論文 88-25-4-1

Recognition of Korean Isolated Digits Using a Pole-Zero Model

(Pole-Zero 모델을 이용한 한국어 단독 숫자음 인식)

金 淳 協,* 朴 圭 泰,** Alan C. Bovik***

(Soon Hyob Kim, Kyu Tae Park and Alan Conrad Bovik)

要 約

본 논문에서는 우리말 단독 숫자음의 어두를 유·무성으로 구분한 다음 pole-zero 모델을 이용하여 3개의 포르만트 주파수를 구하고, 이를 인식 파라메타로하는 인식시스템에 대하여 논했다. 맨 처음 시행되는 알고리즘은 대수에너지와 영교차율을 유·무성음을 구분하는 파라메타로 해서 두 개의 영역으로 완전 분리시켰다. 다음단계로 유·무성음으로 나누어진 숫자음을 인식하기 위하여 정규화된 DTW 알고리즘을 3개의 포르만트 주파수를 파라메타로 하여 사용하였다. 우리는 각 프레임에서 3개의 포르만트 주파수를 구하기 위하여서는 pole-zero 모델 방법이 pole 모델 방법보다 더 정확하게 포르만트 주파수를 구하고 있음을 입증하고 97.3%의 인식율을 얻었다.

Abstract

In this paper, we describe an isolated words recognition system for Korean isolated digits based on a voiced-unvoiced decision algorithm and a frequency domain analysis. The algorithm first performs a voiced-unvoiced decision procedure for the beginning part of each uttered word using the normalized log energy and zero crossing rate as decision parameters. Based on this decision, each word is assigned to one of two classes. In order to identify the uttered word within each class, a dynamic time warping algorithm is applied using formant frequencies as the basis for the distance measure. We exploit a pole-zero analysis to measure formant frequencies in each frame . We have observed that pole-zero analysis can provide more accurate estimation of formant frequencies than analysis based on poles only. Experimental Recognition rates of 97.3% illustrating the performance of the recognition system was achieved.

*正會員, 光云大學校 電子計算機工學科
(Dept. of Comp. Eng., Kwangwoon Univ.)
**正會員, 延世大學校 電子工學科
(Dept. of Elec. Eng., Yonsei Univ.)
***非會員, 美國 텍사스·어스틴大學校 電氣 및 電子計算機工學科
(Dept. of Electrical and Comp. Eng., Univ. of Texas at Austin)

接受日字 : 1987年 11月 21日
(※ 이 연구는 한국과학재단의 연구비로 수행되었음.)

I. Introduction

In this paper, we describe an isolated word recognition algorithm for Korean digits, based on a voiced-unvoiced decision algorithm and frequency domain analysis. The algorithm first applies a voiced-unvoiced decision procedure for the initial part of each uttered word using the normalized log energy and zero crossing rate as decision parameters. Based on this decision, each word is assigned to one of two classes. In order to identify the uttered word within each class, a dynamic

warping algorithm is applied using formant frequencies as the basis for the distance measure. We exploit a pole-zero analysis to measure formant frequencies in each frame, as we have found that pole-zero analysis allows for more accurate estimation of formant frequencies than analysis based on poles only. Experimental results illustrating the performance of the recognition system are also given.

We first describe those properties of Korean isolated digits which are cogent to the development of an effective feature-based recognition algorithm. Articulated Korean isolated digits are have the following approximate English pronunciations: zero/yung/; one/il/; two/i/; three/sam/; four/sa/; five/o/; six/yuk/; seven/tsil/; eight/pal/; and nine/kwu/. For the Korean isolated digits zero/yung/to nine/kwu/, there are several important properties which must be considered:

i) the inital sound of some isolated digits is an unvoiced consonant, e.g.,/s/,/ts/,/p/,/k/.

ii) the termination of some isolated digits takes the form of a voiced consonant such as the liquid/l/or nasal sounds/m/,/ng/.

iii) the final consonant/k/of digit "six/yuk/" is a stop consonant similar to silence.

iv) Korean isolated digits are all single-syllable words. The computer recognition of Korean isolated digits is performed by analysis of formant frequency and by exploiting the above properties.

In Section II, a logical combination of the normalized log energy (NLE) and zero-crossing rate (ZCR) is used to provide an initial classification of the unvoice and voice of the inital digit sound. In Section III, we derive the power spectral density (PSD) function via an autoregressive moving average (ARMA) model. The most important factor in the computer recognition of isolated digits is the development of accurate means for extraction of the short term PSD of the spoken digits. Previously, autoregressive (AR) models have been proposed for estimation of the relevant parameters and extraction of the formant frequencies [4]. However, an all-pole, AR or an all-zero, moving average (MA) model can lead to very poor spectral estimates for some consonants. Since a method based on a pole-zero ARMA model is theoretically superior to one based on either an AR or MA process in the spectral analysis of most consonants (especially the nasal, fricative, and stop consonants), we use the Cadzow-Kay [1], [3] closed form ARMA spectral estimation method with optimal order (p=q=8) for analysis of consonants and optimal order (p=q=10) for analysis of vowels [1-3,5,6]. Use of the optimal order provides for very accurate location of the formant frequencies.

