Mathematics and Music

Dave Benson

Department of Mathematics, University of Georgia, Athens GA 30602, USA

Home page: ftp://byrd.math.uga.edu/pub/html/index.html

E-mail address: djb@byrd.math.uga.edu

Date: January 4, 2002

To Christine Natasha

Ode to an Old Fiddle

From the Musical World of London (1834);[1]

THE POOR FIDDLER'S ODE TO HIS OLD FIDDLE

Torn
Worn
Oppressed I mourn
B a d
S a d
Three-quarters mad
Money gone
Credit none
Duns at door
Half a score
Wife in lain
Twins again
Others ailing
Nurse a railing
Billy hooping
Betsy crouping
Besides poor Joe
With fester'd toe.
Come, then, my Fiddle,
Come, my time-worn friend,
With gay and brilliant sounds
Some sweet tho' transient solace lend,
Thy polished neck in close embrace
I clasp, whilst joy illumines my face.
When o'er thy strings I draw my bow,
My drooping spirit pants to rise;
A lively strain I touch—and, lo!
I seem to mount above the skies.
There on Fancy's wing I soar
Heedless of the duns at door;
Oblivious all, I feel my woes no more;
But skip o'er the strings,
As my old Fiddle sings,
"Cheerily oh! merrily go!
"PRESTO! good master,
"You very well know
"I will find Music,
"If you will find bow,
"From E, up in alto, to G, down below."
Fatigued, I pause to change the time
For some *Adagio*, solemn and sublime.
With graceful action moves the sinuous arm;
My heart, responsive to the soothing charm,
Throbs equably; whilst every health-corroding care
Lies prostrate, vanquished by the soft mellifluous air.
More and more plaintive grown, my eyes with tears o'erflow,
And Resignation mild soon smooths my wrinkled brow.
Reedy Hautboy may squeak, wailing Flauto may squall,
The Serpent may grunt, and the Trombone may bawl;
But, by Poll,* my old Fiddle's the prince of them all.
Could e'en Dryden return, thy praise to rehearse,
His Ode to Cecilia would seem rugged verse.
Now to thy case, in flannel warm to lie,
Till call'd again to pipe thy master's eye.

*Apollo.

[1]Quoted in Nicolas Slonimsky's *Book of Musical Anecdotes*, reprinted by Schirmer, 1998.

Contents

Introduction

This volume is a much expanded and augmented version of the course notes for the undergraduate course "Mathematics and Music", given at the University of Georgia. The prerequisites for the first time I gave the course were differential and integral calculus, as well as an elementary knowledge of music notation. The second time, I also required either calculus of several variables or ordinary differential equations. Many parts of the notes require a little more mathematical background. Exactly what parts of these notes are taught in the course depend on who turns up, what they are interested in, and the extent of their mathematical background.

These notes are not really designed for sequential reading from end to end, although it is possible to read them this way. In particular, the level of mathematical and musical sophistication required of the reader varies dramatically from section to section. I have tried to write in such a way that sections which use mathematics beyond the reach of the reader can be skipped without compromising possibly more elementary later sections.

It is quite possible, and the reader is encouraged to do so, to read sections of interest as though they were independent of each other, and then refer back, and to the index to fill in the gaps. The index has been made unusually comprehensive with this in mind.

I particularly encourage the reader to skip some of the later parts of the first chapter, because to be honest, the later chapters are more interesting.

> *Why do we consider sine waves to represent "pure" notes, and all other tones to be made up out of sine waves?*

The answer to this question begins with a discussion of the human ear. This will lead to a discussion of harmonic motion, and explain the relevance of sine waves. This discussion gives rise to another more challenging question.

> *How is it that a string under tension can vibrate with a number of different frequencies at the same time?*

This question will focus our attention on the analysis of musical notes into sine waves, or Fourier analysis. Our discussion of Fourier series does not include proofs, but it does include rigorous statements of the relevant theorems. We discuss the distinction between convergence and uniform convergence, through the example of the Gibbs phenomenon. Fourier analysis, together with d'Alembert's remarkably elegant solution to the wave equation in one dimension, answers the question of the vibrating string.

> *Why does the vibration of a drum result in a spectrum which does not consist of multiples of a fundamental frequency in the way that the spectrum of a vibrating string does?*

We discuss Bessel functions as an example of Fourier analysis, for two purposes. First, the Bessel functions describe the motion of a vibrating circular

membrane such as a drum, and are essential to understanding why the sound of a drum results in an inharmonic spectrum. The other purpose is as preparation for the discussion of frequency modulation synthesis in a later chapter.

> *Why does the modern western scale consist of twelve equally spaced notes to an octave?*

Spectrum and Fourier analysis will form the point of departure for our discussion of the development of scales. The emphasis is on the relation between the arithmetic properties of rational approximations to irrational numbers, and musical intervals. We concentrate on the development of the standard Western scales, from the Pythagorean scale and just intonation, through the meantone scale, to the irregular temperaments of Werckmeister and others, and finally the equal tempered scale. We also discuss a number of other scales. These include scales not based on the octave, such as the Bohlen–Pierce scale based on odd harmonics only, and the alpha, beta and gamma scales of Wendy Carlos.

> *How can a bunch of zeros and ones on a computer represent music? How does this affect the way we understand and manipulate music?*

In the final chapter, we discuss digital signal processing, synthesis and computer music. The emphasis is on how the ideas involved in synthesis and signal processing reflect back into an understanding of structural elements of sound. Interesting sounds do not have a static frequency spectrum, and the goal is to understand the evolution of spectrum with time. We discuss the relevance of Bessel functions to FM synthesis, as well as Nyquist's theorem on aliasing above half the sample rate, the MIDI protocol, internet resources, and so on.

> *Why do rhythms and melodies, which are composed of sound, resemble the feelings, while this is not the case for tastes, colors or smells? Can it be because they are motions, as actions are also motions? Energy itself belongs to feeling and creates feeling. But tastes and colors do not act in the same way.*
>
> Aristotle, *Prob.* xix. 29

Deryck Cooke, who reconstructed from Mahler's sketches a performing edition of his tenth symphony, has written a wonderful book [**16**] in which he describes the musical vocabulary and how it conspires to transmit mood. That is not the subject of this text, as mathematics has little to say about the correspondence between mood and musical form. We are interested instead in the mathematical theory behind music. Nonetheless, we can never lose sight of the evocative power of music if we are to reach any understanding of the context for the theory.

Books

I have included an extensive annotated bibliography, and have also indicated which books are still in print. This information may be slightly out of date by the time you read this.

There are a number of good books on the physics and engineering aspects of music. Dover has kept some of the older ones in print, so they are available at relatively low cost. Among them are Backus [**2**], Benade [**6**], Berg and Stork [**7**], Campbell and Greated [**11**], Fletcher and Rossing [**30**], Hall [**39**], Helmholtz [**43**], Jeans [**49**], Johnston [**51**], Morgan [**73**], Nederveen [**76**], Olson [**78**], Pierce [**84**], Rigden [**91**], Roederer [**96**], Rossing [**97**], Rayleigh [**89**], Taylor [**108**].

Books on psychoacoustics include Buser and Imbert [**10**], Cook (Ed.) [**15**], Deutsch (Ed.) [**24**], Helmholtz [**43**], Howard and Angus [**46**], Moore [**71**], Sethares [**105**], Von Békésy [**5**], Winckel [**113**], Yost [**116**], and Zwicker and Fastl [**117**]. A decent book on physiological aspects of the ear and hearing is Pickles [**83**].

Books including a discussion of the development of scales and temperaments include Asselin [**1**], Barbour [**4**], Blackwood [**8**], Daniélou [**21, 22**], Deva [**25**], Devie [**26**], Helmholtz [**43**], Hewitt [**44**], Isacoff [**48**], Jorgensen [**52**], Lattard [**56**], Lindley and Turner-Smith [**62**], Lloyd and Boyle [**63**], Mathieu [**67**], Moore [**72**], Neuwirth [**77**], Padgham [**79**], Partch [**80**], Pfrogner [**82**], Rameau [**88**], Ruland [**101**], Vogel [**110**], Wilkinson [**112**] and Yasser [**115**]. Among these, I particularly recommend the books of Barbour and Helmholtz. The Bohlen–Pierce scale is described in Chapter 13 of Mathews and Pierce [**66**].

There are a number of good books about computer synthesis of musical sounds. See for example Dodge and Jerse [**27**], Moore [**72**], and Roads [**93**]. For FM synthesis, see also Chowning and Bristow [**13**]. For computers and music (which to a large extent still means synthesis), there are a number of volumes consisting of reprinted articles from the Computer Music Journal (M.I.T. Press). Among these are Roads [**92**], and Roads and Strawn [**95**]. Other books on electronic music and the role of computers in music include Cope [**17, 18**], Mathews and Pierce [**66**], Moore [**72**] and Roads [**93**]. Some books about MIDI (Musical Instrument Digital Interface) are Rothstein [**100**], and de Furia and Scacciaferro [**32**]. A standard work on digital audio is Pohlmann [**85**].

