B. Gold

A History of Vocoder Research at Lincoln Laboratory

During the 1950s, Lincoln Laboratory conducted a study of the detection of pitch in speech. This work led to the design of voice coders (vocoders)—devices that reduce the bandwidth needed to convey speech. The bandwidth reduction has two benefits: it lowers the cost of transmission and reception of speech, and it increases the potential privacy. This article reviews the past 30 years of vocoder research with an emphasis on the work done at Lincoln Laboratory.

If I could determine what there is in the very rapidly changing complex speech wave that corresponds to the simple motions of the lips and tongue, I could then analyze speech for these quantities, I would have a set of speech defining signals that could be handled as low-frequency telegraph currents with resulting advantages of secrecy, and more telephone channels in the same frequency space as well as a basic understanding of the carrier nature of speech by which the lip reader interprets speech from simple motions.

—Homer Dudley, 1935

Historical Background

More than 200 years ago, the Hungarian inventor W. von Kempelen was the first to build a mechanical talking contrivance that could mimic, albeit in a rudimentary fashion, the various sounds of human speech [1, 2]. The device, constructed around 1780, proved that human speech production could be replicated.

The next major development in the field was undoubtedly the most important: Alexander Graham Bell's invention of the telephone. Although there is no need to chronicle the well-known events that led to the invention, nor any need to discuss its subsequent impact on human communications, it is interesting to note that Bell's primary profession was that of a speech scientist, i.e., someone with knowledge of the intricacies of the human vocal apparatus. Reference 3, which contains a description of Bell's Harp telephone, reveals the inventor's keen understanding of the rudiments of the spectral envelope of speech.

But the telephone that Bell invented was basically a transmitter of waveforms. In fact, telephone technology to date has been almost totally oriented toward transmission while the subject of speech modeling has remained of peripheral practical interest, in spite of years of research by engineers and scientists from Lincoln Laboratory, Bell Laboratories, and other research centers.

Modern methods of speech processing really began in the United States with the development of two devices: the voice operated demonstrator (voder) [4] and channel voice coder (vocoder) [5]. Both devices were pioneered by Homer Dudley, whose philosophy of speech processing is quoted at the beginning of this article.

The appearance of the voder at the 1939 World's Fair in San Francisco and New York City elicited intense curiosity. The device was controlled by a human operator who operated a console similar to a piano keyboard (Fig. 1). The keys on the keyboard switched the appropriate bandpass filters into the system; by depressing two or three of the keys and setting the wrist bar to the buzz (i.e., voicing) condition, an operator could produce vowel and nasal sounds. If the wrist bar was set to the hiss (i.e., voiceless) condition, sounds such as the voiceless fricatives could be generated. Special keys were used to produce the plosive and affricative sounds (the *ch* in *cheese* and *j* in *jaw*), and a pitch pedal was used to control the intonation of the sound.

While the voder was the first electronic synthesizer, the channel vocoder was the first

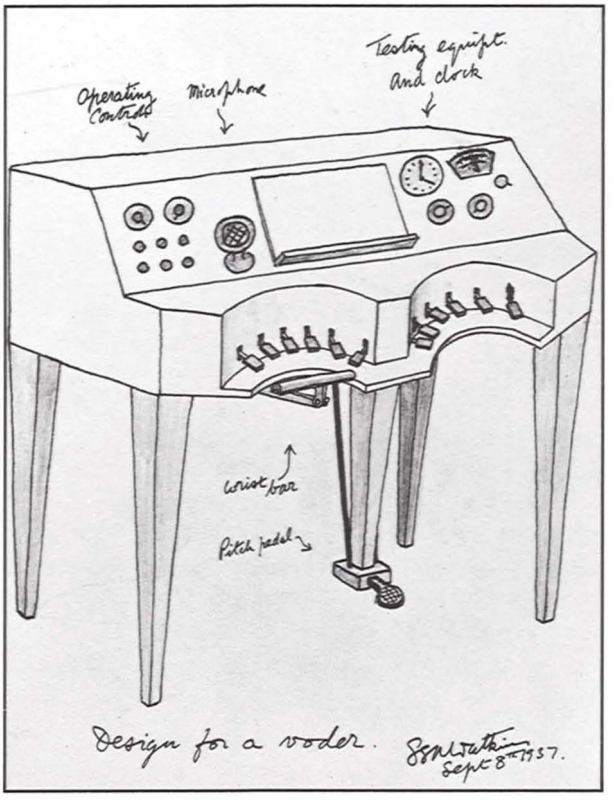


Fig. 1—Original sketch of voder by S.W. Watkins (8 September 1937).

analysis-synthesis system: that is, the channel vocoder derived certain parameters from a speech wave, and the parameters were then used to control a synthesizer that reproduced the speech.

To paraphrase Dudley, vocoders could lead to advantages of more secure communications, and a greater number of telephone channels in the same frequency space. Both of Dudley's predictions were correct, but the exact ways in which they came to pass (or are coming to pass) probably differed from what he imagined.

In 1929, digital communications was unknown. When digitization did become feasible, researchers realized that the process could provide a means of making communications less vulnerable to eavesdropping. Digitization, however, required wider transmission bandwidths. For example, the current 3-kHz bandwidth of a local telephone line cannot transmit a pulse-coded modulation speech signal that is coded to 64 kb/sec (the present telephone standard). Thus the channel vocoder was quickly recognized as a means of reducing the speech bit rate to some number that could be handled through the average telephone channel. Eventually the military set the standard rate at 2.4 kb/sec.

To understand Dudley's concept of more telephone channels in the same frequency space, it is important to realize that human speech production depends on relatively slow changes in the vocal-tract articulators such as the tongue and the lips. Thus, if we could develop accurate models of the articulators and estimate the parameters of their motion, we could create an analysis-synthesis system that has a low data rate. Dudley managed to bypass the difficult task of modeling the individual articulators by realizing that he could lump all articulator motion into one time-varying spectral envelope. Furthermore, Dudley understood that, to a first approximation, the excitation produced by vocal-cord vibrations (for voiced sounds) and regions of vocal-tract constriction (for consonants) could be separated from the actual speech sounds that are produced by the motion of the tongue, the lips, and the participation of the nasal passage. Both the spectrum envelope and excitation parameters must vary slowly (because the articulators themselves vary slowly) and thus transmission of these derived parameters requires only several hundred hertz of bandwidth rather than the usual 3 to 10 kHz needed to transmit speech via normal telephony methods.

In an interesting aside, an informative and entertaining paper by W.R. Bennett [6] gives a historical survey of the X-System of secret telephony that high-level Allied personnel used extensively during World War II. For example, during the invasion of Normandy in 1944, the system was used for communications between London and Washington. Details about X-System have only recently become declassified, and from that information X-System appears to have been a sophisticated version of Dudley's channel vocoder. X-System included such features as pulse-coded modulation transmission, logarithmic encoding of the channel signal, and, of course, enciphered speech.

Advances in Telephony

In an article that appeared in *National Geographic* magazine [7], F.B. Colton gives an engrossing account of the history of telephone transmission. Figure 2, which is from that article, shows the telephone wires on lower Broadway in New York City in 1887. It is clear that progress in telephony could easily have been brought to a halt if not for such improvements as underground cables and multiplexing techniques. At present, fiber optical transmission and lasers are allowing another great leap forward in capacity as telephone traffic continues to increase.

Such a technological advance brings to mind the following question: Will optical fibers with their enormous bandwidth make vocoding technology obsolete? The author believes the answer is no because of the great potential growth of radio telephones, e.g., cellular phones. The permissible bandwidths of such devices are limited by nature; thus it seems certain that the growth of cellular phone traffic will require vocoding techniques.

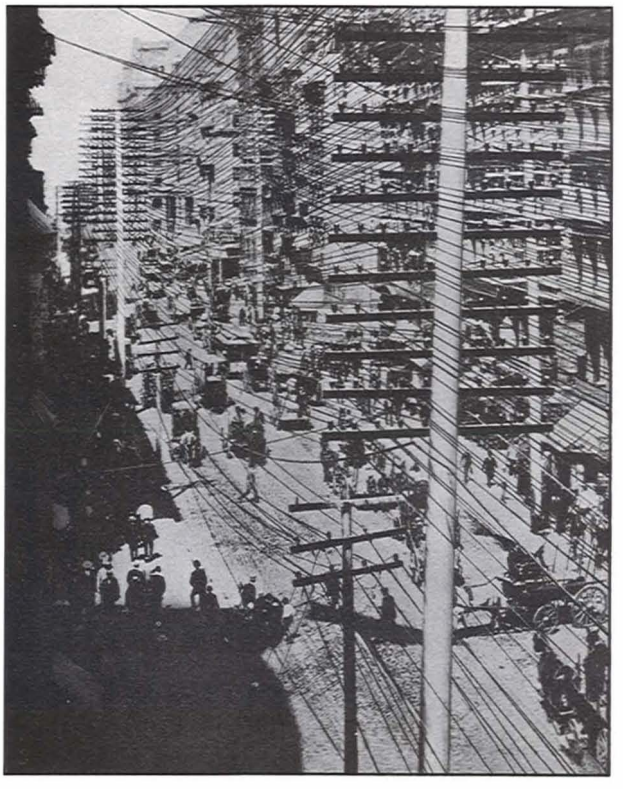


Fig. 2—Telephone wires on lower Broadway in New York City in 1887. (Courtesy of National Geographic.*)*

Work at Lincoln Laboratory

Lincoln Laboratory's history in vocoding commenced in the 1950s with extensive research in pattern recognition, which led to work on detecting the pitch of speech. Some of the Laboratory's contributions in vocoding include the

1. development of pitch-detection algorithms and hardware that have served as standards for almost 30 years,
2. development of the first vocoder hardware for satellite and aircraft narrowband speech communications,
3. creation of the first real-time computer programs that used linear predictive coding (LPC) and homomorphic vocoding,
4. use of vocoders for packet speech communications,
5. use of high-frequency (HF) channels for vocoding,
6. invention of a novel analysis-synthesis system (the Sine Transform Coder), and
7. introduction of methods that allow vocoder research and development to be carried out in real time on general-purpose computers.

This article discusses the first, second, third, and seventh contributions, and gives references to the remaining items.

The Vocoder

Figure 3 is a schematic diagram of a vocoder. In the left of the figure, the vocal-source analyzer finds the correct time-varying excitation parameters of the input speech (discussed in the following section, "Pitch Detection"). A parameter is modeled as either a quasi-periodic pulse train or a noise source. The vocal-tract analyzer finds the time-varying shape of the spectral envelope of the input speech by means of an appropriate algorithm (discussed in the subsequent section, "Speech Synthesis"). The excitation parameters and spectral envelope are then transmitted and the received signals are used to

control a speech-synthesis system, as shown in the right portion of Fig. 3.

Pitch Detection

In the late 1950s, vocoder researchers realized that the lack of an effective pitch detector was a major stumbling block to adequate vocoding. Pitch detection is defined in the following way. A speech signal can be modeled as the output of a time-varying linear filter that is excited by one or more source functions. The human vocal tract acts as the time-varying filter and the excitation signals derive from either the quasi-periodic vibrations of the vocal cords, the turbulence created by points of stricture, or the pressure buildup and sudden discharge at some point of closure in the vocal tract. Thus analysis of the speech signal involves finding parameters related to both the vocal-cord and vocal-tract models. The vocal-cord vibrations control the time-varying fundamental frequency of the speech signal. The analysis of this parameter is referred to as *pitch detection.*

Early research at Lincoln Laboratory used a parallel-processing pattern-recognition algorithm to model the pitch-detection process. (Details of the research will be discussed later.) This section begins with an overview of pitch detection, discusses some of the difficulties, summarizes the connections between human pitch perception and automatic pitch detection, and reviews the ideas that have been applied to the latter problem.

Overview of Pitch Detection

During the 1950s, Lincoln Laboratory was researching pattern-recognition algorithms for automatic Morse code detection and the automatic recognition of handwritten alphanumerics. One of the interesting concepts in this area was the use of parallel processing to increase the probability of making the right decision. During the same time period, the computer group at Lincoln Laboratory developed the TX-2 computer, a powerful and highly interactive machine that greatly facilitated advanced research on difficult pattern-recognition problems. This fortuitous symbiosis between the parallel-processing concept and the TX-2 computer led to an innovative contribution to the art of pitch detection and created the environment for Lincoln Laboratory's entry into the vocoder field.

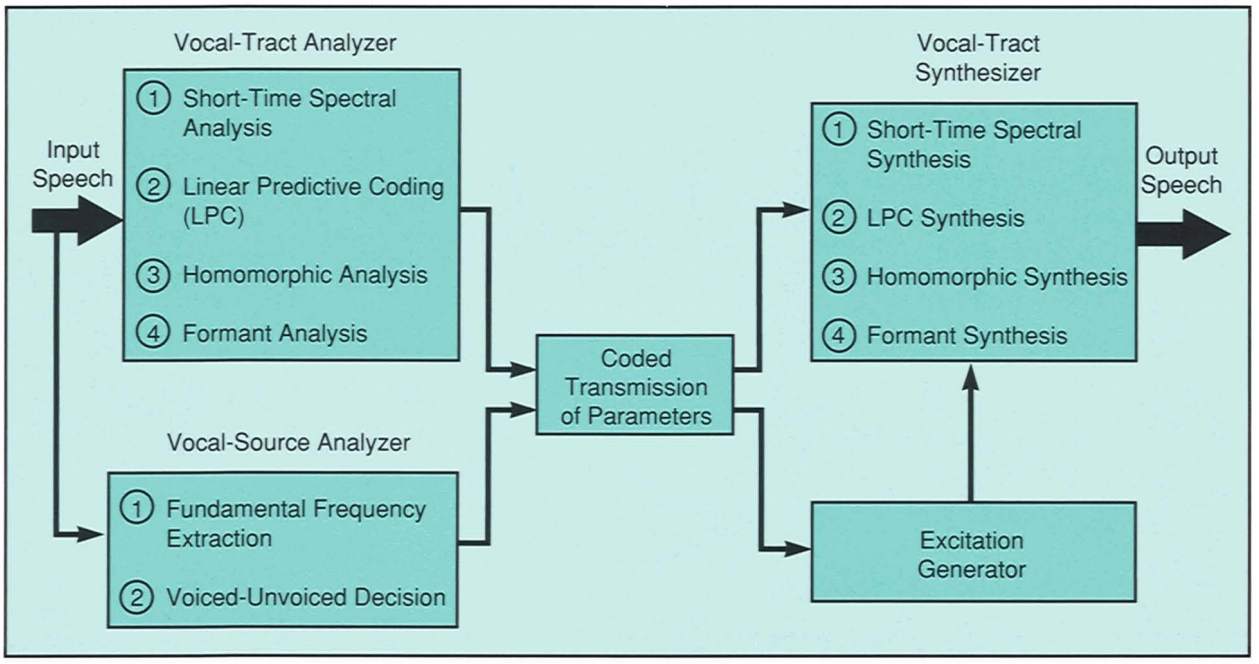


Fig. 3—Schematic of a vocoder.

In the music world, pitch is the fundamental frequency of a note, a definition that is not ambiguous as long as the note consists of evenly spaced harmonics. But what if the signal contains aharmonic frequencies? In addition, the perception of pitch by people (see the box, "Human Pitch Perception: The Debate over Place versus Periodicity") is a function of various parameters such as intensity and duration as well as fundamental frequency. For these reasons, audiologists prefer to define pitch as a perceptual entity. For the purposes of this article, however, the terms pitch and voice fundamental frequency will be used interchangeably. Thus pitch can be interpreted as the response of a device called a pitch detector to a speech signal. And correct pitch corresponds to speech synthesis during which no error is detected by the human auditory system.

All speech sounds require the positioning of the human vocal cords. During voiced sounds (e.g., vowels and nasals), the vocal cords vibrate quasi-periodically; during the voiceless sounds, the vocal cords are open. The vibrations of the vocal cords create quasi-periodic pressure waves that can be modeled as the input to a time-varying linear filter (i.e., the vocal tract). When a speaker wishes to stress a word or a syllable, the speaker's pitch will usually rise; that is, the frequency of the vocal-cord vibration will increase. Speech analysis-synthesis systems depend greatly on the pitch detector's ability to distinguish between voiced and unvoiced sounds, and its ability to determine the time-varying fundamental frequency of the voiced sounds.

Signal Processing, Pattern Recognition, and Editing

Insight is gained by studying the mechanism of human speech production. During voicing, the vocal cords open and close at the fundamental frequency. When we look at a speech oscillogram we see this periodicity, but the precise instants at which opening and closing take place are generally masked by the filtering effects of the vocal tract. It is thus appealing to investigate ways of processing the speech signal so as to bring into evidence some obvious event that signifies the start of a period. This strategy leads to approaches that are best categorized as signal processing.

After signal processing, it is necessary to make explicit voiced-unvoiced (or buzz-hiss) decisions and to assign specific pitch numbers to the voicing periods. (The pitch numbers denote either the periods or frequencies of the pitches.) This part of the pitch-detection process can rightly be called pattern recognition.

Lastly, we can use *a priori* knowledge to fix mistakes. Such knowledge includes information about the range of pitch values to be detected and the observation that the duration of voicing rarely lasts less than 50 msec.

Thus pitch detection can be divided into three cascaded operations: signal processing, pattern recognition, and editing.

Difficulties of Estimating Pitch and Making the Voiced-Unvoiced Decision

A basic task of vocoder systems is the separation of speech into a source and a vocal-tract filter. The separation implies that the pitch and spectrum can be individually manipulated prior to synthesis. Exciting the synthesizer with a pure hiss source can result in a sound like whispered speech; exciting with constant-period pulses produces a strong monotone effect. Even a single error, such as a missing or added pulse, can be perceived if it occurs during a steady-state vowel. If the error occurs near the beginning or end of voicing, however, the error may not be perceived [8, 9]. Excessive smoothing of the measured fundamental frequency creates a monotone effect.

The fundamental frequency of an adult voice can vary from 60 to 600 Hz. The large range often results in the pitch estimate being off by a factor of two to three. Such errors are annoying to listeners, but even more annoying is a hiss decision during voicing. The inverse of that error (a buzz decision during hissing) appears, on the average, to be less annoying.

In developing an algorithm for voiced-unvoiced decision making, one should note that most English sounds are pure buzz (quasi-

Human Pitch Perception: The Debate over Place versus Periodicity

Although the automatic extraction of the fundamental frequency of speech is a relatively recent discipline, the study of human pitch perception has been a lively field for many years. E. de Boer wrote an excellent review of the subject [1].

H. von Helmholtz [2] conceived of the auditory system as a bank of many overlapping bandpass filters. Near the entrance to the cochlea, the membrane and associated hair cells respond to high frequencies. As the vibrations penetrate more deeply into the cochlea (Fig. A[1]), the response becomes more sluggish, corresponding to filters with lower center frequencies. Thus, for example, a pure tone would cause a specific place on the basilar membrane to vibrate most vigorously, which would lead to perception of that tone. Tones of different frequencies stimulate different places on the membrane. From an engineering point of view, this model corresponds to a filter bank that covers the audio range (Fig. A[2]). The pitch of the tone can be determined by locating those filters which contain energy.

