Comparative Study of Several Distortion Measures for Speech Recognition

N. Nocerino
F. K. Soong
L. R. Rabiner
D. H. Klatt

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT. In this study we compared several different spectral distortion measures including the Itakura-Saito (IS), the log likelihood ratio (LLR), the likelihood ratio (LR), the cepstral (CEP), and two perceptually based distortion measures, the weighted likelihood ratio (WLR) and the weighted slope metric (WSM) distortion measures, in terms of their effects on the performance of a standard dynamic time warping (DTW) based, isolated word, speech recognizer. Two modifications of the basic forms of each measure were also investigated, namely a Bark-scale frequency warping and the incorporation of suprasegmental energy information. All distortion measures and their modifications were tested on an alpha-digit vocabulary, 4-talker, telephone recording data base. The results can be summarized as: (1) All LPC-based distortion measures performed reasonably well. The LLR and WSM distortion measures gave the highest recognition accuracy, while the IS distortion measure gave the lowest score; (2) Whereas the addition of suprasegmental energy information helped the recognition performance, the use of gain and absolute loudness degraded the performance; (3) Bark-scale frequency warping did not perform as well as its unwarped counterpart; (4) The WLR distortion measure did not perform as well as its unweighted counterpart.

I. Introduction

Since it was first introduced, the Itakura-Saito distortion measure [1] has played a key role in speech coding, analysis, synthesis and recognition. Several studies were conducted to investigate the relationship between different LPC-based distortion measures and to study their properties from a theoretical point of view [2,3].

It is the goal of this research to compare several basic distortion measures (including two recently proposed, perceptually based measures [4,5]) and to study their influence on the performance of an isolated word, DTW based, speech recognition system. We also tested two modifications of the basic distortion measures: Bark-scale frequency warping of the LPC-derived distortion measure, and incorporation of suprasegmental energy information.

II. Spectral Distortion Measures

2.1 Itakura-Saito Distortion Measure

The maximum likelihood distortion measure, also known as the Itakura-Saito distortion measure, was first used for short-time spectral estimation of speech signals. The measure, denoted as d_{IS}, is:

$$d_{\mathrm{IS}}(S_{in},f) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S_{in}(\lambda)}{f(\lambda)} + \ln\frac{f(\lambda)}{S_{in}(\lambda)} - 1\right]\frac{d\lambda}{2\pi} \qquad (1)$$

where $S_{in}(\lambda)$ is the short-time spectral density (or periodogram) of an input speech signal, and $f(\lambda)$

$$f(\lambda) = \frac{\sigma^2}{|1 + a_1 e^{-j\lambda} + \cdots + a_p e^{-jp\lambda}|^2} = \frac{\sigma^2}{|A|^2} \qquad (2)$$

is the spectral density function of a corresponding pth-order all-pole model. Defining d as the log spectral distance between $S_{in}(\lambda)$ and $f(T)$, at frequency λ, i.e.

$$d = \ln\frac{S_{in}(\lambda)}{f(\lambda)} \qquad (3)$$

we can rewrite the distortion measure of Eq. (1) as

$$d_{\mathrm{IS}}(S_{in},f) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[e^d - d - 1\right]\frac{d\lambda}{2\pi} \qquad (4)$$

It is clear that the Itakura-Saito distortion measure is asymmetrical with respect to the positive and negative values of d. The asymmetrical nature of the Itakura-Saito measure makes it subjectively meaningful since the perceptually important formant (spectral peaks) information is emphasized.

Given two all-pole model spectra, $f(\lambda)$ and $f'(\lambda)$, the corresponding Itakura-Saito distortion measure between $f(\lambda)$ and $f'(\lambda)$ is:

$$d_{\mathrm{IS}}(f(\lambda),f'(\lambda)) = \frac{\sigma^2}{\sigma'^2}\,\frac{\mathbf{a}'^T\mathbf{R}\mathbf{a}'}{\mathbf{a}^T\mathbf{R}\mathbf{a}} + \ln\frac{\sigma'^2}{\sigma^2} - 1 \qquad (5)$$

where

$$\mathbf{a}^T = [1,a_1,a_2,\ldots,a_p]$$

$$\mathbf{a}^T = [1,a_1',a_2',\ldots,a_p'] \qquad (6)$$

are the vector representations of the impulse responses of the two all-zero filters and $\mathbf{R}$ is the $(p+1) \times (p+1)$ input sample autocorrelation symmetric Toeplitz matrix whose first row consists of $(p+1)$ autocorrelation values of the input from zero to p time lags, i.e. $[r(0),r(1),\ldots,r(p)]$.