Books on random music and fractal music include Xenakis [**114**], Johnson [**50**] and Madden [**64**].

Popular magazines about electronic and computer music include "Keyboard" and "Electronic Musician" which are readily available at magazine stands.

Acknowledgements

I would like to thank Manuel Op de Coul for reading an early draft of these notes, making some very helpful comments on Chapters 5 and 6, and making me aware of some fascinating articles and recordings (see Appendix R). Thanks to Paul Erlich and Herman Jaramillo for emailing me various corrections and other helpful comments. Thanks to Robert Rich for responding to my request for information about the scales he uses in his recordings (see §6.1 and Appendix R). Thanks to Heinz Bohlen for taking an interest in these notes and for numerous email discussions regarding the Bohlen–Pierce scale §6.7. Thanks to my students, who patiently listened to my attempts at explanation of this material, and who helped me to clean up the text by understanding and pointing out improvements, where it was comprehensible, and by not understanding where it was incomprehensible.

This document was typeset with AMSLaTeX. The musical examples were typeset using MusicTeX, the graphs were made as encapsulated postscript (eps) files using MetaPost, and these and other pictures were included in the text using the graphicx package.

Essays

During the term, I shall expect each student to write one essay, on a topic to be chosen by the student and approved by me. This will be collected during or before the tenth week of the semester. For undergraduates, the essay will consist of between 5 and 20 typed pages. For graduate students, I shall expect between 10 and 40 pages. Some examples of topics which the student may like to consider are as follows. This should be regarded as an indication of what sort of topics are likely to be regarded as acceptable.

These essays will be graded for grammar, style and use of English, as well as mathematical content. If you use a mathematical formula, make sure it is part of a complete sentence. In general, as a matter of style, a sentence should not begin with a mathematical symbol. Comprehensibility is a key issue too. To ascertain whether what you have written makes sense, I would recommend asking a friend to read through what you have written. If your friend asks you what something means, that's probably an indication that you should include more explanation.

Psychoacoustics

The Ear and Cochlear Mechanics

Concert hall acoustics

Sound compression

CSound

MIDI

Digital synthesis algorithms

History of scales from some particular culture
(e.g., Indian, Greek, Arabic, Chinese, Balinese, etc.)

Bessel functions

Combinatorics of twelve tone music

Relation of spectrum to scale

The Fourier transform

Wavelets (requires a strong mathematics background)

Formants and the human voice

Cross interleaved Reed–Solomon codes and the compact disc

The physics of some class of musical instruments
(stringed, woodwind, brass, percussive, etc.)

The phase vocoder

Get hold of a technical article from the *Computer Music Journal*, *Acustica*, or the *Journal of the Acoustical Society of America* and explain it assuming only a mathematical background at the level of this course.

CHAPTER 1

Waves and harmonics

1.1. What is sound?

The medium for the transmission of music is sound. A proper understanding of music entails at least an elementary understanding of the nature of sound and how we perceive it.

Sound consists of vibrations of the air. To understand sound properly, we must first have a good mental picture of what air looks like. Air is a gas, which means that the atoms and molecules of the air are not in such close proximity to each other as they are in a solid or a liquid. So why don't air molecules just fall down on the ground? After all, Galileo's principle states that objects should fall to the ground with equal acceleration independently of their size and mass.

The answer lies in the extremely rapid motion of these atoms and molecules. The mean velocity of air molecules at room temperature under normal conditions is around 450–500 meters per second (or somewhat over 1000 miles per hour), which is considerably faster than an express train at full speed. We don't feel the collisions with our skin, only because each air molecule is extremely light, but the combined effect on our skin is the air pressure which prevents us from exploding!

The mean free path of an air molecule is 6×10^{-8} meters. This means that on average, an air molecule travels this distance before colliding with another air molecule. The collisions between air molecules are perfectly elastic, so this does not slow them down.

We can now calculate how often a given air molecule is colliding. The collision frequency is given by

$$\text{collision frequency} = \frac{\text{mean velocity}}{\text{mean free path}} \sim 10^{10} \text{ collisions per second.}$$

So now we have a very good mental picture of why the air molecules don't fall down. They don't get very far down before being bounced back up again. The effect of gravity is then observable just as a gradation of air pressure, so that if we go up to a high elevation, the air pressure is noticeably lower.

So air consists of a large number of molecules in close proximity, continually bouncing off each other to produce what is perceived as air pressure. When an object vibrates, it causes waves of increased and decreased pressure. These waves are perceived by the ear as sound, in a manner to be investigated in the next section, but first we examine the nature of the waves themselves.

Sound travels through the air at about 340 meters per second (or 760 miles per hour). This does not mean that any particular molecule of air is moving in the direction of the wave at this speed (see above), but rather that the local disturbance to the pressure propagates at this speed. This is similar to what is happening on the surface of the sea when a wave moves through it; no particular piece of water moves along with the wave, it is just that the disturbance in the surface is propagating.

There is one big difference between sound waves and water waves, though. In the case of the water waves, the local movements involved in the wave are up and down, which is at right angles to the direction of propagation of the wave. Such waves are called *transverse waves*. Electromagnetic waves are also transverse. In the case of sound, on the other hand, the motions involved in the wave are in the same direction as the propagation. Waves with this property are called *longitudinal waves*.

Longitudinal waves

$\longrightarrow$ Direction of motion

Sound waves have four main attributes which affect the way they are perceived. The first is *amplitude*, which means the size of the vibration, and is perceived as loudness. The amplitude of a typical everyday sound is very minute in terms of physical displacement, usually only a small fraction of a millimeter. The second attribute is *pitch*, which should at first be thought of as corresponding to frequency of vibration. The third is *timbre*, which corresponds to the shape of the frequency spectrum of the sound. The fourth is *duration*, which means the length of time for which the note sounds.

These notions need to be modified for a number of reasons. The first is that most vibrations do not consist of a single frequency, and naming a "defining" frequency can be difficult. The second related issue is that these attributes should really be defined in terms of the perception of the sound, and not in terms of the sound itself. So for example the perceived pitch of a sound can represent a frequency not actually present in the waveform. This phenomenon is called the "missing fundamental", and is part of a subject called psychoacoustics.

Attributes of sound

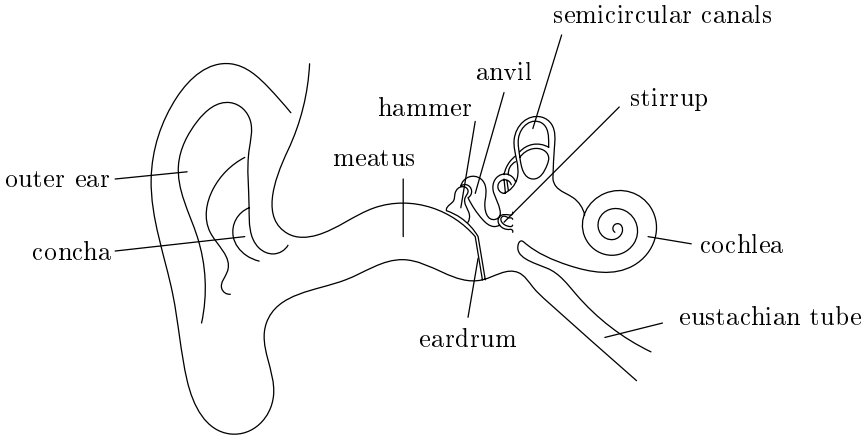
Physical	Perceptual
Amplitude	Loudness
Frequency	Pitch
Spectrum	Timbre
Duration	Length

In order to get much further with understanding sound, we need to study its perception by the human ear. This is the topic of the next section.

1.2. The human ear

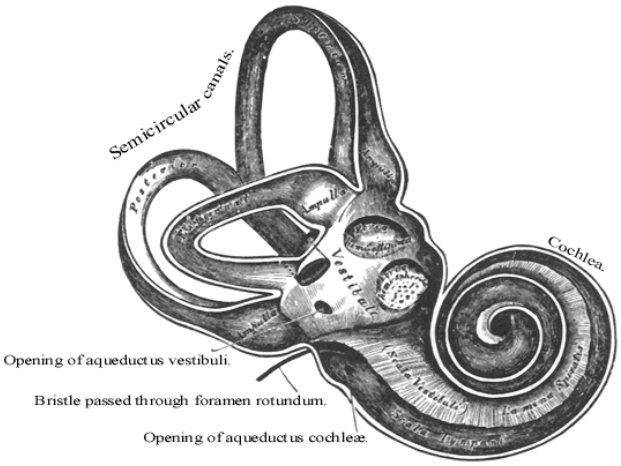
In order to understand the origins of the mathematical construction of scales, we must begin by understanding the physiological structure of the human ear. I have borrowed extensively from Gray's Anatomy for this description.

