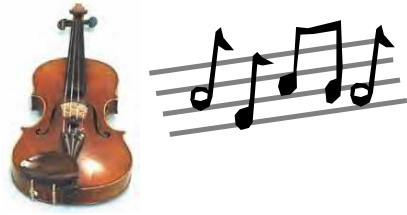


12



MUSIC SIGNAL PROCESSING

- 12.1 Introduction
- 12.2 Musical Instruments
- 12.3 A Review of Basic Physics of Sound
- 12.4 Music Signal Features and Models
- 12.5 Ear: Hearing of Sounds
- 12.6 Psychoacoustics of Hearing
- 12.7 Music Compression
- 12.8 High Quality Music Coding: MPEG
- 12.9 Stereo Music
- 12.10 Music Recognition

Music instruments and systems are some of the earliest human inventions that intuitively made use of the relationships between harmonics before the mathematics of such relations were understood and formalised. Similarly the notes and the lay out of the keys of musical instruments, such as a piano, were ‘matched’ to that of the layout and the frequency resolution of human auditory system well before the development of a formal scientific knowledge of the anatomy and frequency analysis functions of the cochlear of human ear.

Musical signal processing has a wide range of applications including; digital compression and coding of music for efficient storage and transmission on mobile phones and portable music players, modelling and reproduction of the acoustics of music instruments and music halls, digital music synthesisers, digital audio editors, digital audio mixers, spatial-temporal sound effects for home entertainment and cinemas, music content classification and indexing and music search engines for Internet.

This chapter begins with an introduction to the applications of music signal processing and the methods of classification of different types of musical instruments. The way that musical instruments such as guitar and violin produce vibrations (or sound) is explained. This is followed by a review of the basic physics of vibrations of string and pipe musical instruments, the propagation of sound waves and the frequencies of musical notes.

The human auditory system comprising of the outer, the middle and inner parts of ear are studied. The factors that affect the perception of audio signals and pitch and the psychoacoustics of hearing, and how these

Music Signal Processing

psychoacoustic effects are utilised in audio signal processing methods, are considered.

Digital music processing methods are mostly adaptations and extensions of the signal processing methods developed for speech processing. Harmonic models of music signals, source-filter models of music instruments, and probability models of the distribution of music signals and their applications to music coding methods such as MP3 and music classification are studied.

12.1 Introduction

Music signal processing methods – that is the methods used for coding/decoding, synthesis, composition and content-indexing of music signals – facilitate some of the essential functional requirements in a modern multimedia communication system. Widespread digital processing and dissemination of music began with the introduction of CD format digital music in the 1980s, increased with the popularity of MP3 Internet music and the demand for music on multimedia mobile/portable devices such as iPod and continues with the ongoing research in automatic transcription and indexing of music signals and the modeling of music instruments.

Some of the applications of music signal processing methods include the followings:

- Music coding for efficient storage and transmission of music signals. Examples are MP3, and Sony's adaptive transform acoustic coder.
- Noise reduction and distortion equalization such as Dolby systems, restoration of old audio records degraded with hiss, crackles etc., and signal processing systems that model and compensate for non-ideal characteristics of loudspeakers and music halls.
- Music synthesis, pitch modification, audio mixing, audio morphing, audio editing and computer music composition.
- Music transcription and content classification, music search engines for Internet.
- Music sound effects as in 3-D spatial surround music and special effect sounds in cinemas and theatres.

Music processing can be divided into two main branches: music signal modelling and music content creation. The music signal modelling approach is based on a well-developed body of signal processing techniques using

Music Signal Processing

signal analysis and synthesis tools such as filter banks, Fourier transform, cosine transform, harmonic plus noise model, wavelet transform, linear prediction models, probability models, hidden Markov models, hierarchical models, decision-tree clustering, Bayesian inference and perceptual models of hearing. Its typical applications are music coding, music synthesis, modelling of music halls and music instruments, music classification and indexing and creation of spatial sound effects.

Music creation has a different set of objectives concerned with the methods of composition of music content. It is driven by the demand for electronic music instruments, computer music software, digital sound editors and mixers and sound effect creation. In this chapter we are mainly concerned with music signal modelling and transformations.

Instrument type	Examples	Excitation type	Pitch changing method
String	Violin, viola, violoncello, bass viol, Cello, Guitar, piano, banjo, harp, sitar, balalaika, koto, mandolin, kanoon, zither, lyre hammered, dulcimer, berimbau.	String vibrations by plucking, hitting, or bowing strings.	Using different lengths, thickness, density or tension of strings.
Woodwind (not always made of wood)	Saxophone, clarinet, oboe, flute, piccolo, English horn, bagpipes krummhorn, shawm, flute, recorder, tin whistle, slide whistle.	Blowing air across: an edge, as in the flute; between a reed and a fixed surface, as in the clarinet and saxophone; between two reeds, as in oboe.	Opening and closing holes along the instrument's length with fingers.
Brass	Trumpet, trombone, french horn, tuba, bugle, digeridu, conch shell.	The sound comes from a vibrating column of air inside the tube. The air vibrates in resonance with the vibrating lips of the player, who presses her or his lips to the mouthpiece and forces air out.	Varying the speed of vibration of lips, varying the effective length of the tube, as on the trombone, or play through different lengths of tubing, as on brass instruments with valves.
Percussion	Drums, tambourine xylophone, marimba, vibraphone, hand-bells, chimes, gamelan Cymbals, gong, spoons, log drum, woodblock, triangle, maracas, rhythm sticks.	Sound source is a vibrating membrane or vibrating piece of solid material. The instrument is made to vibrate by hitting, shaking, rubbing.	Most percussion instruments do not have a definite pitch. The pitch of others like bell or drum depends on the material, its thickness and tension.
Keyboards	Harpsichord, Clavichord Piano, Pipe organ, Celesta, accordion	Strings (piano) or pipes (organ)	Varying string length and tension, varying the pipe length diameter and density.

Table 12.1 A popular classification of music instruments

Music Signal Processing

Aerophones	Idiophones
Instruments whose tone is generated by means of air set in vibration. The vibrating air is usually contained within the body of the instrument, like a pipe, as is the case for flutes and trumpets.	Instruments whose sound is produced by the material of the instrument itself, which is stiff and elastic enough to vibrate. Cymbals and bells are good examples of such instruments.
Chordophones	Membranophones
Instruments with strings as tone-producing elements that are stretched between fixed points. The strings vibrate when they are plucked, struck or scraped, like the violin or harp.	Instruments from which sound is produced mainly by the vibration of a stretched membrane, such as the drum

Table 12.2 Sachs-Hornbostel Classification of musical instruments.

12.2 Musical Instruments

There are a number of different systems for the classification of musical instruments. In a popular classification system, musical instruments are divided, according to how they set the air into vibrations, into four types of instruments: (1) strings, (2) woodwind, (3) brass and (4) percussion instruments.

Table 12.1 gives examples of each type of instrument and the excitation form for each instrument. In an alternative classification system, used in some academic literature, musical instruments are classified according to four major categories, based on the vibrating material that produces the sound. This system of classification named after its inventors as the Sachs-Hornbostel system is shown in table12.2. There is an additional class of electronic instruments called electrophones such as Theremin, Hammond organ, and electronic and software synthesizers.

12.2.1 Acoustic and Electric Guitars

A guitar has four main parts: (1) a number of strings, usually six strings, with different thickness and density, (2) the head of the guitar containing the tuning pegs for changing the tension of the strings, (3) the neck of the guitar with frets embedded on its face for changing the effective length of the strings by pressing them against the frets and (4) a means of amplifying and shaping the spectrum of the sound of the guitar. The main difference between an acoustic guitar and an electric guitar is the way that the

vibrations of strings are ‘picked up’, amplified and spectrally shaped. In an acoustic guitar the vibrations are picked up and transmitted via the saddle-bridge mechanism to the guitar’s wooden sound box that amplifies and spectrally shapes the vibrations.

The Body of Acoustic Guitar

The wooden body of an acoustic guitar amplifies the vibrations of strings and changes the timber and quality of the sound by shaping the amplitude of the harmonics of vibrating strings. Guitar strings are thin and have a small surface area and hence on their own, when they vibrate, they move only a small amount of air and produce little sound. To amplify the sound of strings, the vibrations of strings are transferred via the saddle and bridge on which the strings rest, to the larger surface area of the sound board which is the upper part of the body of the wooden box of guitar with a circular hole. The circular hole on the guitar acts as a Helmholtz resonator and affects the overall sound of the guitar. Note that the classical experiment on Helmholtz resonance is to blow air over a bottle which makes the air in the bottle resonate.

The body of an acoustic guitar has a waist, or a narrowing. The two widening are called bouts. The upper bout is where the neck connects, and the lower bout which is usually larger is where the bridge attaches. The size and shape of the body and the bouts affect the tone and timber that a given guitar produces. The top plate of a guitar is made so that it can vibrate up and down relatively easily. It is usually made of spruce or some other light, springy wood, about 2.5 mm thick. On the inside of the plate there is a series of braces that strengthen the plate and keep the plate flat. The braces also affect the way in which the top plate vibrates. The back plate is less important acoustically for most frequencies, partly because it is held against the player's body. The sides of a guitar do not radiate much sound.

The Guitar Strings

There are six strings on most guitars (bass guitars have 4 strings and some guitars have more than six strings) and they are tuned from the lowest string

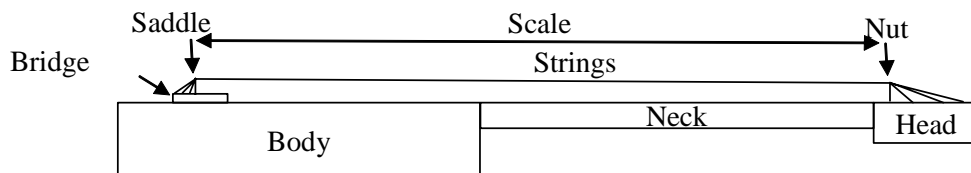


Figure 12.1 – Illustration of the main parts of an acoustic guitar.

Music Signal Processing

- the string closer to the top of the guitar as it rests on player's lap - to the highest string as: E, A, D, G, B, E. The pitch of a vibrating string depends on four factors:

- The *mass of the string*: heavier strings vibrate slower. On steel string guitars, the strings get thicker from high to low pitch. On acoustic guitars, the size change is complicated by a change in density: the low density nylon strings get thicker from the E to B to G; then the higher density wire-wound nylon strings get thicker from D to A to E.
- The frequency of vibration can be changed by changing the tension in the string using the tuning pegs: tighter gives a higher pitch.
- The frequency also depends on the length of the string that is free to vibrate. A player can change the length by holding a string firmly against the fingerboard with a finger. Shortening the string, by stopping it on a higher fret, gives higher pitch.

Bridge, Saddle and Nut

Attached to the soundboard of a guitar is a piece called the bridge which acts as the anchor for one end of the six strings. The bridge has a thin, hard piece embedded in it called the saddle, which is the part that the strings rest on. The other end of each string rests on the nut which is between the neck and the head of a guitar, the nut is grooved to hold the strings. The saddle and the nut hold the two effective vibrating ends of the string. The distance between these two points is called the scale length of the guitar. The vibrations of the strings are transmitted via the saddle and the bridge to the upper part of the body of guitar which acts as a sound board for amplification of the sound.

The Head, Neck and Frets of Guitar

The head of a guitar is the part that contains the tuning pegs. The neck of a guitar is the part that connects the head to the main body of the guitar. The face of the neck, containing the frets, is also called the fingerboard. The frets are metal pieces cut into the fingerboard at specific intervals. By pressing a string down onto a fret, the effective vibrating length of the string and therefore its fundamental frequency of vibration or tone it changes.

Music Signal Processing

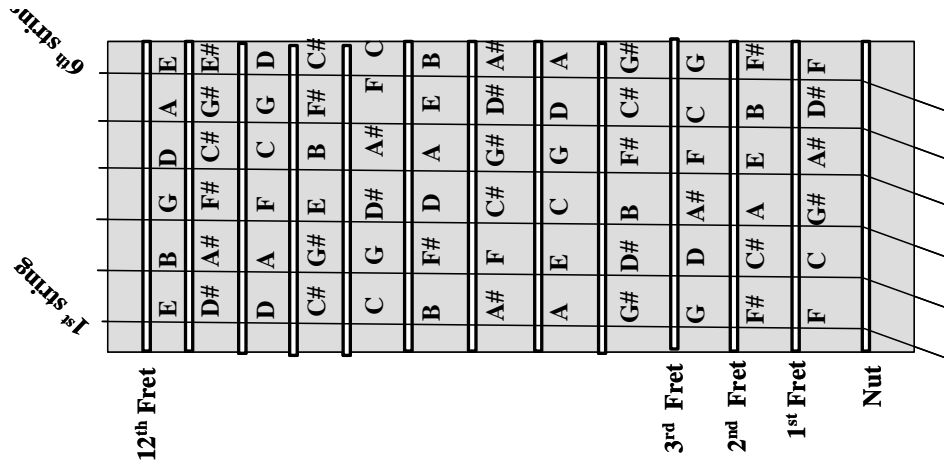


Figure 12.2 - An Illustration of the notes of strings when the effective length of a string is changed by pressing them on different frets.

Note	Fret	Frequency (1st string)	Fret position from saddle
E4	open	329.6	26.00
F4	1	349.2	24.54
F4#	2	370.0	23.16
G4	3	392.0	21.86
G4#	4	415.3	20.64
A4	5	440.0	19.48
A4#	6	466.1	18.38
B4	7	493.8	17.35
C4	8	523.2	16.38
C4#	9	554.3	15.46
D4	10	587.3	14.59
D4#	11	622.2	13.77
E5	12	659.2	13.00

Table 12.3 the frequencies of the notes of the 1st string with the string pressed on different frets. It is assumed that the scale length of 26 inches, note as the length of string halves the frequency of its pitch doubles.

Electric Guitars

Electric guitars do not need a hollow vibrating box to amplify the sound. Hence, the body of an electric guitar can be made of a solid of any shape. In an electric guitar the mechanical vibrations of the strings are picked up by a series of electro-magnetic coils placed underneath the strings. The coil-wrapped magnets convert the mechanical vibrations of strings into electric vibrating currents which are then band pass filtered and amplified by an electronic amplifier. The guitar amplifiers can do more than amplification; they can be operated in their non-linear distortion region to create a variety of rich sounds.

Electromagnetic pickups work on the principle of variable magnetic reluctance. The pickup consists of a permanent magnet wrapped with many turns of fine copper wire. The pickup is mounted on the body of the instrument, close to the strings. When the instrument's metal strings vibrate in the magnetic field of the permanent magnet, they alter the reluctance of the magnetic path. This changes the flux in the magnetic circuit which in turn induces a voltage in the winding. The signal created is then carried for amplification. Electric Guitars usually have several rows of pickups, including the humbucking pickups, placed at different intervals. A humbucking pickup comprises two standard pickups wired together in series. However, the magnets of the two pickups are reversed in polarity, and the windings are also reversed. Hence, any hum or other common mode electro-magnetic noise that is picked up is canceled out, while the musical signal is reinforced.

12.2.1 The Violin

The violin evolved from earlier string instruments such as the Rebec; a Middle Eastern bow-string instrument, the Lira da braccio and the fiddle. In its modern form the violin, shown in Figure 12.2, emerged in Italy around 1550. The most renowned violins were made by Cremonese violin-makers, like Amati, Stradivari and Guarneri, dating from about 1600 to 1750.

Violin sounds are produced by drawing a bow across one or more of four stretched strings. The string tensions are adjusted by tuning pegs at one end of the string, so that their fundamental frequencies are about 200, 300, 440 and 660 Hz corresponding to the notes G, D, A and E respectively. The strings vibrations produce little sound on their own. To amplify the sound and to shape its spectrum, energy from the vibrating string is transferred to the wooden sound box. The main plates of the violin's wooden box vibrate,

Music Signal Processing

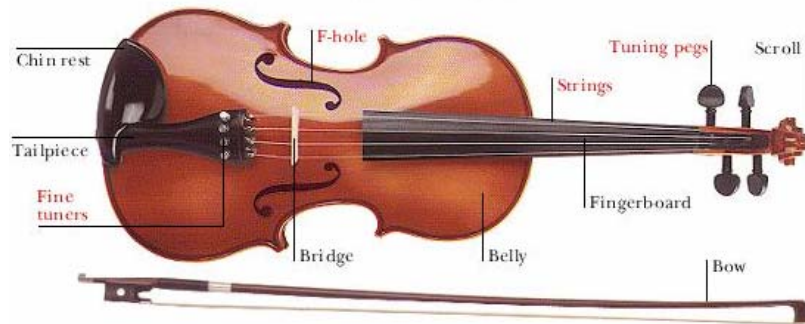


Figure 12.2 - A violin

amplify and shape the frequency spectrum of the sound. The strings are supported by the “bridge”, shown in Figure 12.3, which defines the effective vibrating length of the string, and acts as a mechanical transformer. The bridge converts the transverse forces of the strings into the vibrations of the sound box. The bridge has its own resonant modes and affects the overall tone and the sound of the instrument.

The front plate of the violin is carved from a fine-grained pinewood. Maple is usually used for the back plate and pine for the sides. Two *f*-shaped holes cut into the front plate affect its vibrations at high frequencies, and boost the sound output at low frequencies. The resonant frequency is affected by the area of the *f*-holes and the volume of the instrument.

The output of violin is increased by wedging a solid rod, the sound post, between the back and front plates, close to the feet of the bridge. The force exerted by the bowed strings causes the bridge to rock about this position, causing the plates to vibrate. This increases the volume of the output sound. The violin has a *bass bar* glued underneath the top plate to dampen its response at higher frequencies and prevent the dissipation of vibrations energy into acoustically inefficient higher frequencies.

As Hermann von Helmholtz observed when a violin string is bowed, it vibrates in a different way from the linear sinusoidal waves setup when the strings of a guitar are plucked. The repeated plucking of a guitar’s strings sets in motion a set of sinusoidal waves and harmonics on the strings that can be modelled by the linear system theory. This linearity and superposition principles imply

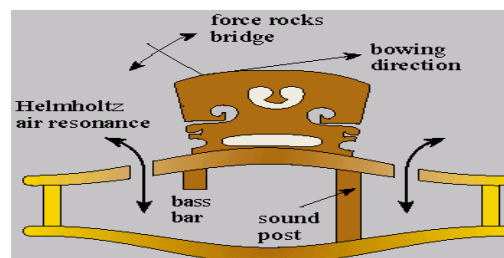


Figure 12.3 Cross section of violin at the bridge.

