Distance Measures Based Upon Adaptive Filtering For Robust Speech Recognition In Noise

잡음 환경하에서의 음성 인식을 위한 적응필터링 거리 척도에 관한 연구

(W. G. Chung*, C. K. Un*)

정 원 국*, 은 종 관*

ABSTRACT

When background noise is present, speech recognizers tend to yield performance degradation. In this paper, we propose distance measures which are robust to the background noise. We assume that the feature vectors of degraded speech can be considered as the output of a finite impulse response(FIR) system whose input is the feature vectors of corresponding clean speech. The FIR system represents the effect of noise. The unknown system parameters are estimated using the recursive least squares(RLS) algorithm. We define distance measures in terms of the estimation errors. The single channel first-order FIR structure shows the best recognition performance and thus allows to make a computationally efficient algorithm. Our speaker-independent isolated word recognition tests at various signal-to-noise ratio(SNR) levels show that our proposed distance measure performs better than the Euclidean distance measure and the projection distance measures.

요 약

잡음이 있는 환경하에서는 음성인식의 성능이 현저하게 떨어지게 된다. 본 논문에서는 이러한 잡음의 영향에 강한 거리척도를 제안하고자 한다. 우리는 잡음이 더해진 음성신호의 특징벡터를 깨끗한 음성신호의 특징벡터가 FIR 시스템을 거쳐 변형된 것이라고 가정한다. 여기서 FIR 시스템은 잡음의 영향을 모델링한 것이라고 할 수 있다. 미지의 FIR 시스템계수들은 RLS 적응 알고리즘을 이용하여 구한다. 제안된 거리척도는 적응 여과기의 예측 오차에 관한 식으로 표시되어진다.

여러가지 적응 여과기의 구조중 단일 채널 일차 FIR 구조가 가장 좋은 음성 인식 성능을 보이며, 이 경우 효과적인 거리척도 알고리즘을 구할 수 있다. 여러가지 신호대 잡음비에 관하여 화자독립 격리단어 인식 실험을 DTW 알고리즘을 이용하여 수행하여 본 결과 제안된 거리척도가 거의 모든 신호대 잡음비에 대하여 우수한 성능을 보였다.

*Communications Research Laboratory
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology

I. Introduction

Recently the desirability of achieving robust

speech recognition in noise has attracted a great deal of interest. Sometimes there occurs a situation for which there is a mismatch in the training and the testing conditions. This mismatch condition is one major source of performance degradation for speech recognizers. A speech recognizer designed to perform well under noise-free conditions often shows remarkable degradation in performance when background noise is present. Consequently, efficient methods for minimizing the effects of the background noise are needed.

A number of solutions to this problem involve ways to effectively remove noise from the degraded speech. This is usually in the form of enhancement techniques such as spectral subtraction, Wiener filtering, and microphone arrays[1]. However, such methods have shown limited performance improvements and can only be implemented under conditions where an estimate of the noise statistics is known. The approach taken in this paper requires no estimate of the noise statistics. The focus of interest is on developing distance measures which are robust to the effects of noise rather than trying to remove noise from the degraded speech. Recently, similar approaches have been taken by several researchers to find both robust distance measures, such as projection distance[2][3] and frequency-weighted Itakura spectral distance[1], and robust feature representations, such as the IMELDA representation[5] and the short-time modified coherence representation[6].

In this paper, our aim is to model the effects of noise and to develop new distance measures based upon this modeling. It is well known that the effect of additive white noise is to produce a smoothed spectrum[7] and to reduce the norm of linear predictive coding(LPC) cepstral vectors[2][3]. We assume that the feature vectors of degraded speech is the output of a finite impulse response(FIR) system whose input signal is the feature vector of corresponding clean speech. And the FIR system is assumed to represent the effect of noise. Hence, the effect of noise is modeled to transform the feature vector of clean speech into that of degraded speech by the FIR system.

