achieved in lip reading. When the tactile information is used in combination with lip reading, the ability to identify spoken words is considerably increased. For example, in one measurement of discrimination among 12 words, the lip reading response was about 60% correct. When supplemented by the tactile information, the response increased to 85% (PICKETT).

As in the visible speech method, the vocoder apparatus for tactile display is relatively complex. A much simplified tactile device is shown in Fig. 7.29 (KRINGLEBOTN). This device employs only five vibrators

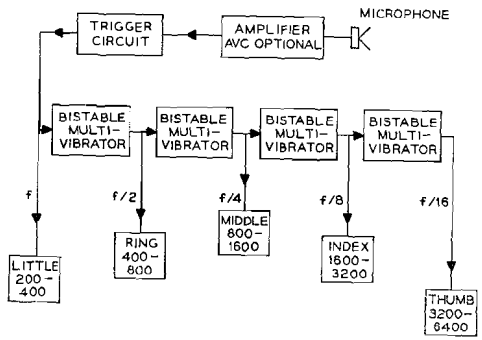


Fig. 7.29. A frequency-dividing tactile vocoder. (After KRINGLEBOTN)

applied to one hand. No filters are used, but stages of frequency division are arranged to divide five frequency ranges of the speech signal so as to vibrate the fingers individually. The vibrations on each finger are felt most strongly in the frequency range 200 to 400 cps. Because of the successive frequency divisions, this sensitivity range corresponds to successively higher frequency ranges in the input signal when distributed over the fingers, going from little finger to thumb. This method probably transmits some frequency information about the speech signal in terms of tactile frequency and other frequency information in terms of tactile location. Training tests with this system have been carried out with deaf children (KRINGLEBOTN).

A number of other efforts in kinesthetic and tactile communication are in progress. Although many of these aim toward machine aids for the blind rather than for the deaf, the presentation of sensory information involves problems common to both areas (BLISS; LINVILL).

7.83. Low Frequency Vocoder

The conventional electronic hearing aid is an amplifying and frequency shaping device. It facilitates the use of whatever residual

hearing a deafened person may have. In severe cases, however, the residual hearing is often confined to a very small bandwidth, usually at the low-frequency end of the audible spectrum. For example, a typical audiogram might show 60 to 80 db loss from 30 to 400 cps and 110 db above 500 cps.

One proposal is to make maximal use of such residual hearing. Slowly varying signals that describe the short-time speech spectrum (such as vocoder channel signals) are modulated either onto sinusoidal carriers of very low frequency, or onto distinctive complex signals of relatively small bandwith (PIMONOW). In one implementation, seven spectrum channels extending to 7000 cps are used. The rectified, smoothed outputs amplitude modulate the same number of low-frequency, sinusoidal carriers. The carriers are spaced from 30 to 300 cps. The modulated carriers are summed and presented as an auditory signal. In an alternative arrangement, the modulated signals are non-sinusoidal and include a low-frequency noise band, a periodic pulse train, and a band of actual speech. In one series of experiments, deafened subjects who could not use ordinary hearing aids apparently learned to discriminate well among a limited ensemble of words (PIMONOW).

Various devices for spectrum shifting, transposing or dividing have also been considered (JOHANSSON; GUTTMAN, and NELSON; LEVITT and NELSON). These devices generally aim to recode high-frequency information into a lower-frequency range where residual hearing exists. Like visible speech displays, their value appears to lie more in articulatory training than in speech reception. Like the other sensory aids discussed in this section, frequency scaling devices are still in the research stage. Extended experimentation and technical development will determine their potential as practicable aids to hearing.

VIII. Systems for Analysis-Synthesis Telephony

The discussions in Chapters III and IV considered the basic physics of the mechanisms for speech production and hearing. The topics of Chapters V, VI and VII set forward certain principles relating to the analysis, artificial generation, and perception of speech. The present and final chapter proposes to indicate how the foregoing results, in combination, may be applied to the efficient transmission of speech.

Efficient communication suggests transmission of the minimum information necessary to specify a speech event and to evoke a desired response. Implicit is the notion that the message ensemble contains only the sounds of human speech. No other signals are relevant. The basic problem is to design a system so that it transmits with maximum efficiency only the perceptually significant information of speech.

One approach to the goal is to determine the physical characteristics of speech production, perception and language and to incorporate these characteristics into the transmission system. As such, they represent information that need not be transmitted. Ideally, the characteristics are described by a few independent parameters, and these parameters serve as the information-bearing signals. Transmission systems in which a conscious effort is made to exploit these factors are generally referred to as *analysis-synthesis* systems.

In the ideal analysis-synthesis system, the analysis and synthesis procedures are presumably accurate models of human speech production. To the extent this is true, the resulting signal is coded and synthesized in a distortionless form. Additional economies in transmission can accrue from perceptual and linguistic factors. The pure analysis-synthesis system therefore has the greatest potential for bandsaving, and its analysis and synthesis processings typically require complex operations.

In contrast, other transmission systems aim for modest or little bandsavings, with terminal apparatus which is simple and inexpensive. Such systems typically exploit fewer properties of speech, hearing and language than do the pure analysis-synthesis systems. Nevertheless, they are of considerable interest and importance, and their potential applications range from mobile radio and scatter links to various commercial wire circuits. Although emphasis in this chapter is given to analysis-synthesis techniques, systems of the latter category are also brought in for discussion, especially in the context of digital coding and transmission.

The results of Chapter III and VI showed that speech signals can be described in terms of the properties of the signal-producing mechanism, that is, the vocal tract and its excitation. This characterization suggests important possibilities for efficient encoding of speech. In fact, it forms the common basis for a large class of bandwidth-compression systems. The idea is schemetized in Fig. 8.1. Three operations are involved. First, the automatic analysis of the signal into quantities that describe the vocal excitation and mode structure; second, the multiplexing and transmission of these parameters; and finally, the reconstruction of the original signal from them.

In a parallel manner, the discussion in Chapter IV suggested that the ear performs a kind of short-time frequency analysis at its periphery. The analysis includes a mechanical filtering, the equivalent of a rectification, and a neural encoding which—apparently at an early stage—involves an integration. In such a process, certain details of the original speech wave are lost and are not perceptually significant. Presumably a transmission system might also discard this information without noticeably influencing the preceived signal. It might thereby effect an economy in requisite channel capacity.

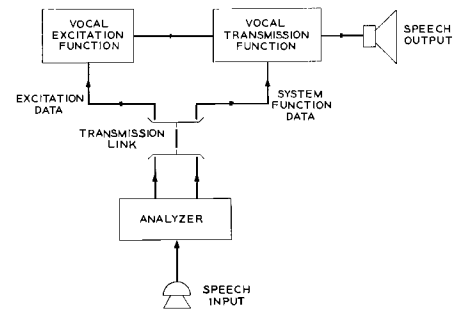


Fig. 8.1. Source-system representation of speech production

In a similar fashion, other aspects of the signal—for example, the sequential constraints on the sounds of a given language, or the natural pauses in connected speech—might be used to advantage. In short, practically all aspects of speech production, hearing and language have relevance to analysis-synthesis telephony. The following sections propose to discuss complete analysis-synthesis systems, and a number of these factors will be put in evidence.

8.1. Channel Vocoders

Analysis-synthesis telephony came of age, so to speak, with DUDLEY's invention of the Vocoder. This was more than three decades ago. In recent years, the name Vocoder (for *Voice Coder*) has become largely a generic term, commonly applied to analysis-synthesis systems in which the excitation and system functions are treated separately (see Fig. 8.1). The original Vocoder—now referred to as a spectrum channel vocoder—has probably been described in the literature more times than any other single system. Nevertheless, for the sake of completeness, as a convenient point of departure, and because it set forth, at such an early time, an important philosophy in voice transmission, a brief description of the idea will be repeated once more.

Following the coding scheme illustrated in Fig. 8.1, the Vocoder incorporates one important constraint of speech production and one of perception. It recognizes that the vocal excitation can be a broad-spectrum, quasi-harmonic sound (voiced), or a broad-spectrum, random signal (unvoiced). It also recognizes that perception, to a large degree, is dependent upon preservation of the shape of the short-time amplitude spectrum. A block diagram of an early Vocoder is shown in Fig. 8.2 (DUDLEY, 1939b).

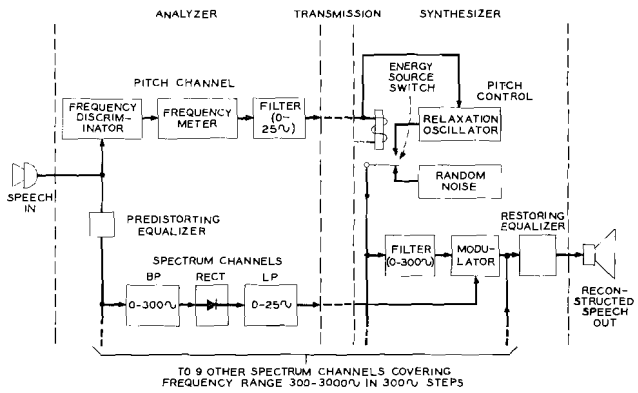


Fig. 8.2. Block diagram of the original spectrum channel vocoder. (After DUDLEY, 1939b)

The excitation information is measured by the top branch of the circuit. A frequency discriminator and meter measure the fundamental component of the quasi-periodic voiced sounds. Values of the fundamental frequency and its temporal variations are represented by a proportional electrical voltage from the meter. This "pitch" signal is smoothed by a 25 cps low-pass filter. Unvoiced sounds normally have insufficient power in the fundamental frequency range to operate the frequency meter. Nonzero outputs of the pitch meter therefore indicate voicing as well as the value of the pitch.

Ten spectrum channels in the lower part of the circuit measure the short-time amplitude spectrum at ten discrete frequencies. Each channel includes a band-pass-filter (300 cps wide originally), a rectifier and a low-pass filter (25 cps). The measured spectrum is therefore precisely that described in Section 5.1, Chapter V. The predistorting equalizer pre-emphasizes the signal to produce nearly equal average powers in the spectrum-analyzing filters. The spectrum-defining channel signals consequently have about the same amplitude ranges and signal-to-noise ratios for transmission. The eleven 25-cps wide signals occupy a total bandwidth of less than 300 cps and must be multiplexed in frequency or time for transmission.

At the receiver, the speech spectrum is reconstructed from the transmitted data. Excitation, either from a pitch-modulated, constant average power pulse-generator, or from a broadband noise generator, is applied to an identical set of band-pass filters. The outputs from the filters are amplitude modulated by the spectrum-defining signals. A short-time spectrum, approximating that measured at the transmitter, is recreated.

With proper design the synthesized speech can be made surprisingly intelligible. An example of speech transmitted by a 15-channel vocoder is shown by the spectrograms in Fig. 8.3. Important features such as formant structure and voiced-unvoiced excitation are relatively well preserved.

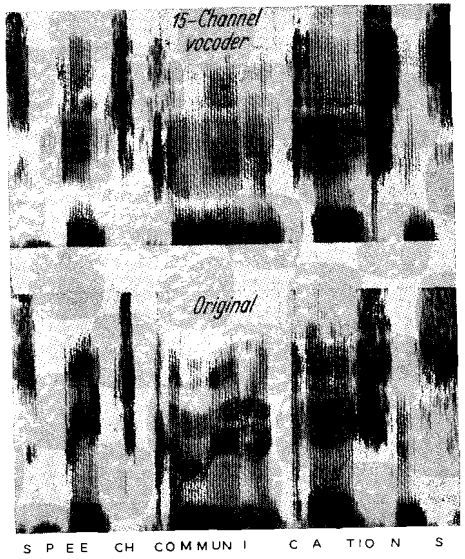


Fig. 8.3. Spectrogram of speech transmitted by a 15-channel vocoder

8.11. Design Variations in Channel Vocoders

Since the original development of the Vocoder many different versions and variations have been constructed. Number and spacing of the analyzing filters along the frequency scale, their bandwidths, degree of overlap, and selectivity are all matters of design variation. Similarly, many different pitch extraction and voiced-unvoiced detection circuits have been examined, as well as the characteristics of the rectifier and low-pass filter. The number of channels used has ranged from as few as eight to as many as 100, and the filter characteristics have ranged from broad, steep, flat-topped responses to narrow, simple-tuned circuits. Space does not permit a detailed discussion of all these investigations. However, typical analog hardware implementations include those of R. L. MILLER, 1953; DAVID, 1956; VILBIG and HAASE, 1956a, b; SLAYMAKER; SHEARME, and HOLMES; COOPER, PETERSON, and FAHRINGER; WERNER and DANIELSSON; YAGGI; YAGGI and MASON; and STEELE and CASSEL, 1963a, b. Digital implementations have been nearly equally

varied, and include the work of GOLDEN; FREUDBERG *et al.*; GOLD; and GOLD and RADER. In particular, Fast Fourier Transform techniques have been found advantageous in digital implementations.

Although intelligibility may be high, practical realizations of conventional channel vocoders generally exhibit a perceptible degradation of speech naturalness and quality. The synthetic speech possesses a machine-like quality which is characteristic of the device. Several factors seem to be responsible. One is the coding of excitation data. Voiced-unvoiced discriminations often are made with noticeable errors. Relevant structure in the pitch signal may not be preserved, and, under certain conditions, octave errors may be made in the automatic pitch extraction. Voiced sounds are synthesized from a pulse source whose waveform and phase spectrum do not reflect certain details and changes of the real vocal cord wave. The spectral analysis also has a granularity, or lack of resolution, imposed by the number, bandwidth and spacing of the analyzing filters. A given speech formant, for example, might be synthesized with too great a bandwidth. Further, the large dynamic range of the amplitude spectrum may not be covered adequately by practical rectifiers and amplifiers.

The basic channel vocoder design can be improved in several ways. The important excitation problems can be obviated to a large extent by the voice-excitation technique to be discussed in a following section. Also sophisticated pitch extraction methods, such as the cepstrum method described in Section 5.3, Chapter V, provide more precise pitch and voiced-unvoiced data. The spectral representation problems can be lessened by careful filter design, or by the use of digital techniques such as the Fast Fourier Transform.

8.12. Multiplexing Channel Vocoders

8.121. Frequency-Space Multiplexing. The customary techniques for transmitting a multiplicity of simultaneous signals are frequency-space multiplexing and time-division multiplexing. In the former, the requisite amount of spectrum bandwidth is allocated to each signal. The individual signals are modulated onto separate carriers, which are transmitted simultaneously within the allocated channels and are demodulated at the receiver. In the latter, the several signals time-share a single transmission path of appropriate bandwidth.

Frequency multiplexing of vocoder signals is attractive from the standpoint of circuit simplicity and existing analog communication links. Certain relations can be observed to conserve spectrum space and, at the same time, provide accurate transmission. Since the vocoder signals normally contain a dc component, the modulation method must

be chosen to preserve this response. Conventional double-sideband (DSB) amplitude modulation would satisfy the response requirement, but would not be economical of bandwidth. Conventional single-sideband (SSB) modulation with suppressed carrier, although taking only half the bandwidth, would not reliably preserve the low-frequency components of the modulation. Vestigial sideband transmission might suffice. However, a two-phase (or quadrature) modulation method has been advanced as the best solution (HALSEY and SWAFFIELD).

A pair of channel signals DSB modulate separate carriers of the same frequency but differing in phase by $\pi/2$ radians. The two double-sideband signals then occupy the same frequency band. Provided the transmission path has attenuation and phase characteristics symmetrical about the carrier frequency, either signal-complex can be rejected at the receiver by demodulating (multiplying and integrating) with a synchronous quadrature carrier. Frequency and phase synchrony of the carriers at the transmitter and receiver are of course critical.

The quadrature method is generally not satisfactory for transmission of conventional voice signals. Practical stabilities are such that the crosstalk between circuits cannot be kept low enough. For vocoder signals, however, a crosstalk attenuation between spectrum channels of about 25 db seems adequate[1]. This figure is within the practical limits of the quadrature method. The signal-to-crosstalk ratio is the cotangent of the phase error between the modulating and demodulating carriers. Therefore, a crosstalk attenuation of 25 db, or more, requires a phase error of about 3.3 degrees, or less.

8.122. Time-Division Multiplexing. Time-division multiplexing involves the transmission of sample values of the channel signals taken in time sequence. According to the sampling theorem, the rate of sampling must be at least twice the highest frequency contained in the channel signals. The vocoder signals are typically bandlimited to about 20 cps, hence sampling rates on the order of 40 sec^{-1}, or higher, are indicated. Practically, to provide adequate channel separation in the desampling (distributing) operation, a total transmission bandwidth about twice the sum of the input signals, that is, the same as for DSB frequency-multiplex, is required (BENNETT, 1941). Even then, the crosstalk between channels may be only marginally acceptable. For example, in a 12-channel system the signal-to-crosstalk ratio is only on the order of 20 db. Without further coding, therefore, this multiplexing method appears somewhat less attractive from the fidelity standpoint than the quadrature frequency-space multiplex. On the other hand, its simplicity, and the

[1] The pitch channel is more sensitive to crosstalk. For it, an attenuation on the order of 40 db is desirable.

possibility for analog smoothing of the spectral shape, make it of interest.

One vocoder developed on the time-multiplex principle is called the Scan Vocoder (VILBIG and HAASE, 1956a, b). It is illustrated in Fig. 8.4. One hundred spectrum channels, using high frequency (130 kc) magnetostriction rods as the filters, produce a short-time spectrum. The filter outputs are scanned at 30 sec^{-1} and the time-multiplexed spectral

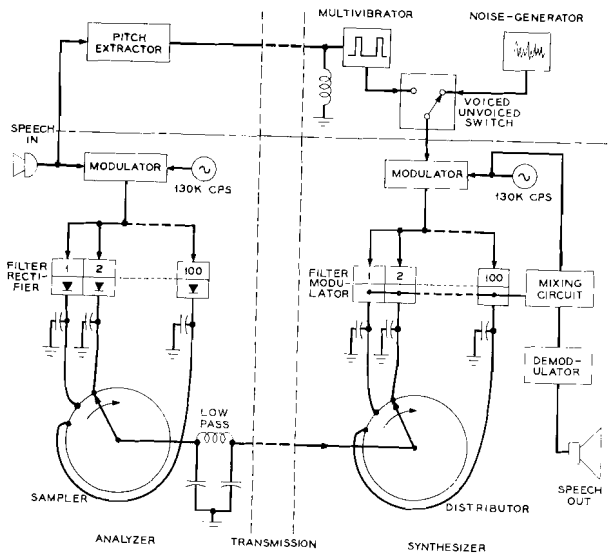


Fig. 8.4. Channel vocoder utilizing time-multiplex transmission. (After VILBIG and HAASE, 1956)

envelope is smoothed by a 200 cps low-pass filter. The envelope signal is demultiplexed by a synchronously scanning distributor at the receiver. The pitch information is transmitted in a separate channel.

8.123. Digital Transmission of Vocoder Signals. Transmission of signals in the form of binary pulses has a number of advantages. One is the possibility for repeated, exact regeneration of the signal. Noise and distortion do not accumulate as they do in analog amplification. Quality of the digital signal can, within limits, be made independent of transmission distance. Another advantage is the freedom to "scramble" the message in complex ways for secure or private communication. The price paid for these important advantages is additional transmission bandwidth. Time-divison multiplexing, coupled with pulse code modulation (PCM) of the channel signals, is consequently an attractive means

for vocoder transmission. The signal value in each sampled channel is represented by a sequence of binary pulses. The ordered and "framed" pulses are transmitted over a relatively broadband channel, synchronously distributed at the receiver, and reconverted from digital to analog form.

Although the digital signal requires comparatively greater bandwidth, the vocoded speech signal makes feasible full digital transmission over about the same bandwidth as normally used for nondigital conventional telephony. An important question is how many binary pulses are sufficient to represent each sample of the channel signals. The answer of course depends upon the quality of received signal that is acceptable. Current technology has used pulse rates from 1200 to 4800 bits/sec in particular applications (YAGGI and MASON). A typical design, for example, uses 18 spectrum channels which are sampled at 40 sec^{-1} and which are normalized in amplitude. The number of binary digits used to specify the sampled values of channels 1 through 14 is three bits; for channels 15 through 18, two bits; for the over-all amplitude level, three bits, and for pitch and voiced-unvoiced indication, seven bits. Therefore, 60 bits are included in one scan or "frame", and 2400 bits/sec is the transmitted data rate. Numerous variations in the design for digital transmission can be found.

8.13. Vocoder Performance

Although voice quality and naturalness normally suffer in transmission by vocoder, the intelligibility of the synthesized speech can be maintained relatively high, often with a vocoder having as few as ten

Table 8.1. *Consonant intelligibility for a vocoder. Percent of initial consonants heard correctly in syllables (logatoms).* (After HALSEY and SWAFFIELD)

b — 90%	l — 97%	r — 100%	w — 100%
f — 74	m — 85	s — 94	sh — 100
h — 100	n — 99	t — 91	th — 43
k — 85	p — 77	v — 96	none — 70

channels. For a high-quality microphone input and a fundamental-component pitch extractor, typical syllable intelligibility scores for a ten-channel (250 to 2950 cps) vocoder are on the order of 83 to 85 per cent (HALSEY and SWAFFIELD). Typical intelligibility scores for initial consonants range over the values shown in Table 8.1.

Weak fricatives such as *th* are not produced well in this system. The 30 per cent error indicated for no initial consonant (i.e., for syllables beginning with vowels) indicates imprecision in the voiced-unvoiced

switching. Such syllables were heard as beginning with consonants when in fact they did not. Even so, the consonant intelligibilities are reasonably good.

Comparable performances can also be obtained when the vocoder signals are time-sampled (scanned), quantized and encoded in terms of binary pulses. An early model 10-channel vocoder, arranged for digital transmission, gave typical consonant intelligibility scores shown in Table 8.2. The data rates are exclusive of pitch information. Four different quantizing levels were used (R. L. MILLER and D. K. GANNETT, unpublished; quoted in DAVID, 1956).

Table 8.2. *Vocoder consonant intelligibility as a function of digital data rate.* (After DAVID, 1956)

	Number of quantizing levels			
	6	5	4	3
Binary pulse rate (bits/sec)	1 300	1 160	1 000	788
Consonant intelligibility (%)	82	79	79	69

More elaborate designs provide somewhat higher intelligibilities. For example, a consonant intelligibility of approximately 90 per cent is typical of a 16-channel vocoder whose channel signals are sampled 30 sec^{-1} and quantized to three bits (i.e., 1440 bits/sec) (DAVID, 1956).

8.2. Reduced Redundancy Channel Vocoders

It is generally recognized that vocoder channel signals are not completely independent, and that possibilities exist for further processing the signals to orthogonalize them. Several investigations have considered methods for further eliminating redundancy.

8.21. "Peak-Picker"

The results of the vocal-tract computations in Chapter III show that the values of the speech spectrum at adjacent frequency positions are closely related. In a vowel sound, for example, the entire vocal transmission spectrum is specified by the formant frequencies. Usually, therefore, the neighboring channel signals in a vocoder are strongly correlated. One transmission system, called a peak-picking vocoder, attempts to eliminate this dependence. It operates by transmitting a few—three to five—channel signals which at any instant represent local maxima of the short-time spectrum.

The circuitry employed is modeled upon that described for the formant-extracting system in Section 5.2, Chapter V. Inhibitory connections prevent two adjacent channels from being selected. The identities of the "picked" maximum channels and their amplitudes are signaled to a conventional 18-channel vocoder synthesizer. A pitch signal is also sent. Thus at any one time only a few channels of the synthesizer are activated. Intelligibility scores as high as 70 per cent are reported for nonsense syllables, and a digital transmission rate of about 1 000 bits/sec is estimated to be required (PETERSON and COOPER).

8.22. Linear Transformation of Channel Signals

A related approach attempts to discover the dependence among the channel signals and to eliminate this redundancy in a smaller number of signals (KRAMER and MATHEWS). For n channel signals, a set of m signals, where $m \leqq n$, are formed which are a linear combination of the original n. The coefficients of the linear transformation constitute an $(m \cdot n)$ matrix of constants. The transformation matrix is realized practically with an $(m \cdot n)$ array of fixed resistors. Decoding of the m signals to retrieve an approximation to the original n is also accomplished by a linear transformation, namely, the transpose of the $(m \cdot n)$ matrix. The coefficients of the transformation are obtained to minimize the mean square difference between the original n signals and the reconstructed n signals.

The technique was applied to the spectrum signals of a 16-channel vocoder (i.e., $n=16$). For a reduction to $m=6$, it was reported that the output was almost completely understandable, although quality was substantially less than that of the 16-channel vocoder. For $m=10$, the quality was judged to be better than existing, conventional 10-channel vocoders. In the latter condition, the additional saving in channel capacity is estimated to be in the ratio of 3 to 2.