2.2 The Log Likelihood Ratio (LLR) and the Likelihood Ratio (LR) Distortion Measures

An alternative choice of gain for the IS measure, is derived by choosing a value β to minimize the IS distance between $f(\lambda)$ and $\beta f'(\lambda)$, i.e. $\min_{\beta} d_{\mathrm{IS}}(f(\lambda),\beta f'(\lambda))$ and the resultant distortion measure is

$$d_{\mathrm{LLR}} = \min_{\beta} d_{\mathrm{IS}}(f,\beta f') = \ln\left[\frac{\mathbf{a}'^T\mathbf{R}\mathbf{a}'}{\alpha}\right] \qquad (7)$$

This log likelihood ratio measure, d_{LLR}, was proposed by Itakura [6] for speech recognition; hence, it is also commonly referred to as the Itakura distortion measure.

Another alternative is to set the gain terms, σ and σ', so that the test and reference patterns are compared with each other solely on the basis of their spectral shapes i.e. set $\sigma = \sigma'$. The resulting distortion measure is called the likelihood ratio distortion measure and is represented as

$$d_{\mathrm{LR}}(f,f') = \frac{\mathbf{a}'^T\mathbf{R}\mathbf{a}'}{\alpha} - 1 \qquad (8)$$

2.3 LPC Cepstral Distortion Measure

The L_2 norm of the log spectral distortion measure between two time series, $x(n)$ and $x'(n)$, can be approximated by an N term cepstral distortion measure as

CH2118-8/85/0000-0025 $1.00 © 1985 IEEE

$$d^N_{\mathrm{CEP}}(x,x') = \sum_{\ell=-N}^{N} (c_\ell - c'_\ell)^2 \qquad (9)$$

2.4 Weighted Likelihood Ratio (WLR) Distortion Measure and Weighted Slope Metric (WSM) Distortion Measure

The weighted likelihood ratio distortion measure, d_{WLR}, has the form [4]

$$d_{\mathrm{WLR}}(f,f') = \frac{1}{2}\int_{-\pi}^{\pi} \left[\frac{f}{r_0} - \frac{f'}{r'_0}\right] (\log(f) - \log(f'))\, \frac{d\lambda}{2\pi} \qquad (10)$$

The integration can be approximated efficiently in the time domain by a truncated version

$$d^N_{\mathrm{WLR}}(f,f') = \sum_{i=1}^{N} \left[\frac{r(i)}{r(0)} - \frac{r'(i)}{r'(0)}\right] (c(i) - c'(i)) \qquad (11)$$

The weighted slope metric (WSM) has the form [5]

$$d_{\mathrm{WSM}}(f,f') = k_E|E_f - E_{f'}| + \sum_{i=1}^{Q} k_s(i)(s_f(i) - s_{f'}(i))^2 \qquad (12)$$

where Q is the number of critical frequency bands, k_E is a weighting coefficient on the absolute energy difference, $|E_f - E_{f'}|$, between f and f', $k_s(i)$ is a weighting coefficient for the difference, $s_f(i) - s_{f'}(i)$, between the two critical band spectral slopes of f and f'. The slope weighting function we used in our experiments was

$$k_s(i) = \frac{1}{2}\,[k^f_s(i) + k^{f'}_s(i)] \qquad (13)$$

where

$$k^f_s(i) = \left[\frac{k_{\mathrm{LMAX}}}{k_{\mathrm{LMAX}} + \Delta_{\mathrm{LMAX}}(i)}\right] \cdot \left[\frac{k_{\mathrm{GMAX}}}{k_{\mathrm{GMAX}} + \Delta_{\mathrm{GMAX}}(i)}\right], \qquad (14)$$

and $\Delta_{\mathrm{LMAX}}(i)$ and $\Delta_{\mathrm{GMAX}}(i)$ are the log spectral differences (in decibels) between the spectral values at the ith critical band and the nearest local maximum (LMAX) spectral peak and the global maximum (GMAX) spectral peak, respectively. The coefficients, k_{LMAX} and k_{GMAX}, are used here not only to prevent $k^f_s(i)$ and $k^{f'}_s(i)$ from becoming singular but also to distribute the weighting proportionately between the local and the global spectral behavior. To illustrate the weighting used in the weighted slope metric a typical critical frequency band vowel spectrum and its corresponding $\Delta_{\mathrm{GMAX}}(i)$ and $\Delta_{\mathrm{LMAX}}(i)$ are depicted in Fig. 1. The slope weighting coefficient, $k^f_s(i)$, given by Eq. (14) has a larger value at the spectral peaks, especially the global peak, than at the spectral valleys. As a consequence the spectral slope difference, $s_f(i) - s_{f'}(i)$, is emphasized more at the spectral peak locations.

