TR-H-264

A Mathematical Framework for Auditory Processing: A Mellin Transform of a Stabilised Wavelet Transform?

Toshio IRINO and Roy D. PATTERSON (Univ. Cambridge)

1999.1.29

ATR人間情報通信研究所
〒619-0288　京都府相楽郡精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Telephone: +81-774-95-1011
Fax : +81-774-95-1008

A MATHEMATICAL FRAMEWORK FOR AUDITORY PROCESSING: A MELLIN TRANSFORM OF A STABILISED WAVELET TRANSFORM?

Toshio Irino and *Roy D. Patterson*

ATR Human Information Processing Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun,
619-0288, Kyoto, JAPAN
irino@hip.atr.co.jp
http://www.hip.atr.co.jp/~irino/

Centre for Neural Basis of Hearing, Dept. of Physiology, Univ. of Cambridge
Downing Street, Cambridge,
CB2 3EG, UK
roy.patterson@mrc-cbu.cam.ac.uk
http://www.mrc-apu.cam.ac.uk/personal/roy.patterson/

INTRODUCTION

In the conventional psychoacoustic description, sounds have a **pitch**, a **loudness** and a **timbre**. Pitch is the psychological correlate of the repetition rate of the sound source, loudness is the psychological correlate of the intensity of the sound source, and timbre, or sound quality, is the psychological correlate of everything else. It is still commonly believed that timbre can be explained by the short term power spectrum of the sound, but this has proved a singularly unfruitful hypothesis, limited largely to explaining that sounds with proportionately more high-frequency energy will sound 'brighter'. Despite 30 years of computer-assisted signal processing, there is still no practical specification of the short-term power spectrum for use in timbre research, and there are no computational models for the timbre of musical instruments and the human voice, despite the wealth of articles on details of the spectra of music and speech sounds. There are increasingly sophisticated spectral vocoders for capturing and resynthesising speech sounds (Kawahara et al., 1998) but they do not explain timbre, or how the auditory system might code timbre.

The main progress over the past decade in this area has been in time-domain models of auditory processing; they simulate the neural activity pattern that appears in the auditory nerve in response to a complex sound, and convert it into a stabilised representation of the **auditory image** that we hear in response to a complex sound. The auditory images of tonal sounds contain elaborate time-interval patterns referred to as **auditory figures**, whose **shape** varies with the timbre of the sound. More importantly, the shapes of the auditory figures produced by sources with similar timbres, like cellos and violins, are similar; the auditory figures differ in size and they occupy different positions in the auditory image. But the shape is similar, which suggests that we should look to the shape of the **auditory figure** as a basis for a model of timbre perception, and a means of determining the elements of timbre that identify a family of instruments or the human voice.

Sound sources in the world have **size** and **shape** and it would appear to be fairly natural to map physical size and shape to the size and shape of auditory figures. This has been an implicit assumption in our auditory image research for some years now. The purpose of the research is to move computational hearing forward beyond pitch and loudness to investigate the elements of timbre revealed by the shapes of auditory figures, and to establish measures of the properties that we hear.

Recently, we have discovered a mathematical transform, the Mellin transform, which represents signals, not like the Fourier transform, in terms of the energy at a specific physical frequency, but rather in terms of the size and shape of the signal. The mathematics indicates that size is a physical property of a sound, just like repetition rate and intensity, and that it is separate from **shape** information and independent. The optimal spectral preprocessor for the Mellin transform is a wavelet transform which agrees well with the current understanding of the auditory filterbank. The Mellin transform is time variant, however, and so it cannot be applied to the wavelet transform of the signal directly. Briefly, there must be a process between the wavelet and Mellin transforms to identify the appropriate start point for the Mellin transform at all moments in time. The strobed temporal integration process used in the auditory image model (AIM) to construct stabilised auditory images establishes the start point of each cycle of periodic sounds, and so can provide the necessary start point for the Mellin transform at all points in time. This suggests that pre-cortical auditory processing may be a form of Mellin transform to extract information about the size and shape of a source from a stabilised version of the wavelet transform. The analysis suggests that the auditory image and auditory figures are an intermediate representation in this Mellin transform process.

In this paper we describe a computational version of the auditory Mellin transform. This involves 1) specifying the role of existing processes and representations, such as the **auditory filter bank**, the **auditory image**, and the **auditory figure,** in the larger Mellin transform framework, 2) developing an auditory form of the extensions required to complete the Mellin transform, and 3) specifying the relationship between the products of the Mellin transform and the components of auditory perception.

Conventional Analysis

Many conventional techniques use the Fourier transform or linear predictive coding as the initial stage of processing to produce a time-frequency representation of the signal. This time-frequency representation is a better representation for the features speech or music than the waveform, because it removes phase information between the harmonics of a tonal sound, and the loss of phase, at least between low-frequency harmonics, is characteristic of auditory processing. It is also a very efficient and well established method of calculating a time-frequency representation of signals.

Most applications optimise their system in terms of the initial spectral transformation and do so in the extreme. The end product of this optimisation, however, still results in systems with rather limited performance when compared with humans. For example, speech recognition systems trained on either male speech or female speech perform badly when presented with a child's speech. Basically the problem is that the formant structure and pitch are different for children than adults. To resolve these problems people have tried to train speech recognisers on a mixture of speech from children and adults, or to build recognisers with parallel competing models for the speech of men, women and children. Unfortunately, there are no large scale data bases for children, and the production of such data bases is very time consuming and expensive. Moreover, current performance of such systems still falls far short of that required for machine communication and translation. The problem arises from the fact that the vocal tract and the vocal cords are considerably smaller in children and it is difficult to normalise for such changes using traditional spectrographic data. This is characteristic of many physical scaling problems in source-filter systems. Other examples are families of musical instruments and families of engines.

I. GENERAL DESCRIPTION OF THE SOLUTION TO THE NORMALISATION PROBLEM

Briefly, the mathematics of the Mellin transform suggest that it is possible to solve the normalisation problem by replacing the time-frequency representation with a 'log-time-interval'-'log-frequency' representation in which the pattern of activity does not change with source size except for a vertical shift of the entire pattern; all the information about source size is summarised in a single valued parameter, the vertical position of the pattern.

A. The Mellin transform

A scale-invariant representation of the 'log-time-interval'-'log-frequency' representation can be produced using the Mellin transform (Titchmarsh, 1948), the equations for which are presented in Appendix A.

B. The physics of loss less acoustic tubes

A solution to the propagating wave of a lossless acoustic tube can be obtained using a plane-wave approximation. Any physics text shows the analytic solution of the propagating wave for horns with uniform cross section. When the cross section is varying in time, as with the vocal tract, we can still solve the propagating wave numerically using an approximation involving a set of micro-cylinders, that is, a set of cylinders with varying diameter, each with arbitrarily small height. Any text describing speech production presents such methods for solving the propagating wave equation (Nakata, 1995).

We consider the impulse of such an acoustic tube. The import feature is the correspondence between the size of the acoustic tube as it expands or contracts and the duration of the impulse response which expands or contracts linearly in time with the size of the acoustic tube. The physical length of the acoustic tube is linearly correlated with the length of the impulse response.

We hear phonemes pronounced by men, women and children as approximately the same although the length of the vocal tract varies considerably from group to group and from speaker to speaker. Any phonetics text book describes the shape of the vocal tract in terms of the position of the tongue for the speech sounds of a particular language, and they emphasise that it is the relative size of the two main cavities (mouth and throat) that matter rather than the absolute length. This means that the relative proportions of the cross-area functions are preserved when the length of the vocal tract varies. Therefore, for a given speech sound, the impulse responses of the vocal tracts of men, women and children vary primarily in the overall length of the impulse response in time; the shape of the wave within the impulse response varies relatively less. This is clearly a description of the ideal case, but it does present a reasonable approximation for the case of speech sounds and the sounds of musical instruments.

C. The General Problem

It seems likely that the auditory system has a way of understanding the scaling of the impulse response of the vocal tract, in which case, we do not need to consider the Fourier representation of the signals and the problem of normalising in the Fourier domain. With an invariant representation, we would readily recognise that a phoneme is the same when produced by men, women and children. Such a representation would appear to be available by judicious application of the Mellin transform to a stabilised version of the wavelet transform of the sound.

The Mellin transform cannot be applied directly to real sounds because it is not shift invariant in time. This problem will be described shortly. To date, the Mellin transform does not appear to have been applied in any signal processing devices, or even in signal processing research with a few minor exceptions (see for examples, Altes, 1976; Cohen 1993; Umesh et al., 1997, 1998).

For successful application of the Mellin transform, we require a stabilised representation of the speech signal to avoid the shift-varying properties of the Mellin transform and circumvent the bottleneck in speech recognition performance with current Fourier, Cepstral and LPC techniques. This paper describes a relatively simple method for calculating a size-invariant feature representation using the Mellin transform, provided a stabilised representation of the signal exists.

II. SOLUTION TO THE SHIFT-VARYING PROBLEM OF THE MELLIN TRANSFORM

The shift-varying problem can be solved by providing the appropriate start point for the Mellin transform at all moments in time. The outline of the solution is presented in Figure 1 where the signal is converted into a stabilised wavelet transform before we compute the Mellin transform. Source information is normalised in the Mellin Image so that it does not change with the temporal dilation or compression associated with changes in the size of the source, or with changes in the repetition rate of the sound. Any kind of conventional signal processing or pattern recognition can be applied to the Mellin transform to utilise the normalised source information.

A. Stabilised Wavelet Transform

Figure 2 is a block diagram of the calculation of the Mellin transform. The equations for the processes are presented in Appendices B and C. The terms used in this description are explained in Appendix D. The numbers in brackets refer to components of Figure 2.

The signal, which is often periodic (Eq. B1), is analysed by a wavelet transform (Combes et. al, 1989); the carrier frequencies of the wavelet kernels (Eq. B2) are spaced linearly on a logarithmic frequency scale, as illustrated by the gammachirp kernel in (Eq. 2). This is the optimal kernel in terms of minimal uncertainty for the Mellin transform (Irino and Patterson, 1997). The wavelet transform with a gammachirp kernel is constant-Q, that is, the 'best-frequency to bandwidth' ratio is constant, and so, the wavelet filters are dilated or compressed versions of the wavelet kernel (Eq. B3). The wavelet transform is the set of convolutions of the signal with all the wavelet filters (Eq. B4). The signal analysed by the wavelet transform is compressed separately in each channel using either log compression (Eq. B5) or power compression (Eq. B6). When a signal is expanded or compressed in time outside the system, the wavelet filterbank passes the signal without distortion; the signal shifts to filters with proportionately higher or lower centre frequencies, but the response itself is just an expanded or compressed version of the original signal. In this sense the wavelet transform is transparent to expansion and compression of the signal.

B. Preconditions for the use of the Mellin Transform

Equation A1 shows that it is necessary to specify the start point for the Mellin transform because the Mellin transform is a shift-varying transform. This is one reason that the Mellin transform has only rarely been used in processing of mechanical vibrations. This stands in contrast to the Fourier transform which is shift invariant. As noted earlier, the Mellin transform has the advantage that it normalises the physical scale of the source; that is it maps similar sources with different sizes to the

same distribution. We can overcome the shift-varying problem of the Mellin transform if we can identify the appropriate start point for the analysis of a given sound.

The signal is always flowing in time, and so the output of the wavelet transform is always flowing in time, and it is not in general possible to identify the appropriate start point for the Mellin transform. For periodic and quasi-periodic sounds, the individual wavelet-filter outputs have one maximum per cycle of the wave and the source information is in the wave shape within the pattern. Thus, we detect the repetition of the wave and initialise the start point for the Mellin transform to the maximum of each cycle. Examples of maxima detection are presented in Irino and Patterson (1996) and Patterson and Irino (1998). In each channel, the time of the maximum is taken to be the start time, or strobe point, for temporal integration, and temporal integration means adding a copy of the wavelet-filter output into the corresponding channel of a new representation, the stabilised wavelet transform of the sound. The wavelet-filter output is added point-for-point with what is already in the corresponding channel of the stabilised wavelet transform. The mechanism is referred to as strobed temporal integration (Patterson, Allerhand and Giguére, 1995). The origin of this representation is always zero since the abscissa is time interval from the previous peak. The stabilised wavelet transforms of periodic and quasi-periodic sounds (Eq. 7) preserve source information in what is referred to as source figures (Eq. 8), which appear as stabilised time-interval patterns within individual periods of the stabilised wavelet transform. We can apply the Mellin transform to source figures because the shift varying problem has been solved in this representation by positioning successive cycles of the wavelet transform output at the same place relative to the origin. So strobed temporal integration establishes the proper conditions for analysing source information using the Mellin transform.

III. CALCULATION OF THE MELLIN TRANSFORM

The Mellin transform can be rewritten in terms of the operators used in quantum mechanics, and in this case, the Mellin operator is the product of the standard time and frequency operators described by Gabor (1946) and Cohen (1993). So the product of time and frequency is essential to the concept of the Mellin transform. We perform the Mellin transform (Eq. A1) on any of the source figures (Eq. B8) in the stabilised wavelet transform, by integrating along lines where the product of time-interval and frequency is constant (Eq. B9). The form of the Mellin transform is shown in Eq. B10. The parameter, p, is complex (Eq. B11), and when expanded produces a complex version of Eq. B10, as shown in Eq. B12. This enables us to produce a rectangular representation of the Mellin transform of the source figure whose abscissa is the product of time-interval and frequency. This representation is referred to as a Mellin Image and source information is normalised so that it does not change with the temporal dilation or compression associated with changes in the size of the source, or with changes in the repetition rate of the sound. Any kind of conventional signal processing or pattern recognition can be applied to the Mellin transform to utilise the normalised source information. The calculation of the Mellin transform will be described in more detail when examples are presented in later sections.

To this point, the calculation of the Mellin Image is for one frame of the stabilised wavelet transform of the sound; that is, for one moment in time. Sounds are dynamic and so the Mellin Image is computed at evenly spaced times and the resulting sequence of Mellin Images represents the source information of an extended dynamic sound.

The Mellin Image is invariant when the signal source changes size, despite the fact that the length of the impulse response changes with source size. This is not true for the Fourier transform or auto-regressive analyses like LPC analysis. Despite invariance with source size, the Mellin Image is

sensitive to source differences other than dilation or compression of the impulse response. In the case of speech, sounds from different people are normalised for vocal tract length, which makes other differences, such as phoneme differences, more readily identifiable. As a result, we can expect to achieve an improvement in performance when a recogniser trained on adult male speech is applied to the speech of children. In addition, we can expect improved performance when the Mellin Image is used as a preprocessor for speech recognition machines.

IV. THE MELLIN TRANSFORM OF A CLICK TRAIN

Figure 3 is a block diagram of a hypothetical speech recognition system based on the wavelet and Mellin transforms. First the signal is converted into a Stabilised Auditory Image (SAI) which is the auditory equivalent of the stabilised wavelet transform. The SAI is converted into a size-shape image and finally a Mellin Image. The Mellin Image is the auditory equivalent of the Mellin transform.

A. The Stabilised Auditory Image

The speech sound is analysed by an auditory filterbank; the impulse responses of the individual filters have gamma envelopes and chirping carriers (Eq. C1), and the filterbank is essentially constant-Q above 500 Hz (Eq. C2), meaning that the 'best-frequency'-to-'bandwidth' ratio is constant. So the auditory filterbank is basically a wavelet transform (Eq. C3 and C4) with a gammachirp kernel (Eq. C1) (Irino and Patterson, 1997), whose parameter values are set to simulate cochlear filtering. The output of the auditory filterbank is converted into a neural activity pattern (NAP); the process includes half-wave rectification, log or power compression of amplitude (Eq. C5 and C6), and adaptation, which together enhance the onset of the signal and sharpen features in the filterbank output. The activity in each channel is monitored to identify local maxima in the activity which are used to control temporal integration. The process operates on the envelope of the activity and specifically on the derivative of the envelope, referred to as 'delta gamma' (Irino and Patterson, 1996). The local maxima occur regularly when the signal is periodic or quasi-periodic as in the voiced parts of speech and sustained musical notes. Temporal integration is strobed on each of the local maxima and temporal integration (Eq. C7) consists of taking the current segment of the neural activity pattern (about 35 ms in duration) and adding it into the corresponding channel of the auditory image, with whatever is currently in that channel. This strobed temporal integration (STI) process converts the time dimension of the NAP into a time-interval dimension (Eq. C7). The log-best-frequency dimension is the same in the new representation (Eq. C3). STI is applied separately to each channel of filterbank output; the stabilised auditory image (SAI) (Eq. C7) is the array of all stabilised neural patterns for all channels in the auditory filterbank. The auditory image decays continuously with a half life of 30 ms. The spectral profile of the SAI is similar to a spectral vector of a traditional spectrogram, and the spectral profiles of sequences of auditory images have been used to construct auditory spectrograms for speech recognition (for examples, see Patterson et al., 1989, 1992, 1996; Robinson et al. 1990).

B. The Size-Shape Image

The stabilised auditory image is most often presented in a rectangular form with a linear time-interval axis oriented horizontally. There is also a spiral form of the auditory image (Patterson, 1987) but that is for musical pitch comparisons at a higher level in the system. The SAI is converted into a

size-shape image in four stages as illustrated in Figure 4, and then into a Mellin Image. The construction of the size-shape image from the SAI is illustrated in Figure 4.

An example of a stabilised auditory image from the standard AIM model (Patterson et al., 1995) is presented in Figure 5. It shows just under three cycles of the pattern produced in the auditory image by a click train with a click rate of 100 Hz. The click train produces a pitch like that of a male speaker with a 'deep' voice. The ordinate is channel best frequency in Hertz and it is a quasi-logarithmic frequency axis. The abscissa is time-interval in milliseconds from the local maximum that initiated temporal integration; it is a linear axis in this representation. The main verticals are spaced by the period of the original wave. The zero on the abscissa is the point to which the local maxima are mapped during temporal integration. The local maxima identify the individual cycles of periodic signals, and the start points for features in non-periodic signals. In this way, STI identifies the cycles of wave and produces candidate start points, or zeros, for the Mellin transform at each multiple of the period of the wave.

The Mellin transform includes the assumption that the channels of the initial wavelet filterbank are aligned such that the start points of the individual impulse responses are all at zero. STI introduces a negative phase shift to the low-frequency channels which is not strictly in agreement with the mathematics of the Mellin transform. The misalignment is illustrated by the curve of pulses to the left of each main vertical in the auditory image in Figure 5. The SAI can be realigned simply by shifting each channel to the right by one period of the centre frequency of the auditory filter as indicated in Eq. C8 before auditory figure extraction. The aligned version of the click-train SAI in Figure 5 is presented in Figure 6. The main verticals now provide a very good approximation to the correct start point for the Mellin transform. Note, however, the alignment has little effect on the final values in the Mellin transform of the signal.

The STI process stabilises the repeating time-interval patterns produced by tonal sounds in the NAP, and generates concentrations of activity on vertical lines that divide the auditory image into frames (Eq. C9) whose width is the period of the original sound. The pattern of time intervals produced by a tonal sound in any one frame is referred to as an auditory figure (AF) (Eq. C9) of the source, and the frame itself is referred to as a frame of the auditory image. The time-interval profile of the SAI (Figure 5) is referred to as the 'summary SAI', and it is used to identify the period of the sound, which is then used to identify the boundaries of the auditory figures for extraction from the auditory image.

After extraction, the time-interval axis of the auditory figure is transformed into a log-time-interval axis (Eq. C10) to convert the curved impulse response lines of the auditory figure into parallel, regularly spaced lines that are essentially straight in the region above 500 Hz. Figure 7 shows the transformed version of the leftmost Auditory Figure (AF) from the SAI in Figure 5 with a logarithmic time-interval axis. It was derived from Figure 6 using a spline interpolation of the sampling points. The vertical solid line is the upper boundary of the AF. The activity associated with the ringing of the auditory filters in response to an impulse falls along the dashed lines which are straight in the region above 500 Hz in this representation of the AF. The slope of the lines is that of the negative diagonal across the AF frame. The dimensions of the representation are log-time-interval and log-frequency, and this is the form that facilitates the calculation of the Mellin transform.

The calculation of the Mellin transform, and the shape of the sound source are more easily understood if the AF with log-time-interval axis (Eq. C10) (Figure 7) is time-interval aligned to reorient the diagonals to the vertical as shown in Figure 8 (Eq. C11). To each channel, we make a time-interval shift equal to the log of the best frequency of the channel. The verticals that bound the region of the AF in Figure 7 now appear as diagonals in Figure 8. The new abscissa is the product of time-interval and best-frequency, designated h (Eq. B9), on a logarithmic frequency scale. The ordinate is, as before, best-frequency on a logarithm frequency scale. The left-most dashed vertical shows points in

the AF where the product of time-interval and best-frequency, h, is unity. For the click train, which produces impulse responses in all of the wavelet filters, the activity is concentrated on verticals at integer multiples of h. This is emphasised in Figure 9 which is a version of Figure 8 with a linear h axis; the response in each channel is a transformed and aligned version of the wavelet kernel. This figure is produced directly from Figure 5; the sample points in each channel is re-sampled proportionally to the best frequency of the channel and, thus, the activity on the line of h=0 is again presented in the figure. The solid curve is the boundary of the AF. The shape of an AF in this representation does not change with the size of the source; the AF just moves up or down the verticals as source size decreases of increases, respectively. Accordingly, this representation is referred to as the Size-Shape Image, or SSI. It is particularly useful for visualising the shapes associated with the AFs of vowel sounds, as will be illustrated in a later subsection.

The auditory figure in Figure 8 is derived from the leftmost figure frame in the SAI of Figure 5, but this need not be the case. The origin for the AF in the SSI can be the start point of any of the AFs in the SAI; that is, any multiple of the period of the sound (Eq. C9). For example, there is a concentration of activity on the 10-ms vertical in the SAI, corresponding to the period of the click train, and it is equally reasonable, to use the AF beginning at this vertical, as the start point for the alignment, log transform, and rotation that produce the SSI. Indeed, when tonal signals occur in noise, the characteristics of the tone are better represented by the second AF of the SAI; the pattern of activity in the first AF of the SAI contains proportionately more of the noise component.

One of the profiles of the SSI emphasises information about the impulse response of the wavelet kernel, and deviations from the impulse response produced by a signal, and so it is referred to as an impulse profile (Eq. C12). The other profile of the SSI is the auditory spectrum of the auditory figure, and is referred to as the spectral profile of the SSI (Eq. C13). The impulse profile contains source information that is unlike that in the traditional spectrogram vector. This suggests the generation of a new form of spectrogram for speech recognition in which the spectral profile and the impulse profile of each frame of the SSI are concatenated to produce the vector associated with that frame of the signal in the spectrogram. Thus would be an SSI spectrogram.

C. Construction of the Mellin Image

The SSI is dominated by the impulse response of the wavelet transform. In order to extract the information like the resonances of the source, it is necessary to deconvolve the impulse response of the click train from other source information in the SSI. In Figures 8 and 9, the peaks on any one vertical of constant h are all similar in height, indicating that, for a click train, the distribution is largely independent of best-frequency. This suggests that a form of deconvolution can be accomplished by taking the Fourier transform of the SSI along each vertical, in which case, the majority of the impulse response will appear at very low spatial frequencies. Indeed, the Mellin Image is a transformation of the SSI in which each vertical vector is replaced by the magnitude of the Fourier transform (FT) of the activity on the SSI vector. It corresponds to the application of a complex sinusoidal weighting, (Eq. C14), along the log-frequency. The result is another two-dimensional image in which each vertical is a magnitude spectrum for the corresponding line of the SSI. The new representation (Eq. C15) is referred to as the Mellin Image (MI). It has the same abscissa, h, as the SSI, but the ordinate is a new variable, c, which is the spatial radius frequency of the Fourier transform. The vertical position of an auditory figure in the SSI is converted into phase information in the Fourier transform, and as such does not appear in the magnitude spectrum. Thus, the MI version of the auditory figure presents shape information about the source in a form that does not change with the size of the source or the repetition rate of the excitation of the source.

The MI of the AF of the click train (Figures 9) is presented in Figure 10. The click train is observed to produce minimal activity in the MI except at the very lowest frequencies as described above. This is because the response in the SSI on any vertical line is essentially flat in the case of a click; in fact, the amplitude of the response rises slowly with best frequency because the bandwidth increases with best frequency and the spectrum of a click is flat. The form of the Mellin image does not change with repetition rate except for the upper limit on the frame of auditory figure which decreases with the period of the sound.

Before proceeding with the vowel examples, consider the relationship between the MI presented in this section which is written in terms of frequency-domain integration (Eq. C15) and the Mellin transform presented earlier which is written in terms of time-interval-domain integration (Eq. B12). The time-frequency constraint that is fundamental to the Mellin transform is presented in Eq. C16 in logarithmic terms, and the derivative of this equation is presented in Eq. C17. Substituting first for frequency in Eq. 15, and then for h using Eqs C10 and C11, produces Eq. C18 which has the same integral as the Mellin transform of Eq. B12. The constant differs somewhat but this does not matter to the form of the Mellin Image.

V. THE SAI, SSI AND MI OF A DAMPED SINUSOID

The SAI for a repeating damped sinusoid with an decaying exponential envelope is presented in Figure 11. In this case, the half-life of the damped envelope is 2 ms and the carrier frequency is 2 kHz and the repetition rate is 100 Hz. With these parameter values, the damped sinusoid is like a single-formant vowel. The repeating onsets of the damped sinusoid produce a click-like response in frequency regions away from 2 kHz, and the spacing of the verticals shows the period of the sound. In the region of 2 kHz, the impulse response is accentuated and lengthened by the resonance associated with the decaying exponential envelope. This is a common feature of natural sounds including speech. The SSI of the auditory figure of the damped sinusoid is presented in Figure 12 and much of the activity in channels away from 2 kHz is just like that of the click train. In the 2 kHz channel, however, the activity is extended to higher h values and there is a progressive rotation of the portion of the impulse lines indicating that the instantaneous frequency in this region is neither that of the wavelet kernel, nor the carrier frequency of the channel (except in the 2 kHz channel).

