IMPLEMENTATION OF SHORT-TIME HOMOMORPHIC DEREVERBERATION

by

James Lewis Caldwell

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREES OF

BACHELOR OF SCIENCE

and

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1971

Signature of Author _____
Department of Electrical Engineering, August 16, 1971

Certified by _____
Thesis Supervisor (Academic)

Certified by _____
Thesis Supervisor (VI-A Cooperating Company)

Accepted by _____
Chairman, Departmental Committee on Graduate Students

IMPLEMENTATION OF SHORT-TIME HOMOMORPHIC DEREVERBERATION

by

James Lewis Caldwell

ABSTRACT

For short reverberation times ($\sim$ 10 msec), and over short intervals ($\sim$ 50 msec), speech waveforms reverberated by small rooms can be approximated by the convolution of a function interpreted as the "room impulse response" with the uncorrupted speech. It is possible, therefore, to remove the effects of reverberation by applying a deconvolution process to 50-msec sections of reverberated waveform. Homomorphic deconvolution, which consists of transforming the convolved signal components into a sum of corresponding signals, linear-filtering the sum, and reconstructing dereverberated speech from the result, is well-suited to this application. Computer simulations discussed in this paper show that such a process is effective for artificially-reverberated speech. New results include filtering techniques that are useful when the reverberation is not precisely known before processing.

THESIS SUPERVISOR: Alan V. Oppenheim
TITLE: Associate Professor of Electrical Engineering

ACKNOWLEDGEMENT

I welcome this opportunity to thank J. L. Flanagan and R. W. Schafer of Bell Laboratories, whose encouragement, technical advice, and amazing patience have made this research both possible and rewarding. Many valuable conversations with L. R. Rabiner and R. C. Lummis of Bell Labs have also contributed to results presented here. Professor A. V. Oppenheim of M.I.T., who sparked my initial interest in this topic, provided enthusiastic supervision, and Miss Bonnie Wilson provided invaluable clerical assistance in final thesis preparation. My greatest debt, which can only be repaid in love, is to my wife Bobbi for the hundreds of hours she sacrificed while I labored to finish this paper.

TABLE OF CONTENTS

I. Introduction

A. Spectral Coloration

Small-room reverberation poses a significant problem in the quality improvement of Speakerphone and Conference Room Telephony Systems. Speech waveforms reach the microphone by a multiplicity of paths, one direct and the others involving reflections from walls, tables, or other objects. The dimensions of a typical office or conference room are such that interpath delays are likely to fall between 1 and 10 ms. As a result of the shortness of these delays, one perceives a tonal distortion of the speech (referred to as "spectral coloration"), rather than distinct echoes. The speaker's voice often sounds as if it is coming from within a hollow barrel. This unnatural and annoying effect is far more pronounced to a person listening via the microphone than to one actually in the room. The auditory system of a listener in the room tends to adjust to, and discriminate against, echoes of sounds. This is not possible for one listening by artificial means. In this case, therefore, the effects of short-time reverberation must be removed by signal processing.

A few interesting approaches to dereverberation have been developed, not all applicable to the removal of spectral coloration. In the process of Mitchell and Berkley, a set of bandpass filters divides reverberant speech into several channels, and each filter output is center-clipped at a level anticipated to eliminate echo in its band[1]. This is particularly effective for suppression of long-time reverberative tails, but hardly affects coloration. Flanagan and Lummis experimented with another process capable of reducing spectral coloration, but not the effects of long-time reverberation[2]. Small-room reverberation produces an effect

upon speech which is similar to that of a linear comb-filter with "valleys" in its frequency response typically spaced at intervals of 100-500 Hz. Valley spacing varies with microphone placement, so positioning several microphones at different locations in a room produces "filters" which pass different parts of the frequency spectrum. By allowing each of several microphones to contribute to a composite signal those parts of the spectrum it receives more faithfully than the other microphones, Flanagan and Lummis partially eliminate the valleys. The resulting speech sounds considerably less colored than any individual microphone output.

No one process has proven suitable for all types of reverberation. Dereverberation by homomorphic deconvolution, the subject of this paper, is no exception. Several factors make this method impractical for long-time dereverberation; however, its applicability to removal of spectral coloration makes it of interest in the present problem.

B. Past Development of Homomorphic Dereverberation and Contributions of this Research

The theory of homomorphic dereverberation has been developed considerably in the work of others. Oppenheim[5] and Schafer[3] are primarily responsible for developing the theory of the complex cepstrum and its application to deconvolution. Their work includes a rigorous mathematical justification for homomorphic systems, techniques for the analysis and computation of complex cepstra, and the foundations of short-time echo-removal. With regard to the latter area, little attention was given to finding filtering methods that are practical when little is known about the reverberation, or to the possibility of using the cepstrum (as opposed to the complex cepstrum) for dereverberation. Oppenheim and Schafer have also been closely associated with the homomorphic analysis

and synthesis of speech.[7,8] Of particular significance to the present research are the principles of Oppenheim's homomorphic vocoder[7] (based upon the cepstrum), which play an important role in the discussions of Section III. Most recently, Flanagan[4] has suggested the use of the cepstrum for the removal of spectral coloration due to short-time reverberation. His proposal includes a scheme for averaging several cepstra. To a large extent, the ideas of Flanagan and Schafer served as a starting point for this work.

The cepstrum (or complex cepstrum) of a reverberant speech waveform is the sum of a component corresponding to the unreverberated speech and an unwanted component due to the reverberation alone. If the reverberative component is precisely known, its removal is straightforward. In this thesis, new techniques are introduced for discriminating between components when the reverberative component is <u>not</u> precisely known. As initially suggested by Flanagan[4], these methods center about the advantages obtained when several cepstra of differently-reverberated speech waveforms are combined. One technique involves computing the average and difference of two such cepstra. The difference yields information about the reverberative components, which can be used to remove these components from the average cepstrum. Computer simulations indicate that this is quite effective in reducing the effects of artificial reverberation in voiced speech when the interpath reverberation delays are in the 3-10 msec range, as for small-room reverberation. Results for delays outside this range are not as satisfactory. In a second filtering technique, the use of memory and minimum mean-square error estimation is proposed to reduce the speech distortion produced by removal of the reverberative component. Another objective of this method is the possibility of

extending echo-removal capability to include echo delays down to 1 msec. No speech-processing experiments have been done using the second technique, and it is set forth primarily as a stimulus to future research. Furthermore, the processing of unvoiced speech, female speech, or naturally-reverberated speech has not been attempted. Although the artificial reverberation used in simulations was reasonably severe, it is not certain that either filtering method would work as well for natural reverberation.

A separate problem treated in the thesis is the choice of an appropriate method of resynthesizing the dereverberated speech waveform from its recovered cepstrum. This is complicated by the lack of speech phase information in the cepstrum. The basic result is to demonstrate the compatability of the filtering process with the resynthesis procedure used by Oppenheim in the homomorphic vocoder[7]. Certain interesting side issues, such as the influence of the input sectioning window upon the accuracy of pitch detection as well as upon echo removal, are treated. As achievement of sufficiently high-quality synthetic speech has been a major problem in this work, it is felt that a discussion of some of the important factors influencing this quality will be useful to the reader. For ease of reading, however, many details have been omitted.

The remainder of this paper is arranged in the following manner. Section II deals with important background material on homomorphic deconvolution, treated in terms of the complex cepstra of speech waveforms and room "impulse responses." Section III covers the speech-synthesis aspects of the homomorphic dereverberation process, and introduces the cepstrum. Section IV begins the discussion of cepstrum filtering techniques, including the average-difference process mentioned above; most of the important experimental results are treated here. Section V contains

the proposal for the memory/estimation filtering technique, also mentioned above. Section VI consists of conclusions and suggestions for further research.

II. Principles of Echo Removal by Homomorphic Deconvolution

A. Linear Time-Invariant Model for the Reverberation Process

A reverberated waveform may be usefully and not unreasonably modelled as a superposition of weighted and delayed replicas of an unreverberated source waveform. On the basis of this idealization, a reverberant speech waveform detected by a stationary microphone can be expressed as

$$x(t) = \int_{-\infty}^{t} s(\tau)p(t,\tau)d\tau \ ,$$

where $s(t)$ is the unreverberated speech waveform and $p(t,\tau)$ is the "impulse response" of the room. Although $p(t,\tau)$ is generally made non-stationary by source motion, it is sufficient for our purposes to view it as stationary:

$$p(t,\tau) = p(t-\tau, \tau_o) = p(t-\tau) \ .$$

In the processing to be discussed, speech is processed in "sections" of about 50 msec in duration, over which period source motion is typically small. Hence, we shall treat the reverberated speech waveform as a convolution:

$$x(t) = \int_{-\infty}^{t} s(\tau)p(t-\tau)d\tau = s(t) * p(t) \qquad (1)$$

The dereverberation problem can then be interpreted as one of deconvolution.

The case of a speech waveform distorted by a single echo of itself illustrates the above notions. In this case, $x(t)$ would take the form

$$x(t) = s(t) + a_1 s(t-t_1) = s(t) * p(t),$$

from which $p(t)$ is seen to be a "train" of two impulses:

$$p(t) = u_o(t) + a_1 u_o(t-t_1), \qquad (2a)$$

where t_1 is the interpath delay. Similarly, if there are M sharply

defined echoes, then

$$p(t) = u_o(t) + \sum_{i=1}^{M} a_i u_o(t-t_i). \tag{2b}$$

If the i^{th} path includes dispersive media, then the i^{th} echo will not be "sharply defined"; that is, p(t) takes on the more general form

$$p(t) = u_o(t) + \sum_{i=1}^{M} h_i(t-t_i). \tag{2c}$$

Such "echoes" will turn out to be more difficult to remove by the signal processing described in this paper than sharply defined echoes. Unnecessary complication will be avoided here by limiting the discussion of process theory to the case of sharply defined echoes.

Given the above time-invariance approximation, the extraction of s(t) from x(t) by any method is equivalent to inverse filtering, or passing x(t) through a linear system having time-invariant impulse response $p^{-1}(t)$, where

$$p^{-1}(t) * p(t) = u_o(t).$$

To do this exactly requires complete knowledge of p(t), an undesirable requirement from the point of view of practicality. A practical approach to deconvolving s(t) and p(t) is based on using limited information about these signals to greatest possible advantage. This problem has three facets: (1) identification of a set of characteristics distinguishing s(t) from p(t); (2) development of signal processing which can utilize and/or enhance these differences; (3) development of a method of reconstructing an acceptable speech waveform from a minimal amount of information about s(t), allowing maximum elimination of p(t). All three aspects, as they apply to homomorphic dereverberation, will be treated here.

B. Homomorphic Dereverberation of Speech

At this point it is necessary to switch from continuous-time notation to discrete-time notation. For the type of system to be discussed below, a continuous-time interpretation leads to serious mathematical difficulties. In particular, signals of infinite amplitude are encountered. In a discrete interpretation, however, this does not occur. The theory is accordingly phrased in discrete terms and implemented using a digital computer.

Let us assume that a reverberated speech waveform is indeed a convolution:

$$x(t) = s(t) * p(t) ,$$

where

$$p(t) = u_o(t) + \sum_{i=1}^{M} a_i u_o(t-t_i).$$

Define the discrete-time sequence x(n) by

$$x(n) \triangleq x(t)\Big|_{t=nT} , \tag{3a}$$

where n takes on integer values only; then x(n) is the value of a "sample" of x(t) taken at time nT, where T is an appropriately chosen sampling period. Similarly, define

$$s(n) \triangleq s(t)\Big|_{t=nT} . \tag{3b}$$

To remain consistent with the interpretation of x(t) as a convolution, we must define a p(n) such that x(n) is the discrete convolution of s(n) with p(n):

$$x(n) = \sum_{k=-\infty}^{+\infty} s(k)p(n-k) = s(n) * p(n). \tag{4}$$

Assume that p(t) has the form of Equation 2b, that s(t) is bandlimited, and that T is chosen to be less than or equal to the Nyquist sampling period. Then from

$$x(t) = s(t) + \sum_{i=1}^{M} a_i s(t-t_i)$$

and the bandlimited interpolation formula

$$s(t-t_i) = T \sum_{m=-\infty}^{+\infty} s(m) \frac{\sin \frac{\pi}{T}(t-t_i-mT)}{\pi(t-t_i-mT)} ,$$

it follows that

$$p(n) = u_o(n) + \sum_{i=1}^{M} a_i \frac{\sin \frac{\pi}{T}(nT-t_i)}{\frac{\pi}{T}(nT-t_i)} ,$$

where $u_o(n)$ is the unit sample:

$$u_o(n) = 1 , \quad n = 0$$
$$= 0 , \quad n \neq 0 .$$

If all the t_i's are integer multiples of T, then

$$p(n) = u_o(n) + \sum_{i=1}^{M} a_i u_o(n-n_i), \quad n_i = \frac{t_i}{T} ,$$

an ideal correspondence with p(t); but if the t_i's are not integer multiples of T, the relation is more complicated. For example, suppose

$$p(t) = u_o(t) + u_o(t- \frac{33T}{2}) .$$

Then

$$p(n) = u_o(n) + \frac{(-1)^{n+1}}{\pi(n- \frac{33}{2})} ,$$

as shown in Figure 1. As in the case of dispersive media, this spreading of energy is an undesirable effect. However, no special treatment of the problem is given below; the simpler form of Equation 6 is assumed.

Figure 1

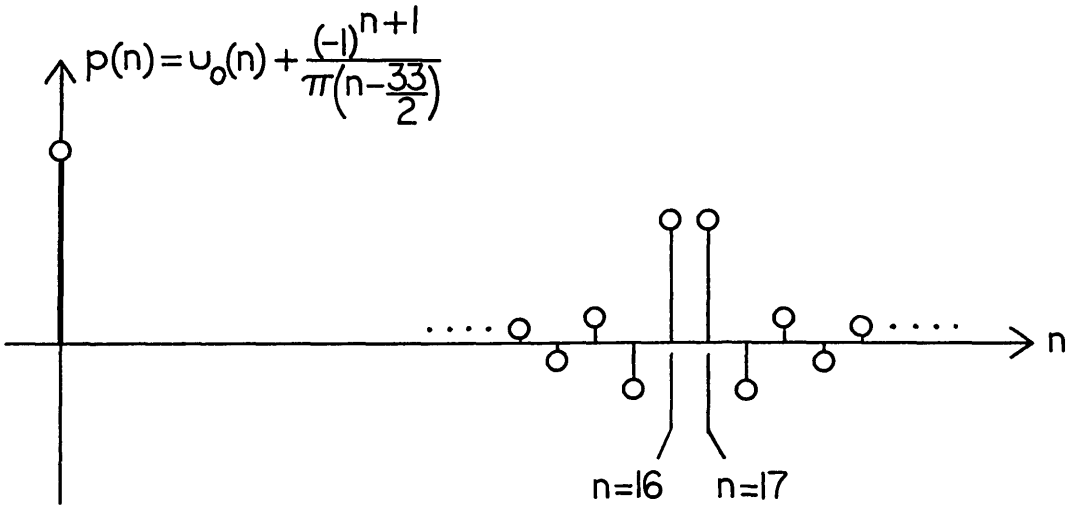
Sample sequence p(n), defined such that

$$s(n) * p(n) = [s(t) * p(t)]_{t = nT} ,$$

where

$$p(t) = u_o(t) + u_o(t - \frac{33T}{2})$$

and s(t) is appropriately bandlimited.


$p(n) = u_0(n) + \frac{(-1)^{n+1}}{\pi\left(n - \frac{33}{2}\right)}$
n
n=16
n=17

Much is known about sophisticated methods for separating added signals by linear filtering, particularly when only partial information about one or both signals is available. Not nearly as much can be said about convolved pairs of signals, such as $s(n) * p(n)$. It is possible, though, to transform a convolved signal pair into an added pair in such a way as to allow the use of linear filtering techniques. The resulting approach to separating $s(n)$ and $p(n)$ is most easily expressed symbolically. Let $A[-]$ be a transformation having the property

$$A[s(n) * p(n)] = \hat{s}(n) + \hat{p}(n) \tag{7a}$$

where

$$\begin{aligned} \hat{s}(n) &= A[s(n)] \\ \hat{p}(n) &= A[s(n)] \ . \end{aligned} \tag{7b}$$

If $p(n)$ were known (which is only partially true in practice), the ideal linear separation of $\hat{s}(n)$ and $\hat{p}(n)$ could be accomplished:

$$[\hat{s}(n) + \hat{p}(n)] - \hat{p}(n) = \hat{s}(n) \ . \tag{8}$$

If A is invertible (i.e., a one-to-one mapping), $s(n)$ can be recovered:

$$s(n) = A^{-1}[\hat{s}(n)]. \tag{9}$$

Hence, the ideal overall process can be expressed as

$$\underbrace{s(n)}_{\text{output}} = A^{-1}\big[A[\underbrace{s(n) * p(n)}_{\text{input}}] - A[p(n)]\big] \ . \tag{10}$$

The transformation A can be "realized" as a sequence of three invertible operations:

(1) Evaluation of the Fourier transform, $S(e^{j\omega})P(e^{j\omega})$, of $s(n) * p(n)$. The Fourier transform of a sequence is its z-transform evaluated on the unit circle, $z=e^{j\omega}$. This transform always exists for the finite-duration sequences with which we shall be concerned.

(2) Evaluation of the complex logarithm of $S(e^{j\omega})P(e^{j\omega})$. The object of this is to separate the product into a sum:

$$\log S(e^{j\omega})P(e^{j\omega}) = \log S(e^{j\omega}) + \log P(e^{j\omega}) \ , \tag{11a}$$

ultimately leading to satisfaction of Equation 7a. Equation 11a must hold for real and imaginary parts:

$$\log|S(e^{j\omega})||P(e^{j\omega})| = \log|S(e^{j\omega})| + \log|P(e^{j\omega})| \tag{11b}$$

$$\angle\{S(e^{j\omega})P(e^{j\omega})\} = \angle\{S(e^{j\omega})\} + \angle\{P(e^{j\omega})\} \tag{11c}$$

where

$$X(e^{j\omega}) = |X(e^{j\omega})|e^{j\angle X(e^{j\omega})} \ .$$

Equation 11b is always satisfied; but Equation 11c does not necessarily hold. The value of the "phase angle", $\angle X(e^{j\omega})$, is only determinate to within an integer multiple of 2π, and one's natural inclination is to resolve this ambiguity by always specifying the principal value[3] of the phase, which lies between 0 and 2π, when computing the complex logarithm. Equation 11c does not generally hold for the principal value, since the left-hand side is bounded by 2π while the right-hand side is only bounded by 4π. Satisfaction of Equation 11a can be accomplished if the phase is computed in a certain manner[3], but this leads to practical difficulties that will be discussed in Part III.

(3) Evaluation of the inverse Fourier transform of $\hat{S}(e^{j\omega}) + \hat{P}(e^{j\omega})$, where $\hat{S}(e^{j\omega})$ and $\hat{P}(e^{j\omega})$ denote $\log S(e^{j\omega})$ and $\log P(e^{j\omega})$, respectively:

$$\hat{S}(e^{j\omega}) + \hat{P}(e^{j\omega}) \longrightarrow \hat{s}(n) + \hat{p}(n).$$

Note that since each of these steps is invertible, the inverse transformation, A^{-1}, is also defined.

It is conceptually helpful to represent the above sequence of operations as blocks in a system, as shown in Figure 2a. Figure 2b

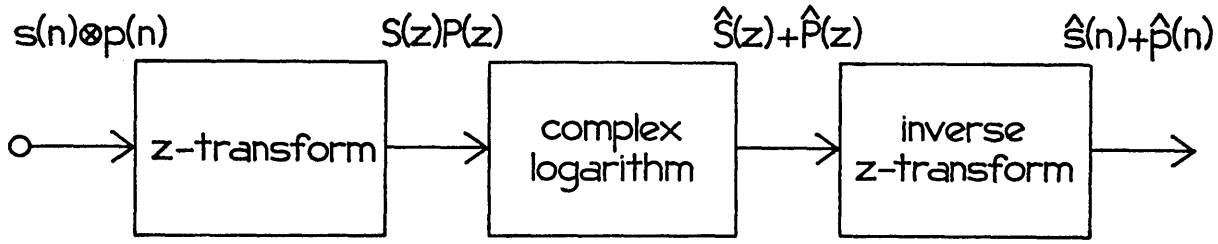
Figure 2

(a) Block diagram representation of operation sequence used to transform $s(n) * p(n)$ into $\hat{s}(n) + \hat{p}(n)$ (transformation 'A').

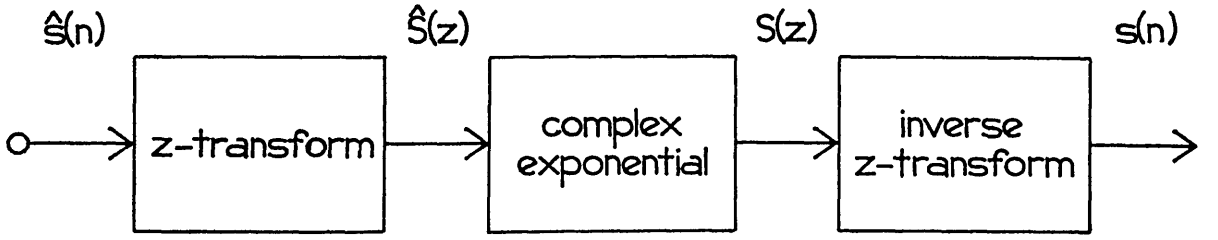
(b) Similar representation for operation sequence which transforms $\hat{s}(n)$ back into $s(n)$ (transformation 'A^{-1}').

(c) Canonic form of homomorphic deconvolution system, where operator L represents a linear filter.

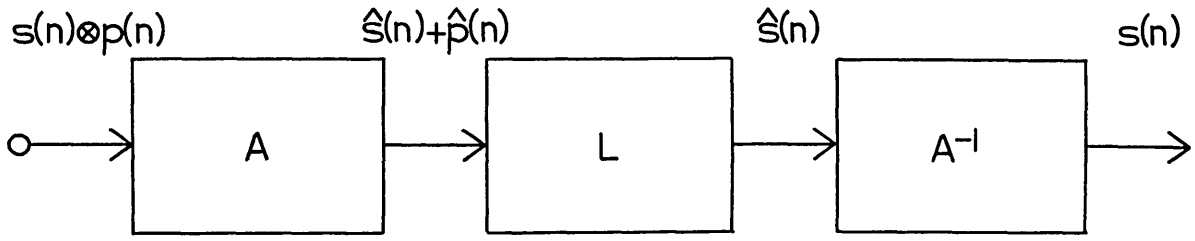
Note: Symbol '$\otimes$' in drawing, corresponding to '*' in text, denotes discrete convolution.


s(n)⊗p(n)
S(z)P(z)
Ŝ(z)+P̂(z)
ŝ(n)+p̂(n)
z-transform
complex logarithm
inverse z-transform

(a)


ŝ(n)
Ŝ(z)
S(z)
s(n)
z-transform
complex exponential
inverse z-transform

(b)



(c)

illustrates the inverse sequence of operations. These "systems" are representative of a class known as homomorphic systems, reflecting interpretation of the inputs and outputs as elements of vector spaces and the relation between the spaces as a homomorphism[5]. In simple terms, this means that A and A^{-1} each satisfy a "generalized linearity property", which, in this case, is

$$A\{[s_1(n)]^{a_1} * [s_2(n)]^{a_2}\} = a_1 A[s_1(n)] + a_2 A[s_2(n)],$$

with the inverse relation for A^{-1}.

Since A transforms the convolution $s(n) * p(n)$ into the sum $\hat{s}(n) + \hat{p}(n)$, a possibility is to use selective linear filtering to recover $\hat{s}(n)$ from this sum. The ideal linear filter for this purpose can be represented in terms of the linear operator L[-], where

$$L[\hat{s}(n) + \hat{p}(n)] = L[\hat{s}(n)] + L[\hat{p}(n)] = \hat{s}(n).$$

The composite operator $A^{-1}LA$, therefore, will produce s(n) from $s(n) * p(n)$; for any L, it is easily shown that

$$A^{-1}LA[s(n) * p(n)] = A^{-1}LA[s(n)] * A^{-1}LA[p(n)],$$

and the above choice for L determines that

$$A^{-1}LA[s(n)] = s(n)$$

$$A^{-1}LA[p(n)] = u_o(n) .$$

This operator also satisfies a generalized linearity property, and so may be viewed as representing in canonic form a class of "homomorphic deconvolution systems" illustrated symbolically in Figure 2c.

In general, if

$$\hat{x}(n) = A[x(n)],$$

then $\hat{x}(n)$ is called the complex cepstrum of x(n). For example, $\hat{s}(n)$, $\hat{p}(n)$, and $\hat{s}(n) + \hat{p}(n)$ are the complex cepstra of s(n), p(n), and $s(n) * p(n)$,

respectively. We next discuss some basic properties of these complex cepstra which are useful in selecting an approach to the linear separation problem.

C. Complex Cepstra of Speech Waveforms and Small-Room Impulse Responses

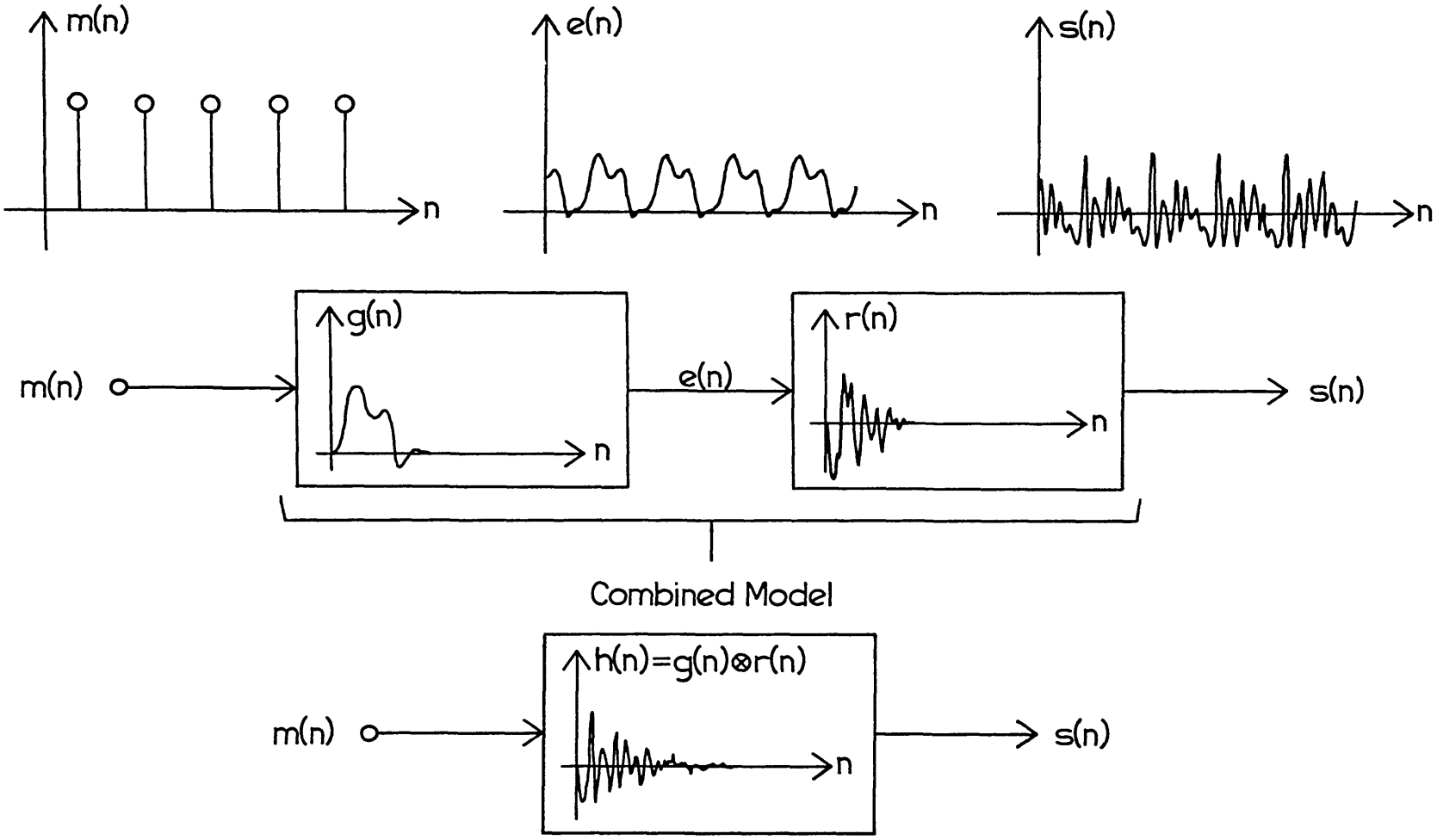
It is often observed that $\hat{s}(n)$ and $\hat{p}(n)$ each tend to consist largely of a sequence of narrow, separated peaks, except for a broader concentration of energy near the origin in $\hat{s}(n)$. When the peaks of $\hat{p}(n)$ do not overlap those of $\hat{s}(n)$, the possibility exists for removing them by selectively multiplying by zero the regions where these peaks lie. Such an approach to filtering complex cepstra has been called "frequency-invariant" filtering (of the log spectrum)[3], because it is the dual of linear time-invariant filtering. Simple models of speech waveforms and room "impulse responses" are helpful in understanding why this is a reasonable approach.

A linear time-varying system excited by a train of pulses, representing the vocal tract as excited by puffs of air, provides a useful model for the production of voiced speech[6]. As shown at the top of Figure 3, the pulse train, e(n), can be thought of as the output of a linear system with impulse response g(n), excited by a series of impulses, m(n). In turn, e(n) excites a system having the vocal tract impulse response, r(n). Although g(n) and r(n) are more accurately time-varying impulse responses, the typically slow variation of vocal tract and vocal cord characteristics in actual speech production allow us to model these as approximately time-invariant over short analysis intervals of about 50 msec. For convenience, g(n) and r(n) can be convolved and interpreted as the impulse response, h(n), of a single linear time-invariant system.

Figure 3

Simple linear time-invariant model for speech waveform production, approximately valid over short intervals. Sequence g(n) is glottal pulse and r(n) is vocal tract impulse response.

Note: Symbol '⊗' in drawing, corresponding to '*' in text, denotes discrete convolution.


m(n)
n
e(n)
n
s(n)
n
m(n)
g(n)
n
e(n)
r(n)
n
s(n)
Combined Model
m(n)
h(n)=g(n)⊗r(n)
n
s(n)

The speech waveform, s(n), is then the convolution of the impulse train, m(n), with h(n). The spacing between impulses in m(n) determines the short-time fundamental frequency, or pitch, of s(n).

As pitch and vocal tract characteristics change slowly, a 50-msec segment of a speech waveform appears as a section of a periodic signal. This is reflected in the frequency spectrum of the segment. Figure 4 shows the magnitude spectrum of the short piece of speech waveform illustrated in Figure 3. Prior to spectral analysis, this segment was weighted with a Hamming window

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right), \quad 0 \leq n \leq N-1$$
$$= 0 \text{ otherwise} \quad ,$$

which, because of the low sidelobes of its transform, reduces the spectral distortion resulting from truncation of the speech waveform. The spectrum of Figure 4 exemplifies two basic features typical of most short-time voiced speech spectra. One is the series of narrow peaks spaced at equal intervals in frequency, a result of the periodicity of the speech segment. The other is a slowly-varying spectral envelope, which is very nearly $|H(e^{j\omega})|$, the magnitude spectrum of h(n). This envelope is not greatly distorted when the segment is weighted by a Hamming window.

From the nature of its spectrum, the form of the complex cepstrum, $\hat{s}(n)$, of a short section of speech can be deduced. The spectrum of s(n) can be viewed as a product of two components: $H(e^{j\omega})$, which provides the spectral envelope, and $M(e^{j\omega})$, the transform of the impulse train excitation m(n).