The ear is divided into three parts, called the outer ear, the middle ear or *tympanum* and the inner ear or *labyrinth*. The outer ear is the visible part on the outside of the head, called the *pinna* (plural *pinnæ*) or *auricle*, and is ovoid in form. The hollow middle part, or *concha* is associated with focusing and thereby magnifying the sound, while the outer rim, or *helix* appears to be associated with vertical spatial separation, so that we can judge the height of a source of sound.



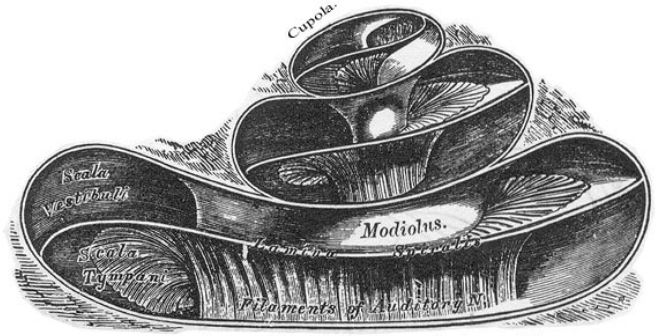
The concha channels the sound into the auditory canal, called the *meatus auditorius externus* (or just *meatus*). This is an air filled tube, about 2.7 cm long and 0.7 cm in diameter. At the inner end of the meatus is the ear drum, or *tympanic membrane*.

The ear drum divides the outer ear from the middle ear, or *tympanum*, which is also filled with air. The tympanum is connected to three very small bones (the *ossicular chain*) which transmit the movement of the ear drum to the inner ear. The three bones are the hammer, or *malleus*, the anvil, or *incus*, and the stirrup, or *stapes*. These three bones form a system of levers connecting the ear drum to a membrane covering a small opening in the inner ear. The membrane is called the *oval window*.



The osseous labyrinth laid open. (Enlarged.)

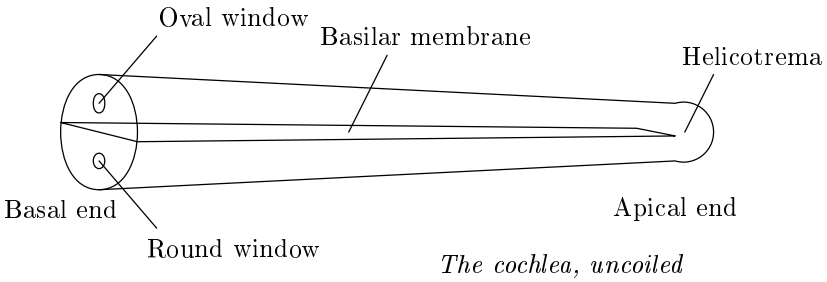
The inner ear, or *labyrinth*, consists of two parts, the *osseous labyrinth*,[1] consisting of cavities hollowed out from the substance of the bone, and the *membranous labyrinth*, contained in it. The osseous labyrinth is filled with various fluids, and has three parts, the *vestibule*, the *semicircular canals* and the *cochlea*. The vestibule is the central cavity which connects the other two parts and which is situated on the inner side of the tympanum. The semicircular canals lie above and behind the vestibule, and play a role in our sense of balance. The cochlea is at the front end of the vestibule, and resembles a common snail shell in shape. The purpose of the cochlea is to separate out sound into various components before passing it onto the nerve pathways. It is the functioning of the cochlea which is of most interest in terms of the harmonic content of a single musical note, so let us look at the cochlea in more detail.



The cochlea laid open. (Enlarged.)

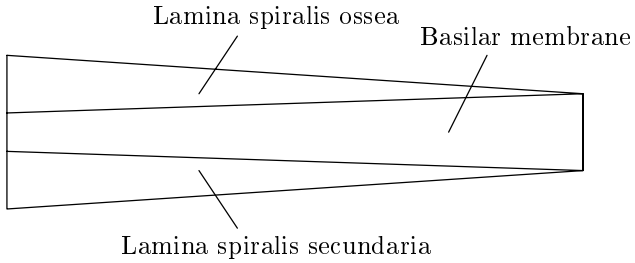
The cochlea twists roughly two and three quarter times from the outside to the inside, around a central axis called the *modiolus* or *columnella*. If it could be unrolled, it would form a tapering conical tube roughly 30 mm (a little over an inch) in length.

[1](Illustrations taken from the 1901 edition of *Anatomy, Descriptive and Surgical*, Henry Gray, F.R.S.)



The cochlea, uncoiled

At the wide (*basal*) end where it meets the rest of the inner ear it is about 9 mm (somewhat under half an inch) in diameter, and at the narrow (*apical*) end it is about 3 mm (about a fifth of an inch) in diameter. There is a bony shelf or ledge called the *lamina spiralis ossea* projecting from the modiolus, which follows the windings to encompass the length of the cochlea. A second bony shelf called the *lamina spiralis secundaria* projects inwards from the outer wall. Attached to these shelves is a membrane called the *membrana basilaris* or *basilar membrane*. This tapers in the opposite direction than the cochlea, and the bony shelves take up the remaining space.



The basilar membrane divides the interior of the cochlea into two parts with approximately semicircular cross-section. The upper part is called the *scala vestibuli* and the lower is called the *scala tympani*. There is a small opening called the *helicotrema* at the apical end of the basilar membrane, which enables the two parts to communicate with each other. At the basal end there are two windows allowing communication of the two parts with the vestibule. Each window is covered with a thin flexible membrane. The stirrup is connected to the membrane called the *membrana tympani secundaria* covering the upper window; this window is called the *fenestra rotunda* or *oval window*, and has an area of 2.0–3.7 mm^2. The lower window is called the *round window*, with an area of around 2 mm^2, and the membrane covering it is not connected to anything apart from the window. There are small hair cells along the basilar membrane which are connected with numerous nerve endings for the auditory nerves. These transmit information to the brain via a complex system of neural pathways.

Now consider what happens when a sound wave reaches the ear. The sound wave is focused into the meatus, where it vibrates the ear drum. This causes the hammer, anvil and stirrup to move as a system of levers, and so the stirrup alternately pushes and pulls the membrana tympani secundaria

in rapid succession. This causes fluid waves to flow back and forth round the length of the cochlea, in opposite directions in the scala vestibuli and the scala tympani, and causes the basilar membrane to move up and down.

Let us examine what happens when a pure sine wave is transmitted by the stirrup to the fluid inside the cochlea. The speed of the wave of fluid in the cochlea at any particular point depends not only on the frequency of the vibration but also on the area of cross-section of the cochlea at that point, as well as the stiffness and density of the basilar membrane. For a given frequency, the speed of travel decreases towards the apical end, and falls to almost zero at the point where the narrowness causes a wave of that frequency to be too hard to maintain. Just to the wide side of that point, the basilar membrane will have to have a peak of amplitude of vibration in order to absorb the motion. Exactly where that peak occurs depends on the frequency. So by examining which hairs are sending the neural signals to the brain, we can ascertain the frequency of the incoming sine wave. This description of how the brain "knows" the frequency of an incoming sine wave is due to Hermann Helmholtz, and is known as the place theory of pitch perception.

The phenomenon of *masking* is easily explained in terms of Helmholtz's theory. Alfred Meyer (1876) discovered that an intense sound of a lower pitch prevents us from perceiving a weaker sound of a higher pitch, but an intense sound of a higher pitch never prevents us from perceiving a weaker sound of a lower pitch. The explanation of this is that the excitation of the basilar membrane caused by a sound of higher pitch is closer to the basal end of the cochlea than that caused by a sound of lower pitch. So to reach the place of resonance, the lower pitched sound must pass the places of resonance for all higher frequency sounds. The movement of the basilar membrane caused by this interferes with the perception of the higher frequencies.

The extent to which the ear can discriminate between frequencies very close to each other is not completely explained by the mechanics of the cochlea alone. It appears that a sort of psychophysical feedback mechanism sharpens the tuning and increases the sensitivity. In other words, there is information carried both ways by the neural paths between the cochlea and the brain, and this provides active amplification of the incoming acoustic stimulus. If the incoming signal is loud, the gain will be turned down to compensate. If there is very little stimulus, the gain is turned up until the stimulus is detected. An annoying side effect of this is that if mechanical damage to the ear causes deafness, then the neural feedback mechanism turns up the gain until random noise is amplified, so that singing in the ear, or *tinnitus* results. The deaf person does not even have the consolation of silence.

Further reading:

Moore, *Psychology of hearing* [**71**].

Pickles, *An introduction to the physiology of hearing* [**83**].

