sections, lend further support to the link. Psychological and physiological experimentation continue to serve jointly in expanding knowledge about the processes involved in converting the mechanical motions of the inner ear into intelligence-preserving neural activity.

The physiological-psychoacoustic correlations which have been put forward here have involved only the simplest of signals – generally, signals that are temporally punctuate or spectrally discrete, or both. Furthermore, the correlations have considered only gross and salient features of these signals, such as periodicity or time of occurrence. The primary aim has been to outline the peripheral mechanism of the ear and to connect it with several phenomena in perception. Little has been said about classical psychoacoustics or about speech perception. As the stimuli are made increasingly complex – in the ultimate, speech signals – it seems clear that more elaborate processing is called into play in perception. Much of the additional processing probably occurs centrally in the nervous system. For such perception, the correlations that presently can be made between the physiological and perceptual domains are relatively rudimentary. As research goes forward, however, these links will be strengthened.

The literature on hearing contains a large corpus of data on subjective response to speech and speech-like stimuli. There are, for example, determinations of the ear's ability to discriminate features such as vowel pitch, loudness, formant frequency, spectral irregularity and the like. Such data are particularly important in establishing criteria for the design of speech transmission systems and in estimating the channel capacity necessary to transmit speech data. Instead of appearing in this chapter, comments on these researches have been reserved for a later, more applied discussion where they have more direct application to transmission systems.

V. Techniques for Speech Analysis

The earlier discussion suggested that the encoding of speech information might be considered at several stages in the communication chain. On the transmitter side, the configuration and excitation of the vocal tract constitute one description. In the transmission channel, the transduced acoustic waveform is a signal representation commonly encountered. At the receiver, the mechanical motion of the basilar membrane is still another portrayal of the information. Some of these descriptions exhibit properties which might be exploited in communication.

Efforts in speech analysis and synthesis frequently aim at the efficient encoding and transmission of speech information[1]. Here the goal is the transmission of speech information over the smallest channel capacity adequate to satisfy specified perceptual criteria. Acoustical and physiological analyses of the vocal mechanism suggest certain possibilities for efficient description of the signal. Psychological and physiological experiments in hearing also outline certain bounds on perception. Although such analyses may not necessarily lead to totally optimum methods for encoding and transmission, they do bring to focus important physical constraints. Transmission economies beyond this level generally must be sought in linguistic and semantic dependencies.

The discussions in Chapters II and III set forth certain fundamental relations for the vocal mechanism. Most of the analyses presumed detailed physical knowledge of the tract. In actual communication practice, however, one generally has knowledge only of some transduced version of the acoustic signal. (That is, the speaker does not submit to measurements on his vocal tract.) The acoustic and articulatory parameters of the preceding chapters must therefore be determined from the speech signal if they are to be exploited.

This chapter proposes to discuss certain speech analysis techniques which have been found useful for deriving so-called "information-bearing elements" of speech. Subsequent chapters will consider synthesis of speech from these low information-rate parameters, perceptual criteria appropriate to the processing of such parameters, and application of analysis, synthesis and perceptual results in complete transmission systems.

5.1. Spectral Analysis of Speech

Frequency-domain representation of speech information appears advantageous from two standpoints. First, acoustic analysis of the vocal mechanism shows that the normal mode or natural frequency concept permits concise description of speech sounds. Second, clear evidence exists that the ear makes a crude frequency analysis at an early stage in its processing. Presumably, then, features salient in frequency analysis are important in production and perception, and consequently hold promise for efficient coding. Experience supports this notion.

Further, the vocal mechanism is a quasi-stationary source of sound. Its excitation and normal modes change with time. Any spectral measure applicable to the speech signal should therefore reflect temporal features of perceptual significance as well as spectral features. Something other then a conventional frequency transform is indicated.

[1] Other motivating objectives are: basic understanding of speech communication, voice control of machines, and voice response from computers.

5.11. Short-Time Frequency Analysis

The conventional mathematical link between an aperiodic time function $f(t)$ and its complex amplitude-density spectrum $F(\omega)$ is the Fourier transform-pair

$$F(\omega)=\int_{-\infty}^{\infty} f(t)\,e^{-j\omega t}\,dt$$
$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\,e^{j\omega t}\,d\omega\,. \tag{5.1}$$

For the transform to exist, $\int_{-\infty}^{\infty}|f(t)|\,dt$ must be finite. Generally, a continuous speech signal neither satisfies the existence condition nor is known over all time. The signal must consequently be modified so that its transform exists for integration over known (past) values. Further, to reflect significant temporal changes, the integration should extend only over times appropriate to the quasi-steady elements of the speech signal. Essentially what is desired is a *running* spectrum, with real-time as an independent variable, and in which the spectral computation is made on weighted past values of the signal.

Such a result can be obtained by analyzing a portion of the signal "seen" through a specified time window, or weighting function. The window is chosen to insure that the product of signal and window is Fourier transformable. For practical purposes, the weighting function $h(t)$ usually is the impulse response of a physically-realizable linear system. Then, $h(t)=0$; for $t<0$. Generally $h(t)$ is desired to be unipolar and is essentially the response of a low-pass filter. The Fourier transform (5.1) can therefore be modified by transforming that part of the signal seen through the window at a given instant of time. The desired operation is

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

or,

$$F(\omega,t)=e^{-j\omega t}\int_{0}^{\infty} f(t-\lambda)\,h(\lambda)\,e^{j\omega\lambda}\,d\lambda\,. \tag{5.2}$$

The signal, with its past values weighted by $h(t)$, is illustrated for a given instant, t, in Fig. 5.1.

The short-time transform, so defined, is the convolution

$$[f(t)\,e^{-j\omega t}*h(t)],\quad \text{or alternatively,}\quad e^{-j\omega t}[f(t)*h(t)\,e^{j\omega t}].$$

If the weighting function $h(t)$ is considered to have the dimension $\sec^{-1}$ (i.e., the Fourier transform of $h(t)$ is dimensionless), then $|F(\omega,t)|$ is a

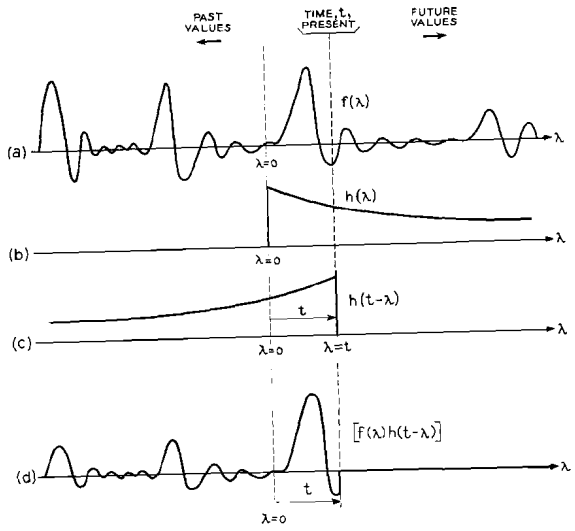


Fig. 5.1. Weighting of an on-going signal $f(t)$ by a physically realizable time window $h(t)$. λ is a dummy integration variable for taking the Fourier transform at any instant, t

short-time amplitude spectrum with the same dimension as the signal. Like the conventional Fourier transform, $F(\omega,t)$ is generally complex with a magnitude and phase, namely $|F(\omega,t)|\,e^{-j\vartheta(\omega,t)}$, where $\vartheta(\omega,t)$ is the short-time phase spectrum.

By definition, the inverse relation also holds

$$[f(\lambda)\,h(t-\lambda)]=\frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega,t)\,e^{j\omega\lambda}\,d\omega\,.$$

Note that at any time $t=t_1$, the product $[f(\lambda)\,h(t_1-\lambda)]$ is determined for all $\lambda\leqq t_1$. If the window function $h(t_1-\lambda)$ is known, then the original function over the interval $-\infty\leqq\lambda\leqq t_1$ can be retrieved from the product. For a value of λ equal to t_1

$$[f(t)\,h(0)]=\frac{1}{2\pi}\int F(\omega,t_1)\,e^{j\omega t_1}\,d\omega\,,$$

or in general for $\lambda=t$

$$f(t)=\frac{1}{2\pi\,h(0)}\int_{-\infty}^{\infty} F(\omega,t)\,e^{j\omega t}\,d\omega\,.$$

The short-time transform is therefore uniquely invertible if one non-zero value of the window function is known. Typically $h(t)$ can be

chosen so that $h(0)=1$ and

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega, t)\, e^{j\omega t}\, d\omega$$

which bears a pleasing parallel to the conventional infinite-time Fourier transform.

The inversion implies that $f(t)$ can be determined for the same points in time that $F(\omega, t)$ is known, provided $F(\omega, t)$ is known as a continuous function of frequency. However, in cases where the product function $[f(\lambda)\, h(t-\lambda)]$ is of finite duration in λ (say owing to a finite duration window) then samples of the waveform $f(t)$ may be recovered exactly from samples in ω of $F(\omega, t)$ (WEINSTEIN). Discrete-frequency, continuous-time values of the short-time transform, $F(\omega_n, t)$, are of particular interest and will find applications in later discussions.

5.12. Measurement of Short-Time Spectra

We notice that (5.2) can be rewritten

$$\begin{aligned} F(\omega, t) &= \int_{-\infty}^{t} f(\lambda)\cos\omega\lambda\, h(t-\lambda)\, d\lambda - j\int_{-\infty}^{t} f(\lambda)\sin\omega\lambda\, h(t-\lambda)\, d\lambda \\ &= [a(\omega, t)-jb(\omega, t)]. \end{aligned} \tag{5.3}$$

Further,

$$\begin{aligned} |F(\omega, t)| &= [F(\omega, t)\, F^*(\omega, t)]^{\frac{1}{2}} \\ &= (a^2+b^2)^{\frac{1}{2}} \end{aligned} \tag{5.4}$$

and

$$\vartheta(\omega, t)=\tan^{-1} b/a,$$

where $F^*(\omega, t)$ is the complex conjugate of $F(\omega, t)$. Note that $|F(\omega, t)|$ is a scalar, whereas $F(\omega, t)\, F^*(\omega, t)$ is formally complex, and that $|F(\omega, t)|^2$ is the short-time power spectrum. The measurement of $|F(\omega, t)|$ can therefore be implemented by the operations shown in Fig. 5.2.

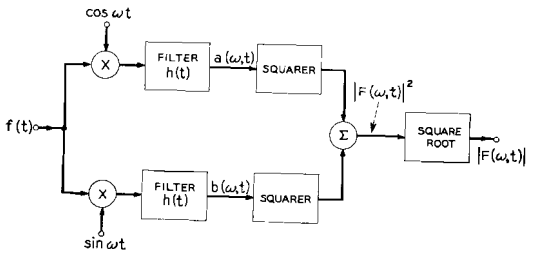


Fig. 5.2. A method for measuring the short-time amplitude spectrum $|F(\omega, t)|$

The frequency-domain interpretation of these operations is apparent. The heterodyning (or multiplication by $\cos\omega t$ and $\sin\omega t$) shifts (or translates) the spectrum of $f(t)$ across the pass-band of filter $h(t)$. The latter is normally a low-pass structure. Frequency components of $f(t)$ lying close to ω produce difference-frequency components inside the low-pass band and yield large outputs from the $h(t)$ filter. Quadrature versions of the shifted signals are squared and added to give the short-time power spectrum $|F(\omega, t)|^2$.

Alternatively, Eq. (5.2) can be written

$$\begin{aligned} F(\omega, t) &= e^{-j\omega t}\left\{\int_0^{\infty} f(t-\lambda)\, h(\lambda)\cos\omega\lambda\, d\lambda + j\int_0^{\infty} f(t-\lambda)\, h(\lambda)\sin\omega\lambda\, d\lambda\right\} \\ &= [a'(\omega, t)+jb'(\omega, t)]\, e^{-j\omega t}. \end{aligned} \tag{5.5}$$

The alternative measurement of $|F(\omega, t)|=[a'^2+b'^2]^{\frac{1}{2}}$ can therefore be effected by the operations in Fig. 5.3.

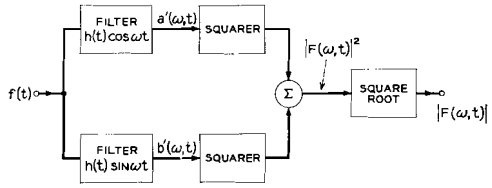


Fig. 5.3. Alternative implementation for measuring the short-time amplitude spectrum $|F(\omega, t)|$

Again, in terms of a frequency-domain interpretation, the measurement involves filtering by phase-complementary band-pass filters centered at ω and having bandwidths twice that of the low-pass $h(t)$ function. The outputs are squared and added to produce the short-time power spectrum $|F(\omega, t)|^2$. Both filters have impulse responses whose envelopes are the time window, $h(t)$. As many pairs of filters are required as the number of frequency values for which the spectrum is desired. Notice, too, that for both methods of measurement (i.e., Figs. 5.2 and 5.3) if the input signal $f(t)$ is a unit impulse the short-time amplitude spectrum is simply $h(t)$, the weighting function.

It is common, in experimental practice, to minimize equipment complexity by making an approximation to the measurements indicated in Figs. 5.2 and 5.3. The desired measurement $|F(\omega, t)|=[a'^2(\omega, t)+b'^2(\omega, t)]^{\frac{1}{2}}$ is essentially the time envelope of either $a'(\omega, t)$ or $b'(\omega, t)$.

The time envelope of a Fourier-transformable function $u(t)$ can be defined as

$$e(t)=[u^2(t)+\hat{u}^2(t)]^{\frac{1}{2}}, \quad \text{where } \hat{u}(t)=\left[u(t)*\frac{1}{\pi t}\right]$$

is the Hilbert transform of $u(t)$. One can show that $\widehat{u(t)\,v(t)}=u(t)\,\hat{v}(t)=\hat{u}(t)\,v(t)$, provided the spectra of $u(t)$ and $v(t)$ do not overlap.

Making use of these relations, and the possibilities for interchanging orders of integration in the convolutions, one notices that

$$\begin{aligned} a'(\omega,t)&=[f(t)*h(t)\cos\omega t] \\ \hat{a}'(\omega,t)&=\left[a'(\omega,t)*\frac{1}{\pi t}\right] \\ &=f(t)*\left[h(t)\cos\omega t*\frac{1}{\pi t}\right] \\ &=f(t)*[h(t)\sin\omega t] \\ &=b'(\omega,t), \end{aligned} \tag{5.6}$$

provided the spectrum of $h(t)$ does not overlap ω. The quantity $|F(\omega,t)|$ is therefore essentially the time envelope of either $a'(\omega,t)$ or $b'(\omega,t)$. The envelope can be approximated electrically by developing the envelope of either filter branch in Fig. 5.3. This is conventionally done by the linear rectification and low-pass filtering indicated in Fig. 5.4. If the impulse response of the low-pass filter is appropriately chosen, the output $|f(t)*p(t)|*q(t)$ approximates $|F(\omega,t)|$.

The measurement method of Fig. 5.4 is precisely the one used in the well-known Sound Spectrograph and in most filter-bank spectrum analyzers. In particular, it is usually the method used to develop the short-time spectrum in vocoders and in several techniques for automatic formant analysis. All of these applications will be discussed in further detail subsequently. As a present example, however, Fig. 5.5 shows successive short-time spectra of a voiced speech sample as produced by a bank of 24 filters. The filters are approximately 150 cps wide, and cover the frequency range 150 to 4000 cps. Each filter is followed by a rectifier and an R-C network. The filter bank is scanned every 10 msec and the short-time spectrum plotted. High-frequency emphasis is used on the input signal to boost its level in the high-frequency end of the

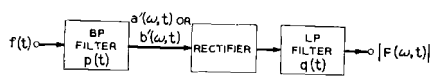


Fig. 5.4. Practical measurement of the short-time spectrum $|F(\omega,t)|$ by means of a band-pass filter, a rectifier and a smoothing network

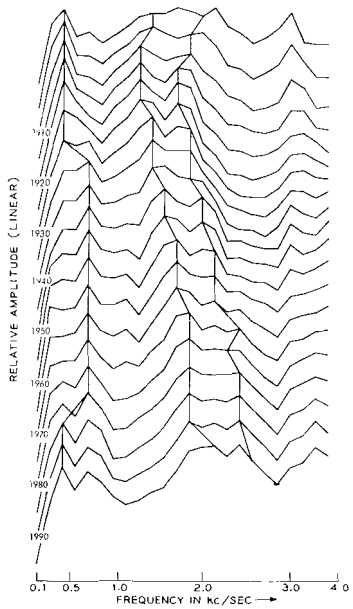


Fig. 5.5. Short-time amplitude spectra of speech measured by a bank of 24 band-pass filters. A single filter channel has the configuration shown in Fig. 5.4. The spectral scans are spaced by 10 msec in time. A digital computer was used to plot the spectra and to automatically mark the formant frequencies. (After FLANAGAN, COKER and BIRD)

spectrum. The filter-bank output is fed into a digital computer through an analog-to-digital converter, and the spectral scans are plotted automatically by the computer (FLANAGAN, COKER, and BIRD). The lines connecting the peaks represent speech formant frequencies which were automatically determined by computer processing of the short-time spectrum.

5.13. Choice of the Weighting Function, $h(t)$

In speech applications, it usually is desirable for the short-time analysis to discriminate vocal properties such as voiced and unvoiced excitation, fundamental frequency, and formant structure. The choice of the analyzing time window $h(t)$ determines the compromise made between temporal and frequency resolution. A time window short in duration corresponds to a broad band-pass filter. It may yield a spectral analysis in which the temporal structure of individual vocal periods is resolved. A window with a duration of several pitch periods, on the other hand, corresponds to a narrower bandpass filter. It may produce

an analysis in which individual harmonic spectral components are resolved in frequency.

In order to illustrate applicable orders of magnitude for filter widths and time windows, imagine the analyzing bandpass filter to be ideal (and nonrealizable) with a rectangular amplitude-frequency response and with zero (or exactly linear) phase response. Let the frequency-domain specification be

$$\begin{aligned} P(\omega)&=1; \quad (\omega_0-\omega_1)\leqq\omega\leqq(\omega_0+\omega_1)\\ &=1; \quad -(\omega_0+\omega_1)\leqq\omega\leqq-(\omega_0-\omega_1)\\ &=0; \quad \text{elsewhere}. \end{aligned} \tag{5.7}$$

The impulse response of the filter is therefore

$$\begin{aligned} p(t)&=\left(\frac{2\omega_1}{\pi}\right)\left(\frac{\sin\omega_1 t}{\omega_1 t}\right)\cos\omega_0 t\\ &=h(t)\cos\omega_0 t, \end{aligned} \tag{5.8}$$

and the time window for this ideal filter is the sin x/x envelope of the impulse response. If the time between initial zeros of the envelope is arbitrarily taken as the effective duration, D, of the time window, then $D=2\pi/\omega_1=4\pi/\Delta\omega$, where $\Delta\omega=2\omega_1$ is the bandwidth of the filter[1]. The D's corresponding to several $\Delta\omega$'s are

Condition	$\Delta\omega/2\pi$ (cps)	D (msec)
(1)	50	40
(2)	100	20
(3)	250	8

Condition (1) is an analyzing bandwidth commonly used to resolve the harmonic spectral components in voiced portions of speech. For this bandwidth, the duration of the time window spans about four or five pitch periods of a man's voice.

