the input signal to produce the time-varying spectrum of the output signal.

VI. Conclusions

In this correspondence, we have derived the Wigner distribution relation between the two input and output time-varying spectra of the linear time-variant system; this mathematical relationship can tell us how a time-variant system modifies the time-varying spectrum of the input to determine that of the output, and this result will be very useful for time-variant filtering in mixed time-frequency domain. A linear time-variant digital filtering example is used to show the effect of time-variant filtering by using the Wigner distribution.

Appendix

In this appendix, we will derive the new WD relations (11) and (13) for the linear time-variant system:

$$\begin{aligned}
&\frac{1}{2\pi}\iint W_h(t,t',w,w')\cdot W_x(t',-w')\,dt'\,dw' \\
&\quad=\frac{1}{2\pi}\iint\left[\iint h(t+\tau_1/2,t'+\tau_2/2)\right. \\
&\qquad\left.\cdot h^*(t-\tau_1/2,t'-\tau_2/2)\,e^{-j(w\tau_1+w'\tau_2)}\,d\tau_1\,d\tau_2\right] \\
&\qquad\cdot\left[\int x(t'+\tau_3/2)\cdot x^*(t'-\tau_3/2)\,e^{jw'\tau_3}\,d\tau_3\right]dt'\,dw' \\
&\quad=\frac{1}{2\pi}\iiiint\left[\int e^{-jw'(\tau_2-\tau_3)}\,dw'\right]h(t+\tau_1/2,t'+\tau_2/2) \\
&\qquad\cdot h^*(t-\tau_1/2,t'-\tau_2/2) \\
&\qquad\cdot x(t'+\tau_3/2)\,x^*(t'-\tau_3/2)\,e^{-jw\tau_1}\,d\tau_1\,d\tau_2\,d\tau_3\,dt' \\
&\quad=\iiiint\delta(\tau_2-\tau_3)\,h(t+\tau_1/2,t'+\tau_2/2) \\
&\qquad\cdot x(t'+\tau_3/2)\cdot h^*(t_1-\tau_1/2,t'-\tau_2/2) \\
&\qquad\cdot x^*(t'-\tau_3/2)\,e^{-jw\tau_1}\,d\tau_1\,d\tau_2\,d\tau_3\,dt' \\
&\quad=\iiint h(t+\tau_1/2,t'+\tau_2/2)\,x(t'+\tau_2/2) \\
&\qquad\cdot h^*(t-\tau_1/2,t'-\tau_2/2) \\
&\qquad\cdot x^*(t'-\tau_2/2)\,e^{-jw\tau_1}\,d\tau_1\,d\tau_2\,dt'.
\end{aligned}$$

Now let $t'+\tau_2/2=\alpha_1$ and $t'-\tau_2/2=\alpha_2$, then $t'=(\alpha_1+\alpha_2)/2$, $\tau_2=\alpha_1-\alpha_2$

$$J=\begin{vmatrix}\frac{\partial t'}{\partial\alpha_1} & \frac{\partial t'}{\partial\alpha_2}\\ \frac{\partial\tau_2}{\partial\alpha_1} & \frac{\partial\tau_2}{\partial\alpha_2}\end{vmatrix}=\begin{vmatrix}\frac{1}{2} & \frac{1}{2}\\ 1 & -1\end{vmatrix}=-1$$

$$\begin{aligned}
\therefore|J|&=1\\
&=\iiint h(t+\tau_1/2,\alpha_1)\,x(\alpha_1)\,h^*(t-\tau_1/2,\alpha_2) \\
&\quad\cdot x^*(\alpha_2)\,e^{-jw\tau_1}\,d\alpha_1\,d\alpha_2\,d\tau_1 \\
&=\int\left[\int h(t+\tau_1/2,\alpha_1)\,x(\alpha_1)\,d\alpha_1\right] \\
&\quad\cdot\left[\int h^*(t-\tau_1/2,\alpha_2)\,x^*(\alpha_2)\,d\alpha_2\right]e^{-jw\tau_1}\,d\tau_1 \\
&=\int y(t+\tau_1/2)\,y^*(t-\tau_1/2)\,e^{-jw\tau_1}\,d\tau_1 \\
&=W_y(t,w).
\end{aligned}$$

So $W_y(t,w)=1/2\pi\iint W_h(t,t',w,w')\cdot W_x(t',-w')\,dt'\,dw'$.