The MI for the damped sinusoid is presented in Figure 13. The activity associated with the repeating onsets of the damped sinusoid appears as activity on and near the abscissa in the same position as activity produced by a click train. The activity associated with the resonance in the 2 kHz region of the SSI produces extra vertical bands in the MI of the damped sinusoid, indicating a broad spatial-frequency response at greater h values than for the click train. The vertical bands in the MI become progressively wider as h increases because, in the SSI, the lines in the fine structure of the feature tilt progressively more as h increases. This is the characteristic of a source with a single resonance, or formant. The banding in the Mellin Images of other damped sinusoids is essentially the same independent of the frequency of the carrier, the half life of the envelope, or the repetition rate of the sound, and this is what is meant by establishing the shape of the source independent of size and repetition rate. The level and extent of the vertical bands increase slowly with the half life of the damped sinusoid. In the next subsection, the example is extended to synthesised vowels based on an area-function model of the vocal tract.

VI. CALCULATION OF THE SSI AND MI FOR FOUR VERSIONS OF THE VOWEL 'a'

A set of four synthetic 'a' vowels were constructed to illustrate the invariance properties of the SSI and Mellin Image. The vowels were all produced by a typical vocal tract model with cross-area functions from one specific male speaker (Yang and Kasuya, 1995). In the case of speech, it is the vocal tract shape that the SSI and MI are intended to capture. One pair of the four vowels had the original vocal tract length, and differed by being excited by streams of glottal pulses with different rates (100 Hz and 160 Hz). Their auditory images are presented in Figures 14 and 15, respectively. The resonances in the vocal tract extend the impulse response in the auditory figure in the frequency region of the resonance. In speech research, these resonances are referred to as formants. The second and third formants of the vowels have centre frequencies of about 1000 and 2200 Hz, respectively. Note that although the main verticals in Figure 15 are closer together than those in Figure 14, reflecting the difference in glottal rate, the positions of the formants do not change with glottal rate. The second pair of 'a' vowels are produced from a vocal tract having the same set of cross-area functions but with the length of the tract shortened by 2/3. The glottal rates of the two vowels are 100 and 160 Hz as before. The auditory images of these vowels are presented in Figures 16 and 17, respectively. The positions of the second and third formants are the same in the two figures, but the positions have moved up by a factor of 3/2 to 1500 and 3300 Hz, respectively, due to the shortening of the vocal tract. The spacing of the main verticals in Figures 16 and 17 are the same as those in Figures 14 and 15, respectively.

The SSI's for the four vowels are presented in Figures 18-21 in the same order as the SAI's. The distinction between the response to the glottal pulses towards the left of the AF and the formants towards the right of the AF is enhanced in these auditory figures. The pattern of information for the vowels with the longer vocal tract (Figures 18 and 19) is essentially the same; the only real difference is in the position of the right-hand boundary of the AF which is determined by the repetition rate of the wave and so is more restrictive in Figure 19 for the vowel with the higher pitch. Similarly, the SSI's for the vowels with the shorter vocal tract (Figures 20 and 21) are essentially the same except for the right-hand boundary of the AF. Moreover, comparing the SSI's for longer and shorter vocal tracts, reveals that the pattern produced by the first four formants is very similar, except for the fact that the pattern in Figures 20 and 21 for the shorter vocal tract is shifted up relative to that in Figures 18 and 19 for the longer vocal tract. The fifth and sixth formants visible in the SSI's of the vowels from the longer vocal tract shift up to frequencies beyond the range of Figures 20 and 21, but they too would be seen to have been shifted by the same fixed amount if the range of the Figure were extended.

The MI's for the four vowels are presented in Figures 22-25 in the same order as the SAI's and SSI's. The ordinate in the MI is the Mellin coefficient $c/2\pi$ and the unit is cycles/frequency-range which means that an ordinate value of unity in the MI corresponds to a spatial frequency in the SSI whose period is the full frequency range of the SSI ordinate from 100 to 6000 Hz. The MI's values for a specific value of h show the spatial frequencies that best fit the activity in the corresponding column of the SSI. For the first few integer multiples of h in the SSI of the vowel 'a', the activity is broadband in response to the glottal pulse and so there is activity at low spatial frequencies below about $c/2\pi$ values of 4. As the value of h rises above 2, the formats begin to appear as separate bands in the SSI and the best fitting spatial frequency rises from around 6 to 18 as h increases from 2 to 8. For values of h above 8, there is only one formant in the SSI and so there are several extra, relatively broad vertical bands in the MI. These are the main characteristics of the vowel 'a' in the MI.

VII. CALCULATION OF THE SSI AND MI FOR THE JAPANESE VOWELS 'a, e, i, o, u'

A set of the five Japanese vowels 'a, e, i, o, u' were synthesised to illustrate how vowel differences appear in the Size-Shape Image and Mellin Image. The same vocal tract model was used with the same male speaker but different cross-area values for the different vowels as specified by

Yang and Kasuya (1995). All five vowels had the original vocal tract length and were excited by 100-Hz streams of glottal pulses. The SAI's for the five vowels 'a, e, i, o, u' are presented in Figures 26-30, respectively; the corresponding SSI's are presented in Figures 31-35; and the corresponding MI's are presented in Figures 36-40. Comparison of the SAI's and SSI's for individual vowels shows that the log transform changes the emphasis on formant duration; for example, in the SAI of the vowel 'a', (Figure 26) the second formant is about three times the duration of the fourth formant, but in the SSI, (Figure 31), the two have about the same extent on the frequency-weighted, time-interval dimension, h. Without the log transform, the higher formants would play very little role in determining the pattern in the MI. The alignment of channels in the SSI also makes it easier to determine when the wavelet impulse response gives way to the resonance properties of the source. Comparison of the SAI's and SSI's for the vowels 'o' (Figures 29 and 34) and 'e' (Figures 27 and 32), shows that the narrow second and third formants in the 'o' produce a deep cancellation valley between them that cuts back through all but the first impulse response line, whereas the cluster of higher formants in 'e' prevents the formation of cancellation valleys and there are three unbroken impulse lines in this region.

The SSI and MI of the vowel 'a' (Figures 31 and 36) were discussed in the previous section. Comparison with the SSI and MI of the vowel 'e' (Figure 32 and 37) shows that the higher formants in the SSI of 'e' (Figure 32) are more closely clustered than for the 'a' (Figure 31), and the formants extend to higher h values in the 'e'. The clustering in the 'e' leads to stronger low spatial frequencies in the MI of 'e' around $c/2\pi$ values of 4. It also leads to stronger higher spatial frequencies in the region of $c/2\pi$ values of 12 and 16. Moreover, the effect extends to much higher h values because the second fourth formants in 'e' both persist to greater h values. The vowel 'i' (Figures 33 and 38) is similar to the vowel 'e' in having a cluster of high frequency formants and in this case they are even more closely bunched. For h values in the range 2-6 this results in activity being concentrated around $c/2\pi$ values of 8, rather than around 5 or 3 as it was for 'e' and 'i', respectively. For h values above 4, the activity shifts up to the region of $c/2\pi$ values of 15 and 20, whereas it was 12 and 16 in 'e'. In 'i', the long fourth formant in the SSI leads so broad vertical bands in the MI extending up to h values beyond 15.

The SSI of the vowel 'o' (Figure 34) shows a large frequency separation between the first and second formants on the one hand, and the remaining set of three upper formants on the other. As a result there is less activity in the MI (Figure 39) at the lowest spatial frequencies (below $c/2\pi$ values of 4). While the first formant is present (for values of h up to 5 in Figure 34), there is activity at $c/2\pi$ values of 5 and 8 in the MI reflecting the spacing between the first and second formants. As these formants die away, the dominant spatial frequencies shift up to the around $c/2\pi$ values of 12 and 20, reflecting the spacing of the higher formants for h values of 4 and above. The lingering cluster of high formants in 'o', leads to diffuse activity at low spatial frequencies extending out to high h values, which distinguishes this vowel from all of the others. The vowel 'u' (Figures 35 and 40) is in many way the simplest because the formant resonances are relatively broad and so they do not extend far out into the SSI or MI. This in itself many be the defining characteristic of this vowel, its lack of distinctive features at large h and $c/2\pi$ values. For h values in the range 2-5 there is strong activity for $c/2\pi$ values around 7, and for h values in the range 4-5 there is strong activity for $c/2\pi$ values around 14. There is also some diffuse vertical banding which is restricted to values of h less than about 9, which among the other vowels only occurs for 'a' (Figure 36).

VIII. SUMMARY

In this paper, we have described a framework of "Stabilised Wavelet-Mellin Transform" and a computational version of the auditory Mellin transform. This involves 1) specifying the role of existing processes and representations, such as the auditory filter bank, the auditory image, and the auditory

figure, in the larger Mellin transform framework, 2) developing an auditory form of the extensions required to complete the Mellin transform, and 3) specifying the relationship between the products of the Mellin transform and the components of auditory perception. The resulting representation is referred to as a 'Mellin Image'; the transform normalises for size and so the image presents source shape information independent of source size. The Mellin Image preserves both spectral and temporal information, so it is likely to be useful in many signal processing applications as well as in auditory research.

APPENDIX A: The Mellin Transform

The Mellin transform (Titchmarsh, 1948) of a signal, $s(t)$ ($t > 0$), is defined as

$$S(p) = \int_0^\infty s(t)\, t^{p-1} dt, \tag{A1}$$

where p is a complex argument. One of the important properties is

$$\text{if } s(t) \Rightarrow S(p), \text{ then } s(at) \Rightarrow a^{-p} S(p), \tag{A2}$$

where the arrow ($\Rightarrow$) indicates "is transformed into" and a is a real dilation constant. That is, the distribution $S(p)$ is just multiplied with a constant a^{-p} when the function $s(t)$ is scaled in time. If p is denoted by $p = p_r + jp_i$,

$$a^{-p} = a^{-(p_r + jp_i)} = a^{-p_r} a^{-jp_i} = a^{-p_r} \exp(-jp_i \ln a) \tag{A3}$$

where $j = \sqrt{-1}$, and exp and ln are the exponential and natural logarithmic operators. Since $|a^{-p} S(p)| = |a^{-p_r}| \cdot |S(p)|$, the absolute distribution $|S(p)|$ id not affected by a scaling of the signal, except for the constant that specifies the scale of the current signal; nor is it affected when the distribution is normalised.

APPENDIX B: Stabilised Wavelet-Mellin Transform

Signal: $s(t)$

Periodicity of signal: $s(t - kt_p) = s(t), \quad \{t > 0, k = 1,2,....,\}$ (B1)

Wavelet kernel: $g(t) = \gamma(t)\exp(j2\pi f_b t + jc_1 \ln t + j\phi), \quad \{t > 0\}$ (B2)

Wavelet filter:
$$\begin{aligned} g_W(\alpha f_b, t) &= g(\alpha t) = \gamma(\alpha t)\exp(j2\pi f_b \alpha t + jc_1 \ln \alpha t + j\phi) \\ &= \gamma(\alpha t)\exp(j2\pi \alpha f_b t + jc_1 \ln t + j\phi_\alpha), \quad \{^\forall \alpha \,|\, \alpha_{\min} \le \alpha \le \alpha_{\max}\} \\ \phi_\alpha &= \phi + c_1 \ln \alpha \end{aligned} \tag{B3}$$

Wavelet transform:

$$S_W(\alpha f_b, t) = \mathcal{W}[s(t)] = \int_0^\infty g_W(\alpha f_b, \tau_1) s(t - \tau_1)\, d\tau_1 \tag{B4}$$

Compression:

$$\begin{aligned} S_W(\alpha f_b, t) &= \log |S_W(\alpha f_b, t)| \cdot \max\{S_W(\alpha f_b, t), 0\} \\ S_W(\alpha f_b, t) &= \log |S_W(\alpha f_b, t)| \cdot \operatorname{sgn}\{S_W(\alpha f_b, t)\} \end{aligned} \tag{B5}$$

$$\begin{aligned} S_W(\alpha f_b, t) &= |S_W(\alpha f_b, t)|^\beta \cdot \max\{S_W(\alpha f_b, t), 0\}, \quad \{\beta \,|\, 0 < \beta \le 1\} \\ S_W(\alpha f_b, t) &= |S_W(\alpha f_b, t)|^\beta \cdot \operatorname{sgn}\{S_W(\alpha f_b, t)\}, \quad \{\beta \,|\, 0 < \beta \le 1\} \end{aligned} \tag{B6}$$

Stabilised Wavelet Transform:

$$A_I(\alpha f_b, \tau) = \sum_{k=0}^{\infty} S_W(\alpha f_b, \tau + kt_p) e^{-\xi\tau} e^{-k\eta t_p}, \quad \{^\forall \tau \,|\, \tau_{\min} \le \tau \le \tau_{\max}\} \tag{B7}$$

Source Figure:

$$A_F(\alpha f_b,\tau) = A_I(\alpha f_b,\tau_2), \quad \{^{\exists}k,\tau,\tau_2 \mid kt_p \le \tau_2 \le (k+1)t_p,\ 0 \le \tau \le t_p\} \tag{B8}$$

Constraint for integration: $\alpha f_r \cdot \tau = h$ (B9)

Mellin transform:

$$M_I(h,p) = \mathcal{M}[A_F(h/\tau,\tau)] = \int_0^{t_p} A_F(h/\tau,\tau)\,\tau^{p-1}d\tau$$
$$= \int_0^{t_p} A_F(h/\tau,\tau)\,e^{(p-1)\ln\tau}d\tau \tag{B10}$$

$$p = -jc + (\mu + \tfrac{1}{2}) \tag{B11}$$

$$M_I(h,c) = \int_0^{t_p} A_F(h/\tau,\tau)\,e^{(-jc+\mu-\frac{1}{2})\ln\tau}d\tau \tag{B12}$$

APPENDIX C: Stabilised Wavelet-Mellin Transform based on Auditory Image Model

Auditory filter:

$$g_c(t) = at^{n-1}\exp(-2\pi b\mathrm{ERB}(f_r)t)\exp(j2\pi f_r t + jc_1\ln t + j\phi), \quad \{t > 0\}$$
$$\mathrm{ERB}(f_r) = 24.7 + 0.108 f_r \tag{C1}$$

$$f_r \cong \alpha f_b \quad \{f_r \mid f_r > 500\ \mathrm{Hz}\} \tag{C2}$$

$$g_c(f_r,t) \cong g_W(\alpha f_b,t), \quad \{^{\forall}\alpha \mid \alpha_{\min} \le \alpha \le \alpha_{\max}\} \tag{C3}$$

Auditory wavelet transform:

$$S_W(\alpha f_b,t) = \mathcal{W}[s(t)] = \int_0^{\infty} g_W(\alpha f_b,\tau_1)s(t-\tau_1)\,d\tau_1 \tag{C4}$$

Adaptation / Compression :

$$S_W(\alpha f_b,t) \cong \log|S_W(\alpha f_b,t)| \cdot \max\{S_W(\alpha f_b,t),0\} \tag{C5}$$

$$S_W(\alpha f_b,t) \cong |S_W(\alpha f_b,t)|^{\beta} \cdot \max\{S_W(\alpha f_b,t),0\}, \quad \{\beta \mid 0 < \beta \le 1\} \tag{C6}$$

Stabilised Auditory Image:

$$A_I(\alpha f_b,\tau) = \sum_{k=0}^{\infty} S_W(\alpha f_b,\tau + kt_p)e^{-\xi\tau}e^{-k\eta t_p}, \quad \{^{\forall}\tau \mid \tau_{\min} \le \tau \le \tau_{\max}\} \tag{C7}$$

Filter alignment:

$$A_{IA}(\alpha f_b,\tau) = A_I(\alpha f_b,\tau_2 - k/\alpha f_b), \quad \{^{\exists}k \mid k = 0,1,2,\ldots\} \tag{C8}$$

Auditory Figure (AF):

$$A_F(\alpha f_b,\tau) = A_{IA}(\alpha f_b,\tau_2), \quad \{^{\exists}k,\tau,\tau_2 \mid kt_p \le \tau_2 \le (k+1)t_p,\ 0 \le \tau \le t_p\} \tag{C9}$$

AF with log time-interval:

$$A_{FL}(\alpha f_b,\ln\tau) = A_F(\alpha f_b,\tau) \tag{C10}$$

Impulse alignment in Log AF & Size Shape Image (SSI):

$$A_{SSI}(\alpha f_b,h) = A_{FL}(\alpha f_b,\ln\tau + \ln\alpha f_b) \tag{C11}$$

Impulse Profile of SSI:

$$P_{AI}(h) = \int_{\alpha_{\min}}^{\alpha_{\max}} W(\alpha f_b,h)\cdot A_{SSI}(\alpha f_b,h)\,d\alpha \tag{C12}$$

Spectral Profile of SSI:

$$P_{AS}(\alpha f_b) = \int_{h_{\min}}^{h_{\max}} W(\alpha f_b, h) \cdot A_{SSI}(\alpha f_b, h)\, dh \tag{C13}$$

Sinusoidal weighting along log-frequency axis:

$$W(\alpha f_b, h, c) = e^{\{jc-(\mu-\frac{1}{2})\}\ln \alpha f_b} \tag{C14}$$

Mellin Image:

$$M_I(h,c) = \int_{\alpha_{\min} f_b}^{\alpha_{\max} f_b} A_{SSI}(\alpha f_b, h) e^{\{jc-(\mu-\frac{1}{2})\}\ln \alpha f_b}\, d(\alpha f_b) \tag{C15}$$

Showing Equivalence:

Constraint from Eq. (B9): $\ln \alpha f_b + \ln \tau = \ln h$ (C16)

$$\frac{1}{\alpha}\frac{d\alpha}{d\tau} = -\frac{1}{\tau} \tag{C17}$$

$$\begin{aligned} M_I(h,c) &= \int_{\alpha_{\min}}^{\alpha_{\max}} A_{SSI}(\alpha f_b, h) e^{\{jc-(\mu-\frac{1}{2})\}\ln \alpha f_b}\, f_b d\alpha \\ &= \int_{\alpha_{\min}}^{\alpha_{\max}} A_{SSI}(\alpha f_b, h) e^{\{jc-(\mu+\frac{1}{2})\}\ln \alpha f_b}\, (1/\alpha) d\alpha \\ &= \int_0^{t_p} A_F(\alpha f_b, \tau) e^{\{jc-(\mu+\frac{1}{2})\}(\ln h - \ln \tau)}\, (-1/\tau) d\tau \\ &= \{-e^{\{jc-(\mu+\frac{1}{2})\}\ln h}\} \int_0^{t_p} A_F(h/\tau, \tau) e^{\{-jc+(\mu-\frac{1}{2})\}\ln \tau} d\tau \end{aligned} \tag{C18}$$

APPENDIX D: Terminology for the components of an Auditory Mellin Transform

Pitch is the psychological correlate of the **repetition rate** of the sound source.

Loudness is the psychological correlate of the **intensity** of the sound source.

Timbre, or **sound quality**, is the psychological correlate of what enables one to distinguish sounds that have the same pitch and loudness. (This is a paraphrase of the ANSI official definition of timber.)

Auditory Filter: A wavelet filter is one whose impulse response dilates or compresses linearly in time with the peak frequency of the filter (Combes et. al, 1989). An **auditory filter** is a wavelet filter whose envelope is a gamma function and whose is fine-structure is 1) a sinusoid at the peak frequency of the filter (gammatone auditory filter, Patterson et al., 1988, 1995), or 2) for optimality in time-frequency trading, a chirp that asymptotes to the peak frequency of the filter (gammachirp auditory filter, Irino and Patterson, 1997). If the bandwidth of the filter is about 10% of the peak frequency and the order of the filter, n, is 4, the filter can explain a wide variety of physiological and psychophysical data on hearing (Patterson et al., 1992).

Auditory Filterbank: A wavelet transform is a multi-channel array of wavelet filter outputs spaced evenly along a log-best-frequency axis (Combes et. al, 1989). A wavelet transform is a 'constant-Q' filter system throughout the frequency range of the transform. An **auditory filterbank** is a wavelet transform in which the bandwidths and best frequencies of the filters are set to simulate the spectral analysis performed by the basilar partition in the cochlea; for humans with normal hearing the values

are specified by the ERB and ERB-rate functions described by Glasberg and Moore (1990). It is a 'constant-Q' system with filters spaced in accordance with peak frequency in range above about 500 Hz. Below 500 Hz, the decrease in filter bandwidth is retarded, relative to constant-Q, and the function asymptotes to a constant 24.7 Hz. The same deceleration is applied to filter spacing. In auditory models like AIM, the auditory wavelet transform is referred to as an 'auditory filterbank' and it is used to simulate the motion of the basilar membrane in response to a sound.

Neural Activity Pattern (NAP): The version of the wavelet transform in the auditory nerve at the level of the cochlear nucleus includes amplitude compression, adaptation and lateral suppression or inhibition and it is referred to as the NAP. In AIM, these processes are simulated with log amplitude compression and two-dimensional adaptive thresholding (2D-AT) (Holdsworth and Patterson, 1993; Patterson and Holdsworth, 1996). The source characteristics appear as a distinctive pattern of time-intervals in the NAP and the pattern repeats once per cycle of the sound.

Stabilised Auditory Image (SAI): The repeating patterns in the NAPs of tonal sounds are stabilised by a process of strobed temporal integration which preserves the time-interval patterns in the NAP (Patterson and Holdsworth, 1993; Patterson et al. 1992). The result is the **stabilised auditory image (SAI).**

Auditory Image Frame: In the computational model, the current state of the simulated auditory image is calculated regularly every m milliseconds, and each of the successive displays of the auditory image is referred to as a frame of the auditory image, or auditory image frame. The internal representation of the auditory image exists continuously and is updated asynchronously in each channel.

Auditory Figure (AF): When a tonal sound is mixed with background noise, the repeating time-interval pattern produced by the tonal component of the combined sound appears, in the auditory image, as a figure standing out against a background of irregular time intervals produced by the noise component of the sound. The 'auditory figure' concept is central to the concept of the source shape image (SSI), and the SSI is the basis of the impulse profile and the resonance profile.

Auditory Figure Frame (AFF) or Auditory Image Frame: The frame that limits the extent of the auditory figure in the auditory image is produced by repetition rate of the source, and its width is the period of the source. STI identifies the period and creates the auditory figure frame. It is any one of the regions bounded by the verticals associated with the period of the sound.

Spectral and Temporal Profiles of auditory images and figures: The dimensions of the auditory image are linear time and log frequency. We can integrate over time, across the full width of the auditory image to produce a **spectral profile** (**SP**) of the auditory image, or across an auditory figure frame to produce a **spectral profile** of the auditory figure. Conversely, we can integrate across frequency to produce **temporal profiles** (**TP**) of the auditory image or the auditory figure. The auditory figure reveals properties of the impulse responses of the auditory filters and the excitation of the sound source such as glottal pulses. This suggests information about excitation might be obtained form the temporal profile of the auditory figure, and Meddis and Hewitt (1992) showed that, despite their strange appearance, these temporal profiles in the form of summary correlograms had information about vowel quality that improved vowel identification with a primitive recogniser.

This form of the temporal profile, however, can be expected to be of limited use because integrating across filter channels with different carrier frequencies smears the fine structure (see, for example, the temporal profile of a damped sound in Patterson, 1994, or Irino and Patterson, 1996). The Mellin transform provides a solution to this problem.

Impulse Lines of the auditory image: The auditory figure of a click train is a set of sloping curved lines connecting peaks in the impulse responses of the individual filters across channels. We refer to these as **impulse lines** of the auditory image, or figure, and note the product of 'filter best frequency' and 'time interval' is constant for any given impulse line.

Size-shape Image (SSI): Tonal sources with differing size but the same shape produce auditory figures with the same shape but different sizes and positions on the impulse lines of the auditory image. It is possible to produce a straightened, aligned version of the auditory figure frame in which the impulse lines are vertical and parallel; that is, the kernels of the wavelet filters are scaled and aligned. In this representation, the figure shape remains fixed and only the vertical position changes with source size. This is accomplished by 1) applying a log transform to the time-interval dimension which straightens the lines and renders them parallel with uniformly spacing, and then 2) aligning the lines to be vertical. The operation is limited to one auditory figure frame. The resulting version of the auditory figure is referred to as **a size-shape image (SSI)**.

Spectral profile and **Impulse profile** of the SSI: The vertical profile of the SSI is a pitch-synchronous **spectral profile** of the signal that may provide a better basis as a frame of a spectrogram than the Fourier magnitude spectra or mel-frequency cepstral coefficients. The horizontal profile of the SSI is referred to as the **impulse profile** of the size-shape image. The impulse profile of the auditory figure as it appears in the SSI is invariant with source size (although it is not itself the Mellin transform). The spectral profile varies with source size, inasmuch as it shifts vertically on the impulse lines, but it does not change its shape. So the centroid of the spectral profile might be a good measure of source size. Moreover, an image that is completely invariant with source size could be computed by plotting the figure relative to its vertical centroid. The **Impulse Profile** is a more promising profile of the auditory figure than the temporal profile of the auditory image. Integration along impulse lines preserves the distinction between auditory figure activity that is associated with the gammachirp filterbank that did the analysis (which appears as activity on impulse lines of integer value), and activity due to the source (which appears on impulse lines of integer and non-integer value).