The broad filter condition (3), on the other hand, produces a weighting function comparable in duration with a single pitch period of a man's voice. The time resolution of this analysis therefore resolves amplitude fluctuations whose temporal courses are of the order of a pitch period. Filter conditions analogous to both (1) and (3) are employed in the well-known Sound Spectrograph which will be discussed in the following section.

[1] Sometimes one-half this value is taken as the effective window duration.

The middle condition (2) is a sort of time-frequency compromise for speech. It is a filter width which has been found useful in devices such as vocoders and formant trackers. The short-time spectra already shown in Fig. 5.5 are representative of this resolution.

In passing, it is relevant to estimate the effective time window for the mechanical short-time analysis made by the basilar membrane in the human ear. From the earlier discussion in Chapter IV[1], a reasonably good approximation to the displacement impulse response of the basilar membrane, at a point maximally responsive to radian frequency β, is

$$\begin{aligned} p(t)&=(\beta t)^2 e^{-\beta t/2}\sin\beta t\\ &=h_{bm}(t)\sin\beta t. \end{aligned} \tag{5.9}$$

The time window for the basilar membrane, according to this modeling[2], is the "surge" function plotted in Fig. 5.6. One notices that the

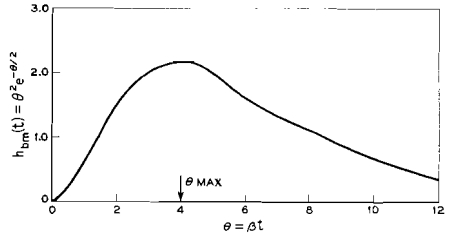


Fig. 5.6. The effective time window for short-time frequency analysis by the basilar membrane in the human ear. The weighting function is deduced from the ear model discussed in Chapter IV

time window has a duration inversely related to β. It has its maximum at $t_{max}=4/\beta$. If, as a crude estimate, $2t_{max}$ is taken as the effective duration D of the window, then for several membrane places:

$\beta/2\pi$ (cps)	$D=2t_{max}$ (msec)
100	12.0
1000	1.2
5000	0.2

For most speech signals, therefore, the mechanical analysis of the ear apparently provides better temporal resolution than spectral resolu-

[1] See also the "third" model described in FLANAGAN (1962a).

[2] Eq. (5.9) does not include the effects of the middle ear. See Chapter IV for these details.

tion. Generally, the only harmonic component resolved mechanically is the fundamental frequency of voiced segments. This result is borne out by observations on the models described in Chapter IV.

5.14. The Sound Spectrograph

Spectral analysis of speech came of age, so to speak, with the development of the Sound Spectrograph (KOENIG, DUNN and LACEY). This device provides a convenient means for permanently displaying the short-time spectrum of a sizeable duration of signal. Its method of analysis is precisely that shown in Fig. 5.4. Its choice of time windows (see preceding section) is made to highlight important acoustic and perceptual features such as formant structure, voicing, friction, stress and pitch. Many other devices for spectrum analysis have also been developed, but the relative convenience and ease of operation of the sound spectrograph has stimulated its wide acceptance in speech analysis and phonetic science. Because it is such a widely used tool, this section will give a brief description of the device and its principles of operation.

Fig. 5.7 shows a functional diagram of one type of sound spectrograph (commonly known as the Model *D* Sonagraph). With the microphone switch (SW1) in the *record* position, a speech sample (generally about 2.5 sec in duration) is recorded on a magnetic disc. The microphone switch is turned to *analyze*, and a spectral analysis of the sample is made by playing it repeatedly through a bandpass filter. Upon successive playings the bandpass filter is, in effect, scanned slowly across the frequency band of the signal. The result is therefore equivalent to an analysis by many such filters. For practical reasons it is more convenient to use a fixed bandpass filter and to "slide" the spectrum of the signal past the filter. This is accomplished by modulating the signal onto a high frequency carrier and sliding one sideband of the signal past the fixed bandpass filter. The translation is accomplished by varying the

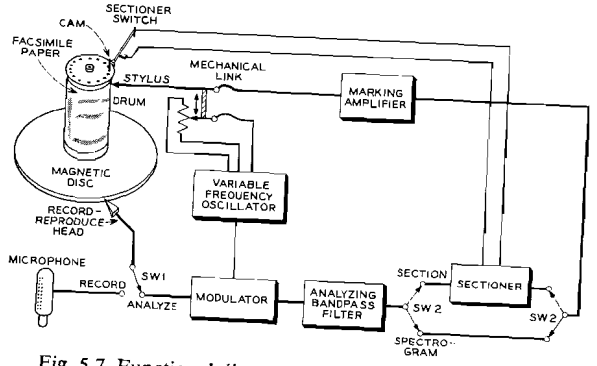


Fig. 5.7. Functional diagram of the sound spectrograph

frequency of the carrier. The carrier frequency control is mechanically geared to the magnetic disc so the signal spectrum is progressively analyzed upon repeated rotations of the disc.

With SW2 in the *spectrogram* position, the output current of the bandpass filter is amplified and passed to a stylus whose vertical motion is geared to the magnetic disc and the carrier control (or to the effective frequency position of the bandpass filter). The stylus is in contact with an electrically sensitive facsimile paper which is fixed to a drum mounted on the same shaft as the magnetic disc. Electrical current from the stylus burns the paper in proportion to the current magnitude. The paper therefore acts as the full-wave rectifier of Fig. 5.4, and the finite size and spreading of the burned trace perform the low-pass filtering. The density of the burned mark is roughly proportional to the logarithm of the current magnitude. Because of the mechanical linkage, the stylus and carrier move slowly across the frequency range of the signal as the magnetic disc rotates, and a time-intensity-frequency plot of the signal is "painted" on the paper.

Two widths of the bandpass filter are conventionally used with the instrument, 300 cps and 45 cps. The time-frequency resolution of the analysis is essentially determined by these widths. As discussed in the preceding section, the wide pass-band provides better temporal resolution of speech events, while the narrow band yields a frequency resolution adequate to resolve harmonic lines in voiced utterances. A typical spectrogram made with the 300 cps wide analyzing filter is shown in the upper diagram of Fig. 5.8. As previously indicated, the abscissa is time, the ordinate is frequency, and darkness of the pattern represents in-

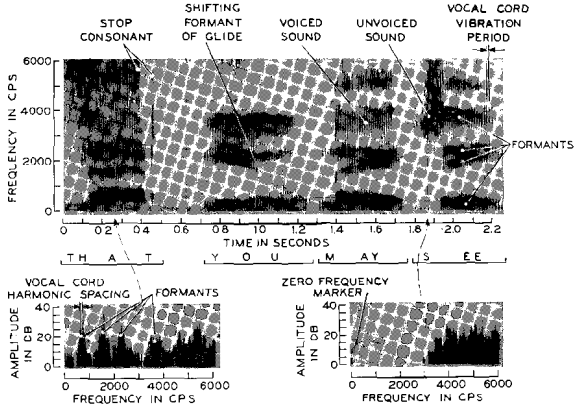


Fig. 5.8a and b. (a) Broadband sound spectrogram of the utterance "That you may see". (b) Amplitude *vs* frequency plots (amplitude sections) taken in the vowel portion of "that" and in the fricative portion of "see". (After BARNEY and DUNN)

tensity. Several speech features are indicated. Note that the time resolution is such that vertical striations in the voiced portions show the fundamental period of the vocal cords.

The facsimile paper is capable of depicting an intensity range (from lightest gray to darkest black) of only about 12 db (PRESTIGIACOMO, 1957). It often is desirable to examine amplitude spectra over a greater intensity range. A means is therefore provided for making a frequency-versus-amplitude portrayal at any given instant along the time scale. For this operation, SW2 in Fig. 5.7 is put to the *section* position. A cam is placed on the drum periphery at the time of occurrence of the sound whose amplitude section is desired. The functions of the carrier and stylus are as previously described.

The sectioner contains a full-wave rectifier, an R-C integrator and a biased multivibrator. In one version of the apparatus, as the magnetic disc and drum rotate, the cam closes the section switch at the desired instant in the utterance. The value of the short-time spectrum at this instant is effectively "read" and stored on a capacitor in the input circuit of a biased multivibrator. The multivibrator is held on (i.e., free runs) until the capacitor charge decays to a threshold value. The multivibrator then turns off. During its on-time, it delivers a marking current to the stylus and (because of the exponential decay) the length of the marked trace is proportional to the logarithm of the smoothed output of the analyzing filter. Because the stylus is scanning the frequency scale with the filter, an amplitude (*db*)-versus-frequency plot is painted for the prescribed instant.

Amplitude sections are usually made with the 45 cps (narrow band) filter. Typical sections taken in a vowel and in a fricative are shown in the lower half of Fig. 5.8.

Because the speech sample must be played repeatedly as the analyzing filter scans its band, the time to produce the complete spectrogram is appreciable. Common practice is to shorten the analyzing time by playing back at several times the recording speed. A typical value, for example, is a speed-up of three-to-one. A recorded bandwidth of 100 to 4000 cps is therefore multiplied to 300 to 12000 cps. If the analyzing bandpass filter is centered at, say, 15000 cps, then the carrier oscillator may scan from 15000 to 27000 cps. Depending upon frequency range and technique, one to several minutes may be required to analyze a 2.5 sec speech sample. In the course of the analysis the sample may be played back several hundred times. A common figure for the filter advance is of the order of 20 cps/playback.

The manner in which broadband spectrograms highlight vocal modes, or formants, for various articulatory configurations is illustrated in Fig. 5.9. Articulatory diagrams for four vowels, /i, æ, a, u/ and their cor-

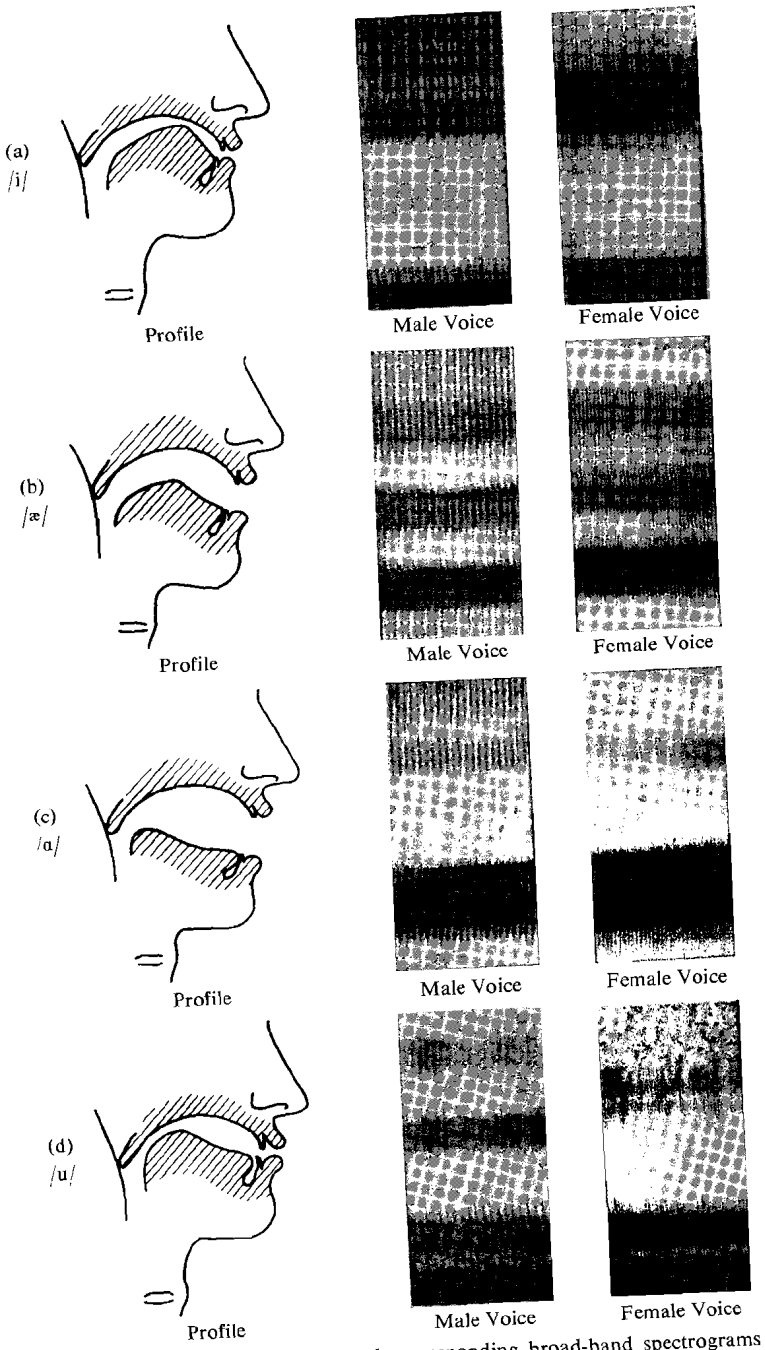


Fig. 5.9 a–d. Articulatory diagrams and corresponding broad-band spectrograms for the vowels /i, æ, ɑ, u/ as uttered by adult male and female speakers. (After POTTER, KOPP and GREEN)

responding broadband (300 cps) spectrograms are shown. The dark bands indicate the spectral energy concentrations and reflect the vocal modes for a given configuration. (These spectrograms can be compared with the calculated mode patterns for similar vowels in Figs. 3.28 and 3.29 of Chapter III.)

Typical of the research uses to which this type of spectrographic display has been put is a large-scale study of vowel formant frequencies, amplitudes, and pitches for a number of different speakers (PETERSON and BARNEY). The results of this study for 33 men give the mean formant frequencies for the English vowels as plotted in Fig. 5.10. The vowels were uttered in an /h−d/ environment.

Numerous "relatives" of the sound spectrograph – both predecessors and successors – have been designed and used, each usually with a specific purpose in mind. These devices range from scanned filter banks to correlation instruments. In a short space it is not possible to mention many of them. One variation in the spectrographic technique is the so-called "resonagraph" (HUGGINS, 1952). This device is designed to delineate formant frequencies and to suppress nonformant energy. Another modification displays the time derivative of the spectral amplitude, rather than simply the amplitude (MEYER-EPPLER, 1951; KOCK and MILLER). The effect is to emphasize dynamic time changes in the spectrum and to suppress quasi-steady portions. Features such as stop consonants or formant transitions are therefore more sharply delineated.

An even closer relative is the so-called visible speech translator (DUDLEY and GRUENZ; RIESZ and SCHOTT) in which the conventional sound spectrogram is painted electronically in real time, either on a moving belt coated with luminescent phosphor, or on a rotating cathode

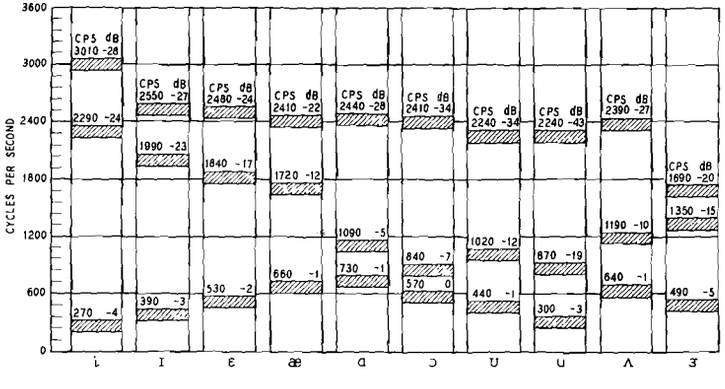


Fig. 5.10. Mean formant frequencies and relative amplitudes for 33 men uttering the English vowels in an /h–d/ environment. Relative formant amplitudes are given in db *re* the first formant of /ɔ/. (After PETERSON and BARNEY as plotted by Haskins Laboratories)

ray tube. A still different variation is the correlatograph (BENNETT, 1953; BIDDULPH) which plots the magnitude of the short-time autocorrelation function of the signal in trace density, the delay parameter on the ordinate, and time along the abscissa.

Several schemes for quantizing the intensity dimension of the conventional spectrogram have also been described (KERSTA, 1948; PRESTIGIACOMO, 1957). The result is to yield a "topological map" of the signal in which intensity gradients are indicated by the closeness of the contour lines.

5.15. Short-Time Correlation Functions and Power Spectra

If $x(t)$ is an on-going stationary random signal, its autocorrelation function $\varphi(\tau)$ and its power density spectrum $\Phi(\omega)$ are linked by Fourier transforms (WIENER; LEE).

$$\varphi(\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}x(t)\,x(t+\tau)\,dt$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(\omega)\,e^{j\omega\tau}\,d\omega$$

and

$$\Phi(\omega)=\int_{-\infty}^{\infty}\varphi(\tau)\,e^{-j\omega\tau}\,d\tau\,. \tag{5.10}$$

[Note that $\varphi(0)$ is the mean square value, or average power, of the signal.]

For an aperiodic Fourier-transformable signal, $y(t)$, parallel relations link the autocorrelation function $\psi(\tau)$ and the energy density spectrum $\Psi(\omega)$.

$$\psi(\tau)=\int_{-\infty}^{\infty}y(t)\,y(t+\tau)\,dt$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Psi(\omega)\,e^{j\omega\tau}\,d\omega \tag{5.11}$$

$$\Psi(\omega)=\int_{-\infty}^{\infty}\psi(\tau)\,e^{-j\omega\tau}\,d\tau\,,$$

where

$$\Psi(\omega)=Y(\omega)\,Y^{*}(\omega),\quad\text{and}\quad Y(\omega)=\int_{-\infty}^{\infty}y(t)\,e^{-j\omega t}\,dt\,.$$

[Note that

$$\psi(0)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Psi(\omega)\,d\omega$$

is the total energy of the signal.]

In both cases the correlation functions are real and even functions of the delay parameter τ, and the spectra are real and even functions of the frequency ω. All of the transforms can therefore be written as cosine transforms. These transform-pairs suggest the possibility of determining short-time spectral information by means of correlation techniques, provided the latter can be extended to the short-time case.