Section IV describes a further improvement of the recognition algorithm using the normalized Euclidean distance method as local distance measurement. In Section V and VI the experimental results are presented and discussed and conclusions are offered.

II. Voiced-Unoviced Decision Procedure

The first step in the proposed recognition system involves the assignment of each uttered digit into one of two classes. This classification is based on a voiced-unvoiced (V-U) decision for the initial part of each uttered digit. First, the digitized speech signal is divided into equally-spaced frames of a fixed size. In making the V-U decision, two computationally efficient temporal measurements of the speech signal are employed: the zero crossing rate and the normalized log energy measured over each frame.

In the following, let x_n denote the digitized speech signal. The zero crossing rate Z_k (ZCR) in frame k is then given by

$$Z_x = \sum_{n=N_k}^{N_k+N-1} [1-\mathrm{sgn}(x_n + 1)\ \mathrm{sgn}(x_n)]/2, \tag{1}$$

where N is the size of a frame and where N_k denotes the position of the first sample in the k^{th} frame. The log energy in each frame, LE_k, is then

$$LE_k = 10 \log \left[\sum_{n=N_k}^{N_k+N-1} x_n^2 \right]. \tag{2}$$

It is often convenient to use a normalized energy instead of the log energy. The normalized log energy (NLE), E_k, is computed using

$$E_k = 100\ LE_k/\max_{1 \leqslant j \leqslant \ell} LE_j, \qquad (3)$$

where ℓ is the index of the last frame in a given uttered word. We have found the ZCR and NLE to be useful features for the V-U decision. In general, the ZCR is much higher in an unvoiced sound region than in a voiced region, while the opposite is true for the NLE. For example, Fig. 1 depicts the NLE and ZCR for the digit "two/i/" which is composed of a single vowel, and "seven/tsil/" which includes a voiceless fricative consonant/ts/at the beginning of the utterance. In Fig. 1, E and Z indicate the value of the NLE and ZCR for each frame, respectively. In the graph for "seven/tsil/", the unvoiced sound /ts/ occupies the first four frames.

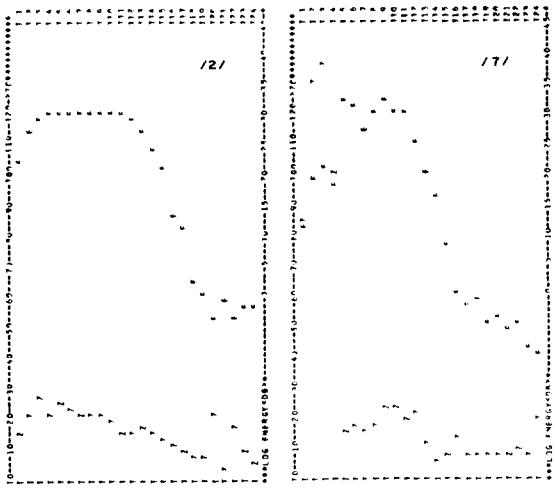


Fig.1. Normalized log energy (NLE) and zero-crossing rate (ZCR) of Korean digits: (a) "two/i/" (b) "seven/tsil/."

In this implementation the V-U decision procedure exploits the ZCR and NLE in the following way. First, we define three parameters: the ZCR difference DZ_k, the NLE difference DE_k, and the difference S_k between DE_k and DZ_k, as follows:

$$DZ_k = Z_{k+1} - Z_k, \qquad (4)$$

$$DE_k = E_{k+1} - E_k, \qquad (5)$$

$$S_k = DE_k - DZ_k, \qquad (6)$$

for k = 1,2, ..., $\ell/3$.

We have observed that when an uttered word begins with an unvoiced sound, S_k is maximized at the boundary between the unvoiced beginning part and the remainder of the word. This is easily understood since, at the boundary, DZ_k is decreasing while DE_k is increasing. This allows the identification of the boundary frame, M, between the two regions. If the given input word begins with a voiced sound, the variation of S_k in the first $\ell/3$ frames is relatively small, while S_M is also small. This observation allows one to use S_M as a feature for the V-U decision at the beginning part of each uttered digit.