Music Signal Processing

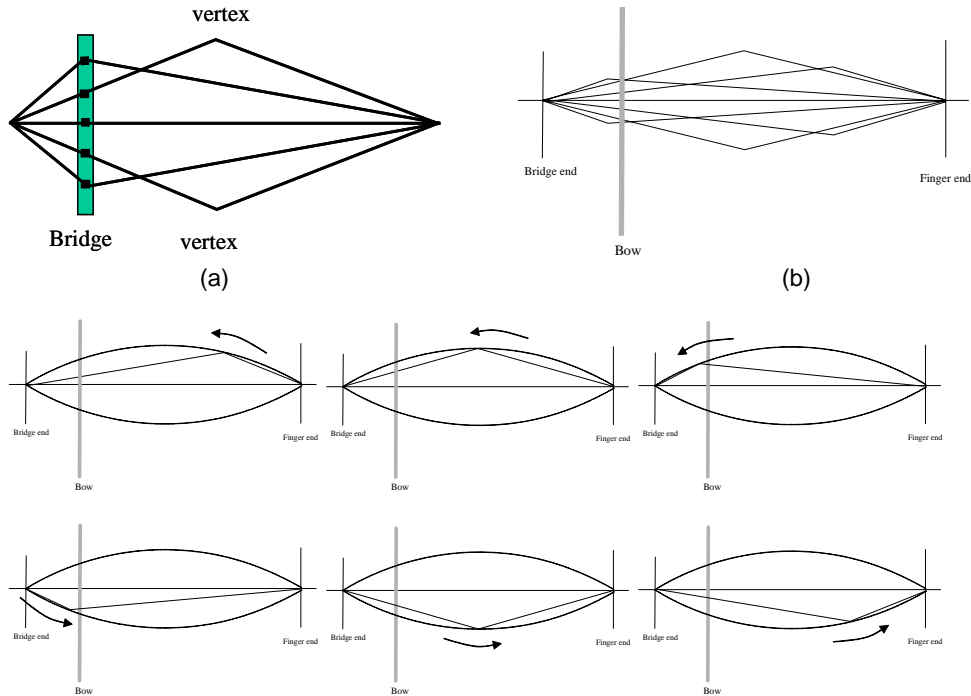


Figure 12.4 Sawtooth movement of violin strings: (a) shows movement of several different strings, (b) shows how one string may move back and forth, (c) show the 'snap shot' of movements of one string in (b).

that the sounds produced by plucking two strings of a guitar is the sum of their individual sounds and when a string is hit harder a greater sound with the expected pitch of the string is produced. The behaviour of a violin and bow system is nonlinear in that, for example, a greater amount of force applied via the bow does not simply produce a bigger or longer sound but it may produce altogether a different (perhaps scratchy) sound. It is this linear vs. nonlinear behaviour that underlies the fact that for a beginner it is usually easy to play the strings of a guitar in a musical-sounding way even when the wrong sequence of notes are played, whereas in contrast for a beginner it is difficult to play the strings of a violin in a musical and pleasant way.

Although the strings of a violin vibrate back and forth parallel to the bowing direction, other transverse modes of vibrations of the string are also excited, made up of straight-line sections in the form of a V-shaped waveform known as Helmholtz waves. The correct bowing action excites a Helmholtz mode with a single vertex separating two straight sections as shown in Figure 12.4. When the vertex of a Helmholtz wave is between the

bow and the fingered end of the string, the string moves at the same speed and direction as the bow. Only a small force is needed to lock the two motions together. This is known as the *sticking phase*. But as the vertex of V-shaped wave moves past the bow on its way to the bridge and back, the string slips past the bow and starts to move in the opposite direction to it. This is known as the *slipping phase*.

Although the sliding friction of bow and string is relatively small in the slipping phase, energy is continuously transferred from the strings to the vibration modes of the instrument at the bridge. Each time the vertex reflects back from the bridge and passes underneath the bow, the bow has to replace the lost energy by exerting a short pulse on the string so that it moves again at the same velocity as the bow. This process is known as the “slip-stick” mechanism of string excitation.

It turns out that for the stick-slip mechanism of the Helmholtz waves to work in a proper fashion and to produce sustained and pleasant sounds the bow force exerted on the strings must be within a certain maximum and minimum bounds that depend on the distance of the bow from the bridge, as shown in Figures 12.5. Assuming that the bow is at a distance of βL from the bridge end, where L is the length of the string between the finger and the bridge, then the minimum force is proportional to β^2 whereas the maximum force is proportional to β^1 .

The saw-tooth signal generated on the top of the violin bridge by a bowed string has a rich harmonic content. The amplitude of each frequency component of the saw-tooth signal is modified by the frequency response of

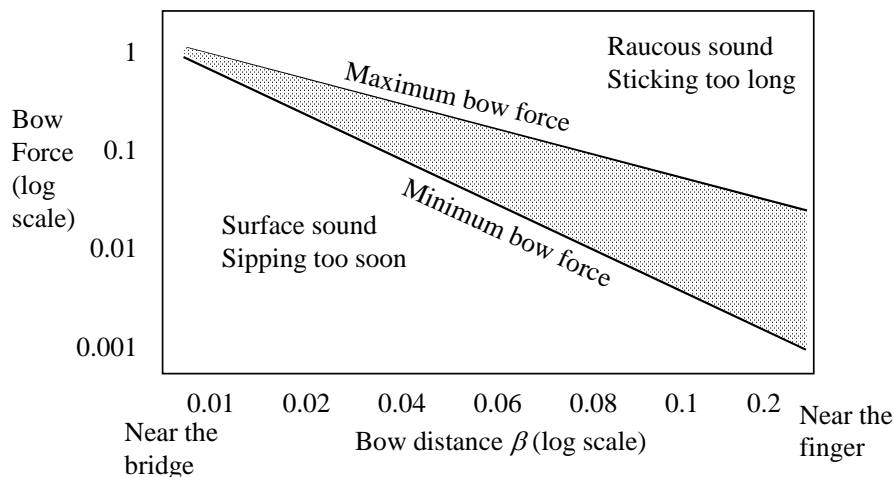


Figure 12.5 - The Schelleng diagram of bow force versus position for a long steady bow stroke of a violin.

Music Signal Processing

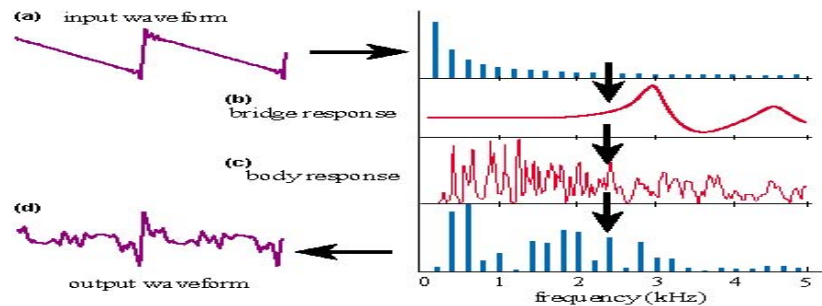


Figure 12.6 – Illustration of the general shape of the input and output waveforms of a violin and their respective spectra (a,d), together with the frequency responses of the bridge and the violin body (b,c).

the instrument, which is determined by the mechanical resonance of the bridge and by the vibrations of the body of the violin Figure 12.6. At low frequencies the bridge acts as a mechanical lever. However, between 2.5 and 3 kHz the bowing action excites a strong resonance of the bridge, with the top of the bridge rocking about its narrowed waist. This boosts the signal intensity in this frequency range, where the ear is most sensitive. Another resonance occurs at about 4.5 kHz in which the bridge bounces up and down on its feet. Between these two resonances there is a dip in the frequency response.

12.2.2 Wind Instruments

Wind instruments include different forms, shapes and arrangements of brass and wooden cylindrical tubes and also the human voice production system described in Chapter xx. To study the working of wind instruments we consider one of the simplest examples of a wind instrument a pennywhistle; a cylindrical instrument open at both ends with holes cut along it, and with a flat narrow tube at one end as the mouthpiece. The mouthpiece directs an air stream at a slanted hole with a sharp edge that splits the air stream causing air currents to excite the tube.



Assume the whistle has all its finger holes covered. Consider the propagation of a sudden change in air pressure at one end of the tube, such as the lowering of the pressure by taking some air out of one end. The adjacent air molecules will move to fill in the vacuum, leaving behind a new vacuum, which will in turn be filled by the neighboring air molecules and so on. In this way a pulse of low-pressure air will propagate along the tube. When the pulse arrives at the open end of the tube it will attract air from the room and will be reflected back with a changed polarity as a high-pressure pulse. A

Music Signal Processing

cycle of low and high-pressure air along the tube forms the fundamental period of the sound from the tube with a wavelength of $\lambda=2L$ where L is the length of the tube. Assuming the speed of propagation of sound is c , the fundamental frequency f_1 of an open-end tube is

$$f_1 = \frac{c}{2L} \quad (12.1)$$

The quality of resonance of the pipe's sound depends on the reflection and loss properties of the tube. In an open-ended tube there is no effective containment at the ends of the tube other than the room pressure that form pressure nodes. For the fundamental note, there are two pressure nodes at the ends and a pressure anti-node in the middle of the tube. The boundary condition of two pressure nodes at the ends of the tube, is also satisfied with all integer multiples of the fundamental frequency, hence integer multiples of the fundamental note exist with different intensities. In addition, the finger holes of a pennywhistle can be used to change its effective length and hence the wavelength and the fundamental frequency of the sound.

Closed-end Tubes A closed-end tube behaves like an open-end tube with the exception that at the closed end the pressure for the reflected wave must remain the same as that of the incoming wave, hence the reflected wave has the same polarity as the incoming wave. This implies that the wavelength of the fundamental note is four times (two round trips in the tube) the length of the tube. Hence the fundamental frequency of a closed-end tube is one half of a similar open-end tube or equivalently an octave lower. Due to the same-polarity of reflection at the closed end, a closed-end tube will generate a harmonic (or overtone) series based on odd integer multiples of the fundamental frequency. A closed-end cylindrical tube will produce an unusual set of musical notes.

Effect of Closed-end Tube Shape on the Harmonic Series

Consider a harmonic series built on an open-end pipe with its fundamental note a low C with a frequency of 130.81 Hz. The first 11 harmonics for this pipe is show in table 12.4.

As a trumpet is a closed-end pipe - the player's lips on the mouthpiece closes one end of the pipe. Since closed-end cylindrical pipes only produce the odd harmonics, this should exclude octaves, which follow powers of two multiples of the fundamental frequency. However, due to the design of the shape of trumpet - only a small section of a trumpet is actually cylindrical - trumpets produce overtone series that include the octaves. Most trumpets have gently tapered lead pipes. The most non-cylindrical part of the horn is

Music Signal Processing

C3 130.81 Hz	B-flat-5 at 915.67 Hz
C4 261.63 Hz	C6 1046.5 Hz
G4 392.00 Hz	D6 1177.29 Hz
C5 523.25 Hz	E6 1308.1 Hz
E5 659.26 Hz	F6 1438.91 Hz
G5 783.99 Hz	G6 1569.72 Hz

Table 12.4 – The harmonics of an open pipe

the bell and the mouthpiece. To analyze the effect of adding taper to a cylindrical closed-end pipe, consider the following closed end harmonic series in table 12.5.

1	C2	11	F sharp 5
3	G3 (slightly sharp)	13	A5
5	E4 (slightly sharp)	15	C6 (flat, though)
7	B flat	17	D6 (quite flat)
9	D5 (slightly sharp)	19	E6 (again, quite flat)

Table 12.5 – The harmonics of a closed-end pipe

The addition of a mouthpiece lowers the top six harmonics number 9, 11,13,15,17,19 yielding the following new adjusted series in Table 12.6.

1 -- C2	F5
3 -- G3 (slightly sharp)	G5 (slightly sharp, though)
5 -- E4 (slightly sharp)	B-flat 6
7-- B-flat	C6
D5	D6

Table 12.6 – The harmonics of a ‘trumpet’ pipe with mouthpiece added

The addition of a bell section by flaring the end of the tube moves up the lower modes. The modes raised are 1, 3 and 5. The new series, which is the standard overtone series for a low B-flat trumpet is shown in table 12.7.

E2 82.41 Hz	F5 698.46
B-flat 3	G5 783.99
F 4 349.23	B-flat 6
B-flat 4 466.16	C6 1046.5
D5 587.33	D6 1177.29

Table 12.7 – The harmonics of a closed-end pipe with the addition of mouth piece and bell section.

12.2.3 Examples of Spectrograms of Musical Instruments

Figure 12.7 shows some typical spectrograms of examples of string, brass, pipe and percussion instruments, for single note sounds. The figure reveals the harmonic structure and or shaped-noise spectrum of different instruments.

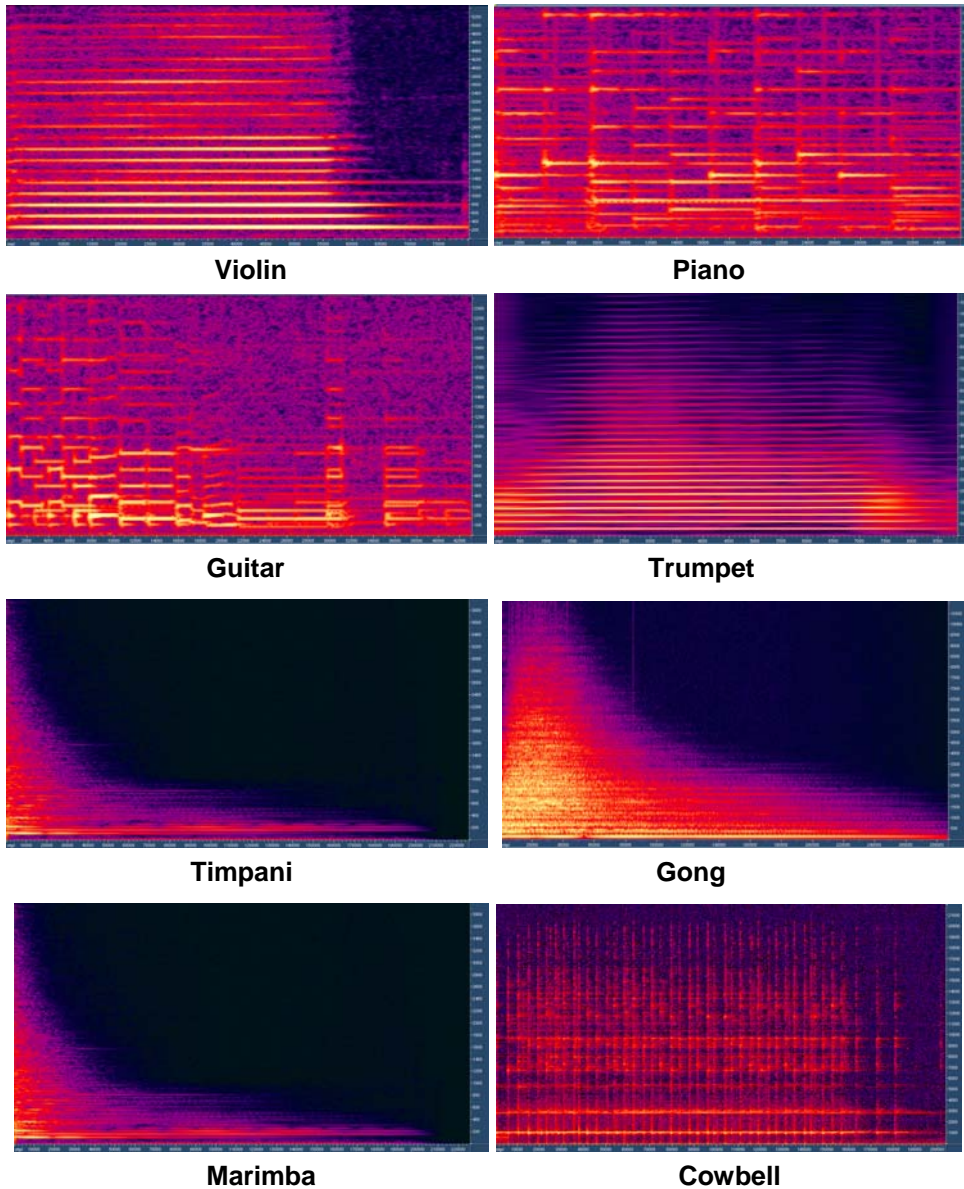


Figure 12.7 – Examples of spectra of some musical instruments.

12.3 A Review of Basic Physics of Sounds

Sound is the audible effect of air pressure variations caused by the vibrations, movement, friction or collision of objects. In this section we review the basic physics, properties and propagation of sound waves.

12.3.1 Sound Pressure, Power Intensity Levels, and Speed

Sound Pressure Level. The minimum audible air pressure variations (i.e. the threshold of hearing) p_0 is only 10^{-9} the atmospheric pressure or 2×10^{-5} N/m² (Newton/meter²). Sound pressure is measured relative to p_0 in decibels as

$$SPL(p) = 20 \log_{10}(p / p_0) \quad \text{dB} \quad (12.2)$$

From Equation (12.2) the threshold of hearing is 0 dB. The maximum sound pressure level (the threshold of pain) is $10^6 p_0$ (10^{-3} the atmospheric pressure) or 120 dB. Hence the range of hearing is about 120 dB, although the range of comfortable and safe hearing is less than 120 dB.

Sound power level. For a tone with a power of w watts this is defined in decibels relative to a reference power of $w_0 = 10^{-12}$ watts (or 1 pico watts) as

$$PL = 10 \log_{10}(w / w_0) = 10 \log_{10} w - 120 \quad \text{dB} \quad (12.3)$$

Sound intensity level. This is defined as the rate of energy flow across a unit area as

$$IL = 10 \log_{10}(I / I_0) = 10 \log_{10} I - 120 \quad \text{dB} \quad (12.4)$$

where $I_0 = 10^{-12}$ watts/m².

Speed of Sound Propagation Sound travels with a speed of

$$c = 331.3 + 0.6t \quad \text{m/s} \quad (12.5)$$

where t is the temperature of the air in degrees Celsius. Hence, at 20 °C the speed of sound is about 343.3 meters per second or about 34.3 cm per ms. Sound propagates faster in liquids than in air and faster in solids than in liquids. The speed of propagation of sound in water is 1500 m/s in metals can be 5000 m/s.

12.3.2 Frequency, Pitch, Harmonics, Overtones and Intervals

Sound waves are produced by vibrating objects and instruments. The frequency of a sound is the same as that of the source and is defined as the number of oscillations per second in the units of Hertz. Since a sound wave is a pressure wave, the frequency of the wave is also the number of oscillations per second from a high pressure (compression) to a low pressure (rarefaction) and back to a high pressure.