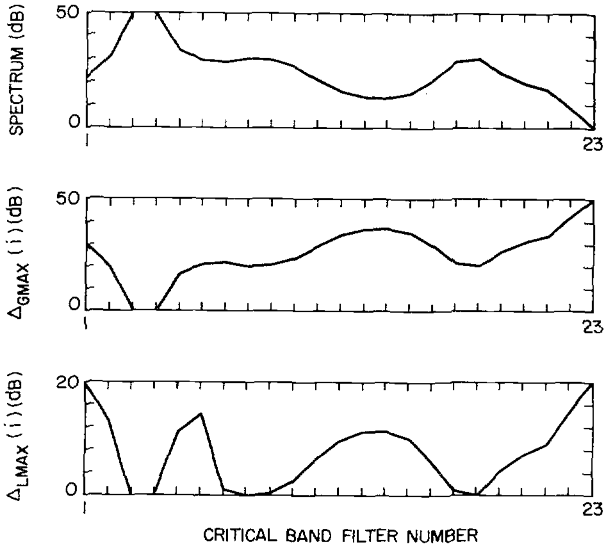

Fig. 1 Plots explaining the weighted slope metric functions.

III. Modifications of the LPC Distortion Measures — Bark-Scale Frequency Warping and Additional Temporal Energy Information

3.1 Bark-Scale Frequency Warping

The results of a wide variety of psychophysical experiments have indicated that a Bark or Mel scale spacing along the frequency axis is subjectively more meaningful than a linear frequency scale. In order to exploit this perceptual information, here we propose an efficient procedure to warp the frequency scale of the distortion measure from a linear (Hz) scale to a critical band (Bark) scale. Using the cepstral distortion measure, as an example, we compute the warped cepstral distortion measure, d_{WCEP}, as

$$d_{\mathrm{WCEP}}(f,f') = \frac{1}{2\bar{B}} \int_{-\bar{B}}^{\bar{B}} [\log f(\lambda(B)) - \log f'(\lambda(B))]^2\, dB \qquad (15)$$

where B is frequency in Barks and $\log(f(\lambda(B)))$ is the LPC inverse spectrum of f on a Bark-scale. The value $\bar{B}$ in Eq. (15) is the Nyquist frequency (i.e., one half of the sampling frequency $0.5 f_s$) in Barks. We can rewrite Eq. (15), by using the cepstral representation, as

$$d_{\mathrm{WCEP}}(f,f') = \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} (c_\ell - c'_\ell)(c_m - c'_m)\, w_{\ell,m} \qquad (16)$$

where we define the warping function $w_{\ell,m}$ as

$$w_{\ell,m} = \frac{1}{2\bar{B}} \int_{-\bar{B}}^{\bar{B}} e^{j\lambda(B)(\ell - m)}\, dB \qquad (17)$$

The warping function $w_{\ell,m}$ is a symmetric Toeplitz form. Thus Eq. (16) can be approximated by a finite sum as

$$d^N_{\mathrm{WCEP}} = \sum_{\ell=-N}^{N} \sum_{m=-N}^{N} (c_\ell - c'_\ell)(c_m - c'_m)\, w_{|\ell - m|} \qquad (18)$$

The same warping technique can be applied to other LPC-based spectral distortion measures.

3.2 Addition of Temporal Energy Information

Temporal energy information has been shown to be a useful parameter for improving the performance of an isolated word recognizer. In this study we used an energy distance defined as:

$$d_E(f,f') = k \cdot g(|\log \bar{r}_0 - \log \bar{r}'_0|) \qquad (19)$$

where $\bar{r}_0$ and $\bar{r}'_0$ are the relative energies (in decibels) with respect to the maximum energies, $E_{\max}$ and $E'_{\max}$, of the test and reference utterances. In this manner, the global absolute loudness effects are removed from the energy distance computation. The nonlinear function, $g(\cdot)$ in Eq. (19), is defined as

$$g(E) = \begin{cases} 0 & |E| \leqslant E_{LO} \\ |E| - E_{LO} + E_{OF} & E_{LO} < |E| \leqslant E_{HI} + E_{LO} - E_{OF} \\ E_{HI} & E_{HI} + E_{LO} - E_{OF} < |E| \end{cases} \qquad (20)$$

A plot of $g(E)$ versus E is given in Fig. 2. The constant k in Eq. (20) is a weighting coefficient that is used to combine the energy distortion with the spectral distortion. We have found experimentally that $k = 0.1$ is a good value for the energy distortion weighting.