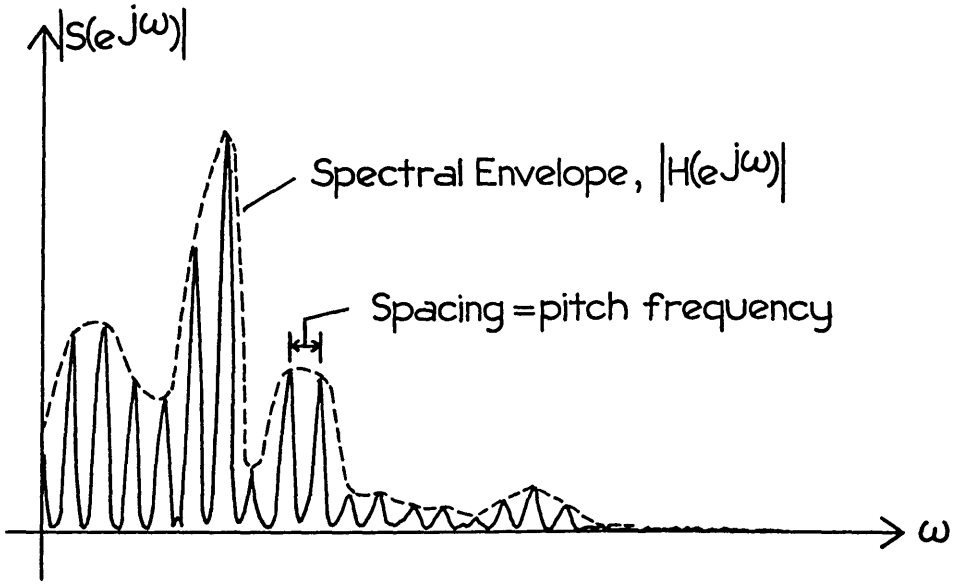
$$S(e^{j\omega}) = H(e^{j\omega})M(e^{j\omega})$$

The transform of $\hat{s}(n)$ is the logarithm of $S(e^{j\omega})$:

$$\log S(e^{j\omega}) = \log H(e^{j\omega}) + \log M(e^{j\omega})$$

Figure 4

Short-time magnitude spectrum of segment of speech waveform s(n) shown in Figure 3. Segment was pre-weighted with Hamming window to reduce spectral distortion.


|S(e^jω)|
Spectral Envelope, |H(e^jω)|
Spacing = pitch frequency
ω

or

$$\hat{S}(e^{j\omega}) = \hat{H}(e^{j\omega}) + \hat{M}(e^{j\omega}).$$

The complex cepstrum of a section of voiced speech is thus the sum of two distinctly different components, $\hat{h}(n)$ and $\hat{m}(n)$:

$$\hat{s}(n) = \hat{h}(n) + \hat{m}(n).$$

Since $H(e^{j\omega})$ is a slowly-varying function, $\hat{H}(e^{j\omega})$ is also slowly-varying. Therefore, $\hat{h}(n)$, which corresponds to the spectral envelope, occupies the low-time region of $\hat{s}(n)$, as shown in Figure 5. On the other hand, $\hat{M}(e^{j\omega})$, like $M(e^{j\omega})$, exhibits peaks equally spaced in frequency.* Hence, $\hat{m}(n)$ consists of a series of peaks spaced periodically in time, at multiples of the pitch period. As shown in Figure 5, $\hat{s}(n)$ extends into both positive and negative time, even if $s(n) = 0$ for $n < 0$. This is because the real and imaginary parts of $\log S(e^{j\omega})$, which are equivalent to the log magnitude and phase of $S(e^{j\omega})$, respectively, are not generally Hilbert transforms of each other, as required for $\hat{s}(n)$ to be "causal".[3] Furthermore, even if a short section of $s(n)$ is analyzed, $\hat{s}(n)$ is generally infinite in duration, although it tends to die out faster than $1/|n|$.[3]

Next, let us consider the complex cepstrum of a simple room impulse response, $p(n)$. Let $p(n)$ consist of two impulses, corresponding to the case of a single echo of amplitude a_1:

$$p(n) = u_o(n) + a_1 u_o(n-n_1).$$

Then

$$P(e^{j\omega}) = 1+a_1 e^{-j\omega n_1},$$

which is recognized as a periodic function of frequency. Figure 6 shows a

---

*The frequency of this periodic variation in frequency is sometimes called the "quefrency", to distinguish it from the frequency itself. Quefrency, of course, has the dimensions of time.

Figure 5

Typical complex cepstrum of speech waveform segment.

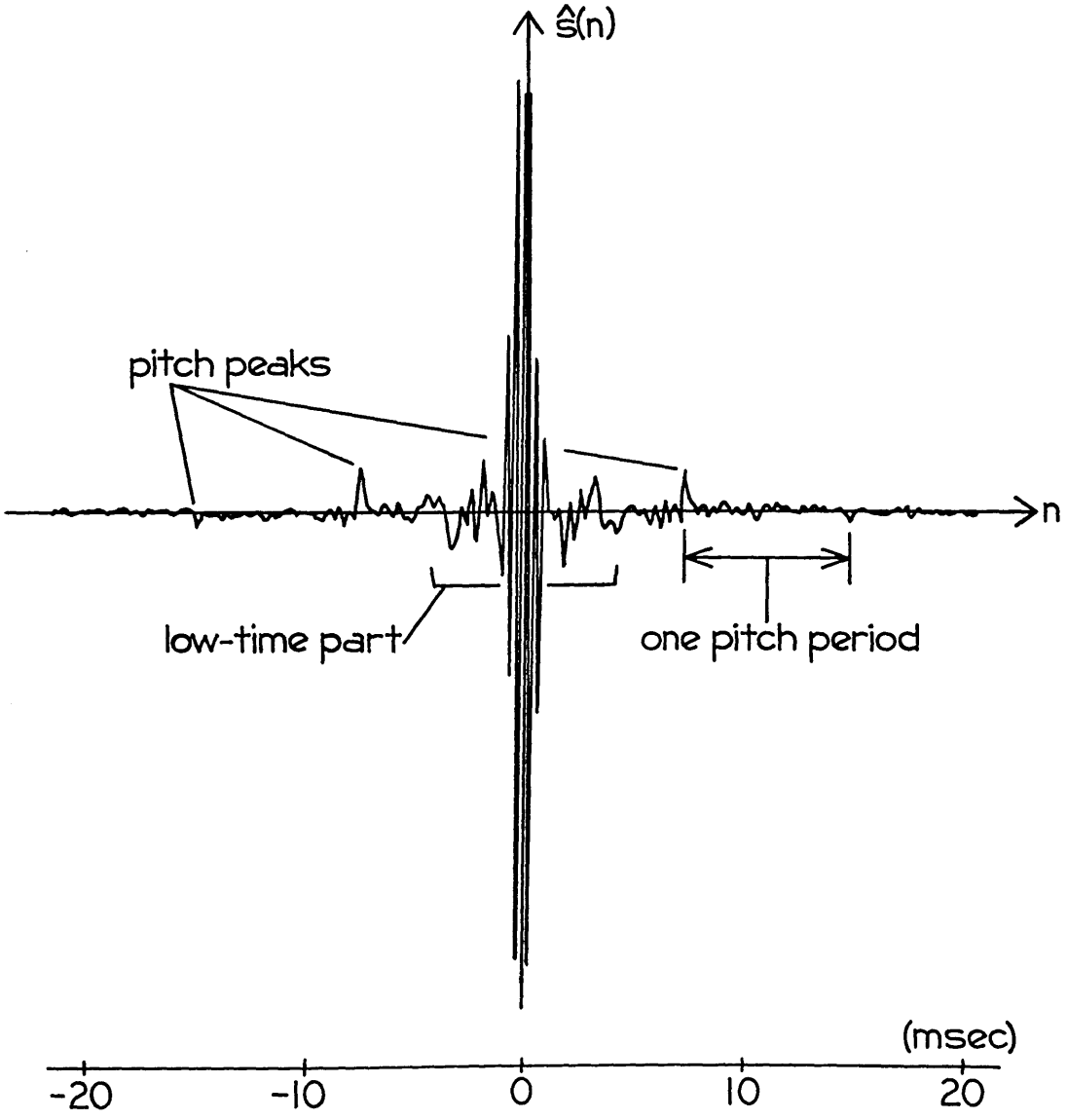
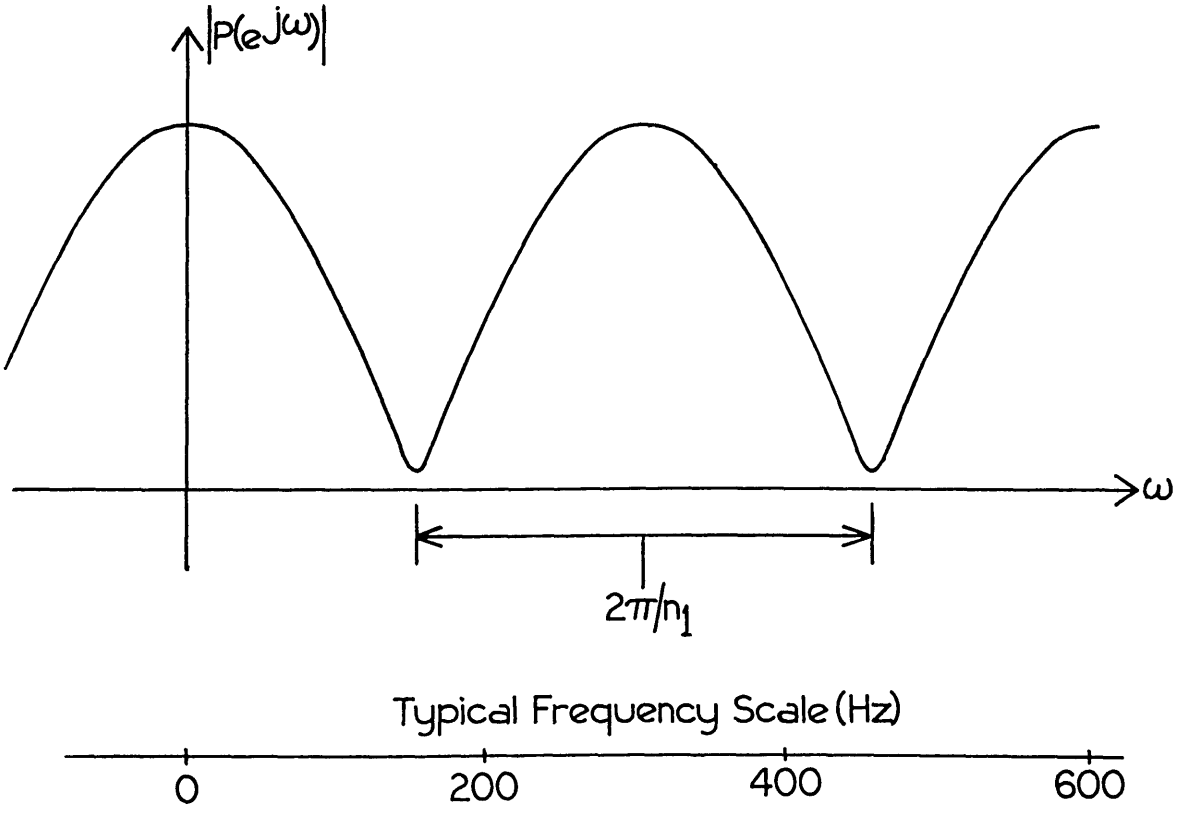

ŝ(n)
pitch peaks
n
low-time part
one pitch period
(msec)
-20
-10
0
10
20

Figure 6

Magnitude spectrum of simple room "impulse response" $p(n)$, where

$$p(n) = u_o(n) + a_1 u_o(n-n_1).$$

The typical "valleys" in $|P(e^{j\omega})|$ are responsible for the "comb-filtering" effect of simple reverberation.


|P(ejω)|
ω
2π/n1

Typical Frequency Scale (Hz)
0
200
400
600

plot of the magnitude of $P(e^{j\omega})$. The valleys in the room frequency response, as can be seen from Figure 6, would cause considerable distortion of the sound reaching the microphone.

In this case, log $P(e^{j\omega})$ (or $\hat{P}(e^{j\omega})$) is periodic in frequency, its period of variation being $2\pi/n_1$. Therefore, $\hat{p}(n)$, like $\hat{m}(n)$, consists of a series of peaks, these spaced at intervals of n_1 in time, as shown in Figure 7. For this example, $\hat{p}(n)$ occupies only positive time, though in general $\hat{p}(n)$ extends into negative time also.

The complex cepstrum of $s(n) * p(n)$ is simply the sum of $\hat{p}(n)$ and $\hat{s}(n)$. Figure 8 illustrates the desirable situation in which $\hat{p}(n)$ does not overlap the major concentrations of energy in $\hat{s}(n)$. The arrangement of peaks in Figure 8 is not unreasonable, since values for the echo time, n_1, typically fall between 1 and 10 ms, while the pitch period of a speech waveform ordinarily falls between 5 and 20 ms.

Realistic room impulse responses contain more than one echo, of course. It is not generally feasible to determine analytically the form of $\hat{p}(n)$ for more complicated $p(n)$. Usually many more peaks appear in a pattern not simply related to $p(n)$, and the peaks may not be as sharp as for the case of a simple echo. Since the overlap of $\hat{p}(n)$ with $\hat{s}(n)$ increases, it becomes more difficult to recover $\hat{s}(n)$.

If it is known where to expect to find peaks of $\hat{p}(n)$, then a reasonable method of eliminating $\hat{p}(n)$ might be to set the complex cepstrum to zero at these points. This is easy to do when $\hat{s}(n) + \hat{p}(n)$ is stored in an array or register of a digital computer. Of course, such an operation may cause varying degrees of "damage" to $\hat{s}(n)$, depending on the values of $\hat{s}(n)$ at the zeroed locations.

Figure 7

Complex cepstrum, $\hat{p}(n)$, of

$$p(n) = u_o(n) + a_1 u_o(n-n_1).$$

Note, in this instance, that

$$\hat{p}(n) = 0, \quad n < 0,$$

although this is generally not true. Also, $\hat{p}(n)$ decays faster than $1/|n|$ for $|a_1| < 1$.

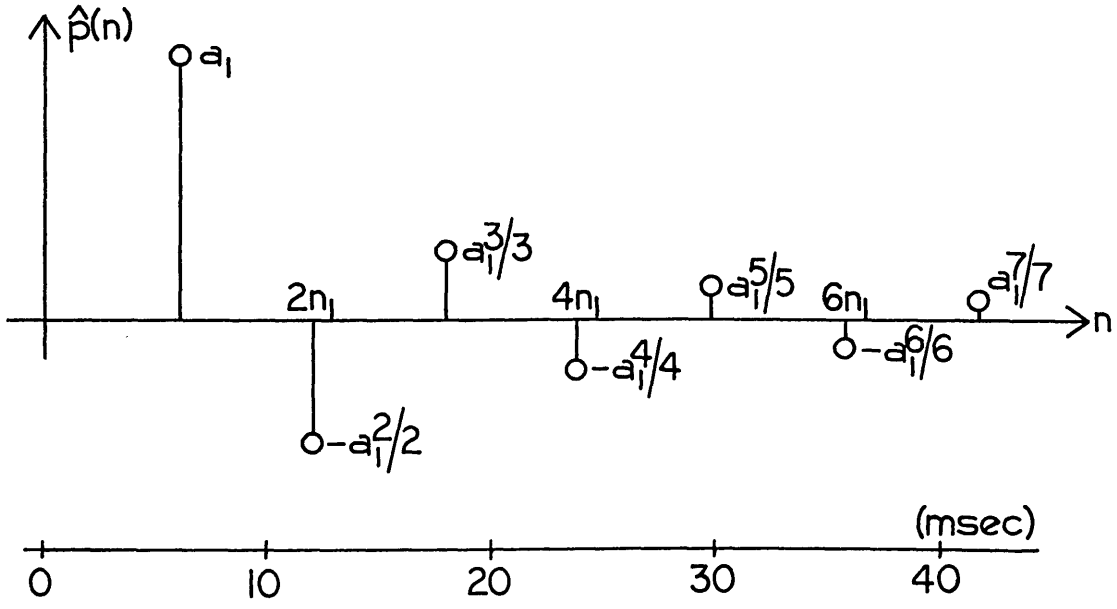
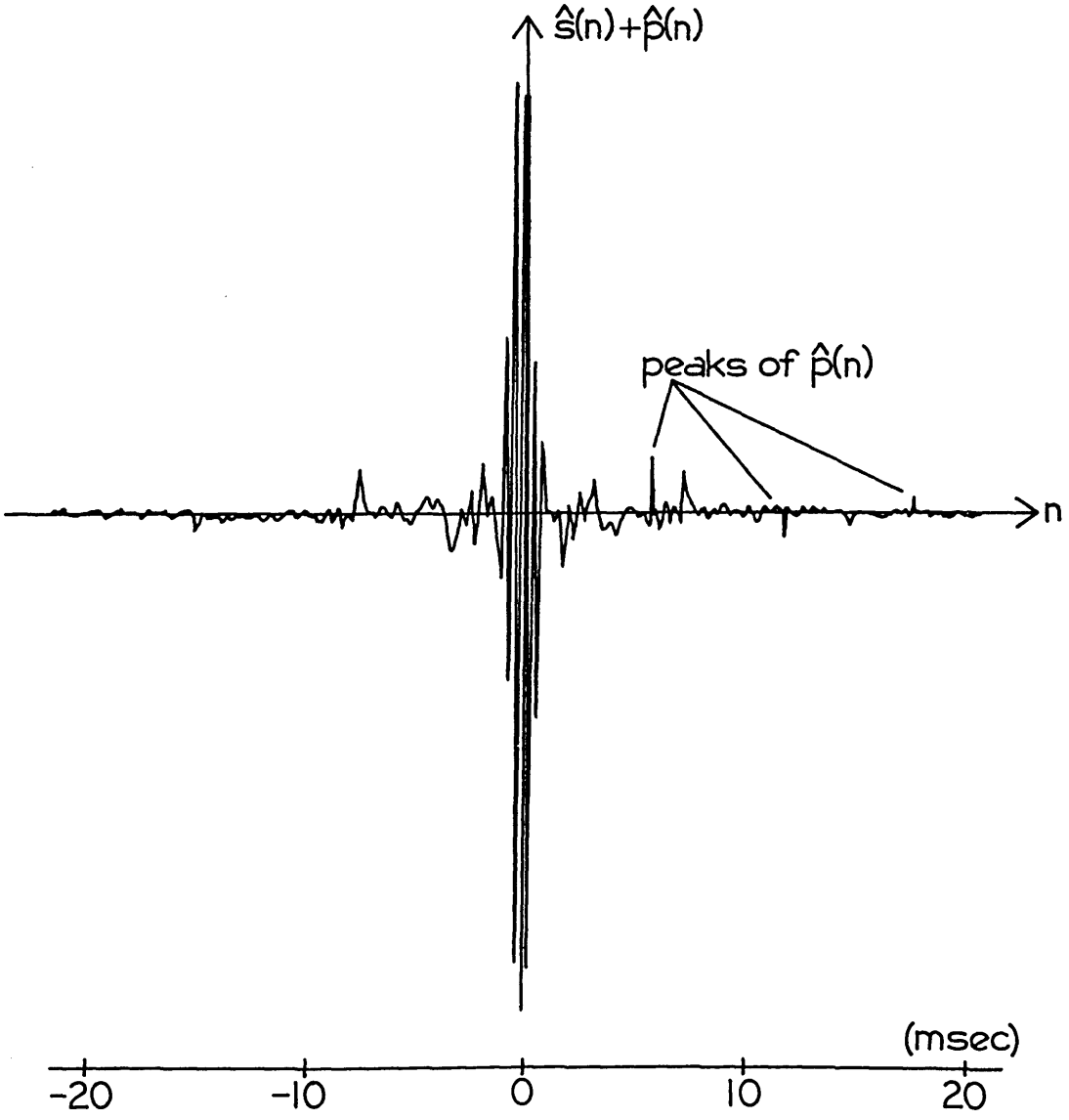

p̂(n)
a1
2n1
-a1^2/2
a1^3/3
4n1
-a1^4/4
a1^5/5
6n1
-a1^6/6
a1^7/7
n
(msec)
0
10
20
30
40

Figure 8

Complex cepstrum of reverberated speech waveform, $s(n) * p(n)$, showing "reverberation peaks" which must be removed to recover $\hat{s}(n)$.


ŝ(n)+p̂(n)
peaks of p̂(n)
n
(msec)
-20
-10
0
10
20

Given only $\hat{s}(n) + \hat{p}(n)$, however, the information needed to identify the peaks of $\hat{p}(n)$ is not readily available. Another approach might be to set to zero all parts of the complex cepstrum in which the values of $\hat{s}(n)$ are expected to be so small as to be insignificant. The more extensive the zeroed regions are, the more likely it is that they will include all the peaks of $\hat{p}(n)$. Unfortunately, determining which regions of $\hat{s}(n)$ are significant poses a problem similar to that of finding the peaks of $\hat{p}(n)$. It is difficult to distinguish peaks of $\hat{s}(n)$ from those of $\hat{p}(n)$ by automatic means.

A more effective although less efficient approach to eliminating $\hat{p}(n)$ is to identify unwanted peaks by "comparing" the complex cepstra of two differently reverberated versions of the same section of speech. That is, compare $\hat{s}(n) + \hat{p}_1(n)$ with $\hat{s}(n) + \hat{p}_2(n)$; where they differ significantly, there is a peak in either $\hat{p}_1(n)$ or $\hat{p}_2(n)$. These two complex cepstra are obtainable from the outputs of two differently-positioned microphones; since the reverberation would be different in each case, the $\hat{p}(n)$'s will likewise be substantially different. This idea will be treated in more detail in Section IV.

D. Summary

A reverberated speech waveform is conveniently represented as a convolution $s(n) * p(n)$, where $s(n)$ is the unreverberated speech waveform and $p(n)$ is the "impulse response" of the reverberant room. It is proposed to recover $s(n)$ from $s(n) * p(n)$ by homomorphic deconvolution, which consists of:

(1) transforming $s(n) * p(n)$ into a sum of corresponding signals, $\hat{s}(n) + \hat{p}(n)$;

(2) linear-filtering the sum to eliminate $\hat{p}(n)$;

(3) transforming the result, $\hat{s}(n)$, back into $s(n)$.

The characteristics of the "complex cepstra", $\hat{s}(n)$ and $\hat{p}(n)$, of $s(n)$ and $p(n)$ often make it possible to remove $\hat{p}(n)$ from $\hat{s}(n) + \hat{p}(n)$ without excessive distortion of $\hat{s}(n)$, if the peaks of $\hat{p}(n)$ can be accurately identified. One way of doing this is to compare two complex cepstra of the form $\hat{s}(n) + \hat{p}_1(n)$ and $\hat{s}(n) + \hat{p}_2(n)$, where $\hat{p}_1(n)$ and $\hat{p}_2(n)$ are substantially different.

In the next section it is seen that the cepstrum, or even part of the complex cepstrum, is more practically dealt with than the complex cepstrum. However, its use complicates the final step of the deconvolution process, recovery of $s(n)$ from the filtering result.

III. Aspects of Voiced Speech Resynthesis from the Even Part of the Complex Cepstrum

A. The Cepstrum

Recent vocoder work[7,9] has indicated that speech of reasonable quality can be resynthesized from spectral magnitude data, through the introduction of a minimum phase characteristic[7] which resembles the phase of natural speech. This suggests a possible simplification of the homomorphic filtering process. In Section II it was seen that

$$\hat{S}(e^{j\omega}) = \text{F.T.}[\hat{s}(n)] = \log\{\text{F.T.}[s(n)]\} = \log S(e^{j\omega}),$$

where "F.T." denotes "Fourier transform." The real and imaginary parts of the above are

$$\hat{S}_R(e^{j\omega}) = \log|S(e^{j\omega})|$$

$$\hat{S}_I(e^{j\omega}) = \angle\{S(e^{j\omega})\} .$$

Since the inverse transform of $\hat{S}_R$ is the even part of $\hat{s}(n)$,

$$\hat{s}_{ev}(n) \longleftrightarrow \hat{S}_R(e^{j\omega}) ,$$

it follows that if acceptably good speech can be synthesized from $|S(e^{j\omega})|$ alone, then it should be necessary to recover only $\hat{s}_{ev}(n)$ from the reverberated input speech. This means that instead of computing $\hat{s}(n) + \hat{p}(n)$ from $s(n) * p(n)$ and removing $\hat{p}(n)$, only $\hat{s}_{ev}(n) + \hat{p}_{ev}(n)$ will be computed to start with, necessitating removal of $\hat{p}_{ev}(n)$. This process is, in most respects, similar to the process described in Section II, except that resynthesis of speech cannot be accomplished from direct "inverse transformation" of $\hat{s}_{ev}(n)$ as it could for $\hat{s}(n)$. In the author's work, the resynthesis technique employed follows closely that of Oppenheim's homomorphic vocoder[7].

The even part of the complex cepstrum is referred to as the cepstrum. Aside from the reason discussed above, certain practical considerations have favored the use of the cepstrum. One problem in computing the complex cepstrum is evaluation of the complex logarithm such that the relation

$$\log S(e^{j\omega})P(e^{j\omega}) = \log S(e^{j\omega}) + \log P(e^{j\omega})$$

of Equation 11a is satisfied. As discussed in Section II, the problem stems not from the real part of the equation,

$$\log|S(e^{j\omega})||P(e^{j\omega})| = \log|S(e^{j\omega})| + \log|P(e^{j\omega})|,$$

which is always true and suffices for determination of $\hat{s}_{ev}(n) + \hat{p}_{ev}(n)$, but the imaginary part,

$$\angle\{S(e^{j\omega})P(e^{j\omega})\} = \angle\{S(e^{j\omega})\} + \angle\{P(e^{j\omega})\}.$$

It is true that for any piecewise-continuous integer-valued function $N(\omega)$,

$$e^{j\angle X(e^{j\omega})} = e^{j\{2\pi N(\omega) + \angle X(e^{j\omega})\}};$$

thus, the actual phase functions produced by a computational algorithm for $S(e^{j\omega})P(e^{j\omega})$, $S(e^{j\omega})$, and $P(e^{j\omega})$ can generally have the form

$$\angle\{S(e^{j\omega})\} = \theta_s(\omega) + 2\pi N_s(\omega)$$

$$\angle\{P(e^{j\omega})\} = \theta_p(\omega) + 2\pi N_p(\omega)$$

$$\angle\{S(e^{j\omega})P(e^{j\omega})\} = \theta_s(\omega) + \theta_p(\omega) + 2\pi N_{sp}(\omega),$$

where $\theta(\omega)$ is the principle value of the phase and $N_{sp}(\omega)$ may not equal $N_s(\omega) + N_p(\omega)$. Schafer[3] has argued that one way of guaranteeing satisfaction of Equation 11c (the only way yet discovered) is to choose the unique $N(\omega)$ such that $\theta(\omega) + N(\omega)$ is a continuous function.* This involves finding

*This is always possible for a finite-duration sequence $x(n)$, for which $X(e^{j\omega})$ is continuous, implying that there is a unique continuous function from which the phase (as computed) may differ only by some multiple of 2π at any value of ω.

the discontinuities of 2π in $\theta(\omega)$, which can be viewed as the "continuous phase" in modulo-2π form. All of the algorithms proposed for "unwrapping" this phase function have required a fine-resolution spectral analysis or have otherwise necessitated an excessive amount of computation. This has been the most serious barrier to use of the complex cepstrum.

Another potential problem with the complex cepstrum arises in a filtering process involving the "comparison" of two or more complex cepstra. The benefits of comparison are realized only if the complex cepstra have a common component corresponding to the unreverberated speech, and differ only in the components due to reverberation, e.g.

$$\hat{x}_1(n) = \hat{s}(n) + \hat{p}_1(n)$$
$$\hat{x}_2(n) = \hat{s}(n) + \hat{p}_2(n) \qquad \hat{p}_1(n) \neq \hat{p}_2(n) \ .$$

Such complex cepstra can be derived from the outputs of two differently-placed microphones. But if the source-to-microphone distances are not equal, then there is a "differential delay" between $x_1(n)$ and $x_2(n)$; i.e.,

$$x_1(n) = s(n) * p_1(n)$$
$$x_2(n) = s(n-M) * p_2(n), \quad M \neq 0$$

which does not result in the desired condition.

However, the cepstrum provides a possible solution to this problem. Let the short-time spectrum of s(n) be defined by

$$S_w(e^{j\omega},r) = \sum_{n=0}^{N-1} w(n-r)s(n)e^{-j\omega n}$$
$$= \text{F.T.}[w(n-r)s(n)],$$

where w(n) is a duration-limited weighting window:

$$w(n) = 0, \quad n < 0, \quad n \geq N.$$

For window durations on the order of 40-60 msec, it is observed that the short-time magnitude spectra of speech waveforms tend to vary slowly with window position, r, while the short-time phase spectra vary considerably more rapidly. Thus, although the odd part of a complex cepstrum computed from S_w may be substantially altered by small changes in r, the even part or cepstrum will not, since it depends only on the magnitude spectrum. Applied to the above problem, this means that the short-time cepstrum of s(n-M) is approximately equal to the short-time cepstrum of s(n) if M is a number of samples corresponding to a delay of less than about 7.5 msec. This is roughly the delay that would be produced by a 7.5-foot difference in speaker-to-microphone distances. Since speech can be resynthesized from these short-time cepstra, use of the cepstrum provides a potentially feasible solution to the differential-delay problem. Experimental verification of this idea is discussed in Section IV.

It must be emphasized that these arguments against the complex cepstrum are not based on enough conclusive evidence to rule it out entirely. Most of the present investigation centered around the cepstrum because of the simplifications it allowed. Parallel experiments were not carried out using the complex cepstrum.

B. The Resynthesis Problem

It was seen above that the cepstrum has certain practical advantages over the complex cepstrum. In addition, $\hat{p}_{ev}(n)$ is often more easily separable from $\hat{s}_{ev}(n)$ than $\hat{p}(n)$ is from $\hat{s}(n)$, because the odd part of $\hat{s}(n)$ frequently causes a worse time-overlap with p(n) than the even part. Therefore the filtering problem is simplified by use of the cepstrum.

On the other hand, resynthesis of s(n) from $\hat{s}_{ev}(n)$ is complicated by the cepstrum's lack of speech phase information. In recovery of s(n)

from $\hat{s}(n)$ by "direct inverse transformation",*

$$s(n) = A^{-1}[\hat{s}(n)],$$

the odd part of $\hat{s}(n)$ produces the phase spectrum of s(n), while the even part produces the magnitude spectrum. The phase is necessary to preserve important temporal characteristics of the waveform: proper amplitude modulation, periodicity (or quasi-periodicity), incidence of voiced-unvoiced and unvoiced-voiced transitions, etc. In short, the magnitude spectrum determines how the signal energy is distributed among the sinusoidal components of the waveform, but is insufficient to determine the temporal energy distribution, for which phase information is also required. It follows that $A^{-1}[\hat{s}_{ev}(n)]$ is not s(n), and generally does not even have the same temporal properties.

Exact recovery of s(n) from $\hat{s}_{ev}(n)$ is therefore not possible. An interesting question is whether speech resynthesized from the processed magnitude spectrum but using the phase spectrum of the reverberated waveform sounds adequately dereverberated. The net effect of such a process is essentially equivalent to passing the unreverberated speech waveform through an all-pass filter having a phase equal to the phase spectrum of p(n). This was tried experimentally, and the results compared with the unreverberated input and the "completely" reverberated waveform. Unfortunately, there was little difference between the completely reverberated speech and that with only phase distortion. It can be concluded, therefore, that the phase of a reverberated speech waveform must either be processed to remove the effects of reverberation or replaced by a suitable "artificial" phase.

---

*Note that this inverse transformation process applies whether we mean a section of s(n) or s(n) in its entirety.

There are several possible approaches to approximate speech resynthesis using artificial phase. The most simple-minded is to synthesize a section of output waveform (corresponding to the analyzed input section) directly from the magnitude spectrum produced from $\hat{s}_{ev}(n)$ and an artificial phase. This is equivalent to creating an artificial complex cepstrum by adding an odd-symmetry sequence, $\hat{\underset{\sim}{s}}_{od}(n)$, to $\hat{s}_{ev}(n)$ and "inverse transforming":

$$\begin{matrix}\text{Section}\\ \text{of}\\ \text{Output}\end{matrix} = \underset{\sim}{s}(n) = A^{-1}[\hat{s}_{ev}(n) + \hat{\underset{\sim}{s}}_{od}(n)].$$

Synthesized sections would then be butted together to form the processed output signal.

The main problem is to choose a phase (or an $\hat{\underset{\sim}{s}}_{od}(n)$) which results in a satisfactory waveform. A convenient choice is minimum phase.[3] The minimum phase function, $\phi_{min}(\omega)$, where

$$S_{min}(e^{j\omega}) = |S(e^{j\omega})|e^{j\phi_{min}(\omega)},$$

is the Hilbert transform[3] of $\log |S(e^{j\omega})|$, and can therefore be produced by choosing

$$\begin{aligned}\hat{\underset{\sim}{s}}_{od}(n) &= \hat{s}_{ev}(n), \quad n > 0\\ &= 0 \qquad , \quad n = 0\\ &= -\hat{s}_{ev}(n), \quad n < 0.\end{aligned}$$

The resulting "minimum-phase complex cepstrum" is

$$\hat{s}_{min}(n) = \hat{s}_{ev}(n)k(n)$$

where

$$\begin{aligned}k(n) &= 2 \ , \quad n > 0\\ &= 1 \ , \quad n = 0\\ &= 0 \ , \quad n < 0.\end{aligned}$$

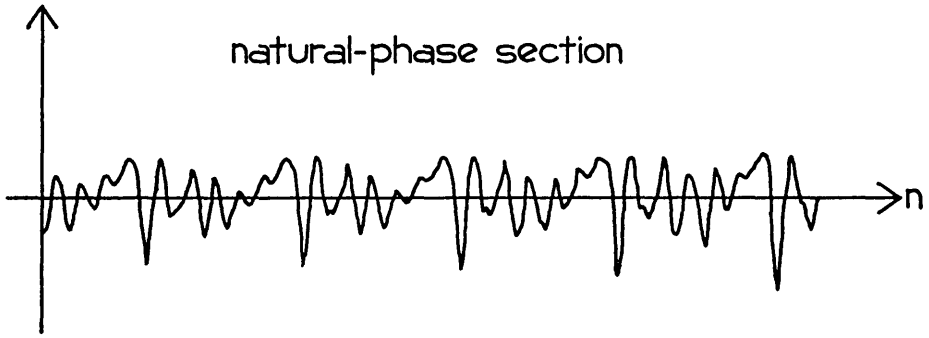
Sections of speech synthesized with minimum phase often closely resemble the original, natural-phase sections, as the example of Figure 9 illustrates. However, this case also exemplifies some of the undesirable characteristics of minimum-phase sections. These sections almost invariably decay as time increases, and generally appear time-shifted relative to the original section in such a way that the maximum peak occurs near $n = 0$. Therefore, even if the minimum-phase sections resemble their natural-phase counterparts in certain aspects (such as quasi-periodicity), successive sections do not "match" at their boundaries. Furthermore, since each section decays, a waveform produced by butting together minimum-phase sections has a "periodic" amplitude modulation not characteristic of the original waveform. The possibility exists for weighting the sections to compensate for this modulation; section boundaries could perhaps be matched by appropriate time-shifting, and "smoothed" together by overlapping sections and interpolating sample values. However, no successful method of doing this has been found.