Human ear is a sensitive detector of the fluctuations of air pressure, and is capable of hearing sound waves in a range of about 20 Hz to 20 kHz. The sensations of prominent sound frequencies are referred to as the *pitch* of a sound. *A high pitch sound corresponds to a high fundamental frequency and a low pitch sound corresponds to a low fundamental frequency.* The harmonics of a fundamental frequency F_0 are its integer multiples kF_0 .

Certain sound waves which when played simultaneously produce a pleasant sensation are said to be in *consonant*. Such sound waves form the basis of the *intervals* in music. For example, any two sounds whose frequencies make a 2:1 ratio are said to be separated by an *octave* and two sounds with a frequency ratio of 5:4, are said to be separated by an *interval* of a third. Examples of other music sound wave intervals and their respective frequency ratios are listed in table 12.8.

Wavelength of Sounds

The wavelength λ of a sound wave depends on its speed of propagation c and frequency of vibration f through the equation $\lambda=c/f$. For example at a speed of 344 meters/ second, a sound wave at a frequency of 10 Hz has a

Interval	Frequency Ratio	Examples
Octave	2:1	512 Hz and 256 Hz
Third	5:4	320 Hz and 256 Hz
Fourth	4:3	342 Hz and 256 Hz
Fifth	3:2	384 Hz and 256 Hz

Table 12.8 - Musical sound intervals and their respective frequency ratios.

wavelength of 34.4 meters, at 1 kHz it has a wavelength of 34.4 cm and at 10 kHz it has a wavelength of 3.44 centimeters.

Bandwidths of Music and Voice

The bandwidth of unimpaired hearing is normally between 10 Hz to 20 kHz, although some individuals may have a hearing ability beyond this range of frequencies. Sounds below 10 Hz are called infra-sounds and above 20 kHz are called ultra-sounds. The information in speech (i.e. words, speaker identity, accent, intonation, emotional signals etc.) is mainly in the traditional telephony bandwidth of 300 Hz to 3.5 kHz.

The sound energy above 3.5 kHz mostly conveys quality and sensation essential for high quality applications such as broadcast radio/tv, music and film sound tracks. Singing voice has a wider dynamic range and a wider bandwidth than speech and can have significant energy in the frequencies well above that of normal speech. For music the bandwidth is from 10 Hz to 20 kHz. Standard CD music is sampled at 44.1 kHz or 48 kHz and quantized with the equivalent of 16 bits of uniform quantization which gives a signal to quantization noise ratio of about 100 dB at which the quantization noise is inaudible and the signal is transparent.

12.3.3 Frequencies of Musical Notes

There are two musical pitch standards, the American pitch standard which takes A in the fourth piano octave (A4) to have a frequency of 440 Hz (Table 12.9), and the International pitch standard, which takes A4 to have a frequency of 435 Hz. Both of these pitch standards define *equal tempered chromatic scales*. This means that each successive pitch is related to the previous pitch by a factor of the twelfth root of 2 ($\sqrt[12]{2} = 1.05946309436$) known as a half-tone. Hence there are twelve half-tones (black and white keys on a piano), or steps, in an octave which corresponds to a doubling of pitch.

The frequency of the intermediate notes, or pitches, can be found by multiplying (or dividing) a given starting pitch by as many factors of the twelfth root of 2 as there are steps up to (or down to) the desired pitch. For example, the G above A4 (that is, G5) in the American Standard has a frequency of $440 \left(\sqrt[12]{2}\right)^{10} = 783.99$ Hz. Likewise, in the International standard, G5 has a frequency of 775.08 Hz. G#5 (G5 sharp) is another factor of the 12th root of 2 above these, or 830.61 and 821.17 Hz, respectively. Note when counting the steps that there is a single half-tone (step) between B and C, and E and F.

Music Signal Processing

These pitch scales are referred to as 'equal tempered' or 'well tempered.' This refers to a compromise built into the use of the 12th root of 2 as the factor separating each successive pitch. For example, G and C are a fifth apart. The frequencies of notes that are a perfect fifth apart are exactly in the ratio of 1.5. G is seven chromatic steps above C, so, using the 12th root of 2, the ratio between G and C on either standard scale is $\left(\sqrt[12]{2}\right)^7 = 1.49830707688$, which is slightly less than the 1.5 required for a perfect fifth. This slight reduction in frequency is referred to as *tempering*. Tempering is necessary on instruments such as the piano that can be played in any key because it is impossible to tune all 3rds, 5ths, etc. to their exact ratios (such as 1.5 for fifths) and simultaneously have, for example, all octaves come out exactly in the ratio of 2.

Figure (12.8) shows the frequencies of the keys on a piano. Note that the keys are arranged in groups of 12. Each set of 12 keys spans an octave which is the doubling of frequency. For example the frequency of A_N is $2^N A_0$ or N octaves higher than A_0 .

Music Signal Processing

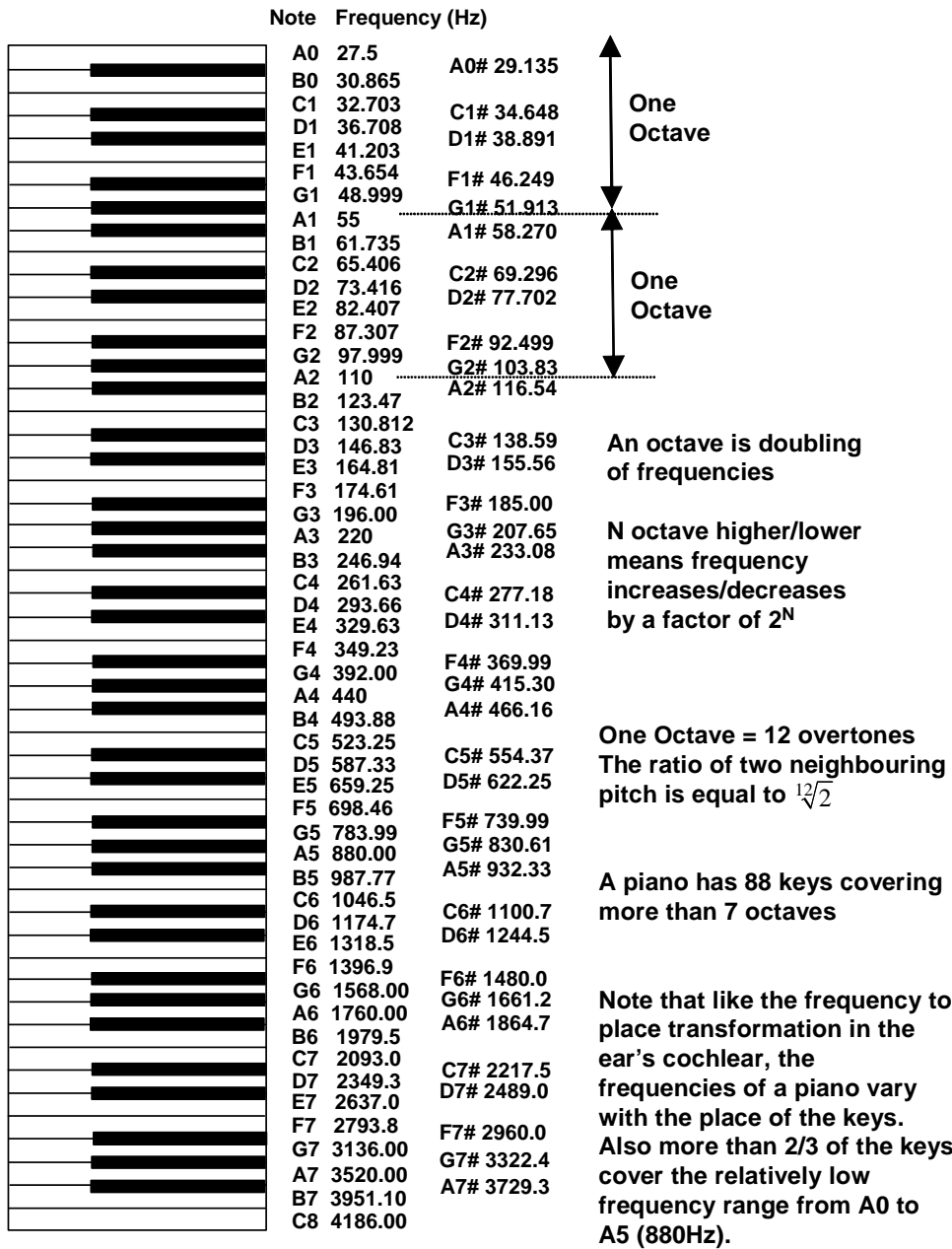


Figure 12.8 – The frequencies of the keys on a piano. Note they keys are arranged in groups of 12. Each set o 12 keys spans an octave which is the doubling of frequency. For example the frequency of A_N is $2^N A_0$ or N octave higher than A_0 .

Music Signal Processing

Table 12.9 Music frequencies for equal-tempered scale $A_4 = 440$					
Note	Frequency (Hz)	Note	Frequency (Hz)	Note	Frequency (Hz)
C ₀	16.35	B ₂	123.47	F [#] ₅ /G ^b ₅	739.99
C [#] ₀ /D ^b ₀	17.32	C ₃	130.81	G ₅	783.99
D ₀	18.35	C [#] ₃ /D ^b ₃	138.59	G [#] ₅ /A ^b ₅	830.61
D [#] ₀ /E ^b ₀	19.45	D ₃	146.83	A ₅	880.00
E ₀	20.60	D [#] ₃ /E ^b ₃	155.56	A [#] ₅ /B ^b ₅	932.33
F ₀	21.83	E ₃	164.81	B ₅	987.77
F [#] ₀ /G ^b ₀	23.12	F ₃	174.61	C ₆	1046.50
G ₀	24.50	F [#] ₃ /G ^b ₃	185.00	C [#] ₆ /D ^b ₆	1108.73
G [#] ₀ /A ^b ₀	25.96	G ₃	196.00	D ₆	1174.66
A ₀	27.50	G [#] ₃ /A ^b ₃	207.65	D [#] ₆ /E ^b ₆	1244.51
A [#] ₀ /B ^b ₀	29.14	A ₃	220.00	E ₆	1318.51
B ₀	30.87	A [#] ₃ /B ^b ₃	233.08	F ₆	1396.91
C ₁	32.70	B ₃	246.94	F [#] ₆ /G ^b ₆	1479.98
C [#] ₁ /D ^b ₁	34.65	C ₄	261.63	G ₆	1567.98
D ₁	36.71	C [#] ₄ /D ^b ₄	277.18	G [#] ₆ /A ^b ₆	1661.22
D [#] ₁ /E ^b ₁	38.89	A [#] ₃ /B ^b ₃	233.08	A ₆	1760.00
E ₁	41.20	B ₃	246.94	A [#] ₆ /B ^b ₆	1864.66
F ₁	43.65	C ₄	261.63	B ₆	1975.53
F [#] ₁ /G ^b ₁	46.25	C [#] ₄ /D ^b ₄	277.18	C ₇	2093.00
G ₁	49.00	D ₄	293.66	C [#] ₇ /D ^b ₇	2217.46
G [#] ₁ /A ^b ₁	51.91	D [#] ₄ /E ^b ₄	311.13	D ₇	2349.32
A ₁	55.00	E ₄	329.63	D [#] ₇ /E ^b ₇	2489.02
A [#] ₁ /B ^b ₁	58.27	F ₄	349.23	E ₇	2637.02
B ₁	61.74	F [#] ₄ /G ^b ₄	369.99	F ₇	2793.83
C ₂	65.41	G ₄	392.00	F [#] ₇ /G ^b ₇	2959.96
C [#] ₂ /D ^b ₂	69.30	G [#] ₄ /A ^b ₄	415.30	G ₇	3135.96
D ₂	73.42	A ₄	440.00	G [#] ₇ /A ^b ₇	3322.44
D [#] ₂ /E ^b ₂	77.78	A [#] ₄ /B ^b ₄	466.16	A ₇	3520.00
E ₂	82.41	B ₄	493.88	A [#] ₇ /B ^b ₇	3729.31
F ₂	87.31	C ₅	523.25	B ₇	3951.07
F [#] ₂ /G ^b ₂	92.50	C [#] ₅ /D ^b ₅	554.37	C ₈	4186.01
G ₂	98.00	D ₅	587.33	C [#] ₈ /D ^b ₈	4434.92
G [#] ₂ /A ^b ₂	103.83	D [#] ₅ /E ^b ₅	622.25	D ₈	4698.64
A ₂	110.00	E ₅	659.26	D [#] ₈ /E ^b ₈	4978.03
A [#] ₂ /B ^b ₂	116.54	F ₅	698.46		

% This program generates and plays sine waves corresponding to the musical notes.

```
function MusicalNotes()
```

12.3.4 Sound Propagation: Reflection, Diffraction, Refraction and Doppler Effect

The propagation of sound waves affects the sensation and perception of music. Sound propagates from the source to the receiver through a combination of four main propagation modes namely; (1) direct propagation path, (2) reflection from walls, (3) diffraction around objects or through openings and (4) refraction due to temperature differences in the layers of air. In general, in propagating through different modes sound is delayed and attenuated by different amounts.

Reflection happens when a sound wave encounters a medium with different impedance from which it is travelling in, for example when the sound propagating in the air hits the walls of a room as shown in Figure 12.9. Sound reflects from walls, objects, etc. Acoustically, reflection results either in sound reverberation for small round-trip delays (less than 100 ms), or in echo for longer round-trip delays.

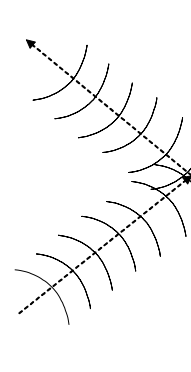


Figure 12.9 Reflection

Diffraction is the bending of waves around objects and the spreading out of waves beyond openings as shown in Figure 12.10. In order for this effect to be observed the size of the object or gap must be comparable to or smaller than the wavelength of the waves. When sound waves travel through doorways or between buildings they are diffracted, so that the sound is heard around corners. If we consider two separate ‘windows’ then each ‘window’ acts as a new source of sound, and the waves from these secondary sources can act constructively and destructively. When the size of the openings or obstacles is about the same as the wavelength of the sound wave, patterns of maxima and minima are observed. If a single opening is divided into many small sections, each section can be thought of as an emitter of the wave. The waves from each piece of the opening are sent out in phase with each other; at some places they interfere constructively, and at others they interfere destructively.

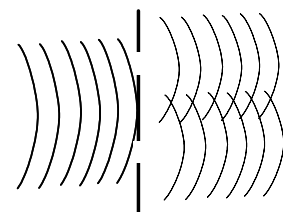


Figure 12.10 Diffraction

Refraction, shown in Figure 12.11, is the bending of a wave when it enters a medium where its speed of propagation is different. For sound waves refraction usually happens due to

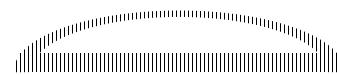


Figure 12.11 Refraction

temperature changes in different layers of air as the speed of sound increases with temperature so during day when the higher layers of air are cooler sound is bent upward (it takes longer for sound to travel in the upper layers) and during night when a temperature inversion happens the sound is bent downwards.

Doppler Effect is the perceived changes in the received frequency of a waveform resulting from relative movements of the source (emitter) towards or away from the receiver. As illustrated in Figure 12.12, the received pitch of the sound increases (sound wave fronts move towards each other) when there is a relative movement of the source towards the receiver and decreases when there is a relative movement of the source away from the receiver. When a sound source approaches a receiver with a relative speed of v , the perceived frequency of the sound is raised by a factor of $c/(c-v)$, where c is the speed of sound. Conversely, when a sound source moves away from the receiver with a relative speed of v , the perceived frequency of the sound is lowered by a factor of $c/(c+v)$. When $c=v$ (or $c > v$) sound barrier is broken and a sonic is due to reinforcement of densely packed wave fronts is heard. The sound barrier speed is 761 mph.

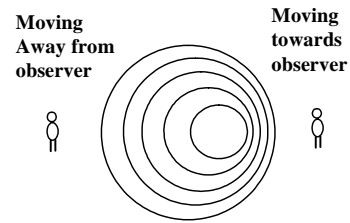


Figure 12.12 Doppler effect

Assuming that the sound source is moving with speed $\pm v_{sr}$ (where + is towards receiver and - away from it) and the receiver is moving with speed $\pm v_{rs}$ (where + is towards source and - away from it) then the relationship between the perceived frequency f_r and the source frequency f_s is given by

$$f_r = \frac{c + v_{rs}}{c - v_{sr}} f_s = \frac{1 + \frac{v_{rs}}{c}}{1 - \frac{v_{sr}}{c}} f_s \quad (12.6)$$

Where as explained v_{sr} and v_{rs} can be positive or negative depending on the relative direction of movement.

Doppler effects also happen with electromagnetic waves such as light waves. For example if a light source is moving towards the observer it seems bluer (shifted to a higher frequency) this is known as *blue shift* and if a light source is moving away from an observer it seems redder (shifted to a lower frequency) and this is known as *red shift*. The fact that the light from distant galaxies is red shifted is considered as major evidence that the universe is expanding.

12.3.5 Motion of Sound Waves on Strings

A wave travelling along a string will bounce back at the fixed end and interfere with the part of the wave still moving towards the fixed end. When the wavelength is matched to the length of the string, the result is standing waves. For a string of length L , the wavelength is $\lambda=2L$, the period, that is time taken to travel one wavelength, is $T=2L/c$, where c is the speed of the wave and the *fundamental frequency* of the string is

$$f_1 = \frac{c}{2L} \quad (12.7)$$

In general, for a string fixed at both ends the harmonic frequencies are the integer multiples of the fundamental given as:

$$f_n = \frac{nc}{2L} \quad n=1,2,3,4, \dots \quad (12.8)$$

For example, when a guitar string is plucked, waves at different frequencies will bounce back and forth along the string. However, the waves that are not at the harmonic frequencies will have reflections that do not interfere constructively. The waves at the harmonic frequencies will interfere constructively, and the musical tone generated by plucking the string will be a combination of the harmonics.

Example 12.1 The fundamental frequency of a string depends on its mass, length and tension. Assume a string has a length of $L=63$ cm, a mass of $m=30$ g, and a tension of $S=87$ N. Calculate the fundamental frequency of this string.