Another related study used a Hadamard matrix transformation to reduce the redundancy among the channel signals of a 16-channel vocoder (CROWTHER and RADER). The Hadamard transformation produces unit-weight linear combinations of the channel signals. It therefore requires no multiplications, but only additions and subtractions. This technique, implemented digitally in a computer, was applied to two different 16-channel vocoders. The results showed that the quality provided by the vocoders when digitized for 4000 bits/sec could be retained in the Hadamard transformation for a data rate as low as 1 650 bits/sec. The Hadamard transformation is therefore suggested as a simple, useful means for improving the quality of low bit-rate vocoders (CROWTHER and RADER).

8.23. Pattern-Matching Vocoders

Another variation of the vocoder involves classification of the frequency *vs* amplitude spectral information of the channel signals into a limited number of discrete patterns (SMITH, 1957). In one such study (DUDLEY, 1958), spectral pattern are associated with phonetic units of speech. The sound analysis is carried out according to the pattern recognition scheme described in Section 5.5, Chapter V. At any instant, the best match between the short-time speech spectrum and a set of stored spectral patterns is determined. A code representing the matching pattern is signaled to a vocoder synthesizer, along with conventional pitch and voiced-unvoiced data. New information is signalled only when the phonetic pattern changes. At the receiver, a set of spectral amplitude signals, approximating the signalled pattern, are applied to the modulators of the synthesizer. The pitch signal supplies the appropriate excitation. Filter circuits are included to provide smooth transitions from one sound pattern to the next.

An early version of the device used a ten-channel vocoder and only ten stored patterns. It is illustrated in Fig. 8.5. The stored patterns corresponded to the steady-state spectra of four consonant continuants and six vowels (s, f, r, n, and i, ɪ, ɛ, ɑ, o, u, respectively). For one speaker (from whose speech the spectral patterns were derived), digits uttered in isolation were recognized by two listeners with scores of 97 and 99 per cent correct, respectively. On common monosyllables, however, the intelligibility fell to around 50 per cent. The addition of six more patterns increased the score by a small amount. The bandwidth required for transmission was only on the order of 50 cps, or around 60 times less

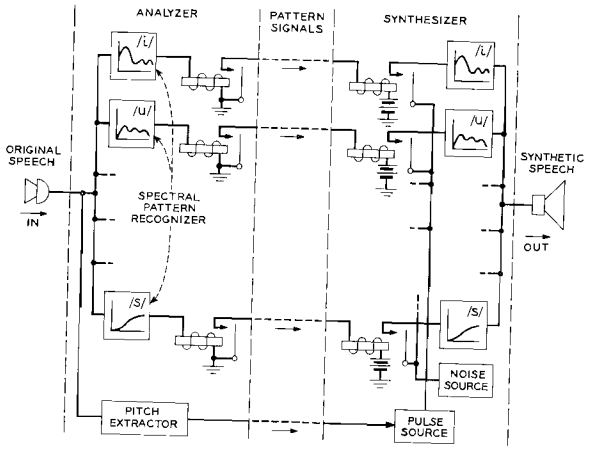


Fig. 8.5. Phonetic pattern-matching vocoder. (After DUDLEY, 1958)

than that for a conventional voice channel! While the intelligibility and quality of the speech processed by the device are clearly inadequate for most applications, the implementation does indicate the possibilities of narrow-band transmission for restricted message ensembles and limited speaker populations.

The obvious question suggested by the rather surprising performance with only ten stored patterns is how many stored spectral patterns would be needed to approach the performance of the conventional vocoder? At least one investigation has aimed to examine the question (SMITH, 1957, 1963). The outputs of the analyzer of a channel vocoder are sampled at 50 sec^{-1}, normalized in amplitude, and quantized. The digital description of the short-time spectrum is then compared to a large library of digital patterns stored in a rapid-access memory. No requirement is imposed that these spectral patterns correspond to specific phonetic units of speech. Using digital processing techniques, the best fitting pattern is selected and its code transmitted. The objective is to determine the smallest population of patterns necessary to meet given performance criteria. The processing cannot, of course, result in better speech quality than provided by the conventional vocoder. It may, however, afford a useful bandsaving beyond that the of channel vocoder. Digital data rates for the transmission of the spectral patterns and excitation are estimated to be on the order of 400 to 800 bits/sec (SMITH, 1957a, 1963).

8.3. Voice-Excited Vocoders

Despite their high potential for transmitting intelligible speech with bandwidth savings on the order of ten-to-one, or more, vocoders have been applied only in special communication situations. Little or no commercial use has been made, largely because speech quality and naturalness suffer in the processing[1]. The resulting synthetic speech tends to have a "machine accent", and its naturalness is less than that of a conventional voice circuit.

The seat of the difficulty is largely the extraction of excitation information—that is, the pitch measurement and the voiced-unvoiced discrimination. The difficult problem of automatic pitch extraction is well known. The device must faithfully indicate the fundamental of the voice over a frequency range of almost a decade (if male and female voices are to be handled) and over a large range of signal intensity. Practically, the pitch extractor must cope with unfavorable conditions where the speech signal may be produced in noisy and reverberant

[1] Other considerations include the cost of terminal equipment compared to the cost of bandwidth.

environments. In addition, the signal may suffer band limitation that eliminates the first several lowest harmonics, requiring that the fundamental frequency be generated from some non-linear operation. These difficulties are compounded by the human ear's ability to detect small imprecisions in pitch data. (See Section 7.24, Chapter VII.)

Some of the many approaches that have been made to the pitch extraction problem have been briefly outlined in Section 5.3, Chapter V. It suffices here to say that solutions are yet to be implemented to bring the quality of the spectrum channel vocoder up to the quality of conventionally-coded voice circuits. The same general remark applies to the voiced-unvoiced discrimination which is also signalled in the pitch channel.

One method for avoiding the difficulties inherent in automatic analysis of excitation data is the voice-excited vocoder (SCHROEDER and DAVID; DAVID, SCHROEDER, LOGAN, and PRESTIGIACOMO). In this device excitation information is transmitted in an unprocessed, subband of the original speech. At the receiving end, this baseband is put through a nonlinear distortion process to spectrally flatten and broaden it. It is then used as the source of excitation for regular vocoder channels covering the frequency range above the baseband. A block diagram of the arrangement is shown in Fig. 8.6.

The flattened excitation band reflects the spectral line structure of the quasi-periodic voiced sounds and the continuous spectral character of the unvoiced sounds. Because it is derived from a real speech band, it inherently preserves the voiced-unvoiced and pitch information. At some sacrifice in bandwidth, the overall quality of the processed signal can be made comparable to conventional voice circuits. A higher quality signal is therefore realized together with a part of the band-saving advantage of the channel vocoder.

In one implementation of the device the baseband is taken as 250 to 940 cps. The frequency range 940 to 3650 cps, above the baseband, is covered by 17 vocoder channels. The first 14 of these channels have

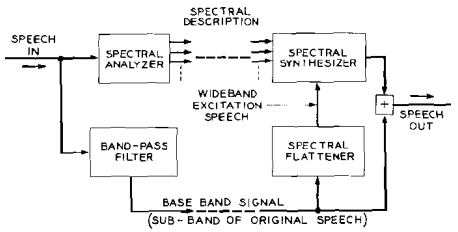


Fig. 8.6. Block diagram of voice-excited vocoder. (After DAVID, SCHROEDER, LOGAN and PRESTIGIACOMO)

analyzing bandwidths of 150 cps, and the upper three are slightly wider. The total transmission band occupancy is 1000 to 1200 cps, yielding a bandwidth compression of about three-to-one. The method of spectral flattening is shown in Fig. 8.7. The transmitted baseband is rectified and applied to the bandpass filters of the vocoder synthesizer. The filter outputs are peak-clipped to remove amplitude fluctuations. They are then applied as inputs to amplitude modulators which are controlled by the vocoder channel signals.

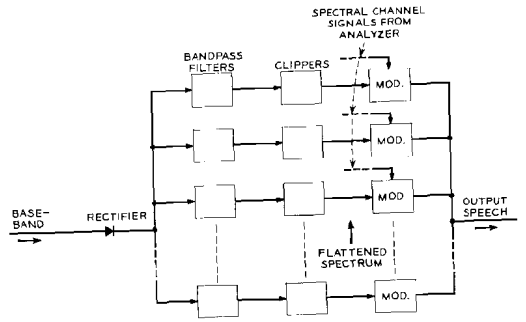


Fig. 8.7. Block diagram of spectral flattener. (After DAVID, SCHROEDER, LOGAN and PRESTIGIACOMO)

Intelligibility and speech quality tests, using speech from a carbon button microphone, were carried out to compare the voice-excited vocoder to telephone handset speech bandlimited to the same frequency range (DAVID, SCHROEDER, LOGAN, and PRESTIGIACOMO). To provide a more sensitive test, and to keep intelligibility substantially below 100%, masking noise was added to provide an 18 db speech-to-noise ratio. Phonetically balanced (PB) words were used in articulation tests (see Section 7.6, Chapter VII). For male speakers, the intelligibility of the voice-excited vocoder was found to be 6.1% less than the carbon-microphone speech of the same bandwidth. For female speakers the intelligibility was 10.1% less than the carbon-microphone speech.

Over-all speech quality of the voice-excited vocoder was assessed, along with that for three other transmission methods, by presenting listeners with sentences in isolation. The subjects were asked to rate each sentence "as good as normal telephone" or "worse than normal telephone". In 72% of the cases, the voice-excited vocoder was rated as good as normal telephone. In the same test, for comparison, a long distance carrier telephone circuit rated 82%, an 1800 cps lowpass circuit rated 36%, and a regular 18-channel vocoder rated 17%. The results

show the voiced-excited system to be better than the spectrum channel vocoder and to approach the quality of conventional voice circuits. Its application, as with similar methods, depends upon desired trade-offs between cost of terminal equipment, amount of bandsaving and signal quality.

8.31. Multiplexing and Digitalization

The problems in multiplexing the voice-excited vocoder are essentially similar to those discussed in Section 8.2 for the channel vocoder. The main difference is the unprocessed baseband. For economical transmission in a frequency multiplex system, it should be left unaltered or produced as a single sideband modulation. Transmission of the spectrum-defining channel signals can be the same in both cases.

One design of a voice-excited vocoder uses 500 cps of unprocessed baseband and 13 spectrum channels above the baseband (Howell, Schneider, and Stump, 1961a, b). The baseband is transmitted by single sideband modulation, and the channel signals are transmitted by vestigial sideband. Another analog implementation uses an unprocessed baseband of 250 to 925 cps and 10 vocoder channels covering the range to approximately 3000 cps (Golden, MacLean, and Prestigiacomo). The channel signals are double-sideband amplitude modulated onto 10 carriers spaced by 60 cps in the range 925 to 1630 cps. A bandwidth compression of approximately two-to-one is thereby realized.

Digital simulation and current computer techniques have also been used to design and study a complete voice-excited vocoder (Golden). To realize the digital simulation, the sampled-data equivalents of all filters and all circuits of an analog 10-channel voice-excited vocoder were derived (see, for example, Section 6.26, Chapter VI). Transformation of the continuous system into the sampled-function domain permits its simulation in terms of discrete operations which can be programmed in a digital computer. In the present instance, the entire vocoder was represented inside the computer, and sampled-quantized input speech signals were processed by the program.

The immense advantage that this technique offers for research and design of signal-processing systems cannot be overemphasized. The entire transmission system can be simulated and evaluated before constructing a single piece of hardware. The usual price paid is non-real time operation of the system. The time factor for the present simulation was 172 to 1, or 172 sec of computation to process one second of speech. However, as digital techniques develop and as computers become even faster, this time factor will shrink proportionately.

Another vocoder development has resulted in a time-multiplexed, fully digitalized voice-excited vocoder (Yaggi; Yaggi and Mason).

The device is designed to operate at a data rate of 9600 bits/sec and to use PCM encoding. The system operates with a baseband whose upper cutoff is, optionally, either 800 cps or 950 cps. For the former, 12 vocoder channels cover the range to 4000 cps; for the latter, 11 channels are used. The baseband signal is sampled at the Nyquist rate and quantized to 5 bits. The spectrum channels are sampled at 50 $\sec^{-1}$ (64 $\sec^{-1}$ for the 950 cps baseband); the lower three are quantized to 3 bits, and the higher ones to 2 bits. Amplitude normalization of the spectrum signals is also used. Comparable choices have been made in alternative digital implementations (Gold and Tierney).

Other coding techniques which, like the voice-excited vocoder, avoid the pitch tracking problem include the phase vocoder, the vobanc and the analytic rooter. These methods are discussed in later sections.

8.4. Correlation Vocoders

The channel vocoder demonstrates that speech intelligibility, to a large extent, is carried in the shape of the short-time amplitude spectrum. Any equivalent specification of the spectral shape would be expected to convey the same information. One equivalent description of the squared amplitude spectrum is the autocorrelation function. The correlation function can be obtained strictly from time-domain operations, and a spectral transformation is not required. Time-domain processing therefore offers simplicities in implementation. The relations linking these quantities have been discussed in detail in Section 5.1, Chapter V. A short-time autocorrelation specification of the speech signal might therefore be expected to be a time-domain equivalent of the channel vocoder.

In Chapter V, a short-time autocorrelation function of the function $f(t)$ was defined for the delay parameter, τ, as

$$\varphi(\tau, t) = \int_{-\infty}^{t} f(\lambda) f(\lambda+\tau) k(t-\lambda)\, d\lambda, \tag{8.1}$$

where $k(t)=0$ for $t<0$ and is a weighting function or time apperture [usually the impulse response of a physically realizable low-pass filter, see Eq. (5.15)]. Under the special condition $k(t)=2\alpha\, e^{-2\alpha t}=h^2(t)$, $\varphi(\tau,t)$ can be related to the measurable short-time power spectrum

$$\Psi(\omega, t) = |F(\omega, t)|^2,$$

where

$$F(\omega, t) = \int_{-\infty}^{t} f(\lambda) h(t-\lambda) e^{-j\omega\lambda}\, d\lambda. \tag{8.2}$$

In fact, it was shown that

$$\varphi(\tau, t) = \frac{e^{\alpha|\tau|}}{2\pi} \int_{-\infty}^{\infty} \Psi(\omega, t)\, e^{j\omega\tau}\, d\tau; \tag{8.3}$$

and

$$\Psi(\omega, t) = \int_{-\infty}^{\infty} e^{-\alpha|\tau|}\, \varphi(\tau, t)\, e^{-j\omega\tau}\, d\tau. \tag{8.4}$$

In this case the measurable short-time power spectrum—which is essentially the quantity dealt with in the channel vocoder (or rather the square root of it)—is the Fourier transform of the product of the weighting $e^{-\alpha|\tau|}$ and the short-time autocorrelation function $\varphi(\tau, t)$. The spectral information might therefore be specified in terms of the correlation. Several transmission methods for utilizing this relation have been examined (HUGGINS, 1954; SCHROEDER, 1959, 1962; KOCK; BIDDULPH).

One method for applying the principle is shown in Fig. 8.8. In the top branch of the circuit, the input speech is submitted to pitch extraction. This information is derived and employed in a manner identical to the channel vocoder. In the lower branch of the circuit, the input signal is put through a spectral equalizer which, in effect, takes the square root of the input signal spectrum. The basis for this operation is that the ultimate processed signal is going to be a correlation function whose Fourier transform is the power spectrum (or, the squared amplitude spectrum) of the input signal. Although spectrum-squared speech is generally quite intelligible, it has an unnatural intensity or stress variation. Since the spectrum squaring is inherent in the process, it is taken into account at the outset.

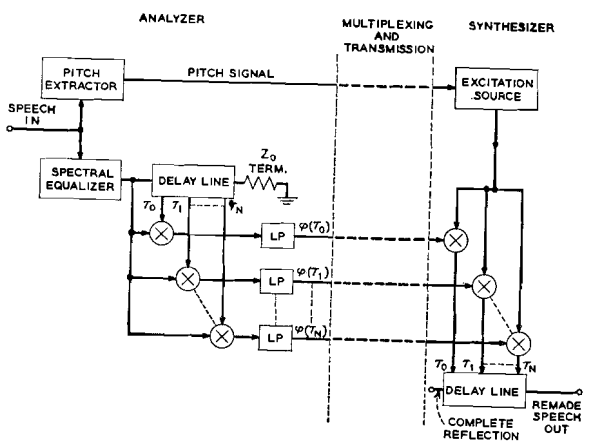


Fig. 8.8. Autocorrelation vocoder. (After SCHROEDER, 1959, 1962)

After spectral square-rooting, the short-time autocorrelation function of the signal is computed for specified delays. This is done by multiplying the appropriate output of a properly terminated delay line with the original input, and low-pass filtering the product (in this case with a 20-cps low-pass filter). The impulse response of the low-pass filter is the $k(t)$ as given in Eq. (8.1). Since the autocorrelation function is band-limited to the same frequency range as the signal itself, the correlation function is completely specified by sampling at the Nyquist interval (i.e., $1/2\,BW$). For a 3000 cps signal, therefore, a delay interval $\Delta\tau = 0.167$ msec is sufficient. The greatest delay to which the function needs to be specified, practically, is on the order of 3 msec (SCHROEDER, 1962). Thus a total of 18 delay channels—each using about 20 cps bandwidth—are required. The total bandwidth is therefore 360 cps and is about the same as required by the channel vocoder.

At the synthesizer, voiced sounds are produced by generating a periodic waveform in which the individual pitch period is the correlation function described by the values existing on the $n\tau$-channels at that instant. The waveform is generated by letting the pitch pulses of the excitation "sample" the individual τ-channels. The sampling is accomplished by multiplying the excitation and each channel signal. The samples are assembled in the correct order by a delay line, and are low-pass filtered to yield the continuous correlation function. Since the correlation function is even, the synthesized wave is made symmetrical about the τ_0 sample. This can be done practically with the delay line correctly terminated at its output end, but unterminated and completely reflecting at the far end, as shown in Fig. 8.8. Low-pass filtering of the samples emerging from the line recovers the continuous signal.

Because a finite delay is used in the analysis, the measured correlation function is truncated, and discontinuities will generally exist in the synthesized waveform. This leads to a noticeable distortion. The distortion can be reduced by weighting the high-delay correlation values so that they have less influence in the synthesized wave. The discontinuities are thereby smoothed, and the processed speech obtained approaches that from channel vocoders of the same bandwidth compression[1].

8.5. Formant Vocoders

The results of the acoustic analyses in Chapter III suggest that one efficient way to code speech is in terms of the vocal mode pattern. The

[1] This truncation distortion in synthesis can be avoided if the correlation data are used to control a recursive filter. See the technique devised for the Maximum Likelihood Vocoder (ITAKURA and SAITO, 1968) in Section 8.8.

results show, for example, that adjacent values of the short-time amplitude spectrum are not independent, but are closely correlated. In fact, specification of the complex poles and zeros is equivalent to specifying the spectrum at all frequencies. The formant vocoder aims to exploit this fact and to code the speech signal in terms of the mode pattern of the vocal tract. Because it does not use multiple control signals to describe strongly correlated points in the speech spectrum, the formant-vocoder hopes to achieve a band-saving in excess of that accomplished by the channel vocoder. The practicability of formant vocoders depends upon how well formant-mode data, or the equivalent, can be automatically derived. In addition, excitation information must be provided as in the channel vocoder.

A number of formant-vocoder systems have been designed and instrumented. Although it is not possible to treat each in detail, this section proposes to indicate typical circuit realizations and the results obtained from them.

Formant-vocoders generally divide into two groups—essentially defined by the synthesis philosophies set forth in Chapter VI. That is, the classification relates to the cascade and parallel connections of the synthesis circuits. The cascade approach strives to reconstruct the signal by simulating, usually termwise, the perceptually significant pole and zero factors of the vocal transmission. The complex frequencies of the poles and zeros, and the excitation data (pitch and voiced-unvoiced) are the coding parameters.

The parallel connection attempts to reconstruct the same signal in a different, but equivalent, way—namely, from information on the frequencies of the formants (poles) and their spectral amplitudes (residues). Ideally, the mode frequencies and their residues are specified in complex form. The complex residues are equivalent to specification of the spectral zeros. The discussion of Section 6.2, Chapter VI, has set down in some detail the relations between the cascade and parallel representations of the speech signal. If the requisite data for either synthesis arrangement can be obtained automatically and with sufficient accuracy, the formant vocoder has the potential for producing intelligible speech of perhaps better quality than that of the channel vocoder. Because it attempts to duplicate the vocal mode structure, it innately has the potential for a better and more natural description of the speech spectrum.

One of the earliest, complete formant-vocoder systems was a parallel arrangement (MUNSON and MONTGOMERY). It is illustrated in Fig. 8.9. At the analyzer, the input speech band is split into four subbands. In each band, the average frequency of axis-crossings,

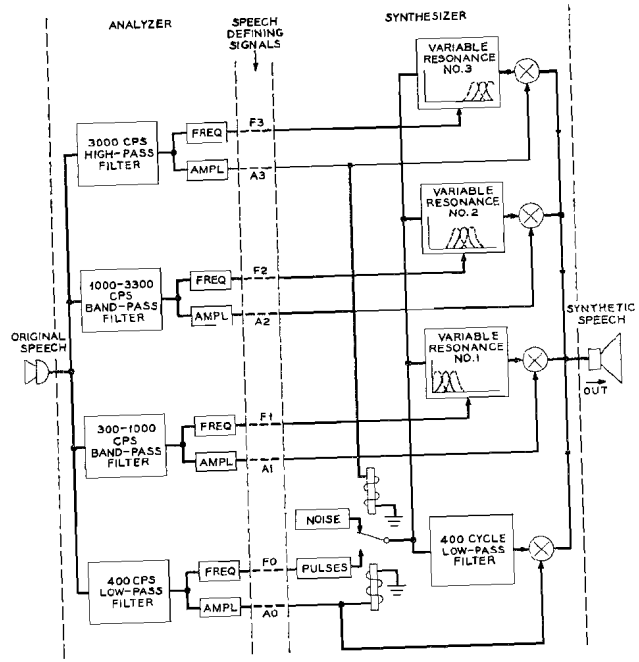


Fig. 8.9. Parallel-connected formant vocoder. (After MUNSON and MONTGOMERY)

F, and the average rectified-smoothed amplitude, A, are measured[1]. Signal voltages proportional to these quantities are developed. These eight parameters, which approximate the amplitudes and frequencies of the formants and of voicing, are transmitted to the synthesizer.

The synthesizer contains excitation circuitry, three variable resonators connected in parallel, and a fourth parallel branch with a fixed low-pass filter. Voiced (pulse) excitation of the parallel branches is signalled by the voicing amplitude, $A0$. The $A0$ control also determines the amplitude of the signal passing the fixed low-pass branch of the circuit. As in the channel vocoder, the frequency of the pulse source is prescribed by $F0$. Unvoiced (noise) excitation of the parallel branches is determined by amplitude $A3$. The amplitudes and frequencies of the three formant branches are continuously controlled and their outputs combined.

Intelligibility scores reported for the system were approximately 100% for vowel articulation and about 70% for consonant articulation.

[1] Note in this design the highest two bands normally contain more than a single formant. Their amplitude and frequency measures primarily reflect the most prominent formants in these ranges.

The total bandwidth occupancy of the eight control signals was about 300 cps, or about the same as for the channel vocoder. A number of different versions of parallel-connected formant vocoders have subsequently been constructed (for example, CHANG; CAMPANELLA; AYERS; STEAD and JONES; HOWARD). Two of these will receive further comment in the following section on digitalizing and multiplexing.

An early effort at realizing a cascade system also investigated the effects of severe band-limitation of the control signals (FLANAGAN and HOUSE). One synthesizer configuration considered in the study is shown in Fig. 8.10. The control data employed were pitch $F0$; amplitude of

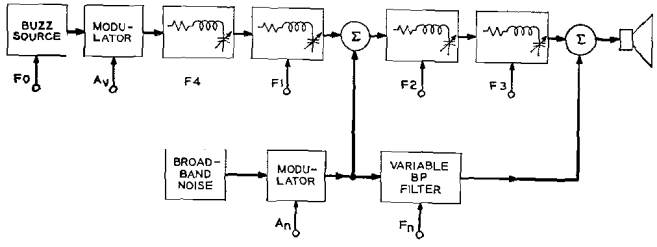


Fig. 8.10. Cascade-connected formant vocoder. (After FLANAGAN and HOUSE)

voicing AV; three formant frequencies $F1$, $F2$, $F3$ (covering the range approximately 100 to 3000 cps); a single, relatively-broad, fricative noise resonance FN (the major resonance in the range 3000 to 7000 cps); and the amplitude of noise excitation AN.

The formant frequency data were obtained from a peak-picking analyzer as described in Section 5.2, Chapter V. The amplitude of voicing was determined from the rectified-smoothed signal in a low-pass band of the original speech, and the amplitude of noise excitation was determined from the rectified-smoothed signal in the 3000 to 7000 cps band. Pitch was measured with a fundamental-extracting circuit, as in the channel vocoder. Each of the seven control signals was band-limited to slightly less than 10 cps by a low-pass filter, so that the total bandwidth occupancy was on the order of 60 cps.