SSI Spectrogram: This is new form of spectrogram for speech recognition. Each 'spectral' vector of the SSI spectrogram is the concatenation of the spectral profile and the impulse profile of the auditory figure in the current SSI frame.

REFERENCES

Altes (1978). "The Fourier-Mellin transform and mammalian hearing," J. Acoust. Soc. Am., 63,pp.174-183.

Cohen (1993)."The scale transform," IEEE Trans. Acoust. Speech and Signal Processing.41,pp.3275-3292.

Combes, Grossmann and Tchamitchian Eds. (1989) "Wavelets," Springer-Verlag, Berlin.

Gabor (1946)."Theory of communication," J. IEE (London), 93,42-457.

Irino (1995),"An optimal auditory filter," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, NY.

Irino (1996),"A 'gammachirp' function as optimal auditory filter with the Mellin transform," IEEE ICASSP-96.

Irino, T. and Patterson, R.D. (1994). A computational theory of auditory event detection. J. Acoust. Soc. Am., 95, 2943.

Irino, T. and Patterson, R.D. (1996). "Temporal asymmetry in the auditory system," J. Acoust. Soc. Am. 99, 2316-2331.

Irino, T. and Patterson, R.D. (1997). "A time-domain. level-dependent auditory filter: the gammachirp," J. Acoust. Soc. Am. 101, 412-419.

Kawahara, H., Katayose, H., Patterson, R.D., and de Cheveigne, A. (1998) 'Equilibrium points of frequency-to-instantaneous-frequency mapping and its application to accurate F0 extraction,' ICSLP98, Sydney.

Glasberg and Moore (1990). "Derivation of auditory filter shapes from notched-noise data," Hear. Res., 47, pp. 103-138.

Nakata, K. (1995). "Speech, (revised)", Corona, Tokyo.

Patterson, R.D. (1994)." The sound of a sinusoid: Time-interval models," J. Acoust. Soc. Am., 96, pp. 1419-1428.

Patterson, R.D. (1987)." A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am., 82, pp. 1560-1586.

Patterson, R.D. and Hirahara, T (1989). HMM speech recognition using DFT and auditory spectrograms. ATR Technical Report, Kyoto, Japan.

Patterson, R.D., Holdsworth, J. and Allerhand, M. (1992) 'Auditory Models as preprocessors for speech recognition', In: The Auditory Processing of Speech: From the auditory periphery to words, M. E. H. Schouten (ed), Mouton de Gruyter, Berlin, 67-83.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand M. (1992) 'Complex sounds and auditory images', In: Auditory physiology and perception, Proceedings of the 9h International Symposium on Hearing, Y Cazals, L. Demany, K. Horner (eds), Pergamon, Oxford, 429-446.

Patterson, R.D., Anderson, T., and Allerhand, M. (1994). "The auditory image model as a preprocessor for spoken language," in Proc. Third ICSLP, Yokohama, Japan, 1395-1398.

Patterson, R.D. and J. Holdsworth (1996). A functional model of neural activity patterns and auditory images. In: Advances in Speech, Hearing and Language Processing , (W. A. Ainsworth, ed.), Vol 3. Part B. JAI Press, London. (in press from 1991- 1996).

Patterson, R.D., Allerhand, M., and Giguére, C., (1995). "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," J. Acoust. Soc. Am. 98, 1890-1894.

Patterson, R.D., Anderson, T.R and Francis K. (1996) "Binaural auditory images and a noise-resistant, binaural auditory spectrogram for speech recognition," The Auditory Basis of Speech Perception, Eds. W. Ainsworth and S. Greenberg, ESCA Workshop, Keele, July 1996.

Patterson, R.D. and Irino, T. (1998). "Auditory Temporal Asymmetry and Autocorrelation," In: Psychophysical and physiological advances in hearing, Eds. A. Palmer, A. Rees, Q. Summerfield and R. Meddis. Whurr, London, 554-562.

Patterson, R.D. and Irino, T. (1998). "Modelling temporal asymmetry in the auditory system," J. Acoust. Soc. Am. 104, 2967-2979.

Robinson, A., Holdsworth, J., Patterson, R. and Fallside, F. (1990). A comparison of preprocessors for the Cambridge recurrent-error-propagation-network speech recognition system. Proceedings of First ICSPL, Kobe, Japan.

Umesh, Cohen, and Nelson (1997)."Frequency-warping and speaker-normalization," IEEE ICASSP-97.

Umesh, Cohen, and Nelson (1998). "Improved scale-cepstral analysis in speech," IEEE ICASSP-98.

Titchmarsh (1948)."Introduction of the Theory of Fourier Integrals, "Oxford University Press, London.

Tsuzaki, M. and Patterson, R.D. (1998). "Jitter Detection: A brief review and some new experiments," In: Psychophysical and physiological advances in hearing, Eds. A. Palmer, A. Rees, Q. Summerfield and R. Meddis. Whurr, London, 546-553.

Yang and Kasuya (1995)."Dimension differences in the vocal tract shapes measure from MR images across boy, female and male subjects," J. Acoust. Soc. Jpn (E), 16, pp.41-44.

FIGURE CAPTIONS

Figure 1: Block diagram for Stabilised Wavelet Mellin Transform.

Figure 2: Block diagram for Stabilised Wavelet Transform.

Figure 3: An example of speech recognition system based on Auditory Stabilised Wavelet Mellin Transform, or Mellin Image.

Figure 4: Block diagram for Auditory Figure and Size-Shape Image.

Figure 5: Stabilised Auditory Image (SAI) of a click train with the repetition rate of 100 Hz.

Figure 6: Alignment for the auditory filter response.

Figure 7: Transformed version of the leftmost Auditory Figure (AF) from the SAI in Figure 5 with in logarithmic time-interval axis, i.e. produced from Figure 6. The vertical solid line is the upper boundary of the AF.

Figure 8: Size-Shape Image (SSI) produce from Figure 7 by aligning to reorient the diagonal of log-time-interval to the vertical. The abscissa is h value, i.e. the product of time-interval and the best frequency of the channel. The solid curve is the upper boundary of the AF.

Figure 9: Size-Shape Image (SSI) with linear h axis. This is equivalent to Figure 8 but is directly produced from Figure 5 by resampling of sample points in each channel. The solid curve is the upper boundary of the AF.

Figure 10: Mellin Image (MI) calculated from Figure 9. The ordinate is the Mellin coefficient $c/2\pi$ and the unit is cycles/frequency-range which means that an ordinate value of unity in the MI corresponds to a spatial frequency in the SSI whose period is the full frequency range of the SSI ordinate from 100 to 6000 Hz.

Figure 11: Stabilised Auditory Image (SAI) of a damped sinusoid with the half-life of 2 ms, the carrier frequency of 2000 Hz, and the repetition rate of 100 Hz.

Figure 12: Size-Shape Image (SSI) of the damped sinusoid in Figure 11.

Figure 13: Mellin Image (MI) of the damped sinusoid in Figure 11.

Figure 14: Stabilised Auditory Image (SAI) of a synthetic vowel 'a' produced by a vocal-tract model using the original vocal-tract area function of one male subject and the glottal-pulse rate of 100 Hz.

Figure 15: Stabilised Auditory Image (SAI) of a synthetic vowel 'a' with the original vocal-tract length and the glottal-pulse rate of 160 Hz.

Figure 16: Stabilised Auditory Image (SAI) of a synthetic vowel 'a' with the 2/3 compressed vocal-tract length and the glottal-pulse rate of 100 Hz.

Figure 17: Stabilised Auditory Image (SAI) of a synthetic vowel 'a' with the 2/3 compressed vocal-tract length and the glottal-pulse rate of 160 Hz.

Figure 18: Size-Shape Image (SSI) of the synthetic vowel 'a' in Figure 14.

Figure 19: Size-Shape Image (SSI) of the synthetic vowel 'a' in Figure 15.

Figure 20: Size-Shape Image (SSI) of the synthetic vowel 'a' in Figure 16.

Figure 21: Size-Shape Image (SSI) of the synthetic vowel 'a' in Figure 17.

Figure 22: Mellin Image (MI) of the synthetic vowel 'a' in Figure 14.

Figure 23: Mellin Image (MI) of the synthetic vowel 'a' in Figure 15.

Figure 24: Mellin Image (MI) of the synthetic vowel 'a' in Figure 16.

Figure 25: Mellin Image (MI) of the synthetic vowel 'a' in Figure 17.

Figure 26: Stabilised Auditory Image (SAI) of a synthetic vowel 'a' produced by a vocal-tract model using the original vocal-tract area function of one male subject and the glottal-pulse rate of 100 Hz. This is the same as Figure 14.

Figure 27: Stabilised Auditory Image (SAI) of a synthetic vowel 'e' by using the vocal-tract area function of the same male subject and the glottal-pulse rate of 100 Hz.

Figure 28: Stabilised Auditory Image (SAI) of a synthetic vowel 'i' by using the vocal-tract area function of the same male subject and the glottal-pulse rate of 100 Hz.

Figure 29: Stabilised Auditory Image (SAI) of a synthetic vowel 'o' by using the vocal-tract area function of the same male subject and the glottal-pulse rate of 100 Hz.

Figure 30: Stabilised Auditory Image (SAI) of a synthetic vowel 'u' by using the vocal-tract area function of the same male subject and the glottal-pulse rate of 100 Hz.

Figure 31: Size-Shape Image (SSI) of the synthetic vowel 'a' in Figure 26. This is the same as Figure 18.

Figure 32: Size-Shape Image (SSI) of the synthetic vowel 'e' in Figure 27.

Figure 33: Size-Shape Image (SSI) of the synthetic vowel 'i' in Figure 28.

Figure 34: Size-Shape Image (SSI) of the synthetic vowel 'o' in Figure 29.

Figure 35: Size-Shape Image (SSI) of the synthetic vowel 'u' in Figure 30.

Figure 36: Mellin Image (MI) of the synthetic vowel 'a' in Figure 26. This is the same as Figure 22.

Figure 37: Mellin Image (MI) of the synthetic vowel 'e' in Figure 27.

Figure 38: Mellin Image (MI) of the synthetic vowel 'i' in Figure 28.

Figure 39: Mellin Image (MI) of the synthetic vowel 'o' in Figure 29.

Figure 40: Mellin Image (MI) of the synthetic vowel 'u' in Figure 30.

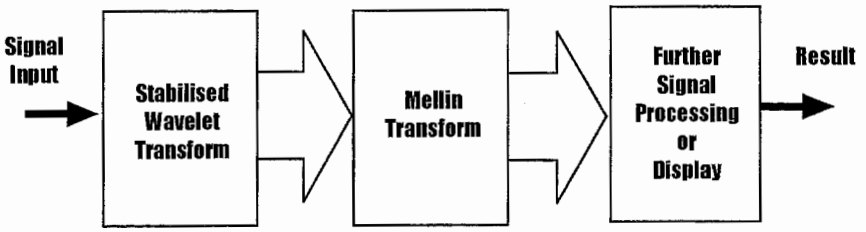
Signal Input
Stabilised Wavelet Transform
Mellin Transform
Further Signal Processing or Display
Result

Fig.1

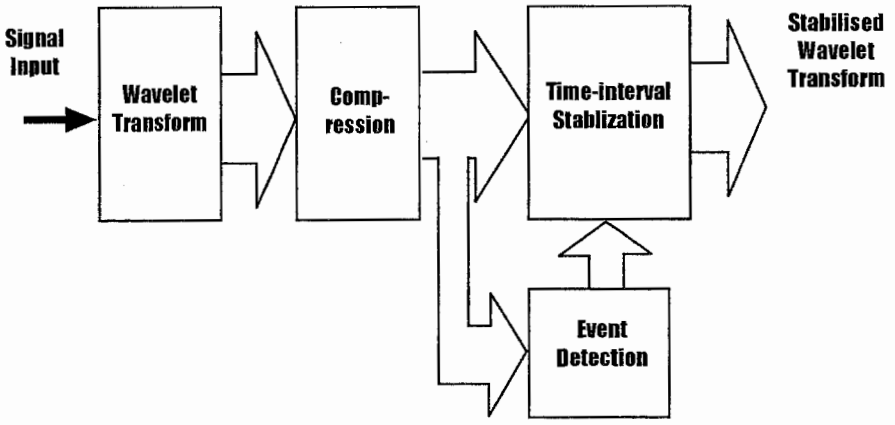
Signal Input
Wavelet Transform
Comp-ression
Time-interval Stablization
Stabilised Wavelet Transform
Event Detection

Fig. 2

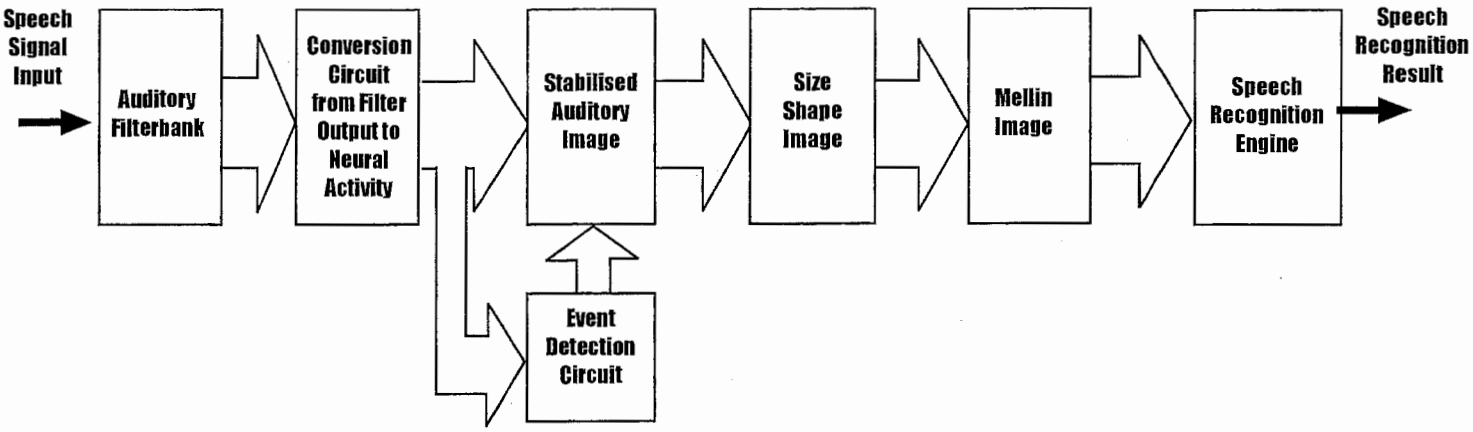
Speech Signal Input
Auditory Filterbank
Conversion Circuit from Filter Output to Neural Activity
Stabilised Auditory Image
Size Shape Image
Mellin Image
Speech Recognition Engine
Speech Recognition Result
Event Detection Circuit

Fig.3

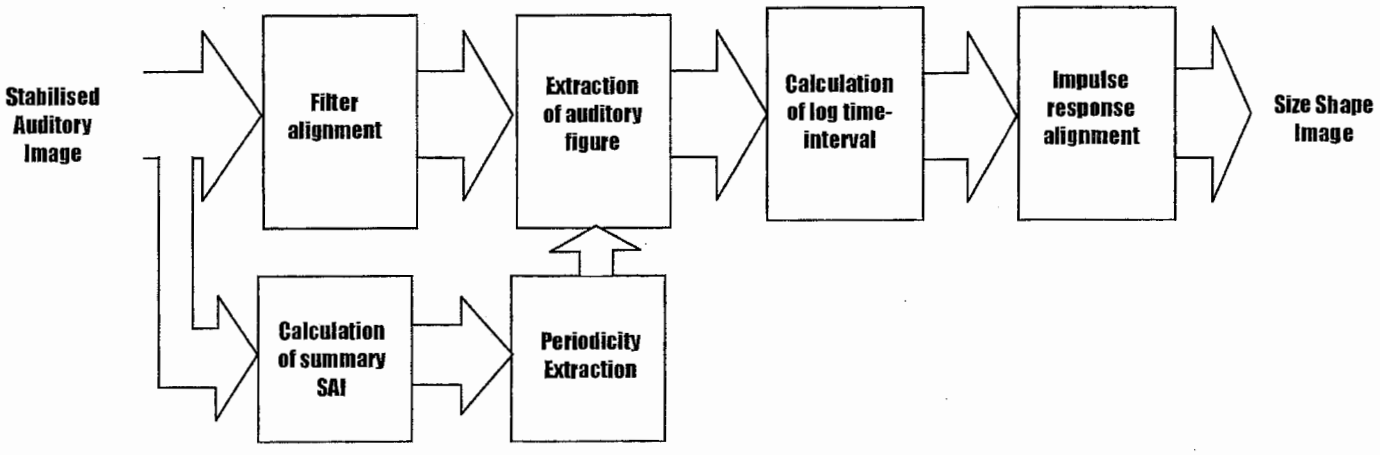
Stabilised Auditory Image
Filter alignment
Extraction of auditory figure
Calculation of log time-interval
Impulse response alignment
Size Shape Image
Calculation of summary SAI
Periodicity Extraction

Fig. 4

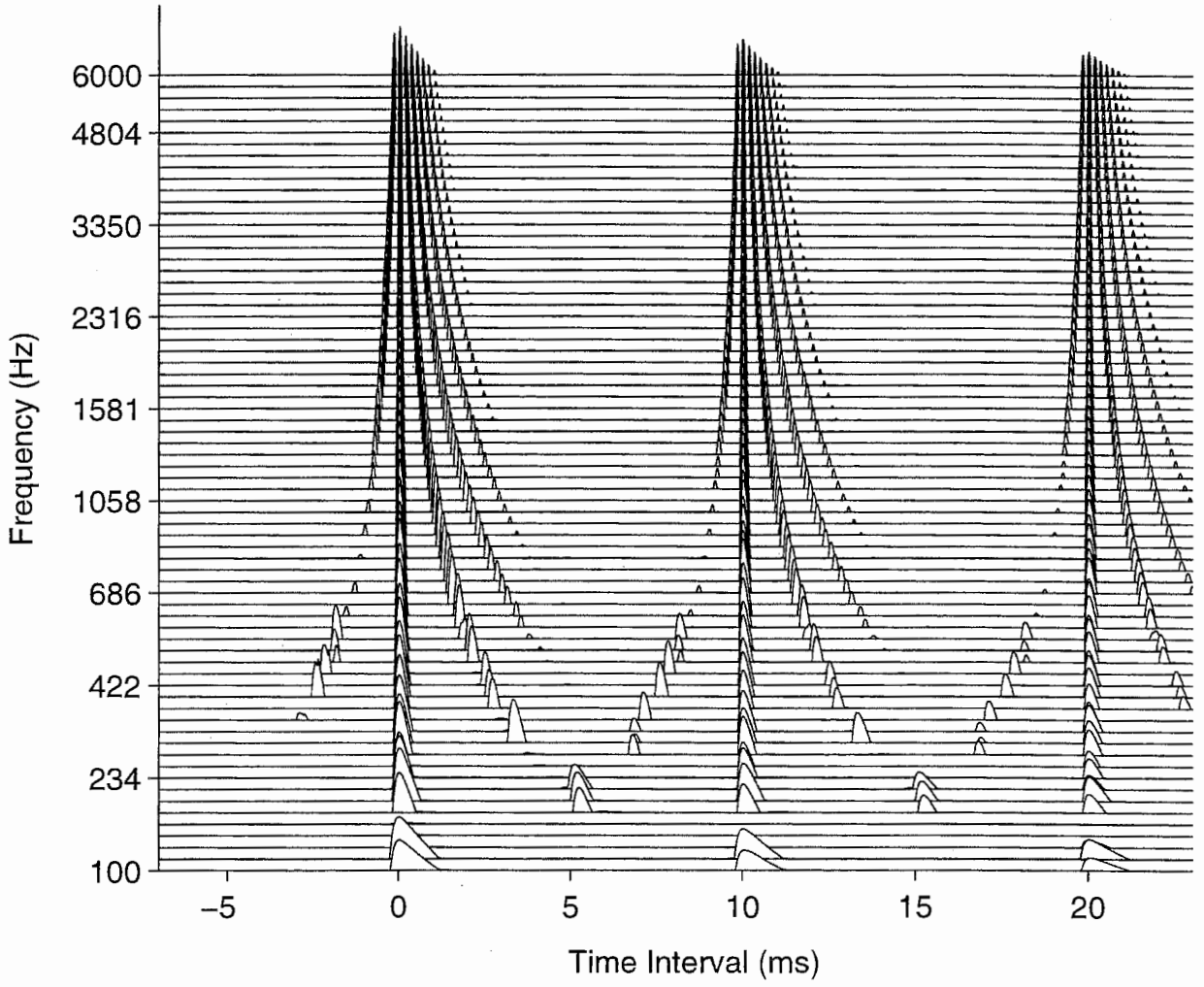
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
−5
0
5
10
15
20
Time Interval (ms)

Fig. 5

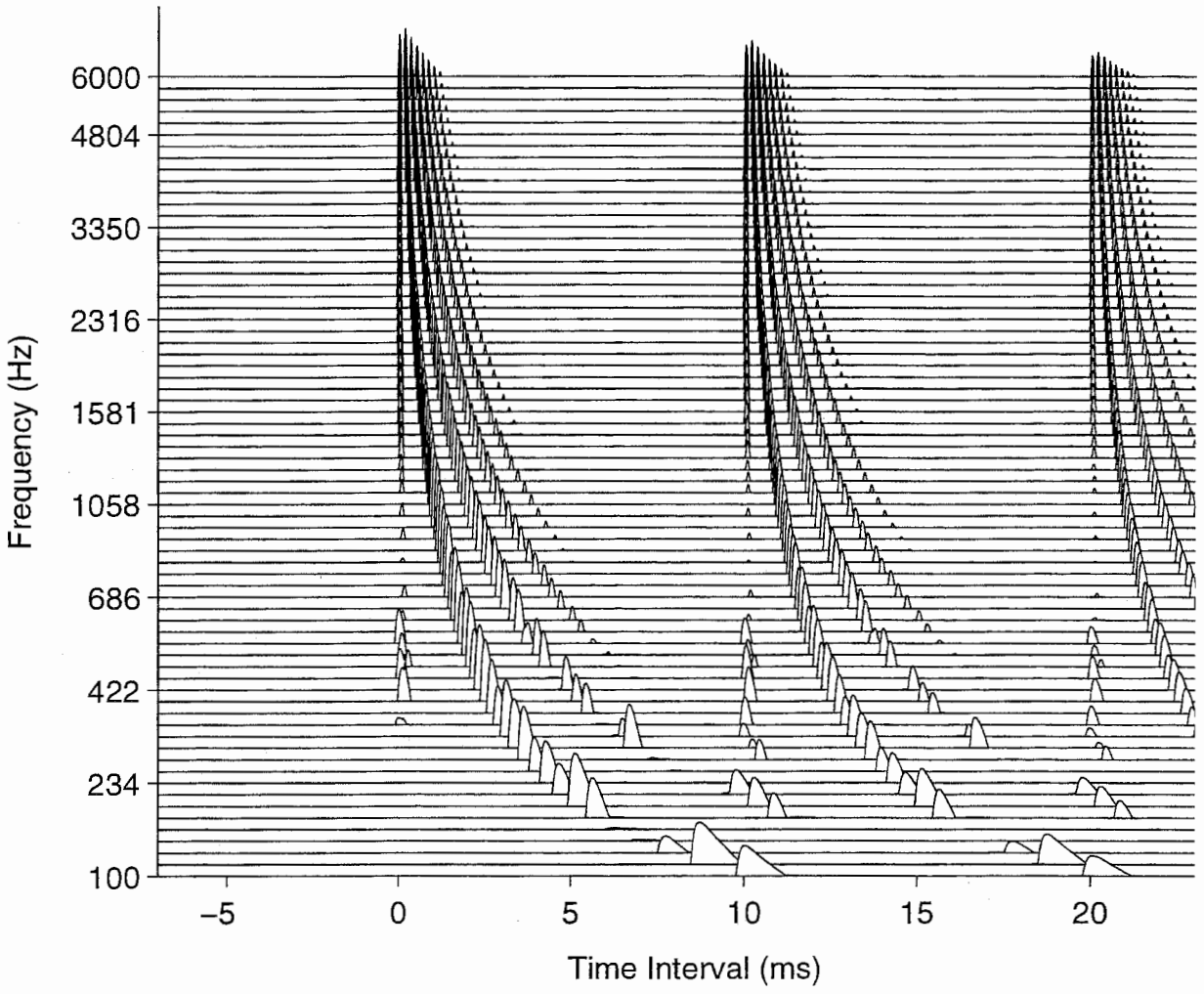
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
−5
0
5
10
15
20
Time Interval (ms)