For the sake of convenience, we divide the space of feature of speech utterances into two spaces:clean and noisy space. We try to define distance measures in the noisy space as an effort for matching the training and the testing conditions. Thus, the feature vector of clean reference template is first transformed into that of the noisy space by the FIR system which represents the effect of noise. And then a distance measure is defined between the degraded test feature vector and the transformed reference feature vector. The unknown system parameters can be found using adaptive filtering techniques, in which the desired signal is the noisy test speech and the input signal is the clean reference template. We can identify the unknown FIR system by minimizing the square of estimation errors, as can be seen in Fig. 1. This is a kind of

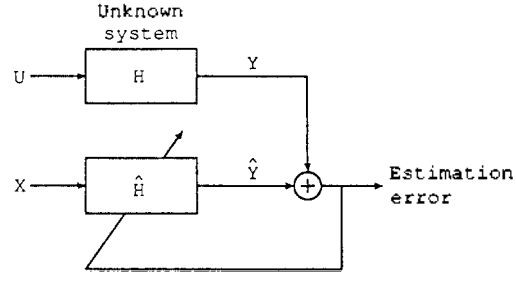

Fig. 1. Identification of unknown system

the system identification problem. We define the distance measure in terms of the estimation error. Note that since the FIR filtering is done in linear fashion, the feature vectors should keep the Euclidean properties. If the input signal X has the same phonetic contents as the desired signal Y, we expect the proposed distance to be small. And, if X is different from Y in phonetic contents, we expect the distance to be large.

Recently, a family of distance measures based on projection operation has been proposed by Mansour and Juang[2][3]. They showed analyti-

cally and empirically that additive white noise reduces the norm of LPC cepstral vectors. Using this observation, they compensated the norm reduction by a first-order equalization. We will show that the projection distance measure is a special case of our proposed distance measure, which is a first-order realization Although we here confine ourselves in considering only the LPC cepstral vectors and additive white noise, our concept of distance can be expandable to other interfering noise and other feature representations which have the Euclidean properties.

This paper is organized as follows. First, the formulation of the proposed distance measures is presented. Following this, recognition resultare presented using the proposed distance measures at various signal-to-noise ratio(SNER) levels. Last, a summary of the results is given.

Ⅱ. Formulation of Distance Measures

Let us denote X to be a reference word and Y to be a test word which has been contaminated with noise. They are sequences of feature vectors as follows.

$$\begin{aligned} X &= \{x(0), x(1), \cdots, x(L-1)\} \\ Y &= \{y(0), y(1), \cdots, y(L-1)\} \end{aligned} \tag{1}$$

where $x(i)$ and $y(i)$ are P dimensional feature vectors, and L is the number of frames in X and Y. For the simplicity of description, we assume that X and Y have the same number of frames. Actually, however, since they have different numbers of frames, we need a kind of alignment. It has been shown that the dynamic time warping(DTW) is an efficient algorithm to do the time alignment.

It searches for optimal path using dynamic programming. For the ease of decsription, we assume that X and Y have already time-aligned.

As described in the previous section and can be seen in Fig. 1, Y is considered as the outputs of an FIR system which represents the effect of noise. We try to identify the unknown FIR system and develop a new distance measure in terms of the estimation errors. The unknown system parameters can be found using adaptive filtering technique, in which the desired signal is the test word Y and the input signal is the reference word X. The adaptive filter is adjustable by minimizing some measure of error between the desired signal Y and the filter output $\hat{\mathbf{y}}$. First, we consider a general transversal multichannel FIR linear filter structure: the time-varying impulse response of the filter is denoted by the $P\times NP$ weight matrix $\Phi_N(t)$, where N is the filter order ane t is the frame index. The error signal at time t can be expressed as

$$\begin{aligned} \varepsilon_N(t) &\triangleq \mathbf{y}(t) - \hat{\mathbf{y}}(t) \\ &= \mathbf{y}(t) - \Phi_N(t)\,\mathbf{x}_N(t) \\ &= \mathbf{y}(t) - [\mathbf{W}_0(t)|\mathbf{W}_1(t)|\cdots|\mathbf{W}_{N-1}(t)]\,\mathbf{x}_N(t) \end{aligned} \tag{2}$$

where $\mathbf{W}_i(t)$ denotes the weight matrix for the delayed input vector by i samples, $\mathbf{x}_N(t)=[\mathbf{x}^T(t) \cdots \mathbf{x}^T(t-N+1)]^T$ is the input regressor, and the superscript T denotes the transpose of a vector. The estimate $\hat{\mathbf{y}}(t)$ is the weighted linear combination of the past input signals and in this case it reflects all correlations between channels.