All voiced sounds were produced by the upper resonator string of the circuit, following strictly the cascade approach. The unvoiced sounds were produced by a cascade-parallel connection which introduced zeros, as well as poles, into the transmission. Data on frequencies of zeros, as such, were not transmitted.

Although the band saving was high, detailed articulation testing of the system showed its performance to be relatively poor. In nonsense

monosyllables, the vowel articulation was on the order of 82%. For the consonants, the mean score was 27%. Confusion-matrix analysis of listener responses showed that voiced-unvoiced errors were few. Errors in discriminating voiced-stops and nasals, however, were relatively numerous, the synthesizer being congenitally incapable of simulating these sounds. In addition, errors in discriminating affricates and stops were due in large part to temporal imprecision resulting from the severe band-limitation of the control signals.

A more recent, digital computer simulation of an improved version of the synthesizer corrects some of the shortcomings (FLANAGAN, COKER, and BIRD). It provides for an additional pole-zero pair in the voiced branch and a controllable zero in the unvoiced branch (see Fig. 6.38 and Section 6.2, Chapter VI). When combined with a sophisticated digitally-simulated formant analyzer, the performance as a complete real-time formant vocoder is unusually good (COKER). The formant analysis in the computer is accomplished by a detailed matching of the real speech spectrum by a pole-zero model spectrum, similar to the analysis-by-synthesis procedure. (See Section 5.21.) The digital processing provides much greater accuracy than can be obtained with analog equipment. The precision in the formant tracking, and the more detailed accounting for system and excitation characteristics by means of the additional pole-zero pair, contribute significantly to the quality of the synthetic speech.

A further word may be appropriate concerning the relative merits of parallel versus cascade connections, and about the approach which may result in the most efficient and practical set of parameters. The vocal transmission for vowel sounds contains only poles. The residues in these poles are therefore functions only of the pole frequencies. Given the formant frequencies, any formant amplitude specification is redundant because the amplitudes are implied by the frequencies. In this case, the cascade synthesizer provides correct formant amplitudes automatically from formant frequency data alone. For nonvowel sounds the vocal transmission can have zeros, one or two of which may prove to be perceptually significant. To simulate these factors, the cascade synthesizer requires controllable antiresonances. Again, given the proper pole *and* zero frequencies, spectral amplitudes are automatically accounted for.

The parallel synthesizer, on the other hand, requires the significant pole frequencies and, ideally, the complex residues in these poles. The residues, in effect, specify the spectral zeros. The contribution to perception of the residue phases is modest but not negligible (FLANAGAN, 1965). (See Section 6.223.) A relevant question about formant synthesis is then "Which is easier to analyze automatically, the frequencies of spectral zeros or the amplitudes and phases of spectral maxima?" The question is complicated by one other matter—the excitation source.

What are its perceptually important characteristics? Are they easier to include in one model than in the other? At the present stage of study, the ultimate practical choice is not clear.

8.51. Multiplexing and Digitalization of Formant Vocoders

One real-time formant vocoder that has been given extensive tests is the parallel configuration shown in Fig. 8.11 (STEAD and JONES). Besides being tested in the unmultiplexed "back-to-back" connection, this system has also been examined in a fully digitalized version using time-division PCM techniques (STEAD and WESTON). The components of the system have several similarities with devices discussed previously. In one version, the synthesizer is based upon an earlier development (LAWRENCE, 1953). The formant-frequency extractor is based upon the peak-picking technique described in Section 5.2, Chapter V. The overall implementation and circuit design are unusually refined, and considerable effort is made to insure adequate dynamic range for the extraction of frequency and amplitude data. In the analog form, low-pass filters confine the eight control parameters to approximately 20 cps each, resulting in a total bandwidth occupancy of about 160 cps. Typical intelligibility scores for phonetically-balanced words and for relatively naive listeners are reported to average approximately 70%.

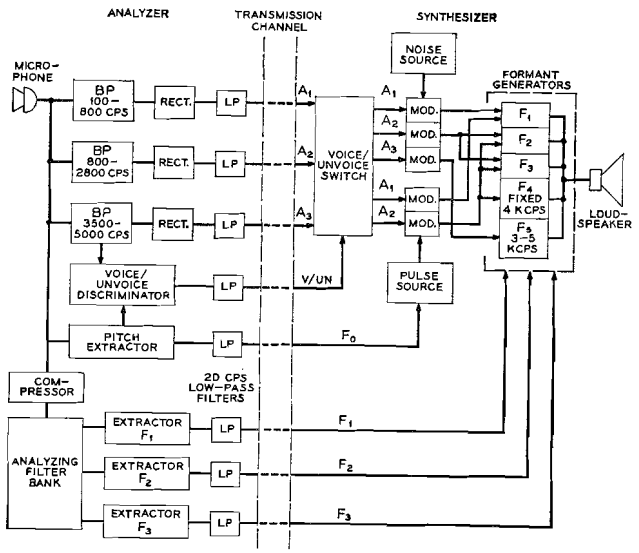


Fig. 8.11. A complete formant-vocoder system utilizing analog and digital transmission techniques. (After STEAD and JONES; STEAD and WESTON)

As mentioned earlier, the advantages of digital transmission are several. Not least is the ability to regenerate the signal repeatedly—essentially free of accumulating distortion. Problems in switching, time sharing and security are also amenable to straightforward solutions with the signal in digital form. One difficulty, however, is that transmission of the digital signal requires more bandwidth than does the analog form. For example, a 3000 cps speech band sampled at the Nyquist rate (6000 sec^{-1}) and quantized to 6 or 7 bits may require—without further coding—a bandwidth on the order of 50000 cps. If, through appropriate coding, the data rate could be brought down to the order of 1000 bits/sec, the digital signal could easily be transmitted over ordinary 3000 cps voice channels. The formant vocoder holds promise for providing such a coding.

In the formant vocoder of Fig. 8.11, the control parameters were band-limited to 20 cps. For digitalizing the control signals, however, a sampling rate of 32 sec^{-1} was found to be a safe working minimum. This rate suggests that the control parameters have little significant energy above about 16 cps. The amplitude quantization found acceptable for digitalizing the control data of this system is shown in Table 8.3.

Table 8.3. *Quantization of formant-vocoder signals.* (After STEAD and WESTON)

Parameter	Number of levels	Bits
$F1$:	16	4
$F2$:	16	4
$F3$:	8	3
$A1$:	8	3
$A2$:	8	3
$A3$:	8	3
V/UN:	2	1
$F0$:	64[a]	6
		27

[a] Estimated for linear coding of fundamental frequency.

In evaluating the digital transmission, 16 levels were thought too generous for the first formant frequency, but 8 levels were too coarse. For the three amplitude parameters, the 8 levels each were also thought too generous and that additional saving could be effected by coding the functions on a log-amplitude scale[1].

[1] These observations have been confirmed, extended and quantified in greater depth by computer-perceptual experiments on the bandlimitation and quantization of formant data in synthesis (ROSENBERG, SCHAFER and RABINER).

It is relevant to compare the practical figures of Table 8.3 with earlier estimates of the precision necessary in quantizing similar parameters (FLANAGAN, 1957b). The earlier estimates were based upon the just-perceptible changes which listeners could detect in the formant parameters of synthetic vowels (see Section 7.2, Chapter VII). The quantizing accuracy estimated to be necessary is given in Table 8.4.

In view of the limitations of the perceptual data on which the estimates are based, the correspondence with the numbers in Table 8.3 is surprisingly close. It suggests that psychoacoustic measures of the type discussed in Chapter VII might be judiciously applied with some confidence to estimate system performance.

Table 8.4. *Estimated precision necessary in quantizing formant-vocoder parameters*
The estimates are based upon just-discriminable changes in the parameters of synthetic vowels. (After FLANAGAN, 1957b)[a].

Parameter	Number of levels	Bits
F 1:	14	3.8
*F*2:	14	3.8
*F*3:	9	3.2
*A*1:	3	1.6
*A*2:	3	1.6
*A*3:	2	1.0
*F*0:	40	5.3
		20.3

[a] The amplitude parameters are considered to be logarithmic measures.

After sampling and quantizing, the data of Fig. 8.11 are PCM encoded for transmission. At a sampling rate of 32 sec^{-1}, the control data, exclusive of pitch, are specified with 672 bits/sec. A 6-bit pitch parameter produces a total rate of 864 bits/sec, a rate which could be transmitted by conventional 3000 cps channels. Although detailed testing has not been carried out, the digitally transmitted signal is reported to differ only slightly in intelligibility and quality from the analog connection. One interesting observation about the system is that the spectrum of the quantizing noise associated with digitalizing the control signals does not lie in the audio band. Rather, the noise corresponds to a sort of quasi-random uncertainty in the synthesis process. Its subjective effects have not been fully explored.

A preliminary study has considered the effects of digital errors in the PCM-encoded parameters of a formant vocoder (CAMPANELLA, COULTER, and IRONS). The system on which the tests were performed is similar to that shown in Fig. 8.11, except the voiced-unvoiced decision is transmitted by the pitch signal. The total bandwidth occupancy of the control signals is 140 cps. The formant parameters are quantized to 3 bits. Pitch is quantized to 5 bits. A 43.5 sec^{-1} scan rate in the time multiplexing produces a data rate of 1000 bits/sec. Under error free conditions, articulation scores of about 80% on PB words are claimed for this bit rate. A digital error rate of 3% degrades the articulation score by 15%. This impairment is found to be equivalent to reducing the signal-to-noise ratio of the analog parameters to 9.5 db.

8.52. Voice-Excited Formant Vocoders

The voice-excitation technique described in Section 8.3 has also been applied to a parallel-connected formant vocoder (FLANAGAN, 1960b). The circuit arrangement is shown in Fig. 8.12. In this implementation, a baseband of about 400 cps (300 to 700 cps) is transmitted in unprocessed form. Three formant vocoder channels cover the frequency range 800 to 3200, and the amplitude and frequency of three spectral maxima in this range are transmitted. Formant extraction is accomplished according to the maximum-picking technique described in Section 5.2, Chapter V. All control signals are low-passed to 17 cps. The total bandwidth occupancy is therefore slightly more than 500 cps.

At the synthesizer the baseband is spectrally broadened. It is peak-clipped, differentiated, half-wave rectified and used to trigger a one-shot multivibrator. The pulse output of the multivibrator provides the excitation source for the formant channels. Unvoiced sounds create shot noise from the multivibrator. Voiced sounds produce periodic pulse trains which sometimes may have more than one pulse per fundamental

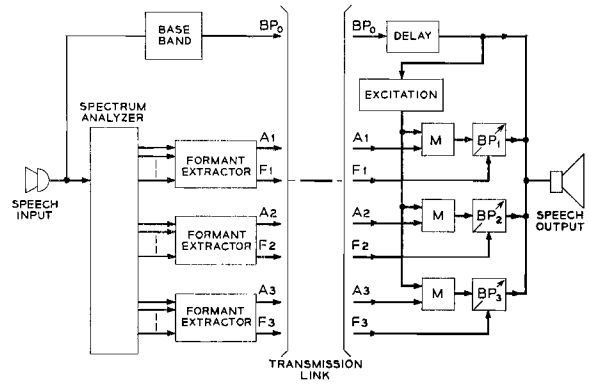


Fig. 8.12. Voice-excited formant vocoder. (After FLANAGAN, 1960b)

period. The technique generally provides an improvement in the quality and naturalness of the formant vocoder transmission. However, because the baseband is such a large percentage of the total bandwidth, it is almost as economical to use conventional vocoder channels above the baseband.

A related voice-excited technique uses the spectral shape of the first formant region to shape the second and third formants (DE JAGER, personal communication, 1961). A baseband about 300 to 800 cps is separated and transmitted in unprocessed form. In two other (formant) bands, 800 to 2000 cps and 2000 to 3200 cps, zero-crossing counters and rectifier-integrator circuits determine signals representing the amplitudes and frequencies of the formants. These four signals are low-passed to 40 cps each, and are sent with the baseband to the receiver.

The synthesizer reconstructs a spectrum in which the baseband (essentially the first formant) is produced in its original position. A second formant is synthesized in a separate parallel branch by heterodyning the baseband to the measured second formant frequency position. A third is generated in a similar fashion. The output speech is obtained by adding the three parallel branches in accordance with the measured amplitudes. The spectral components of the heterodyned bands generally become inharmonic, and the pitch frequency is preserved in them only to the extent of line spacing. Perceptually, the degradation of pitch information is less than might be expected, since the baseband is retained in original form with its correct line structure, and it is an effective masker.

8.6. Orthogonal Function Vocoders

One approach to describing a signal with the fewest independent parameters is to approximate the signal, in some sense, by a series of orthogonal functions. The coefficients of the expansion then become the information-bearing quantities. The orthogonal functions chosen for the representation presumably should capitalize upon some known characteristic of the signal.

The orthogonal function approach has been considered for describing both the speech waveform and the amplitude spectrum. A precise waveform description holds relatively small potential for bandwidth reduction – unless information such as phase is discarded and use is made of voiced-unvoiced and pitch tracking measurements. The spectral description , or its time-domain equivalent, promises more. The relationships between short-time spectral analysis and correlation analysis suggest techniques for efficient description of the speech spectrum.

8.61. Expansion of the Speech Waveform

A general method has been described in the literature for representing signal waveforms by orthogonalized, exponential functions (HUGGINS, 1957; KAUTZ). The method has been applied to the analysis of single pitch periods of voiced sounds (DOLANSKY, 1960). If $f(t)$ is a single pitch period, then the approximation

$$f(t) \simeq \sum_m c_m g_m(t) \tag{8.5}$$

is made, where the $g_m(t)$ are the set of orthogonalized, exponential functions. Their Laplace transforms of odd and even orders are given by

$$\begin{aligned} G_{2n-1}(s) &= \sqrt{2\alpha_n}\,\frac{s+|s_n|}{(s-s_n)(s-s_n^*)}\prod_{j=1}^{n-1}\frac{(s+s_j)(s+s_j^*)}{(s-s_j)(s-s_j^*)} \\ G_{2n}(s) &= \sqrt{2\alpha_n}\,\frac{s-|s_n|}{(s-s_n)(s-s_n^*)}\prod_{j=1}^{n-1}\frac{(s+s_j)(s+s_j^*)}{(s-s_j)(s-s_j^*)} \end{aligned} \tag{8.6}$$

where

$$s_n = (-\alpha_n + j\beta_n)\,.$$

The inverse transforms of Eq. (8.6) are

$$\begin{aligned} g_{2n-1}(t) &= \sum_{k=1}^{n}\frac{1}{\beta_k}\,|\mathscr{K}_{2n-1}(s_k)|\,e^{-\alpha_k t}\sin[\beta_k t - \vartheta_{2n-1}(s_k)] \\ g_{2n}(t) &= \sum_{k=1}^{n}\frac{1}{\beta_k}\,|\mathscr{K}_{2n}(s_k)|\,e^{-\alpha_k t}\sin[\beta_k t - \vartheta_{2n}(s_k)] \end{aligned} \tag{8.7}$$

where

$$\mathscr{K}_m(s_k) = \{G_m(s)\,[(s+\alpha_k^2)+\beta_k^2]\}_{s=s_k}$$

and

$$\vartheta_m(s_k) = \frac{\operatorname{Re}\mathscr{K}_m(s_k)}{|\mathscr{K}_m(s_k)|}\,.$$

The first two $g_m(t)$'s, therefore, are simple damped sinusoids which differ in amplitude and phase. The product-series components of $G_m(s)$ are seen to be all-pass functions. An n of 7 (or an m of 14) is considered to be adequate for the speech wave approximation (DOLANSKY). The critical frequencies s_n are fixed and are chosen to span the voice frequency range, typically in intervals of a few hundred cps.[1]

Assuming $f(t)$ is zero for $t<0$, and since

$$\begin{aligned} \int_0^\infty g_p(t)\,g_q(t)\,dt &= 1; \quad p=q \\ &= 0; \quad p \neq q\,, \end{aligned}$$

[1] A relevant question might inquire as to the potential of this technique if the s_n could be derived in an adaptive way; that is, if the s_n could be varied to match the signal.

the k-th coefficient of the orthonormal series is given by

$$c_k = \int_0^\infty f(t)\, g_k(t)\, dt. \tag{8.8}$$

One straightforward, but impractical, means for measuring the coefficients is apparent. Suppose the signal $f(t)$ is filtered with a realizable filter whose impulse response is $g_k(t)$, the result is

$$O(t) = \int_0^\infty g_k(\tau)\, f(t-\tau)\, d\tau. \tag{8.9}$$

If, however, the time-reversed signal $f(-t)$ is filtered, the result is

$$O(t) = \int_0^\infty g_k(\tau)\, f(t+\tau)\, d\tau. \tag{8.10}$$

The value $O(0)$, that is, the result at the instant when the time reversed $f(t)$ ends, is the value of c_k. This measurement, performed for all the $g_m(t)$'s, provides the requisite coefficients.

A perhaps more practicable, real-time application of the orthogonal function for speech waveform transmission is shown by the system in Fig. 8.13a (MANLEY and KLEIN). For voiced sounds the input speech is led to a pitch extractor which generates an impulse train at the fundamental frequency. These impulses produce the set of orthogonal functions $g_m(t)$ by exciting realizable networks having the functions as their impulse responses. Approximations to the coefficients of the series (8.5) are obtained by calculating.

$$c_k = \int_0^T g_k(t)\, f(t)\, dt, \tag{8.11}$$

where T is a given pitch period. The calculation is carried out by the multipliers, the reset integrators and the sample-and-hold elements shown in the diagram. The pitch pulses reset the integrators and trigger the sampling circuits to read and store the value of the integral at the end of period T. Before multiplexing and transmission, the pitch pulse-frequency is converted into an analog signal by a frequency meter, and the time-varying coefficients $c_1(t)$, $c_2(t) \ldots c_m(t)$ are further smoothed by low-pass filtering.

At the receiver, in Fig. 8.13b, the signal is reconstructed, pitch period by pitch period, according to Eq. (8.5). A pitch-modulated pulse generator excites an identical set of $g_m(t)$ networks and their outputs are respectively modulated by the $c_m(t)$ coefficients. The sum is an approximation to the original voiced sound.

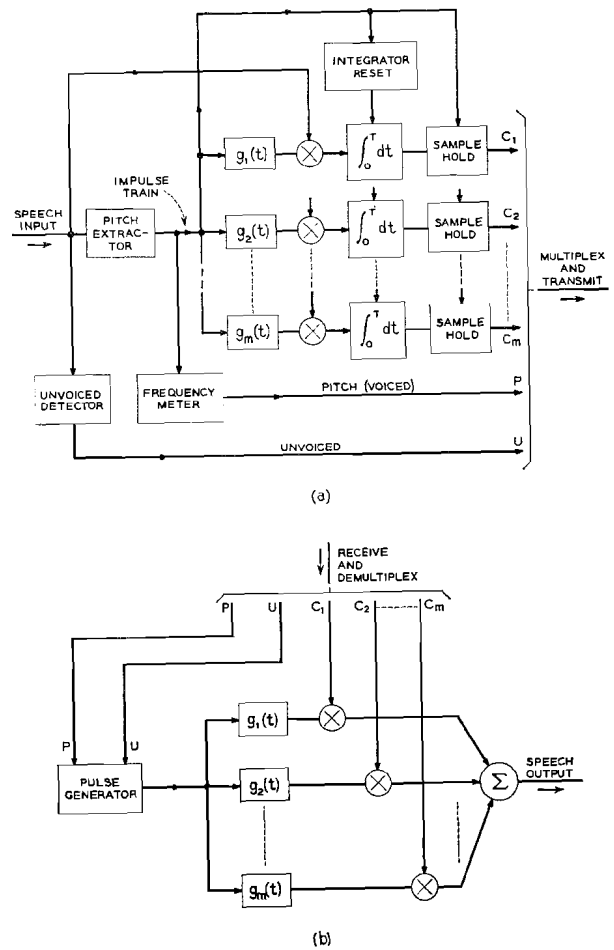


Fig. 8.13a and b. System for transmitting speech waveforms in terms of orthogonal functions. (After MANLEY and KLEIN.) (a) Analyzer. (b) Synthesizer

The processing of unvoiced, aperiodic sounds is slightly different. Ideally they are treated as if their waveforms constituted one pitch period. The onset of an unvoiced sound is detected and, if the unvoiced sound is relatively brief, as in a stop, only one pitch pulse is generated in the transmitter and in the receiver. The unvoiced indication is signalled to the receiver by the $u(t)$ parameter. If the unvoiced sound is sustained (for example, a fricative), the pulse generators are made to continue generating pulses with periods long enough that the periodicity is not significant perceptually.

8.62. Expansion of the Short-Time Amplitude Spectrum

At least one orthogonal-function description of the short-time amplitude spectrum has been proposed as a bandsaving means for coding speech (PIROGOV). The approach is particularized to a Fourier series description where, in effect, a spectrum of the amplitude spectrum is obtained. The technique is illustrated in Fig. 8.14.

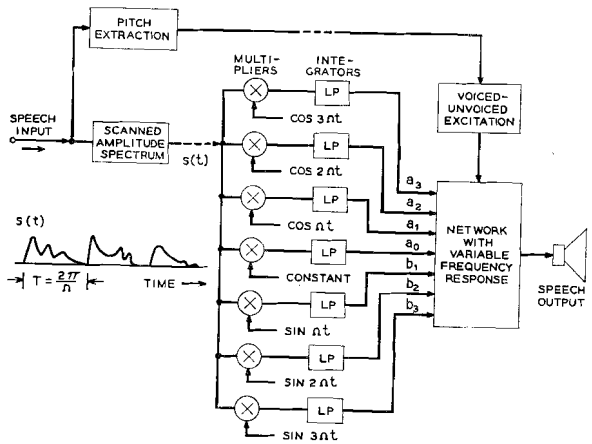


Fig. 8.14. Method for describing and synthesizing the short-time speech spectrum in terms of Fourier coefficients. (After PIROGOV)

A short-time amplitude spectrum is produced as a time function by scanning at a frequency $1/T$. The operation can be implemented as in the formant extractor described in Section 5.2, or in the manner of the "scan vocoder" discussed in Section 8.1, or even with a single scanning filter. The frequency $1/T$ would normally range from 25 to 50 cps, depending upon the requirements imposed on the quality of transmission. As in the "scan vocoder", the spectral description $s(t)$ is transmitted over a restricted bandwidth channel. A bandwidth between 75 and 250 cps is reported to be adequate. Excitation information, that is, pitch and voiced-unvoiced indications, must also be transmitted. As in the conventional vocoder, a bandwidth of 25 to 50 cps is expected to be adequate for these data. Synchronizing information about the scanning must also be made known to the receiver.

At the receiver, a Fourier series description of the amplitude spectrum is computed, namely,

$$s(t)=\frac{a_0}{2}+\sum_{n=1}^{N}\left[a_n\cos n\Omega t+b_n\sin n\Omega t\right], \tag{8.12}$$

where, as usual, the coefficients are

$$a_n=\frac{2}{T}\int_0^T s(t)\cos n\Omega t\,dt$$

$$b_n=\frac{2}{T}\int_0^T s(t)\sin n\Omega t\,dt,$$

and $\Omega=2\pi/T$. Practically, the Fourier coefficients are obtained by multiplying $s(t)$ by the outputs of several harmonic oscillators each synchronized to the scanning frequency Ω. An $N=3$ to 5 is claimed to provide an adequate spectral description (PIROGOV).

The coefficients vary relatively slowly with time. They are used to control an electrical network so that its frequency response is approximately the same as the measured spectral envelope of the speech signal. The network is then excited in a manner similar to the conventional vocoder, that is, either by periodic pulses or by noise. The reconstructed speech is the output of the variable network shown in Fig. 8.14.