Fig. 6

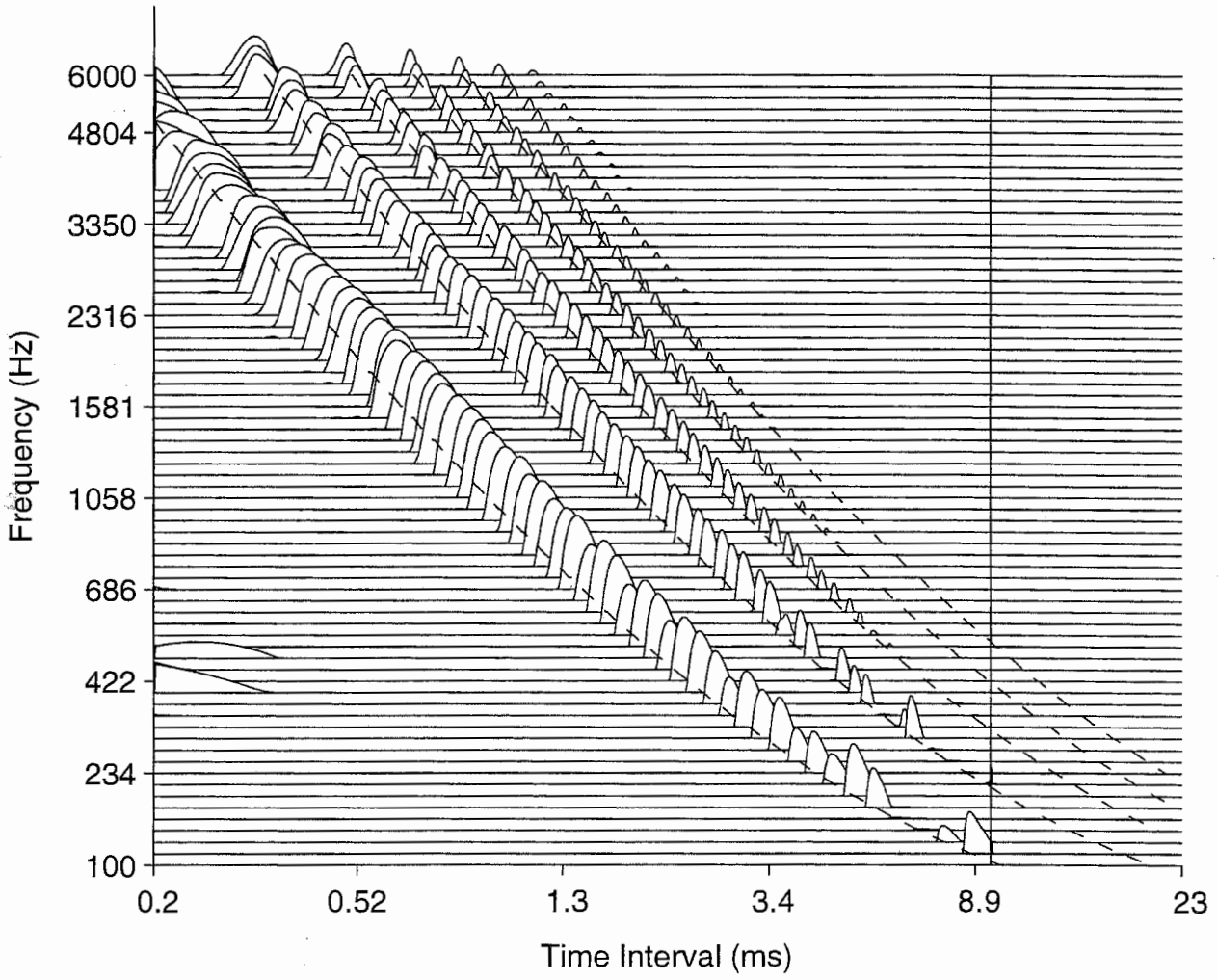
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
0.2
0.52
1.3
3.4
8.9
23
Time Interval (ms)

Fig. 7

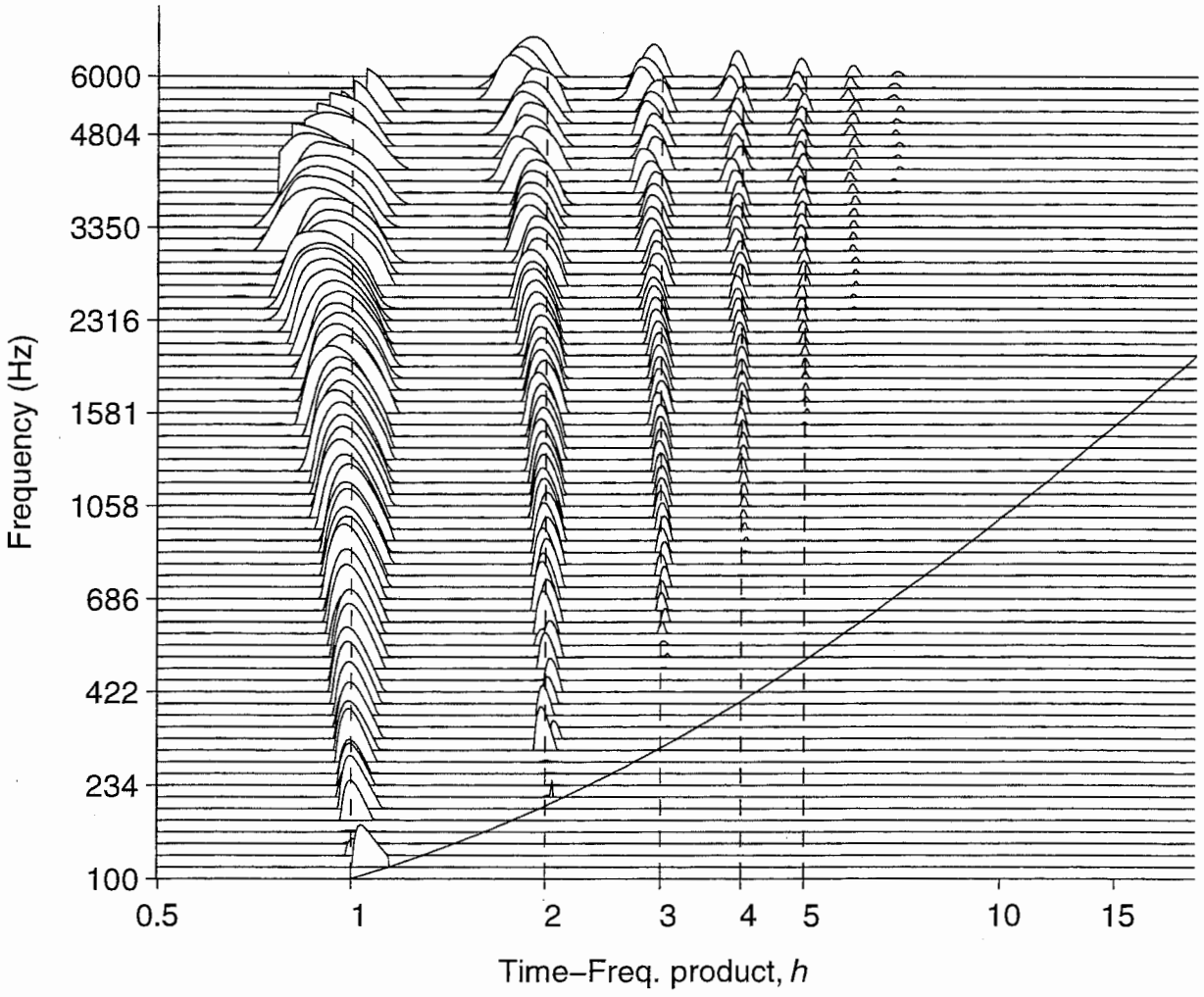
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
0.5
1
2
3
4
5
10
15
Time–Freq. product, h

Fig. 8

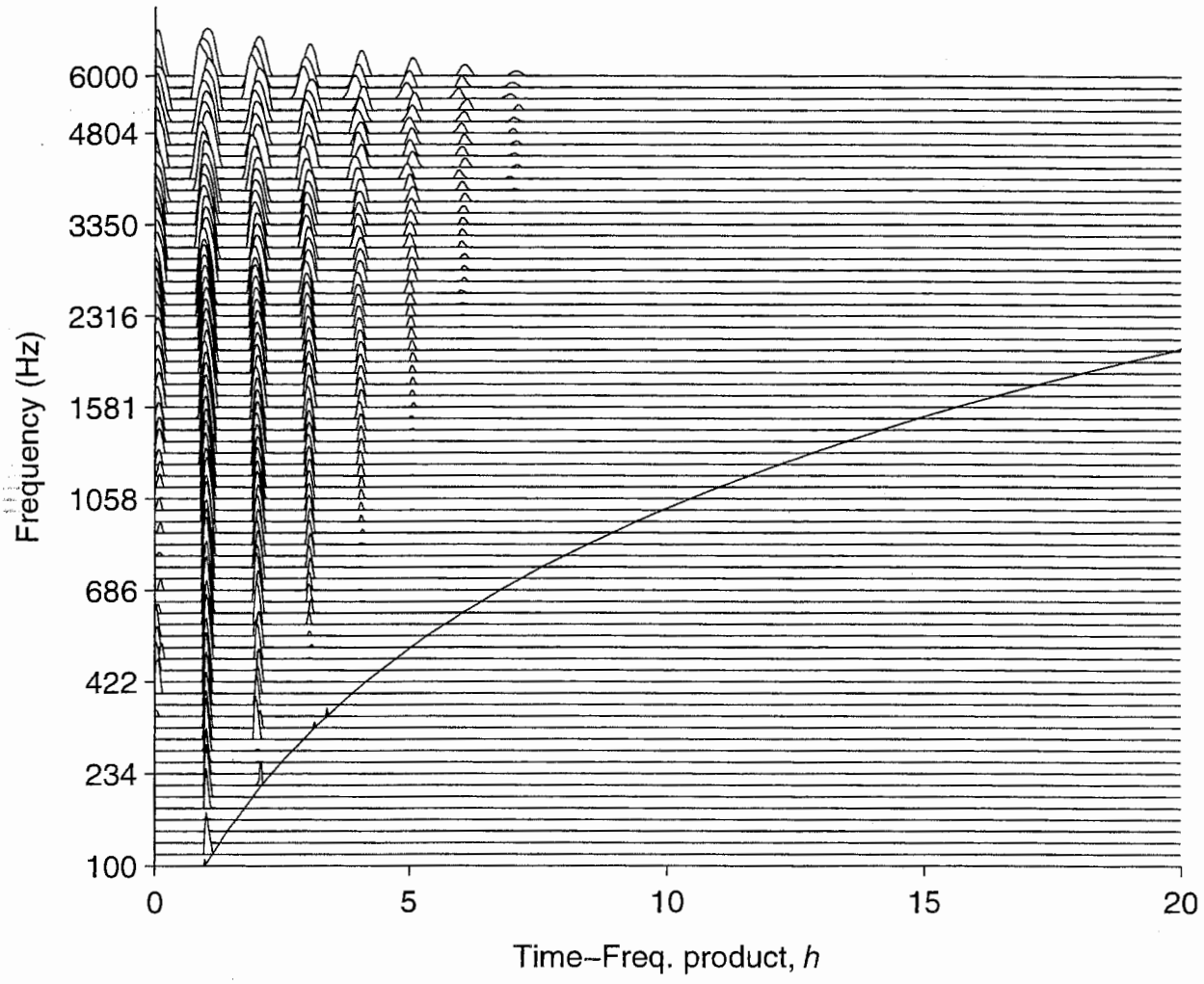
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
0
5
10
15
20
Time–Freq. product, h

Fig. 9

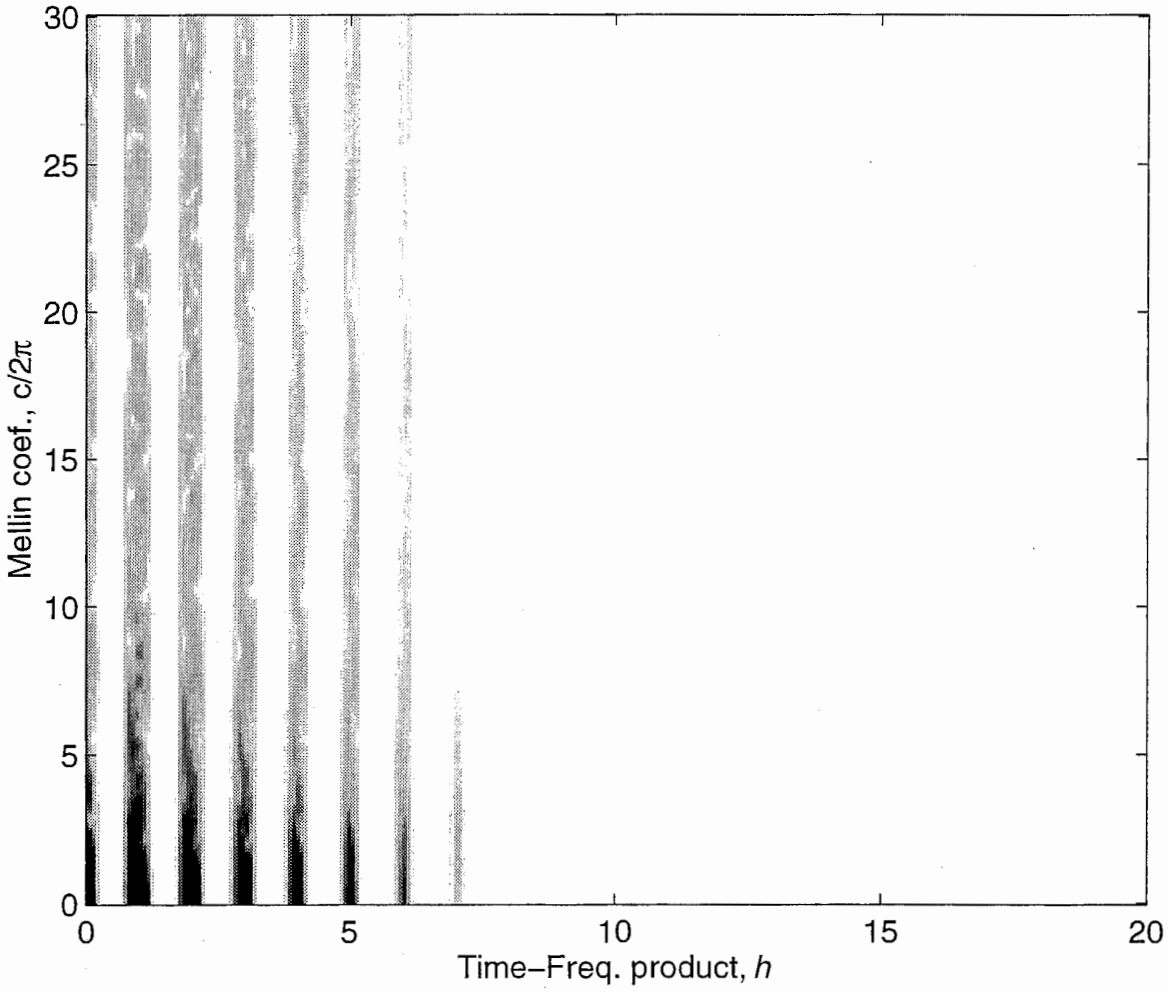

Fig. 10

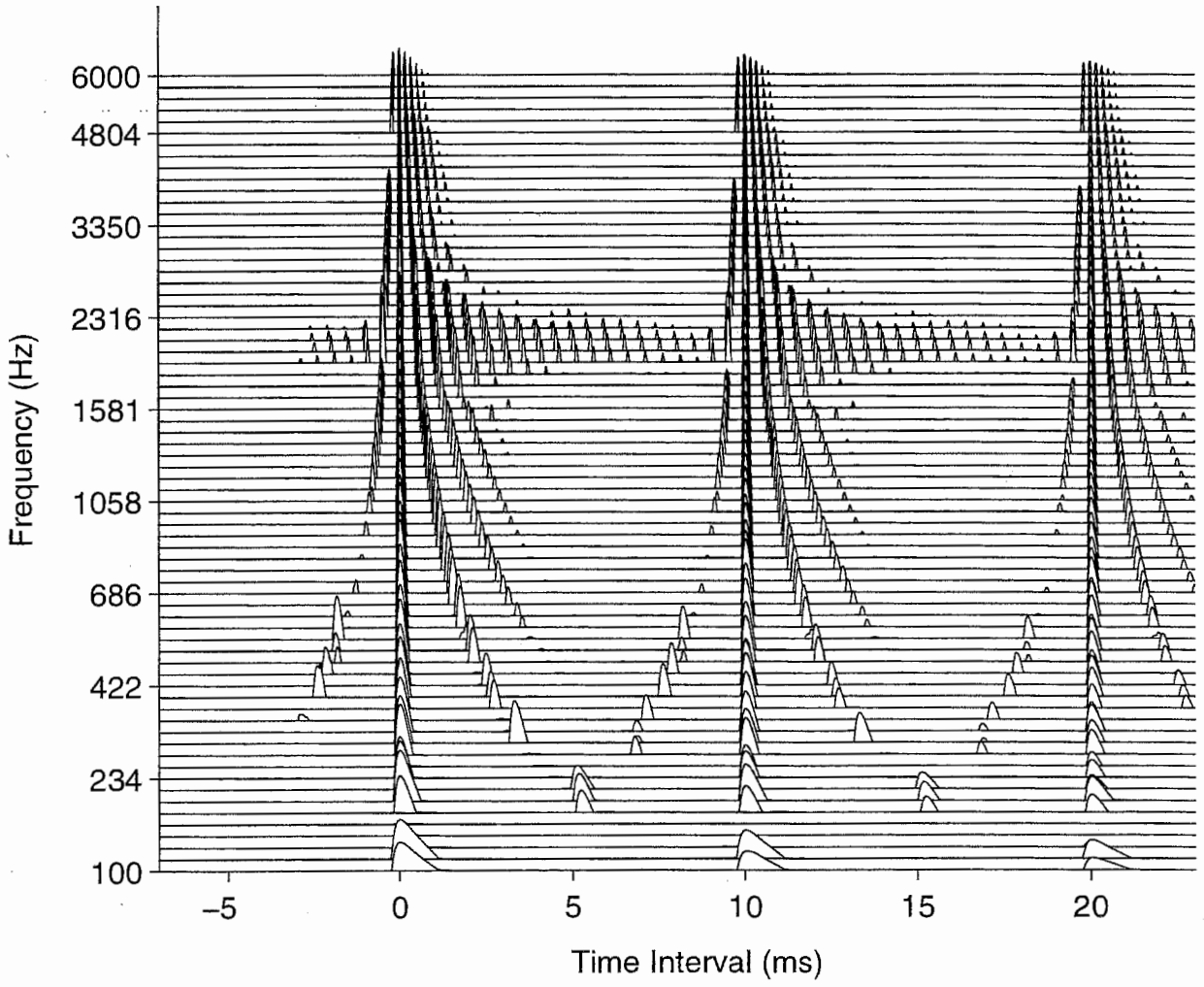
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
–5
0
5
10
15
20
Time Interval (ms)

Fig. 11

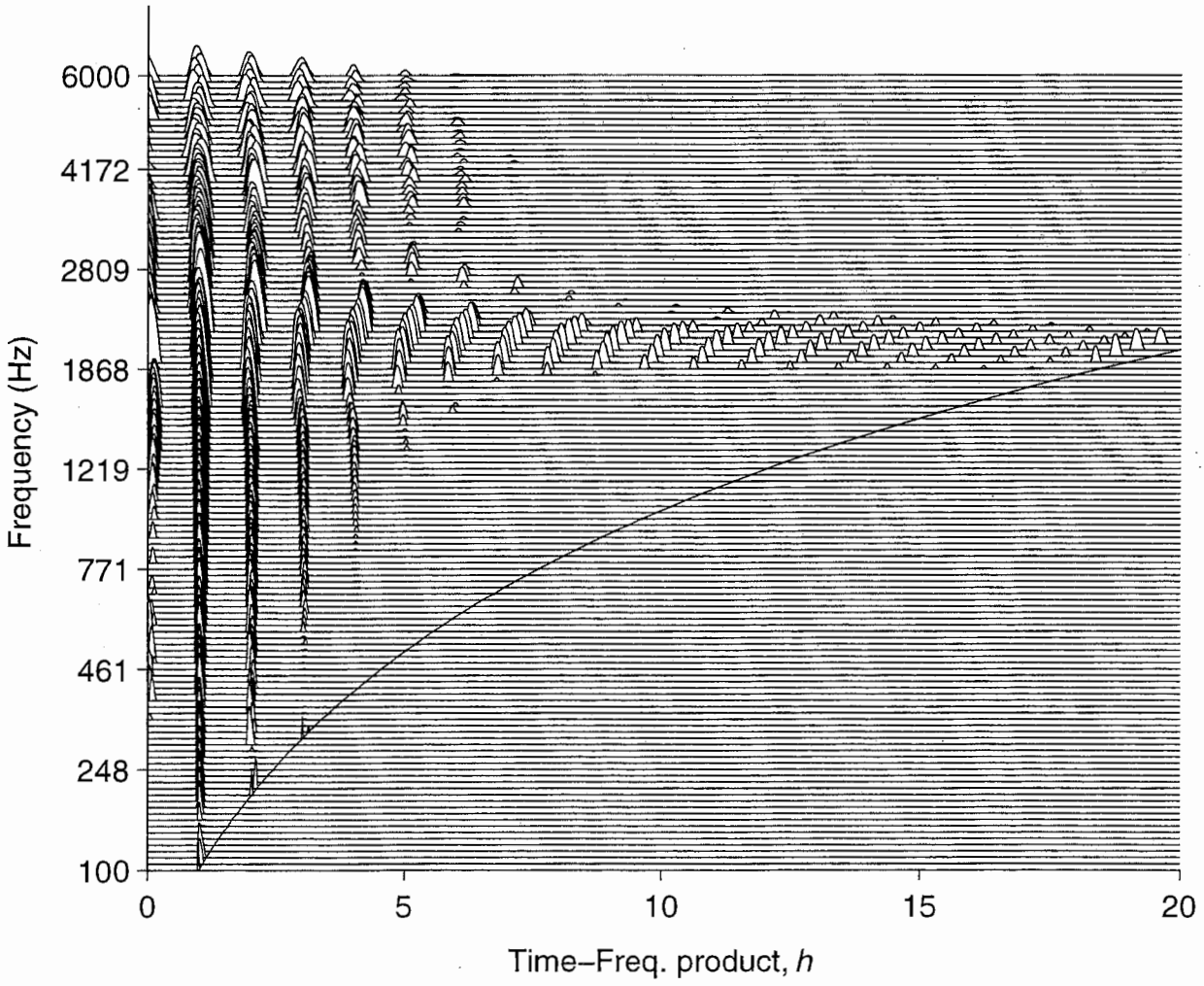
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 12

