Chapter 10

Devices for Spread-Spectrum Communications

In digital communications systems, the term "spread spectrum" refers to a system in which the bandwidth of the transmitted signal is much greater than the bandwidth of the data being transmitted. This feature gives a number of advantages over a conventional system, notably that it provides security and is less sensitive to interference from unwanted signals. Surface-wave devices have been developed for a variety of applications in this area, in particular for correlation of waveforms with large time-bandwidth products, that is, for matched filtering. We have already seen in Chapter 9 that surface-wave chirp filters are used as matched filters in radar systems. The applications in spread-spectrum systems are similar in some respects, but there are some important differences. For example, the waveforms involved are usually phase-shift-keyed (PSK) waveforms rather than chirps, and there is a strong emphasis on programmability, that is, the ability to change the device response electronically.

Another important distinction here lies in the basic physical principles of many of the devices. In previous chapters, nearly all of the devices considered relied on interdigital transducers, multi-strip couplers or reflective groove arrays. In this chapter some new principles are introduced. For example the acoustic convolvers of Section 10.3 exploit the weak acoustic non-linearity of lithium niobate to mix two surface-wave beams, and thus rely on an effect that is avoided in other devices. This gives a very versatile and quite simple method for correlating complex waveforms. Other types of convolver use non-linear interactions in *semiconductors*, using a semiconductor in close proximity to the surface of a piezoelectric and thus exploiting the electric field accompanying the surface wave. Alternatively, the surface wave may propagate on a semiconductor substrate, with a piezoelectric film to generate an electric field. The use of semiconductors in connection with surface waves also introduces a range of other possibilities, including the ability to store an acoustic signal and thus provide an acoustic memory, as described in Section 10.4.2. In fact, semiconductor interactions open up the possibility of an "acoustic chip", in which integrated circuitry and surface-wave devices are combined on the same semiconductor substrate. Some preliminary devices of this type are mentioned in Section 10.2.1. The performance to date has been rather limited owing to

technological difficulties but, in view of the wide range of possibilities such chips would offer, there could well be substantial future developments in this area.

Section 10.1 below gives a brief introduction to the principles of spread-spectrum systems. Section 10.2 is concerned with several linear surface-wave devices, including fixed and programmable matched filters for PSK waveforms. The acoustic convolver is introduced in Section 10.3, which includes devices using surface-wave waveguides in order to improve the efficiency of the non-linear interaction. Section 10.4 is mainly concerned with non-linear semiconductor devices, including convolvers and storage devices. Finally, Section 10.5 gives a brief description of surface-wave oscillators which are included here for convenience, though their applications are not of course limited to spread-spectrum systems.

10.1 PRINCIPLES OF SPREAD-SPECTRUM SYSTEMS

Spread-spectrum communication systems occur in a very wide variety of forms, as described for example in References [349–351]. In this section we review the basic principles briefly. The applications of surface-wave devices in these systems will be described in subsequent sections, and are also discussed in References [352–354].

A common type of spread-spectrum transmitter is illustrated in Figure 10.1. The input data is taken to be a stream of binary digits, each of length T, and these are applied to a balanced modulator. A C.W. waveform is also applied, and the modulator either passes this directly or inverts it, depending on whether the current input digit is a "one" or a "zero". The modulator output is thus a *phase-shift keyed*, or PSK, waveform, with relative phase 0 or 180° corresponding to the data. In a

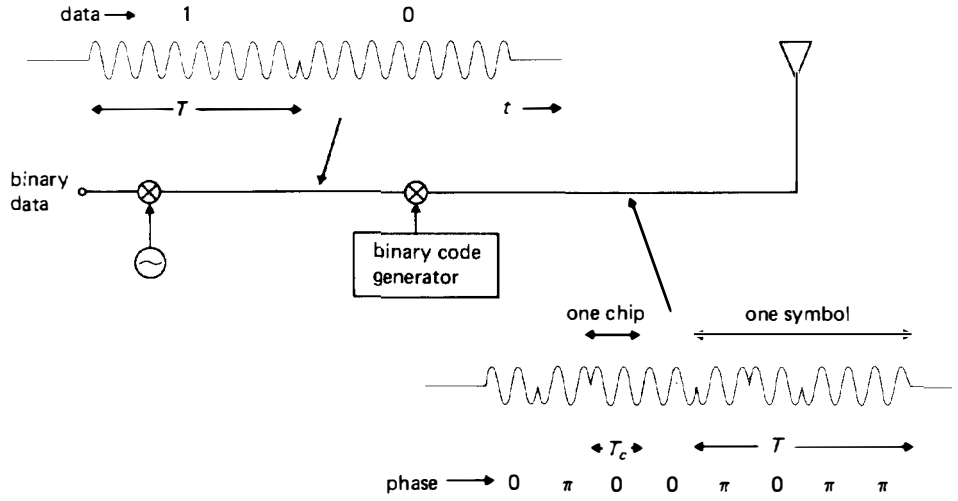


FIGURE 10.1. Spread-spectrum transmitter using PSK waveform.

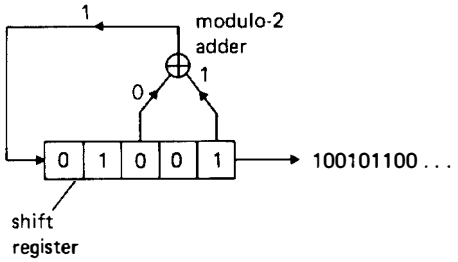


FIGURE 10.2. Pseudo-noise code generator. For the case shown the code repeats after 31 digits.

conventional communication system a waveform of this type would be transmitted, after amplification and up-conversion. In a spread-spectrum system an additional stage of bi-phase modulation is employed, using a binary code generated by a code generator. This second stage introduces phase changes much more frequently than the original data, with the consequence that the signal bandwidth is substantially increased. The waveform corresponding to one data digit, of length T, is called a "symbol" and can be regarded as a sequence of contiguous pulses, each of length T_c and phase 0 or 180°, known as "chips". The spectrum of the output waveform depends in a complex manner on the coding employed; however it is usual to choose the code sequence in a quasi-random manner, and it is then found that the power spectrum of the output has a shape similar to that of an individual chip, and hence its bandwidth is approximately $1/T_c$. In contrast, the waveform prior to code modulation has a bandwidth of approximately $1/T$. Hence the bandwidth has been increased by a factor T/T_c, equal to the number of chips in each symbol. Typically, each symbol will contain 50 to 500 chips. Symbol lengths vary widely according to the application, and are generally in excess of 10μsec.

The code to be used may be simply stored in a digital memory and read out as required, but a convenient alternative is to use a digital shift register as shown in Figure 10.2. Here two of the register stages are combined by a modulo-2 adder to provide a feedback to the register input; the adder gives a "one" output if its two inputs are different and a "zero" if they are the same. As the register is clocked, a binary code emerges at the right. Provided an appropriate feedback logic is used, a register with n stages can give $2^n - 1$ output digits before repeating, and the code is then called a "pseudo-noise" code. Such a code is found to have quasi-random properties, and is thus effective for spreading the bandwidth of the signal. Furthermore, the number of stages required in the shift register is generally much less than the number of digits in one cycle of the code.

In the receiver the phase changes produced by the code generator must be removed in order to extract the data, and Figure 10.3 shows a receiver using a matched filter for this purpose. The matched filter is the commonest application for surface-wave technology in spread-spectrum systems. Suppose initially that the same code is used for each symbol, as in Figure 10.1, so that the symbols are all the same except that some of them are inverted. The matched filter (Section A.3) has an impulse response corresponding to the time-reverse of one symbol. For each input symbol the filter

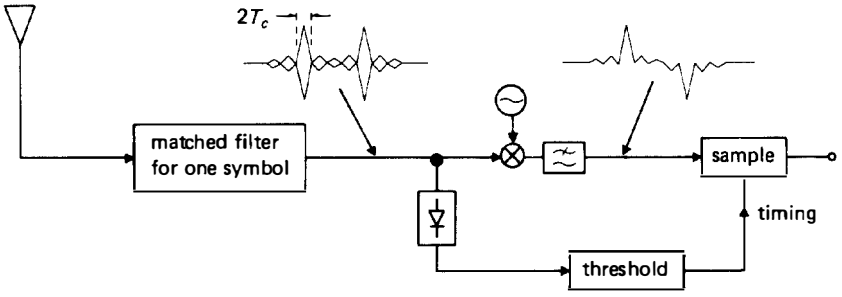


FIGURE 10.3. Spread-spectrum receiver using matched filter.

output typically gives a triangular correlation peak, with basewidth $2T_c$, and a series of relatively small time-sidelobes on either side. The relative phases of the peaks correspond to the data, and this is recovered by demodulating the filter output using a synchronous detector and sampling at the peak times; a comparator is used to determine the polarity of the sampler output, which corresponds to the original data. The required sampling times can be obtained by using an envelope detector followed by a threshold circuit.

Compared with a conventional communication system, a spread-spectrum system offers several advantages. It has increased security, since the signal can only be demodulated if the code is known and its wide bandwidth makes its presence more difficult to detect. The receiver responds preferentially to the waveform that it is matched to, and so is relatively insensitive to narrow-band interference. If the transmitter timing is known, the timing of the narrow correlation peak in the receiver can be used to give a high-resolution measure of the range between the transmitter and receiver. Also, the narrow correlation peak enables the system to reject multi-path signals with different delays, due to reflections from buildings for example. On the other hand, a spread-spectrum system does *not* give an improved signal-to-noise ratio when wide-band noise is present at the receiver input; the output signal-to-noise ratio of the matched filter is $2P^sT/N_i$, as shown by equation (9.1) of Section 9.1, and is therefore independent of the signal bandwidth. All of these features are in strong contrast with the advantages of matched filters in pulse compression radar systems, discussed in Section 9.1.

Before the advent of surface-wave devices, the implementation of matched filters placed heavy demands on existing technology. Consequently another type of receiver, shown in Figure 10.4, has been more commonly used. Here the signal is correlated by an *active correlator*, which multiplies it by a code identical to that used in the transmitter, thus converting each symbol to an unmodulated pulse of length T. The latter is down-converted to baseband and integrated over the symbol duration, and a polarity decision then gives the data. Ideally this receiver gives the same performance as the matched filter receiver described above, and moreover it is easier to implement. However, it has the significant drawback that the code generator must be accurately synchronised with the code of the received signal, typically with an accuracy of $T_c/4$. In general the signal timing will be initially unknown, so the receiver must first perform an acquisition procedure: trial correlations are done with the code timing

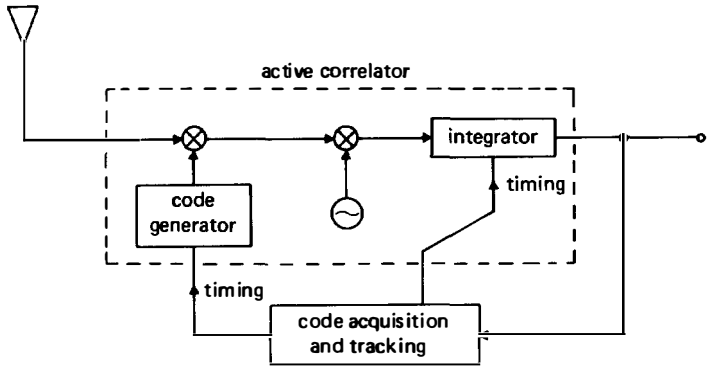


FIGURE 10.4. Spread-spectrum receiver using active correlator.

changed slightly after each symbol length, continuing until a correlation is observed at the output. Typically, the number of trials needed is about twice the number of chips in the symbol, so that the acquisition procedure can be quite lengthy. In contrast, the matched filter receiver does not require this lengthy acquisition procedure, and this is its main advantage.

It was assumed above that the same coding was used for each symbol, but in practice it is quite common to vary the coding in the interest of greater security. In the transmitter, this may be done by using a memory with a library of codes. This does not substantially complicate the operation of an active correlator receiver. For a matched filter receiver a bank of matched filters can be envisaged, but generally the number of codes used would make this approach inconvenient. Thus a *programmable* matched filter, in which the coding can be changed electronically, is preferable. It will be seen later that there are several surface-wave techniques for realising programmable matched filters. An alternative, and often used, method of varying the coding is simply to use a pseudo-noise code generator in the transmitter, with a repetition period much greater than the symbol length.

10.2. LINEAR DEVICES

We have seen above that the the main application for surface-wave devices in spread-spectrum systems lies in matched filtering of PSK waveforms. Section 10.2.1 below describes surface-wave devices for this purpose, including programmable devices. The most important limitations of these devices arise from phase errors due to velocity errors or temperature changes, and these are considered in Section 10.2.2. In addition to PSK waveforms a variety of other waveforms can be considered for spread-spectrum systems, and in Section 10.2.3 we consider surface-wave devices for use with MSK waveforms. Devices for frequency-hopped waveforms are considered in Section 10.2.4.

The devices described here are all linear. In Sections 10.3 and 10.4 below some

non-linear devices, particularly convolvers, will be described, and it will be shown that these offer a very different and effective approach for programmable matched filtering.

10.2.1. Matched Filters for PSK Waveforms

We first consider matched filters in which the coding is fixed. It was seen in Section 8.1 that an interdigital transducer can be designed simply by sampling the required impulse response, apart from the minor complication of the presence of the element factor. This principle can be applied to PSK filters, but has the disadvantage that electrode interaction effects are severe if a single-electrode design is used, since the impulse response is usually many wavelengths long [355]; also, this method would require the other transducer to have few electrodes, which could make it difficult to match efficiently if a wide bandwidth were required (Section 7.1.1). For these reasons most devices have used a modified configuration, in which the required impulse response is synthesised by the cooperative action of two transducers. As shown in Figure 10.5, one transducer is uniform with length vT_c, where v is the velocity. The other transducer is essentially an array of short uniform transducers, often called taps, all connected to two bus-bars. The taps have uniform spacing vT_c, and are all identical except that some are inverted in accordance with the required code; this inverts the corresponding contribution to the device output waveform. In response to a short impulse, the uniform transducer at the left generates a rectangular surface-wave pulse which propagates along the device, exciting the taps in sequence, so that a bi-phase PSK waveform is produced as shown in the figure. The principle is very similar to thinning (Section 8.3), but here the distortion due to thinning is compensated by the response of the uniform transducer. The device response is in practice distorted somewhat by the element factor, and by the fact that the taps have finite lengths, but these distortions are usually small enough to be acceptable for spread-spectrum applications.