The speed of the wave on a string is given by

$$c = \left(\frac{\text{Tension}}{\text{Mass / Length}} \right)^{1/2} = \left(\frac{87}{(0.03/0.63)} \right)^{1/2} = 42.74 \text{ m/s} \quad (12.9)$$

From eq(12.7) the fundamental frequency is obtained as

$$f_1 = \frac{c}{2L} = \frac{42.74}{2 \times 0.63} = 33.9 \text{ Hz} \quad (12.10)$$

The harmonic frequencies are given by nf_1 .

12.3.6 Longitudinal Waves in Wind instruments and Pipe Organs

A main differences between the sound waves in pipes and on strings is that while strings are fixed at both ends, a tube is either open at both ends or open at one end and fixed at the other. In these cases the harmonic frequencies are given by:

$$\text{Tube open at both ends} \quad f_n = \frac{nc}{2L} \quad n=1,2,3,4, \dots \quad (12.11)$$

$$\text{Tube open at one end} \quad f_n = \frac{nc}{4L} \quad n=1,2,3,4, \dots \quad (12.11)$$

Hence, the harmonic frequencies of a pipe are changed by varying its effective length. A pipe organ has an array of different pipes of varying lengths, some open-ended and some closed at one end. Each pipe corresponds to a different fundamental frequency. For an instrument like a flute, on the other hand, there is only a single pipe. Holes can be opened along the flute to reduce the effective length, thereby increasing the frequency. In a trumpet, valves are used to make the air travel through different sections of the trumpet, changing its effective length; with a trombone, the change in length is obvious.

Example 12.2 Calculation of Formants: Resonance of vocal tract

The vocal tract tube can be modelled as a closed-end tube with an average length of 17 cm for a male speaker. Assume that the velocity of sound is $c=343.3$ m/s at a temperature of 20 °C. The n^{th} harmonic of the resonance of vocal tract tube is given by

$$f_n = nc/4L = n343.3 / (4 \times 0.17) = 505 n \quad (12.13)$$

Hence the fundamental resonance frequency of vocal tract, the first resonance (aka formant) of speech, is 505 Hz. Since this formant varies with temperature it is usually rounded 500 Hz. The higher formants occur at odd multiples of the frequency of the first formant at 1500, 2500, 3500 and 4500. Note that this is a simplified model. In reality the shape of the vocal tract is affected by the position of articulators and the formants are a function of the phonetic content of speech and the speaker characteristics.

12.3.7 Wave Equations for Strings

In this section we consider the question of how a force such as plucking or hammering a string sets up a pattern of wave motions on a string. The wave

Music Signal Processing

equation for a vibrating string can be derived from Newton's second law of motion which states that: *force=mass×acceleration*.

Consider a short length Δx of an ideal string under tension as illustrated in Figure 12.11. The net vertical force can be expressed in terms of the string tension T and the angular displacements as

$$F_y = T \times (\sin(\phi_1) - \sin(\phi_2)) = T \times (\tan(\phi_1) - \tan(\phi_2)) \quad (12.14)$$

where it is assumed that for small values of ϕ , $\sin(\phi) = \tan(\phi) = \frac{\partial y}{\partial x}$. Note $\tan(\phi_1)$ and $\tan(\phi_2)$ are the displacement slopes at x and $x+\Delta x$ respectively given by

$$\tan(\phi_1) = \left. \frac{\partial y}{\partial x} \right|_x \quad \tan(\phi_2) = \left. \frac{\partial y}{\partial x} \right|_{x+\Delta x} \quad (12.15)$$

and

$$\tan(\phi_2) = \frac{\partial y}{\partial x} + \frac{\partial^2 y}{\partial x^2} \Delta x \quad (12.16)$$

From Equations (12.14-12.16) we obtain the tension force as

$$F_y = T \left(\frac{\partial^2 y}{\partial x^2} \right) \Delta x \quad (12.17)$$

Assuming that the string has a uniform mass density of ε per unit length, then the mass of a length Δx is $\varepsilon \Delta x$. Using the Newton's second law describing the relationship between force, mass and acceleration we have

$$T \left(\frac{\partial^2 y}{\partial x^2} \right) \Delta x = \varepsilon \Delta x \left(\frac{\partial^2 y}{\partial t^2} \right) \quad (12.18)$$

or

$$c^2 \frac{\partial^2 y}{\partial x^2} = \frac{\partial^2 y}{\partial t^2} \quad (12.19)$$

where $c = \sqrt{T/\varepsilon}$ has the dimension of velocity. From Equation (12.19) we can obtain the following types of solutions for waves travelling in time t in positive and negative x directions

$$y^+(x,t) = f(x - ct) \quad (12.20)$$

Music Signal Processing

and

$$y^-(x,t) = f(x+ct) \quad (12.21)$$

The sum of two travelling waves is also a solution and gives a standing wave as

$$y(x,t) = A f(x-ct) + B f(x+ct) \quad (12.22)$$

A discrete-time version of equation (12.21) can be obtained as

$$y(n,m) = A f(n-cm) + B f(n+cm) \quad (12.23)$$

where n and m represent the discrete-space and the discrete-time variables.

12.3.8 Wave Equation for Acoustic-Tubes

The wave equations for ideal acoustic tubes are similar to the wave equations for strings with the following differences:

- (a) The motion of a vibrating string can be described by a *single* 2-dimensional variable $y(x,t)$ whereas an acoustic tube has *two* 2-dimensional variables: the pressure gradient $p(x,t)$ and the volume velocity $u(x,t)$. Note that in reality the motion of string and pressure waves are functions of the 3-dimensional space.
- (b) String vibrations are perpendicular, or transverse, to the direction of the wave propagation and string waves are said to be *transversal*, whereas in a tube the motion of waves are in the same direction as the wave oscillations and the waves are said to be *longitudinal*.

In an acoustic tube the pressure gradient and the velocity gradient interact. Using the Newton's second law, describing the relationship between force, mass and acceleration, with an analysis similar to that for deriving the wave equation for strings, the equations expressing pressure gradient and velocity gradient functions can be described as

$$\text{Pressure gradient: } c^2 \left(\frac{\partial^2 p}{\partial x^2} \right) = \left(\frac{\partial^2 p}{\partial t^2} \right) \quad (12.24)$$

$$\text{Velocity gradient: } c^2 \left(\frac{\partial^2 u}{\partial x^2} \right) = \left(\frac{\partial^2 u}{\partial t^2} \right) \quad (12.25)$$

The wave velocity c can be expressed in terms of the mass density of air ρ and the compressibility of air κ as

$$c = \frac{1}{(\rho\kappa)^{0.5}} \quad (12.26)$$

The solution for pressure and velocity gradient are

$$p(x,t) = Z_0[u^+(x-ct) + u^-(x+ct)] \quad (12.27)$$

$$u(x,t) = u^+(x-ct) - u^-(x+ct) \quad (12.28)$$

where Z_0 is obtained as follows. Using Newton's second law of motion:

$$\frac{\partial p}{\partial x} = -\frac{\rho}{A} \left(\frac{\partial u}{\partial t} \right) \quad (12.29)$$

where A is the cross-section area of the tube. From Equations (12.27-29) we obtain

$$Z_0 = \frac{\rho c}{A} \quad (12.30)$$

12.4 Music Signal Features and Models

The signal features and models employed for music signal processing are broadly similar to those used for speech processing. The main characteristic differences between music and speech signals are as follows:

- (a) The essential features of music signals are pitch (i.e. fundamental frequency, timber (related to spectral envelope), slope of attack, slope of sustain, slope of decay and beat.
- (b) The slope of the attack at the start of a note or a segment of music, the sustain period, the fall rate and the timings of notes are important acoustic parameters in music. These parameters have a larger dynamic range than those of speech.

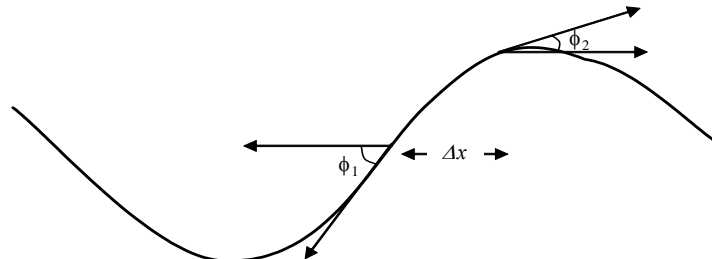


Figure 12.13 - Displacement movement of a vibrating string.

Music Signal Processing

- (c) Beat and rhythm, absent in normal speech, are important acoustic features of musical signals.
- (d) Music signals have a wider bandwidth than speech extending up to 20 kHz and often have more energy in higher frequencies than speech.
- (e) Music signals have a wider spectral dynamic range than speech. Music instruments can have sharper resonance and the excitation can have a sharp harmonic structure (as in string instruments).
- (f) Music signals are polyphonic as they often contain multiple notes from a number of sources and instruments played simultaneously. In contrast speech is usually a stream of monophonic events from a single source. Hence, music signals have more diversity and variance in their spectral-temporal composition.
- (g) Music signals are mostly stereo signals with a time-varying cross-correlation between the left and right channels.
- (h) The pitch and its temporal variations play a central role in conveying sensation in music signals, the pitch is also important in conveying prosody, phrase/word demarcation, emotion and expression in speech.

The signal analysis and modelling methods used for musical signals include:

- (a) Harmonic plus noise models.
- (b) Linear prediction models.
- (c) Probability models of the distribution of music signals.
- (d) Decision-tree clustering models.

In the followings we consider different methods of modelling music signals.

12.4.1 Harmonic Plus Noise Model (HNM)

The harmonic plus noise model describes a signal as the sum of a periodic component and a spectrally-shaped random noise component as

$$x(m) = \underbrace{\sum_{k=1}^N \left[A_k(m) \cos(2\pi k f_0(m) m) + B_k(m) \sin(2\pi k f_0(m) m) \right]}_{\text{Fourier Series Harmonics}} + \underbrace{e(m)}_{\text{Noise}} \quad (12.31)$$

where $f_0(m)$ is the time-varying fundamental frequency or pitch, $A_k(m)$ and $B_k(m)$ are the amplitudes of the k^{th} sinusoidal harmonic components and $e(m)$ is the non-harmonic noise-like component at time discrete-time m .

The sinusoids model the main vibrations of the system. The noise models the non-sinusoidal energy produced by the excitation and any non-sinusoidal system response such as breath noise in wind instruments, bow noise in strings, and transients in percussive instruments. For example, for wind instruments, the sinusoids model the oscillations produced inside the pipe and the noise models the turbulence that takes place when the air from the player's mouth passes through the narrow slit.

For bowed strings the sinusoids are the result of the main modes of vibrations of the strings and the sound box, and the noise is generated by the sliding of the bow against the string plus other non-linear behaviour of the bow-string-resonator system.

The amplitudes, and frequencies of sinusoids vary with time and their variations can be modelled by a low-order polynomial. For example $A_k(m)$ can be modelled as a constant, a line, or a quadratic curve as

$$A_k(m) = a_k(m_i) \quad (12.32)$$

$$A_k(m) = a_k(m_i) + b_k(m_i)(m - m_i) \quad (12.33)$$

$$A_k(m) = a_k(m_i) + b_k(m_i)(m - m_i) + c_k(m_i)(m - m_i)^2 \quad (12.34)$$

where m_i is the beginning of the i^{th} segment of music. Similar equations can be written for $B_k(m)$. The rate of variations of a music signal are state-dependent and different set of polynomial coefficients are required during attack, sustain and fall periods of a musical note. The noise component $e(m)$ is often spectrally-shaped and may be modelled by a linear prediction filter as

$$e(m) = \sum_{k=1}^P a_k e(m) + \varepsilon(m) \quad (12.35)$$

For the harmonic plus noise model we need to estimate the fundamental frequency f_0 , the amplitudes of the harmonics and the parameters of the noise-shaping filter.

12.4.2 Linear Prediction Models for Music

Music Signal Processing

Linear prediction analysis can be applied to the modelling of music signals in two ways:

- (1) to model the music signal within each signal frame,
- (2) the model the correlation of the signal across speech frames, e.g. to model the correlation of harmonics and noise across successive frames. A linear predictor model is described as

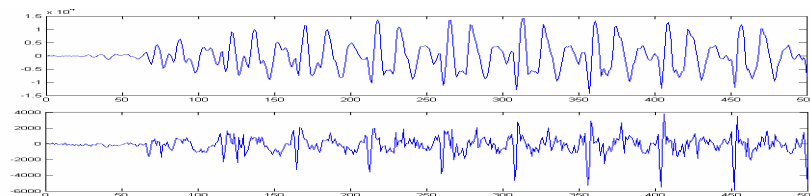
$$x(m) = \sum_{k=1}^P a_k x(m-k) + e(m) \quad (12.36)$$

where a_k are the predictor coefficients and $e(m)$ is the excitation. For music signal processing linear prediction model can be combined with harmonic noise model (HNM), so that linear predictor models the spectral envelop of the music whereas HNM model the harmonic plus noise structure of the excitation. Combination of linear predictor and HNM are described in chapter xx

12.4.3 Sub-band Linear Prediction for Music Signals

A main assumption of linear prediction theory is that the input signal has a flat spectrum that is shaped as it is filtered through the predictor. Due to the wider spectral dynamic range and the sharper resonance of music signals compared to speech, and due to the non-white harmonically-structured spectrum of the input excitation for string instruments, a linear prediction system has difficulty modelling the entire bandwidth of music signal and capturing its spectral envelope such that the residue or input is spectrally flat.

The problem can be partly mitigated by a sub-band linear prediction system introduced in section xx. A further reason for using a sub-band based method with music signals is the much larger bandwidth of musical signals. The signal can be divided into N sub-bands and then each sub-band signal can be down-sampled prior to LP modelling. Each sub-band signal having a smaller dynamic range will be better suited to LP modelling. Figure 12.14-15 show examples of linear prediction analysis of music.



Music Signal Processing

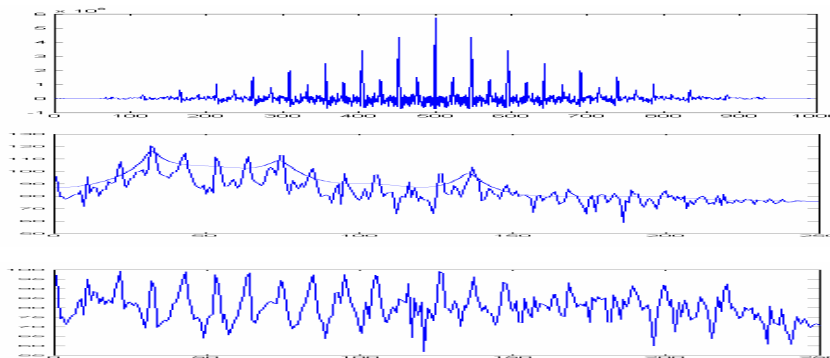


Figure 12.14 - (a) Speech, (b) output of inverse linear predictor, (c) correlation of speech, (d) DFT and LP spectra of (a), (e) spectrum of input signal in (b).

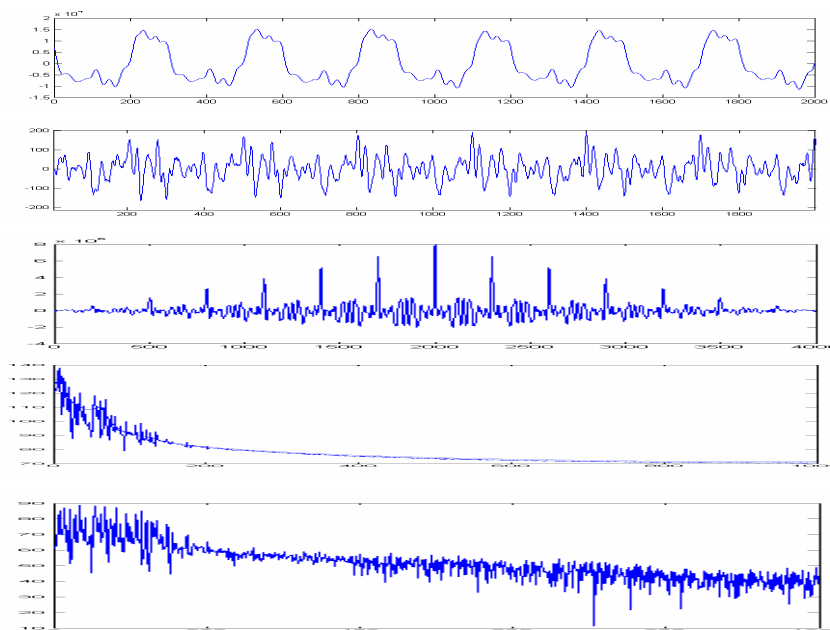


Figure 12.15 - (a) A segment of music signal, (b) output of inverse linear predictor, (c) autocorrelation of signal in (a), (d) DFT and LP spectra of (a), (e) spectrum of predictor's input signal in (b).

12.4.4 Statistical Models of Music

As in speech processing, statistical and probability models are used for music coding and classification. For example, the entropy coding method, where the length of a code assigned to a sample value depends on the

Music Signal Processing

probability of occurrence of that sample value (i.e. the more frequent sample values, or symbols, are assigned shorter codes) is used in music coders such as in MP3 music coders described in section 12.8.

Music compression, music recognition, and computer music composition benefit from probability models of music signals. These probability models describe the signal structures at several different levels:

- (a) At the level of sample or parameter values, the probability models describe the distribution of different parameters of music such as the pitch, the number of harmonics, the spectral envelop, amplitude variation with time, onset/offset times of music events.
- (b) At the level of grammar, a finite state Markovian probability model describes the concurrency and the sequential dependency of different notes in the contexts of chords, parts and rhythmic and melodic variations.
- (c) Hierarchical structures of music can be modelled using structured multi-level finite state abstractions of music generation process.

12.5 The Ear and the Hearing of Sounds

Sound is the auditory sensation of air pressure fluctuations picked up by ears. In this section we study how the ears work and act as transducers that transform the variations of air pressure into electrical firings of neurons decoded by brain. We also study aspects of the psychoacoustics of hearing such as the threshold of hearing, critical bandwidth and auditory masking in frequency and time domains

The ear is a transducer that converts the air pressure variations onto the eardrums into electrical firings of neurons which are transmitted to brain and decoded as different sounds. The presence of an ear on each side of the head allows stereo hearing and the ability to find the direction of arrival of sound from an analysis of the relative intensity and phase (delay) of the sound waves reaching each ear. The ear is composed of three main parts:

- (1) The outer ear picks up air vibrations and directs it to the ear drum.
- (2) The middle ear translates air vibrations into the mechanical vibrations of the bones in the middle ear which impinge on inner tubes.
- (3) The inner ear transforms mechanical vibration in the middle ear into hydraulic vibrations of fluid-filled cochlear tubes that set off neural

firings of hair cells. The anatomy of ear is described in the followings.