In the preceding discussion on short-time spectral analysis, the approach was to analyze a Fourier-transformable "piece" of the signal obtained by suitably weighting the past values. The correlation relations for aperiodic functions can be similarly extended to this description of the speech signal. According to the earlier derivations, at any instant t the following transforms are presumed to hold for the speech signal $f(t)$,

$$F(\omega, t)=\int_{-\infty}^{t} f(\lambda)\, h(t-\lambda)\, e^{-j\omega\lambda}\, d\lambda$$
$$[f(\lambda)\, h(t-\lambda)]=\frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega, t)\, e^{j\omega\lambda}\, d\omega, \tag{5.12}$$

where $h(t)$ is the weighting function. Then, formally,

$$\psi(\tau, t)=\int_{-\infty}^{t} f(\lambda)\, h(t-\lambda)\, f(\lambda+\tau)\, h(t-\lambda-\tau)\, d\lambda$$
$$\psi(\tau, t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi(\omega, t)\, e^{j\omega\tau}\, d\omega \tag{5.13}$$
$$\Psi(\omega, t)=[F(\omega, t)\, F^{*}(\omega, t)]=\int_{-\infty}^{\infty} \psi(\tau, t)\, e^{-j\omega\tau}\, d\tau.$$

Practically, for real time measurement at time t, $f(t+\tau)$ for $\tau>0$ is not known. [For a fixed over-all delay (comparable to the window duration) τ may be considered to be a differential delay.] However, $\psi(\tau, t)$ is formally an even function of τ. It can therefore be defined in terms of negative τ so that

$$\Psi(\omega, t)=\int_{-\infty}^{\infty} \psi(\tau, t)\, e^{-j\omega\tau} d\tau=2\int_{-\infty}^{0} \psi(\tau, t)\cos\omega\tau\, d\tau, \tag{5.14}$$

where $\Psi(\omega, t)$ is also an even function of ω.

Thus a short-time autocorrelation measure, related to the short-time power spectrum $|F(\omega, t)|^2$ by the aperiodic transform, can be made. Techniques for the measurement of $|F(\omega, t)|^2$ have already been described in Section 5.12. Measurement of $\psi(\tau, t)$ for negative τ can be effected by the arrangement shown in Fig. 5.11. The individual output taps from the delay lines are weighted according to $h(t)$. Corresponding points (in the running variable λ) are multiplied, and the integration is

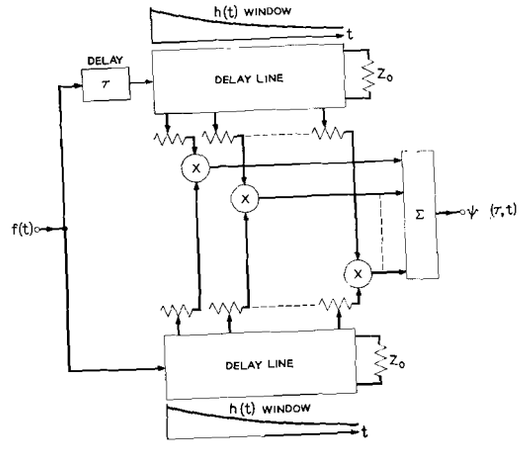


Fig. 5.11. Method for the measurement of the short-time correlation function $\psi(\tau, t)$

approximated as a finite sum[1]. $\psi(\tau, t)$ is therefore a running correlation which is related to $|F(\omega, t)|^2$ or $\Psi(\omega, t)$ by a Fourier transform.

It is also possible to define a short-time correlation function produced by weighting the product of the original signal and the signal delayed (FANO). The defining relation is

$$\varphi(\tau, t)=\int_{-\infty}^{t} f(\lambda)\, f(\lambda+\tau)\, k(t-\lambda)\, d\lambda, \tag{5.15}$$

where $k(t)=0$, $t<0$ is the weighting function. The measure is easily implemented for $\tau\leqq 0$ by the circuit shown in Fig. 5.12. This technique has been used experimentally to measure correlation functions for speech sounds (STEVENS, 1950; KRAFT; BIDDULPH).

In general, no simple transform relation exists between $\varphi(\tau, t)$ and a measurable short-time power spectrum. Under the special condition

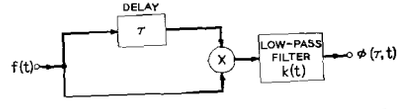


Fig. 5.12. Circuit for measuring the running short-time correlation function $\varphi(\tau, t)$

[1] The operations of Fig. 5.11 compute

$$\psi(\tau, t)=\int_{0}^{\infty} f(t-\lambda)\, h(\lambda)\, f(t-\lambda-\tau)\, h(\lambda+\tau)\, d\lambda,$$

for negative τ, instead of the form given in Eq. (5.13).

$k(t)=2\alpha e^{-2\alpha t}=[h(t)]^2$, however, $\varphi(\tau, t)$ can be related to $\Psi(\omega, t)=|F(\omega, t)|^2$.

$$\begin{aligned}\psi(\tau, t)&=\int_{-\infty}^{t} f(\lambda)\,h(t-\lambda)\,f(\lambda+\tau)\,h(t-\lambda-\tau)\,d\lambda\\&=e^{\alpha\tau}\int_{-\infty}^{t} 2\alpha\, f(\lambda)\,f(\lambda+\tau)\,e^{-2\alpha(t-\lambda)}\,d\lambda\\&=e^{\alpha\tau}\varphi(\tau, t);\quad \tau\leqq 0.\end{aligned}\tag{5.16}$$

But as previously argued, $\psi(\tau, t)$ is an even function of τ, and if $\varphi(\tau, t)$ is defined as an even function, then $\psi(\tau, t)=e^{-\alpha|\tau|}\varphi(\tau, t)$ for all τ, or

$$\begin{aligned}\varphi(\tau, t)&=e^{\alpha|\tau|}\psi(\tau, t)\\&=\frac{e^{\alpha|\tau|}}{2\pi}\int_{-\infty}^{\infty}\Psi(\omega, t)\,e^{j\omega\tau}\,d\omega,\end{aligned}$$

and

$$\begin{aligned}\Psi(\omega, t)&=\int_{-\infty}^{\infty} e^{-\alpha|\tau|}\varphi(\tau, t)\,e^{-j\omega\tau}\,d\tau\\&=\int_{-\infty}^{\infty} e^{-\alpha|\tau|}\varphi(\tau, t)\cos\omega\tau\,d\tau.\end{aligned}\tag{5.17a}$$

It also follows that

$$\begin{aligned}\Psi(\omega, t)&=\frac{1}{2\pi}\left[\mathcal{F}\{e^{-\alpha|\tau|}\}*\mathcal{F}\{\varphi(\tau, t)\}\right]\\&=\frac{1}{2\pi}\left[\left(\frac{2\alpha}{\alpha^2+\omega^2}\right)*\Phi(\omega, t)\right]\\&=\frac{1}{2\pi}\left[|H(\omega)|^2*\Phi(\omega, t)\right],\end{aligned}\tag{5.17b}$$

where $\mathcal{F}$ denotes the Fourier transform.

Thus the short-time power spectrum $\Psi(\omega, t)$ is the real convolution of the power spectrum $\Phi(\omega, t)$ with the low-pass energy spectrum $(2\alpha/\alpha^2+\omega^2)$. $\Psi(\omega, t)$ therefore has poorer spectral resolution than the Fourier transform of $\varphi(\tau, t)$ [i.e., $\Phi(\omega, t)$]. Note also that for $h(t)=(2\alpha)^{\frac{1}{2}}e^{-\alpha t}$, $|F(\omega, t)|$ is essentially measured by single-resonant circuits with impulse responses $[(2\alpha)^{\frac{1}{2}}e^{-\alpha t}\cos\omega t]$ and $[(2\alpha)^{\frac{1}{2}}e^{-\alpha t}\sin\omega t]$. (See Fig. 5.3.)

Weighting functions different from the exponential just discussed do not lead to simple transform relations between $\varphi(\tau, t)$ and a power spectrum. Other definitions, however, can be made of measurable correlations and short-time power spectra, and these can be linked by specially defined transforms (SCHROEDER and ATAL). For example, one

can define a short-time spectrum

$$\Omega(\omega, t)=\int_{-\infty}^{\infty}\varphi(\tau, t)\,m(|\tau|)\cos\omega\tau\,d\tau,\tag{5.18a}$$

in which $\varphi(\tau, t)$, as given in Eq. (5.15), is defined as an even function of τ (but is measured for delays only) so that,

$$\varphi(\tau, t)=\int_{-\infty}^{t} f(\lambda)\,f(\lambda-|\tau|)\,n(t-\lambda)\,d\lambda,\tag{5.18b}$$

where $m(t)$ and $n(t)$ are physically realizable weighting functions and are zero for $t<0$[1]. $\Omega(\omega, t)$ and $\varphi(\tau, t)$ are then linked by the definitions (5.18). $\varphi(\tau, t)$ can be measured according to Fig. 5.12, and a straightforward measure of $\Omega(\omega, t)$ can also be made. Substituting for $\varphi(\tau, t)$ in the definition of $\Omega(\omega, t)$ gives

$$\begin{aligned}\Omega(\omega, t)&=2\int_{-\infty}^{t} f(\lambda)\,n(t-\lambda)\,d\lambda\int_{0}^{\infty} f(\lambda-\tau)\,m(\tau)\cos\omega\tau\,d\tau\\&=2\{n(t)*f(t)[f(t)*m(t)\cos\omega t]\}.\end{aligned}\tag{5.19}$$

The operations indicated in (5.19) are a filtering of the signal $f(t)$ by a (normally bandpass) filter whose impulse response is $[m(t)\cos\omega t]$; a multiplication of this output by the original signal; and a (normally low pass) filtering by a filter whose impulse response is $n(t)$. The measurement is schematized in Fig. 5.13.

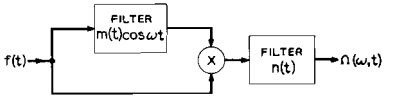


Fig. 5.13. Arrangement for measuring the short-time spectrum $\Omega(\omega, t)$. (After SCHROEDER and ATAL)

For the case $m(t)=n(t)=e^{-\alpha t}$, $\Omega(\omega, t)$ reduces to $\Psi(\omega, t)$. From the definition of $\Omega(\omega, t)$, the inverse relation follows

$$\varphi(\tau, t)=\frac{1}{2\pi\, m(|\tau|)}\int_{-\infty}^{\infty}\Omega(\omega, t)\cos\omega\tau\,d\omega.\tag{5.20}$$

The defining relations of Eq. (5.18) also imply that

$$\Omega(\omega, t)=M(\omega)*\Phi(\omega, t),\tag{5.21}$$

[1] If $\Omega(\omega, t)$ is to be a positive quantity, some further restrictions must be placed on $n(t)$.

where

$$M(\omega)=\int_{-\infty}^{\infty} m(|\tau|)\,e^{-j\omega\tau}\,d\tau,$$

and

$$\Phi(\omega,t)=\int_{-\infty}^{\infty}\varphi(\tau,t)\,e^{-j\omega\tau}\,d\tau.$$

This result can be compared with Eq. (5.17), where

$$|H(\omega)|^2=\int_{-\infty}^{\infty} e^{-\alpha|\tau|}e^{-j\omega\tau}\,d\tau$$

$$H(\omega)=\int_{0}^{\infty}(2\alpha)^{\frac{1}{2}}e^{-\alpha\tau}e^{-j\omega\tau}\,d\tau=\int_{0}^{\infty}h(\tau)\,e^{-j\omega\tau}\,d\tau.$$

Since $\Omega(\omega,t)$ is obtained from $\Phi(\omega,t)$ by convolution with the (low pass) spectrum $M(\omega)$, it has poorer spectral definition than $\Phi(\omega,t)$.

5.16. Average Power Spectra

The spectral measuring schemes of the previous discussion use windows which are relatively short in duration to weight past values of the signal. They yield spectra in which brief temporal fluctuations are preserved. A long-term mean value of the spectrum, say $\overline{|F(\omega,t)|^2}$, might also be of interest if average spectral distribution is of more importance than short-time variations. Such an average can be written as

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}F(\omega,t)\,F^*(\omega,t)\,dt=\overline{|F(\omega,t)|^2}=\overline{\Psi(\omega,t)}$$

$$=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}dt\int_{-\infty}^{t}f(\lambda)\,h(t-\lambda)\,e^{-j\omega\lambda}\,d\lambda\int_{-\infty}^{t}f(\eta)\,h(t-\eta)\,e^{j\omega\eta}\,d\eta. \tag{5.22}$$

Changing variables and rearranging

$$\overline{|F(\omega,t)|^2}$$

$$=\int_{0}^{\infty}d\lambda\,h(\lambda)\,e^{j\omega\lambda}\int_{0}^{\infty}d\eta\,h(\eta)\,e^{-j\omega\eta}\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f(t-\lambda)\,f(t-\eta)\,dt. \tag{5.23}$$

According to Eqs. (5.10), the latter integral is simply $\varphi(\lambda-\eta)$, which is the Fourier transform of $\Phi(\omega)$. That is,

$$\varphi(\lambda-\eta)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(\delta)\,e^{j\delta(\lambda-\eta)}\,d\delta$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(\delta)\,e^{-j\delta(\lambda-\eta)}\,d\delta,$$

because $\Phi(\omega)$ is real and even. Then

$$\overline{|F(\omega,t)|^2}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(\delta)\,d\delta\int_{0}^{\infty}h(\lambda)\,e^{j\lambda(\omega-\delta)}\,d\lambda\int_{0}^{\infty}h(\eta)\,e^{-j\eta(\omega-\delta)}\,d\eta$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(\delta)\,H(\omega-\delta)\,H^*(\omega-\delta)\,d\delta \tag{5.24}$$

$$\overline{|F(\omega,t)|^2}=\frac{1}{2\pi}\left[\Phi(\omega)*|H(\omega)|^2\right].$$

Therefore, the long-time average value of the power spectrum $|F(\omega,t)|^2$ is the real convolution of the power density spectrum $\Phi(\omega)$ and the energy density spectrum of the time window $h(t)$. The narrower the $|H(\omega)|^2$ spectrum, the more nearly $\overline{|F(\omega,t)|^2}$ represents the power density spectrum $\Phi(\omega)$. A narrow $H(\omega)$ corresponds to a long time window and to narrow bandpass filters in the circuits of Figs. 5.3 and 5.4. In the limit $H(\omega)$ is an impulse at $\omega=0$, the time window is a unit step function and $\overline{|F(\omega,t)|^2}$ has the same spectral characteristics as $\Phi(\omega)$. For any value of ω, $\overline{|F(\omega,t)|^2}$ is the integral of the power density spectrum "seen" through the aperture $|H(\omega)|^2$ positioned at ω. It is therefore the average power of the signal in the pass band of the filter in Fig. 5.4.

It was previously demonstrated [Eq. (5.17)] that for the special condition $h(t)=[(2\alpha)^{\frac{1}{2}}e^{-\alpha t}]$,

$$\Psi(\omega,t)=\frac{1}{2\pi}\left[|H(\omega)|^2*\Phi(\omega,t)\right].$$

Notice that for this situation, the long-time average is

$$\overline{\Psi(\omega,t)}=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\int_{-\infty}^{\infty}e^{-\alpha|\tau|}\,\varphi(\tau,t)\cos\omega\tau\,d\tau\,dt$$

$$=\int_{-\infty}^{\infty}e^{-\alpha|\tau|}\,\overline{\varphi(\tau,t)}\cos\omega\tau\,d\tau. \tag{5.25}$$

Substituting for $\varphi(\tau,t)$ from (5.15) and interchanging variables leads to

$$\overline{\Psi(\omega,t)}=\int_{0}^{\infty}e^{-\alpha|\tau|}\,\varphi(\tau)\cos\omega\tau\,d\tau\int_{0}^{\infty}k(\beta)\,d\beta. \tag{5.26}$$

Since

$$\int_{0}^{\infty}k(t)\,dt=\int_{0}^{\infty}h^2(t)\,dt=1,$$

then

$$\overline{\Psi(\omega,t)}=\frac{1}{2\pi}\left[|H(\omega)|^2*\Phi(\omega)\right],$$

which corresponds to the result (5.24).

5.17. Measurement of Average Power Spectra for Speech

A number of experimental measurements of the average power spectrum of speech have been made (for example, SIVIAN; DUNN and WHITE). The technique frequently used is essentially the bandpass filter arrangement shown previously in Fig. 5.4, with the exception that a square-law rectifier and a long-time integrator (averager) are used. This arrangement is shown is Fig. 5.14. If the switch closes at time $t=0$ and

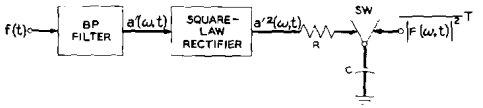


Fig. 5.14. Circuit for measuring the long-time average power spectrum of a signal

remains closed for T sec, the accumulated capacitor voltage is an approximation to $\overline{|F(\omega, t)|^2}$ and is,

$$V_c(T)=\int_0^T a'^2(\omega,\lambda)\,\frac{1}{RC}\,e^{-\frac{1}{RC}(T-\lambda)}\,d\lambda. \tag{5.27}$$

If $RC \gg T$, then the exponential is essentially unity for $0 \leqq \lambda \leqq T$, and

$$V_c(T)\cong\frac{1}{RC}\int_0^T a'^2(\omega,\lambda)\,d\lambda \sim \overline{|F(\omega,t)|^2}. \tag{5.28}$$

The measurement described by (5.28) has been used in one investigation of speech spectra. Bandpass filters with bandwidths one-half octave wide below 500 cps and one octave wide above 500 cps were used. The integration time was $\frac{1}{8}$ sec (DUNN and WHITE). Distributions of the absolute root-mean-square speech pressure in these bands – measured 30 cm from the mouth of a talker producing continuous conversational speech – are shown in Fig. 5.15. The data are averages for six men. The distribution for the unfiltered speech is shown by the marks on the left ordinate.

If the integration time is made very long, say for more than a minute of continuous speech (all natural pauses between syllables and sentences being included), or if many short-time measurements are averaged, one obtains a long-time power spectrum in which syllabic length variations are completely smoothed out. Assuming that the speech power is uniformly distributed in the octave and half-octave filter bands the measured longtime power density spectrum, $\Phi(\omega)$, for speech is shown in

Fig. 5.16. The ordinate here is given in terms of mean-square sound pressure per cycle. In both Figs. 5.15 and 5.16, the detailed formant structure of individual sound is averaged out.

