SPEECH RECOGNITION SYSTEM BASED ON WALSH FUNCTION

F.Y. KORKMAZSKY

Computer Design Office
Chernovtsy, USSR 274030

ABSTRACT

The principles of designing a speech recognition system based on Walsh functions are described. For the initial speech signal description the energy spectrum of the Walsh transform with Hadamard ordering is proposed. The advantages of the above signal representation system are its invariance under cyclic shift of signals and a high processing speed of computing the energy spectrum of signals. It is shown that a posteriori informativeness coefficients used in training and recognition procedures give a considerable increase in speech signal recognition rate. A method of reference speech pattern correction at the recognition stage is proposed which should also contribute to a higher speech recognition rate.

INTRODUCTION

Most of the present speech recognition systems using the spectral speech signal representation are designed on the basis of Fourier transform. It has gained its popularity among researchers due to development of fast Fourier transform (FFT) algorithms which had helped to increase the processing speed for computation of signal spectra. Thus, to obtain a complex Fourier spectrum using the FFT, $N\log_2 N$ of complex additions and $N/2\log_2 N$ of complex multiplications are required. The principal advantage of the Fourier transform method is the invariance of Fourier energy spectrum under the cyclic shift of the input signal. This property of the Fourier transform enables one to obtain energy spectra which are independent of the phase of the processed signal. However, the realization of the FFT algorithm would require either special equipment performing the FFT, or the use of integrated circuits for processing of spectra by means of FFT. A microprocessor-based program realization of FFT is difficult because real time performing of the FFT algorithm is not possible, especially when the number of processing points is large. On the other hand, there are some orthogonal transforms which neither require processing in the complex plane, nor the use of multiplication. Therefore, such transforms are performed much faster than FFT, and lend themselves to microprocessor-based realization. An important transformation of this kind is the Walsh transform. A Walsh function is a complete orthonormal set of functions assuming either +1 or -1 values. Thus, the Walsh transform which cosists in a convolution of input signals with Walsh functions, requires only two operations, i.e. addition and subtraction, and does not require multiplication. There are several varieties of Walsh transforms/1/ for which corresponding fast Walsh transform (FWT) algorithms can be used. The FWT algorithms usually require $N/2\log_2 N$ real additions or subtractions which allows real time realization of these algorithms on the basis of modern microcomputers.

Most of the present speech recognition systems using the Walsh functions for describing speech signals give a high speed of obtaining the spectral description of speech signals. However, the vocabulary for recognition amounts at best to several scores of words/2/. The level of recognition errors is also unsatisfactory. We believe that there are two reasons for low quality of such systems. The first is that the energy spectra of speech signals are described using those varieties of Walsh transforms which are not invariant under the cyclic shift. The second reason is that in these systems the problems of speech processing at higher levels, especially training and decision making procedures, have not been given proper consideration. The purpose of the present paper is to fill in these gaps and to demonstrate the possibilities of the Walsh function method in designing of high quality speech recognition systems.

SPECTRAL DESCRIPTION OF SPEECH SIGNALS

To obtain a spectral description of speech signals, the energy spectrum of the Walsh transform with Hadamard ordering is proposed/1/. Such a description possesses two important advantages over other methods of spectral description of speech signals. The first is that the energy spectrum of the Walsh transform with Hadamard ordering is invariant under the cyclic shift. The second is that the realization of the corresponding fast transform algorithm requires $2(N-1)$ operations of real addition (subtraction). Since the processing speed for computations involving the FWT algorithm is determined by $N\log_2 N$, one can consider computations using the proposed algorithm to be $1/2\log_2 N$ as fast as those involving the FWT algorithms. Thus, for example, when the number of data points N is 256, the proposed algorithm gives a 4-fold increase in the processing speed.

As was mentioned earlier, the energy spectrum of the Walsh transform with Hadamard ordering is invariant under the cyclic shift. This means that for any periodic sequence $X(m)$, its energy spectrum of the Walsh transform with Hadamard ordering will coincide with that for a sequence $X(m+n)$, m and n being integers. As for other varieties of the Walsh transform, such as the Walsh transform with Walsh ordering, only the invariance of energy spectrum under dyadic shifts is valid/1/. This means that the corresponding energy spectra of sequences $X(m)$ and $X(m \oplus n)$ will coincide. The $\oplus$ operation means modulo 2 addition. The $E_h(r)$ components of the Walsh transform energy spectrum with Hadamard ordering are calculated according to the formula/1/:

$$E_h(0) = a_h^2(0), \quad E_h(r) = \sum_{k=2^{r-1}}^{2^r-1} a_h^2(k) \qquad (1)$$