Remark: If the definition domain is real, then follow the above proof, this relation becomes

$$W_y(t,w)=\frac{1}{2\pi}\iint W_h(t,w,t',w')\,W_x(t',w')\,dt'\,dw'.$$

References

[1] T. A. C. M. Claasen and W. F. G. Mecklenbräuker, "The Wigner distribution—A tool for time-frequency signal analysis. Part I: Continuous-time signals; Part II: Discrete-time signals; Part III: Relation with other time-frequency signal transformations," *Phillips J. Res.*, vol. 35, pp. 217–250, pp. 276–300, and pp. 372–389, 1980.

[2] B. E. A. Saleh, and N. S. Subotic, "Time-variant filtering of signals in the mixed time-frequency domain," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1479–1485, Dec. 1985.

[3] G.-F. Boudreaux-Bartels and T. W. Parks, "Time-varying filtering and signal estimation using Wigner distribution synthesis techniques," *IEEE Trans. Acoust., Speech, Signal Processing.*, vol. ASSP-34, pp. 442–451, June 1986.

[4] K. B. Yu and S. Cheng, "Signal synthesis from pseudo-Wigner distribution and applications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1289–1302, Sept. 1987.

[5] B. V. K. Vijaya Kumar, C. P. Neuman, and K. J. DeVos, "Discrete Wigner synthesis," *Signal Processing*, vol. 11, no. 3, pp. 277–304, Oct. 1986.

[6] L. Jacobson and H. Wechsler, "A theory for invariant object recognition in the frontoparallel plane," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 325–331, May 1984.

[7] L. A. Zadeh, "Frequency analysis of variable networks," *Proc. IRE*, vol. 32, pp. 291–299, Mar. 1950.

[8] N. C. Huang and J. K. Aggarwal, "On linear shift-variant digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 672–679, Aug. 1980.

Recognition of Chinese Diphones

LING Y. WEI, JONG P. LEE, AND CHI C. LEE

***Abstract*—We have studied recognition of isolated Chinese words with more emphasis on diphones. Under the nearly same test conditions, we obtained speaker-dependent recognition rates as follows: 1) 76.3 percent for 59 Chinese phonetic units; 2) 95 percent for 40 monophones; and 3) 99.5 percent for 100 diphones. With three speakers, the recognition rate of 100 diphones becomes 94 percent. The very high recognition rate of diphones is consistent with our experience in spoken Chinese, and it points to a new direction toward machine understanding of Chinese language in the future.**

I. Introduction

According to linguistics, a word should be a unit in the spoken language characterized by syntactic and semantic independence and integrity [1]. This notion of "word" is generally agreed upon by renowned linguists such as Chao Yuen-ren [2]. In the Chinese language, a word may consist of one character, two characters, three characters, or more. Since each Chinese character is a monosyllable, or a monophone to be precise (a Chinese monophone takes one of the forms: C, V, CV, and CVN), we shall classify Chinese words as monophone, diphone (two monophones in concatenation and with a perspicuous meaning), triphone, quadraphone, etc. Lexical statistics [3], [4] show that diphones comprise about two-thirds of

Manuscript received September 24, 1987; revised March 29, 1988.

The authors are with the Department of Computer Engineering, Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.

IEEE Log Number 8822846.

all Chinese words. Similar relative proportions are observed in spoken and written Chinese. The popularity of diphones stems from their superiority in a phonetic and semantic union that gives much less ambiguity or confusion than monophones.

In recognition of Mandarin speech, most of the work has been concerned with four tones [5]–[7] and subword units [8]–[15]. Some excellent results have been obtained in these studies using various techniques. Work on recognition of Chinese words is barely beginning. Lee and Huang used VQ and HMM techniques for connected digit recognition [16]. They produced recognition rates such as: 76 percent for single digits (monophones); 92.7 percent for 2 connected digits (diphones); 88.5 percent for 3 connected digits (triphones); and 88.75 percent for 4 connected digits (quadraphones). Huang *et al.* had embarked on recognizing 1000 Chinese words of which 300 were monophones [17]. With two well-trained speakers, they obtained a recognition rate of 70 percent for monophones and 95 percent for polyphones (mainly diphones).