The operation of the controllable network is based upon the fact that $s(t)$ is actually a spectral amplitude $S(\omega)$, $0\leqq\omega\leqq\omega_{\max}$. Hence, $\Omega=2\pi/T=2\pi/\omega_{\max}$, so that Eq. (8.12) can be rewritten as

$$S(\omega)=\frac{a_0}{2}+\sum_{n=1}^{N}a_n\cos\frac{2\pi n\omega}{\omega_{\max}}+b_n\sin\frac{2\pi n\omega}{\omega_{\max}}. \tag{8.13}$$

If the excitation amplitude spectrum is $G(\omega)$, then the output of the variable network should be $S(\omega)\cdot G(\omega)$. Assuming the excitation spectrum is flat and of unity amplitude, a given sine component ω_1 in the excitation spectrum should produce an output time function

$$\begin{aligned}f_1(t)=\frac{a_0}{2}\sin\omega_1 t&+\sin\omega_1 t\sum_{n=1}^{N}a_n\cos\frac{2\pi n\omega_1}{\omega_{\max}}\\&+\sin\omega_1 t\sum_{n=1}^{N}b_n\sin\frac{2\pi n\omega_1}{\omega_{\max}}.\end{aligned} \tag{8.14}$$

Expanding the second and third terms as sums and differences of angles gives

$$\begin{aligned}2f_1(t)=a_0\sin\omega_1 t&+\sum_{n=1}^{N}a_n\left[\sin\left(\omega_1 t-\frac{2\pi n\omega_1}{\omega_{\max}}\right)\right.\\&\left.+\sin\left(\omega_1 t+\frac{2\pi n\omega_1}{\omega_{\max}}\right)\right]\\&+\sum_{n=1}^{N}b_n\left[\cos\left(\omega_1 t-\frac{2\pi n\omega_1}{\omega_{\max}}\right)-\cos\left(\omega_1 t+\frac{2\pi n\omega_1}{\omega_{\max}}\right)\right].\end{aligned} \tag{8.15}$$

The second terms of the arguments, i.e., $\frac{2\pi n \omega_1}{\omega_{max}}$ correspond to time advances and delays of

$$n\tau = n \cdot \frac{2\pi}{\omega_{max}}.$$

The time function can therefore be constructed by the circuit shown in Fig. 8.15. The cosine terms of Eq. (8.15) are obtained by Hilbert transforming a difference of sine terms (i.e., by incurring a broadband $\pi/2$ phase shift). Although (8.15) is particularized for a given spectral component of excitation, namely ω_1, the process is the same for all other components. It is reported that with a spectral description of $N=4$ or 5, the synthesized speech quality is natural enough to satisfy the requirements of ordinary voice channels.

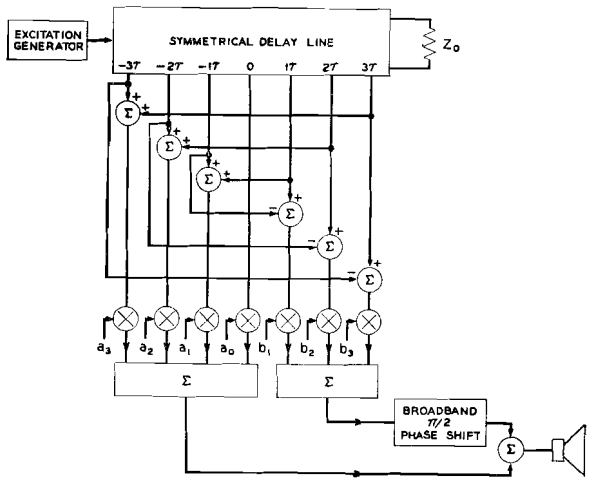


Fig. 8.15. Technique for realizing the variable electrical network of Fig. 8.14

8.63. Expansion of the Short-Time Autocorrelation Function

For an on-going time function $f(t)$, the discussion of Chapter V derived the relation between the short-time autocorrelation function (defined for positive delays)

$$\varphi(\tau, t) = \int_{-\infty}^{t} f(\lambda) f(\lambda-\tau) k(t-\lambda)\, d\lambda, \qquad \tau \geqq 0, \tag{8.16}$$

and the measurable short-time amplitude spectrum

$$F(\omega, t) = \int_{-\infty}^{t} f(\lambda) h(t-\lambda) e^{-j\omega\lambda}\, d\lambda. \tag{8.17}$$

For the specific weighting function

$$k(t) = h^2(t) = 2\sigma e^{-2\sigma t},$$

the short-time correlation and spectrum are linked by the weighted Fourier cosine transform

$$\begin{aligned} |F(\omega, t)|^2 &= \int_{-\infty}^{\infty} e^{-\sigma|\tau|} \varphi(\tau, t) \cos\omega\tau\, d\tau \\ &= \frac{1}{2\pi} |H(\omega)|^2 * \Phi(\omega, t), \end{aligned} \tag{8.18}$$

where $H(\omega)$ and $\Phi(\omega, t)$ are the Fourier transforms of $h(t)$ and $\varphi(\tau, t)$, respectively. The transform-pair (8.18) implies that $\varphi(\tau, t)$ is an even function of τ.

The preceding section described a technique for representing the short-time amplitude spectrum $|F(\omega, t)|$ in terms of an orthogonal function expansion. Since the correlation function and power spectrum are uniquely linked, it might be expected that a related orthonormal expansion can be written for the correlation function. This expansion leads to an alternative time-domain representation of the signal. In particular, Laguerre functions have been found a convenient expansion for this description (Y. W. Lee; Manley; Kulya).

Suppose the short-time correlation function of $f(t)$ for positive delays is expanded in terms of a realizable function set $\{\xi_i(\tau)\}$, orthonormal on the internal $0 \leqq \tau \leqq \infty$ and zero for $\tau < 0$. Then

$$\varphi(+\tau, t) = \sum_{i=0}^{\infty} a_i(t)\, \xi_i(\tau), \qquad \tau \geqq 0. \tag{8.19}$$

Because of the orthogonal properties

$$\begin{aligned} a_i(t) &= \int_0^{\infty} \varphi(+\tau, t)\, \xi_i(\tau)\, d\tau \\ &= \int_0^{\infty} \xi_i(\tau)\, d\tau \int_{-\infty}^{t} f(\lambda) f(\lambda-\tau) k(t-\tau)\, d\lambda. \end{aligned} \tag{8.20}$$

Changing the order of integration and substituting $\gamma = (\lambda - \tau)$ gives

$$a_i(t) = \int_{-\infty}^{t} f(\lambda) k(t-\lambda)\, d\lambda \int_{-\infty}^{\lambda} f(\gamma)\, \xi_i(\lambda-\gamma)\, d\gamma. \tag{8.21}$$

The coefficients $a_i(t)$ are therefore obtained by first filtering $f(t)$ with a network whose impulse response is $\xi_i(t)$, multiplying the result by $f(t)$ and then filtering the product with a network whose impulse response is $k(t)$. The operations are illustrated in Fig. 8.16.

The $a_i(t)$ coefficients obtained from (8.21) describe $\varphi(\tau, t)$ for positive delays ($\tau \geqq 0$). If, as defined and as discussed in Chapter V, $\varphi(\tau, t)$ is an even function of τ, the correlation for negative delay may be written

$$\varphi(-\tau, t) = \sum_{i=0}^{\infty} a_i(t)\,\xi_i(-\tau), \qquad \tau < 0, \tag{8.22}$$

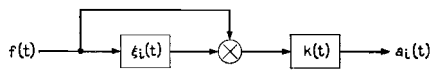


Fig. 8.16. Expansion coefficients for the short-time auto-correlation function

and the correlation function for all τ is

$$\begin{aligned}\varphi(\tau, t) &= \varphi(+\tau, t) + \varphi(-\tau, t)\\ &= \sum_{i=0}^{\infty} a_i(t)\,[\xi_i(\tau) + \xi_i(-\tau)]\,.\end{aligned} \tag{8.23}$$

The Fourier transform of $\varphi(\tau, t)$ is the power spectrum

$$\begin{aligned}\Phi(\omega, t) &= \sum_{i=0}^{\infty} a_i(t) \int_{-\infty}^{\infty} [\xi_i(\tau) + \xi_i(-\tau)]\, e^{-j\omega\tau}\, d\tau\\ &= \sum_{i=0}^{\infty} a_i(t)\,\{\Xi_i(\omega) + \Xi_i^*(\omega)\}\end{aligned} \tag{8.24}$$

where $\Xi_i(\omega)$ is the Fourier transform of $\xi_i(\tau)$.

The spectrum $\Phi(\omega, t)$ is related to the measurable power spectrum of Eq. (8.18) such that

$$|F(\omega, t)|^2 = \sum_{i=0}^{\infty} a_i(t)\,\{\Xi_i'(\omega) + \Xi_i'^*(\omega)\}\,, \tag{8.25}$$

where $\Xi_i'(\omega)$ is the Fourier transform of $[e^{-\sigma|\tau|}\,\xi_i(\tau)]$.

Writing $\Xi_i(\omega)$ in terms of its magnitude and phase,

$$\Xi_i(\omega) = \alpha_i(\omega)\, e^{-j\beta_i(\omega)}. \tag{8.26}$$

Then

$$\begin{aligned}\Phi(\omega, t) &= \sum_{i=0}^{\infty} a_i(t)\,\alpha_i(\omega)\,[e^{-j\beta_i(\omega)} + e^{+j\beta_i(\omega)}]\\ &= 2\sum_{i=0}^{\infty} a_i(t)\,\alpha_i(\omega)\cos\beta_i(\omega)\,.\end{aligned} \tag{8.27}$$

Thus the coefficients $a_i(t)$ of an orthonormal expansion of the autocorrelation function [Eq. (8.19)] are also the coefficients of a Fourier series expansion of the power spectrum.

So far, the orthogonal filter functions $\xi_i(t)$ have not been particularized. They have only been assumed to be physically realizable impulse responses. One simple set of orthonormal filters—and one that leads to a familiar result—is an ideal delay line with radian bandwidth B and with delay taps spaced at the Nyquist interval $1/2B$. The frequency response at the i-th tap is

$$\begin{aligned}\Xi_i(\omega) &= e^{-j\left(\frac{i\omega}{2B}\right)}, \quad 0 \leqq \omega \leqq B\\ &= e^{j\left(\frac{i\omega}{2B}\right)}, \quad -B \leqq \omega \leqq 0\\ &= 0, \quad \text{elsewhere}.\end{aligned} \tag{8.28}$$

The impulse response at the i-th tap is therefore

$$\xi_i(t) = \frac{B}{\pi}\,\frac{\sin\left(Bt - \frac{i}{2}\right)}{\left(Bt - \frac{i}{2}\right)}. \tag{8.29}$$

As prescribed by Eq. (8.28), the amplitude response is $\alpha_i(\omega) = 1$, and the phase response is

$$\beta_i(\omega) = \left(\frac{i\omega}{2B}\right).$$

The power spectrum expansion of Eq. (8.27) is therefore the Fourier series

$$\Phi(\omega, t) = 2\sum_i a_i(t)\cos\left(\frac{i\omega}{2B}\right). \tag{8.30}$$

The $a_i(t)$, on the other hand, which are computed according to the operations of Fig. 8.16, are simply values of the short-time autocorrelation function $\varphi(\tau, t)$ for $\tau = (i\omega/2B)$. These coefficients could be supplied directly to the left side of the synthesizer of Fig. 8.15 and used to generate the spectrum $\Phi(\omega, t)$. In this case, one has a correlation-vocoder synthesizer as described in Section 8.4.

Ideal broadband delay lines are neither physically wieldy nor particularly easy to construct. It is consequently of interest to consider other orthonormal function sets which might be useful in representing the short-time autocorrelation function or the power spectrum. Preferably, the functions should be realizable with simple lumped-element networks. The choice of Laguerre functions has advantages in this connection (Y. W. Lee).

Such an orthogonal set is

$$\{\xi_i(t)\}=\{l_i(t)\},$$

where the $l_i(t)$ are described by

$$l_i(t)=(2\lambda)^{\frac{1}{2}}\,e^{-\lambda t}\sum_{n=0}^{i}\frac{(-1)^n(2\lambda t)^{i-n}(i!/n!)}{[(i-n)!]^2}. \tag{8.31}$$

Its frequency transform is

$$\begin{aligned}L_i(\omega)&=\frac{(2\lambda)^{\frac{1}{2}}}{2\pi}\cdot\frac{(\lambda-j\omega)^i}{(\lambda+j\omega)^{i+1}}\\&=(-1)^i\frac{1}{\pi(2\lambda)^{\frac{1}{2}}}\left(\frac{\lambda}{j\omega+\lambda}\right)\left(\frac{j\omega-\lambda}{j\omega+\lambda}\right)^i\\&=A_i[u(\omega)]\,[v(\omega)]^i.\end{aligned} \tag{8.32}$$

The function (8.32) can be realized by cascading RC circuits of the type shown in Fig. 8.17, together with an amplification A_i.

If (8.32) is put in the form

$$L_i(\omega)=\alpha_i(\omega)\,e^{-j\beta_i(\omega)}, \tag{8.33}$$

then

$$L_i(\omega)=\frac{(2\lambda)^{\frac{1}{2}}}{2\pi}\cdot\frac{1}{(\omega^2+\lambda^2)^{\frac{1}{2}}}\cdot e^{j\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]}. \tag{8.34}$$

Further,

$$[L_i(\omega)+L_i^*(\omega)]=\frac{(2\lambda)^{\frac{1}{2}}}{\pi}\,\frac{1}{(\omega^2+\lambda^2)^{\frac{1}{2}}}\cos\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]. \tag{8.35}$$

The spectrum $\Phi(\omega, t)$ according to (8.24) and (8.27) is

$$\begin{aligned}\Phi(\omega,t)&=2\sum_{i=0}^{\infty}a_i(t)\,\alpha_i(\omega)\cos\beta_i(\omega)\\&=\left(\frac{2}{\pi^2\lambda}\right)^{\frac{1}{2}}\sum_i a_i(t)\,\frac{\cos\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]}{\left(1+\frac{\omega^2}{\lambda^2}\right)^{\frac{1}{2}}}.\end{aligned} \tag{8.36}$$

To show how the positive frequency domain is spanned by the Laguerre functions, the first several terms of the final factor in (8.36) are plotted in Fig. 8.18 (MANLEY). The functions are seen to have the desirable feature that they attenuate with increasing frequency, as does the speech spectrum.

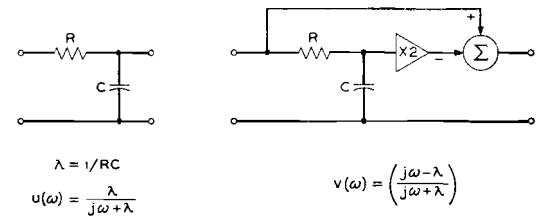


Fig. 8.17. Realization of Laguerre functions by RC networks [see Eq. (8.32)]

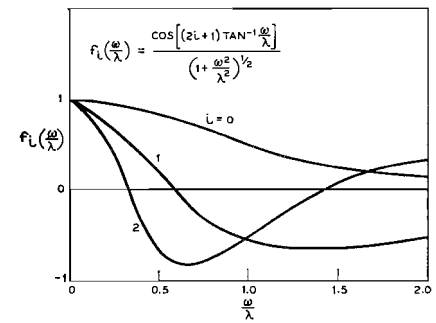


Fig. 8.18. Plot of the final factor in Eq. (8.36) showing how the positive frequency range is spanned by the first several Laguerre functions. (After MANLEY)

A transmission system based upon these relations can be constructed. The assumption is that the spectrum-squaring operation, that is, the synthesis of a signal having a spectrum $\Phi(\omega, t)$, is perceptually acceptable. (See Sections 8.1 and 8.4 for other comments on spectrum squaring.) Such a signal is

$$\varphi(\tau,t)=\sum_{i=0}^{\infty}a_i(t)\,[l_i(\tau)+l_i(-\tau)],$$

having the spectrum

$$\Phi(\omega,t)=\sum_{i=0}^{\infty}a_i(t)\,[L_i(\omega)+L_i^*(\omega)]. \tag{8.37}$$

The correlation $\varphi(\tau, t)$ is an even function of τ and is produced from $l_i(\tau)$, $\tau\geqq 0$. But with the circuits of Fig. 8.17, it is not possible to generate $l_i(-\tau)$. However, the ear is relatively insensitive to modest phase differences, and it suffices perceptually to generate a spectrum whose modulus is the same as $\Phi(\omega, t)$. Such a spectrum can be obtained from the odd function $[l_{m-i}(\tau)+l_{m+i+1}(\tau)]$ (KULYA). The corresponding spectrum is then

$$\Phi'(\omega,t)=\sum_{i=0}^{\infty}a_i(t)\,[L_{m-i}(\omega)+L_{m+i+1}(\omega)],$$

where, from Eq. (8.34),

$$[L_{m-i}(\omega)+L_{m+i+1}(\omega)] = \frac{(2\lambda)^{\frac{1}{2}}}{\pi}\frac{1}{(\omega^2+\lambda^2)^{\frac{1}{2}}}\left[e^{j2(m+1)\tan^{-1}\frac{\omega}{\lambda}}\right]\cos\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]. \quad (8.38)$$

Except for the phase angle $\left[e^{j2(m+1)\tan^{-1}\frac{\omega}{\lambda}}\right]$, Eq. (8.38) is identical to Eq. (8.35). The complete transmission system is therefore the circuit shown in Fig. 8.19. In Fig. 8.19a, the Laguerre expansion coefficients are developed according to Eq. (8.37) and after the fashion of Fig. 8.16. A pitch signal $p(t)$ is also extracted. The coefficients and pitch data are multiplexed and transmitted to the synthesizer in Fig. 8.19b. As in the vocoder, the synthesizer excitation is either wide-band noise or pitch-modulated pulses. By resorting to the odd function $[l_{m-1}(\tau)+l_{m+i+1}(\tau)]$, the synthesizer imposes the spectrum $\Phi'(\omega, t)$ upon the broadband excitation. Similar results can be obtained from an orthonormal expansion of the correlation function in terms of Tschebyscheff polynomials (KULYA).

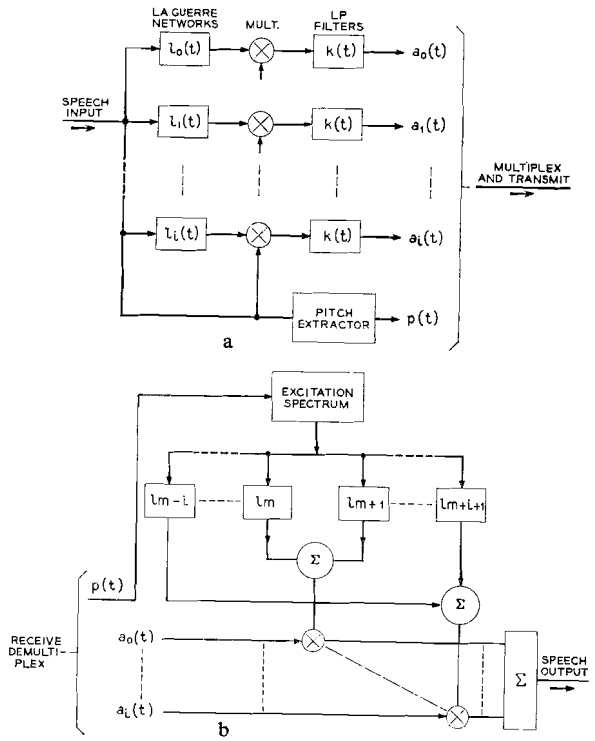


Fig. 8.19a and b. A Laguerre function vocoder. (a) Analyzer. (b) Synthesizer. (After KULYA)

8.7. Homomorphic Vocoders

In a further approach toward exploiting the source-system distinction in the speech signal, a processing technique called homomorphic filtering has been applied to vocoder design (OPPENHEIM; OPPENHEIM and SCHAFER). The approach is based on the observation that the mouth output pressure is approximately the linear convolution of the vocal excitation signal and the impulse response of the vocal tract. Homomorphic filtering[1] is applied to deconvolve the components and provide for their individual processing and description.

The analyzer and synthesizer operations for a complete homomorphic vocoder are shown in Fig. 8.20. Fig. 8.20a illustrates the analysis. At successive intervals (typically every 20 msec), the input speech signal is multiplied by a data window (a 40 msec Hamming

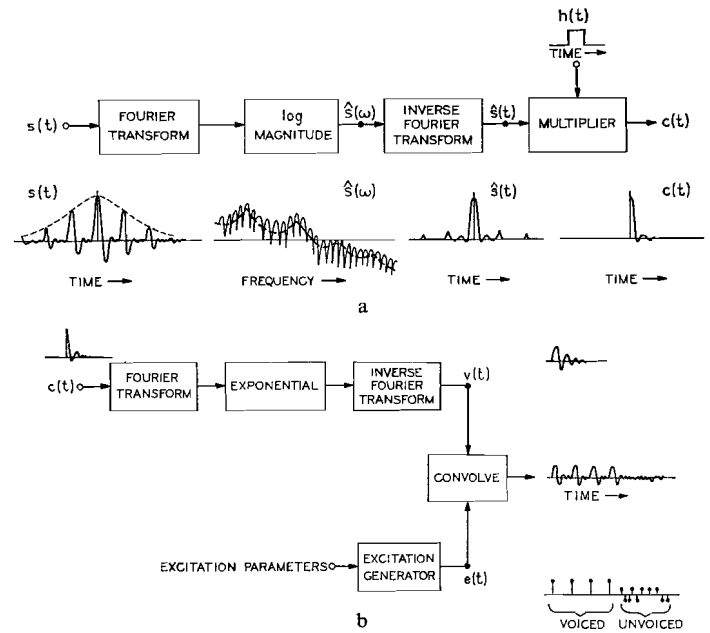


Fig. 8.20a and b. Analysis and synthesis operations for the homomorphic vocoder. (After OPPENHEIM)

[1] Homomorphic filtering is a generic term applying to a class of systems in which a signal-complex is transformed into a form where the principles of a linear filtering may be applied (OPPENHEIM). In the case of a speech signal, whose spectrum is approximately the product of the excitation spectrum and the vocal-tract transmission, a log-taking operation produces an additive combination of the source and system components. The cepstrum technique is therefore a special form of homomorphic filtering (see Sections 5.214 and 5.3).

window in this case) and the short-time Fourier transform is computed[1]. For each analysis interval the logarithm of the spectral magnitude is taken to produce the log-spectrum $\hat{S}(\omega)$. A further inverse Fourier transform produces the real, even time function $\hat{s}(t)$ which is defined as the cepstrum (see Sections 5.214 and 5.3). The low-time parts of $\hat{s}(t)$ characterize the slow fluctuations in $\hat{S}(\omega)$ due to the vocal-tract resonances, and the high-time parts of $\hat{s}(t)$ characterize the rapid fluctuations in $\hat{S}(\omega)$ due to vocal excitation properties. The high-time part of $\hat{s}(t)$ is used for voiced-unvoiced analysis and for pitch extraction, in accordance with the techniques described in Sections 5.214 and 5.3.

The final step in the analysis is to derive an equivalent minimum-phase description of the vocal-tract transmission by truncating and saving the positive low-time part of the cepstrum[2]. This is accomplished by multiplication with the time window $h(t)$. The result is $c(t)$ which together with the excitation information constitute the transmission parameters. The transform of $c(t)$ has a spectral magnitude illustrated by the dashed curve in $\hat{S}(\omega)$.

Synthesis is accomplished from $c(t)$ and the excitation information as shown in Fig. 8.20b. Periodic pulses, generated at the analyzed pitch, are used for synthesis of voiced sounds, and uniformly spaced pulses of random polarity are used for unvoiced sounds. The transmitted $c(t)$ is Fourier transformed, exponentiated (to undo the log-taking of the analysis), and an inverse transform yields a minimum-phase approximation to the vocal-tract impulse response. This impulse response is convolved with the excitation pulses to produce the output signal.

The system of Fig. 8.20 was implemented digitally on a general-purpose computer. Fast Fourier transform techniques and fast convolution techniques were used for the computations. The spectral analyses consisted of 512-point discrete Fourier transforms corresponding to a spectral resolution of approximately 20 cps. Cepstrum computations also consisted of 512-point inverse transforms. Spectra and cepstra were computed at 20-msec intervals along the input speech waveform and $c(t)$ was described by the first 32 points of the cepstrum. Linear interpolation over the 20 msec intervals was used for the excitation and impulse response data. Listening tests performed on the system in a back-to-back mode yielded judgments of good quality and natural sound. In a separate experiment the $c(t)$ data were reduced to 26 in number and quantized to six bits each for a transmission rate of 7800 bits/sec. At this bit rate no noticeable degradation was reported (OPPENHEIM).

[1] See Section 5.11 for properties of the short-time Fourier transform.

[2] The minimum-phase properties of this function are not obvious. A proof can be found in OPPENHEIM *et al.*

A further study of the homomorphic vocoder utilized a time-varying data window for analysis and a digital implementation for transmission at 3700 bits/sec (HAMMETT). At this bit rate, a signal of good quality was reported, with some reduction in naturalness.

Another study has applied predictive coding (see Section 8.13) to the transmission of the homomorphic vocoder signals. Using this technique, transmission of spectral information was digitally implemented for a data rate of 4000 bits/sec with modest impairment in quality. Listening tests concluded that spectral information digitized to around as 5000 bits/sec permits a quality indistinguishable from the unquantized system (WEINSTEIN and OPPENHEIM).

8.8. Maximum Likelihood Vocoders

All vocoder devices attempt to represent the short-time spectrum of speech as efficiently as possible. Exact reproduction of the waveform is not necessary. Some devices, such as channel vocoders, depend upon a frequency-domain transformation of the speech information, while others, such as correlation vocoders (Section 8.4) and orthogonal function vocoders (Section 8.6), use strictly a time-domain representation of the signal.