Although the *place* theory described above supplies a credible explanation for pure tones, the theory has trouble accommodating the perception of complex periodic signals that consist of many harmonics. An important fact to note in these cases is that

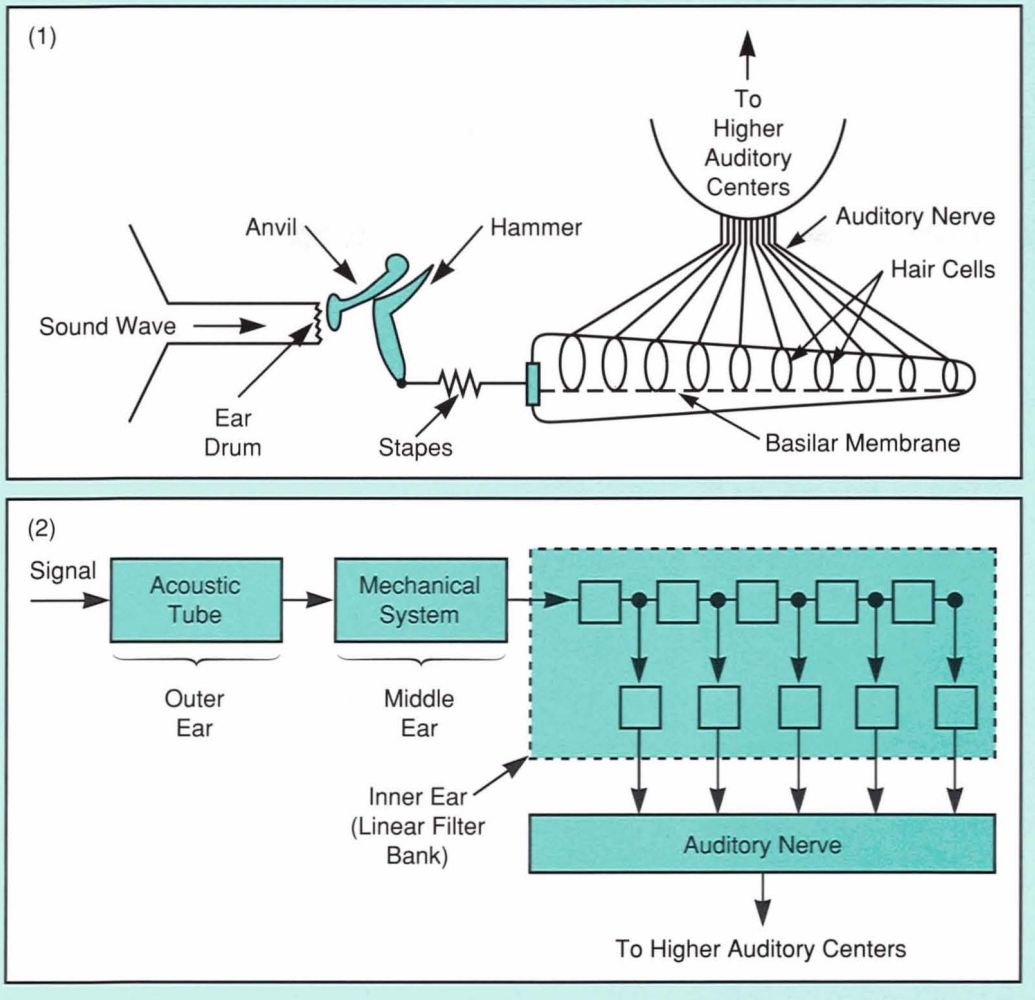


Fig. A—The auditory system: (1) schematic of the ear and (2) engineering model.

even when the fundamental frequency is missing, the perceived pitch is that of the fundamental frequency. Fletcher [3] referred to this phenomenon as the *missing fundamental*, while J.F. Schouten [4] named it the *residue*. Results of this sort have led to theories of human pitch perception as being influenced by the actual period of the signal rather than by a place mechanism. These theories are thus called *periodicity* theories.

For the perception of the missing fundamental to be explained by a place theory would require some nonlinear phenomenon in the ear to cause the place in the basilar membrane that corresponds to the fundamental frequency to vibrate despite the physical absence of the fundamental. J.C.R. Licklider [5] proved that the above did not happen via the following clever experiment: he alternately played a pure tone (at the fundamental) and a harmonic series of the same fundamental frequency but with the actual fundamental physically absent. The listener then perceived two sounds of equal pitch but different timbre. Then noise with a bandwidth centered at the fundamental was added to the above sequence and the following was discovered: the pure tone was completely masked while the harmonic series was still perceived to have the same pitch. If perception of the harmonic series were dependent on the combination tones appearing at the fundamental-frequency place in the basilar membrane, then the combination tones would have been masked, too.

Licklider's experiments cast a vote in favor of Schouten's theory of periodicity. The experiments, however, did not lead to an answer to the question "How does the auditory system work?"

Further insight was obtained from the experiments of A.J.M. Houtsma and J.L. Goldstein [6]. In Houtsma and Goldstein's first experiment, musically trained subjects were asked to recognize the *interval* between two successively played signals. (An interval is defined as the difference between the fundamental frequencies of two signals.) Each signal contained two successive harmonics of a given fundamental frequency. When the two signals were presented to both ears, the trained subjects had no trouble identifying the intervals.

The second experiment was a repeat of the first but with one notable exception; for each signal, only one harmonic was presented to one ear while the other harmonic was presented to the other ear. Again, pitch intervals were correctly identified. The result indicates that the perception of pitch is centrally located; that is, perception took place after the auditory signals from the two ears had been combined. What actually happens physiologically is still a mystery, but these experiments must now be incorporated into any proposed model. Also very significant is the fact that an altered version of the place theory sneaks back in. We can imagine that the appropriate places on the basilar membrane vibrate for all existing harmonics of the signal and the central processor, or brain, judiciously combines this knowledge to produce a pitch value.

A straightforward interpretation of Houtsma and Goldstein's results is that the ear and brain perform a high-resolution spectrum analysis followed by a pattern-recognition procedure to detect pitch based on this spectral structure. J.L. Goldstein [7] proposed a model of this sort and H. Duifhuis, L.F. Willems, and R.J. Sluyter [8] implemented the model to extract pitch from a speech wave.

References

1. E. de Boer, "On the 'Residue' and Auditory Pitch Perception," *Handbook of Sensory Physiology* **5,** ed. W. Keidel (Springer-Verlag, New York, 1976) pp. 479–583.
2. H. von Helmholtz, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik* (Vieweg, Braunschweig, 1862) (5th edition 1896); *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, 1st English ed. (1897), paperback ed. (Dover, New York, 1954).
3. H. Fletcher, *Speech and Hearing in Communications* (D. Van Nostrand, Princeton, NJ, 1953).
4. J.F. Schouten, "The Residue, a New Component in Subjective Sound Analysis," *Proc. Kon. Acad. Wetensch, The Netherlands* **43,** p. 356 (1940a).
5. J.C.R. Licklider, "Periodicity Pitch and Place Pitch," *J. Acoust. Soc. Am.* **26,** 945 (A) (1954).
6. A.J.M. Houtsma and J.L. Goldstein, "The Central Origin of the Pitch of Complex Tones: Evidence from Musical Interval Recognition," *J. Acoust. Soc. Am.* **51,** 520 (1972).
7. J.L. Goldstein, "An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones," *J. Acoust. Soc. Am.* **54,** 1496 (1973).
8. H. Duifhuis, L.F. Willems, and R.J. Sluyter, "Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception," *J. Acoust. Soc. Am.* **71,** 1568 (1982).

periodic signals) or pure hiss (noiselike signals). The exceptions are the voiced fricatives (e.g., the *v* in *van*, the *z* in *zero* and *azure*, and the *th* in *the*) and the voiced affricative (e.g., the *dg* in *edge*), in which the constriction creates aperiodic excitation while the vibrating vocal cords simultaneously generate quasi-periodic excitation. Thus the speech source (or excitation) function is

neither strictly pure buzz nor pure hiss. Nevertheless, most vocoder systems work on the premise that the voiced-unvoiced decision is binary. The binary assumption is based on the conviction (perhaps implicit) that the resultant vocoded speech quality is not perceived to degrade.

With good-quality speech in a reasonably quiet environment, the occasional errors made by sophisticated pitch- and voicing-detection algorithms are generally not a problem. Major difficulties, however, arise when the environment is less benign. For example, in vocoding speech spoken in a high-speed jet aircraft, the background noise is loud enough to cause many buzz-to-hiss errors. Also, when speech is passed through the carbon-button microphone of an ordinary telephone handset and then through a typical telephone channel, enough noise and distortion are added to produce substantial excitation errors.

The annoyance level of perceived pitch and voicing errors depends not only on the pitch estimator and voicing detector, but also on the vocoder synthesizer. (This point will be discussed in the section "Speech Synthesis.")

Figure 4 illustrates why automatic pitch detection is difficult. Figure 4(a) shows two speech waveforms: the top signal has a long period, about four times that of the bottom signal. The example illustrates the large dynamic range of the human voice's fundamental frequency. In fact, the pitch of some male voices can be as low as 60 Hz while the pitch of children's voices can be as high as 800 Hz. Any vocoder that is dedicated to vocoding all human voices must be able to process this order-of-magnitude pitch range. In many practical situations, however, the range can be greatly reduced. For example, the pitch of most male voices ranges from 100 to 200 Hz.

Figure 4(b) shows how the period of a speech signal can fluctuate drastically and almost instantaneously. Note that the leftmost periods of the curve are short, and the following periods are abruptly more than twice as long. The rightmost periods then snap back to shorter durations. This kind of behavior makes pitch tracking difficult.

Similarly, Fig. 4(c) shows a rapid change in the spectrum. For example, a sudden closure as in a vowel-to-nasal transition can cause such a change. Although the fundamental frequency in Fig. 4(c) has not changed drastically, pitch detection based on waveform analysis can suffer. On the other hand, a spectrally based pitch extractor would presumably not be sensitive to such a perturbation.

Figure 4(d) shows a transition region from aperiodic excitation (hiss) to quasi-periodic excitation (buzz). It would appear that a fast-acting time-domain pitch extractor would be best to catch the precise transition instant. Finally, Figs. 4(e) and 4(f) respectively show how telephone transmission and added acoustic noise can degrade the speech signal. It is appreciably more difficult to extract the pitch from such signals.

Modern Methods of Pitch Estimation (Signal Processing)

R.J. Ritsma [10] demonstrated that low frequencies dominate in the perception of pitch. Indeed, it seems established that human pitch perception of speech and probably for all periodic sounds is most strongly influenced by the frequencies below about 1500 Hz.

It is interesting to note that many pitch-detection algorithms use a low-pass-filtered version of the speech as their input. Thus vocoder developers, perhaps without being aware of Ritsma's result, found that similar regions of dominance hold for both pitch detection (by vocoder analysis) and pitch perception (by people).

Although low-pass filtering is effective in removing extraneous influences, it still requires a formidable pattern-recognition algorithm to extract pitch. This requirement has led researchers in the field to develop more complex signal processing methods, including spectral flattening and correlation [11], inverse filtering [12], and *cepstral analysis* [13].

The use of cepstral analysis as the signal processing portion of a pitch estimator arises from the algorithm's ability to deconvolute two signals. As discussed earlier, speech can be

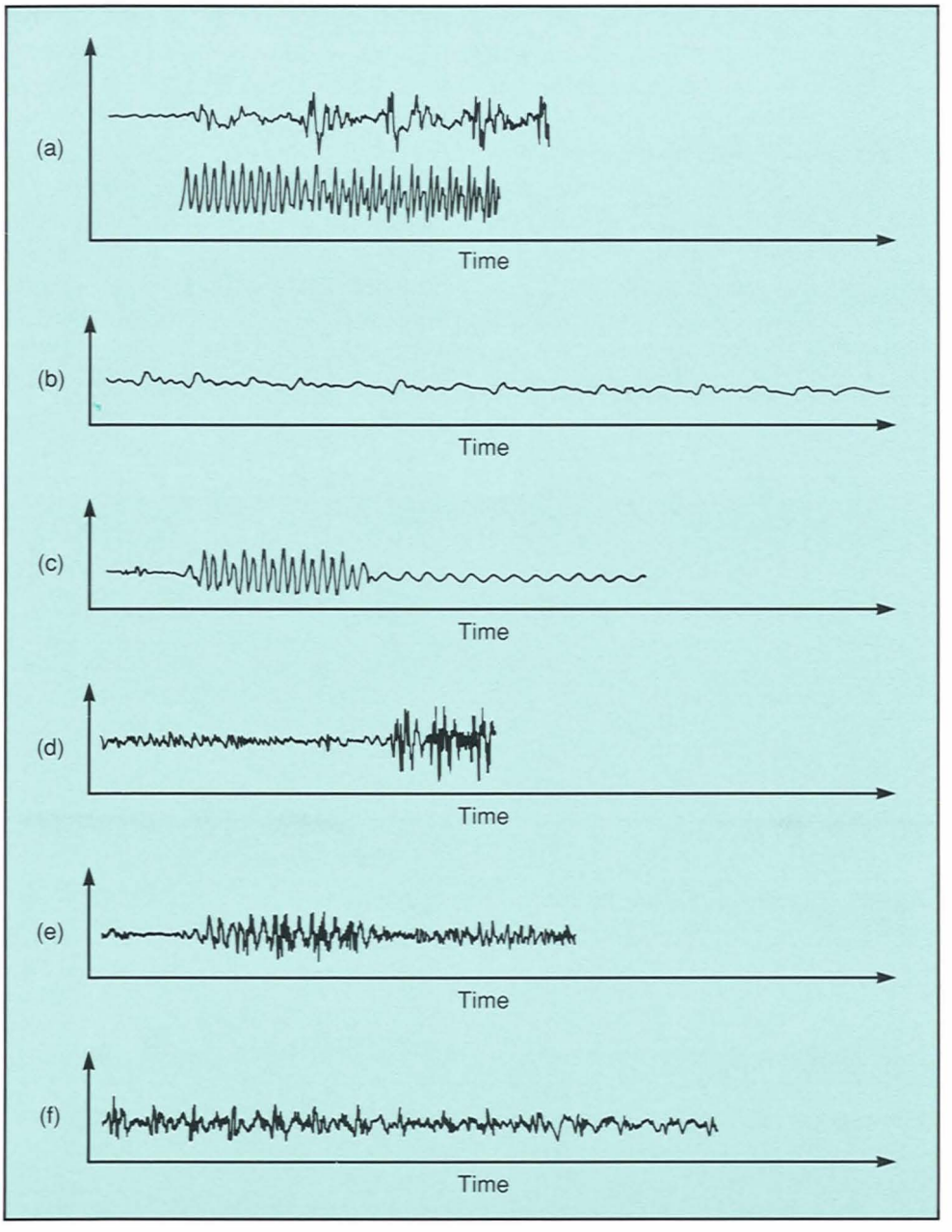


Fig. 4—The detection of pitch in speech is difficult because (a) the pitch of the human voice can range from 60 Hz for some adult males to 800 Hz for some children, (b) fluctuations in the pitch of a person speaking can be dramatic and abrupt, (c) variations in the vocal tract can cause sudden changes in the speech spectrum, (d) although the transition from voiced (hiss) to unvoiced (buzz) excitation might not alter the speech's fundamental frequency, the transition does greatly alter the waveform of the signal, (e) telephone transmission degrades the speech signal, and (f) other acoustic background noise can also degrade the speech signal.

modeled as the excitation function convolved with the vocal-tract function, and cepstral analysis can in principle separate these two functions.

It should be stressed that all vocoder algorithms incorporate the concept of deconvolution because any technique that can separate out the excitation function from the vocal-tract (or spectral) function is, in a sense, performing deconvolution. The unique property of cepstral analysis is that deconvolution is attained almost entirely through linear filtering and a simple memory-

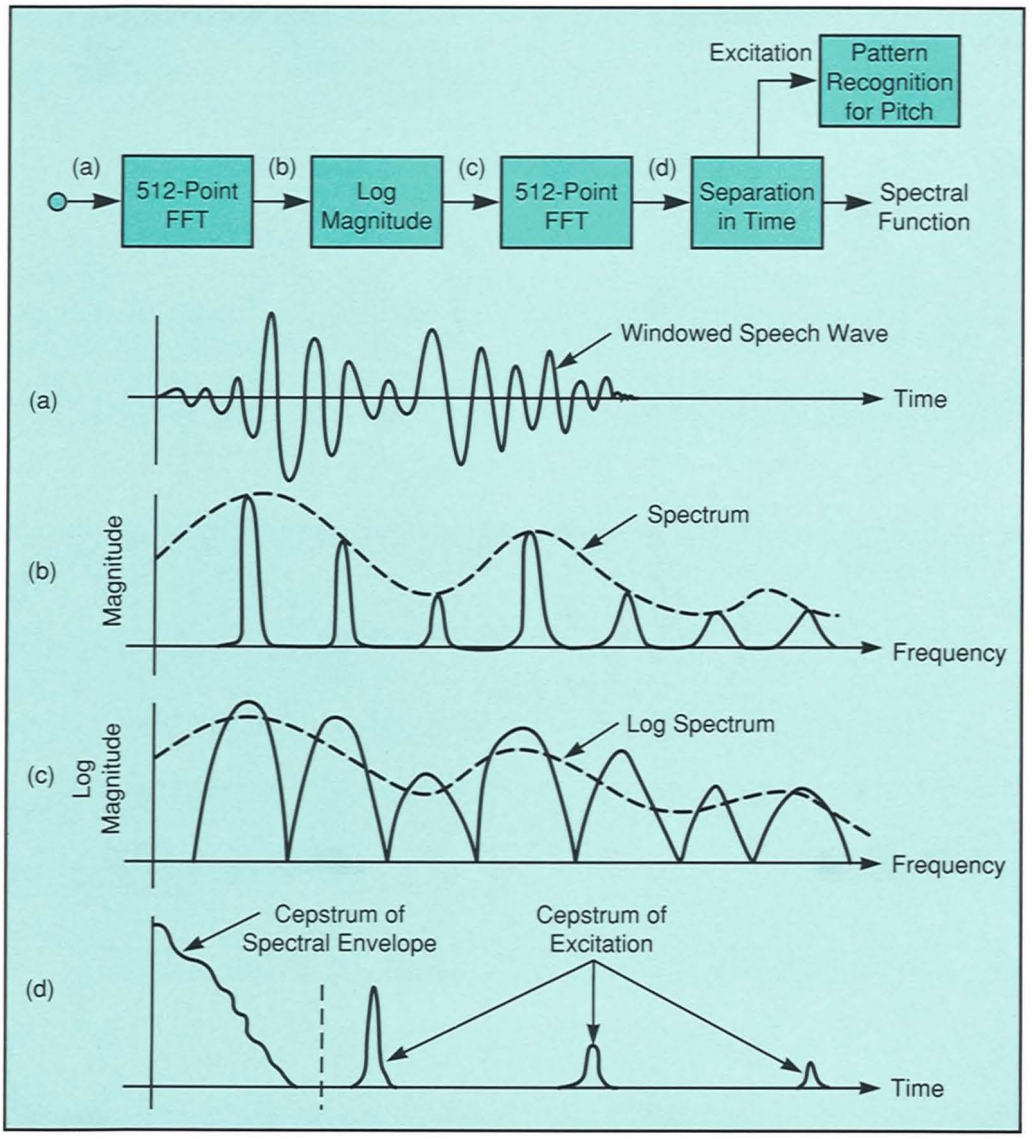


Fig. 5—Cepstral analysis.

less nonlinear operation (in our case, a logarithmic function).

The technique follows straightforwardly from the following argument. The Fourier transform of the convolution of two time functions yields the product of two frequency functions, and the logarithm of this product yields the sum of the logarithms of the two frequency functions. The excitation function, thus transformed, is still a rapidly changing, quasi-periodic function, whereas the transformed spectral-envelope function varies relatively slowly. Therefore, these two transformed functions can be separated by the appropriate linear filtering of the log spectrum. One method of accomplishing this separation is through another Fourier transform, as shown in Fig. 5. Thus cepstral analysis brings about the desired deconvolution. A.V. Oppenheim pioneered early work on deconvolution by cepstral methods; his work will be discussed in the section "Speech Synthesis."

Pattern-Recognition Methods for Pitch Detection

In the early 1960s, Lincoln Laboratory designed an algorithm that used parallel process-

ing to estimate the voice fundamental period [14]. The algorithm (Fig. 6[a]) consisted of four major steps:

1. a filter for the speech signal,
2. a processor that generated six functions of the peaks of the filtered speech signal,
3. six identical elementary pitch-period estimators (PPE), each working on one of the six functions, and
4. a global statistically oriented computation based on the results of step 3.

Figure 6(b) depicts the six measurements on the filtered speech. Measurement m_1 follows the largest peak in the pitch period so that the pitch of any signal with a high positive peak factor is accurately tracked. Since we also want to track the peak factor even when the polarity of the signal is reversed, m_4 has been introduced to accommodate high negative peak factors. In many instances in which the peak factors are not high, we can still perceive large peak-to-peak swings by measuring m_2 and m_5. Finally, strong second harmonics in the signal can create two peaks of nearly equal size per period; the measurements m_3 and m_6 have been introduced to deal with this effect.