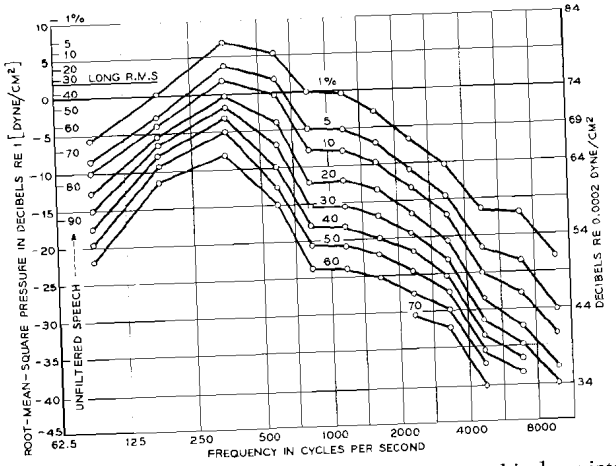


Fig. 5.15. Root mean square sound pressures for speech measured in $\frac{1}{8}$ sec intervals 30 cm from the mouth. The analyzing filter bands are one-half octave wide below 500 cps and one octave wide above 500 cps. (After DUNN and WHITE.) The parameter is the percentage of the intervals having levels greater than the ordinate

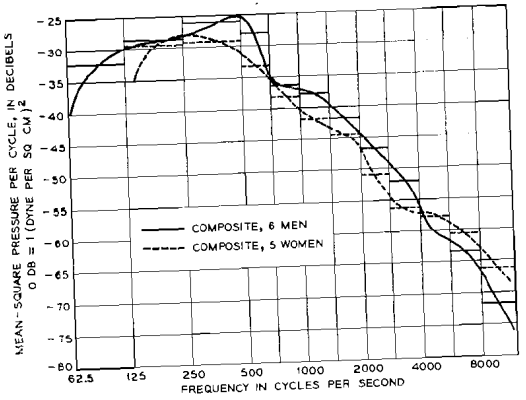


Fig. 5.16. Long-time power density spectrum for continuous speech measured 30 cm from the mouth. (After DUNN and WHITE)

5.2. Formant Analysis of Speech

Formant analysis of speech can be considered a special case of spectral analysis. The objective is to determine the complex natural frequencies of the vocal mechanism as they change temporally. The changes

are, of course, conditioned by the articulatory deformations of the vocal tract. One approach to such analysis is to consider how the modes are exhibited in the short-time spectrum of the signal. As an initial illustration, the temporal courses of the first three speech formants are traced in an idealized form on the spectrogram of Fig. 5.17. Often, for bandwidth compression application, an automatic, real-time determination of these data is desired.

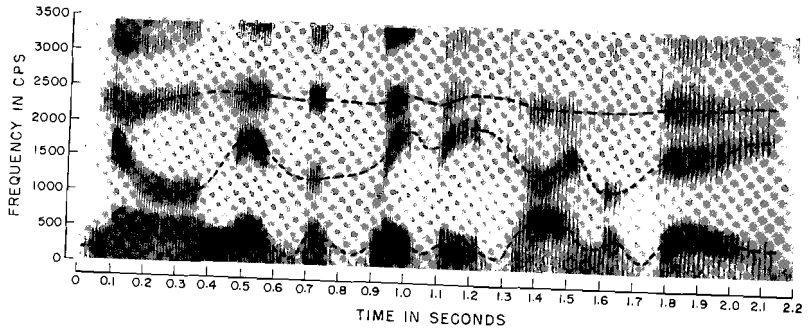


Fig. 5.17. Sound spectrogram showing idealized tracks for the first three speech formants

As certain of the results in Chapter III imply, the damping or dissipation characteristics of the vocal system are relatively constant and predictable, especially over the frequency range of a given formant. Generally, therefore, more interest attaches to the temporal variations of the imaginary parts of the complex formant frequencies than to the real parts. Nevertheless, an adequate knowledge of the real parts, or of the formant bandwidths, is important both perceptually and in spectral analysis procedures.

The "system function" approach to speech analysis, as discussed in Chapter III, aims at a specification of the signal in terms of a transmission function and an excitation function. If the vocal configuration is known, the mode pattern can be computed, and the output response to a given excitation can be obtained. In automatic analysis for encoding and transmission purposes, the reverse situation generally exists. One has available only the acoustic signal and desires to analyze it in terms of the properties of the source and the modes of the system. One main difficulty is in not knowing how to separate uniquely the source and the system.

The normal modes of the vocal system move continuously with time, but they may not, for example, always be clearly manifest in a short-time spectrum of the signal. A particular pole may be momentarily

obscured or suppressed by a source zero or by a system zero arising from a side-branch element (such as the nasal cavity). The short-time spectrum generally exhibits the prominent modes, but it is often difficult to say with assurance where the low-amplitude poles or significant pole-zero pairs might lie.

Further complicating the situation is the fact that the output speech signal is generally not a minimum-phase function (that is, it may not have all its zeros in the left half of the complex frequency plane). If it were, its phase spectrum would be implied by its amplitude spectrum. The vocal-tract transmission is, of course, minimum phase for all conditions where radiation takes place from only one point, i.e., mouth or nostril. For simultaneous radiation from these points it is not. It can be shown that the glottal source, provided the volume velocity wave is zero at some time during its period, possesses only finite-frequency zeros and no poles (MATHEWS, MILLER and DAVID, 1961 b). Further, it can be shown that the zeros can lie in either the right or left half planes, or in both (DUNN, FLANAGAN and GESTRIN). These factors conspire to make accurate automatic formant analysis a difficult problem. The present section outlines a number techniques for the automatic measurement of formant frequency and formant bandwidth, and indicates the performance they achieve.

5.21. Formant-Frequency Extraction

In its simplest visualization, the voiced excitation of a vocal resonance is analogous to the excitation of a single-tuned circuit by brief, periodic pulses. The output is a damped sinusoid repeated at the pulse rate. The envelope of the amplitude spectrum has a maximum at a frequency equal essentially to the imaginary part of the complex pole frequency. The formant frequency might be measured either by measuring the axis-crossing rate of the time waveform, or by measuring the frequency of the peak in the spectral envelope. If the bandwidth of the resonance is relatively small, the first moment of the amplitude spectrum,

$$\bar{f} = \frac{\int f A(f)\, df}{\int A(f)\, df}$$

might also be a reasonable estimate of the imaginary part of the pole frequency.

The resonances of the vocal tract are, of course, multiple. The output time waveform is therefore a superposition of damped sinusoids and the amplitude spectrum generally exhibits multiple peaks. If the individual resonances can be suitably isolated, say by appropriate filter-

ing, the axis-crossing measures, the spectral maxima and the moments might all be useful indications of formant frequency. If, on the other hand, the more subtle properties of the source and the system are to be accounted for—say the spectral zeros produced by the glottal source or by a sidebranch resonator—a more sophisticated measure of the normal modes generally is necessary. One such approach is the detailed fitting of an hypothesized spectral model to the real speech spectrum. For analyses of this type, it is often advantageous to employ the storage and rapid logical operations of a digital computer.

5.211. Axis-Crossing Measures of Formant Frequency. One of the earliest attempts at automatic tracking of formant frequencies was an average zero-crossing count (E. PETERSON). The idea was to take the average density of zero-crossings of the speech wave and of its time derivative as approximations to the first and second formants, respectively. The reasoning was that in the unfiltered, voiced speech the first formant is the most prominent spectral component. It consequently is expected to have the strongest influence upon the axis-crossing rate. In the differentiated signal, on the other hand, the first formant is de-emphasized and the second formant is dominant. The results of these measures, however, were found to be poor, and the conclusion was that the method did not give acceptable precision.

A number of refinements of the zero-crossing technique have been made. In one (MUNSON and MONTGOMERY; DAVIS, BIDDULPH, and BALASHEK), the speech signal is pre-filtered into frequency ranges appropriate to individual formants. The axis-crossing rate and the amplitude are measured for the signal in each of the bands. A remaining disadvantage, however, is that the method is still subject to the overlapping of the formant frequency ranges.

A more elaborate implementation of the same basic idea, but with a feature designed to minimize deleterious overlap, has also been made (CHANG). The notion is to employ an iterative measure of the average rate of zero-crossing in a given frequency range and to successively narrow the frequency range on the basis of the measured rate. The expectation is for rapid convergence. Fig. 5.18 illustrates the method. The signal is pre-filtered by fixed filters into ranges roughly appropriate to the first two formants. An axis-crossing measure, ρ_0, of the lower band is made and its value is used to tune automatically a narrower, variable band-pass filter. The axis-crossing output of this filter is, in turn, taken as an indication of the first formant frequency ($F1$). Its value is used to adjust the cut-off frequency of a variable HP filter. The average axis-crossing output of the latter is taken as an estimate of the second formant frequency ($F2$).

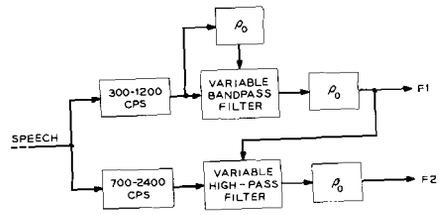


Fig. 5.18. Automatic formant measurement by zero-crossing count and adjustable pre-filtering. (After CHANG)

If the spectral distribution of the signal is continuous, as in the case of unvoiced sounds, the average axis-crossing rate for a given spectral element is approximately twice the first moment of the spectral piece (CHANG, PIHL and ESSIGMAN). However, other more direct methods for measuring spectral moments have been considered.

5.212. Spectral Moments. The n-th moment of an amplitude spectrum $A(\omega)$ is $M_n = \int \omega^n A(\omega)\, d\omega$, where ω is the radian frequency. If a suitable pre-filtering or partitioning of the spectrum can be made, then a formant frequency can be approximated by

$$\bar{\omega} = \frac{M_1}{M_0} \cong \frac{\sum_i \omega_i A(\omega_i)}{\sum_i A(\omega_i)}.$$

A number of formant measures based upon this principle have been examined (POTTER and STEINBERG; GABOR; SCHROEDER, 1956; CAMPANELLA). The spectral partitioning problem remains of considerable importance in the accuracy of these methods. However, certain moment ratios have been found useful in separating the frequency ranges occupied by formants (SUZUKI, KADOKAWA and NAKATA). Another difficulty in moment techniques is the asymmetry or skewness of the spectral resonances. The measured formant frequency may be weighted toward the "heavier" side of the spectrum, rather than placed at the spectral peak.

5.213. Spectrum Scanning and Peak-Picking Methods. Another approach to real-time automatic formant tracking is simply the detection and measurement of prominences in the short-time amplitude spectrum. At least two methods of this type have been designed and implemented (FLANAGAN, 1956a). One is based upon locating points of zero slope in the spectral envelope, and the other is the detection of local spectral maxima by magnitude comparison. In the first—illustrated in Fig. 5.19—

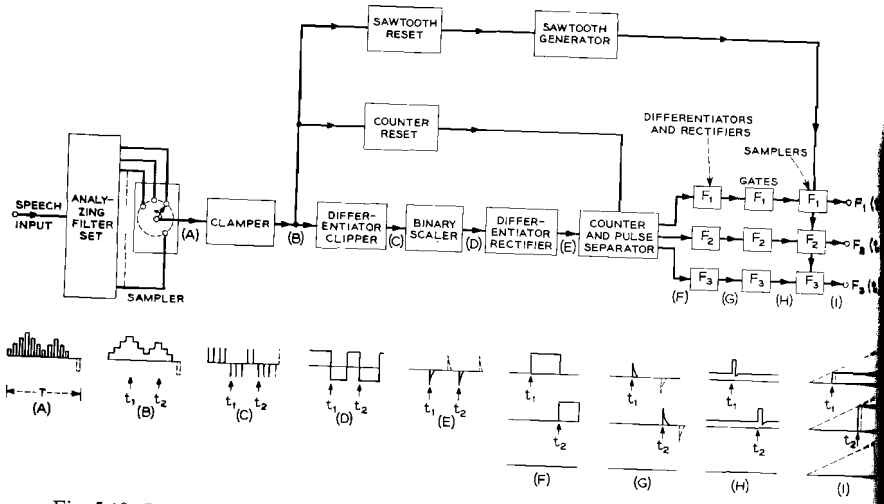


Fig. 5.19. Spectrum scanning method for automatic extraction of formant frequencies. (After FLANAGAN, 1956a)

a short-time amplitude spectrum is first produced by a set of bandpass filters, rectifiers and integrators. The analysis is precisely as described earlier in Section 5.12. The outputs of the filter channels are scanned rapidly (on the order of 100 times per second) by a sample-and-hold circuit. This produces a time function which is a step-wise representation of the short-time spectrum at a number (36 in this instance) of frequency values. For each scan, the time function is differentiated and binary-scaled to produce pulses marking the maxima of the spectrum. The marking pulses are directed into separate channels by a counter where they sample a sweep voltage produced at the scanning rate. The sampled voltages are proportional to the frequencies of the respective spectral maxima and are held during the remainder of the scan. The resulting stepwise voltages are subsequently smoothed by low-pass filtering.

The second method segments the short-time spectrum into frequency ranges that ideally contain a single formant. The frequency of the spectral maximum within each segment is then measured. The operation is illustrated in Fig. 5.20. In the simplest form the segment boundaries are fixed. However, additional control circuitry can automatically adjust the boundaries so that the frequency range of a given segment is contingent upon the frequency of the next lower formant. The normalizing circuit "clamps" the spectral segment either in terms of its peak value or its mean value. This common-mode rejection enables the

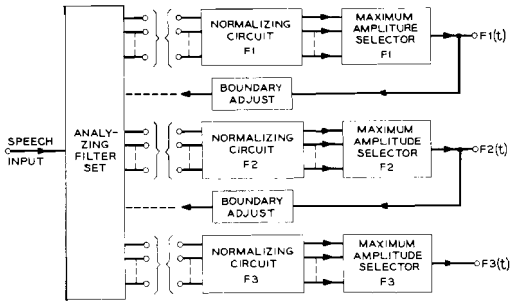


Fig. 5.20. Peak-picking method for automatic tracking of speech formants. (After FLANAGAN, 1956a)

following peak-selecting circuitry to operate over a wide range of amplitudes. The maxima of each segment are selected at a rapid rate—for example, 100 times per second—and a voltage proportional to the frequency of the selected channel is delivered to the output. The selections can be time-phased so that the boundary adjustments of the spectral segments are made sequentially and are set according to the measured position of the next lower formant. A number of improvements on the basic method have been made by providing frequency interpolation (SHEARME, 1959), more sophisticated logic for adjusting the segment boundaries (HOLMES and KELLY), and greater dynamic range for the peak selectors (STEAD and JONES). The objective in all these designs has been the realization of a real-time, practicable hardware device for direct application in a transmission system.

A typical output from the device of Fig. 5.20, using fixed boundaries, is shown in Fig. 5.21. It is clear that the operation is far from perfect.

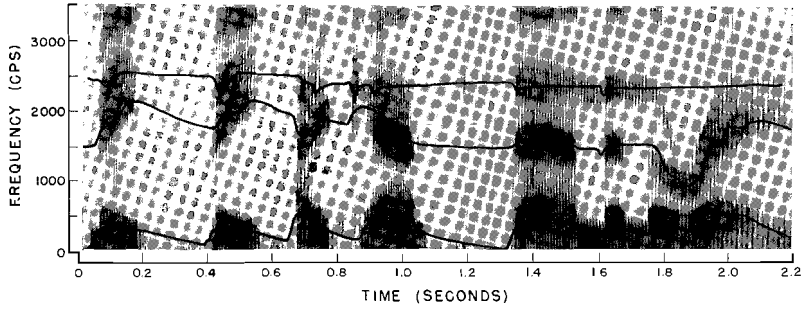


Fig. 5.21. Formant outputs from the tracking device shown in Fig. 5.20. In this instance the boundaries of the spectral segments are fixed

In this example a large third formant error occurs in the /r/ of "rain." Automatic control of the $F2$–$F3$ boundary, however, eliminates this error. As a rough indication of the performance, one evaluation shows that its output follows $F1$ of vowels within ± 150 cps greater than 93% of the time, and $F2$ within ± 200 cps greater than 91% of the time (FLANAGAN, 1956a). Although one desires greater precision, this method – because of its simplicity and facility for real-time analysis – has proved useful in several investigations of complete formant-vocoder systems (FLANAGAN and HOUSE; STEAD and JONES; SHEARME, SMITH and KELLY).

5.214. Digital Computer Methods for Formant Extraction. The development of digital computers has enabled application of more sophisticated strategies to speech processing. The more esoteric processings are made possible by the ability of the computer to store and rapidly manipulate large quantities of numerical data. A given data sample can be held in the machine while complex tests and measures are applied to analyze a particular feature and make a decision. This advantage extends not only to formant tracking, but to all phases of speech processing. The relations between sampled-data systems and continuous systems (see, for example, RAGAZZINI and FRANKLIN) permit simulation of complete transmission systems within the digital computer. This is a topic in itself, and we will return to it in a later chapter.

The digital analyses which have been made for speech formants have been primarily in terms of operations on the spectrum. The spectrum either is sampled and read into the computer from an external filter bank, or is computed from a sampled and quantized version of the speech waveform. One approach along the latter line has been a pitch-synchronous analysis of voiced sounds (MATHEWS, MILLER and DAVID, 1961b). Individual pitch periods are determined by visual inspection of the speech oscillogram. The computer then calculates the Fourier series for each pitch period as though that period were one of an exactly periodic signal. The envelope of the calculated spectrum is then fitted by a synthetic spectrum in successive approximations and according to a weighted least-square error criterion. A pole-zero model for the vocal tract and the glottal source, based upon acoustic relations for the vocal tract (see Chapter III), produces the synthetic spectrum.

The fitting procedure is initiated by guessing a set of poles and zeros appropriate to the calculated real spectrum. The computer then successively increments the frequency and damping of each individual pole and zero to minimize the weighted mean-square error (in log-amplitude measure). After about 10 to 20 complete cycles, a close fit to the speech spectrum can be obtained. Typical rms log-amplitude errors range from about 1.5 to 2.5 db. A typical result of the fitting procedure is shown in

Fig. 5.22. The measured formant frequencies and bandwidths are then taken as the frequencies and bandwidths of the best fitting spectral model.

A computer system for non-pitch-synchronous formant analysis, in which spectral data are produced external to the computer, can also be summarized (HUGHES; FORGIE and HUGHES). A bank of 35 contiguous bandpass filters with rectifiers and integrators produces a short-time spectrum of the running speech. The filter outputs are scanned at a rapid rate ($180\ \text{sec}^{-1}$) to produce a framed time function which represents successive spectral sections (essentially the same as that shown in Fig. 5.5). This time function is sampled every 154 μsec and quantized to 11 bits by an analog-to-digital converter. A certain amount of the data is then held in the computer storage for processing.

One analysis procedure for which the computer is programmed (1) locates the fricative sounds in a word and classifies them; (2) locates the first and second formants in voiced segments; and (3) calculates the overall sound level. The formant tracking procedure is basically a peak-picking scheme similar to that shown previously in Fig. 5.20. However, a number of detailed, programmed constraints are included to exploit vocal tract characteristics and limitations. In principle, the procedure for a given spectral scan is as follows. Find the peak filter in the frequency range appropriate to the first formant. Store the frequency and ampli-

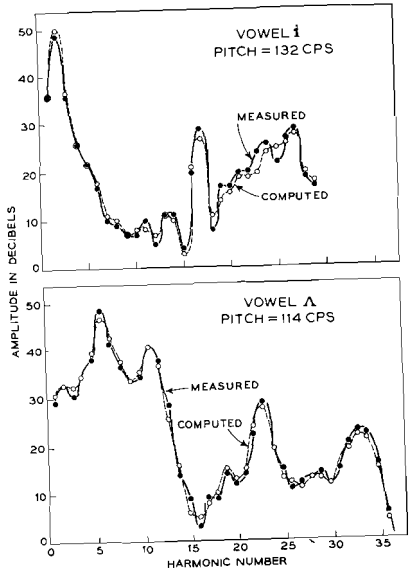


Fig. 5.22. Spectral fit computed for one pitch period of a voiced sound. (After MATHEWS, MILLER and DAVID, 1961b)

tude values of this channel. On the basis of the $F1$ location, adjust the frequency range for locating $F2$. Locate the peak filter in the adjusted $F2$ range and store its frequency and amplitude values. Finally, examine the next spectral scan and find $F1$ and $F2$, subject to continuity constraints with previously determined values. Large, abrupt changes in $F1$ and $F2$ of small time duration are ignored. Typical results, described as "good" and "average" from this procedure are shown in Fig. 5.23.