$$r = 1,2,\ldots,n; \quad n = \log_2 N$$

Here $a_h(k)$ are the coefficients of the Walsh transform with Hadamard ordering, N is the number of input sequence points. It can be seen from eq.(1) that the number of components in the energy spectrum is $(n+1)$. Each component $E_h(r)$ of the energy spectrum represents the energy content of a group of sequencies rather than a single sequency, as is the case for the energy spectrum of the Walsh transform with Walsh ordering. The set of sequencies contained in each component of the energy spectrum of the Walsh transform with Hadamard ordering is calculated according to the formula/1/:

$$\begin{cases} F[E_h(0)] = 0 \\ F[E_h(1)] = N/2 \\ \vdots \\ F[E_h(k)] = 2^k, 3\cdot 2^k, 5\cdot 2^k, \ldots, N/2 - 2^k \\ \vdots \\ F[E_h(1)] = 2, 6, 10, \ldots N/2 - 2 \\ F[E_h(n)] = 1, 3, 5, \ldots N/2 - 1 \end{cases} \qquad (2)$$

To represent speech signals by means of the energy spectrum of the Walsh transform with Hadamard ordering, N =128 data points have been used for every time slot of 10 milliseconds. Accordingly, the number of components of the energy spectrum with Hadamard ordering amounted to 7 (n =7). The quantity $E_h(0)$ represents the energy of the direct component of the input sequence, and is not considered here.

To calculate the energy spectrum, a 16-bit computing device has been used whose computing speed characteristics were as follows. Execution time of the basic register arithmetic operations is 0.36 μsec., main memory read/write time is about 2 μsec. The device has 16 general registers. To calculate the energy spectrum of a speech segment 10 ms long by means of the above device, required about 2.5 ms (the quadratic coefficients are calculated using the table of squares contained in the main memory). After the energy spectrum is determined, the speech signal is represented by the quantities $\log_2(E_h(r)/E_0)$, r =1, 2, ... n, and $\log_2 E_0$, where $E_0 = \sum_{r=1}^{n} E_h(r)$. Thus, the input speech signal is represented by a 8-dimensional vector for each 10-ms slot. The number of bits representing components of a given vector was equal to 8. Input speech signal points were also represented using 8 bits per point. To store the reference speech patterns, 4 bits per each component of the 8-component vector are used. Experiments on speech pattern recognition have been carried out using the Walsh energy spectrum with Hadamard ordering for spectral description of speech signals. The recognition rate for vocabularies containing 50, 100 and 250 words was 99%, 96% and 92%, respectively.

TRAINING AND RECOGNITION

To increase the recognition rate of the system, a special training and recognition procedure has been developed. It is based on the following assumptions. An important method for improving the recognition rate is to represent each of the speech patterns to be recognized by several reference patterns obtained from corresponding clusterization of training samples. However, the present clusterization methods usually require optimization of some clusterization quality functional which depends on relative cluster position only within one group of patterns, irrespective of clusters formed within other groups. Meanwhile, in the process of recognition one has to relate distances obtained for different groups. In fact, the recognition rate is determined by relative positions of clusters belonging to different groups. Furthermore, one has to consider the dynamics of each cluster informativeness which depends

on the type of input speech signal. This is especially true for speaker-independent speech recognition systems in which speech variations from speaker to speaker necessitates adoptation to a speaker voice at the recognition stage.
We now assume that the clusterization procedure is carried out for each recognition pattern (i.e. a word or a phonem), and for each of the clusters a reference pattern is formed by averaging the training samples wich have fallen into a given cluster. Suppose then that a whole set of reference patterns is represented by J groups, each group containing one reference pattern for each recognition pattern. Then the adoptation to a speaker's voice within a given system will mean an automatic selection of a group of reference patterns for recognition according to "a posteriori" informativeness coefficients which are dynamically calculated for each group of reference patterns. For each of J groups of reference patterns, the following quantities are proposed as the informativeness coefficients:

$$\lambda(j) = \log_2 \frac{\rho_{i_1(j)}(j)}{\rho_{i_2(j)}(j)}, \quad j=\overline{1,J} \qquad (3)$$

where $\rho_{i_1(j)}(j)$ is the distance between the speech pattern being identified and the nearest reference pattern in the j-th group of reference patterns $(i_1(j)=\arg\min_{i=\overline{1,N}} \rho_i(j))$, N being the number of recognition patterns, $\rho_{i_2(j)}(j)$ is the distance to the next nearest reference pattern in the j-th group of reference patterns $(i_2(j)=\arg\min_{i\neq i_1(j)} \rho_i(j))$. The quantity of $\lambda(j)$ is the contrast range of the decision $i_1(j)$ obtained within the j-th group of reference patterns which directly determines the recognition rate of this decision. The decision i^* concerning the identity of the speech pattern being recognized is made on the basis of a set of decisions for all groups of reference patterns according to the formula

$$\begin{cases} j^{(1)} = \arg\max_{j=\overline{1,J}} \lambda(j) \\ i^* = i_1(j^{(1)}) \end{cases} \qquad (4)$$