Figure 7 shows the operation of an elementary PPE, a device whose function has much in common with the firing pattern of an auditory nerve fiber. In an elementary PPE, a sufficiently large pulse will cause the detector to fire. A blanking interval (called a refractory interval in

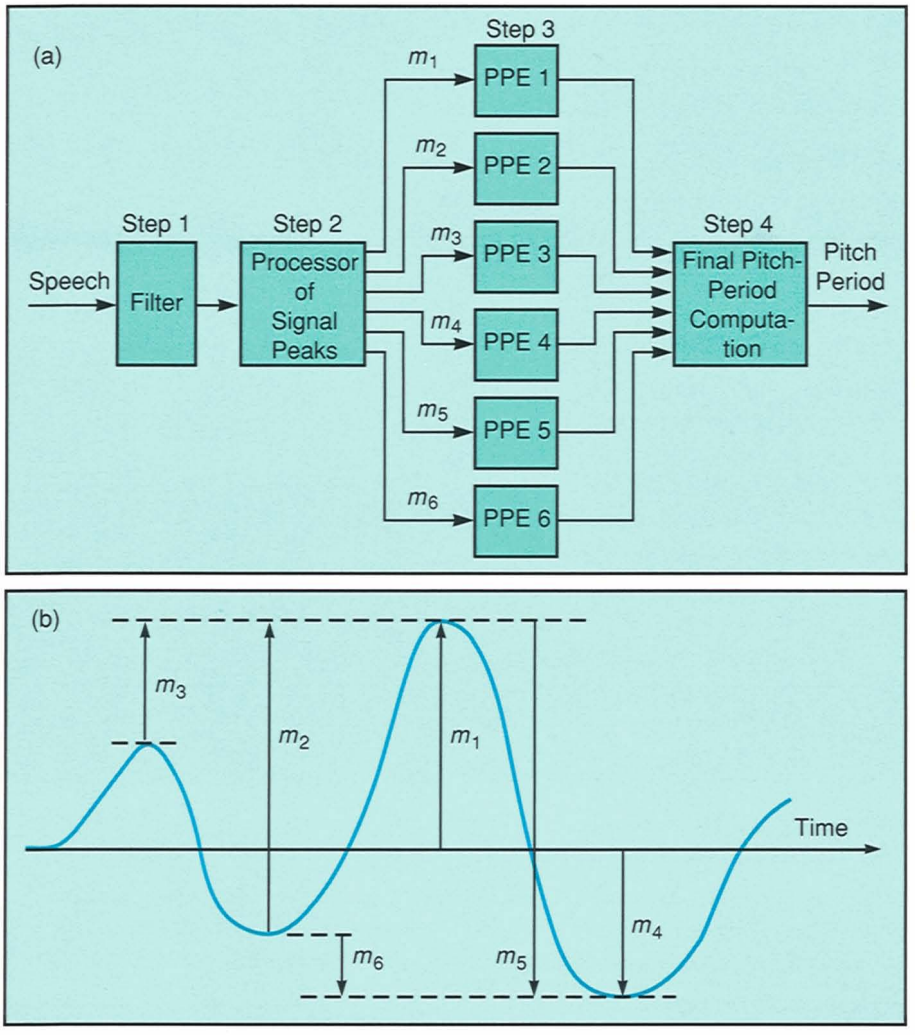


Fig. 6—Estimation of pitch period: (a) schematic of algorithm, and (b) diagram showing basic measurements (m_1, m_2, . . . , m_6) made on the filtered speech.

neural nomenclature) follows, during which subsequent large inputs have no effect. After the blanking interval, the device becomes asymptotically more sensitive to external stimuli and will fire again when any input exceeds the diminishing threshold. Lincoln Laboratory's device incorporates a form of learning in that both blanking intervals and rundown times adapt to the measured periods.

Six elementary pitch detectors are employed, one for each of the measurements of Fig. 6(b). The three most recent measured periods in addition to all possible sums of the three are saved to insure that the true period is found. Thus each elementary detector presents six measurements to the final estimator (as shown in Fig. 8[a]), and a total of 36 measurements are available for further processing. However, to minimize delays the final estimator considers only six candidates, namely, the most recent periods measured by each detector.

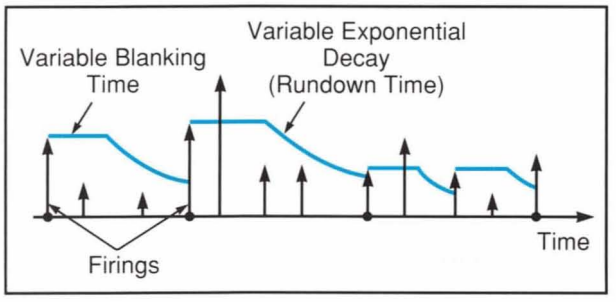


Fig. 7—Operation of the detection circuit. After a firing, a variable blanking time occurs during which subsequent pulses are ignored. The blanking time is followed by a variable exponential decay (rundown time) during which the device fires whenever any pulse exceeds the threshold.

The final estimator compares each of the six candidates (the first row of the matrix in Fig. 8[b]) with the other 35 measurements, and the number of *coincidences* are noted. A coincidence occurs when the absolute difference between a candidate and a measurement does not exceed certain empirically determined thresholds. The candidate with the highest score (i.e., the largest number of *biased* coincidences) is declared the winner. (Reference 14 describes this algorithm in greater detail.) If the score of a winner is below a certain empirically determined threshold, then the signal is considered unvoiced. (Reference 15 provides a detailed description of the buzz-hiss detection rule.)

Lincoln Laboratory's pitch-detection algorithm has an interesting history. Its earliest implementation was on the Laboratory's TX-2 computer. Some speech scientists initially thought the algorithm too complex to be of practical interest. But N.L. Daggett [16] built a hardware version with emerging LSI technology, and by 1980 the entire algorithm was incorporated on a single chip [17]. With current VLSI technology, a *complete* vocoder can be implemented on a single digital signal processing (DSP) chip on which the pitch detector would account for a small percentage of the total processing.

Lincoln Laboratory's algorithm analyzes the temporal properties of the filtered speech signal. Some researchers assert, however, that pitch perception by people is based on a frequency (as opposed to a temporal) analysis of the signal. Thus designers of pitch detectors have also looked at frequency-based approaches. While at Lincoln Laboratory, S. Seneff designed a pitch detector based on the harmonics obtained from a short-time, high-resolution spectral analysis of the speech wave [18]. Quoting Seneff,

> the algorithm described here uses an iterative technique which begins by considering only the two largest peaks. It then adds each peak in turn, from largest to smallest, and after the addition of a new peak determines a new list of potential pitches as the distance between adjacent peaks under consideration. Such a technique results in a built-in weighting mechanism, whereby the largest peak is included in every iteration, but the smallest only in the last. The final decision algorithm determines the pitch from a list which includes all of the estimates from each iteration.

Although it did not work as well as the time-domain system described previously, Seneff's frequency-domain algorithm was more robust, especially when it analyzed speech degraded by telephone-line transmission.

Editing of the Pitch Contour

Constraints of the human-speech-production mechanism dictate that pitch should be a reasonably smooth function of time. This piece of knowledge has been applied in various ways to remove ambiguities in the pitch contour. J.W. Tukey [19] proposed a nonlinear type of smoothing best classified as median smoothing, a technique that Seneff describes in Ref. 18.

Neural Encoding of Pitch

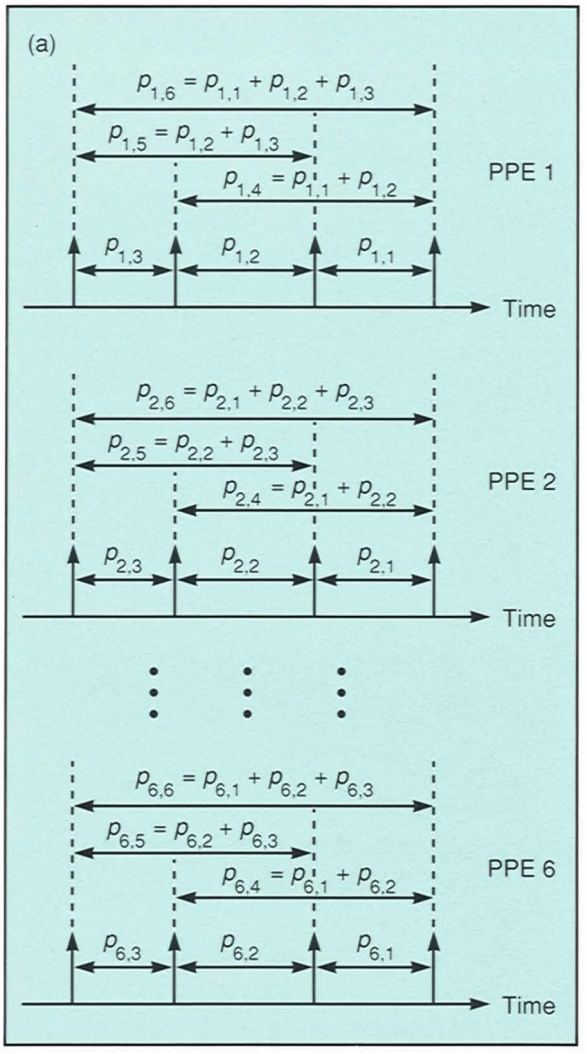
M. Miller and M. Sachs [20] showed that the auditory neurons in cats will behave either as time- or frequency-domain pitch detectors, depending on the relationship between the fundamental frequency of the signal and the characteristic frequency of the auditory fiber. Thus it appears that the next step in designing an accurate and robust pitch estimator should encompass both domains.

Speech Synthesis

G. Fant [21] gives a concise statement that defines a model of human speech: "The speech wave is the response of the vocal tract to one or more excitation signals." In engineering terms, Fant's definition means that human speech production can be modeled as a time-varying filter excited by buzzes, hisses, or a combination of the two. J.L. Flanagan [3] and L. Rabiner and R. Schafer [22] supply in-depth discussions of the subject.

From this starting point, we can describe the spectrum of the speech wave as the product of an excitation spectrum and a vocal-tract spectrum. In particular, during speech sounds that involve vocal-cord vibration, the excitation spectrum can be approximated by a harmonic spectrum. This approximation requires that the speech spectrum itself be resolved into a fine structure and a spectral envelope as shown in Fig. 9.

Many problems are hidden in Fig. 9. First, we note that the fine structure is time variable as the pitch changes. In fact, given a sufficiently high variation in pitch, it is wrong to assume



(b)

Estimates of Period	PPE No. 1	2	3	4	5	6
	$p_{1,1}$	$p_{2,1}$	$p_{3,1}$	$p_{4,1}$	$p_{5,1}$	$p_{6,1}$
	$p_{1,2}$	$p_{2,2}$	$p_{3,2}$	$p_{4,2}$	$p_{5,2}$	$p_{6,2}$
	$p_{1,3}$	$p_{2,3}$	$p_{3,3}$	$p_{4,3}$	$p_{5,3}$	$p_{6,3}$
	$p_{1,4}$	$p_{2,4}$	$p_{3,4}$	$p_{4,4}$	$p_{5,4}$	$p_{6,4}$
	$p_{1,5}$	$p_{2,5}$	$p_{3,5}$	$p_{4,5}$	$p_{5,5}$	$p_{6,5}$
	$p_{1,6}$	$p_{2,6}$	$p_{3,6}$	$p_{4,6}$	$p_{5,6}$	$p_{6,6}$

Fig. 8—Final estimation of pitch period: (a) outputs of the six individual pitch-period estimators, and (b) matrix of the outputs. Each of the entries in the first row of the matrix is a candidate for the final estimate.

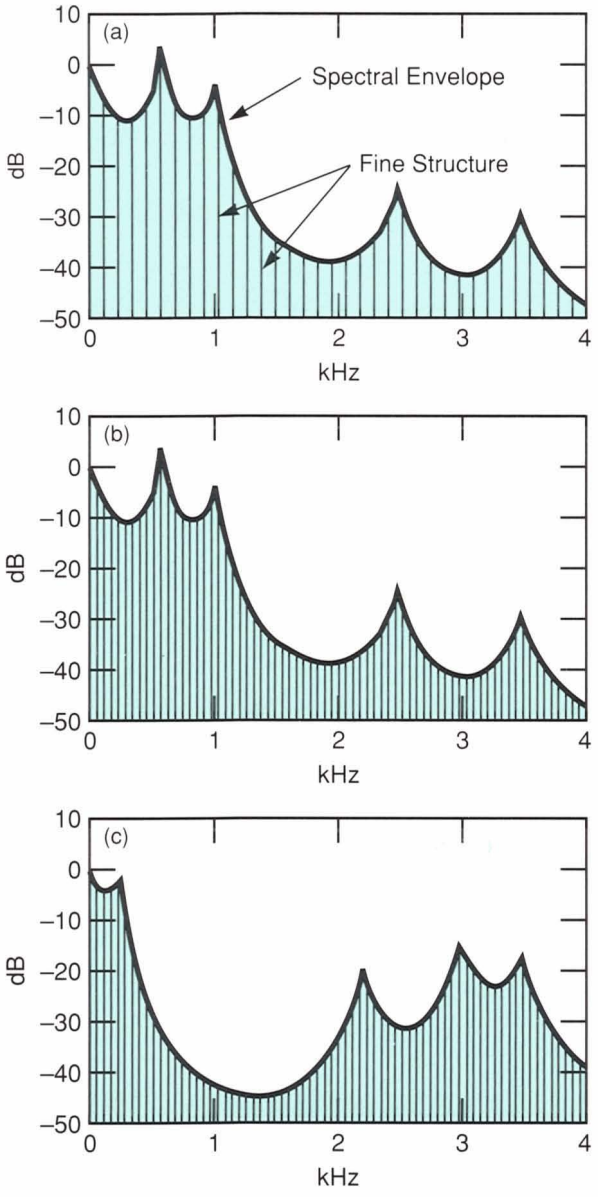


Fig. 9—Spectral envelope and fine structure of several vowels: (a) vowel at relatively high pitch, (b) same vowel as part a *but with lower pitch, and (c) different vowel but with the same pitch as part* b.

that the spacing between adjacent harmonics is constant. Also, the exact harmonic structure may be compromised by irregularities in the vocal-cord vibration and by various forms of external sources of speech degradation such as noise or the speech's transmission through a telephone system. Because of these effects, but perhaps more so because the pitch of speech varies greatly (from 60 Hz to 800 Hz), it is difficult to define the spectral envelope consistently. In Fig. 9, we have drawn imaginary curves that pass through the peaks of the harmonics but we cannot say what the true spectral envelope is, especially at frequencies away from the harmonics.

Furthermore, the term spectrum in our case necessarily means short-time spectrum. Thus our results are critically dependent on how long an interval of speech is processed to produce a single spectrum.

Issues such as the above problems tend to get resolved empirically. The analysis of speech yields valuable insight but even more important is the experience gained by actually listening to a vocoder. Dudley's first vocoder contained 10 channels and its quality was found greatly wanting; a later version that employed 30 channels was far more satisfactory.

Speech-Synthesis Models

The ideas introduced in the previous section lead to a simplified model of speech synthesis. According to the model, speech synthesizers consist of three major blocks: a buzz generator that generates a quasi-periodic pulse train, a hiss generator that produces white noise, and a linear filter with time-varying parameters. In this section we will focus on the filter and describe several of its variations.

Speech synthesizers can be divided into two broad classes: *universal* in which the synthesizer structure remains the same for all sounds (i.e., only the parameters vary), and *sound specific* in which significant structural as well as parametric changes occur that depend on the speech itself. Let us first describe the universal structures. Since such structures are linear filters, they can be described in terms of their poles and zeros. We distinguish among the following types of synthesizers: (a) fixed poles and variable zeros, (b) variable poles, (c) variable zeros, and (d) variable poles and zeros.

Figure 10 shows a classic channel-vocoder synthesizer. Each of the fixed filters shown is typically a 4-pole or a 2-pole filter. Zeros are created by the parallel addition of all filter outputs. The poles of the overall transfer function are simply the poles of all the individual filters,

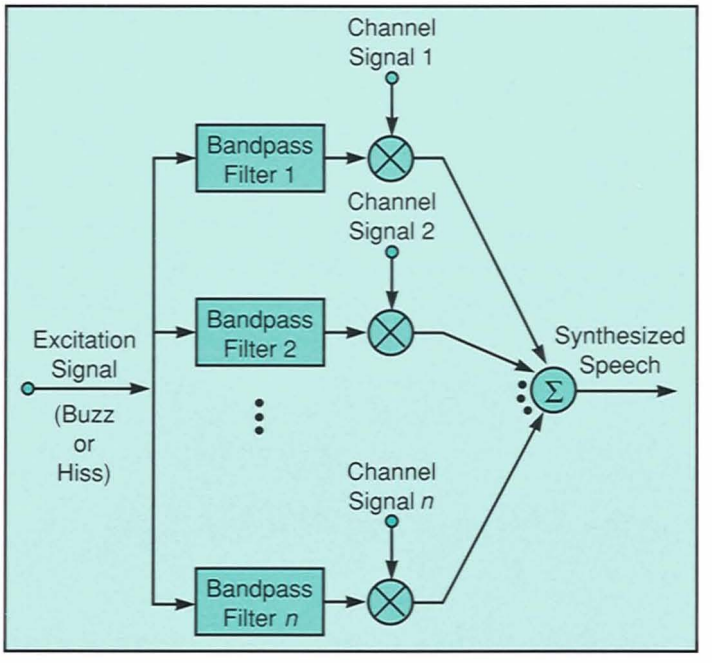


Fig. 10—The classical channel-vocoder synthesizer, which has fixed poles and variable zeros.

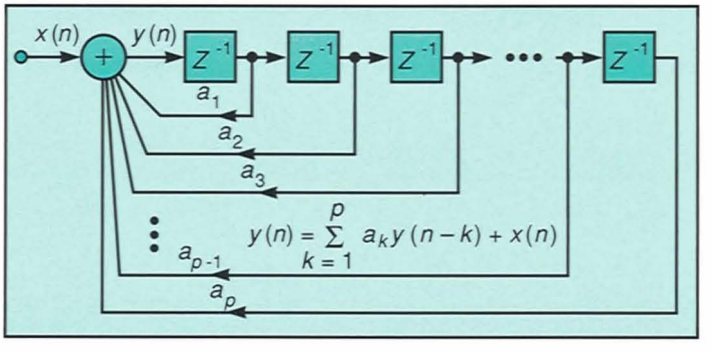


Fig. 11—Linear predictive coding (LPC) vocoder synthesizer, which has variable poles (variable a *coefficients).*

while the zeros vary with the time-varying parameters applied to the modulators. Thus the channel-vocoder synthesizer has fixed poles and variable zeros.

Figure 11 shows a variable-pole speech synthesizer in which the pole variations are controlled by the parameters shown.

Figure 12 shows a variable-zero filter. Structurally, this is a tapped delay line or, in digital terms, a finite impulse response (FIR) filter. Such a synthesizer requires many more variable parameters than that of Figs. 10 or 11, but, as we shall see when discussing cepstral methods, these parameters can be derived from a smaller number of more basic parameters.

Figure 13 shows an example of a synthesizer in which both poles and zeros are time varying. This particular device is called a parallel-formant synthesizer; the variations are created by adjusting both the poles of the filters and the modulation signals.

Sound-specific models (as distinguished from universal models) are really the results of research on synthesizers with human rather than automatic analysis. One example is Fant's OVE II synthesizer [23] (Fig. 14). Note that there

are three separate networks: the top network generates vowels and semivowels, the middle network generates nasals, and the bottom network generates fricatives and plosives. Also, the excitation functions are connected in a variety of ways so that, for example, whispered vowels can be generated. OVE II can produce good synthetic speech when a skilled human analyzer performs the analysis. The usual procedure is to enter parameters based on a spectrogram of the utterance to be synthesized. Thus far no one has succeeded in designing a completely automatic analysis-synthesis system based on OVE II.