Other phases are of course possible. Zero phase, ($\phi(\omega) = 0$), for example, can be used, but zero-phase sections tend to be even more distorted than minimum-phase sections. A typical zero-phase section is illustrated in Figure 10. Zero phase and minimum phase are mentioned here primarily because they are easily generated using the cepstrum. Some experimentation with other phases has been done; the results have not been satisfactory and are omitted here.

One might suspect that if successive sections do not overlap, then information common to two such sections,which is perhaps essential to make the synthesized sections compatible with each other, is lost. This would suggest that section overlap should be a large fraction of section length.

Figure 9

Comparison of natural-phase speech waveform section and minimum-phase section having same magnitude spectrum. Minimum-phase section was constructed by computing Discrete Fourier Transform[10] (DFT) of natural-phase section, replacing natural phase with minimum phase, and computing Inverse Discrete Fourier Transform (IDFT) of resulting spectrum.


natural-phase section
n

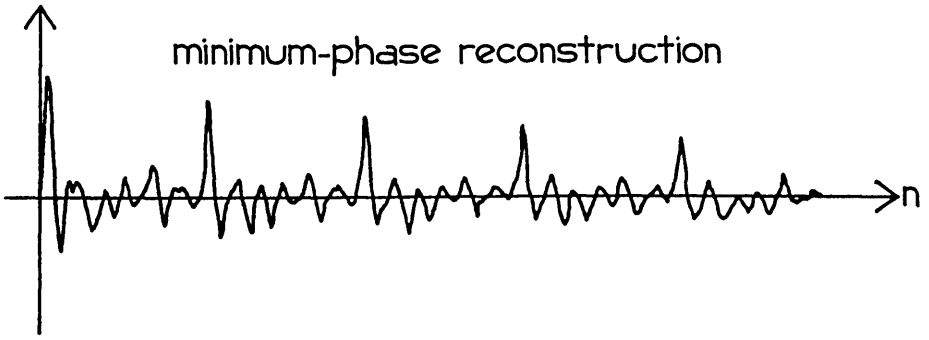
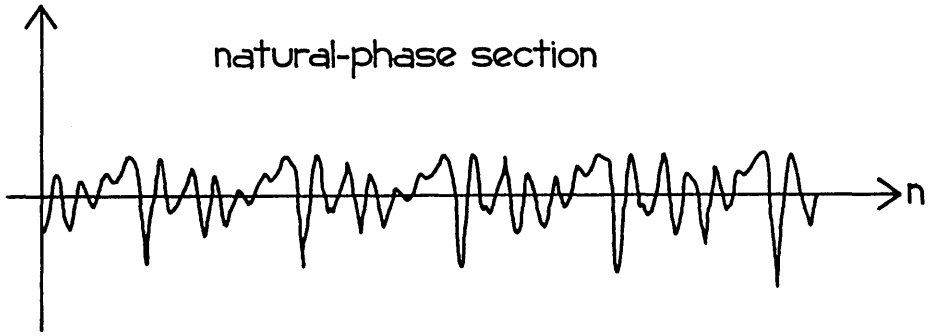
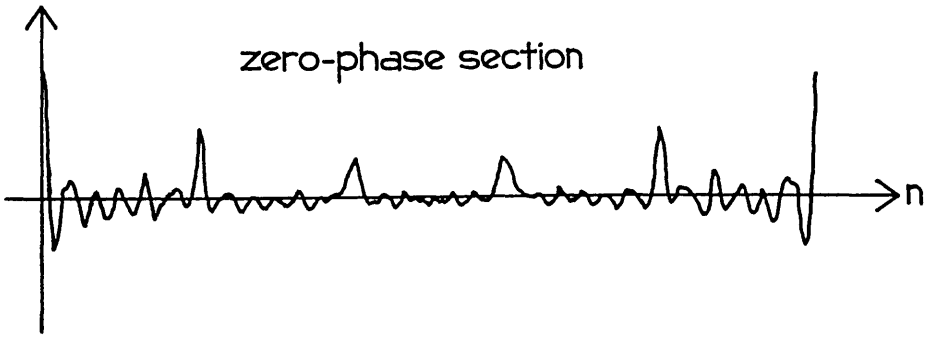

minimum-phase reconstruction
n

Figure 10

Comparison of natural-phase speech waveform section and zero-phase section with same magnitude spectrum. Zero-phase section was computed in same manner as minimum-phase section of Figure 9.


natural-phase section
n


zero-phase section
n

In the extreme case, sections of N-sample duration could be overlapped by N-1 points. For each input section, a corresponding processed output section (with artificial phase) would be produced. Then a complete waveform could be constructed as a composite of these sections, either by interpolating sample values in overlapping regions or allowing each section to contribute only a few samples to the output.

A difficulty in successfully accomplishing this is that time-shift information is lost with the discarded phase. For example, if successive N-sample sections are overlapped by N-1 samples, then the corresponding processed sections should look almost identical to one another except for a time-shift of one sample. But often this time-shift is completely lost, as is illustrated by the minimum-phase and zero-phase examples. Resynthesis techniques can introduce the shift artificially, but serious problems still remain.

Before considering an example, let us define the short-time "running" spectrum of a waveform x(n) by

$$X_w(n, e^{j\omega}) = \sum_{m=0}^{N-1} w(m)x(m+n)e^{-j\omega m}.$$

Computationally, we are limited to a discretely-sampled version of the above:

$$X_w(n, k) = X_w(n, e^{j\omega})\Big|_{\omega = \frac{2\pi k}{N}} = \sum_{m=0}^{N-1} w(m)x(n+m)e^{-j\frac{2\pi km}{N}}.$$

This is equivalent to the Discrete Fourier Transform[10] (DFT) of a weighted N-point section of x(n), or may be interpreted as the spectrum produced by a bank of N filters having impulse responses*

---

*I.e., the output of the k^{th} filter is seen to be

$$y_k(n) = X_w(n-N+1,k).$$

$$h_k(n) = w(N-1-n)e^{j\frac{2\pi(n+1)k}{N}}, \quad 0 \leq n < N$$

$$= 0 \quad \text{otherwise.}$$

The discrete spectrum is specified by N successive samples, k = 0, ... , N-1, because

$$e^{-j\frac{2\pi(k-rN)m}{N}} = e^{-j\frac{2\pi km}{N}}, \quad r \text{ integer.}$$

These N samples are sufficient to recover the section of x(n) between $n = n_o$ and $n = n_o+N-1$ according to the inversion formula

$$x(n) = \frac{1}{Nw(n-n_o)} \sum_{k=0}^{N-1} X_w(n_o, k)e^{j\frac{2\pi kn}{N}}, \quad n_o \leq n \leq N-1,$$

which is essentially equivalent to the Inverse Discrete Fourier Transform (IDFT) of $X_w(n_o, k)$.

We now return to the resynthesis problem. The procedure described below has been proposed as a possible solution. Difficulties encountered here are typical of those encountered with other "direct inverse transformation" approaches to synthesizing the processed speech waveform using artificial phase.

First, an artificial-phase running spectrum,

$$\underset{\sim}{S}(n, k) = |S_w(n, k)|e^{j\underset{\sim}{\phi}(n, k)},$$

is computed, where $|S_w(n, k)|$ is the magnitude spectrum of the weighted section of s(n), as recovered from $\hat{s}_{ev}(n)$. Next, consistent with the fact that

$$s(n) = \frac{1}{Nw(0)} \sum_{k=0}^{N-1} S_w(n, k)e^{j\frac{2\pi kn}{N}}, \quad \text{all } n,$$

the output waveform is synthesized from the running spectrum according to

$$\underset{\sim}{s}(n) = \frac{1}{Nw(0)} \sum_{k=0}^{N-1} \underset{\sim}{S}(n, k) e^{j\frac{2\pi kn}{N}}, \text{ all } n.$$

Note that this is <u>not</u> just a section of output waveform, but a complete waveform.

It is possible to interpret $\underset{\sim}{s}(n)$ as a composite of artificial-phase sections. Let $g_n(m)$ denote the sequence

$$g_n(m) = \frac{1}{Nw(0)} \sum_{k=0}^{N-1} \underset{\sim}{S}(n, k) e^{j\frac{2\pi km}{N}},$$

where n is treated here as constant. Thus $g_n(m)$ has period N:

$$g_n(m) = g_n(m+N), \text{ all } m.$$

Suppose that $\underset{\sim}{S}(n, k)$ is a minimum-phase spectrum; for example, let $|\underset{\sim}{S}(n, k)|$ equal the magnitude spectrum of the speech waveform section of Figure 9, and let the phase be the corresponding minimum-phase spectrum. Then the minimum-phase section of Figure 9 is one period of $g_n(m)$.* It follows that $g_n(m)$ is a periodically-decaying waveform with its maxima near multiples of N. Nearly all $g_n(m)$'s synthesized from minimum-phase sampled spectra are found to have these same characteristics. Now, observe that

$$\underset{\sim}{s}(n) = g_n(m)\Big|_{m=n} ;$$

i.e., $\underset{\sim}{s}(n)$ is the n^{th} sample of the n^{th} element of a sequence of periodically-decaying $g_n(m)$'s. Because all of the $g_n(m)$'s have similar characteristics, $\underset{\sim}{s}(n)$ itself will also exhibit these properties. That is, $\underset{\sim}{s}(n)$ will decay periodically (although it will not generally be periodic), and will have maxima near multiples of N! Similar phenomena can be

---

*This section was actually computed from the formula given for $g_m(n)$.

expected to occur whenever an artificial phase, which is entirely dependent on the magnitude spectrum, produces $g_m(n)$'s having common properties highly characteristic of that phase.

This occurs for zero phase as well as minimum phase. In an experiment, the sentence "The light flashed the message to the eyes of the watchers" was synthesized according to the above procedure using zero phase. As expected, the resulting waveform was characterized by a strongly periodic modulating envelope, with the envelope period dominating the sound of the sentence. The sentence was still intelligible, but quality was entirely unacceptable.

C. Pitch-Synchronous Synthesis

Probably the most serious problem with the above resynthesis approaches is that the proper quasi-periodic structure of voiced speech waveforms is not preserved. To solve this problem, the pitch-synchronous synthesis method used by Oppenheim in the homomorphic vocoder[7] was adopted. This, in principle, is the literal implementation of the speech-production model of Figure 3. Three basic operations are involved:

1. The low-time part of $\hat{s}_{ev}(n)$, $\hat{h}_{ev}(n)$, corresponding to the spectral envelope $|H(e^{j\omega})|$, is separated from the rest of the cepstrum. An odd-symmetry part,

$$\begin{aligned} \hat{h}_{od}(n) &= \hat{h}_{ev}(n), \quad n > 0 \\ &= 0, \qquad n = 0 \\ &= -\hat{h}_{ev}(n), \quad n < 0, \end{aligned}$$

is added to $\hat{h}_{ev}(n)$ to produce an artificial "vocal tract impulse response,"

$$\underset{\sim}{h}(n) = A^{-1}[\hat{h}_{ev}(n) + \hat{\underset{\sim}{h}}_{od}(n)]$$

which is the minimum-phase counterpart of the h(n) shown in Figure 3.

2. The current pitch period of the input waveform is measured by locating the first pitch peak of the cepstrum. Recall from Section II that these peaks are produced by the approximate periodicity of $M(e^{j\omega})$, which is the transform of the impulse train m(n) in Figure 3.

3. Using the pitch data obtained in Step 2, the effective convolution of m(n) with $\underset{\sim}{h}(n)$ is carried out, producing a speech waveform

$$\underset{\sim}{s}(n) = m(n) * \underset{\sim}{h}(n).$$

Quality synthesis requires that new cepstra be computed approximately once every 10 msec, producing new $\underset{\sim}{h}(n)$'s and pitch measurements often enough to follow the natural changes of h(n) and m(n).

It should be pointed out that the Fourier transform and inverse Fourier transform operations involved in the above are computationally performed using the DFT and IDFT:

$$\text{DFT:}\quad X(k) = X(e^{j\omega})\Big|_{\omega = \frac{2\pi k}{N}} = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}},\ 0 \le k \le N-1$$

$$\text{IDFT:}\quad x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{j\frac{2\pi kn}{N}},\ 0 \le n \le N-1.$$

Both x(n) and X(k) are periodic with period N. "Negative time", as in the definition of $\hat{\underset{\sim}{h}}_{od}(n)$, corresponds computationally to $\frac{N}{2} \le n \le N-1$, and "positive time" corresponds to $0 \le n \le \frac{N}{2} - 1$. Thus, the "artificial complex cepstrum" is actually computed according to

$$\hat{h}_{ev}(n) + \hat{\underset{\sim}{h}}_{od}(n) = \hat{h}_{ev}(n)k(n),\ 0 \le n \le N-1$$

where

$$\begin{aligned} k(n) &= 2,\ 1 \le n \le \tfrac{N}{2} - 1 \\ &= 1,\ n = 0 \text{ and } n = \tfrac{N}{2} \\ &= 0,\ \tfrac{N}{2} + 1 \le n \le N-1\ . \end{aligned}$$

How much of $\hat{s}_{ev}(n)$ should be included in the "low-time part" is determined by the amount needed to maintain synthesis quality, the necessity to exclude any pitch peaks from the low-time part, and the desirability of maximizing the amount of $\hat{p}_{ev}(n)$ eliminated in the process of isolating the low-time part. A reasonable choice is found to be 3-4 msec of the cepstrum, as illustrated by the typical cepstrum of Figure 11. (For example, at a sampling rate of 10KHz, this corresponds to about 30-40 samples.) Thus, $\hat{h}_{ev}(n)$ is obtained from $\hat{s}_{ev}(n)$ by multiplying the latter by a "truncation window," $\ell(n)$:

$$\ell(n) = 1, \quad 0 \leq n \leq L, \; N-L \leq n \leq N-1,$$

$$= 0, \quad L < n < N-L \; ,$$

where L = 30-40. Actually, since "negative time" samples of $\hat{h}_{ev}(n)$ are not used, the actual $\ell(n)$ can be

$$\ell(n) = 1, \quad 0 \leq n \leq L$$

$$= 0, \quad L < n \leq N-1.$$

In practice, the convolution of $\underset{\sim}{h}(n)$ with m(n) is performed by simply delaying successively-computed $\underset{\sim}{h}(n)$'s by amounts equal to the speech period and adding them to form a running output waveform. This process is illustrated in Figure 12. The quality of the resynthesized waveform can be improved by smoothing the transition from one $\underset{\sim}{h}(n)$ to the next. This can be accomplished by interpolation. The interpolation method involves a weighted sum of two $\underset{\sim}{h}(n)$'s, the weighting depending upon the time a given synthesized period* begins relative to the interval between production of the two $\underset{\sim}{h}(n)$'s. Let M samples be this interval and let $\underset{\sim}{h}_o(n)$ and $\underset{\sim}{h}_1(n)$ be produced at the beginning and end of the interval

---

*I.e., a new period begins at each impulse of m(n).

Figure 11

A typical voiced-speech cepstrum, $\hat{s}_{ev}(n)$, illustrating its similarity to the corresponding complex cepstrum. The low-time part is defined such that it excludes the pitch peaks and to include the smallest fraction of the cepstrum from which quality $\underset{\sim}{h}(n)$'s can be recovered.

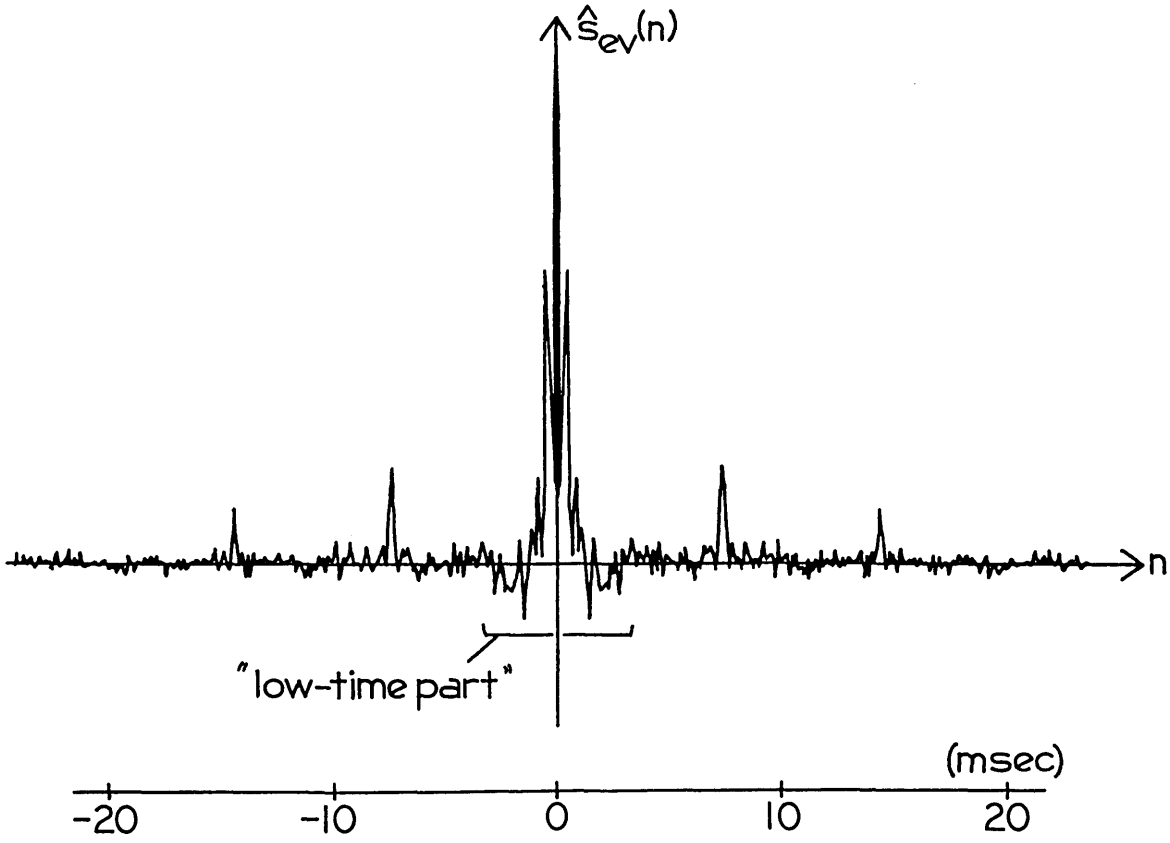
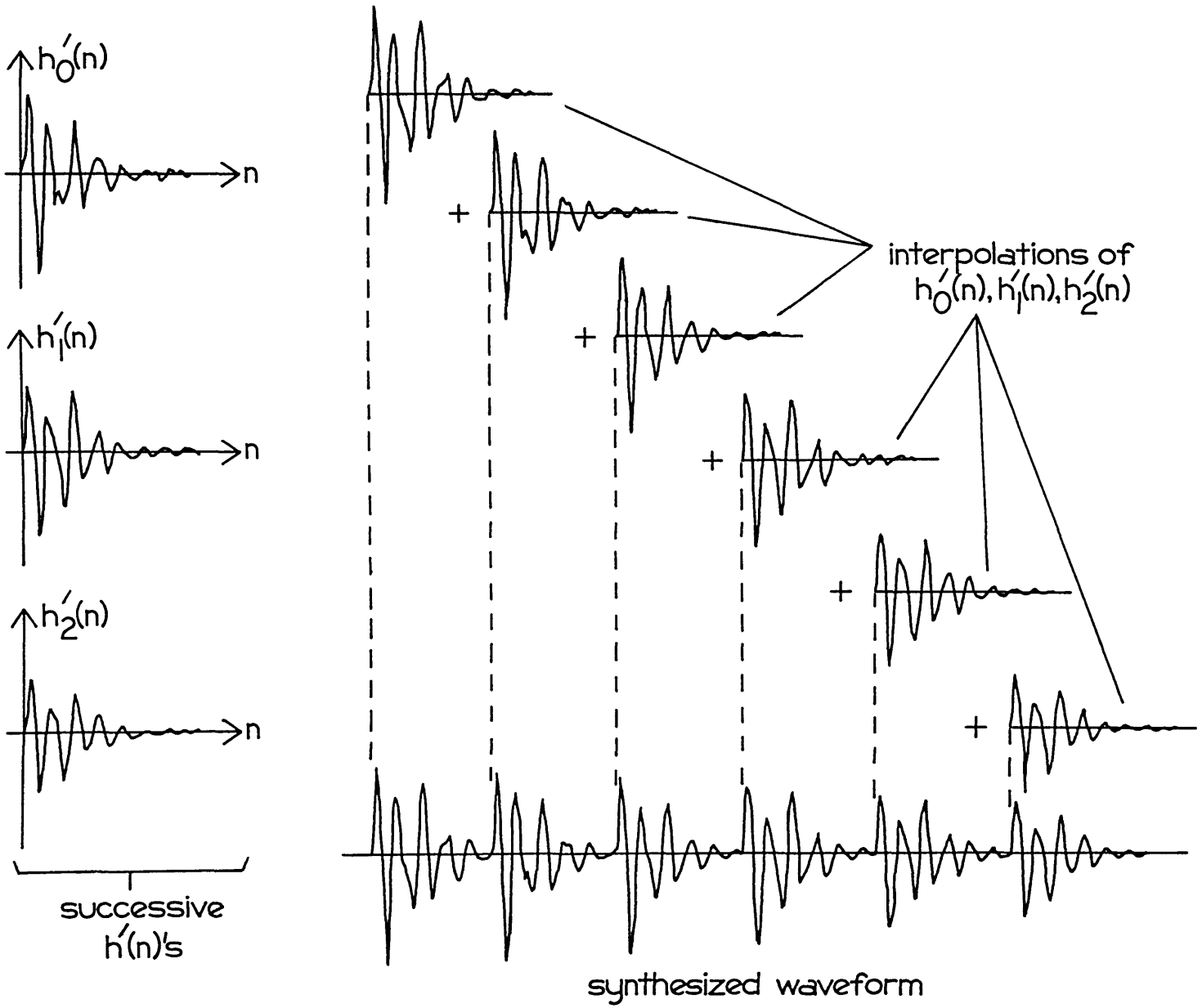

$\hat{s}_{ev}(n)$
n
"low-time part"
(msec)
-20
-10
0
10
20

Figure 12

Running speech waveform construction from $\underset{\sim}{h}(n)$'s produced each 10 msec. Delays of $\underset{\sim}{h}(n)$'s, which determine period of waveform, are based upon pitch period values found by detection of cepstral pitch peaks. Each period is a weighted average of $\underset{\sim}{h}(n)$'s, with the weights adjusted according to the time a given period begins relative to the times that the $\underset{\sim}{h}(n)$'s are computed from their cepstra. These relative times are measured according to the vertical dimension in the drawing. (See also Figure 13.)

Note: Symbol 'h'(n)' in drawing corresponds to $\underset{\sim}{h}(n)$ in text.


$h'_0(n)$
n
$h'_1(n)$
n
$h'_2(n)$
n
successive
$h'(n)$'s
interpolations of
$h'_0(n), h'_1(n), h'_2(n)$
synthesized waveform

respectively. Then an $\underset{\sim}{h}(n)$ to be positioned in the running speech output Q samples into the interval is computed as

$$\underset{\sim}{h}(n) = (1-\frac{Q}{M})\underset{\sim}{h}_o(n) + \frac{Q}{M}\underset{\sim}{h}_1(n),$$

as illustrated in Figure 13.

The length and shape of the input sectioning window, w(n), also influence the quality of the synthesis. The length of the window should be short enough that h(n) and the pitch do not vary significantly over the window duration. Then

$$w(n)s(n) \simeq w(n)[h(n) * m(n)].$$

A further requirement is that

$$w(n)[h(n) * m(n)] \simeq [w(n)m(n)] * h(n)$$

so that the low-time part of the cepstrum of $\underline{s(n)w(n)}$ is approximately $\hat{h}_{ev}(n)$. In terms of transforms, this necessitates that

$$|W(e^{j\omega}) * [H(e^{j\omega})M(e^{j\omega})]| \simeq |W(e^{j\omega}) * M(e^{j\omega})||H(e^{j\omega})|.$$

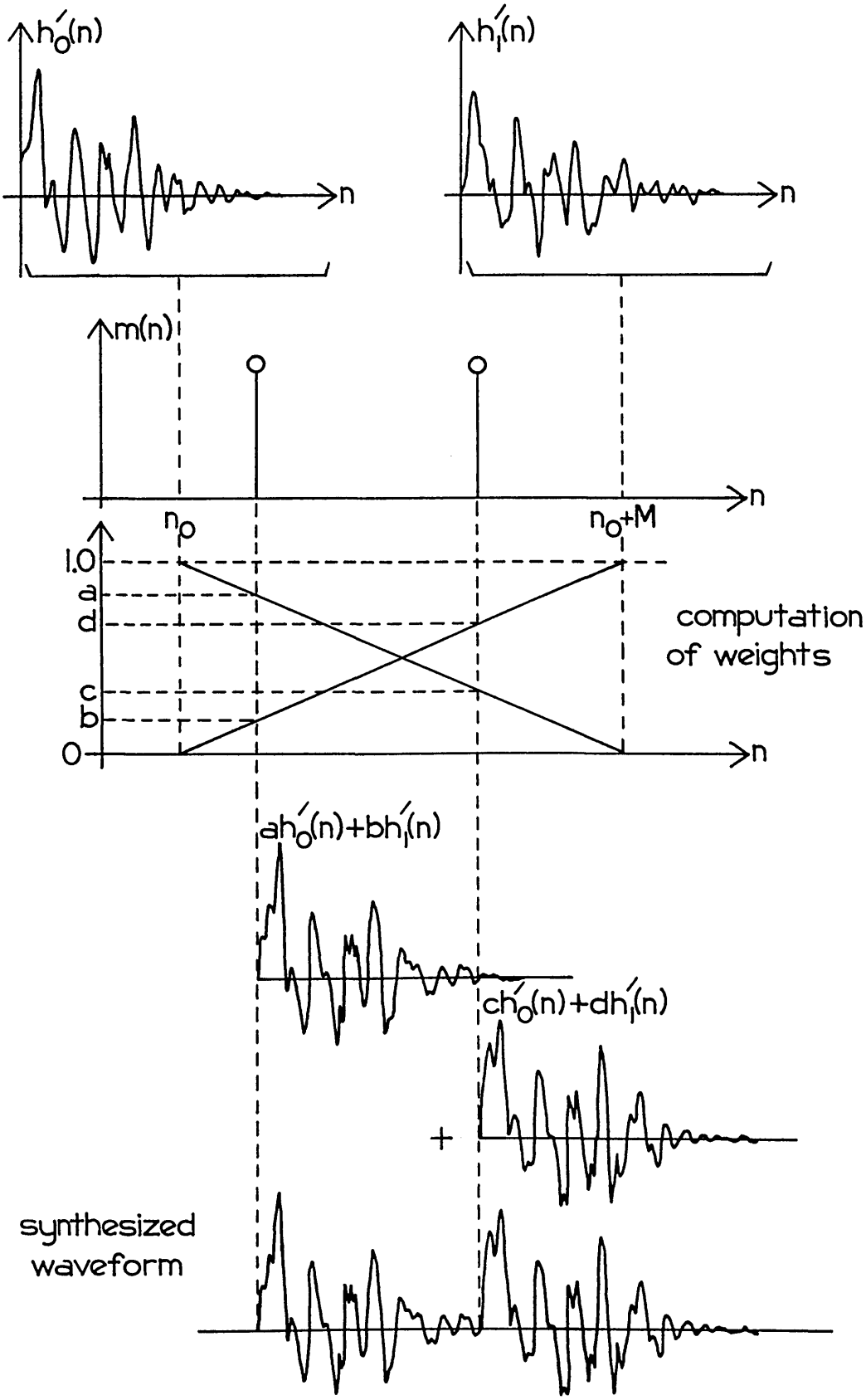
This approximation is good if $|W(e^{j\omega})|$ is sufficiently narrowband relative to variations in $|H(e^{j\omega})|$. Since the bandwidth of $|W(e^{j\omega})|$ varies as the inverse of the duration of w(n), a lower limit is placed on section length, in conflict with the first requirement above. Thus only a restricted range of window durations is acceptable.

Another factor is that the window should enhance the amplitude of the first cepstral pitch peak so that it is easily identifiable as a local maximum in the cepstrum. This is accomplished by choosing w(n) such that $W(e^{j\omega})$ has low sidelobes, making $|W(e^{j\omega}) * M(e^{j\omega})|$ more nearly "sinusoidal" than if the sidelobes are large. Figures 14 and 15 illustrate the effect of sidelobes with a comparison between a 512-point square window (Figure 14) and a 512-point Hamming window, given by

Figure 13

Detailed procedure for weighted averaging of $\underset{\sim}{h}(n)$'s, which is done to smooth the transitions between successively-computed $\underset{\sim}{h}(n)$'s. This partially offsets roughness of quality due to computation of a new $\underset{\sim}{h}(n)$ only once every 10 msec. In drawing, $h'_o(n)$ is viewed as being computed "at" $n=n_o$, and $h'_1(n)$ "at" $n=n_o+M$.

Note: $h'(n)$ of drawing corresponds to $\underset{\sim}{h}(n)$ of text.


$h'_0(n)$
$h'_1(n)$
n
m(n)
n_0
n_0+M
1.0
a
d
c
b
0
computation of weights
$ah'_0(n)+bh'_1(n)$
$ch'_0(n)+dh'_1(n)$
+
synthesized waveform

Figures 14 and 15

Comparison of square and Hamming window performance.

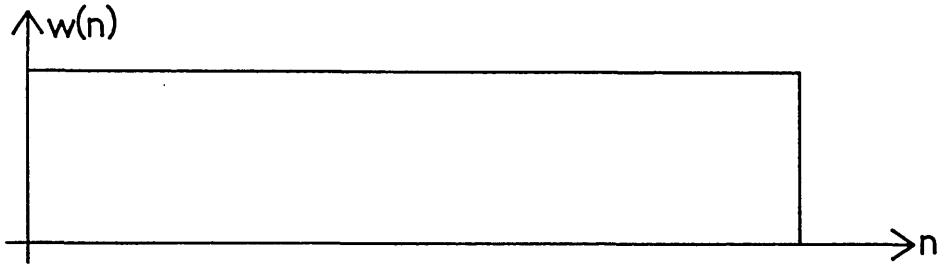
(14a) 512-point square window

(14b) Magnitude spectrum of this window: note high sidelobes

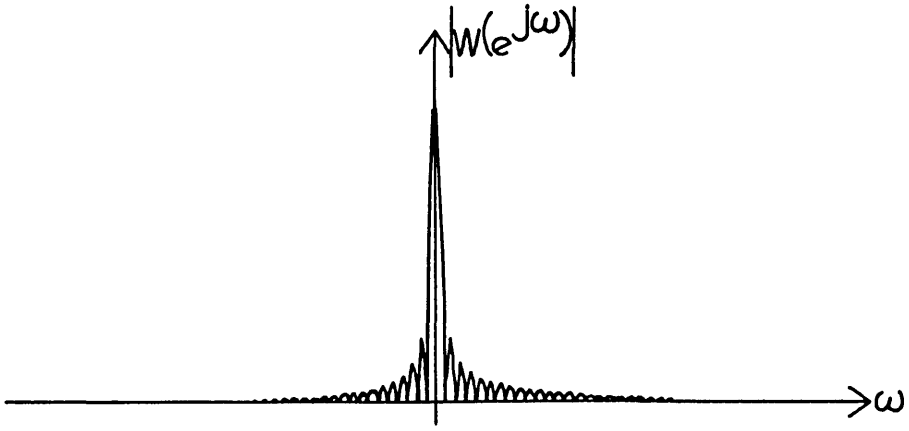
(14c) Magnitude of speech waveform section weighted by square window. Low amplitude, narrow spikes are due to spectral sidelobes of window.

(14d) Cepstrum computed from above spectrum; pitch peak is not clearly resolved.