Yost, *Fundamentals of hearing. An introduction* [**116**].

Zwicker and Fastl, *Psychoacoustics: facts and models* [**117**].

1.3. Limitations of the ear

In music, frequencies are measured in Hertz (Hz), or cycles per second. The approximate range of frequencies to which the human ear responds is usually taken to be from 20 Hz to 20,000 Hz. For frequencies outside this range, there is no resonance in the basilar membrane, although sound waves of frequency lower than 20 Hz may often be *felt* rather than heard.[2] For comparison, here is a table of hearing ranges for various animals.[3]

Species	Range (Hz)
Turtle	20–1,000
Goldfish	100–2,000
Frog	100–3,000
Pigeon	200–10,000
Sparrow	250–12,000
Human	20–20,000
Chimpanzee	100-20,000
Rabbit	300–45,000
Dog	50–46,000
Cat	30–50,000
Guinea pig	150–50,000
Rat	1,000–60,000
Mouse	1,000–100,000
Bat	3,000–120,000
Dolphin (*Tursiops*)	1,000–130,000

Sound intensity is measured in *decibels* or dB. Zero decibels represents a power intensity of 10^{-12} watts per square meter, which is somewhere in the region of the weakest sound we can hear. Adding ten decibels (one *bel*) multiplies the power intensity by a factor of ten. So multiplying the power by a factor of b adds $10\log_{10}(b)$ decibels to the level of the signal. This means that the scale is logarithmic, and n decibels represents a power density of $10^{(n/10)-12}$ watts per square meter.

Often, decibels are used as a relative measure, so that an intensity *ratio* of ten to one represents an increase of ten decibels. As a relative

[2]But see also: Tsutomi Oohashi, Emi Nishina, Norie Kawai, Yoshitaka Fuwamoto and Hiroshi Imai, *High-frequency sound above the audible range affects brain electric activity and sound perception*, Audio Engineering Society preprint No. 3207 (91st convention, New York City). In this fascinating paper, the authors describe how they recorded gamelan music with a bandwidth going up to 60 KHz. They played back the recording through a speaker system with an extra tweeter for the frequencies above 26 KHz, driven by a separate amplifier so that it could be switched on and off. They found that the EEG (Electroencephalogram) of the listeners' response, as well as the subjective rating of the recording, was affected by whether the extra tweeter was on or off, even though the listeners denied that the sound was altered by the presence of this tweeter, or that they could hear anything from the tweeter played alone. They also found that the EEG changes persisted afterwards, in the absence of the high frequency stimulation, so that long intervals were needed between sessions.

Another relevant paper is: Martin L. Lenhardt, Ruth Skellett, Peter Wang and Alex M. Clarke, *Human ultrasonic speech perception*, Science, Vol. 253, 5 July 1991, 82-85. In this paper, they report that bone-conducted ultrasonic hearing has been found capable of supporting frequency discrimination and speech detection in normal, older hearing-impaired, and profoundly deaf human subjects. They conjecture that the mechanism may have to do with the *saccule*, which is a small spherical cavity adjoining the scala vestibuli of the cochlea.

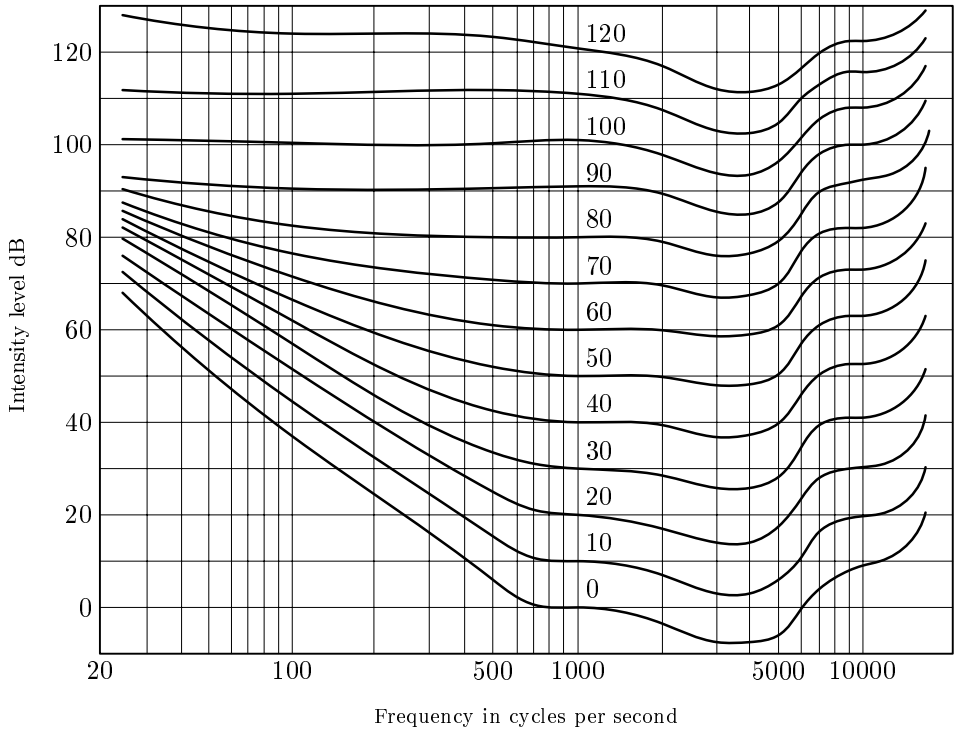
[3]Taken from R. Fay, *Hearing in Vertibrates. A Psychophysics Databook.* Hill-Fay Associates, Winnetka, Illinois, 1988.

measure, decibels refer to ratios of powers whether or not they directly represent sound. So for example, the power gain and the signal to noise ratio of an amplifier are measured in decibels. It is worth knowing that $\log_{10}(2)$ is roughly 0.3 (to five decimal places it is 0.30103), so that a power ratio of 2:1 represents a difference of about 3 dB. To distinguish from the relative measurement, the notation dB SPL (Sound Pressure Level) is sometimes used to refer to the absolute measurement of sound described above. It should also be mentioned that rather than using dB SPA, use is often made of a weighting curve, so that not all frequencies are given equal importance. There are three standard curves, called A, B and C. It is most common to use curve A, which has a peak at about 2000 Hz and drops off substantially to either side. Curves B and C are flatter, and only drop off at the extremes. Measurements made using curve A are quoted as dBA, or dBA SPA to be pedantic.

The *threshold of hearing* is the level of the weakest sound we can hear. Its value in decibels varies from one part of the frequency spectrum to another. Our ears are most sensitive to frequencies a little above 2000 Hz, where the threshold of hearing of the average person is a little above 0 dB. At 100 Hz the threshold is about 50 dB, and at 10,000 Hz it is about 30 dB. The average whisper is about 15–20 dB, conversation usually happens at around 60–70 dB, and the threshold of pain is around 130 dB.

The relationship between sound pressure level and perception of loudness is frequency dependent. The following graph, due to Fletcher and Munson[4] shows equal loudness curves for pure tones at various frequencies.

[4]H. Fletcher and W. J. Munson, *Loudness, its definition, measurement and calculation*, J. Acoust. Soc. Am. 5 (1933), 82–108.



The unit of loudness is the *phon*, which is defined as follows. The listener adjusts the level of the signal until it is judged to be of equal intensity to a standard 1000 Hz signal. The phon level is defined to be the signal pressure level of the 1000 Hz signal of the same loudness. The curves in this graph are called *Fletcher–Munson curves*, or *isophons*.

The amount of power in watts involved in the production of sound is very small. The clarinet at its loudest produces about one twentieth of a watt of sound, while the trombone is capable of producing up to five or six watts of sound. The average human speaking voice produces about 0.00002 watts, while a bass singer at his loudest produces about a thirtieth of a watt.

The *just noticeable difference* or *limen* is used both for sound intensity and frequency. This is usually taken to be the smallest difference between two successive tones for which a person can name correctly 75% of the time which is higher (or louder). It depends in both cases on both frequency and intensity. The just noticeable difference in frequency will be of more concern to us than the one for intensity, and the following table is taken from Pierce [**84**]. The measurements are in cents, where 1200 cents make one octave (for further details of the system of cents, see §5.4).