Spectral Flattening of the Excitation Function

Figure 9 leads one to believe that our model assumes a clear separation of the excitation and the spectrum. Thus for voiced speech the excitation is postulated to consist of equally spaced harmonics, all of the same amplitude. The spec-

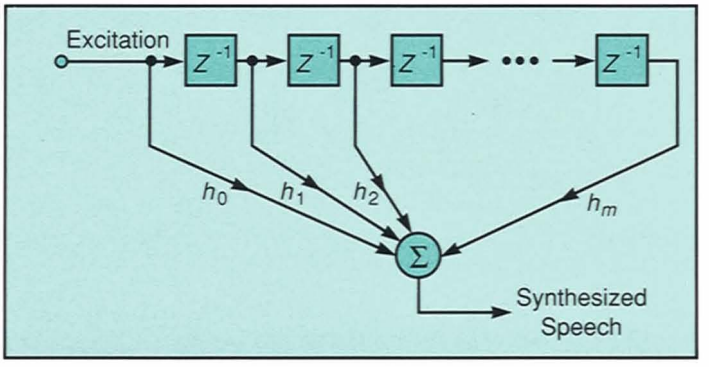


Fig. 12—Homomorphic vocoder synthesizer, which has variable zeros.

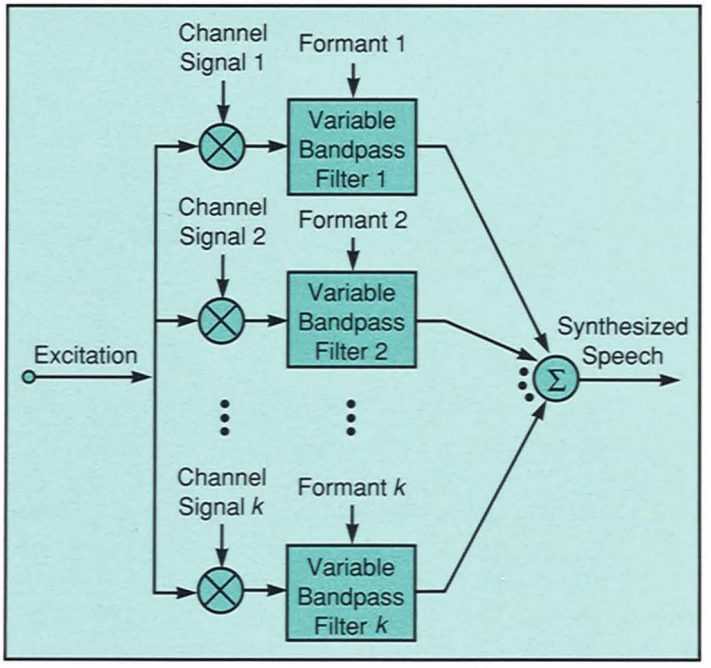


Fig. 13—Parallel-formant vocoder synthesizer, which has variable poles and zeros.

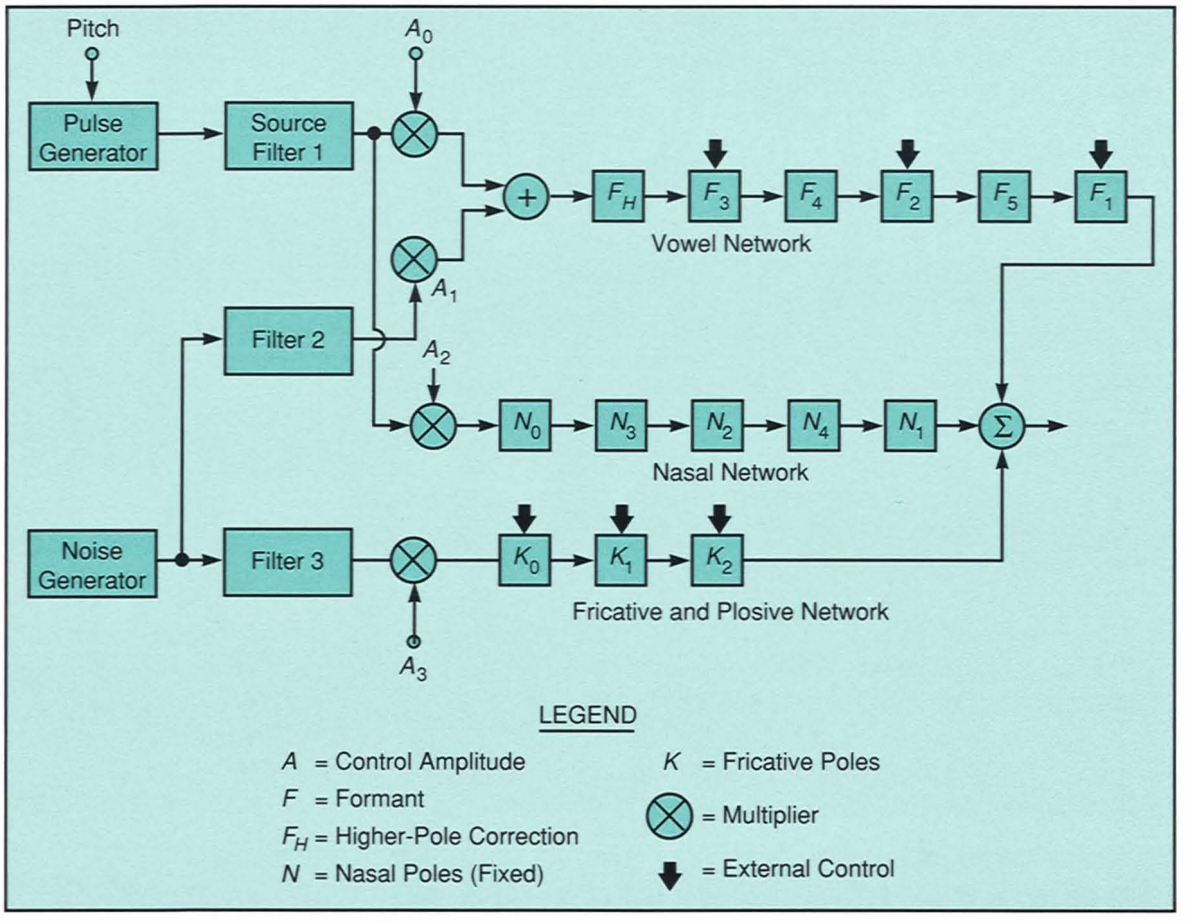


Fig. 14—The OVE II synthesizer of Gunnar Fant [23].

trum, or more explicitly the spectral envelope, is assumed to be a smooth curve. Much evidence exists that the physical situation is far from this ideal.

Figure 15 shows a few examples of glottal pulses obtained by various methods. Note that the pressure wave is often not perfectly periodic nor is the waveshape identical from period to period. Therefore, an excitation signal can have a spectrum that deviates greatly from the ideal, and any spectral analysis on the speech signal needs to incorporate the deviation.

Let us denote the excitation spectrum by $E(f)$ and the vocal-tract spectrum by $V(f)$. Then the measured speech spectrum is given by $S(f) = E(f)V(f)$. Now, let us imagine that by some magical trick the system were capable of knowing $E(f)$ and thereby capable of knowing the precise shape of the excitation wave. Then all synthesizers that we have thus far described would synthesize speech with the spectrum $E(f)S(f) = E(f)^2V(f)$. This result is clearly wrong

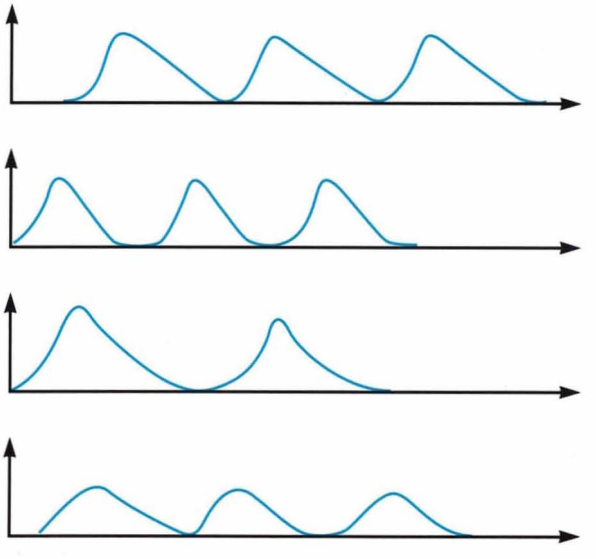


Fig. 15—Samples of glottal pulses.

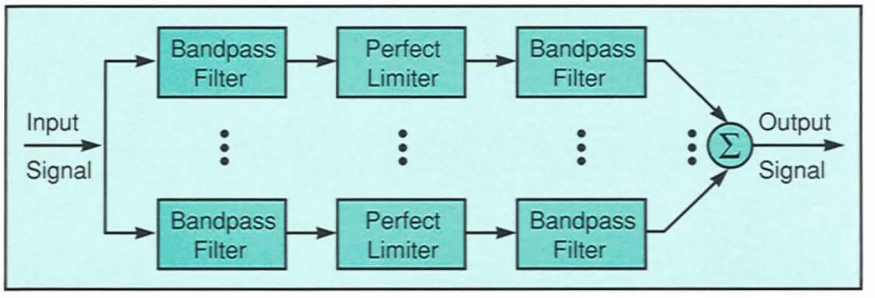


Fig. 16—Spectral flattening.

unless the spectrum $E(f)$ is constant with frequency, which for voiced speech means equality at the harmonics.

The solution to this problem is *spectral flattening*, in which all harmonics of the excitation signal are set equal when the voicing block is used. When the voiceless (or hiss) block is used, a short-term whitening of the spectrum is performed (i.e., the spectrum is made constant with respect to frequency). Figure 16 shows one technique for approximate spectral flattening. The technique was derived from vocoder research at Lincoln Laboratory [24].

Filter-Bank Spectral Analysis

A first design might consist of a set of bandpass filters, each of the same fixed bandwidth and with crossovers at the 3-dB points. Each bandpass-filter output is applied to either a full-wave or half-wave linear rectifier that, in turn, feeds a low-pass filter. All low-pass filters are also of the same bandwidth, and the bandwidth is fixed in advance. Such a system is perhaps simplest to design and implement. For a digital implementation, the system has the advantage that the gains of all filters are equal, so that no gain adjustment is needed for the different channels.

Given the imposition of equal bandwidth, we still have several parameters to consider. First, there is the overall bandwidth of the spectrum analyzer. This disarmingly simple parameter actually has led to much soul searching on the part of system builders. Consider, for example, that channel capacity is directly proportional to bandwidth. Thus, if the telephone system insisted on transmitting all frequencies in the speech signal, the frequencies would encompass about 10 kHz, or about three times as much as the present American telephone system allows.

Another important parameter is the actual width of each bandpass filter; this width in turn specifies the number of filters. Here we are dealing with the question of frequency resolution. If we desire to resolve individual harmonics of all speech signals, we need to design narrowband filters. On the other hand, if we are interested in the spectrum envelope of the speech, the filters would be wider. Figure 17 shows how the measured spectrum changes as the bandwidth of the uniform-filter bank changes (compare Fig. 17[c] to Fig. 17[e]).

Finally, the designer has to specify the type of filter desired. A wide variety of filter designs exist:

a. *Bessel filters* have a desirable response of linear phase versus frequency. Such filters have fairly shallow cutoff characteristics. (Bessel filters are so named because the denominator of the analog transfer function is a Bessel polynomial.)
b. *Elliptic filters* are the opposite extreme of Bessel filters. Elliptic filters have sharp cutoff characteristics, but because of high reverberations the filters are generally not recommended for speech-spectrum analysis.
c. *Chebyshev filters* have undesirable phase properties similar to elliptic filters.
d. *Butterworth filters* appear to be a decent compromise between sharp cutoff and linear-phase response. Here the parameter to consider is the order of the filter: too high an order can produce serious phase distor-

tion. Second- or third-order Butterworth filters are good analysis filters.

e. Lincoln Laboratory discovered *Lerner filters* by adding suitably weighted resonator outputs to obtain both an excellent linear phase and sharp amplitude cutoff.

f. *Frequency-sampling filters* are derived by cascading a comb filter with suitably weighted resonators in parallel, such that the resonator poles are canceled by the comb-filter zeros. Frequency-sampling filters have perfectly linear phase and excellent cutoff characteristics.

g. *Digital filters* derived from window functions are called nonrecursive, or FIR, filters. The phase of such filters can always be made perfectly linear while many different amplitude characteristics can be derived.

h. The bandwidths of *other filters* can be derived from auditory system analysis. Both physiological and psychophysical data indicate that the ear performs some sort of spectrum analysis. The physiological data include G. von Bekesy's measurements of basilar membrane motion [25] and N.Y.S. Kiang's tuning curves for the cat's auditory nerve [26].

Filters *a* through *d* are of the minimum-phase variety; i.e., there are no zeros in the right-half *s* plane (for analog filters) or no zeros outside the unit circle (for digital filters). Dispensing with these constraints makes other designs (filters *e* through *h*) interesting. The author and C.M. Rader [27] give detailed descriptions of filters *a* through *f*. Most of the filter designs were first tested as digital implementations in a vocoder context at Lincoln Laboratory.

As the center frequency increases, the effective bandwidth of the ear's auditory filters also increases. Figure 18 shows three examples of proposed filter-bank designs based on various psychoacoustic data.

It is important to remember that the ear's frequency resolution at higher frequencies is poorer than at lower ones. Taking advantage of this fact is easily accomplished in a channel vocoder by the use of fewer analysis channels for higher frequencies.

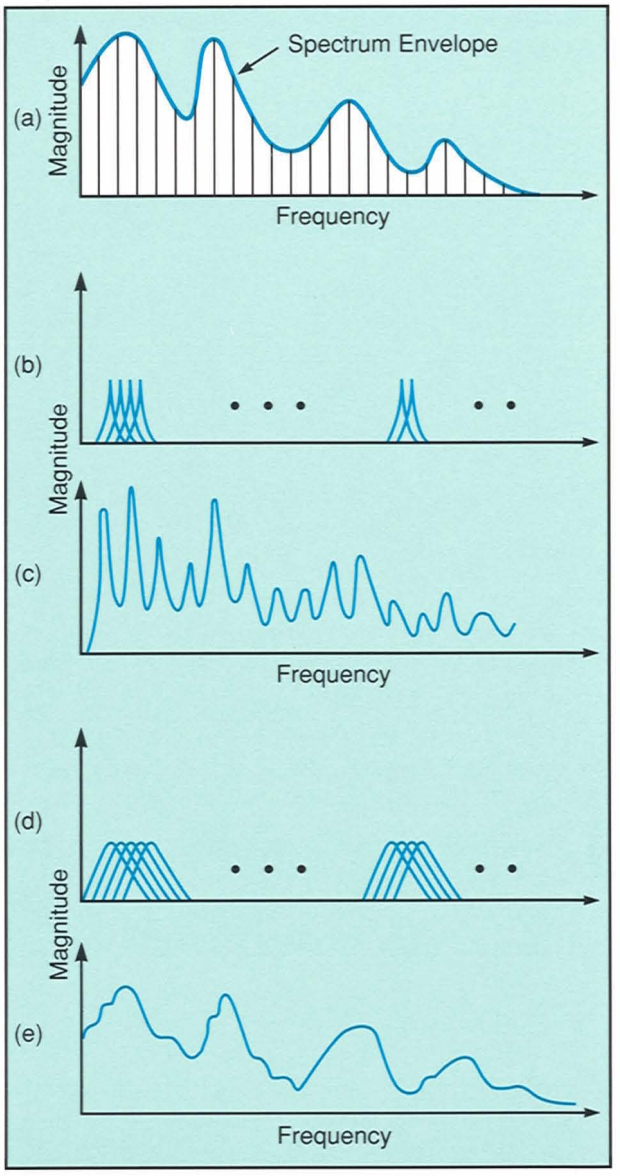


Fig. 17—Spectra obtained with different filter banks: (a) idealized steady-state speech spectrum, (b) bank of narrowband filters, (c) spectrum obtained from narrowband-filter bank, (d) bank of wideband filters, and (e) spectrum obtained from wideband-filter bank.

Discrete Fourier transforms (DFT) of the speech signal can be processed to create a filter bank. It is well known that there are important mathematical relationships between the filter bank and the DFT spectrum analysis. As an example, any filter bank consisting of FIR filters can be exactly emulated by a sliding DFT (in which a new DFT is taken for every sample) with a proper window function.

From an implementational viewpoint, filter-

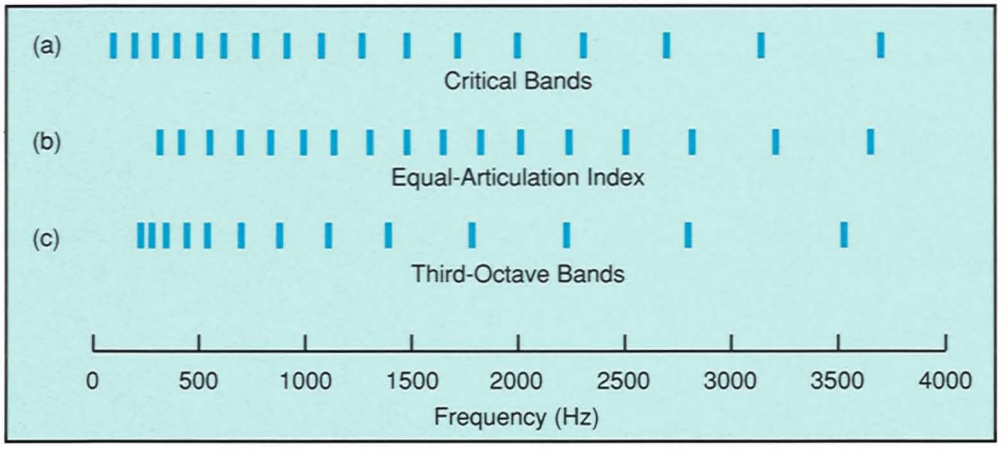


Fig. 18—Various concepts for defining filter bandwidth versus center frequency: (a) critical bands, (b) equal-articulation index, and (c) third-octave bands.

bank analysis seems more straightforward than the use of DFT. If, however, a very high-resolution analysis is desired, fast Fourier transform (FFT) methods can save considerable time. And if one desires to resolve the harmonics of the speech spectrum, DFT is the appropriate mechanism. For example, if the bandwidth of interest is 7 kHz and the desired frequency resolution is 7 Hz, then no less than a 2048-point spectrum is required.

Linear Prediction

References 28, 29, and 30 provide detailed discussions of linear predictive coding (LPC). In this article, we summarize the concept of LPC without the inclusion of formulas.

LPC commences by predicting the *n*th sample of a signal from a linear, weighted sum of the previous *p* samples. Developing a good method of computing the coefficients of the linear predictor is the essence of LPC. Note that the predictor is really a linear filter and because all operations are linear the true signal can be recovered from the error signal with inverse filtering. An error can be defined as the difference between the predicted and measured *n*th sample.

B.S. Atal and S.L. Hanauer [29] developed LPC as an alternative method of speech analysis-synthesis for bandwidth compression. Soon after, E.M. Hofstetter created the first real-time implementation of an LPC vocoder. LPC became a popular technique; in 1975 the U.S. government recognized the LPC vocoder as the standard vocoder system. One reason for the device's popularity was its appreciably smaller size compared, for example, to Dudley's traditional channel vocoder. Recent LPC research has been directed toward efficient coding of the error signal with the intent of designing a high-quality vocoder capable of operating in the range of 2.4 to 4.8 kb/sec.

Since the original work of Atal and Hanauer, many variants of the original algorithm have been proposed and implemented. In this article, we briefly describe the autocorrelation method of J.D. Markel and A.H. Gray [30].

In an LPC vocoder, transmission of the error signal and the coefficients is sufficient to reconstruct the exact speech signal at the receiver. Let us assume that we can faithfully transmit these parameters. If our prediction method is any good, much of the speech information will be transferred to the coefficients and therefore the error signal should contain substantially less information than the original signal. This reduction in information can, in principle, lead to the desired bandwidth reduction.

