Paparella: Volume I: Basic Sciences and Related Principles

Section 2: Physiology

Part 1: Ear

Chapter 5: Physics of Sound

Juergen Tonndorf

The following account is written for people whose background is not in physics and who, therefore, have little if any training in mathematics. Mathematics will be used with respect to only one point (*impedance*), but it will be very elementary, not going beyond simple calculus. Those desiring a more formal treatment are referred to other textbooks, such as are listed in the bibliography.

Sound is a form of physical energy. The ear and the larynx are mechanical devices (receiver of sound and generator, respectively); in this respect they must both obey the appropriate laws of physics, i.e., of *acoustics,* as the particular subfield is called that deals with sound.

A good cause can be made for the fact that the larynx operates like certain types of wind instruments. However, to simply compare the action of the ear to that of a microphone is quite misleading. Microphones are so designed that they only measure, but do not disturb, an existing sound field, and especially so that they do not draw acoustic power from it. In contrast, the ear does consume acoustic power, although admittedly in quite small quantities.

It has been customary for a long time to describe sound in terms of tones (or mixtures of such) of very long durations, mainly because of certain powerful mathematical descriptions that can be applied to such cases. However, sounds in our everyday experience are usually not of the latter type. More typically, they are short, of the order of 0.1 sec, and often much shorter; for example, speech sounds in running speech. Therefore, we must also consider short-lasting, so-called *transient,* sounds. In counterdistinction, the long-duration sounds are then referred to as *steady-state* events.

It may be noted in this connection that the information-carrying capacity of steady-state sounds is essentially nil. It takes sound signals that change rapidly with time and have little predictability to transmit information.

The Decibel (dB)

Before entering the discussion of acoustics proper, we must define the *decibel (dB),* a measure that has found widespread acceptance not only in acoustics but also in electrical engineering, optics, and many other fields. As will be pointed out in detail later, sound signals may vary in intensity from low to high, corresponding to variations in loudness from soft to loud. Intensity is a physical measure and is defined as the average rate of energy flow through a unit area (1 cm^2) at a given point in a specified direction. Its unit is the watt/cm^2. Since power = energy per sec, it may also be considered a measure of "power density".

In acoustics, one is frequently interested in mere relative changes, ie, how much more intense one given sound is compared with another. Moreover, more often than not, effects such as the attenuation afforded by a wall are independent of the actual level. Note that the term "level" refers to absolute power values in watts/cm^2, but expressed in dB; there are also energy levels, sound pressure levels, and so forth. Typically, such a wall may attenuate sound to one-tenth its original value. If the original value is Lo and the attenuated one La, we may have

Lo	10,	100,	1000 ...	
La	1	10	100	(I)

In all three cases, the result is a *ratio of 10.* We can rewrite the values of L as

Lo	10^1	10^2	10^3 ...	
La	10^0	10^1	10^3 ...	(II)

We see then that the ratio, being 10^1, can be conveniently expressed by its exponent, 1 in the present example. These exponents, as is well known, are *logarithms to the base of ten.* This is the basis of the decibel notation, which also contains a scaling factor of 10:

$$N_{dB} = 10 \log (I'/I'') \quad (1)$$

The result is a number (N), since the ratio between two values (I-intensity) that have the same dimension (eg watts/cm^2) is dimensionless.

In actual practice, sound is more conveniently measured in terms of pressure than in terms of intensity. Sound pressure (p) is given as dyne/cm^2 (ie force per area), and is related to sound intensity (I) as in

$$p^2 = I \quad (2)$$

Therefore,

$$N_{dB} = 10 \log (p'/p'')^2 = 20 \log (p'/p'') \quad (3)$$

Hence, a ratio of 10 in terms of intensity equals 10 dB, whereas the same ratio in terms of sound pressure equals 20 dB; doubling an intensity value means an increment of 3 dB, whereas doubling a pressure value means an increment of 6 dB. The 5-dB step of a clinical audiometer is thus seen to represent almost a doubling of the sound pressure value. The short table (III) gives a few representative values.

dB	10	20	30	40
Sound intensity ratios	10	100	1000	10,000
Sound pressure ratios	3,162	10	31.62	100

(III)

The acceptance of the decibel was aided by another fact. Communication engineers had found quite early that the range of intensities between the hearing threshold (the level at which one barely hears a sound) and those sounds painful to the ear is quite large. In the middle frequency range (1000 Hz to 3000 Hz), in which a normal ear is most sensitive, this range is close to 1:10.000.000. Moreover, it was commonly believed until recently that the response of the ears varies as a logarithm of the signal magnitude, the so-called *Weber-Fechner law*. We know now that this law is only an approximation, and that the responses of all senses vary as power functions of signal magnitude whereby the exponent of such functions is a characteristic of each particular sense - the power law of Stevens and of Plateau. Engineers had already found it convenient at the time when radios were first introduced to use volume controls that were logarithmically rather than linearly tapered. Such controls are nothing but variable resistors. In this manner the user has the impression that he is actually controlling "loudness" in a linear manner.

If one wishes to express the results of a given measurement (of a noise level, for example) in absolute terms but still use the dB, one simply has to form the dB ratio between the power (or sound pressure) level in watts/cm^2 or dynes/cm^2, respectively, to a given *standard reference level*. For many years, the values of these reference levels have been set at 10^{-10} watts/cm^2 or 0.0002 dyne/cm^2, respectively. The arbitrariness of such reference levels is reflected by the fact that in underwater acoustics a level of 1 dyne/cm^2 is used as a reference. Still other values are used in audiometry, where the zero reference level is determined as a sort of an average threshold for healthy young ears. It varies with frequency to account for the frequency dependence of human hearing. The 1951-ASA (American Standards Association, now ANSI (American National Standards Institute)) levels have in 1969 been replaced by the 1964 ISO (International Standards Organization) levels, which differ slightly, but consistently, from the former.

Simple Harmonic Motion

Sound is usually produced by structures that are set into vibration by mechanical, electromagnetic, or a host of other means.

The simplest sustained tone that can be produced - for example, by a tuning fork - is based upon very uniform, pendulum-like motions of its tines. If we plot such to-and-fro motions against time, we obtain a *time course*, a sections of which is shown in Figure 1. This is called a *simple harmonic motion* (for a reason that will become obvious presently) or, because of its derivation from *sine* (or *cosine*) functions, a *sinusoidal waveform.* In a true sinusoidal wave, the time duration of one cycle, the *period*, is rigidly maintained. One cycle may be counted from any given starting point until the waveform has returned to the same point having the same tendency of motion. For example, starting at the zero line on the left of Figure 1, the trace goes upward to the positive maximum, downward again to zero, passes the negative maximum, and returns to zero. Such count could be started at any other point. The numbers of cycles per second is a measure of the frequency (unit: Hertz (Hz)). The closest psychophysical equivalent to frequency is pitch.

The excursions of the tracing from zero in either direction represent the *displacements* of the vibrating structure from its resting position. The height is called amplitude. Starting from the zero line at any point in time, this may be the *instantaneous* amplitude, and it may be either positive or negative; at the point of maximum displacement it is known as the *peak* amplitude; that from a positive peak to a negative one, the *peak-to-peak* amplitude.

One another important measure is the so-called *root-mean square* (RMS) amplitude. It represents a statistical average and is registered by many measuring instruments. For sinusoidal events, it is equal to the peak amplitude divided by sq root of 2, ie approximately to 0.707 times the value of the peak amplitude. The importance of RMS values lies in *power* considerations. A unidirectional (DC) current of electricity passing through a wire dissipates energy in the form of heat. It is clear that an alternating (AC) current that changes direction, reducing even to zero at regular intervals (like the tracing in Fig. 1), heats the wire to a lesser extent. The RMS value the designates the magnitude of a DC current that has the same heating capacity as an AC current of a given peak value. Since power considerations are principally alike in electrical, mechanical, and other physical processes, the RMS concept has universal validity. The nearest psychophysical equivalent to the displacement amplitude of a sound generator (at least for a given frequency) is *loudness.*

Superposition of Sine-Wave Events

A given structure may be subjected to more than one sinusoidal event at the same time. Obviously, the structure cannot execute two different vibrations at the same time, but will follow their *resultant* from instant to instant. We will limit consideration to two such events applied simultaneously and will start out with a special case in which both are of the same frequency.