The performance of these devices is reviewed by Bell *et al.* [356, 357], and further experimental results are given in References [352, 358–364]. The substrate material is usually quartz because good temperature stability is usually needed, as explained below. Most experimental devices had 30 to 150 taps with a chip rate ($1/T_c$) in the range 5 to 20 MHz, though a chip rate of 200 MHz has been demonstrated [364]. Typically, the devices were coded in accordance with pseudo-noise codes. A common experimental test of the device fidelity is to examine the output waveform when the

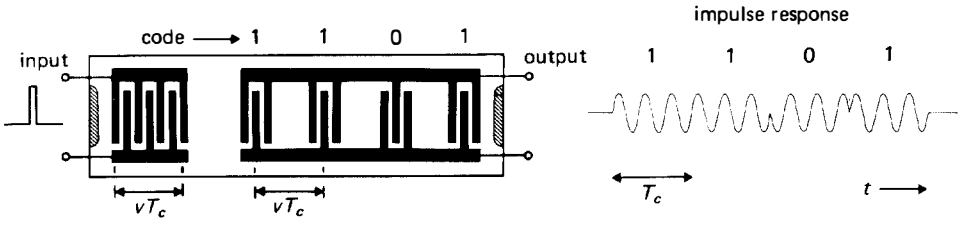


FIGURE 10.5. Fixed-coded matched filter for PSK.

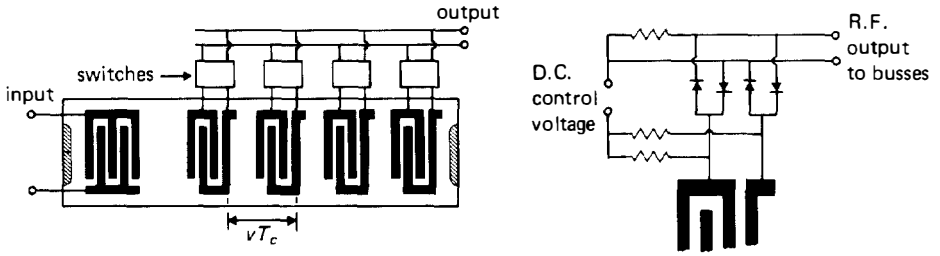


FIGURE 10.6. Programmable matched filter for PSK, with one type of switch circuit shown at right.

device is used to correlate the waveform that it is matched to; the relative levels of the correlation peak and time-sidelobes are sensitive to amplitude and phase errors in individual taps. For typical devices, the peak-to-sidelobe ratios were within 1 or 2 dB of the ideal values.

As already seen in Section 10.1, spread-spectrum systems often change the coding employed, so that a *programmable* matched filter would be required rather than the fixed-coded devices considered above. Programmability can be implemented by connecting the individual taps to a bank of electronic switches, as illustrated in Figure 10.6; each switch has a D.C. control input whose polarity determines the phase (0 or 180°) of the contribution from the corresponding tap. Arrangements of this type have been demonstrated by several authors [365, 366], using either hybrid or integrated circuitry; however, the need to accommodate the physical size of the circuit and the bonding pads for the wire interconnections implies a minimum tap spacing, corresponding to a minimum chip length T_c of typically 100 nsec.

To overcome this limitation, and the inconvenience of a large number of wire bonds, Hickernell *et al.* [367, 368] have developed a very different approach. Here the surface wave propagates on a *silicon* substrate, and is tapped by an array of field-effect transistors making use of the piezoresistive effect, that is, the change of resistivity accompanying an acoustic strain. The transistor bias can be used to control the amplitude and phase. The wave is launched by an interdigital transducer, with a piezoelectric zinc oxide overlay. This approach enables integrated circuitry to be fabricated in the *same* substrate, eliminating the need for individual wire bonds to each tap; Hickernell's device included a shift register, so that the code required could be read in serially, and a variety of other circuits. The device demonstrated programmable correlation of 31-chip waveforms with 10 MHz chip rate, showing that the approach is practically feasible even though the technology is rather demanding.

This type of technology, in which surface-wave and semiconductor devices are integrated on the same substrate, could well be very significant in the future development of surface-wave devices. It implies that the well known flexibility of integrated circuits can be combined directly, on the same substrate, with almost any type of surface-wave device. However, the technology needed is at present relatively immature, and consequently the results obtained to date are rather limited.

An alternative approach to integration is that of Hagon [369], who demonstrated a programmable PSK filter using a piezoelectric aluminium nitride film on a sapphire

substrate. The taps in this case were interdigital transducers, and the switches needed to control the phase were fabricated in a silicon film deposited on the same sapphire substrate. More recently, Grudkowski *et al.* [370] have shown that field-effect transistors on gallium arsenide may be used as programmable surface-wave taps, and clearly these could also be combined directly with integrated circuitry. These two technologies are further possibilities for fully integrated devices, though neither has to date been developed to the extent of the silicon device described above.

10.2.2. Output Waveform and Effect of Phase Errors

It is shown here that, when a PSK waveform is applied to an appropriate surface-wave matched filter, the output waveform is generally quite sensitive to errors in the surface-wave velocity and to temperature changes, and also to doppler shifts. For comparison, we first consider the ideal output waveform.

(a) Ideal Output Waveform. We consider a PSK waveform $s(t)$, with centre frequency ω_0, given by

$$s(t) = \sum_{n=1}^{N} a_n v_c(t - nT_c) \exp(j\omega_0 t) + \text{c.c.}, \tag{10.1}$$

where "c.c." indicates complex conjugate, N is the number of chips and $a_n = \pm 1$ are coefficients corresponding to the coding. The function $v_c(t)$ is the envelope of one chip, given by

$$v_c(t) = \text{rect}(t/T_c), \tag{10.2}$$

which is a rectangular pulse of duration T_c. The impulse response of the filter is taken to be

$$h(t) = \sum_{n=1}^{N} b_n v_c(t - nT_c) \exp(j\omega_0 t) + \text{c.c.}, \tag{10.3}$$

where $b_n = \pm 1$. Apart from a phase constant, which is inconsequential, the filter is matched to the waveform $s(t)$ if the coefficients b_n are given by

$$b_n = a_{N+1-n}. \tag{10.4}$$

We also define $r_c(t)$ as the correlation function of the chip envelope $v_c(t)$, so that

$$r_c(t) = \int_{-\infty}^{\infty} v_c(\tau) v_c(\tau - t)\, d\tau. \tag{10.5}$$

Using equation (10.2), this is found to be a triangular function of height T_c and basewidth $2T_c$.

When the waveform $s(t)$ is applied to the filter, the output waveform $g(t)$ is the convolution of $s(t)$ with $h(t)$, and $g(t)$ is the correlation function of $s(t)$ if the filter is matched. In practice, $s(t)$ and $h(t)$ can be taken to be bandpass functions, and it follows that it is sufficient to convolve the positive-frequency parts of $s(t)$ and $h(t)$ (the terms proportional to $\exp(j\omega_0 t)$), and then add the conjugate. With some

re-arrangement this gives

$$g(t) = \sum_{k=2}^{2N} c_k r_c(t - kT_c) \exp(j\omega_0 t) + \text{c.c.} \tag{10.6}$$

where the coefficients c_k are given by

$$c_k = \sum_{n=1}^{k-1} a_{k-n} b_n, \qquad \text{for } k \leqslant N + 1,$$

$$= \sum_{n=k-N}^{N} a_{k-n} b_n, \quad \text{for } k \geqslant N + 1. \tag{10.7}$$

Thus the output waveform is essentially a sum of delayed versions of the chip correlation function $r_c(t)$, with amplitudes proportional to c_k. Since $r_c(t)$ has duration $2T_c$, the envelope at times $t = kT_c$ is simply $2c_kT_c$. For a matched filter the correlation peak occurs at $k = N + 1$, and using equations (10.4) and (10.7) its amplitude is given by $c_{N+1} = N$. The amplitudes of the time-sidelobes, given by the other c_k, depend on the coding.

(b) Velocity Errors and Temperature Changes. We now suppose that the response of the surface-wave filter is somewhat different from ideal. The effects of velocity errors and temperature changes were considered in Section 6.4, in terms of a small quantity ε defined as the fractional change of delay, as in equations (6.35) or (6.39). The actual impulse response $h'(t)$ of the device is related to the ideal response $h(t)$ by

$$h'(t) = h\left(\frac{t}{1 + \varepsilon}\right) \approx h[t(1 - \varepsilon)]. \tag{10.8}$$

This is substituted into equation (10.3) to find $h'(t)$. It can be assumed that ε is too small to have any significant effect on the envelope, and it follows that the only effect of the error is to multiply the waveform by a phase term $\exp(-j\omega_0\varepsilon t)$. To a good approximation this term can be taken to be constant over the duration of each chip, so that the impulse response becomes

$$h'(t) \approx \sum_{n=1}^{N} b'_n v_c(t - nT_c) \exp(j\omega_0 t) + \text{c.c.}, \tag{10.9}$$

where $b'_n = b_n \exp(-j\omega_0 nT_c\varepsilon)$. This is the same as the ideal response, equation (10.3), except that b_n has been replaced by b'_n. The filter output waveform can therefore be obtained from the earlier analysis for the ideal case, equation (10.6). Thus the amplitudes of the correlation peak and time-sidelobes are given by the c_k of equation (10.7), with b_n replaced by b'_n. For the correlation peak $k = N + 1$ and equation (10.7) is readily summed as a geometric progression, so that the peak

amplitude is found to be given by

$$|c_{N+1}| = \left| \frac{\sin(\pi M\varepsilon)}{\sin(\pi M\varepsilon/N)} \right|, \tag{10.10}$$

where $M = N\omega_0 T_c/(2\pi)$ is the number of cycles in the waveform. The peak amplitude is thus zero for $\varepsilon = \pm 1/M$. A typical requirement is that it should not fall by more than 1 dB, and for $N \gg 1$, which is usually the case, this implies $|\varepsilon| < 0.26/M$. This is generally more stringent than the requirement for linear chirp filters, considered in Section 9.5.2.

Carr *et al.* [360] give examples for a variety of codes, showing that the error affects the sidelobe amplitudes as well as the correlation peak, though the sidelobe changes are not generally significant if the reduction of the peak is acceptable. The analysis for the temperature sensitivity has been confirmed experimentally for *Y*, *X* quartz substrates [360] and for *ST*, *X* quartz substrates [357].

(c) Doppler Shifts. The effect of a doppler shift of the input waveform can be found in a similar way. As stated in Section 9.5.3, if $S(\omega)$ is the spectrum of the ideal input waveform, the doppler-shifted waveform can be taken to have a spectrum $S'(\omega) = S(\omega - \omega_d)$, where ω_d is the doppler frequency. Using the shifting theorem, equation (A.11), it is found that the doppler-shifted waveform $s'(t)$ is given by the ideal waveform $s(t)$ of equation (10.1), with ω_0 replaced by $\omega_0 + \omega_d$. There is therefore a phase error $\omega_d t$, and if this is taken to vary little over the duration of one chip we have

$$s'(t) \approx \sum_{n=1}^{N} a'_n v_c(t - nT_c) \exp(j\omega_0 t) + \text{c.c.}, \tag{10.11}$$

where $a'_n = a_n \exp(j\omega_d n T_c)$. This has the same form as the ideal input waveform, equation (10.1). Thus, when the doppler-shifted waveform is applied to an ideal matched filter, the output peak and sidelobe amplitudes are given by the coefficients c_k of equation (10.7), with a_n replaced by a'_n. The correlation peak amplitude is found to be given by equation (10.10) with ε replaced by ω_d/ω_0, and thus falls to zero for $\omega_d = \pm 2\pi/T$, where $T = NT_c$ is the length of the waveform. A 1 dB reduction of the peak occurs for $\omega_d = \pm 1.6/T$, showing that PSK waveforms are much more doppler sensitive than chirp waveforms (Section 9.5.3). Experimental confirmation is given by Carr *et al.* [360].

10.2.3. Devices for MSK Waveforms

Although we have assumed so far that the waveform transmitted in a spread-spectrum system is a PSK waveform, there are in fact several other possibilities. Here we consider *minimum-shift keyed* (MSK) waveforms. Compared with PSK, MSK waveforms have spectral sidelobes whose amplitudes fall more rapidly with the distance from the centre frequency, as will be shown below. This feature enables MSK waveforms with different centre frequencies to be spaced more closely in frequency,

without appreciable mutual interference. As will be seen, there are several surface-wave devices that may be used for generation or correlation of these waveforms.

An MSK waveform may be written in the form

$$y(t) = \sum_m y_m(t),$$

with

$$y_m(t) = \cos\left(\omega_0 t + d_m \frac{\pi t}{2T_c} + \phi_m\right), \quad \text{for } mT_c \leqslant t \leqslant (m+1)T_c \quad (10.12)$$

and $y_m(t) = 0$ for other t. Here $d_m = \pm 1$ are the coding coefficients, T_c is the chip period, ω_0 is the centre frequency and ϕ_m are constants. It is usual to constrain the ϕ_m such that the waveform is continuous for all t, and this requires $\phi_m - \phi_{m-1} = \frac{1}{2}\pi m(d_{m-1} - d_m)$. The MSK waveform is clearly a form of frequency-shift keying, where the two frequencies involved are separated by $\Delta\omega = \pi/T_c$ and the choice of frequency for each chip is determined by the d_m. An example is shown in Figure 10.7.

Some methods for generating an MSK waveform are discussed by Amoroso and Kivett [371] who show, in particular, that it can be generated by applying a *PSK* waveform to a filter with a rectangular impulse response of length T_c. Such a filter is easily realised using surface-wave technology [372, 373]; as shown in Figure 10.7 the filter simply has a uniform transducer with length corresponding to the length of impulse response required, and another short uniform transducer which does not appreciably influence the response. This method enables MSK waveforms to be generated very conveniently.

To show that this method can indeed give an MSK waveform, the input PSK waveform is written as

$$v(t) = \sum_n v_n(t) = \sum_n a_n \text{ rect } (t/T_c - n) \exp(j\omega_1 t) + \text{c.c.}, \quad (10.13)$$

where $v_n(t)$ respresents chip n and $a_n = \pm 1$ gives the PSK coding. The filter impulse response $h(t)$ is a rectangular pulse of length T_c, written as

$$h(t) = \text{rect } (t/T_c) \exp(j\omega_2 t) + \text{c.c.} \quad (10.14)$$

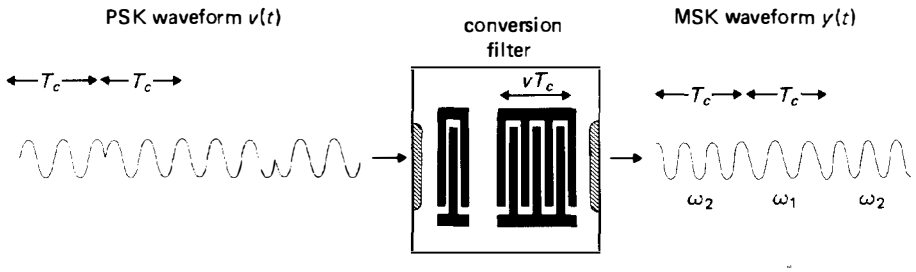


FIGURE 10.7. Conversion of PSK to MSK.