In all vocoder devices, the greatest step toward band conservation derives from observing the source-system distinctions in the production of speech signals (see Fig. 8.1). Vocal excitation information and system function data are treated separately, and decisions about voiced-unvoiced excitation and pitch-period measurement are typically made. Devices which do not make the source-system distinction and which do not perform pitch extraction—such as the voice-excited vocoder and some transmission methods described in later sections of this chapter—derive their bandsaving solely from the ear's acceptance of a signal having a short-time spectrum similar to that of the original speech. Their representation of the signal is commensurately less efficient.

Differences among vocoder devices lie in how they attempt to represent the perceptually-important information in the short-time speech spectrum. The channel vocoder merely samples the spectrum at prescribed frequency intervals and transmits these values. An orthonormal expansion of the amplitude spectrum aims to give adequate definition of the spectrum through a few coefficients of a prescribed set of basis functions. The formant vocoder assumes a pole-zero model for the vocal transmission and aims to locate the first few formant frequencies to effect an efficient description of the whole spectrum. The time-domain approach of the correlation vocoder transmits samples of the correlation function and synthesizes a wave composed of the even, truncated correlation funciton.

The Laguerre vocoder, another time-domain method, uses an orthonormal expansion of the short-time correlation function and attempts to represent it by a few coefficients.

Another technique, called the Maximum Likelihood Method (ITAKURA and SAITO, 1968), attempts to combine the advantages of time-domain processing and formant representation of the spectrum. The method is also amenable to digital implementation.

An all-pole model of the power spectrum of the speech signal is assumed. Zeros are omitted because of their lesser importance to perception and because their effect can be represented to any accuracy by a suitable number of poles. The synthesizer includes a recursive digital filter, shown in Fig. 8.21, whose transmission function in z-transform notation is

$$T(z)=\frac{1}{1+H(z)} =\left[\frac{1}{1+a_1 z^{-1}+a_2 z^{-2}+\cdots+a_p z^{-p}}\right], \tag{8.39}$$

where $z^{-1}=e^{-sD}$ is the delay operator, D is the sampling interval and s is the complex frequency (see, for example, Section 6.26).

The complex roots of the denominator polynomial are the complex formants (bandwidths and frequencies) used to approximate the speech signal. The coefficients, a_i, of the denominator polynomial are obtained from time-domain calculations on samples of a short segment of the speech waveform; namely, $s_1, s_2 \ldots s_N$, where $N \gg p$. Under the assumption that the waveform samples s_i are samples of a random gaussian process, a maximum likelihood estimate is obtained for the a_i's. This estimate corresponds to minimization of a function of the logarithmic

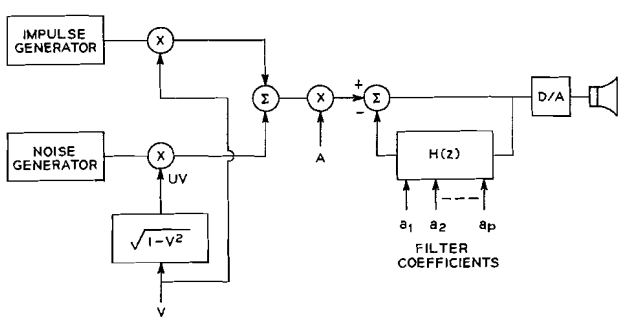


Fig. 8.21. Synthesis method for the maximum likelihood vocoder. Samples of voiced and voiceless excitation are supplied to a recursive digital filter of p-th order. Digital-to-analog (D/A) conversion produces the analog output. (After ITAKURA and SAITO, 1968)

difference between the power spectrum of the filter $|T(z)|^2$ and the short-time power spectrum of the signal samples

$$S(\omega)=\frac{1}{2\pi N}\left|\sum_{n=1}^{N} s_n e^{-jn\omega D}\right|^2. \tag{8.40}$$

The minimization results in a fit which is more sensitive at the spectral peaks than in the valleys between formants. Perceptually this is an important feature of the method. The fit of the all-pole model to the envelope of the speech spectrum is illustrated in Fig. 8.22.

The maximum likelihood estimate of the filter coefficients is obtained from the short-time correlation function

$$\Phi_i=\frac{1}{N}\sum_{j=1}^{N-i} s_j s_{j+i}, \qquad (i=0, 1, \ldots, N-1) \tag{8.41}$$

by solving the set of equations

$$\sum_{i=1}^{p} \Phi_{|i-j|}\, a_i=-\Phi_j, \qquad (j=1, 2, \ldots, p). \tag{8.42}$$

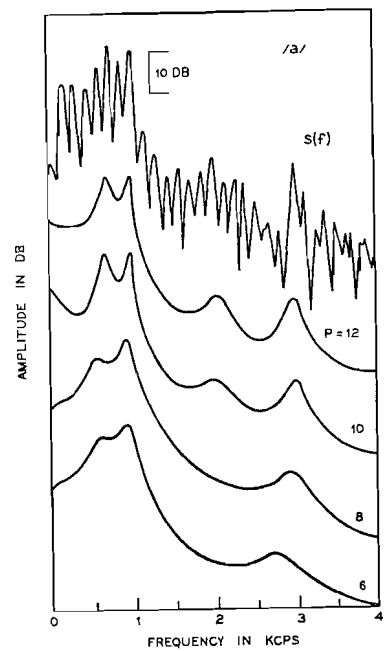


Fig. 8.22. Approximations to the speech spectrum envelope as a function of the number of poles of the recursive digital filter. The top curve, $S(f)$, is the measured short-time spectral density for the vowel /a/ produced by a man at a fundamental frequency 140 cps. The lower curves show the approximations to the spectral envelope for $p=6$, 8, 10 and 12. (After ITAKURA and SAITO, 1970)

The maximum likelihood estimate also produces the amplitude scale factor for matching the speech signal power spectrum, namely

$$A^2 = \sum_{i=-p}^{p} A_i \Phi_i ,$$

where

$$A_i = \sum_{j=0}^{p} a_j a_{j+|i|}; \quad a_0 = 1, \quad a_k = 0 (k > p). \tag{8.43}$$

As shown in Fig. 8.21, excitation of the synthesizer follows vocoder convention and uses a pulse generator and a noise generator of the same average power. Extraction of pitch period T is accomplished by a modified correlation method which has advantages similar to the cepstrum method, but relies strictly upon time domain techniques and does not require transformation to the frequency domain. A voicing amplitude signal, V, is also derived by the pitch extractor. The voiced and voiceless excitations are mixed according to the amplitude of the voicing signal, V. The unvoiced (noise) excitation level is given by $UV = \sqrt{1-V^2}$. The mixing ratio therefore maintains constant average excitation power. Overall control of the mixed excitation, by amplitude signal A, completes the synthesis[1].

Typical parameters for the analysis and synthesis are: sampling rate of input speech, $1/D = 8$ kcps; number of poles, $p = 10$; and number of analyzed samples, $N = 240$ (i.e., 30 msec duration). For transmission purposes, the control parameters are quantized to: 9 bits for each of the 10 a_i's, and 6 bits for each of the three excitation signals. Sampling these quantized parameters at 50 $\sec^{-1}$ yields a 5400 bit/sec encoding of the signal for digital transmission. The technique is demonstrated to be substantially better than digitized channel vocoders (ITAKURA and SAITO, 1968).

Furthermore, the maximum likelihood method has been shown to be valuable for automatic extraction of formant frequencies and formant bandwidths. The complex roots z_i of $[1+H(z)]$ in (8.39) give the real and imaginary parts of the formant frequencies, i.e., their bandwidths and center frequencies. Given $H(z)$ as defined by the coefficients a_i, a root-finding algorithm is applied to determine the z_i. Formant tracking tests on real speech show that the method with $p = 10$ produces accurate estimates of formant bandwidths and frequencies. An example of automatic formant tracking for a five-vowel sequence is shown in Fig. 8.23 (ITAKURA and SAITO, 1970).

[1] Note that while the coefficients a_i are derived from the short-time correlation function Φ_i, the synthesis method utilizes a recursive filter and avoids the "truncation" distortion present in the open-loop synthesis of the correlation vocoder (see Section 8.4).

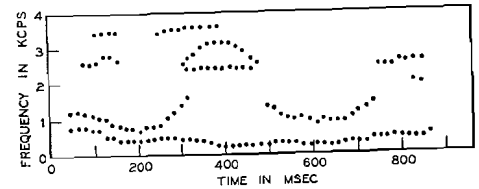


Fig. 8.23. Automatic tracking of formant frequencies determined from the polynomial roots for $p = 10$. The utterance is the five-vowel sequence /a, o, i, u, e/. (After ITAKURA and SAITO, 1970)

8.9. Linear Prediction Vocoders

Another time-domain vocoder method for speech analysis and synthesis employs the properties of linear prediction (ATAL and HANAUER, 1971). This method also utilizes an all-pole recursive digital filter excited either by a pitch-modulated pulse generator or a noise generator to synthesize the signal. The filter coefficients in this case represent an optimum linear prediction of the signal[1]. The coefficients are determined by minimizing the mean square error between samples of the input signal and signal values estimated from a weighted linear sum of past values of the signal. That is, for every sample of the input signal, s_n, an estimate $\hat{s}_n$ is formed such that

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k} .$$

The filter coefficients, a_k, are determined by minimizing $\overline{(s_n - \hat{s}_n)^2}$ over an analysis interval that is typically a pitch period, but which may be as small as 3 msec for $p = 12$ and a sampling rate of 10 Kcps. The a_k's are given as a solution of the matrix equation

$$\Phi \boldsymbol{a} = \boldsymbol{\psi} , \tag{8.44}$$

where $\boldsymbol{a}$ is a p-dimensional vector whose k-th component is a_k, Φ is a $(p \times p)$ covariance matrix with term φ_{ij} given by

$$\varphi_{ij} = \sum_n s_{n-i} s_{n-j} , \quad \begin{matrix} (i = 1, \ldots, p) \\ (j = 1, \ldots, p) \end{matrix} \tag{8.45}$$

and $\boldsymbol{\psi}$ is a p-dimensional vector with the j-th component $\psi_j = \varphi_{j0}$, and the sum extends over all speech samples N in a given analysis interval. Since the matrix Φ is symmetric and positive definite, Eq. (8.44) can be solved without matrix inversion. These relations are similar to those obtained from the Maximum Likelihood method [See Eq. (8.42).] except for the difference in the matrix Φ. The two solutions approach each other for the condition $N \gg p$.

[1] A general discussion of the theory of optimum linear prediction of signals is given in Section 8.13.

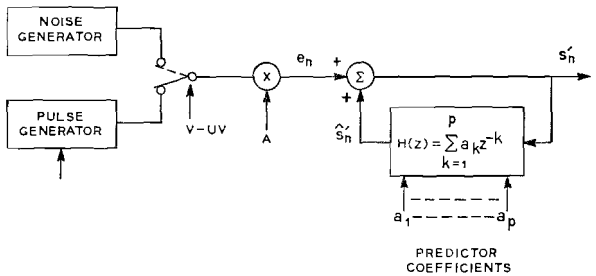


Fig. 8.24. Synthesis from a recursive digital filter employing optimum linear prediction. (After ATAL and HANAUER, 1971)

Synthesis is accomplished as shown in Fig. 8.24. Excitation either by pitch-modulated pulses or by random noise is supplied to a recursive filter formed from the linear predictor. The amplitude level, A, of the excitation is derived from the rms value of the input speech wave. The filter transmission function, is

$$T(z)=\frac{1}{1-H(z)}, \tag{8.46}$$

where

$$H(z)=\sum_{k=1}^{p} a_k z^{-k},$$

which, except for the sign convention, is the same as the Maximum Likelihood method (Section 8.8). The filter coefficients a_k account both for the filtering of the vocal tract and the spectral properties of the excitation source. If e_n is the n-th sample of the excitation, then the corresponding output sample of the synthesizer is

$$s'_n=e_n+\sum_{k=1}^{p} a_k s'_{n-k},$$

where the primes distinguish the synthesized samples from the original speech samples. The complex roots of $[1-H(z)]$ in (8.46) therefore include the bandwidths and frequencies of the speech formants. The filter coefficient data can be transmitted directly as the values of the a_k, or in terms of the roots of $[1-H(z)]$. The latter requires a root-finding calculation. Alternatively, the coefficient data can be transmitted in terms of the correlation functions φ_{ij}. Further, it can be shown that the recursive filter function describes an equivalent hard-walled pipe composed of right-circular sections in cascade. Its area is expected to be similar to that of the real vocal tract. The area data therefore provide an equivalent

form for the coefficient information. Because they can insure a stable filter function, the area and correlation functions are attractive for transmission and interpolation of the control data.

In one implementation, extraction of the pitch period, T, is accomplished by calculating the short-time autocorrelation function of the input speech signal after it has been raised to the third power. This exponentiation emphasizes the pitch periods of voiced passages. The voiced-unvoiced decision, $V-UV$, is based on the peak amplitude of the correlation function and on the density of zero crossings in the speech wave. Another implementation uses the error $(s_n-\hat{s}_n)$ and a peak-picking algorithm to determine the pitch period. Good-quality synthesis at a digital bit rate as low as 3600 bps has been reported for $p=12$ (ATAL and HANAUER, 1971).

Because the roots of polynomial $[1-H(z)]$ describe the complex formant frequencies, the linear prediction method is also effective for extracting formant bandwidths and center frequencies. For $p=12$ the accuracy in obtaining formant frequencies is considered to be within perceptual tolerances. An example of formant extraction from a voiced sentence is shown in Fig. 8.25 (ATAL and HANAUER, 1971).

The Linear Prediction Vocoder and the Maximum Likelihood Vocoder implement their analysis-synthesis procedures in much the same way. MARKEL has pointed out that they are fundamentally similar, and that both utilize an analysis technique devised earlier by Prony

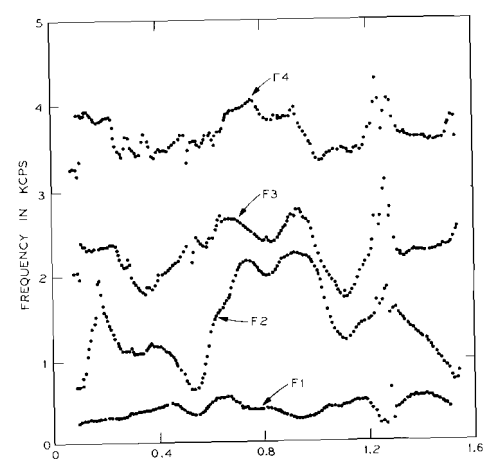


Fig. 8.25. Formant frequencies determined from the recursive filter coefficients. The utterance is the voiced sentence "We were away a year ago" produced by a man at an average fundamental frequency of 120 cps. (After ATAL and HANAUER, 1971)

(MARKEL, 1971). Further, an inverse digital filter, designed along the principles of Eqs. (8.39) and (8.46) and Figs. 8.21 and 8.24, has also been found useful for automatic formant extraction (MARKEL, 1972).

8.10. Articulatory Vocoders

An attractive approach to the general vocoder problem is to code speech in terms of articulatory parameters. Such a description has the advantage of duplicating, on a direct basis, the physiological constraints that exist in the human vocal tract. Nondiscontinuous signals that describe the vocal transmission would then produce all sounds, consonants and vowels.

The idea is to transmit a set of data which describes the tract configuration and its excitation as functions of time. The nature of the analysis—although neither its completeness nor sufficiency—is exemplified by the articulatory-domain spectral matching techniques described in Section 5.4, Chapter V. The synthesizer could be a controllable vocal tract model, such as described in Section 6.2, Chapter VI, or some equivalent device. At the present time no complete vocoder system based upon these principles has been demonstrated. However, the approach appears to be promising and to have much to recommend it. Its success will depend largely upon the precision with which articulatory data can be obtained automatically from the acoustic signal. As the discussion of Chapter V has indicated, computer techniques may provide the requisite sophistication for the analysis.

8.11. Frequency-Dividing Vocoders

A class of vocoder devices aims to avoid the difficult problems of pitch tracking and voiced-unvoiced switching that conventional vocoders use. The intent is to settle for modest savings in bandwidth in order to realize a simpler implementation and a synthetic signal of higher quality. The voice-excited vocoder described in Section 8.3 represents one such effort. The band saving which accrues is due primarily to the ear's criteria for representing the short-time spectrum of the signal.

Frequency division is a well-known process for reducing the bandwidth of signals whose spectral widths are determined primarily by large-index frequency modulation. While speech is not such a signal, subbands of it (for example, formant bands or individual voice harmonics) have similarities to large-index frequency modulation. Frequency division by factors of two or three are possible before intelligibility deteriorates substantially.

Frequency division generally implies possibilities for frequency multiplication. Similarly, spectral division-multiplication processes suggest possibilities for compression and expansion of the signal's time scale. Reproduction of a divided signal at a rate proportionately faster restores the frequency components to their original values, and compresses the time scale by a factor equal to the frequency divisor.

8.111. Vobanc

Various methods—including electrical, mechanical, optical and digital—have been used to accomplish division and multiplication. All cannot be described in detail. Several, however, serve to illustrate the variety of designs and applications.

One frequency-division method for bandwidth conservation is the Vobanc (BOGERT, 1956). Although constructed practically using heterodyne techniques, the principle involved is shown in Fig. 8.26. The speech

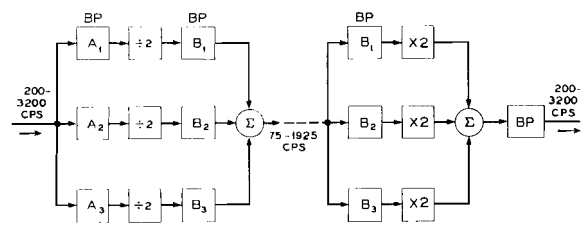


Fig. 8.26. Block diagram of the Vobanc frequency division-multiplication system. (After BOGERT, 1956)

band 200 to 3200 cps is separated into three contiguous band-pass channels, A_1, A_2, A_3. Each channel is about 1000 cps wide and normally covers the range of a speech formant. Using a regenerative modulator, the signal in each band is divided by two and limited to one-half the original frequency range by BP filters B_1, B_2, B_3. The added outputs of the filters yield a transmission signal which is confined to about one-half the original bandwidth.

At the receiver, the signal is again filtered into the three bands, B_1, B_2, and B_3. The bands are restored by frequency doubling and are combined to provide the output signal. In consonant articulation tests with 48 listeners and 10 talkers, the Vobanc consonant articulation was approximately 80 per cent. In the same test, an otherwise unprocessed channel, band-limited to 200 to 1700 cps, scored a consonant intelligibility of about 66 per cent.

Other systems similar in band-division to Vobanc have been investigated (SEKI; MARCOU and DAGUET). One proposal, called Codimex, considers potential division by factors as high as eight (DAGUET), although practical division appears limited to factors of two or three.

8.112. Analytic Rooter

Another technique for frequency division of formant bands of speech is called analytic rooting (SCHROEDER, FLANAGAN and LUNDRY). The processing is done in terms of the analytic signal. This approach avoids characteristic degradations that frequency division methods such as used in the Vobanc introduce.

The analytic signal $\sigma(t)$ of a real, bandlimited signal $s(t)$ is defined as

$$\sigma(t)=s(t)+j\hat{s}(t), \tag{8.47}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$. In polar form the analytic signal is

$$\sigma(t)=a(t)\,e^{j\Phi(t)}, \tag{8.48}$$

where

$$a(t)=[s^2(t)+\hat{s}^2(t)]^{\frac{1}{2}}$$
$$\Phi(t)=\tan^{-1}[\hat{s}(t)/s(t)].$$

It follows that

$$s(t)=a(t)\cos[\Phi(t)], \quad \text{and} \quad \hat{s}(t)=a(t)\sin[\Phi(t)]. \tag{8.49}$$

A real signal $s_{1/n}(t)$ corresponding to the n-th root of the analytic signal can be defined as

$$\begin{aligned} s_{1/n}(t)&=\mathrm{Re}[\sigma(t)]^{1/n}\\ &=\mathrm{Re}[s(t)+j\hat{s}(t)]^{1/n}\\ &=[a(t)]^{1/n}\cos[\Phi(t)/n]. \end{aligned} \tag{8.50}$$

The analytic signal rooting therefore implies division of the instantaneous frequency by a factor n, and taking the n-th root of the signal envelope[1]. For the case $n=2$ the relations are particularly tractable for computer simulation.

$$\begin{aligned} s_{\frac{1}{2}}(t)&=[a(t)]^{\frac{1}{2}}\cos[\tfrac{1}{2}\Phi(t)]\\ &=[a(t)]^{\frac{1}{2}}[\tfrac{1}{2}(1+\cos\Phi(t))]^{\frac{1}{2}}. \end{aligned} \tag{8.51}$$

[1] Note that for those cases where perceived pitch is determined by the envelope of the signal waveform, this process leaves the pitch unaltered. This method is therefore attractive for restoring speech distorted by a helium atmosphere, such as breathed by a deep-sea diver.

Since $a(t)\cos\Phi(t)=s(t)$, one may write (8.51) as

$$s_{\frac{1}{2}}(t)=(\tfrac{1}{2})^{\frac{1}{2}}[a(t)+s(t)]^{\frac{1}{2}}. \tag{8.52}$$

Similarly, it can be shown that the Hilbert transform $\hat{s}_{\frac{1}{2}}(t)$ of $s_{\frac{1}{2}}(t)$ is

$$\hat{s}_{\frac{1}{2}}(t)=(\tfrac{1}{2})^{\frac{1}{2}}[a(t)-s(t)]^{\frac{1}{2}}. \tag{8.53}$$

Eq. (8.53) also follows from (8.52) by the observation that multiplication of $s(t)$ by -1 is equivalent to a phase shift of π and that, according to (8.51), this corresponds to a phase shift of $\pi/2$ in $s_{\frac{1}{2}}(t)$, i.e., a Hilbert transformation.

Eq. (8.52) is a simple relation which is easy to simulate on a computer and amenable to straight-forward instrumentation—except for one difficulty: the sign of the square root and therefore of $s_{\frac{1}{2}}(t)$, according to (8.52), is indeterminate.

The proper sign can be recovered by changing the sign of the square root in (8.52) every time the phase $\Phi(t)$ of the original signal $s(t)$ goes through 2π (or an integer multiple of 2π). According to (8.49) this is the case when $\hat{s}(t)=0$, while $s(t)<0$.

A remaining phase ambiguity of π in $s_{\frac{1}{2}}(t)$ is unavoidable and is a direct consequence of the 2π phase ambiguity in the original signal $s(t)$. This phase ambiguity has no practical consequence.

The inverse operation of analytic-signal rooting is given by

$$s_n(t)=\mathrm{Re}[s(t)+j\hat{s}(t)]^n. \tag{8.54}$$

By writing

$$s_n(t)=[a(t)]^n\cos[n\,\Phi(t)], \tag{8.55}$$

and by comparing (8.55) with (8.50), the inverse relationship is evident.

For $n=2$, (8.54) yields

$$s_2(t)=\mathrm{Re}[s^2(t)+2js(t)\hat{s}(t)-\hat{s}^2(t)], \tag{8.56}$$

or

$$s_2(t)=s^2(t)-\hat{s}^2(t).$$

If process (8.56) is applied to $s_{\frac{1}{2}}(t)$, the original signal $s(t)$ is recovered. This can be verified by substituting $s_{\frac{1}{2}}(t)$ and $\hat{s}_{\frac{1}{2}}(t)$ from (8.52) and (8.53) into (8.56):

$$s_2(t)=\tfrac{1}{2}\{[a(t)+s(t)]-[a(t)-s(t)]\},$$

or

$$s_2(t)=s(t). \tag{8.57}$$

The Hilbert transform of the original signal can be recovered by multiplying $s_{\frac{1}{2}}(t)$ and $\hat{s}_{\frac{1}{2}}(t)$:

$$\begin{aligned} 2s_{\frac{1}{2}}(t)\cdot\hat{s}_{\frac{1}{2}}(t) &= \{[a(t)+s(t)][a(t)-s(t)]\}^{\frac{1}{2}} \\ &= \{a^2(t)-s^2(t)\}^{\frac{1}{2}} \\ &= \hat{s}(t). \end{aligned} \tag{8.58}$$

For a signal whose bandwidth is narrow compared to its center frequency, the original signal can be approximately recovered by squaring $s_{\frac{1}{2}}(t)$ and subsequent bandpass filtering. From (8.52),

$$2s_{\frac{1}{2}}^2(t) = a(t)+s(t). \tag{8.59}$$

If the spectrum of $a(t)$ does not overlap that of $s(t)$, which is approximately true for narrowband signals, then $s(t)$ can be recovered by bandpass filtering.