Once the boundary frame M is identified, the following three quantities are computed:

$$P = \left(\sum_{k=1}^{M} E_k/Z_k\right)/M, \qquad (7)$$

$$Z_f = \left(\sum_{k=1}^{M} Z_k\right)/M, \qquad (8)$$

$$Z_b = \left(\sum_{k=M+2}^{\ell} Z_k\right)/(\ell-M-1) \qquad (9)$$

In view of the previously mentioned properties of the ZCR and NLE, we may expect P to take relatively large values whenever the first M frames are occupied by a voiced sound; similarly, the ratio of Z_f/Z_b will also be small. Using these two observations and S_M, we may define the following V-U decision rule:

V-U Decision Rule: Given an uttered input digit and constant thresholds, α, β, and γ, if the three conditions

$$\alpha \leqslant P, \quad Z_f/Z_b \leqslant \beta, \quad S_M \leqslant \gamma$$

are simultaneously satisfied, conclude that the digit begins with a voiced sound. Otherwise, conclude that the word begins with an unvoiced sound.

In this experiment, α, β, and γ were predetermined empirically as follows: α = 1.60, β = 2.2, and γ = 38.0.

Once the V-U decision is made for the uttered word, the word is classified into one of two classes (viz., either a class containing zero/yung/, one/il/, two/i/five/o/, six/yuk/, or a class containing three/sam/, four/sa/, seven/tsil/, eight/pal/, nine/kwu/).

III. Estimation of Formant Frequencies via Pole-zero Analysis

It has been noted that short-time speech signals can be compactly represented using a low-order (typically, less than the 10th order) pole-zero (ARMA) model. This model is also closely related to the speech production model [4], [5], [6]. Using this model, at each frame the speech signal x_n is expressed as

$$x_n = -\sum_{i=1}^{p} a_i x_{n-i} + \sum_{j=0}^{q} b_j u_{n-j}, \tag{10}$$

where a_i and b_j are the model coefficients and u_n is the source signal, which is typically modeled as either a pulse train (for a voiced sound) or a pseudo-random signal (for an unvoiced sound). In representing a speech signal using this model, the first problem is to find model coefficients a_i and b_j. In order to find these coefficients, we employed methods proposed by Cadzow [1] and Kay [3]. We briefly review these estimation procedures in Section. 1.

1. Determination of the model coefficients

We first note that the time series elements x_n and u_n are uncorrelated for $n > m$ (viz., $E[u_n x^*_m]=0$ for $n > m$). This fact can be used in determining the coefficient estimate a_k as follows. Multiplying each side of eq. (10) by x^*_{m-n}, we obtain

$$x_m x^*_{m-n} = \sum_{j=0}^{q} b_j u_{m-j} x^*_{m-n} - \sum_{i=1}^{p} a_i x_{m-i} x^*_{m-n}. \tag{11}$$

By taking expectations we obtain the following system of linear equations:

$$\sum_{i=1}^{p} a_i r_x^n(n-i) + r_x^n(n) = e(n) \tag{12}$$

for $q < n < N$, where e(n) denotes the residual error resulting from the correlation between u_n and x^*_n. The autocorrelation estimates for $q < n < N$ and $0 \leqslant i \leqslant p$ are given by:

$$r_x^n(n-i) = \sum_{m=n'+1}^{N} x_{m-i} x^*_{m-n}/(N-n'), \tag{13}$$

where n'= max (n,p). The system of eq. (12) can be expressed in vector-matrix form as

$$\mathbf{e} = \mathbf{Ra} + \mathbf{r}, \tag{14}$$

where **e** is a (N-q-1)x1 matrix and **a** is a px1 matrix. The error estimates **e** are composed of a sum of terms having zero expected value. The px1 vector **a** which minimizes the residual error can be found by solving $d\ (\mathbf{e}^T\mathbf{e})/d\mathbf{a} = 0$, which yields

$$-\mathbf{r} = \mathbf{Ra}. \tag{15}$$

Finally, the estimate **a** is obtained by taking the matrix pseudo-inverse of R. Thus:

$$\hat{\mathbf{a}} = -(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{r}. \tag{16}$$

In order to determine the MA parameters b_j, we then apply a method proposed by Kay using spectral factorization. If x_n is an ARMA (p,q) process as given by eq. (10), the power spectral density $p_x(z)$ of x_n is given by

$$p_x(z) = B(z)B(z^{-1})/[A(z)A(z^{-1})], \tag{17}$$

where

$$B(z) = \sum_{j=0}^{q} b_j z^{-j} \text{ and } A(z) = \sum_{i=0}^{p} a_i z^{-i} \text{ with } a_0 = 1.$$

Denoting the inverse z-transform of $B(z)\,B(z^{-1})$ as the sequence $[c_k]$ (note that $C_{-k}=C_k$), then

$$p_x(e^{j\omega}) = \sum_{k=-q}^{q} [c_k \cos(\omega k)/|A(e^{j\omega})|^2], \tag{18}$$

where c_k is the autocorrelation function of the

residual time series

$$e_t = \sum_{k=0}^{p} a_k x_{t-k}.$$

The estimate of the sequence c_k is then:

$$\hat{c}_k = \begin{cases} w_k \left[\sum_{t=p+1}^{N-k} e_t e_{t+k} + \sum_{t=1}^{N-p-k} d_t d_{+k} \right] / \\ (N-p); \; k = 0,1,\ldots, q \\ C_{-k} \qquad ; k = -q, \ldots,-1, \end{cases} \tag{19}$$

where w_k is the positive-semidefinite lag window

$$w_k = 1 - k/(q+1); \; k=0, 1, \ldots, q \tag{20}$$

and e_t, d_t are forward and backward residual time series.