In our project, we aim at recognizing continuous Chinese speech of a very large vocabulary. Our strategy is to recognize diphones in the very first stage using the simplest possible techniques. Determination of monophones will be next. Because of homonyms, monophones will be treated with greater care using syntactic and semantic analysis if necessary. Our philosophy is to go easy first and face hardship later. The present study is only the first step moving in that direction.

II. Method

A. Important Parameters for Acoustic Analysis

1) Frequency range: 0–10 kHz
2) Sampling frequency: 20 kHz
3) A/D resolution: 12 bits
4) Hamming window: 25.6 ms wide
5) Frame rate: 15 ms.

Spectral analysis and recognition were performed with IBM PC/AT.

B. Spectral Energy Ratio Coding (SERC) [18]

The frequency range (0–10 kHz) of speech sound is divided into two groups of bands: 1) the primary bands: $A1$ = 0–0.4 kHz, $A2$ = 0.4–4 kHz, and $A3$ = 4–10 kHz; 2) the secondary bands: $B1$–$B12$ where $B10$ = 4–6 kHz, $B11$ = 6–8 kHz, and $B12$ = 8–10 kHz. The bands $B1$–$B9$ in the 0–4 kHz range are spaced according to the critical band scheme or mel scale [19], [20].

Let $A0$ be the spectral energy of speech found in the 0–10 kHz range, and let the above notations for the various bands also stand for the spectral energies in those bands, respectively. Then the spectral energy ratio (SERC) is defined by [18]

$$ai = (Ai/A0) \times 100, \qquad i = 1, 2, 3 \tag{1}$$

$$bj = (Bj/A0) \times 100, \qquad j = 1, 2, \cdots, 12. \tag{2}$$

The SERC takes the form of $A0$, $a1$, $a2$, $a3$; $b1$, $b2$, $\cdots$, $b12$ where ($a1$, $a2$, $a3$) is called the primary code and ($b1$, $b2$, $\cdots$, $b12$) is the secondary code. We can use the primary code for coarse classification which can reduce time for screening the database and use the secondary code for final recognition.

C. Endpoint Detection

We use the following features of Mandarin speech for endpoint detection: 1) short time energy $A0$; 2) in the short time analysis, every Chinese monophone word contains a single peak; and 3) the average length of speaking a Chinese monophone is about 360 ms. Assume $E(n)$ is the energy of frame n. We set a dynamic threshold ET as

$$ET = (1/5)\left[\sum_{n=1}^{5} E(n)\right] \times \text{FACTOR} \tag{3}$$

where the FACTOR is 1.5 in our system. Using $A0$, ET, and other pertinent conditions, we can pick up a special signal from background noise. For diphone endpoint detection, we have two methods. In Method A, we treat a diphone as a monophone. In Method B, we take a diphone as composed of two monophones and cut it into two parts.

D. Pattern Matching and Decision Rule

Calculation of distance measures (City Block) for pattern matching is very simple. The matching between a test frame (m) and a reference frame (n) is performed in the following two stages.

1) Coarse matching:

$$d(a) = \sum_i \left| ai(m) - ai(n) \right| \tag{4}$$

where i = 1, 2, 3 and a's, the primary codes.

2) Fine matching:

$$d(b) = \sum_i \left| bj(m) - bj(n) \right| \tag{5}$$

where j = 1, 2, 3, $\cdots$, 12 and b's, the secondary codes.

We do not use DTW, but do use the linear time matching which is simpler and better than DTW for small variations of speech periods (±30 percent) [21]. What we do is to match frames from a few fixed points (u_k) in the utterance duration (T) such that $u_k = c_k T$ where c_k is a constant depending on position and is between 0 and 1. One set of c_k's is used for monophones (input and reference) and another set for diphones (input and reference). This is equivalent to linear time alignment and is in principle not much different from DTW in the cases of monophones and diphones. For the choices of c_k(u_k) values, we follow the suggestion by Matsuda *et al.* that the points chosen are the optimal points for recognizing consonants and vowels in a monosyllable [22].