Phase. In order to determine the resultant, we must know in what part of its cycle one event happens to be with respect to the other. Because of their derivation from sine functions (see previous discussion), the tracing of a full "cycle" may be said to represent the circumference of a circle with a radius equal to the peak amplitude. Thus, each point along the cycle, or its projection upon the time axis, may be expressed in terms of an angle between zero and 360 degrees, starting at any point - a zero-crossing, for example. Such an angle, which then determines uniquely the relative state of the vibratory event, is called a *phase angle.* Since the circumference of a circle may also be expressed in *radians*, one can also employ multiples or fractions of pi to quantify a phase angle (2 pi radians - 260°).

Figure 2 gives two pairs of events that have different phase relationships. It is seen that for either pair the phase relation determined at any point in time holds for any other point. This arises from the fact that we had assumed that both events have the same frequency. In such a case it takes exactly the same time for either event to complete one cycle or any fraction thereof. In Figure 2(a), the phase relation is 90 degrees (1/2 pi, with event B *leading* event A, or event A *lagging* behind event B. In Figure 2(b), the phase relation is exactly 180 degrees (1 pi); lead or lag cannot be determined in this case. The latter situation is often referred to as *phase opposition.* If there is no phase difference, the events are said to be *in phase* (0° phase angle).

Once the phase angle is determined, the resultant waveform of the two events can be determined by adding their instantaneous amplitudes, so-called *linear superposition*, with proper regard to their signs in the manner of Figure 3. Of special interest are cases a and b. In both of them, amplitudes of the primary events are equal. When the phase angle is 0 degrees (case a), the resultant R has twice the amplitude of each single event since amplitude is simply doubled at all times (*reinforcement*). When the phase angle is 180 degrees (case b), the resultant R is a straight line, amplitudes of the primary events being equal and of opposite signs at all times (*cancellation*). In all other cases (for example, case c), that is, when phase angles are neither 0 degrees nor 180 degrees or when amplitudes of the primaries are not equal to each other, there is partial reinforcement or partial cancellation. For the case of equal amplitude of the primary signals, the limiting phase angle is 120 degrees. If the latter is smaller than this value, there is partial reinforcement; if it is larger, there is partial cancellation.

It is noted that all resultants, with the exception of the case of complete cancellation, are also sine waves having the same frequency as the primary events.

Beats. If we now allow the two events to have different frequencies, although such differences should be small (500 Hz and 510 Hz, for example), an interesting phenomenon develops. In contrast to the cases just discussed, the phase relation between the two events does not stay put, but alters continuously. Inspection of Figure 4C indicates that whenever a period of time elapses that is the reciprocal of the difference of the two signal frequencies (10 Hz = 1/0.1 sec in the present example) a given phase relation will repeat itself, only to change once more at the next instant. The superposition of two such events must lead to a waveform such as given in Figure 4 D. Its amplitude no longer stays uniform; it fluctuates in a sinusoidal manner at a rate once more equal to the difference between the signal frequencies. Furthermore, as can be seen from Figure 4, the waveform of the resultant has a period different from that of either of the two signal frequencies. It is equivalent to their average (1/2(500 + 510) = 505 Hz in the present example).

When the amplitudes of the two primaries are equal, there is one instant of complete cancellation followed by a later one at which amplitude is exactly twice that of each primary (see Fig. 4 D). This had to be expected from the results of Figure 3, that is, from the principle of linear superposition. Since beats are heard especially well under this latter condition, they are known as "best beats". When the amplitudes of the primaries are not equal, neither cancellations nor reinforcements are complete. Moreover there are slight variations of frequency in addition to those of amplitude. However, the resultant beats sound less distinct to the ear.

The example of Figure 4 concerned two primaries that were only slightly apart in frequency. Musicians who tune their instruments while listening to the disappearance of beats call this type an *imperfect unison.* Physicians refer to them as simple beats.

Beats may also appear between primaries that are not quite in *harmonic relationship* (discussed later), that is, when their frequency ratio is not quite an integral number such as 1:2, 1:3, 2:3, and so on. For example, the combination of 500 Hz and 1010 Hz leads also to a 10 Hz beat. The correct harmonic relationship would be 500 Hz and 100 Hz (1:2). This is then known either as a *mistuned consonance* or as *complex beats* in musical or physical terminology, respectively.

Complex Harmonic Motion. Finally, we must examine the waveforms resulting from superposition of events having frequencies that are related to one another by exact integral numbers, eg 500 Hz and 1000 Hz (1:2); 500 Hz and 1500 Hz (1:3); 500 Hz and 750 Hz (2:3); 750 Hz and 1000 Hz (3:4); and so on. Beats, obviously, cannot occur in such cases. Although in the case of 1:2 relationship, for example, one primary waveform completes two cycles at the same time the other one completes only one, this relationship is strictly maintained at all times, and there is no gradual shift in phase. Figure 5 shows the waveform resulting from such a combination for two different phase relationships between the primaries. In contrast to the simple sinusoidal waveform of Figure 1, those of Figure 5 are known as *complex* waveforms. To be sure, Figure 5 still appears relatively "simple", but when several frequencies are superpositioned, the resultant waveform will soon become quite complex, especially since its shape depends not only upon the number of components but also upon their relative strengths and their phase relationships (Figs. 5 and 6).

In music, such whole number relationships play a special role. For example, the 1:2 ratio represents an octave, the 2:3 ratio a fifth, the 3:4 ratio a fourth, and so forth. For this reason, the term *harmonic relationship* has been in use for a long time. In a given harmonic series, one refers to the lowest common divisor (eg 100 Hz of the series, 100, 200, 300, 400, ... Hz) as the *fundamental* or the *first partial.* The 200 Hz component would be the first overtone or the second partial, and so forth. The terms "fundamental" and "overtones" are most commonly used in musical terminology, whereas "partials" is employed more commonly in physical terminology. The term "basic frequency" is occasionally employed for the first partial. We realize now that the relationship between beating primaries is an *inharmonic one.*

Fourier Analysis. Actually, the *synthesis* of complex waveforms as described in the foregoing has become possible only relatively recently after suitable devices had become available. The knowledge that complex waveforms show a *periodicity,* that is, the periodic repetition of a characteristic waveform, however complex, is somewhat older. This important discovery is credited to the French mathematician Fourier. *Fourier analysis* is a powerful mathematical tool and is used today in many different fields, not only in acoustics. Fourier first described his theorem for the problem of heat transfer in 1811. Its application to acoustics, and particularly to the performance of the ear, was first suggested by G. S. Ohm in 1843. This is known as *Ohm's law of acoustics,* in contrast to his better known *electrical law.*

The point must be stressed that for any given waveform there is only one solution in terms of the contained frequencies and their amplitude and phase relationships. It is then possible to record this information in the form of a *spectrum*, ie relative amplitude versus frequency (including phase information). In such a plot, each frequency appears as a line of a given height. For an example, see Fig. 6. All complex sounds, including inharmonic sounds, have *discrete line spectra.* It is a curious fact, however, that as one gains information about the frequency composition of a given complex sound one loses sight of its waveform. There is no denying that by experience one learns to some degree to associate some waveforms of lesser complexity with their spectra, but that is an *acquired* faculty not one that is inherent.

Either notation, waveform or spectrum, is complete and unique. When information is being stored for future playback (phonograph records, recording tape) it is more convenient to record waveforms. For purposes of analyzing the performance (potential or real) of electroacoustics systems, Fourier transformation is an indispensable and powerful tool. Ohm, as just mentioned, had postulated that the ear performs a Fourier analysis on incoming sound. This hypothesis was based upon the observation that the ear can differentiate complex tones to some degree, a faculty that can be improved by training. Moreover, the ear distinguishes musical instruments and voices partly by recognizing their characteristic *timbre* or *quality.* The word "timbre" is used in musical terminology and the word "quality" in physical terminology. Instruments do not produce pure tones (ie simple sinusoidal sounds); their timbre depends upon the number and distribution of the higher harmonics they invariably contain. This is so not only for given types of instruments, but also for instruments of the same kind, making it possible, for example, to distinguish high-quality violins from those of lesser quality. The same difference in timbre applies to human voices, ie to typical sopranos, altos, tenors, and basses.