12.5.1 The Outer Ear

The outer ear consists of pinna, ear canal, and the outer layer of the eardrum.

The Pinna shown in Figure 12.16 is composed of a cartilage. The pinna and the ear canal are shaped to facilitate efficient transmission of sound pressure waves to the eardrum. The total length of the ear canal in adults is approximately two and a half centimeter, which for a closed-end tube gives a resonance frequency of approximately $f=c/4L$ or 3400 Hz, where c is the speed of propagation of sound (assumed 340 m/s) and L is the length of the ear canal. Note that this frequency also coincides with the frequency of maximum sensitivity of hearing.

Tympanic Membrane (eardrum) - At the end of the ear canal at the tympanic membrane, the energy of air vibrations are transformed into the mechanical energy of eardrum vibrations. The tympanic membrane, or eardrum, is approximately 1 cm in diameter and has three layers, with the outer layer continuous with the skin of the outer ear canal. The central portion of the tympanic membrane provides the active vibrating area in response to sound pressure waves.

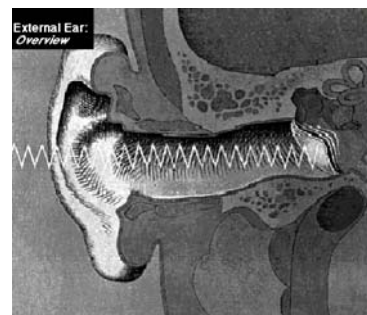


Figure 12.16 External Ear is composed of pinna and ear canal terminates at ear drum where the middle ear starts.

12.5.2 The Middle Ear

The middle ear serves as *an impedance-matching transformer, and also an amplifier*. The middle ear matches the impedance of air in the ear canal to the impedance of the perilymph liquid in the cochlear of the inner ear. The middle ear, shown in Figure 12.17, is composed of a structure of three bones, known as ossicles, which are the smallest bones in the body. The ossicles transmit the vibrations of the sound pressure waves

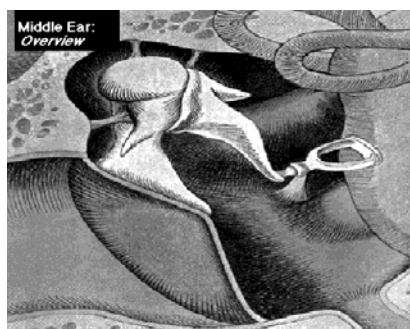


Figure 12.17 The middle ear contains three small bones that transmit eardrum vibrations to the oval window.

Music Signal Processing

from the eardrum to the oval window. Due to a narrowing of the contact area of transmission bone structure from the eardrum to the oval window, amplification of pressure is achieved at the oval window. The three bones of the middle ear are known as malleus, incus and stapes.

Malleus is the nearest, of the three ear bones in the middle ear, to the eardrum. The malleus is attached to the inner layer of the tympanic membrane and vibrates with it.

Incus is attached to the malleus, and so vibrates with it. The incus is also attached to the head of the stapes. As the cross section of the incus is less than that of the malleus at the ear drum, the incoming sound is given a small boost in energy of about 2.5 dB.

Stapes has a footplate seated in the oval window which separates the middle ear from the inner ear. As the incus vibrates, so does the footplate of the stapes. As the vibrating area of the tympanic membrane is larger than the area of the stapes, the incoming sound is given amplification in energy of over 20 dB.

Round Window The round window is the most basal end of the scala tympani, and allows release of hydraulic pressure of the fluid perilymph that is caused by vibration of the stapes within the oval window.

Eustachian Tube This tube connects the middle ear with the nasopharynx of the throat. This tube opens with swallowing or coughing to equalize pressure between the middle ear and ambient pressure that is found in the throat.

12.5.3 The Inner Ear

The inner ear is the main organ of hearing. It transforms the mechanical vibration of the middle ear into a travelling wave pattern on the basilar membrane and then to the neural firings of hair cells. The inner ear is comprised of two main sections, Figure 12.18, vestibular labyrinth and cochlea labyrinth. In the cochlea, the motion is due to vibrations of air; in the vestibular system, the motions transduced arise from head



Figure 12.18 The *inner ear* is composed of cochlear, a labyrinth of fluids and inner and outer hair cells.

Music Signal Processing

movements, inertial effects due to gravity, and ground-borne vibrations. The labyrinth is buried deep in the temporal bone and consists of the two organs the utricle and the sacculus and the semicircular canals. The utricle and sacculus are specialized primarily to respond to *linear accelerations* of the head and *static head position*, whereas the semicircular canals, as their shapes suggest, are specialized for responding to *rotational accelerations* of the head. The *scala tympani*, *scala media* and *scala vestibuli* make up the cochlea which is the organ that converts sound vibrations into neural signals.

Cochlea (derived from the Greek word *kochlias*, for snail), shown in Figure 12.16, is a spiral snail-shaped structure that contains three fluid-filled tubes. The cochlea converts sounds delivered to it as the mechanical vibrations of middle ear at oval window, into electrical signals of neurons. This transduction is performed by specialized sensory cells within the cochlea. The vibration patterns initiated by movements of the stapes footplate of the middle ear on the oval window set up a traveling wave pattern within the cochlea's fluid-filled tubes. This wavelike pattern causes a shearing of the cilia of the outer and inner hair cells. This shearing cause hair cell *depolarization* resulting in neural impulses that the brain interprets as sound. The neural signals, which code the sound's characteristics, are carried

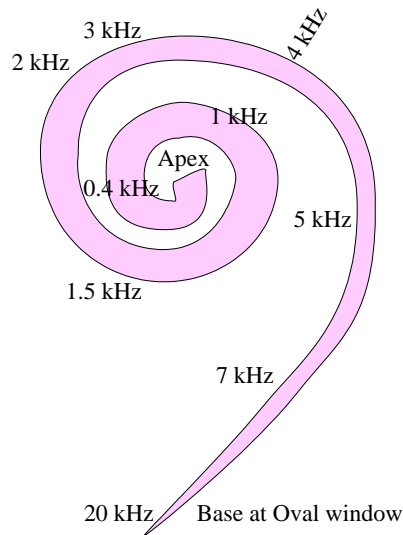


Figure 12.19 Illustration of frequency-place transformation along the length of cochlear.

Width at apex = 0.5 mm
Width a base = 0.04 mm
Length = 32 mm

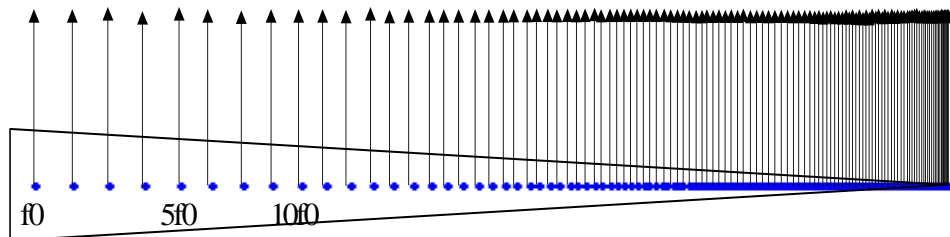


Figure 12.20 Illustration of the frequency to place distribution of the harmonics of a periodic waveform along the length of the basilar membrane of the cochlear. Note that each unit length (e.g. mm) of the basilar membrane at the apex-end where the frequency resolution is higher analyses a smaller bandwidth (is less crowded with frequency components) than at the other end of the basilar membrane where the frequency resolution is lower.

to the brain by the auditory nerve.

The vibrations of the fluids in cochlea affect a *frequency to place transformation* along the basilar membrane. The higher frequencies excite the part of cochlea near the oval window and the lower frequencies excite the parts of cochlea further away from the oval window. Hence, as shown in Figure 12.19, distinct regions of the cochlea and their neural receptors respond to different frequencies. Figure 12.20 illustrates how a periodic waveform with uniformly spaced harmonic frequencies may be registered along the length of basilar membrane.

The Structure of Cochlea - The cochlea's coiled shell contains a bony core and a thin spiral bony shelf (osseous spiral lamina) that winds around the core and divides the bony labyrinth of the cochlea into upper and lower chambers. There is another middle membranous tube in-between these two, Figure 12.21. These three compartments are filled with fluids that conduct the travelling wave patterns.

The upper compartment, called the *scala vestibuli*, leads from the oval window to the apex of the spiral. Hence the mechanical vibrations of stapes on the oval window are converted to travelling pressure waves along the scala vestibuli which at the apex connects to the lower compartment, called the *scala tympani*, that extends from the apex of the cochlea to a membrane-covered opening in the wall of the inner ear called the round window, which acts as a pressure release window. These compartments constitute the bony labyrinth of the cochlea and they are filled with perilymph which has a low potassium K^+ concentration and a high sodium Na^+ concentration. The perilymphatic chamber of the vestibular system has a wide connection to scala vestibuli, which in turn connects to scala tympani by an opening called the *helicotrema* at the apex of the cochlea. Scala tympani is then connected to the cerebrospinal fluid (CSF) of the subarachnoid space by the *cochlear aqueduct*.

The membranous labyrinth of the cochlea is represented by the cochlea *scala media* also known as the cochlear duct. It lies between the two bony compartments and ends as a closed sac at the apex of the cochlea. The cochlear duct is separated from the scala vestibuli by a vestibular membrane called *Reissner's*

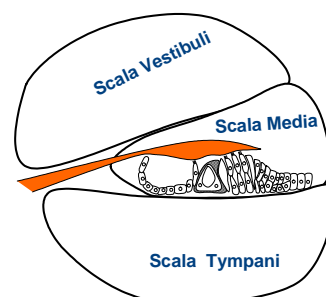


Figure 12.21 – A schematic drawing of a cross section of the cochlear tubes.

Music Signal Processing

membrane and from the scala tympani by a *basilar membrane*. Scala media is filled with endolymph. In contrast to perilymph, endolymph has a high potassium K^+ concentration and a low sodium Na^+ concentration. The endolymphatic system of the cochlea (*scala media*) is connected to the saccule by the *ductus reuniens* and from there connects to the *endolymphatic sac*, which lies in a bony niche within the cranium. The endolymph of the utricle and semi-circular canals also connects to the endolymphatic sac.

Basilar membrane is a ribbon like structure on which rests the organ of corti, the main organ of hearing. It extends from the bony shelf of the cochlea and forms the floor of the cochlear duct. It contains many thousands of fibres, whose lengths and stiffness vary becoming progressively longer and more compliant from the base of the cochlea to its apex. Because of these two gradients of size and stiffness, high frequencies are coded at the basal end with low frequencies progressively coded toward the apical end.

Vibrations entering the perilymph at the oval window travel along the scala vestibuli and pass through the vestibular membrane to enter the endolymph of the cochlear duct, where they cause movements in the basilar membrane. After passing through the basilar membrane, the sound vibrations enter the perilymph of the scala tympani, and their forces are dissipated to the air in the tympanic cavity by movement of the membrane covering the round window.

Travelling Waves in Cochlea

The travelling wave in cochlear fluid can be modeled by a one-dimensional transmission-line. A sound stimulus vibrates the tiny hammer, anvil and stirrup bones that lean against the oval window at the entrance to the cochlea, thus setting the cochlear fluid in motion Figure (12.22). Owing to the incompressibility of the fluid, variation in the longitudinal flow are accompanied by lateral motion of the basilar membrane. This movement is caused by the pressure difference that develops between the fluid ducts as a result of the fluid flux. These mutual interactions between the fluid and the membrane generate a slow wave that travels from the base towards the apex.

As the basilar membrane is elastic, and has very little longitudinal rigidity, adjacent sections of the membrane can move almost independently

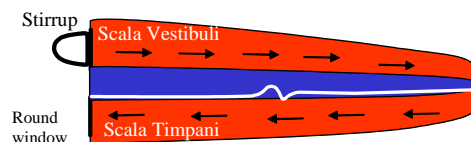


Figure 12.22 – An illustration of traveling wave and the response of basilar membrane.

Music Signal Processing

of one another, being coupled only through the fluid. Moreover, the membrane's lateral stiffness varies greatly along its length, decreasing by about two orders of magnitude from the base to the apex of the cochlea. This changing stiffness means that the wave propagation is dispersive. As the wave advances, its wavelength decreases and it slows down. In regions where the damping is negligible (i.e. near the base) the wave must grow in amplitude to conserve the flow of energy. At some point, however, the motion of the basilar membrane becomes fast enough for viscous drag to become significant. This characteristic place is near the base of the cochlea for higher frequencies. Beyond this point, the damping steals energy from the wave and its amplitude quickly declines.

Organ of Corti shown in figure 12.23, is the main receptor organ of hearing and resides within the scala media. The organ of Corti contains the hearing receptors (hair cells) and is located on the upper surface of the basilar membrane and stretches from the apex to the base of the cochlea. Its receptor cells, which are called hair cells, are arranged in rows and they possess numerous hair like processes that extend into the endolymph of the cochlear duct. As sound vibrations pass through the inner ear, the hairs shear back and forth against the tectorial membrane, and the mechanical deformation of the hairs stimulates the receptor cells. Various receptor cells, however, have different sensitivities to such deformation of the hairs. Thus, a sound that produces a particular frequency of vibration will excite certain receptor cells, while a sound involving another frequency will stimulate a different set of cells.

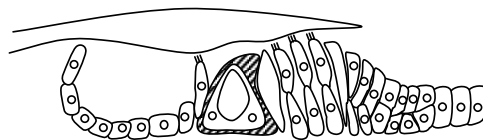


Figure 12.23 Organ of Corti transforms vibrational waves in the fluids of cochlear to neural firings of hair cells.

The outer and inner hair cells of the organ of Corti transform vibrations into neural firing transmitted via auditory nerve to the brain.

Tunnel of Corti is a space filled with endolymph that is bordered by the pillars of Corti and the basilar membrane.

Pillars of Corti are supporting cells that bound the tunnel of Corti. The tunnel of Corti runs the entire length of the cochlear partition.

Tectorial Membrane is a flexible, gelatinous membrane overlying the sensory receptive inner and outer hair cells. The cilia of the outer hair cells are embedded in the tectorial membrane. For inner hair cells, the cilia may or may not be embedded in the tectorial membrane. When the cochlear partition changes position in response to the travelling wave, the shearing of

Music Signal Processing

the cilia is thought to be the stimulus that causes depolarization of the hair cells to produce an action potential.

Hair Cell Receptors- The auditory receptor cells are called hair cells because they possess stereo cilia, which participate in the signal transduction process. The hair cells are located between the basilar (base) membrane and the reticular lamina, a thin membrane that covers the hair cells. The stereo cilia extend beyond the reticular lamina into the gelatinous substance of the tectorial (roof) membrane. Two types of hair cells are present in the cochlea: inner hair cells are located medially, and are separated from the outer hair cells by rods of Corti, Figure (12.24). Hair cells synapse upon dendrites of neurons whose cell bodies are located in the spiral ganglion. Signals detected within the cochlea are relayed via the spiral ganglia to the cochlear nuclei within the brainstem via the auditory nerves VIII. The outer hair cells consist of

three rows of approximately 12000 hair cells. Although they are much greater in number than the inner hair cells, they receive only about 5% of the innervations of the nerve fibres from the acoustic portion of the auditory nerve. These cells contain muscle-like filaments that contract upon stimulation and fine tune the response of the basilar membrane to the movement of the traveling wave. Because of their tuned response, healthy outer hair cells will ring following stimulation. The inner hair cells is one row of approximately 3500 inner hair cells (i.e about 10% of the number of outer hair cells) . These cells receive about 95% of the innervations from the nerve fibres from the acoustic portion of the auditory nerve. These cells have primary responsibility for producing our sensation of hearing. When lost or damaged, a severe to profound hearing loss usually occurs.

Synapses of Hair Cells The stereocilia (hairs) of the hair cells are imbedded in the gelatinous tectorial membrane, which has a relatively high inertial resistance to movement, so that shearing forces, caused by travelling waves in cochlea, bend the hairs. Bending of the cilia causes the hair cells to either depolarize or hyperpolarize, depending upon the direction of the bend. The deflection of the hair-cell stereocilia opens mechanically gated ion channels that allow any small, positively charged ions, primarily potassium and calcium, to enter the cell. The influx of positive ions results in a receptor potential within the cell, which triggers the cell's specific signaling activity.

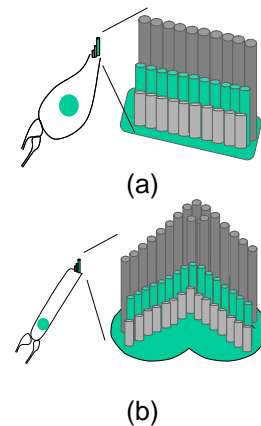


Figure 12.24 (a) Inner and (b) outer hair cells.

Music Signal Processing

Although the inner hair cell itself does not fire an action potential, it has synaptic contacts with auditory nerve fibers that do. The cilia directly regulate the flow of potassium from the endolymph into the hair cell. At rest, there is a small influx of potassium. As the cilia bend in one direction, the K⁺ channel is opened, allowing potassium to enter the cell, thereby depolarizing the cell and increasing the influx of calcium via voltage-sensitive calcium channels. The influx of calcium causes the release of neurotransmitter onto the spiral ganglion cells. Bending in the opposite direction shuts off the potassium influx and hyperpolarizes the hair cells.

In mammalian inner hair cells, the receptor potential triggers the release of neurotransmitters, mainly glutamate, at the basal end of the cell, by opening voltage gated calcium channels. The neurotransmitters diffuse across the narrow space between the hair cell and a nerve terminal, where they then bind to receptors and thus trigger action potentials in the nerve. In this way, the mechanical sound signal is converted into an electrical nerve signal.

Inner hair cells are primarily responsible for sending auditory signals to the brain via the VIII cranial nerve. The outer hair cells play a mechanical role in regulating the responsivity of the basilar membrane to incoming sound vibrations. Motor proteins within the outer hair cells change the cellular structure to either dampen or amplify the incoming sound vibrations. Because of these properties, the outer hair cells are referred to as the cochlear amplifier.

Within the cochlea, hair cell sensitivity to frequencies progresses from high frequencies at the base to low frequencies at the apex. The cells in the single row of inner hair cells passively respond to deflections of sound-induced pressure waves. Cells in the rows of outer hair cells can elongate or shorten in response to motion of the basilar membrane to actively produce amplification or attenuation of the response of the inner hair cells. The scala vestibuli and scala tympani are filled with perilymph, similar in composition to extracellular fluid, while the cochlear duct is filled with endolymph, similar in composition to intracellular fluid. The hair cells synapse on processes of neurons of the spiral ganglia, initiating signal transmission into the central nervous system via the eighth cranial nerve.