We also note that the predictor is an FIR filter, which contains only zeros in the complex *z*-plane. Thus the inverse filter must be an all-pole filter (Fig. 11).

The next step is to introduce a criterion that leads to a reasonable computational procedure for finding the coefficients; the procedure must

also meet certain standards for accurate prediction. One such procedure is the minimum mean-square criterion, formulated as follows. A time interval is chosen during which the original signal is sufficiently stationary (from 20 to 40 msec for speech). Next, a set of coefficients is found to minimize the mean-squared value of the error signal over that interval. Such a formulation leads to a set of p linear equations. Standard mathematical methods are then applied to find the desired coefficients of the original linear predictor.

Let us summarize the results and ideas that have been developed up to this point. We will then determine what further steps are needed. First, we have associated a specific structure with a speech wave such that a new speech sample can be predicted on the basis of p previous measured samples. (In statistics, this type of prediction model is called autoregressive.) The parameters of the model can thus be found by solving a set of linear equations.

We now discuss several consequences of this version of the LPC algorithm. Most of our statements will be made without proof.

a. When we compute the autocorrelation function of the impulse response of the LPC synthesis filter, the first p samples of the function are equal to the corresponding samples of the autocorrelation function of the speech itself. Therefore, LPC analysis can be viewed as an approximate method of matching the correlation function of the output speech to that of the input speech, much as a channel vocoder matches output and input spectra.

b. It can be shown that minimizing the z-transform of the error signal leads to the same set of equations as that obtained by minimizing the error signal. Thus the mathematical formalism can be carried out in the spectral domain and leads to the same synthesis structure as before. Furthermore, the autocorrelation function can be computed by performing an inverse Fourier transform on the measured square of the spectrum.

c. The formants, or resonances, of the vocal tract are faithfully preserved, whereas the nulls in the spectrum are less accurately tracked by the LPC approximation.

d. Another point of interest concerns audio preprocessing. For example, let us choose to filter the speech with a sharp-cutoff low-pass filter at 3 kHz. Then, prior to our LPC computation, let us sample the signal at 12 kHz. The LPC algorithm will valiantly try to create a good spectral match not only of the speech spectrum but also of the sharp cutoff filter. The filter thus degrades the ability of the analysis to represent the speech spectrum accurately. The relatively narrow 3-kHz bandwidth will cause the resultant autocorrelation curve to stretch with respect to the x-axis. Therefore, the initial $p + 1$ points of the measured correlation will change more slowly, which could lead to numerical errors in the solution of the autocorrelation-matrix equation.

Homomorphic Vocoding

In the early 1960s, researchers at Bell Laboratories studied the process of deconvolution. Given a signal that was assumed to be created by the convolution of two other signals, by what technique could the two other signals be separated and examined individually? An important practical application of this problem is earthquake analysis if one assumes that observed seismic signals are responses of the earth to sudden movements of faults [31]. A.V. Oppenheim independently explored the mathematics of a specific class of nonlinear transformations [32]. Interestingly, the *cepstral analysis* developed by the Bell Labs researchers and the *homomorphic filtering* of Oppenheim turned out to be the same scheme.

During a two-year stay at Lincoln Laboratory in the late 1960s, Oppenheim developed the homomorphic vocoder algorithm [33]. The algorithm takes the log magnitude of the DFT of the original speech and thus transforms the deconvolution problem into a quasi-linear problem. This transformation allows for good separation of the excitation function from the vocal-tract transfer function. Figure 19 is a block diagram

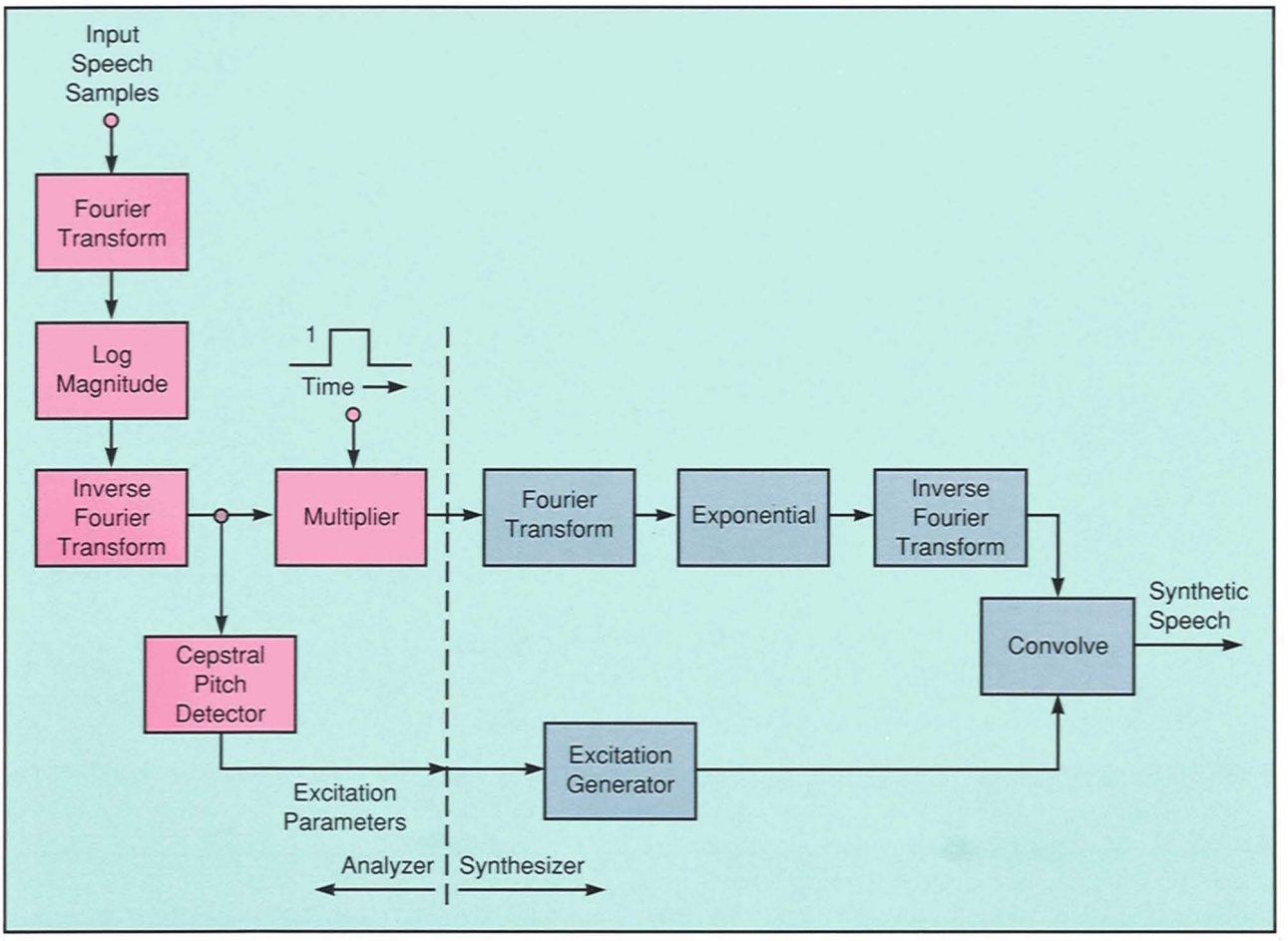


Fig. 19—Schematic of the homomorphic vocoder.

of the homomorphic vocoder.

In the 1970s, P. Blankenship implemented the first real-time simulation of the homomorphic vocoder on the Lincoln Digital Signal Processor (LDSP), which is described in the section "Speech-Processing Facilities."

As stated earlier, all vocoder systems perform deconvolution in that they separate the excitation from the vocal-tract filter. The channel vocoder uses wide-bandpass filters and energy detectors that effectively remove most of the excitation signal's contributions. LPC performs the same function by finding parameters of a vocal-tract (all-pole) model. One interesting feature of homomorphic analysis is its use of a high-resolution DFT. Thus with homomorphic vocoders, as distinct from either channel or LPC vocoders, the spectral envelope and fine structure are both available for subsequent spectral and pitch-detection processing.

Vocoder Hardware

By the early 1960s, many laboratories were actively designing and building channel vocoders. In November 1967, the Air Force Cambridge Research Laboratories (AFCRL) sponsored a speech analysis-synthesis survey that invited participants to exhibit their vocoder hardware. There were nine displays of 2400-b/sec channel vocoders, from Collins Radio Company, L.M. Ericsson (Sweden), Lincoln Laboratory, Bell Laboratories, Texas Instruments Inc., the U.S. Army, AFCRL, and Philco, which displayed two systems. It is interesting to note that many of the vocoders employed spectral flattening.

The Lincoln Laboratory channel vocoder (Fig. 20) performed spectral analysis with a bank of 16 analog Bessel bandpass filters, 16 half-wave rectifiers, and 16 fifth-order low-pass Bessel filters that had 20-Hz cutoff frequencies. The

pitch and voicing hardware was based on the author's algorithm that is described both in the section "Pitch Detection" and in Ref. 15. The digital hardware could produce a new estimate every few milliseconds. A single 8-bit period and two 4-bit increments were transmitted for each 20-msec time frame. The spectral information was sampled every 20 msec, log-encoded to 5 bits, and transformed via a Hadamard matrix. The information rate was thus reduced to 33 bits per 20-msec frame.

During the 1970s, a noteworthy change occurred in the philosophy of real-time algorithmic implementations. In the 1950s, a special-purpose implementation of a vocoder algorithm could be about 100 times faster than a computer simulation of the same algorithm. But during the late 1960s and early 1970s, the concurrent development of microprocessor technology and DSP algorithms radically changed that ratio. By the mid-1970s, the most efficient hardware approach was with a programmable microprocessor.

Using such microprocessors, Hofstetter, J. Tierney, and O. Wheeler [34] implemented real-time LPC hardware. After some study, the chip set chosen was a group of bit-slice-oriented components. Although the original design called for three separate microprocessors that would each perform a subtask, the final design required just a single microprocessor. Hofstetter, Tierney, and Wheeler felt that the lone microprocessor could most efficiently satisfy all signal processing requirements if it were augmented by a hardware multiplier. The overall device, called the Linear Predictive Coding Microproces-

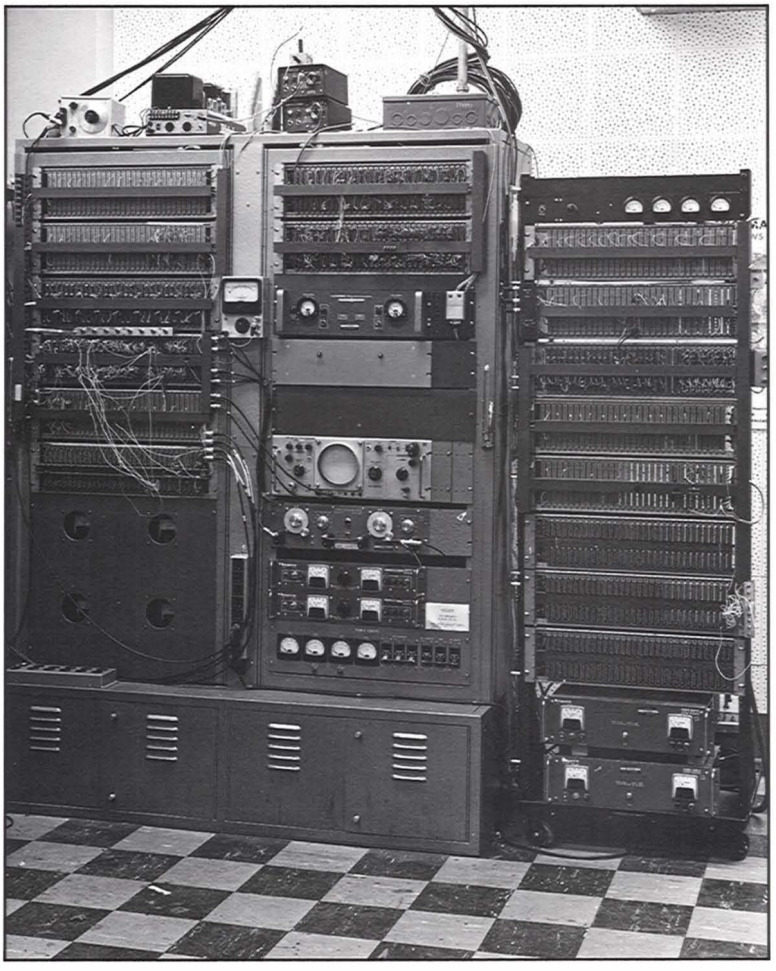


Fig. 20—Early channel vocoder built at Lincoln Laboratory.

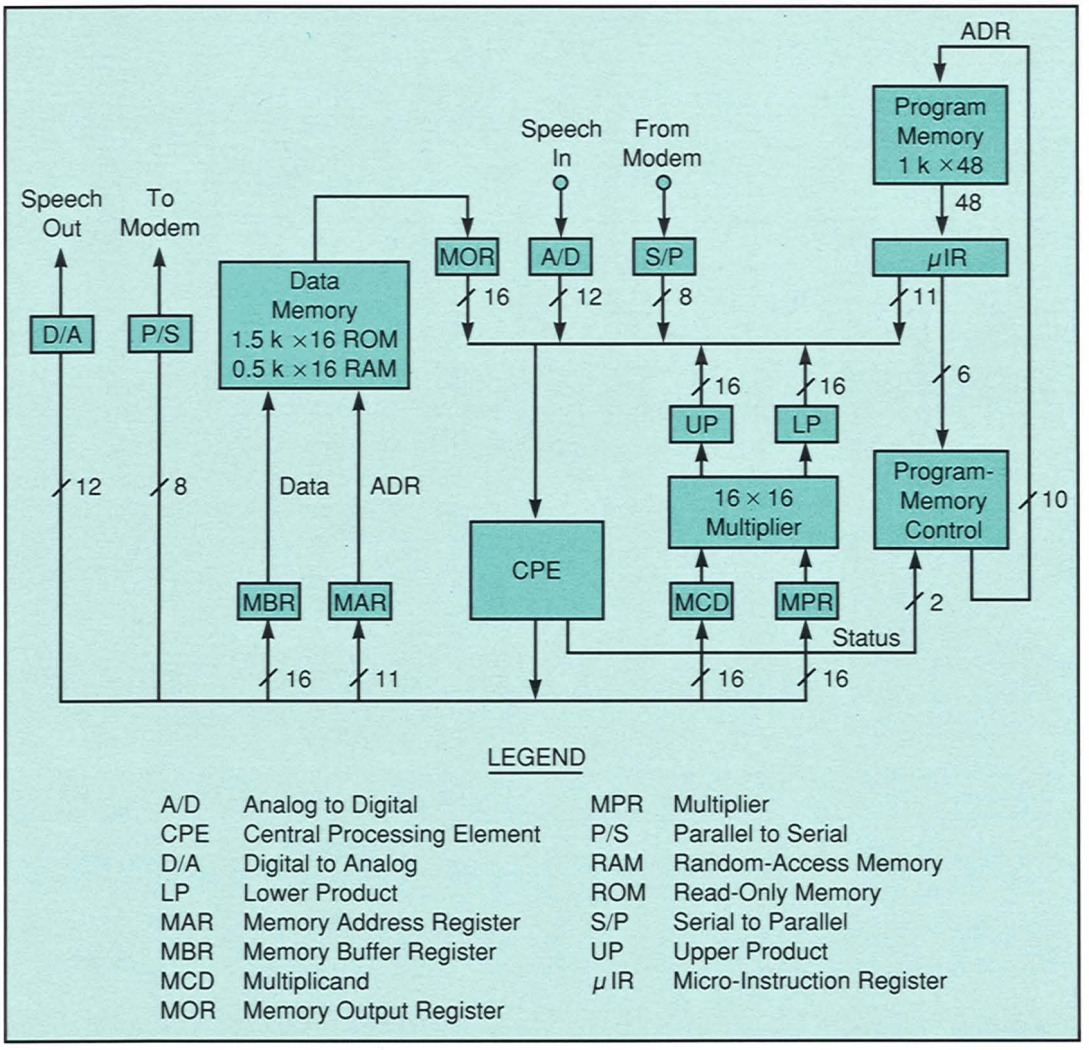


Fig. 21—Schematic of Linear Predictive Coding Microprocessor (LPCM).

sor (LPCM), was then the most compact real-time version of the LPC algorithm. Figure 21 is a block diagram of the LPCM and Fig. 22 is a photograph of the completed hardware. Indeed, LPCM was a leap forward in compactness when compared to the channel-vocoder hardware of the 1960s.

The box of Fig. 22 contains two standard 16" × 7" universal wirewrap boards. The boards were chosen to accommodate 14- to 40-pin packages and the total hardware consisted of 162 dual in-line packages located on approximately 1½ boards. The power consumption was less than 45 W.

It appears that as hardware becomes more compact, the equipment superstructure required to test the hardware gets larger. When Dudley built his vocoder, he did so without the aid of computer simulation. In contrast, the LPCM algorithm was first programmed by Hofstetter on the Lincoln Fast Digital Processor (FDP) (the first real-time LPC program) [35] and then programmed again on the Lincoln Digital Voice Terminal (LDVT). Furthermore, an important subtask of the LPCM program was the design and fabrication of the LPCM tester, which was umbilically connected to the LPCM during the debugging phase. The main component of the tester was a 1024-word × 48-bit random-access memory that effectively replaced the programmable read-only memory (PROM) destined to reside in the LPCM. The tester also

duplicated some of the functions of the LPCM program-control chip so that the LPCM could be extensively tested.

Another means of testing the hardware and firmware was with a simulation program on Lincoln Laboratory's general-purpose facility, which was centered on a Univac 1219 computer. In addition, an assembler program that understood LPCM mnemonics and symbolic addresses was written for the Univac machine. The binary output produced by the assembler could be loaded into the LPCM tester, and the PROMs were later burned with the LPCM's program memory.

By the end of the 1970s, technological advances allowed even smaller vocoders. A design by J.A. Feldman and Hofstetter [17] of the same LPC algorithm used in the LPCM required only 16 ICs (as opposed to the LPCM's 165 ICs), occupied one-half of a 7" × 7" wirewrap board, and dissipated 8.6 W (compared to 45 W). Three signal processing chips (NEC's PD7720s) were actually individual microprocessors, but they were each used for specific functions: one implemented the LPC analysis, another the pitch detection, and a third the LPC synthesis. The remaining ICs included an Intel 8085-based 8-bit microprocessor chip set that performed the control and communications functions. Figures 23 and 24 show Feldman and Hofstetter's device.

Thus Lincoln Laboratory's pitch detector, which used the same algorithm that had been used in previous implementations and that had been invented by the author 20 years ago, could now fit onto a single programmable chip. This feat was especially fulfilling to the author, who had observed the original real-time implementation by V.J. Sferrino. The implemention had occupied a complete rack of equipment (the middle rack of Fig. 20).

It is interesting to note that vocoder hardware was once described as taking up a substantial portion of the space of a large boat [6]. Today the same functions can be embodied on a single silicon chip.