A complete transmission system based upon the foregoing principles has been simulated on a digital computer. In the simulation, the speech spectrum is first divided into four contiguous passbands, each nominally containing no more than one formant. Each bandpass signal is then analytically rooted, band-limited, and recovered in accordance with the previous explanation.

To accomplish square rooting of the signal, and a band reduction of 2-to-1, a typical channel in the flow diagram for the simulation program is shown in Fig. 8.27. The bandpass filter BPF1 separates a spectral segment which nominally contains no more than one formant. The Hilbert transform of this signal is formed by a transversal filter HT1. Since the Hilbert transform filter ideally has a response which is neither time-limited nor band-limited, an approximation is made to the transform which is valid over the frequency range of interest and which is truncated in time.

In a parallel path, the bandpass signal $s(t)$ is delayed by an amount DEL1 equal to one-half the duration of the impulse response of the Hilbert filter. It, too, is squared and $(s^2+\hat{s}^2)$ is formed by ADD1. The square root of this result yields $a(t)$ in accordance with (8.48), and the addition of the delayed $s(t)$ in ADD2 gives $[a(t)+s(t)]$. Multiplication by $\frac{1}{2}$ and the subsequent square rooting form $s_{\frac{1}{2}}(t)$, according to (8.52).

Selection of the sign of $s_{\frac{1}{2}}(t)$ is accomplished by the following logical decisions in SWITCH. The algebraic sign of $s_{\frac{1}{2}}(t)$ is changed whenever $\hat{s}(t)$ goes through zero while $s(t)<0$. The signal $s_{\frac{1}{2}}(t)$, so signed, is then applied to BPF$\frac{1}{2}$, having cutoff frequencies, and hence bandwidth, equal to one-half the values for BPF1.

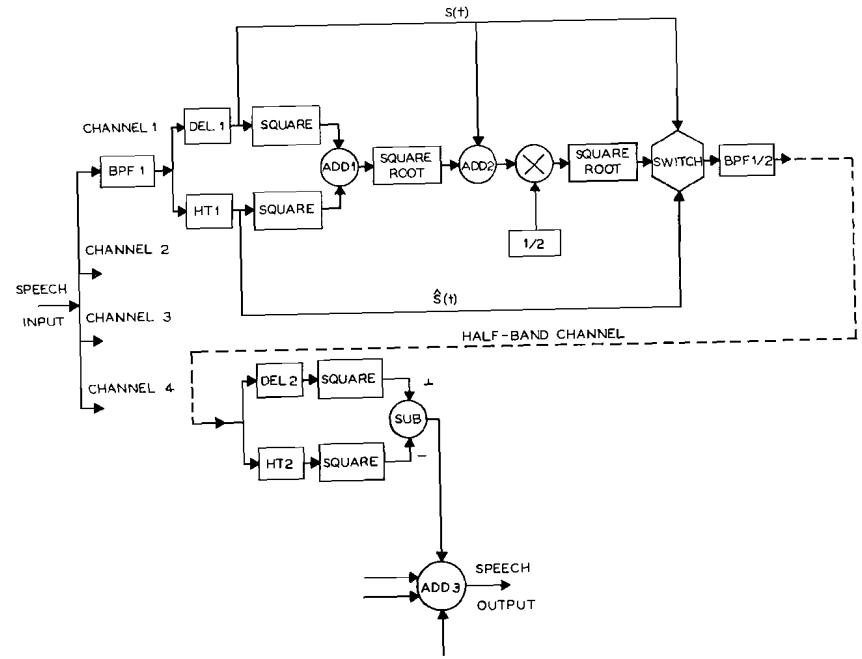


Fig. 8.27. Diagram for computer simulation of the analytic rooter. (After SCHROEDER, FLANAGAN and LUNDRY)

Analytic squaring of this band-limited version of $s_{\frac{1}{2}}(t)$ is accomplished in accordance with (8.56). The Hilbert transform is produced by HT2, which is similar to HT1 except that the duration of the impulse response of the former is twice that of the latter. Subtracting $\hat{s}^2(t)$ from $s^2(t)$ recovers an approximation to the original bandpassed signal $s(t)$.

The programmed operations in all four channels are identical except that the bandpass filters, Hilbert transform filters, and delays are chosen in accordance with the desired passband characteristics. In the computer implementation, eighth-order Butterworth filters with cutoff frequencies listed in Table 8.5 are used for the bandpass filters.

Table 8.5. *Eighth-order Butterworth filter cutoff frequencies in cps*

	BPF1	BPF $\frac{1}{2}$	Formants nominally in passband
Channel 1	238–714	119–357	$F1$
Channel 2	714–1428	357–714	$F1$ or $F2$
Channel 3	1428–2142	714–1071	$F2$ or $F3$
Channel 4	2142–2856	1071–1428	$F3$

The Hilbert filters are realized from a band-limited and time-limited approximation to the Hilbert transform. Ideally, the impulse response (inverse) of the Hilbert transform is $h(t)=1/\pi t$, and the magnitude of the transform is unity at all frequencies. Truncating the spectrum of the transform at frequency ω_c produces an impulse response $\tilde{h}(t)=(\cos\omega_c t-1)/\pi t$, which although band-limited is not time-limited. The function $\tilde{h}(t)$ is asymmetric and its even Nyquist samples are identically zero. Odd Nyquist samples have the value $2/\pi nT$, where n is the sample number and T is the Nyquist interval. The response $\tilde{h}(t)$ can be truncated (limited) in time at a sufficiently long duration so that over the frequency range of interest the transform remains acceptable.

For programming ease, the transform is realized by an asymmetric transversal filter whose even (Nyquist) coefficients are zero and whose odd coefficients are $2/\pi nT$, weighted with a Hamming window of duration τ. Specifically,

$$\tilde{h}(nT)=\frac{2}{\pi nT}\left\{0.54-0.46\cos\left(\frac{2\pi(nT+\tau/2)}{\tau}\right)\right\}, \qquad (8.60)$$

where $n=1, 2, 3, \ldots, \tau/2T$ represents values for one-half the coefficients of the asymmetrical filter. The simulation is for a 10-kHz bandwidth (ω_c) and $T=0.5\times10^{-4}$ second. The values of the Hamming window used for each of the four bands are given in Table 8.6.

Table 8.6. *Impulse response durations for the Hilbert filters*

	τ in ms	
	HT1	HT2
Channel 1	5.0	10.0
Channel 2	2.5	5.0
Channel 3	1.3	2.5
Channel 4	0.9	1.7

A typical result from the system, with the BPF$\frac{1}{2}$ filters included in the transmission path, is shown by the spectrograms in Fig. 8.28. The upper spectrogram shows an input sentence to the system. The lower spectrogram shows the signal recovered from the half-bandwdith transmission. As the spectrograms show, original formant structure and pitch information is preserved relatively well in the recovered signal. The result is a transmission of respectable quality over a channel bandwidth equal to one-half that of the original signal.

At least one practical hardware implementation of the analytic-rooter, using solid-state circuitry, has been constructed and tested (SASS and MACKIE).

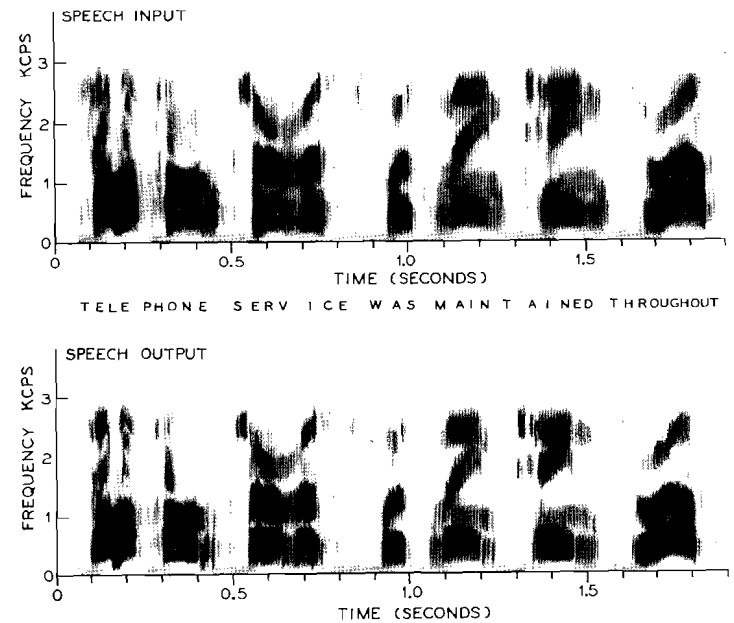


Fig. 8.28. Sound spectrograms of speech analyzed and synthesized by the analytic rooter. The transmission bandwidth is one-half the original signal bandwidth. (After SCHROEDER, FLANAGAN and LUNDRY)

8.113. Harmonic Compressor

Another complete division-multiplication transmission system, designed with a sufficient number of filters to operate on individual voice harmonics, has been investigated by digital simulation (SCHROEDER, LOGAN and PRESTIGIACOMO). The method, called the "harmonic compressor", uses 50 contiguous bandpass filters, each 60 cps wide, covering the range 240 to 3240 cps. The circuit is shown in Fig. 8.29. It is designed to achieve a bandwidth reduction of two-to-one. On the transmitter side, the signals from the bandpass filters are divided by two and combined for transmission over one-half the original bandwidth. At the receiver the components are again separated by filtering and restored by multiplication by two. All filters and operations are simulated in a large digital computer. From informal listening tests, the quality and intelligibility of the transmitted speech are judged to fall between that of a voice-excited vocoder with a 700 cps baseband and an unprocessed signal of the same bandwidth. A time speed up by a factor of two can also be applied to the transmitted signal to restore it to the original frequency range.

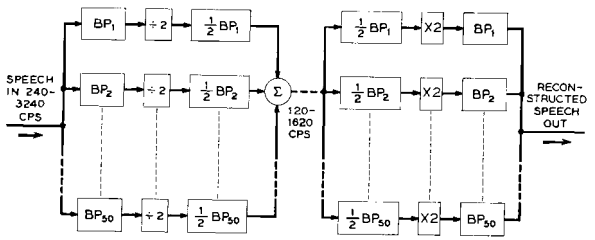


Fig. 8.29. Block diagram of "harmonic compressor". (After SCHROEDER, LOGAN and PRESTIGIACOMO)

A related investigation in which attention is focused upon the individual harmonic components of the signal has considered optical methods, mechanical string-filter methods, and ultrasonic storage devices for frequency division-multiplication (VILBIG, 1950, 1952; VILBIG and HAASE, 1956a, b). A part of this same effort produced an electrical "speech stretcher" (GOULD). The idea is to expand the time scale of speech by the arrangement shown in Fig. 8.30. The speech signal is filtered by 32 contiguous BP-filters covering the range 75 to about 7000 cps. The filter bandwidths are approximately 100 cps wide up to 1000 cps, and increase logarithmically to 7000 cps. Full-wave rectification doubles the frequency components of each band. Band-pass filtering at twice the original bandwidth eliminates much of the harmonic distortion. Recording the combined signal and playing back at one-half speed restores the components to their original frequency positions. The time scale of the signal, however, is expanded by two.

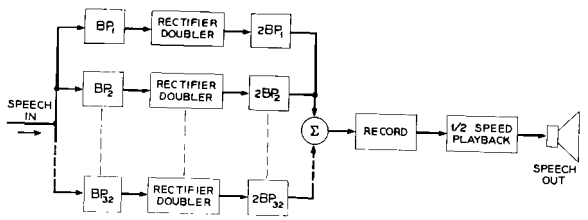


Fig. 8.30. A "speech stretcher" using frequency multiplication to permit expansion of the time scale. (After GOULD)

8.114. Phase Vocoder

A final frequency division-multiplication method makes use of the short-time phase derivative spectrum of the signal to accomplish the band saving. The method permits non-integer divisions as well as

integer values. It can be applied either to single voice harmonics or to wider subbands which can include single formants. It also permits a flexible means for time compressing or expanding the speech signal. The method is called Phase Vocoder (FLANAGAN and GOLDEN).

If a speech signal $f(t)$ is passed through a parallel bank of contiguous band-pass filters and then recombined, the signal is not substantially degraded. The operation is illustrated in Fig. 8.31, where $BP_1 \ldots BP_N$ represent the contiguous filters. The filters are assumed to have relatively

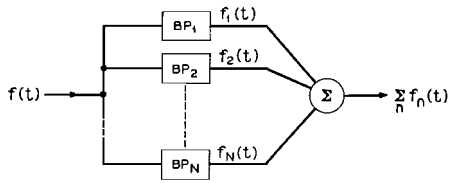


Fig. 8.31. Filtering of a speech signal by contiguous band-pass filters

flat amplitude and linear phase characteristics in their pass bands. The output of the n-th filter is $f_n(t)$, and the original signal is approximated as

$$f(t) \cong \sum_{n=1}^{N} f_n(t). \tag{8.61}$$

Let the impulse response of the n-th filter be

$$g_n(t) = h(t) \cos \omega_n t, \tag{8.62}$$

where the envelope function $h(t)$ is normally the impulse response of a physically-realizable low-pass filter. Then the output of the n-th filter is the convolution of $f(t)$ with $g_n(t)$,

$$\begin{aligned} f_n(t) &= \int_{-\infty}^{t} f(\lambda)\, h(t-\lambda) \cos[\omega_n(t-\lambda)]\, d\lambda \\ &= \mathrm{Re}\left[\exp(j\omega_n t) \int_{-\infty}^{t} f(\lambda)\, h(t-\lambda) \exp(-j\omega_n \lambda)\, d\lambda\right]. \end{aligned} \tag{8.63}$$

The latter integral is a short-time Fourier transform of the input signal $f(t)$, evaluated at radian frequency ω_n. It is the Fourier transform of that part of $f(t)$ which is "viewed" through the sliding time aperture $h(t)$. If we denote the complex value of this transform as $F(\omega_n, t)$, its magnitude is the short-time amplitude spectrum $|F(\omega_n, t)|$, and its angle is the short-time phase spectrum $\varphi(\omega_n, t)$. Then

$$f_n(t) = \mathrm{Re}[\exp(j\omega_n t) F(\omega_n, t)]$$

or

$$f_n(t) = |F(\omega_n, t)| \cos[\omega_n t + \varphi(\omega_n, t)]. \tag{8.64}$$

Each $f_n(t)$ may, therefore, be described as the simultaneous amplitude and phase modulation of a carrier ($\cos \omega_n t$) by the short-time amplitude and phase spectra of $f(t)$, both evaluated at frequency ω_n.

Experience with channel vocoders shows that the magnitude functions $|F(\omega_n, t)|$ may be band-limited to around 20 to 30 Hz without substantial loss of perceptually-significant detail. The phase functions $\varphi(\omega_n, t)$, however, are generally not bounded; hence they are unsuitable as transmission parameters. Their time derivatives $\dot{\varphi}(\omega_n, t)$, on the other hand, are more well-behaved, and may be band-limited and used to advantage in transmission. To within an additive constant, the phase functions can be recovered from the integrated (accumulated) values of the derivatives. One practical approximation to $f_n(t)$ is, therefore,

$$\tilde{f}_n(t) = |F(\omega_n, t)| \cos[\omega_n t + \tilde{\varphi}(\omega_n, t)], \tag{8.65}$$

where

$$\tilde{\varphi}(\omega_n, t) = \int_0^t \dot{\varphi}(\omega_n, t)\, dt.$$

The expectation is that loss of the additive phase constant will not be unduly deleterious.

Reconstruction of the original signal is accomplished by summing the outputs of n oscillators modulated in phase and amplitude. The oscillators are set to the nominal frequencies ω_n, and they are simultaneously phase and amplitude modulated from band-limited versions of $\dot{\varphi}(\omega_n, t)$ and $|F(\omega_n, t)|$. The synthesis operations are diagrammed in Fig. 8.32.

These analysis-synthesis operations may be viewed in an intuitively appealing way. The conventional channel vocoder separates vocal excitation and spectral envelope functions. The spectral envelope functions of the conventional vocoder are the same as those described

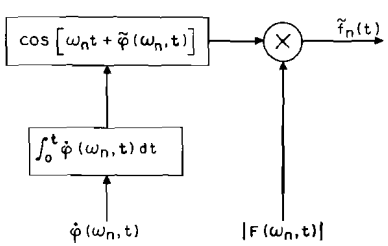


Fig. 8.32. Speech synthesis from short-time amplitude and phase-deviative spectra. (After FLANAGAN and GOLDEN)

here by $|F(\omega_n, t)|$. The excitation information, however, is contained in a signal which specifies voice pitch and voiced-unvoiced (buzz-hiss) excitation. In the phase vocoder, when the number of channels is reasonably large, information about excitation is conveyed primarily by the $\dot{\varphi}(\omega_n, t)$ signals. At the other extreme, with a small number of broad analyzing channels, the amplitude signals contain more information about the excitation, while the $\dot{\varphi}$ phase signals tend to contain more information about the spectral shape. Qualitatively, therefore, the number of channels determines the relative amounts of excitation and spectral information carried by the amplitude and phase signals. If good quality and natural transmission are requisites, the indications are that the $\dot{\varphi}(\omega_n, t)$ signals require about the same channel capacity as the spectrum-envelope information. This impression seems not unreasonable in view of experience with voice quality in vocoders.

A complete phase vocoder analyzer and synthesizer has been simulated on a digital computer. In the analyzer, the amplitude and phase spectra are computed by forming the real and imaginary parts of the complex spectrum

$$F(\omega_n, t) = a(\omega_n, t) - j\, b(\omega_n, t),$$

where

$$a(\omega_n, t) = \int_{-\infty}^{t} f(\lambda)\, h(t-\lambda) \cos \omega_n \lambda\, d\lambda$$

and

$$b(\omega_n, t) = \int_{-\infty}^{t} f(\lambda)\, h(t-\lambda) \sin \omega_n \lambda\, d\lambda. \tag{8.66}$$

Then,

$$|F(\omega_n, t)| = (a^2 + b^2)^{\frac{1}{2}}$$

and

$$\dot{\varphi}(\omega_n, t) = \left(\frac{\dot{a}\, b - \dot{b}\, a}{a^2 + b^2}\right).$$

The computer, of course, deals with sampled-data equivalents of these quantities. Transforming the real and imaginary parts of (8.66) into discrete form for programming yields

$$a(\omega_n, mT) = T \sum_{l=0}^{m} f(lT)\, [\cos \omega_n lT]\, h(mT - lT)$$

$$b(\omega_n, mT) = T \sum_{l=0}^{m} f(lT)\, [\sin \omega_n lT]\, h(mT - lT), \tag{8.67}$$

where T is the sampling interval. In the simulation, $T=10^{-4}$ sec. From these equations, the difference values are computed as

$$\Delta a = a[\omega_n, (m+1)T] - a[\omega_n, mT]$$

and

$$\Delta b = b[\omega_n, (m+1)T] - b[\omega_n, mT]. \tag{8.68}$$

The magnitude function and phase derivative, in discrete form, are computed (8.67) and (8.68) as

$$|F[\omega_n, mT]| = (a^2+b^2)^{\frac{1}{2}}$$

$$\frac{\Delta\varphi}{T}[\omega_n, mT] = \frac{1}{T}\frac{(b\Delta a - a\Delta b)}{a^2+b^2}. \tag{8.69}$$

Fig. 8.33 shows a block diagram of a single analyzer channel as realized in the program. This block of coding is required for each channel.

In the simulation, a sixth-order Bessel filter is used for the $h(lT)$ window. The simulation uses 30 channels ($N=30$) and $\omega_n = 2\pi n(100)$ rad/sec. The equivalent passbands of the analyzing filters overlap at their 6 dB down points, and a total spectrum range of 50 to 3050 cps is analyzed.

Programmed low-pass filtering is applied to the amplitude and phase difference signals as defined by Fig. 8.33. Simulation of the whole system is completed by the synthesis operations for each channel performed

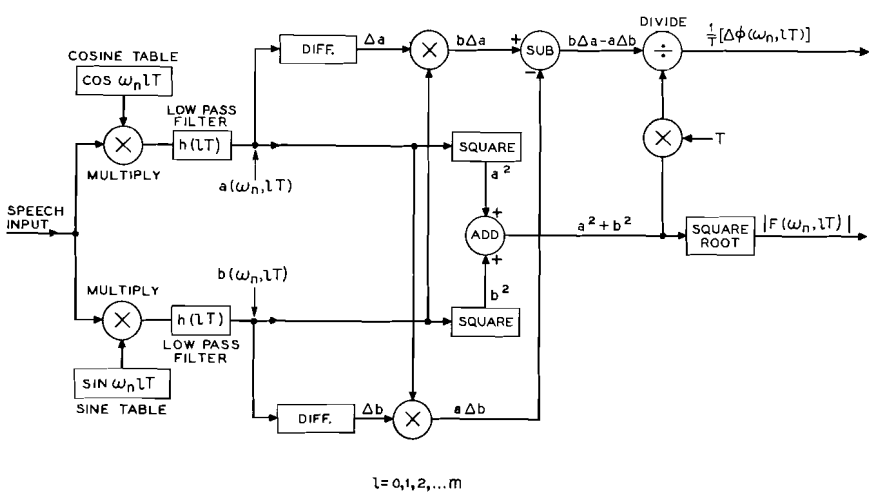


Fig. 8.33. Programmed analysis operations for the phase vocoder. (After FLANAGAN and GOLDEN)

according to

$$f_n(mT) = |F(\omega_n, mT)| \cos\left[\omega_n mT + T\sum_{l=0}^{m}\frac{\Delta\varphi(\omega_n, lT)}{T}\right]. \tag{8.70}$$

Adding the outputs of the n individual channels, according to (8.61), produces the synthesized speech signal.

As part of the simulation, identical (programmed) low-pass filters were applied to the $|F(\omega_n, lT)|$ and $(1/T)\,[\Delta\varphi(\omega_n, lT)]$ signals delivered by the coding block shown in Fig. 8.33. These low-pass filters are similar to the $h(lT)$ filters except they are fourth-order Bessel designs. The cut-off frequency is 25 cps, and the response is -7.6 dB down at this frequency. This filtering is applied to the amplitude and phase signals of all 30 channels. The total bandwidth occupancy of the system is therefore 1500 cps, or a band reduction of 2:1.

After band-limitation, the phase and amplitude signals are used to synthesize an output according to (8.70). The result of processing a complete sentence through the programmed system is shown by the sound spectrograms in Fig. 8.34[1]. Since the signal band covered by the analysis and synthesis is 50 to 3050 cps, the phase-vocoded result

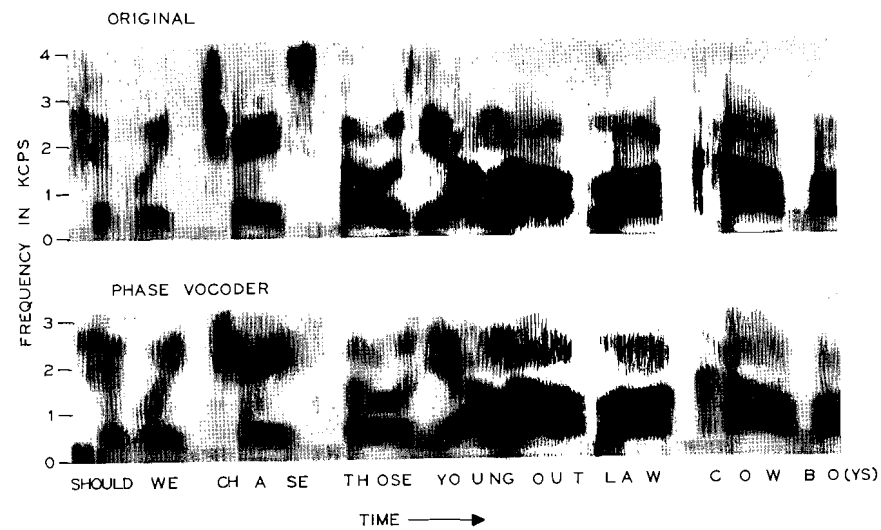


Fig. 8.34. Speech transmitted by the phase vocoder. The transmission bandwidth is one-half the original signal bandwidth. Male speaker: "Should we chase those young outlaw cowboys." (After FLANAGAN and GOLDEN)

[1] The input speech signal is band limited to 4000 cps. It is sampled at 10000 cps and quantized to 12 bits. It is called into the program from a digital recording prepared previously.

is seen to cut off at 3050 cps. In this example, the system is connected in a "back-to-back" configuration, and the band-limited channel signals are not multiplexed.

Comparison of original and synthesized spectrograms reveals that formant details are well preserved and pitch and voiced-unvoiced features are retained to perceptually significant accuracy. The quality of the resulting signal considerably surpasses that usually associated with conventional channel vocoders.