Our proposed distance is defined in terms of the estimation errors, i.e., at time t the distance between $\mathbf{x}(t)$ and $\mathbf{y}(t)$ is defined by

$$d(\mathbf{x}(t), \mathbf{y}(t)) \triangleq (\mathbf{y}(t) - \hat{\mathbf{y}}(t))^T \ (\mathbf{y}(t) - \hat{\mathbf{y}}(t)) \tag{3}$$

However, using (3) and (2) as a distance requires computational burden and resulted in poor recognition performance in preliminary test in this study, and therefore was not pursued further. In order to circumvent this disadvantage, we impose two constraints on the filter structure.

i) constrained multichannel FIR filter

If we neglect the channel correlations, then $\mathbf{W}_i(t)$ in (2) becomes a diagonal matrix and $\hat{\mathbf{y}}(t)$

$=[\hat{y}_0(t)\cdots\hat{y}_{P-1}(t)]^T$ is thus expressed as

$$\hat{y}_k(t) \triangleq \sum_{i=0}^{N-1} \mathbf{W}_{i,k}(t)\, x_k(t-i),\ k=0,\ 1,\ \cdots,\ P-1 \quad (4)$$

where $\mathbf{W}_{i,k}(t)$ is the k-th diagonal term of $\mathbf{W}_i(t)$, and $x_k(t)$ is the k-th element of the input feature vector $\mathbf{x}(t)$. Here, P elements of the vectors are fed into each channel and processed independently.

ii) single channel FIR filter

In this structure, all P channels have the same weight vector and $\hat{\mathbf{y}}(t)$ is given by

$$\begin{aligned}\hat{y}(t) &\triangleq X_N(t)\ w(t)\\ &=[x(t)\ x(t-1)\cdots x(t-N+1)][w_0(t)\ w_1(t)\cdots\\ &\quad w_{N-1}(t)]^T \end{aligned} \quad (5)$$

where $X_N(t)$ is the data matrix.

There are two well-known methods of solving the minimum-mean-square error problem to obtain the weight vector. One solution is the recursive least squares(RLS) algorithm where the solution can be computed recursively in time. The other solution is the least mean squares(LMS) algorithm which attempts to minimize the mean square of the error signal and involves an instantaneous estimate of gradient. Since the stationarity of speech signal is weak, we obtain the weight vector adaptively using the RLS algorithm which shows the fastest convergence and the smallest steady-state error[8].

Although computationally efficient RLS adaptive algorithms have been proposed, the computational burden is still high to apply the RLS algorithms to the proposed distance measure. Fortunately, the preliminary noisy speech recognition test(Table 1) showed that when the FIR filter order was one, the recognition performance was as good as when the filter order was larger than one. This can be understood if one considers Mansour's observation[2][3] that the additive white noise causes the cepstral vector norm reduced and this norm reduction effect can be modeled and reduced by a first-order equalization. This is an encouraging result because it means we can easily implement the RLS algorithm. If the filter order N is equal to one, our proposed distance measure for the constrained multichannel FIR filter structure can be easily derived as follows.

$$\begin{aligned}r_k(t) &\triangleq \sum_{i=0}^{t} \lambda^{t-i} x_k^2(i)\\ &=\lambda r_k(t-1) + x_k^2(t)\end{aligned} \quad (6)$$

$$\begin{aligned}p_k(t) &\triangleq \sum_{i=0}^{t} \lambda^{t-i} x_k(i)\, y_k(i)\\ &=\lambda p_k(t-1) + x_k(t)\, y_k(t)\end{aligned} \quad (7)$$

$$W_{0,k}(t) = r_k^{-1}(t)\, p_k(t) \quad (8)$$

$$\hat{y}_k(t) \triangleq W_{0,k}(t)\, x_k(t) \quad (9)$$

$$d_{multi}(\mathbf{x}(t),\mathbf{y}(t)) \triangleq \sum_{k=0}^{P-1} (y_k(t)-\hat{y}_k(t))^2 \quad (10)$$

where $r_k(t)$ is the autocorrelation of the k-th channel at t-th frame of $\mathbf{X}$, $p_k(t)$ is the cross-correlation between $\mathbf{X}$ and $\mathbf{Y}$, and the forgetting factor λ is employed in order to track slowly time-varying conditions. Note that $r_k(t)$ and $p_k(t)$ are only scalars in this case and they are easily obtained by the recursive relationships. Furthermore, from (6) one can see that $r_k(t)$ is always positive and thus $W_{0,k}(t)$ is always well-defined.