A decision is not made in two cases. The first is for

$$\lambda(j^{(1)}) < \tilde{\lambda}_1 \qquad (5)$$

where $\tilde{\lambda}_1$ is a positive quantity. The second is for

$$\begin{cases} i_1(j^{(1)}) \neq i_1(j^{(2)}) \\ \lambda(j^{(1)}) - \lambda(j^{(2)}) < \tilde{\lambda}_2 \\ j^{(2)} = \arg\max_{j=\overline{1,J}, j\neq j^{(1)}} \lambda(j), \; i_1(j^{(2)}) = \arg\min_{i=\overline{1,N}} \rho_i(j^{(2)}) \end{cases} \qquad (6)$$

where $\tilde{\lambda}_2$ is a positive quantity.
The above method can be extended to include the training procedure. Suppose that within each of J groups of reference patterns, initial approximations to reference pattens $\tilde{e}_i^{(0)}(j)$ $(i=\overline{1,N}, j=\overline{1,J})$ are formed for each of N recognition patterns as a result of some self-training procedure. For the v-th training sample of the K-th recognition pattern, the following quantities are calculated:

$$\mathcal{L}_K^{(v)}(j) = \log_2 \frac{\rho_K^{(v)}(j)}{\rho_{K_1(j)}(j)}, \quad j=\overline{1,J} \qquad (7)$$

where $\rho_K^{(v)}(j)$ is the distance between the v-th training sample of the K-th recognition pattern and its reference pattern $\tilde{e}_K^{(0)}(j)$ in the j-th group of reference patterns, $\rho_{K_1(j)}^{(v)}(j)$ is the distance between the v-th training sample of the K-th recognition pattern and its nearest reference pattern in the j-th group of reference patterns $(K_1(j)=\arg\min_{i\neq K} \rho_i^{(v)}(j))$, besides the reference pattern of the K-th recognition pattern $e_K^{(0)}(j)$. This sample is involved in the formation of a reference pattern of the K-th recognition pattern only in the $j_K^{(v)}$ group such that

$$j_K^{(v)} = \arg\max_{j=\overline{1,J}} \mathcal{L}_K^{(v)}(j) \qquad (8)$$

Since the quantity of $\mathcal{L}_K^{(v)}(j_K^{(v)})$ represents the contrast range of the v-th training sample representation within some favourable group of reference patterns, the quality of training of the K-th pattern formed by V training samples can be evaluated from the training quality functional

$$\Lambda_K^{(1)} = \sum_{v=1}^{V} \mathcal{L}_K^{(v)}(j_K^{(v)}) \qquad (9)$$

To improve the reference patterns of the K-th recognition pattern, a multiple iteration procedure can be initiated. The r-th iteration will give the reference patterns of the K-th recognition pattern $\tilde{e}_K^{(r)}(j)$ $(j=\overline{1,J})$ by averaging the training samples of the K-th recognition pattern within the corresponding groups of reference patterns, in which they are placed according to eq.(8). The iteration procedure stops when

$$\Lambda_K^{(r)} \leq \Lambda_K^{(r-1)} \qquad (10)$$

where $\Lambda_K^{(r)}$ and $\Lambda_K^{(r-1)}$ are the training quality functionals for the r-th and $(r-1)$-th iteration, respectively. The procedure also stops when the number of iterations reaches its limiting value R_{max}:

$$r = R_{max} \qquad (11)$$

It should be noted that the training quality functional $\Lambda_K^{(r)}$ used in formation of reference speech patterns enables one to correct the vocabulary taking into account the training results, and replace "difficult" words in a vocabulary prior to recognition. The experimental check-up of the above method verified its potential usefulness in increasing the recognition rate. The experiments were carried out using two groups of reference patterns ($J=2$). Significant increase in the recognition rate has been achieved, which amounted to 99.5%, 99% and 98% for vocabularies containing 50, 100 and 250 words, respectively.

CORRECTION OF REFERENCE PATTERNS

As it was mentioned earlier, the present methods of reference pattern formation are not fully adequate since they do not provide optimum separation of speech pattern groups. We propose a new approach to reference pattern formation which cosists in automatic correction of reference patterns formed at the training stage in the process of recognition. Suppose that the training procedure resulted in the formation of reference patterns for N recognition patterns. Let us represent the recognition patterns together with their reference patterns by points in R-dimensional space. To calculate the distance between the recognition patterns and the reference patterns we employ the Chebyshev's metric:

$$\rho_i = \sum_{r=1}^{R} |X_r - \tilde{X}_{ir}| \qquad (12)$$

Here ρ_i is the distance between the speech pattern being recognized $\{X_r\}_{r=\overline{1,R}}$ and the reference pattern $\{\tilde{x}_{ir}\}_{r=\overline{1,R}}$ of the i-th recognition pattern. Let ρ_{i_1} be the distance between the speech pattern being recognized and its nearest reference pattern. Consider the following quantities:

$$\Delta_i = \rho_i - \rho_{i_1}, \; i \neq i_1, \; i=\overline{1,N} \qquad (13)$$