"Hawaii" (a) "Yowie" (b)

Fig. 5.23a and b. Tracks for the first and second formant frequencies obtained from a computer-analysis of real-time spectra. The speech samples are (a) "Hawaii" and (b) "Yowie" uttered by a man. (After HUGHES)

A real-time spectral input to a computer has also been applied in a spectral-fitting technique for formant location (BELL *et al.*). The procedure—termed "analysis-by-synthesis" by its originators—is illustrated in Fig. 5.24. As before, a filter bank produces a short-time spectrum which is read into the digital computer via an analog-to-digital converter. Inside the computer, speech-like spectra are generated from a pole-zero model of the vocal tract and its excitation. (The filter bank characteristics are also applied to the synthetic spectra.) As in the pitch-synchronous analysis, the model is based upon the acoustical principles discussed in Chapter III. The real and synthetic spectra at a given instant are compared, and a weighted square error is computed. The nature of the comparison is illustrated in Fig. 5.25. The effect of an error in formant frequency is indicated by Fig. 5.25a. An error in formant bandwidth is illustrated in Fig. 5.25b.

On the basis of error computations for the immediate and for adjacent spectral samples, a programmed automatic control strategy determines the procedure for adjusting the pole-zero positions of the fitting

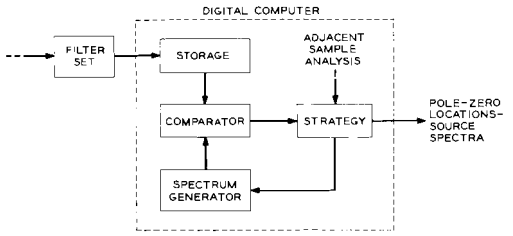


Fig. 5.24. Computer procedure for formant location by the "analysis-by-synthesis" method. (After BELL *et al.*)

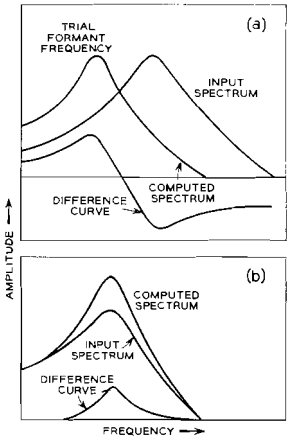


Fig. 5.25a and b. Idealized illustration of formant location by the "analysis-by-synthesis" method shown in Fig. 5.24

synthetic spectrum to reduce the weighted error. When a minimum-error fit is obtained, the computer automatically stores the pole-zero locations of the vocal tract model and the source characteristics chosen for that spectrum. Five operations are carried out by the computer: (1) storage of real input speech spectra; (2) generation of synthetic spectra; (3) control and adjustment of the synthetic spectra; (4) calculation of spectral difference according to a prescribed error criterion; and (5) storage and display of the parameters which yield minimum error. Provisions are made so that, if desired, the comparison and control functions can be performed by an human operator instead of by the automatic procedure.

In principle the programmed matching procedure is applicable both to vowel and consonant spectra, but the matching model for consonants is generally more complex. A typical result of the procedure is shown for the first three formants in Fig. 5.26. The (a) part of the figure shows a sound spectrogram of the utterance /h ə b I b/ with sample intervals laid off along the top time axis. The (b) part of the figure shows the computer-determined formant tracks for essentially the vowel portion of the second syllable (i.e., /I/). The sample numbers on the abscissa of the (b) part correspond with those at the top of (a). The top diagram in part (b) is the square error for the spectral fit. The "analysis-by-synthesis" technique has also been implemented using a gradient-climbing calculation for matching the short-time spectrum (OLIVE). Other implementations have used sequential algorithms for fitting the spectrum (FUJISAKI).

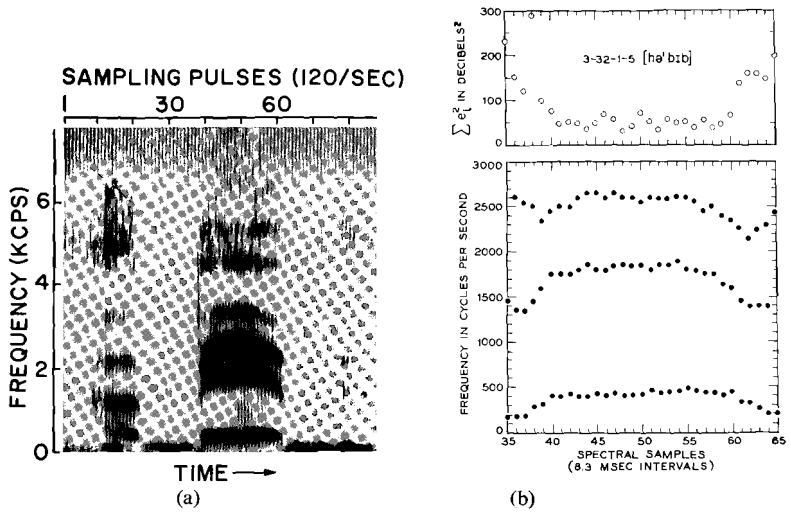


Fig. 5.26a and b. Computer-determined formant tracks obtained by the "analysis-by-synthesis" method. (a) Spectrogram of original speech. (b) Extracted formant tracks and square error measure. (After BELL *et al.*)

Another computer formant tracker uses a principle related to the pole-zero model of speech (COKER). The analyzing strategy is a combined peak-picking and spectral fitting approach. A filter bank, associated rectifiers and lowpass filters produce a short-time spectrum. The filter outputs are scanned by an electronic commutator, and the time waveform representing the spectral sections is led to an analog-to-digital converter. The output digital signal describing the successive spectra is read into the computer, and the short-time spectra are stored in the memory.

The automatic analyzing procedure, prescribed by a program, first locates the absolute maximum of each spectrum. A single formant resonance is then fitted to the peak. The single resonance is positioned at a frequency corresponding to the first moment of that spectral portion lying, say, from zero to 6 db down from the peak on both sides. The single formant resonance is then inverse filtered from the real speech spectrum by subtracting the log-amplitude spectral curves. The operation is repeated on the remainder until the required number of formants are located. Since the peakpicking is always accomplished on the whole spectrum, the problem of formant segmentation is obviated! Proximate formants can also be resolved and accurate results can be obtained on running speech. The formant selections can be displayed directly on the spectral sections in a manner similar to that shown in Fig. 5.5. Again, the ability of the computer to store large amounts of data and to perform relatively complex operations at high speed permits a detailed fitting of the

spectrum. The analysis is easily accomplished in real time, and the computer can essentially be used as the formant-tracking element of a complete formant-vocoder system (COKER and CUMMISKEY).

A still different method for formant analysis (SCHAFER and RABINER) makes use of a special digital transform—the Chirp-Z transform (RABINER, SCHAFER and RADER). The method also incorporates Fast Fourier Transform methods for spectral analysis (COOLEY and TUKEY). In its complete form, the method depends upon relations prescribed by a 3-pole model of voiced sounds and a single pole-zero model of voiceless sounds.

The point of departure is a short-time transform of the speech waveform for both voiced an voiceless sounds. The steps in the spectral analysis are depicted in Fig. 5.27.

The upper part of the figure shows the analysis of voiced speech. The waveform at the top left is a segment of voiced speech of approximately 40 msec duration, which has been multiplied by a Hamming window[1]. Over such a short time interval, the speech waveform looks

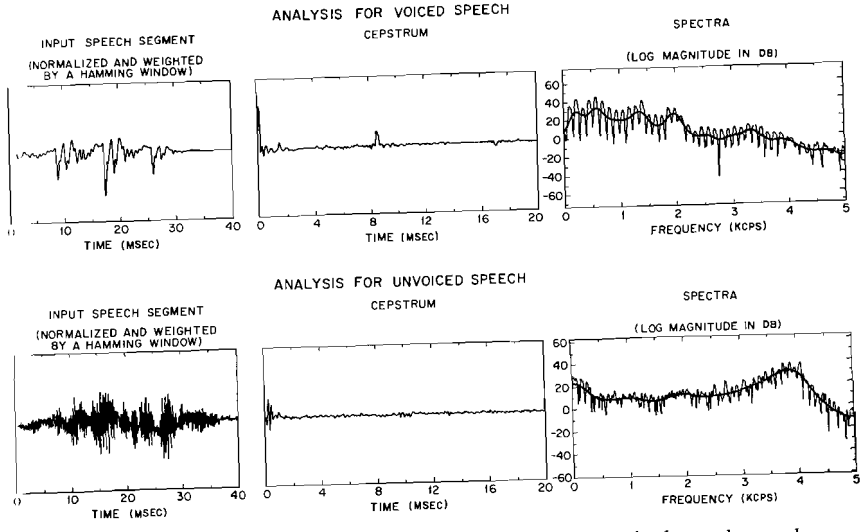


Fig. 5.27. Spectrum and cepstrum analysis of voiced and unvoiced speech sounds. (After SCHAFER and RABINER)

[1] The Hamming window is specified by the function

$$h(t)=\left\{0.54+0.46\cos\left(\frac{2\pi t}{\tau}\right)\right\} \quad \text{for} \quad -\frac{\tau}{2}\leqq t\leqq\frac{\tau}{2},$$

where τ is the window duration. This data window is attractive because the side lobes of its Fourier transform remain more than 40 db down at all frequencies (BLACKMAN and TUKEY).

like a segment of a periodic waveform. The detailed time variation of the waveform during a single period is primarily determined by the vocal tract response, while the fundamental period (pitch period) reflects the vocal-cord vibration rate.

The logarithm of the magnitude of the Fourier transform of this segment of speech is the rapidly-varying spectrum plotted at the top right of Fig. 5.27. This function can be thought of as consisting of an additive combination of a rapidly-varying periodic component, which is associated primarily with the vocal-cord excitation, and a slowly-varying component primarily due to the vocal-tract transmission function. Therefore, the excitation and vocal-tract components are mixed and must be separated to facilitate estimation of formant values. The standard approach to the problem of separating a slowly-varying signal and a rapidly-varying signal is to employ linear filtering. Such an approach applied to the log magnitude of the short-time Fourier transform leads to the computation of the *cepstrum* (BOGERT, HEALY and TUKEY).

The cepstrum is a Fourier transform of a Fourier transform. To compute the cepstrum the Fourier transform of the time waveform is computed. The logarithm is taken of the magnitude of this transform. Inverse Fourier transformation of this log-magnitude function produces the cepstrum. (See also Section 5.3.)

The cepstrum is plotted in the middle of the top row of Fig. 5.27. The rapidly-varying component of the log-magnitude spectrum contributes the peak in the cepstrum at about 8 msec (the value of the pitch period). The slowly-varying component corresponds to the low-time portion of the cepstrum. Therefore, the slowly-varying component can be extracted by first smoothly truncating the cepstrum values to zero above about 4 msec, and then computing the Fourier transform of the resulting truncated cepstrum. This yields the slowly-varying curve which is superimposed on the short-time spectrum, shown at the right of the top row in Fig. 5.27.

The formant frequencies correspond closely with the resonance peaks in the smoothed spectrum. Therefore, a good estimate of the formant frequencies is obtained by determining which peaks in the smoothed spectrum are vocal tract resonances. Constraints on formant frequencies and amplitudes, derived from a three-pole model of voiced sounds, are incorporated into an alogrithm which locates the first three formant peaks in the smoothed spectrum.

The analysis of unvoiced speech segments is depicted in the bottom row of Fig. 5.27. In this case, the input speech resembles a segment of a random noise signal. As before, the logarithm of the magnitude of the Fourier transform of the segment of speech can be thought of as

consisting of a rapidly-varying component, due to the excitation, plus a slowly-varying component due to the spectral shaping of the vocal-tract transfer function. In this case, however, the rapidly-varying component is not periodic but is random. Again the low-time part of the cepstrum corresponds to the slowly-varying component of the transform, but the high-time peak present in the cepstrum of voiced speech is absent for unvoiced speech. Thus, the cepstrum can also be used in deciding whether an input speech segment is voiced or unvoiced, and if voiced, the pitch period can be estimated from the location of the cepstral peak. Low-pass filtering of the logarithm of the transform, by truncation of the cepstrum and Fourier transformation, produces the smoothed spectrum curve which is again superimposed on the short-time transform at the lower right of Fig. 5.27. In this case, an adequate specification of the spectrum shape can be achieved by estimating the locations of a single wide-bandwidth resonance and a single anti-resonance, i.e., a single pole and zero.

Continuous speech is analyzed by performing these operations on short segments of speech which are selected at equally-spaced time intervals, typically 10–20 msec apart. Fig. 5.28 illustrates this process for a section of speech which, as evidenced by the peaks in the cepstra, is voiced throughout. The short-time spectrum and smoothed spectrum corresponding to each cepstrum are plotted adjacent to the cepstrum. In going from top to bottom in Fig. 5.28, each set of curves corresponds to the analysis of segments of speech selected at 20 msec increments in time. The formant peaks determined automatically by the program are connected by straight lines. Occasionally the formants come close together in frequency and pose a special problem in automatic extraction.

In the third and fourth spectra from the top, the second and third formants are so close together that there are no longer two distinct peaks. A similar situation occurs in the last four spectra where the first and second formants are not resolved. A procedure for detecting such situations has been devised and a technique for enhancing the resolution of the formants has been developed. An example of the technique is shown in Fig. 5.29.

The curve shown in Fig. 5.29a is the smooth spectrum as evaluated along the $j\omega$-axis of the complex frequency s-plane. (The lowest three vocal tract eigen-frequencies corresponding to this spectrum are depicted by the x's in the s-plane at the left.) Because formants two and three ($F2$ and $F3$) are quite close together, only one broad peak is observed in the conventional Fourier spectrum. However, when the spectrum is evaluated on a contour which passes closer to the poles, two distinct peaks are in evidence, as shown in Fig. 5.29b. The Chirp z-transform alogrithm facilitates this additional spectral analysis by allowing a fast

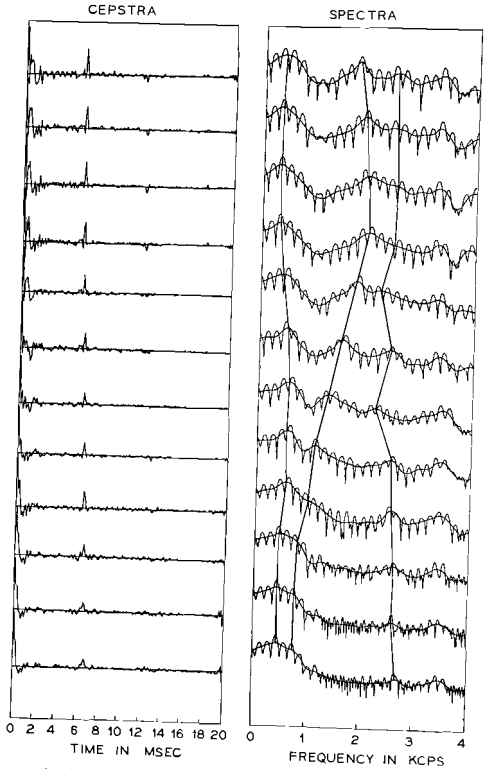


Fig. 5.28. Cepstrum analysis of continuous speech. The left column shows cepstra of consecutive segments of speech separated by 20 ms. The right column shows the corresponding short-time spectra and the cepstrally-smoothed spectra

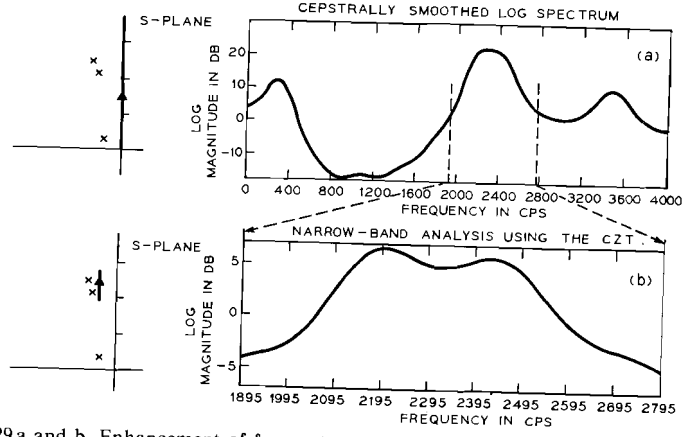


Fig. 5.29a and b. Enhancement of formant resonances by the Chirp-z transform: (a) Cepstrally-smoothed spectrum in which F_2 and F_3 are not resolved. (b) Narrow-band analysis along a contour passing closer to the poles. (After SCHAFER and RABINER)

computation of the spectrum along an s-plane contour shown at the left of Fig. 5.29b.

Once the vocal excitation and formant functions are determined, they can be used to synthesize a waveform which resembles the original speech signal. (Systems for speech synthesis from formant data are discussed in Section 6.2.) Comparison of the formant-synthesized signal with the original speech signal is an effective means for evaluating the automatic formant tracking. Fig. 5.30 shows a typical result of automatic analysis and synthesis of a voiced sentence. The upper curves show the pitch period and formant parameters as automatically estimated from a natural utterance whose spectrogram is also shown in the figure. The bottom of the figure shows the spectrogram of speech synthesized from the automatically estimated pitch and formant parameters. Comparison of the spectograms of the original and synthetic speech indicates that the spectral properties are reasonably well preserved.

Another approach using computer processing is the analysis of real speech spectra in terms of a model of articulation (HEINZ, 1962a, b).