Tables I and II summarize the distortion measures that will be used in our experimental investigations. Table I reviews the mathematical formulation of the six basic distortion measures; Table II gives

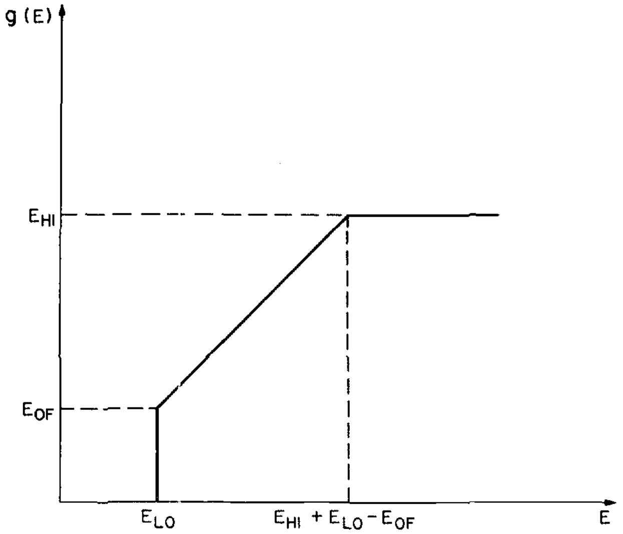

Fig. 2 The nonlinear function, g, applied to the log energy.

Distortion Measure	Notation	Expression
Itakura-Saito	d_{IS}	$\frac{\sigma^2}{\sigma'^2}\,\frac{\mathbf{a}'^T\mathbf{R}\mathbf{a}'}{\mathbf{a}^T\mathbf{R}\mathbf{a}} + \ln\frac{\sigma'^2}{\sigma^2} - 1$
Log Likelihood Ratio (Itakura)	d_{LLR}	$\ln\frac{\mathbf{a}'^T\mathbf{R}\mathbf{a}'}{\mathbf{a}^T\mathbf{R}\mathbf{a}}$
Likelihood Ratio	d_{LR}	$\frac{\mathbf{a}'^T\mathbf{R}\mathbf{a}'}{\mathbf{a}^T\mathbf{R}\mathbf{a}} - 1$
Weighted Likelihood Ratio	d^N_{WLR}	$\sum_{i=1}^{N}\left(\frac{r_i}{r_0} - \frac{r_i'}{r_0'}\right)(c_i - c_i')$
Weighted Slope Metric	d_{WSM}	$k_E\lvert E_f - E_f'\rvert + \sum_{i=1}^{Q} k_s(i)\lvert s_f(i) - s_{f'}(i)\rvert^2$
Cepstrum	d^N_{CEP}	$\sum_{i=1}^{N}(c_i - c_i')^2$

Table I. The six basic distortion measures

expressions for the modifications to the distortion measures for frequency warping and for the addition of temporal energy.

IV. Experimental Results

4.1 Database, LPC Spectral Analysis and DTW

We chose a four-talker (2 male and 2 female) 39-word, alpha-digit telephone recording database for our benchmark comparisons. The database was recorded over normal dialed-up telephone lines in a sound booth and the analog input was sampled at 6.67 kHz. For each talker, there were 5 to 7 utterances of the same word used for training and 10 utterances used for testing. The training and test tokens were recorded several weeks apart.

An 8^{th}-order LPC analysis was performed on each frame of speech data. A 45 ms (300 samples) Hamming window was used to compute the autocorrelation coefficients of preemphasized speech samples. The 1st order preemphasis filter has the transfer function $H(z) = 1 - 0.95z^{-1}$. The overlapping period between adjacent analysis frames was 30 ms. All speech recognition experiments were performed on a speaker-trained basis.

Distortion Measure	Notation	Expression
Bark-scale Warped Cepstrum	d^N_{WCEP}	$\sum_{m=-N}^{N}\sum_{\ell=-N}^{N}(c_\ell - c_\ell')w_{\lvert\ell-m\rvert}(c_m - c_m')$
Likelihood Ratio with Temporal Energy	d_{LRE}	$d_{\mathrm{LR}} + kg(\lvert\log\bar{r}_0 - \log\bar{r}_0'\rvert)$

Table II. Two examples of the modified distortion measures.

