STANFORD ARTIFICIAL INTELLIGENCE PROJECT
MEMO AIM-175

STAN-CS-307

HADAMARD TRANSFORM FOR SPEECH WAVE ANALYSIS

BY

HOZUMI TANAKA

SUPPORTED BY

ADVANCED RESEARCH PROJECTS AGENCY
ARPA ORDER NO. 457
AUGUST 1972

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

STANFORD ARTIFICIAL INTELLIGENCE PROJECT
MEMO AIM-175

AUGUST 1972

COMPUTER SCIENCE DEPARTMENT
REPORT NO. CS-307

HADAMARD TRANSFORM FOR SPEECH WAVE ANALYSIS

by

Hozumi Tanaka

Abstract: Two methods of speech wave analysis using the Hadamard transform are discussed. The first method is a direct application of the Hadamard transform for speech waves. The reason this method yields poor results is discussed. The second method is the application of the Hadamard transform to a log-magnitude frequency spectrum. After the application of the Fourier transform the Hadamard transform is applied to detect a pitch period or to get a smoothed spectrum. This method shows some positive aspects of the Hadamard transform for the analysis of a speech wave with regard to the reduction of processing time required for smoothing, but at the cost of precision. A formant tracking program for voiced speech is implemented by using this method and an edge following technique used in scene analysis.

This research was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defence (SD-183)

Acknowledgements

The author would like to express deepest thanks to Dr. A.L. Samuel and Mr. J.K. Sicerz. Dr. A.L. Samuel suggested this research to the author. Mr. K.L. Siberz developed a speech wave analysis system which has been used for this research as a basic tool. The author is grateful to Dr. R. Thosar for his many valuable suggestions to implement a formant tracking program. The author takes this opportunity of expressing his appreciation to Prof. John McCarthy for the opportunity to work at Artificial Intelligence Project as a visiting scholar from 1971 August to 1972 August. I would also like to thank Mr. and Mrs. Paul and Mr. J.E. Gips for their help in preparing this report.

Table of contents

1 Introduction.

Recently people in various fields have paid much attention to the Hadamard transform and have obtained results from its application in such fields as filter design, voice analyzer/synthesizer and multiplexer equipment [1]. The Hadamard (or discrete Walsh) transform is one of the orthogonal transforms using discrete Walsh functions and has a fast algorithm similar to the Fourier transform [2],[3].

There are many reasons why the Hadamard transform is attractive. Two major reasons are as follows. First, the Fast Hadamard Transform algorithm -FHT- uses only add / subtract operation. Multiplication is not neccessary for the FHT. This makes the calculation of the FHT extremely simple and faster than the Fast Fourier Transform - FFT. In the Fourier transform case one needs multiplication for the sine-cosine coefficients, sometimes even with irrational numbers. The FHT offers quite a simple and an appropriate algorithm when using a digital computer.

Secondly, the discrete Walsh functions give us a general bask for signal analysis, namely the concept of sequency rather than that of frequency. The sequency of discrete Walsh functions is defined by one half of the average number of zero crossings per second. This concept enables us to replace the concept of frequency of the sine-cosine functions.

Because of this feature of the Hadamard transform one may well think of the possibility that all problems which have been solved using the Fourier transform might be re-interpreted by the Hadamard transform. Furthermore, one might hope for some interesting new discoveries since the Hadamard transform might reveal some new aspect of the problem concerned.

From this optimistic standpoint, the author has attempted an analysis of the speech wave using the Hadamard transform. Similar attemots have been made in the past [4], and they have suggested some possibilites about the application of the Hadamard transform to the speech wave by showing some correspondence between the frequency spectrum and the sequency spectrum. This report will show two methods of speech wav3 analysis using the Hadamard transform, the direct and the indirect methods. These two methods show both the advantages and disadvantages of the Hadamard transform for speech wave analysis.

Section 2 will explain the direct method by an application of the Hadamard transform to a speech wave. This method gives a poor result due to the strong shift sensitivity of the Hadamard sequency spectrum. Some shift invariant terms of the sequency power spectrum are known but they are complicated to calculate or too simple to provide enough information. A few experimental results are shown in this section to demonstrate these facts.

Section 3 will explain the indirect method named the "hapstrum" technique. The hapstrum technique is a similar technique to the so called cepstrum technique [5] except that the FHT is applied to the log-magnitude frequency spectrum. This technique is indirect in the sense that at first the FFT (not FHT) is applied to a short span of a speech wave and then the FHT is used to detect the pitch period or to get a smoothed spectrum. This technique shows some positive aspect of the Hadamard transform for the analysis of a speech wave with regard to smoothing of a spectrum. Some experimental results will demonstrate this.

A formant tracking program has been implemented using the technique of an edge follower in scene analysis combined with the hapstrum technique. However, such an approach always contains a pitfall, namely the problem of wrong way entrance. This will be discussed in section 3.3.

Finally, in section 4 a tentative evaluation will be made of the Hadamard transform for analyzing speech waves.

2 Direct application of the Hadamard transform to speech wave analysis.

In this section the Hadamard transform will be directly applied to a speech wave to get the sequency power spectrum. The existence of some correspondence between frequency spectrum and sequency spectrum has been reported on [4]. As a given vocalic sound can be characterized by the location of its first three formant frequencies, it is worth investigating the existence of formant "sequencies" in the Hadamard sequency spectrum instead of formant frequencies. A few experiments will demonstrate poor results and the reason will be discussed.

2.1 Definition of sequence and sequency.

The definition of sequence was introduced by H. F. Harmuth [3] and it gives a new basis from which to investigate the characteristic of signals. A sequence number of a walsh function is defined by the number of sign changes per unit time. Let $N = 2^n$ consecutive real numbers $a(j)$, $0 \le j < N$, be represented by a 1 x N matrix $[a(j)]$. The Hadamard transform of $[a(j)]$ is

$$[A(k)] = (1/N)[a(j)]H(n) \qquad (1)$$

where the N x N Hadamard matrix $H(n)$ is defined recursively in the equation (2).

$$H(n+1) = \begin{Vmatrix} H(n) & H(n) \\ H(n) & -H(n) \end{Vmatrix} \qquad (2)$$

$$H(0) = \|1\|$$

Each column of $H(n)$ represents one of discrete Walsh functions [7]. The examples of $H(3)$ and $H(4)$ are shown in the Fig. 2.1.

The sequency is defined by the average number of zero crossings per unit time divided by 2. Let $p(k)$ be the number of sign changes (zero related to frequence or sequency it is desirable to calculate defined by eq. (3)

$$q(k) = [(p(k)+1)/2] \qquad (3)$$

where $[x]$ represents the largest integer which does not exceed x (see H(3) of the Fig. 2.1). It is known that $p(k)$ takes on all values between zero and $N-1$ and $q(k)$ takes all values between zero and $N/2$.

$$
H(3) = \begin{Vmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{Vmatrix}
$$

$$
\begin{array}{cccccccc}
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow
\end{array}
$$

--- sequence of each column: p

-- sequency of each column: q.

$$
H(4) = \begin{Vmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\
 & & & & & & & & & & & & & & & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1
\end{Vmatrix}
$$

Fig. 2.1 Examples of the Hadamard matrix.

Let us introduce two notations, $A(c,p(k))$ and $A(s,p(k))$ for $A(k)$.

$$A(k) = \begin{cases} A(c,q(k)) & \text{if } p(k) \text{ is even} \\ A(s,q(k)) & \text{if } p(k) \text{ is odd.} \end{cases}$$

In analogy of frequency power spectrum, sequency power spectrum is defined as follows.

$$\begin{cases} A^2(c,0) & \\ A^2(c,q) + A^2(s,q) & 0 < q < N/2 \\ A^2(s,N/2) & \end{cases} \tag{4}$$

The Parseval's relation is preserved on the coefficients $A(k)$ and $a(k)$.