Auditory Nerve transmits neural signals from the cochlear and vestibular labyrinth to our brains. It consists of the cochlear nerve, carrying information about hearing, and the vestibular nerve, carrying information about balance. The auditory nerve is also known as the *acoustic nerve*.

Music Signal Processing

The auditory nerve is made up of approximately 30,000 nerve fibres with afferent (sensory) portions projecting from cell bodies in the spiral ganglion and efferent axons coming from cells in the olivary complex. The afferent neurons are divided into type I which synaps on inner hair cells (90-95% of hair cells) and type II which synaps on outer hair cells (5-10% of hair cells). Due to their larger size and myelination type I fibres transmit information to the brainstem 1-2 msec faster than type II counterparts. The connectivity of these two fibre types is different. Type I fibres innervate inner hair cell in a many to one relationship. Type II fibres typically innervate outer hair cells in a many to many relationship. They are sometimes referred to as carrying convergent information since the response of many outer hair cells converges in only a few type II fibres.

Functional Properties of Ear

Frequency to Place Principle - Von Bekesy showed that the cochlea's basilar membrane vibrates much more at its base near the oval window than at the apex in response to high frequencies. He also showed that the opposite is true when low frequency sounds are used. This early work suggested that specific places in the cochlea may be responsible for hearing specific frequencies of sound. This initial work lead to what was called the place principle. Essentially this said that each hair cell and neuron in the cochlea is tuned to respond to a specific frequency and that that frequency can be determined based on it position in the cochlea. This theory proposes that the brain is able to tell which frequencies are being heard based on which set of neurons are firing. In section we consider this process in more detail

Amplitude Discrimination Human hearing has a dynamic range of approximately 110dB; about some 20 dB more than most high end recording equipment which have a dynamic range of 90 dB. Individual neurons, however, only has a dynamic range of 40dB. However, various neuron groups have different threshold such that the threshold of the more sensitive group is approximately 40dB above the less sensitive group. In this way one set of neurons are just beginning to fire above their spontaneous rate when the other group is beginning to fire at their maximal rate. Additionally, at very high sound levels, the frequency tuning properties of the cochlea break down. Thus, we are able to tell that a 103 dB sound is louder than a 100dB sound, but its difficult time telling whether it has a different pitch.

Stereo Hearing and Direction of Arrival of Sounds

Music Signal Processing

Our two ears provide us with stereo hearing ability that allows us to locate the spatial direction of arrival of the sources of sounds. This is achieved by calculation the difference in the intensity and in the time of arrival (TOA) of sound at each ear. The time difference between the sound arriving at left and right ears can be as much as a maximum of $t_{max} = \text{distance between the ears} / \text{speed of sound}$. Assuming the average distance between ears is about 15 cm and the speed of sound is 340 m/s then t_{max} would be about 0.44 ms which is equivalent to one cycle of a 2273 Hz signal. Note that the time differences that allows the brain to locate the direction of arrival of sounds are minute. The phase locking principle seems to also play a role in localizing sounds, particularly at low frequencies. We are able to tell the difference in the position of a sound source based on time delay between when it reaches our right and left ears. In order to tell which direction a sound is coming from parts of the hearing pathway in the brainstem (Superior Olivary Complex and Cochlear Nucleus) detect delays as small as 20 microseconds.

12.6 Psychoacoustics of Hearing

The Psychoacoustics of hearing explains the relationship between the sensation of sounds and the measurable physical properties of the sounds. Audio signal processing methods utilise models of the psychoacoustics and the sensitivity of hearing; i.e. how the perception of the frequencies in a sound is affected by its frequency-time distribution and by its proximity in time and frequency to other signals. In this relation, three basic concepts were developed from experimental observations. These are:

- (a) Absolute threshold of hearing.
- (b) Auditory critical bandwidths.
- (c) Spectral and temporal masking.

12.6.1 Absolute Threshold of Hearing

The absolute threshold of hearing (ATH) is the minimum pressure level of a pure tone that can be detected by a person of acute hearing in noiseless conditions. The frequency dependency of the absolute threshold of hearing, shown in Figure 12.25, was obtained in a series of experiments on human auditory systems in the 1940s [Fletcher]. The variations of the absolute threshold of hearing with frequency may be modelled as

Music Signal Processing

$$ATH(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \text{ (dB SPL)} \quad (12.37)$$

Note that the maximum sensitivity of hearing (i.e. the minimum ATH) occur around 3 to 4 kHz and is about -5 dB SPL. Sound pressure levels just detectable at the frequency of the maximum sensitivity of hearing are not detectable at other frequencies.

In general, two frequency tones of equal power but significantly different frequencies will not sound equally loud. The loudness of a sound is the perception of its power and depends on its frequency and the sensitivity of hearing at that frequency. The perceived loudness of a sound may be expressed in *sones*, where 1 sone is defined as the loudness of a 40 dB tone at 1 kHz.

In music and speech coding the absolute threshold of hearing can be interpreted as the maximum allowable coding noise in frequency (note that this argument does not take into account the noise masking process introduced shortly). In designing coding systems algorithm designers do not know the playback level, hence it is common practice to equate the energy of one bit to that of the sound pressure level at the minimum of threshold of

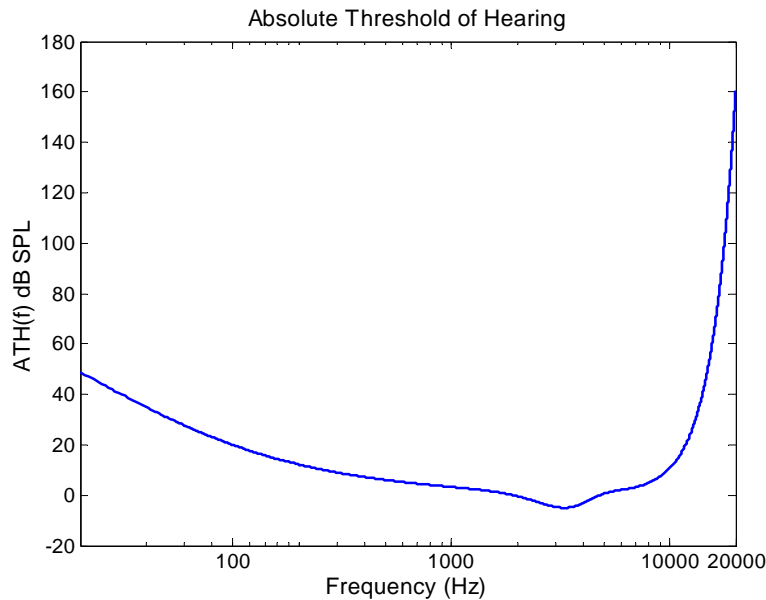


Figure 12.25 – Variations of absolute threshold of hearing (ATH) with frequency.

hearing curve at 4 kHz.

% PlotATH plots the audio threshold of hearing curve

Music Signal Processing

function PlotATH()

```
f=20:20000
```

```
ATH=(3.64*((.001*f).^0.8))-(6.5*(exp(-0.6*(-3.3+0.001*f).^2)))+(0.001*((.001*f).^4));
```

```
plot (ATH);
```

function HearingAudioResponse()

Generates, plays back and displays a sine wave of constant magnitude with its frequency sweeping the range 0 to 20000 Hz. The wave is played back. Played at low volume this program demonstrates the variable perceived loudness of hearing of a constant amplitude sinewave as a function of frequency.

12.6.2 Critical Bands of Hearing

Critical bands of hearing are segments of cochlea, about 1.3 mm long, which act like band-pass filters. The audible effect of two or more tones residing within a critical band would be different than if the same signals were in different critical bands.

The concept of critical bands of hearing is based on the observations of the perception of audio signals along the basilar membrane of the cochlea where a frequency to place transformations takes place Figure 12.24. Experiments show that at any frequency along the cochlear the ear behaves like a series of band-pass filters known as critical bands which have frequency dependent-bandwidth. Signals within a critical band affect each other' perception far more than signals in different bands. Critical bandwidth increases with the centre frequency of the band. Below 500 Hz bandwidths are constant at about equal 100 Hz. Over 500 Hz the bandwidth of each critical band is 20% larger than the preceding band.

As stated, the effects of two tones or narrowband noise within a critical band are substantially different than the effect of two tones or narrowband noise in different critical bands. For example the ear is not able to resolve two pure tones within a critical band and perceives their combined effect as a beating of the tones. Whereas two tones with their frequencies at different bands can be resolved and heard as distinct tones.

Critical bandwidth around a frequency can be defined as the bandwidth at which the subjective response of hearing changes abruptly. For example, the perceived loudness of a narrowband noise at a constant sound pressure level remains constant as the bandwidth of noise is increased up to the critical bandwidth. However beyond the critical bandwidth the loudness increases.

Music Signal Processing

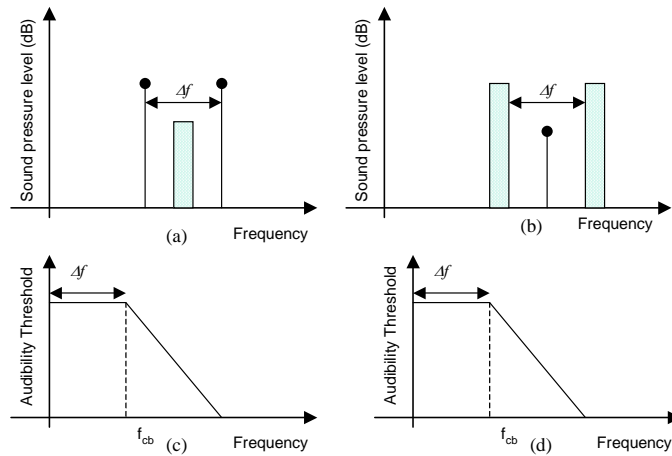


Figure 12.26 Illustration of critical bands, (a) a narrow band noise with two masking tones, (b) a tone with two masking narrowband noise, (c) and d) show the thresholds of hearing as a function of the distance between the maskers.

Figures 12.26 illustrate two experiments used to demonstrate critical bandwidth. In Figure 12.26.a and c the detection threshold of a narrowband noise flanked between two masking tones remains constant as long as the separation between the two masking tones remains within a critical band. In Figure 12.26.b and d the detection level of a tone positioned between two narrowband noise signals remains constant as long as the separation between the two narrow band noise signals remains within a critical band. The detection threshold decreases rapidly as the separation of maskers increases beyond a critical bandwidth.

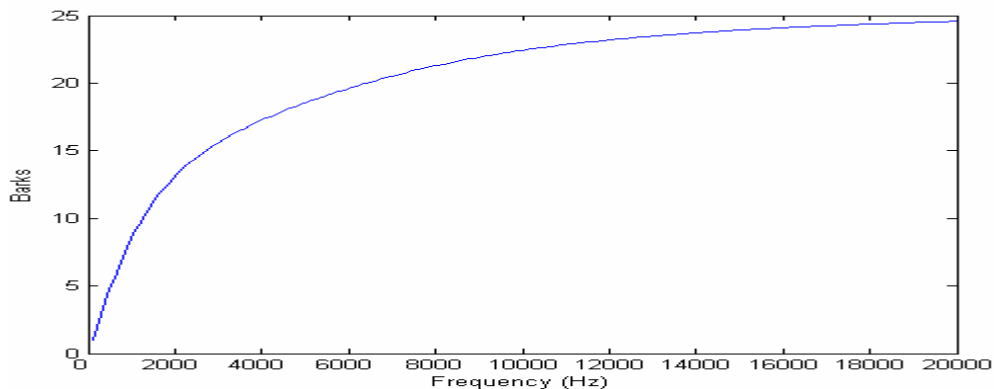


Figure 12.27 Illustration of the relationship between linear frequency units (Hz) and critical band units of Bark.

Music Signal Processing

Critical bandwidth around a frequency f remains constant at (about 100 Hz bandwidth) up to 500 Hz and then increases at about 20% of the centre frequency above 500 Hz. The variation of critical bandwidth BW_c with frequency can be modelled as

$$BW_c(f) = 25 + 75 \left[1 + 1.4 \left(\frac{f}{1000} \right)^2 \right]^{0.69} \quad (\text{Hz}) \quad (12.38)$$

A distance of one critical band is termed as one *bark* and the formula for mapping linear frequency to bark scale is given by

$$z_c(f) = 13 \arctan(0.00076f) + 3.5 \arctan \left[\left(\frac{f}{7500} \right)^2 \right] \quad (\text{Bark}) \quad (12.39)$$

Critical bandwidth varies continuously with frequency as shown in Figure 12.27. However in audio processing application it is useful to define a discrete set of critical bands as shown in table 12.10.

12.6.3 Audio Masking

The concept of masking, i.e. a signal drawn inaudible by a louder signal in its time/space proximity, is a familiar experience. In general a large amplitude signal that happens to be in time, space and frequency proximity of a smaller amplitude signal can partially or totally mask the latter.

Simultaneous masking occurs when two sounds occur at the same

Critical Band/centre (Hz)	Frequency (Hz)			Critical Band/centre(Hz)	Frequency (Hz)		
	Low	High	Width		Low	High	Width
0 / 50	0	100	100	13 / 2150	2000	2320	320
1 / 150	100	200	100	14 / 2500	2320	2700	380
2 / 250	200	300	100	15 / 2900	2700	3150	450
3 / 350	300	400	100	16 / 3400	3150	3700	550
4 / 450	400	510	110	17 / 4000	3700	4400	700
5 / 570	510	630	120	18 / 4800	4400	5300	900
6 / 700	630	770	140	19 / 5800	5300	6400	1100
7 / 840	770	920	150	20 / 7000	6400	7700	1300
8 / 1000	920	1080	160	21 / 8500	7700	9500	1800
9 / 1175	1080	1270	190	22 / 10500	9500	12000	2500
10 / 1370	1270	1480	210	23 / 13500	12000	15500	3500
11 / 1600	1480	1720	240	24 / 19500	15500	22050	6550
12 / 1850	1720	2000	280				

time, such as when a conversation (the masked signal) is rendered inaudible by a noisy passing vehicle (the masker).

Masking can also happen with non-simultaneous signals. Backward masking occurs when the masked signal ends before the masker begins, i.e. the masker masks a signal preceding it. Forward masking occurs when the masked signal begins after the masker has ended, that is the masker masks a signal succeeding it. Hence a signal at time t , $x(t)$, can be masked by another signal $y(t+\tau)$ or $y(t-\tau)$. Note that maskings of non-simultaneous signals are in part due to delays in the transmission and processing of audio signals from the ear to brain.

Masking becomes stronger as the sounds get closer together in time, space and frequency. For example, simultaneous masking is stronger than either forward or backward masking because the sounds occur at the same time. In this section we consider temporal and spectral masking effects.

12.6.4 Spectral Masking

Spectral masking is related to the concept of critical bands of hearing. For music coding two types of spectral masking are defined, tone-masking-noise and noise-masking tones. A tone masking noise masks a noise within a critical bandwidth provided that the noise spectrum is below a threshold that depends on the strength of the masking tone. A noise masking tone masks a tone within a critical bandwidth provided that the tone strength is below a threshold dependent on the strength of the noise.

The masking effect of a tone (or noise) is not confined to within the critical bands. Inter-band masking is also observed and depends on the strength of the masking tone or noise. Figure 12.28 illustrates the masking effect of a pure tone. The spread of masking is often modelled by a triangular shape filter with the slopes of +25 dB and -10 dB per bark. The shape of the spreading function can be approximated as

$$SF(x) = 15.81 + 7.5(x + .474) - 17.5\sqrt{1 + (x + .474)^2} \quad \text{dB} \quad (12.40)$$

where x has units of barks.

```
Nbarks=2; f=-Nbarks: .01 : Nbarks;
```

```
SF=15.81+7.5*(x+0.474)-17.5*(1+(x+0.474).^2).^0.5;
```

```
plot(SF);
```

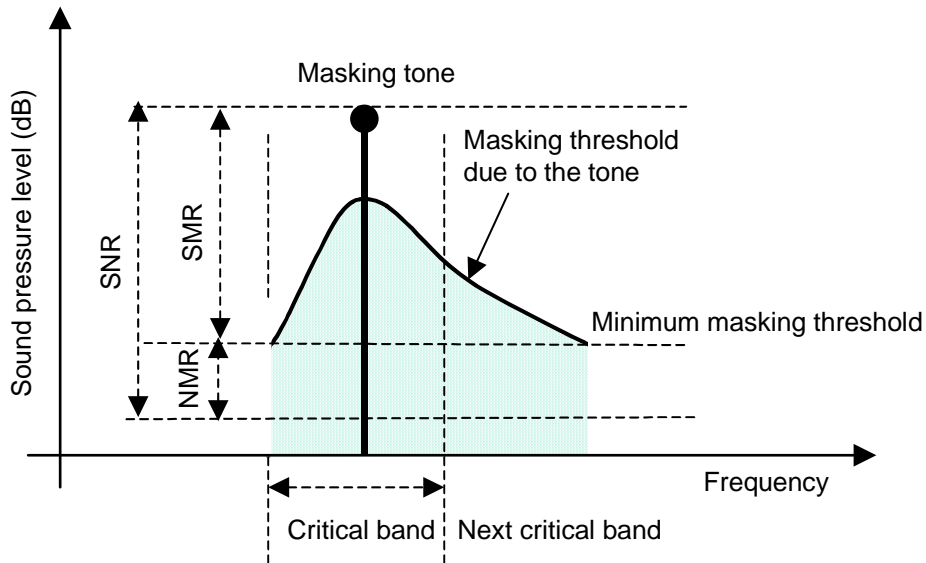


Figure 12.28 Illustration of the masking effects of a tone.

In psychoacoustic coders, after accounting for the shape of the spread function, the masking thresholds for noise and tone components are obtained as

$$TH_N = E_T - 14.5 - B \quad \text{dB} \quad (12.41)$$

$$TH_T = E_N - 14.5 - K \quad \text{dB} \quad (12.42)$$

where E_T and E_N are energy levels of the tone and noise maskers and B is the critical band number. The parameter K is typically selected between 3 to 5 dB. Masking thresholds are also referred to as the just noticeable distortion (JND) levels. For perceptual coding, masking signals are first classified as tones or noise maskers and then the JND values are calculated to shape the spectrum and keep the coding noise below audible levels.