The centre frequencies of these two waveforms are made equal to the two frequencies of the MSK waveform, so that $(\omega_1 + \omega_2)/2 = \omega_0$, the MSK centre frequency, and $\omega_2 - \omega_1 = \pi/T_c$. The filter output waveform $y(t)$ is the convolution of $v(t)$ and $h(t)$, and is written as

$$y(t) = \sum_n g_n(t), \tag{10.15}$$

where $g_n(t) = v_n(t) * h(t)$ is the filter output due to chip n of the input waveform. Equations (10.13) and (10.14) are used to find $g_n(t)$, and the waveforms are taken to be bandpass functions so that it is sufficient to convolve the positive-frequency terms and then add the conjugate. After some manipulation, this gives

$$g_n(t) = \frac{2a_nT_c}{\pi}(-j)^n \exp(j\omega_0 t)\cos[\tfrac{1}{2}\pi(t/T_c - n)] + \text{c.c.}, \quad \text{for } |t - nT_c| \leqslant T_c \tag{10.16}$$

and $g_n(t) = 0$ for other t. Thus, $g_n(t)$ is a waveform of duration $2T_c$, with a cosine-shaped envelope falling to zero at each end. Alternatively, writing the cosine as a sum of exponentials shows that $g_n(t)$ is a sum of two rectangular pulses, with carrier frequencies ω_1 and ω_2. The total output waveform $y(t)$ is the sum of the $g_n(t)$, as in equation (10.15), but it can also be expressed as a sum of the $y_m(t)$ as in equation (10.12). The term $y_m(t)$ is the chip, of length T_c, commencing at $t = mT_c$, and for this time interval only $g_m(t)$ and $g_{m+1}(t)$ are non-zero. We thus find, from equation (10.16),

$$y_m(t) = \frac{T_c}{\pi}[(a_m + a_{m+1})\,e^{j\omega_1 t} + (-1)^m(a_m - a_{m+1})\,e^{j\omega_2 t}] + \text{c.c.},$$

$$\text{for } mT_c \leqslant t \leqslant (m+1)T_c. \tag{10.17}$$

Since $a_m = \pm 1$, this expression has the same form as equation (10.12) and is therefore an MSK waveform; the coefficient d_m is 1 when a_m and a_{m+1} are different, and -1 when they are the same. It also follows that $y(t)$ is continuous for all t.

Some further deductions can be made if we assume that the centre frequency is restricted such that $\omega_0 T_c = 2\pi M + \pi/2$, where M is some integer. In this case it follows from equation (10.16) that the waveform can be written

$$y(t) = \frac{2T_c}{\pi}\sum_n a_n g_c(t - nT_c), \tag{10.18}$$

where

$$g_c(t) = \exp(j\omega_0 t)\cos(\tfrac{1}{2}\pi t/T_c)\,\text{rect}(\tfrac{1}{2}t/T_c) + \text{c.c.} \tag{10.19}$$

It follows that an alternative method of generation is to use a filter whose impulse response has the form of $g_c(t)$ and apply impulses with spacing T_c and with polarities corresponding to the a_n; this method has been demonstrated using a surface-wave

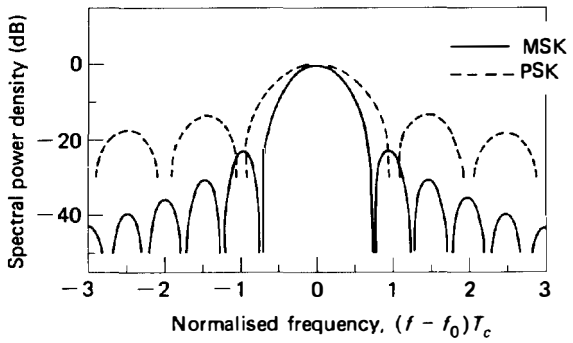


FIGURE 10.8. Spectral power density for one chip of MSK and one chip of PSK.

filter [374]. It is also clear that a matched filter for MSK can be realised by a method similar to the PSK filter of Figure 10.5, except that the uniform input transducer is replaced by an apodised transducer designed to give an impulse response proportional to $g_c(t)$. The output given by this filter in response to an input MSK waveform can be obtained from the analysis of Section 10.2.2 above, replacing $v_c(t)$ by the envelope of $g_c(t)$. It can also be anticipated that, for a waveform containing a large number of chips with coding chosen in a quasi-random manner, the power spectral density will be similar to that of $g_c(t)$. This conclusion is confirmed elsewhere [375]. Figure 10.8 compares the spectrum of $g_c(t)$ with that of a PSK chip, showing that the sidelobes of the MSK waveform fall more rapidly with frequency.

10.2.4. Frequency Hopping

Frequency hopping is a rather different approach to the problems of increasing security and reducing interference effects, and is applicable to both radar and communication systems. In its basic form the transmitter simply makes occasional changes of the waveform centre frequency, according to a pre-arranged pattern known to the receiver. This requires the use of a frequency synthesiser in the transmitter. In the receiver a similar synthesiser can be used, in conjunction with a balanced modulator, to remove the frequency hops.

A simple method of frequency synthesis using surface-wave devices [376, 377] is shown in Figure 10.9. Two chirp filters, with the same chirp slope, are impulsed at slightly different times. The resulting chirp waveforms, which overlap, are mixed together and the sum-frequency term, with constant frequency, is extracted by bandpass filtering. The output frequency is linearly related to the spacing τ of the input pulses, and can therefore be easily varied. By using two such systems and impulsing repetitively, a continuous waveform can be generated. Frequency synthesis can of course be done by more conventional methods, using multipliers and dividers; the main advantage of the surface-wave method is that a very large range of different frequencies can easily be obtained, since chirp filters with large time-bandwidth products are available (Chapter 9). Furthermore, in a communication system frequency hopping is most effective when there are many hops within each data digit, and this requires the frequency changes to be made rapidly and phase coherence to

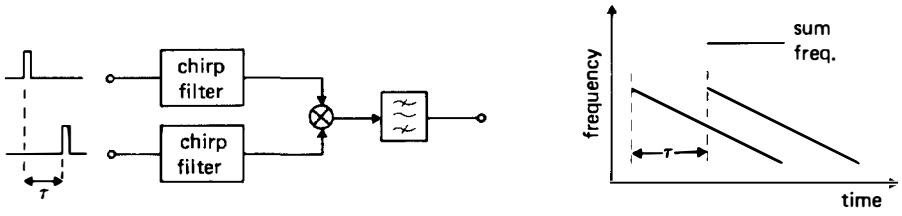


FIGURE 10.9. Frequency synthesis using chirp filters.

be maintained over many hops. These requirements are difficult to meet by conventional methods but can be met by the surface-wave method, which can change the frequency in a few nsec. The phase coherence has been demonstrated by correlating the waveform using a surface-wave convolver [377], described in Section 10.3.

Another surface wave method of frequency synthesis makes use of a filter bank, already described in Section 8.5. A repetitive comb waveform is generated using a stable oscillator, and the filter bank is used to filter the harmonics. An electronic switch is connected to each output of the filter bank so that the required frequency can be selected [378]. This technique gives better spectral purity than the chirp method, but is more limited in the number of frequencies obtainable.

Hunsinger *et al.* [352, 363] point out that a limited number of frequencies can be obtained simply by impulsing surface-wave devices with different centre frequencies. They used four surface-wave PSK filters, thus generating waveforms employing both PSK modulation and frequency hopping. This method of introducing frequency hopping provides a useful increase of signal bandwidth quite simply.

10.3. ACOUSTIC CONVOLVERS

In Section 10.2 we have considered some devices for correlating PSK and MSK waveforms, including programmable devices since programmability is a common requirement in spread-spectrum systems. The surface-wave convolver [379–382] offers another approach for programmable correlation, with the significant advantage of being much simpler to implement. The convolver differs markedly from all the devices described so far in this book, in that its operation relies on a non-linear effect and thus the basic physical principles are quite different. In this section we consider "acoustic" convolvers, in which the non-linearity involved is a property of the substrate material. In its basic form the acoustic convolver is structurally one of the simplest of surface-wave devices, consisting of two interdigital transducers and a uniformly metallised area between them. Despite this simplicity it performs one of the most sophisticated signal processing operations, correlating complex waveforms with a very high degree of programmability, subject only to constraints on the waveform duration and bandwidth. In addition the convolver is much less sensitive to temperature and velocity changes than most surface-wave devices.

The basic principles common to all convolvers are described in Section 10.3.1 below, and in Section 10.3.2 the performance of the basic acoustic convolver is considered. In recent years many devices have used waveguides to confine the surface-wave energy, thus increasing the energy density and hence the strength of the non-linear effect, and these devices are described in Section 10.3.3. Section 10.3.4 shows how the signal processing operation of a convolver can be analysed using a two-dimensional frequency response.

Some other types of convolver, using non-linear effects in semiconductors, will be described in Section 10.4.

10.3.1. Principles of Non-linear Convolvers

In this section the basic principles of non-linear convolvers are described, illustrating by referring to surface-wave devices in which the non-linearity is a property of the substrate material. While such devices are of most interest here, it should be borne in mind that the principles are in fact quite general. Thus, other sources of non-linearity can be exploited, as will be seen in Section 10.4, and other types of waves, such as bulk acoustic waves, can be used.

We first consider the non-linear interaction of two surface waves in the convolver shown in Figure 10.10(a), which is essentially a simple interdigital delay line. At this stage the two input waveforms $f_1(t)$ and $f_2(t)$ are taken to be C.W. waveforms, with frequencies ω_1 and ω_2 respectively, though more general waveforms will be considered later. For low power levels, such that non-linear effects are insignificant, the corresponding surface-wave amplitudes have the forms

$$u_1(t, x) = A_1 \cos(\omega_1 t - k_1 x)$$

and

$$u_2(t, x) = A_2 \cos(\omega_2 t + k_2 x), \tag{10.20}$$

where A_1 and A_2 are constants, $k_1 = \omega_1/v$ and $k_2 = \omega_2/v$ are the wavenumbers and v is the wave velocity. Attenuation and diffraction are assumed to be negligible. The

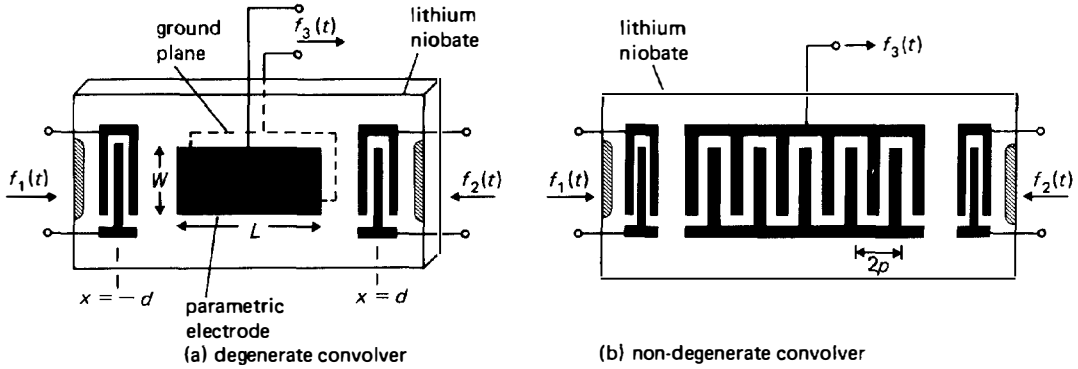


FIGURE 10.10. Surface-wave acoustic convolvers.

convolver makes use of non-linearity to mix the two waves. A variety of non-linear mechanisms can be used, and for the device of Figure 10.10(a) the non-linearity is a property of the substrate material itself. For this reason the device is called an "acoustic" convolver. As already seen in Section 6.3, acoustic non-linearity causes a surface wave to be accompanied by harmonics, and also causes attenuation of the fundamental. The convolver is operated at relatively low power levels, such that the fundamental components of the waves are not significantly affected and are therefore given by equations (10.20). However, the non-linearity will generate terms proportional to the squares $u_1^2(t, x)$ and $u_2^2(t, x)$ of the fundamentals, and a product term proportional to

$$u_1(t, x)u_2(t, x) = \tfrac{1}{2}A_1A_2[\cos\{(\omega_1 + \omega_2)t + (k_2 - k_1)x\} + \cos\{(\omega_1 - \omega_2)t - (k_1 + k_2)x\}]. \quad (10.21)$$

For low power levels, any further higher-order terms can be assumed to be negligible.

For a piezoelectric material, each of the various terms present has in general an associated electric field, and the convolver makes use of metal electrodes to selectively sense one of the terms. In the convolver of Figure 10.10(a) a uniform metal film called the "parametric electrode" is used, acting in conjunction with a ground plane; the latter may be provided simply by the metal carrier that the substrate is mounted on. The parametric electrode selectively senses the component of the electric field invariant with x. The sum-frequency component of the product term, equation (10.21), gives such a field when the input frequencies ω_1 and ω_2 are equal. Assuming that ω_1 and ω_2 are non-zero, the only other spatially-invariant fields arise from the difference-frequency components of the squares $u_1^2(t, x)$ and $u_2^2(t, x)$, but these terms have zero frequency and are suppressed because there is no D.C. path through the parametric electrode. Thus, ideally the output voltage arises only from the sum-frequency component of the product term, and exists only when the input frequencies are equal. With $\omega_1 = \omega_2 = \omega$, the output voltage has the form $\frac{1}{2}A_1A_2 \cos(2\omega t)$, and has a frequency equal to twice the input frequency. Note that the output amplitude is proportional to the product of the input amplitudes; a device giving this relation is said to be *bilinear*. This type of surface-wave mixing was first observed by Svaasand [383] using a quartz substrate, though most later devices have used lithium niobate, which gives a stronger interaction.

The device of Figure 10.10(a) is called a *degenerate* convolver because it responds most strongly when the input frequencies are equal. If the frequencies of the C.W. input waveforms are different the sum-frequency component of the product term, equation (10.21), has a spatial periodicity $2\pi/(k_2 - k_1)$. This component can be selectively sensed by introducing this periodicity in the parametric sensor [384]. As shown in Figure 10.10(b) this can take the form of an interdigital array of electrodes with pitch $2p$, giving an output at the sum frequency $(\omega_1 + \omega_2)$ when $\omega_1 - \omega_2 = \pm \pi v/p$. This device is therefore a *non-degenerate* convolver. The parametric sensor is essentially the same as a single-electrode transducer designed for a centre frequency of $|\omega_1 - \omega_2|$, though there is of course no surface wave present at this frequency. Apart from the frequency difference involved the non-degenerate

device operates in a manner very similar to the degenerate device. However, it has been found necessary to space the parametric sensor from the surface in order to prevent it causing substantial bulk wave excitation [379], and since this makes the fabrication inconvenient the non-degenerate device has received relatively little attention.