For matching of monophones, the fixed points are set at u_1 = 0.33 T, u_2 = 0.67 T, u_3 = 0.89 T. Each is measured from the beginning point of the utterance. For matching of diphones with Method A, take six fixed points at u_1 = 0.17 T (starting point for matching), u_2 = 0.33 T, u_3 = 0.44 T, u_4 = 0.67 T, u_5 = 0.83 T, and u_6 = 0.94 T. With Method B, we treat a diphone as two monophones. At each point, take 6 frames backward (toward the left). Sum 6 corresponding frame distances. The number of 6 frames was empirically determined for its duration (90 ms) which was adequate for feature condensation, thus saving time in computation. Then sum distances from all fixed points to get the total distance.

After the total distance $D(a)$ for primary codes has been computed, we can use it to make screening. The screening rule is very simple. First, if $D(a)$ exceeds a preset threshold value, there is no need to compute $D(b)$ for the reference word fetched for matching the input word. The former is then removed from the list of candidates. If $D(a)$ is lower than the preset threshold, $D(b)$ is computed. The decision rule is to choose the one with the lowest $D(b)$ of all candidates as the final recognition result.

The two stage matching can save computation time by as much as 83 percent if the threshold of $D(a)$ is preset to eliminate 80 percent of words. This indicates the advantage of feature condensation with a code and linear matching at fixed points.

III. Result

When listening to Chinese speech, our experience is that the degree of intelligibility (hence recognition rate) increases in the following order: 1) phonemes, 2) monophones, and 3) diphones. In this study, we set out to determine how a machine would respond in the listening test. For this purpose, we have to minimize the effect of human factor so that only the "machine" plays the dominant role. In Experiments 1, 2, and 3, a single speaker (JPL) did all the talking. Experiment 4 was designed to test the relative merits of Method A and Method B for recognizing 100 diphones uttered by three speakers.

1) Experiment 1—Recognition of Phonetic Units: There are 59 phonetic units (21 consonants and 38 vowel-containing units) in the Chinese language. JPL spoke twice: once for test and again for reference. The recognition rate is 76.3 percent. This low rate is

due to the fact that there are two groups of very confusable units: 1) the fricative group containing 9 units (s, z, sh, tz, ch, j, ji, shi, chi); and 2) the nasal group containing 17 units (m, n, an, en, ang, eng, ian, in, ing, uan, uen, uang, ueng, iuan, iun, iang, iung). To determine which group performs better, we tested the nasal group and another group containing other consonants. The recognition rate of the nasal group was 83.3 percent and that of the other group was 75 percent. This gives us a little consolation because 186 of all 416 basic tones are nasals which fortunately are found not to be the diehard for recognition. If they were, 44 percent of all Chinese words would be troublesome in Mandarin conversation.

2) Experiment 2—Recognition of Monophones: JPL spoke 40 monophone words 5 times: 2 for test and 3 for reference. The 40 monophone words were chosen mostly by their high frequency of occurrence in daily use. The recognition rate is 95 percent in each of two tests. The 40 monophones contain 3 groups of confusable words: 1) 9 words each beginning with "s" or "sh," 2) 5 words each ending with "*I*," and 3) 16 words each ending with a nasal. Without much drilling in pronunciation by JPL, the recognition rate would not reach 95 percent, which is unusually high for monophones.