12.6.5 Temporal Masking

Temporal masking, due to the simultaneous occurrence of two signals, is a common experience; for example when a passing vehicle's noise drowns a conversation. However temporal masking can also be non-simultaneous. A signal can mask the audibility of another signal even before the masker begins (an effect known as backward masking or pre-masking) and also sometime after the masker is ended (an effect known as forward masking or

post-masking). Figure 12.29 shows a sketch of the shape of temporal masking in time. Pre-masking and post-masking clearly point to delays in the processing of sounds in the auditory-brain system. Note that the backward masking is much shorter than the forward making effect. Backward masking lasts about 5 ms whereas forward masking can last up to 300 ms.

12.6.6 Use of Psycho-acoustic Model in Music Coding

In this section we consider an example of calculation of audio masking thresholds in MPEG-1 layer-1 audio coder. The aim is to calculate the time-varying maximum allowable quantisation noise at each frequency that is masked and rendered inaudible by signal activity in other frequencies. The sampling rate is 44100 samples per second with 16 bits per sample.

Step 1 - Spectral Analysis and SPL Normalisation

The psycho-acoustic model of hearing uses a DFT of 512 points. The signal is divided into segments of 512 samples (about 11.6 ms) and windowed with an overlap of 32 samples between successive windows, therefore each window contains 480 new samples. At a sampling rate of 44100 samples/second this gives a spectral resolution of $\Delta f = F_s/N = 44100/512 = 86.13$ Hz.

The audio signal amplitude is normalised as

$$x_n(m) = \frac{x(m)}{N(2^{b-1})} \quad (12.43)$$

where N is the DFT length and b is the number of bits per sample ($b=16$ bits)

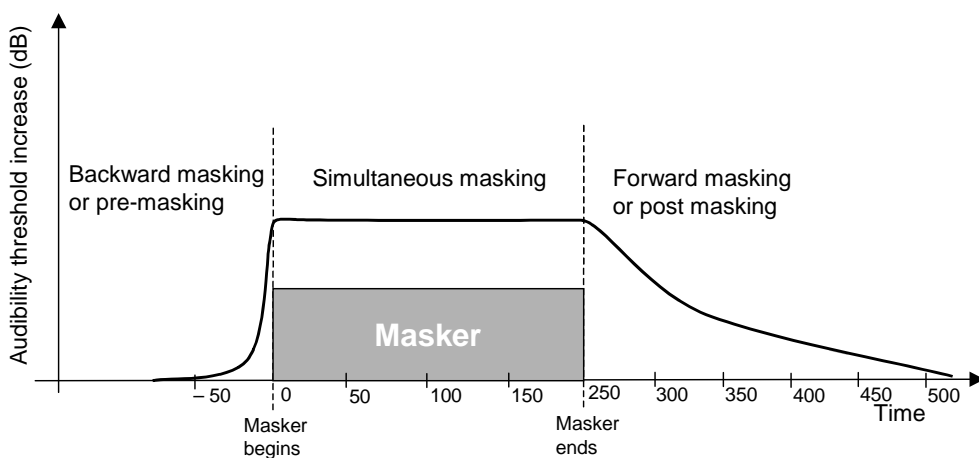


Figure 12.29 Illustration of temporal making effects.

. The adjusted short-time power spectrum of the signal is obtained as

$$P(k) = PN + 10\log\left(|X(k)|^2\right) \quad (12.44)$$

where $PN = 20\log_{10}(2^{b-1}) = 90$ dB which is the full amplitude of a 16 bit quantiser;. For a normalized full-scale sine wave of normalised amplitude $A=1$ we have $|X(k)|^2=0.5$ and $90+20\log_{10}0.5=84$ dB.

Step 2 – Identification of Tonal and Noise Maskers

Local maxima in the frequency components of the spectrum that exceed the neighbouring spectral components within a predefined frequency distance (in bark scale) by more than 7 dB are classified as tonal frequency components. The rule for classification of frequency components as tonal is as follows:

$$S_T = \left\{ P(k) \text{ if } \begin{cases} P(k) > P(k \pm 1) \text{ and} \\ P(k) > P(k \pm \Delta_k) + 7 \text{ dB} \end{cases} \right. \quad (12.45)$$

where

$$\Delta_k \in \begin{cases} 2 & 2 < k < 63 & (0.17 - 5.5 \text{ kHz}) \\ [2,3] & 63 \leq k < 127 & (5.5 - 11 \text{ kHz}) \\ [2,6] & 127 \leq k < 63 & (11 - 20 \text{ kHz}) \end{cases} \quad (12.46)$$

Tonal maskers are computed from the peaks listed in the tonal set S_T as

$$P_{TM}(k) = 10\log_{10} \sum_{j=-1}^1 10^{0.1P(k+j)} \quad (\text{dB}) \quad (12.47)$$

The noise masker values for each critical band are obtained from the non-tonal signal components as

$$P_{NM}(\bar{k}) = 10\log_{10} \sum_j 10^{0.1P(j)} \Big|_{\forall P(j) \notin P(k \pm \Delta k)} \quad (\text{dB}) \quad (12.48)$$

where \bar{k} is the geometric mean of the spectral lines within a critical band given by

$$\bar{k} = \left(\prod_{j=l}^u j \right)^{1/(u-l+1)} \quad (\text{dB}) \quad (12.49)$$

Where l and u are the lower and upper frequency (DFT) bin indices of the critical band. The tonal and noise maskers that have a value less than the

Music Signal Processing

absolute threshold of hearing are discarded. Then any two tonal values that fall within a half bark of each other are replaced by the bigger of the two. The remained are sub-sampled without loss of information according to the following formula

$$P_{TM, NM}(i) = P_{TM, NM}(k) \quad (12.50)$$

$$i = \begin{cases} k & 1 \leq k \leq 48 \\ k + (k \bmod 2) & 49 \leq k \leq 96 \\ k + 3 - ((k - 1) \bmod 4) & 97 \leq k \leq 232 \end{cases} \quad (12.51)$$

The effect of eq(12.50) is a 2:1 decimation of the masker bins in the critical bands 18-21 and a 4:1 decimation in the critical bands 22-25 with no loss of the masking components.

Calculation of Masking Thresholds

The masking threshold $T_{TM}(i, j)$ at frequency bin i due to a tone masker at frequency bin j , $P_{TM}(j)$, is modelled by

$$T_{TM}(i, j) = P_{TM}(j) - 0.275 z(j) + SF(i, j) - 6.025 \quad \text{dB SPL} \quad (12.52)$$

where $z(j)$ is the bark frequency of bin j . $SF(i, j)$ is a spreading function model of the masking effect at bin i due to a masking tone $P_{TM}(j)$ at bin j modelled by a piecewise linear function as

$$SF(i, j) = \begin{cases} 17\Delta_z - 0.4P_{TM}(j) + 11 & -3 \leq \Delta_z < -1 \\ (0.4P_{TM}(j) + 6)\Delta_z & -1 \leq \Delta_z < 0 \\ -17\Delta_z & 0 \leq \Delta_z < 1 \\ (0.15P_{TM}(j) - 17)\Delta_z - 0.15P_{TM}(j) & 1 \leq \Delta_z < 8 \end{cases} \quad (\text{dB SPL}) \quad (12.53)$$

where $\Delta_z = z(i) - z(j)$ is the separation in barks between the maskee and the masker. Similarly, $T_{NM}(i, j)$, the masking threshold at frequency bin i due to a noise masker $P_{NM}(j)$ at frequency bin j is given by

$$T_{NM}(i, j) = P_{NM}(j) - 0.175 z(j) + SF(i, j) - 2.025 \quad \text{dB} \quad (12.54)$$

Music Signal Processing

$SF(i,j)$ for noise masker is obtained by replacing $P_{TM}(j)$ with $P_{NM}(j)$ in Equation(xx). The global masking threshold at bin i due to all tone and noise maskers is given by

$$T_g(i) = 10 \log_{10} \left(10^{0.1ATH(i)} + \sum_{l=1}^L 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^M 10^{0.1T_{TM}(i,m)} \right) \text{ dB} \quad (12.55)$$

where L and M are the number of tonal and noise maskers.

12.7 Music Coding (Compression).

The transmission bandwidth and the storage capacity requirement for digital music depend on the sampling rate and the number of bits per sample. A stereo music with left and right channels sampled at 44100 Hz and quantised with 16 bits per sample generates data at a rate of

$$2 \times 44100 \times 16 = 1,411,200 \text{ bits per second} \quad (12.56)$$

and requires about 5 Gigabits or 635 Mega bytes of storage per hour of music.

The objective of music compression is to reduce the bit rate as far as possible while maintaining high fidelity. This is usually achieved through decomposition of the music signal into a series of de-correlated time-frequency components or a set of source-filter parameters of a synthesizer model of music. Using a psycho-acoustic model, the various components of the decomposed music signal are each allocated the minimum number of bits required to maintain the quantization noise masked and inaudible and achieve high fidelity of reconstructed signal.

The Signal Structures Used in Music Compression

In general music coders utilise three aspects of the signals to reduce the bit rate and simultaneously maintain high fidelity, these are:

- (a) The correlation structure of music signals; this means the part of a sample $x(m)$ that is predictable from past samples can be modelled and need not be transmitted.
- (b) The psychoacoustics of hearing, this means that the inaudible level of quantisation noise, at each time-frequency component, is time-varying and depends on the amount of noise that can be asked.

Music Signal Processing

- (c) The statistical distribution of music signals. This means the probability distribution can be used for efficient non-uniform quantisation and for efficient transmission in variable length coding.

Specifically, the signal structures utilised in music coding are as follows:

- Short-term correlations of successive samples of music, within a frame of say 5 to 30 ms, can be modelled by principle component analysis (e.g. discrete cosine transform (DCT)) or by a linear prediction model.
- Long-term inter-frame correlation and periodic patterns can be modelled by a pitch model and an amplitude envelope model or by modelling the correlation of successive frames of music with a DCT or linear prediction model.
- Non-uniform probability distribution of the signal variables can be utilised in the quantisation process and also in entropy coding to assign variable length codes to different signal values according to their probability, e.g. Huffman code where more frequently occurring values are assigned shorter code lengths.
- Cross correlation between the left and the right channels of a stereo music can be used in joint channel coding to reduce the bit rate.
- The masking effects of the auditory hearing can be used to allocate the minimum number of bits to each part of the signal such that the quantisation noise remains masked below the just noticeable distortion levels.

In this section we consider adaptive transform coding and MPEG (MP3) music coding methods.

12.7.1 The Basic Principles of Music Compression

The quantization noise of a music coder depends on a number of factors that include; (a) the number of bits per sample, (b) the efficiency of utilisation of the distributions of the music signal in time and frequency domains and (c) the efficiency of utilisation of the psychoacoustics of hearing. The goal of audio coding is to utilise the time-frequency distribution of the signal and to shape the time-frequency distribution of the quantisation noise such that the quantisation noise is made inaudible and the reconstructed signal is indistinguishable from the original signal.

Music Signal Processing

In general, audio coders operate by decomposing the signal into a set of units; each unit corresponds to a certain range in time and frequency. Using this time-frequency distribution, the signal is analysed according to psychoacoustic principles. This analysis indicates which set of frequency spectral components are critical to hearing and must be coded with higher precision and assigned more bits, and which set of frequency components are less important and can tolerate relatively more quantization noise without degrading the perceived sound quality. Based on this information, the available bits are distributed and allocated to the time-frequency groups and the spectral coefficients in each frequency group are quantized with the allocated bits. In the decoder, the quantized spectra are reconstructed according to the bit allocation pattern and then synthesized into an audio signal.

12.7.2 Adaptive Transform Coding

A transform coder, Figure (12.30), consists of the following sections:

- (a) Buffer and window, divides the signal into overlapping segments of length N samples. The segment length may be variable as it controls the time and frequency resolutions and affects the severity of pre-echo distortion described in section 12.20.
- (b) Frequency analysis transforms the signal into frequency. Discrete cosine transform is often used due to its ability to compress most of the signal energy into a relatively limited number of principal components. Fourier transform may also be used.
- (c) A pre-echo detector detects abrupt changes (e.g. attacks) in signal energy which can cause an audible spread of the quantisation noise of the high energy part of a frame of music into the low energy part of the frame and hence produce a pre-echo distortion.
- (d) Psycho-acoustic model calculates the tonal and non-tonal components of the signal and then estimates the just noticeable

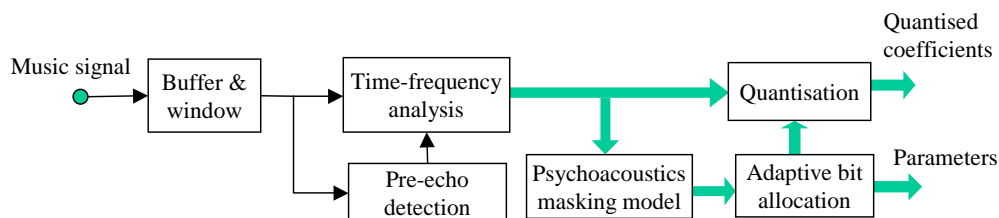


Figure 12.30 – Outline of adaptive transform coding.

distortion (JND) levels for each frequency band of each signal frame.

- (e) The quantizer represents each frequency component with k bits. One quantiser may be used for all frequencies. However, it may be advantageous to use a set of quantisers to span the frequency range; one quantiser for each group of frequency bins.
- (f) The bit allocation modules, known as rate-distortion loop, allocates bits to each quantiser in order to satisfy two requirements: (a) to keep the total number of bits within the intended bit rate, (3) to keep the distortion of each frequency partial below the calculated JND levels.

12.7.3 Time-Frequency Analysis

The time-frequency analysis of music can be performed using a number of different methods:

- (a) Frequency transformation of frames of music using a DCT or FFT.
- (b) A bank of band-pass filters to split the signal into sub-bands
- (c) A combination of filter-bank followed by finer frequency analysis using a transform such as DCT as in MPEG and Sony's Atrac.

Time-frequency analysis is a central part of music coders such as MP3 and Atrac. In MP3 the signal is split into 32 equal-width bands of frequency. The sub-band signals are then transformed into finer frequency components using the modified discrete cosine transform (MDCT). The MDCT allows up to 50% overlap between time-domain windows, leading to improved frequency resolution while maintaining critical sampling.

In Sony's Atrac coder the signal is split into three sub-bands which span the frequency range 0-5.5 kHz, 5.5-11 kHz, and 11-22 kHz. The sub-band decomposition is performed using quadrature mirror filters (QMF). The input signal is divided into upper and lower bands by the first QMF, and the lower band is divided again by a second QMF. The use of QMF's ensures that time-domain aliasing caused by the sub-band decomposition will be cancelled during reconstruction.

12.7.4 Variable Length Transform: Pre-Echo Control

A change in the value of a frequency component of a signal segment would have an affect on the values of every time domain sample of that segment and vice versa. When a signal window, that contains a fast rising "attack"

Music Signal Processing

part of a music track, is transformed to frequency domain, quantized and then transformed back to time, Figure(12.31), the quantisation noise in the high energy attack part of the signal spreads over the entire length of the window including the relatively lower energy pre-attack part of the signal window. This results in an audible noise in the low energy part of the window before the attack part of the signal, this noise is known as pre-echo noise.

To prevent pre-echo noise a signal ‘attack’ detector is used and a shorter duration window is employed during the ‘attack’ portions of the signal. This would limit the spread of quantisation noise over time. However, as frequency resolution is inversely proportional to the duration of the signal window, a shorter duration window gives a lower frequency resolution. The compromise solution is to use two or more windows of different durations depending on the time-varying characteristics of music. For example, at a sampling rate of 44100 Hz, we can switch between three windows: a long window of length 512 (11.6 ms) samples, a medium window of length 128 samples (2.9 ms) and a short window of 64 samples (1.45 ms).

Note that a short window (necessary for the attack part of a signal) is not necessary for signal decay, because the quantization noise will be masked by forward masking which persists longer than backward masking.

12.7.5 Quantization of Spectral Values

The spectral coefficients of each frame of music signal are quantized using two parameters: (1) a word length and (2) a scale factor. The scale factor

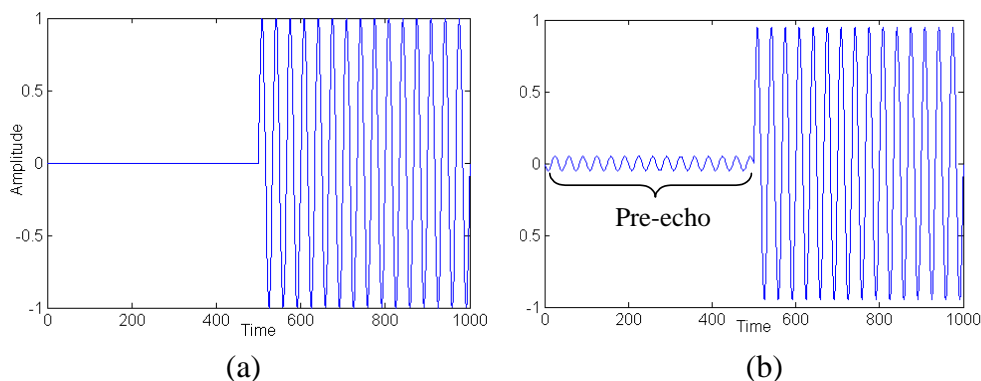


Figure 12.31 Illustration of pre-echo in transform coding of a sinewave: (a) original signal, (b) quantized signal with pre-echo.

Music Signal Processing

defines the full-scale range of the quantization, and the word length defines the precision within that scale. Furthermore, the quantization steps may be spread non-uniformly to reflect the non-uniform probability distribution of sample values.

Each group of frequencies has the same word length and scale factor, reflecting the psycho-acoustic similarity of the frequencies within a group. The scale factor is chosen from a fixed list of possibilities, and reflects the magnitude of the spectral coefficients in each group of frequencies. The word-length is determined by the bit allocation algorithm. For each audio frame the following signal information is stored:

- MDCT block size mode (long or short).
- Word length data for each group of frequencies.
- Scale factor code for each group of frequencies.
- Quantized spectral coefficients.

Coding of Quantized Spectral Values

The quantised spectral values are encoded with a binary number. Efficient entropy coding of spectral values can be achieved using Huffman coding to assign variable-length binary codes to different quantisation levels. Shorter codewords are assigned to more frequently occurring values and longer codewords to infrequently occurring values.

The Huffman coder and the quantisation process can interact in a rate-distortion loop as increasing the quantisation step size will yield a higher proportion of samples with smaller quantisation levels requiring smaller length codes and a consequent reduction in the bit rate but at the expense of more quantisation noise and vice versa.