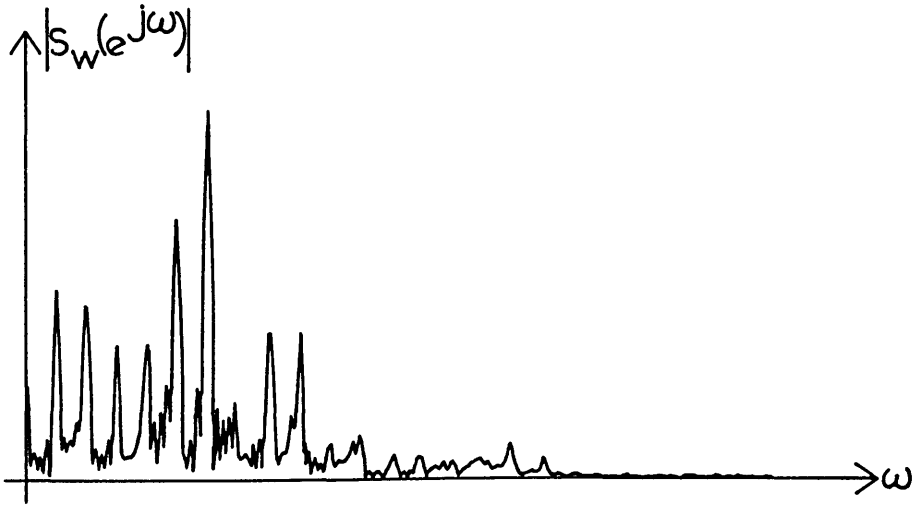
(15a, b, c, d) Corresponding functions for a 512-point Hamming window. Note clear resolution of cepstral pitch peak in Figure 15d.



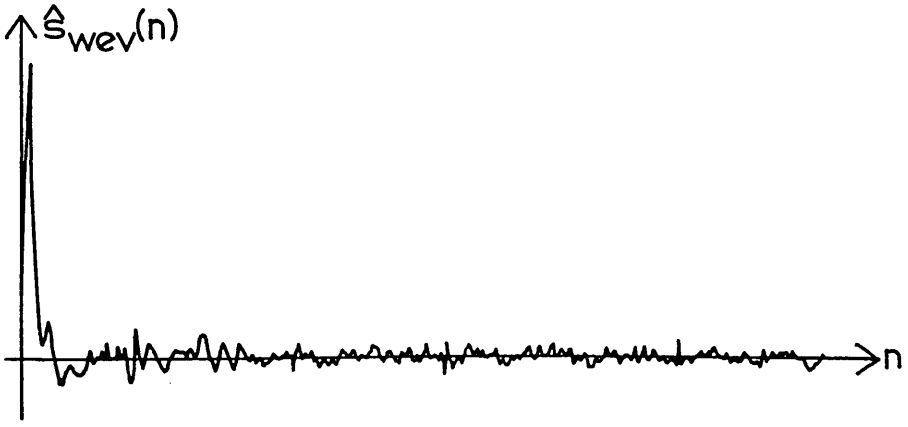
(a)



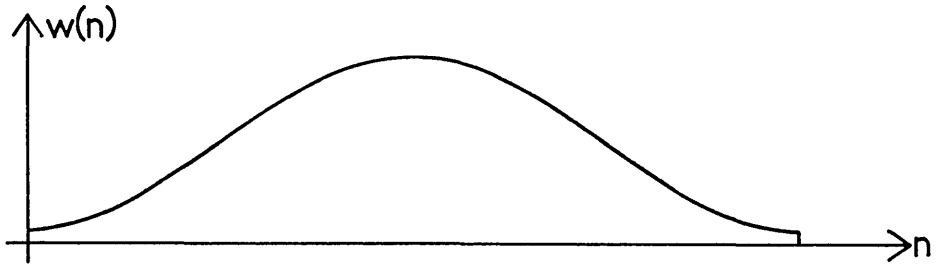
(b)



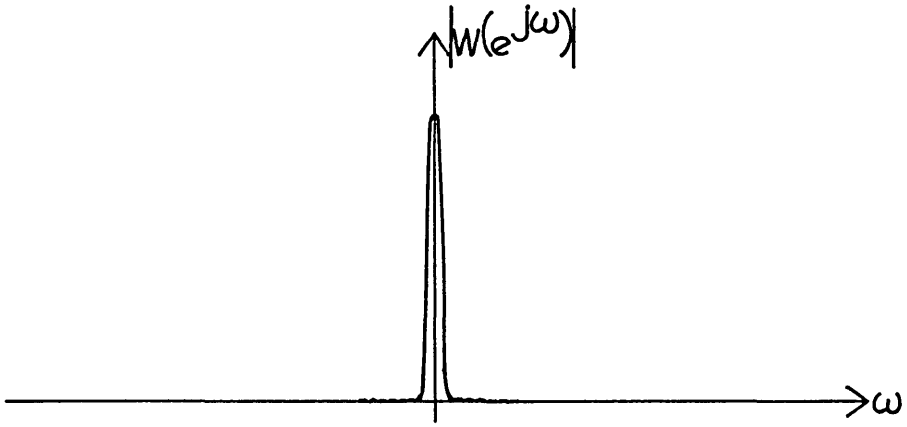
(c)



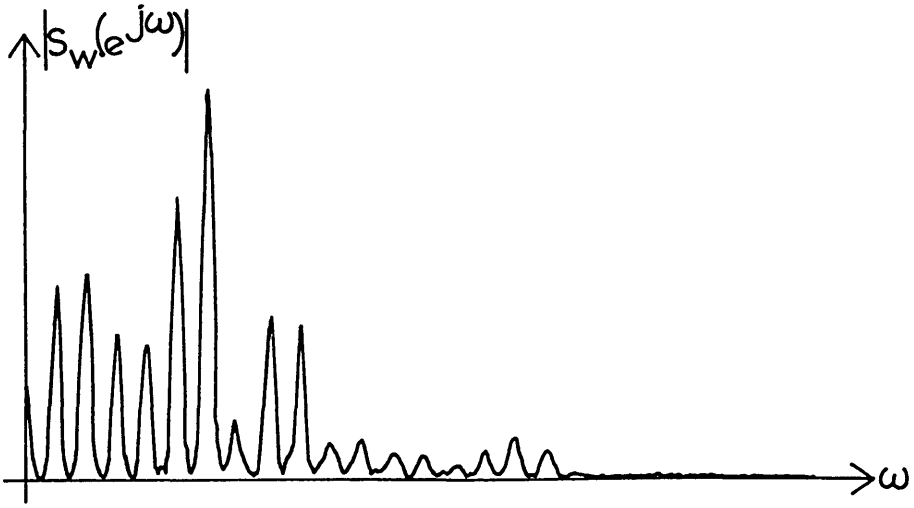
(d)



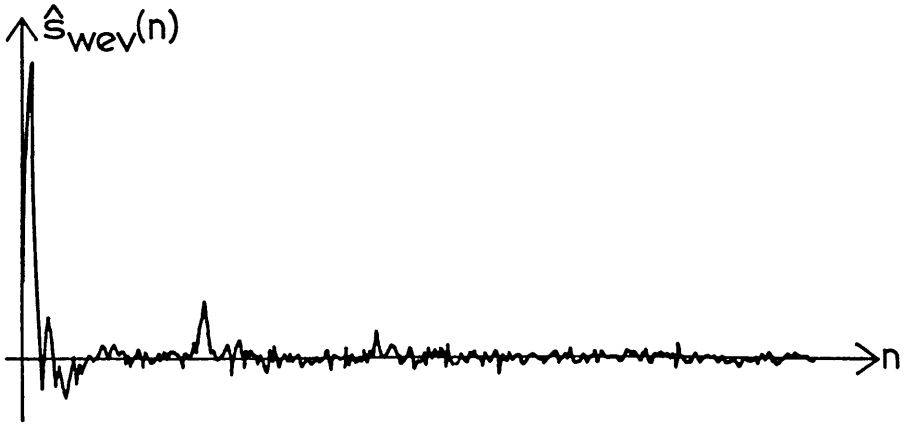
(a)



(b)



(c)



(d)

$$w(n) = 0.54 - 0.46 \cos \frac{2\pi n}{N}, \quad 0 \leq n \leq N-1.$$

As the drawings show, the performance of the Hamming window is far superior. In addition, the low sidelobes improve the approximation

$$w(n)s(n) \simeq [w(n)m(n)] * h(n)$$

as mentioned in the preceding paragraph.

It has been observed that, for a Hamming window, pitch peak amplitude varies substantially with the position of the window relative to the analyzed speech waveform. This results from the fact that $w(n)\cdot s(n)$ looks more or less like a portion of a periodic sequence depending upon the position of the pitch periods within the window. As the number of periods captured by the window increases, the pitch peak variation decreases. Figure 16 shows the results of an experiment in which a hypothetical m(n) was weighted by a shiftable 512-point Hamming window w(n):

$$m(n) = \sum_{r=-\infty}^{+\infty} u_o(n-rM)$$

$$m_w(n) = m(n)w(n-\eta).$$

The window shift is allowed to vary over the range $0 \leq \eta \leq M-1$, or one period of m(n). The first cepstral pitch peak amplitude as a function of η, determined from

$$\text{Peak Amplitude} = \text{Max}[\hat{m}_{wev}(n)], \quad M-5 \leq \eta \leq M+5,$$

is plotted for M = 150 and M = 170. Note that there is a substantial increase in variation as the number of periods captured by the window changes from 3.4 (M = 150) to 3.0 (M = 170). A similar plot, Figure 16c, was made using a portion of a speech waveform in place of m(n), with approximately 4.2 periods under the window. Figures 16d, e, and f show the optimum positioning of the window as determined from the plot maxima

Figure 16

Computer-generated plots showing effect position of Hamming window (relative to speech waveform) can have on pitch peak amplitude. Plots a, b, and c show peak amplitude as a function of the window-shift index, η, for two impulse trains and a voiced speech waveform; Figures 16d, e, and f show the weighted waveform when the window is in the optimal position. Window length is 512 samples.

(a) Cepstral pitch peak amplitude for weighted impulse train, impulse spacing of 150 samples.

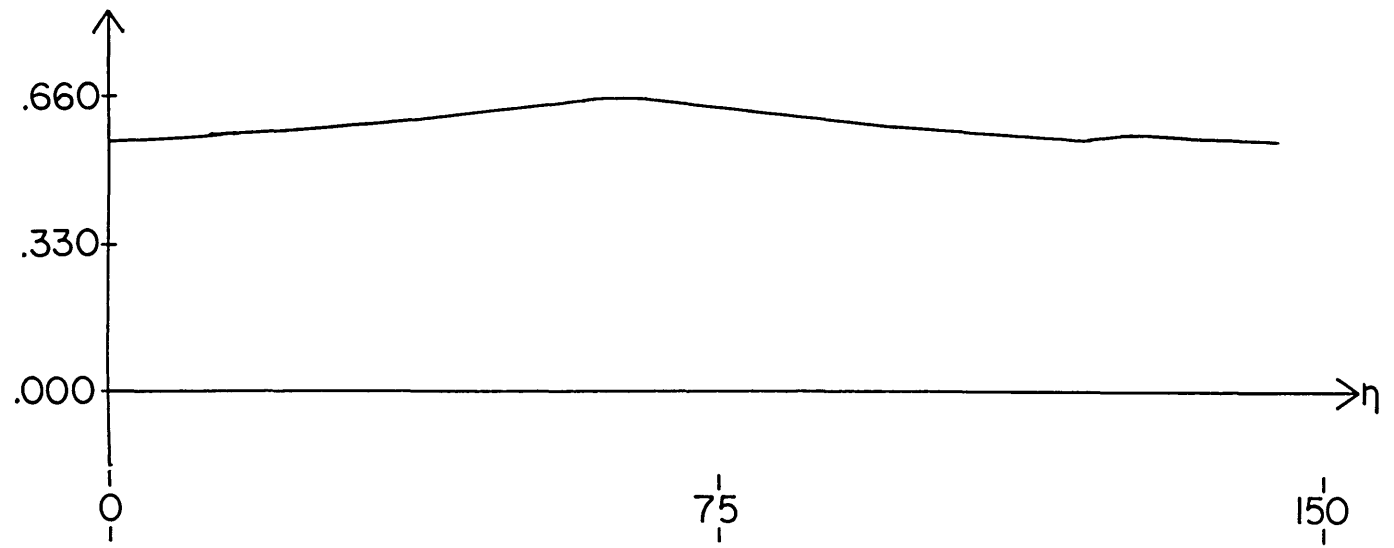
(b) Same, except impulse spacing of 170 samples.

(c) Pitch peak amplitude for Hamming-weighted speech waveform, with 4.2 periods under window.

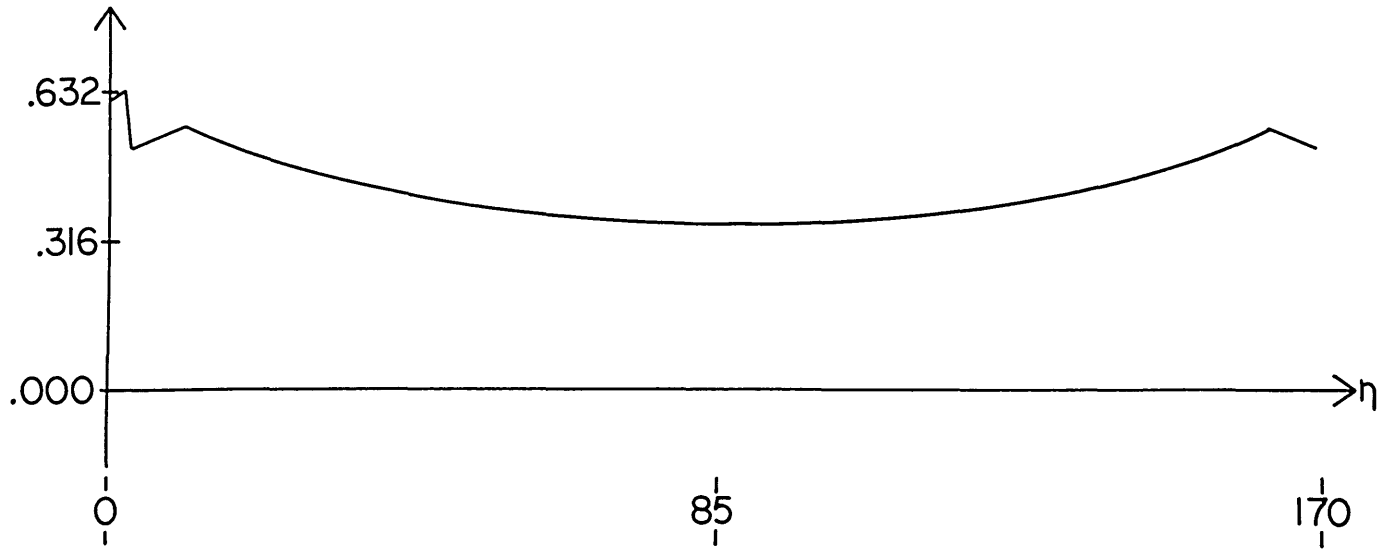
(d) Impulse train of part (a) weighted by optimally-positioned window.

(e) Same, for impulse train of part (b).

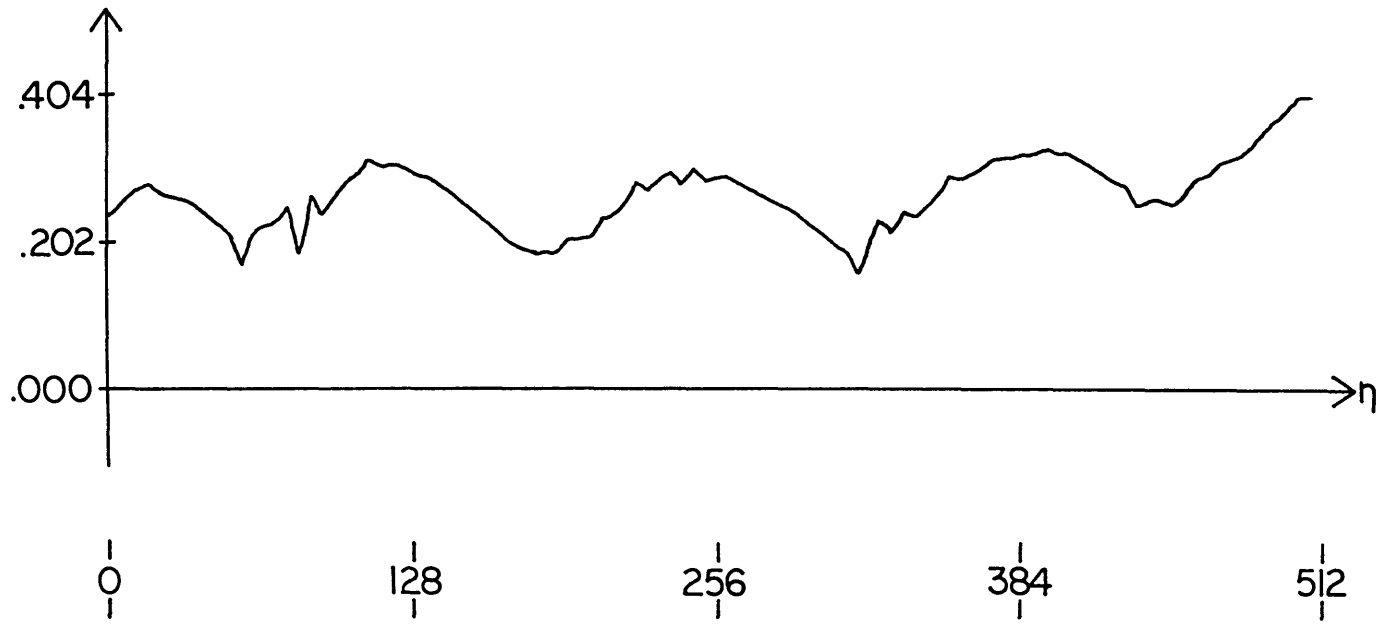
(f) Same, for speech waveform of part (c).



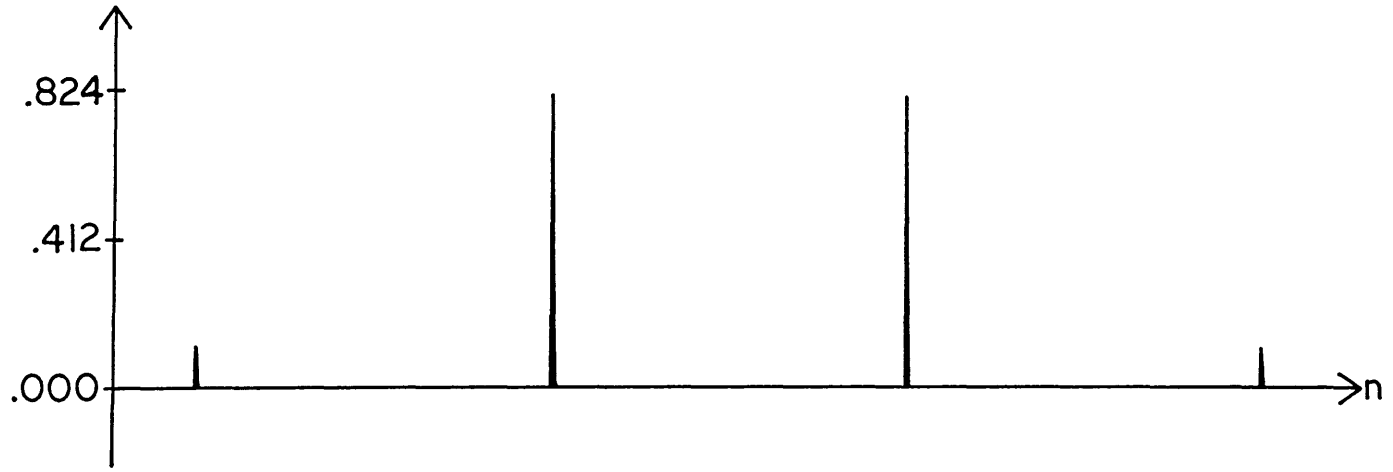
(a)



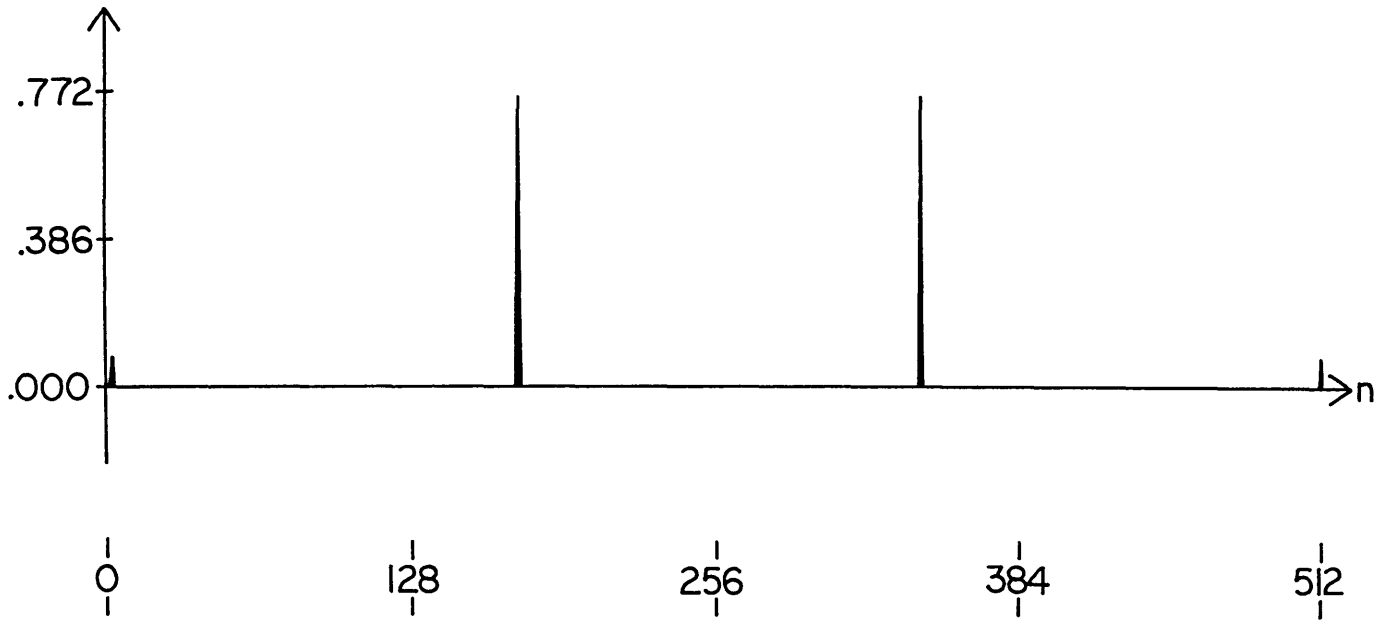
(b)



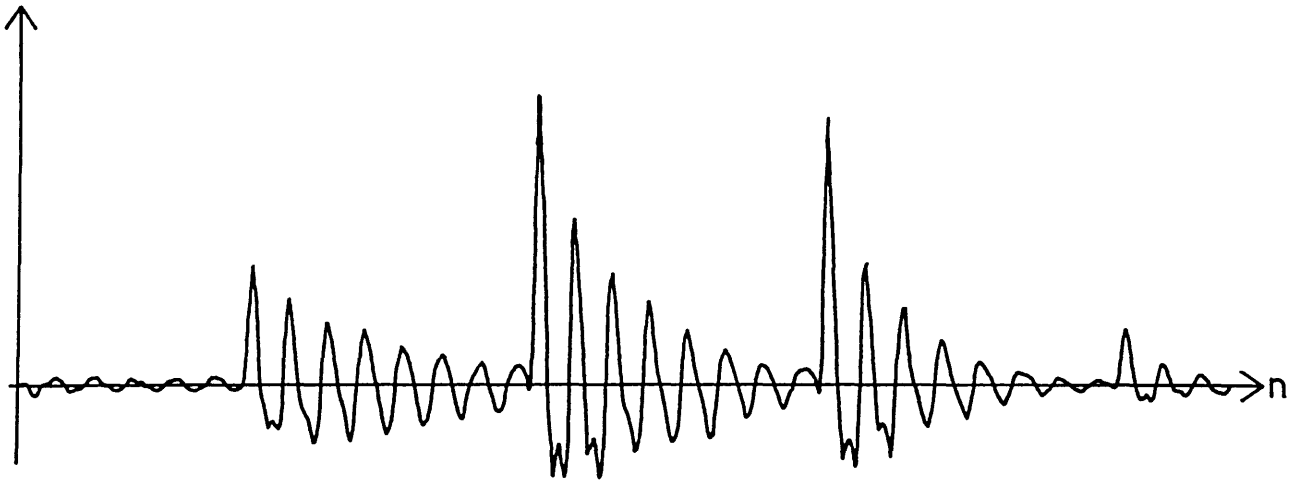
(c)



(d)



(e)



(f)

of Figures 16a, b, and c, respectively. Not too surprisingly, maximum peak height occurs when the maximum amount of "periodic energy" is under the window. A method for automatic positioning of the window to take advantage of the maximum peak height for a given window length might substantially improve the performance of cepstral pitch detection algorithms.

For typical speech waveforms, the requirement that h(n) and the pitch period do not vary much within the window allows a window length not much longer than about 50 msec. This may conflict with the need to have enough pitch periods under the window to produce a consistently strong pitch peak, if the period is long. This problem is encountered with low-pitched male speech. For example, if the window length is 40 ms and the pitch period is 15 ms, then there are only 2.67 periods under the window, so that for some window positions the cepstral pitch peak may be very small. Due to pitch-detection errors thereby caused, the synthesized speech may sound very rough in low-pitched areas. A possible solution to this problem is the automatic adjustment of window length to some minimum number of pitch periods, according to the most recent pitch measurement, whenever the normal window length is less than this minimum number of periods. The improvement in pitch-detection accuracy would probably be worth the sacrifice of short window length.

As will be discussed in Section IV, window length and shape also have importance with respect to the effectiveness of cepstral filtering for eliminating reverberation. It is shown there also that a Hamming window is a good choice for these requirements.

D. Speech Quality with Pitch-Synchronous Synthesis

Rough or strident quality has been the main difficulty with

pitch-synchronous synthesis as implemented. This distortion is judged by some to be less tolerable than the reverberation being removed! Probably the fundamental limit upon the quality of speech synthesized in this way is determined by the extent to which the model of Figure 3 accurately represents natural voiced speech production.* If the model is assumed valid, several secondary factors still detract from the quality, some of which were discussed above. Two others have been seen as contributing to roughness: the minimum phase used in the synthesis of $\underset{\sim}{h}(n)$, and the cepstral truncation involved in isolating $\hat{h}_{ev}(n)$ from the rest of the cepstrum. Both of these affect the temporal distribution of energy in $\underset{\sim}{h}(n)$.

In particular, minimum-phase $\underset{\sim}{h}(n)$'s tend to have a high peak factor and often decay more rapidly than natural-phase h(n)'s, even when they have the same magnitude spectra. This does not seem to affect the ready identification of voiced sounds by a listener, but may strongly influence quality. The duration of $\underset{\sim}{h}(n)$ may be increased and the peak factor decreased by adding to the phase shift in $\underset{\sim}{H}(e^{j\omega})$. An addition to the phase other than a linear term results in a time-shift of certain sinusoidal components of $\underset{\sim}{h}(n)$ relative to others, which can produce a spreading of the energy distribution in $\underset{\sim}{h}(n)$. Zero-phase $\underset{\sim}{h}(n)$'s (corresponding to "zero delay" of all sinusoidal components) generally have the highest peak factor and shortest duration of all. This is because a sum of undelayed cosines has maximum constructive interference at n=0 and destructive interference increases very rapidly as $|n|$ increases. Minimum-

---

*It should be pointed out that this model can be extended to allow synthesis of voiced sounds.[7] However, only voiced speech is considered in this paper.

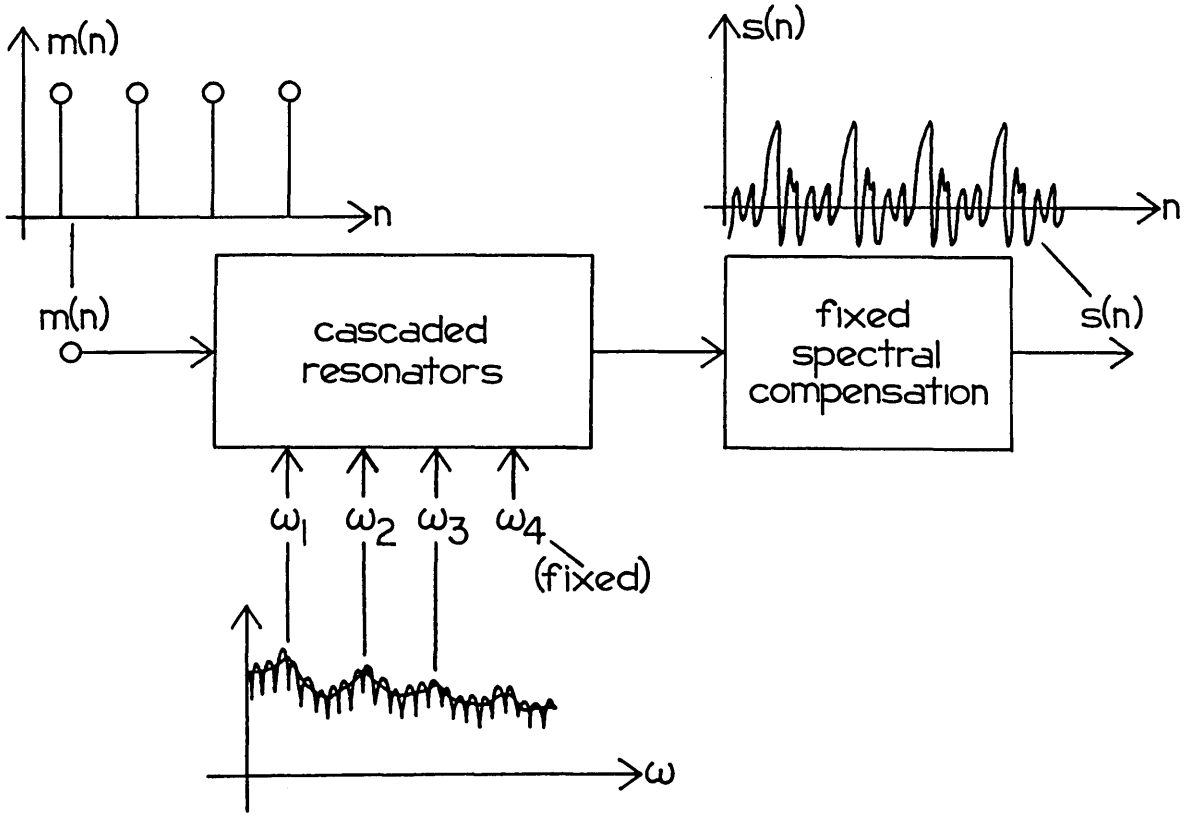
phase $\underset{\sim}{h}(n)$'s have the property that of all causal $\underset{\sim}{h}(n)$'s having the same magnitude spectrum, the minimum-phase $\underset{\sim}{h}(n)$ has the smallest phase shift. Since this is as close to zero phase as a causal $\underset{\sim}{h}(n)$ can get, minimum phase also produces relatively high peak factor and short duration. The concurrent appearance of rough or "buzzy" speech quality with these characteristics suggests the possible importance of the choice of artificial phase to accomplishing good quality.

A few experiments were carried out in which the phase was increased by making $\hat{\underset{\sim}{h}}_{od}(n)$ larger than for minimum phase. A marked change in the temporal energy distribution of the $\underset{\sim}{h}(n)$'s was produced, but other complications resulted in a net degradation of speech quality. One interesting consequence was that as the duration of $\underset{\sim}{h}(n)$ was increased beyond a certain point, initially unreverberated input speech was made to sound reverberated in the synthesis!

While minimum-phase, pitch-synchronous synthesis has often resulted in undesirably rough quality as implemented above, Rabiner and Schafer have not encountered this degree of roughness with minimum-phase formant synthesis.[9] Their technique also assumes the voiced speech production model of Figure 3, but the linear system representing the vocal tract, glottal pulse, and radiation characteristics is realized as a ten-pole digital network. The multiple resonances of this network are periodically adjusted to correspond to the naturally-occurring resonant peaks in $|H(e^{j\omega})|$ (Figure 17), which is computed from approximately the first four milliseconds of the cepstrum $\hat{s}_{ev}(n)$. The network frequency response, therefore, is actually an approximation to $|H(e^{j\omega})|$. Significantly, this ten-pole model is devised in such a way that although the network impulse response is minimum-phase, speech synthesized by impulse-

Figure 17

Illustration of essentials of formant synthesis (voiced speech) by method of Rabiner and Schafer. Resonant frequencies of three cascaded digital resonators are adjusted at periodic intervals to correspond to natural formant peaks detected from envelope of log magnitude spectrum of input speech waveform; envelope is computed from low-time part of cepstrum.



log magnitude spectrum and envelope obtained from speech waveform analysis

train excitation of the network sounds more natural than if the $\underset{\sim}{h}(n)$'s are derived directly from the cepstrum. On the other hand, the amount of computation required for formant synthesis by present techniques is greater than for synthesis with direct-computed $\underset{\sim}{h}(n)$'s.