By substituting eq. (16) and eq. (19) into eq. (18), the power spectral density of x_n is obtained.

2. *Determination of the optimal order of the ARMA model*

In applying the estimation procedure to the analysis of speech signals, it is very important to select the optimal order (p and q), both to obtain the exact formant frequencies and to expedite computation. In the following, we give a number of observations on this problem. To obtain a practical understanding of this problem, we applied the Cadzow-Kay closed form ARMA spectral estimation technique to speech signals of "two/i/" which is a vowel and to the initial part of "seven/tsil/", viz., the unvoiced consonant/ts/. Fig. 2 illustrates the estimated power spectrum using pole-zero analysis and the spectrum obtained via direct FFT of the signals.

In the case of (p=q=8) for /i/, there is a little difference between the first and second formant frequencies obtained by the two method. However, for the consonant/ts/, there is a good match between the results obtained by the two methods. The spectrum obtained by pole-zero analysis exhibits three sharp peaks at correct

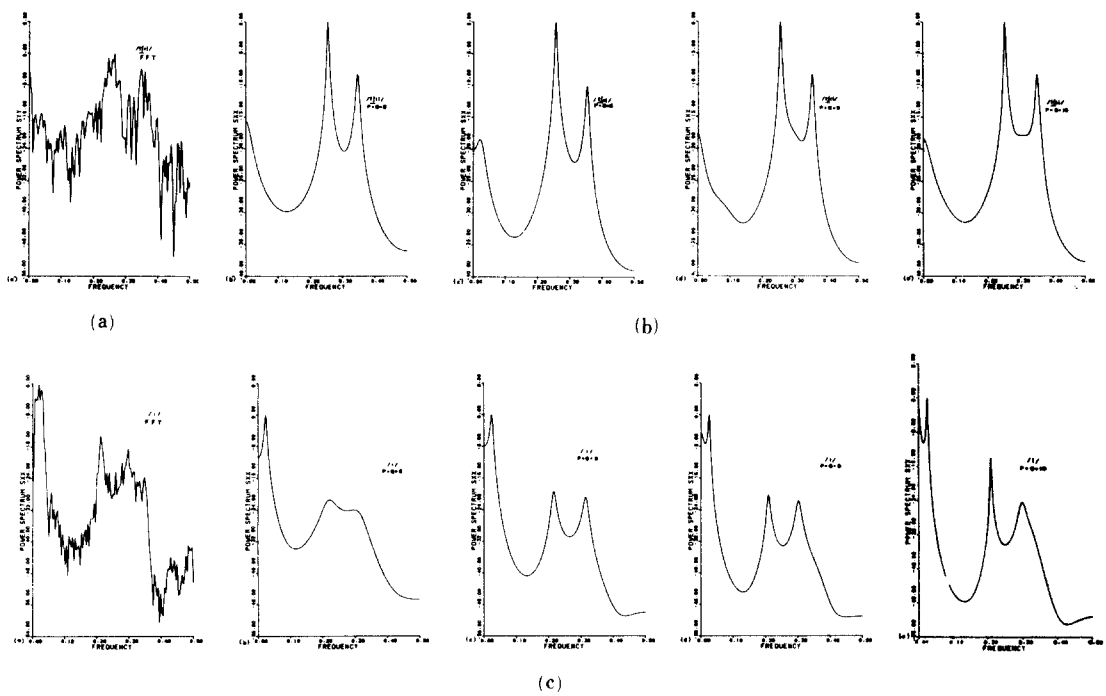


Fig.2. Estimation of the formant frequencies of input signal (FFT shown at the top) for various orders of the ARMA model:
(a) FFT spectrum, (b) consonant/ts/spectrum, (c) vowel/i/spectrum.

locations with almost no spurious peaks.. It should be noted that the order (p=q=8) is needed to achieve accurate estimation of formant frequencies of a consonant.

In the case of (p=q=10), for/i/, there is a close simularity in the peak locations obtained by the two methods. However, for/ts/, there is a small offset in the positions of the poles obtained by the pole-zero analysis. We have observed similar phenomena for other vowels and consonants. Consequently, in order to obtain the correct estimation of three formant frequencies, we choose the order of ARMA model as (p=q=10) for vowels and (p=q=8) for consonants.

Figures 3 and 4 illustrate the results of the analysis for spoken Korean digits. The estimated formant frequency (f1, f2, f3, f4) trajectories from raw speech data by ARMA model (p=q=8) and (p=q=10) are shown in Fig. 3, and the smoothed spectra of the isolated Korean digits using the ARMA model are shown in Fig. 4.

