{S}(\omega) = \frac{P_s(\omega)}{P_x(\omega)} X(\omega) \tag{5}$$

where $P_x(\omega)$ and $P_s(\omega)$ are the corresponding power density spectra (PDS). $P_s(\omega)$ is obtained through spectral subtraction as

$$P_s(\omega) = p_x(\omega) - \gamma P_v(\omega). \tag{6}$$

The parameter γ adds considerable flexibility in the algorithm and is required to obtain a smooth PDS of speech. $\gamma > 1$ corresponds to an over-correction which may be used to eliminate the remaining peaks and valleys in the PDS computed from simple subtraction [12]. An experimentally obtained suitable value of γ is found as 2. Following the approach of [20], $P_x(\omega)$ and $P_v(\omega)$ are obtained by averaging the squared amplitude spectrum of a few adjacent frames, assuming local stationarity. Implementation details are given in [14].

4. EXPERIMENTAL PROCEDURE

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)

Four utterances of every word from each of twenty male speakers were recorded in a normal laboratory environment. 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 speaker-independent recognition system another 130 utterances were created from six additional speakers with different accents and dialects who were not used for the creation of the training set. The analog utterances were band pass filtered between 80–3500 Hz. A 2.5 s portion of each utterances embedded in silence was sampled to obtain 20 000 samples.

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

$$p(z) = 1 - 0.95\, z^{-1}. \tag{7}$$

A 12th order LPC modeling with Hamming window was incorporated. An overlapping fixed frame length of 32 ms with increment of 16 ms was used. In order to reduce the number of templates to a manageable size, clustering was applied to the training utterances. The algorithm of Rabiner and Wilpon [13] was used for this purpose. The number of clusters per digit ranged between 8 and 13, with the average being close to 10. For the clustering of training data and the recognition of test utterances, a fixed end point dynamic time warping was employed [16]. The decision rule to classify the unknown utterances was based on a $k-nn$ rule with $k = 3$. Details of the word recognition system is given in [21–23].

For the enhancement algorithms, 32 ms frames, 16 ms increment, and Hanning window were used. The filter order for the T-MMSE algorithm was set to 12. The noise statistics were obtained from the average of the first five "silent" frames.

4.1. End Point Detection of Noisy Speech

End point detection is an essential step in speech recognition. The success of end point detection directly affects the recognition accuracy. A number of end point detection algorithms well suited for speech with high S/N ratio are available in the literature [24, 25]. They typically use the short time energy and zero crossing rates. However, in the case of noisy speech they require some modifications.

For the pre-processor based recognition of noisy speech two approaches can be adopted. End points may be detected either from the noisy speech or from the processed (enhanced) speech. Our investigation has shown that in either of these approaches the zero crossing rate does not show any discriminating pattern when the speech is corrupted with broad band noise. On the other hand, the short time energy continues to show a discriminating pattern. This is demonstrated in Figure 1, where the enhancement was done using the T-MMSE algorithm.

Therefore, only the short time energy was used for the end point detection in both the training and test utterances. This approach failed to detect the weak fricatives at the end. But as the same approach was taken for the training and the test utterances this problem did not have any detrimental effect on the recognition. The short time energy threshold as suggested in [24] was used for the end point detection. It was observed that accurate end point detection was possible from the enhanced speech by the algorithm of [18] without any major modification when the T-MMSE and F-MMSE were used for enhancement.

However, the end point detection from the enhanced speech was much in error when the spectral subtraction methods were used. This can be attributed to the heuristics employed in these approaches (such as use of an artificial spectral floor, residual noise removal, forced attenuation *etc*). For the spectral subtraction methods, end point detection from the noisy speech yielded a better result, although the short time energy threshold required modification to adapt to noisy speech.