Frequency	Intensity (dB)										
(Hz)	5	10	15	20	30	40	50	60	70	80	90
31	220	150	120	97	76	70					
62	120	120	94	85	80	74	61	60			
125	100	73	57	52	46	43	48	47			
250	61	37	27	22	19	18	17	17	17	17	
550	28	19	14	12	10	9	7	6	7		
1,000	16	11	8	7	6	6	6	6	5	5	4
2,000	14	6	5	4	3	3	3	3	3	3	
4,000	10	8	7	5	5	4	4	4	4		
8,000	11	9	8	7	6	5	4	4			
11,700	12	10	7	6	6	6	5				

It is easy to see from this table that our ears are much more sensitive to small changes in frequency for higher notes than for lower ones. When referring to the above table, bear in mind that it refers to *consecutive* notes, not simultaneous ones. For simultaneous notes, the corresponding term is the *limit of discrimination*. This is the smallest difference in frequency between simultaneous notes, for which two separate pitches are heard. We shall see in §1.7 that simultaneous notes cause beats, which enable us to notice far smaller differences in frequency. This is very important to the theory of scales, because notes in a scale are designed for harmony, which is concerned with clusters of simultaneous notes. So scales are much more sensitive to very small changes in tuning than might be supposed.

Vos[5] studied the sensitivity of the ear to the exact tuning of the notes of the usual twelve tone scale, using two-voice settings from Michael Praetorius' *Musæ Sioniæ*, Part VI (1609). His conclusions were that scales in which the intervals were not more than 5 cents away from the "just" versions of the intervals (see §5.5) were all close to equally acceptable, but then with increasing difference the acceptability decreases dramatically. In view of the fact that in the modern equal tempered twelve tone system, the major third is about 14 cents away from just, these conclusions are very interesting. We shall have much more to say about this subject in Chapter 5.

Exercises

1. Power intensity is proportional to the square of amplitude. How many decibels represent a doubling of the amplitude of a signal?

1.4. Why sine waves?

What is the relevance of sine waves to the discussion of perception of pitch? Could we make the same discussion using some other family of periodic waves, that go up and down in a similar way?

[5]J. Vos, *Subjective acceptability of various regular twelve-tone tuning systems in two-part musical fragments*, J. Acoust. Soc. Am. **83** (1988), 2383–2392.

The answer lies in the differential equation for simple harmonic motion, which we discuss in the next section. To put it briefly, the solutions to the differential equation

$$\frac{d^2y}{dt^2} = -\kappa y$$

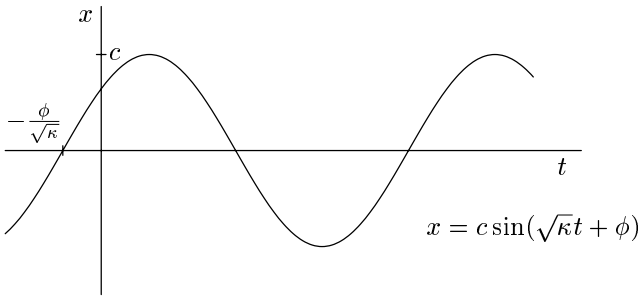
are the functions

$$y = A\cos\sqrt{\kappa}t + B\sin\sqrt{\kappa}t,$$

or equivalently

$$y = c\sin(\sqrt{\kappa}t + \phi)$$

(see §1.7 for the equivalence of these two forms of the solution).



The above differential equation represents what happens when an object is subject to a force towards an equilibrium position, the magnitude of the force being proportional to the distance from equilibrium.

In the case of the human ear, the above differential equation may be taken as a close approximation to the equation of motion of a particular point on the basilar membrane, or anywhere else along the chain of transmission between the outside air and the cochlea. Actually, this is inaccurate in several regards. The first is that we should really set up a second order partial differential equation describing the motion of the surface of the basilar membrane. This does not really affect the results of the analysis much except to explain the origins of the constant κ. The second inaccuracy is that we should really think of the motion as *forced damped harmonic motion* in which there is a damping term proportional to velocity, coming from the viscosity of the fluid and the fact that the basilar membrane is not perfectly elastic. In §§1.9–1.10, we shall see that forced damped harmonic motion is also sinusoidal, but contains a rapidly decaying transient component. There is a *resonant frequency* corresponding to the maximal response of the damped system to the incoming sine wave. The third inaccuracy is that for loud enough sounds the restoring force may be nonlinear. This will be seen to be the possible origin of some interesting acoustical phenomena. Finally, most musical notes do not consist of a single sine wave. For example, if a string is plucked, a periodic wave will result, but it will usually consist of a sum of sine waves with various amplitudes. So there will be various different peaks of amplitude of vibration of the basilar membrane, and a more complex signal is sent

to the brain. The decomposition of a periodic wave as a sum of sine waves is called Fourier analysis, which is the subject of Chapter 2.

1.5. Harmonic motion

Consider a particle of mass m subject to a force F towards the equilibrium position, $y = 0$, and whose magnitude is proportional to the distance y from the equilibrium position,

$$F = -ky.$$

Here, k is just the constant of proportionality. Newton's laws of motion give us the equation

$$F = ma$$

where

$$a = \frac{d^2y}{dt^2}$$

is the acceleration of the particle and t represents time. Combining these equations, we obtain the second order differential equation

$$\frac{d^2y}{dt^2} + \frac{ky}{m} = 0. \tag{1.5.1}$$

We write $\dot{y}$ for $\dfrac{dy}{dt}$ and $\ddot{y}$ for $\dfrac{d^2y}{dt^2}$ as usual, so that this equation takes the form

$$\ddot{y} + ky/m = 0.$$

The solutions to this equation are the functions

$$y = A\cos(\sqrt{k/m}\,t) + B\sin(\sqrt{k/m}\,t). \tag{1.5.2}$$

The fact that these are the solutions of this differential equation is the explanation of why the sine wave, and not some other periodically oscillating wave, is the basis for harmonic analysis of periodic waves. For this is the differential equation governing the movement of any particular point on the basilar membrane in the cochlea, and hence governing the human perception of sound.

Exercises

1. Show that the functions (1.5.2) satisfy the differential equation (1.5.1).

2. Show that the general solution (1.5.2) to equation (1.5.1) can also be written in the form

$$y = c\sin(\sqrt{k/m}\,t + \phi).$$

Describe c and ϕ in terms of A and B. (If you get stuck, take a look at §1.7).

1.6. Vibrating strings

Consider a vibrating string, anchored at both ends. Suppose at first that the string has a heavy bead attached to the middle of it, so that the mass m of the bead is much greater than the mass of the string. Then the string exerts a force F on the bead towards the equilibrium position, and whose magnitude, at least for small displacements, is proportional to the distance y from the equilibrium position,

$$F = -ky.$$

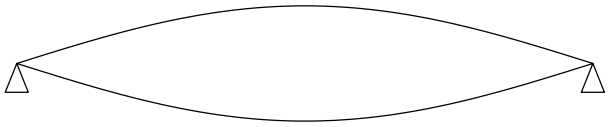
According to the last section, we obtain the differential equation

$$\frac{d^2y}{dt^2} + \frac{ky}{m} = 0.$$

whose solutions are the functions

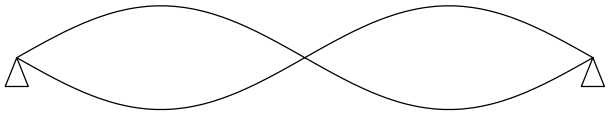
$$y = A\cos(\sqrt{k/m}\,t) + B\sin(\sqrt{k/m}\,t),$$

where the constants A and B are determined by the initial position and velocity of the string.



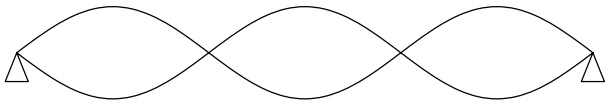
If the mass of the string is uniformly distributed, then more vibrational "modes" are possible. For example, the midpoint of the string can remain stationary while the two halves vibrate with opposite phases. On a guitar, this can be achieved by touching the midpoint of the string while plucking and then immediately releasing. The effect will be a sound exactly an octave above the natural pitch of the string, or exactly twice the frequency. The use of harmonics in this way is a common device among guitar players. If each half is vibrating with a pure sine wave then the motion of a point other than the midpoint will be described by the function

$$y = A\cos(2\sqrt{k/m}\,t) + B\sin(2\sqrt{k/m}\,t).$$



If a point exactly one third of the length of the string from one end is touched while plucking, the effect will be a sound an octave and a perfect fifth above the natural pitch of the string, or exactly three times the frequency. Again, if the three parts of the string are vibrating with a pure sine wave, with the middle third in the opposite phase to the outside two thirds, then the motion of a non-stationary point on the string will be described by the function

$$y = A\cos(3\sqrt{k/m}\,t) + B\sin(3\sqrt{k/m}\,t).$$

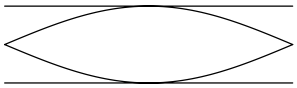


In general, a plucked string will vibrate with a mixture of all the modes described by multiples of the natural frequency, with various amplitudes. The amplitudes involved depend on the exact manner in which the string is plucked or struck. For example, a string struck by a hammer, as happens in a piano, will have a different set of amplitudes than that of a plucked string. The general equation of motion of a typical point on the string will be

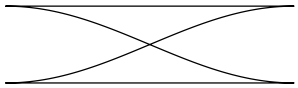
$$y = \sum_{n=1}^{\infty} \left(A_n \cos(n\sqrt{k/m}\, t) + B_n \sin(n\sqrt{k/m}\, t) \right).$$

This leaves us with a problem, to which we shall return in the next chapter. How can a string vibrate with a number of different frequencies at the same time? This forms the subject of the theory of Fourier series and the wave equation. Before we are in a position to study Fourier series, we need to understand sine waves and how they interact. This is the subject of the next section. We shall return to the subject of vibrating strings in §3.1, where we shall develop the wave equation and its solutions.