To appreciate the signal processing applications of the convolver it is necessary to consider more general input waveforms. Suppose that input waveforms $f_1(t)$ and $f_2(t)$ are applied to the input transducers of the degenerate convolver of Figure 10.10(a). Assuming that no distortion arises from the transducer responses or from propagation effects, and that the power levels are low enough, the surface-wave amplitudes will be proportional to $f_1(t - x/v)$ and $f_2(t + x/v)$. For simplicity, a delay corresponding to the transducer separation is ignored here. Since the surface-wave amplitudes must be oscillatory it can be concluded that, as in the C.W. case, the output will be mainly due to the product term; this will be justified formally below. The parametric electrode gives an output voltage $f_3(t)$ proportional to the spatial integral of the product, and thus

$$f_3(t) = \int_{-\infty}^{\infty} f_1(t - x/v) f_2(t + x/v)\, dx. \tag{10.22}$$

This assumes that the input waveforms have finite duration so that the product exists only for a finite region of x, and that the parametric electrode extends at least over this region. Writing $\tau = t - x/v$, equation (10.22) becomes

$$f_3(t) = v \int_{-\infty}^{\infty} f_1(\tau) f_2(2t - \tau)\, d\tau. \tag{10.23}$$

This equation is the *convolution* of $f_1(t)$ and $f_2(t)$, apart from the factor of 2 which causes a contraction in the time-scale. Apart from this contraction the convolver output is formally the same as that given by a *linear* filter, even though it relies on a non-linear effect. For a linear filter (Section A.2) the output waveform is given by the convolution of the input waveform with the filter impulse response. In the convolver, one of the input waveforms, called the *reference*, has the role of the "impulse response", and since this can be varied at will the convolver is clearly an exceptionally versatile device. In principle it can behave as any type of linear filter, for example as a bandpass filter or a matched filter for chirp or PSK waveforms, subject only to constraints on the bandwidth and duration of the reference. In practice a convenient method of generating the reference is needed and in consequence the convolver has been used mainly as a matched filter, as discussed in more detail later. The time-contraction given by the convolver arises physically because the two waves have a relative velocity of $2v$, in contrast to the velocity v which applies for a linear surface-wave device; this is also associated with the fact that for C.W. input waveforms the output frequency is twice the input frequency, as seen earlier.

Further insight into the operation of the degenerate convolver can be obtained from a frequency-domain analysis. The Fourier transform of the input waveform $f_1(t)$ is denoted $F_1(\omega)$, and since $f_1(t)$ is real we have $F_1(-\omega) = F_1^*(\omega)$. This enables the transform to be written as an integral over positive frequencies. Writing $F_1(\omega) = A_1(\omega) \exp [j\phi_1(\omega)]$, and also substituting $(t - x/v)$ for t, we find

$$f_1(t - x/v) = \frac{1}{\pi}\int_0^\infty A_1(\omega_1)\cos[\phi_1(\omega_1) + \omega_1 t - k_1 x]\,\mathrm{d}\omega_1, \qquad (10.24)$$

where $k_1 = \omega_1/v$. A similar relation applies for $f_2(t + x/v)$, with $A_1(\omega_1)$ and $\phi_1(\omega_1)$ replaced by $A_2(\omega_2)$ and $\phi_2(\omega_2)$, and with $-k_1$ replaced by $k_2 = \omega_2/v$. The spatial integral required to obtain the output waveform is readily done with the aid of the relation

$$\int_{-\infty}^{\infty}\cos(a + kx)\,\mathrm{d}x = 2\pi\delta(k)\cos a = 2\pi v\delta(\omega)\cos a, \qquad (10.25)$$

where a is a constant and $k = \omega/v$. Equation (10.25) follows from the transform of $\cos(\omega_0 t)$, equation (A.27), evaluated to $\omega = 0$. In evaluating the convolver output waveform, it is noted that the surface-wave waveforms must be bandpass waveforms, a restriction imposed by the transducers. Hence $A_1(\omega)$ and $A_2(\omega)$ are zero at $\omega = 0$, and the integrands involved are finite only when k_1 and k_2 are positive and non-zero. It follows from this that the spatial integrals of the linear and square-law terms are identically zero, apart from a D.C. term which is eliminated because there is no D.C. path through the parametric electrode. The output due to the product term, equation (10.22), involves a product of cosines, giving sum- and difference-frequency components. The spatial integral of the latter is found to be identically zero, and the output waveform $f_3(t)$, due only to the sum-frequency term, is found to be

$$f_3(t) = \frac{v}{2\pi}\int_{-\infty}^{\infty} F_1(\omega)F_2(\omega)\,\mathrm{e}^{2j\omega t}\,\mathrm{d}\omega. \qquad (10.26)$$

From the convolution theorem, equation (A.19), this is seen to be the time-contracted convolution of $f_1(t)$ and $f_2(t)$, agreeing with equation (10.23).

The same method can be used for the case of the non-degenerate convolver of Figure 10.10(b). For illustration, the spatial sensitivity of the parametric sensor can be taken to be sinusoidal, with period $2p$, so that the output due to the product term has the form

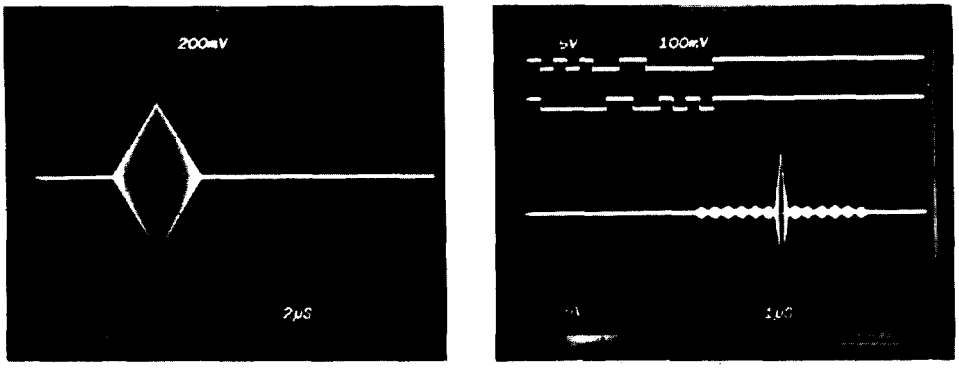
$$f_3(t) = \int_{-\infty}^{\infty} f_1(t - x/v)f_2(t + x/v)\cos(\pi x/p)\,\mathrm{d}x.$$

The input waveforms are assumed to be confined to frequency bands which do not overlap, and are such that only the sum-frequency component can have "wavenumber" equal to $2p$. The output waveform is found to be

$$f_3(t) = \frac{v}{4\pi}\,\mathrm{e}^{j\Omega t}\int_0^\infty F_1(\omega)F_2(\omega + \Omega)\,\mathrm{e}^{2j\omega t}\,\mathrm{d}\omega + \text{conjugate},$$

where $\Omega = \pm \pi v/p$, the sign depending on which of the input waveforms has the higher centre frequency. Thus, $f_3(t)$ is the time-contracted convolution of the input waveforms, but now with a frequency shift of Ω in one of the inputs and in the output.

The Convolver as a Matched Filter. Since the convolver behaves essentially as a linear filter, with the "impulse response" given by the applied reference



(a) (b)

FIGURE 10.11. Output waveforms produced by a degenerate acoustic convolver similar to that of Figure 10.10(a), for input waveforms that are (a) rectangular pulses; (b) PSK waveforms coded according to the 13-chip Barker code and its time-reverse. In both cases the output waveforms are 4.6 μsec long. (Courtesy J. H. Collins, University of Edinburgh)

waveform, it may in particular be used as a matched filter. Thus the convolver and its reference waveform generator may replace the matched filter in the spread-spectrum receiver illustrated in Figure 10.3. For this application the required reference waveform must correspond to the time-reverse of the ideal input waveform (for one symbol). As we have seen earlier, PSK and MSK waveforms can be generated quite readily, and hence the convolver is particularly suited for matched filtering of these waveforms. Note that the convolver enables the coding to be changed at will, and also the bandwidth and centre frequency if necessary; it is therefore highly programmable.

For illustration, Figure 10.11 shows some output waveforms produced by a degenerate acoustic convolver similar to that of Figure 10.10(a), with an input centre frequency of 110 MHz and a parametric region 4.6 μsec long. In Figure 10.11(a) the inputs are both rectangular pulses 4.6 μsec long, with carrier frequency 110 MHz. The output waveform has a triangular envelope, as given by the convolution of a rectangle with itself, with carrier frequency 220 MHz. The output waveform virtually disappears if either of the two inputs is removed. In Figure 10.11(b) one input is a PSK waveform coded according to the 13-chip Barker code 0101001100000, and the other input is the time-reverse. The binary coding is shown in the upper part of the figure, prior to bi-phase modulation. For this case the ideal output waveform has six sidelobes of equal amplitude on each side of the main peak, and the peak amplitude is 13 times the sidelobe amplitude. It can be seen that the convolver output waveform corresponds very closely to this ideal.

Note that, despite the use of a non-linear mechanism, the convolver processes an input waveform in an essentially linear fashion, as shown by equation (10.23), and this applies even for an input signal accompanied by noise. Ideally, the output signal-to-noise ratio will be the same as that given by a linear matched filter. There is however the limitation that the signal is properly processed only if it arrives at the same, or almost the same, time as the reference is applied, so that the two waveforms overlap only in the parametric region of the device. If the time of arrival of the signal

is unknown, an obvious strategy is to apply the reference waveform repetitively. A detailed examination of this process [385] shows that, apart from a time distortion due to the time contraction, an ideal convolver can give *exactly* the same output as a matched filter, irrespective of the signal timing, and this applies even if the signal is accompanied by noise or interference. To obtain this result, the delay along the convolver parametric region must be at least twice the duration of the reference.

Signal Processing Using Idler Wave Generation. An alternative type of signal processing can be obtained by applying input waveforms to one transducer and to the parametric port, instead of to both transducers. With an appropriate choice of frequencies, the non-linearity in this case generates a second surface wave, travelling in the opposite direction to the incident wave. This "idler" wave can be sensed by the input transducer, or by another transducer at the same end. If the input at the parametric port is a short pulse the output waveform is esentially the *time-reverse* of the transducer input waveform. This has been demonstrated using the non-degenerate convolver of Figure 10.10(b) [379, 384], and also using the non-degenerate diode convolver, considered later. For degenerate devices, the use of the idler wave is not usually practicable because of the presence of spurious signals. Time-reversal is of some practical interest since a reference of this form is needed for a convolver using contra-directed input waves. Thus a coded signal can be correlated using a reference of the *same* form if the reference is first time-reversed in another convolver.

10.3.2. Performance of Basic Convolvers

The simplest type of surface-wave convolver is the degenerate acoustic device of Figure 10.10(a). The principles of this device were discussed in the previous section, and here we consider its practical performance. Acoustic convolvers using waveguides to improve the efficiency will be described in Section 10.3.3 below, and convolvers using semiconductors will be described in Section 10.4.

The first convolvers to be demonstrated were similar in principle to the device of Figure 10.10(a), but used bulk acoustic waves rather than surface waves [386]. Convolution using surface waves, with the degenerate and non-degenerate structures of Figure 10.10, was first demonstrated by Luukkala and Kino [384]. In all of these devices the non-linear propagation medium was lithium niobate.

(a) Efficiency of Degenerate Acoustic Devices. As explained in Section 10.3.1, the convolver exploits a non-linear interaction of propagating surface waves to produce an output voltage proportional to the spatial integral of the product. A primary concern is the efficiency of this interaction, and this has been examined theoretically by several authors [387–390]. It is found that, for C.W. input waveforms with the same frequency ω, the open-circuit output voltage at frequency 2ω may be expressed in the form

$$[V_{oc}]_{rms} = \frac{M}{W}\sqrt{P_{s1}P_{s2}}, \tag{10.27}$$

where P_{s1} and P_{s2} are the surface-wave powers, taken to be small, and W is the width

of the parametric electrode, assumed to be equal to the transducer apertures. The constant M depends only on the substrate material and orientation, and is generally defined such that the equation gives the r.m.s. value of the open-circuit voltage. The derivation assumes ideal propagation conditions, so that diffraction, dispersion and attenuation are negligible. Note that the output voltage is independent of the input frequency and of the length of the parametric electrode.

The constant M is related in a complex manner to the linear and non-linear bulk constants of the material [387–390]. In principle this enables M to be calculated for any orientation of interest, but in practice many of the constants required are unknown, or are not sufficiently accurate. It is thus necessary to determine M experimentally, by measuring the output voltages of practical devices, though theoretical predictions for lithium niobate are in fair agreement with experiment [390]. Measurements have been reported for a variety of materials [391], showing that Y, Z lithium niobate gives the relatively large value of $M = 1.2 \times 10^{-4}$ V m/watt. This is the usual choice of substrate material since, in addition to the strong non-linearity, it also gives low attenuation and diffraction spreading. The temperature sensitivity is of little consequence in convolvers, as explained below. The only material known to give a larger M-value is PZT ceramic, but this is generally unacceptable because of the acoustic propagation loss.

The analysis [387, 388] also leads to the conclusion that the convolver output port can be represented by a simple equivalent circuit consisting of a voltage generator V_{oc} in series with a capacitance. The latter is simply the capacitance measured between the parametric electrode and the ground plane. The absence of any resistance here appears to imply that infinite power can be extracted, but in fact the analysis only applies when the interaction is too weak to cause any appreciable reduction of the surface-wave amplitudes, as is usually the case in practice. A small amount of resistance is in fact present because of the resistivity of the parametric electrode, and because the voltage on the electrode causes some excitation of bulk waves.

A more practical measure of the device efficiency is the *bilinearity factor* C, relating the external powers of the input and output waveforms. We define P_1 and P_2 as the available powers of the generators supplying the C.W. input waveforms, and P_0 as the output power delivered to the load. These powers can if appropriate take account of any matching networks, which are usually included in practical devices in order to maximise the efficiency. Since the device is bilinear P_0 is proportional to P_1P_2, and the efficiency can therefore be characterised by defining the bilinearity factor as

$$C = 10 \log_{10}\left(\frac{P_0}{P_1P_2}\right), \quad \text{(dBm)}, \tag{10.28}$$

where P_1, P_2 and P_0 are measured in mW. This definition can be applied to any bilinear device, including several types of convolver, described later, which do not use the acoustic non-linearity; it also applies if the performance is affected by propagation effects such as attenuation or diffraction. Generally, C will depend on the input frequency ω; for an ideal device, C would be independent of ω over the band occupied by the input waveforms to be used. With appropriate matching, the degenerate

acoustic convolver of Figure 10.10(a) with an input centre frequency of, say, 100 MHz gives typically $C = -95$ dBm. For practical applications a better efficiency is desirable, though not always essential, and some modified types of convolver giving better efficiencies will be described in Sections 10.3.3 and 10.4.

(b) Second-order Effects. In practice, degenerate convolvers deviate in several ways from the ideal behaviour described above and in Section 10.3.1, though some imperfections can be minimised quite simply [380, 392]. In particular, it is usual to restrict the input frequency band such that it does not overlap the output band, and to use external bandpass filters to reject some unwanted components.

It has been assumed so far that the power levels of the input surface waves are low, so that the non-linearity has a negligible effect on the fundamental waves and causes only square-law and product terms to be generated. At high power levels this will not be true, so that the device will no longer be bilinear; the output amplitude will no longer be proportional to the product of the input amplitudes, and will thus exhibit saturation. In this situation the output voltage is no longer given by equation (10.27), and the bilinearity factor of equation (10.28) is not valid. In practice, saturation is not usually observed at realistic power levels, as might be expected from the low device efficiencies. For example, if C.W. waveforms with $+20$ dBm available power are applied to a device with a typical $C = -95$ dBm, the output power is -55 dBm, which is many orders of magnitude below the input powers.