For the adaptive MMSE algorithms end point detection was done from the enhanced speech using the following energy threshold:

$$\begin{aligned} \mathrm{I1} &= \Gamma * (\mathrm{IMX} - \mathrm{IMN}) + \mathrm{IMN} \\ \mathrm{I2} &= 4 * \mathrm{IMN} \\ \mathrm{ITL} &= \min(\mathrm{I1}, 12) \\ \mathrm{ITU} &= 5 * \mathrm{ITL}. \end{aligned} \tag{8}$$

where IMX and IMN are the peak and "silence" energies of an utterance. ITL and ITU are the lower and upper energy thresholds respectively [24]. Γ was 0.023 for the T-MMSE and 0.032 for the F-MMSE. For the other enhancement algorithms, the noisy speech was used for the end point detection with the following energy thresholds:

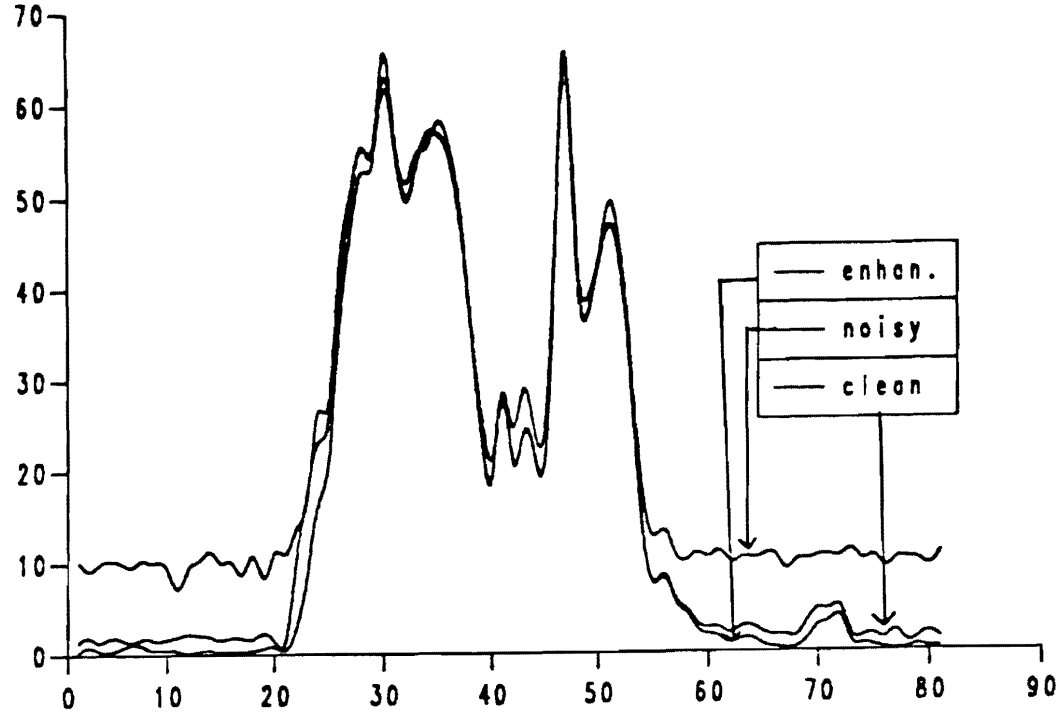

Figure 1(a). *Short Time Energy of the Word* Wahd.