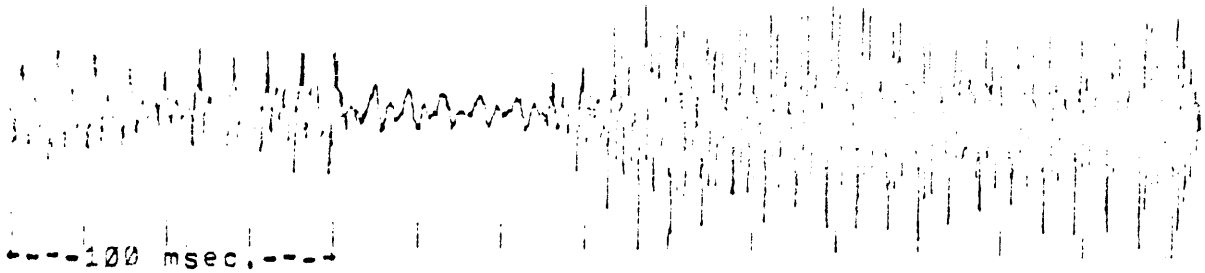
$$(1/N)\sum_{k=0}^{N-1} a^2(k) = A^2(c,0) + \sum_{q=1}^{(N/2)-1} [\, A^2(c,q) + A^2(s,q)] + A^2(s,N/2) \tag{5}$$

2.2 Strong shift-sensitivity of the Hadamard sequency spectrum.

It is interesting to investigate the formant structure in the sequency spectrum of a speech wave. The most convenient way is to change many consecutive spectra into a visual form, that is the sonogram of sequencies. A short time span (12.8 ms.) of a digitized speech wave (sample rate = 20000 Hz.) is directly transformed into sequency spectrum. Then the log_magnitude of this spectrum is taken. Many short time sequency spectra are calculated in this way, are accumulated, and eventually output to a video screen.

Experimental results are shown in Fig. 2.2. The upper part shows a speech wave to be analyzed, the middle part a sonogram of frequency spectra of this speech wave and the lower part a sonogram of sequency spectra. It is easy to see that the sonogram of sequency spectra (the lowest one) is rougher than that of the frequencies (the middle one). The formant sequency structure is not clear and it appears to be very difficult to build a speech wave analysis system based on the extraction of formant components using the Hadamard sequency spectrum.

The reason why the sonogram of sequency spectra becomes so rough and irregular is made clear by the following experiment. The Hadamard sequency spectrum is calculated for a fixed time span (12.8 msec. long) of a speech wave. The time span is shifted right by 100 microseconds for each successive calculation of the sequency spectrum. In other words, calculation of a sequency spectrum is made each 100 microsecond time-shift. A frequency spectrum of the Fourier transform is calculated in the same way to make comparision with sequency spectrum. The results are shown in Fig. 2.3.



Speech wave

Frequency spectrum

Sequency spectrum

Fig. 2.2 Sonogram of sequency and frequency spectra.

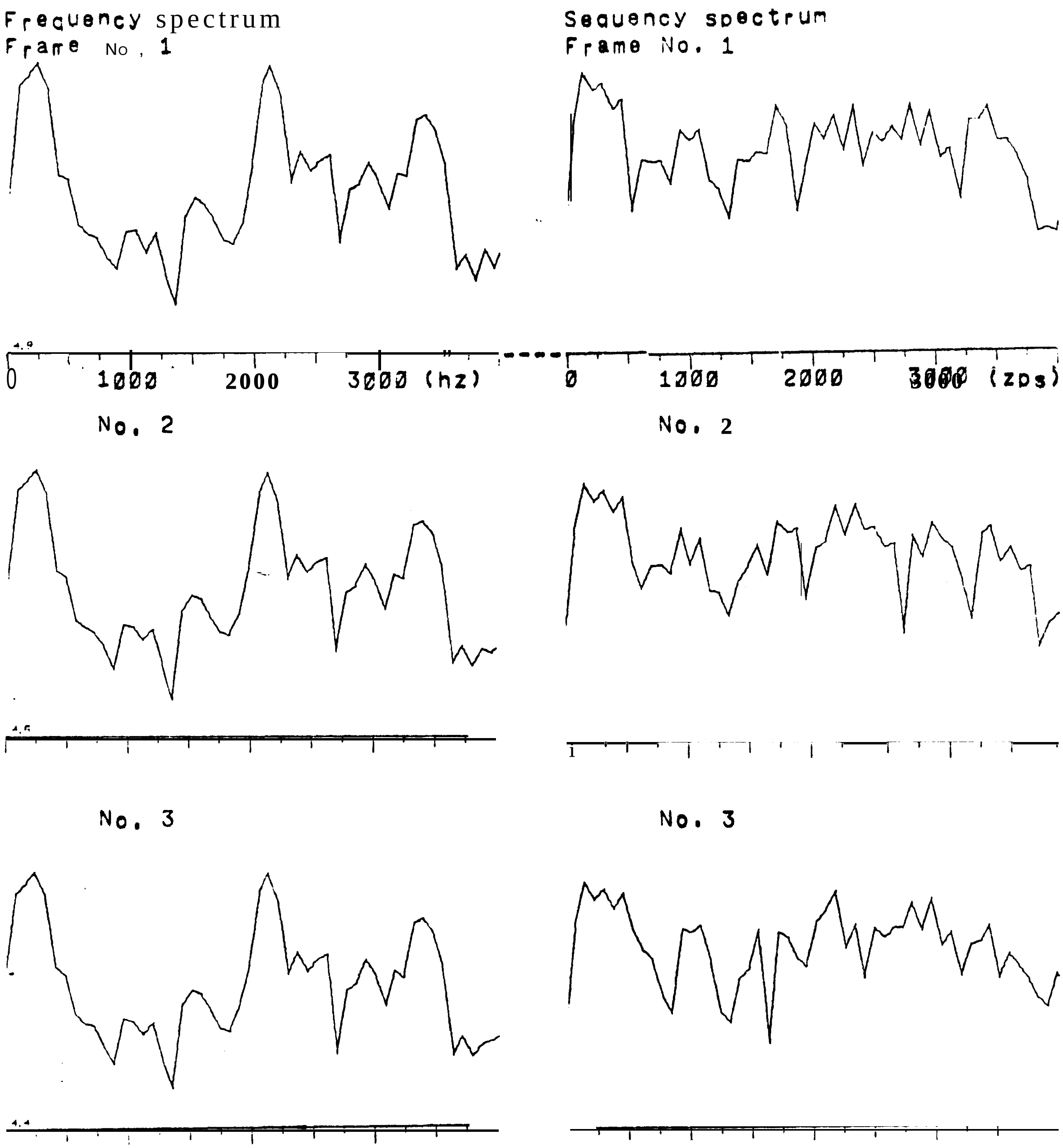


Fig. 2,3 Strong shift-sensitivity of the Hadamard sequency spectrum, Each frame is calculated each 100 microsecond time-shift.

From Fig. 2.3 we can easily understand that although the time-shift is limited to this small value, the shape of consecutive sequency spectra changes rapidly. The location of a peak which appears to represent a formant component changes drastically in the next sequency spectrum. One cannot expect these rapid changes from observation of the original speech wave since the speech wave does not appreciably change its shape during 100 microseconds. In contrast, in the Fourier case, a frequency spectrum does not change its shape so much during 100 microseconds. This strbns shift-sensitivity of the Hadamard sequency spectrum causes the irregularity or rough pattern of a sequency sonogram and makes impossible the application of the pitch-synchronous method.

The strong time-shift sensitivity of a sequency spectrum also can be explained theoretically. Pichler [6] shows the Hadamard sequency spectrum is invariant under the dyadic time-shift:

[b(j)] is obtained by the dyadic time-shift t

$$[b(j)] = [a(j \oplus t)]$$

where $j \oplus t$ stands for component-wise modulo two addition (no carry) for the binary representation of j and t. Pichler's result is written as follows,

$$B^2(c,q) + B^2(s,q) = A^2(c,q) + A^2(s,q) \tag{6}$$

Unfortunately the Hadamard sequency spectrum is not invariant under circular time-shift of the input [a(j)]. If [a(j)] is shifted by t circularly forming [c(j)] we obtain:

$$[c(j)] = [a((j + t))]$$

where ((j + t)) is the principal value of j + t modulo N. In general

$$C^2(c,q) + C^2(s,q) \neq A^2(c,q) + A^2(s,q) \tag{7}$$

The experiment shown in Fig. 2.3 is not the case of circular time-shift but one can easily understand that the relation of eq (7) causes the strong shift sensitivity in the Hadamard sequency spectrum. Note that in contrast to the Hadamard sequency spectrum a frequency spectrum of the discrete Fourier transform is invariant under circular time-shift since absolute value of a shift operator is one.

2.3 Difficulties in calculating shift invariants for the Hadamard transform.

Some attempts have been made to define circular time-shift invariants for the Hadamard transform. Ohnsorg has defined a complete set of circular time-shift invariants of the Hadamard transform and also has shown intermediate forms which are invariant to both circular time-shift and dyadic time-shift. For more detailed derivation of a complete set of circular time-shift invariants and its intermediate forms see [7].