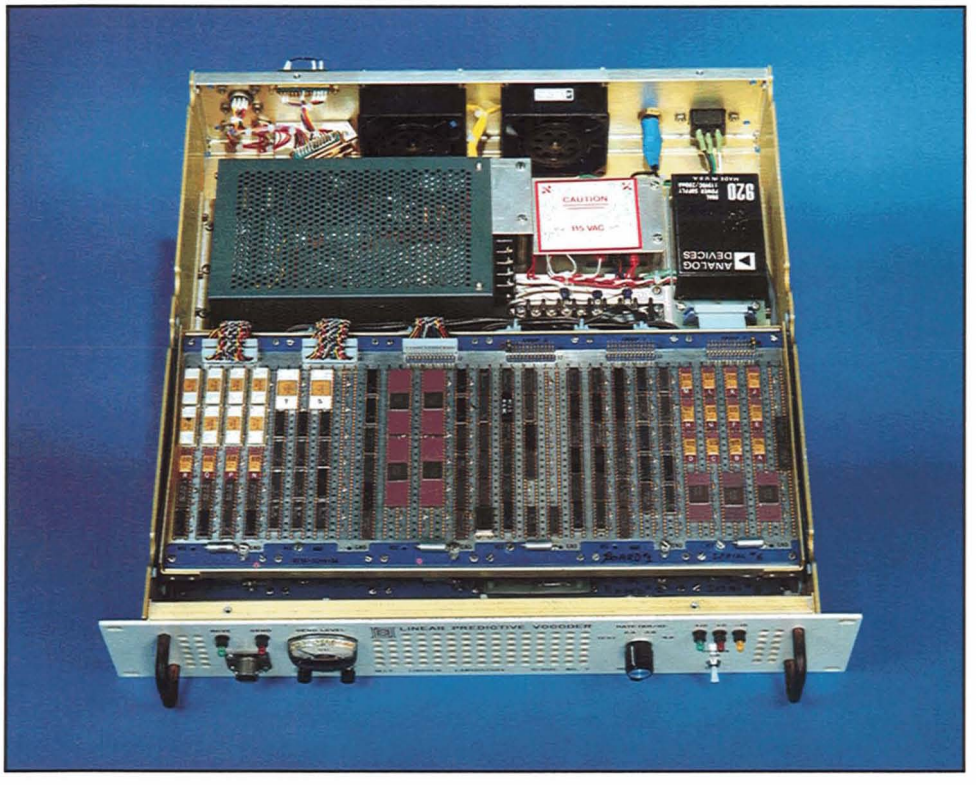


Fig. 22—The Linear Predictive Coding Microprocessor (LPCM) of Fig. 21.

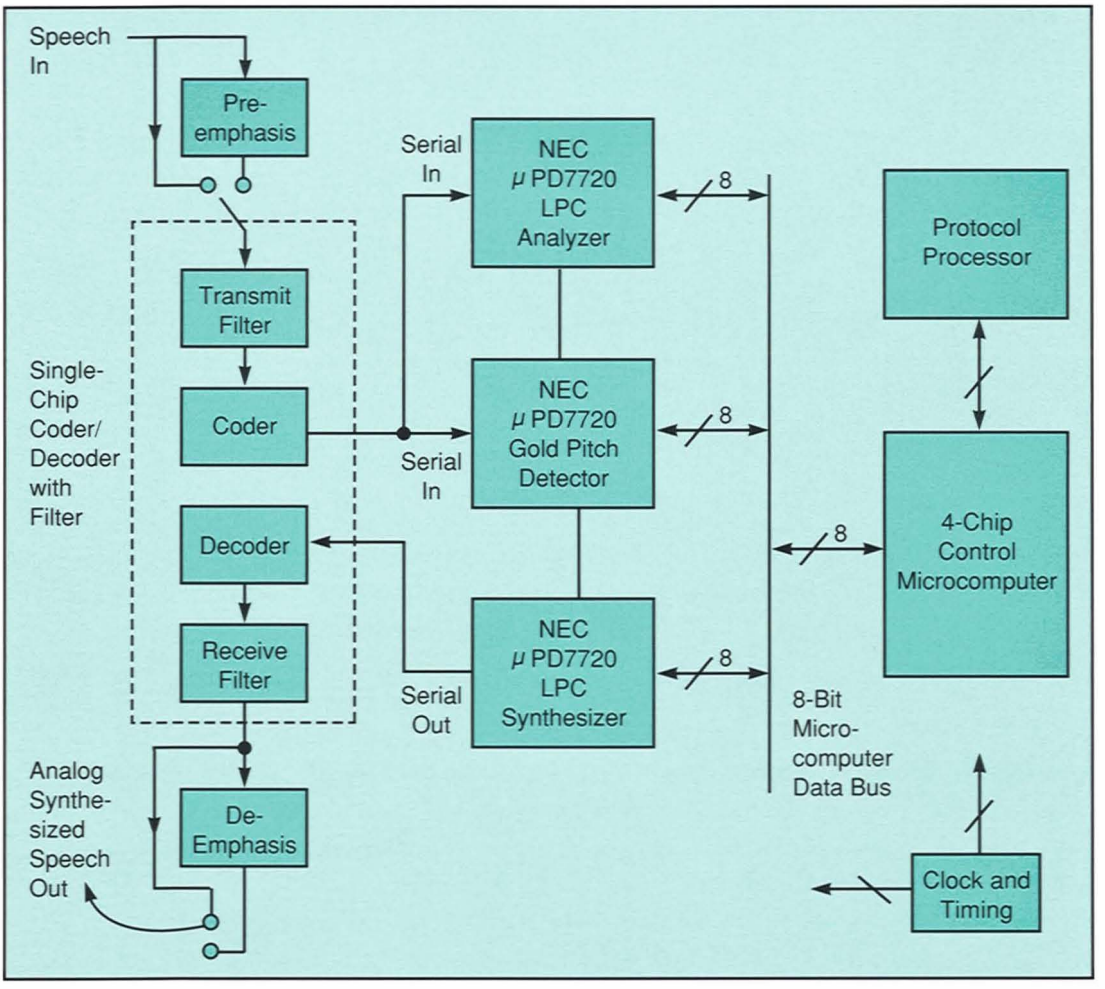


Fig. 23—Implementation of linear predictive coding (LPC) with three programmable signal processing chips.

Low-Rate Vocoder Systems

In an informal exercise, the author recently timed himself reading a paragraph aloud for 20 seconds and then counted the total number of letters, numbers, spaces, and punctuations in the paragraph. By considering each of these to be equivalent to a 5-bit Teletype character, the author counted a total of 258 characters, or 1290 bits. A Teletype machine transmitting the same paragraph at 75 b/sec would have completed the job in 17.2 sec. Thus, to a first approximation, a Teletype system can transmit textual information at almost the same rate as a person speaking.

Speech, however, has much more information than text. The speaker's identity, emotional state, and prosodic nuances all count as information. Although it is not clear how much of such extra information actually exists or is really pertinent, it seems reasonable to assume that a good 2400-b/sec vocoder can contain nearly all of this extra information. Thus we can conservatively deduce that the limit of bandwidth compression is somewhere between 75 and 2400 b/sec. This section will study systems with bit rates within that range.

An obvious way to reduce the rate of a 2400-b/sec vocoder is to lower the frame rate and quantize the parameters more coarsely. However, if a brute-force approach is taken, quality and intelligibility deteriorate rapidly. The frame rate can be effectively reduced and fewer bits can be allocated for parameters, but some degree of

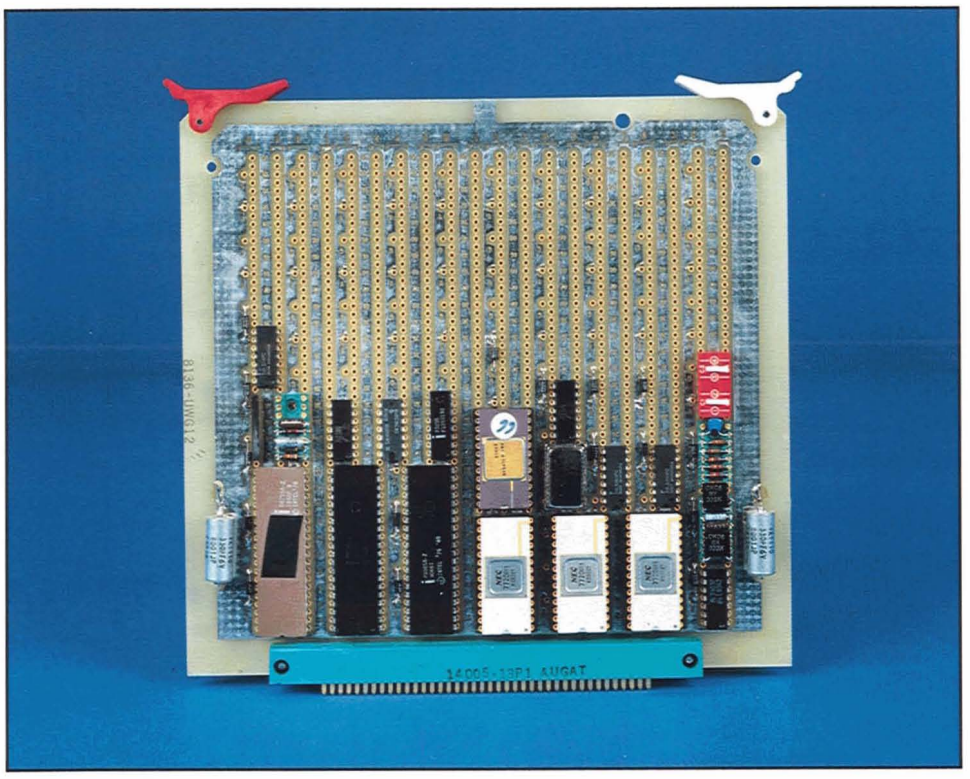


Fig. 24—Hardware for three-chip linear predictive coding (LPC) vocoder of Fig. 23.

sophistication is required. We shall first consider the frame-rate problem and then the parameter-quantization problem. Such methods have proved to be useful in achieving rates as low as 1200 b/sec.

For a further reduction, the methods become more powerful. Notable among these methods is the concept of pattern matching that was invented by C.P. Smith [36, 37] and reinvented more recently under the name *vector quantization* by A. Buzo, A.H. Gray, Jr., R.M. Gray, and J.D. Markel [38]. Pattern-matching techniques appear to give good results in the range of 600 to 1000 b/sec. Further reductions entail some form of phonetic vocoding, which implies some type of speech recognition, and the analysis problem becomes correspondingly more difficult.

It is also worth mentioning formant-tracking methods, which are oriented more toward pattern recognition than toward pattern matching. Here the game is to analyze the speech into fewer parameters than the number used by either channel or LPC vocoders. Formant-tracking methods have a long and complicated history. To summarize, many versions of the approach have been proposed and tried for the past 30 years, but they have failed in moving from the research environment into a more practical setting.

Reduction of Bit Rate by Effective Lowering of Frame Rate

E. McLarnon's fixed-rate method for lowering the frame rate [39] has been widely used in the speech research group at Lincoln Laboratory. In this method, only alternate frames are transmitted. (A variation of the method calls for transmitting every third frame.) A 2-bit code that is also transmitted commands the receiver in the following way:

> Assume frame m is the frame that is not to be transmitted. Then if frame m is sufficiently similar to frame $(m - 1)$, the receiver will reproduce frame $(m - 1)$ as frame m. On the other hand, if frame m is similar to frame $(m + 1)$, then frame $(m + 1)$ is reproduced as frame m. Finally, if frame m occurs dur-

ing a rapid spectral change, frame m might be more similar to the interpolated value between frames $(m+1)$ and $(m-1)$. Since these two frames are available at the receiver, the necessary interpolation can be computed there.

In summary, the above method, which is called *frame fill-in*, achieves a savings of nearly two to one. Blankenship and M.L. Malpass applied the method to both a channel vocoder and an LPC vocoder [40]. Here are their summary and conclusions:

> Methods have been described based on the principle of frame fill-in for developing reduced-rate transmission systems from standard 2400 bits-per-second backbones. It was shown that both channel vocoders and LPC types of vocoders could be adapted with virtually no increase in computational complexity to operate at 1200 or 800 bits-per-second.
>
> It was found through formal and informal evaluation methods that both channel vocoder and LPC-based systems perform quite well at 1200 b/s and would probably be usable in most environments where the 2400-b/s parent could be successfully operated. At 800 b/s both systems were considered marginal and usable only in limited circumstances. However, the channel vocoder was seen to perform incrementally better, probably due to the uniquely efficient parameter coding scheme employed which tends to be less sensitive to quantization inaccuracies.

Reduction of Data Rate by Reduction of the Number of Bits per Frame

The British were the first to implement a very simple and quite successful transformation that used *differential coding* [41]. Their resulting Belgard vocoder coded the lowest-frequency channel with 3 bits and all other channels with 2-bit differential coding. Since the major correlations in a channel vocoder are between adjacent channels, the differential coding was very effective. The technique was also employed by the author to develop a new spectrally flattened channel-vocoder algorithm [42]. Differential coding improved the quality of the algorithm and led to a coding strategy that used 6 bits for the lowest channel and 3-bit differential coding for subsequent channels.

Pattern Matching

In the late 1950s, C.P. Smith [36] introduced the concept of pattern matching for bandwidth reduction. His reasoning was very straightforward. Imagine that you have a channel vocoder with 16 channels and you arbitrarily devote 3 bits to quantize each channel. Thus a total of 48 bits is used. This arrangement implies that your coding strategy is capable of distinguishing among 2^{48} spectral measurements. It is clear that the human ear cannot tell that many patterns apart. Therefore, even if a person could tell any two patterns apart from a total of 2^{20} (i.e., 1,048,576) patterns, 3 bits per channel for a 16-channel system is still highly redundant. Now, given a strategy that permits the identification of the storage location of any one of the million or so patterns, we see that the transmitter needs to transmit only that storage location. Because the receiver has the same set of stored patterns as the transmitter, the receiver will generate the correct spectrum upon receiving the address.

At Lincoln Laboratory, these concepts were applied to D.B. Paul's Spectral Envelope Estimation Vocoder (SEEVOC) system [43, 44] and to the author's channel vocoder [45]. Thus, in the late 1970s and early 1980s, there were at least three implementations of vector quantization. Each used a different vocoder configuration, a different comparison mechanism for measuring differences (e.g., the use of a weighted rms measure), and a different strategy for building a system.

Adaptive Pattern Matching

To deal with the large number of patterns, Paul developed another approach [44] that could adapt to new speakers by continuously altering the pattern table to match the current speaker and environment. In Paul's approach, an incoming spectrum is compared with all existing reference spectra. If the best match fails to satisfy a fixed criterion, the new spectrum is incorporated into the pattern table. The new spectrum replaces that reference spectrum which has not been matched for

the longest time period.

The performance of Paul's system was impressive. When a new speaker began to talk, a brief period of quasi-intelligible vocoded speech was followed by adaptation: the system quickly tuned in on the new speech. It should be noted that the system required the periodic transmission of updated reference-pattern sets to the receiver. Also, the maintenance of a low bit rate required the detection of silence intervals during which new pattern sets could be sent. The system performed well at 800 b/sec. Lincoln Laboratory's SEEVOC algorithm [43] was used in the experiment.

Formant Analysis and Synthesis for Bit-Rate Reduction

Thus far our methods for reducing the vocoder data rate have depended on finding a good parametric description of the speech short-time spectrum, followed by attempts to remove the redundancies in the parameters without unduly disturbing the synthesized speech. If the attempts to remove the redundancies are in any way successful, it will imply the existence of a parameter set that is less redundant than the parameter sets that can be derived, for example, from channel-vocoder, LPC, or homomorphic analyses. Indeed, such a parameter set, consisting of formants, has long been a popular topic of speech research. Using formants, J.L. Flanagan built a device that he called a *terminal-analog synthesizer* along with an associated analyzer [46].

Despite many ingenious efforts, however, formant analysis and synthesis has not led to any practical bandwidth-compression systems. J.N. Holmes [47], R.J. McAulay [48], and G.S. Kang and D.C. Coulter [49] have spearheaded some of the more sophisticated recent research.

Speech-Processing Facilities

More than 30 years ago, researchers in speech processing knew that the digital computer would one day play a major role in speech research. At the time, however, the large and relatively slow digital computer could make only limited contributions in the field. Since then two major events—the development of a comprehensive theory of digital signal processing and rapid progress in the integrated circuit field—have propelled digital technology and computer processing methods to their current status as integral aspects of both speech-processing research facilities and specialized speech hardware devices.

During the 1940s and 1950s, the human-speech-production mechanism was successfully modeled via transmission-line and analog-network theory. Spectrum analysis of the speech wave was conceived as the outputs of a bank of analog bandpass filters and associated energy detectors. Formants were analog tuned circuits. It seemed natural to assume that devices such as the channel vocoder and various formant synthesizers were intrinsically analog devices. Yet it was recognized very quickly that the digital computer did have a role to play in speech research.

By the mid-1950s, in fact, digital computers were used as tools for speech-recognition research at Lincoln Laboratory and at the MIT Research Laboratory of Electronics. In early experiments, the speech had to be spectrum-analyzed first—a task that seemed too formidable for computers—so an analog filter bank had to be incorporated into the system (Fig. 25). In the figure, each output channel is a relatively narrowband signal that has significant spectral components in the region of 0 to 25 Hz. Therefore, sampling each output at 50 Hz preserved spectral information. If, for example, the spectrum analyzer contained 36 channels (as was the case for an early analyzer that was connected to the Lincoln TX-2 computer, which will be discussed in the next section), then the A/D converter would need to receive the incoming samples at a rate of 1800 samples/sec with perhaps 8-bit accuracy. This requirement could be fulfilled in the mid-1950s.

Once the speech samples were safely in the computer, a variety of speech-recognition algorithms could be tested. Using the combination of analog spectrum analyzer and digital computer to perform research on vowel and fricative recog-

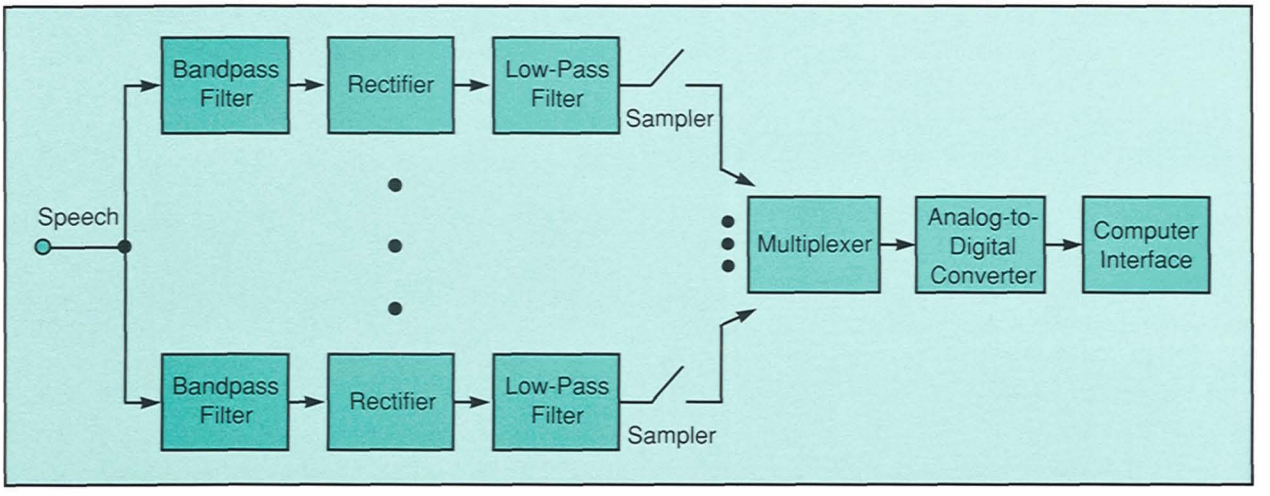


Fig. 25—Analog spectrum analyzer feeding a digital computer.

nition, J.W. Forgie and C.D. Forgie [50–52] made inroads into the massive problem of automatic recognition of connected speech.

In the 1960s, computer methods greatly augmented research and development on channel vocoders. Lincoln Laboratory and MIT were also heavily involved in the development of DSP theories and techniques, and Lincoln Laboratory engineers took advantage of the latest IC technology to build computers that would be more effective as speech research tools. Four distinct systems were developed: the TX-2, the Fast Digital Processor (FDP), the Lincoln Digital Voice Terminal (LDVT), and the Lincoln Digital Signal Processor (LDSP). Each computer had an impact on speech research and each was in turn influenced by the continuing development of speech algorithms.

The TX-2 Computer

The TX-2 computer [53–55], designed at Lincoln Laboratory, followed previous efforts that led to computers such as Whirlwind [56] and the Memory Test Computer [57]. TX-2 was a very large computer dedicated primarily to Lincoln Laboratory staff members for research in computer technology and in algorithmic work in such fields as pattern recognition and the simulation of nervous-system functions. Figure 26 shows the console of TX-2. It is interesting that the entire system could be run by a single user.