Generation of Sinusoidal Vibrations

We may now raise the question of the factors that let a tuning fork or a similar instrument execute its vibrations. As everyone knows, one has only to strike a tuning fork once in order to activate it. It will then vibrate for quite a while, say 1 to 2 minutes, at slowly diminishing amplitudes. The latter point will be ignored for the time being.

Events in mechanical systems are governed by their physical properties and by forces that are either external or inherent. When the tines of a tuning fork have been displaced by an external force (eg after having been struck by a rubber mallet), a force is evoked that tends to restore the fork to its previous equilibrial state. In most instances, such *restoring forces* are given by the *elastic* properties of the material, their magnitude being governed by Hooke's law, ie that the resultant stress equals the strain. This law holds only for relatively small displacements. Beyond certain limits, things get more complicated, as we will see later. Under the effect of the elastic, restoring force, the tines return to their resting position. However, in doing so they acquire velocity. Because of their inherent mass, this means an increase in their *momentum* (mass x velocity = momentum). The momentum is highest, but the restoring force lowest, just when the tines reach their resting position. Therefore, the movement is carried right through that point, initiating a displacement of opposite sign. In turn, the growing displacement evokes a restoring force, also of opposite sign, slowing down the motion that eventually, at some maximal amplitude, comes to a standstill. Now the elastic force takes over, tending again to bring the tines back to their equilibrial position. Once more the momentum increases and the whole event repeats itself, only running in the opposite direction.

It is clear then that the vibration is maintained by alternate effects of the elastic restoring force and the momentum. If the event is sinusoidal, both vary sinusoidally with time. The elastic force reaches it maximum at the point of maximal displacement, and its minimum (it actually becomes zero) in the resting position. The momentum, on the other hand, is zero at the point of reversal, that is, the position of maximal displacement, and maximal when the tines pass through their resting position. Inspection of Figure 7 indicates that such a condition is fulfilled when and only when the phase relationship between the displacement or the restoring force on the one hand and the velocity or the momentum on the other is 90 degrees, with the velocity leading the displacement.

The rate of velocity is not uniform. In other words, acceleration and deceleration alternate with each other. Both are highest when the tines come to their standstill at the point of maximal displacement and then begin again in the opposite direction. This means that displacement and acceleration reach their respective maxima at the same time but are in phase opposition. The phase relationship between the displacement, the velocity, and the acceleration of a sinusoidal vibratory event is depicted in Figure 7. It may be mentioned here that all three entities have "amplitudes". Thus, to avoid confusion one should always specify an amplitude of displacement, or of velocity, or of acceleration.

Tuning forks are known to maintain their frequency very precisely. In fact, some modern chronometers employ small tuning forks as time-keeping devices. Thus, the frequency is a built-in feature of these forks; it is known as their *natural* frequency. The latter does not depend upon the amplitude of vibration or the force with which the tuning fork is struck. It depends upon (1) the elastic coefficient of the material (the stiffer, the higher the natural frequency) and (2) its mass (the heavier, the lower the natural frequency). Obviously, then, the natural frequency is related to the two factors that maintain the vibrations, the elastic restoring force, and the inertia (f0 = natural frequency; E = elasticity; M = mass):

f0 = (1/2 pi) sq root ((E/M). (4)

Instruments like tuning forks start their vibrations very easily, "resonating" with another nearby fork of the same frequency. Therefore, their natural frequency is also known as the *resonance frequency.*

Damping. Earlier in this chapter we disregarded the fact that tuning forks do not actually maintain their vibratory amplitude but show a gradual decrement with time. This decrement results from the effect of *friction,* both external (eg air resistance) and internal (eg friction within the crystal lattice of the material). Any form of movement, one directional (DC) or alternating (AC), is opposed by friction, its magnitude being proportional to the velocity of such movement. Thus, it leads the displacement by 90 degrees in phase (Fig. 7). Its effect on vibratory events is known as *damping.*

Since the velocity of sinusoidal vibrations is always in some proportion to their displacement amplitude (see previous discussion), it is clear that the absolute magnitude of the effect of damping decreases proportionally as the displacement amplitude is diminished. Expressed differently, in a given event the effect of damping is always at a fixed ratio to the displacement amplitude. It is only at the start or at the end of such an event that it becomes appreciably higher as a result of the effect of *static* friction. This fact will make the assessment of damping a relatively simply task. All one has to do is to measure the amplitude ratio of any two successive periods. This ratio is called the *decrement* and defines damping uniquely. In practice, the logarithmic decrement is usually used, that is, the natural logarithm of the above ratio. The higher the damping, the less the number of cycles a vibratory event goes through before coming to a complete stop. Figure 8 shows the effect of such variations in damping. It covers a range of slightly more than one order of magnitude.

It is evident that the damping of tuning fork is very much less than that of any of the examples shown as its decrement is much slower. The value of delta = 6.8 is of special interest. It is exactly at this damping value that a structure, after having been displaced, is unable to execute even a single vibratory cycle. Instead, it is returning asymptotically to its resting position. This damping value is known as *critical damping.* If damping is even higher (overdamping), the return to the resting position is executed in a creeping manner, making its duration longer than that of a half cycle of the corresponding vibratory event. This retarding effect of damping makes itself already felt at a level below critical damping. It is for that reason that the natural frequency of a given system goes slightly with increased damping as we shall see later (see Fig. 10).

Free Vibrations, Maintained Vibrations, Forced Vibrations. As was just mentioned, tuning forks struck once vibrate for a long time. Thus, by definition, their damping must be very low. Since they receive but on initial impulse and none thereafter, their vibrations are an example of so-called *free vibrations.* The latter, as we know from the example of tuning forks, occur always at the natural frequency of the system in question. There are ways of driving tuning forks continuously, for example, by having one tine of a steel fork act as the arm of an electromagnetic interrupter. Forks of this or of similar construction were used extensively as generators of steady tones before the advent of electronic instruments, since they can be made to maintain their amplitude at any desired level. Such vibrations are known as *maintained vibrations.* Finally, an earphone or a loudspeaker can be driven at many different frequencies, not only at their natural frequencies. Otherwise, they could not reproduce speech signals or music, both of which change rapidly in frequency with time. Such vibrations are known as *forced vibrations.* When the driving force is withdrawn, there is usually a very brief period of free vibrations before the system comes to rest.

Laryngeal Voice Production. One generator that is of special interest to otolaryngologists is the larynx. Simultaneously, it may serve as an example of how vibrations are generated by a unidirectional force. During expiration, when voice production usually takes place, the air current flowing through the organ is unidirectional, but the vocal cords vibrate in an alternating mode. The underlying process is as follows: The laryngeal muscles must first position the vocal cords, preferably in the midline position, and put them under proper longitudinal tension. This tension provides the "tuning" of the cords. A chest contraction now increases the air pressure in the subglottic space. This pressure, finally, overcomes the muscular opposition and forces the glottic chink open. At this moment, an air current gets under way and the subglottic pressure decreases accordingly. Owing to the latter change, the vocal cords approximate each other once more. This latter event is aided by the fact that an air current flowing through a narrow channel (where current velocity and channel width are reciprocally related) exerts a negative pressure upon the channel walls, sucking them into the channel (the so-called Bernoulli effect). After the glottis is closed once more, the subglottic pressure rises again, and the next cycle is started. The restoring force in this case is given partially by muscular tension and partially by the Bernoulli effect. The momentum is that of the moving air. Both factors are again seen to be 90 degrees out of phase with each other. Varying the muscular tension of the vocal cords varies the frequency produced.

Actually, it is not necessary that the vocal cords be tightly approximated. Even when the glottis is open but the vocal cords are tensed and the rate of the expiratory air current is increased, the vocal cords are set into motion by the air current flowing through the glottic chink. This is another application of the Bernoulli effect leading once more to an oscillatory motion. In this latter case, a voiceless "whispered" sound is produced, whereas the output in the first case if of the voiced type.