As a first step, consider intermediate forms, a set {P(k)} which is a sum of groups of components in [A(k)] squared such that

$$P^2(0) = A^2(0)$$
$$P^2(1) = A^2(1)$$
$$P^2(2) = A^2(2) + A^2(3) \qquad (8)$$
$$\cdots\cdots\cdots$$

In general

$$P^2(m) = \sum_k A^2(k)$$

where $2^{m-1} \le k < 2^m$ for $1 \le m \le n$.

Examples of calculations of a set {P} for various input waves are shown in Fig. 2.4. In the figure the short time span of the speech wave for the Hadamard transform is fixed to 12.8 msec. Each component of a set {P} is shown as a function of time in the Fig. 2.4. Overlap of the time span for the next Hadamard transform is 6.4 msec. The case of a sinusoidal wave indicates the filtering characteristic of a set {P} because the position of each peak moves to the left as k increases in P(k). In other words, the smaller the value of k in eq (8), the more likely it is that the component P(k) will pass the higher frequency component since frequency increases with time passing in the original input wave. However, as the band of each filter is determined by the number N, which is the dimension of an array [A(k)], we lose flexibility. Although the calculation of a set {P} from N components of [A(k)] is straightforward, we can get only $1 + n (= \log_2 N)$ components of P. For instance, if N = 256 one can get only 9 components of P and one of them is d.c. component. This means a great deal of information reduction is made and it is doubtful if a set {P} contains enough information to perform speech wave analysis.

Ohnsorg has defined another complete set of the Hadamard transform which has exactly (N/2) + 1 invariants for a circular time-shift. (The discrete Fourier transform -DFT- gives a (N/2) + 1 point spectrum.) However it is not a straightforward way to calculate the invariants since it includes many matrix multiplications. According to [7] if we let {J} be a quadratic invariant set of the Hadamard transform, then

in the case when N = 8

$$J^2(0) = A^2(0)$$
$$J^2(1) = A^2(1)$$
$$J^2(2) = A^2(2) + A^2(3) \qquad (9)$$
$$J^2(3) = A^2(4) + A^2(6) - A(4)A(7) + A(5)A(6)$$
$$J^2(4) = A^2(5) + A^2(7) + A(4)A(7) - A(5)A(6)$$

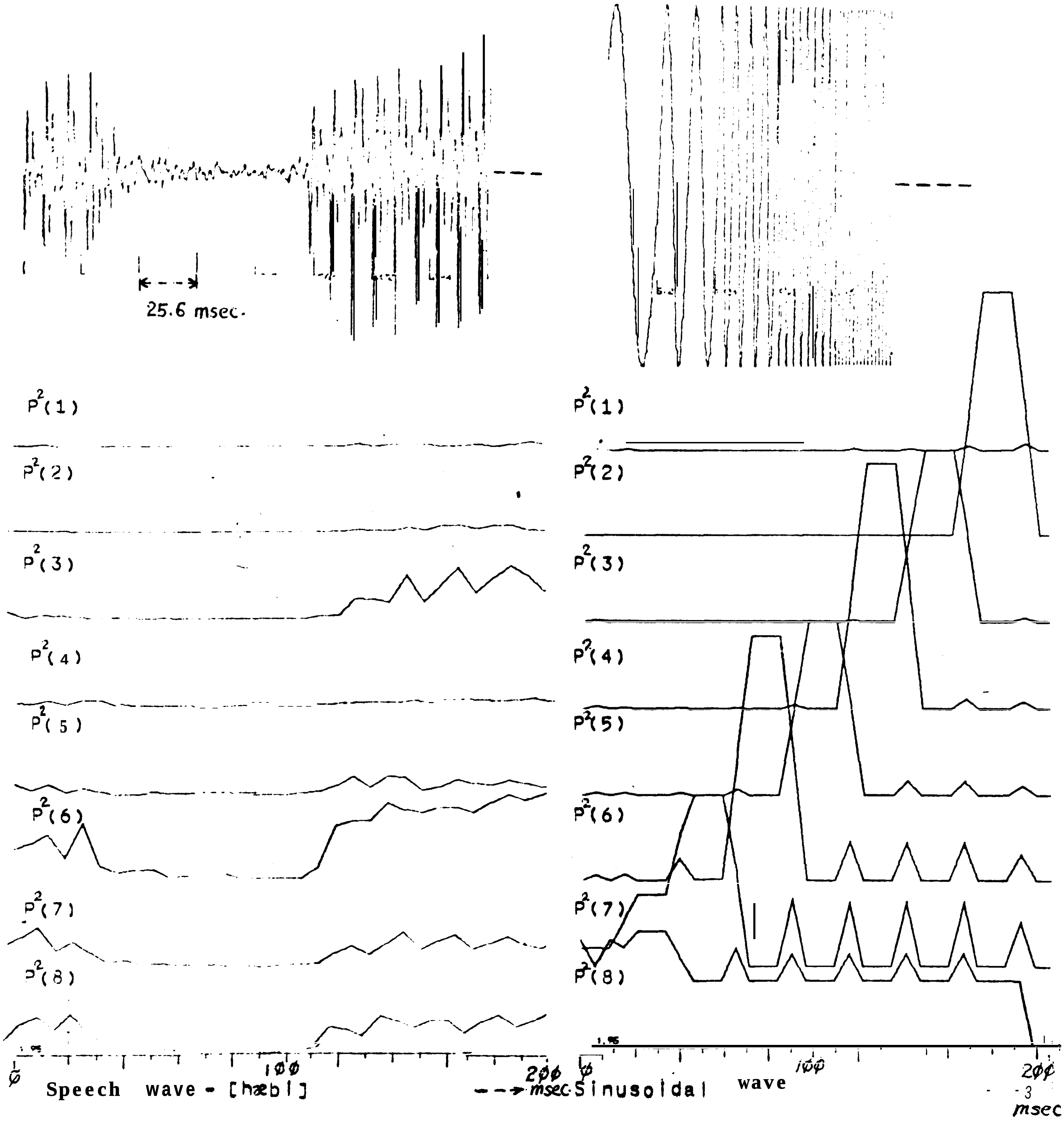


Fig. 2.4 Calculation of eq(8) for various inputwaves.

Although there is no explanation about how these terms (J) are related to frequence or sequU
Ohnsorg's [7] invariants. As Ohnsorg suggests that the prominant energylines of the discrete Fourierspectrum tend to be exaggerated in the quadratic spectrum (J).

N. Ahmed et al [8] found an efficient algorithm to calculate these terms, however multiplication by an irrational number is included in the algorithm and it is more complicated than that of Hadamard transform.

3 Hapstrum technique.

In this section the "hapstrum" technique is introduced. The hapstrum technique is a similar technique to the cepstrum technique except that the inverse fast Hadamard transform -IFHT- is applied to the log-magnitude frequency spectrum and the output is called "hapstrum." This technique is indirect in the sense that at first the FFT (not FHT) is applied to a short time span of a speech wave to obtain the spectrum and then the FHT is used to extract pitch period or to get smoothed spectrum. The strong time-shift sensitivity of the Hadamard transform is removed by the first application of Fourier transform to speech waves.

This technique illustrates a positive aspect of the Hadamard transform for the analysis of a speech wave, especially with regard to the smoothing of a spectrum. A formant tracking program has been implemented using this technique.

3.1 Outline.

To show both the advantages and disadvantages of the hapstrum technique we will depict the outline of both the cepstrum and the hapstrum techniques. Although there is more than one definition of the cepstrum technique we give a typical application in the upper part of Fig 3.1. The hapstrum technique is shown in the lower part of Fig. 3.1.

From Fig. 3.1, one can easily understand the difference between both techniques. The frequency spectrum of a short time span of a speech wave filtered by a Hamming window is obtained by the discrete Fourier transform -DFT. Then the log-magnitude of this spectrum is taken. After the processing, in the case of the cepstrum technique the inverse discrete Fourier transform -IDFT- and DFT are applied to get pitch period and smoothed spectrum. On the other hand, in the case of the hapstrum technique the IDFT and DFT are replaced by the IFHT and FHT, respectively. A hapstrum, which is ordered in sequence (not sequency), is obtained by the IFHT of a log-magnitude spectrum. From the replacements one gets the advantage of the fast calculation of the Hadamard transform. Due to the elimination of linear filtering, computing cost is even further reduced by the method.