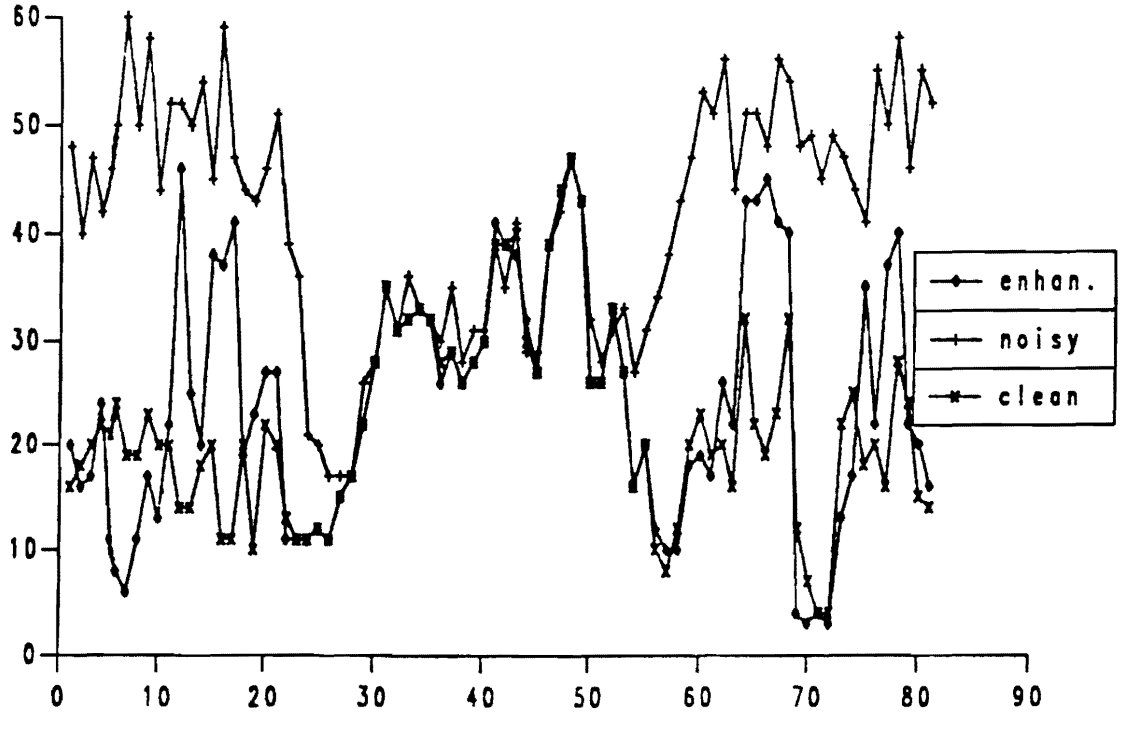

Figure 1(b). *Zero Crossing Rates of the Word* Wahd.

$$\begin{aligned} I1 &= 0.027 * (IMX - IMN) + IMN \\ I2 &= 4 * IMN \\ ITL &= \min(I1, I2) \\ ITU &= 0.15 * (IMX - IMN) + IMN. \end{aligned} \quad (9)$$

5. RESULTS

Table 3 shows the recognition result for the noise free test utterances. It shows the recognition, misrecognition, and rejection scores. The rejection threshold for each digit was set equal to the mean of the distance between the reference utterances of the same digit. It can be observed that SWLR attained the highest recognition rate. This may be explained from the fact that spoken Arabic has more vowel content compared to English and SWLR assigns higher weight at the spectral peaks which are the most important characteristics in distinguishing vowel phonemes. This is consistent with Shikano's [8] claim that WLR is a better distance measure for vocabularies with high vowel contents. Another important observation is the fact that symmetric measures performed better compared to the asymmetrical ones. This is typical for clean speech recognition, where the role of reference and test utterances may be reversed.

Figure 2 depicts the effect of noise on the recognition accuracy when the test utterances were corrupted with white noise. Although the noise has detrimental effect on the recognition rate, the superiority of the WLR measures is clearly evident. As the presence of noise distorts the LPC spectral model of the test utterance, the relative distortion is more in the valleys compared to the peaks. This mismatch at the valleys causes the LLR and CEP distances to perform poorly. On the other hand, WLR gives more weight to the peaks and largely overcomes the effect of distortion. It can be also observed that asymmetric measures that assign weights proportional to the reference (clean) spectrum performed better compared to the symmetric measures. Thus UWLR and LLR1 performed better than SWLR and SLLR respectively.

Table 3. Clean Speech Recognition (%).