As in other surface-wave devices, the convolver can be affected by several propagation effects. Consider first the effect of *attenuation*, assuming the wave propagation to be otherwise ideal. We consider the degenerate device of Figure 10.10(a) and allow for the delay due to the transducer separation, which was neglected in the previous section. If waveforms $f_1(t)$ and $f_2(t)$ are applied to the input transducers, and if the power levels are low, the fundamental waves have amplitudes of the form

$$u_1(t, x) \propto f_1(t - x/v - d/v) \exp[-\alpha(d + x)],$$
$$u_2(t, x) \propto f_2(t + x/v - d/v) \exp[-\alpha(d - x)], \tag{10.29}$$

where the transducers are taken to be located at $x = \pm d$ and α is the attenuation coefficient, taken to be independent of frequency. The parametric electrode is taken to be of length L, occupying the region $|x| < \frac{1}{2}L$. The output waveform $f_3(t)$ is found by integrating the product $u_1(t, x)u_2(t, x)$ over this region, giving

$$f_3(t) = \gamma\, e^{-2\alpha d} \int_{-L/2}^{L/2} f_1(t - x/v - d/v) f_2(t + x/v - d/v)\, dx, \tag{10.30}$$

where γ is a constant. We thus have the important conclusion that the form of the output waveform is not affected by exponential attenuation. This is in marked contrast to a linear surface-wave device, in which attenuation causes distortion of the output waveform. In practice α varies a little with frequency (Section 6.5), causing some distortion, but this is usually very small and has a form such that it can be compensated by external trimming [393].

Equation (10.30), with α constant, can be taken as the form of output waveform given by an *ideal* degenerate convolver. The convolution relation is obtained if the

input waveforms have finite durations, less than L/v, and are timed such that the product exists only in the parametric region $|x| < \frac{1}{2}L$. Equation (10.30) is then unaffected if the limits are changed to $\pm \infty$, and with $\tau = t + x/v - d/v$ gives

$$f_3(t) = v\gamma\, e^{-2\alpha d} \int_{-\infty}^{\infty} f_1(2t - 2d/v - \tau)\, f_2(\tau)\, d\tau. \tag{10.31}$$

This is the convolution of $f_1(t)$ and $f_2(t)$, with a time-contraction of 2 and a delay of d/v. Equation (10.31) shows that the output is not distorted by a change of the *velocity* v (due to fabrication errors for example), since this simply changes the delay d/v. In addition *temperature* changes cause no distortion; d and v are both affected, but this simply changes the delay d/v in accordance with the temperature coefficient α_T of Section 6.4. The insensitivity of the convolver to these propagation effects is an important advantage and is particularly relevant when processing PSK waveforms, since PSK filters are very sensitive to velocity and temperature changes (Section 10.2).

Another relevant propagation effect is *dispersion*, which can significantly affect the convolver fidelity. As seen in Section 6.5, dispersion arises from the mass loading due to the parametric electrode, and depends on the electrode thickness. The effect is conveniently expressed in terms of the spectrum of the output waveform. If $V_0'(\omega)$ is the output spectrum, allowing for dispersion, it can be shown that for a degenerate device [393, 394]

$$V_0'(\omega) \approx V_0(\omega) \exp\,[-jLk(\omega/2)], \tag{10.32}$$

where $V_0(\omega)$ is the ideal output spectrum, not allowing for a delay $L/(2v)$, and $k(\omega)$ is the surface-wave wavenumber at frequency ω. Equation (10.32) is valid for arbitrary input waveforms. It shows that the phase distortion in the output spectrum, at frequency ω, is the same as the phase distortion for a surface wave at frequency $\omega/2$ travelling the length L of the parametric electrode. Note that, for no distortion, k must be a linear function of ω. This requires the group velocity $v_g = d\omega/dk$ to be independent of ω, though the phase velocity ω/k need not be constant. For PSK waveforms, theoretical simulations [394] lead to the conclusion that the dispersion is generally acceptable if $|dv_g/d\omega|$ is less than $v_g^2 T_c^2/(8L)$. For the basic convolver this requirement is usually satisfied quite easily, though it becomes of some concern for waveguide convolvers, as will be seen in Section 10.3.3.

In many devices the most serious imperfection is *fold-over convolution*, due to the acoustic reflectivity of the transducers. A wave generated at one end of the device is reflected by the transducer at the other end, and the reflected and primary waves are mixed in the parametric region in the usual way. This gives an output signal when only one input waveform is applied, so that the device is not strictly bilinear; thus the bilinearity factor of equation (10.28) is meaningful only if the fold-over convolution is well suppressed. A simple test for this is to measure the output level when C.W. waveforms of the same power are applied to the inputs, and compare with the output level observed when one input is disconnected. Typically, a degenerate device such as that of Figure 10.10(a) gives a fold-over rejection of 25 to 30 dB in this test, and this is not generally acceptable. As in linear devices the transducers could be electrically mis-matched in order to reduce their reflection coefficients (Section 7.1.3), but for convolvers this is not generally attractive as it reduces the efficiency, which is already

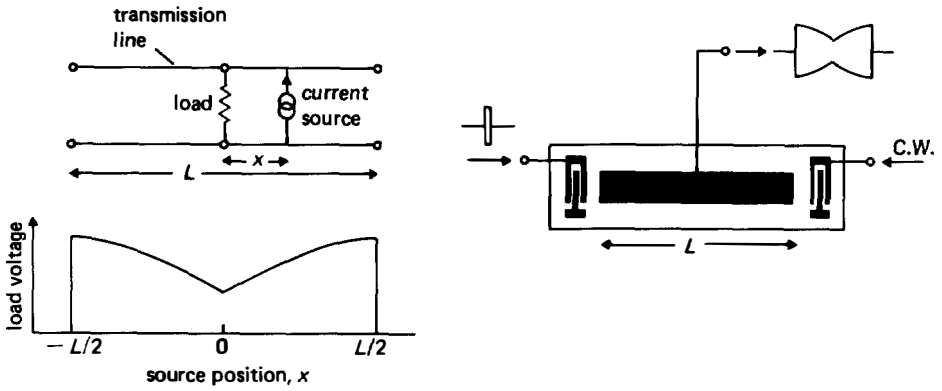


FIGURE 10.12. Transmission-line effect in convolver. Left: analysis. Right: simple experimental test.

rather low. A better method is to adopt a duplicated arrangement, explained later in Section 10.3.3; this typically enables a fold-over rejection of 40 dB to be obtained. The non-degenerate convolver of Figure 10.10(b) gives better fold-over rejection because the product due to the reflected wave and the primary wave does not have a spatial periodicity equal to that of the parametric sensor.

Some devices are significantly affected by electromagnetic wave propagation along the parametric electrode, which is thus found to behave as a *transmission line*. When this is the case, the output voltage due to the product term at location x is found to be dependent on x, so that the device is spatially non-uniform. Thus the ideal form of the output, equation (10.30), is no longer produced even if the product is formed correctly at each x. The effect becomes significant when the length of the parametric electrode exceeds about one quarter of the electromagnetic wavelength, at the output frequency. A simple analysis [395, 396] treats the parametric electrode as a transmission line, with a C.W. current source at the output frequency representing a localised excitation, as in Figure 10.12. The calculated output voltage, as a function of source position, is shown for a transmission line one third of a wavelength long, appropriate for a degenerate device with 260 MHz output frequency and a parametric electrode 25 μsec long [395]. A simple experimental test for this is to apply a C.W. waveform to one input of the convolver, and a short pulse to the other input; as the surface-wave pulse scans along the parametric electrode, the output waveform shows the spatial sensitivity as a function of time. This generally gives good agreement with the analysis, though strictly speaking it is not compatible since the analysis assumes C.W. excitation.

The electrical length of the parametric electrode can be reduced, thus reducing the distortion, by mounting the substrate over a dielectric layer with lower permittivity [395]. Some other methods are mentioned in Section 10.3.3 below, and a more rigorous experimental test is described in Section 10.3.4.

10.3.3. Waveguide Convolvers

In view of the low efficiency of the basic acoustic convolver, several more efficient types of convolver have been developed. In this section we consider a modified type of acoustic convolver, while some other devices are described in section 10.4.

Assuming that the input transducers are well matched and the input power levels are specified, equation (10.27) shows that the only strategy available for increasing the output voltage of a degenerate acoustic convolver is to reduce the width W of the parametric electrode. It is found that a width of only a few wavelengths can be used, and for such a narrow width the electrode behaves as a surface-wave *waveguide*. This is a useful feature since it eliminates the effects of diffraction spreading and beam steering, which would cause severe difficulties if the beam were not guided. On the other hand, the presence of a waveguide introduces some new complications, so we first consider briefly the waveguide behaviour.

A metal strip on a lithium niobate substrate acts as a waveguide because the surface-wave velocity for a uniformly metallised surface is lower than that for a free surface. The velocity reduction is mainly due to electrical loading, and consequently the strip is called a "$\Delta v/v$ waveguide". The behaviour has been analysed by Schmidt and Coldren [397], and experimental results [397, 398] are in good agreement. The waveguide can support a series of propagating modes, each of which is dispersive. The phase and group velocities of the first few modes are shown in Figure 10.13 for a Y, Z lithium niobate substrate. Here the analysis [397] has been used with accurate velocity data obtained from optical probing measurements (Section 6.2.4), since the velocity anisotropy affects the solutions somewhat. The abscissa gives the guide width a divided by λ_0, the wavelength for straight-crested waves propagating in the Z-direction on a free surface. For the present we consider only the continuous curves, referring to a film of zero thickness.

In a convolver it is desirable that only the fundamental mode should be present, since the presence of two or more modes causes distortion. The higher modes can be excluded by operating at frequencies below their cut-off points. However, the excitation of these modes is generally quite weak at higher frequencies, because their transverse amplitude distributions are very different from the distribution of the wave incident at the end of the guide. This is true in particular for the first higher mode, which has an amplitude anti-symmetric about the guide centre line. For this reason, higher-mode excitation is generally small if $a/\lambda_0 < 4$, the cut-off point of the second higher mode. At the same time, since dispersion causes distortion of the convolver response (Section 10.3.2), it is advisable to avoid the region $a/\lambda_0 < 1.5$ where the fundamental is highly dispersive. Thus a suitable range for the guide width is $1.5 < a/\lambda_0 < 4$.

In this region of a/λ_0, the group velocity of the fundamental mode is almost independent of frequency, as shown by the corresponding continuous line in Figure 10.13. This implies that the dispersion causes little distortion of the convolver output waveform, as noted earlier. However, for a finite film thickness the dispersion is increased, as shown by the broken line which refers to a typical experimental value. The added dispersion here is due to mechanical loading, and is acceptable for many applications. In fact, the extra dispersion can be advantageous because it reduces the

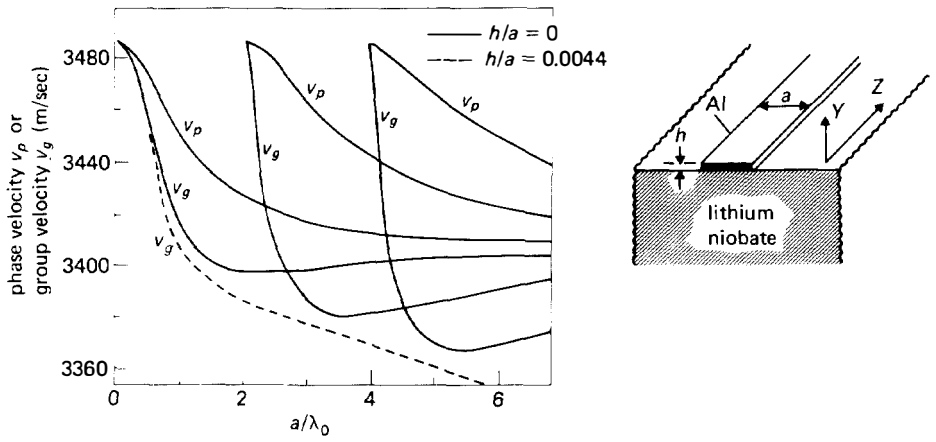


FIGURE 10.13. Phase and group velocities for modes in a $\Delta v/v$ waveguide (Courtesy, Plessey Research).

surface-wave attenuation associated with the non-linearity, as noted in Section 6.3. This is confirmed by the fact that saturation is observed in convolvers using very thin films [399] but not in convolvers using typical film thicknesses, for input power levels of about +20 dBm.

To use a waveguide effectively, some efficient method of launching the narrow surface-wave beams must be found. Narrow-aperture uniform transducers are not generally attractive because their high impedances make them difficult to match efficiently. At high frequencies the effective impedance may be reduced by parasitic capacitance (Section 7.1.2), but this also increases the Q-factor and so limits the bandwidth obtainable. A better approach is to use a wide aperture transducer and compress the beam width using either a multi-strip coupler (Section 5.7) or a waveguide horn [400]. Focussing transducers, in which the electrodes are curved, have also been shown to be effective [401]. Another method is to use a narrow-aperture *chirp* transducer illuminating the end of the waveguide directly, as shown in Figure 10.14. We have already seen, in Section 9.3.2, that the admittance of a chirp transducer is larger than that of a uniform transducer with the same bandwidth and aperture. The use of a chirp thus enables the impedance of a narrow-aperture transducer to be reduced without changing its aperture, as shown experimentally using a time-bandwidth product of about 10 [402]. In a convolver, the dispersion introduced by the transducers is cancelled by arranging one transducer to be an up-chirp and the other a down-chirp, as in Figure 10.14. A further advantage is that the transducer chirps can be made slightly different to compensate for small amounts of dispersion in the waveguide [402, 403].

The first waveguide convolvers, demonstrated by Defranould and Maerfeld [404], used multi-strip beam compressors and gave a bilinearity factor of $C = -71$ dBm, much better than the -95 dBm typical of the basic degenerate convolver. To obtain a suitable capacitance a ground electrode was located on the *upper* surface, adjacent to the waveguide, instead of on the rear surface as in Figure 10.10(a). Here we

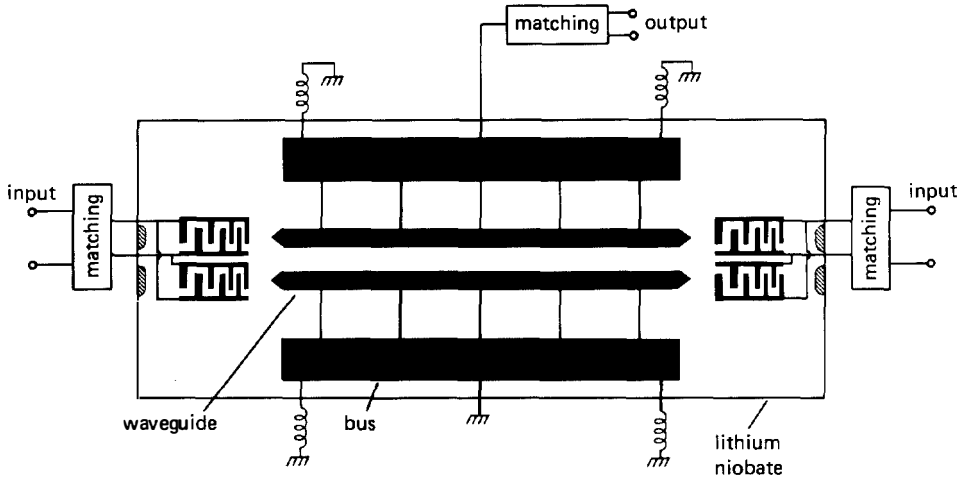


FIGURE 10.14. Waveguide convolver using chirp transducers.

consider as an example the device using chirp transducers [402], shown in Figure 10.14. This device had an input centre frequency and bandwidth of 300 MHz and 120 MHz respectively, and the delay along the parametric region was 16 μsec; thus signals with time-bandwidth products up to 1,920 could be correlated. As shown in the figure, the basic structure is duplicated so that there are two input transducers at each end, and the electrode polarities in one of the four transducers are reversed. This implies that, ideally, the waves launched at one end do not excite the tranducers at the other end; thus the acoustic reflection coefficient is reduced, and the fold-over suppression is improved. The waveguide width was 34 μm, giving $a/\lambda_0 = 3$ at 300 MHz, the input centre frequency. The wide bus-bars, connected at intervals to the waveguides, are included to minimise losses due to the resistance of the narrow waveguides. The small inductors reduce the spatial non-uniformity due to the transmission-line effect [405]; alternatively, it has been shown that the non-uniformity can be reduced by using an external transmission-line network to equalise the phase [406]. A photograph of the device, including simple matching networks, is shown in Figure 10.15.