IV. Matching Via Dyamic Time Warping

Once each word is classified into one of two classes, the next problem is to identify the word within the assigned class. In this implementation, the matching is performed using the dynamic time warping (DTW) algorithm. As usual, we will refer to the stored model data of each word as the reference pattern, and to the input data of the given word as the test pattern. We assume throughout that, prior to application of the matching algorithm, the endpoints (beginning and ending frames) of the unknown isolated digit have been accurately located, e.g., using the technique suggested by Rabiner and Sambur [8]. We also assume that the endpoints of each reference pattern are accurately known. Now the matching can be cast as a path finding problem over a finite grid, as shown in Fig. 5(b). The DTW algorithm [7] has been popularly used in speech recognition. We briefly describe the technique in the following.

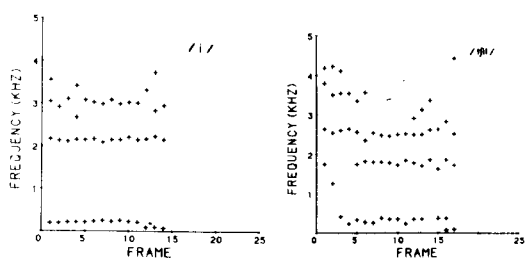


Fig.3. The estimated formant frequency (f_1, f_2, f_3, f_4) trajectories from raw data using ARMA model (p=q=8 or 10). (a) two/i/. (b) seven/tsil/.

Let R(i) be a feature vector for the i^{th} frame of the reference pattern, and T(j) be a feature vector in the j^{th} frame of the test pattern. Let m and n denote the total number of frames in the reference and test patterns, respectively. Then the matching problem reduces to finding an optimal path from the origin (1,1) to the termination point (N,M) on the grid, which minimizes the total matching cost. Let d(i,j) be the local cost function for matching the i^{th} and j^{th} frames in the reference and test patterns, respectively. The DTW algorithm formulates the matching problem in the following way:

Subject to the path constraints which will be defined later, find the path m=w(n), $1 \leqslant n \leqslant N$, $1 \leqslant m \leqslant M$, which minimizes the total cost D(**R**,**T**) defined by

$$D(\mathbf{R},\mathbf{T}) = \min_{(i(k),j(k),k)} \left[\sum_{k=1}^{K} d(i(k), j(k))w(k) \right] / n(w) \tag{21}$$

where k is a parameter of the path, w(k) is a weight function associated with k, and n(w) is a weight function associated with w.

Here the path constraints consist of the endpoint constraints and the local path constraints. The endpoint constraints restrict the optimal path w(n) to begin at the point (1,1) and to end at (N,M). The local continuity constraints guarantee that an excessive compression or expansion in the time scales can be avoided. The local constraints consist of the monotonicity constraint:

$$i(k+1) \geqslant i(k) \text{ and } j(k+1) \geqslant j(k), \tag{22}$$

and a continuity constraint which restricts the local range of the path in the vicinity of the point (m,n) as shown in Fig. 5(a). The valid paths to the point (n,m) come from either (n-1, m-2), or (n-1, m-1), or (n-1, m). We further restrict that the path from (n-2, m) should not go to the