In the extreme right of Fig. 26 is a panel of lights that are associated with the arithmetic element. Between the two people in the photo are 24 toggle-switch registers that could be accessed by the computer as if they were part of the main memory.

Unlike most computers of that period, TX-2 was very flexible and was often adapted to the needs of an individual user. It was designed with a large number of available input and output ports. Thus, in addition to accommodating conventional peripheral devices such as tape units, drums, and printers, one could also develop special hardware to interface with TX-2. During its time, TX-2 served as a test bed for new memory developments, new computers, experimental display facilities, and various other specialized user-interface devices.

TX-2 had many properties and features that made the computer useful for speech processing. The following is a partial list.

a. TX-2 was managed in a nonbureaucratic way. A user could interact directly with TX-2 without any help from an intermediary. The degree of interaction was equivalent to present-day workstations (such as those manufactured by Sun Microsystems [58]) and the SPIRE system [59].

b. Results could be given to a user on a CRT, on a printer, or by acoustic means.

c. A large core memory permitted the storage of about 10 seconds of speech.

d. Programs could be quickly modified on line via the toggle switches and knobs mentioned above. This feature permitted users to set parameters and control programs while the programs were running.

The TX-2 facility played a key role in vocoder research at Lincoln Laboratory and enabled the following accomplishments:

a. development of the parallel-processing pitch-detection algorithm described in the section "Pitch Detection,"

b. testing of the pitch-detection algorithm by the connection of TX-2 to real-time vocoder hardware, and

c. simulation of a complete channel vocoder. With the development of DSP theory, it became feasible to simulate a complete channel vocoder on TX-2. Real-time simulation was out of the question but the enormous increase in algorithmic speed turned simulation into a practical alternative for research on a complete vocoder. In 1959, M. Mathews [60] computed the time to process one second of speech to be several hours. With TX-2, the time was reduced to several minutes.

Of course, DSP algorithms, such as digital filtering and FFT, improved the vocoding performance of all computers. But the advantage of the TX-2 facility as a vehicle for speech-processing research stemmed from its mode of operation as a user-interactive, private workstation. TX-2's large memory, excellent display facility, and flexible I/O, coupled with new DSP algorithmic methods, resulted in a very sophisticated speech-processing facility for that era (the early 1960s).

However, perhaps the most critical property of an adequate facility—that of real-time simulation—had not been attained. Experience has taught us that a new speech-processing system has not really been fully evaluated until it has processed an immense amount of speech material of different speakers and different environments. The only practical method of fulfilling such criteria is to build real-time hardware or to have available a facility capable of real-time simulation.

Fig. 26—TX-2 computer.

The Fast Digital Processor (FDP)

During the 1960s, researchers at Lincoln Laboratory felt that recent advances in high-speed circuits coupled with DSP theoretical advances made it possible to perform real-time simulation of software, such as vocoding algorithms, that required great computational speeds. Thus in 1967 Lincoln Laboratory initiated a study that led to the design and construction of the signal processing computer named the Fast Digital Processor [35]. FDP was the first of the class of computers that are now called *array processors*.

The FDP circuitry was second-generation emitter-coupled logic (ECL) with gate-switching times of about 3.5 nsec. Even with such fast switching times, Lincoln Laboratory researchers discovered that many useful speech algorithms could not be simulated in real time if a straightforward sequential computer structure was adopted. Consequently, computing strategies that used parallelism and pipelining were implemented. Figure 27 is an architectural sketch of FDP that helps to illustrate that

a. four parallel arithmetic elements and dual data memories were implemented to speed up complex arithmetic operations for FFT programs. Digital filtering and autocorrelation programs were also speeded up.

b. program memory and data memory were physically separated so that both of the system's memories could be simultaneously accessed (referred to as the Harvard architectural style).

c. a horizontal microcode structure allowed for the simultaneous manipulation of address and data.

d. instructions were pipelined. A typical instruction required three cycles (each cycle took 150 nsec) so that three instructions were usually in the pipeline at any instant.

Because FDP's input and output capabilities were limited, a Univac-1219 computer, to which FDP served as a peripheral, was in charge of I/O control. For real-time processing of analog signals such as speech, A/D and D/A converters were incorporated. Later, a core memory of 128k 18-bit words and a display were connected. The display was designed for fast scanning so that a real-time spectrogram could be implemented.

If we estimate the degree of parallelism in FDP to be about eight times that of a standard sequential computer, then we can approximate the FDP throughput to be about 50 MIPS.

The FDP-Univac Signal Processing Facility

FDP was a physically large device. Figure 28, in which the computer stands in the foreground of the photo, gives a slightly exaggerated view of FDP's size. Inside the two sets of large double doors are the four arithmetic elements and the control element. To the right of the doors are drawers that house the memory, and surrounding the drawers are the power supplies. The Univac-1219 computer (the main frame to which the FDP was attached) is just to the right of the FDP.

At the assembly-language level, it was appreciably more difficult to program FDP than it was to program a simple Von Neumann computer. Furthermore, array processors are not only more difficult to program but also more difficult to debug because the user must keep in mind many more states. No attempt was made to write a high-level language for the FDP even though such an effort might have led to more widespread use of the facility. Nevertheless, valuable research was conducted at the facility:

- the first real-time simulation of an LPC vocoder was implemented [61],
- a real-time speech spectrogram was computed on FDP and displayed with the help of special-purpose hardware [62], and
- a high-speed non-real-time spectrally flattened channel vocoder was simulated on FDP [63].

The facility also proved to be useful in several radar projects.

The next generation of ECL was more than twice as fast as the circuitry used for FDP. This increase led to new ideas toward improving the facility; one of the evolving concepts was that of a speech-processing device that could serve both as a facility and as a field device. The concept led to the design and construction of

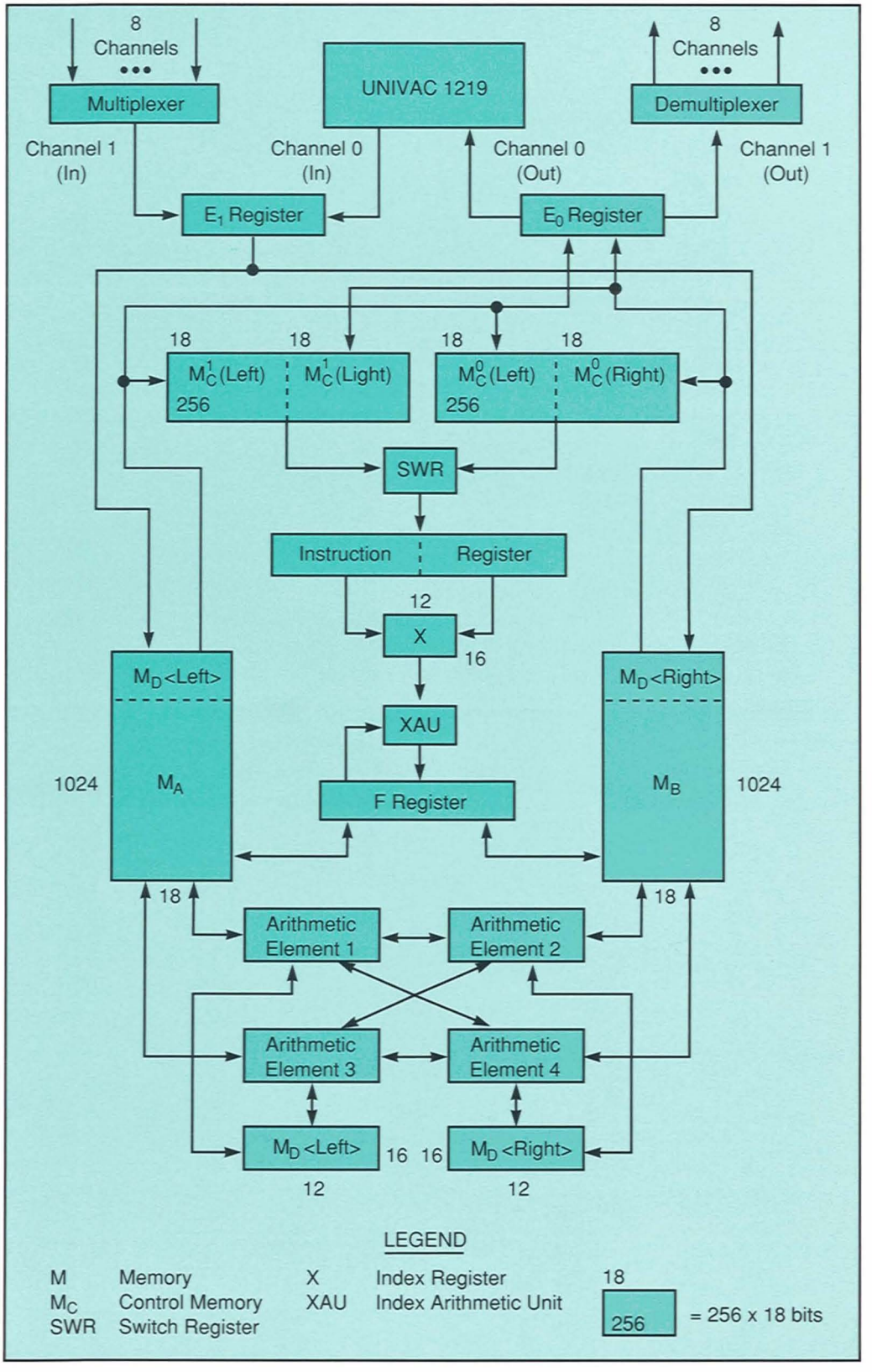


Fig. 27—Schematic of Fast Digital Processor (FDP).

several Lincoln Digital Voice Terminals.

The Lincoln Digital Voice Terminals (LDVT)

Early design studies [64] showed that with ECL technology and a careful design, a simple Harvard-style uniprocessor could support a basic cycle time of 55 nsec. The fast cycle time meant that with appropriate pipelining the raw speed of an LDVT would be approximately three times greater than that of FDP. It was also estimated that the proposed structure could result in real-time simulation of many types of vocoders, including LPC. And, in fact, a wide variety of vocoder algorithms was implemented in real-time LDVTs, both in the laboratory and in the field. Hofstetter, Blankenship, Malpass, and Seneff describe the algorithms and their implementations [65], and Ref. 66 describes the LDVT architecture.

LDVTs had several advantages over FDP: they were substantially smaller and cheaper, they were much easier to program, and they performed better for most problems. Size constraints, however, limited the amount of memory in an LDVT. Another important limitation was that of I/O capability. (Because the mission of LDVTs was to be real-time hardware devices, there was no need to include a truly flexible system that could easily communicate with various peripherals, including other computers.) However, our experience with LDVTs and the programs they implemented brought to light the many advantages of small, fast, easy-to-program computers. Thus, when the opportunity arose to construct a new generation of signal processors, we used LDVT as the starting point and added more memory along with a more capable I/O system.

The Lincoln Digital Signal Processor (LDSP)

Figure 29 shows the architecture of the successor to LDVT: the LDSP. Figure 30 is a photo of the rack-mounted LDSP hardware, which consisted of four LDSPs controlled by a PDP-11

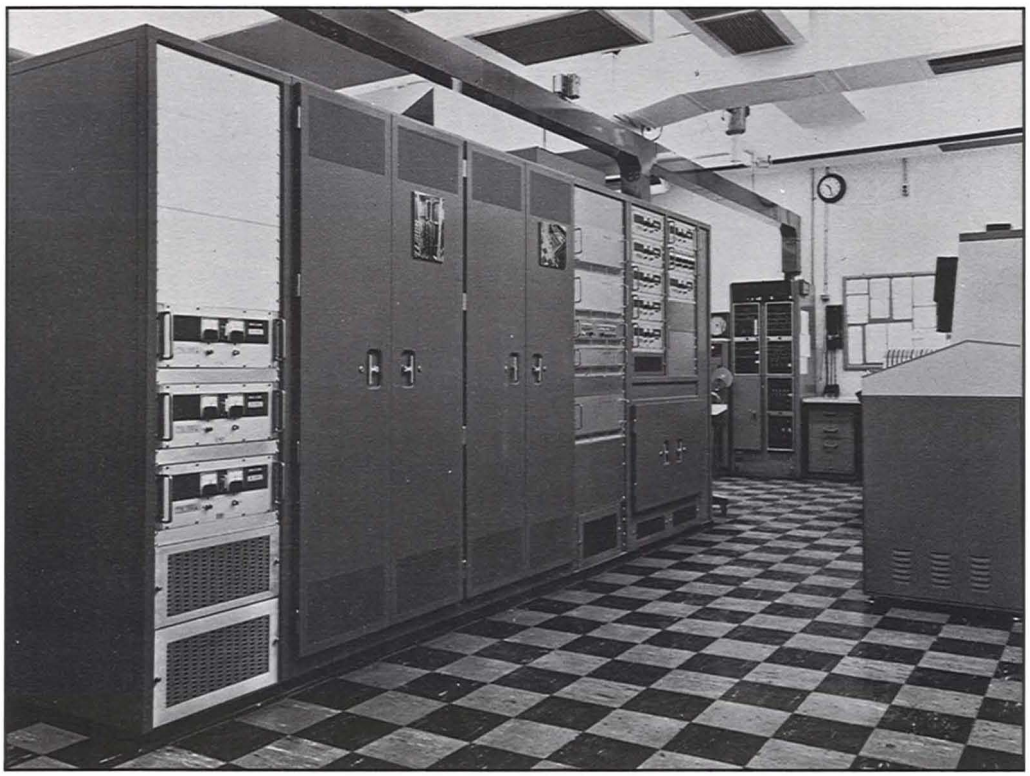


Fig. 28—Fast Digital Processor (FDP) facility.

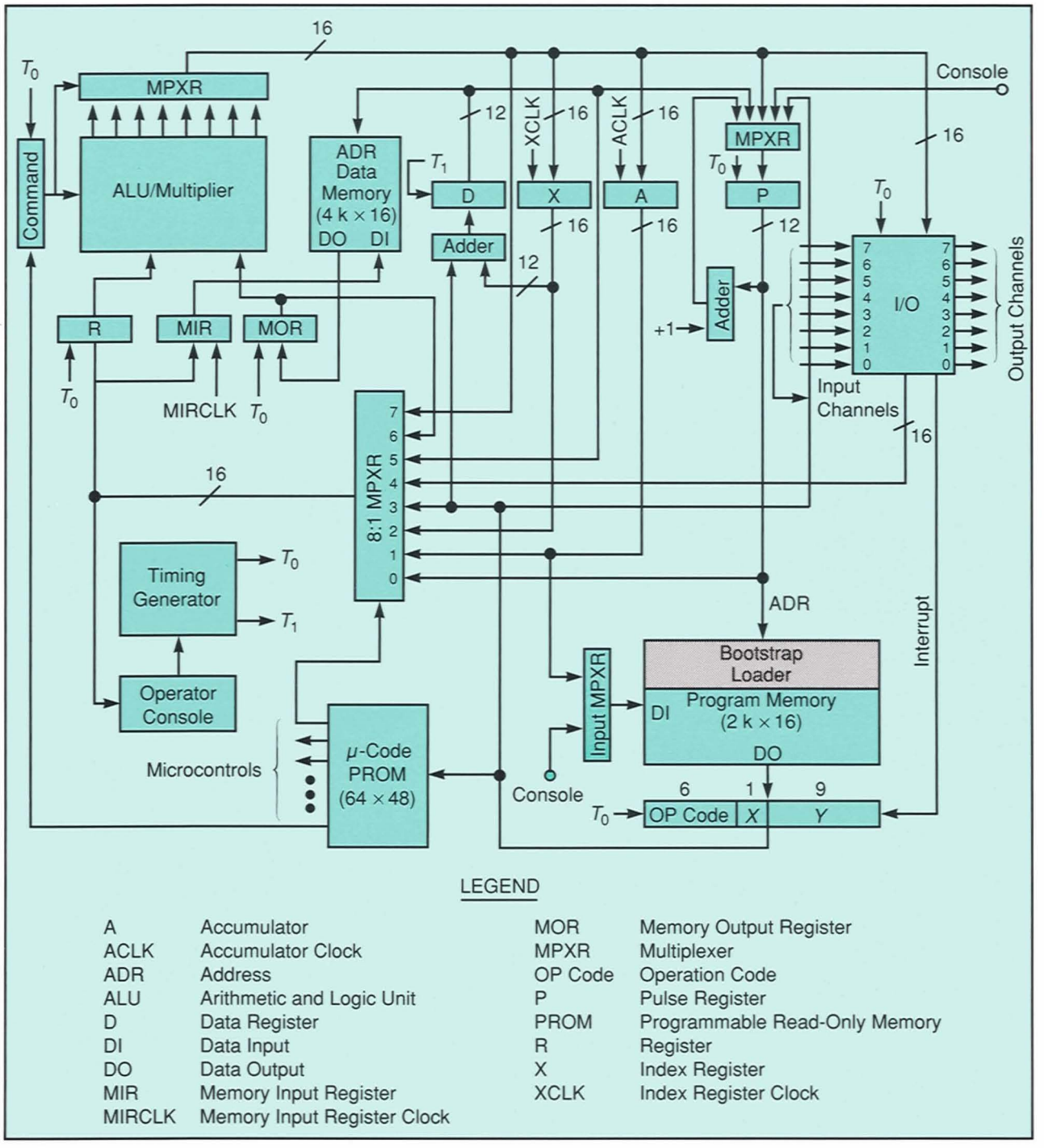


Fig. 29—Schematic of Lincoln Digital Signal Processor (LDSP).

computer, and a signal processor for each LDSP.

As a result of the development of the LDVTs and LDSP, Lincoln Laboratory established a facility where the signal processing machines were connected to a single general-purpose computer. LDVT or LDSP operation in a real-time mode required downloading a program from the general-purpose computer. The new facility used time sharing but also functioned effectively for real-time simulation of vocoder algorithms. For fast, non-real-time processing, speech had to be entered into the disk system in

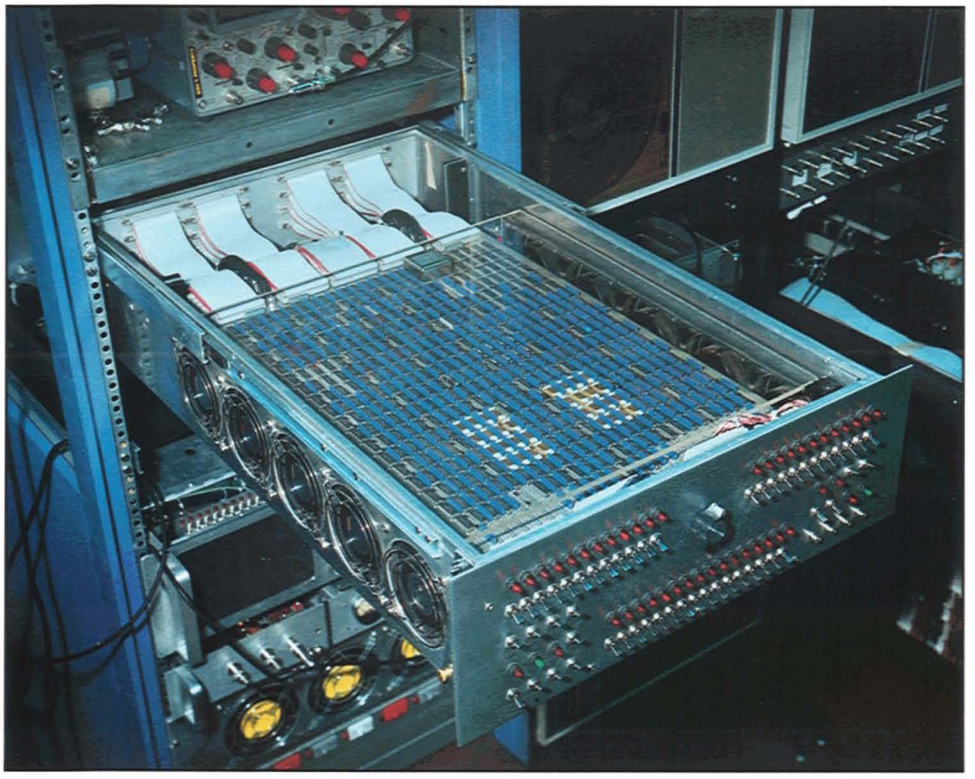


Fig. 30—The Lincoln Digital Signal Processor (LDSP).

real time at a sampling rate as high as 15,000 samples/sec. Lincoln Laboratory designed an intricate piece of real-time software to perform this task [67]. Another important addition was audio signal processing such as filtering, sampling, preemphasis and deemphasis of high frequencies, and A/D and D/A conversion [68, 69].