Let us note that in the cepstrum case after the application of the inverse discrete Fourier transform we need low-pass filtering of the log-magnitude of the discrete Fourier transform. By means of low-pass filtering a smoothed spectrum is obtained due to the elimination of the fine structure of the spectrum. This is accomplished by multiplying the cepstrum by a low-pass filter function.

In contrast to the cepstrum technique, the hapstrum technique uses an ideal filter as a low-pass filter in the sequency domain of the hapstrum. Therefore one needs no multiplication to cut higher

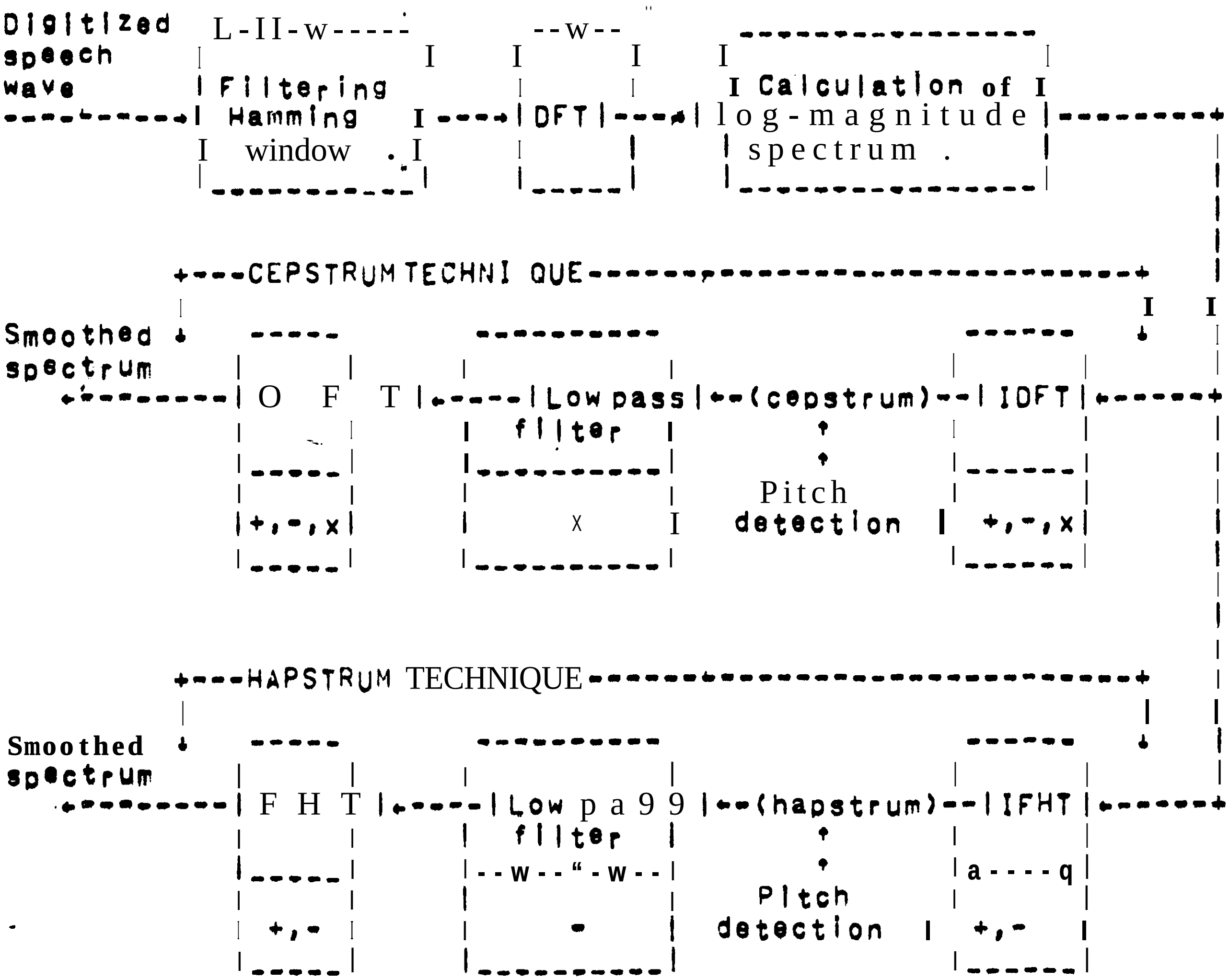


Fig. 3.1 The outline of the cepstrum technique and the hapstrum technique.

sequence components, The higher sequence components are simply made zero. This also reduces computing cost(the symbols +/- and x in the figure indicate the necessity of add/subtract operations or multiplications).

From the author's experience the calculation of the FHT is ten times as fast as that of the FFT. This suggests that by using the hapstrum technique we can make the calculation of spectrum smoothing at most three times as fast as that of spectrum smoothing using the cepstrum technique.

However, we should be aware that smoothing by the cepstrum gives us a better approximation far an original log-magnitude spectrum in the sense of least-square error criterion and that smoothing by the hapstrum degrades resolution of peak position of log- magnitude spectrum. The theoretical reason for this will be discussed in section 3.3.

3.2 Pitch detection.

To extract a pitch period we have to take a sufficient time-span of a speech wave to calculate a lag-magnitude spectrum, namely long enough to Include at least two glottal pulses.

In our experiments the duration I9 taken to be 25.6 msec corresponding to 512 samples of a digitized speech wave since the sampling rate of a speech wave is 20000 Hz.

Flg. 3.2-a shows a series of cepstrum plots. A series Of cepstrum are calculated for each consecutive segment of speech wave one half of which overlaps the previous segment. In the case of the cepstrum, to get a higher resolution 512 zeros are added to the next 512 samples of a digitized speech wave. Thls means the IDFT and DFT are calculated on 1024 points.

Fig. 3.2-b shows a series of hapstrum plots. The hapstrum is calculated Under the same condition as the cepstrum of Fig. 9. To calculate a hapstrum we do not add Zero to the next 512 samples of a speech wave, since one cannot get higher resolusion of the hapstrum by adding zeros (see 3.3 in this section). If 512 zeros are added to the next 512 samples of a digitized speech wave on8 will get a hapstrum such that the component of the sequence (not sequency) 2i and 2i + 1 becomes the same value, where i is a positive integer. In other words a hapstrum of a speech wave segment with added zeros is easily calculated from one wlthout added zeros. This special feature of the Hadamard transform Is utilized by the smoothing of the log-magnitude spectrum in the next section. The proof is shbwn in the APPENDIX in more a generalized form.

Comparing fig. 3.2-a with Fig. 3.2-b, we observe that ln the cepstrum a sharp Peak appears at approximately 4.5 msec but in the case of the