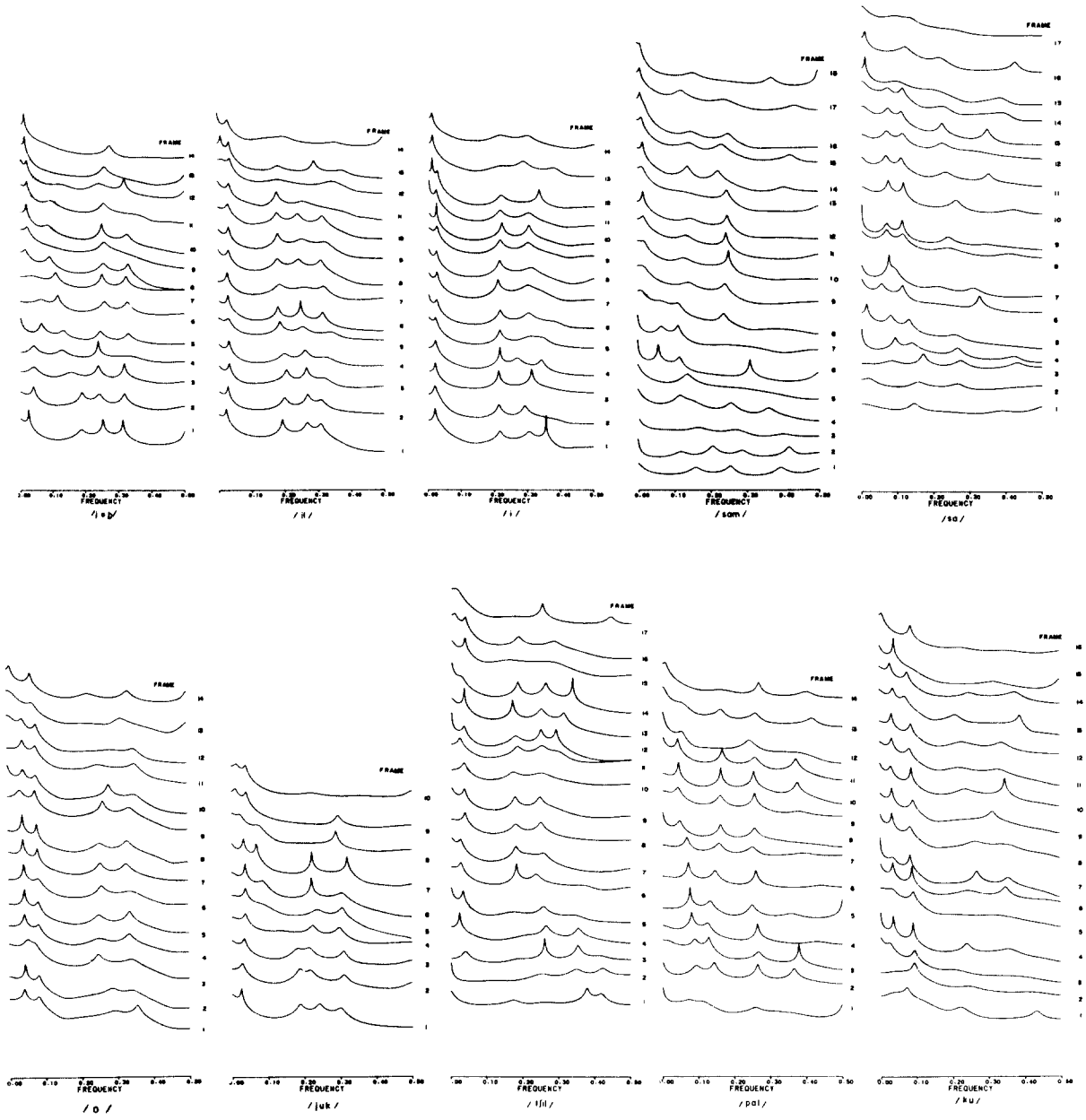


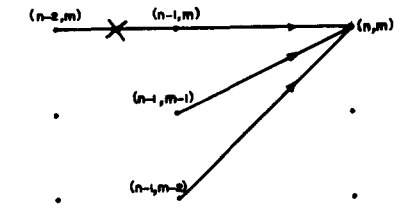
Fig.4. The smoothed spectra of isolated Korean digits using ARMA model (p=q=8 or 10).

point (n,m) through (n-1, m). Also, all points on any path must lie within the allowable regions of the (n,m)-plane, as indicated in Fig. 5(b). These allowable regions can be expressed as

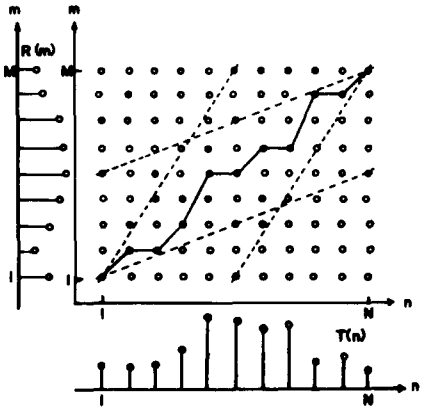
$$1 + (i(k)-1) \leqslant j(k) \leqslant 1+2\,(i(k)-1), \qquad (23a)$$

$$M+2(i(k)-N) \leqslant j(k) \leqslant M+ (i(k)- N\ ). \qquad (23b)$$

There are several frequently used local distane measures d(i,j). In this implementation, we have used a normalized Euclidean distance measure defined on the tridimensional space (f_1, f_2, f_3), where f_j indicates the j^{th} formant frequency in



(a)



(b)

Fig.5. Dynamic time warping : (a) one path set of possible transitions to the grid point (n,m), (b) an example illustrating the grid for warping **T**(n) to **R**(m) via the path m=w(n).

each pattern. The measure is defined by

$$ND(i,j) = \left[\sum_{k=1}^{3} \left[(f^{(T)}_{i,k} - f^{(R)}_{j,k})/f_k\right]^2\right]^{1/2} \quad (24)$$

where $f^{(T)}_{i,k}$ is the k^{th} formant frequency of the i^{th} test pattern, $f^{(R)}_{j,k}$ is the k^{th} formant frequency of the j^{th} reference pattern, and f_k is the mean of $f^{(T)}_{i,k}$ and $f^{(R)}_{j,k}$. Generally, the variation in the first and second formant frequencies is small relative to that in the third formant, as shown in Fig.3. Thus, if an unnormalized distance is used, the variation in the third formant frequency will dominate the distance measure. This phenomenon is undesirable in recognition, since the first and second formant frequencies are more important than the third in the sense of the amount of information contained in each formant frequency. Thus we have employed the normalized distance measure to avoid any bias toward the variation in the third formant frequency.