The above ten-pole model provides a convenient idealization through which to discuss the effects of cepstral truncation. Specifically, the formant network has the system function (expressed as a z-transform)

$$H_f(z) = \frac{A}{(1-az^{-1})(1+bz^{-1})} \prod_{k=1}^{4} \left(\frac{1}{(1-c_k z^{-1})(1-c_k^* z^{-1})} \right) ,$$

where A, a, and b are real constants and

$$c_k = e^{-\alpha_k + j\omega_k} ,$$

for which α_k and ω_k are the k^{th} formant bandwidth and frequency. Each pole corresponds in the time domain to an exponentially-decaying sequence; the network impulse response $h_f(n)$ is the convolution of these sequences. The decay rates are controlled by a, b, and the α_k's. In particular, as the α_k's, or formant bandwidths, are increased, the rate of decay increases. Now, an important consequence of the ten-pole model is that the complex cepstrum $\hat{h}_f(n)$ is necessarily infinite in duration. As computed from[3]

$$\hat{h}_f(n) = \frac{1}{2\pi j} \oint_C \log H_f(z) z^{n-1} dz ,$$

where C is the unit circle, $\hat{h}_f(n)$ is given by

$$\hat{h}_f(n) = \frac{1}{n} \left(a^n + (-b)^n + \sum_{k=1}^{4} e^{-\alpha_k n} \cos \omega_k n \right) , \quad n > 0 ,$$

$= \log A, \quad n=0$

$= 0, \quad n < 0.$

The infinite duration of $\hat{h}_f(n)$ contrasts with the truncated complex cepstrum from which the $\underset{\sim}{h}(n)$'s are computed.

Consider the effect of truncating $\hat{h}_f(n)$:

$$\hat{h}_{ft}(n) = \ell(n)\hat{h}_f(n),$$

where

$$\ell(n) = 1 , \quad |n| \leq L$$

$$= 0 , \quad |n| \geq L .$$

Since multiplication in the time domain corresponds to convolution in the frequency domain, the effect of this truncation will be to increase the formant bandwidths. Therefore, $h_f(n)$ will tend to decay more rapidly. The lower is L, the higher will be the rate of decay. We would expect a similar effect with the truncation of $\hat{s}_{ev}(n)$ to produce $\hat{h}_{ev}(n)$. Actually, what is really produced is only a truncated approximation to the "true" $\hat{h}_{ev}(n)$; although this leaves the positions (in frequency) of the formant peaks essentially unchanged, the width of the peaks is increased, with the result that $\underset{\sim}{h}(n)$ decays more rapidly than it should. Formant synthesis escapes this effect because the formant bandwidths are preset at values which produce natural-sounding speech, and thus are not influenced by cepstral truncation.

It must be emphasized, however, that to maximize the elimination of reverberation, truncation of $\hat{h}_{ev}(n) + \hat{\underset{\sim}{h}}_{od}(n)$ at the lowest possible point is necessary unless reverberant "clutter" can be reduced to an insignificant level without distortion of $\hat{h}_{ev}(n) + \hat{\underset{\sim}{h}}_{od}(n)$. An approach to accomplishing this is suggested in Section V.

A final possibility is that the effects of truncation might be reduced somewhat by the smoothing of the truncation window at its ends. An end-tapered truncation window is illustrated in Figure 18. This tapering tends to reduce the sidelobes of the spectrum of the truncating window, although it also increases the width of the central lobe. Such a window was used in the experiments to be described in Section IV.

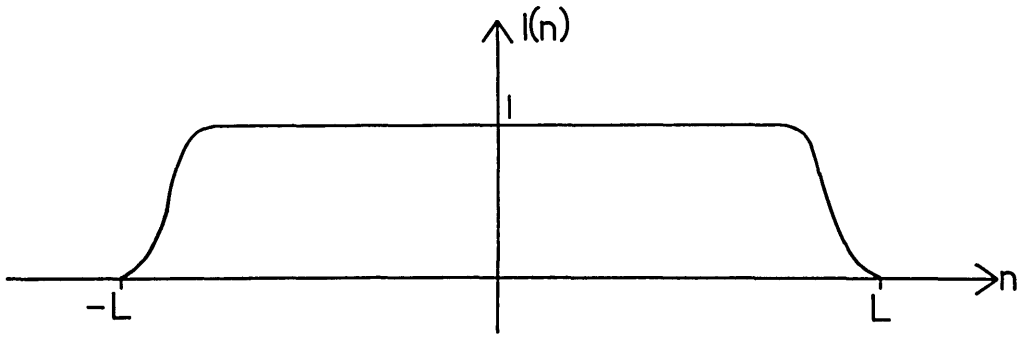
E. Summary

The cepstrum, or even part of the complex cepstrum, is seen to have advantages over the full complex cepstrum with respect to computational feasibility, the problem of differential delays between the multiple inputs required for "comparison" of complex cepstra, and the ease of removing the reverberant component of the cepstrum. On the other hand, synthesis of speech from the cepstrum is complicated by its lack of speech phase information. A feasible solution to this problem is use of the pitch-synchronous synthesis technique of Oppenheim's homomorphic vocoder.[7] Speech synthesized by this technique has been slightly rough, although quite intelligible. Some important factors influencing synthesis quality seem to be speech model validity, the effects of sectioning the input waveform, the use of minimum phase in resynthesis, and truncation of the cepstrum.

In the next two sections, methods of filtering the cepstra of reverberated speech waveforms to remove the reverberant component are discussed. The pitch-synchronous synthesis technique is found to be realistically compatible with these filtering methods.

Figure 18

Truncation window, $\ell(n)$, with tapered ends to reduce broadening of formant peaks caused by cepstral truncation.


I(n)
1
n
-L
L

IV. Filtering of Cepstra

It will be useful before beginning this discussion to define certain sequences clearly. The weighted section of reverberated speech waveform from which our cepstra are computed can be expressed as

$$w(n)[s(n) * p(n)],$$

where w(n) is the weighting window. The cepstrum of this section will be denoted by $\hat{x}_{ev}(n)$, and often referred to as the "reverberated cepstrum". The reverberated cepstrum will be viewed as the sum of two components, the first to be "recovered" and the second to be "removed":

(1) $\hat{s}_{wev}(n)$, the cepstrum of s(n)w(n);

(2) $\hat{p}_{ev}(n)$, defined by

$$\hat{p}_{ev}(n) = \hat{x}_{ev}(n) - \hat{s}_{wev}(n) \ .$$

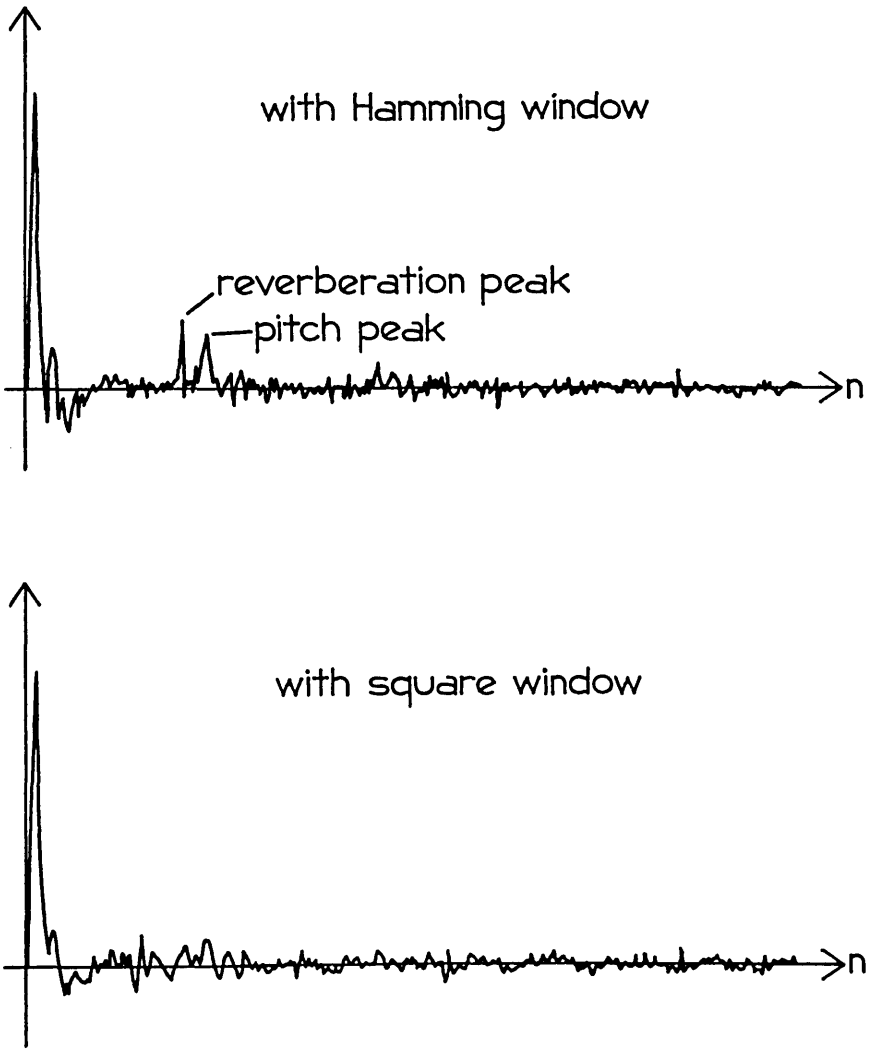
Both $\hat{s}_{wev}(n)$ and $\hat{p}_{ev}(n)$ depend upon w(n); careful choice of this window is important because it affects the approximation of $\hat{s}_{wev}(n)$ to $\hat{s}_{ev}(n)$ and the amount of additive "clutter" due to $\hat{p}_{ev}(n)$. It is to be noted, in this regard, that $\hat{p}_{ev}(n)$ is <u>not</u> equal to $\hat{p}_{ev}(n)$.

The pitch-synchronous synthesis method discussed in Section III requires the recovery of $\hat{h}_{ev}(n)$, the cepstrum of the short-time vocal tract impulse response, and accurate determination of pitch period from the location of the first cepstral pitch peak. Therefore, two important requirements upon w(n) are that it allows the low-time part of $\hat{s}_{wev}(n)$ to approximate $\hat{h}_{ev}(n)$ closely and that it enhances pitch peak amplitude. We saw in Section III that the Hamming window satisfies these requirements.

A significant result is that the Hamming window can also reduce the amount of "clutter" from $\hat{p}_{ev}(n)$ and sharpen the resolution of "reverberation peaks", as compared to results obtained without Hamming weighting of the input. An example is shown in Figure 19. This is important

Figure 19

Comparison of reverberated cepstra, $\hat{x}_{ev}(n)$, produced with Hamming weighting and square weighting of the input speech section, respectively. Notice, for the Hamming window, that not only is the pitch peak amplitude enhanced as compared to the square window cepstrum, but also the reverberation peak.


with Hamming window
reverberation peak
pitch peak
n
with square window
n

because the effectiveness of cepstral filtering is reduced if the reverberant clutter is widely spread in the cepstrum rather than being concentrated in narrow peaks. Improved peak resolution makes identification and specific removal of reverberation peaks simpler.

Of course, the Hamming window does not make $\hat{p}_{wev}(n)$ any less "cluttery" than $\hat{p}_{ev}(n)$ itself. What the window actually accomplishes is to improve the approximation

$$w(n)[s(n) * p(n)] \simeq [w(n)s(n)] * p(n) = s_w(n) * p(n)$$

or

$$w(n) \sum_{i=1}^{M} a_i s(n-n_i) \simeq \sum_{i=1}^{M} a_i w(n-n_i)s(n-n_i),$$

where

$$p(n) = \sum_{i=1}^{M} a_i u_o(n-n_i) \; ; \; n_M > n_i \; , \; i \neq M \; ,$$

so that

$$\hat{x}_{ev}(n) \simeq \hat{s}_{wev}(n) + \hat{p}_{ev}(n).$$

Since this requires that

$$w(n) \simeq w(n-n_M), \text{ all } n,$$

it is seen that the tapering of window ends is an important factor in achieving the approximation. Note, however, that the quality of approximation decreases as the duration, n_M, of p(n) increases, regardless of window shape. Hence, the ability of cepstral filtering to recover $\hat{s}_{wev}(n)$ generally decreases as reverberation times increase.

Comparison of the plots of $\hat{p}_{ev}(n)$ for square versus Hamming windows in Figure 20 illustrates the improvement that can be afforded by the Hamming window. For this example,

$$p(n) = u_o(n) + 0.5u_o(n-64).$$

The sampling rate was 10KHz and the window length was 512 samples.

Figure 20

Comparison of $\hat{\underset{\sim}{p}}_{ev}(n)$ for Hamming and square windowing of the input. These cepstra were generated by computing separately $\hat{x}_{ev}(n)$ and $\hat{s}_{wev}(n)$, and then subtracting the latter from the former.

(a) ("Ideal") cepstrum of p(n), where

$$p(n) = u_o(n) + 0.5u_o(n-64).$$

(b) $\hat{\underset{\sim}{p}}_{ev}(n)$ for square window.

(c) $\hat{\underset{\sim}{p}}_{ev}(n)$ for Hamming window.

All cepstra were computed from the same section of speech.

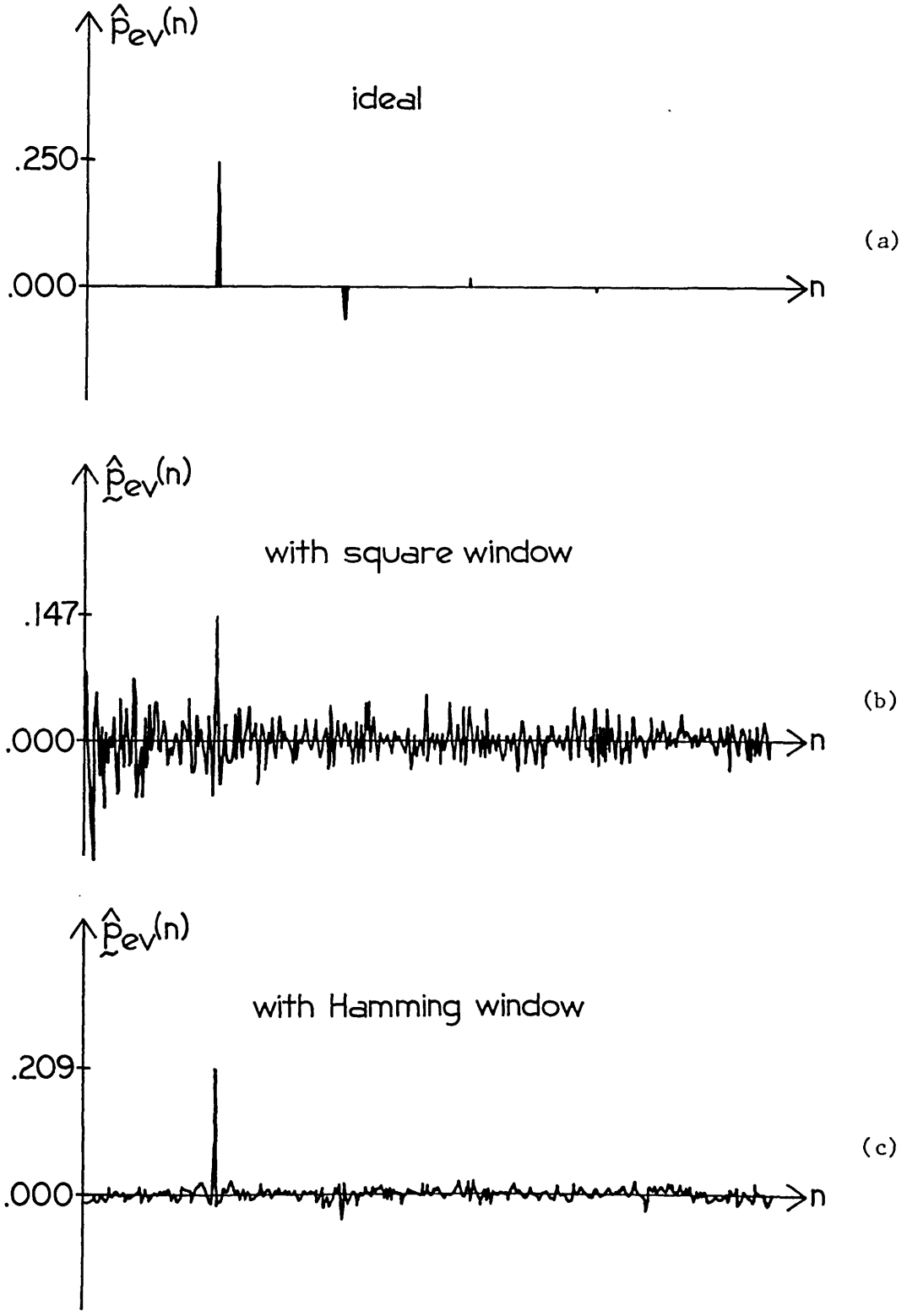

$\hat{P}_{ev}(n)$
ideal
.250
.000
n
(a)
$\hat{P}_{ev}(n)$
with square window
.147
.000
n
(b)
$\hat{P}_{ev}(n)$
with Hamming window
.209
.000
n
(c)

Figure 20a shows the ideal $\hat{p}_{ev}(n)$, given by

$$\hat{p}_{ev}(n) = \frac{1}{2} \sum_{k=1}^{\infty} \frac{(0.5)^k}{k} u_o(n-64k) \ ;$$

Figures 20b, c show $\hat{\underset{\sim}{p}}_{ev}(n)$ for square and Hamming windows, respectively.* The improvement is not always so dramatic, but is usually quite substantial.

For reference, Figure 21 shows the $\hat{\underset{\sim}{p}}_{ev}(n)$ of Figure 20c added to the cepstrum $\hat{s}_{wev}(n)$. Below this, $\hat{\underset{\sim}{p}}_{ev}(n)$ is shown alone on a matching scale. In this particular instance, the reverberation peak does not exceed the pitch peak in amplitude.

Note from Figure 20 that the amplitude of the main reverberation peak is lower for square than Hamming windowing. In other instances the difference is even greater. This may lead to the somewhat mistaken speculation that the square window has "reduced" the reverberative information in the cepstrum. Such an assumption is incorrect because what has really occurred is that the reverberative information is just more spread out and more difficult to remove than for the Hamming window. Indeed, nothing changes the fact that Figure 20b is the cepstrum of a reverberated speech section! Significantly, the main reverberation peak of Figure 20c is such a dominant feature of this $\hat{\underset{\sim}{p}}_{ev}(n)$ that its removal alone very nearly eliminates reverberation in the processed output speech. This single peak can be removed with little damage to $\hat{s}_{wev}(n)$, because it is so narrow.

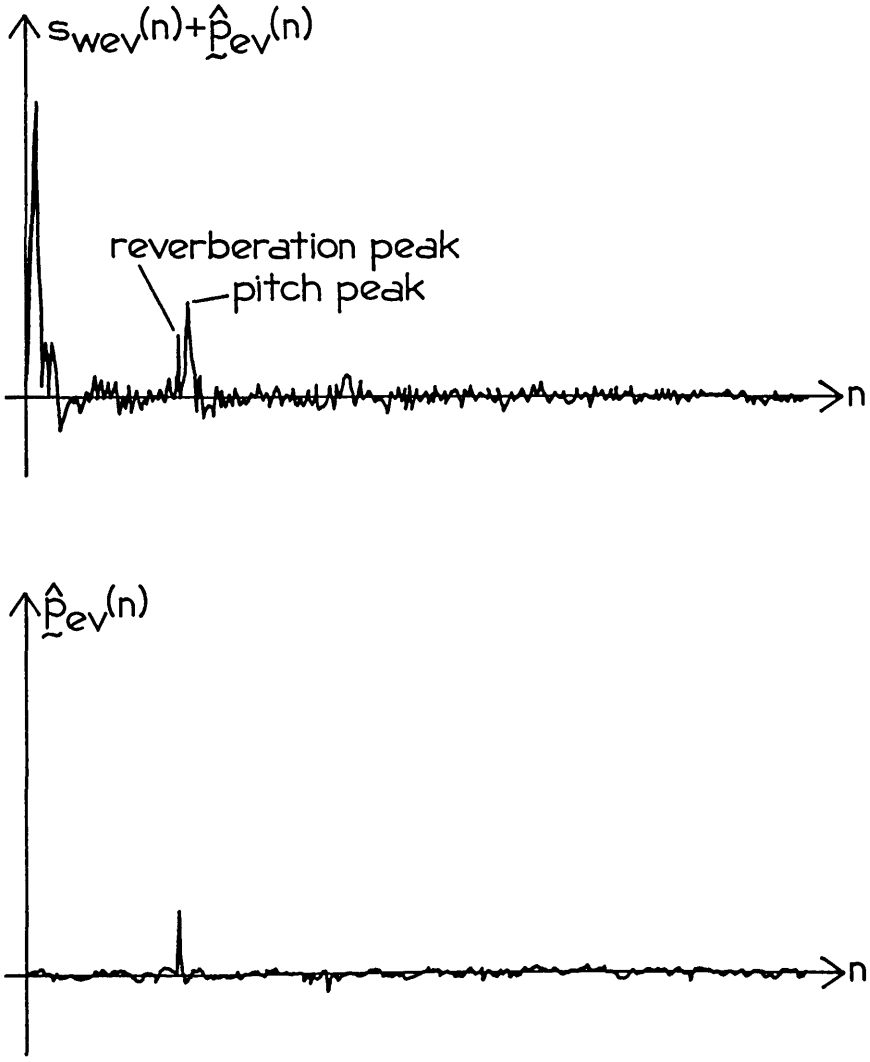
Some window shapes other than Hamming or square were tried, but the performance of the Hamming window was consistently better. In particular, exponential shapes of various decay rates and product windows,

---

*Note, however, the different amplitude scales.

Figure 21

Cepstrum $\hat{p}_{ev}(n)$ of Figure 20c added to $\hat{s}_{wev}(n)$ to form reverberated cepstrum (above), and $\hat{p}_{ev}(n)$ shown alone for reference. The latter is the component that must be removed from the reverberated cepstrum.


$s_{wev}(n)+\hat{p}_{ev}(n)$
reverberation peak
pitch peak
n
$\hat{p}_{ev}(n)$
n

consisting of a Hamming window multiplied by an exponential window, were used.

A. Single-Cepstrum Processing

Cepstral filtering methods can be classified according to whether or not they involve the comparison of two or more cepstra to identify the reverberative information. If only one cepstrum is available, comparison is not possible. This has the most severe effect upon pitch detection, since there is little way to distinguish between reverberation peaks and pitch peaks without cross-reference between the cepstra of two differently-reverberated versions of the same speech waveform section. Other than comparison, four relatively unreliable characteristics can be used to discriminate between peak types:

1. The first pitch peak is always positive; hence negative peaks can be rejected.
2. For reverberation which is not severe, the pitch peak amplitude is often larger than any of the reverberation peaks.
3. Pitch peak location often changes little from a given cepstrum to its successor, if the time between cepstra is on the order of 10 msec. Therefore, if the pitch period from the preceding cepstrum is accurately known, it is known where the pitch peak in the present cepstrum is most likely to be located.
4. The main reverberation peaks, corresponding to dominant impulses in p(n), tend to move even less than the pitch peaks from cepstrum to cepstrum. Changes in their positions result only from motion of the speaker with respect to the microphone. After several cepstra have been examined, therefore, it should be possible to decide which peak is the first pitch peak, because it

would have moved relative to the other peaks. The degree of certainty should increase with the number of cepstra examined. This perhaps will allow "tracking" of reverberation peaks as they slowly move, a procedure which would be very useful in eliminating confusion if the pitch peak location "crosses" a reverberation peak.

Little use can be made of the second characteristic, since high reverberation peaks and/or low pitch peaks sometimes occur. The first characteristic is always useful but often not sufficient, because reverberation peaks are not always negative. An obvious drawback to use of the third characteristic is the possible unreliability of past pitch period estimates. An algorithm which selects one of two peaks on the basis of past pitch data may become erroneously "stuck" on a reverberation peak if it once makes an error in selection. A combination algorithm based on the third and fourth characteristics would have the best chances of success, but would require considerable time to identify and "lock onto" all peaks with certainty.

The recovery of $\hat{h}_{ev}(n)$ is also difficult. If no reverberation peaks lie in the low-time part of the cepstrum, the job is simple. However, this is certainly not always the case. Data obtained by the above methods indicating the positions of reverberation peaks would be helpful in determining which parts of the cepstrum to remove, but the low-time part of $\hat{s}_{wev}(n)$ often contains peaks of its own which can be confused with reverberation peaks. A center-clipping process which excludes all cepstrum samples below a given threshold amplitude, or its inverse, which excludes samples above a given threshold, tend to perform poorly due to the difficulty of making accurate statements about expected $\hat{h}_{ev}(n)$

amplitude or reverberation peak amplitude. A peak of $\hat{h}_{ev}(n)$ erroneously removed from the cepstrum will produce the same effect in the processed output speech as will a reverberation peak which is not removed.

B. Cepstral Averaging

A technique which is not as effective as cepstral comparison but is worthy of discussion is cepstral averaging. This process, and all others involving averaging and/or comparison, requires two or more cepstra computed from corresponding sections of differently-reverberated versions of the input speech waveform. These reverberated waveforms can be obtained from the outputs of substantially-separated microphones.* A necessary property of the reverberation cepstra for averaging to be useful is that the peaks of $\hat{\underset{\sim}{p}}_{1ev}(n)$ do not coincide with those of $\hat{\underset{\sim}{p}}_{2ev}(n)$, $\hat{\underset{\sim}{p}}_{3ev}(n)$, etc. This will generally be the case if the reverberation impulse trains $p_1(n)$, $p_2(n)$, ... are sufficiently different. Then if the k^{th} reverberated cepstrum is denoted by

$$\hat{x}_{kev}(n) = \hat{s}_{wev}(n) + \hat{\underset{\sim}{p}}_{kev}(n),$$

the average of K such cepstra is

$$\frac{1}{K} \sum_{k=1}^{K} \hat{x}_{kev}(n) = \hat{s}_{wev}(n) + \frac{1}{K} \sum_{k=1}^{K} \hat{\underset{\sim}{p}}_{kev}(n).$$

Hence, although there are now more reverberation peaks to contend with, they may be reduced in amplitude by as much as a factor of 1/K. This does <u>not</u> imply that the level of reverberation in the output is 1/K of that in the input. Rather, the magnitude spectrum of each p(n) is exponentiated:

---

*It will be assumed, for simplicity, that there is no "overall delay" between the waveforms, in the sense discussed in Section III.

$$|P_{out}(e^{j\omega})| \simeq |P_{in}(e^{j\omega})|^{1/K} .$$

Because cepstral averaging reduces reverberation peak amplitude, it improves the chances of accurate discrimination between pitch peaks and reverberation peaks. However, it still cannot be guaranteed that reverberation peak amplitude will not exceed pitch peak amplitude. Averaging can also reduce the amplitude of whatever reverberative information overlaps the low-time part of $\hat{s}_{wev}(n)$. Therefore, even if some of this information is not filtered out, the effect may not be as severe as if averaging were not done. It is possible that if the reverberation peak amplitudes are reduced enough, $\hat{s}_{wev}(n)$ would be less distorted than if removal of this information were attempted. This suggests that enough cepstra should be averaged to reduce reverberation peak amplitudes to a level at least as small as $\hat{s}_{wev}(n)$. The obvious drawback to such an idea is the large amount of computation required and the unfeasibility of an excessive number of microphones.

Figure 22 shows the results of several cepstral averaging experiments. In each of these, three cepstra are averaged. The reverberating impulse trains each contain three impulses. The cepstra are shown in pairs, the unreverberated cepstrum $\hat{s}_{wev}(n)$ above, and the averaged, reverberated cepstrum below. Figure 22a, for which the reverberating impulses were spaced with no particular pattern, is one of the best results. Small cepstral reverberation peaks at times corresponding to reverberation impulse delays, however, can be found upon close inspection. These peaks are more clearly identifiable in Figure 22b. In Figure 22c, the reverberation impulse spacing was the same as that in Figures 22a, b, but here a reverberation peak exceeds the pitch peak in amplitude. The reverberating

Figure 22

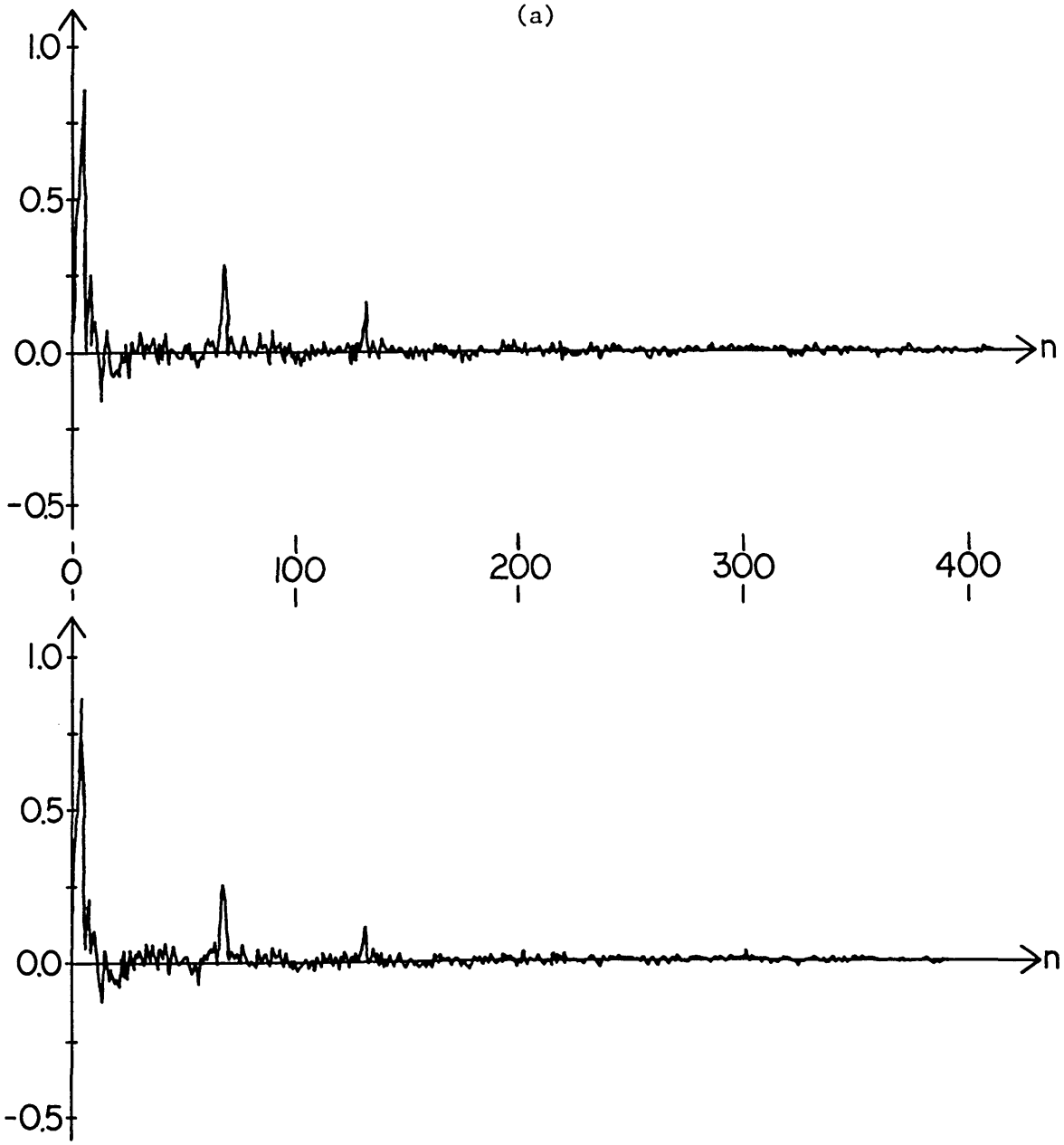
Six examples of cepstral averaging results. First of each pair of cepstra is $\hat{s}_{wev}(n)$ without any reverberation components added. Second is average of three reverberated cepstra. Each p(n) is of the form

$$p(n) = a_0 u_o(n) + a_1 u_o(n - n_1) + a_2 u_o(n - n_2) .$$

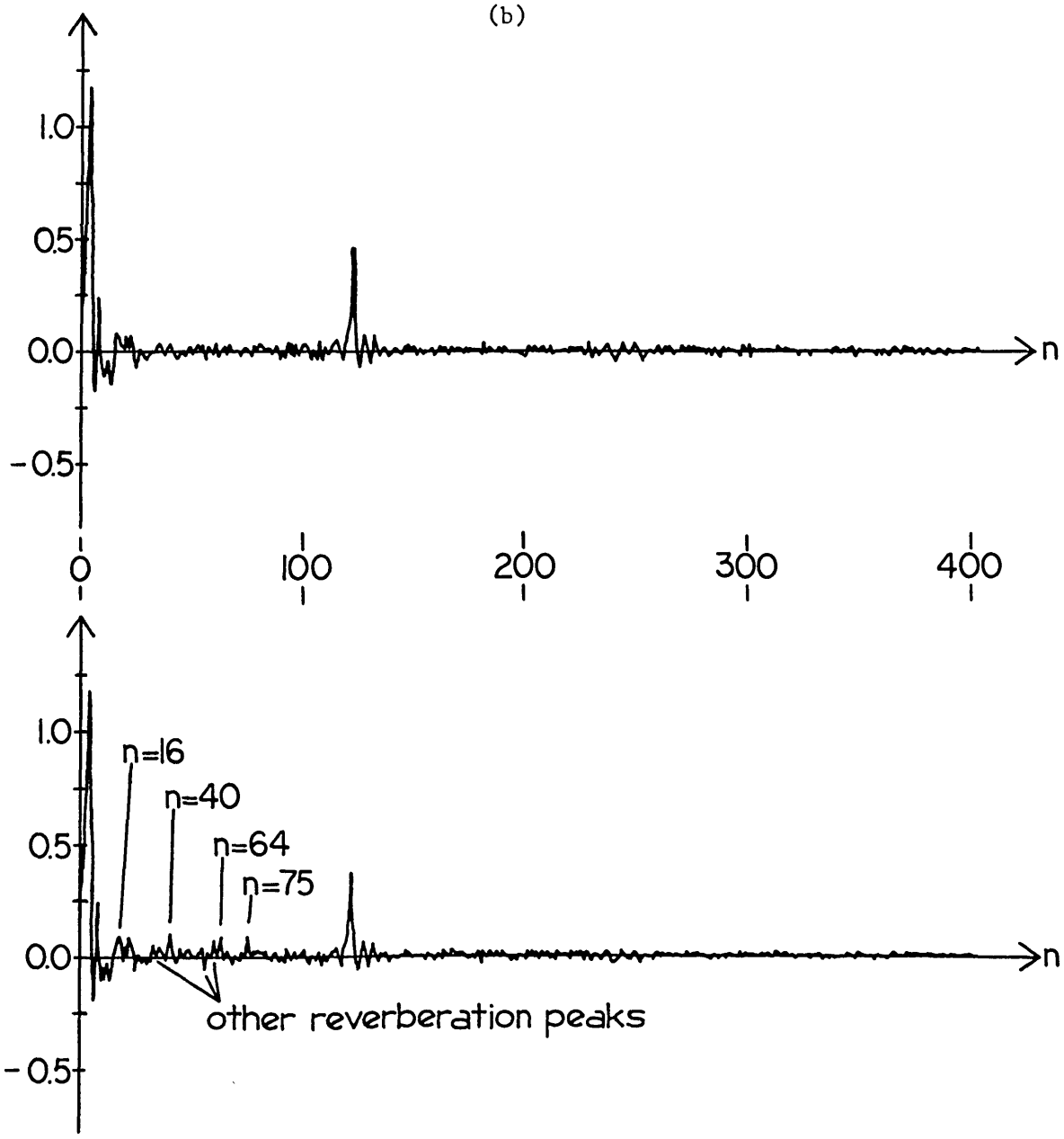
The values for the impulse amplitudes and delays for each example are given below:

		a_0	a_1	a_2	n_1	n_2
(a)	$p_1(n)$:	1.0	0.9	0.5	64	75
	$p_2(n)$:	1.0	0.8	0.76	16	40
	$p_3(n)$:	1.0	0.75	0.86	45	77
(b)	$p_1(n)$:					
	$p_2(n)$:	same as example (a)				
	$p_3(n)$:					
(c)	$p_1(n)$:					
	$p_2(n)$:	same as example (a)				
	$p_3(n)$:					
(d)	$p_1(n)$:	0.35	0.70	0.35	35	70
	$p_2(n)$:	-0.35	0.70	-0.35	61	122
	$p_3(n)$:	0.35	0.70	0.35	50	100
(e)	$p_1(n)$:					
	$p_2(n)$:	same as example (d)				
	$p_3(n)$:					
(f)	$p_1(n)$:	1.0	0.90	0.81	35	70
	$p_2(n)$:	1.0	0.90	0.81	61	122
	$p_3(n)$:	1.0	0.90	0.81	50	100