For the single channel FIR filter structure, if $N=1$, the proposed distance is similarly derived as follows.

$$\begin{aligned}\mathrm{R}(t) &\triangleq \sum_{i=0}^{t} \lambda^{t-i} \mathbf{x}^{\mathrm{T}}(i)\mathbf{x}(i)\\ &=\lambda \mathrm{R}(t-1) + \mathbf{x}^{\mathrm{T}}(t)\mathbf{x}(t)\end{aligned} \quad (11)$$

$$\begin{aligned}\mathrm{P}(t) &\triangleq \sum_{i=0}^{t} \lambda^{t-i} \mathbf{x}^{\mathrm{T}}(i)\mathbf{y}(i)\\ &=\lambda \mathrm{P}(t-1) + \mathbf{x}^{\mathrm{T}}(t)\mathbf{y}(t)\end{aligned} \quad (12)$$

$$w_0(t)=\mathrm{R}^{-1}(t)\,\mathrm{P}(t)$$

$$=\frac{\lambda\,\mathrm{P}(t-1)+\mathbf{x}^{\mathrm{T}}(t)\mathbf{y}(t)}{\lambda\,\mathrm{R}(t-1)+\mathbf{x}^{\mathrm{T}}(t)\mathbf{x}(t)} \tag{13}$$

$$\hat{\mathrm{y}}(t)\ \triangleq\ w_0(t)\mathrm{x}(t) \tag{14}$$

$$d_{single}(\mathbf{x}(t),\mathbf{y}(t))\ \triangleq\ (\mathbf{y}(t)-\hat{\mathbf{y}}(t))^{\mathrm{T}}(\mathbf{y}(t)-\hat{\mathbf{y}}(t)). \tag{15}$$

The autocorrelation and the crosscorrelation in the crosscorrelation in this case can also be easily obtained by the recursive realtionships. As mentioned earlier, Mansour and Junang proposed a family of distance measures based on the projection operation. They observed that the cepstral vectors have the norm reduction effect when they were noise-contaminated, and obtained the optimal weighting factor by the orthogonality principle. The projection distance is given by

$$d_{proj}(\mathbf{x}(t),\mathbf{y}(t))\ \triangleq\ (\mathbf{y}(t)-\hat{\mathbf{y}}(t))^{\mathrm{T}}(\mathbf{y}(t)-\hat{\mathbf{y}}(t))$$
$$=(\mathbf{y}(t)-\omega(t)\mathbf{x}(t))^{\mathrm{T}}(\mathbf{y}(t)-\omega(t)\mathbf{x}(t)) \tag{16}$$

where the weighting factor $\omega(t)$ is

$$\omega(t)=\frac{\mathbf{x}^{\mathrm{T}}(t)\mathbf{y}(t)}{\mathbf{x}^{\mathrm{T}}(t)\mathbf{x}(t)} \tag{17}$$

Comparing (17) with (13), it can be seen that if the forgetting factor λ is zero, then (17) is identical to (13). Thus, the projection distance measure is considered as a special case of our proposed distance measure when $\lambda=0$. While the projection distance measure deals with only current frame, our proposed distance measure deals with all of the past frames as well as the current frame.

Ⅲ. Simulation Results

The effectiveness of our proposed distance measures was conducted through speaker-independent isolated word recognition experiments at different global SNR levels. The noisy speech was simulated by adding zero-mean white Gaussian noise to the clean test speech. The analog speech signals are first lowpass filtered with a cutoff frequency of 4.5 KHz, and then digitized at a sampling rate of 10 KHz. The digitized speech is then manually end-pointed and analyzed with a 12-th order LPC analyzer. The LPC vectors are converted into corresponding cepstral vectors. A 30 msec Hamming window is used in computing the first 13 autocorrelation coefficients every 10msec. In the recognition phase, test templates are compared to reference templates through a standard dynamic time warping(DTW) procedure. The database is 72 Korean phonetically balanced isolated words. The reference templates used in our recognition experiments were generated from 4 different speakers. And, the test database was obtained from 3 different speakers who did not participate in the training phase.