The resultant waveform of the vocal cords in either case is not sinusoidal, although for a voiced output it is periodic when sustained sounds are produced. In the first case (closed glottis = voiced output), there are brief pulses corresponding to the opening phases separated from another by longer lasting quiescent intervals corresponding to glottic closure.

It is clear then from waveform considerations that the output of the larynx is not a pure tone but a complex tone. In the case of an unvoiced output it is essentially a noise, which we will define later. We have yet to describe the role of the respiratory tract in voice production (see the section on filtering of sound later in this chapter).

It may be noted that the present brief account of laryngeal voice production follows the so-called *aerodynamic* concept. The so-called *neuromuscular theory* that was in vogue briefly in the early 1950s is untenable from the acoustic standpoint. The aerodynamic function of the larynx has great similarity to the working mechanism of brass musical instruments - bugles, trumpets, and the like. The role of the vocal cord is then played by the lips of the player.

Another way of converting a DC motion into AC vibrations is shown by the action of a bow upon a violin string. The horsehair of the bow is covered with resin to make it sticky. When the bow is drawn over the strings with some force, the friction resulting from the sticky resin will take the string along for a short distance. After a while the elastic restoring force must overcome the driving force and the string returns fast to its resting position and usually beyond it. The resulting waveform is of the so-called *sawtooth* type, that is, the first slope (that corresponding to the forced motion) is more gradual than the second one (that corresponding to the return motion). Thus, in this case the output is not sinusoidal. In general, generators of this kind are known as *relaxation oscillators.*

Acoustic Transients

With the exception of the larynx in its unvoiced mode, all generators described so far produce simple harmonic motions or complex harmonic motions. The latter, as we know, are nothing but combinations of sine waves according to the theorem of Fourier. However, as was stated in the introduction, long-lasting vibrations (a basic requirement of true sine-wave motions and their analysis by Fourier series) are not the rule in the production of sound. More typically, sounds are short lasting. Consider, for example, all plosive speech sounds. Even the voiced consonants and the vowels of running speech are not really "long lasting".

We have seen that a generator that has very little damping, a tuning fork, for example, will execute free vibrations for a long time after a stimulating force is withdrawn. It is evident, then, that short-lasting sounds can be produced only by generators that have high damping. If one taps the diaphragm of a good quality loudspeaker, ie one that is well damped, one hears a short "plop". Although the output of a smaller speaker activated the same way sounds somewhat higher in pitch (such sounds are generally referred to as "clicks"), it is hard, and practically impossible, to assign real pitch values to such short-lasting sounds, ordering them along a musical scale, for example. Their character is not that of a tone but rather of a short-lasting *noise.*

Fourier Analysis. When forming the Fourier spectrum of such transient sounds, one will find a fundamental difference with respect to sine-wave events. Spectra of long-lasting periodic events were said to consist of discrete lines, each line representing one sine-wave component. Spectra of short-lasting events are *continuous*, consisting of the frequency *bands* of varying widths. Mathematically, the two cases differ also from each other. With steady-state events, one forms a *Fourier series,* ie one considers one single period of an event that is understood never to change and to last infinitely long. With transients, one forms a *Fourier integral*, ie one considers the event as a whole. The relation between time duration (delta tau) and frequency band width (delta f) is given by the following equation:

$$Dt \times Df = 1. \qquad (1)$$

In other words, the shorter the duration, the wider the band width, and vice versa. For a hypothetical transient of zero duration, the band width would be unlimited. Figure 9 contains some examples of transients. It shows that a short section of a sine wave does not actually have a line spectrum; it has a composite band spectrum, and as its duration is increased, the component around the nominal frequency, f0, gains more and more prominence. A true line spectrum is not established until the duration becomes infinite. For this reason, the point was made repeatedly in the foregoing discussion that genuine sine-wave events are of very long (theoretically, infinite) duration. Actually, the ear does not note any change in quality of a "tone" once it is held longer than for a few seconds, a fact justifying the use of the word "sinusoidal events".

It is noted that the two events - a pulse of zero duration, having an infinite band width, and an infinitely long sine wave, having a band width of zero cycles - are reciprocally related to each other, representing the two extremes of equation 5. What is the waveform of one case represents the spectrum of the other.

In extreme cases (Fig. 9), transient signals may become *aperiodic,* that is, their waveform does not even show a single zero crossing, as shown in the middle tracings of Figure 9, for example. Such events are termed *aperiodic,* since there is no periodic repetition.

When a sinusoidal event is suddenly started or ended or even when only the amplitude is stepped up or down from one level to another, one hears a click; that is to say, the sudden change involved produces a transient. Once more, the longer the transient, ie the more gradually one lets the change take place, the narrower is the bandwidth of the transient. It is this relationship that makes it possible to eliminate the click as an audible signal. Band-width elimination usually takes place at the expense of the high frequency portion of the band, and it is in the higher frequencies (around 3000 Hz) that the human ear is most sensitive. Toward low frequencies, its sensitivity falls off with about 12 dB/octave. Thus, increasing the duration of the transient by letting the signal grow (or decline) gradually will shift its acoustic energy into the low frequency region, making it essentially inaudible. For this reason, the ANSI standardized audiometers specify the duration of signal onset ante decay (the signal envelope).

Acoustic transients produced during the onset or termination of tones have importance in another respect. They are characteristics of different musical instruments. Earlier, the quality of the tones produced was said to be different from instrument to instrument, and this quality was said to depend upon the number, relative amplitude, and distribution of higher harmonics. Actually, it turns out that the initial (or terminal) transients are even more important criteria for telling one instrument apart from another when both of them are playing the same note.

Interestingly enough, when the initial transient was removed from the recording of notes played by various instruments and the mutilated recording presented to a panel of musical experts, they were unable to tell instruments apart, such as cello and a trumpet, which ordinarily (ie when their transients are present) are easily distinguished. Another way of demonstrating the importance of the initial transient is to listen to a tape recording of piano music while the tape is being rewound. Piano notes, being produced by a mallet hitting a string, have strong initial transients. When played backward, the tones cannot be recognized as those of a piano at all.

Noise. So far we have considered only single transients. The question is what type of sound will be produced when such transients are repeated. There are two possibilities: (1) The transients, eg pulses, are repeated in a periodic manner. In that case the spectrum is of the discrete line variety, having numerous harmonics. The repetition rate gives rise to the fundamental. The output of the human larynx producing a sustained sound in the voiced mode as was described earlier may serve as an example for this case. (2) The transients are repeated in a random manner, ie their repetition rate, their duration, and their amplitude are completely randomized. In that case, when there is no periodicity, the spectrum remains of the broad band type, and what one hears is a "whooshing", noiselike sound. The hissing of steam and, above all, the sound of jet engines are good examples. Because of their broad frequency content, the analogy to white light has suggested the term *white noise*, an expression that has found wide acceptance. Because of its ability to mask tones of any frequency, white noise is used in audiometry for masking purposes. Actually, since it is only a narrow band of frequencies around that of the test tone that is required for masking (the so-called critical band), more recently narrow bands of white noise are being employed. The total energy the patient is exposed to depends upon both amplitude and frequency band width. Thus, by limiting the band width the patient is exposed to less sound energy and has a lesser overall loudness sensation.

Impedance

We have learned that free vibrations (and also maintained vibrations) are associated with low damping and with narrow tuning; that is to say, devices such as tuning forks are capable of vibrating only within a narrow range around their natural frequency. Forced vibrations for which the frequency can be varied over a wider range (broad tuning) are associated with higher damping (usually less than critical).

When a tuning fork is driven at its natural frequency (maintained vibrations), a minimal effort is needed. In other words, the fork offers only a small "opposition" to such a driving signal. However, if one tries to drive the same fort at a frequency that is only moderately different from its natural one, the opposition to such an effort has risen very sharply. This opposition, which is thus shown to be frequency dependent, is known as the *impedance (Z)* or, more specifically, the mechanical impedance (Zm) in the case under consideration.