An asymmetric end-point constrained dynamic time warping (DTW) procedure was used with the local slope constraints of 1/2 and 2 on the warping path. There were 5 to 7 training tokens per word for each talker. The reference pattern of each word was chosen as the token in the training set which had the smallest DTW average distortion to all other training tokens.

4.2 Global Performance Comparison

Table III gives recognition error percentages of the six basic LPC-derived distortion measures, d_{IS}, d_{LLR}, d_{LR} d_{CEP}, d_{WLR}, and d_{WSM}. As it is shown in Table III, almost all of the distortion measures performed reasonably well. The average recognition error percentages for the best 5 distortion measures were within 0.7% however the Itakura-Saito distortion measure gave an error rate that was 2.9% worse than the log likelihood ratio error rate. It can also be seen in Table III that the weighted likelihood ratio distortion measure, d_{WLR}, did not perform as well as its unweighted counterparts, d_{LR} and d_{LLR}, although the differences in performance were small.

4.3 Effects of Varying Weighted Slope Metric Parameters

As seen in Table III, the weighted slope metric distortion measure, d_{WSM}, performed as well as the best LPC distortion measure, d_{LLR}. This result is in contrast to previous speech recognition experiments with the weighted slope distortion measure [7]. However it should be noted that the d_{WSM} used here is a degenerate form since the weighting coefficients, k_E, k_{LMAX} and k_{GMAX} provided virtually no weighting (i.e., $k_E = 0$ and $k_{\mathrm{LMAX}} = k_{\mathrm{GMAX}} = \infty$). The weighted slope metric distortion measure was therefore an unweighted slope metric. We experimented with several different values of k_{LMAX} and k_{GMAX} and the results of these experiments are summarized in Table IV. It can be seen in Table IV that the recognition error percentage decreases monotonically with increasing values of k_{LMAX} and k_{GMAX}.

Distortion Measure	Error Percentage %					Comments
	Tlkr. 1	Tlkr. 2	Tlkr. 3	Tlkr. 4	Ave.	
d_{LLR}	5.6	4.4	10.5	13.3	8.45	
d_{WSM}	5.6	4.4	11.5	12.3	8.45	$k_E = 0$, $k_{\mathrm{LMAX}} = k_{\mathrm{GMAX}} = \infty$
d_{LR}	5.1	5.6	10.5	13.3	8.63	
d_{CEP}	5.1	5.6	9.7	15.1	8.88	$c_0 = c_0 = 0$, $N = 32$
d_{WLR}	8.2	6.9	7.4	14.1	9.15	$N = 16$
d_{IS}	7.7	5.9	11.3	20.5	11.35	

Table III. Error percentages of the six basic distortion measures.

Parameters			Error Percentage %				
k_E	k_{LMAX}	k_{GMAX}	Tlkr. 1	Tlkr. 2	Tlkr. 3	Tlkr. 4	Ave.
0	1	1	9.0	6.4	22.3	23.8	15.4
0	10	10	6.7	4.4	11.0	17.2	9.8
0	50	8	7.2	3.8	11.5	15.1	9.4
0	10	250	5.6	3.6	13.3	12.6	8.8
0	250	250	5.9	4.4	11.5	12.1	8.5
0	∞	∞	5.6	4.4	11.5	12.3	8.5

Table IV. Error percentages of the weighted slope metric measure.

4.4 Effects of Adding Energy or Gain Information

Table V shows the recognition error percentages when gain, absolute energy, and normalized energy were used in the various distortion measures. The results indicate that gain sensitive d_{IS} is worse than normalizing the LPC gain to a constant level, e.g., setting $c_0 = c_0'$ in d_{CEP} or setting $\sigma = \sigma'$ in d_{IS} (i.e., d_{LR}). It is also shown in the table, that the recognition performance of d_{IS} and d_{LR} was improved by 1.27% and 1.98% when the temporal energy information was incorporated.

4.5 Bark-scale Frequency Warping Results

The results of using both the Bark-scale warping for both the cepstral distortion measure and the weighted likelihood distortion measure are given in Table VI. The warped distortion measure performed worse than their unwarped counterparts in both cases.