In order to determine the optimum FIR filter order, a series of experiments with the filter order from 1 to 5 was conducted. The single channel filter structure was used in the experiments, i.e., (5) and (3) were used as the distance measure. Table 1 shows the test results at various SNR levels. As can be seen in Table 1, when the filter order is one, the recognition performance is as good as when the filter order is larger than one. This result allows to make computationally efficient algorithms, i.e., the first-order realization, (10) and (15). We use these first-order distance measures in the remaining experiments.

Table 1. Recognition performance in percentage as a function of FIR filter order when a single channel filter structure is employed.

filter order	Singnal-to-Noise Ratio(dB)					
	0	10	20	30	40	∞
1	36.11	81.94	91.67	95.83	93.06	93.06
2	37.50	80.56	91.67	97.22	93.06	93.06
3	37.50	79.17	91.67	97.22	93.06	93.06
4	34.72	80.56	90.28	94.44	93.06	91.67
5	36.11	75.00	88.89	91.67	93.06	93.06

Table 2 shows the recognition performance as a function of the forgetting factor λ at various SNR levels. The adaptive filter has the single channel first-order structure and thus the distance used is the same as (15). The advantage at large λ is consistent with almost all SNR levels, confirming our proposed distance measure is superior to the projection distance measure because the latter is a special case of the former when λ is equal to zero.

Table 2. Recognition performance in percentage as a function of forgetting factor λ when a single channel first-order filter structure is employed.

	Singnal-to-Noise Ratio(dB)					
λ	0	10	20	30	40	∞
0.0	36.11	70.83	84.72	86.11	88.89	86.11
0.2	37.50	69.44	84.72	86.11	88.89	86.11
0.4	40.28	70.83	83.33	86.11	87.50	86.11
0.6	40.28	70.83	83.33	84.72	87.50	86.11
0.8	37.50	75.00	87.50	88.89	93.06	90.28
1.0	33.33	81.94	91.67	95.83	95.83	94.44

We compared the performance of the proposed distance measure with other distance measures and techniques to gain some perspective of the merit of the new measure. The distance measures used for the recognition test are

i) the proposed constrained multichannel first-order distance measure defined in (15)

ii) the proposed single channel first-order distance measure defined in (10)

iii) the conventional Euclidean distance measure

iv) the projection distance measure defined in (16)

v) the modified projection distance measure [2][13] defined by

$$d_{mproj}(\mathbf{x}(t),\mathbf{y}(t)) \triangleq \|\mathbf{x}(t)\|(1-\cos\beta)$$

$$\cos\beta = \frac{\mathbf{x}^{\mathrm{T}}(t)\,\mathbf{y}(t)}{\|\mathbf{x}(t)\|\|\mathbf{y}(t)\|} \tag{18}$$

where $\|\mathbf{x}\|$ denotes the norm of a vector $\mathbf{x}$.

We repeated the same experiments with other distance measures to assess how the proposed distance measures perform in comparison to them. The results of these experiments are given in Table 3 and Fig. 2. The single channel distance measure performs better than the multichannel distance measure. This result may be conjectured that when each channel is processed independently, which occurs in the multichannel case, not only the noise effect is reduced by the adaptive filtering, but also the phonetic difference can be masked. The modified projection distance measure works better than the projection distance measure. This result coincides with that of Mansour and Juang. The recognition experiments in this paper show that the projection distance measures work well at low SNR conditions, but it does not work as well as the Euclidean distance measure at high SNR conditions. Our proposed single channel distance measure performs well at all SNR conditions, although it performs slightly worse than the Euclidean distance measure for clean speech data.

Table 3. Comparison of recognition performance in percentage
(a) proposed multichannel distance measure, Eq.(11)
(b) proposed single channel distance measure, Eq. (17)
(c) Euclidean distance measure
(d) projection distance measure, Eq. (18)
(e) modified projection distance measure, Eq. (20).