Mechanical Impedance. Resistance, in the sense just described, is only one component of the mechanical impedance. It is designated by the letter *R*, and its a frequency-dependent entity. The other frequency-dependent component is known as the *reactive* component and is designated by the letter *X*. Like the natural frequency, it is determined by the elastic and inertial properties of the vibrating structure. The elastic force was said to be proportional to acceleration. Since these two factors are in phase opposition with respect to each other (see Fig. 7), their resultant effect can be determined simply by forming their numerical difference. The total reactance is thus either mass dominated (by convention then considered positive) or stiffness dominated (then considered negative).

Once more, according to Figure 7, the phase relation between the two reactive factors (corresponding to displacement and acceleration, respectively) and the resistance (corresponding to velocity) is 90 degrees. Numerically such two components can be added by *vectorial summation*, which for this special case follows the well-known theorem of Pythagoras, ie

$$Z^2 = R^2 + X^2. \qquad (6)$$

In the case of a simple vibrating system (for example, a tuning fork, in which all masses, elastic components, and frictional factors may be conceptually lumped into one mass, etc, each) the impedance (Z_m) may be written in detail as follows:

$$Zm^2 = R^2 + (2 \text{ pi } fM - (E/2 \text{ pi } f)^2 \qquad (7)$$

(f = frequency; M = mass; E = elasticity; R= resistance). As was the case with the definition of the natural frequency (equation 4), it is seen that frequency is inversely related to mass and directly to the elastic property, We now solve equation 7 for the case of the natural frequency by substituting equation 4 into it and obtain:

$$Z_m^2 = R^2 + (M \text{ sq root } (E/M) - E \text{ sq root } (M/E))^2. \qquad (8)$$

The reactance part (X) of equation 8, after appropriate simplification, can be rewritten as:

$$X = \text{sq root } EM - \text{sq root } EM = 0. \qquad (9)$$

In other words, when the frequency is equal to the resonant frequency, the reactive component becomes zero, and the impedance is determined solely by the resistive component. This explains why tuning forks offer a minimal impedance when driven at their natural frequency, as was stated previously.

It also follow from equation 7 that for all $f > f^0$, the impedance increases with the mass of the system as frequency goes higher; whereas for all $f < f^0$, it increases with the elasticity as frequency becomes lower. In other words, it is mass controlled above the natural frequency and stiffness controlled below that point. In either case, it is always higher than that at the exact point of resonance.

Tuning forks were said to be narrowly tuned, and this was found to be associated with low damping. Inspection of equation 7 indicates that when the resistance R is small, the reactance X becomes dominant, making the impedance Z_m strongly dependent upon frequency f.

The way one assesses such a situation quantitatively is to determine, frequency by frequency and for a constant input, the output of the system under consideration. The results are then plotted as amplitude versus frequency. Figure 10 gives some examples of such *frequency response curves*, as they are called. In particular, Figure 10 shows that when damping is low, the system is very narrowly tuned. As damping is increased, the tuning becomes broader and broader. Thereby, the resonant point moves slightly to the left, as was mentioned previously. When damping is critical, the curve does not display a resonant point anymore, sloping gradually as frequency goes higher. In the region of the former resonant point, it is already 6 dB down from its initial value. The flattest curve (which is obviously the most desirable from the standpoint of an optimal transducer) is reached when damping is somewhat less than critical, specifically at a logarithmic decrement of 3.14. Above the resonant point, all curves, regardless of damping, slope down approximately with the square of frequency. This fact follows once more from equation 7. When mass *M* becomes dominant, and elasticity *E* can be neglected, the impedance *Z* must vary with f^2, resistance *R* then being a constant for a given decrement. In order to have a good high frequency response, one has to push the resonant point of a system as high as possible, and this means low mass and relatively high stiffness.

The mechanical impedance may be defined as the complex ratio (complex because of the phase relationship between the resistive and reactive components) between the effective force acting upon a given area and the resulting linear velocity of displacement through that area. Its unit is the mechanical ohm (dyne x sec x cm^{-1}). The term "effective" in this context is the same as the root-mean-square (RMS) value described previously.

Acoustic Impedance. So far we have considered only the mechanical impedance, which manifests itself when a system is driven mechanically in some form or another. The situation is slightly, but not principally, different when a system is considered that is driven acoustically or has an acoustic output. In both of the latter cases, one can determine the *acoustic impedance,* either "looking in" (system being driven) or "looking out" (system having an output). The unit is the acoustic ohm (dyne x sec x cm^{-5}). The exact definition is "the complex ratio between the effective sound pressure averaged over the surface of the system to the effective volume velocity through it".

The reason one may be interested in the outgoing (looking out) impedance lies in the problem of *impedance matching* with the system the acoustic energy is being fed into (eg the surrounding air). The problem of impedance matching will be discussed later.

Characteristic Impedance. There is a third kind of impedance that is exhibited by large (theoretically unbounded) media, eg air or water. The mass of such large bodies is, at least theoretically, infinitely large. Consequently, according to equation 4, their resonant frequency must approach zero Hz; that is to say, this type of impedance, which is known as the *characteristic impedance,* is frequency independent. The unit is the mechanical ohm/cm (dyne x sec x cm^{-3}). It is more simply determined as the product of the sound velocity (c) within the particular medium (see later) and its density (ro),

$$Z = c \times ro. \qquad (10)$$

The characteristic impedance of air is quite low. At 0°C and at barometric pressure of 760 mm Hg it is 42.86 ohms/cm, varying slightly with atmospheric pressure and with temperature. At 20°C, it is approximately 41.5 ohms/cm. For water, the characteristic impedance is much higher. At 25°C, it is 149,000 ohms/cm. In sea water (salinity of 3.6 per cent), also at 25°C, it is 157,000 ohms/cm, the increase being due to the minerals in solutions. In solids, it reaches its highest value. In saline, for example, it has a value of 4,570,000 ohms/cm. As its definition (ohms/cm) implies, the total value of the characteristic impedance increases linearly with the depth of penetration of the energy into a given medium.

Sound Transmission

Before continuing our discussion on impedance and impedance matching, we must first describe how sound is being transmitted within a given medium, especially through air. Consider a loudspeaker diaphragm vibrating under the effect of an electrical signal applied to its voice coil. The speaker will radiate energy into the surrounding air. The underlying mechanism is as follows: as the speaker diaphragm moves out, it pushes the adjacent air particles forward, thus raising the air pressure very slightly above its static value, typically one atmosphere. As the diaphragm moves back again, and then beyond its resting position, it pulls the adjacent air particles with it, thus creating an equally slight decrease of the air pressure. In other words, alternating compressions and rarefactions of the surrounding air are being set up. These latter disturbances do not stay put within a continuous medium but move away from their source. Let us for a moment consider the speaker as a point-shaped source radiating in all directions, a hypothetical case known as a "pulsating sphere". For such a case, the two-dimensional pattern created along the surface of a large pond into which a stone has been dropped provides a good illusion. Circular wave crests emerge from the site of excitation and, as they move outward away from the source, their diameters continually increase with time and distance. Each crest is followed by a trough and that, in turn, by another crest, and so forth, until eventually a whole system of concentric crests and throughs is established, continually moving away from the source.

The speed of propagation of such a disturbance, the velocity of sound in our case, is uniform and, as we have seen, a characteristic property of the medium in question. Incidentally, this is not correct for surface waves on water; their speed of propagation increases with distance.

There is another difference between sound waves in air and other types, and that concerns the *mode* of particle motion. We have seen that acoustic energy is propagated through air in the form of alternating compressions and rarefactions. As schematically shown in Figure 11 (top), these pressure changes take place in the same plane as that in which the waves are being propagated. Therefore, this mode is known as a *longitudinal* form of particle motion. A *transversal* mode (Fig. 11, bottom) is found in the propagation of light if one considers its wave property. Polarization means to restrict the transversal motion to one particular plane. Incidentally, surface waves on water move transversally and longitudinally at the same time, so-called *trochoidal* wave motion. That means a particle, as it is bobbing up and down, is simultaneously moving to and fro.