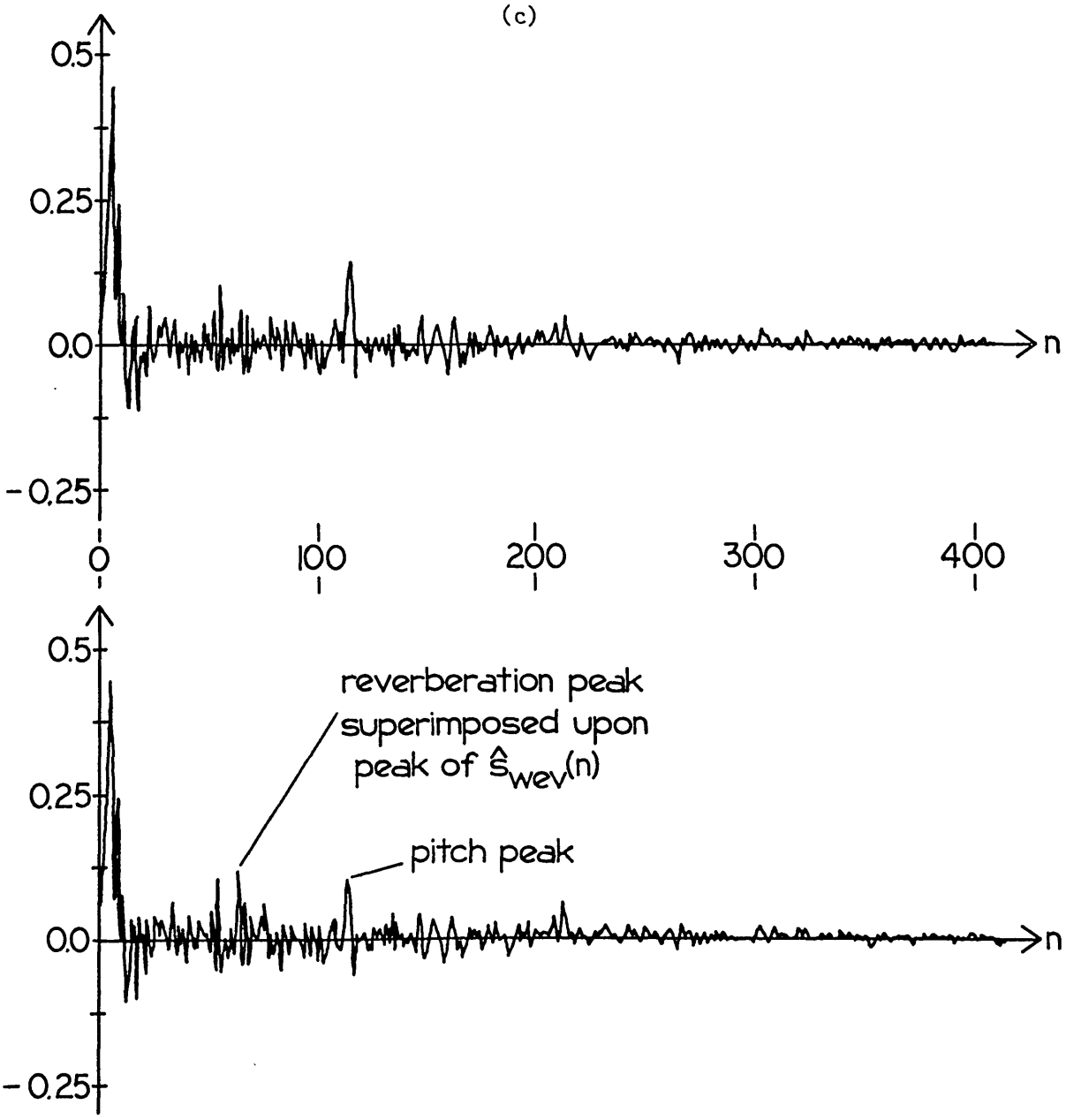
(a)



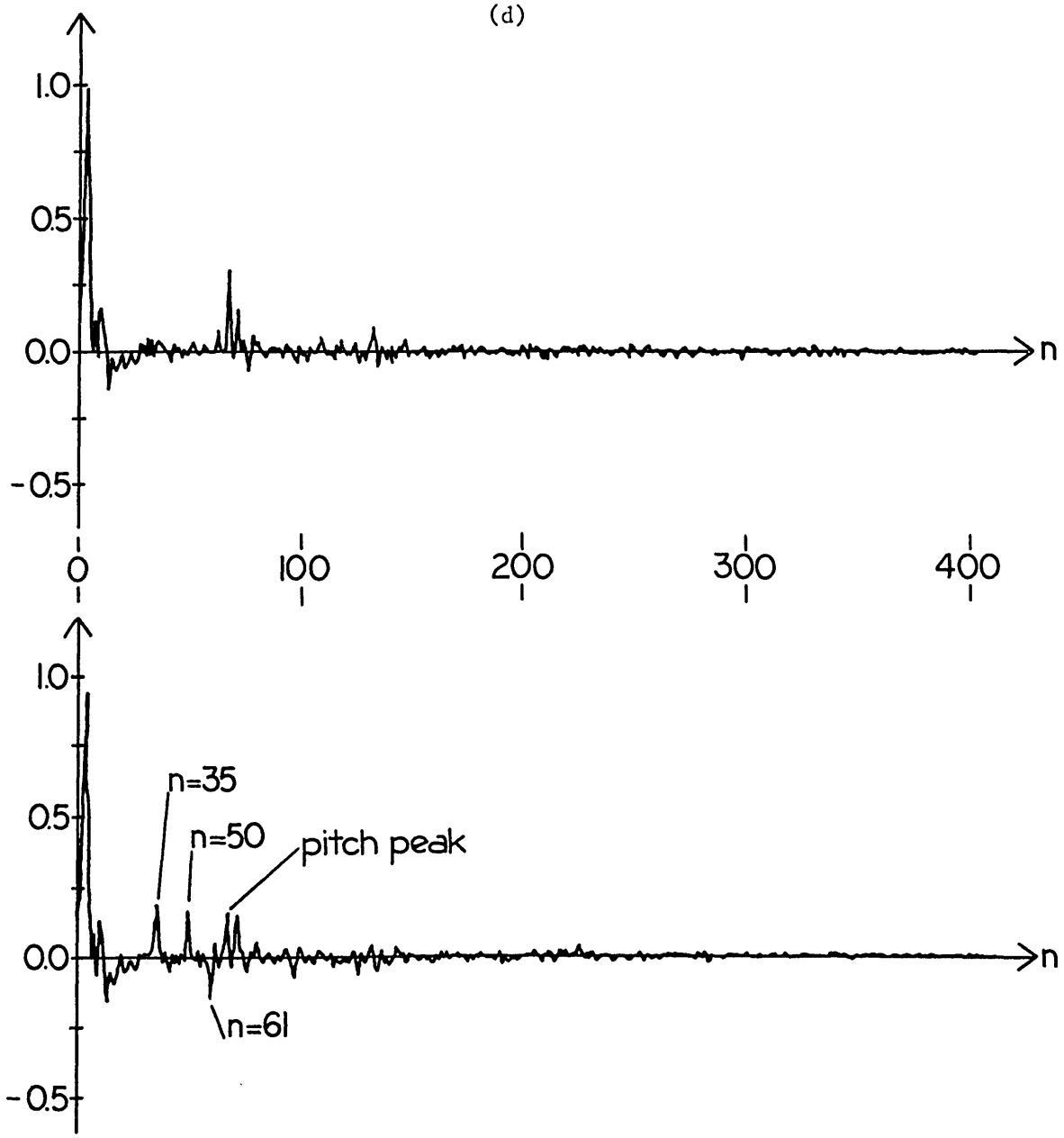
(b)


1.0
0.5
0.0
-0.5
n
0
100
200
300
400
1.0
0.5
0.0
-0.5
n=16
n=40
n=64
n=75
n
other reverberation peaks

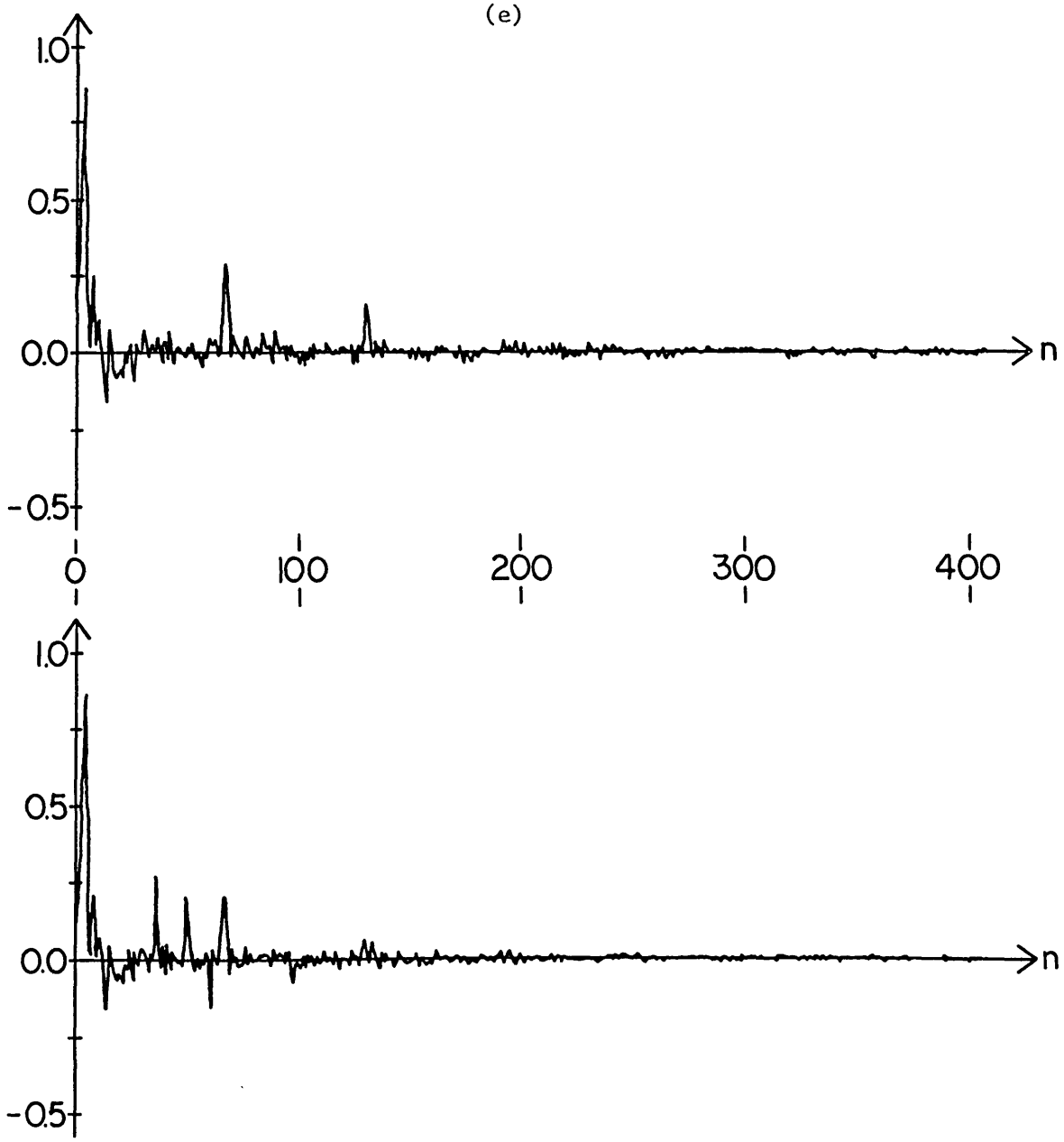
(c)


0.5
0.25
0.0
-0.25
n
0
100
200
300
400
0.5
0.25
0.0
-0.25
n
reverberation peak
superimposed upon
peak of $\hat{s}_{wev}(n)$
pitch peak

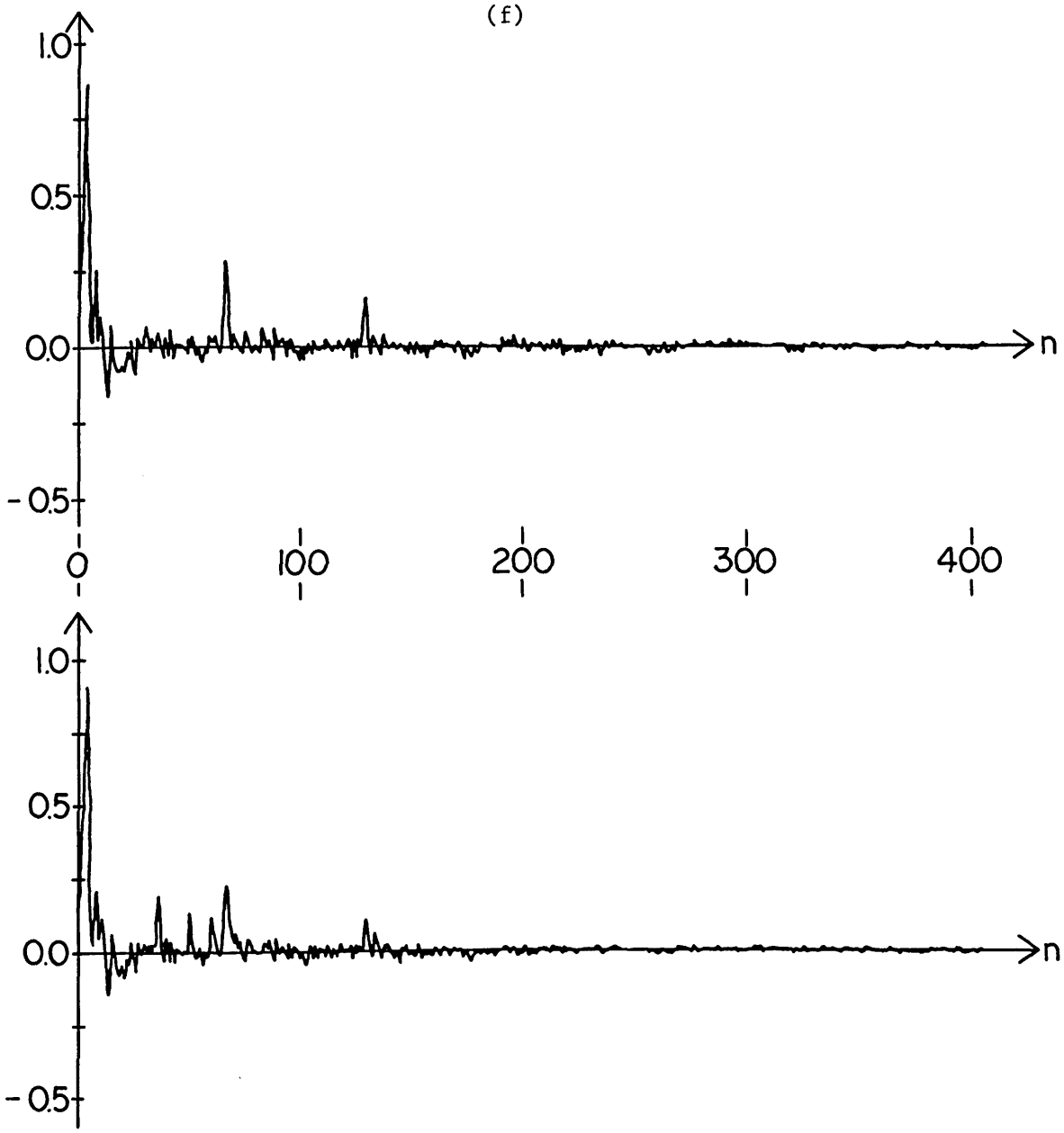
(d)



(e)



(f)



impulses for Figures 22 d, e, and f are equally-spaced, which produces more serious reverberation than random spacing. The reason for this is that an impulse train of the form

$$p(n) = u_o(n) + \sum_{k=1}^{M} a_k u_o(n-kP)$$

can be expressed as a convolution of M two-impulse trains:

$$p(n) = \left[u_o(n) + b_1 u_o(n-P)\right] * \ldots * \left[u_o(n) + b_M u_o(n-P)\right].$$

Each two-impulse train has a magnitude spectrum like that in Figure 6, with the "valleys" occurring at a spacing of $2\pi/P$. Since the spectrum of p(n) is the product of M such spectra, and since the valleys in each spectrum correspond, the valleys in $P(e^{j\omega})$ will be very "deep", resulting in serious reverberative distortion. If, in the case of three impulses, p(n) is of the form

$$p(n) = au_o(n) \pm 2au_o(n-P) + au_o(n-P),$$

spectral zeroes are produced, and the reverberation peaks are very large. This was done in the examples of Figures 22d, e. Figure 22d is a good example of a case in which several reverberation peaks could be confused with the pitch peak. Note, however, that the negative reverberation peak could be rejected, because the first pitch peak must be positive. In Figure 22e, the reverberation peak amplitudes again exceed the pitch peak amplitude. They are again so large that if they were not removed, the output speech would remain reverberant-sounding. The impulse weighting for Figure 22f is not as severe as for Figures 22d, e, but the reverberation peaks are still unacceptably large even after averaging.

These examples indicate that cepstral averaging alone is probably insufficient for reverberation reduction for feasible numbers of averaged

cepstra. In addition, it does not solve the problem of interference with pitch detection. Therefore, a method of removing the reverberation peaks before pitch detection is necessary.

C. Cepstral Averaging with Adaptive Comb-Filtering

The following discussion is phrased in terms of the comparison of two cepstra, for simplicity, but may be extended to any number of cepstra if desired.

It is desirable, when removing reverberant peaks from the cepstrum, to minimize the damage to $\hat{s}_{wev}(n)$ incurred in the process. Simple threshold clipping is unsatisfactory in this respect, because it is insufficiently selective. It is possible, however, to identify reverberation peaks in the average of two cepstra by computing the difference, $d_{12}(n)$, of the cepstra in addition to averaging. The common $\hat{s}_{wev}(n)$ components cancel:

$$d_{12}(n) \triangleq \hat{x}_{1ev}(n) - \hat{x}_{2ev}(n) = \hat{s}_{wev}(n) + \hat{p}_{1ev}(n) - [\hat{s}_{wev}(n) + \hat{p}_{2ev}(n)]$$

$$= \hat{p}_{1ev}(n) - \hat{p}_{2ev}(n).$$

Therefore, if the dominant peaks of $\hat{p}_{1ev}(n)$ and $\hat{p}_{2ev}(n)$ are never coincident, $d_{12}(n)$ will exhibit a large peak wherever $\hat{p}_{1ev}(n)$ or $\hat{p}_{2ev}(n)$ have large peaks.* The peaks may be removed from the average cepstrum using a multiplicative "comb-filter" c(n), defined by

$$c(n) = 1, \quad |d_{12}(n)| < t(n)$$

$$= 0, \quad |d_{12}(n)| > t(n)$$

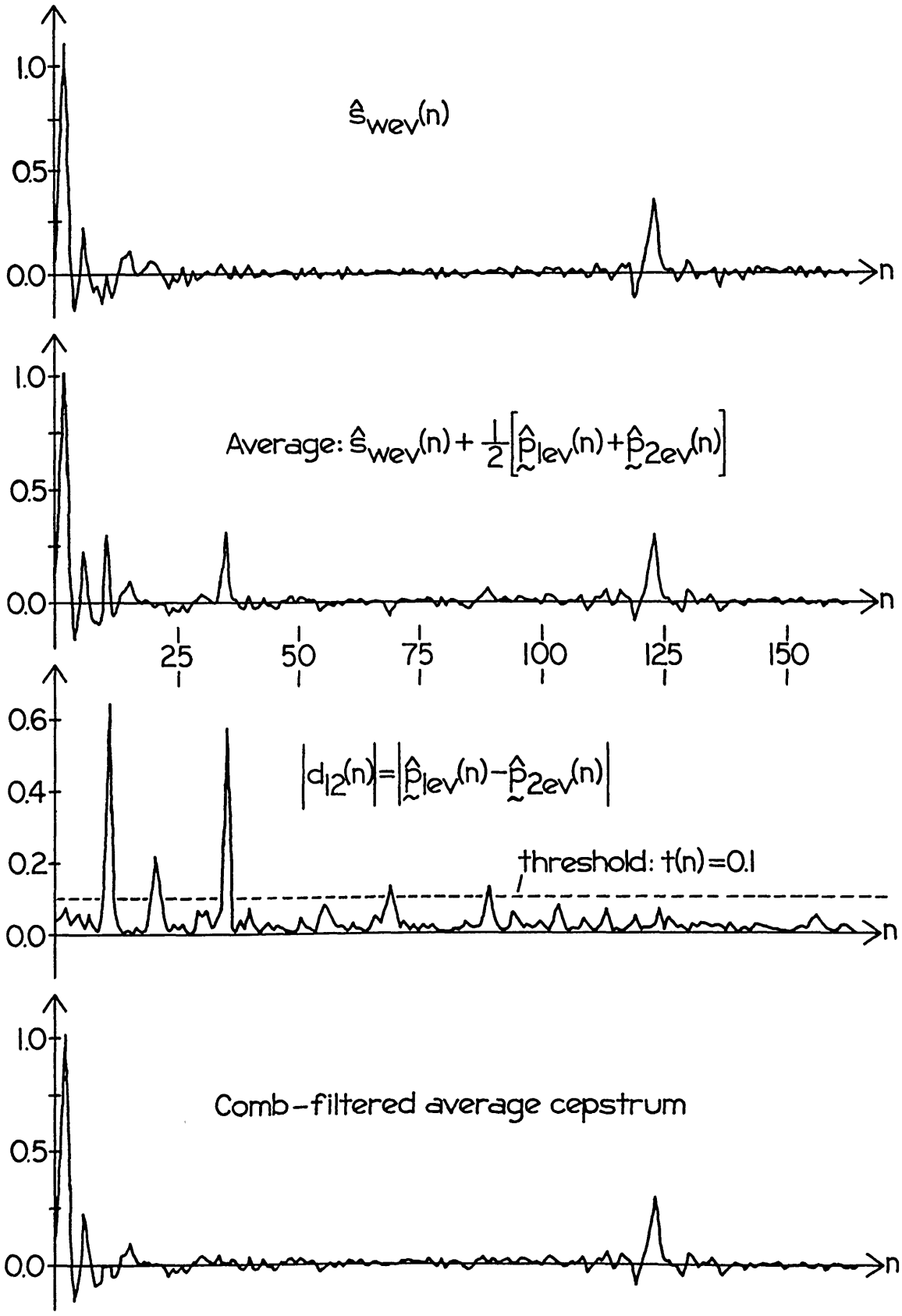
where t(n) is a suitably-defined threshold function. The filtered

---

*Of course, if $\hat{p}_{1ev}(n)$ and $\hat{p}_{2ev}(n)$ have coincident peaks of comparable amplitude and the same sign, this technique of peak identification fails. It is assumed this occurs infrequently.

Figure 23

Example of identification and removal of reverberation peaks by locating large peaks in the difference of two reverberated cepstra and comb-filtering the average cepstrum. The "comb-filtering" is accomplished by multiplying (in the time domain) the average cepstrum by zero at every sample where the cepstral difference, $|d_{12}(n)|$, exceeds the threshold, t(n).


1.0
0.5
0.0
$\hat{s}_{wev}(n)$
n
1.0
0.5
0.0
Average: $\hat{s}_{wev}(n) + \frac{1}{2}\left[\hat{\underset{\sim}{p}}_{1ev}(n) + \hat{\underset{\sim}{p}}_{2ev}(n)\right]$
n
25
50
75
100
125
150
0.6
0.4
0.2
0.0
$|d_{12}(n)| = |\hat{\underset{\sim}{p}}_{1ev}(n) - \hat{\underset{\sim}{p}}_{2ev}(n)|$
threshold: t(n)=0.1
n
1.0
0.5
0.0
Comb-filtered average cepstrum
n

average is then

$$\left[\frac{1}{2}\,\hat{x}_{1ev}(n) + \frac{1}{2}\,\hat{x}_{2ev}(n)\right]\, c(n) \simeq \hat{s}_{wev}(n).$$

An example of the filtering process is shown in Figure 23.

Proper choice of the threshold t(n) is important. Clearly it should be high enough to avoid making c(n) equal zero for too many values of n. A reasonable criterion for choosing t(n) is that it equal the average absolute value of $\hat{s}_{wev}(n)$ for each value of n. Then, if a reverberation peak is larger than the threshold, the chance is good that removal of the peak is worth the distortion of $\hat{s}_{wev}(n)$ caused by removal. If the reverberation peak falls below the threshold, it is probably smaller than $\hat{s}_{wev}(n)$ and its removal would likely worsen the distortion. A hypothetical "best shape" for t(n), excluding consideration of pitch peaks, is illustrated in Figure 24a, and a step approximation to it in Figure 24b. It may be desired to make the threshold extra high in the low-time region, if preventing filtering distortion of $\hat{h}_{ev}(n)$ is considered more important than removing low-time reverberative peaks. Such a procedure may be justified by the observation that large reverberation peaks in the low-time region generally correspond to short-delay reverberation, which often is usually more tolerable than long-delay reverberation. Also, if cepstral pitch detection will be done after multiplicative filtering, it is important that the filtering leave the pitch peak undistorted even if there is a reverberation peak near to or coincident with the pitch peak. "Protection" of the pitch peak can be accomplished by raising the threshold in the area where the pitch peak is expected, according to the last cepstral pitch period measurement. Note that the pitch peak can fall outside the protected area and still be detected, however. A threshold which

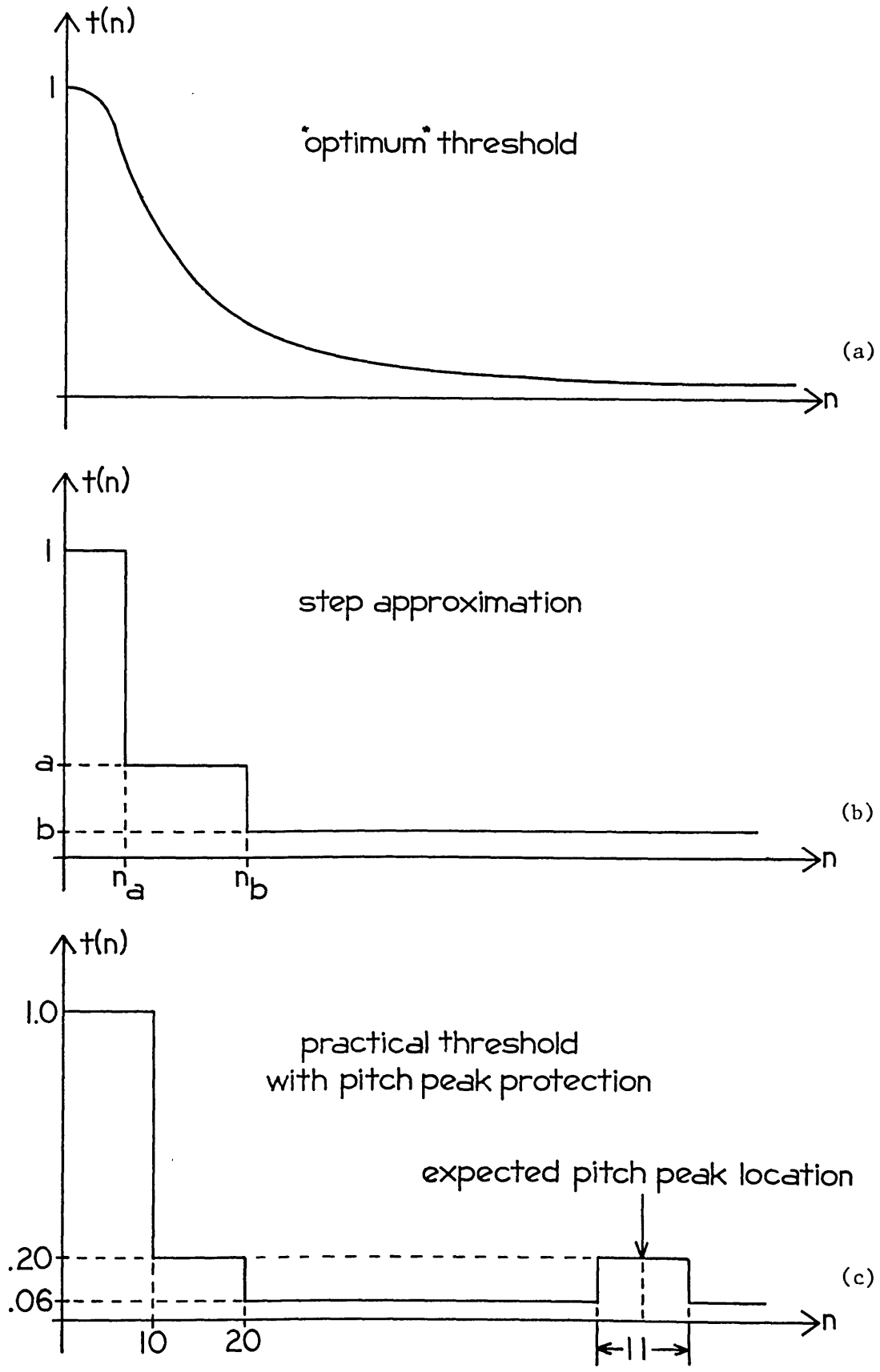
Figure 24

Better threshold functions, t(n), which could be used in the filtering process of Figure 23. Thresholds account for expected amplitude of $\hat{s}_{wev}(n)$.