V. Experimental Results

The overall recognition system is illustrated in Fig. 6. In this experiment, we used a set of isolated words spoken by three 21-year-old males having a standard Korean accent. Uttered words were recorded in a soundproof room using a high-quality audio tape-recording system. The recorded words were processed by a VAX 11/750 digital computer equipped with 12-bit A/D and D/A converters. The recoded analog speech signal was first filtered by a low-pass filter with a cutoff frequency of 4.8 kHz. The filtered analog sgnal was sampled at 10 kHz sampling rate and converted into a digital signal using a 12-bit resolution A/D converter. The digital signals thus obtained was divided into equally spaced frames of fixed size of 25.6 ms, or equivalently 256 samples.

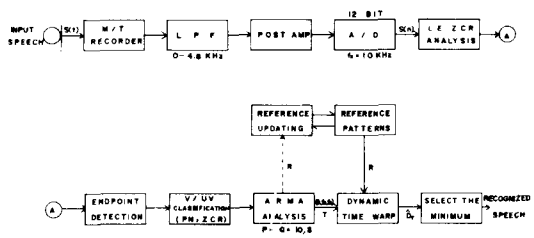


Fig.6. Overall block diagram of the proposed Korean isolated digit recognition system.

Fig. 1 illustrates the fact that we are able to accurately find the frame f_M which keeps an initial sound separated from the medial vowel of a Korean orthographic syllable. Here we use the property that a direction of slope of ZCR difference DZ_k and NLE difference DE_k are opposite at the separating frame f_M. The ZCR of the initial sound of unvoiced digits (70/frame) is nearly twice that of voiced digits (40/frame). The ZCR can subsequently be seen to be a useful parameter for classifying voice and unvoice sounds. The logical combination of the NLE and the ZCR is useful to find a separating frame.

Fig.7. shows that ARMA model proposed here

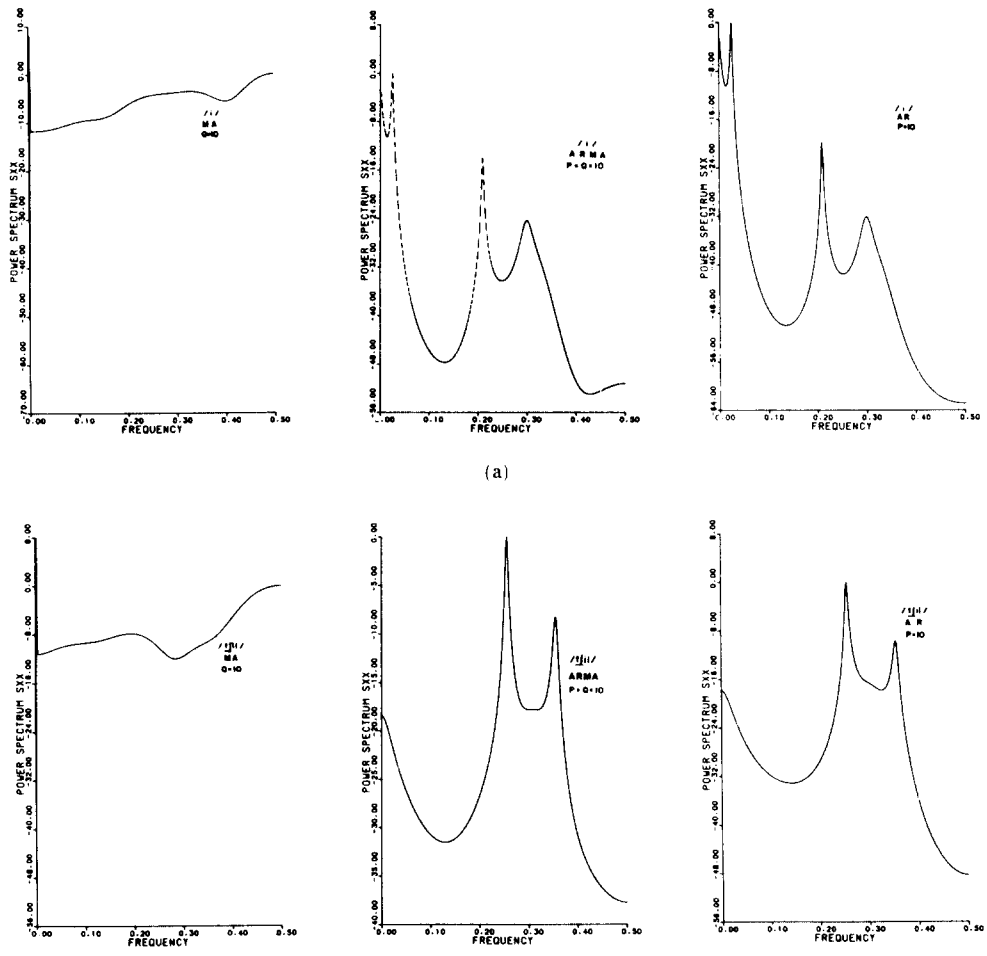


Fig.7. Comparison of AR, MA, and ARMA spectral analysis performance: (a) vowel/ i/spectrum. (b) consonant/ts/spectrum.