A decision which identifies an input speech pattern as belonging to a certain reference pattern will be the most reliable when the speech pattern is closest to the reference pattern of its group and farthest from reference patterns of those groups, to which it does not belong. Therefore, an optimum placement of reference speech patterns will be such that Δ_i are at their maxima. We propose an heuristic solution of this problem. The procedure consists in a successive correction using the results of reference speech pattern $\{\tilde{x}_{ir}\}_{r=\overline{1,R}}$ recognition. For each recognized speech pattern we select from Δ_i values those values Δ_K which satisfy the relation

$$\Delta^{(1)} < \Delta_K < \Delta^{(2)}, \; \Delta^{(1)} > 0, \; \Delta^{(2)} > 0 \qquad (14)$$

The correction of reference speech patterns will not be made if among Δ_i values there are no values which satisfy condition (14). Otherwise, the correction is applied both to the reference pattern of the i_1 recognition pattern, and the reference patterns of any K recognition patterns for which relation (14) is satisfied. Let $\{\tilde{x}_{ir}(0)\}_{r=\overline{1,R}}$ be the reference pattern of the i-th recognition pattern obtained as a direct result of the training procedure, and let $\{\tilde{x}_{ir}(v)\}_{r=\overline{1,R}}$ be the reference pattern of the i-th recognition pattern obtained as a result of the v-th correction of the initial reference pattern $\{\tilde{x}_{ir}(0)\}_{r=\overline{1,R}}$. Then, if a decision to correct reference patterns is made in the process of the next speech pattern $\{x_r\}_{r=\overline{1,R}}$ recognition, the correction will be carried out as follows. For a reference speech pattern recognized as i_1, the co-ordinates of its new corrected reference pattern $\{\tilde{x}_{i_1r}(v+1)\}_{r=\overline{1,R}}$ are obtained from the formula

$$\tilde{x}_{i_1r}(v+1) = \begin{cases} \tilde{x}_{i_1r}(v) + \Delta\tilde{x}_{i_1r}(v), & \text{if } x_r \geq \tilde{x}_{i_1r}(v) \\ \tilde{x}_{i_1r}(v) - \Delta\tilde{x}_{i_1r}(v), & \text{if } x_r < \tilde{x}_{i_1r}(v) \end{cases} \quad r=\overline{1,R} \qquad (15)$$

The $\Delta\tilde{x}_{i_1r}(v)$ values are calculated from the formula

$$\Delta\tilde{x}_{i_1r}(v) = \beta |x_r - \tilde{x}_{i_1r}(v)|, \; 0<\beta<1, \; r=\overline{1,R} \qquad (16)$$

The correction of co-ordinates for the reference patterns of those K recognition patterns for which relation (14) is satisfied, is made using the formula

$$\tilde{x}_{Kr}(v+1) = \begin{cases} \tilde{x}_{Kr}(v) - \Delta\tilde{x}_{Kr}(v), & \text{if } x_r \geq \tilde{x}_{Kr}(v) \\ \tilde{x}_{Kr}(v) + \Delta\tilde{x}_{Kr}(v), & \text{if } x_r < \tilde{x}_{Kr}(v) \end{cases} \quad r=\overline{1,R} \qquad (17)$$

The $\Delta\tilde{x}_{Kr}(v)$ values are calculated from the formula

$$\Delta\tilde{x}_{Kr}(v) = \gamma |x_r - \tilde{x}_{Kr}(v)|, \; \gamma \geq 0, \; r=\overline{1,R} \qquad (18)$$

The principal advantage of the above method is that training (reference pattern correction) and recognition procedures are performed simultaneously, and it does not require the use of a large number of training samples, which would be the case if these procedures were separated.

SUMMARY

It is shown that a speech recognition system with a large vocabulary and high recognition rate can be developed on the basis of Walsh functions. For vocabularies containing 50, 100 and 250 words, recognition rates up to 99.5%, 99% and 98%, respectively, have been obtained.
The proposed method of reference pattern correction in the process of recognition is expected to further increase the speech pattern recognition rate.

REFERENCES

/1/ N.Ahmed, K.R.Rao, "Orthogonal Transforms for Digital Signal Processing", Berlin, Heidelberg, New York, 1975.

/2/ P.A.Lee, I.Seymour, Evaluation of the Fast Walsh-Hadamard Transform for Speech Recognition by an 8-bit Microprocessor, Acoustica, vol.59, pp.274-278, 1986.