Bit Allocation Method

The bit allocation algorithm divides the available bits between the various frequency bands through adjusting the quantisation step size. Frequency bands allocated a large number of bits will have less quantization noise; bands with fewer bits will have more noise. For high fidelity the bit allocation algorithm must be adaptively matched to the masking properties of signal to ensure that sub-bands that are critical to the hearing process have sufficient bits, and that the noise in non-critical bands is not perceptually significant.

Rate-Distortion Loop

An iterative rate-distortion loop is used to adjust the number of quantisation levels for each group of frequency bins until the quantisation noise at each frequency remains below the just noticeable distortion (JND) threshold level, subject to the constraint that the rate remains within the maximum number of bits available for each frame.

The rate part of the rate-distortion loop adjusts the overall quantisation parameters to maintain the overall bit rate within the required level. The distortion part of the loop adjusts the quantisation levels (and hence the bit rate) within each group of frequencies to maintain the quantisation noise below noticeable distortion level. The rate control and distortion control form an iterative optimization loop, the goal of which is to achieve the required target bit rate and inaudible distortion. There is a maximum number of iteration at which the loop will be terminated in order to prevent the system from iterating infinitely for a frame for which it is not possible to obtain the required audio transparency within the available number of bits.

12.8 High Quality Audio Coding: MPEG-1/2 layer-3 (MP3)

MPEG is the acronym for the **m**oving **p**icture **e**xperts **g**roup established in 1988 to develop open standards for development of coders for moving pictures and audio. Open standards are specifications that are available to developers interested in implementing the standard. Usually an implementation example is provided to avoid misinterpretation of the text of the standard.

The audio coding standard developed by the MPEG group is used in many applications including digital audio broadcasting, internet audio, portable audio, DVD and audio storage. A popular version of MPEG is MP3 developed at the Fraunhofer Institute, a German audio research laboratory. MP3 (short for MPEG-1/2, audio layer) is a subset of MPEG compression that can take a two-minute, 22-megabyte, music track on a CD, and reduce it by a factor of 16 to about 1.4 megabytes, the size of a floppy disk.

There are a few versions of MPEG standard, each version includes a higher level of quality, flexibility and application and offers different choices..

MPEG-1 audio consists of three operating modes, called layers, with increasing complexity, delay and quality from layer-1 to layer-3. MPEG-1

Music Signal Processing

defines audio compression at sampling rates of 32 kHz, 44.1 kHz and 48 kHz. It works with both mono and stereo signals, and a technique called joint stereo coding can be used for efficient coding of the left and right channels. Available bit rates are 32-192 kbps for mono and 64-384 kbps for stereo music. MPEG-1 Layer-3 provides high quality audio at about 128 kbps for stereo signals.

MPEG-2 introduced new concepts for video coding and digital TV. It extends MPEG-1 audio sampling rates to half rates to include 16 kHz, 22.05 kHz, and 24 kHz. It also includes the 3/2channel format comprising right centre left channels together with right and left surround channels.

MPEG-3 was to define video coding for high definition television (HDTV) applications. However, as MPEG-2 contains all that is needed for HDTV, MPEG-3 was rolled into MPEG-2.

MPEG-4 is more concerned with new functionalities than better compression efficiency. The major applications of MPEG-4 are mobile and fixed terminals, database access, communications and interactive services. MPEG-4 audio consists of audio coders spanning the range from 2 kbps low bit rate speech, up to 64 kbps/channel high quality audio.

MPEG-7 is a content representation standard for multimedia information search engines, filtering, management and processing of data.

MPEG Bit Rates

MPEG audio allows flexible compression ratio. Within the prescribed limits of the open standard, the selection of the bit rate is left to the implementer and/or the operator of the audio coder. The standard bit rate is defined in the range from 8 kbps to 320 kbps. The open standard also enables the use of variable bit-rate coding and fixed bit-rate coding within the prescribed limits.

12.8.1 MPEG Structure

Figure 12.32 illustrates a block diagram structure of an MP3 coder. This is frequency domain coder in that segments of 576 time domain samples are transformed into 576 frequency domain samples. The available bits are then allocated non-uniformly among various frequency components depending on the just noticeable distortion (JND) levels at different frequencies calculated from the psychoacoustic model.

It consists of the following subsystems.

The Filter bank can be a uniformly spaced filter bank or a non-uniformly spaced filter bank where the filters' bandwidth are matched to the critical bands of hearing i.e. high resolution at lower frequencies and lower resolutions at higher frequencies (see table 12.10)

In this example we assume the filter bank consists of 32 equal-bandwidth poly-phase filters of length 512 taps each. The filters are equally spaced and split a bandwidth of 24 kHz, at a sampling rate of 48 KHz, to 32 bands of width 750 Hz. The output of each filter is down sampled by a factor of 32:1. After down sampling each subband has a sampling rate of 1500 Hz, and the total numbers of samples across 32 subbands is 48000 samples per second; the same as the input sampling rate before band splitting.

Modified discrete cosine transform Each sub-band output is segmented into segments of 18 samples long – corresponding to a segment length of $18 \times 32 = 576$ samples of the original signal before down sampling or 12 ms duration– and transformed by a modified discrete cosine transform (MDCT). Hence the 750 Hz width of each subband is further decomposed into 18 frequency bins with a frequency resolution of $750/18 = 41.7$ Hz.

The auditory perceptual model is based on critical bands of hearing and masking thresholds as described in section 12.6. A 1024-samples FFT of the music signal (with a frequency resolution $F_s/N = 48000/1204 = 46.875$ Hz) is used to calculate the noise masking thresholds, the so called just noticeable distortion (JND) levels; this is the amount of quantization noise in each frequency band that would be masked and made inaudible by the signal energy at and around that band as explained in section 12.6. The frequency bands for calculation of the masking thresholds are based on the critical bands of hearing. If the quantisation noise energy can be kept below the masking threshold then the compressed signal would have the same

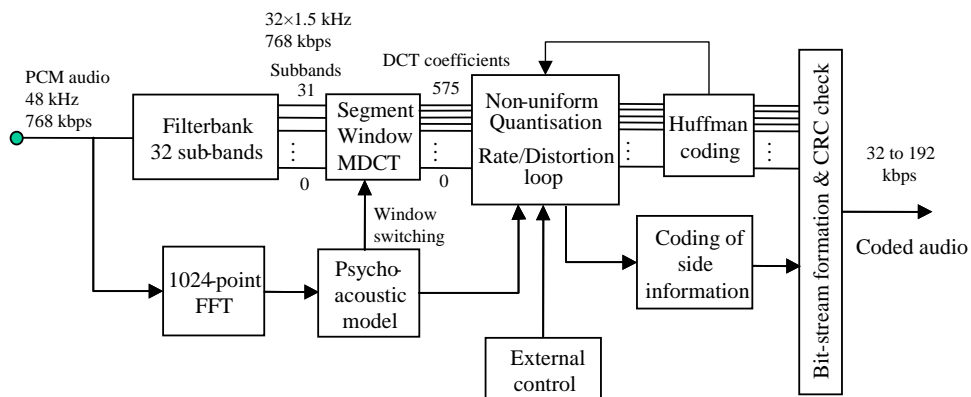


Figure 12.32 Illustration of MPEG1 layer-3 system.

Music Signal Processing

Symbol Probability	Huffman Code
$P_8 = 0.05$	00011
$P_7 = 0.1$	0000
$P_6 = 0.15$	001
$P_5 = 0.2$	10
$P_4 = 0.2$	11
$P_3 = 0.15$	010
$P_2 = 0.1$	011
$P_1 = 0.05$	00010

Figure 12.33 Illustration of Huffman coding of the quantisation levels of a coder. In this case the minimum bits/symbol indicated by the entropy is 2.8464. The Huffman code shown achieves 2.9 bits/sample.

transparent perceptual audio quality as the original signal.

Quantisation and coding processes aim to distribute the available bits among the DCT coefficients such that the quantisation noise remains masked. This is achieved through an iterative two-stage optimization loop. A power law quantiser is used so that large spectral values are coded with a larger quantization step size, as a higher signal energy masks more quantisation noise. The quantised values are then Huffman coded Figure (12.33). To adapt the coder to the local statistics of the input audio signal the best Huffman coding table is selected from a number of choices.

The Huffman coder is a probabilistic coding method that achieves coding efficiency through assigning shorter length codewords to more probable (i.e. more frequent) signal values and longer length codewords to less frequent values. Consequently, for audio signals smaller quantised values, which are more frequent, are assigned shorter length codewords and larger values, which are less frequent, are assigned longer length codewords.

Quantisation consists of two loops an inner loop that adjusts the rate to keep the overall bit rate within the required limit and an outer loop that aims to keep the distortion in each critical band masked. The rate-loop interacts with Huffman coder as explained next.

Rate Loop (inner loop) adjusts the bit rate, and maintains it at or below the target rate, through control of the relative proportion of samples with smaller and larger values using a global gain. If the number of available bits is not enough to Huffman code a block of data within the prespecified rate

then the global gain is adjusted to result in a larger quantisation step size and a hence higher proportion of smaller-valued quantised samples requiring shorter length codewords. This is repeated with different values of quantisation step size until the bit rate requirements is no more than the target bit rate for each block.

Distortion Loop (outer loop) When the quantisation distortion in a band is noticeable then the number of quantisation levels needs to be increased by adjusting the scale factors. The scale-factors are applied to each scaleband to shape the quantisation noise so that it remains below the masking threshold. Initially the scale factor for each band is set to 1. If the quantisation noise in a band exceeds the masking threshold then the scale factor for that band is adjusted (increased) to reduce the quantisation step size and keep the noise below the masking threshold. However this adjustment and control of distortion noise can result in a higher bit rate than allowed since to reduce quantisation noise in each band the number of quantisation levels is increased. So the rate adjustment loop has to be repeated each time scale-factors are adjusted. Therefore the rate loop is nested in the distortion loop. The rate and distortion loops may fail to converge and may go on forever, however several conditions can be checked to stop the iterations in an early stage.

12.9 Stereo Music Coding

The relationship between the signals in the left and the right channels of a stereo music may be modelled as

$$x_L(m) = fn[x_R(m)] + x_{L\perp R}(m) \quad (12.57)$$

$fn[x_R(m)]$ is the part of the left channel $x_L(m)$ which is correlated with the right channel $x_R(m)$ and can be predicted from it, fn is a mapping function, and $x_{L\perp R}(m)$ is the part of the left channel $x_L(m)$ uncorrelated with $x_R(m)$, \perp denotes orthogonal. The objective of stereo coding is to exploit inter-channel correlation and redundancies so that the part of the music signal that is common to both channels is not coded twice.

Figure 12.34 shows an example of signals in the left and right channels of a section of stereo music. The spectrograms and the time waveforms show considerable similarity between the two signals. The normalised cross correlation of the left right channels, shown in figure 12.x.d, fluctuates with time. However, there are periods when the normalised cross-correlation approaches the maximum value of one indicating identical signals. In stereo music coding it is assumed that:

Music Signal Processing

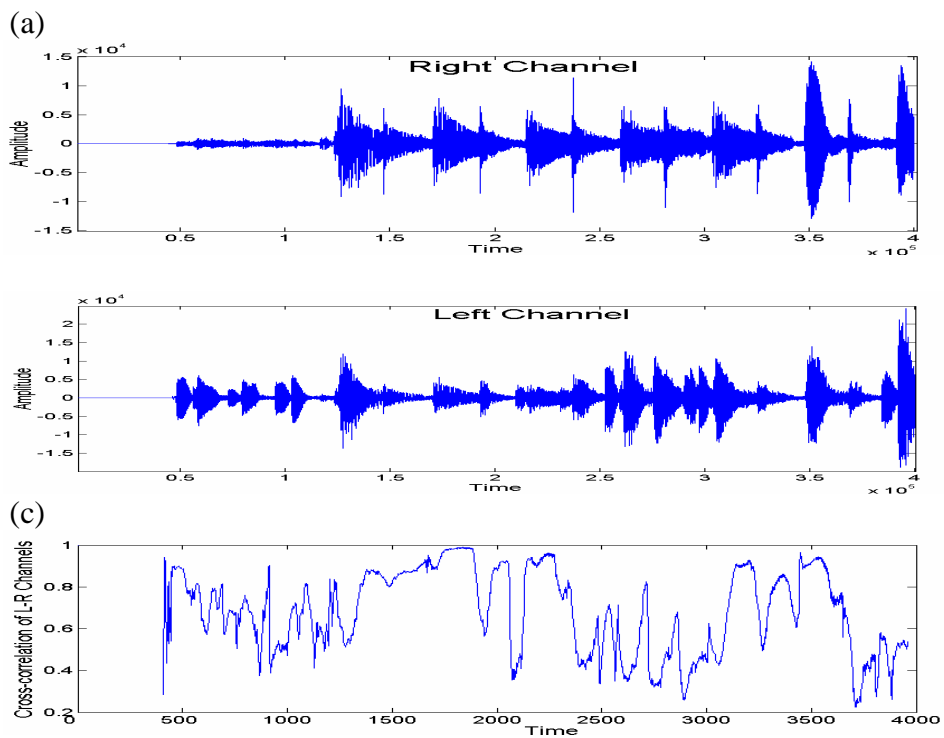
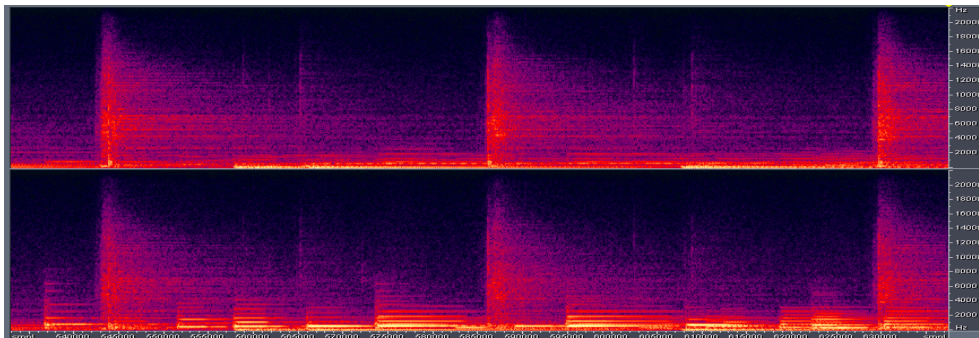
- The *masking threshold* of hearing is the same for both channels. Thus, the mean of the two channels is fed to the masking threshold estimator.
- The audio is centred, i.e. approximately equal in both channels.

In practice, joint stereo coding systems encode the sum ($L+R$) and the difference ($L-R$) of the left (L) and right (R) channels. From an analysis of the FFT spectra of $L+R$ and $L-R$ signals a combined set of thresholds of the just noticeable distortions (JND) levels is calculated for both $L+R$ and $L-R$ channels. The sum and difference signals are then coded in the same way as a mono signal using the common set of JNDs for both channels.

Music Signal Processing

More complex stereo coding models may use an inter-channel linear predictor model, or inter-channel orthogonalisation transforms.

12.10 Music Recognition and Transcription



(d) **Figure 12.34** - The spectrogram of 2 seconds of stereo Jazz, (b), (c) and right and left channels of a segment of music of duration 10 seconds and their cross correlation respectively.

Music Signal Processing

The main interest in music instrument recognition and music score transcription stems from the need to index and retrieve the audio content of multimedia files on the internet for multimedia search engines. Music transcription is the conversion of a digital music signal stream to the corresponding symbolic representation or music score. Music instrument identification and music content transcription are in principle similar to speaker identification and speech recognition.

The recognition of monophonic music is not a difficult task, however the problem is that music is often multi-instrument and polyphonic; it includes notes from several instruments played simultaneously plus vocal singing. Hence, in practice music recognition is a complex and difficult task to accomplish. Although there is a relatively simple relationship between different musical notes, the entropy or perplexity of music is far greater than speech, because music has a larger bandwidth, it is not constrained by a well-defined grammar, it is constrained mainly by aesthetics and perception, and most importantly a musical performance can contain any number and variety of instruments plus sound effects.

Ideally the transcription of music will involve the followings:

- (a) Determination of the number of musical instruments.
- (b) Identification of the instruments.
- (c) Modelling the characteristics of each instrument.
- (d) Identification of the timing, pitch and spectral and temporal parameters of musical sounds from individual instruments.

Despite many years of concentrated international research there are still significant unsolved problems in the development of reliable speech transcription systems. It is therefore reasonable to expect that the more complex problem of music transcription, which in many cases includes singing voice, is unlikely to be solved in the foreseeable future. However, as is the case with speech recognition systems, there can be useful limited-task music classification systems for a music search engine.

Music Signal Processing

where for music signals the features may include the values of spectral envelop, amplitude envelop, beat, pitch, jitter, shimmer, low to high frequency energy, brightness, noise, slope of attack, slope of sustain, slope of fall etc.

How useful a binary question is and what set of values of the thresholds should be chosen, can be answered through monitoring the 'quality' of split that the question yields. The quality of split depends on its ability to split and separate the music signals into different music instruments or different types of music or different musical notes. One good criterion for the binary split of a node into two clusters is the minimum average entropy defined as

$$\bar{H}(Y) = \sum_{i=1}^2 p(l_i) H_i(Y) \quad (12.58)$$

where $p(l_i)$ is the probability of node i and $H_i(Y)$, the entropy of node i is defined as

$$H_i(Y) = \sum_{j=1}^N p(c_j | l_i) \log_2(p(c_j | l_i)) \quad (12.59)$$

where c_j is a sample feature. Minimum entropy criterion, as the name implies, will yield clusters with minimum possible randomness.

Bibliography

A H Benade 1976 *Fundamentals of Musical Acoustics* (Oxford University Press) .

L Cremer 1984 *The Physics of the Violin* (MIT Press, Cambridge, Massachusetts) translation J S Allen .

R. Veldhuis, M. Breeuwer and R. van der Wall, "Subband coding of digital audio signals without loss of quality," *Proc. 1989 International Conference on Acoustics, Speech and Signal Processing*, Glasgow, pp. 2009-2012.

A. Sugiyama, F. Hazu, M. Iwaware and T. Nishitani, "Adaptive transform coding with an adaptive block size (ATCABS)," *Proc. 1990 International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, pp. 1093-1096.

Music Signal Processing

G. Davidson, L. Fielder and M. Antill, "High-quality audio transform coding at 128 kbits/s," *Proc. 1990 International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, pp. 1117-1120.

G. Davidson, L. Fielder and M. Antill, "Low-complexity transform coder for satellite link applications," Audio Engineering Society 89th Convention preprint 2966, Sept. 1990.

J. S. Tobias, Ed., *Foundations of Modern Auditory Theory, Vol. 1*, Academic Press, New York, 1970.