Metric	Recognized	Misrecognized	Rejected
CEP	94.6	5.40	0.00
LLR1	90.0	9.23	0.77
LLR2	91.5	8.50	0.00
SLLR	94.6	5.40	0.00
SWLR	96.2	3.80	0.00
UWLR	94.6	5.40	0.00

In the recognition of noisy speech with enhancement, it was observed that all of the methods are generally capable of improving the recognition rate at high to moderate S/N ratio. Figure 3 shows a comparison of different distance measures when the noise reduction was carried out with T-MMSE algorithm. The improvement on the over all recognition accuracy and the superiority of the UWLR measure is clearly evident. Figure 4 depicts a similar comparison but with noise reduction being done by Berouti's algorithm ($\theta = 1$). In this figure the LLR-1 and UWLR out-performed the other distance measures.

Figures 5 and 6 compare different enhancement algorithms when the UWLR and LLR-1 respectively were used as the distance measures. The superiority of the MMSE algorithms can be observed from these illustrations. It was observed that MMSE algorithms yielded most satisfactory results with all the distance measures. For the LLR-1 measure, Berouti's algorithm produced comparable results. Although results

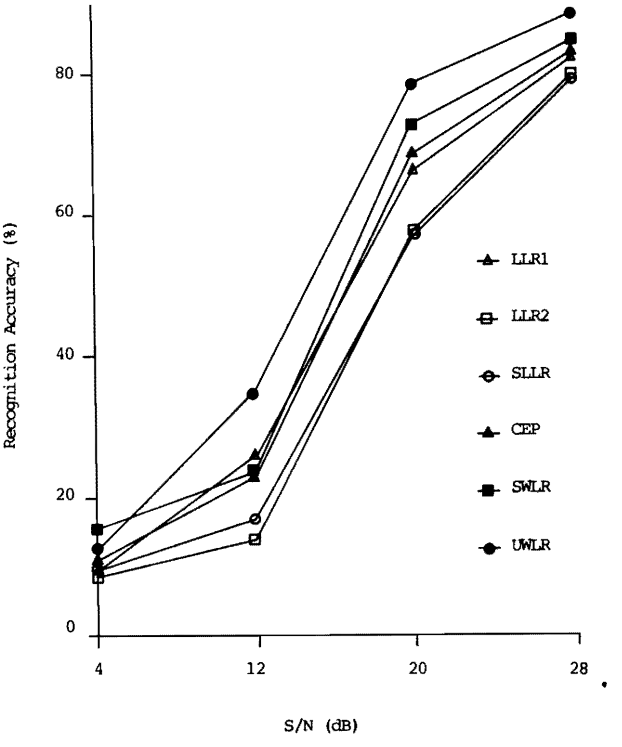

Figure 2. Recognition Accuracy in the Presence of White Noise.

for white additive noise are presented here, high pass and low pass additive noise have shown a similar trend. In general, the WLR measures outperformed the LLR & CEP measures and MMSE algorithms outperformed the other enhancement algorithms. In addition, the UWLR and LLR1 performed better compared to their symmetrical counterparts. Among the spectral subtraction methods, Berouti's algorithm [7] (with $\theta = 1$) performed best. A combination of Berouti's algorithm and LLR-1 distance measure yielded a high recognition score.

6. CONCLUSIONS

The performance of different LPC distance measures and a number of noise reduction algorithms for the recognition of noise corrupted speech was investigated on a LPC based, speaker independent Arabic digit recognizer. The noise was additive and normally distributed. White and colored noise were used. It was observed that, while symmetrical distance measures yield better recognition for clean test utterances, asymmetrical ones yield better results for noisy test utterances. In general, the WLR measures outperformed the CEP and LLR measures when the test utterances are noisy. Among the noise reduction algorithms, MMSE filtering is found to yield better recognition compared to the spectral subtraction methods. Among the spectral subtraction methods Berouti's algorithm yielded the best results. Combinations of "MMSE (time or frequency domain) and UWLR" and "Berouti and LLR-1" yielded the highest recognition scores. Further studies are underway on implicit noise reduction where the noise reductions are applied directly to the features extracted from the noisy speech.