The parameters of Figure 11 are amplitude and distance, indicating that an event that is sinusoidal in time (see Fig. 1) propagates also as a sinusoidal system of waves. The distance from crest to crest, or from any other point to its nearest equivalent, is called a *wave-length.* The information conveyed by Figure 11, top and bottom, is essentially the same, except for the difference in the mode of particle motion. For purposes of illustration, the transversal form (Fig. 11, bottom) is usually preferred both for temporal and spatial representation. It is more easily drawn and recognized.

Sound Velocity. The propagation velocity of sound can be measured. In air it is relatively slow, varying somewhat with temperature. At 20°C (68°F) for example, it is 344 meters (1120 ft) per sec. Everyone has made the following observation: when watching a man from a distance who is cutting a tree, one *first sees* him swinging his ax and sometime *later hears* the impact. In water (fresh water of 30°C (68°F)), the velocity is higher (as was also the characteristic impedance), namely 1493.2 m/sec (4554.3 ft/sec), that is, roughly four times the velocity in air. In steel, it is about 16 times that in air, or four times that in water, namely, 5000 m/sec (16,200 ft/sec).

The following two statements are pertinent to the present problem: (1) When forced vibrations are being propagated, their frequency remains constant; and (2) within a given medium the wave propagation is independent of frequency. It follows from those two statements that in a given situation the wavelength (lambda) must be in a reciprocal relation to the frequency (f) with respect to the propagation velocity (c):

$$c = f \times \text{lambda} \qquad (11)$$

In air, then, wavelengths of sound waves between 100 Hz and 10,000 Hz vary from 3.44 m (11.2 ft) to 3.44 cm (1.36 in). Compared with the wavelength of visible light (4000 to 8000 Å), the wavelength of sound waves is much larger because of the difference in propagation velocity.

Inverse Square Law. The example of a stone being dropped into water may help to illustrate another point. We may say that each crest carries a certain amount of energy away from the source. Disregarding frictional losses for the time being, we see that this amount of energy is spread thinner and thinner as the circumference of the crest increases with distance from its source. It is recalled that intensity is a measure of "power density". Since the circumference of a circle varies as the square of the radius, it follows that the *"power density"* must vary as the square of the distance. This *square-of-the distance law* is a good approximation of the attenuation of sound in a *free field* situation, ie, when there are no obstacles in the way of a propagating sound wave. Even when the distribution is not strictly spherical, it still holds reasonably well.

Reflection of Sound, Diffraction, and Refraction. When the medium is not unbounded. as would be the case in all real situations, the energy cannot be propagated away from its source in an unlimited manner. For the time being we are still disregarding frictional losses. Sooner or later it will meet an obstacle - a wall, for example. In such a case, some of the energy will be bounced of the obstacle, ie, it will be reflected very much like a beam of light is reflected from a given surface under comparable circumstances. Reflections of sound produces echos. The same principles apply to acoustic as well as to optical reflections, ie, the angle of reflection must be equal to the angle of incidence and so forth. Moreover, it is the ratio of the magnitude of the surface roughness to the wavelength that determines whether the energy is reflected as a beam or is scattered (ie diffracted). To be optically flat, a surface must be planed and polished to a high degree of perfection. To be acoustically flat, tolerances can be less by many orders of magnitude. On the other hand, it is the ratio of the size of the obstacle to the wavelength that determines whether all energy is reflected or some of it is *refracted* around the obstacle.

For visible light, obstacles that cause refractions are very small; for example, dust particles suspended in the air. For sound, obstacles must be much larger to cause refractions. Moreover, the range of wavelengths of audible sound (3.4 cm to 3.4 m for a range of 100 to 10.000 Hz; see above) is several orders of magnitude larger than that of visible light (4000 to 8000 Å), which represents only a range of a factor two. Consequently, for sound waves there must be a pronounced frequency-dependent effect with respect to diffraction and refraction. Low frequencies (long wavelengths) are more easily refracted, ie "heard around corners", than high frequencies. The latter are reflected in toto and, hence, prevented from going around corners. In general, it is the magnitude of the wavelengths of sound waves that causes refraction and diffraction to such a degree that it is virtually impossible to "beam" sound, except at very high frequencies.

Standing Waves. There is one special phenomenon we must mention. Whenever the path lengths between the walls of a room, into which sound is being fed, are an integral multiple of the wavelength, a sort of spatial resonance phenomenon is set up. The reflected sound returns in a direction opposite to that of the incident sound. At a hard reflecting surface, the sound pressure is maximal at the wall and changes phase abruptly by 180 degrees, so that the incident wave and the reflected one are in phase opposition. This does not lead to a cancellation of amplitude as in the time-domain situation of Figure 3, but to a *cancellation of wave travel.* The result is a system of so-called *standing waves.* For the sake of illustration, they can be easily reproduced in a bathtub. At some point along their pathway the sound pressure changes with time from maximal positive to maximal negative (this value being twice that of the incident wave alone) in a sinusoidal fashion. The first such point is located directly at the wall. With distance from the wall, these points alternate with others at which the pressure is held constant at its resting value. The latter points are known as pressure *nodes* and the former as pressure *antinodes.* The particle velocity is shifted by 180 degrees with respect to the pressure. That is, antinodes of sound pressure correspond to nodes of velocity, and vice versa.

Standing waves are easily noticed in an enclosed room into which a high frequency sound is being transmitted, for example, the 9 kHz interference tone produced by two radio stations on adjacent frequency bands. On hearing such a tone one has only to move one's head slightly (lambda = 3.8 cm; 1/2 lambda = 1.9 cm) in order to go from an antinode where the tone sounds loud to a node where it may vanish altogether. Standing wave situations may also be used to tell whether a given microphone responds to sound pressure or to its velocity. Both types are being used. Such microphones will indicate in an opposite number, either directly on the wall (where there is a velocity antinode) or at 1/2 lambda in front of it (where there is a pressure antinode).

Strange as it may sound, reflection (or formation of standing waves if conditions are right) takes place not only when the walls are hard (eg when aerial sound hits the water surface), but also when they are completely yielding (eg when sound propagating in water hits a water/air boundary). In the latter case, there is a pressure node at the wall, ie a velocity antinode, the latter changing phase by 180 degrees on reflection.

In all other cases, when only some energy is being reflected and some is admitted into the second medium, *partial* standing waves will develop.

Impedance Matching. The condition for reflection is given whenever there is a large ratio between the impedances on both sides of the boundary, eg 41.5 ohms/cm to 149.000 ohms/cm for the case of an air-water boundary. The point must be made once more that it does not matter whether the energy goes from the medium of the lower impedance to that of the higher, or vice versa. A large impedance ratio, which is of course detrimental to the transmission of sound energy from one medium to another (or from one structure to another), is referred to as an *impedance mismatch.* One way of assessing the degree of impedance match or mismatch is to measure the ratio of the incident energy to the reflected energy and their relative phase angle. Especially when the impedance of one system is known, that of the other can be determined in this fashion. *Impedance bridges* that are built upon this principle have been in use for some time. One special, and quite useful, application is in the clinical assessment of the impedance of the tympanic membrane.

One might think of compensating for the loss at the boundary by delivering a large amount of energy to ensure that a sufficient amount is transmitted into the second medium in spite of the high rate of rejection at the boundary. The example of the air/water boundary serves to illustrate the futility of such an effort. In this particular case, approximately 0.1 per cent of the incident energy is admitted into the second medium, while 99.9 per cent is being reflected. For small mismatches, say of the order of 1:2 or 1:3, such strategy may be defensible, but with large mismatches it is simply too wasteful to even be considered.

There is a better way, one that does not waste energy, ie, the use of an *impedance-matching transformer.* Electrical power transformers that serve exactly the same purpose in electrical or electronic circuits are well-known illustrations of this method of transferring power in analogous situations. Mechanical and acoustic transformers are also used. The middle ear is a good example of a mechanical transformer. Its operating principles rely upon simple leverage systems.