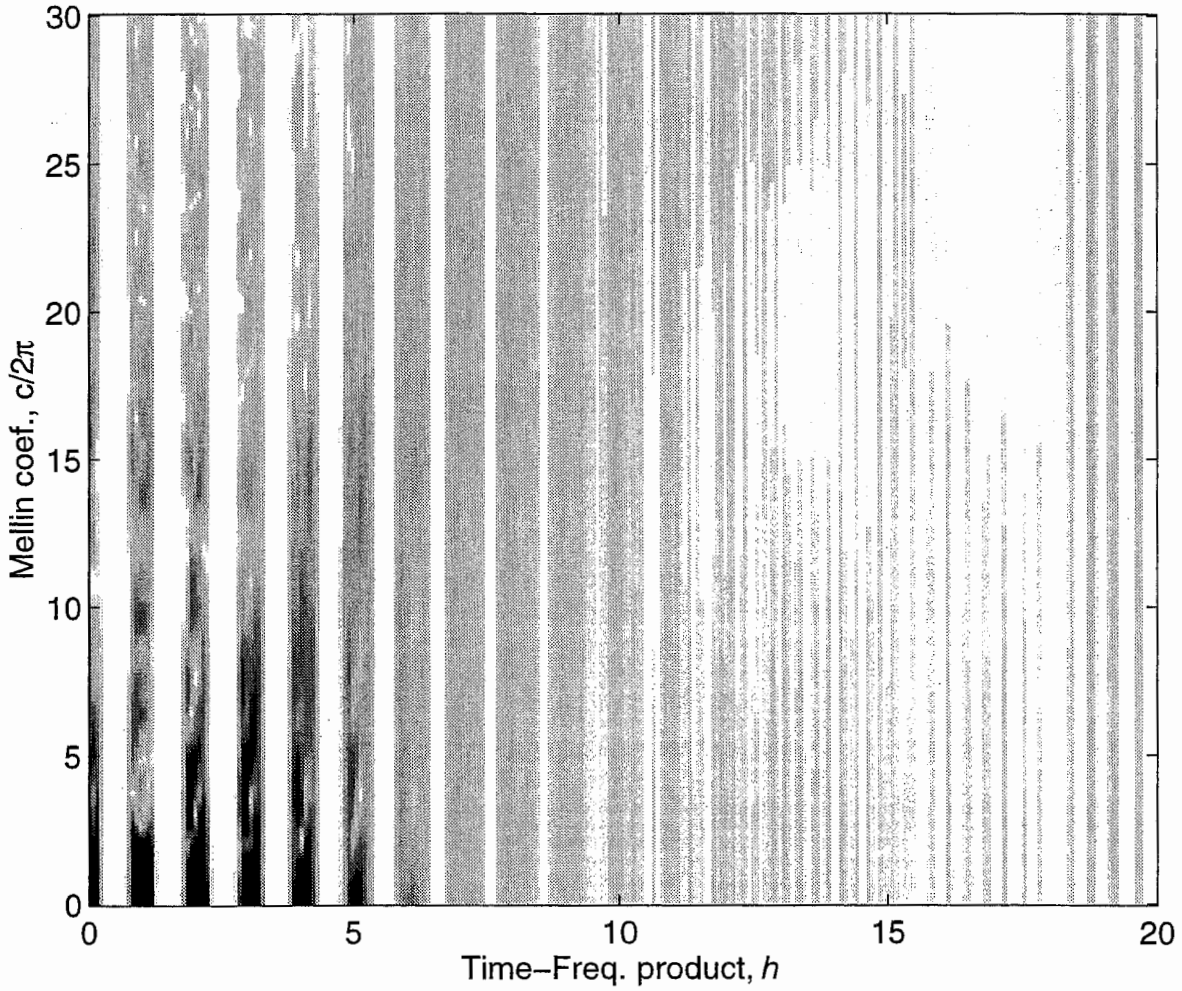

Fig. 13

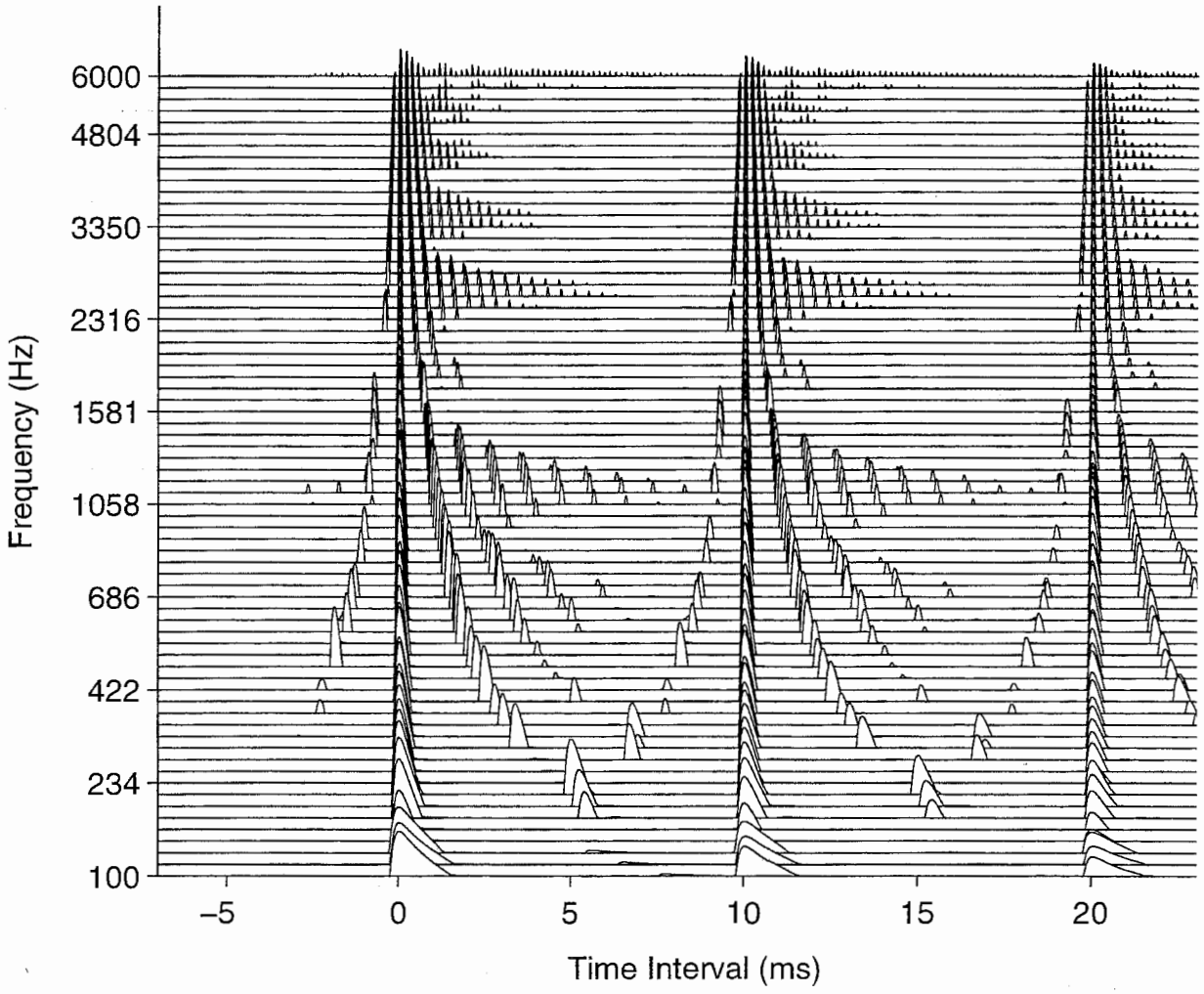
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
–5
0
5
10
15
20
Time Interval (ms)

Fig. 14

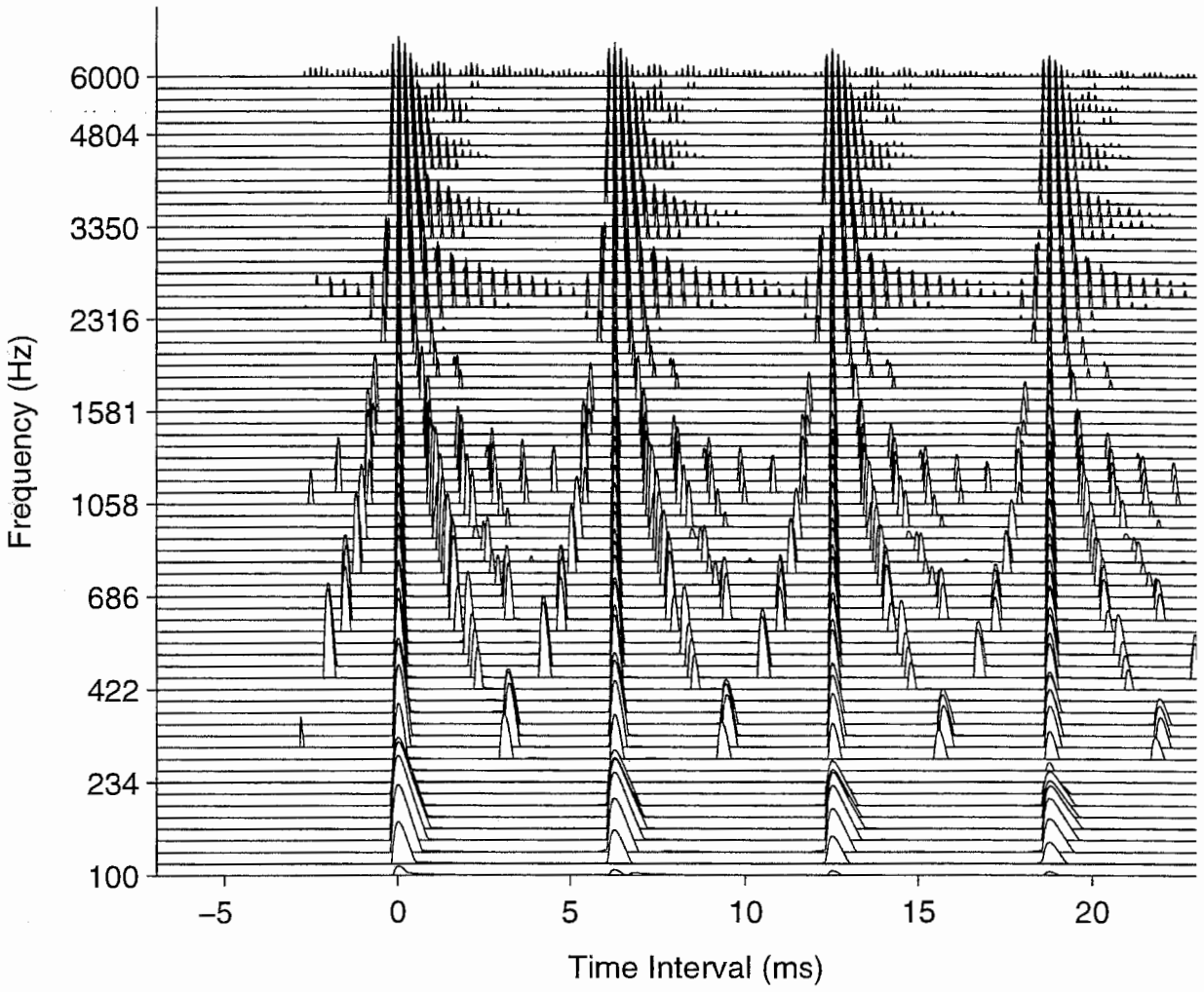
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
–5
0
5
10
15
20
Time Interval (ms)

Fig. 15

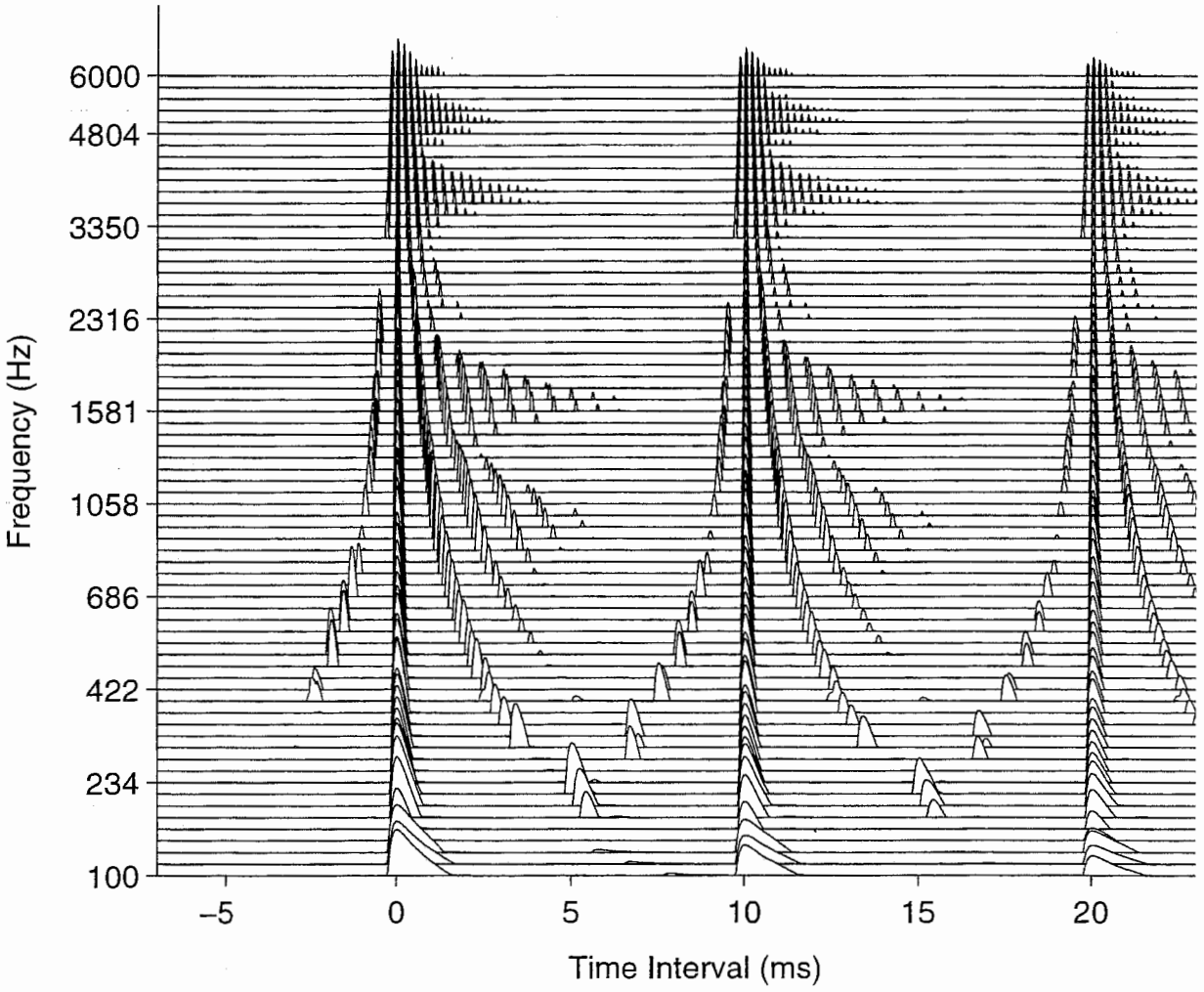
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
−5
0
5
10
15
20
Time Interval (ms)

Fig. 16

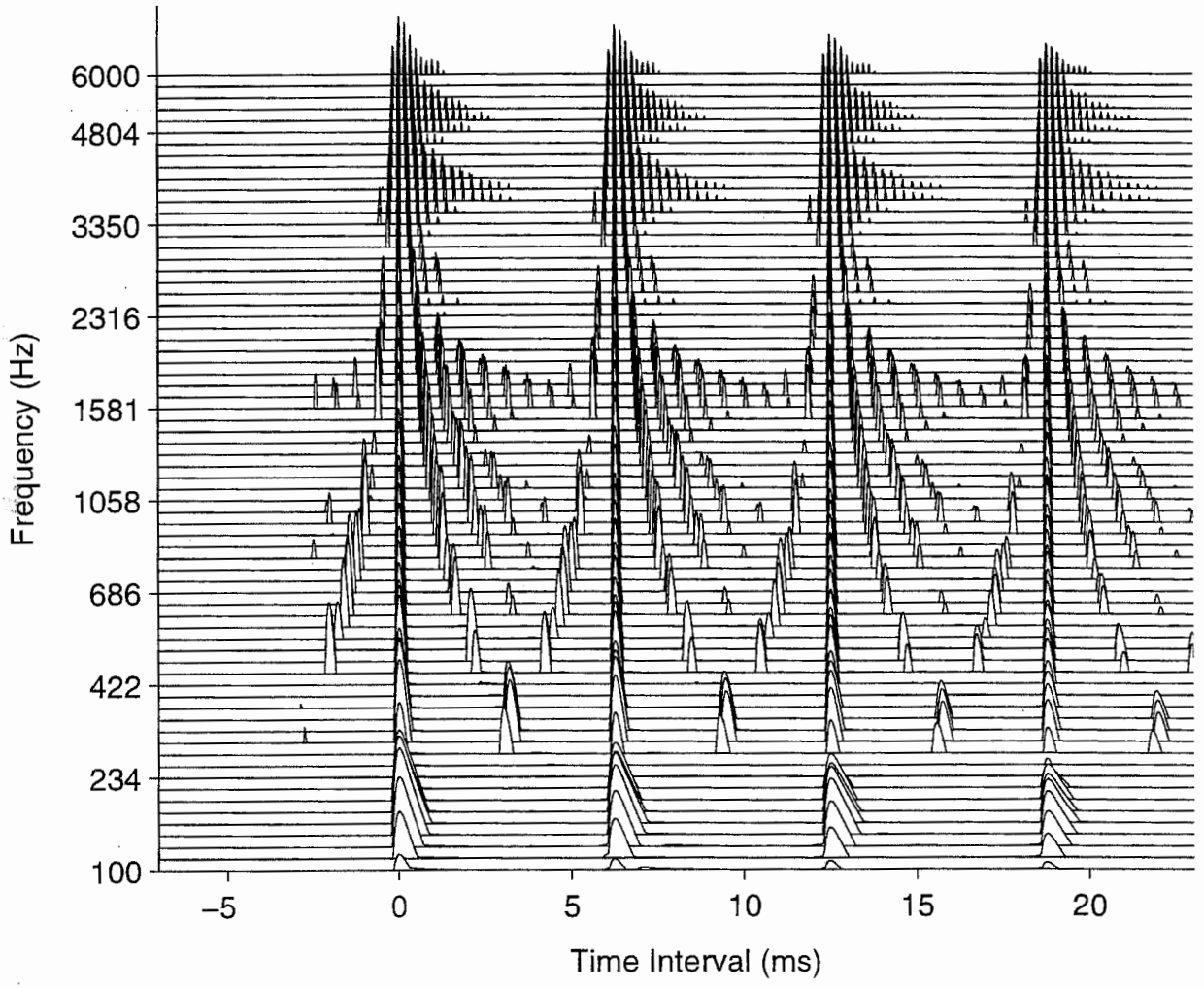
Frequency (Hz)
6000
4804
3350
2316
1581
1058
686
422
234
100
–5
0
5
10
15
20
Time Interval (ms)

Fig. 17

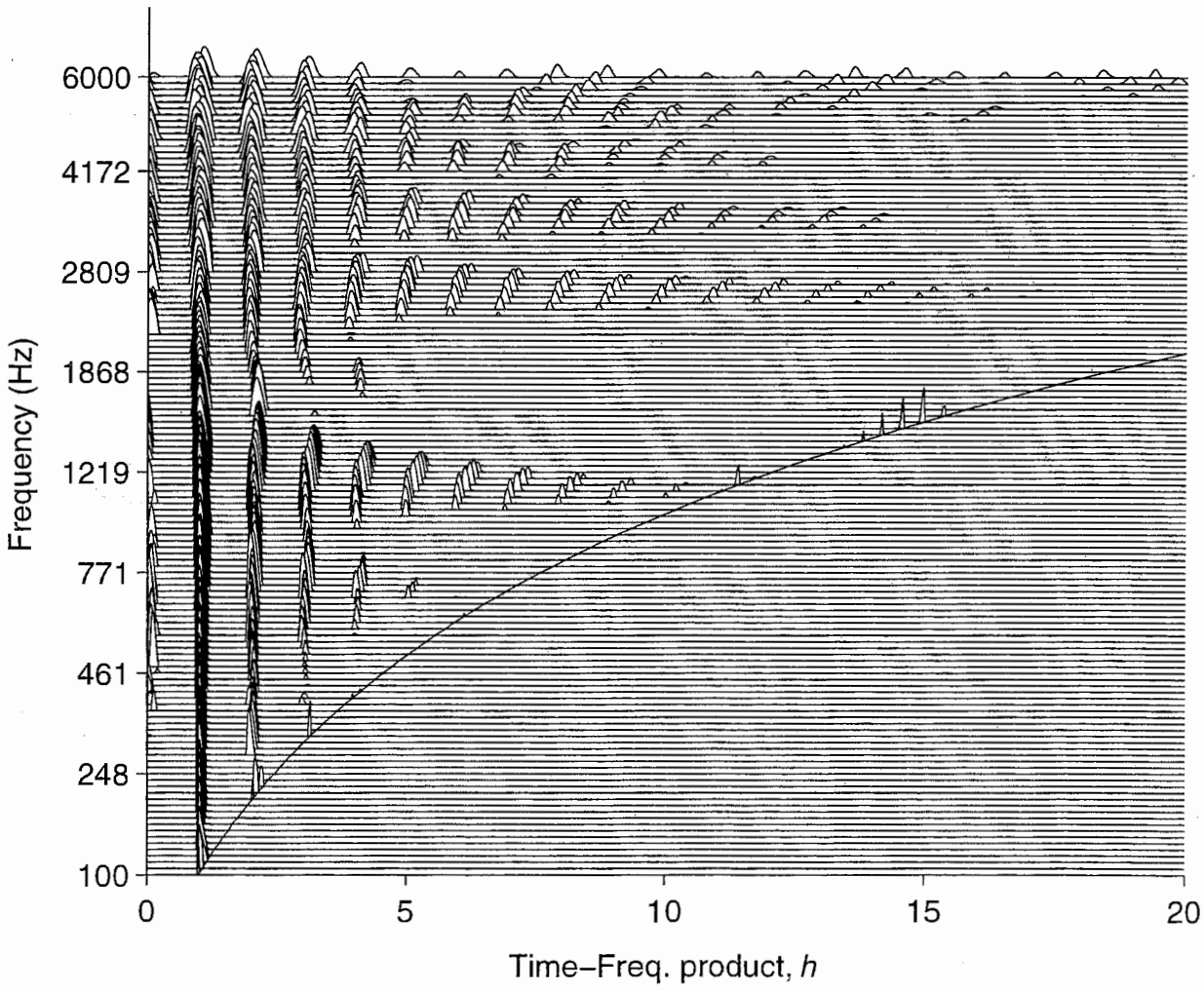
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 18

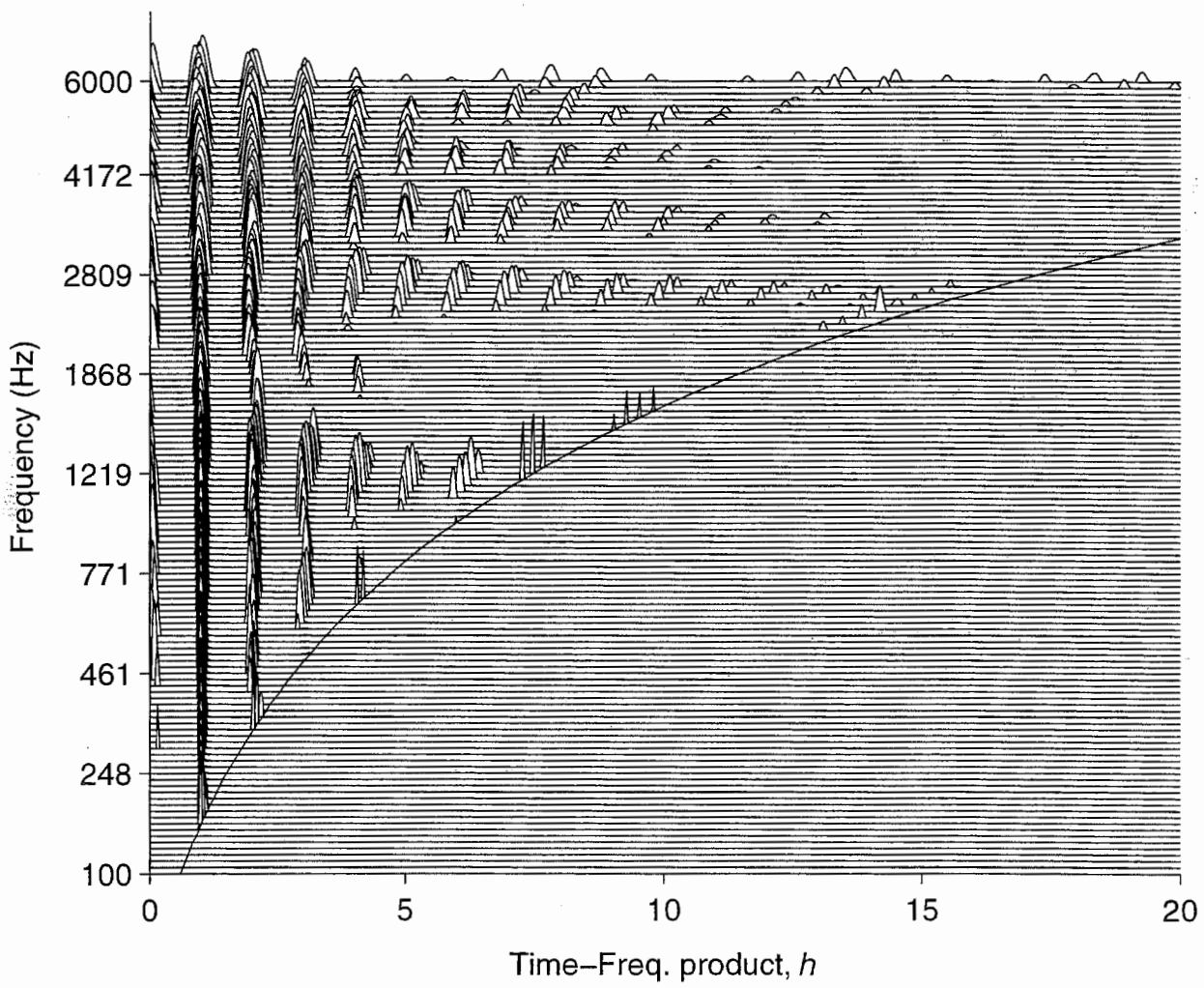
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 19

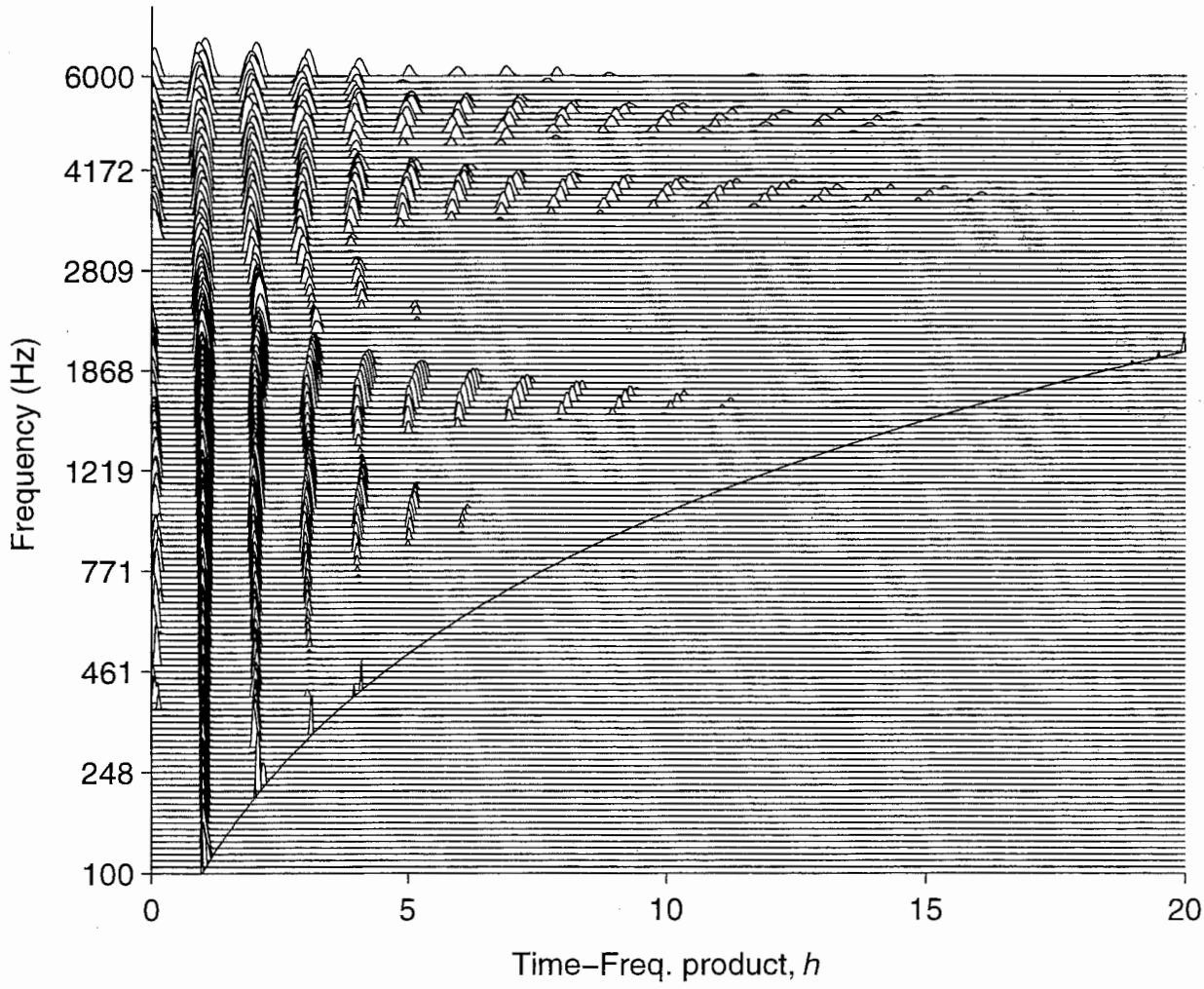
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 20