provides better spectral estimates than either the AR or MA models for analyzing the consonant/ ts/. The pole frequencies in the ARMA model are very similar to those of the AR model for the voiced sound/i/, since the antiformant frequencies in the MA model are somewhat smooth. However, the pole frequencies in the ARMA case are shifted compared to those of the AR case resulting in antiformant frequencies in the MA case for the consonant /ts/. Futhermore, for consonants, it is clear that spectral analysis methods based on ARMA process models are superior to those based on AR process models.

Fig. 2. shows that the optimal order ARMA method can be used to trace both the pole and zero frequencies with high accuracy for the consonant/ts/and the vowel/i/, respectively. This has motivated our use of the Cadzow-Kay closed form ARMA model for speech analysis.

Table. 1 shows the results of a recognition experiment performed on Korean isolated digits. It shows that 100% classification of voiced and unvoiced sounds was achieved, while a recognition rate of 97.3% was achieved on 150 digits for three speakers. Typical errors were substitutions of the digits "one/il/" with "two/i/", "three/sam/" with "eight/pal/", "four/sa/" with "seven/tsil/", and "six/yuk/" with "five/o/." The recognition performance was further improved by a adapting a normalized Euclidean distance method as a local

Table.1 Confusion matrices for Korean digits.

In \ Out	0	1	2	3	4	5	6	7	8	9
0	15									
1		14	1							
2			15							
3				14					1	
4					14			1		
5						15				
6						1	14			
7								15		
8									15	
9										15

Correct rate; 97.3% $\left(\frac{146}{150}\right)$

distance measurement and via the selection of an optimal order ARMA model for consonants. Each digit can be recognized with five seconds using pole-zero (p=Q= 8 or 10) model.

VI. Conclusion

We have proposed three methods for the improvement of Korean isolated digit recognition. First, we reduce recognition time to 50 percent via the complete classification of the unvoiced initial sound digits and voiced initial sound digits by a logical combination of normalized log energy and zerocrossing rates before frequency analysis. Secondly, using Cadzow-Kay's closed form ARMA spectral estimation method, it is shown that when the order (p=q) is 6, 8, 9, 10, 14, 16, and 25, in order to extract three formant frequencies, $(8,8)^{th}$ and $(10, 10)^{th}$ order ARMA spectral estimates are necessary to provide the desired frequency resolution. This result shows that the optimal order for the analysis of consonants is smaller than that for the vowel sounds. It has been demonstrated that the ARMA model provides spectral estimates superior to those of the more specialized AR model and MA models. Fianlly, it is shown that the DTW algorithm is made more effective for the recognition of Korean isolated digits by using a normalized Euclidean distance as local distance measurement in order to reduce the weights to lead arbitary frequency to bias. A recognition rate of 97.3% on 150 digits for three speakers is obtained. The errors are attributable to untrained speakers and confusion of the formant trajectory of the phoneme; overall, the recognition method and three proposed algorithms are found to be very effective.

References

[1] Cadzow, J.A. "High performance spectral estimation-a new ARMA method," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 5. Oct. 1980.

[2] Han, H., Kim, S.H., and Park, K.T. "A study on the automatic recognition of Korean basic spoken digit using energy of special bandwidth," *Proc. Korean Inst. Electron. Engrs.*, vol. 19, no. 3, 5-12, Feb. 1982. (in Korean).

[3] Kay, S.M. "A new ARMA spectral estimator," *IEEE Trans. Acoust., Speech, Signal Process.* vol. ASSP-28, no. 5, 585-588, Oct. 1980.

[4] Markel, J.D. and Suzuki, H. *Linear Prediction of Speech.* New York: Springer-Verlag. 1976.

[5] Miyanag, Y., Miki, N., and Nagai, N. "Adaptive identification of a time-varying ARMA speech model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 3, 423-433, June 1986.

[6] Morikawa, H. and Fijisaki, H. "Adaptive analysis of speech based on a pole-zero representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 1, 77-87, Mar. 1982.

[7] Myers, C., Rabiner, L.R., and Rosenberg, A.E. "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 6, 623-634, Dec. 1980.

[8] Rabiner, L.R. and Sambur, M.R. "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, 297-315, Feb. 1975. *