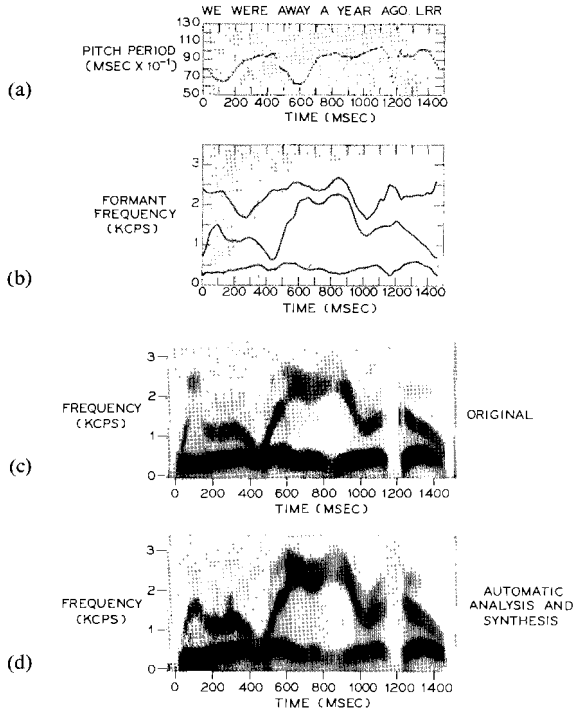


Fig. 5.30a–d. Automatic formant analysis and synthesis of speech. (a) and (b) Pitch period and formant frequencies analyzed from natural speech. (c) Spectrogram of the original speech. (d) Spectrogram of synthesis speech. (After SCHAFER and RABINER)

This approach differs from the preceding techniques essentially in the spectrum-generation and control strategy operations. The vocal tract poles and zeros are obtained from an articulatory or area function specification of the tract. These are obtained by solving the Webster horn equation (see Chapter III). A spectrum corresponding to the computed poles and zeros is generated and compared to the real speech spectrum. The error in fit is used to alter the synthetic spectrum by adjusting, on the articulatory level, the vocal tract area function. A modification of a three-parameter description of vocal configuration is used to specify the area function (DUNN, 1950; STEVENS and HOUSE, 1955; FANT, 1960).

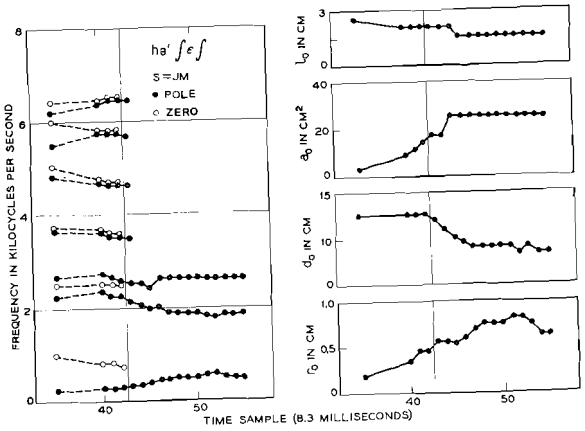
This formulation, provided the area function can be specified accurately enough, offers an important advantage over pole-zero models of the vocal system. The latter have as their input parameters the locations in the complex plane of the poles and zeros of the vocal transmission. The poles of the system are independent of source location and depend only on the configuration (see Chapter III). They move in a continuous manner during the production of connected speech, even though the source may change in character and location. The zeros, however, depend upon source location as well as upon tract configuration. They may move, appear and disappear in a discontinuous fashion. This discontinuous behavior of the zeros—and the resulting large changes in the speech spectrum—makes pole-zero tracking difficult.

An articulatory description of the signal obviates these difficulties to a considerable extent. More realistic continuity constraints can be applied to the articulators. The location of the unvoiced source is generally implied by the configuration, and the vocal zero specification is an automatic by-product of the specification of configuration and excitation. In terms of articulatory parameters, the spectra of consonants and consonant-vowel transitions can be matched with little more difficulty than for vowels. A typical result of this articulatory fitting procedure is shown in Fig. 5.31.

The left diagram shows the temporal courses of the poles and zeros in the /ʃ ɛ/ portion of the bisyllabic utterance /h ə′ ʃ ɛ ʃ/ (the time scale is the sample number multiplied by 8.3 msec). The vertical line, where the zero tracks disappear, represents the consonant-vowel boundary. (Only the first three formants are computed in the vowel part of the utterance.) The diagram to the right shows the corresponding temporal courses of the four articulatory parameters that were adjusted to make the spectral matches. They are:

r_0, the effective radius at the tongue constriction,

d_0, the location of the tongue constriction measured from the glottis,



(a) (b)

Fig. 5.31 a and b. Pole-zero computer analysis of a speech sample using an articulatory model for the spectral fitting procedure. The (a) diagram shows the pole-zero positions calculated from the articulatory model. The (b) diagram shows the articulatory parameters which describe the vocal tract area function. (After HEINZ, 1962a)

a_0, the cross-sectional area of the mouth opening, and

l_0, the length of the lip tube (or mouth section).

Their trajectories are essentially continuous as the match proceeds across the consonant-vowel boundary. In going from the fricative /ʃ/ to the vowel /ɛ/, the mouth section becomes shorter and more open. The position of the constriction moves back toward the glottis, and the radius of the constriction becomes larger. The position of the unvoiced sound source during the fricative is taken 2.5 cm anterior to the constriction (i.e., $d_0+2.5$). The manner in which these relatively simple motions describe the more complicated pole-zero pattern is striking. Success of the method depends directly upon the accuracy with which the articulatory parameters describe the vocal-tract shape. Derivation of sophisticated articulatory models is an important area for research. (See Section 5.4.)

5.22. Measurement of Formant Bandwidth

The bandwidths of the formant resonances—or the real parts of the complex poles—are indicative of the losses associated with the vocal system. Not only are quantitative data on formant bandwidths valuable in corroborating vocal tract calculations (for example, those made in Chapter III for radiation, viscous, heat-conduction, cavity-wall and glottal losses), but a knowledge of the damping is important in the proper synthesis of speech.

A number of measurements have been made of vocal tract damping and formant bandwidth[1]. The measurements divide mainly between two techniques; either a measure of a resonance width in the frequency domain, or a measure of a damping constant (or decrement) on a suitably filtered version of the speech time waveform. In the former case the formant is considered as a simple resonance, and the half-power frequencies of the spectral envelope are determined. In the latter case the formant is considered a damped sinusoid, having amplitudes A_1 and A_2 at times t_1 and t_2. The damping constant, σ, for the wave and its half-power bandwidth, Δf, are related simply as

$$\sigma=\pi\Delta f=\frac{\ln A_2/A_1}{(t_2-t_1)}.$$

The results of one of the more extensive formant bandwidth studies are summarized in Fig. 5.32 (DUNN, 1961). Part (a) of the figure shows the formant bandwidths measured by fitting a simple resonance curve to amplitude sections of vowels uttered in an /h–d/ syllable. The data are averages for 20 male voices producing each vowel. The second

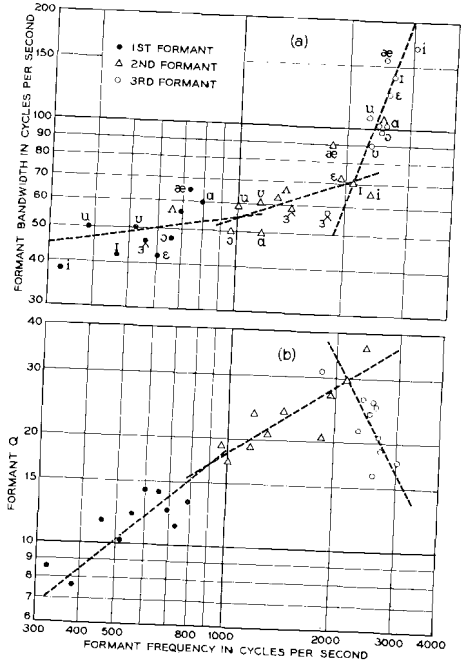
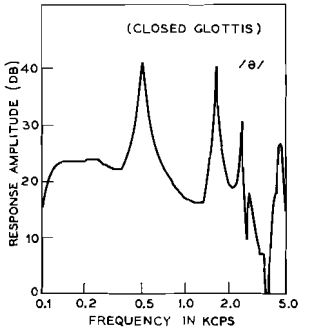


Fig. 5.32a and b. Measured formant bandwidths for adult males. (After DUNN, 1961)

[1] For a good summary and bibliography of earlier investigations, see DUNN (1961), Also, see FANT (1958, 1959a, b).

curve (b) represents the same data plotted in terms of $Q=f/\Delta f$. The upper graph shows that over the frequency ranges of the first and second formants, the nominal bandwidths are generally small—on the order of 40 to 70 cps. Above 2000 cps the bandwidth increases appreciably. The lower plot of formant-Q *vs* formant frequency shows that resonant Q's are largest in the frequency region around 2000 cps.

Formant bandwidths can also be effectively measured from a frequency response of the actual vocal-tract (FUJIMURA). A sine wave of volume velocity is introduced into the vocal-tract at the glottal end by means of a throat vibrator. The pressure output at the mouth is measured as the input source is changed in frequency. A typical vocal-tract frequency response is shown in Fig. 5.33a. The variation in first-formant bandwidth, as a function of first-formant frequency, is shown in 5.33b.



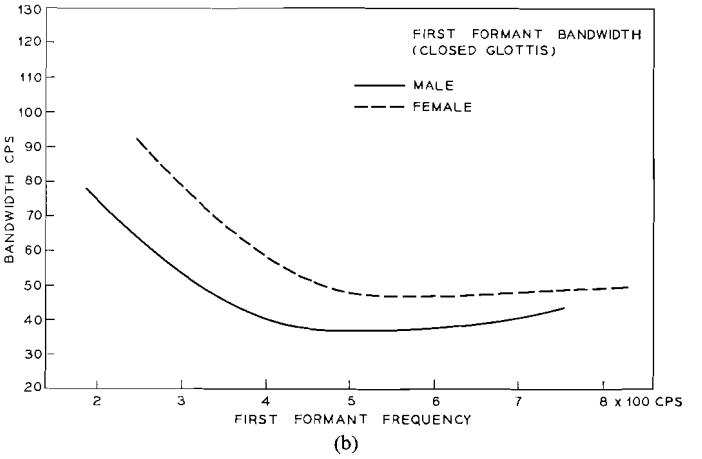


Fig. 5.33a and b. (a) Vocal-tract frequency response measured by sine-wave excitation of an external vibrator applied to the throat. The articulatory shape is for the neutral vowel and the glottis is closed. (After FUJIMURA and LINDQUIST). (b) Variation in first-formant bandwidth as a function of formant frequency. Data for men and women are shown for the closed-glottis condition. (After FUJIMURA and LINDQUIST)

These data are for a closed-glottis condition. The bandwidth is seen to increase as first formant frequency diminishes, owing primarily to the influence of cavity-wall loss. (See calculations of cavity-wall loss in Section 3.37.)

The origins of the principal contributions to vocal-tract damping have already been indicated by the theory derived in Chapter III. These are glottal loss and cavity-wall loss for the lower formants, and radiation, viscous and heat-conduction loss for the higher formants.

5.3. Analysis of Voice Pitch

Fundamental frequency analysis – or "pitch extraction" – is a problem nearly as old as speech analysis itself. It is one for which a complete solution remains to be found. The main difficulty is that voice pitch has yet to be adequately defined. Qualitatively, pitch is that subjective attribute that admits of rank ordering on a scale ranging from low to high. The voiced excitation of the vocal tract is only quasi-periodic. Not only does the exciting glottal waveform vary in period and amplitude, but it also varies in shape. Precisely what epochs on the speech waveform, or even on the glottal waveform, should be chosen for interval or period measurement is not clear. Furthermore, the relation between an interval, so measured, and the perceived pitch is not well established.

Most pitch-extracting methods take as their objective the indication of the epoch of each glottal puff and the measurement of the interval between adjacent pulses. Still, exactly how this relates to the pitch percept with all the random jitter and variation of the glottal wave is a question worthy of inquiry.

Most automatic or machine pitch extractors attempt either to describe the periodicity of the signal waveform (GRÜTZMACHER and LOTTERMOSER; GRUENZ and SCHOTT; DOLANSKY, 1955; GILL) or to measure the frequency of the fundamental component if it is present (DUDLEY, 1939b). Computer efforts at pitch extraction essentially do the same, but usually more elaborate constraints and decisions are applied (INOMATA; GOLD; SUGIMOTO and HASHIMOTO).

One particularly useful method for machine pitch extraction utilizes properties of the cepstrum to reveal signal periodicity (NOLL; OPPENHEIM, SCHAFER and STOCKHAM). As described in Section 5.214, the cepstrum is defined as the Fourier transform of the logarithm of the amplitude spectrum of a signal. Since it is a transform of a transform, and since the resulting independent variable is reciprocal frequency, or time, the terms "cepstrum" and "quefrency" were coined by its inventors

(BOGERT, HEALY and TUKEY) to designate the transform and its independent variable.

The log-taking operation has the desirable property of separating source and system characteristic (at least to the extent that they are spectrally multiplicative). If the output speech wave, $f(t)$, is the convolution of the vocal tract impulse response, $v(t)$, and the vocal excitation source, $s(t)$, the magnitudes of their Fourier transforms are related as

$$|F(\omega)| = |V(\omega)| \cdot |S(\omega)|,$$

where all the amplitude spectra are even functions. Taking the logarithm of both sides gives

$$\ln|F(\omega)| = \ln|V(\omega)| + \ln|S(\omega)|.$$

Similarly, taking the Fourier transform[1] of both sides yields

$$\mathscr{F}\ln|F(\omega)| = \mathscr{F}\ln|V(\omega)| + \mathscr{F}\ln|S(\omega)|.$$

For voiced sounds, $|S(\omega)|$ is approximately a line spectrum with components spaced at the pitch frequency $1/T$. $\mathscr{F}\ln|S(\omega)|$ therefore exhibits a strong component at the "quefrency", T. $|V(\omega)|$, on the other hand, exhibits the relatively "slow" formant maxima. Consequently $\mathscr{F}\ln|V(\omega)|$ has its strongest component at a very low quefrency.

Because of the additive property of the transforms of the log amplitude spectra, the characteristics of the source and system are well separated in the cepstrum. The cepstrum is therefore also a valuable tool for formant analysis as well as pitch measurement (SCHAFER and RABINER). (See Section 5.21.) Measurement of pitch and voiced-unvoiced excitation is accomplished by using a suitable strategy to detect the quefrency components associated with $\mathscr{F}\ln|S(\omega)|$. Because the method does not require the presence of the fundamental component, and because it is relatively insensitive to phase and amplitude factors (owing to the log-magnitude operations) it performs well in vocoder applications. In one test with a complete channel vocoder, it demonstrated superior performance in extracting the pitch and voiced-unvoiced control data (NOLL). Because a large amount of processing is necessary, the method is most attractive for special purpose digital implementations where Fast Fourier Transform hardware can be used. An illustration of pitch determination by cepstrum computation has been shown previously in Figs. 5.28a and 5.30.

Perhaps a more basic measurement of voiced excitation is that of the glottal volume-velocity wave (R. L. MILLER, 1959; FANT, 1959b; MATHEWS, MILLER and DAVID, 1961a; HOLMES, 1962). Approximations

[1] Formally an inverse Fourier transform.

to this function can be obtained by so-called inverse-filtering techniques. The idea is to pass the speech signal through a network whose transmission function is the reciprocal of that of the vocal tract for the particular sound. Zeros of the network are adjusted to nullify vocal tract poles, and the resulting output is an approximation to the input glottal volume current.

The inverse-filtering analysis presumes that the source and system relations for the speech-producing mechanism do not interact and can be uniquely separated and treated independently. This assumption is a treacherous one if the objective is an accurate estimate of the glottal volume velocity. In the real vocal tract they interact to a certain extent (particularly at the first-formant frequency). Another difficulty is that it is not always clear whether to ascribe certain properties (primarily, zeros) to the tract or to the source. The estimate obtained for the glottal wave obviously depends upon the vocal-tract model adopted for the inverse filter. The criterion of adjustment of the inverse filter also influences the answer. Under certain conditions, for example, ripples on the inverse wave which may be thought to be formant oscillations might in fact be actual glottal variations.

One question often raised is "where in the pitch period does the excitation occur." Presumably if such an epoch could be determined, the pulse excitation of a synthesizer could duplicate it and preserve natural irregularities in the pitch period. Because the glottal wave frequently changes shape, such a datum is difficult to describe. One claim is that this epoch commonly is at the close of the cords (R. L. MILLER, 1959), while another (HOLMES, 1962) is that it can occur at other points in the wave. To a first approximation, such an epoch probably coincides with the greatest change in the derivative of the glottal waveform. Often this point can occur just about anywhere in the period. For a triangular wave, for example, it would be at the apex.

A perceptual study has been made of the effects of the glottal waveform on the quality of synthetic speech. The results support the notion that the significant vocal excitation occurs at the point of greatest slope change in the glottal wave (ROSENBERG, 1971 b). Natural speech was analyzed pitch-synchronously. The vocal-tract transmission and the glottal waveform were determined and separated by inverse filtering. Artificial glottal waveforms were substituted and the speech signal was regenerated. Listening tests showed that good quality speech can be obtained from an excitation function fixed in analytical form. The absence of temporal detail, period-to-period, does not degrade quality. A preferred glottal pulse shape has but a single slope discontinuity at closing. It is intrinsically asymmetric, so its spectral zeros never fall on or near the $j\omega$-axis for any combination of opening and closing times (ROSENBERG, 1971 b).

5.4. Articulatory Analysis of the Vocal Mechanism

The discussion of Chapter III showed that if the vocal tract configuration is known, the system response can be computed and the mode structure specified. The cross-sectional area as a function of distance is sufficient to compute the lower eigenfrequencies of the tract. An accurate account of losses along the tract requires knowledge of the cross-sectional shape or the circumference. [See Eq. 3.33).] Because the vocal mechanism is relatively inaccessible, the necessary dimensions are obviously difficult to obtain. Even at best, present methods of measurement yield incomplete descriptions of tract dimensions and dynamics.

X-ray techniques for motion and still pictures have provided most of the articulatory information available to date. The X-ray data generally are supplemented by other measures. Conventional moving pictures can be made of the external components of the vocal system. Palatograms, molds of the vocal cavities, and electromyographic recordings are also useful techniques for "filling in the picture." Much of the effort in X-ray analysis is directed toward therapeutic goals, such as cleft palate repair and laryngeal treatment. Consequently, the results are often left in only a qualitative form. Several investigations, however, have aimed at measuring vocal dimensions and articulatory movements. (FANT, 1960; CHIBA and KAJIYAMA; PERKELL; FUJIMURA *et al.*; HOUDE.)

One of the main problems in obtaining such data is keeping the radiation dose of the subject within safe limits. This usually means that only a very limited amount of data can be taken on a single individual. One ingenious solution to this problem utilizes a computer-controlled X-ray beam which, under program control, is made to irradiate and track only the physiological areas of interest (FUJIMURA *et al.*).

Another problem is the detail of the X-ray photograph. This is particularly a problem in moving X-ray photography, even with the best image-intensifier tubes. Detail which looks deceptively good in the (visually-integrated) moving picture, disappears when one stops the film to study a single frame. Sound recordings are usually made simultaneously for analysis, but often are of poor quality because of the noise of the proximate movie camera.

The detail in still pictures is somewhat better but nevertheless lacking. An example of a typical medical X-ray is shown in Fig. 5.34. The tongue and lips of the subject were coated with a barium compound to make them more visible. The vocal tract position is appropriate to the production of a high-front vowel close to /i/.