Figure 10.16 shows the measured output power, for C.W. input waveforms with the same frequency and with 0 dBm power level; this is equal to the bilinearity factor C defined in equation (10.28). At the input centre frequency, 300 MHz, the bilinearity factor was -67 dBm. The graph on the right shows the phase error, that is, the measured phase of the output with a large linear term subtracted. The graph on the left also shows the output powers observed with one of the inputs disconnected; these unwanted components are due to fold-over convolution, and their powers show that the fold-over suppression is typically 38 dB. Similar performance has been demonstrated by devices using horn [406–408] or multistrip [399, 409] beam compressors instead of chirp transducers. The spatial uniformity of the device is considered further in Section 10.3.4 below.

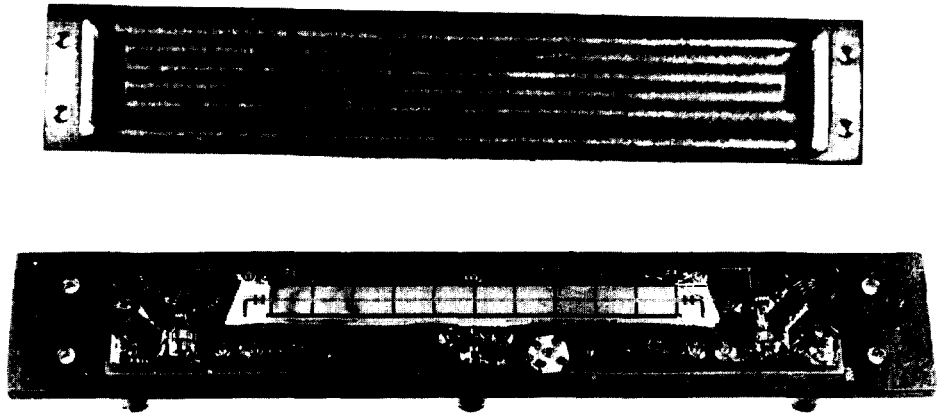


FIGURE 10.15. Photograph of acoustic waveguide convolver, with cover removed. The total length of the package is 12 cm. (Courtesy, Plessey Research).

Required Efficiency and Fold-over Rejection. Assuming typical input power levels of +20 dBm the above convolver, with $C = -67$ dBm, gives an output power level of -27 dBm. This figure applies for C.W. input waveforms, and also applies for coded waveforms if the reference corresponds to the time-reverse of the signal and if the waveform durations are equal to the delay along the parametric region. If the output is applied to an amplifier with 240 MHz bandwidth, an output signal-to-noise ratio of about 58 dB is obtained. This implies that the input signal can be reduced by 58 dB before the output disappears below the noise, thus showing a very useful dynamic range. However, a more detailed examination is needed to deduce the efficiency needed for a given application [410]. The main criterion is that, when correlating a signal accompanied by noise, the thermal noise at the convolver output must not appreciably degrade the output signal-to-noise ratio. It is found that the efficiency needed increases with both the bandwidth and the time-bandwidth product

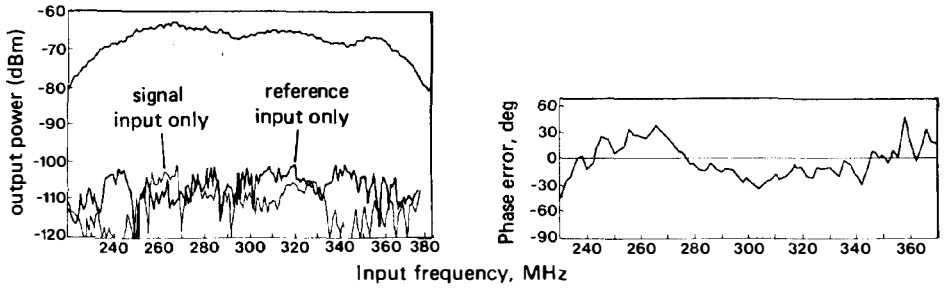


FIGURE 10.16. Performance of the waveguide convolver of Figure 10.15 for C.W. input signals with the same frequency. Left: output powers for 0 dBm input power levels. The upper curve gives the bilinearity factor, C. Right: phase error of output waveform. (Courtesy Plessey Research).

of the signal being correlated; for example, if the signal bandwidth is 120 MHz and the *TB*-product is 2000, the bilinearity factor needs to exceed about −80 dBm. This requirement is satisfied by the waveguide convolver, but not by the simpler acoustic convolver of Figure 10.10(a).

The level of the fold-over convolution is also an important consideration. A particular consequence of this phenomenon is that a spurious output signal, associated with the reference input, is present irrespective of the level of the signal input. This can obscure the required output signal when the input signal level is low, thus reducing the dynamic range. Analysis of this effect [410] shows that, for a device correlating a coded signal, the requirement on the C.W. fold-over rejection depends on the signal *TB*-product, and becomes more stringent for *lower TB* products. For the waveguide convolver it is concluded that the C.W. fold-over rejection of 38 dB is adequate for signals with *TB*-products exceeding 100.

10.3.4. Convolver Fidelity and Frequency Response

We here consider the distortion of the output waveform due to imperfections in the convolver, excluding fold-over convolution. Phenomena such as dispersion, the transmission-line effect and imperfect transducer responses will distort the output waveform, causing for example an increase in the width of the correlation peak and a reduction of signal-to-noise ratio. In addition the peak amplitude will generally depend on the relative timing of the signal and reference inputs, owing to the spatial non-uniformity.

For a *linear* device, any imperfections can be conveniently assessed by considering the frequency response $H(\omega)$, obtained by C.W. measurements. Using $H(\omega)$, the output waveform can be calculated for any input waveform, without explicitly considering the physical processes involved. For a convolver the conventional definition of frequency response is not applicable because the device is not linear. However, it has been shown that a *two-dimensional* frequency response can be used in this case.

To define the convolver frequency response, suppose that C.W. waveforms $\cos(\omega_1 t)$ and $\cos(\omega_2 t)$, with different frequencies, are applied to the two inputs. It is assumed that the device is bilinear, thus neglecting fold-over convolution; this is generally a good approximation for a well-designed device. The output is therefore a sum of two C.W. waveforms at the sum and difference frequencies, and usually the difference-frequency term can be ignored because it is small and is of no interest. The sum-frequency output waveform is written in the form

$$f_3(t) = \tfrac{1}{2} H(\omega_1, \omega_2) \exp[j(\omega_1 + \omega_2)t] + \text{conjugate}, \qquad (10.33)$$

where $H(\omega_1, \omega_2)$ is the two-dimensional frequency response, a complex function whose magnitude and phase give the amplitude and phase of the output waveform. It is convenient to define $H(\omega_1, \omega_2)$ for negative frequencies as well as positive, and this is done by defining $H(-\omega_1, -\omega_2) = H^*(\omega_1, \omega_2)$ and $H(\omega_1, \omega_2) = 0$ when ω_1 and ω_2 have different signs. It can be shown [411] that the function $H(\omega_1, \omega_2)$ gives a complete description of the electrical behaviour of any bilinear device, irrespective of the physical processes involved. For arbitrary input waveforms, the output

waveform can be deduced from the equation

$$F_3(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} F_2(\omega') F_1(\omega - \omega') H(\omega - \omega', \omega') \, d\omega', \tag{10.34}$$

where $F_1(\omega)$ and $F_2(\omega)$ are the Fourier transforms of the input waveforms, and $F_3(\omega)$ is the transform of the output waveform. It is also possible to define a two-dimensional impulse response as the inverse Fourier transform of $H(\omega_1, \omega_2)$, and the output waveform can then be expressed as a two-dimensional convolution involving the input waveforms [411].

For an ideal degenerate convolver, the frequency response takes the form

$$H(\omega_1, \omega_2) = K \operatorname{sinc} [\tfrac{1}{2}(\omega_1 - \omega_2) T_0] \, e^{-j(\omega_1 + \omega_2) T_d} \tag{10.35}$$

for positive frequencies, where K is a constant, $T_0 = L/v$ is the delay along the parametric region, and $2T_d = 2d/v$ is the delay corresponding to the separation of the input transducers. Using equation (10.34), it can be shown that this ideal response leads to an integral with the ideal form of equation (10.30), relating the waveforms in the time domain.

Figure 10.17 shows an experimental arrangement for measuring the frequency response, omitting some amplifiers and filters that are necessary in practice. Two signal generators are used to generate C.W. input waveforms at frequencies ω_1, ω_2, and a vector voltmeter measures the amplitude and phase of the convolver output. The phase reference for the voltmeter is obtained by mixing the two input waveforms, giving a waveform at the sum frequency $\omega_1 + \omega_2$. If the input frequencies are the same, so that $\omega_1 = \omega_2 = \omega$, say, the output amplitude is directly related to the bilinearity factor C at frequency ω.

In order to examine the spatial uniformity of the convolver, it is convenient to define another form of the response. This new function, called the *spatial response* $P(\omega, \tau)$, is defined as the inverse Fourier transform of $H(\omega_1, \omega_2)$ with respect to the difference of the two input frequencies, so that [412–414]

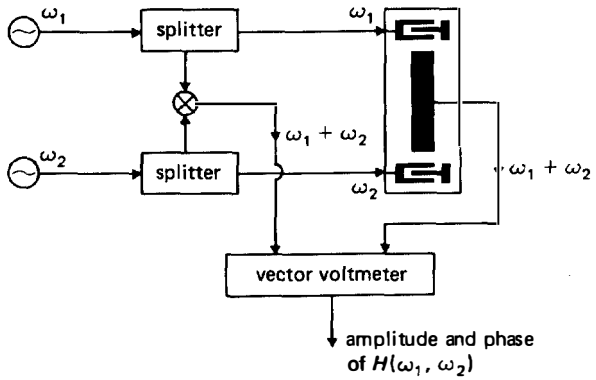


FIGURE 10.17. Measurement of two-dimensional frequency response of convolver.

$$P(\omega, \tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H\left(\frac{\omega + \omega'}{2}, \frac{\omega - \omega'}{2}\right) e^{j\omega'\tau} \, d\omega'. \tag{10.36}$$

For an ideal convolver, $P(\omega, \tau)$ is independent of τ for $|\tau| < T_0/2$ and is also independent of ω over the band occupied by the input waveforms. Further analysis [414] shows that an approximate physical interpretation can be obtained by equating τ with x/v; at each x, the product term is in effect applied to a linear filter with frequency response $P(\omega, \tau)$, and the filter outputs, infinite in number, are summed.

Experimentally, the spatial response can be measured using the arrangement of Figure 10.17, varying the input frequencies such that their sum is equal to a constant ω, and transforming the output data with respect to the difference frequency as in equation (10.36). Figure 10.18 shows the amplitude and phase of $P(\omega, \tau)$ for the waveguide convolver described in Section 10.3.3, taking $\omega = 2\pi \times 600$ MHz, the output centre frequency. The sharp dips observable in the amplitude are associated with the connections between the waveguides and the bus-bars. This method of measuring spatial uniformity is much superior to the method using a short input pulse, mentioned in Section 10.3.2; the spatial resolution obtainable by the latter method is limited by the fact that short input pulses give low signal-to-noise ratios at the device output. In addition, since short pulses have wide bandwidths the output given by the pulse method is in fact a frequency-domain average of the spatial response, and is therefore less informative.

The observed variation of $P(\omega, \tau)$ with τ implies that, for a device correlating a coded waveform, the amplitude of the output correlation peak will depend on the timing of the input signal. In addition, if the input signal is accompanied by noise, the output noise power will vary with time. However, a detailed study of these effect [412] shows that quite large errors can be tolerated, and it can be concluded that the spatial variations shown in Figure 10.18 would be acceptable for most applications.

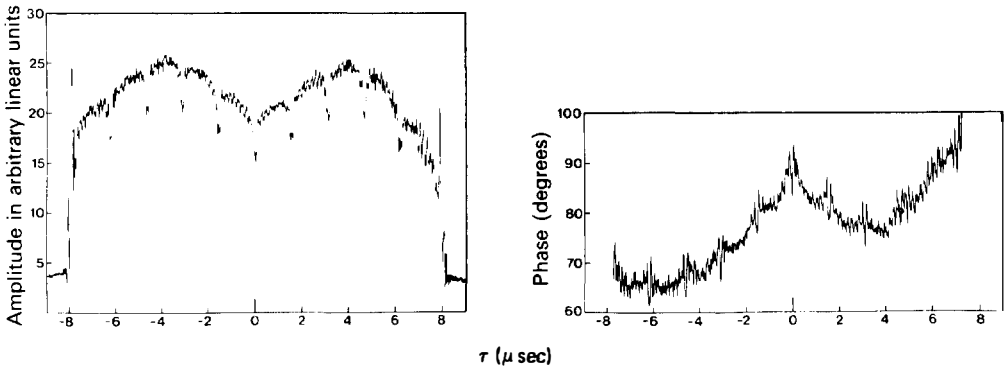


FIGURE 10.18. Amplitude and phase of the spatial response $P(\omega, \tau)$, for the convolver of Figure 10.15. (Courtesy, Plessey Research).

10.4. OTHER NON-LINEAR DEVICES

In addition to the acoustic convolvers described above several other types of convolver have been developed, making use of non-linear effects in semiconductors [381, 382]. Historically, these devices were developed in order to improve on the rather low efficiency of the basic acoustic convolver, and they did indeed give much better efficiencies. However, the later development of the waveguide acoustic convolver has made these novel devices, which are more difficult to fabricate, less attractive. They are nevertheless of some interest, not only because of the performance achieved but also because they have led to the development of several novel signal processing operations. These include the ability to store a reference waveform and to use it to correlate a signal applied later, and also the ability to correlate a very long waveform, with duration much greater than the acoustic delay in the device.