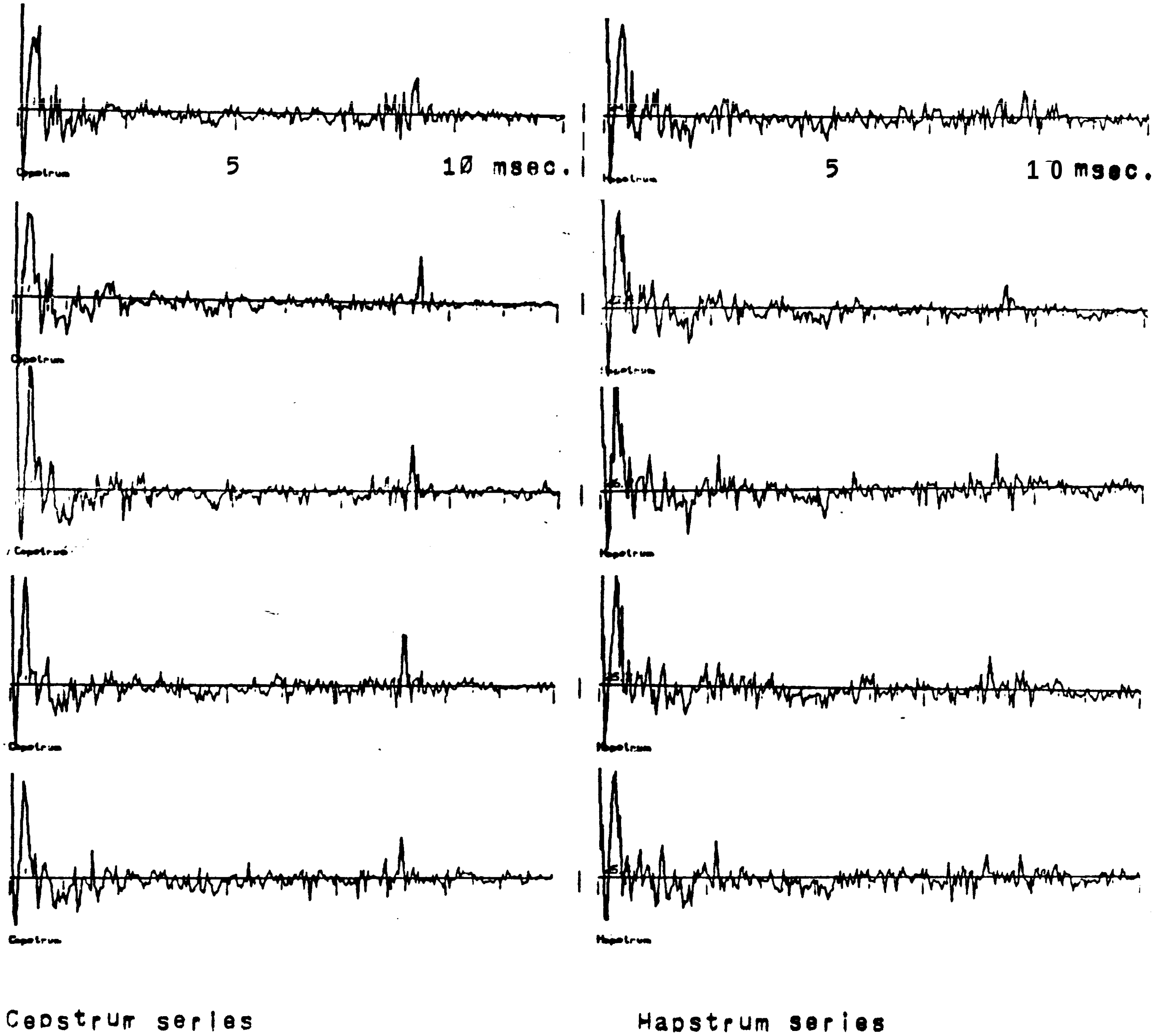


Fig. 3.2 An example of cepstrum series and the hapstrum series.

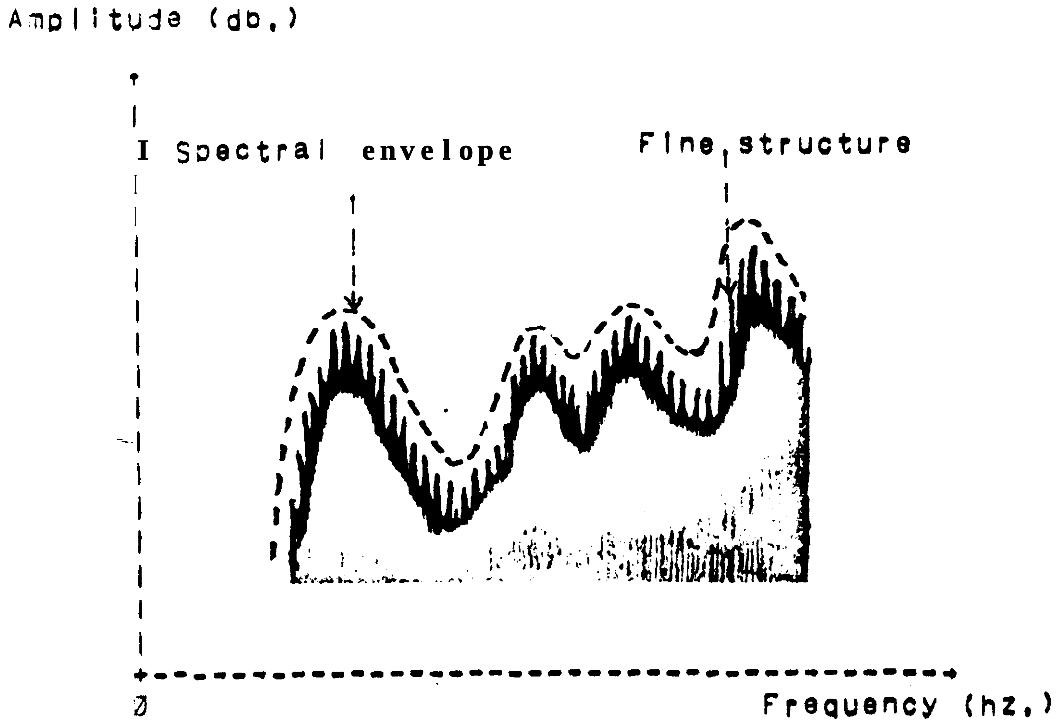


Fig.3.3 Spectral envelope and spectral fine structure of log-magnitude spectrum of a speech wave.

Consider the meanings of filtering by an ideal filter in the sequence domain. Let an array [a(j)] of dimension N (= 2^n) be a digitized signal in which all components such that N/2 ≤ j < N are set to zero. By the application of the FHT including sequence ordering the array [a(j)] is transformed into an array [B(k)] such that each adjacent component becomes the same, namely:

$$
\begin{aligned}
&B(0) = B(1) \\
&B(2) = B(3) \\
&\cdots\cdots\cdots\cdots \\
&B(N-2) = B(N-1)
\end{aligned}
\tag{11}
$$

Furthermore, when all components such that $N/(2^2) \le j < N$ are set to zero the array [a(j)] is transformed into such an array [B(k)] by the application of the FHT including sequence ordering

$$
\begin{aligned}
&B(0) = B(1) = B(2) = B(3) \\
&B(4) = B(5) = B(6) = B(7) \\
&\cdots\cdots\cdots\cdots \\
&B(N-4) = B(N-3) = B(N-2) = B(N-1)
\end{aligned}
\tag{12}
$$

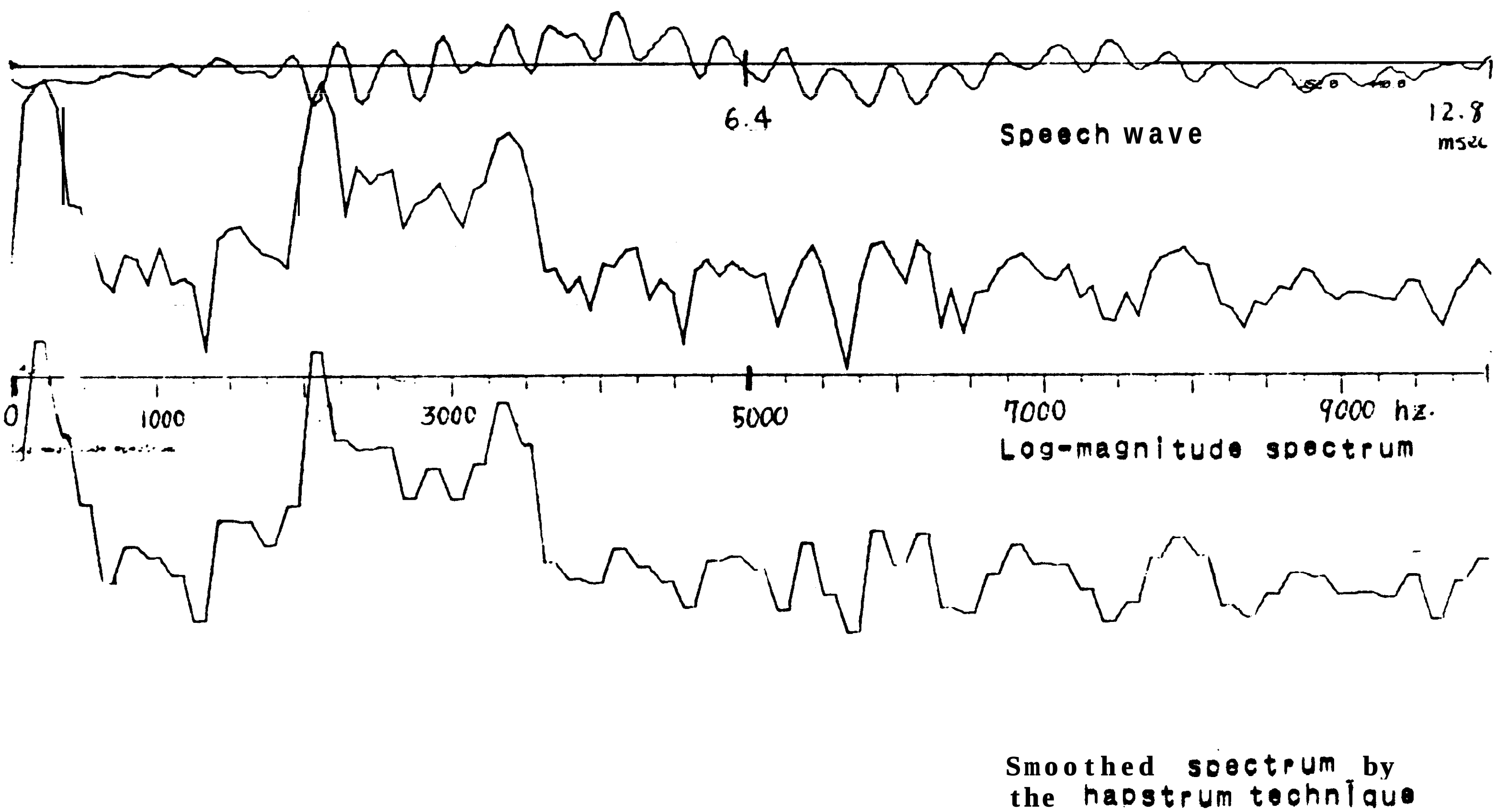
Eq (11) and (12) are generalized in (A) and (B) of APPENDIX.