The typical procedure for obtaining an area function from the X-ray picture can be illustrated. An axial line through the centers of gravity of the cross sectional areas is first located, as shown in Fig. 5.35a

Fig. 5.34. Sagittal plane X-ray of adult male vocal tract

(FANT, 1960). The shape and area of the cross-sections at a number of locations are estimated, as shown in Fig. 5.35 b. The shape estimates are deduced on the basis of all available data, including dental molds of the vocal and nasal cavities, conventional photographs and X-ray photographs from the front. These sections provide anchor points for an estimate of the whole area curve. Intermediate values are established both from the sagittal plane X-ray tracing and from continuity considerations to give the complete area function, as shown in Fig. 5.35 c. Typical results for several sounds produced by one man are shown in Fig. 5.36.

Even under best conditions, some of the vocal dimensions during natural speech are impossible to measure. For example, one often can

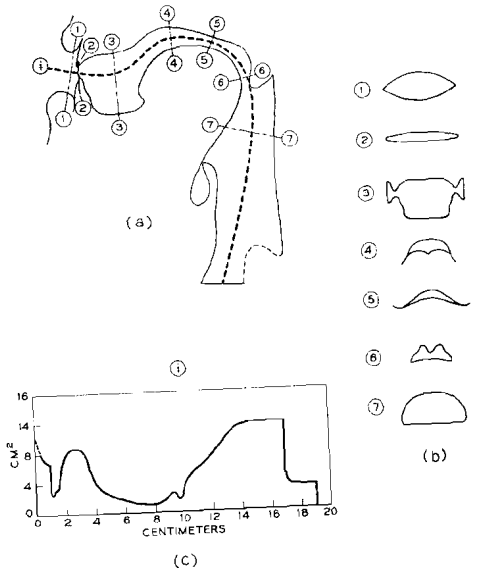


Fig. 5.35 a–c. Method of estimating the vocal tract area function from X-ray data. (After FANT, 1960)

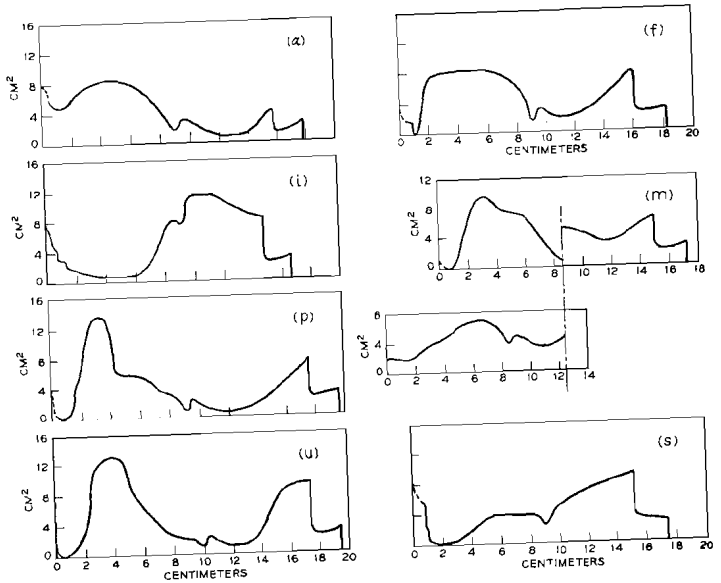


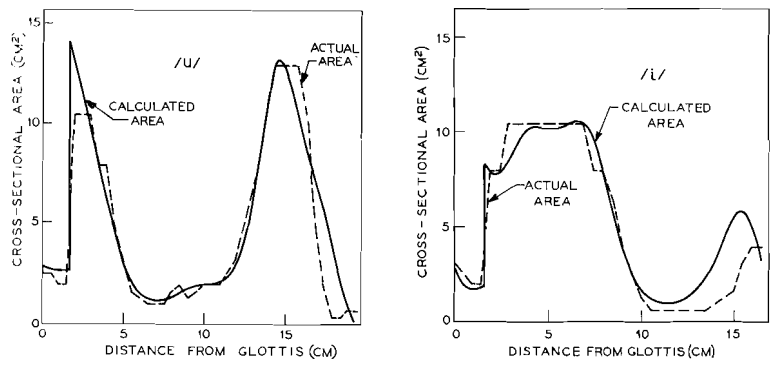
Fig. 5.36. Typical vocal area functions deduced for several sounds produced by one man. (After FANT, 1960)

only make crude estimates of the true shape and lateral dimensions of the pharynx cavity. In the same vein, the true dimensions of the constrictions for fricatives and affricates and the lateral pathways in /l/ are often very uncertain.

Similarly, the vocal source of excitation cannot be studied easily by direct methods. For sustained, open vowels, however, the vocal cord source can be examined by high-speed moving pictures. Measurements of subglottal pressure are also possible and give insight into vocal cord operation. Characteristics of the unvoiced sources, on the other hand, i.e., location, spectral properties and internal impedance, are best inferred from physiological configuration, air flow measurements and spectral analysis of the output sound.

Research interest in better methods for physiological measurements remains high. One active research area centers on the possibilities for relating electromyographic recordings of muscle potentials to the articulator movements observed in X-ray pictures. Several "exotic" schemes for vocal measurement have also been proposed, half humorously. They may, however, hold some promise. For example, a conducting dag loop might be painted around the circumference of the tract at a given position and electrical leads attached. The cross sectional area at that point could be measured by placing the subject in a magnetic field normal to the section and measuring the flux which links the dag loop. Other possibilities might be the attachment of miniature strain gauges at significant points, or the placement of inflatable annular cuffs or catheters at given positions in the tract. Still other possibilities include miniature ultrasonic transducers fixed to the articulators.

Acoustic measurements directly on the vocal-tract also promise useful estimation of the cross-sectional area function (MERMELSTEIN and SCHROEDER; GOPINATH and SONDHI). In one method the acoustic impedance of the tract is periodically sampled at the mouth (GOPINATH and SONDHI). While the subject silently articulates into an impedance tube, pulses of sound pressure are produced periodically (typically at 100 sec^{-1}) and the volume velocity response is measured. The pressure and volume velocity along the tract are assumed to obey WEBSTER'S horn equation [Eq. (3.1)], which is valid for frequencies below about 4000 cps. An asymptotic high-frequency behavior of the tract is assumed. No assumptions are made about the termination at the glottal end or about the length of the tract. Solution of an integral equation yields the integral of the cross-sectional area of an equivalent lossless, hard-walled pipe as a function of distance. Differentiation gives the area function. Typical results, compared to area functions from X-ray measurements, are shown in Fig. 5.37. The impedance tube calculations are made for hard-walled vocal-tracts having the shapes given by the X-ray data.



(a) (b)

Fig. 5.37 a and b. Typical vocal-tract area functions (solid curves) determined from impedance measurements at the mouth. The actual area functions (dashed curves) are derived from X-ray data. (After GOPINATH and SONDHI)

A question of considered importance is the influence of wall-yielding (as is present in the real vocal tract) upon the calculated area function. Present efforts aim to include wall vibration and wall loss into the area determination method. Further research is needed to test the method with real speakers and real speech, and to account for real vocal-tract conditions, including loss, yielding side walls and nasal coupling.

Vocal-tract models, electrical vocal-tract analogs and computational analyses have all been useful in inferring articulatory data and tract dynamics from acoustic measurements of speech sounds and from X-ray data. One articulatory model, which has an important application in synthesis (see Section 6.26), has also been useful in establishing physiological constraints and time constants associated with major articulators (COKER, 1968). The articulatory model describes the vocal area function in terms of seven parameters, shown in Fig. 5.38. The coordinates are: the position of the tongue body, X, Y; the lip protrusion, L; the lip rounding W; the place and degree of tongue tip constriction, R and B; and the degree of velar coupling, N. No nasal tract is incorporated in this version of the model, and velar coupling exerts its influence solely through the tract area function.

The area function described by the model can be used to synthesize connected speech, which in turn can be compared in spectral detail to real speech. Also, because of its correspondence to major vocal elements, the seven-parameter model can be used to duplicate articulatory motions observed from X-ray motion pictures. Further, its description of vocal-tract area can be compared with X-ray area data, as shown in Fig. 5.39. Such comparisons have been useful in analyzing priorities

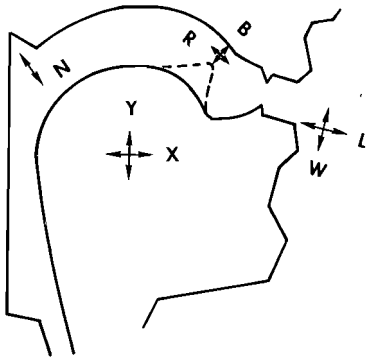


Fig. 5.38. Seven-parameter articulatory model of the vocal tract. (After COKER)

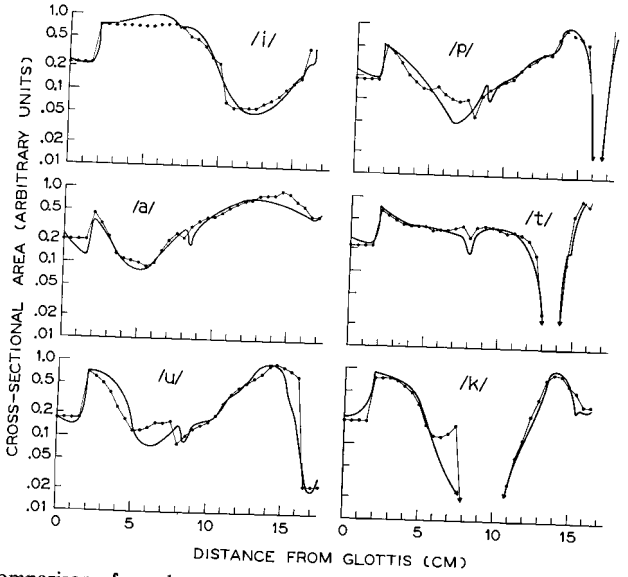


Fig. 5.39. Comparison of vocal tract area functions generated by the articulatory model of Fig. 5.38 and human area data measured from X-rays. (After COKER)

and time-constants for the motions of the articulators in real speech and in quantifying these effects for speech synthesis (COKER, UMEDA and BROWMAN; FLANAGAN, COKER, RABINER, SCHAFER and UMEDA).

5.5. Automatic Recognition of Speech

A human can listen to meaningful speech of a given language and set down a written equivalent of what he hears. He performs a transformation on the acoustic input signal wherein distinctive linguistic

elements (phonemes) are recognized and re-encoded into a sequence of letter symbols. Recognition of the linguistic elements is based upon a knowledge of the contextual, grammatical and semantic constraints of the given language. It does not take much examination of sound spectrograms to convince oneself that a unique relation generally does not exist between a given segment of the acoustic signal and a linguistic element. Neither are phonemic boundaries necessarily apparent in the acoustic signal.

Automatic recognition of speech implies phonemic analysis by machine. It is possible to simulate crudely the initial operations performed on the acoustic signal by the human (see the frequency analysis and neural encoding performed at the ear's periphery in Chapter IV) but, to date, not even the most elaborate mechanical recognizers have been able to apply linguistic constraints comparable in effectiveness to the human. This latter area represents an active field of research in theory of grammar, semantics, and mechanical translation.

The difference (or, more precisely, the gulf) between phoneme recognition for a given language and a straight-forward encoding of the acoustic signal, say in terms of vocal modes and excitation, cannot be overemphasized. The former implies complete linguistic knowledge, the latter only that the signal is produced by the human vocal mechanism. The latter is within the scope of present speech analysis techniques. The former, as yet, is not. If phoneme recognition ultimately proves possible, the import to efficient transmission is, of course, immense. (Recall it was suggested in Section 1.2, Chapter I, that the information rate associated with the utterance of independent, equiprobable phonemes is on the order of 50 bits/sec. A coding exists for transmitting information at this rate over a channel of about 5 cps bandwidth and 30 db signal-to-noise ratio, with as small an error as desired.)

A number of research investigations have treated machines which are capable of recognizing limited ensembles of speech sounds uttered by limited numbers of speakers (often only one). Generally these devices make decisions about either the short-time spectrum of the acoustic signal or about features of the time waveform. The constraints usually employed are ones more appropriate to the vocal mechanism (i.e., acoustical constraints) than to linguistic structure. Without attempting to be exhaustive, the state of the art can be outlined by several examples.

One effort toward a recognizer for a limited ensemble of sounds is a recognizer for spoken digits, called Audrey (DAVIS, BIDDULPH and BALASHEK). The principle of operation is to make a rough measure of the first and second formant frequencies as functions of time, and to compare the measured temporal patterns (in the $F1$–$F2$ plane) with a

set of stored reference patterns. The stored pattern affording the best correlation is then chosen as the uttered digit.

The procedure is illustrated in Fig. 5.40. The speech signal is filtered into two bands, 900 cps low pass and 1000 cps high pass. Limiting amplifiers in both channels peak clip the signals. Axis-crossing measures approximate the frequencies of the first and second formants as

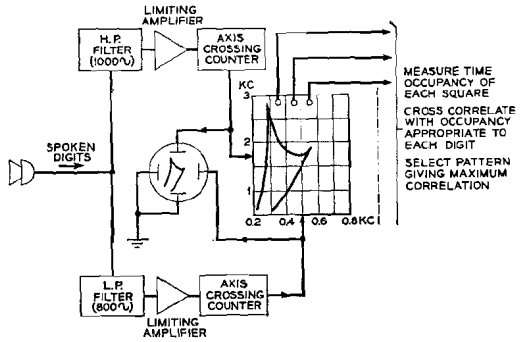


Fig. 5.40. Principle of operation of a spoken digit recognizer. (After DAVIS, BIDDULPH and BALASHEK)

functions of time. The first-formant frequency range (from 200 to 800 cps) is quantized into six 100-cps segments. The second-formant range (from 500 to 2500 cps) is quantized into five 500-cps steps. An $F1$–$F2$ plane with 30 matrix elements is thereby produced. For a given digit utterance, the time that the $F1$–$F2$ trajectory occupies each elemental square is determined.

A reference "time-occupancy" pattern for each digit is stored in the machine. The storage mechanism is 10 weighting resistors associated with each square. Through these resistors, charges are accumulated on 10 separate condensers during the time the square is occupied. A cross correlation of the stored and incoming patterns is effected by weighting the 10 conductances associated with each square according to the average time-occupancy of that square by the respective digits. That is, for each of the 30 squares, there are 10 relays which close charging paths to the 10 fixed condensers. The conductance of a given path is weighted proportional to the time occupancy of that square by a given digit. The condenser left with the greatest charge at the end of the utterance indicates the pattern affording the highest correlation, and hence the spoken digit.

The machine does not have provisions for automatically adjusting its stored patterns to a given speaker's voice. This must be done manually.

When it is done, however, the accuracy in recognizing telephone quality utterances of the digits ranges between 97 and 99% correct.

An extension of this technique is to correlate—on an instant-by-instant basis—a measured short-time amplitude spectrum with stored spectral patterns (DUDLEY and BALASHEK). Instead of the $F1$–$F2$ trackers, a set of bandpass filters (10 in this case, each 300 cps wide) is used to produce a short-time spectrum. Stored spectral patterns (again, 10) are continuously cross-correlated with the short-time spectrum produced by the filters. The maximum correlation is taken as an indication of the particular speech sound being produced. The pattern-matching procedure is illustrated in Fig. 5.41. If $F_0(\omega_n)$ is the short-time amplitude spectrum produced by the n filter channels for a given speech input, and $F_j(\omega_n)$ the j-th stored pattern, the circuit, in principle, approximates the correlation quantity

$$\varphi_{0j}(0)=\frac{1}{\Omega}\int_0^{\Omega} F_0(\omega)\,F_j(\omega)\,d\omega \qquad j=1,2,3,\ldots,$$

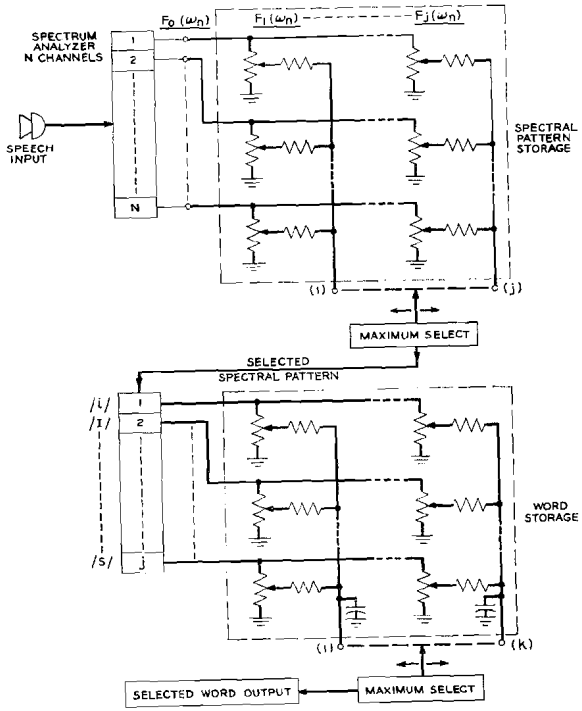


Fig. 5.41. Scheme for automatic recognition of spectral patterns and spoken digits. (After DUDLEY and BALASHEK)

by

$$\varphi_{0j}(0) \cong \frac{1}{n} \sum_n F_0(\omega_n) F_j(\omega_n) \qquad j = 1, 2, 3, \ldots,$$

and selects the j that produces a maximum $\varphi_{0j}(0)$. The 10 sound patterns stored in this particular development are all continuants and are /i, I, ɛ, a, o, u, n, r, f, s/.

A word recognizing device follows the spectral pattern recognizer to recognize the 10 digits. Similar to the Audrey device, each selected spectral pattern is weighted according to its duration in a given digit (see the lower part of Fig. 5.41). Again a maximum selection is made to recognize the uttered digit. The word indication is developed as follows. When a particular spectral pattern is energized, 10 charge paths are set up to 10 fixed condensers. The conductance of a given path is proportional to the average time for which that spectral pattern appears in a given digit. The 10 condensers therefore accumulate charges proprotional to the correlation between the 10 stored word patterns and the measured pattern. At the end of the utterance, a maximum selection indicates the best-fitting word. This device—designed as an elaboration upon the previous one—provides digit recognition with good accuracy when set for a particular voice. In both devices the sequence of spectral patterns and the recognized digits are displayed on electrical panel lights. Despite its early date of conception and implementation, this device and the previously-described digit recognizer, Audrey, still reflect present limitations in automatic speech recognition; namely, one can achieve success if the vocabulary is isolated words, sufficiently small in number, and if the number of speakers is sufficiently constrained.

Another speech recognizing device also compares spectral patterns with stored patterns representative of specific speech phonemes (Fry and Denes). The comparison however, is made in a different way, and the machine types out the identification in terms of special symbols. Selection of a match is asynchronous and is initiated by the rate of change of the spectral patterns. More important, however, an attempt is made to exploit elementary linguistic constraints. A block diagram of the device is shown in Fig. 5.42.