SNR(dB)	(a)	(b)	(c)	(d)	(e)
0	26.85	39.81	18.98	39.81	37.04
10	74.54	81.02	57.41	77.31	81.02
20	87.50	93.06	86.11	84.26	90.28
30	90.74	94.91	93.52	86.57	90.28
40	91.20	93.06	95.37	86.57	91.67
∞	91.67	93.98	94.91	85.65	90.74

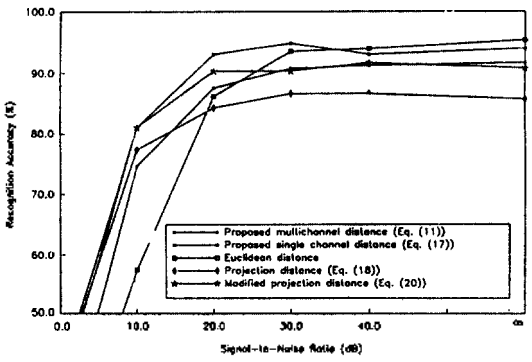

Fig. 2 Speaker-independent isolated word recognition results.

Ⅳ. Conclusion

In this paper, we have proposed distance measures which are robust to background noise. We assume that the effect of noise can be modeled as an FIR system in the feature domain. Hence the feature vectors of the degraded speech can be considered as the outputs of an FIR system which represents the effect of noise, where the input signal is the feature vectors of corresponding clean speech. The unknown FIR system parameters are identified using adaptive filtering techniques, in which the desired signal is obtained from the degraded test speech and the input signal is from the clean reference template. The RLS algorithm is employed as the adaptive algorithm. The adaptive filter is adjustable by minimizing some measure of error between the desired signal and the filter output. We define distance measures in terms of the estimation errors. The single channel first-order FIR structure shows the best recognition performance and thus allows to make a computationally efficient algorithm. The speaker-independent isolated word recognition tests at various SNR levels show that our proposed distance measure performs better than the Euclidean distance measure and the projection distance measures.

References

[1] J.S.Lim, *Speech enhancement*, Prentice-Hall, New Jersey, 1983.

[2] D.Mansour and B.H.Juang, "A family of distortion measures based upon projection operation for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, PP 36-39, 1982.

[3] D.Mansour and B.H.Juang, "A family of distortion measures based upon projection operation for robust speech recognition," in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, Nov. 1989.

[4] F.K.Soong and M.M.Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, Jan. 1988.

[5] M.J.Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 262-265, 1989.

[6] D.Mansour and B.H.Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans, Acoust., Speech, Signal Processing*, vol.37, June 1989.

[7] S.M.Kay, "The effets of noise on the autoregressive spectral estimate," *IEEE Trans, Acoust., Speech, Signal Processing*, pp. 478-485, Oct. 1979.

[8] S.T.Alexander, *Adoaptive Signal Processing: Theory and Applications*, Springer-Verlag, New York, 1986.

▲W.G. Chung

1986. 2. : Department of Electronics Engineering, Seoul National University(B.S.)

1988. 2. : Department of Electrical Engineering, Korea Advanced Institute of Science and Technology(M.S)

1988. 3. : Department of Electrical Engineering, Korea Advanced Institute of Science and Technology(Ph.D. Course)

1988. 3. : Research Engineer at the DigiCom co.

▲Dr. Chong Kwan Un received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Delaware, Newyork, Delaware, in 1964, 1968 and 1969, respectively. From 1969 to 1973 he was Assistant Professor of Electrical Engineering at the University of Maine, Portland, where he taugth communications and did research on synchronization problems. In May 1973, he joined the Staff of the Telecommunication Sciences Center, SRI International, Menlo Park, CA, where he did research on voice digitization and bandwidth compression systems. Since June 1977 he has been with KAIST, where he is Professor of Electrical Engineering and Head of the Communications Research Laboratory, teaching and doing research in the areas of digital communications and digital signal processing. He has authored or coauthored over 200 papers and 70 reports in speech coding and processing, packet voice/data transmission, synchronization, and digital filtering. Also, he holds 7 patents granted or pending. From February 1982 to June 1983 he served as Dean of Engineering at KAIST.

Dr. Un has received a number of awards including the 1976 Best Paper Award from the IEEE Communications Society, the National Order of Merits (Dong Baik Jang) from the Government of Korea, and Achievement Awards from KITE, KICS and ASK. He is a Fellow of IEEE and a member of Tau Beta Pi and Eta Kappa Nu Honor Societies.