Both equations suggest that if [B(k)] is plotted as a function of array index k the curve becomes flat as the value of each adjacent component is the same. Because of this flattening effect it degrades resolution of peak positions in the array [B(k)]. To demonstrate this an example is shown in Fig.3.4. A segment of a speech wave is shown and is analyzed by the hapstrum technique. The two lower curves represent the log-magnitude spectrum and the smoothed result by the hapstrum technique.

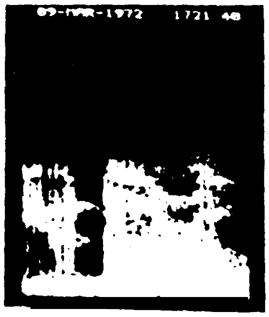
The smoothed log-magnitude spectrum in Fig. 3.4 demonstrates the smoothing effect stated before by eq (11). Many sharp maxima and minima caused by glottal pulses (or pitch frequency) in the original log-magnitude spectrum are diminished in the smoothed spectrum. From the author's experience the number of peaks is decreased to one half that of the original.

The important question is whether or not the prominent peaks caused by resonance of a vocal tract are preserved by the smoothing. From the example shown in Fig. 3.4 we can see that the hapstrum smoothing technique gives good smoothing with regard to preserving the first three formants.

Fig. 3.5 gives another example which suggests the formant components are preserved after the smoothing by the hapstrum technique. The upper is a sonogram of spectra without smoothing and the lower is a sonogram of smoothed results using the hapstrum technique.



Fig, 3.4 Speechwave, the log-magnitude spectrum and the smoothed spectrum by the hapstrum technique, a)



b)

Fig. 3,5 Sonograms of log-magnitude spectra and their smoothed specra. The upper is a sonogram of log-magnitude spectra and the lower is that of smoothed spectra.

3.4 A formant tracking program as an application of the hapstrum technique.

A formant tracking program has been implemented using the hapstrum technique and an edge follower technique as used in scene analysis [11]. In principle, the formant tracking program presented here accepts any kind of smoothing technique such as cepstrum or inverse filtering [12].

Edge followers were first implemented to recognize objects in a scene. An edge follower detects a position where sharp change of contrast occurs and follows it successively. A sonogram is just such a scene with formant trajectories represented as dark stripes. By detecting dark stripes we find the locations of peaks in a spectrum since a sonogram is represented as a sequence of spectra.

There are many difficulties in implementing a formant tracking program based on an edge follower. One problem is that a formant trajectory is not a straightline, but is curved. Some of the edge followers have treated objects composed only of straight lines, such as cubes. This limitation can be of use to an edge follower. For instance we can prevent the following of the wrong path by using the criterion of curvature. We also can forecast the existence of edge, which is hard to detect because of noise, by using straight line interpolation methods. As the production of a speech wave is a dynamic and stochastic process, the human speech wave contains much noise.

A second problem is that it is very difficult to decide a formant frequency from local information. A wide range of overlap exists between the region of the first formant frequency and that of the second, also between the second formant frequency and the third. In the case of a male voice, the first formant frequency ranges from 220 hz to 920 hz., the second from 550 hz. to 2700 hz, and the third from 1120 hz. to 3020 hz.

A third problem is that if we see a sonogram in a microscopic way there exist too many peaks to discriminate the formant components. It is desirable to have a technique to eliminate trivial peaks while preserving prominent peaks caused by the first three formants. Cepstrum is such a technique. Rabiner and Schafer [10] have implemented a formant tracking program based on the cepstrum technique. As their method makes frame-by-frame decisions for the first three formant frequencies, they use only local information in a sonogram. It is desirable to utilize more global information.

Markel [12] has developed a very good technique for getting a smoothed spectrum based on the idea of linear prediction method. He calls it inverse filtering and has developed a formant tracking program [13] which uses information from the previous frame when it is difficult to determine the first three formant frequencies.

The formant tracking program explained here follows Markel's approach but with a backtracking mechanism to recover if a wrong path is followed. If decisions are made frame by frame there is no wrong way entrance problem. Even if we make a wrong decision in a frame, the effect does not propagate to the next. However, if we use the information from just the previous frame the effect of a wrong decision will propagate. To cope with this situation it is necessary to have a recovery technique which utilizes more global information.

3.4.1 Logical structure of a formant tracking program.

Our formant tracking program is composed of four modules named PEAK DETECTOR, CANDIDATE SELECTOR, TRACKER, and RECOVERY. General flow of the program is shown in figure 3.6.

A. PEAK DETECTOR.

PEAK DETECTOR accepts a digitized speech wave of a vocalic sound, calculates a smoothed spectrum by using the hapstrum technique, and determines peaks. It should be noted that the hapstrum technique is used to decrease the processing time required for smoothing. It can easily be replaced by another technique such as inverse filtering or the cepstrum technique.

B. CANDIDATE SELECTOR.

For each region of the first three formant frequencies, CANDIDATE-SELECTOR selects at most three candidates from many peaks detected by PEAK DETECTOR and orders them by amplitude of peaks. The third candidate whose amplitude is 7.5 db less than that of the second candidate is removed by the ordering process. These candidates selected are accumulated and are used by TRACKER and RECOVERY. This routine reduces the search space.

C. TRACKER.

TRACKER takes the results from CANDIDATE SELECTOR and makes a tentative decision for the first three formant frequencies. At first TRACKER looks for a reasonable place to track. There exists a region within which an overlap of two formant components never occurs. In the case of a male voice, only the first formant exists between 220 hz. and 500 hz. and only the second and the third formant exist between 900 hz. and 1100 hz. and between 2700 hz. and 3000 hz.. If the first candidate for a formant frequency is within the first region it is reasonable to assume that this is the peak caused by the formant. After making an initial selection TRACKER begins tracking forward or backward.

TRACKER uses two criteria to determine formant frequencies of the next frame. Basically, TRACKER uses a criterion of minimum shift of peak position from one frame to the next. This nearest neighbour