Distortion Measure				Error Percentage %				
	Gain	Absolute Energy	Normalized Energy	Tlkr. 1	Tlkr. 2	Tlkr. 3	Tlkr. 4	Ave.
d_{CEP}^{16}	yes	no	no	6.2	8.2	10.0	18.5	10.73
	no	no	no	6.4	6.4	10.0	16.2	9.75
d_{IS}	yes	no	no	7.7	5.9	11.3	20.5	11.35
	yes	no	yes	5.6	5.1	11.0	15.6	9.33
d_{LR}	no	no	no	5.1	5.6	10.5	13.3	8.63
	no	no	yes	4.1	4.1	6.7	14.4	7.36
d_{WSM}	no	yes $k_E = 1$	no	6.4	4.6	12.1	12.1	8.80
	no	yes $k_E = 16$	no	6.7	4.9	10.0	14.4	9.00
	no	no $k_E = 0$	no	5.6	4.4	11.5	12.3	8.45

Table V. Effects of energy and normalized temporal energy.

Distortion Measure		Error Percentage %				
	Warping	Tlkr. 1	Tlkr. 2	Tlkr. 3	Tlkr. 4	Ave.
d_{CEP}^{16}	no	6.4	6.4	10.0	16.2	9.75
	yes	7.7	6.2	9.7	17.4	10.25
d_{WLR}^{16}	no	8.2	6.9	7.4	14.1	9.15
	yes	8.2	7.9	10.8	19.0	11.48

Table VI. Recognition error percentages of d_{CEP} and d_{WLR} and their bark-scale warped counterparts.

V. Summary and Discussion

We summarize our results as follows:

1. All of the LPC-derived distortion measures we studied in this study worked reasonably well. Among the six basic forms, the log likelihood ratio and the weighted slope metric distortion measures achieved the highest recognition accuracy, while the Itakura-Saito distortion measure yielded the lowest score.
2. Temporal energy information was useful for improving the recognition accuracy if it was appropriately normalized. On the other hand, use of either the LPC gain and or the absolute loudness level degraded the recognition performance.
3. Contrary to previous results [4], the weighted likelihood ratio did not perform as well as its unweighted counterparts, d_{LR}, and d_{LLR}, in our experiments. However, with some hindsight this result should not be too surprising. The vocabulary we used was a highly confusable one. The most difficult subset of our vocabulary, i.e. the "E" set, {"B", "C", "D", "E", "G", "P", "T", "V", "Z", "3"}, consisted of words which differed only in their initial consonants. Any attempt to emphasize the salient formant (spectral peak) structure of the following vowels in these E-set words, as was done with the weighted likelihood ratio could, and did, lead to degraded performance.
4. The Bark-scale warped distortion measures did not perform as well as their unwarped counterparts. The results were disappointing but in some sense were consistent with previous studies [8]. Also for the highly bandlimited (300 ~ 3000 Hz) telephone data base, expanding the frequency scale at the lower frequency end tended to overemphasize the frequency region where little or no significant speech information existed.
5. The LPC-derived weighted slope metric worked very well. However the best d_{WSM} result was obtained by setting two crucial weighting parameters, k_{LMAX} and k_{GMAX}, to ∞.

REFERENCES

[1] Itakura, F. and Saito, S., "An Analysis-Synthesis Telephony Based on Maximum Likelihood Method," Proc. Int'l Cong. Acoust., C-5-5, 1968.

[2] Gray, A. and Markel, J., "Distance Measures for Speech Processing," IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-24, pp. 380-391, Oct. 1976.

[3] Gray, R., Buzo, A., Gray, A. and Matusyama, Y., "Distortion Measures for Speech Processing," IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-28, No. 4, Aug. 1980.

[4] Shikano, K. and Sugiyama, M., "Evaluation of LPC Spectral Matching Measures for Spoken Word Recognition," Trans. IECE, Vol. J65-D, No. 5, pp. 535-541, May 1982.

[5] Klatt, D. H., "Prediction of Perceived Phonetic Distance from Critical Band Spectra: A First Step," Proc. of ICASSP 1982, Vol. 2, pp. 1278-1281, May 1982.

[6] Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-23, pp. 67-72, Feb. 1975.

[7] Kuo, S. P. F., "A Comparative Study of Spectral Distance Measure Applied to Continuous Speech Recognition," M.S. Thesis, M.I.T., June, 1982.

[8] Blomberg, M., Carlson, R., Elenius, K. and Granstrom, B., "Auditory Nerve in Isolated Word Recognition," Proc. of ICASSP 1984, Vol. 2, pp. 17.9.1~17.9.4.