A filter-bank analyzer (20 channels) produces a short-time amplitude spectrum. Spectral patterns appropriate to a given sound are produced by multiplying the outputs of two channels. The products are scanned by a selector, and the maximum is chosen. The choice is typed out by the machine and is remembered by a storage circuit. On the basis of the choice, the ensemble of stored patterns is biased according to digram statistics for the language. Selection of the next phoneme is biased in favor of its being the most probable one to follow the previous choice.

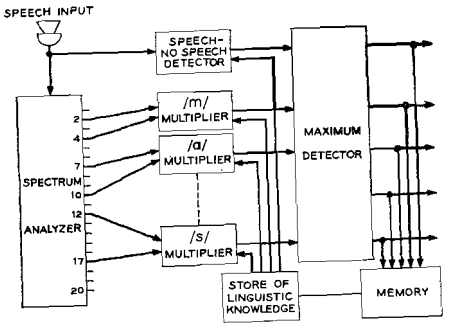


Fig. 5.42. Block diagram of speech sound recognizer employing elementary linguistic constraints. (After Fry and Denes)

In the present machine 14 phonemes are recognized; four vowels, nine consonants and silence. A new selection is made whenever the product voltages have a rate of change greater than a given threshold value. With the machine adjusted for a given speaker, the spoken input and printed output have been compared. When the digram constraints are not used, the percentage correct response on individual sounds and on words is 60% and 24%, respectively. When the digram constraints are connected, these same scores rise to 72% and 44% for the single speaker. For a second and third speaker, without readjusting the machine, the sound articulation scores fall to about 45%.

The linguistic information clearly improves the recognition when scored to give all phonemes equal weight. If scored on the basis of information per phoneme, however, the digram constraints could, under certain conditions, be detrimental. The most probable phoneme is favored, but it is also the conveyor of the least information. The constraints also raise the question of sequential errors and how they might be propagated. A certain level of accuracy in the acoustic recognition is certainly necessary if the use of linguistic constraints is to lead to a decrease, rather than to an increase, in error rate. Sequential errors of course occur in the human listener. A listener, once embarked upon the wrong set of constraints in a particular sequence, may add one error to another for quite a long stretch. In the machine, severe restriction of vocabulary reduces this possibility.

If the linguistic constraints to be incorporated into the recognition process are at all realistic, the storage and processing functions become complex. Also if elaborate processings are to be carried out on the acoustic signal, large storage and rapid computation are requisite. The digital computer is adept at this, and a number of efforts have been

made to capitalize upon its ability. One effort in this direction is the programming of a digit recognizer (DENES and MATHEWS). Short-time amplitude spectra are produced from a filter bank. The filter outputs are scanned sequentially, and the spectral data are read into and stored in the machine. A speech spectrogram—quantized in time, frequency and intensity—is laid down in the storage. Amplitude values are normalized so that the sum of the squares over all time-frequency blocks is unity. The measured time-frequency-intensity pattern is then cross-correlated with stored spectrographic patterns. The correlation is effected by multiplying the amplitude values of corresponding time-frequency elements and summing the products over all elements of the time-frequency plane. The stored pattern yielding the maximum correlation is chosen.

Provisions are made to time-normalize the data if desired. The beginning and the end of the digit utterance are located, and the data are, in effect, stretched to fit a standard time duration (actually 60 scans of the filter bank at $70\,\mathrm{sec}^{-1}$). Without time normalization only the beginning of each utterance is located, and the first 60 scans are used.

The reference pattern for each digit is obtained by averaging the spectral data for three utterances of that digit by five men. These patterns are used to recognize different utterances by the same and by different speakers. For different utterances by the same five speakers, the error rates are found to be 6% with time normalization and 13% without. When the reference patterns are set for a single speaker, the digits uttered by that speaker are recognized essentially with no error.

A more linguistically-based approach, using a large on-line computer facility, performs a feature analysis of segments of the speech waveform (REDDY, 1967). The wave is first divided into minimal segments, 10-msec in duration. Minimal segments which are acoustically similar are grouped to form larger segments representing either sustained parts or transitional parts. Features such as voiced-unvoiced, pitch, intensity, formant frequency and amplitude are used to classify each segment into four phoneme groups: stop, fricative, nasal-liquid and vowel. A very detailed algorithm is then used to assign a phoneme label to each segment of a phoneme group. The object, literally, is a speech to phoneme-like translation. This system, while recognizing the potential advantages of phonetic feature classification and language element probabilities, is nevertheless faced with the same problems of linguistic and semantic constraints that confront all recognizers. Its sophistication pays off, however, in enlarging the speaker population and vocabularly which can be successfully handled. The system has been demonstrated to yield 98% correct recognition on 500 isolated words spoken by one individual (REDDY, 1969).

At least one similar word-recognition experiment has been carried out for the Russian language (VELICHKO and ZAGORUYKO). In this case the energy-time-frequency dimensions of individually spoken words are quantized. A distance functional between the unknown word and the stored references for a word library of 203 words is computed. For two speakers, producing approximately 5000 utterances chosen from the 203 word library, the recognition accuracy was found to be about 95%. Computation time for each utterance was 2 to 4 sec.

The preceding discussion has attempted to indicate by example several stages of development in automatic speech recognition. A sizeable number of related efforts have not been mentioned (for example, SMITH, 1951; BAUMANN, LICKLIDER and HOWLAND; OLSON and BELAR; FORGIE and FORGIE; FRICK; DREYFUS-GRAF; MARTIN; LINDGREN). Most share a common point of departure, namely, the short-time spectrum. It is clear from the discussion that none of the schemes tells us very much about how the human processes speech information, nor about how he recognizes linguistic elements. None of the methods works well on an unrestricted number of voices, nor on a large contextual vocabulary. The human, however, is proficient at handling both. Nevertheless, the investigations do indicate what can be realized in the way of voice-actuated devices for special applications—specifically, applications where vocabulary and number of voices may be suitably restricted. It is clear, too, that for a given accuracy of recognition, a trade can be made between the necessary linguistic constraints, the complexity of the vocabulary, and the number of speakers.

Automatic speech recognition—as the human accomplishes it—will probably be possible only through the proper analysis and application of grammatical, contextual, and semantic constraints. These constraints, as yet, are largely unknown. Perhaps not surprisingly, research in speech synthesis seems to be providing more insight into linguistic constraints than is speech recognition work. One view of speech recognition (PIERCE) makes the point that success will be very limited until the recognizing device understands what is being said with something of the facility of a native speaker.

5.6. Automatic Recognition and Verification of Speakers

The previous discussion pointed up the notion that the spectral patterns of one speaker are not always adequate to recognize the speech of another. This fact suggest that spectral data might be used to recognize or identify different speakers. A number of efforts along these lines have been made—mainly with the use of digital computers. By

way of illustration, one study produced quantized time-frequency-intensity (spectrographic) patterns from a 17-channel filter bank scanned at a rate of 100 sec^{-1} (PRUZANSKY). Ten key words were excerpted from context for 10 different speakers (three women, seven men). For each talker, three utterances of the 10 key words were used to establish the reference patterns for that individual.

For talker identification, the spectrographic pattern of a different key-word utterance by an unknown speaker of the ten-member group was cross-correlated with the reference patterns (again by multiplying amplitudes at each time-frequency element of the spectrogram), and the maximum correlation was taken. Because the utterances varied in length, alignment of patterns was done by matching them at the maximum overall amplitude points. Results showed that among the 10 speakers for whom the reference library was formed, the identification was correct in 89% of the cases.

In the same study, the three dimensional time-frequency-intensity patterns were reduced to two dimensions by summing over the time of the utterance for each filter channel. The summation produces a graph of integrated intensity-versus-frequency for each utterance. It was found that this operation still afforded a recognition score of 89%.

It is of course difficult to draw conclusions about human recognition of speakers from such an experiment. Again, however, for a limited, specific application, where speaker ensemble and vocabulary are restricted, such a technique could be effectively applied.

A few experiments have measured human recognition of speakers from visual inspection of speech spectrograms. In one of these (KERSTA, 1948, 1962a) a group of speakers (either 5, 9 or 12) was asked to utter 10 key words four times. Conventional bar spectrograms and contour spectrograms were made of their utterances (see Section 5.14). For each word a randomized matrix of spectrograms consisting of four utterances of each speaker was displayed. Subjects were asked to identify the utterances of each individual speaker. The errors in grouping the prints according to speaker ranged from 0.35% to 1.0% for bar prints and from 0.37% to 1.5% for contour spectrograms. When the test words were excerpted from context, the error was still about the same order of magnitude.

A second experiment was modeled after fingerprint identification procedures, although the analogy is a tenous one. A file of "voice prints" of five key words was compiled for 12 speakers. Subjects then identified a different set of utterances by an unknown member of the group through comparisons to the reference sets. Using the groups of five cue words, the misidentifications were less than 1%. Identifications based upon two 5-word groups in tandem gave errors of about one-half

percent. Preliminary investigations were also made into the ability to recognize disguised voices. The results suggest that adults have certain invariant linguistic and physiological characteristics which the spectrograph may display even when an effort is made to alter the voice.

These experiments, through a combination of publicity and private development, captured the notice of various law-enforcing organizations, who saw in the method a new means for identifying criminals. Several efforts were made to introduce the technique into legal proceedings with controversial results. Independent experiments were conducted to test the method, and the findings were at variance with the original experiments (YOUNG and CAMPBELL). Most opinion holds that more research is needed to accurately establish the utility of human recognition of speakers from sound spectrograms (FLANAGAN *et al.*; BOLT *et al.*). Subsequent efforts continue in this direction (TOSI). These latter experiments have treated a variety of experimental conditions (for example, closed sets versus open sets) and the error rates in visual identification vary from 1% to 30%, depending upon the experimental constraints. This error range, when analyzed in terms of experimental conditions, appears consistent with previous data.

A problem perhaps more interesting and presently more tractable than speaker recognition is automatic verification of speakers (DODDINGTON; LUMMIS; DAS and MOHN). In the usual context of this problem one has a restricted population of "customers" who want to be verified (i.e., a cooperative situation), and they are willing to state a prearranged phrase (secret if desired) chosen to be advantageous for the machine. (The voice banking, and voice validation of credit cards are applications in point). In the verification situation unknown caller, x, claims to be customer, C_i. The machine must decide to accept or reject x as C_i. The decision can be weighted according to the importance of the verification (for example, whether the sum charged is large or small) and a predetermined mix of error types (i.e., rejecting a true speaker versus accepting a false speaker) can be specified.

The most important aspect of the verification problem, and the one which distinguishes it from the recognition problem, is that no matter what the size of the impostor population, the average percent correct verification tends to be constant. The performance is determined by the average consistencies of the known speakers and by how each of them differs from the average of the impostor population. In a recognition situation, on the other hand, where the unknown must be identified by successive comparisons to all members of a known set, the probability of error is monotonely related to the number of speakers in the set, and the probability of a recognition error approaches unity as the user population becomes large.

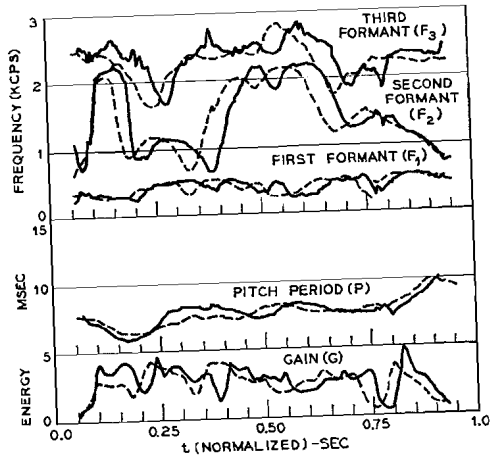
One experiment on verification (DODDINGTON) has made use of pitch, formant and intensity data to form reference patterns for the known speakers. Frequency data (i.e., formants and pitch) were considered attractive because they are resistant to variations in the amplitude-frequency characteristics of a voice communication link. A novel non-linear time-warping of the utterance of an unknown speaker was used to compare (register) it with a stored reference pattern corresponding to the claimed identity. The non-linear warp was achieved on a digital computer by a steepest-ascent algorithm. The algorithm warped the pattern of the unknown speaker to maximize its correlation with the stored reference pattern. A mean square error measure was then made for the registered patterns and the speaker was accepted or rejected depending upon whether the mean square error was less than or greater than a threshold chosen for a specified mix of errors (i.e., reject true versus accept false).

Fig. 5.43 shows how the formant, pitch and intensity (gain) data are compared for a verification phrase; namely, the voiced sentence "We were away a year ago". In Fig. 5.43a the unknown utterance (solid curve) has been given a linear time stretch to make its duration equal to the reference (dashed curve). Poor internal registration is evident. In Fig. 5.44b, the non-linear warp has been applied to register the second formant tracks with maximum correlation. The registration of the other parameters is similarly improved. The remaining differences and the amount of non-linear warp applied are indicative of the similarities of the two patterns. A square error measure is formulated to indicate a "distance" between the registered patterns.

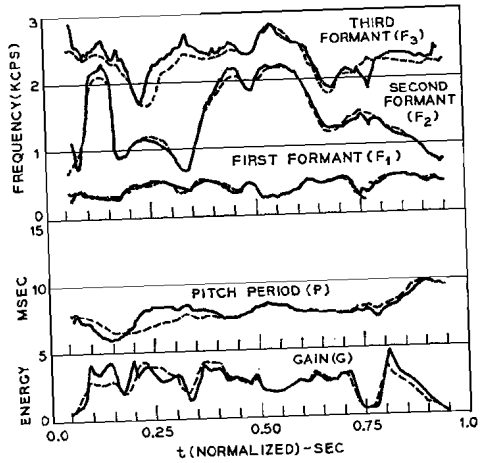
Using this technique, with a population of 40 male speakers, correct verification was achieved 98.5% of the time on the verification phrase "We were away a year ago" used by all subjects. Identical twins included in the experiment were differentiated 100% of the time.

If more sophisticated "distance measures" are used to characterize the differences between the registered patterns for the unknown and reference, a comparable performance can be obtained on simple measures, easily made in real time. A subsequent experiment on the same population of 40 speakers, and using more elaborate distance measures on only intensity, pitch and non-linear warp, achieved 99% correct verification (LUMMIS).

A natural query is "How good would human listeners do in the same task?" To answer this, a completely parallel auditory experiment was conducted with the same 40 speakers, but using human listeners instead of a computer to make the verification decision. The listeners performed with greater error rate than the machine and achieved approximately 96% correct verification (ROSENBERG, 1971 a).



(a)



(b)

Fig. 5.43 a and b. Effects of nonlinear warp in registering speech parameter patterns. The dashed curves are reference data for an individual. The solid curves are a sample utterance from the same individual. (a) Linear stretch to align end points only. (b) Nonlinear warp to maximize the correlation of the *F*2 patterns. (After DODDINGTON)

Results of these and related verification experiments suggest that automatic machine verification may have practical value. An obvious and further question is how easily might accomplished mimics deceive the machine and be erroneously accepted. Continuing research is aimed at this question.

A number of features seem to distinguish one speaker from another. The size and shape of the vocal tract vary considerably among persons. Characteristic damping, mouth and glottal dimensions also vary. Individual nasal coupling, size and damping of the nasal tract are other relevant features. Temporal patterns of intensity (stress) and pitch (inflection) are still others. Vocal obstructions and variations in dental work may contribute still further differences. Some or all these factors might be used to recognize or verify a speaker. It is probable that machine and human do not use the same features to equal effect. The machine, for example, might make use of data the human ear cannot assimilate.

As suggested earlier, the speech-recognition and speaker-identification experiments described here tell us little about the perceptual processing which the human accomplishes. They do not, for example, suggest the temporal span of the recognition unit used by the human. Neither do they indicate subjective techniques for measuring whether the unit is the phoneme, word, sentence, or something larger. The automatic machine methods deal mainly with advantageous processings of essentially the acoustic signal, and not with perception as the human practices it.

The mechanism of human perception of speech is difficult to analyze and present understanding is meager. The discussion of Chapter IV showed that for signals with simple temporal and spectral structure, reasonably close correlations can be made between subjective behavior and the known physiology of the peripheral ear. To a modest extent, similar relations can be established for speech signals. (For example, one can identify features such as voice pitch, formant frequency and voiced-unvoiced excitation in terms of the basilar membrane motion.) But how the neural data are stored and processed after leaving the periphery is a completely open question. Continued research on the electrophysiology of the auditory tract, and on human response to meaningful speech signals, will hopefully provide some of the answers.

VI. Speech Synthesis

Ancient man often took his ability of speech as a symbol of divine origin. Not unnaturally, he sometimes ascribed the same ability to his gods. Pagan priests, eager to fulfill great expectations, frequently tried to make their idols speak directly to the people. Talking statues, miraculous voices and oracles were well known in the Greek and Roman civilizations – the voice usually coming to the artificial mouth via cleverly concealed speaking tubes. Throughout early times the capacity of

"artificial speech" to amaze, amuse and influence its listeners was remarkably well appreciated and exploited.

As the civilized world entered the Renaissance scientific curiosity developed and expanded. Man began to inquire more seriously into the nature of things. Human life and physiological functions were fair targets of study, and the physiological mechanism of speech belonged in this sphere. Not surprisingly, the relatively complex vocal mechanism was often considered in terms of more tractable models. These early models were invariably mechanical contrivances, and some were exceedingly clever in design.

6.1. Mechanical Speaking Machines; Historical Efforts

One of the earliest documented efforts at speech synthesis was by KRATZENSTEIN in 1779. The Imperial Academy of St. Petersburg offered its annual prize for explaining the physiological differences between five vowels, and for making apparatus to produce them artificially. As the winning solution, KRATZENSTEIN constructed acoustic resonators similar in shape to the human vocal tract. He activated the resonators with vibrating reeds which, in a manner analogous to the human vocal cords, interrupted an air stream.

A few years later (1791), VON KEMPELEN constructed and demonstrated a more elaborate machine for generating connected utterances. [Apparently VON KEMPELEN'S efforts antedate KRATZENSTEIN'S, since VON KEMPELEN purportedly began work on his device in 1769 (VON KEMPELEN; DUDLEY and TARNÓCZY).] Although his machine received considerable publicity, it was not taken as seriously as it should have been. VON KEMPELEN had earlier perpetrated a deception in the form of a mechanical chess-playing machine. The main "mechanism" of the machine was a concealed, legless man – an expert chess player.

The speaking machine, however, was a completely legitimate device. It used a bellows to supply air to a reed which, in turn, excited a single, hand-varied resonator for producing voiced sounds. Consonants, including nasals, were simulated by four separate constricted passages, controlled by the fingers of the other hand. An improved version of the machine was built from VON KEMPELEN'S description by Sir CHARLES WHEATSTONE (of the Wheatstone Bridge, and who is credited in Britain with the invention of the telegraph). It is shown in Fig. 6.1.

Briefly, the device was operated in the following manner. The right arm rested on the main bellows and expelled air through a vibrating reed to produce voiced sounds. (See the lower diagram in Fig. 6.1.) The fingers of the right hand controlled the air passages for the fricatives /ʃ/ and /s/, as well as the "nostril" openings and the reed on-off control.