3) Experiment 3—Recognition of Diphones: JPL spoke 100 diphones 5 times: 2 for test and 3 for references. These diphones were chosen from newspapers based on their frequency of occurrence. The recognition rates are as follows:

	Test 1	Test 2	Average
Method A	98 percent	97 percent	97.5 percent
Method B	100 percent	99 percent	99.5 percent

The fact that Method B (treating a diphone as two monophone words) is superior to Method A (treating a diphone as a single word) for a well-trained speaker can be explained as follows. One and the same speaker uttering a diphone, say X, at different times could change not only the total duration $T(X)$ but also the relative durations $T(X_a)$ and $T(X_b)$ of its components X_a and X_b. In Method A, the 6 matching points $u_1 \cdots u_6$ are based on the total duration $T(X)$. When $T(X)$ changes, so will be the u's. Thus, the u-points in a test frame would not linearly match the u-points in the reference frame. This suboptimal matching would lead to a greater error. In Method B, u_1, u_2, and u_3 are based on $T(X_a)$, and u_4, u_5, and u_6 are based on $T(X_b)$. No matter how $T(X_a)$ and $T(X_b)$ change, the u-points will always stay at the same optimal points in the test frame and in the reference frame. The matching is therefore optimal for the utterance by the same speaker at different times. This assures nearly perfect matching in Method B as observed.

4) Experiment 4—Recognition of Diphones: Three speakers (including JPL) spoke 3 times: 2 for test and 1 for reference (using multitemplates, not their average). The results are as follows.

Method A	Test 1	Test 2	Ave.
Speaker 1	98 percent	92 percent	95 percent
2	94 percent	96 percent	95 percent
3	92 percent	92 percent	92 percent
		Total average recognition rate	94 percent
Method B			
Speaker 1	96 percent	94 percent	95 percent
2	82 percent	92 percent	87 percent
3	92 percent	86 percent	89 percent
		Total average recognition rate	90.33 percent

Here, for multispeakers, Method A gives higher recognition rate than Method B, just opposite to that shown in Experiment 3 for a single speaker (JPL). Our explanation is as follows. Method B is based on monophones. The database contains actually 600 monophones templates which could form 600 × 600 diphones of which only 300 are naturally spoken (by the 3 speakers) and the rest are fictitious (i.e., the first monophone spoken by one person and the second by another person). The ratio of fictitious to natural diphones is 1200 : 1. Since fictitious diphones are highly confusable, the recognition rate by Method B is expected to be lower than that by Method A. This does not apply to a single speaker because Method B gives better matching than Method A as explained before.

IV. Discussion

In this study, we have employed some simple methods for recognition of Mandarin speech with more emphasis on diphones. The primary code was designed for feature condensation. Instead of DTW, linear matching was performed at a few fixed points of word duration. From each point backward, only six frames were compared. All these schemes have contributed to considerable saving not only in machine time but also in man's training effort.

As a matter of interest, we include here a table showing the recognition rates of diphones by multispeakers in three different systems.

Systems	Speakers	Diphones	Recognition Rate	Techniques
Lee (16)	15	100*	92.7 percent	LPC, VQ, HMM, LB*
Huang (17)	3	700*	95 percent	LPC, Cepstrum, NLS*, DTW
Wei	3	100	94 percent	FB (SERC), LM*

Note: 100*—100 connected 2-digits.
LB*—Level building
700*—Including some polyphones other than diphones other than diphones.
NLS*—Nonlinear segmentation
LM*—Linear matching.

Further experiments are in progress to test how vocabulary size and number of speakers would affect recognition rate.

References

[1] C. N. Li and S. A. Thompson, *Mandarin Chinese*. Berkeley, CA: University of California Press, 1981.

[2] C. Yuen-ren, *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press, 1968.

[3] Y. Ho, *Dictionary for Mandarin Daily*. Taipei, Taiwan: Mandarin Daily, 1986.

[4] *Source of Chinese Words*. Taipei, Taiwan: Lan-Tern Publishers, 1979.

[5] J. Lee, H. C. Wang, and Y. C. Chang, "Mandarin lexical tone recognition based on vector quantization and hidden Markov models," in *Proc. Int. Comput. Symp.*, Tainan, Taiwan, Dec. 17-19, 1986, pp. 1168-1173.

[6] X. X. Chen, C. N. Cai, P. Guo, and Y. Sun, "A hidden Markov model applied to Chinese four tone recognition," in *Proc. 1987 ICASSP*, Dallas, TX, pp. 797-800, Apr. 16-19, 1987.