Wind instruments behave similarly, but the discussion needs to be divided into two cases, namely open tubes and closed tubes. Also, we need to be careful to distinguish between what is happening to the air pressure and what is happening to the air displacement, because they will have different phases. For a simple open tube, the basic mode of vibration can be represented by the following diagram.

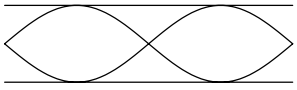


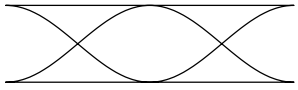
pressure



displacement

Bear in mind that the vertical axis in this diagram actually represents *horizontal* displacement or pressure, and not vertical, because of the longitudinal nature of air waves. Furthermore, the two parts of the graphs only represent the two extremes of the motion. In these diagrams, the nodes of the pressure diagram correspond to the antinodes of the displacement diagram and vice versa. The second and third vibrational modes will be represented by the following diagrams.





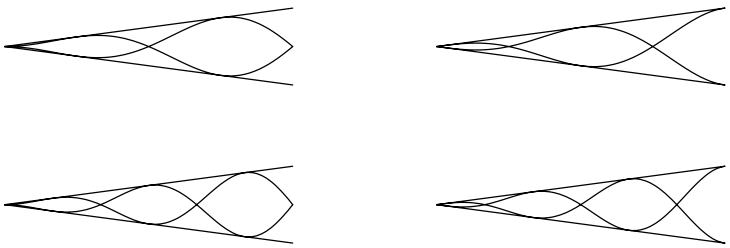
Tubes or pipes which are closed at one end behave differently, because the displacement is forced to be zero at the closed end. So the first two modes are as follows. In these diagrams, the left end of the tube is closed.

pressure displacement

It follows that for closed tubes, odd multiples of the fundamental frequency dominate. For example, the flute is an open tube, so all multiples of the fundamental are present. The clarinet is a closed tube, so odd multiples predominate.

Conical tubes are equivalent to open tubes of the same length, as illustrated by the following diagrams. These diagrams are obtained from the ones for the open tube, by squashing down one end.

pressure displacement



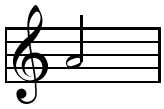
The oboe has a conical bore so again all multiples are present. This explains why the flute and oboe overblow at the octave, while the clarinet overblows at an octave plus a perfect fifth, which represents tripling the frequency. The odd multiples of the fundamental frequency dominate for a clarinet, although in practice there are small amplitudes present for the even ones from four times the fundamental upwards as well.

1.7. Trigonometric identities and beats

Since angles in mathematics are measured in radians, and there are 2π radians in a cycle, a sine wave with frequency ν in Hertz, peak amplitude c and *phase* ϕ will correspond to a sine wave of the form

$$c\sin(2\pi\nu t + \phi). \tag{1.7.1}$$

The quantity $\omega = 2\pi\nu$ is called the *angular velocity*. The role of the angle ϕ is to tell us where the sine wave crosses the time axis (look back at the graph in §1.4). For example, a cosine wave is related to a sine wave by the equation $\cos x = \sin(x + \frac{\pi}{2})$, so a cosine wave is really just a sine wave with a different phase.



440 Hz

For example, modern concert pitch[6] places the note A above middle C at 440 Hz so this would be represented by a wave of the form

$$c\sin(880\pi t + \phi).$$

This can be converted to a linear combination of sines and cosines using the standard formulas for the sine and cosine of a sum:

$$\sin(A + B) = \sin A\cos B + \cos A\sin B \tag{1.7.2}$$

$$\cos(A + B) = \cos A\cos B - \sin A\sin B. \tag{1.7.3}$$

So we have

$$c\sin(\omega t + \phi) = a\cos\omega t + b\sin\omega t$$

where

$$a = c\sin\phi \qquad b = c\cos\phi.$$

Conversely, given a and b, c and ϕ can be obtained via

$$c = \sqrt{a^2 + b^2} \qquad \tan\phi = a/b.$$

What happens when two pure sine or cosine waves are played at the same time? For example, why is it that when two very close notes are played simultaneously, we hear "beats"? Since this is the method by which strings on a piano are tuned, it is important to understand the origins of these beats.

The answer to this question also lies in the trigonometric identities (1.7.2) and (1.7.3). Since $\sin(-B) = -\sin B$ and $\cos(-B) = \cos B$, replacing B by $-B$ in equations (1.7.2) and (1.7.3) gives

$$\sin(A - B) = \sin A\cos B - \cos A\sin B \tag{1.7.4}$$

$$\cos(A - B) = \cos A\cos B + \sin A\sin B. \tag{1.7.5}$$

[6]Historically, this was adopted as the U.S.A. Standard Pitch in 1925, and in May 1939 an international conference in London agreed that this should be adopted as the modern concert pitch. Before that time, a variety of standard frequencies were used. For example, in the time of Mozart, the note A had a value closer to 422 Hz, a little under a semitone flat to modern ears. Before this time, in the Baroque and earlier, there was even more variation. For example, in Tudor Britain, secular vocal pitch was much the same as modern concert pitch, while domestic keyboard pitch was about three semitones lower and church music pitch was more than two semitones higher.

Adding equations (1.7.2) and (1.7.4)

$$\sin(A+B)+\sin(A-B)=2\sin A\cos B \tag{1.7.6}$$

which may be rewritten as

$$\sin A\cos B=\tfrac{1}{2}(\sin(A+B)+\sin(A-B)). \tag{1.7.7}$$

Similarly, adding and subtracting equations (1.7.3) and (1.7.5) gives

$$\cos(A+B)+\cos(A-B)=2\cos A\cos B \tag{1.7.8}$$

$$\cos(A-B)-\cos(A+B)=2\sin A\sin B, \tag{1.7.9}$$

or

$$\cos A\cos B=\tfrac{1}{2}(\cos(A+B)+\cos(A-B)) \tag{1.7.10}$$

$$\sin A\sin B=\tfrac{1}{2}(\cos(A-B)-\cos(A+B)). \tag{1.7.11}$$

This enables us to write any product of sines and cosines as a sum or difference of sines and cosines. So for example, if we wanted to integrate a product of sines and cosines, this would enable us to do so.

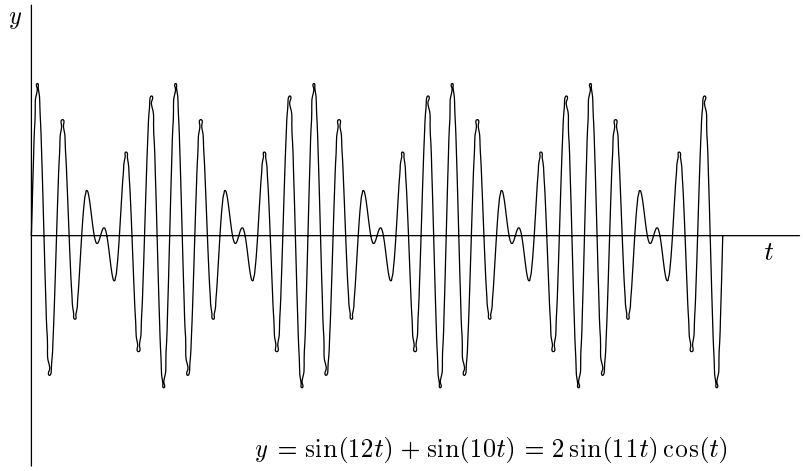
We are actually interested in the opposite process. So we set $u=A+B$ and $v=A-B$. Solving for A and B, this gives $A=\frac{1}{2}(u+v)$ and $B=\frac{1}{2}(u-v)$. Substituting in equations (1.7.6), (1.7.8) and (1.7.9), we obtain

$$\sin u+\sin v=2\sin\tfrac{1}{2}(u+v)\cos\tfrac{1}{2}(u-v) \tag{1.7.12}$$

$$\cos u+\cos v=2\cos\tfrac{1}{2}(u+v)\cos\tfrac{1}{2}(u-v) \tag{1.7.13}$$

$$\cos u-\cos v=2\sin\tfrac{1}{2}(u+v)\sin\tfrac{1}{2}(u-v) \tag{1.7.14}$$

This enables us to write any sum or difference of sine waves and cosine waves as a product of sines and cosines. Exercise 1 at the end of this section explains what to do if there are mixed sines and cosines.