Semiconductor convolvers are described in Section 10.4.1 below, and Section 10.4.2 describes devices for both storage and correlation. Correlation of long waveforms is described in Section 10.4.3.

10.4.1 Semiconductor Convolvers

The surface-wave *diode convolver* [415–417], introduced by Reeder, is shown in Figure 10.19. Here the non-linearity is produced by a set of diodes, each connected to one of a series of uniformly spaced transducers acting as taps. Non-degenerate operation is necessary so that the harmonics of the input waveforms can be rejected by filtering the output. As in the case of the programmable PSK filter, Section 10.2, the need to make many external connections limits the practicable number of taps and their spacing; experimental devices have used up to 150 taps with inter-tap delays down to about 30 nsec, enabling signals with up to 15 MHz bandwidth to be processed. The

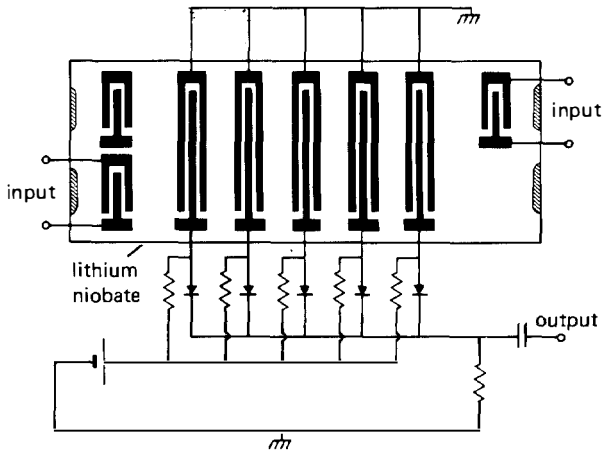


FIGURE 10.19. Diode convolver. The transducer at top left gives the output when idler wave generation is used.

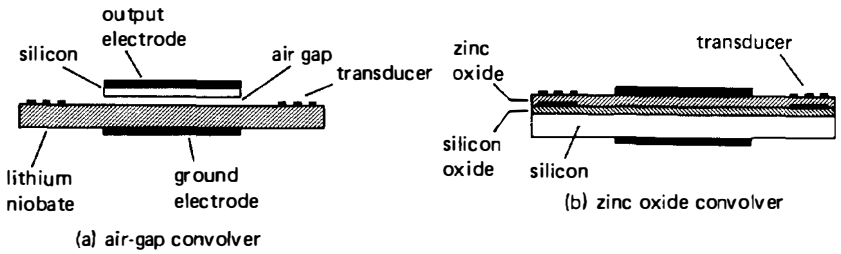


FIGURE 10.20. Semiconductor convolvers.

main advantage is that exceptionally large bilinearity factors, in the region of -30 dBm, can be obtained. Although saturation is observed at relatively low input power levels, in the region of 0 dBm, the large bilinearity factor leads to a very good dynamic range, up to 80 dB.

The efficiency of the diode convolver is affected by the magnitude of a D.C. bias current applied to each diode, being maximised when each diode is forward biassed such that its low-signal impedance is about equal to the tap impedance. The dependence on bias current can be exploited by using different currents in the individual diodes, and this leads to some additional signal processing operations. For example, a transversal filter with adjustable tap weights can be realised [418], and this can operate as a programmable bandpass filter. Alternatively, using linear chirp waveforms applied to the input transducers, it has been shown [419] that the output can give the Fourier transform of the sequence of bias currents. This contrasts with the chirp filter method of Fourier transformation (Section 9.7) in that the input data are in parallel form rather than serial.

Other devices have introduced a non-linear effect by coupling a surface wave more directly to a semiconductor, and two examples that have been investigated extensively are shown in Figure 10.20. In the *semiconductor air-gap* convolver [381, 420], which is degenerate, a silicon slice is held in close proximity to a lithium niobate substrate on which the contra-directed surface waves propagate. The electric fields associated with the surface waves penetrate into the semiconductor, depleting the surface and giving rise to a potential proportional to the square of the total field. The semiconductor must be held very close to the piezoelectric substrate since, as already seen in Section 3.5, the field decays rapidly above the piezoelectric surface. However, if the semiconductor is too close it is found to cause exponential attenuation of the surface waves, sufficient to reduce the overall efficiency of the convolver. There is therefore an optimum spacing, which in practice is typically 0.5 μm. This requirement places considerable demands on the construction of the device, and is its main disadvantage. Usually, the silicon is supported by means of a sparse array of small "posts" on the lithium niobate surface; this does not appreciably perturb the propagating surface waves. The posts are made by ion etching the niobate surface, with the required post areas masked off. Analysis of the device efficiency agrees well with experimental measurements [421, 422]. A device described by Cafarella *et al.* [423] had 100 MHz input bandwidth and 10 μsec interaction length, giving a

bilinearity factor of $C = -66$ dBm, and several other examples are quoted by Reible [422]. This device can also be used to scan an optical image focussed on to the silicon surface, making the lower ground electrode thin enough to be transparent. Several techniques, reviewed by Kino [381], have been used for this purpose, though they have not been found competitive with other technologies such as charge-coupled devices.

A similar principle is used in the *zinc oxide* convolver [424, 425] of Figure 10.20(b). The piezoelectric zinc oxide film, deposited on a silicon substrate, causes the waves to be accompanied by electric fields which penetrate into the silicon, where the non-linear interaction takes place. The film also enables the surface waves to be generated using conventional interdigital transducers. This technology has the advantage of avoiding the inconvenience of a small air gap, though care is needed in the fabrication of the zinc oxide film [424]. An example is the device of Green and Khuri-Yakub [425], which gave an impressive bilinearity factor of -44 dBm and had an input bandwidth of about 5 MHz and a 3.5 μsec interaction length. Several other types of semiconductor convolver are described by Kino [381].

10.4.2. Storage Convolvers

The semiconductor air-gap convolver described above can also be used to store acoustic waveforms by exploiting the phenomenon of charge storage in traps at the semiconductor surface [426, 427]. To obtain reproducible results it has been found advantageous to incorporate diodes in the silicon surface, as in Ingebrigtsen's Schottky-diode device [428] which is illustrated in Figure 10.21(a). Here, a negative voltage applied to the upper electrode causes the diodes to be forward biassed, with a time constant of typically 0.5 nsec. For zero or reversed bias the time constant is

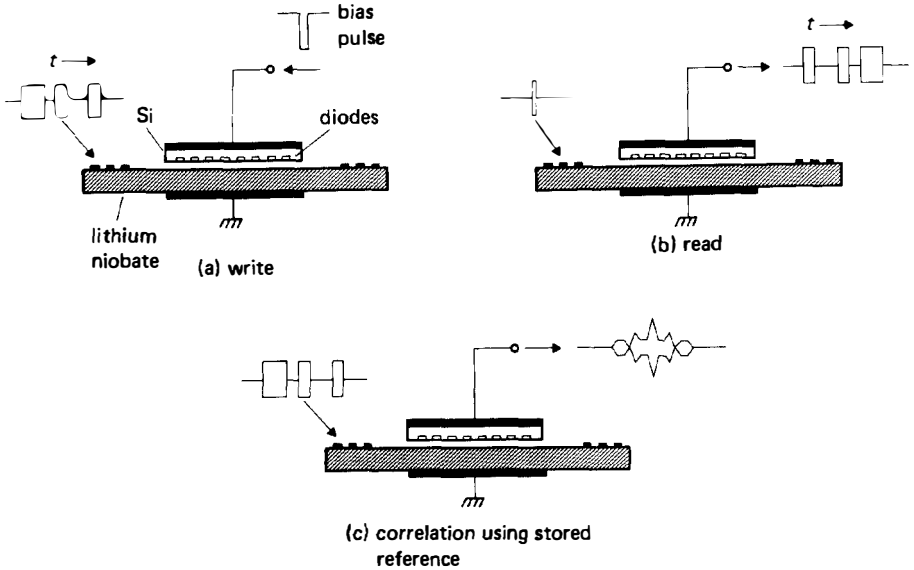


FIGURE 10.21. Air-gap storage convolver.

typically 1 to 100 msec. The waveform $f(t)$ to be stored, which must have a duration not exceeding the acoustic delay along the interaction region, is applied to the transducer at the left. When the surface wave packet is entirely in the interaction region a short bias pulse is applied to the parametric electrode. This forward-biases the diodes so that they rapidly accumulate a charge proportional to the total electric field, which is the sum of the spatially-invariant bias field and the field due to the surface wave. When the bias pulse is turned off the accumulated charges remain, reverse-biasing the diodes so that the decay time for the charges is long. In order that the stored charge pattern should be an accurate representation of the applied waveform the diode spacing must be less than half the surface-wave wavelength (as required by sampling theory), and the length of the bias pulse must be less that one half-period of the surface-wave carrier frequency.

The stored waveform can be read out later by applying a short pulse to the surface-wave transducer, as shown in Figure 10.21(b). The field associated with the travelling surface-wave pulse mixes non-linearly with the field due to the stored charge, and a waveform of the form $f(-t)$, the time-reverse of the stored waveform, emerges at the top electrode. Alternatively a forward output waveform $f(t)$ can be obtained by appling a pulse to the transducer at the right in the figure. The stored signal can be read out very many times without appreciable degradation – over 10^5 times in Ingebrigtsen's case [428].

More generally, if a waveform $s(t)$ is applied to the input transducer at the left, as in Figure 10.21(c), the output waveform $g(t)$ is the cross-correlation of $s(t)$ with the stored waveform $f(t)$, so that

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

where the infinite limits are valid because the duration of $f(t)$ is less than the delay along along the interaction region. The device will therefore correlate a received signal if the stored reference $f(t)$ corresponds to the ideal signal. In contrast with a conventional convolver, the stored reference waveform does not need to be time-reversed and the output waveform is not time-contracted. Moreover, the reference input does not have to be repeated frequently as it does for a conventional convolver (Section 10.3.1). Ingebrigtsen [428] cites a device with 30 MHz input bandwidth and an interaction region 16 μsec long, in which the 3 dB decay time for the stored waveform was about 1 msec. A linear chirp waveform with 30 MHz bandwidth and 1.5 μsec duration was correlated, using a reference stored 1 msec earlier.

Similar devices have use p–n diodes in the silicon in place of Schottky diodes, giving rather longer time constants [429, 430]. Defranould *et al.* [430] describe a device with a 3 dB storage time exceeding 1 sec, and demonstrate correlation of a chirp waveform with 12 MHz bandwidth and 6 μsec duration using a reference stored 10 msec earlier. Storage has also been obtained using zinc oxide convolvers [431, 432], with a configuration as in Figure 10.20(b) except that Schottky or p–n diodes are added in the silicon surface. For example [432], a device using Schottky diodes with an interaction region about 6 μsec long demonstrated a 3 dB storage time of 7.5 msec.

Several other techniques for storage have been investigated [433]. For example Smythe *et al.* [434] have developed a silicon air-gap device in which a charge-coupled

device is built into the silicon. A complex waveform can be entered into the charge-coupled device, and then serves as a reference to correlate a waveform introduced later in the form of a surface wave. Another method, demonstrated before the advent of the semiconductor devices, uses an electron beam to irradiate the surface of a piezoelectric instantaneously [435]. If a surface wave is present this process gives rise to a static charge distribution on the surface, corresponding to the instantaneous surface-wave amplitude. Application of a second electron-beam pulse at a later time causes generation of a new pair of surface waves, corresponding to the original input waveform and its time-reverse. Storage times of several minutes can be obtained, though the complexity of the electron beam technology makes this method unattractive for practical purposes. It has also been shown that a high-energy *optical* pulse can be used to store a surface-wave waveform on lithium niobate [436]. To read out the information the substrate is illuminated with a low-energy continuous optical beam, and a short surface-wave pulse is applied; the stored waveform then appears as modulation on the output optical beam. This method has been used to store surface-wave signals for remarkably long times, up to several weeks, though the need for short optical pulses with energies in the region of 0.1 J makes it of limited practical interest.

10.4.3. Correlation of Long Waveforms

The devices considered so far all have the limitation that they can only correlate waveforms with duration less than the acoustic delay in the device, and this cannot generally exceed about 50 μsec because of the length of substrate material required. In a spread-spectrum communication system it is necessary to correlate the waveform corresponding to one symbol which, as seen in Section 10.1, represents one binary digit of data, and in practice the symbol length can be much longer than 50 μsec. For this reason some special surface-wave techniques have been considered for correlation of waveforms longer than 50 μsec, and these are the subject of this section.

Most of the techniques considered here differ basically from the devices considered earlier in that the required integration is performed as a function of *time* rather than *space*. We have already seen, in Section 10.1, that the simple active correlator operates in this manner, multiplying the received signal by a corresponding reference waveform and applying the product to an integrator. The main disadvantage of this technique is the need for accurate synchronisation, which leads to long acquisition times. An obvious solution to this problem is to employ a *bank* of active correlators, with the reference timing incremented from one correlator to the next. The multiple delays required for the reference can be provided by a surface-wave tapped delay line, and this is the principle of the "tapped delay line correlator" demonstrated by Darby and Maines [437]. A similar principle was used by Menager and Desormiere [438], using thin strips of n on n^+ silicon as taps. In this case the reference and signal propagate as contra-directed surface waves on the same substrate, and each tap is non-linear and therefore mixes the two waves. Correlation is obtained by integrating each tap output with respect to time. However, both of these devices are limited by the requirement for a large number of external connections, as in the case of the programmable PSK filter (Section 10.2.1).

Another method is provided by the *integrating correlator* [439] in which the three basic functions required – multiple delays, multiplication and time integration – are all incorporated in the same surface-wave device. The structure of the device is essentially the same as that of the air-gap storage convolver, Figure 10.21. In the integrating correlator surface waves are introduced at each end, and each diode mixes the two waves to produce a baseband product which appears as a charge on the diode. The charge accumulates with time, so that when coded input waveforms are used each diode charge gives one point of the required correlation function. The accumulated charge distribution is read out by applying a short surface-wave pulse, using the non-linear interaction of the surface wave with the stored charge. The timing of the output correlation peak corresponds to the timing of the signal applied earlier, relative to the reference waveform.

In practice, the performance of this device has been somewhat restricted by the presence of spurious effects. Nevertheless Reible and Yao [440] have demonstrated correlation of a seqeunce of 40 bursts of signal, each with 10 MHz bandwidth and 6 μsec duration, while Ralston and Stern [441] have correlated a 10 msec waveform with 2 MHz bandwidth using a modified scheme to reduce spurious signals. Smythe and Ralston [442] have developed an integrating correlator using a charge-coupled device in the silicon, and have demonstrated correlation of a 200 μsec waveform with 20 MHz bandwidth. Integrating correlators have also been developed using zinc oxide technology, with a structure as in Figure 10.20(b) except that diodes are incorporated in the silicon; in this way Tuan *et al.* [443] correlated a 1 msec waveform with 5 MHz bandwidth. In all of the above examples, quasi-random PSK waveforms were used.