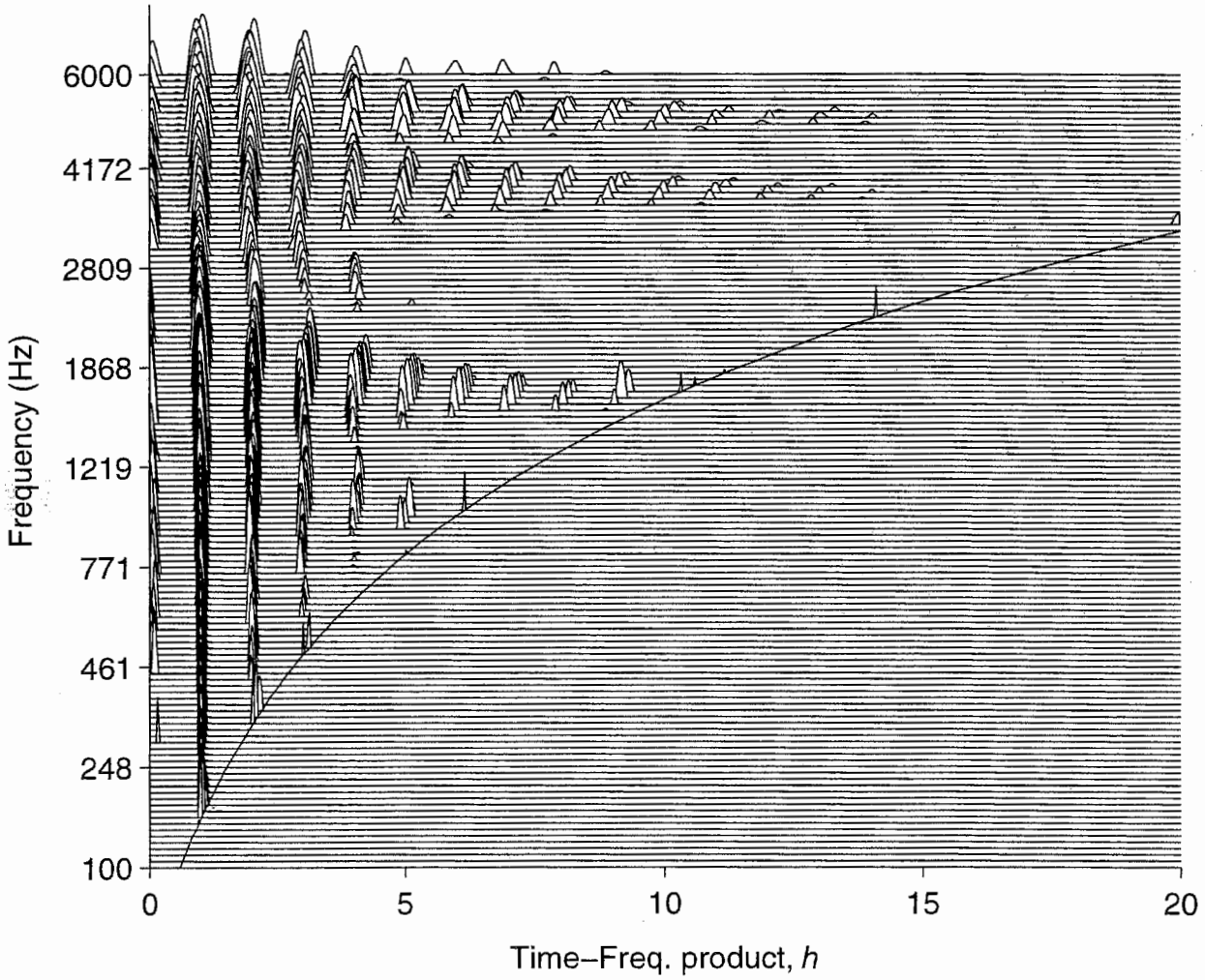
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 21

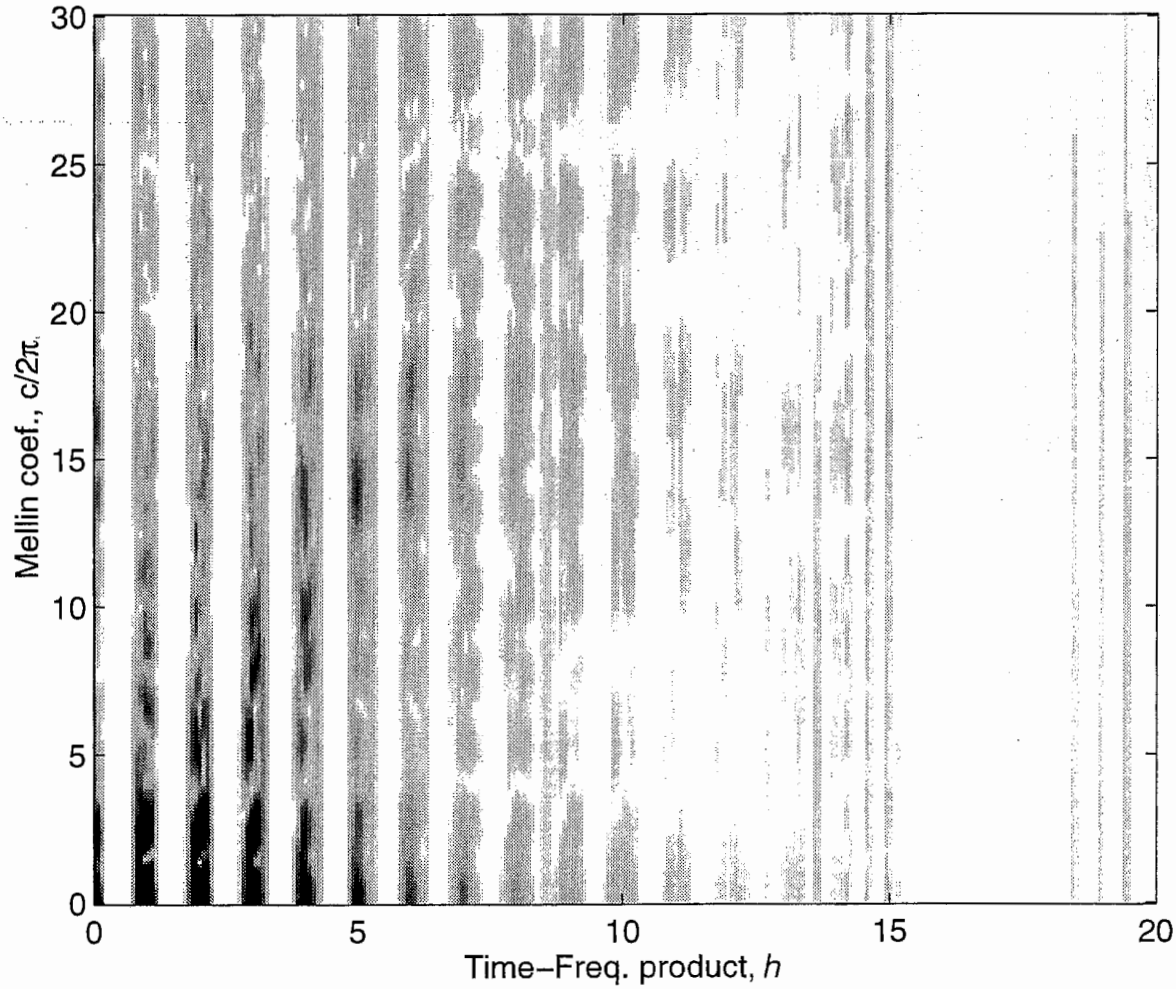

Fig. 22

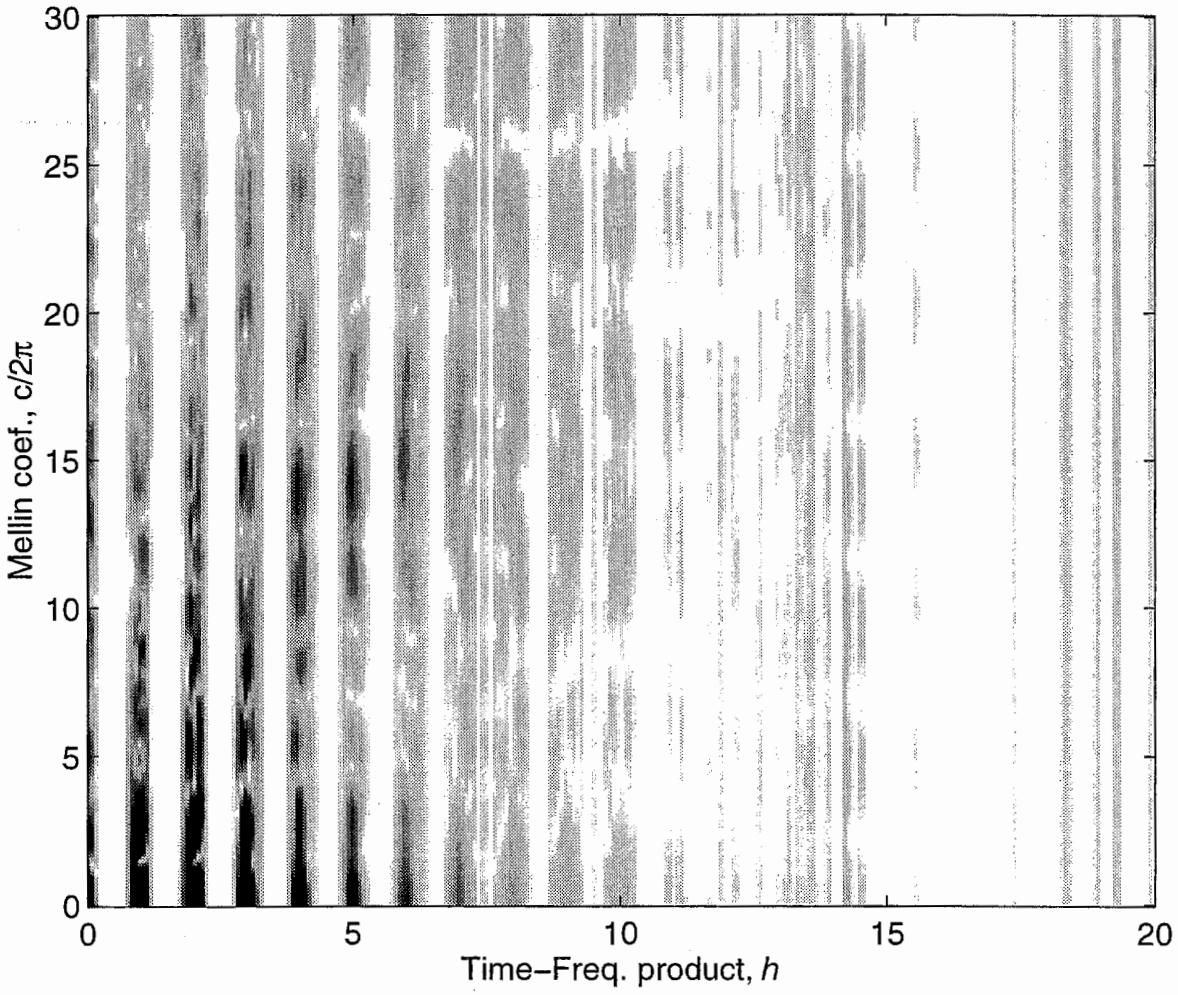

Fig. 23

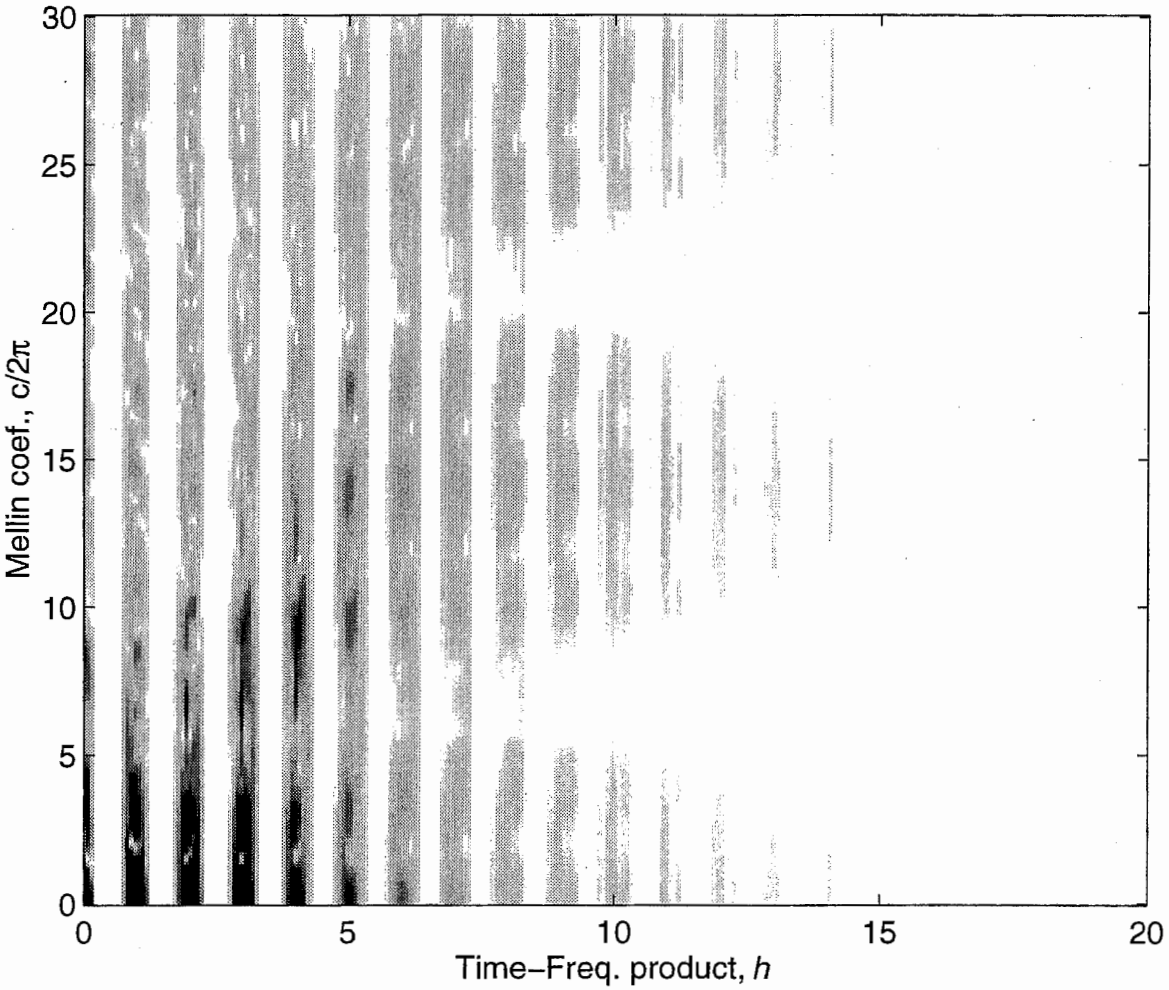

Fig. 24

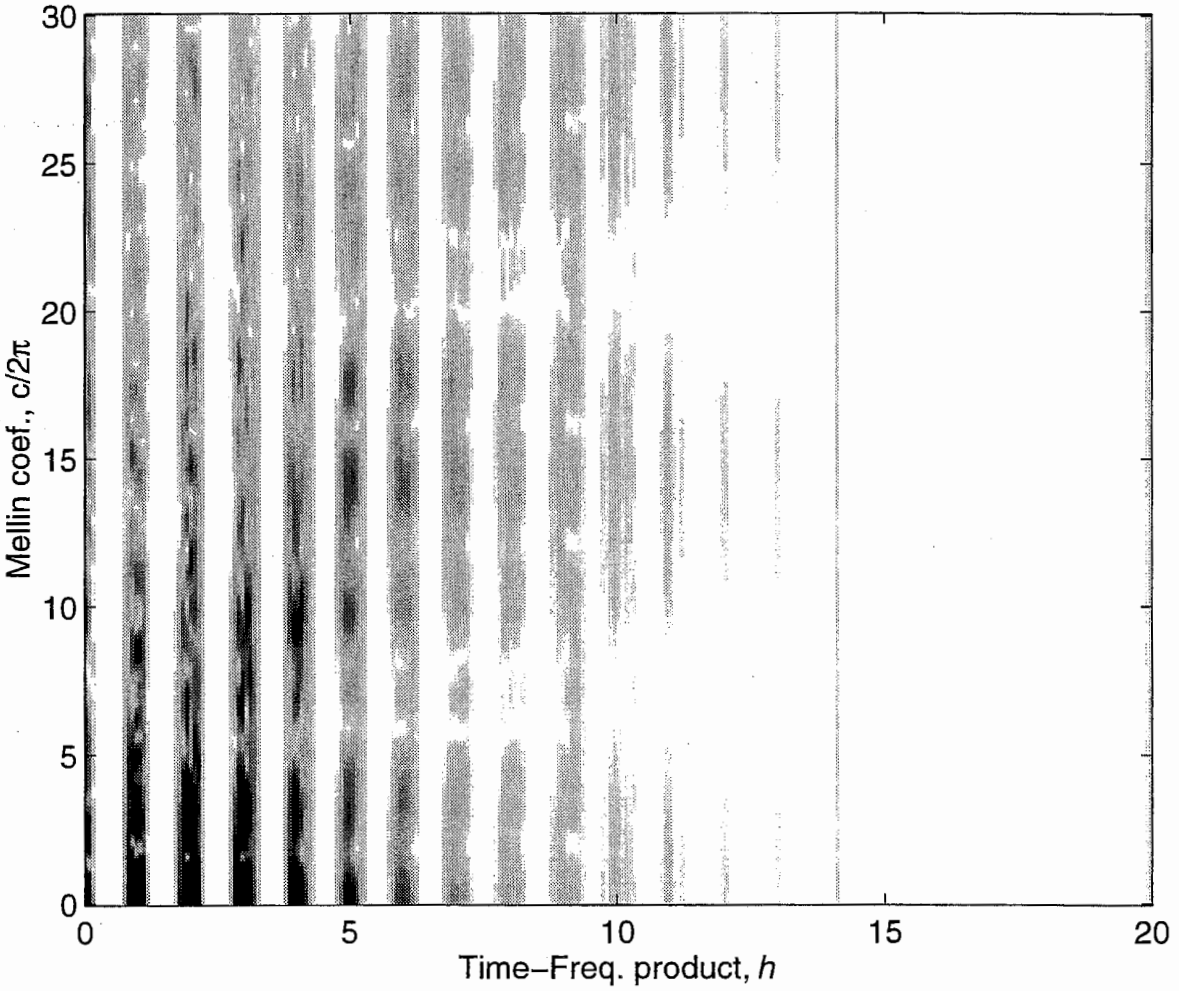

Fig. 25

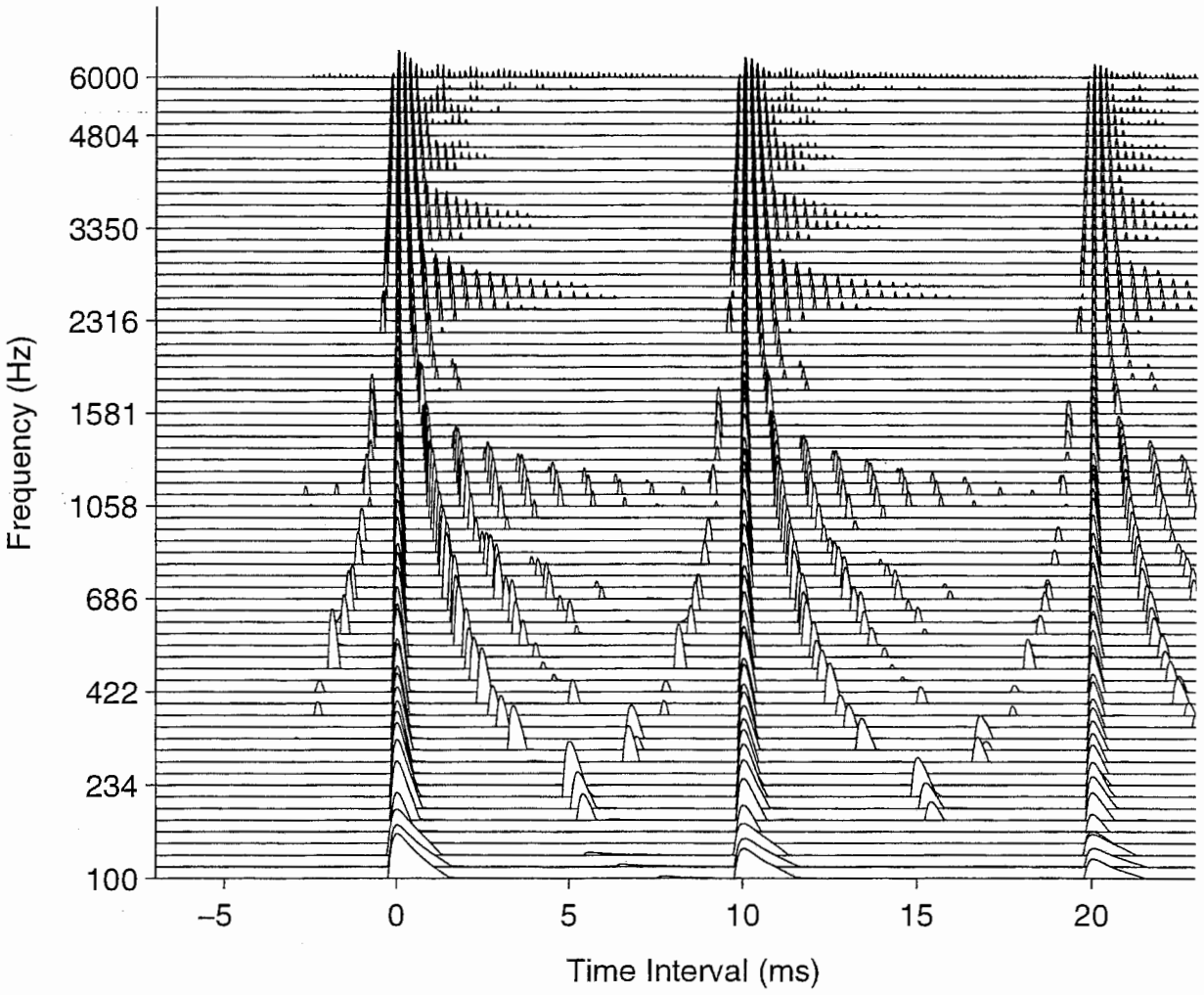
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
–5
0
5
10
15
20
Time Interval (ms)