(a) Hypothetical smooth "optimum" threshold

(b) A step approximation to the threshold of Figure 24a, suitable for computational use.

(c) Modified threshold with amplitude increased in the region where the pitch peak is expected to occur. This threshold was used extensively in speech-processing experiments.


t(n)
1
"optimum" threshold
n
(a)
t(n)
1
step approximation
a
b
n
n_a
n_b
(b)
t(n)
1.0
practical threshold
with pitch peak protection
expected pitch peak location
.20
.06
10
20
11
n
(c)

produced good results in practice is illustrated in Figure 24c, and specified by

$$t(n) = 1 \quad , 0 \leq |n| \leq 10$$
$$= 0.2 \ , 11 \leq |n| \leq 20$$
$$= 0.06, \ |n| > 21,$$

based on a sampling rate of 10KHz, with a pitch peak protection threshold of 0.2 within ± 5 samples of the expected pitch peak location. Better thresholds can perhaps be devised with further experimentation.

Several speech-processing experiments were done using the above techniques. Three male-spoken all-voiced sentences were used:

1. "May we all learn a yellow lion roar." (Speaker LJG)
2. Same sentence (Speaker PDB)
3. "We were away a year ago." (Speaker LRR)

Common parameters in the experiments, except as noted, were:

Sampling rate = 10KHz (4KHz desampling filter)

Interval between successive cepstra = 100 samples (10 msec)

Length of Hamming sectioning window = 500 samples (50 msec)

Number of points in all DFT's = 512 samples

Cepstral truncation point = 32 samples (3.2 msec) .

Minimum-phase resynthesis, as produced by (see Section III)

$$k(n) = 2, \ 0 < n < 256$$
$$= 1, \ n = 0 \text{ and } n = 256$$
$$= 0, \ 256 < n < 511$$

was used in all experiments. Cepstral truncation was smoothed by use of a tapered truncation window $\ell(n)$:

$$\ell(n) = 1, \quad 0 \leq n \leq 30$$

$$= \frac{1}{2}[1 + \cos \frac{2\pi}{8}(n-31)], \quad 31 \leq n \leq 35$$

$$= 0, \quad 36 \leq n \leq 512.$$

Two cepstra were averaged in all cases. All reverberation was artificial.

Results for the first sentence were particularly good. Artificial reverberation, producing spectral zeroes, was used in both input channels:

Channel 1: $p(n) = 0.35u_o(n) + 0.70u_o(n-35) + 0.35u_o(n-70)$

Channel 2: $p(n) = 0.35u_o(n) + 0.70u_o(n-50) + 0.35u_o(n-100)$.

The synthesized output was less rough-sounding than when the other sentences were used. Aural comparison of the processed sentence with each reverberated input showed that reverberation was essentially removed. Comparison of the output with the original unprocessed sentence was also favorable. However, the speaker's voice was very low-pitched (the period ranged as high as 170 samples), and for all but one trial with this sentence, a 400-sample Hamming input window rather than a 500-sample window was used. Numerous pitch detection errors resulted from the consequent low pitch-peak amplitude: serious errors occurred in about 25 of 190 pitch measurements, when 512-point FFT's were used. Most of the pitch errors occurred in the lowest-pitched portions of the sentence. Interestingly, when 1024-point FFT's were tried (still with 400-sample input sections), the number of serious errors in the first half of the sentence, which was all that was processed for the 1024-point DFT's, dropped from six to zero. The use of a 500-sample window reduced the number of pitch errors from 25 to 19.

The largest number of experiments were done using the second sentence. In all trials the synthesized versions of this sentence

sounded noticeably rougher than for the other two sentences, a problem which was never completely resolved. However, except in certain cases in which very short-delay reverberation was used, reverberation in the processed speech was virtually inaudible. Because there was substantial roughness in the synthetic speech, a synthetic standard was used in making this judgment. The standard was produced by passing identical unreverberated speech inputs through the processor. It was very difficult to distinguish between the standard and the dereverberated speech, except in cases that will be noted.

When the severe reverberation discussed two paragraphs above was used, the results were very good. Figures 25a, b, c, d show the narrow-band spectrograms of the original sentence, one of the reverberated inputs (with 35-sample reverberation impulse spacing), the dereverberated speech (Figure 25b processed), and the processed standard, respectively. The corresponding wide-band spectrograms are shown in Figure 26. Figure 25b shows very clearly the spectral zeroes produced by the reverberation, while temporal effects of the reverberation can be seen in Figure 26b. These distortions have been virtually eliminated in Figures 25c and 26c. Comparison of the originals with the standards, however, reveals that the formant bandwidths in some areas have been increased by the processing technique, which contributes to the rapid decay of the h(n)'s in the synthesis. Another effect of the processing was to "smear" the transitions between different voiced sounds.

Pitch detection in this experiment was also notably good. The two serious errors occurred at the end of the sentence. Pitch accuracy, however, was affected when a net delay was introduced in the Channel 1 input relative to the Channel 2 input. Net delays of 10, 22, 40, and 75

Figures 25 and 26

Spectrograms of all-voiced sentence "May we all learn a yellow lion roar" (PDB speaker).

(25a) Narrowband spectrogram, original sentence.

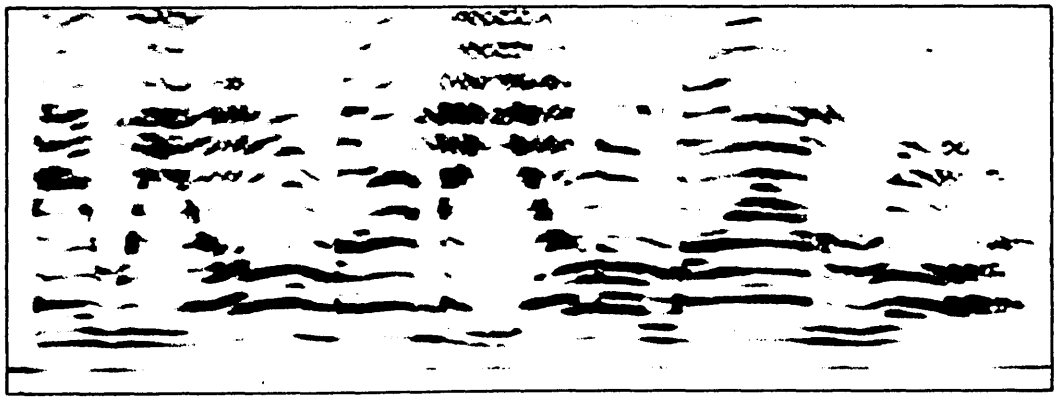
(25b) Narrowband spectrogram, reverberated but unprocessed sentence.

(25c) Narrowband spectrogram, processed reverberated sentence.

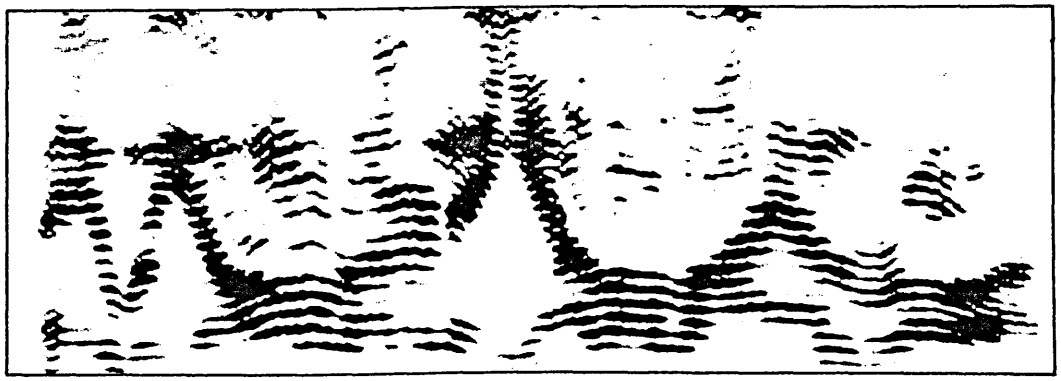
(25d) Narrowband spectrogram, sentence processed but unreverberated before processing.

(26a, b, c, d) Corresponding wideband spectrograms.

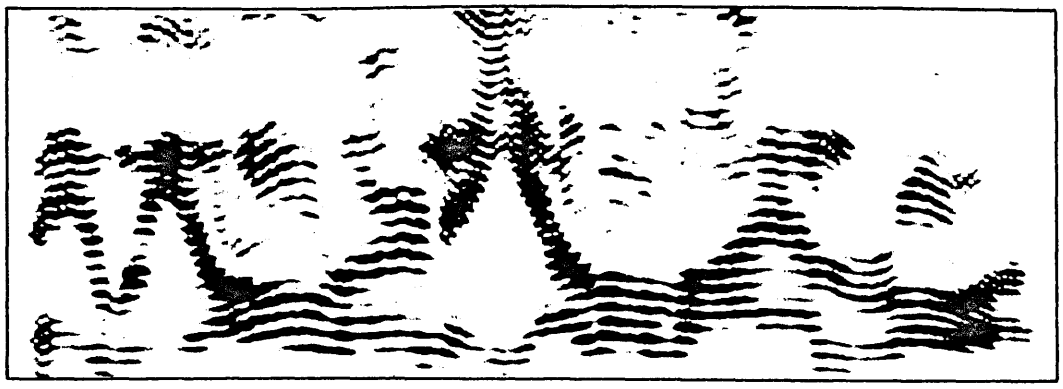
(a)



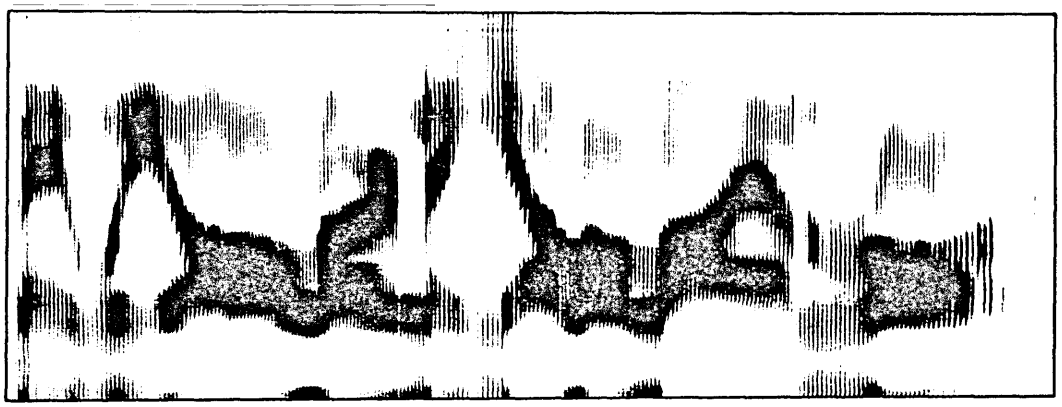
(b)



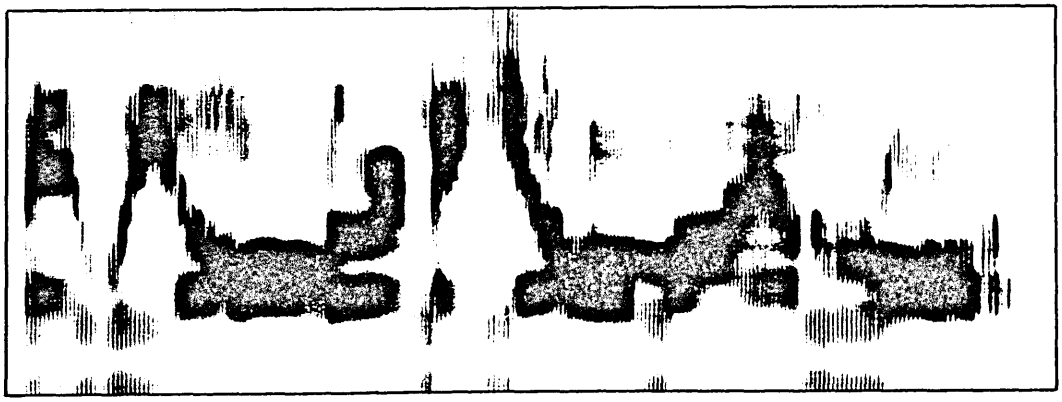
(c)



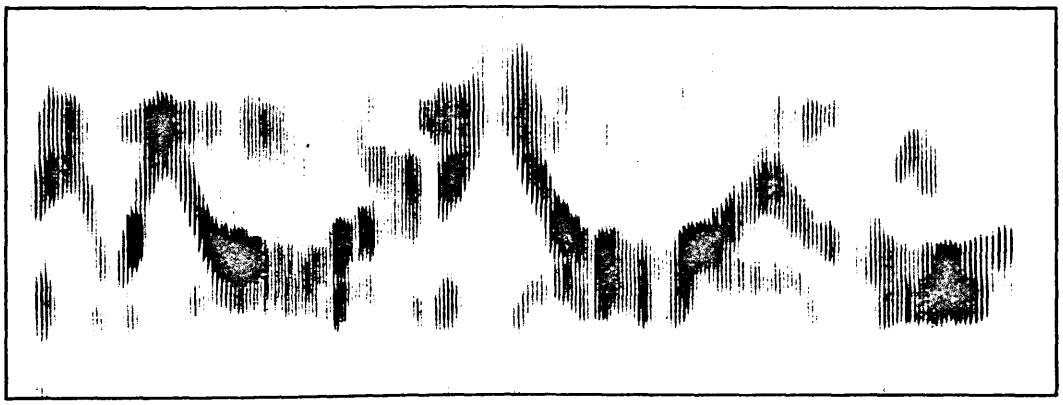
(d)



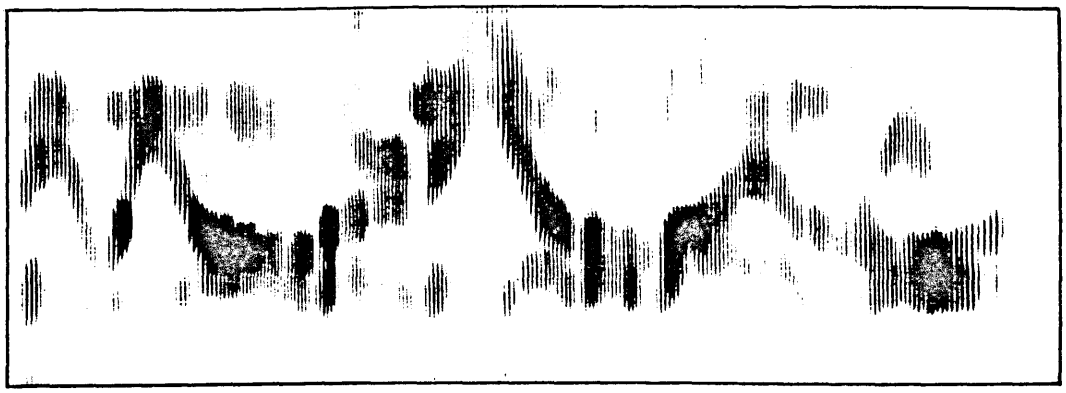
(a)



(b)



(c)



(d)

samples caused 3, 4, 4, and 6 serious pitch errors, respectively, out of a total of 204 pitch measurements. The probable cause of these is the averaging of two slightly different $\hat{s}_{wev}(n)$'s. If the pitch period is rapidly changing in an area of the waveform, the pitch peak in the cepstrum of the delayed waveform may be shifted relative to the pitch peak in the cepstrum of the undelayed input. Therefore, the peaks will not reinforce each other in the average, and low amplitude will result. On the other hand, the several sentences were essentially undistinguishable from the zero-delayed and standard processed outputs, indicating that the quality of dereverberation remained basically unaffected by the delays.

A crucial test of the comb-filtering technique was its effectiveness in minimizing distortion of $\hat{h}_{ev}(n)$. It was not possible to evaluate this aspect of the algorithm's performance from the above experiments, because the cepstrum was truncated at 35 samples, which was the lowest echo delay. Only low-level reverberative information distorted the low-time portion of the cepstrum, and no filtering was done there. An experiment was performed to determine whether the truncation point could be extended to n = 50 or above when the same reverberation was used. It was found that residual reverberance and non-linear distortion in the output increased somewhat as the truncation point was moved higher. In subsequent experiments it was decided to leave the truncation point at 35 samples.

With the truncation point set at 35 samples, the second sentence was processed with the Channel 2 reverberation replaced by

$$p(n) = 0.35u_o(n) + 0.70u_o(n-15) + 0.35u_o(n-30).$$

This caused a reverberation peak to appear at n = 15 in the cepstrum.

After the usual filtering, the processed sentence was comparable in quality to the standard, indicating that $\hat{h}_{ev}(n)$ suffered only slight-to-moderate distortion from removal of this peak. However, in another experiment the Channel 2 reverberation was left as above, and the Channel 1 reverberation replaced with

$$p(n) = 0.35u_o(n) + 0.70u_o(n-19) + 0.35u_o(n-38).$$

The filtering algorithm failed to remove this reverberation without considerable distortion of the speech, although the sentence remained intelligible.

Experiments with the third utterance were performed using

$$p_1(n) = 0.35u_o(n) + 0.70u_o(n-35) + 0.35u_o(n-70)$$

$$p_2(n) = 0.35u_o(n) + 0.70u_o(n-50) + 0.35u_o(n-100)$$

and

$$p_1(n) = 0.35u_o(n) + 0.70u_o(n-15) + 0.35u_o(n-30)$$

$$p_2(n) = 0.35u_o(n) + 0.70u_o(n-19) + 0.35u_o(n-38),$$

repeating two of the above combinations. The results of these were comparable to the corresponding trials with the second sentence.

It can be concluded from these experiments that comb-filtering of the averaged cepstrum prior to pitch detection, according to peaks in the cepstral difference, effectively eliminates confusion between pitch peaks and reverberation peaks, except in the case of low pitch peaks or coincident peaks of $\hat{p}_{1ev}(n)$ and $\hat{p}_{2ev}(n)$. Long pitch period and net delays between the inputs tend to reduce pitch detection performance by causing low pitch peaks to occur. The low-time part of the cepstrum can be recovered without unreasonable distortion if there are no speech waveform echoes with delay times below the cepstral truncation point. Intolerable distortion may result if the echo delays fall below this point.

D. Summary

The Hamming window, which was found in Section III to have desirable properties with respect to pitch-synchronous synthesis of processed speech, is found also to increase the ability of cepstral filtering to remove the reverberative component from the cepstrum. This is important to the compatibility of dereverberation and synthesis requirements.

Single-cepstrum processing appears to be unfeasible due to the difficulty of distinguishing between reverberation peaks and parts of $\hat{s}_{wev}(n)$, given the information provided by a single cepstrum. The averaging of two or more cepstra can reduce the amplitude of reverberation peaks without distorting $\hat{s}_{wev}(n)$, but is probably insufficient to eliminate confusion between pitch peaks and reverberation peaks in the pitch-detection process. Computation of the difference of two reverberated cepstra can yield the information necessary to remove large reverberation peaks prior to pitch detection, reducing the possibility of detection errors. Reverberation peak removal is done by multiplicative comb-filtering of the averaged cepstrum.

Processing results for three artificially-reverberated voiced sentences indicates that for echo times in the range 3-10 msec, the comb-filtering technique essentially eliminates the effects of reverberation. Also, differential delays of up to 7.5 msec between the two input waveforms were found to cause little distortion of the speech. Results were not obtained for naturally-reverberated, unvoiced, or female speech.

V. Weighted Averaging of Cepstra

When the two cepstra are averaged in the above filtering process, substantial amounts of little-distorted information contained in the individual cepstra can be lost. Between the large peaks of $\hat{p}_{1ev}(n)$, for example, it is often true that

$$\hat{x}_{1ev}(n) \simeq \hat{s}_{wev}(n).$$

The same is true for $\hat{p}_{2ev}(n)$ and $\hat{x}_{2ev}(n)$. Sometimes, areas which are badly distorted in one reverberated cepstrum are little distorted in the other. Therefore, when the averaged cepstrum is comb-filtered without regard to this possibility, parts of $\hat{x}_{1ev}(n)$ and $\hat{x}_{2ev}(n)$ are destroyed which might have provided better estimates of $\hat{s}_{wev}(n)$.

It is possible to utilize the individual cepstra to greater advantage by computing a weighted average in which the least distorted portions of $\hat{x}_{1ev}(n)$ and $\hat{x}_{2ev}(n)$ are emphasized and the most distorted portions are suppressed. Let $a_1(n)$ and $a_2(n)$ be two weighting sequences satisfying

$$a_1(n) + a_2(n) = 1, \text{ all } n.$$

Then

$$a_1(n)\hat{x}_{1ev}(n) + a_2(n)\hat{x}_{2ev}(n) = \hat{s}_{wev}(n) + \left[a_1(n)\hat{p}_{1ev}(n) + a_2(n)\hat{p}_{2ev}(n)\right].$$

The sequences $a_1(n)$ and $a_2(n)$ should be chosen to minimize the absolute value of the bracketed term. Since $\hat{p}_{1ev}(n)$ and $\hat{p}_{2ev}(n)$ are unknown, however, the term cannot be made to vanish identically. However, it can be minimized with respect to a probabilistic criterion. Several criteria are possible.

In the discussions below, the following shorthand notation is adopted for simplicity:

$$\hat{s}_w \leftrightarrow \hat{s}_{wev}(n)$$

$$\hat{x}_1,\ \hat{x}_2 \leftrightarrow \hat{x}_{1ev}(n),\ \hat{x}_{2ev}(n)$$

$$\hat{p}_1,\ \hat{p}_2 \leftrightarrow \hat{p}_{1ev}(n),\ \hat{p}_{2ev}(n)$$

$$a_1,\ a_2 \leftrightarrow a_1(n),\ a_2(n).$$

In addition, E[-] is used to denote the expected value or mean of a random variable in the brackets. It is assumed that $\hat{p}_1$, $\hat{p}_2$, $\hat{s}_w$, $\hat{x}_1$, and $\hat{x}_2$ can be treated as random variables if desired.

One approach to the weighted averaging problem is to choose a_1 and a_2 such that the weighted average is the minimum linear mean-square-error estimate of $\hat{s}_w$, given $\hat{x}_1$ and $\hat{x}_2$. $\hat{x}_1$ and $\hat{x}_2$ can be interpreted as known quantities, while $\hat{p}_1$ and $\hat{p}_2$ are treated as random variables dependent upon the random variable $\hat{s}_w$ through the relations

$$\hat{p}_1 = \hat{x}_1 - \hat{s}_w$$

$$\hat{p}_2 = \hat{x}_2 - \hat{s}_w.$$

In other words, the problem is essentially viewed as one of attempting to determine $\hat{p}_1$ and $\hat{p}_2$ in the presence of uncertainty due to $\hat{s}_w$. The weights a_1 and a_2 are found by minimizing the expression

$$\begin{aligned} e(a_1) &= E\{[a_1\hat{p}_1 + a_2\hat{p}_2]^2\} \\ &= E\{[a_1(\hat{x}_1-\hat{s}_w) + (1-a_1)(\hat{x}_2-\hat{s}_w)]^2\} \\ &= E\{[a_1(\hat{x}_1-\hat{x}_2) + \hat{x}_2 - \hat{s}_w]^2\}, \end{aligned}$$

which is the mean square error subject to the condition that $a_1 + a_2 = 1$. This yields (except for the case $\hat{x}_1 = \hat{x}_2$)

$$a_1 = \frac{E(\hat{s}_w) - \hat{x}_2}{\hat{x}_1 - \hat{x}_2},\ a_2 = 1 - a_1 = \frac{\hat{x}_1 - E(\hat{s}_w)}{\hat{x}_1 - \hat{x}_2}. \tag{12}$$

These coefficients are not useful for direct estimation of $\hat{s}_w$, because

the weighted average reduces simply to

$$a_1\hat{x}_1 + a_2\hat{x}_2 = E(\hat{s}_w).$$

However, we shall see that a modified approach can be followed, in which better use can be made of estimation techniques.

The coefficients of Equation 12 serve to justify some "intuitive" conceptions about the proper choice of weights. For example, if $\hat{x}_1$ is close to $E(\hat{s}_w)$ while $\hat{x}_2$ is very different from $E(\hat{s}_w)$, Equation 12 states that a_1 should be chosen much larger than a_2. This supports the notion that, with high probability, $\hat{x}_1$ is a better estimate of $\hat{s}_w$ than is $\hat{x}_2$. Therefore, $\hat{x}_1$ should be weighted more heavily than $\hat{x}_2$. The same example can be interpreted in another way. From Equation 12, the linear minimum mean-square-error estimates of $\hat{p}_1$ and $\hat{p}_2$ given $\hat{x}_1$ and $\hat{x}_2$ are

$$E[\hat{p}_1|\hat{x}_1,\hat{x}_2] = \hat{x}_1 - (a_1\hat{x}_1 - a_2\hat{x}_2) = \hat{x}_1 - E(\hat{s}_w)$$

$$E[\hat{p}_2|\hat{x}_1,\hat{x}_2] = \hat{x}_2 - E(\hat{s}_w). \qquad (13)$$

In this example, the estimate of $\hat{p}_1$ is much smaller in magnitude than the estimate of $\hat{p}_2$, dictating a weighting bias toward $\hat{x}_1$. Another example is the case in which $\hat{x}_1 \simeq \hat{x}_2$. On the basis of these measurements, there is no information to indicate an estimate other than $\hat{s}_w \simeq \hat{x}_1 \simeq \hat{x}_2$. Consistent with this situation, Equation 12 specifies

$$a_1 \simeq a_2.$$

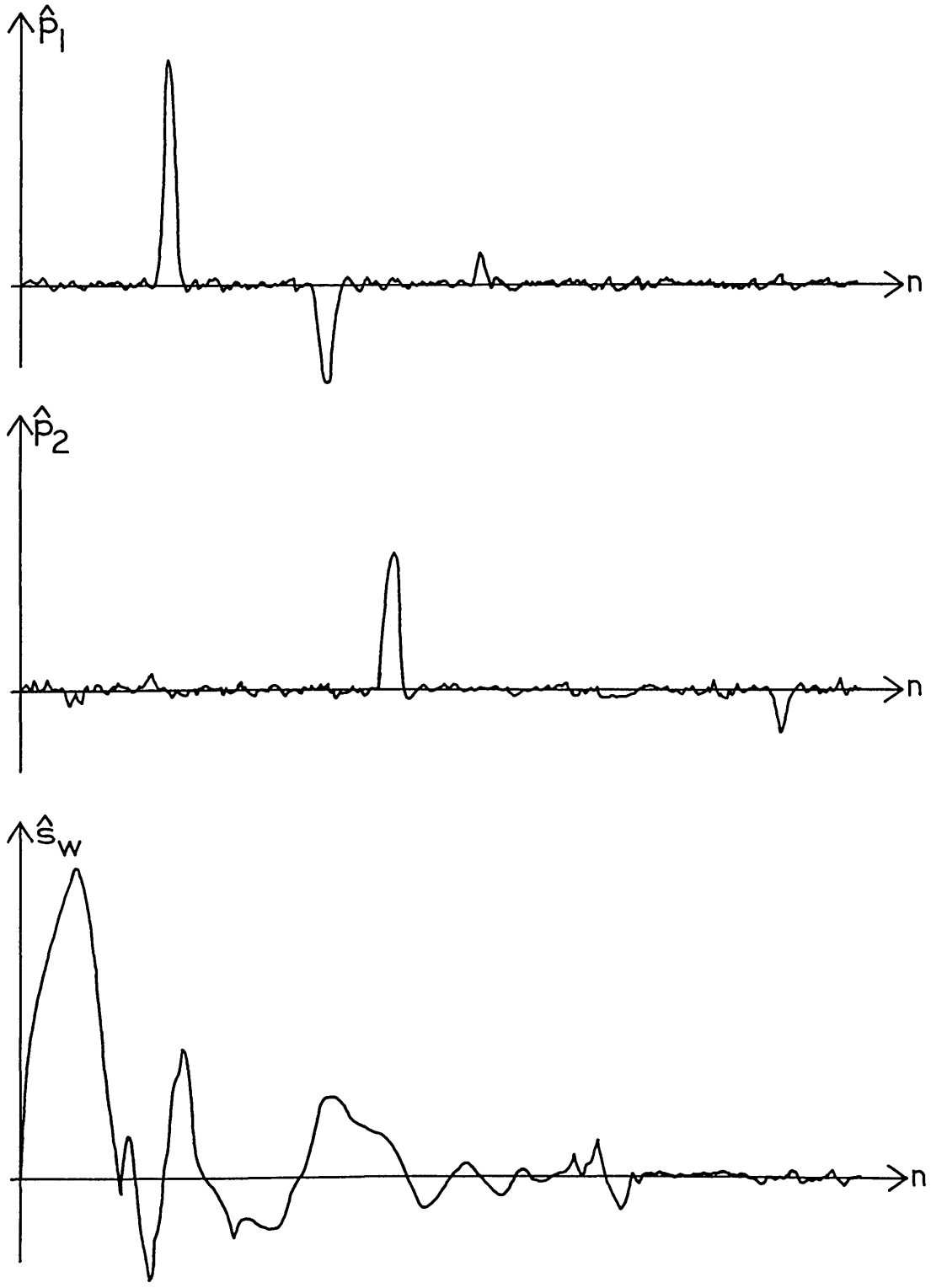
Under some circumstances, however, $\hat{x}_1$ would be chosen as the better estimate of $\hat{s}_w$, while $\hat{x}_2$ may in fact be a far superior one. As illustrated in Figure 27, this occurs when, for example, $E(\hat{s}_w) = 0$, $\hat{p}_1 \simeq -\hat{s}_w$, and $\hat{p}_2 \simeq 0$. Equation 12 would erroneously dictate a preference toward $\hat{x}_1$.* Errors of

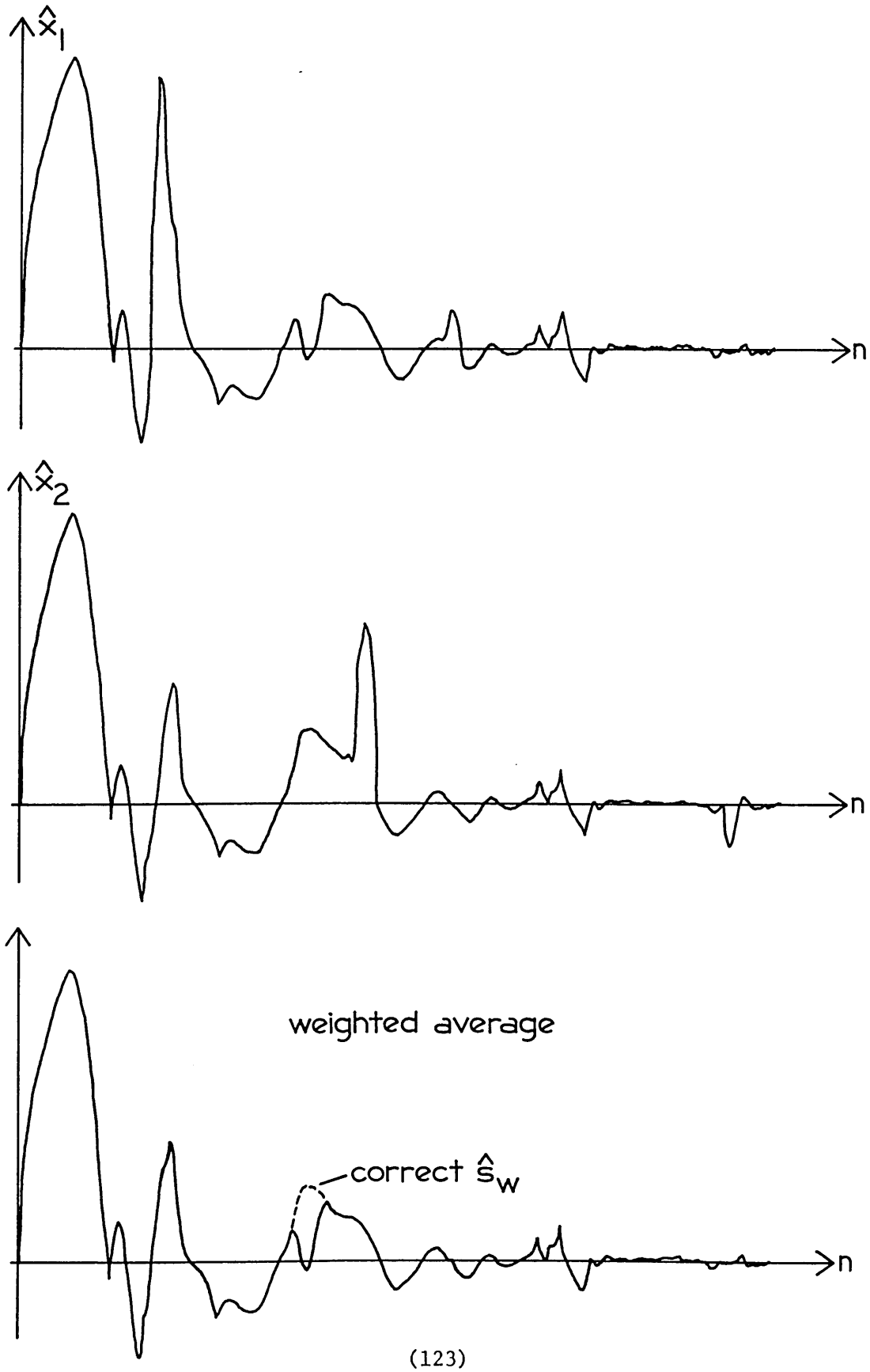
---

*Note that this is also an instance in which comb-filtering would destroy $\hat{s}_w$ at the peak location.