[7] H. Ma, "The four tone recognition of continuous Chinese speech," in *Proc. 1987 ICASSP*, Dallas, TX, Apr. 16-19, 1987, pp. 65-68.

[8] B. C. Wang, C. Y. Tseng, and L. S. Lee, "A three pass Mandarin vowel recognition system," in *Proc. Int. Comput. Symp.*, Tainan, Taiwan, Dec. 17-19, 1986, pp. 1144-1149.

[9] Y. M. Hsu, C. Y. Tseng, and L. S. Lee, "Mandarin vowel recognition based upon a segmental model," in *Proc. Int. Comput. Symp.*, Tainan, Taiwan, Dec. 17-19, 1986, pp. 1150-1154.

[10] C. W. Hwang, C. Y. Tseng, and L.S. Lee, "An efficient Mandarin vowel recognition system based upon multi-section vector quantization and branch-and-bound classification techniques," in *Proc. Int. Comput. Symp.*, Tainan, Taiwan, Dec. 17-19, 1986, pp. 1155-1159.

[11] L. S. Lee, "A three pass approach for Mandarin final recognition," in *Proc. 6th Workshop Comput. Syst., Tech.*, Sun Moon Lake, Taiwan, Aug. 12-15, 1987, pp. 683-700.

[12] H. C. Wang, "A study of confusable words in Mandarin speech," in *Proc. 6th Workshop Comput. Syst. Tech.*, Sun Moon Lake, Taiwan, Aug. 12-15, 1987, pp. 701-720.

[13] L. S. Lee, "A segmental model approach for Mandarin final recog-

nition," in *Proc. 6th Workshop Comput. Syst. Tech.*, Sun Moon Lake, Taiwan, Aug. 12-15, 1987, pp. 745-762.
[14] C. H. Wu, C. Y. Tseng, and L. S. Lee, "New recognition techniques for unvoiced Mandarin consonants based upon hidden Markov models," in *Proc. Nat. Comput. Symp.*, Taipei, Taiwan, Dec. 17-18, 1987, pp. 973-979.
[15] P. Y. Tin, C. Y. Tseng, and L. S. Lee, "A Mandarin consonant recognition system based upon time, frequency domain features and finite state vector quantization technique," in *Proc. Nat. Comput. Symp.*, Taipei, Taiwan, Dec. 17-18, 1987, pp. 980-987.
[16] C. C. Lee and P. Y. Huang, "Speaker independent connected Chinese spoken word recognition," in *Proc. 6th Workshop Comput. Syst. Tech.*, Sun Moon Lake, Taiwan, Aug. 12-15, 1987, pp. 721-743.
[17] X. D. Haung, L. H. Cai, D. T. Fang, B. J. Ci, L. Zhou, and L. Jian, "A large vocabulary Chinese speech recognition system," in *Proc. 1987 ICASSP*, Dallas, TX, pp. 1167-1170.
[18] L. Y. Wei, "Spectral energy ratio coding (SERC) of speech," *Taiwan Telecom. Tech. J.*, vol. 4, pp. 265-270, 1984.
[19] E. Zwicker, "Subdivision of audio frequency range in top critical band," *J. Acoust. Soc. Amer.*, vol. 23, p. 248, 1961.
[20] G. Fant, *Speech Sounds and Features*. Cambridge, MA: M.I.T. Press, 1973.
[21] A. Komatsu, A. Ichikawa, K. Nakata, Y. Asakawa, and H. Matsuzaka, "Phoneme recognition in continuous speech," in *Proc. 1982 ICASSP*, pp. 883-886.
[22] Y. Matsuda, S. Tesuka, M. Kanoh, M. Nishimura, and T. Kanedo, "A method for recognizing Japanese monosyllables by using intermediate cumulative distance," presented at 1984 ICASSP, Paper 9.3.1-4.
[23] L. Y. Wei, "Theoretical basis and strategy for machine understanding of Chinese language," *Taiwan Telecom. Tech. J.*, vol. 6, pp. 383-390, 1987.