Among other duties, the LDSP facility has conducted

a. new algorithm research,
b. diagnostic rhyme testing of vocoder systems,
c. psychophysics testing,
d. testing of new signal processing chips, and
e. systems testing using LDSPs as components.

Several examples of research are described below in more detail.

Packet communications, as exemplified by ARPANET [70], is a new method of transmitting and receiving information. The data to be transmitted are divided into digitized blocks, or packets, and transmitted in a burst mode. The packets are then reassembled at the receiving end.

An attractive feature of packetized speech is its inherent flexibility in regard to data rates for either speech or data. This flexibility inspired the idea that variable-rate vocoders could be of benefit in a packet environment. The goal of one of the earliest simulation efforts on the LDSP was to create a useful variable-rate algorithm. The vehicles chosen were the channel vocoder and the subband coder [71]. This combination was particularly convenient because the bandpass filters in each channel could serve as a carrier of either vocoding information or waveform-coding information.

LPC algorithms and variations were also implemented in real time on one or more LDSPs. The 2400-b/sec government standard [72] and the Lincoln Laboratory 2400-b/sec version were implemented, each on a single LDSP. A variation on the Lincoln LPC added spectral flattening.

To improve the system's performance in the noisy environment of an advanced fighter aircraft, an updated version of the LPC algorithm

was implemented on the Advanced Linear Predictive Coding Microprocessor [73]. In this version, the speech bandwidth was increased to 5 kHz so that more parameters could be transmitted and the analysis rate was doubled [74]. Frame-fill techniques (described in the section "Low-Rate Vocoder Systems") were used to achieve a 2400-b/sec data rate.

McAulay experimented with an LPC analyzer and a spectrally flattened channel-vocoder synthesizer [48]. He also implemented a noise-reduction scheme [75] that has been used as a preprocessor for both LPC vocoders and a simple word-recognition device.

Using software based on the Belgard channel-vocoder algorithm, several Lincoln researchers wrote a program to improve vocoder performance over a high-frequency link. The idea was that the receiving modem would request that the block of data be repeated whenever the channel error rate exceeded a threshold. During the repeat mode the receiving vocoder remained silent. When the channel returned to normal mode the vocoder was made to speak at a faster rate so that the system could catch up with the transmitted signal. Reference 76 describes the work in greater detail.

McAulay and T.F. Quatieri have pioneered a new analysis-synthesis system based on a sinusoidal representation of the speech signal. At high data rates the system produces speech (or music) that is indistinguishable from the input [77, 78]. The incorporation of high-speed programmable microprocessors enabled the system to operate on a real-time basis [79]. Researchers are currently trying to develop a good, low-rate model of the excitation parameters.

In some cases, an algorithm was so computation intensive that real-time implementation of the algorithm on a single LDSP was not possible. Because the Lincoln Laboratory facility comprised four LDSPs connected to a general-purpose processor (the LDSPs were also able to communicate with each other), it was often possible to handle these computation-intensive algorithms with multiple-LDSP real-time simulation.

Summary and Prospects for the Future

Lincoln Laboratory's research on speech bandwidth compression began with the application of parallel processing to the formidable task of detecting the fundamental frequency of speech. This successful effort led to several years of vocoder design and implementation and featured the novel ideas of spectral flattening and the use of banks of linear-phase-bandpass filters. Innovation in vocoder hardware design was best exemplified by the construction of the first pitch detector that used integrated circuits.

Lincoln Laboratory also pioneered the use of computer simulation in vocoder research. This early work culminated in the late 1960s with the Lincoln Experimental Terminal Vocoder (Letvoc), the first practical vocoder for satellite communications [80].

Following a brief period during which the efforts of speech researchers were directed toward the emerging theory and technology of DSP, FDP was designed, built, and used to program the first real-time LPC vocoder. Following that period, speech and DSP research advanced in parallel, and Lincoln Laboratory built LDVT and LDSP and invented and implemented many different vocoder systems via real-time programs. Recent efforts have included the Sine Transform Coder and its implementation on advanced signal processing chips.

The descriptions of facilities in this article have taken us chronologically from some of the very early results of computers as speech processors to the present. The history should help us to determine what kind of facility is most suitable for speech research. In this respect, two items are of major significance:

1. The facility should have the features of a personal workstation, with good displays, flexible I/O, and versatile audio-signal conditioning. The user should be able to enter material easily from tapes or cassettes into computer memory and to display waveforms, spectra, and processing results.

2. The facility should be capable of high-speed processing so that real-time operation can be implemented when necessary. Experience has taught us that a large number of vocoder algorithms cannot be adequately tested unless real-time operation is available.

Electronic circuits are becoming faster and denser, a trend that makes the observer wonder what to expect of future generations of speech-processing facilities. For example, perhaps a single LDSP that is substantially more powerful than the present LDSP would permit implementation of the vocoder algorithm described above on a single LDSP rather than three. Because speed (not memory) is the only limitation that prevents real-time simulation of this particular algorithm from being implemented onto a single LDSP, we see that either faster chips or a more highly parallel architecture is needed. Present VLSI technology has yet to produce faster chips than those already in use in the LDSP. Consequently, greater parallelism is called for, which leads to the tentative conclusion that advances must be made both in parallel-processing hardware and, perhaps most importantly, software to accommodate such hardware.

Although real-time processing is often essential, the detailed diagnosis of speech-processing results is also vital. These two requirements are conflicting because, for example, one cannot simultaneously be looking at all spectral cross sections in a real-time system. There are two possible solutions. With sufficient speed and memory, computers can be programmed so that diagnostic information is compiled in real time while an algorithm is running. A researcher can then scan back and examine the microscopic results. The second approach is to use non-real-time algorithms (identical to the real-time algorithms) to study the results on a non-real-time basis.

Problems persist in the vocoder field. For instance, present-day vocoders have difficulty handling adverse environmental conditions such as acoustically noisy backgrounds. Because the human auditory system does a good job of comprehending speech under such conditions, the question arises as to the feasibility of building a vocoder analyzer that performs comparable functions.

At the very outset of modeling the human auditory system, we are faced with a giant computational problem. The auditory nerve contains about 30,000 fibers, each of which has properties not unlike that of a tuned circuit. The faithful simulation of such a system is a large undertaking, but it would not be surprising if future technology brought us close to such capabilities. The larger challenge is to understand in greater detail the intricate workings of the human auditory system.

Acknowledgments

The contributions made by Lincoln Laboratory for the past 30 years in the art and science of vocoders includes the work and efforts of many staff members. Here is at best an incomplete list: Peter Blankenship, Paul Demko, Norm Daggett, John Harris, Joel Feldman, Ed Hofstetter, Marilyn Malpass, Bob McAulay, Jim Forgie, Al McLaughlin, Al Oppenheim, Doug Paul, Tom Quatieri, Charles Rader, Paul Rosen, Stephanie Seneff, Vinnie Sferrino, and Joe Tierney. The author would like to offer his special thanks to Paul Rosen, whose initiative created the atmosphere for this extended period of research; to Joe Tierney, whose continued involvement contributed much substance to the work; and to Peter Blankenship, whose encouragement and thoughtful editing enhanced this article.

References

1. H. Dudley and T. H. Tarnoczy, "The Speaking Machine of Wolfgang von Kempelen," *J. Acoust. Soc. Am.* **22,** 151 (1950).
2. W. von Kempelen, *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine* (1791).
3. J.L. Flanagan, *Speech Analysis: Synthesis and Perception*, 2nd ed. (Springer-Verlag, New York, 1972).
4. H. Dudley, R.R. Riesz, and S.A. Watkins, "A Synthetic Speaker," *J. Franklin Inst.* **227,** 739 (1939).
5. H. Dudley, "The Vocoder," *Bell Lab. Re.* **17,** 122 (1939).
6. W.R. Bennett, "Secret Telephony as a Historical Example of Spread-Spectrum Communication," *IEEE Trans. Comm.* **COM-31,** 98 (1983).
7. F.B. Colton, "The Miracle of Talking by Telephone," *National Geographic*, 395 (Oct. 1937).
8. P. Lieberman, "Perturbations in Vocal Pitch," *J. Acoust. Soc. Am.* **33,** 597 (1961).
9. P. Lieberman, private communication.
10. R.J. Ritsma, "Frequencies Dominant in the Perception of the Pitch of Complex Sounds," *J. Acoust. Soc. Am.* **42,** 191 (1967).
11. M.M. Sondhi, "New Methods of Pitch Extraction" *IEEE Trans. Audio and Electroacoust.* **AU-16,** 262 (June 1968).
12. J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Trans. Audio Electroacoust.* **AU-20,** 367 (Dec. 1972).
13. A.M. Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.* **41,** 293 (1967).
14. B. Gold and L.R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain" *J. Acoust. Soc. Am.* **46,** 442 (1969).
15. B. Gold "Note on Buzz-Hiss Detection," *J. Acoust. Soc. Am.* **36,** 1659 (1964).
16. N.L. Daggett, "A Computer for Vocoder Pitch Extraction," *Technical Report TN-1966-3*, Lincoln Laboratory (18 Feb. 1966), DTIC #AD-629361.
17. J.A. Feldman, "A Compact Digital Channel Vocoder Using Commercial Devices," *Proc. ICASSP '82* **3,** *Paris, 3–5 May 1982*, p. 1960.
18. S. Seneff "Real-Time Harmonic Pitch Detector," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-26,** 358 (1978).
19. J.W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data," *Proc. EASCON '74, Washington, 7–9 Oct. 1974*, p. 673.
20. M. Miller and M. Sachs, "Representation of Voice Pitch in Discharge Patterns of Auditory-Nerve Fibers," *Hearing Research* **14,** 257 (1984).
21. G. Fant, *Acoustic Theory of Speech Production* (Mouton & Co., The Hague, 1960).
22. L. Rabiner and R. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ, 1978).
23. G. Fant, J. Mártony, U. Rengman, and A. Risberg, "The OVE II Speech Synthesizer," *Speech Communications Seminar, Stockholm, Sept. 1963.*
24. B. Gold and J. Tierney, "Pitch-Induced Spectral Distortion in Channel Vocoders," *J. Acoust. Soc. Am.* **35,** 730 (L) (1963).
25. G. von Bekesy, *Experiments in Hearing* (McGraw-Hill, New York, 1960).
26. N.Y.S. Kiang, "Discharge Patterns of Single Fibers in the Cat's Auditory Nerve," Research Monograph No. 35 (MIT Press, Cambridge, MA., 1965).
27. B. Gold and C.M. Rader, *Digital Processing of Signals* (McGraw-Hill, New York, 1969).
28. E.M. Hofstetter, "An Introduction to the Mathematics of Linear Predictive Filtering as Applied to Speech Analysis and Synthesis," *Technical Note TN-1973-36, Revision 1*, Lincoln Laboratory (12 Apr. 1974), DTIC #AD-777579.
29. B.S. Atal and S.L. Hanauer, "Speech Analysis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.* **50,** 637 (1971).
30. J.D. Markel and A.H. Gray, Jr., "On Autocorrelation Equations as Applied to Speech Analysis," *IEEE Trans. Audio Electroacoust.* **AU-21,** 69 (1973).
31. B. Bogert, M. Healy, and J. Tukey, "The Quefrency Analysis of Time Series for Echoes," *Proc. Symp. on Time Series Analysis*, ed. M. Rosenblott (John Wiley, New York, 1963), p. 209.
32. A.V. Oppenheim, "Superposition in a Class of Non-Linear Systems," *Technical Report 432*, MIT R.L.E. (31 Mar. 1965), DTIC #AD-815344.
33. A.V. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," *J. Acoust. Soc. Am.* **45,** 458 (1969).
34. E.M. Hofstetter, J. Tierney, and O. Wheeler, "Microprocessor Realization of a Linear Predictive Vocoder," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-25,** 379 (1977).
35. B. Gold, I.L. Lebow, P.G. McHugh, and C.M. Rader, "The FDP, a Fast Programmable Signal Processor," *IEEE Trans. Comput.* **E-20,** 33 (1971).
36. C.P. Smith, "Voice Communications Methods Using Pattern Matching for Data Compression," *J. Acoust. Soc. Am.* **35,** 805(A) (1963).
37. C.P. Smith, "Perception of Vocoder Speech Processed by Pattern Matching," *J. Acoust. Soc. Am.* **46,** 1562 (1969).
38. A. Buzo, A.H. Gray, Jr., R.M. Gray, and J.D. Markel, "Speech Coding Based upon Vector Quantization," *IEEE Trans. Acoust. Speech. Signal Process.* **ASSP-28,** 562 (1980).
39. E. McLarnon, "A Method for Reducing the Frame Rate of a Channel by Using Frame Interpolation," *Proc. ICASSP '78, Washington, 10–12 Apr. 1978*, p. 458.
40. P. Blankenship and M.L. Malpass, "Frame-Fill Techniques for Reducing Vocoder Data Rates" *Technical Report 556*, Lincoln Laboratory (26 Feb. 1981), DTIC #AD-A-099395.
41. J.N. Holmes, "The JSRU Channel Vocoder," *IEE Proc. Communications, Radar, and Signal Processing* **127,** *Feb. 1980*, Part F, No. 1, p. 53.
42. B. Gold, P.E. Blankenship, and R.J. McAulay, "New Applications of Channel Vocoders," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-20,** 13 (1981).
43. D.B. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-29,** 786 (1981).
44. D.B. Paul, "An 800 BPS Adaptive Vector Quantization Vocoder Using a Perceptual Distance Measure," *Proc. ICASSP '83* **1,** *Boston, 14–16 Apr. 1983*, p. 73.
45. B. Gold, "Experiments with a Pattern-Matching Channel Vocoder," *Proc. ICASSP '81, Atlanta, 30 Mar.–1 Apr. 1981*, p. 32.
46. J.L. Flanagan, "Note on the Design of 'Terminal-Analog' Speech Synthesizers," *J. Acoust. Soc. Am.* **29,** 306 (1957).
47. J.N. Holmes, "Parallel Formant Vocoders," *IEEE EASCON '78 Washington, 25–27 Sept. 1978*, p. 713.
48. R.J. McAulay, "A Low-Rate Vocoder Based on an Adaptive Subband Formant Analysis," *Proc. ICASSP '81, Atlanta, 30 Mar.–1 Apr. 1981*, p. 28.
49. G.S. Kang and D.C. Coulter, "600 BPS Voice Digitizer,"

Proc. ICASSP '76, Philadelphia, 12–14 Apr. 1976, p. 91.
50. J.W. Forgie and C.D. Forgie, "Results Obtained from a Vowel Recognition Computer Program," *J. Acoust. Soc. Am.* **31,** 1480 (1959).
51. J.W. Forgie and C.D. Forgie, "A Computer Program for Recognizing the English Fricative Consonants |f| and |θ|," Fourth Intl. Congress on Acoustics, Copenhagen, Aug. 1962.
52. J.W. Forgie, personal communication.
53. J.M. Frankovich and H.P. Peterson, "A Functional Description of the Lincoln TX-2 Computer," *Proc. Western Joint Computer Conf., Los Angeles, 26–28 Feb. 1957*, p. 146.
54. J.W. Forgie, "The Lincoln TX-2 Input-Output System," *Proc. Western Joint Computer Conf., Los Angeles, 26–28 Feb. 1957*, p. 156.
55. W.A. Clark, "The Lincoln TX-2 Computer Development," *Proc. Western Joint Computer Conf., Los Angeles, 26–28 Feb. 1957*, p. 143.
56. K.S. Redmond and T.M. Smith, *Project Whirlwind: The History of a Pioneer Computer* (Digital Press, Bedford, MA, 1980).
57. P. Bagley, "Memory Test Computer Programming Reference Manual," *Memo 6M-2527*, Lincoln Laboratory (25 Nov. 1953), revised 9 May 1955.
58. *System Manager's Manual for the Sun Workstation, Models 100U/150U* (Sun Microsystems, Inc., CA, 1983).
59. V. Zue, personal communication, 1983.
60. M. Mathews, "The Effective Use of Digital Simulation for Speech Processing," *Proc. Sem. on Speech Compression and Processing, Air Force Cambridge Research Center, Sept. 1959.*
61. E.M. Hofstetter, personal communication, 1971.
62. T. Bially, personal communication, 1971.
63. P. Demko and J. Tierney, personal communication, 1971.
64. A.J. McLaughlin and P.E. Blankenship, personal communication, 1975.
65. E.M. Hofstetter, P.E. Blankenship, M.L. Malpass, and S. Seneff, "Vocoder Implementations on the Lincoln Digital Voice Terminal," *Proc. EASCON, Washington, 29 Sept.–1 Oct. 1975*, p. 32A.
66. P.E. Blankenship, "LDVT: High Performance Minicomputer for Real-Time Speech Processing," *Proc. EASCON, Washington, 29 Sept.–1 Oct. 1975*, p. 214A.
67. D. James, personal communication, 1974.
68. J. Tierney, personal communication, 1975.
69. D. Chapman, "The LDSP Signal Conditioner," *Lincoln Laboratory Memo* (1978).
70. J.W. Forgie, "Speech Transmission in Packet Switched Store and Forward Networks," *Proc. National Computer Conf. 1975* , p. 137.
71. T. Bially, B. Gold, and S. Seneff, "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks," *IEEE Trans. on Comm.* **COM-28,** 325 (1980).
72. T.F. Tremain, "The Government Standard Linear Predictive Coding Algorithm LPC-10," *Speech Technology* **1,** 40 (1982).
73. E.M. Hofstetter, E. Singer, and J. Tierney, "A Programmable Voice Processor for Fighter Aircraft Applications," *Technical Report 653*, Lincoln Laboratory (18 Aug. 1983), DTIC #AD-A-133780.
74. E. Singer and J. Tierney, "Implementation of a Robust 2400 b/s LPC Algorithm for Operation in Noise Environments," *Technical Report 766*, ESD-TR-86-166, Lincoln Laboratory (1 Apr. 1987), DTIC #AD-A-180644.
75. G. Neben, R.J. McAulay, and C.J. Weinstein, "Experiments in Isolated Word Recognition Using Noisy Speech," *Proc. ICASSP '83* **3,** *Boston, 14–16 Apr. 1983*, p. 1156.
76. B. Gold, J. Lynch, and J. Tierney, "Vocoded Speech through Fading Channels," *Proc. ICASSP '83* **1,** *Boston, 14–16 Apr. 1983*, p. 101.
77. R.J. McAulay and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," *Technical Report 693*, Lincoln Laboratory (17 May 1985), DTIC #AD-A-157023.
78. T.F. Quatieri and R.J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," *Technical Report 717*, Lincoln Laboratory (16 May 1986), DTIC #AD-A-169740.
79. R. J. McAulay, personal communication.
80. J. Tierney, B. Gold, V. Sferrino, J.A. Dumanian, and E. Aho, "Channel Vocoder with Digital Pitch Extractor," *J. Acoust. Soc. Am.* **36,** 1901 (1964).

BERNARD GOLD is a senior staff member of the Machine Intelligence Technology Group, where he specializes in speech processing. Before joining Lincoln Laboratory 37 years ago, Ben worked for Hughes Aircraft Co. in Los Angeles. He received the following degrees in electrical engineering: a bachelor's from the City College of New York, and a master's and Ph.D. from Brooklyn Polytechnic Institute. He is a fellow of the ASA and IEEE, and a member of the National Academy of Engineering.