Figure 27 (two pages)

Example of filtering error that can occur if $\hat{p}_1 \simeq -\hat{s}_w$, $\hat{p}_2 \simeq 0$, and $E(\hat{s}_w)=0$. One possible method of preventing such errors is to utilize past estimates of reverberation peak locations and amplitudes.


$\hat{p}_1$
n
$\hat{p}_2$
n
$\hat{s}_w$
n


$\hat{x}_1$
n
$\hat{x}_2$
n
weighted average
correct $\hat{s}_w$
n

this type due to at least short-term variations in $\hat{s}_w$ can be suppressed by basing the estimates of $\hat{p}_1$ and $\hat{p}_2$ partially upon similar estimates for past cepstra. A convenient method of accomplishing this is to let the "best" estimates of $\hat{p}_1$ and $\hat{p}_2$ for the present pair of cepstra equal the exponential average of the present and all past estimates:

$$\hat{p}_{1k} = b \sum_{m=0}^{\infty} (1-b)^m \left[\hat{x}_{1(k-m)} - E(\hat{s}_w)\right]$$

$$\hat{p}_{2k} = b \sum_{m=0}^{\infty} (1-b)^m \left[\hat{x}_{2(k-m)} - E(\hat{s}_w)\right] \qquad (14)$$

In the above, the subscript 'k' corresponds to the present cepstra, 'k-1' to the immediately preceding cepstra, etc. The advantage of this particular way of averaging estimates is that $\hat{p}_{1k}$ and $\hat{p}_{2k}$ can be computed from only $\hat{p}_{1(k-1)}$ and $\hat{p}_{2(k-1)}$:

$$\hat{p}_{1k} = b\left[\hat{x}_{1k} - E(\hat{s}_w)\right] - (1-b)\hat{p}_{1(k-1)}$$

$$\hat{p}_{2k} = b\left[\hat{x}_{2k} - E(\hat{s}_w)\right] - (1-b)\hat{p}_{2(k-1)}.$$

Therefore, it is necessary to store only one past value to take into account all past estimates. The parameter 'b' controls the extent to which past estimates influence the present estimates.

The choice of b is important to the effectiveness of this procedure. As b approaches zero, Equation 14 approaches a sample mean, and suppression of the effects of changes in $\hat{s}_w$ from cepstrum to cepstrum becomes great. Nevertheless, it is not possible, for $b < 1$, to suppress indefinitely the influence of a long-persistent excursion of $\hat{s}_w$ from its mean, such as might result from an unchanging voiced sound of extended duration. Suppose, for instance, that $\hat{s}_w$ is a step function of k, which although unrealistic, helps to reveal the behavior of the estimates in

response to a long-term change in $\hat{s}_w$:

$$\hat{s}_w = \beta u_{-1}(k).$$

If $\hat{p}_1$ and $\hat{p}_2$ are assumed to be time-invariant, then

$$\hat{p}_{1k} = \hat{p}_1 + \beta - E(\hat{s}_w) - \beta(1-b)^k, \quad k \geq 0$$

$$\hat{p}_{2k} = \hat{p}_2 + \beta - E(\hat{s}_w) - \beta(1-b)^k, \quad k \geq 0.$$

Hence, the estimates eventually change by β. A corresponding measure of the short-term rejection properties of the averaging process is provided by the response of the estimates to an impulse change in $\hat{s}_w$:

$$\hat{s}_w = \beta u_o(k).$$

Under the same conditions as above, this response is

$$\hat{p}_{1k} = \hat{p}_1 - E(\hat{s}_w) + \left[b\beta(1-b)^k u_{-1}(k)\right], \text{ all } k$$

$$\hat{p}_{2k} = \hat{p}_2 - E(\hat{s}_w) + \left[b\beta(1-b)^k u_{-1}(k)\right], \text{ all } k.$$

Note that the maximum excursion of the estimates is proportional to b.

In practice, unfortunately, b could not be made arbitrarily small, because $\hat{p}_{1k}$ and $\hat{p}_{2k}$ exhibit the same response to changes in $\hat{p}_1$ and $\hat{p}_2$ as to changes in $\hat{s}_w$. Variations in $\hat{p}_1$ and $\hat{p}_2$ occur as a result of motion of the speaker relative to the microphones. The cepstral filtering process, to remain effective during such motion, must be able to respond to these variations. Therefore, b must be chosen to "average out" the longest-term changes in $\hat{s}_w$ possible while maintaining adequate response to $\hat{p}_1$ and $\hat{p}_2$. For this choice of b, the exponential average can be expected to produce fairly reliable estimates of $\hat{p}_1$ and $\hat{p}_2$ with high probability.

The following hypothetical example provides additional insight into the factors influencing the choice of b. Assume that $\hat{s}_w$ and $\hat{p}$ (representing $\hat{p}_1$ or $\hat{p}_2$) can be modelled as Gaussian random processes with

autocorrelations $R_{\hat{s}_w}(k_1-k_2)$ and $R_{\hat{p}}(k_1-k_2)$ but generally nonstationary means $\eta_{\hat{s}_w}(k)$ and $\eta_{\hat{p}}(k)$, respectively.* $R_{\hat{s}_w}$ and $R_{\hat{p}}$ are measures of cepstrum-to-cepstrum variations of $\hat{s}_w$ and $\hat{p}$ relative to their means. The broader the distributions of $R_{\hat{s}_w}$ and $R_{\hat{p}}$ are, the smaller are the expected variations of $\hat{s}_w$ and $\hat{p}$ over a given number of cepstra. Of course, the time interval between cepstra directly influences the width of the distributions. As this interval increases, the distributions narrow; in the limit of infinite time intervals, the samples of $\hat{s}_w$ and $\hat{p}$ become entirely uncorrelated. For finite time intervals, in any case, it is advantageous that $R_{\hat{p}}$ be much more broadly distributed than $R_{\hat{s}_w}$, as will presently be demonstrated. This allows the tradeoff discussed in the previous paragraph to be made.

The process of exponential averaging is a linear transformation of the averaged variables. The result of the transformation is equivalent to the action of a linear system with impulse response

$$\lambda(k) = b(1-b)^k u_{-1}(k)$$

upon an input corresponding to the sequence of estimates. It is well-known that if the input to this linear system is a Gaussian random process with mean $\eta_i(k)$ and autocorrelation $R_i(k_1-k_2)$, then the output is also a Gaussian random process, with

$$\eta_o(k) = \lambda(k) * \eta_i(k)$$

and

$$R_o(k) = R_i(k) * \left[\lambda(k) * \lambda(-k)\right], \quad k = k_1 - k_2.$$

---

*Similar results could be obtained if $\hat{p}$ and $\hat{s}_w$ were considered to be deterministic signals.

In the case at hand, the input is

$$\hat{x} - E(\hat{s}_w) = \hat{s}_w + \hat{p} - \eta_{\hat{s}}.$$

Assume that $\hat{s}_w$ and $\hat{p}$ are independent.* Then

$$\eta_i(k) = \eta_{\hat{p}}(k)$$

and

$$R_i(k) = R_{\hat{s}_w}(k) + R_{\hat{p}}(k).$$

It follows that

$$\eta_o(k) = \lambda(k) * \eta_{\hat{p}}(k) \tag{15a}$$

$$R_o(k) = \left[R_{\hat{s}_w}(k) + R_{\hat{p}}(k)\right] * \left[\lambda(k) * \lambda(-k)\right] \tag{15b}$$

$R_o(k)$ and $\eta_o(k)$ as determined above specify the estimates $\hat{p}_k$. Now, a measure of the effectiveness of the cepstral filtering algorithm is the degree to which the process $\hat{p}_k$ is equivalent to the process $\hat{p}$, of which $\hat{p}_k$ is an estimate. According to Equations 15a, b, this requires that

$$\lambda(k) * \eta_{\hat{p}}(k) \simeq \eta_{\hat{p}}(k)$$

and

$$\left[R_{\hat{s}_w}(k) + R_{\hat{p}}(k)\right] * \left[\lambda(k) * \lambda(-k)\right] \simeq R_{\hat{p}}(k).$$

These approximations are good if b is sufficiently <u>large</u> and $R_{\hat{s}_w}$ is relatively narrow compared to $R_{\hat{p}}$. In other words, the impulse response $\lambda(k)$ must be of sufficiently short duration to clearly resolve $\eta_{\hat{p}}$ and $R_{\hat{p}}$, and $\hat{p}$ should tend to vary slowly relative to $\hat{s}_w$. (It is not known whether the latter requirement is met in practice for normal speech and motion patterns.) A second important measure of performance is the variance of

---

*This is not exactly true in practice, because $\hat{p}$, or $\hat{p}_{ev}(n)$, depends upon $s(n)$, and thus upon $\hat{s}_w(n)$. However, due to the effect of the Hamming window, this dependence is considered to be weak.

$\hat{p}_k$ relative to that of $\hat{s}_w$. For the estimates of $\hat{p}_1$ and $\hat{p}_2$ to be accurate with high probability, which is necessary for them to be useful, the variance of $\hat{p}_k$ should be small compared to that of $\hat{s}_w$. I.e., variations in $\hat{s}_w$ must not be allowed to exert a strong influence on $\hat{p}_k$. The variances of $\hat{s}_w$ and $\hat{p}_k$ are given by

$$\sigma^2_{\hat{s}_w} = R_{\hat{s}_w}(0)$$

and

$$\begin{aligned}\sigma^2_{\hat{p}_k} &= R_o(0) \\ &= \sum_{k=-\infty}^{+\infty} \left[R_{\hat{s}_w}(k) + R_{\hat{p}}(k)\right]\left[\lambda(k) * \lambda(-k)\right] \\ &= \sum_{k=-\infty}^{+\infty} \left[R_{\hat{s}_w}(k) + R_{\hat{p}}(k)\right]\left[\frac{b}{2-b}\left\{[1-b]^k u_{-1}(k) + [1-b]^{-k} u_{-1}(-k)\right\}\right]\end{aligned}$$

where b must be in the range

$$0 < b < 1.$$

Therefore, the requirement that $\sigma^2_{\hat{p}} << \sigma^2_{\hat{s}_w}$ necessitates choosing b as small as possible. This implies that, as expected, there must be a compromise between the accuracy with which $\hat{p}_k$ tracks $\hat{p}$ and the accuracy of estimating $\hat{s}_w$ (through estimating $\hat{p}$).

Although extension of this analytic technique to the general case would be formidable, similar results can be expected to apply. It may or may not prove to be sufficiently realistic to model $\hat{s}_w$ and $\hat{p}$ as Gaussian random processes. If it does, such modelling could be a useful tool. In applying the model, $R_{\hat{s}}$ and $R_{\hat{p}}$ could be estimated by measurement of the power density spectra of these signals* for a large set of typical cases.

---

*Keep in mind that these are not, in this case, $\hat{s}_w(n)$ and $\hat{p}(n)$, but rather $\hat{s}_w(k)$ and $\hat{p}(k)$, where k is the cepstrum-to-cepstrum index.

Similarly, sample means could be used to estimate means. If meaningful statistics can be obtained in this way, it may be that they vary in a quite definite manner from speaker to speaker.

The next step is to transform the estimates $\hat{p}_{1k}$ and $\hat{p}_{2k}$ into weights a_1 and a_2. If the estimates were of sufficient accuracy, $\hat{s}_w$ could be determined directly by choosing a_1 and a_2 so that

$$a_1\hat{p}_{1k} + a_2\hat{p}_{2k} = 0.$$

However, due to the above considerations, $\hat{p}_1$ and $\hat{p}_2$ cannot be estimated with arbitrary accuracy. Furthermore, even if $p_1(n)$ and $p_2(n)$ are time-invariant, there will be fluctuations in $\hat{p}_1$ and $\hat{p}_2$ because of their dependence upon $s(n)$. Although Hamming weighting of the input sections tends to suppress this dependence, the variations are enough that $\hat{p}_{1k}$ and $\hat{p}_{2k}$ probably cannot be considered to be sufficiently accurate in an amplitude sense. The main value of these estimates, then, is for the positive identification of large reverberation peaks in one cepstrum or the other. Generally, the locations of these peaks are very constant for time-invariant $p_1(n)$ and $p_2(n)$, and they would also be fairly stable for quasi-time-invariant $p_1(n)$ and $p_2(n)$. Therefore, in a "peak locational" sense, the estimates $\hat{p}_{1k}$ and $\hat{p}_{2k}$ should be potentially quite reliable, with the reliability increasing as the peak amplitude increases relative to the variance of $\hat{s}_w$. This could be particularly helpful in the 1-10 msec low-time range of the cepstrum.

Since $\hat{p}_{1k}$ and $\hat{p}_{2k}$ are not especially reliable in the amplitude sense, $\hat{s}_w$ can perhaps be best estimated by allowing the estimate $\hat{s}_{wk}$ to equal $\hat{x}_2$ if $|\hat{p}_{1k}| >> |\hat{p}_{2k}|$ with high probability and $\hat{x}_1$ if $|\hat{p}_{2k}| >> |\hat{p}_{1k}|$ with high probability. This will usually occur when there is a large peak

in either cepstrum, because coincident large peaks of $\hat{p}_1$ and $\hat{p}_2$ tend to occur with low probability. Due to this fact, it is most likely better to assume that a peak of $\hat{s}_w$ has occurred than coincident peaks of $\hat{p}_1$ and $\hat{p}_2$ if $\hat{x}_1 \simeq \hat{x}_2$. The alternative would be to assume that $\hat{p}_1$ and $\hat{p}_2$ were both large and to set $\hat{s}_{wk}$ equal to its expected value. This is unjustified due to the low relative probability of such an event. Finally, if $\hat{p}_{1k} \simeq -\hat{p}_{2k}$, the best estimate of $\hat{s}_w$ is probably just the average of $\hat{x}_1$ and $\hat{x}_2$.

A suitable method of specifying weights a_1 and a_2 in terms of $\hat{p}_{1k}$ and $\hat{p}_{2k}$, accounting for the reliability of the estimates as a function of the expected variation in $\hat{s}_w$, must, therefore, satisfy the following:

$$a_1 >> a_2 \text{ if } \frac{|\hat{p}_{1k}| - |\hat{p}_{2k}|}{\sigma_{\hat{s}_w}} << 1$$

$$a_2 >> a_1 \text{ if } \frac{|\hat{p}_{1k}| - |\hat{p}_{2k}|}{\sigma_{\hat{s}_w}} >> 1$$

$$a_1 \simeq a_2 \simeq \frac{1}{2} \text{ if } -1 < \frac{|\hat{p}_{1k}| - |\hat{p}_{2k}|}{\sigma_{\hat{s}_w}} < 1 \ ,$$

where $\sigma_{\hat{s}_w}$ is the standard deviation of $\hat{s}_w$.

Figure 28a shows a hypothetical plot of $\sigma_{\hat{s}_w}$ as might be expected from observation of several typical voiced speech cepstra. In Figure 28b is illustrated a reasonable weight-assignment curve as based on the above criteria, and in Figure 28c a stepped approximation to this curve, suitable for computational use.

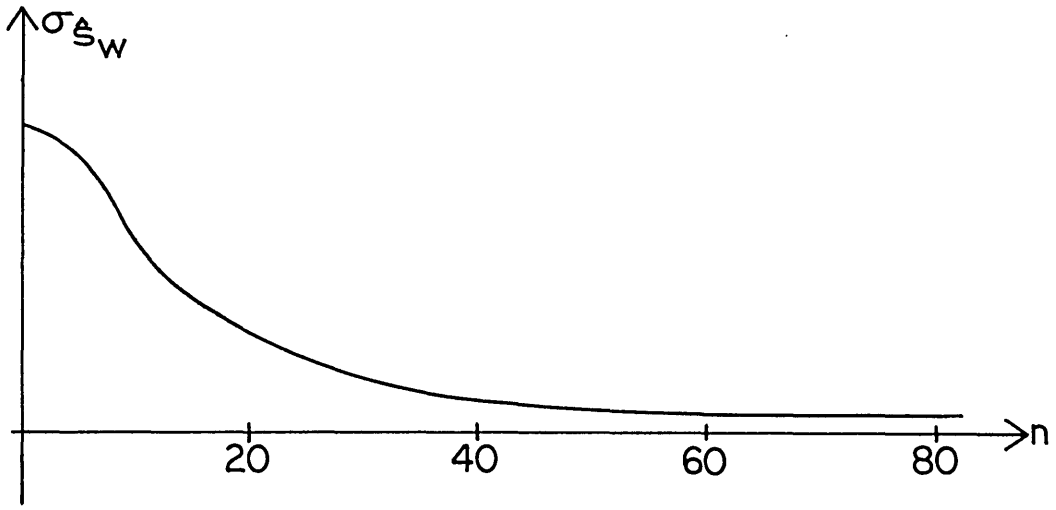
The experimentally-determined sample mean and standard deviation plots of Figure 29 lend credibility to the possibility of this type of filtering. These statistics were computed from an ensemble of 204 short-

Figure 28

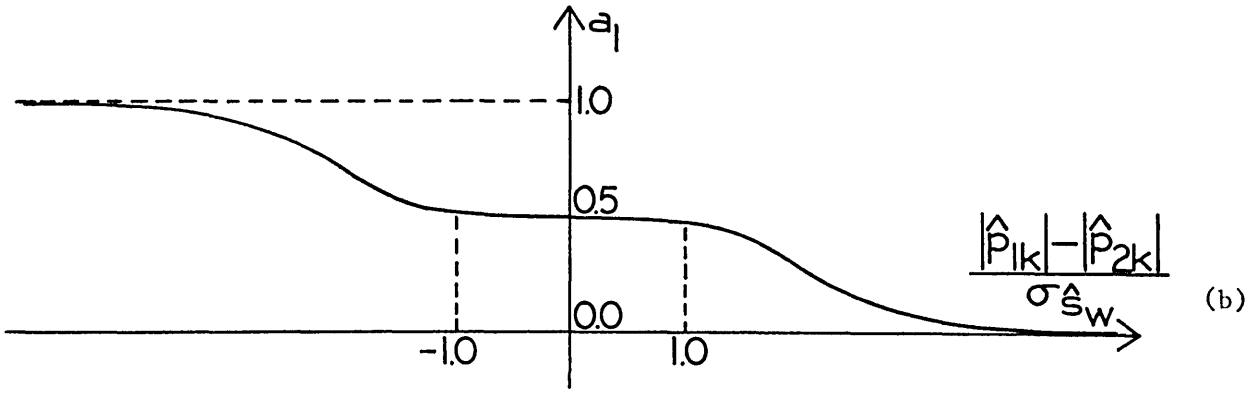
(a) Hypothetical plot of the cepstral standard deviation, $\sigma_{\hat{s}_w}$, as a function of n.

(b) A reasonable (smooth) weight-assignment curve.

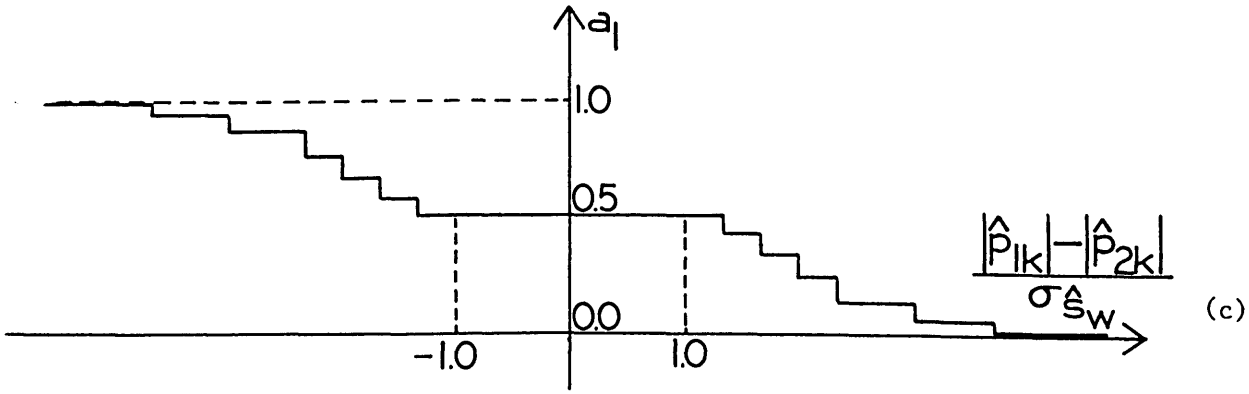
(c) A step approximation to the curve of Figure 28b, suitable for computational use.


$\sigma_{\hat{s}_w}$
20
40
60
80
n

(a)


a_1
1.0
0.5
0.0
-1.0
1.0
$\frac{|\hat{P}_{1k}|-|\hat{P}_{2k}|}{\sigma_{\hat{s}_w}}$

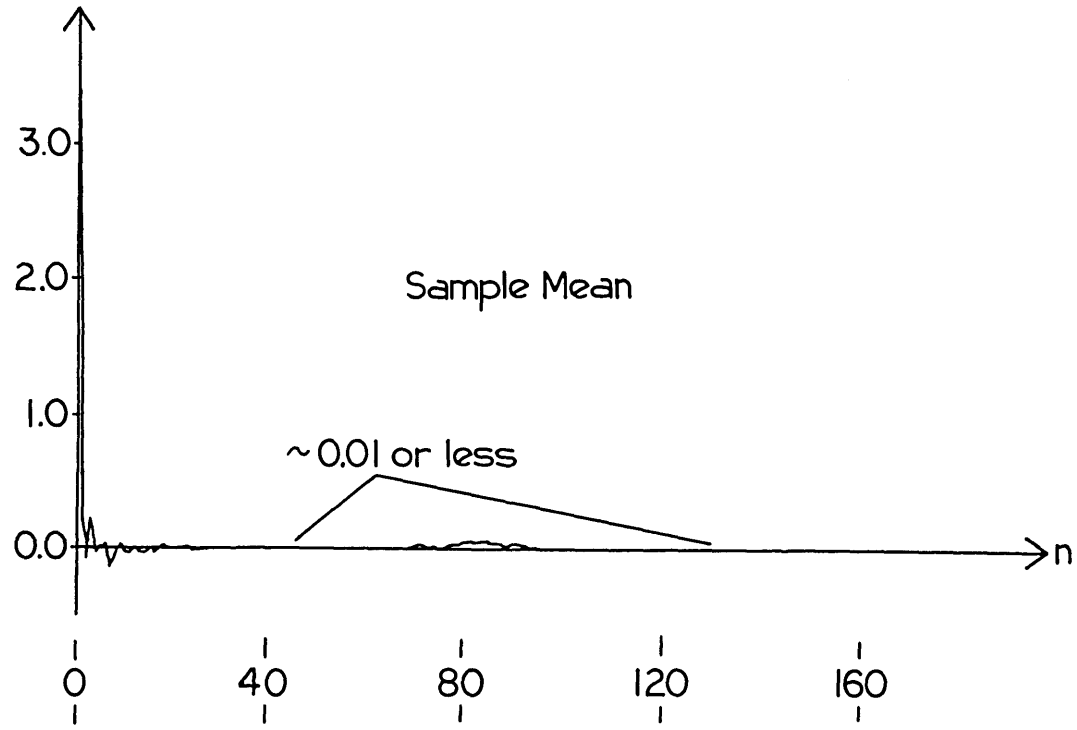
(b)

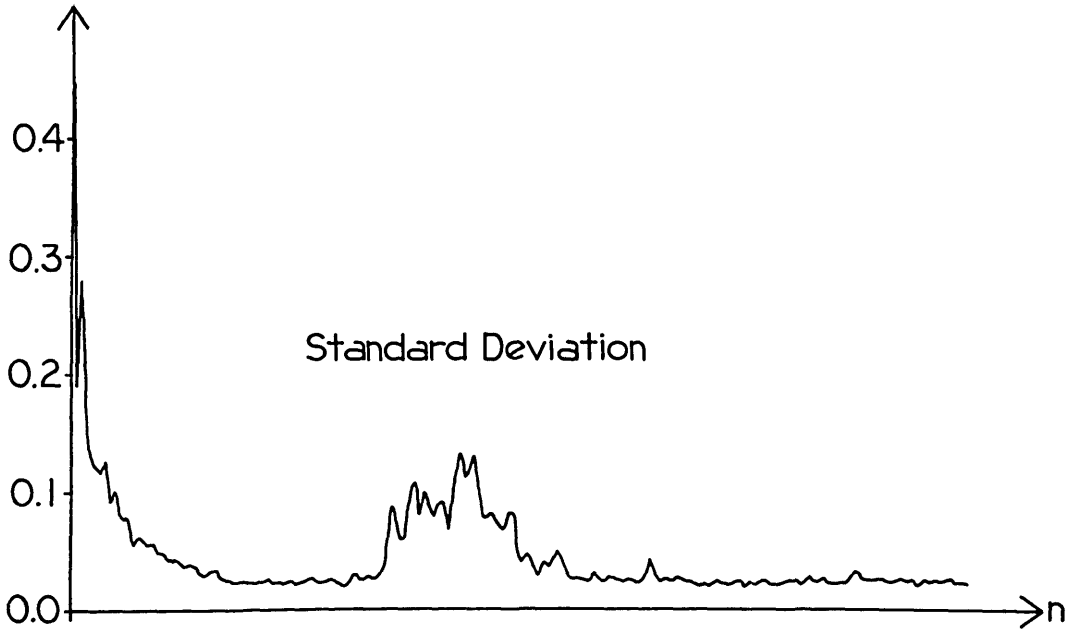

a_1
1.0
0.5
0.0
-1.0
1.0
$\frac{|\hat{P}_{1k}|-|\hat{P}_{2k}|}{\sigma_{\hat{s}_w}}$

(c)

Figure 29

Experimentally-determined cepstral mean and standard deviation as a function of n, as computed from an ensemble of 204 cepstra derived from the sentence "May we all learn a yellow lion roar".


3.0
2.0
1.0
0.0
Sample Mean
~0.01 or less
n
0
40
80
120
160


0.4
0.3
0.2
0.1
0.0
Standard Deviation
n

time cepstra derived from the all-voiced sentence "May we all learn a yellow lion roar" (speaker PDB). This sentence contains a substantially wide range of voiced sounds. The sample mean and standard deviation for each point in the cepstrum are given by

$$\underset{\text{Mean}}{\text{Sample}} = \eta_{\hat{s}_w}(n) = \frac{1}{204}\sum_{k=1}^{204} \hat{s}_{kev}(n)$$

$$\underset{\text{Deviation}}{\text{Standard}} = \sigma_{\hat{s}_w}(n) = \left\{\frac{1}{204}\sum_{k=1}^{204}[\hat{s}_{kev}(n) - \eta_{\hat{s}_w}(n)]^2\right\} \quad ,$$

where $\hat{s}_{kev}(n)$ is the n^{th} point (sample) of the k^{th} cepstrum of the ensemble. The fact that the mean is nearly zero for all but very low values of n is computationally advantageous, because this implies that a special array or register for storage of the mean may be unnecessary. The "hump" in the standard deviation for $n \simeq 65$-100, of course, is due to the occurrance of pitch peaks in that range (sampling rate was 10 KHz).

Unfortunately, time constraints prevented the performance of further experiments to determine the feasibility and effectiveness of the weighted averaging method of filtering described here. This is suggested as an area for future research.

VI. Final Summary and Conclusions

The main findings of this research are listed below.

(1) (Section III) The cepstrum appears to be more practically suited to actual implementation of homomorphic dereverberation than the complex cepstrum.

(2) (Section III) A feasible solution to the problem of voiced (and perhaps unvoiced) speech resynthesis from the cepstrum is the pitch-synchronous synthesis technique of Oppenheim's homomorphic vocoder.[7]

(3) (Section III) Key influences upon the quality of resynthesized speech seem to be the validity of the speech-production model upon which the synthesis procedure is based, the method of realization of the model, the effects of sectioning the input waveform, the "artificial phase" used in resynthesis, and the truncation of the cepstrum. A possible alternative to homomorphic vocoder resynthesis, which has produced slightly rough quality as implemented, is minimum-phase formant synthesis as described by Schafer and Rabiner.

(4) (Section IV) Weighting of input speech sections by a Hamming window improves the resolution of reverberation peaks in the cepstrum, increasing the ability of the cepstral filtering process to remove these peaks. Because a Hamming window also improves the quality of synthesized speech, this fact contributes to the compatibility of the synthesis method with dereverberation requirements.

(5) (Section IV) A technique useful for identifying reverberation peaks in the cepstrum is to locate large peaks in the difference of two cepstra, computed from differently reverberated versions of the same

speech waveform. A peak in this difference indicates a corresponding reverberation peak in one of the cepstra; thus removal of these peaks from the average of the cepstra leaves a result approximating the cepstrum of the unreverberated speech waveform. The primary benefit obtained from this filtering technique is elimination of confusion between cepstral pitch peaks and reverberation peaks in the pitch-detection process.

(6) (Section IV) Reverberation effects are essentially removed from artificially-reverberated waveforms when their cepstra are filtered by the above process. However, except in possibly a few cases (as the 1.5 msec echo discussed in Section IV), the filtering method is not effective for echo delay times less than about 3 msec, due to the increase in cepstral distortion caused by removal of reverberation peaks located in the 0-3 msec range of the cepstrum. Also, an upper limit of 10-15 msec is placed upon the duration of the reverberating impulse train by the length of the input sectioning window.

(7) (Section IV) Differential delays of up to 7.5 msec between the input reverberated waveforms were not found to degrade the quality of dereverberated speech when the cepstrum was used in processing.

(8) (Section V) In the filtering process described above, it is the average of two reverberated cepstra which is filtered. By averaging the cepstra before filtering, it is possible that some "good" information is lost. Stated in a different way, the identification of reverberation peaks by formation of the difference of two cepstra does not, by itself, indicate which cepstrum contains a given reverberation peak. As a result, to insure that all reverberation peaks are eliminated, both cepstra must be overfiltered. A possibly improved method is proposed

which utilizes the mean and standard deviation of the speech component of the cepstrum, along with reverberation peak estimates from past cepstra, to associate each reverberation peak with a particular cepstrum. Then a weighted average of the cepstra is computed, emphasizing the least-distorted portions of each cepstrum and suppressing the most-distorted parts. No experimental speech-processing results are available for evaluation of this method.

Important Conclusions Relating to Future Investigations

If a homomorphic dereverberation process is to be ultimately applied in Speakerphone and/or Conference Room Telephony Systems, improvements in the quality of resynthesized, processed speech will have to be accomplished.

With respect to true evaluation of the effectiveness of cepstral filtering for removal of reverberation, further experiments with naturally-reverberated speech will have to be conducted. Although the artificial reverberation used was severe in the sense that it produced great spectral distortion of the speech and large reverberation peaks in the cepstrum, an important observation is that these peaks have been highly-resolved and easily-identifiable. This may not be so for natural reverberation, complicating the cepstral filtering problem. In addition, the effects of dereverberation processing upon unvoiced speech and female voiced speech are yet to be ascertained.

BIBLIOGRAPHY

1. Mitchell, O. M. M., and D. A. Berkley, "Reduction of Long-Time Reverberation by a Center-Clipping Process," Journal of the Acoustical Society of America, Volume 47, No. 1, Part 1, page 84 (Abstract), January, 1970.

2. Flanagan, J. L., and R. C. Lummis, "Signal Processing to Reduce Multipath Distortion in Small Rooms," Journal of the Acoustical Society of America, Volume 47, No. 6, Part 1, pages 1475-1481, June, 1970.

3. Schafer, R. W., "Echo Removal by Discrete Generalized Linear Filtering," M.I.T. Research Laboratory of Electronics Technical Report No. 466, February 28, 1969.

4. Flanagan, J. L., "Cepstrum Method for Reducing Multipath Distortion in Rooms," Unpublished work.

5. Oppenheim, A. V., "Superposition in a Class of Nonlinear Systems," M.I.T. Research Laboratory of Electronics Technical Report No. 432, March 31, 1965.

6. Fant, C. G. M., Acoustic Theory of Speech Production, Mouton and Co., 's-Gravenhage, The Netherlands, 1960.

7. Oppenheim, A. V., "Speech Analysis-Synthesis System Based on Homomorphic Filtering," Journal of the Acoustical Society of America, Volume 45, No. 2, pages 458-465.

8. Oppenheim, A. V., and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Transactions on Audio and Electroacoustics, June, 1968, pages 221-226.

9. Schafer, R. W., and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," Journal of the Acoustical Society of America, Volume 47, Part 2, pages 634-648, February, 1970.

10. Gold, B., and C. M. Rader, Digital Processing of Signals, McGraw-Hill Book Co., 1969.