A frequency-divided signal may be synthesized by division of the $[\omega_n t+\int \dot{\varphi}_n dt]$ quantities by some number q. This frequency-divided synthetic signal may be essentially restored to its original spectral position by a time speed-up of q. Such a speed-up can be accomplished by recording at one speed and replaying q-times faster. The result is that the time scale is compressed and the message, although spectrally correct, lasts $1/q$-th as long as the original. An example of a 2:1 frequency division and time speed-up is shown by the sound spectrograms in Fig. 8.35.

Time-scale expansion of the synthesized signal is likewise possible by the frequency multiplication $q[\omega_n t+\int \dot{\varphi}_n dt]$; that is, by recording the frequency-multiplied synthetic signal and then replaying it at a speed q-times slower. An example of time-expanded speech is shown by the spectrograms in Fig. 8.36.

An attractive feature of the phase vocoder is that the operations for expansion and compression of the time and frequency scales can be realized by simple scaling of the phase-derivative spectrum. Since the frequency division and multiplication factors can be non-integers, and can be varied with time, the phase vocoder provides an attractive tool for studying non-uniform alterations of the time scale (HANAUER and SCHROEDER).

A number of multiplexing methods may be used for transmission. Conventional space-frequency and time-division methods are obvious techniques. A "self multiplexing" method is also possible in which, say, a two-to-one frequency-divided synthetic signal is transmitted over an analog channel of $\frac{1}{2}$ the original signal bandwidth. Re-analysis, frequency expansion and synthesis at the receiver recovers the signal[1]. Further, at least one digital implementation of the phase vocoder has been made. The phase and amplitude functions were sampled, quantized and framed for digital transmission at digital rates of 9600 bits/sec and 7200 bits/sec.

[1] The greatest number q by which the ω_n and $\dot{\varphi}_n$'s may be divided is determined by how distinct the side-bands about each ω_n/q remain, and by how well each $\dot{\varphi}_n/q$ and $|F_n|$ may be retrieved from them. Practically, the greatest number appears to be about 2 or 3 if transmission of acceptable quality is to be realized.

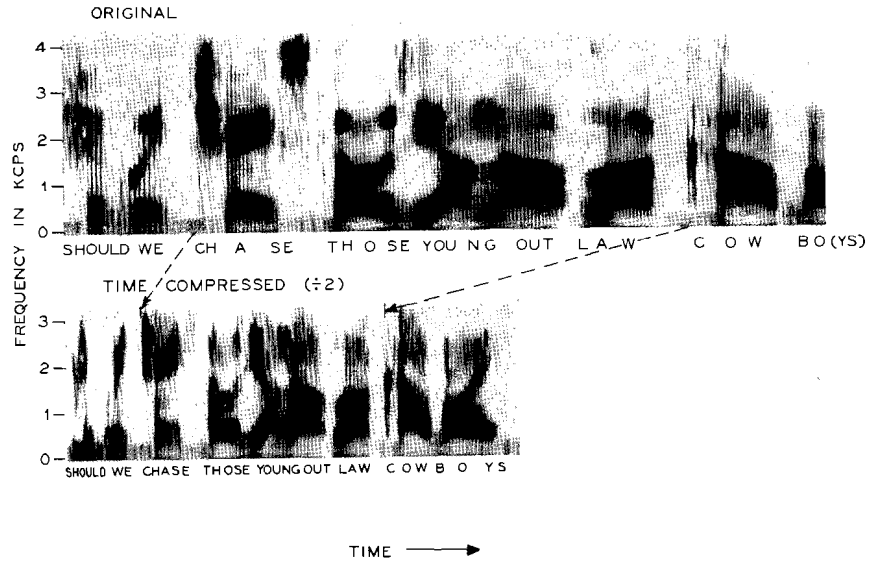


Fig. 8.35. Phase vocoder time compression by a factor of 2. Male speaker

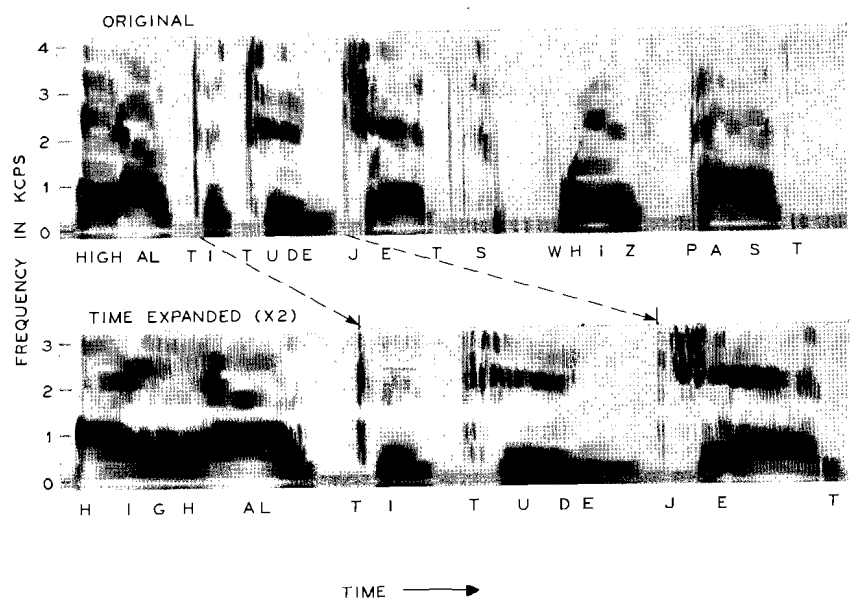


Fig. 8.36. Phase vocoder time expansion by a factor of 2. Female speaker

These transmission rates were compared in listening tests to the same signal coded as log-PCM. The results showed the digital phase vocoder to provide a signal quality comparable to log-PCM at bit rates two to three times higher (CARLSON).

8.12. Time-Assignment Transmission of Speech

In two-way conversation, one party is normally silent and listening on the average of one-half the time. In addition, natural speech has many pauses and silent intervals. A given talker, therefore, transmits a signal only on the order of 35 to 40 per cent of the total time. In long-distance communication, where amplification of the signal is necessary, the two-way communication channels are normally four-wire circuits—or two unilateral transmission paths. Each party has a transmit circuit and a receive circuit. Because of the relative inactivity of each talker, a single one-way channel is not used on the order of 60 to 65 per cent of the time. When a large group of such connections are accessible from single transmit and receive locations, the statistical properties of the conversation ensemble make a significant amount of time and bandwidth available for signal transmission. A method for practicably utilizing this capacity is called Time Assignment Speech Interpolation, or "TASI" (O'NEIL; BULLINGTON and FRASER).

The TASI system has available a group of unilateral transmit and receive circuits—typically the line-pairs in an undersea cable. The system is to serve a greater number of talkers than the number of unilateral circuits. The incoming transmit circuit of each talker is equipped with a fast-acting speech detector, or voice switch. When the detector indicates the presence of speech on its line, an automatic electronic switch connects the line to an available transmit path of the TASI group. Incoming signals for transmission are assigned transmit circuits until all have been filled. When the number of signals to be transmitted exceeds the number of transmit paths, the TASI switch searches the connections to find one that has fallen silent, disconnects it, and assigns that path to a channel which has a signal to transmit.

During pauses and silent intervals, a given talker loses his priority on the transmit link. He is reassigned a channel—often a different one—when he again becomes active. The TASI switch must consequently keep track of who is talking to whom, and it must identify the recipient of each signal presented for transmission. This message "addressing" information can be transmitted in the form of a very short identification signal, either before each talk spurt or over an auxiliary channel that serves the entire system.

A limit obviously exists to the number of incoming signals that can be transmitted by a given group of transmit paths before some "freeze-out" or loss of speech signals occurs. Among other things, this limit is a function of the size of the cable group, the circuit signal-to-noise ratio, and the sensitivity of the speech detectors. Several TASI systems have been put into practical operation on undersea cables. On a 36-channel cable, for example, the effective transmission bandwidth is on the order of two to three times that of the physical circuit.

As mentioned at the beginning of the section, natural pauses of phonemic, syllabic or longer durations occur in a single "one-way" speech signal. These pauses or gaps suggest that the TASI principle might be applied to a single speech channel to realize a band-saving. An experimental circuit, called a "one-man TASI", has considered this point (FLANAGAN, SCHROEDER and BIRD). The system has been tested by simulation in a digital computer. Its principle of operation is illustrated by the schematic sound spectrogram in Fig. 8.37.

As shown in Fig. 8.37, suppose that a speech band of BW cps is to be transmitted, but that a channel width of only BW/2 is available. The natural pauses and gaps in one BW/2 of the signal might be used to transmit information about the other BW/2 band of the signal. If the BW/2 bands are called high band (HB) and low band (LB), four signal possibilities exist. The processing strategies employed in the four situations are illustrated by corresponding letters on Fig. 8.37, and are:

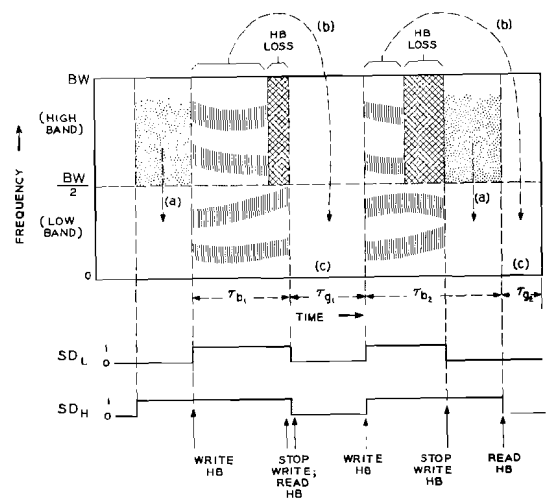


Fig. 8.37. Schematic sound spectrogram illustrating the principle of the "one-man TASI". (After FLANAGAN, SCHROEDER and BIRD)

a) When only HB signal (and no LB signal) is present, the HB is detected, heterodyned down to the LB range, and transmitted immediately over the BW/2 channel.

b) When HB and LB are detected simulataneously, the LB is transmitted immediately, while the HB is heterodyned down and read into a storage for transmission later. (See τ_b intervals in Fig. 8.37).

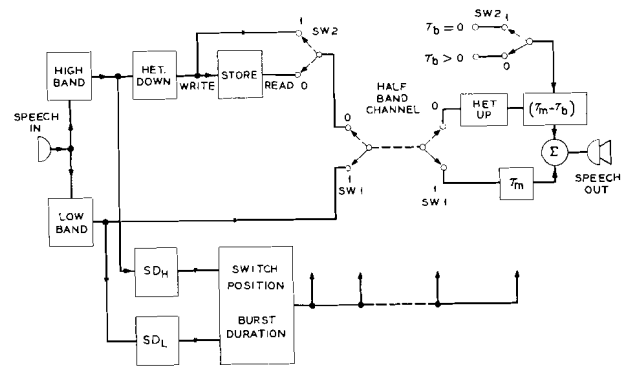
c) When neither HB nor LB signal is detected, a gap exists. (See τ_g intervals in Fig. 8.37.) During this interval, as much of the previously-stored HB is transmitted as there is time for. Generally some trailing edge of the HB will be lost. One set of speech-burst statistics gives average burst durations of about 130 msec followed by average silent intervals of 100 msec (BOLT and MACDONALD). On the basis of these statistics, about 3/13 of the HB signal would be expected to be lost. None of the LB signal is lost.

d) When LB only is present, it is transmitted immediately in the conventional manner.

Two speech detectors, one for each band, are required. In the present study, they were full-wave rectifiers with 15-msec smoothing time constants. Their outputs operated threshold devices with prescribed hysteresis characteristics. The binary output signals from the detectors, shown as SD_L and SD_H in Fig. 8.37, must also be transmitted over a narrow-band channel so that the speech may be properly reassembled at the receiver. Because of the storage on the transmitter side, a fixed transmission delay is incurred before the reassembled signal is available.

The reassembly operations are evident in the block diagram of the complete system in Fig. 8.38. Two delay elements are used at the receiver. One is a fixed, maximum transmission delay τ_m in the LB channel. Its value is equal to or greater than the duration of the longest speech burst to be stored. The other is a variable delay whose value is the difference between τ_m and the last speech-burst duration τ_b. The various switch conditions—corresponding to the SD_L and SD_H signal outputs—are shown in the table.

In testing the system by simulation in a digital computer, the effective size of the HB store was taken as 500 msec. In the unlikely instance of a speech-burst duration longer than 500 msec, the high-band information was discarded, rather than reassembled in the wrong place. Typical operation of the system, as simulated in the computer, is shown by the spectrograms of Fig. 8.39. The utterance is "High altitude jets whiz past screaming". In comparing what the complete system provides over and above a single BW/2 channel, one sees that a substantial amount of the high band is transmitted. All high frequencies from unvoiced bursts are present, and a large percentage of the voiced HB is preserved.



SD_L	SD_H	SW1	SW2	STORE
1	0	1	0	---
0	1	0	1	---
1	1	1	0	WRITE
0	0	0	0	READ

Fig. 8.38. Block diagram of "one-man TASI" system for 2:1 band-width reduction. (After FLANAGAN, SCHROEDER and BIRD)

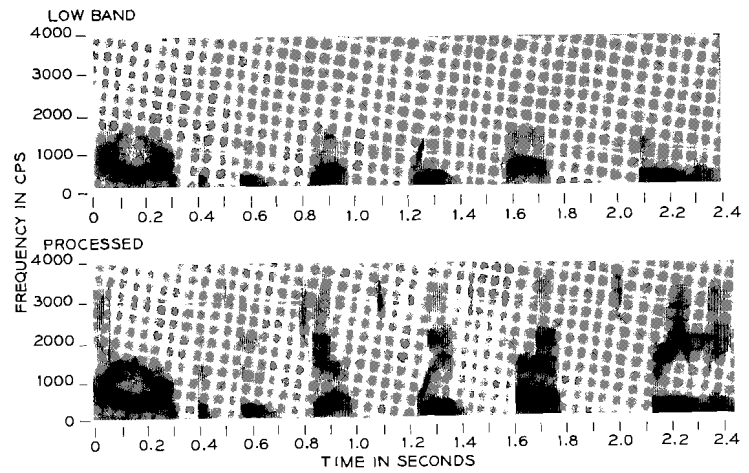


Fig. 8.39. Sound spectrograms illustrating operation of the single channel speech interpolator

The price of the improvement is the complexity of the storage and switching and the 500-msec transmission delay.

Alternatively, the silent gaps in the speech signal may be used to interleave another signal, such as digital data read on demand from a buffer store. In one computer simulation of this technique (SCHROEDER

and HANAUER), the speech envelope was used as a control to switch between speech and data. It was found possible to make available as much as 55% of the speech-signal time for interleaving the alternate information.

8.13. Predictive Coding of Speech

For many classes of information signals, including speech, the value of the signal at a given instant is correlated with its values at other instants, and hence represents redundant information. One theory of data compression in digital systems is therefore based upon forming an error signal, e_i, between the samples of an input sequence, s_i, and linear estimates of those samples, $\hat{s}_i$,

$$e_i = (s_i - \hat{s}_i).$$

Generally, the estimate $\hat{s}_i$ of sample s_i is formed as a weighted linear combination of samples from some portion of the input sample sequence.

The weighting coefficients used for the estimate are computed from statistics of the sample sequence in a manner which is optimum in some sense. If the input sample sequence is not stationary, the weighting coefficients must be updated periodically.

In order to transmit a block of M samples to the receiver, it is necessary that the error samples and the weighting coefficients be transmitted to the receiver. Suppose the desired accuracy of the input sample sequence requires "r" bits per sample. By straightforward quantization, it would take $(M \cdot r)$ bits to transmit the block of M samples. However, if the sample sequence is processed through a data compression system, the number of bits needed to transmit the block is hopefully less. Usually the error signal is transmitted at the same rate as the input sample sequence, but the weighting coefficients are transmitted typically at a rate $1/M$ times the input sequence. Suppose the error signal is quantized to q bits and the N weighting coefficients are coded to w bits per coefficient. The number of bits needed the specify the M samples to the receiver is then $(Mq + Nw)$. In order to obtain a saving, it is required that

$$Mq + Nw < M \cdot r$$

or

$$q + \frac{N}{M} w < r.$$

If the sample sequence is highly correlated, the power in the error signal will be significantly less than the power in the input sample sequence. Hence, fewer bits will be required to describe the error samples

than the input samples. If $M \gg N$, then the term $\frac{N}{M} w$ becomes negligible and the objective can be achieved.

One such method of data compression is linear prediction (ELIAS). Linear prediction has been found to provide significant improvements in picture transmission, speech transmission, and the transmission of telemetry data. A linear predictor forms its estimates of the input samples from past samples in the input sequence. Another method of data compression is linear interpolation. An interpolator forms its estimates of the input samples from both past and future samples in the input sequence.

Linear interpolation has the potential for reducing the power in the error signal beyond that for an equal-order prediction. However, interpolation requires more computation and complex implementation. Also, it looses some of its advantages when the error signal is quantized inside a feedback loop (HASKEW). The present discussion will therefore focus on prediction.

A linear N-th-order predictor estimates the magnitude of the present input sample, s_i, by a linear combination, $\hat{s}_i$, of N weighted past samples.

$$\hat{s}_i = \sum_{j=1}^{N} a_j s_{i-j}, \tag{8.71}$$

where a_j is the weighting coefficient applied to the past sample s_{i-j}.

When the statistics of the input signal are nonstationary (changing as a function of time), the weighting coefficients must be updated periodically. Only the weighting coefficients computed for intervals near the present sample yield accurate estimates of the sample magnitude. In this case, weighting coefficients are updated, for example, every M input samples, where M is usually much larger than the order of the predictor, N.

The output of the predictor, the error e_i, is formed by subtracting the estimated value of the present sample from the actual value of the present sample.

$$e_i = s_i - \sum_{j=1}^{N} a_j s_{i-j}. \tag{8.72}$$

The input signal is now described by the output of the predictor (the error signal) and the weighting coefficients. In z-transform notation

$$e(z) = [1 - P(z)]\, s(z),$$

where

$$P(z) = \sum_{j=1}^{N} a_j z^{-j}. \tag{8.73}$$

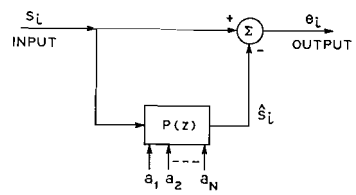
These relations are shown schematically in Fig. 8.40. Recovery of original input signal is obtained from the inverse relation

$$s(z)=e(z)[1-P(z)]^{-1}, \tag{8.74}$$

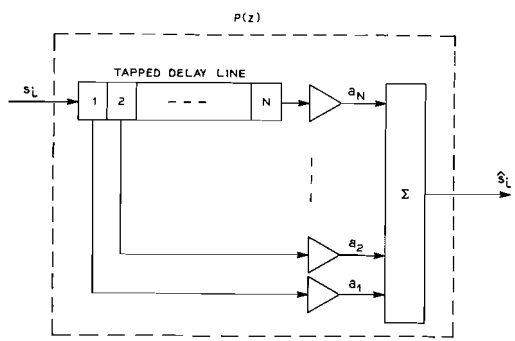
and is given by the operations of Fig. 8.41. Typically, however, the transmitted signals, i.e., the e_i and a_i, are quantized, and the receiver has access only to corrupted versions of them.

The criterion by which the a_i are typically determined is a minimization of the power of the error signal (that is, minimization of the square difference between $\hat{s}_i$ and s_i). For M samples the error power is

$$\varepsilon^2=\frac{1}{M}\sum_{j=1}^{M} e_j^2=\frac{1}{M}\sum_{j=1}^{M}(s_j-\hat{s}_j)^2. \tag{8.75}$$



(a)



(b)

Fig. 8.40 a and b. Block diagram of linear prediction

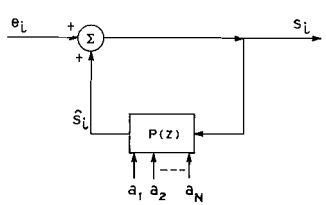


Fig. 8.41. Linear prediction receiver

Substitution for the estimate $\hat{s}_i$ gives

$$\varepsilon^2=\frac{1}{M}\sum_{j=1}^{M}\left[s_j-\sum_{k=1}^{N}a_k s_{j-k}\right]^2$$

or

$$\begin{aligned}\varepsilon^2=&\frac{1}{M}\sum_{j=1}^{M}s_j^2-\frac{2}{M}\sum_{j=1}^{M}\sum_{k=1}^{N}a_k s_j s_{j-k}\\&+\frac{1}{M}\sum_{j=1}^{M}\left[\sum_{k=1}^{N}a_k s_{j-k}\right]\left[\sum_{l=1}^{N}a_l s_{j-l}\right].\end{aligned} \tag{8.76}$$

Interchanging summations and rearranging terms,

$$\begin{aligned}\varepsilon^2=&\frac{1}{M}\sum_{j=1}^{M}s_j^2-2\sum_{k=1}^{N}a_k\left[\frac{1}{M}\sum_{j=1}^{M}s_j s_{j-k}\right]\\&+\sum_{k=1}^{N}\sum_{l=1}^{N}a_k a_l\left[\frac{1}{M}\sum_{j=1}^{M}s_{j-k}s_{j-l}\right].\end{aligned} \tag{8.78}$$

Define the signal power σ^2 and its covariance function r_{kl} as

$$\sigma^2=\frac{1}{M}\sum_{j=1}^{M}s_j^2,$$

and

$$r_{kl}=\frac{1}{M\sigma^2}\sum_{j=1}^{M}s_{j-l}s_{j-k}. \tag{8.79}$$

The error power then becomes

$$\varepsilon^2=\sigma^2\left[1-2\sum_{k=1}^{N}a_k r_{0k}+\sum_{k=1}^{N}\sum_{l=1}^{N}a_k a_l r_{kl}\right]. \tag{8.80}$$

This result can be simplified by matrix notation. Define the column matrix containing the weighting coefficients as

$$A=\begin{bmatrix}a_1\\a_2\\\vdots\\a_N\end{bmatrix}. \tag{8.81}$$

Define the column matrix containing the elements r_{0k} as

$$G=\begin{bmatrix}r_{01}\\r_{02}\\\vdots\\r_{0N}\end{bmatrix}. \tag{8.82}$$

Define the $(N x N)$ matrix containing the elements r_{kl} as

$$R=\begin{bmatrix} r_{11} & r_{12} \cdots r_{1N} \\ r_{21} & r_{22} \cdots r_{2N} \\ \vdots & \vdots \\ r_{N1} & r_{N2} \cdots r_{NN} \end{bmatrix}. \tag{8.83}$$

Note from the equation for r_{kl} that

$$r_{kl}=r_{lk};$$

hence, R is a symmetric matrix. The error power can then be written as

$$\varepsilon^2=\sigma^2[1-2A^T G+A^T RA]. \tag{8.84}$$

To optimize the predictor, the column matrix, A, must be selected such that ε^2 is a minimum. This is accomplished by taking the derivative of ε^2 with respect to A and equating the result to zero.

$$\left.\frac{\partial \varepsilon^2}{\partial A}\right|_{A=A_{\text{opt}}}=0,$$

$$\frac{\partial \varepsilon^2}{\partial A}=2G+2RA=0.$$

Solving the latter equation for A yields

$$A=R^{-1}G. \tag{8.85}$$

The minimum mean-square value of the error signal for the interval of M samples, ε^2, is found by substituting the optimum, A, given by Eq. (8.85) in Eq. (8.84) for ε^2 and simplifying. The result is

$$\{\varepsilon^2\}_{\min}=\sigma^2(1-G^T R^{-1} G), \tag{8.86}$$

where σ^2 is the mean-square value of the input sequence over the interval of M samples.

For practical digital transmission the error samples and the predictor coefficients are quantized to the fewest possible levels. The receiver of the prediction system uses these data to reconstruct estimates of the sample sequence of the original signal. If care is not exercised, quantizing noise may accumulate in the sample sequence.

This difficulty can be simply illustrated. Consider the "open-loop" quantization of the error signal shown in Fig. 8.42. Let tildas represent quantized versions of the signals. The quantizing noise present in the reconstructed received signal is therefore

$$(s_i-\tilde{s}_i)=(e_i-\tilde{e}_i)+(\hat{s}_i-\hat{\tilde{s}}_i),$$

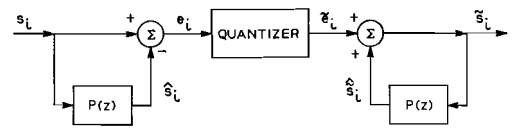


Fig. 8.42. Open-loop quantization of a predictor error signal

where

$$\hat{\tilde{s}}_i=\sum_{j=1}^{N} a_j \tilde{s}_{i-j}. \tag{8.87}$$

The quantizing noise in the received signal is not merely the same as the quantizing noise of the error signal, but also includes the quantizing error in the estimate. Since $\hat{\tilde{s}}_i$ is formed from a sum over N past samples the quantizing noise may accumulate.