Acoustic transformers are employed by musical instruments and also by loudspeakers. The horn of a trumpet is an impedance-matching device from the mouth-piece (high impedance) to the surrounding air (low impedance). Here, the matching is achieved by the gradually increasing inner diameter of the horn, which flares out in an exponential fashion, that is, an increasingly wider volume of air is being driven as distance increases in the mouth of the horn. Megaphones, bullhorns, and horn-type loudspeakers apply the same principles. A man cupping his hands in front of his mouth improves the transmission of his voice in the same manner.

Acoustic Radiation Patterns. An acoustic horn may achieve an efficient power transfer from a loudspeaker into the surrounding air. Yet most conventional loudspeakers create a "hole-in-the-wall" effect from which the sound appears to emanate. To make matters worse, speakers do not radiate all frequencies evenly in all directions. As long as the wavelength is large compared to the diameter of the speaker, the radiated pattern is evenly distributed as a result of the phenomenon of refraction (see earlier discussion). As this ratio becomes smaller with increasing frequency, the pattern becomes gradually pointed like a beam. In addition, side lobes begin to form so that a listener moving perpendicularly to the speaker's axis goes repeatedly from maximum to minimum, back to a maximum, and so forth. For the same reason, a complex tone sounds different from different listening angles. One way to overcome this deficiency is to mount smaller *high frequency tweeters* "co-axially" into the larger *woofer*. The woofer then produces a broad low frequency radiation and the tweeter (or tweeters) does the same for the high frequencies.

There is another point we must consider. Loudspeakers have definite low frequency limitations. With respect to the power radiated by them, we are concerned only with the *resistive* components of their impedances. It is only in resistive elements that power is dissipated, ie, in the form of heat. When one starts out at a high frequency (although at one below the resonant point) and goes down in frequency, the resistive component of the output impedance of a given speaker is reasonably flat down to a point that is determined by the ratio of wavelength to the circumference of the speaker (lambda/2 pi R). At this point, it drops precipitously, ie, with the square of inverse frequency (12 dB/oct).

The so-called *reciprocity theorem* of Helmholtz states that mechano-acoustic and other similar events, in which no power is permanently lost, ie, dissipated, may also occur in a *reversed* manner. Loudspeakers can act as microphones, for example. From this viewpoint the consideration concerning loudspeakers has importance for the case of the tympanic membrane, which receives acoustic power to transmit it to the inner ear. As with the power radiated off by loudspeakers, and essentially for the same reason, the power admitted by the tympanic membrane becomes less with inverse frequency below a cut-off frequency of approximately 2000 Hz, a phenomenon that accounts for the well-known low frequency attenuation of the threshold curve of hearing. This fact is not at all obvious when one neglects power considerations, ie, when one treats the ear as if it were simply a microphone that does not consume power. As was briefly mentioned in the introduction, microphones are analogous to voltmeters in electrical circuits. Neither of them must disturb the existing situation by drawing power. They measure sound pressure (microphone) or voltage (voltmeters), respectively. Sound pressures and voltages are analogous to each other.

Displacement Pattern of Vibrating Membranes. There is one additional point we must discuss with respect to vibrating membranes such as the tympanic membrane. For an example we shall consider the flat circular diaphragm of a magnetic earphone, the reason being that this is the only type of membrane that has been studied in sufficient detail both mathematically and experimentally. The membrane shall be assumed to be clamped around its edge. As long as $f \leq fo$, the membrane will vibrate as a whole, bulging, of course, at its center.

Beyond their first resonant points, such systems usually display a number of additional resonant points. At the first additional one, the membrane ceases to vibrate as a whole, moving instead in opposite phases on its right and left, as shown in Figure 12. The membrane is said to have changed its *mode* of vibrations, this particular one being the first *radial* mode. As further shown in Figure 12, there are a large number of higher modes, some radial, some circular, and some combinations of these two. Figure 12 is by no means exhaustive in this respect. It is evident, then, that when the membrane ceases to operate as a whole, its efficiency must drop. This is especially true for the radial modes. As a matter of fact, the resonance points for radial modes are typically very small, those of the first mode and of the second circular one being usually the highest. This is an additional reason for the fact that the efficiency of a membrane when acting as a radiator of a sound or as a receiver must drop beyond its first resonance point. Principally, the same considerations apply to the tympanic membrane.

Distortion: Amplitude Distortion

It was mentioned in the section on generation of sound that the elastic restoring force is proportional to the displacement only as long as displacement amplitudes are relatively small. If such limits are exceeded (limits that will, of course, differ from system to system), the linear relation between the applied force and the resulting displacement will no longer be maintained. This is equally true for unidirectional (DC) as well as alternating (AC) forces. Figure 13 shows what is known as an *input/output function* of a typical case. Both axes are given in dB. It is seen that at higher driving amplitudes the output increases at a lesser rate than the input until it finally becomes independent of the input, ie, it becomes flat. Thus, the function, which was originally linear, eventually becomes *nonlinear.* Since the result is a *distorted waveform* (as will be shown presently), the process is also known as distortion or, more exactly, *nonlinear distortion.* The particular kind described here is known as an *amplitude distortion* because of the dependence upon signal amplitude. Distortion in the *frequency* and *time* domains (ie frequency and phase distortion as well as "hangover" effects) will be discussed later.

Harmonic Distortion. Suppose the input has a simple sinusoidal waveform. After the limits of linearity are exceeded, the output waveform may become "distorted", as shown in the examples of Figure 14. While in Figure 14 A there are only two little wiggles opposite each other along either slope, in B the peaks on either side of the waveform are clipped off, so-called bilateral *peak clipping.* In C, the peak clipping occurs only unilaterally in addition to further alterations of the waveform. Peak clipping means, of course, a definite limit of displacement amplitudes. In Figure 13, bilateral peak clipping would occur in the flat portion of the input/output function. Unilateral peak clipping is seen when the displacement is limited in one direction long before the other one becomes affected. In a badly designed loudspeaker, for example, the voice coil may hit the bottom of the slot it is moving in but may still have leeway in the opposite direction.

We already know what a distorted waveform means. When it is analyzed according to the Fourier theorem, it will reveal the presence of *higher harmonics that were not part of the original signal.* For this reason, amplitude distortion is also known as *harmonic distortion.* For example, the two wiggles opposing each other in waveform A of Figure 14 indicate a relatively strong third harmonic. Needless to say, the situation becomes more complex as the driving amplitude gets higher.

Intermodulation Distortion. An additional form of distortion is observed in nonlinear systems. Ordinarily, one can feed any two sinusoidal signals into a linear system, and the resultant is simply a linear superposition of the two, ie, the sum of the amplitudes of both systems at any instant of time, in the manner of Figures 3 and 4. An analysis reveals nothing but two original signals. In contrast, when fed into a nonlinear system, the two signals interact, that is, the lower signal *modulates* the higher one in a way similar to the way a program signal modulates the carrier signal of an AM radio station. In this latter case, new frequencies are created, the so-called *combination tones:* the sums and differences between the primaries (first order combination tones) but also second order sums and differences. Let the primary frequencies be 300 and 1000 Hz. The first order sum and differences are 1000 ± 300 = 1300 Hz and 700 Hz. Second order combinations are 1300 + 300 = 1600 Hz; 700 - 300 = 400 Hz; and so forth. There is no limit to such a process, although the relative magnitude of these distortion products decreases rapidly with rank order. This type of distortion, which has basically the same cause as harmonic distortion, is called *intermodulation distortion.*

It cannot be emphasized enough that any system will eventually show distortion after certain limits are exceeded. Commercial power amplifiers, for example, are rated in watts, that is, the amount of power they are capable of dissipating. It is customary to list the magnitude of distortion occurring at the level of maximal power rating. The result can be conveniently expressed as the percentage energy of the original signal that goes into the production of higher harmonics or, similarly, as per cent intermodulation distortion.

One can, of course, hear the results of either form of distortion. Most people when listening to running speech or music find any distortion exceeding 1 to 3 per cent very objectionable. The ear is quite sensitive in this respect, a fact that is somewhat strange since the ear has a low threshold of distortion of its own.

Distortions in the frequency and time domains are a different matter.