$y=\sin(12t)+\sin(10t)=2\sin(11t)\cos(t)$

So for example, suppose that a piano tuner has tuned one of the three strings corresponding to the note A above middle C to 440 Hz. The second string is still out of tune, so that it resonates at 436 Hz. The third is being

damped so as not to interfere with the tuning of the second string. Ignoring phase and amplitude for a moment, the two strings together will sound as

$$\sin(880\pi t) + \sin(872\pi t).$$

Using equation (1.7.12), we may rewrite this sum as

$$2\sin(876\pi t)\cos(4\pi t).$$

This means that we perceive the combined effect as a sine wave with frequency 438 Hz, the average of the frequencies of the two strings, but with the amplitude modulated by a slow cosine wave with frequency 2 Hz, or half the difference between the frequencies of the two strings. This modulation is what we perceive as beats. The amplitude of the modulating cosine wave has two peaks per cycle, so the number of beats per second will be four, not two. So the number of beats per second is exactly the difference between the two frequencies. The piano tuner tunes the second string to the first by tuning out the beats, namely by adjusting the string so that the beats slow down to a standstill.

If we wish to include terms for phase and amplitude, we write

$$c\sin(880\pi t + \phi) + c\sin(872\pi t + \phi').$$

where the angles ϕ and ϕ' represent the phases of the two strings. This gets rewritten as

$$2c\sin(876\pi t + \tfrac{1}{2}(\phi + \phi'))\cos(4\pi t + \tfrac{1}{2}(\phi - \phi')),$$

so this equation can be used to understand the relationship between the phase of the beats and the phases of the original sine waves.

If the amplitudes are different, then the beats will not be so pronounced because part of the louder note is "left over". This prevents the amplitude going to zero when the modulating cosine takes the value zero.

Exercises

1. Use the equation $\cos\theta = \sin(\pi/2 + \theta)$ and equations (1.7.12)–(1.7.13) to express $\sin u + \cos v$ as a product of trigonometric functions.

2. A piano tuner comparing two of the three strings on the same note of a piano hears five beats a second. If one of the two notes is concert pitch A (440 Hz), what are the possibilities for the frequency of vibration of the other string?

3. Evaluate $\displaystyle\int_0^{\pi/2} \sin(3x)\sin(4x)\,dx.$

4. (a) Setting $A = B = \theta$ in formula (1.7.10) gives the double angle formula

$$\cos^2\theta = \tfrac{1}{2}(1 + \cos(2\theta)). \tag{1.7.15}$$

Draw graphs of the functions $\cos^2\theta$ and $\cos(2\theta)$. Try to understand formula (1.7.15) in terms of these graphs.

(b) Setting $A = B = \theta$ in formula (1.7.11) gives the double angle formula

$$\sin^2\theta = \tfrac{1}{2}(1 - \cos(2\theta)). \tag{1.7.16}$$

Draw graphs of the functions $\sin^2\theta$ and $\cos(2\theta)$. Try to understand formula (1.7.16) in terms of these graphs.

5. In the formula (1.7.1), the factor c is called the *peak amplitude*, because it determines the highest point on the waveform. In sound engineering, it is often more useful to know the *root mean square*, or RMS amplitude, because this is what determines things like power consumption. The RMS amplitude is calculated by integrating the square of the value over one cycle, dividing by the length of the cycle to obtain the mean square, and then taking the square root. For a pure sine wave given by formula (1.7.1), show that the RMS amplitude is given by

$$\sqrt{\nu \int_0^{\frac{1}{\nu}} [c\sin(2\pi\nu t + \phi)]^2 \, dt} = \frac{c}{\sqrt{2}}.$$

6. Use equation (1.7.11) to write $\sin kt \sin \frac{1}{2}t$ as $\frac{1}{2}(\cos(k - \frac{1}{2})t - \cos(k + \frac{1}{2})t)$. Show that

$$\sum_{k=1}^{n} \sin kt = \frac{\cos\frac{1}{2}t - \cos(n + \frac{1}{2})t}{2\sin\frac{1}{2}t} = \frac{\sin\frac{1}{2}(n+1)t \sin\frac{1}{2}nt}{\sin\frac{1}{2}t}. \tag{1.7.17}$$

Similarly, show that

$$\sum_{k=1}^{n} \cos kt = \frac{\sin(n + \frac{1}{2})t - \sin\frac{1}{2}t}{2\sin\frac{1}{2}t} = \frac{\cos\frac{1}{2}(n+1)t \sin\frac{1}{2}nt}{\sin\frac{1}{2}t}. \tag{1.7.18}$$

7. Two pure sine waves are sounded. One has frequency slightly greater or slightly less than twice that of the other. Would you expect to hear beats?

1.8. Superposition

Superposing two sounds corresponds to adding the corresponding wave functions. This is part of the concept of *linearity*. In general, a system is linear if two conditions are satisfied. The first, *superposition*, is that two simultaneous independent input signals should give rise to the sum of the two outputs. The second condition, *homogeneity*, says that magnifying the input level by a constant factor should multiply the output level by the same constant factor.

Superposing harmonic motions of the same frequency works as follows. Two simple harmonic motions with the same frequency, but possibly different amplitudes and phases, always add up to give another simple harmonic motion with the same frequency. We saw some examples of this in the last section. In this section, we see that there is an easy graphical method for carrying this out in practise.

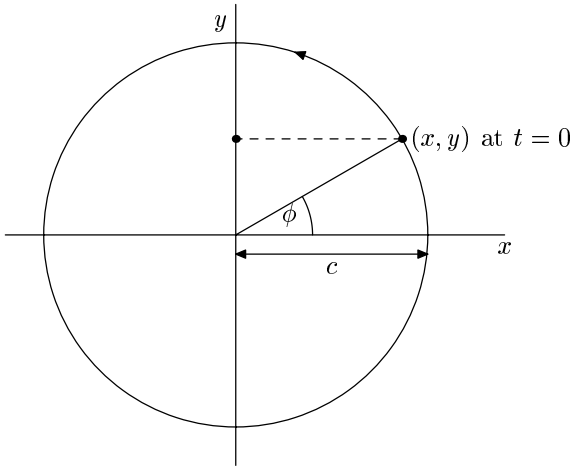
Consider a sine wave of the form $c\sin(\omega t + \phi)$ where $\omega = 2\pi\nu$. This may be regarded as the y-component of circular motion of the form

$$x = c\cos(\omega t + \phi)$$
$$y = c\sin(\omega t + \phi).$$

Since $\sin^2\theta + \cos^2\theta = 1$, squaring and adding these equations shows that the point (x, y) lies on the circle

$$x^2 + y^2 = c^2$$

with radius c, centered at the origin. As t varies, the point (x, y) travels counterclockwise round this circle ν times in each second, so ν is really measuring the number of cycles per second around the origin, and ω is measuring the angular velocity in radians per second. The phase ϕ is the angle, measured counterclockwise from the positive x-axis, subtended by the line from $(0, 0)$ to (x, y) when $t = 0$.



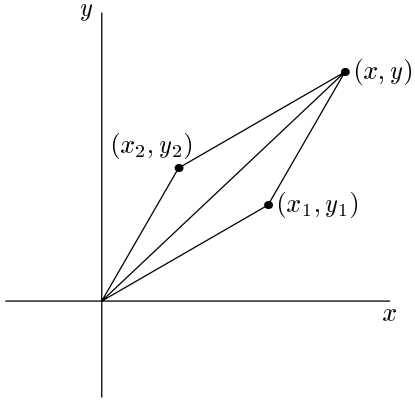
Now suppose that we are given two sine waves of the same frequency, say $c_1 \sin(\omega t + \phi_1)$ and $c_2 \sin(\omega t + \phi_2)$. The corresponding vectors at $t = 0$ are

$$(x_1, y_1) = (c_1 \cos\phi_1, c_1 \sin\phi_1)$$
$$(x_2, y_2) = (c_2 \cos\phi_2, c_2 \sin\phi_2).$$

To superpose (i.e., add) these sine waves, we simply add these vectors to give

$$\begin{aligned}(x, y) &= (c_1 \cos\phi_1 + c_2 \cos\phi_2, c_1 \sin\phi_1 + c_2 \sin\phi_2) \\ &= (c\cos\phi, c\sin\phi).\end{aligned}$$

We draw a copy of the line segment $(0, 0)$ to (x_1, y_1) starting at (x_2, y_2), and a copy of the line segment $(0, 0)$ to (x_2, y_2) starting at (x_1, y_1), to form a parallelogram. The amplitude c is the length of the diagonal line drawn from the origin to the far corner (x, y) of the parallelogram formed this way. The angle ϕ is the angle subtended by this line, measured as usual counterclockwise from the x-axis.