Some other methods of correlating long waveforms make use of more conventional devices. For example, acoustic convolvers can be cascaded to increase the effective interaction length [444]. A more compact method uses a surface-wave delay line in a recirculation loop, as shown in Figure 10.22. Here the loop input signal is repetitive, with repetition period equal to the delay T_d in the delay line. The loop gain is close to unity, so that successive periods of the input waveform are added coherently. In experimental demonstrations the loop input signal was obtained by correlating segments of PSK waveforms using either a PSK filter [352] or a convolver [445]. In the latter case the convolver input signal need not be repetitive; the reference waveform can be coded such that the output gives a sequence of correlation peaks, all with the

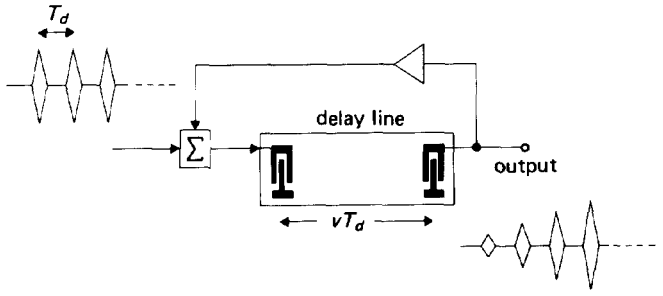


FIGURE 10.22. Recirculation loop for enhancing the SNR of repetitive waveforms.

same phase. The main limitation here lies in the delay line performance; to be effective the delay line frequency response must be very flat in the band of the signal, and spurious responses such as the triple-transit response must be well suppressed. An experimental system [445] with a 35 μsec delay line and a 30 μsec convolver was used to correlate a PSK waveform 2.1 msec long, using 70 circulations in the loop. Using an input waveform with 10 MHz bandwidth and an input signal-to-noise ratio of -41 dB, the output signal-to-noise ratio was $+4$ dB, within a few dB's of the theoretical ideal.

10.5. OSCILLATORS

In this final section we consider surface-wave oscillators, though since this book is concerned mainly with devices for signal processing this topic is considered only briefly. There are two types of surface-wave oscillator, one using a simple delay line and the other using a resonator. The resonator, which has not been described previously, makes use of reflective arrays and can serve as a narrow-band bandpass filter in addition to its oscillator application. For oscillators the important performance criteria are concerned with the frequency stability, and in this respect surface-wave devices do not perform quite as well as the best bulk crystal oscillators, at the present state of development. However, surface-wave devices can operate directly at frequencies up to 2 GHz. Bulk devices cannot generally operate above about 50 MHz, so that for higher frequencies multiplying circuits are necessary; thus at highter frequencies surface-wave devices can eliminate the need for a multiplier, and this generally leads to more compact devices consuming less power. A comparative review of bulk and surface wave oscillators, including fabrication topics affecting the stability, is given by Lukaszek and Ballato [446].

10.5.1. Delay-line Oscillator

As shown in Figure 10.23, the surface-wave delay line oscillator [447, 448] is essentially a simple interdigital delay line with an external amplifier providing feedback from the output to the input. The amplifier small-signal gain is made to exceed the insertion loss of the delay line, so that the circuit oscillates at a frequency such that the total phase change around the loop is a multiple of 2π. The transducer geometries are both symmetrical, and it follows that the delay line frequency response is non-dispersive, with delay T corresponding to the distance between the transducer centres. Thus the phase change due to the delay line is ωT at frequency ω, and the oscillation frequency must obey the relation

$$\omega T + \phi = 2n\pi, \tag{10.37}$$

where n is an integer and ϕ is the phase change in the feedback circuit. To ensure that the device can oscillate at only one frequency a narrow-band delay line is used; as shown in Figure 10.23 this is readily obtained by using a long transducer , with length comparable with vT, and it is usually convenient to thin this transducer (Section 8.3). The delay T_1 in the figure is the inter-tap delay multiplied by the number of taps. The

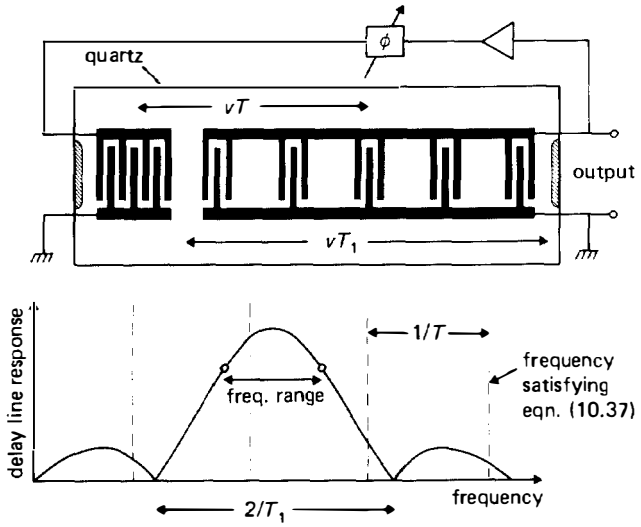


FIGURE 10.23. Delay-line oscillator. Upper: device structure. Lower: frequency selection mechanism. Open circles show typical points for unity loop gain.

amplifier gain is set such that the small-signal loop gain exceeds unity at only one of the frequencies satisfying equation (10.37), as indicated in the lower part of Figure 10.23 where these frequencies are shown by broken lines. It is quite common to include an adjustable phase shifter in the loop to enable the phase ϕ in equation (10.37), and hence the frequency, to be adjusted. This can be used to trim the frequency to compensate for small fabrication errors. Furthermore, if a voltage-controlled phase shifter is used the device becomes a voltage-controlled oscillator, and can generate freqeuncy-modulated waveforms. The substrate material is usually chosen to be ST, X quartz in view of its temperature stability.

10.5.2. Resonators

An alternative type of oscillator uses a surface-wave resonator to stabilise an oscillating circuit, and is thus analogous to the common bulk crystal oscillator mentioned in Chapter 1. A common type of resonator, shown in Figure 10.24(a) consists simply of two parallel reflectors separated by a distance L; this is the basic form for many optical resonators, and for the bulk acoustic resonator. In principle, a surface-wave version of this device could be realised using either tuned interdigital transducers (Section 4.4.4) or multi-strip mirrors (Section 5.5) as the reflectors. However, for these cases the reflection coefficients obtainable are not sufficiently close to unity to give good resonator Q-factors. This is because a small amount of loss arises from the film resistivity, and also from bulk wave excitation associated with the abrupt discontinuity at the edge. A much more effective approach is to use reflective gratings, shown in Figure 10.24(b), as first suggested by Ash [449]. The grating

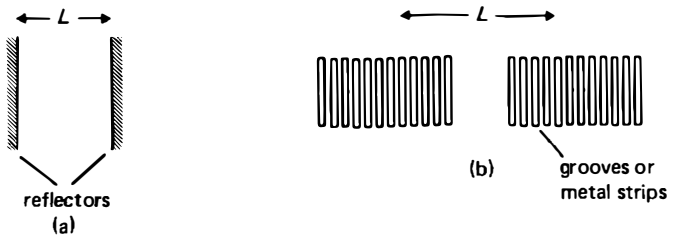


FIGURE 10.24. (a) Basic resonator structure. (b) Surface-wave resonator using reflective gratings.

elements are generally metal strips or grooves, each causing very little perturbation of the wave, and in consequence the loss to bulk waves is very small.

The reflection coefficient of a grating is theoretically given by equation (E.24) of Appendix E, and some examples showing its variation with frequency are given in Figure E.3. The reflection coefficient is close to unity, but only over a narrow frequency band, if $N|r| \gg 1$, where N is the number of elements and r the amplitude reflection coefficient of each element. Typically, $|r|$ will be 0.01 or less, so 100 or more elements are needed. The use of gratings implies that a surface-wave resonator behaves in a rather different manner to the conventional resonator of Figure 10.24(a) since there are now two frequency-selective mechanisms involved – the frequency variation of the grating reflection coefficient, and the cavity resonances. Usually, the device is designed such that all but one of the cavity resonances are suppressed by the frequency variation of the reflection coefficient. Another distinction is that the phase of the grating reflection coefficient varies rapidly with frequency, so that the wave behaves as if it were reflected from a point some distance into the grating; thus the effective cavity length L is considerably larger than the separation of the gratings, as indicated in Figure 10.24(b).

For practical devices, *ST*, *X* quartz is usually chosen as the substrate material in view of its temperature stability, and grooves are usually chosen at the reflecting elements since they are found to give better performance than metal strips. The behaviour of grooves reflecting surface waves through 90° has already been considered in Section 9.6.2, in connection with RAC's. For normal incidence on grooves in *ST*, *X* quartz the behaviour shows similar features: the groove reflection coefficient is found to be given by equation (9.91) with $C = 0.27$, and stored-energy effects are found to perturb the velocity somewhat in accordance with equation (9.99), with $C' = 17.3$ [450, 451]. It is found that the resonator Q-factor, measured by generating a surface wave outside the resonator, can be remarkably high. For example, at 160 MHz a Q-factor of 27,000 can be obtained [452], while a 1.4 GHz resonator gave a Q-factor of 6000 [453]. These figures are quite typical of well-designed resonators, and are substantially higher than the values obtainable using other technologies suitable for the same frequency range [454]. Comparisons with theoretical predictions show that the Q-factors are quite close to the limit imposed by the acoustic propagation loss which, as seen in Section 6.3, increases with frequency.

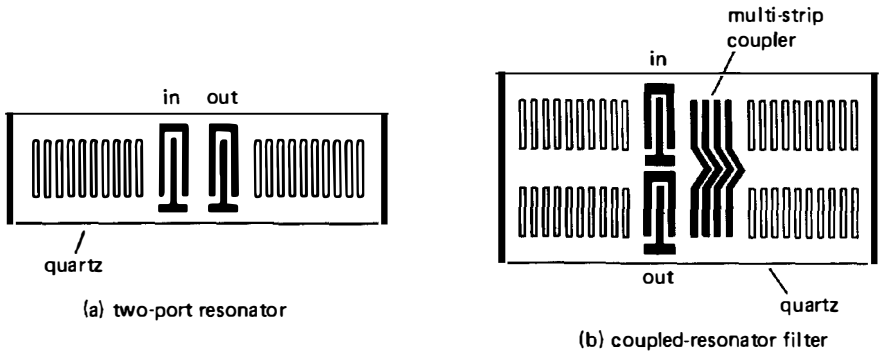


FIGURE 10.25. Resonator configurations.

A practical resonator must of course include one or more transducers to launch and sense the waves. A common arrangement, shown in Figure 10.25(a), uses two transducers between the gratings, forming a two-port resonator. Apart from its use as an oscillator stabilising element, this device can also be regarded as a narrow-band *bandpass filter*, and has been widely investigated in this role [452, 454, 455]. Compared with the interdigital bandpass filter of Chapter 8, the resonator is complementary in that it offers smaller bandwidths, and also has the significant advantage that the insertion loss can be as low as 2 or 3 dB. However, the resonator is somewhat inflexible in that the response is essentially a simple resonance curve, and in particular the stop-band rejection is poor. As in the case of bulk crystal resonators, greater versatility can be obtained by combining several resonators [456], and these may be either separate devices combined in a circuit or several acoustically-coupled resonators on the same substrate. Some experimental examples using the circuit approach [452, 455] show in particular that the pass-band can be flattened and the rejection improved by this method. Acoustic coupling of resonators on the same substrate can be obtained by a variety of methods [454], one of which is illustrated in Figure 10.25(b). Here two resonators on a quartz substrate are coupled by means of a multi-strip coupler; although the coupler cannot efficiently couple two tracks on quartz because of the low piezoelectric coupling (Chapter 5), it can be used effectively in the resonator because only weak coupling is required. A variety of experimental filters are described by Coldren and Rosenberg [454], using several different coupling techniques and up to four coupled resonators.

10.5.3. Oscillator Performance

In both the delay line and the resonator oscillator the short-term and long-term stability are important performance criteria, and any assessment of surface-wave oscillators must take account of their main rival, the well-established bulk crystal oscillator. The stability obtainable using surface-wave techniques is not generally as good as that of the bulk device, but a key advantage is the ability to operate at high frequencies, up to about 2 GHz. Bulk crystals cannot operate in the fundamental

mode above about 50 MHz, because they become too fragile. For higher frequencies overtone operation may be used or the output may be applied to a frequency-multiplying circuit, but in both cases the short-term stability is degraded. Thus some care is needed when comparing surface-wave and bulk-wave devices.

The short-term stability is related to noise generated in the circuit, and can be predicted theoretically quite well for both delay line [448, 457] and resonator [457, 458] oscillators. For optimum stability a high Q-factor is needed (or, for a delay line, a long delay), the insertion loss should be low, and the power applied to the device should be relatively high. For resonators it has been found that too high a power level causes migration of the aluminium metallisation and degrades the long term stability [459]; the power level must be restricted to avoid this, though the phenomenon can be mitigated to some extent by adding a small amount of copper. The short-term stability is often expressed in terms of the power spectral density of the oscillator output, relative to the carrier. Results for a 400 MHz delay line device [457] gave a power density of -140 dBc/Hz at 10 kHz from the carrier and a floor of -160 dBc/Hz at 100 kHz and beyond. For frequencies within about 5 kHz of the carrier the stability is affected by "flicker noise", which is dependent on the surface treatment of the crystal and cannot be predicted accurately. Resonators give better stability, as shown by 120 MHz devices giving about -165 dBc/Hz at 10 kHz and -180 dBc/Hz at 100 kHz and beyond; these results are comparable with bulk crystal oscillators [458]. The short-term stability is also affected by vibration [460].

The long-term stability, that is, the ageing, is known to be related to the treatment of the crystal in preparation, mounting and packaging. These factors are not amenable to analysis, but considerable effort has been applied to optimise the performance, as discussed by Parker [461]. With care taken in fabrication, the frequency of a delay line oscillator can be constant within one part per million over one year, and some of Parker's devices are rather better than this. For resonators, ageing rates of about 3 parts per million per year are reported [459].

Trimming of surface-wave oscillators is an important issue because fabrication tolerances do no generally allow the frequency of the initial devices to be more accurate than about 100 parts per million. Here the delay line oscillator has a distinct advantage: the frequency range accessible by external phase shifting is easily increased by increasing the delay-line bandwidth, and hence the design can allow for anticipated tolerances. For resonators the range obtainable by external phase shifting is limited because of the high Q-factors; however it has been found that accurate trimming over quite large ranges can be done by etching the device with an ion beam [461, 462]. Parker points out that some frequency uncertainty is introduced when the device is subsequently encapsulated, but this can be compensated by external phase shifting [461].

The temperature stability of these devices is mainly determined by the substrate. For *ST, X* quartz, the commonest choice, the temperature stability is given in Section 6.4, and is not as good as the stability obtained with bulk crystals. However, other orientations giving better stability have been found [463], including the SST cut (Section 6.5), while considerably better stability can be obtained using surface-skimming bulk waves in quartz, as described in Appendix F. For delay line

oscillators there is an interesting alternative technique, in which the stability is improved by using two acoustic tracks at different orientations on the same substrate [464]. The principle has been applied to a digitally-compensated oscillator using an AT-cut quartz substrate [465], demonstrating an impressive stability of 5 parts per million over a temperature range of -13 to $+97$°C.