Frequency Distortion

Actually we have already touched upon frequency distortion without mentioning it by name. Whenever a receiving system has a usable frequency range (given by its frequency response curve as defined earlier) that is narrower than the range of signals it is supposed to transmit we call this *frequency distortion.* AM radio stations with their 9-kHz band width "distort" music by cutting off its high frequency components. Hearing aid earphones also have a limited frequency range. There are many more examples of this kind. An ideal system for good reproduction of sound should have at least a flat frequency response from 20 to 20.000 Hz. The significance of the higher frequencies lies not so much in the reproduction of high frequency tones (there are no musical tones that reach that high) but in the correct reproduction of fast transients that require wide frequency bands (see earlier discussion). It is for this reason that piano music is difficult to reproduce in good quality in any but the best systems.

Distortions in the Time Domain

Phase Distortion. Phase distortion is actually a corollary of frequency distortion. In other words, a band-width-limited system cannot reproduce complex signals with the same intercomponent phase relationship they originally had. Throughout the frequency range of such band-limited systems, phase change continually. The reason for this will be discussed later in the section on filtering of sound. Formerly the ear was thought to be insensitive to phase distortion. This erroneous conclusion was based upon a misinterpretation by other writers of one of Helmholtz's original experiments. A change in phase relationship between components of a complex signal affects the waveform (see Fig. 5), and the ear is indeed sensitive to such changes, a fact that was well known to Helmoholtz.

Hangover Effects. In systems that have insufficient damping, each signal is followed by some free vibrations, an effect that lends a "mushy" sound to reproduced speech. Since these hangover effects may interfere with subsequent signals, intelligibility must decrease. Most poor quality systems suffer from this kind of distortion to varying degrees.

Noise Interference. There is no system in which signals are transmitted that does not have *some noise,* either from internal or external sources, although the noise may be quite low with respect to the magnitude of the signal. Since there is interest only in the relative amount of noise, one defines the *signal-to-noise ration.* In fact, a branch of psychophysics called *signal detection theory* considers the signal S/N ration one of the most important determinants of the detection of auditory signals by listeners. However, noise must not necessarily be equated with "distortion". It exists in linear systems as well, and then its effect is a mere *interference,* ie, the noise is superimposed upon the signal. Psychophysically the result is masking, which will not be discussed further in this chapter. If, on the other hand, the system acts nonlinearly, intermodulation distortion occurs, and the results may be very detrimental to the perception of the signal. Cheap tube-type radios with an audible 60-Hz power line hum often show intermodulation distortion between the hum and the signal, especially when the signal is weak and requires a large amount of amplification. In that case, of course, such a radio distorts.

Filtering of Sound

A limitation of frequency band width can be introduced deliberately in a process that by analogy (sorting out of particles beyond a given size) is called *filtering.* There are *low-pass* filters, *high-pass* filters, *band-pass* filters, and *band-rejection* filters. Filters, depending upon their construction, attenuate frequencies outside their pathband to varying degrees. Very simple filters may only attenuate 6 dB/oct, whereas very sharp filters attenuate 30 to 40 dB/oct and more. Although with today's instrumentation it is easier to convert acoustic signals first into electrical ones and then use electric or electronic filters, mechano-acoustic filters are still in use.

Let us consider one common example each of low-pass and high-pass filters and, finally, a variable filter, the upper respiratory tract.

Low-Pass Filter. A car muffler is designed as a low-pass filter. Since the ear is more sensitive to high frequencies (peak sensitivity around 3 kHz) than to low ones, shifting the energy into the low frequency band must reduce the apparent loudness of the exhaust noise. Practically, this filter effect is achieved by having a large number of small, completely closed side chambers along the main duct. These side chambers act like springs, cushioning the sharp impact of the exhaust pulses, that is, limiting their high frequency content.

It may be added here that all transmission systems are essentially low-pass filters. High frequency energy is lost as a result of absorption and scattering, and low frequency energy is transmitted farther.

High-Pass Filter. Suppose one is listening via the telephone to somebody talking. As long as there is good contact between one's ear and the receiver, the voice quality is good. However, if one moves the receiver only slightly away, thus breaking its seal with the ear, the speaker's voice becomes "tinny". It is lacking in low frequency components. One has produced a high-pass filter effect. The explanation is as follows: at the leak the line is "loaded" with masses, ie, the volumes of air that must be moved to and fro through such leaks as the pressure changes when the signal is passing along the line. Mass effects, as we recall, always affect the high frequencies. The cut-off point of such a high-pass filter system is determined by the size of the holes. The resistance to sound varies with particle velocity. Thus, high frequencies are more easily dissipated than low frequencies.

Filter Properties of the Upper Respiratory Tract. The steady-state output of the larynx was said to be either a noise (whispered voice) or a periodic pulse (normal voice). First of all, this generator can be switched on and off, and such switching can be accomplished in a sudden manner (vocal attack) or in a more gradual manner (breathy attack). Thus, the initial and terminal transients can be manipulated. But then the output of a human larynx does not sound very pleasant. It sounds rough, ie, highly distorted, which is not surprising in view of its waveform. As a matter of fact, it does not sound like the voice of a human being at all. What determines the final sound quality is the action of the upper respiratory tract, which, from the acoustic standpoint, may be considered a variable set of filters, mainly of the low-pass and band-pass types. The nasopharynx, for example (a side chamber) serves for low-pass filtering. The mouth, being in essence a resonating cavity in a series, represents a band-pass filter. The cut-off frequencies of both of them can be varied over a fairly wide range simply by changing the enclosed air volume. In addition, changes in the widths of the passages, mainly the throat and the mouth, affect the damping properties as a narrow channel increases the resistance to the passing air stream. This, in turn, affects the slopes of the filters.

The mouth especially, by virtue of the great mobility of the tongue and lips, can change its filter properties in a rapid manner and to an almost unlimited degree, allowing the generation of a great variety of speech sounds. This includes the production of transients as in plosive consonants.

The way the upper respiratory tract affects the laryngeal output is by favoring some frequency bands and attenuating others as a result of its filter properties at a given time. This is what makes the voice of a trained speaker or singer pleasant sounding. In this latter respect, the lower respiratory tract also has its effect. The upper and lower tracts and the larynx form what is known as a coupled system in which one part affects the performance of the others.

This brief account of voice production is by no means exhaustive and is given here only as an illustrative example.

Phase Effects of Filters. Filters not only affect the frequency response but shift the phase/frequency as well. It is for this reason that frequency distortion is invariably accompanied by phase distortion as was outlined previously.

In transmission lines, which in general act like low-pass filters, the inherent phase distortion produces another effect. The transmission velocity decreases with frequency so that the high frequency components of a complex signal arrive at the terminal noticeably later than the low frequencies. Before it was learned how too compensate for this occurrence (simply by delaying the low frequencies proportionally), this phenomenon was the source of the so-called "birdies" heard in long-distance telephone lines, that is, high frequency signals that had become entirely separated from their original signals.

Magnitude Considerations

In the discussion of the dB concept, the reference levels for sound measurements, 10^{-10} microwatts/cm^2 or 0.0002 dyne/cm^2, were cited (1 dyne/cm^2 = 1 microbar = 0.001 millibar, the unit in which barometric pressure is measured). The reader has perhaps wondered about their small magnitudes; 10^{-10}, after all, is 1/10 of 1/1,000,000,000! These values were originally chosen because they are reasonably close to sound intensity and pressure, respectively, at the hearing threshold of a human ear at 1000 Hz. Actually, at the time the agreement was made, it was thought that they represented this threshold accurately. The level of conversational voice in an enclosed room is approximately 60 dB SPL (sound pressure level); this is still only 10^{-4} microwatts/cm^2 or 0.2 dyne/cm^2 (0.2 mbar is less than a barometer is capable of indicating). One hundred-twenty decibels is a noise level that may already be considered as fairly high; but it is a mere 100 microwatts/cm^2 or 200 dynes/cm^2. These two examples may suffice to show that the power and pressures involved in acoustics are quite small compared to those involved in other fields, such as electricity, mechanics (automotive power), and others. By the same token, they indicate that the ear is an extremely sensitive detector of sound. The displacement amplitudes of structures in the ear, such as the tympanic membrane, the stapes, and the basilar membrane, are submicroscopic even at fairly high sound pressure levels.