A key factor in realistic recognition of noisy speech is the difference in the way humans speak in a noisy environment. They tend to speak more slowly and clearly, which may have a major effect on the recognition performance. Further study could be done to investigate this factor by making speech recordings in noisy environments. In addition, in this paper, like most others, the conclusions are drawn from a single experiment (*i.e.* only one noise sequence is added per utterance per S/N ratio), which does not provide any confidence interval on

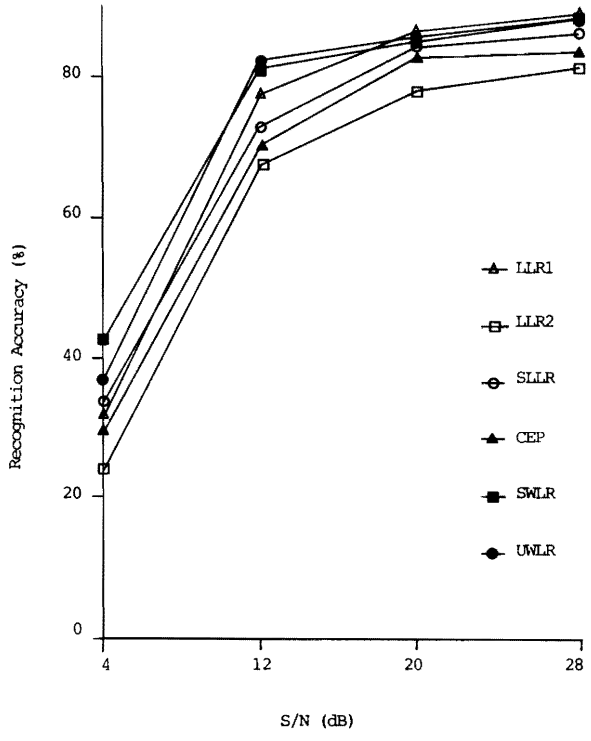

Figure 3. Recognition Accuracy from Speech Enhanced by T-MMSE Algorithm.

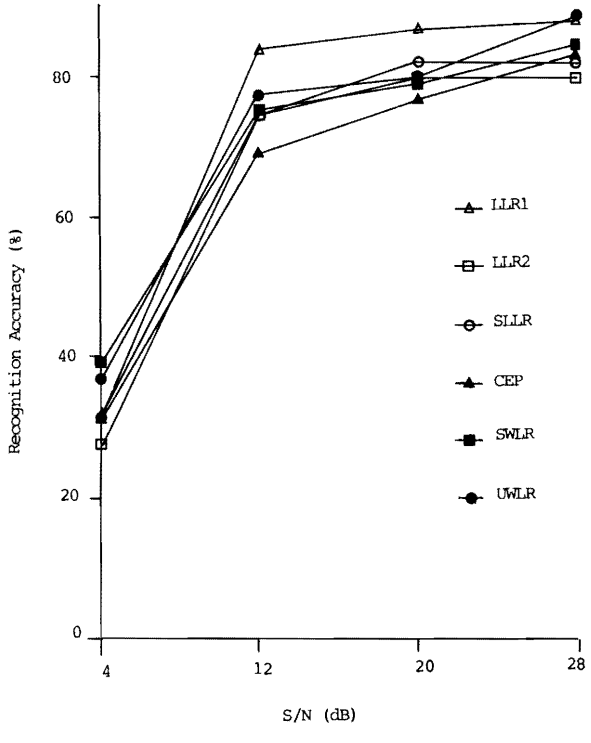

Figure 4. Recognition Accuracy for Speech Enhanced by Berouti's Algorithm.