One encoder arrangement commonly used to avoid this problem is a form generally identified as differential pulse code modulation (DPCM). This system uses feedback around the quantizer to prevent accumulation of quantizing noise. The encoder forms estimates of the input samples from a sample sequence reconstructed after the quantizer.

8.131. Predictive Quantizing; Differential Pulse Code Modulation

Predictive quantizing, or feedback around the quantizer, is a method used in a wide class of digital encoders for reducing the redundancy of a signal. The idea is to form an estimate of the sample of the input signal, and quantize the difference between the signal and its estimate. For accurate estimates, the variance of the difference, or error signal, is less than that of the input and fewer bits are required to transmit the error. Estimators typically include linear prediction networks (both adaptive and nonadaptive) and single or multiple integrators. Differential pulse code modulation (DPCM) and delta modulation (DM) are special cases of predictive quantizing, the latter using merely a 1-bit quantizer for the error signal.

Estimation or prediction of the signal requires knowledge of input signal statistics. In a nonadaptive predictor these data are built into a fixed feedback network. In adaptive prediction, the network is changed as the input signal changes its characteristics.

Digital transmission can be made relatively free of channel errors in well-designed systems. The controlling impairment is consequently noise introduced by the quantization process.

Fig. 8.43 shows a predictive quantizing system (R. A. McDonald). Input signal samples are s_i; the local (transmitter) estimate of the signal is $\hat{\tilde{s}}_i$; the error signal is e_i, which when quantized is $\tilde{e}_i$. The locally

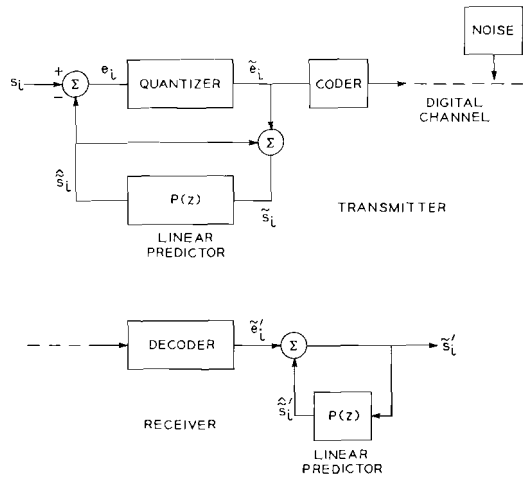


Fig. 8.43. Predictive quantizing system. (After R. A. McDonald)

reconstructed signal is $\tilde{s}_i=(\tilde{e}_i+\hat{\tilde{s}}_i)$. For transmission, $\tilde{e}_i$ is coded into a prescribed digital formant and transmitted. Any digital errors in transmission cause a corrupted version of the error signal, $\tilde{e}_i'$, to be received. Detection produces the reconstructed signal $\tilde{s}_i'$.

This type of differential quantizing has the important feature that the quantization noise in the reconstructed signal is the same as that in the error signal—that is, quantization noise does not accumulate in the reconstructed signal. Quantization noise samples are

$$\begin{aligned} q_i &= (e_i-\tilde{e}_i) \\ &= (s_i-\hat{\tilde{s}}_i-\tilde{e}_i) \\ &= (s_i-\tilde{s}_i). \end{aligned} \tag{8.88}$$

The quantization noise in the transmitted error signal is therefore identical to the quantization noise in the reconstructed signal.

A logical measure of the effectiveness of the predictor in reducing signal redundancy is the amount by which the power of the error signal is reduced below that of the input signal. This ratio is

$$\xi^2 \equiv \frac{E[s_i^2]}{E[e_i^2]} \tag{8.89}$$

where $E[x]$ denotes the expected value of x. To assess this figure one needs to know explicitly the predictor characteristics. Linear prediction represents a well known class of feedback networks. For linear prediction

the signal estimate is formed from a linear combination of past values of the reconstructed input signal. That is,

$$\begin{aligned} \hat{\tilde{s}}_i &= \sum_{j=1}^{N} a_j \tilde{s}_{i-j}{}^{1} \\ &= \sum_{j=1}^{N} a_j [s_{i-j}-(e_{i-j}-\tilde{e}_{i-j})] \\ &= \sum_{j=1}^{N} a_j s_{i-j} - \sum_{j=1}^{N} a_j q_{i-j} \end{aligned} \tag{8.90}$$

for an N-th order predictor. The variance of the error signal is

$$E[e_i^2]=E[(s_i-\hat{\tilde{s}}_i)^2]. \tag{8.91}$$

If the correlation between error samples is vanishingly small (i.e., if the power spectrum of the error is uniform) and if the correlation between input and error samples is negligible, then

$$E[e_i^2] \cong E\left[\left(s_i-\sum_{j=1}^{N} a_j s_{i-j}\right)^2\right] + E[q_i^2]\sum_{j=1}^{N} a_j^2. \tag{8.92}$$

For a given signal, therefore, maximizing ξ^2 is equivalent to minimizing $E[e_i^2]$. Differentiation of $E[e_i^2]$ with respect to a_j and setting the resulting equations to zero gives

$$\begin{aligned} \rho_1 &= (1+1/R)\,a_1 + a_2\rho_1 + a_3\rho_2 + \cdots + a_N\rho_{N-1} \\ \rho_2 &= a_1\rho_1 + (1+1/R)\,a_2 + a_3\rho_1 + \cdots + a_N\rho_{N-2} \\ &\vdots \\ \rho_N &= a_1\rho_{N-1} + a_2\rho_{N-2} + a_3\rho_{N-3} + \cdots + (1+1/R)\,a_N, \end{aligned} \tag{8.93}$$

where $R=E[s_i^2]/E[q_i^2]$ is the signal-to-quantizing noise ratio, and $\rho_j = E[s_i s_{i-j}]/E[s_i^2]$ is the signal autocovariance. The minimum of $E[e_i^2]$ can be written (R. A. McDonald)

$$E[e_i^2]|_{\min} = E[s_i^2]\left[1-\sum_{j=1}^{N} a_j\left(\frac{\rho_j}{(1+1/R)}\right)\right],$$

so that

$$\xi^2|_{\max} = \left[1-\sum_{j=1}^{N} a_j\rho_j/(1+1/R)\right]^{-1}. \tag{8.94}$$

The quantization noise power $E[q_i^2]$ depends upon properties of the quantizer. For example, for a linear quantizer of L steps, of step size

[1] The absence of an a_0 term implies delay around the loop.

Δ_l and step probability P_l, the quantizing noise power can be shown to be (CARLSON)

$$E[q_i^2]=\sum_{l=1}^{L} P_l \frac{\Delta_l^2}{12}. \tag{8.95}$$

For relatively fine quantizing, the quantizer noise is negligible compared to other terms in $E[e_i^2]$.

Historically, a commonly-used feedback network in DPCM systems is a simple integrator or accumulator. For this case $N=1$ and

$$a_1=1,$$
$$a_j=0, \quad j\neq 1$$

and

$$\hat{\tilde{s}}_i=\sum_{j=1}^{\infty} \tilde{e}_{i-j}$$
$$e_i=s_i-(\hat{\tilde{s}}_{i-1}+\tilde{e}_{i-1}). \tag{8.96}$$

The error power from (8.92) is

$$E[e_i^2]=E[s_i^2][2(1-\rho_1)]+E[q_{i-1}^2]. \tag{8.97}$$

Neglecting the quantizing noise,

$$\xi^2 \cong \frac{1}{2(1-\rho_1)}. \tag{8.98}$$

The optimum $N=1$ predictor (in the least error power sense) is however

$$a_1=\frac{\rho_1}{(1+1/R)},$$

for which

$$\xi^2=\frac{1}{(1-\rho_1^2)}. \tag{8.99}$$

The optimum predictor therefore shows a slight advantage (for the case $N=1$) over the simple ideal integrator (R. A. MCDONALD).

Computer studies on speech show that DPCM with a fixed linear predictor network optimized according to the preceding discussion gives approximately $\xi^2=10$ dB. Over 9 dB of this improvement is achieved by an $N=2$ optimum predictor. Compared to a straight PCM encoding, this means that 1 to 2 bits per sample may be saved in the encoding.

Predictive coding and quantizing has been applied in several forms to the digital transmission of speech. Optimum nonadaptive linear predictors for speech have been studied to reduce the bit rate for transmission below that of conventional PCM (R. A. MCDONALD; HASKEW; FUJISAKI). Adaptive predictive coding has also been used in which the predictor is designed to represent the pitch of voiced sounds and the shape of the signal spectrum (ATAL and SCHROEDER; J. M. KELLY *et al.*). Predictive quantizing can be implemented with adaptive quantization as well as with adaptive prediction.

8.132. Adaptive Predictive Coding

Adaptive predictive coding has been used to reduce signal redundancy in two stages: first by a predictor that removes the quasi-periodic nature of the signal, and second by a predictor that removes formant information from the spectral envelope (ATAL and SCHROEDER). The first predictor is simply a gain and delay adjustment, and the second is a linear combination of past values of the first predictor output. The equivalent operations are shown in Fig. 8.44, where

$$P_1(z)=\alpha z^{-k}$$
$$P_2(z)=\sum_{j=1}^{N} a_j z^{-j} \tag{8.100}$$
$$P(z)=\{P_1(z)+P_2(z)[1-P_1(z)]\}.$$

This predictor is used in the DPCM encoder form with a two-level (1 bit) quantizer for the error signal, as shown in Fig. 8.45. The quantizer level is variable and is adjusted for minimum quantization noise power in the error signal. The quantizer representation level Q is set to the average absolute value of the error samples being quantized, i.e.,

$$Q=\frac{1}{N}\sum_{j=1}^{N}|e_j|. \tag{8.101}$$

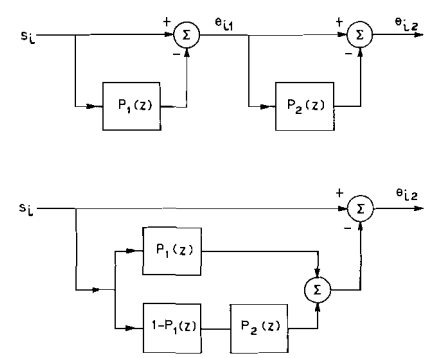


Fig. 8.44. Two stage predictor for adaptive predictive coding. (After ATAL and SCHROEDER)

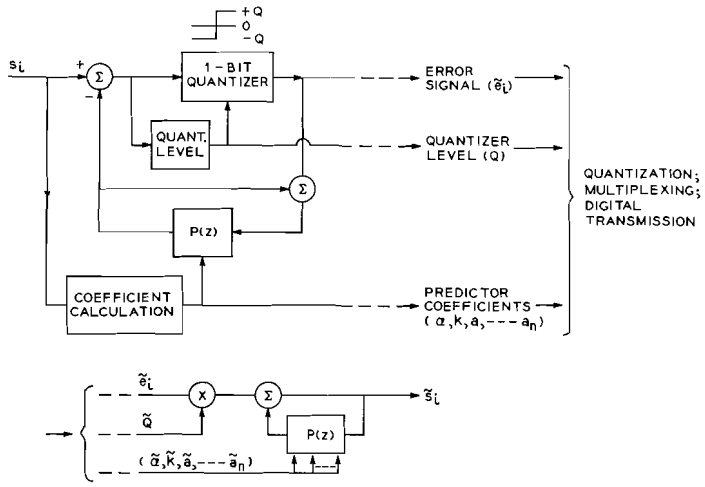


Fig. 8.45. Adaptive predictive coding system. (After ATAL and SCHROEDER)

The coefficients for predictor $P_2(z)$ are calculated as described previously. Those for $P_1(z)$, i.e., α and k, are obtained by minimizing the error power from the first predictor

$$\varepsilon_1^2 = \sum_{j=1}^{N} (s_j - \alpha s_{j-k})^2. \tag{8.102}$$

The minimum is given by

$$\alpha = \sum_{j=1}^{N} (s_j s_{j-k}) \Big/ \sum_{j=1}^{N} s_{j-k}^2 \Bigg|_{k=\text{optimum}}, \tag{8.103}$$

where the optimum k maximizes the normalized correlation

$$\rho = \sum_j s_j s_{j-k} / \{\sum_j s_j^2 \sum_j s_{i-k}^2\}^{\frac{1}{2}}. \tag{8.104}$$

The optimum k is found by a search of computed and tabulated values of ρ.

One implementation of the predictive system has been made for digital transmission at 9600 bps and at 7200 bps (J. M. KELLY *et al.*). The system was optimized in extensive computer-simulation studies. It used the following parameters and quantization to achieve digital transmission at 9600 bps: signal bandwidth = 2950 cps; sampling rate = 6 kcps; prediction optimization interval = 10 msec (N = 60 samples); $P_1(z)$ predictor quantization: $\alpha = 3$ bits, $k = 7$ bits (determined by maximum delay of 20 msec, or 120 samples at 6 kcps, for the computation of ρ); quantizer level = 4 bits; four $P_2(z)$ coefficients at 5 bits each; error

signal = 60 bits/frame (i.e. 60 samples at 6 kcps); parameter normalization = 2 bits (to normalize the $P_2(z)$ coefficients to a range of ± 1 for quantizing accuracy). The transmission coding therefore included a total of 96 bits/frame and a frame rate of 100 sec^{-1}, for a total bit rate of 9600 bps. By sampling at a slower frame rate, and using fewer predictor coefficients [for $P_2(z)$] and fewer bits for the error signal, the total bit rate could be reduced to 7200 bps.

In subjective tests it was found that the 9600 bps predictive coding is equivalent in quality to 4.5 bit log PCM, corresponding to a signal-to-quantizing noise ratio of 16.9 dB. At 7200 bps, the predictive coder was found equivalent in quality to 4.1 bit log PCM, with a corresponding signal-to-quantizing ratio of 14.7 dB. Sensitivity to digital errors in the transmission channel was also studied. Resulting error rates and associated qualities were found to be: 10^{-3} and lower, satisfactory; 10^{-2}, marginal performance; 10^{-1}, unacceptable (J. M. KELLY *et al.*).

8.14. Delta Modulation

Considerable interest attaches to realizing the advantages of digital transmission in economical ways. Multi-bit quantizers, such as used in PCM, are relatively expensive. In telephone communication they normally are not dedicated to individual customers, but typically are shared in time-division multiplex. This requires individual analog transmission to a central point where the digitizing occurs.

In many instances it is desirable to digitize the signal immediately at the source (for example, in some rural telephone systems). Inexpensive digital encoders which can be dedicated to individual customers are therefore required. Delta modulation is one solution.

Delta modulation (DM) may be considered perhaps the simplest form of DPCM. Quantization of the error signal is to one-bit only (i.e., a simple comparator), and a single or double integrator is typically used as the predictor network, as shown in Fig. 8.46a. The transmitted binary samples, $\tilde{e}_i$, are either $+1$ or -1 and represent the sign of the error, $e(t)$. The integrator can be implemented many ways, including a simple analog storage capacitor. A digital implementation, using the terminology employed in the earlier discussion of predictive quantizing, is shown in Fig. 8.46b. The box T is a one-sample delay and $a_1 = 1$ for an ideal integrator. A sample-and-hold converts the discrete samples to a continuous function.

The local estimate provided by the integrator, $\hat{\tilde{s}}(t)$, is the staircase function shown in Fig. 8.47. The step size of the staircase function is determined by the amplifier constant, k. The step-size is typically chosen

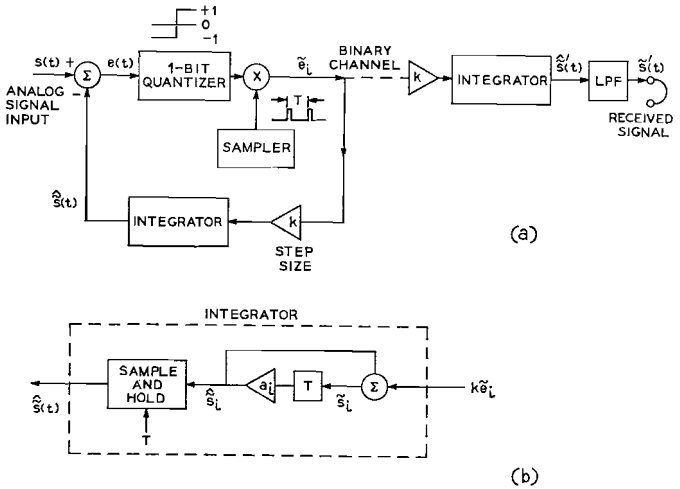


Fig. 8.46a and b. Delta modulator with single integration

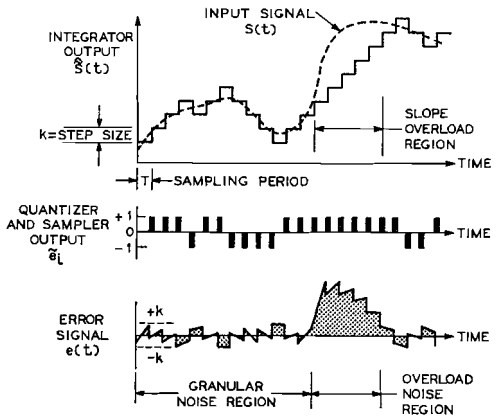


Fig. 8.47. Waveforms for a delta modulator with single integration

small compared to the input signal magnitude. Two types of distortion can occur in the estimate—granular distortion and slope overload. The former is determined by the step size of the quantization (that is, by the amplifier k). The latter is caused by the inability of the encoder to follow the signal when its slope magnitude exceeds the ratio of step size to sampling period,

$$|\dot{s}| > k/T. \tag{8.105}$$

These two types of distortion are indicated in Fig. 8.47.

Granular distortion can be made small by using a small step size. Slope overload can be reduced by using a large step size or by running the sampler (clock) faster. The latter of course increases the transmitted bit rate. In typical designs, for a prescribed bit rate, the step size is selected to effect a compromise "mix" between quantizing distortion and slope overload. Perceptually, more overload noise power is tolerable than granular noise power (JAYANT and ROSENBERG). During granular distortion the samples of the error signal tend to be uncorrelated and the error signal power spectrum tends to be uniform.

For high-quality speech transmission, say with signal-to-noise ratio of the order of 40 dB, the resulting bit rate for simple DM is relatively high, typically greater than 200 Kbps. Tolerable channel error rates are typically 10^{-4}. The signal-to-quantizing noise present in the received signal is strongly dependent upon the final low-pass filter. If simple low-pass filters are used for desampling, the transmission bit rate must be pushed into the Mbps range to achieve high quality. Such high bit rates cannot be supported in many transmission facilities. Consequently, there is strong interest in techniques for reducing the bit rate of DM while at the same time retaining most of its advantages in circuit simplicity. Adaptive delta modulation (ADM) is one such solution.

In ADM the quantizer step size is varied according to a prescribed logic. The logic is chosen to minimize quantizing and slope distortion when the sampler is run at a relatively slow rate[1]. The additional control is typically effected by a step size multiplier incorporated in the feedback loop, as shown in Fig 8.48. As in simple DM, the feedback network may be a single or double integration. The step control logic may be discrete or continuous (JAYANT, GREEFKES, DE JAGER, ABATE), and it may act with a short time constant (i.e., sample-by-sample)

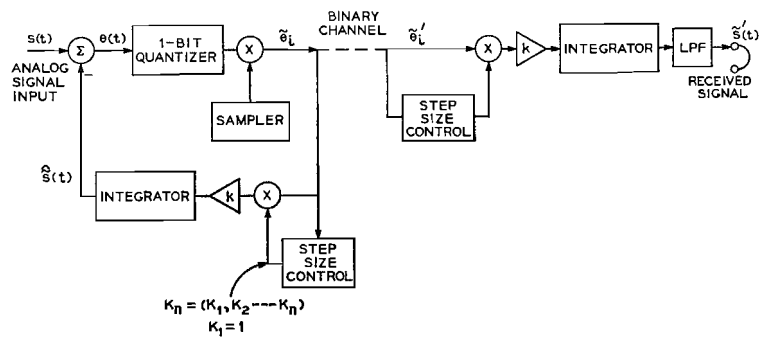


Fig. 8.48. Adative delta modulator with single integration

[1] Adaptation is normally not applied to the feedback network, but this is an attractive possibility for further improvement in the encoding.

or with a time constant of syllabic duration (GREEFKES, TOMASAWA). Normally the step size is controlled by information contained in the transmitted bit stream, but it may be controlled by some feature of the input signal; for example, the slope magnitude averaged over several msec (DE JAGER). In this case, the control feature must be transmitted explicitly along with the binary error signal.

The receiver duplicates the feedback (predictor) branch of the transmitter, including an identical step size control element. In the absence of errors in the transmission channel, the receiver duplicates the transmitter's estimate of the input signal. Desampling by a low-pass filter to the original signal bandwidth completes the detection.

The manner in which discrete adaptation is implemented is illustrated in Fig. 8.49. As long as the slope of the input signal is small enough that the signal can be followed with the minimum step size, k, the multiplier is set to $K_n = K_1 = 1$. When the input signal slope becomes too great, the step size multiplier is increased to permit more accurate following and to minimize slope overload. In the logic illustrated, an increase in step size is made whenever three successive samples of $\tilde{e}_i$ have the same polarity. At the point of greatest input signal slope, a step multiplication by K_3 is attained. Further increases can be accomplished in successive samples, if needed, until a maximum multiplication of K_N is achieved. Any situation where the current channel bit and the past two bits are not the same results in a reduction in step size. Reductions can likewise be accomplished successively until the minimum value $K_n = K_1 = 1$ is again attained.

Exponential adaptation logics have been found valuable for speech encoding (JAYANT). In this case, the multiplier is typically $K_n = P^{n-1}$,

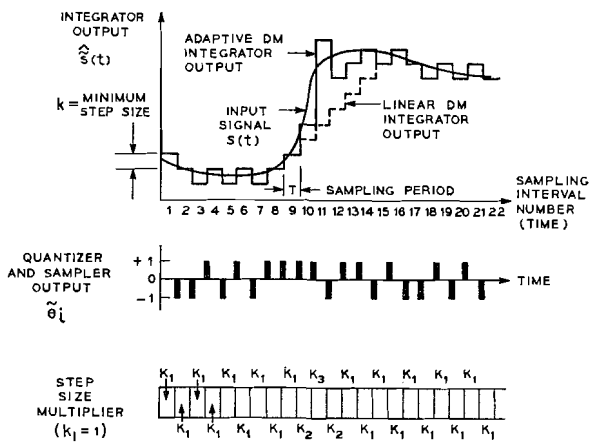


Fig. 8.49. Waveform for an adaptive delta modulator with discrete control of the step size

$n = 1, \ldots, N$. A typical value of P is in the order of 1.5 to 2.0. As few as eight ($N = 8$) discrete multiplier values are found adequate in some applications of exponential ADM.

Because of the ability to "shift gears" automatically, ADM can be designed to yield signal quality comparable to 7-bit log PCM at bit rates commensurate with PCM; typically, 56 Kbps for a 4 KC signal band. At lower bit rates, ADM can surpass PCM in signal-to-noise (S/N) performance. This relation results because S/N for ADM varies roughly as the cube of the sampling rate. For PCM the growth in S/N is 6 dB/bit of quantizing. At low bit rates, ADM wins out. However, the range of normally useful quality is restricted to rates greater than 20 Kbps. A S/N comparison is shown for ADM and PCM in Fig. 8.50.

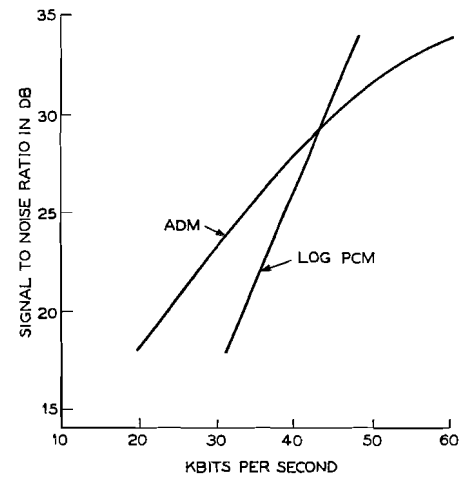


Fig. 8.50. Signal-to-noise ratios as a function of bit rate. Performance is shown for exponentially adaptive delta modulation (ADM) and logarithmic PCM. (After JAYANT)

Because delta modulators can be implemented very economically in digital circuitry, they constitute an attractive means for initial analog-to-digital conversion of signals. However, other formats of digital encoding are frequently used in digital communication systems. Techniques for direct digital conversion from one signal format to another, with no intervening analog detection, are therefore of great interest. Present work in digital communication includes direct digital transformation between simple DM, ADM, linear PCM, log PCM, and DPCM (FLANAGAN and SHIPLEY; GOODMAN; GOODMAN and FLANAGAN). These and related studies aim to establish coding relations which make the transmission system and switching network "transparent" to the signal, regardless of its digital form.