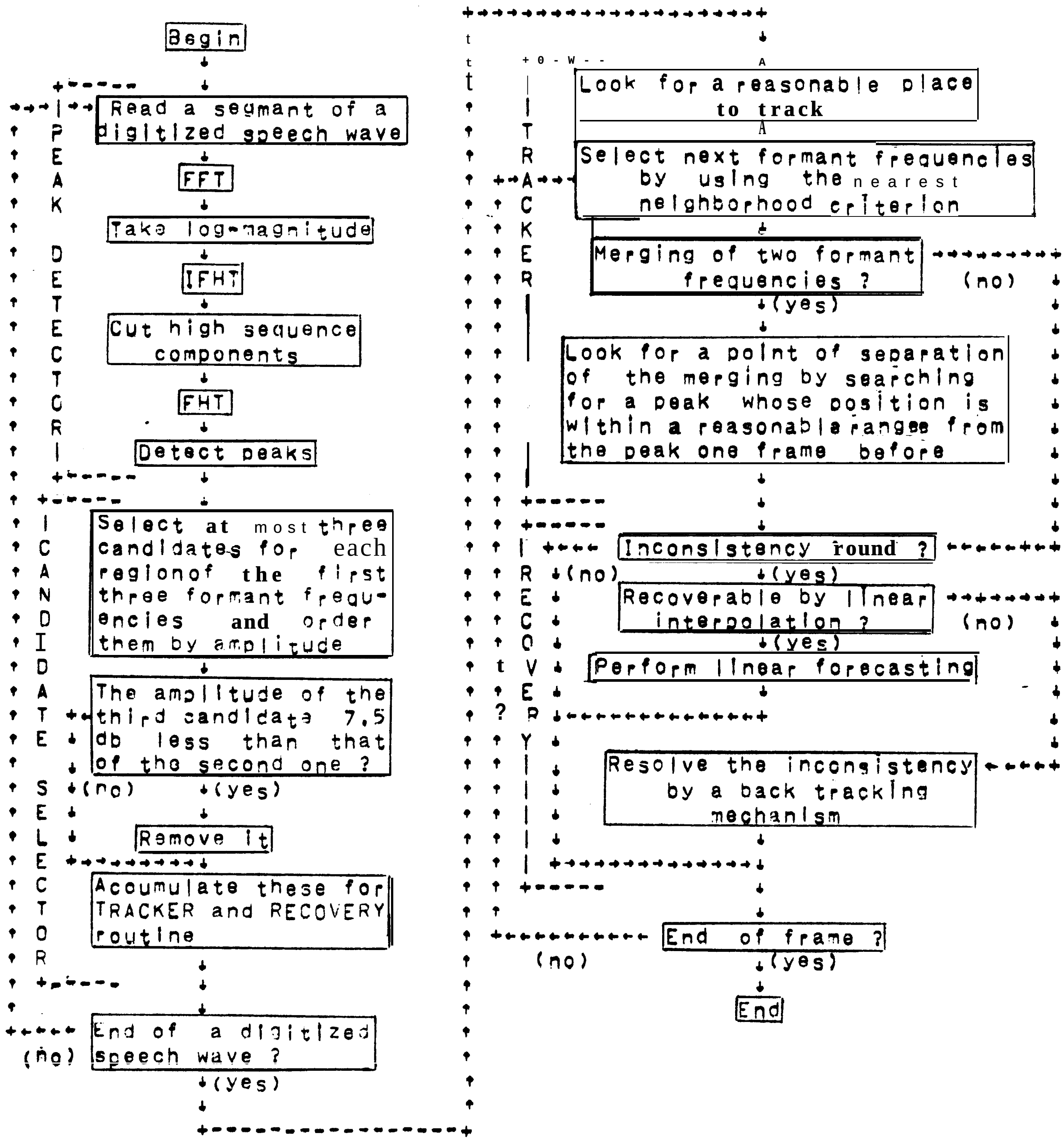


Fig. 3.6 General flow diagram of a formant tracking program.

criterion is used as long as merging of two formant frequencies does not occur. As soon as merging of two formant frequencies occurs, TRACKER makes use of the other criterion to look for a point of separation after merging. After a tentative selection of the next formant components using the first criterion, TRACKER looks for a peak whose position is within a reasonable range from the peak one frame before. If TRACKER can find such a peak it will select the peak as the point of separation of the two trajectories, otherwise the two trajectories remain merged. A wrong decision by TRACKER is corrected by the RECOVERY routine.

D. RECOVERY.

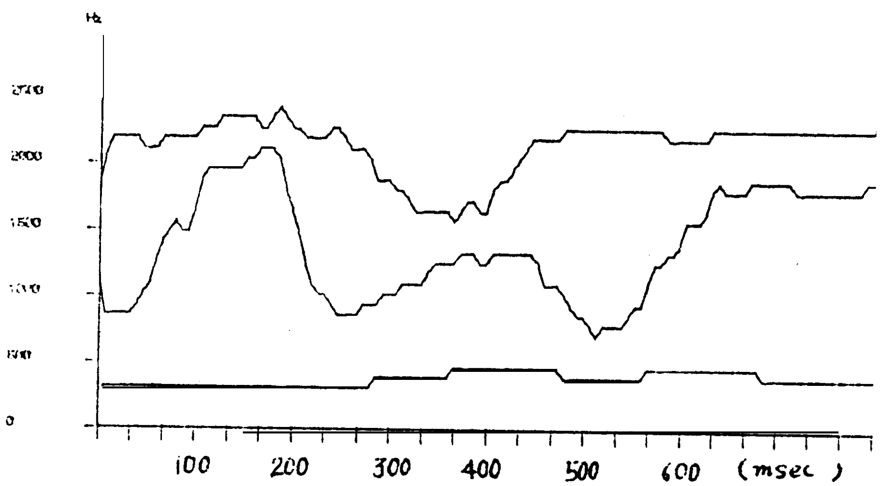
RECOVERY works when some inconsistency is recognized by a formant tracking program. An inconsistency is a discontinuity or a sharp change in following a formant trajectory.

There are two major reasons why TRACKER follows a path that has a sharp change from one frame to the next. The first reason is that a prominent peak caused by a formant component is often lost in a spectrum because of the stochastic movement of glottal pulses. This results in a discontinuity in a formant trajectory if the trajectry is selected in a microscopic way. This can be resolved by using the neighborhood information. A FORECASTER works in this case.

The second reason is that a wrong decision has been made in the past by TRACKER and a wrong path has been followed as a formant trajectory. For example a formant tracking program has mistaken the first formant for the second and the trajectory suddenly enters into the region where the second formant does not exist; namely the region between 220 hz. and 550 hz. The other typical example is the following of a wrong path which is not a formant trajectory and eventually disappears. These are corrected by using a backtracking mechanism in the RECOVERY routine. In the previous example, after the RECOVERY routine has recognized an error, a trajectory followed as the second formant is replaced by the first formant trajectory. Then another peak is selected for the second formant frequency by RECOVERY. The routine extends the new second formant trajectory backward by using TRACKER, and FORECASTER. This is an example of how the back tracking mechanism works. We can see that that a recovery process by back tracking has to have a recursive structure, but in our case the depth of recovery is limited to one.

5.4.2 example.

An example of formant tracking obtained from the program is shown in Fig. 5.7. A spoken sentence is "we were away." More than 90% of running time is devoted to PEAK DETECTOR and CANDIDATE SELECTOR.



Fig, 3,7 An example of the first three fornant trajectories for a sentence of "We were away", 4 Conclusion.

In the paper we have discussed both the advantages and disadvantages of the Hadamard transform, as compared to the Fourier transform, for a speech wave analysis. The experiments in section 2 reveal that application of the Hadamard transform directly to speech waves yields poor results, as it fails to extract important features. The smaller the number of features necessary to accurately represent a speech wave, the better it is. In the Fourier case, in almost all cases for a vocalic sound a speech wave is represented by the first three formant frequencies and the pitch (fundamental) frequency. Only four parameters are needed. However in the Hadamard sequency spectrum, we cannot observe any typical features because of the strong time-shift sensitivity which makes it impossible to apply even a pitch synchronous method. In other words, typical features which are recognizable in the Fourier case are averaged and are scattered away in a wide range of a sequency power spectrum. Some of the experiments in section 2.2 demonstrate it.

Time-shift invariants for the Hadamard sequency spectrum are known. One of these defined by eq. (8) does not bear enough information to perform a speech wave analysis since from a digitized speech wave composed of 256 points, we get only 9 components. Although each of these components has near relationship with an output from a filter bank, its frequency band is determined by the number of points transformed. Ohnsorg [7] has defined another complete set of the Hadamard transform which has exactly the same number of components as a Fourier frequency spectrum and is invariant under a circular time-shift. Ahmed et al [8] found an algorithm to calculate these terms, however multiplication by an irrational number is included and is more complicated than that of the fast Hadamard transform. As Ohnsorg suggests that the prominent energy line of the Fourier spectrum tends to be exaggerated, it is desirable to calculate Ohnsorg's invariants for a speech wave.

In section 3 the hapstrum technique is introduced. This technique is similar to the so called cepstrum technique except that the FHT is applied to the log-magnitude frequency spectrum. After the application of the Fourier transform the Hadamard transform is applied to detect a pitch period or to get a smoothed spectrum. This technique shows some positive aspect of the Hadamard transform for the analysis of a speech wave with regard to the reduction of the processing time required for smoothing. Good smoothing makes it easy to extract the first three formant frequencies in a spectrum. We should note that smoothing by the hapstrum is obtained at the cost of accuracy in determining peak position of a smoothed spectrum. This is explaned by eq (11) or (12) in section 3.3. We can conclude that precise formant frequencies are obtained by the cepstrum technique at the cost of processing time, while a reduction of processing time is obtained by the hapstrum technique at the cost of accuracy in determining the formant frequencies. However, it is often true that to detect a peak causedby a pitch period, is difficult even in the case of a male voice. The author's original optimistic standpoint was that the Hadamard transform might reveal some new aspect of speech waves. However the only gain found from using the Hadamard transform was the reduction of processing time required for smoothing, andthis was obtained at the cost of precision.