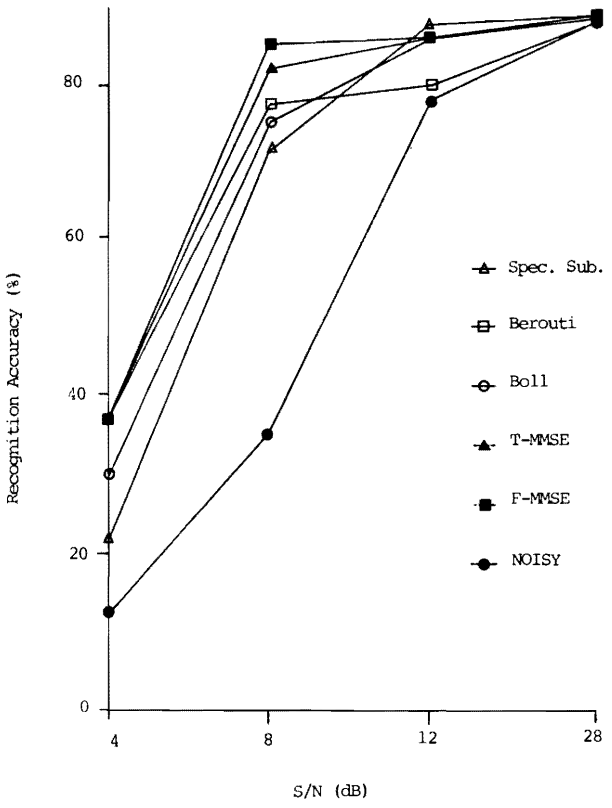

Figure 5. Recognition Accuracy from Enhanced Speech Using the UWLR Measure.

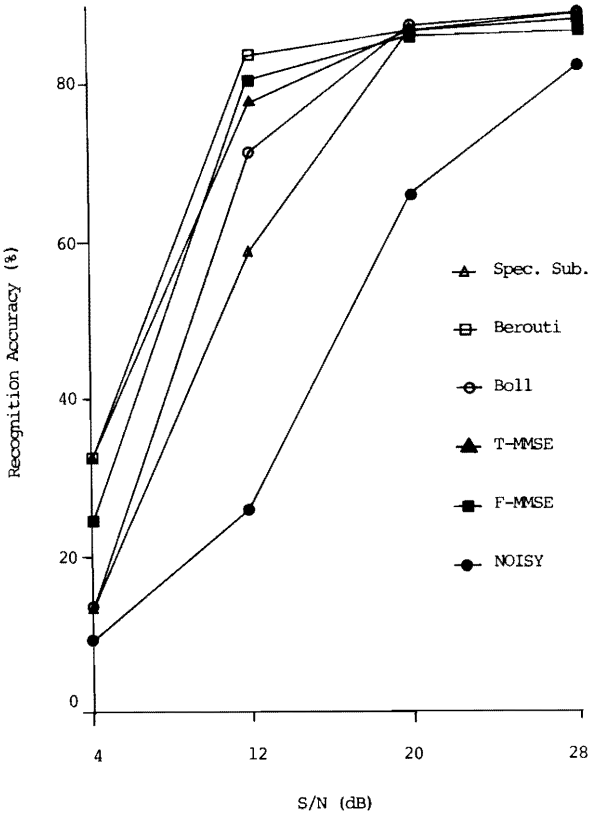

Figure 6. Recognition Accuracy from Enhanced Speech Using the LLRI Measure.

the results. A Monte-Carlo simulation can be done to obtain statistics by performing many (≥ 30) experiments which will provide the confidence interval on the results and will account for the statistical variability. However, to keep the computation to a manageable magnitude, the numbers of enhancement algorithms/ distance measures/ signal-to-noise ratios have to be reduced. A further area of future research would be to investigate the combination of speech enhancement and the Hidden Markov Model (HMM) technique for noisy speech recognition.

ACKNOWLEDGEMENT

The authors acknowledge King Fahd University of Petroleum & Minerals for its support.

REFERENCES

[1] M. S. Ahmed, "Comparison of Noisy Speech Enhancement Algorithms in Terms of LPC Perturbation", *IEEE Trans. Acoust. Speech, Signal Processing*, **ASSP-37** (1989), p. 121.

[2] C. Mokbel and G. Chollet, "Word Recognition in the Car — Speech Enhancement/Spectral Transformation", in *Proceedings: Int. Conf. Acoust., Speech, Signal Processing, Toronto, May 1991*, p. 925.

[3] H. Matsumoto and H. Imai, "Comparative Study of Various Spectrum Matching Measures on Noise Robustness", in *Proceedings: Int. Conf. Acoust., Speech, Signal Processing, Tokyo, April 1986*, p. 769.