Fig. 26

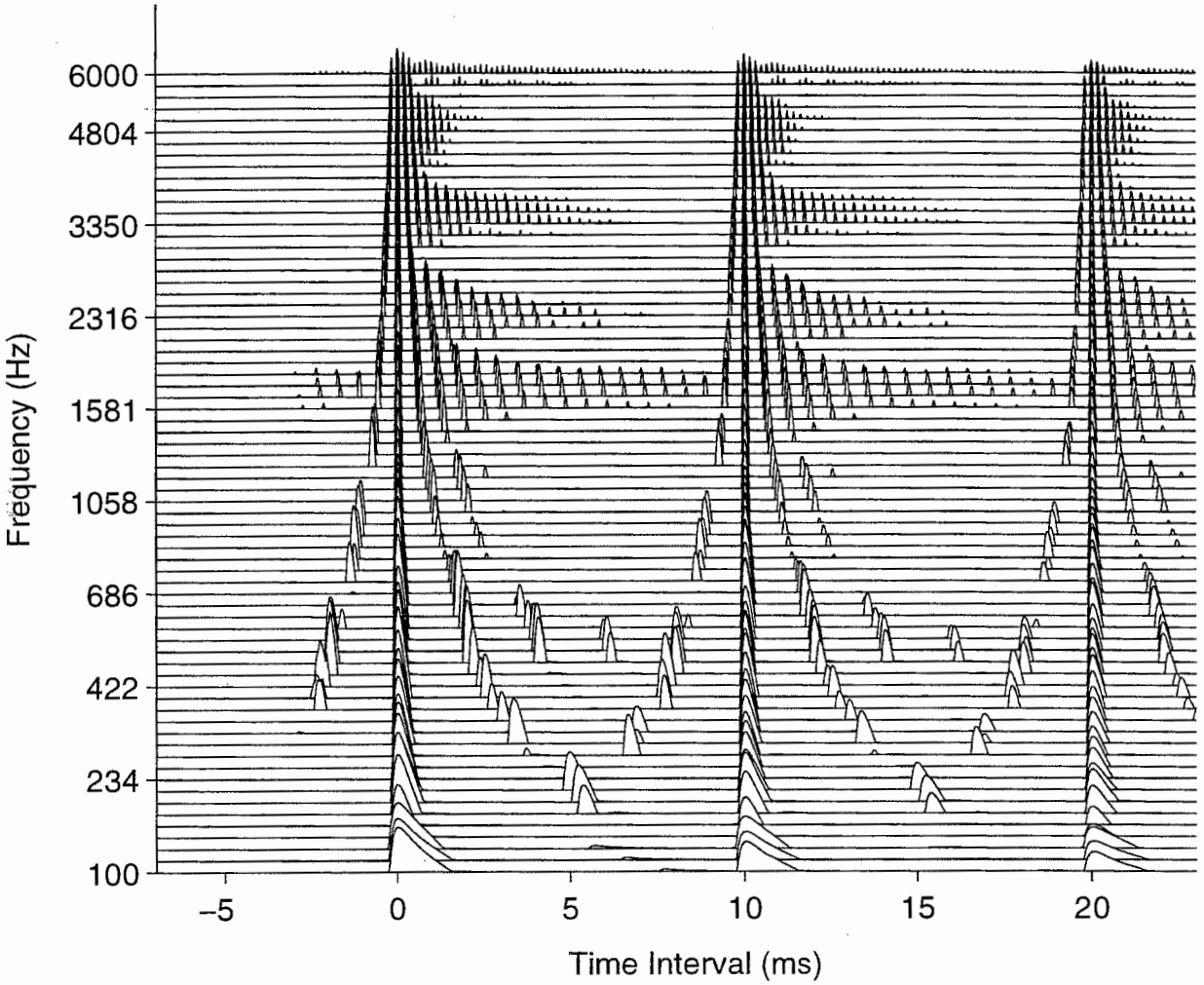
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
−5
0
5
10
15
20
Time Interval (ms)

Fig. 27

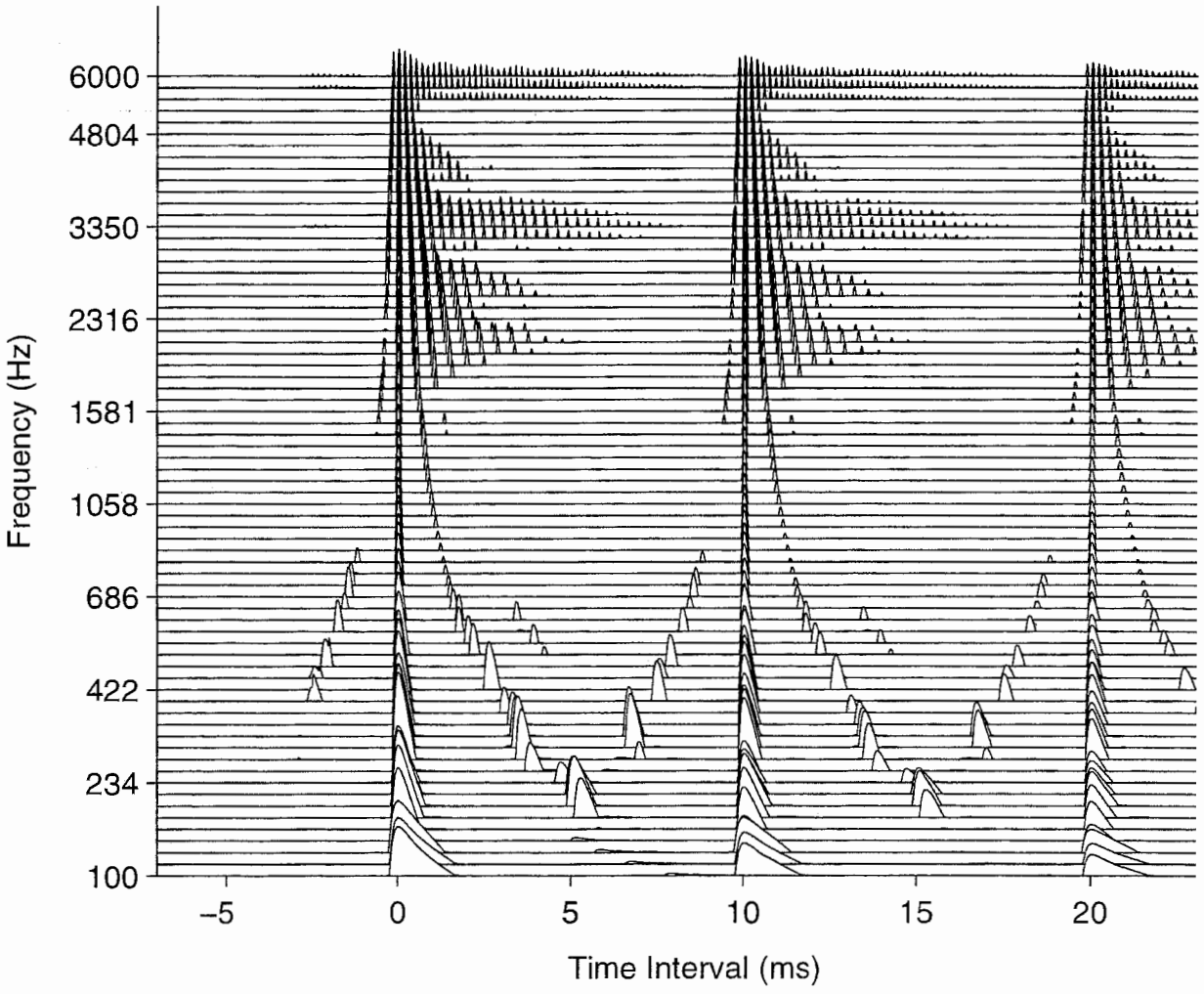
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
–5
0
5
10
15
20
Time Interval (ms)

Fig. 28

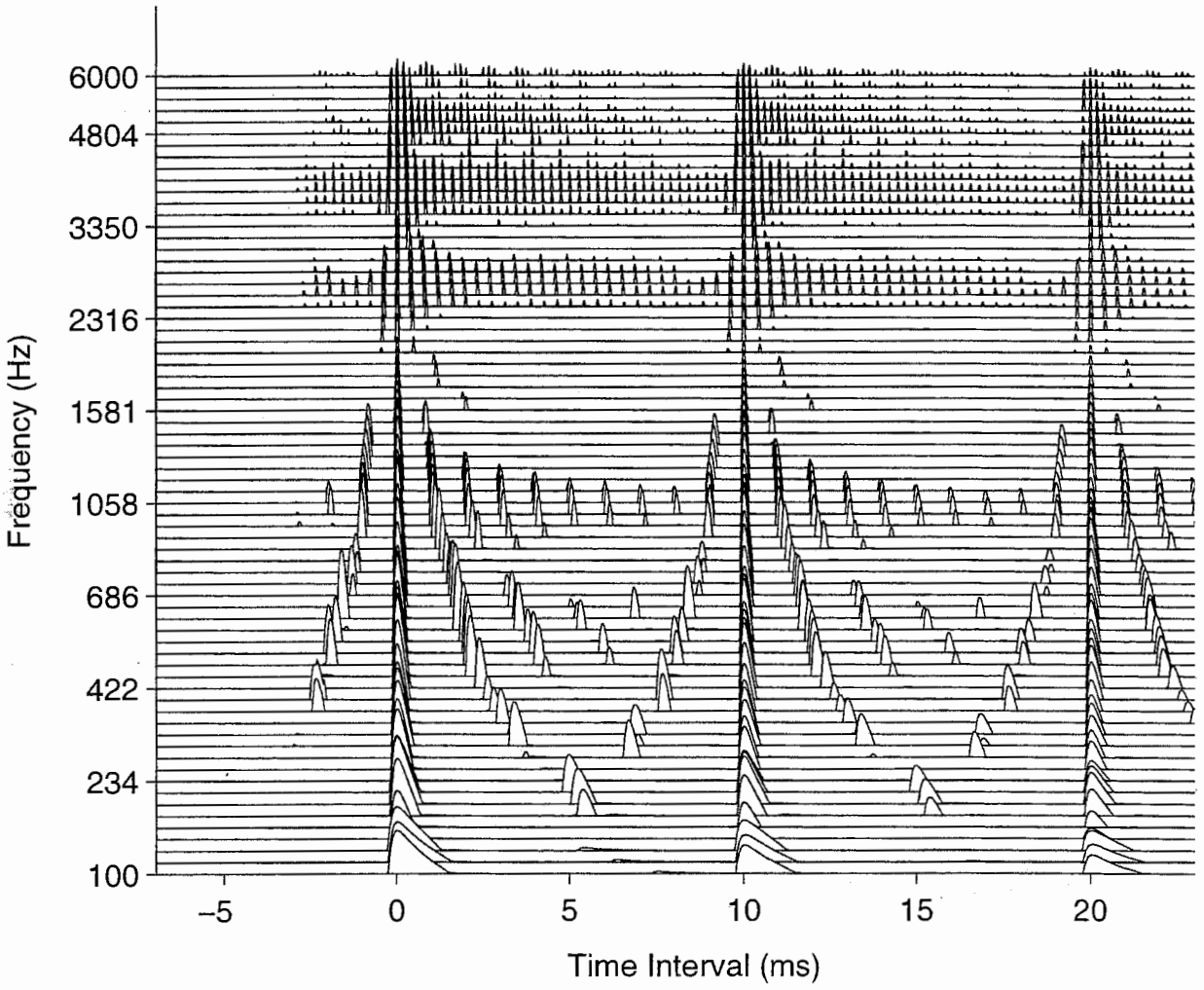
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
–5
0
5
10
15
20
Time Interval (ms)

Fig. 29

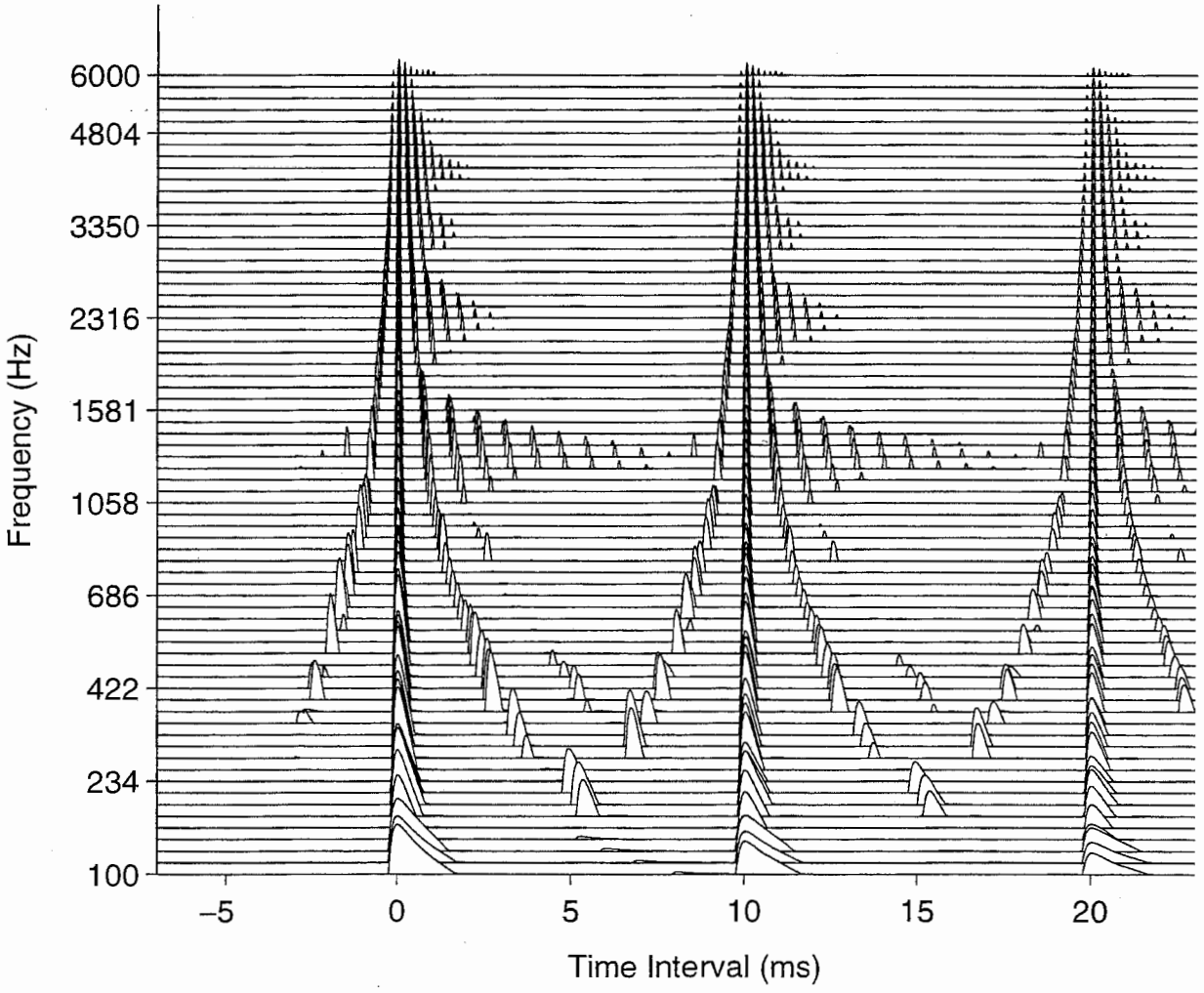
6000
4804
3350
2316
1581
1058
686
422
234
100
Frequency (Hz)
−5
0
5
10
15
20
Time Interval (ms)

Fig. 30

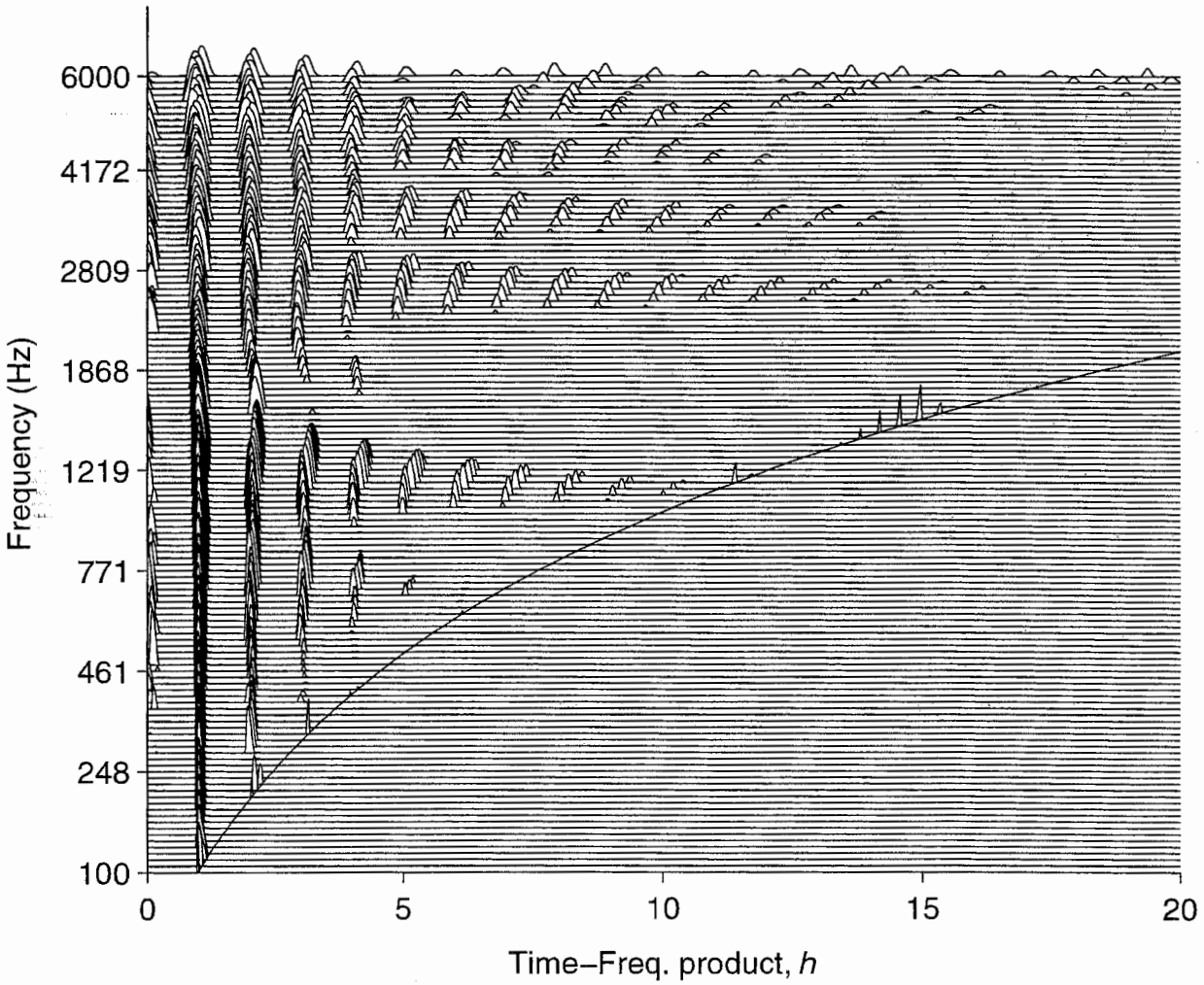
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 31

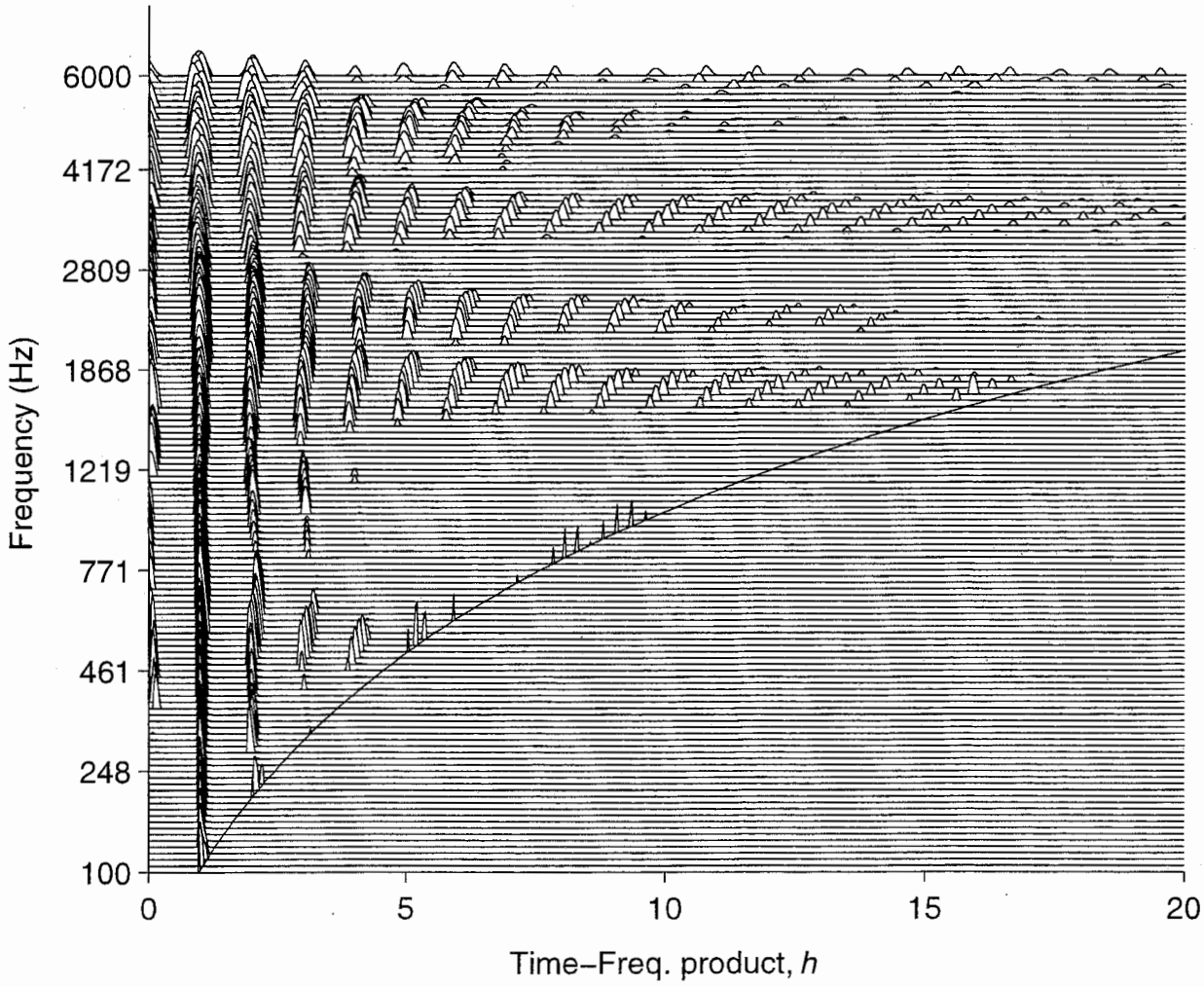
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 32

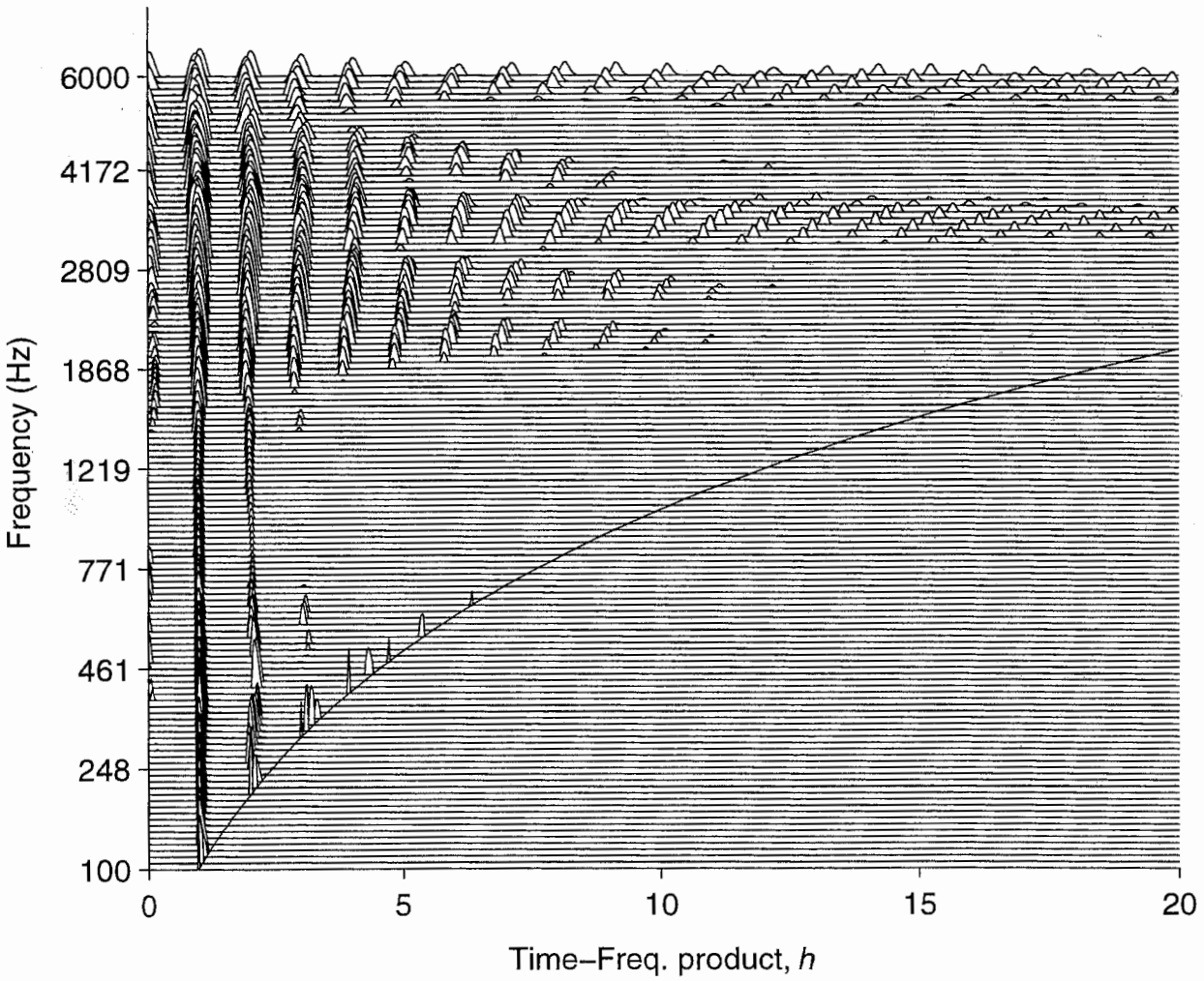
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 33