Comments on "A Two-Stage Representation of DFT and Its Applications"

JA-LING WU AND CHAU-YUN HSU

***Abstract*—This correspondence contains comments on and several corrections to a recently published TRANSACTIONS paper.**

In the above paper,[1] Ersoy developed a two-stage representation in terms of preprocessing and postprocessing of DFT by vector transformation of sines and cosines into new basis functions using Mobius inversion of number theory. This comment points out first that the inversion Mobius transform pair, (A.3) and (A.4), used[1] are valid only when f is a positive rational number [1, p. 208]. Thus, (A.6) should read

$$X_c(f) = \frac{1}{4f} \sum_{\substack{m=1 \\ \text{odd}}}^{\infty} \frac{\mu_m}{m} \left(\sum_{n=-\infty}^{\infty} \left(2x\left(\frac{n}{mf}\right) - x\left(\frac{n}{2mf}\right)\right)\right) \tag{A.6}$$

and $f > 0$. Second, (2.11) should read

$$n' = 0, 1, \cdots, M_1 - 1. \tag{2.11}$$

This range is very important because it determines the size of the circular correlation in the postprocessing matrix equation. The correctness of this new range has been verified by computer simulation.

Third, the two-stage representation form can be applied directly to the computation of discrete Hartley transform (DHT) [3]. Interestingly, since the elements in the preprocessing matrix of DHT are 0, 1, −1, 2, and −2, only shift, addition, and subtraction operations are involved for the preprocessing stage of DHT. It can be shown that the postprocessing matrix of DHT is also in a block-diagonal form, with each block being a circular correlation matrix. For example, with $N = 8$ and using the same notation defined,[1] the preprocessing and postprocessing matrix equations can be obtained respectively as follows:

$$\begin{bmatrix} h(0) \\ h(4) \\ h(2) \\ h(1) \\ h(5) \\ h(6) \\ h(7) \\ h(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 2 & 1 & 0 & -1 & -2 & -1 & 0 \\ 1 & -2 & 1 & 0 & -1 & 2 & -1 & 0 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 0 & -1 & -2 & -1 & 0 & 1 & 2 \\ 1 & 0 & -1 & 2 & -1 & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ x(7) \end{bmatrix}$$

$$\begin{bmatrix} X(0) \\ X(4) \\ X(2) \\ X(1) \\ X(5) \\ X(6) \\ X(7) \\ X(3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b(1,8) & b(5,8) & 0 & 0 & 0 \\ 0 & 0 & 0 & b(5,8) & b(1,8) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & b(1,8) & b(5,8) \\ 0 & 0 & 0 & 0 & 0 & 0 & b(5,8) & b(1,8) \end{bmatrix} \cdot \begin{bmatrix} h(0) \\ h(4) \\ h(2) \\ h(1) \\ h(5) \\ h(6) \\ h(7) \\ h(3) \end{bmatrix}.$$

Finally, some typing errors[1] are listed below.

1) With the substituting of $l = m_1 n$ modulo N, (2.3) should read

$$h(l) = \sum_{k=0}^{N-1} x(k) \left[\mu\left(\frac{lk}{N} + \frac{1}{4}\right) - j\mu\left(\frac{lk}{N}\right)\right]. \tag{2.3}$$

2) The index i used in (2.7) and (2.12) should be replaced by n'.

3) The term $b(4, 16)$ used in the lowest block of (2.13) should be replaced by $b(9, 16)$.

4) Equation (A.16) should read

$$b(m_1, N) = P(m_1(l), N) - P(m_2(N - l), N). \tag{A.16}$$

Manuscript received November 24, 1987; revised March 5, 1988.

J.-L. Wu is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China.

C.-Y. Hsu is with the Department of Electrical Engineering, Tatung Institute of Technology, 40, Chung-Shan North Rd., sec. 3, Taipei, Taiwan, Republic of China.

IEEE Log Number 8822847.

[1] O. K. Ersoy, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 825-831, June 1987.

REFERENCES

[1] M. R. Shroeder, *Number Theory in Science and Communication*. New York: Springer-Verlag, 1983.
[2] R. N. Bracewell, "The fast Hartley transform," *Proc. IEEE*, vol. 72, pp. 1010-1018, Aug. 1984.