[4] A. Errel and M. Weintraub, "Estimating Log Spectral Distance Criterion for Noise Robust Speech Recognition", in *Proceedings: Int. Conf. Acoust., Speech, Signal Processing, Dallas, April 1990*, p. 853.

[5] A. Acro and R. M. Stern, "Robust Speech Recognition by Normalization of the Acoustic Space", in *Proceedings: Int. Conf. Acoust., Speech, Signal Processing, Toronto, May 1991*, p. 893.

[6] A. H. Gray and J. D. Markel, "Distance Measures for Speech Processing", *IEEE Trans. Acoust. Speech, Signal Processing*, **ASSP-24** (1976), p. 380.

[7] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Trans. Acoust. Speech, Signal Processing*, **ASSP-23** (1975), p. 67.

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

[9] J. S. Lim, "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive

White Noise", *IEEE Trans. Acoust. Speech, Signal Processing,* **ASSP-26** (1978), p. 471.

[10] M. R. Weiss, F. Aschkenasy, and T. W. Parsons, "Study and Development of the INTEL Technique for Improving Speech Intelligibility", *Tech. Rep. NSC, FR/4023,* Nicolet Scientific Corp., 1974.

[11] M. R. Weiss, F. Aschkenasy, and T. W. Parsons, "Processing Speech Signals to Attenuate Interferences", *IEEE Symposium on Speech Recognition, April 1974.*

[12] M. Berouti, B. Schartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", in *Proceedings: Int. Conf. Acoust., Speech, Signal Processing, April 1979,* p. 208.

[13] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoust. Speech, Signal Processing,* **ASSP-27** (1979), p. 113.

[14] M. S. Ahmed, "Speech Enhancement by Adaptive MMSE Filtering", *Tech. Rep., Comput. Sci. Dep.,* Carnegie–Mellon University, Pittsburgh, PA, 1986.

[15] L. R. Rabiner and J. G. Wilpon, "Consideration in Applying Clustering Techniques to Speaker Independent Word Recognition", *J. Acoust. Soc. of America,* **66(3)** (1979), p. 663.

[16] L. R. Rabiner, C. S. Myers, and A. E. Rosenberg, "Performance Trade Off in Dynamic Time Warping Algorithm for Isolated Word Recognition", *IEEE Trans. Acoust. Speech, Signal Processing,* **ASSP-28** (1980), p. 622.

[17] R. H. Frazer, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of Speech by Adaptive Filtering", in *Proceedings: Int. Conf. Acoust., Speech, Signal Processing, Philadelphia, April 1976,* p. 251.

[18] M. R. Sambur, "Adaptive Noise Canceling for Speech Signals", *IEEE Trans. Acoust. Speech, Signal Processing,* **ASSP-26** (1978), p. 419.

[19] N. Wiener, *Extrapolation and Smoothing of Stationary Time Series with Engineering Applications.* New York: Wiley, 1949.

[20] Y. Ephrahim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator", *IEEE Trans. Acoust. Speech, Signal Processing,* **ASSP-32** (1984), p. 1109.

[21] M. S. Ahmed and E. M. Hagos, "Implementation of an Arabic Digit Recognition System", *Arabian Journal for Science and Engineering,* **14(1)** (1989), p. 79.

[22] M. S. Ahmed and E. M. Hagos, "Experiment on a Speaker Independent Digit Recognition System", in *Proceedings: Int. Symp. on Signal Processing and Appl., Brisbane, August 1987,* p. 605.

[23] A. M. Al-Marzoug, "LPC Based Recognition of Noisy Speech", *M.S. Thesis, King Fahd University of Petroleum & Minerals, Dhahran,* 1989.

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

[25] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An Improved End Point for Isolated Word Recognition", *IEEE Trans. Acoust. Speech, Signal Processing,* **ASSP-29** (1981), p. 777.

Paper Received 9 November 1991; Revised 20 September 1992.