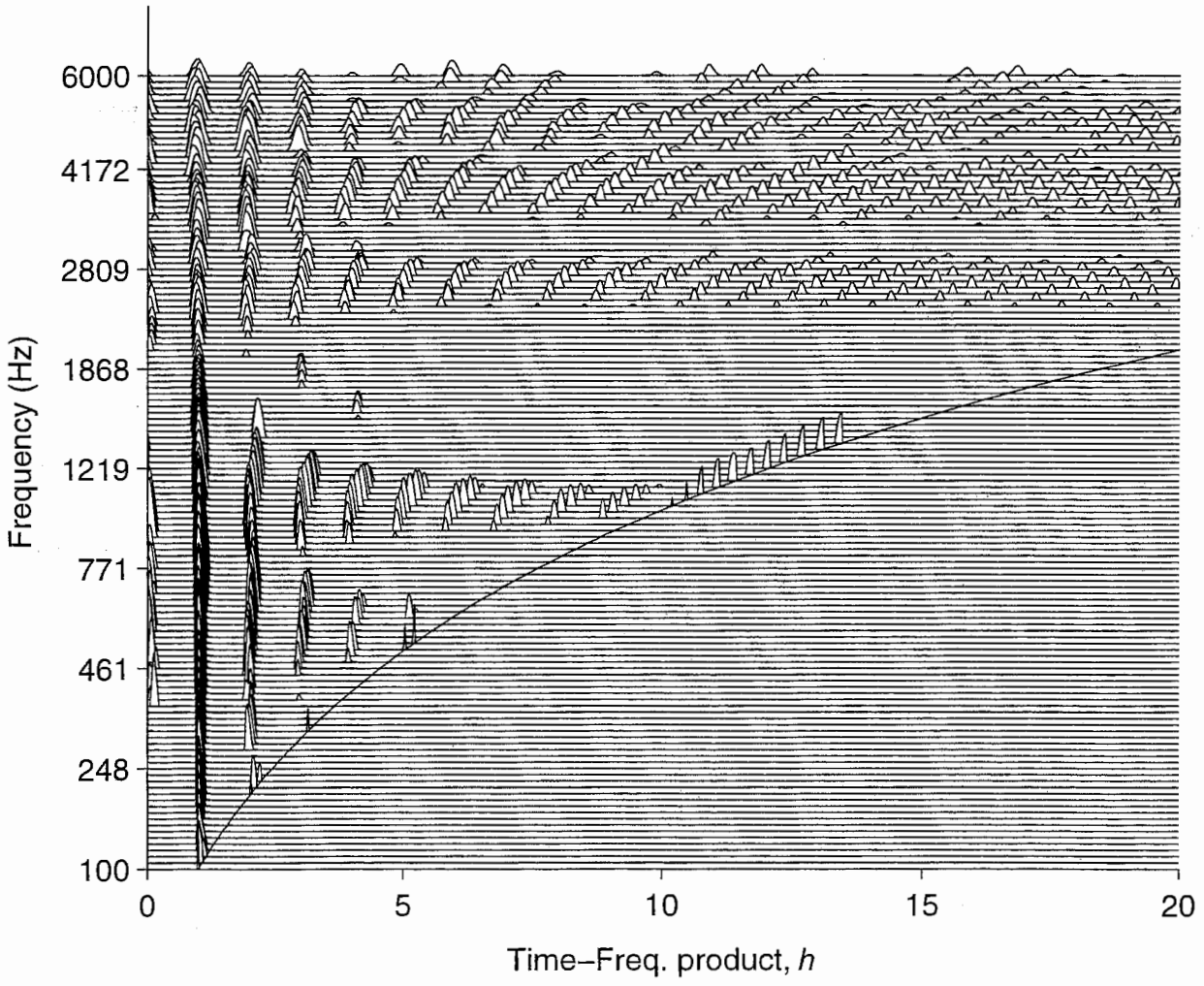
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 34

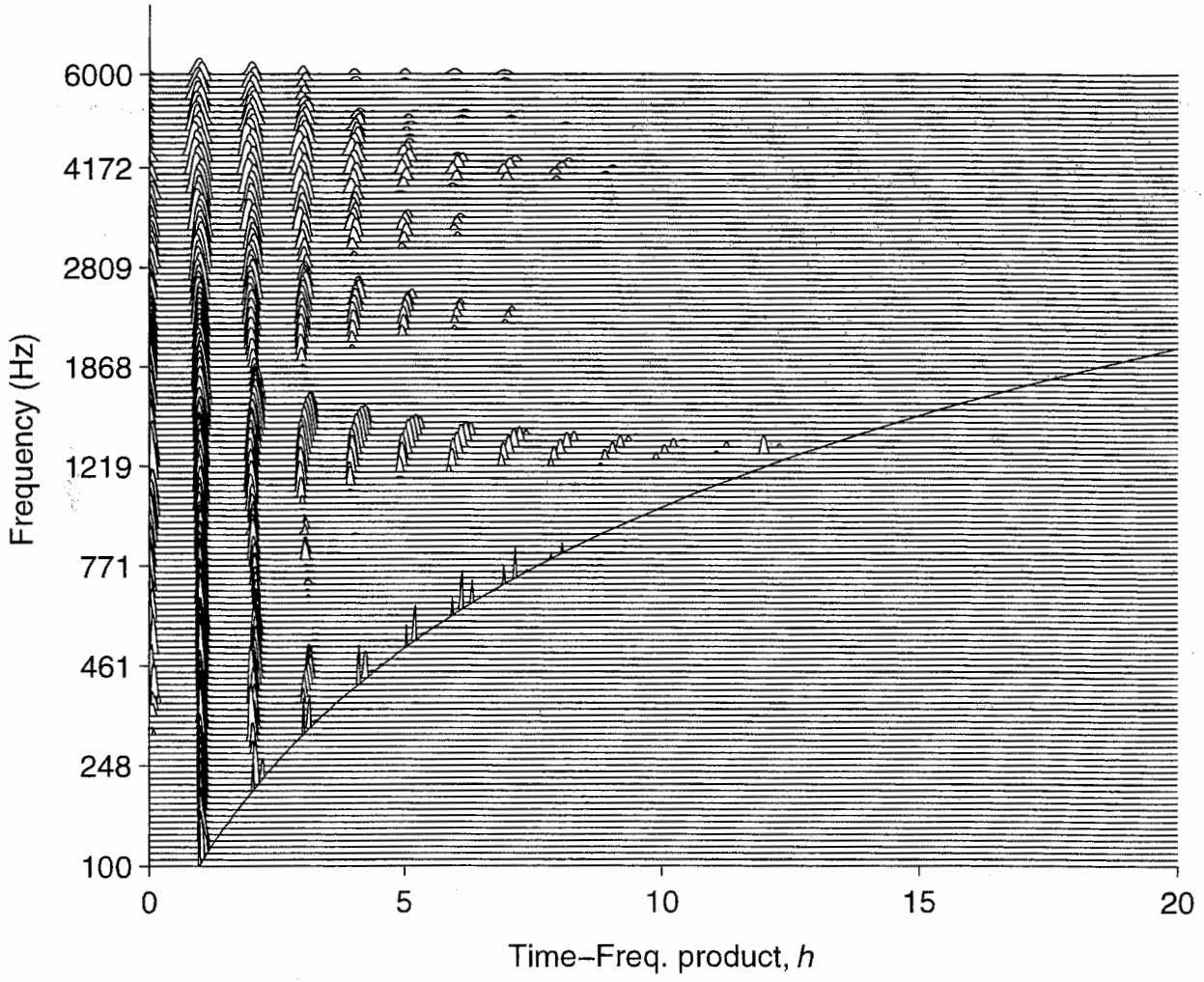
6000
4172
2809
1868
1219
771
461
248
100
Frequency (Hz)
0
5
10
15
20
Time–Freq. product, h

Fig. 35

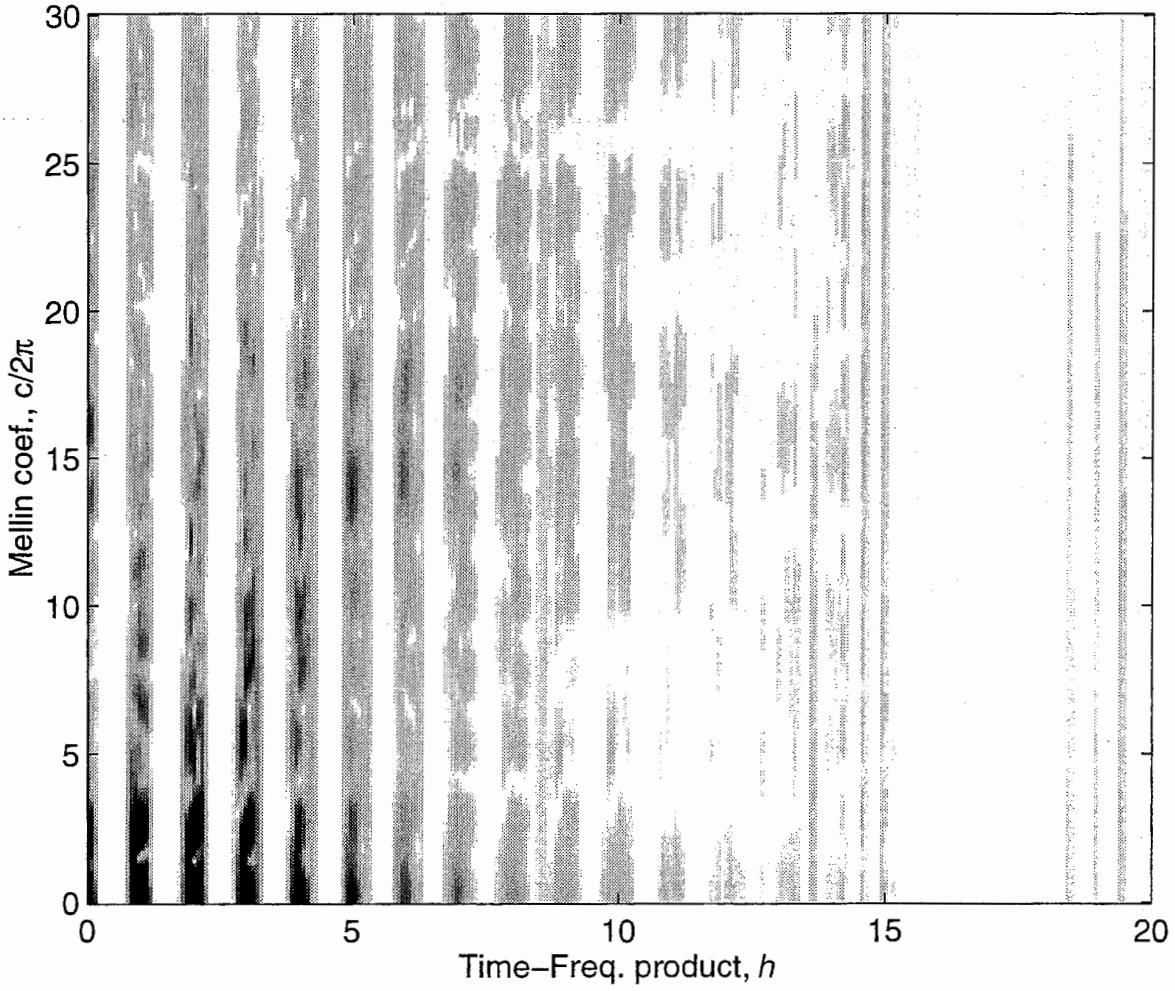

Fig. 36

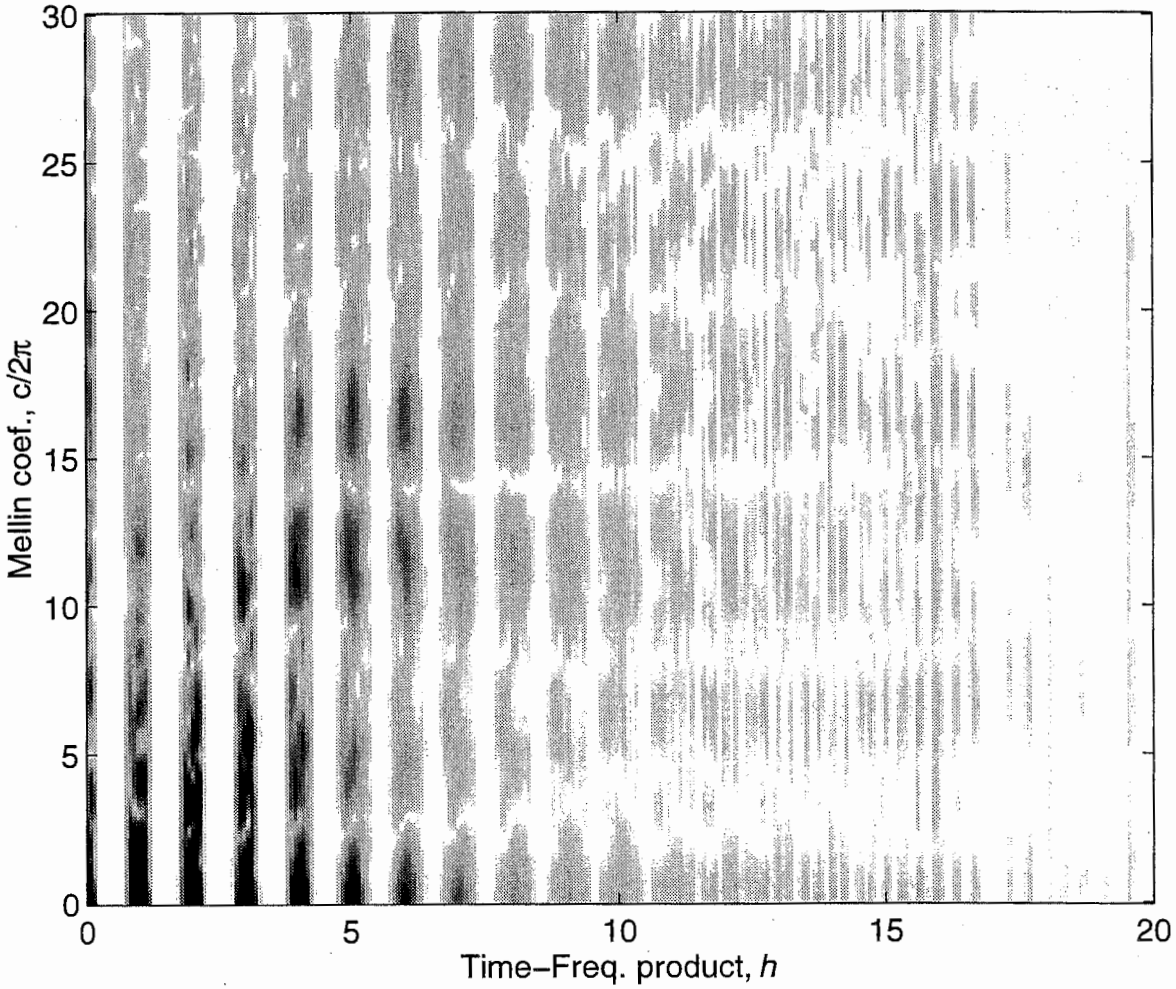

Fig. 37

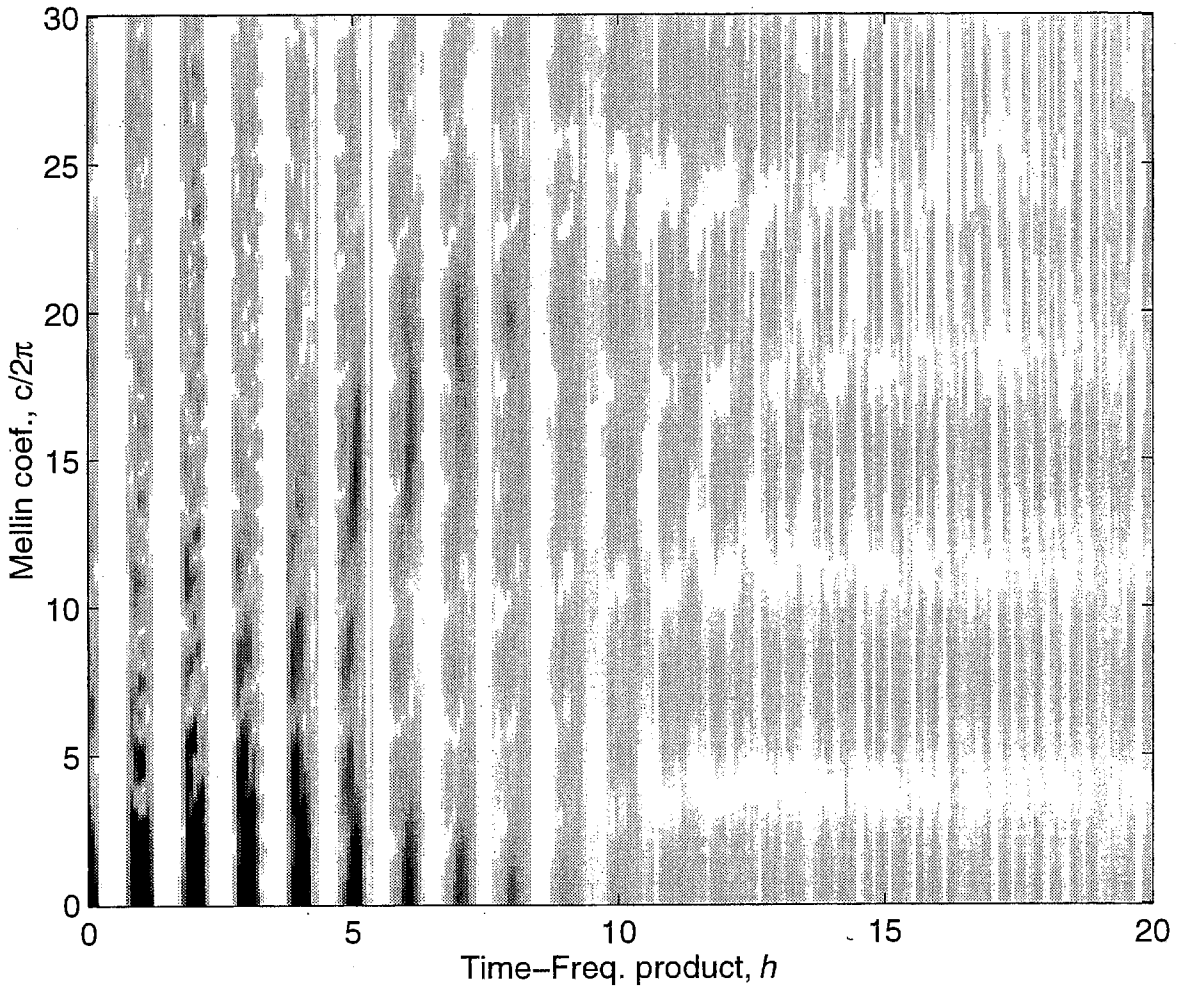

Fig. 38

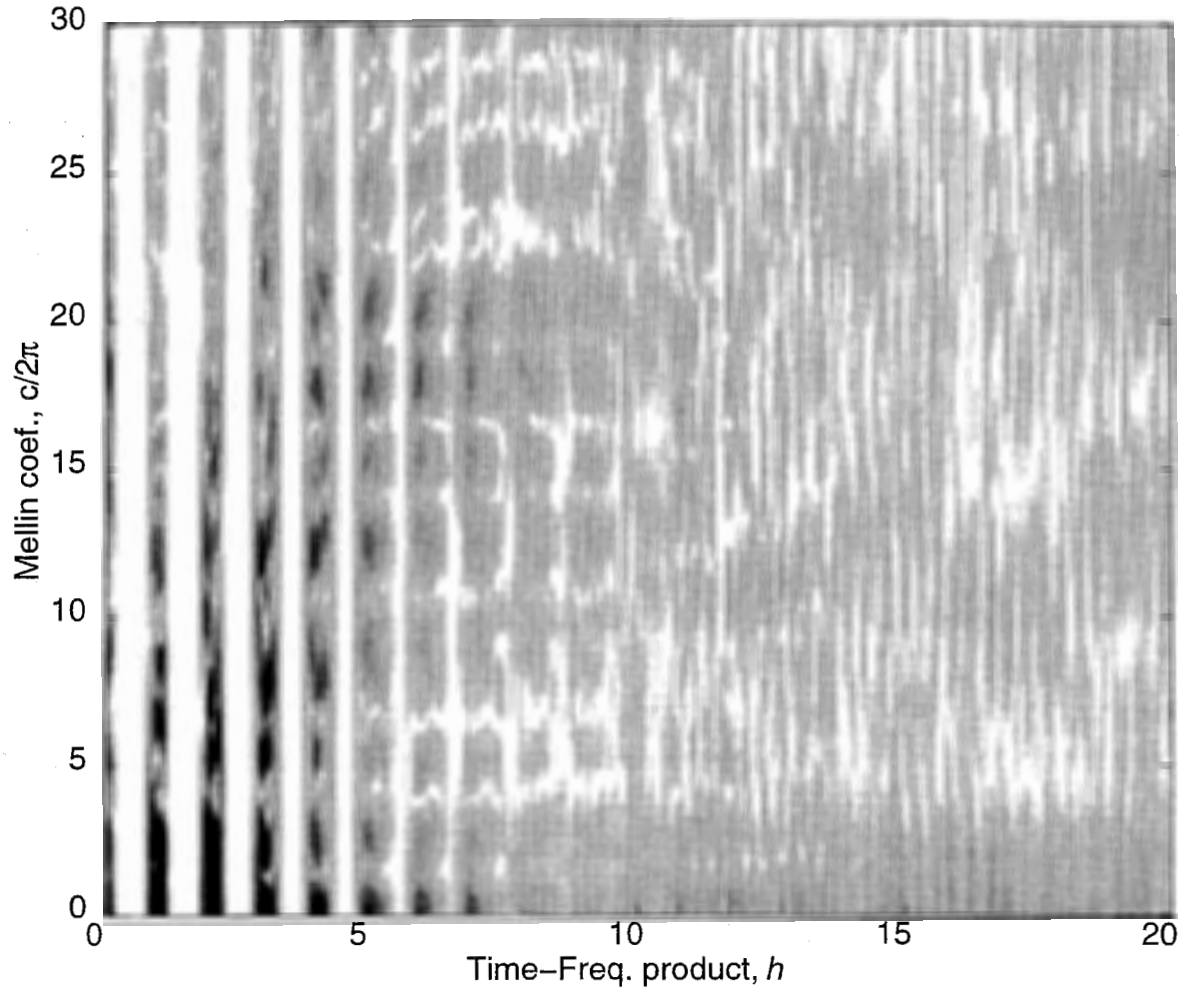

Fig. 39

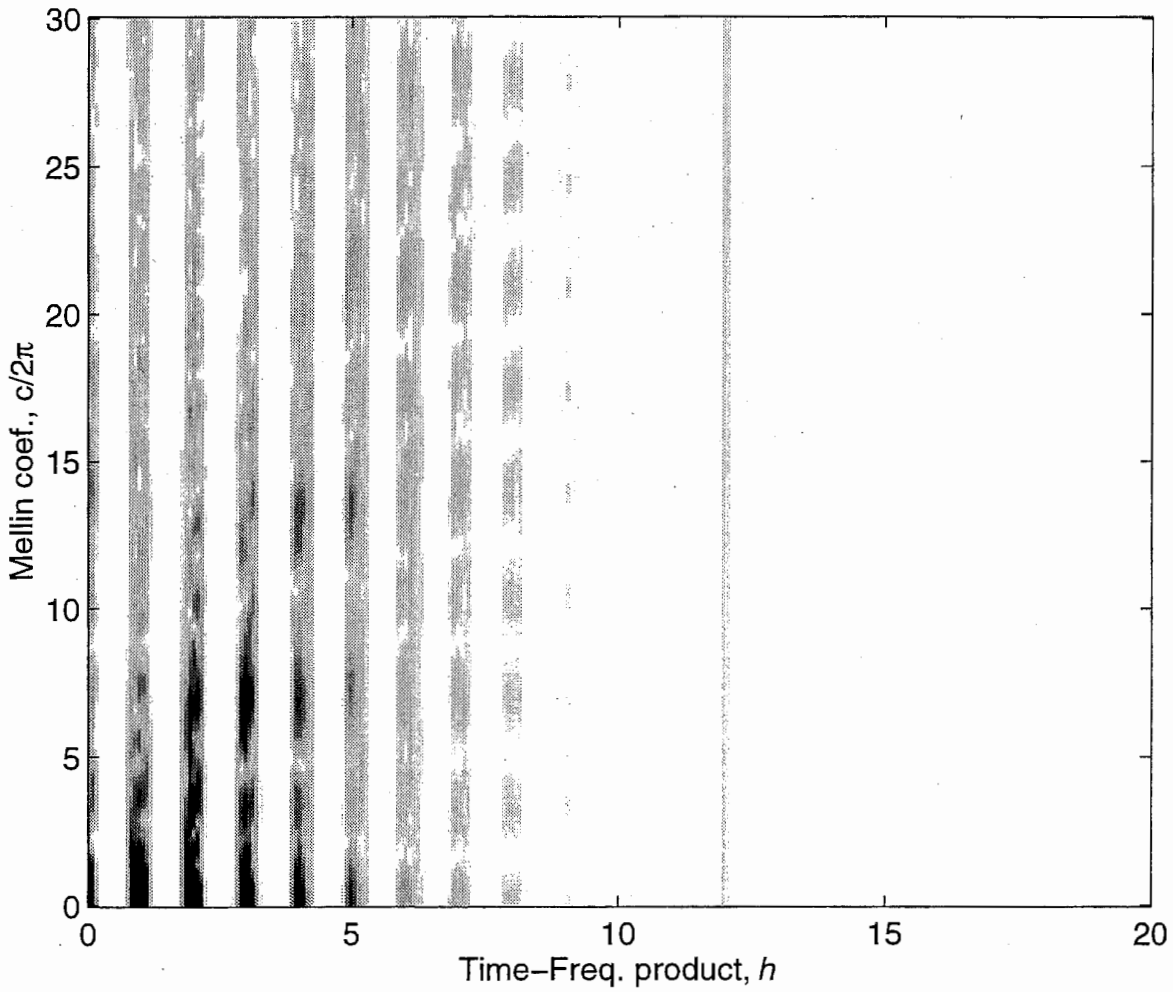

Fig. 40