A formant tracking program using an edge follower has been described in section 3.4. While the algorithm is rather sophisticated, most of the time is still devoted to the smoothing and peak selection procedures.

APPENDIX.

Let us define a few of the functions used here.

Function G(U) is defined as follows:

Let a binary representation of U or G(U) be

$G(U) = G = Gn\text{-}1Gn\text{-}2.....G1G\emptyset$
$U = Un\text{-}1Un\text{-}2.....U1U\emptyset$
Gi and $Ui \in (1,0)$ (for $1 \leq i \leq n-1$)
where
$Gn\text{-}1 = U\emptyset$ XOR $U1$
$Gn\text{-}2 = U1$ XOR $U2$
.... (A-1)
....
$G1 = Un\text{-}2$ XOR $Un\text{-}1$
$G\emptyset = Un\text{-}1$
(XOR stands for exclusive-or)

Let an array $[a(j)]$ be $[e,f]$ such that
$[e,f] = [e\emptyset,e1,...,em\text{-}1,f\emptyset,f1,...,fm\text{-}1]$
and $[e] = [e\emptyset,e1,...,em\text{-}1]$
$[f] = [f\emptyset,f1,...,fm\text{-}1]$.

From the definition of the Hadamard transform

$$[A(j)] = (1/N)[e,f]\begin{Vmatrix} H(n-1) & H(n-1) \\ H(n-1) & -H(n-1) \end{Vmatrix} \qquad (A\text{-}2)$$

where $N = 2^n$ and $m = N/2$.

5.1 Filtering on the sequence domain.

(A) If array $[f]$ is $[\overbrace{\emptyset,\emptyset,...,\emptyset}^{m}]$ than $A(k) = A(l)$ for $l = k + m$, where $0 \leq k \leq (N/2) - 1$, and the difference of sequence number between $A(l)$ and $A(k)$ is one.

Proof:

Suppose the sequence number of the k-th or l-th element of array $[A(j)]$ is s or t, respectively. Then if the binary representation of k, l, s or t is

$k = kn\text{-}1\ kn\text{-}2\ ...\ k1\ k\emptyset$
$l = ln\text{-}1\ ln\text{-}2\ ...\ l1\ l\emptyset$
$s = sn\text{-}1\ sn\text{-}2\ ...\ s1\ s\emptyset$ (A-3)
$t = tn\text{-}1\ tn\text{-}2\ ...\ t1\ t\emptyset$
then $k = G(s)$ and $l = G(t)$ (see [3]) (A-4)

Since $\emptyset \le k \le (N/2) - 1$ and $I = k + (N/2)$
most significant binary digit k_{n-1} and I_{n-1} are

$$k_{n-1} = \emptyset$$
$$I_{n-1} = 1 \text{ and} \qquad (A-5)$$
$$k_i = I_i \quad \text{for } i \ne n-1$$

From eq (A-4) and (A-5)

$$\begin{aligned} I_{n-1} &= t_\emptyset \text{ XOR } t_1 = 1 \\ k_{n-1} &= s_\emptyset \text{ XOR } s_1 = \emptyset \\ k_{n-2} &= s_1 \text{ XOR } s_2 = t_1 \text{ XOR } t_2 \\ &\ldots\ldots\ldots\ldots \qquad (A-6) \\ k_1 &= s_{n-2} \text{ XOR } s_{n-1} = t_{n-2} \text{ XOR } t_{n-1} \\ k_\emptyset &= s_{n-1} = t_{n-1} \end{aligned}$$

We obtain the following relation from eq (A-6),

$s_i = t_i$ for $1 \le i \le n-1$, and

$$s_\emptyset = \emptyset \text{ and } t_\emptyset = 1 \text{ (if } s_1 = \emptyset) \qquad (A-7)$$
$$s_\emptyset = 1 \text{ and } t_\emptyset = \emptyset \text{ (if } s_1 = 1)$$

Eq (A-7) implies that a s or t is in sequence,
In other words the difference of sequence number between A(k) and A(I) is one,

Let [A(J)] be [E,F] where

$$[E,F] = [E_\emptyset, E_1, \ldots, E_{m-1}, F_\emptyset, F_1, \ldots, F_{m-1}] \qquad (A-8)$$

From eq (A-2)

$$E_k = [e](h(k)) + [f](h(k))$$
$$F_k = [e](h(k)) - [f](h(k)) \qquad (A-9)$$

where (h(k)) is the k-th column of matrix H(n-1),

Since in our case $[f] = [\overbrace{\emptyset, \emptyset, \ldots, \emptyset}^{m}]$

$E_k = F_k$, namely $A(I) = A(k)$ for $I = k + (N/2)$ Q.E.D.

(B) We can generalize the result of (A) further,
zero all components of array [a(J)] such that
$2^k \le j \le N - 1$ where $1 \le k \le n-1$, then

$$A(\emptyset) = A(2^k) = A(2\cdot(2^k)) = A(3\cdot(2^k)) = \ldots = A((2^{(n-k)}-1)\cdot 2^k)$$
$$A(1) = A(2^k+1) = A(2\cdot(2^k)+1) = A(3\cdot(2^k)+1) = \ldots = A((2^{(n-k)}-1)\cdot 2^k+1)$$
$$\ldots\ldots\ldots\ldots$$

$$A(i)=A(2^{k}+i)=A(2\cdot(2^{k})+i)=A(3\cdot(2^{k})+i)=,,,=A((2^{(n-k)}-1)\cdot 2^{k}+i)$$

....

$$A(2^{k})=A(2\cdot 2^{k}-1)=A(3\cdot(2^{k})-1)=A(4\cdot(2^{k})-1)=,,,=A(2^{n}-1)$$

and in each group, for example ($A(i)$, $A(2^{k}+i)$, $A(3\cdot(2^{k})+i)$,,,,, $A((2^{(n-k)}-1)\cdot 2^{k}+i)$), $2^{(n-k)}$ consecutive sequence numbers are Included,

proof:

It Is apparent from the recursive definition of the Hadamard transform matrix given in eq (1) and the proof given in (A), 6 References.

[1] H.F. Harmuth: Application of WALSH functions in communication, IEEE Spectrum, Nov., 82-91, 1969,

[2] H.F. Harmuth: TRANSMISSION OF INFORMATION BY ORTHOGONAL FUNCTIONS, Springer-Verlag, 1970,

[3] H.C. Andrews: COMPUTER TECHNIQUES IN IMAGE PROCESSING, Academic Press, 1970,

[4] S.J. Campanella and G.S. Robinson: Digital sequency decomposition of voice signals, Walsh Function Symp., Naval Res, Lab., 230-237, 1970,

[5] A.M. Noll: Cegstrum pitch determination, J. Acoust. Soc. Amer., 41, 2, 293-309, 1967,

[6] F. Pichler: WALSH functions and optimal linear systems, Walsh Function Symp., Naval Res. Lab., 17-22, 1968,

[7] F.R. Ohnsorg: Spectrul modes of the WALSH-HAQAMARD transform, Walsh Function Symp., Naval Res. Lab., 55-59, 1971.

[8] N. Ahmed, A.L. Abdussattar and K.R. Rao: Efficient computation of the WALSH-HADAMARD transform spectral modes, Walsh Function Symp., Naval Res. Lab., 276-279, 1972,

[9] B.S. Atal and S.L. Hanauer: Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Amer., 47, 2, 637-655, 1971,

[10] L.R. Rabiner and R.W. Schafer: System for automatic formant analysis of voiced speech, J. Acoust. Soc. Amer., 47, 2, 634-648, 1970,

[11] A. Herskovits and T.O. Binford: On boundary detection, MIT Project MAC Artificial Intelligence Memo 183, July 1970.

[12] J.D. Markel: Formant trajectory estimation from a linear least-squares Inverse filter formulation, SCRL-Monograph No.7, Oct., 1971.

[13] J.D. Markel: Automatic formant and fundamental frequency extraction from a digital inverse filter formulation, Conf. on Speech Comm. and Processing, 81-84, 1972,