Exercises

1. Write the following expressions in the form $c\sin(2\pi\nu t + \phi)$:

(i) $\cos(2\pi t)$

(ii) $\sin(2\pi t) + \cos(2\pi t)$

(iii) $2\sin(4\pi t + \pi/6) - \sin(4\pi t + \pi/2)$.

2. Read Appendix C. Use equation (C.1) to interpret the graphical method described in this section as motion in the complex plane of the form

$$z = e^{i\omega t + \phi}.$$

1.9. Damped harmonic motion

Damped harmonic motion arises when in addition to the restoring force $F = -ky$, there is a frictional force proportional to velocity,

$$F = -ky - \mu\dot{y}.$$

For positive values of μ, the extra term damps the motion, while for negative values of μ it promotes or forces the harmonic motion. In this case, the differential equation we obtain is

$$m\ddot{y} + \mu\dot{y} + ky = 0. \tag{1.9.1}$$

This is what is called a linear second order differential equation with constant coefficients. To solve such an equation, we look for solutions of the form

$$y = e^{\alpha t}.$$

Then $\dot{y} = \alpha e^{\alpha t}$ and $\ddot{y} = \alpha^2 e^{\alpha t}$. So for y to satisfy the original differential equation, α has to satisfy the *auxiliary equation*

$$mY^2 + \mu Y + k = 0. \tag{1.9.2}$$

If the quadratic equation (1.9.2) has two different solutions, $Y = \alpha$ and $Y = \beta$, then $y = e^{\alpha t}$ and $y = e^{\beta t}$ are solutions of (1.9.1). Since equation (1.9.1) is linear, this implies that any combination of the form

$$y = Ae^{\alpha t} + Be^{\beta t}$$

is also a solution. The *discriminant* of the auxiliary equation (1.9.2) is

$$\Delta = \mu^2 - 4mk.$$

If $\Delta > 0$, corresponding to large damping or forcing term, then the solutions to the auxiliary equation are

$$\alpha = (-\mu + \sqrt{\Delta})/2m$$
$$\beta = (-\mu - \sqrt{\Delta})/2m,$$

and so the solutions to the differential equation (1.9.1) are

$$y = Ae^{(-\mu+\sqrt{\Delta})t/2m} + Be^{(-\mu-\sqrt{\Delta})t/2m}. \tag{1.9.3}$$

In this case, the motion is so damped that no sine waves can be discerned. The system is then said to be *overdamped*, and the resulting motion is called *dead beat*.

If $\Delta < 0$, as happens when the damping or forcing term is small, then the system is said to be *underdamped*. In this case, the auxiliary equation (1.9.2) has no real solutions because Δ has no real square roots. But $-\Delta$ is positive, and so it has a square root. In this case, the solutions to the auxiliary equation are

$$\alpha = (-\mu + i\sqrt{-\Delta})/2m$$
$$\beta = (-\mu - i\sqrt{-\Delta})/2m,$$

where $i = \sqrt{-1}$. See Appendix C for a brief introduction to complex numbers. So the solutions to the original differential equation are

$$y = e^{-\mu t/2m}(Ae^{it\sqrt{-\Delta}/2m} + Be^{-it\sqrt{-\Delta}/2m}).$$

We are really interested in real solutions. To this end, we use relation (C.1) to write this as

$$y = e^{-\mu t/2m}((A+B)\cos(t\sqrt{-\Delta}/2m) + i(A-B)\sin(t\sqrt{-\Delta}/2m)).$$

So we obtain real solutions by taking $A' = A + B$ and $B' = i(A - B)$ to be real numbers, giving

$$y = e^{-\mu t/2m}(A'\sin(t\sqrt{-\Delta}/2m) + B'\cos(t\sqrt{-\Delta}/2m)). \tag{1.9.4}$$

The interpretation of this is harmonic motion with a damping factor of $e^{-\mu t/2m}$.

The special case $\Delta = 0$ has solutions

$$y = (At + B)e^{-\mu t/2m}. \tag{1.9.5}$$

This borderline case resembles the case $\Delta > 0$, inasmuch as harmonic motion is not apparent. Such a system is said to be *critically damped*.

Examples

1. The equation

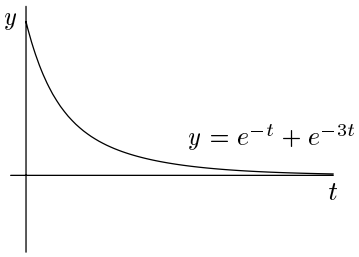
$$\ddot{y} + 4\dot{y} + 3y = 0 \tag{1.9.6}$$

is overdamped. The auxiliary equation

$$Y^2 + 4Y + 3 = 0$$

factors as $(Y+1)(Y+3) = 0$, so it has roots $Y = -1$ and $Y = -3$. It follows that the solutions of (1.9.6) are given by

$$y = Ae^{-t} + Be^{-3t}.$$



2. The equation

$$\ddot{y} + 2\dot{y} + 26y = 0 \tag{1.9.7}$$

is underdamped. The auxiliary equation is

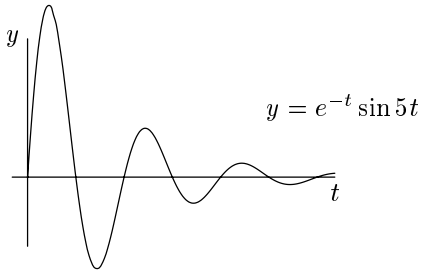
$$Y^2 + 2Y + 26 = 0.$$

Completing the square gives $(Y+1)^2 + 25 = 0$, so the solutions are $Y = -1 \pm 5i$. It follows that the solutions of (1.9.7) are given by

$$y = e^{-t}(Ae^{5it} + Be^{-5it}),$$

or

$$y = e^{-t}(A' \cos 5t + B' \sin 5t). \tag{1.9.8}$$



3. The equation

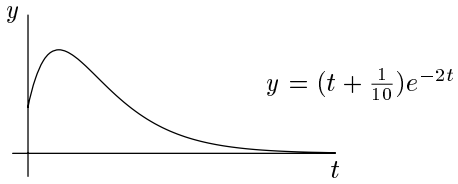
$$\ddot{y} + 4\dot{y} + 4y = 0 \tag{1.9.9}$$

is critically damped. The auxiliary equation

$$Y^2 + 4Y + 4 = 0$$

factors as $(Y+2)^2 = 0$, so the only solution is $Y = -2$. It follows that the solutions of (1.9.9) are given by

$$y = (At + B)e^{-2t}.$$



Exercises

1. Show that if $\Delta = \mu^2 - 4mk > 0$ then the functions (1.9.3) satisfy the differential equation (1.9.1).

2. Show that if $\Delta = \mu^2 - 4mk < 0$ then the functions (1.9.4) satisfy the differential equation (1.9.1).

3. Show that if $\Delta = \mu^2 - 4mk = 0$ then the auxiliary equation (1.9.2) is a perfect square, and the functions (1.9.5) satisfy the differential equation (1.9.1).

1.10. Resonance

Forced harmonic motion is where there is a forcing term $f(t)$ (often taken to be periodic) added into equation (1.9.1) to give an equation of the form

$$m\ddot{y} + \mu\dot{y} + ky = f(t). \tag{1.10.1}$$

This represents a damped system with an external stimulus $f(t)$ applied to it. We are particularly interested in the case where $f(t)$ is a sine wave, because this represents forced harmonic motion. Forced harmonic motion is responsible for the production of sound in most musical instruments, as well as the perception of sound in the cochlea. We shall see that forced harmonic motion is what gives rise to the phenomenon of *resonance*.

There are two steps to the solution of the equation. The first is to find the general solution to equation (1.9.1) without the forcing term, as described in §1.9, to give the *complementary function*. The second step is to find by any method, such as guessing, a single solution to equation (1.10.1). This is called a *particular integral*. Then the general solution to the equation (1.10.1) is the sum of the particular integral and the complementary function.

Examples

1. Consider the equation

$$\ddot{y} + 4\dot{y} + 5y = 10t^2 - 1. \tag{1.10.2}$$

We look for a particular integral of the form $y = at^2 + bt + c$. Differentiating, we get $\dot{y} = 2at + b$ and $\ddot{y} = 2a$. Plugging these into (1.10.2) gives

$$2a + 4(2at + b) + 5(at^2 + bt + c) = 10t^2 + t - 3.$$

Comparing coefficients of t^2 gives $5a = 10$ or $a = 2$. Then comparing coefficients of t gives $8a + 5b = 1$, so $b = -3$. Finally, comparing constant terms gives $2a + 4b + 5c = -3$, so $c = 1$. So we get a particular integral of $y = 2t^2 - 3t + 1$. Adding the complementary function (1.9.8), we find that the general solution to (1.10.2) is given by

$$y = 2t^2 - 3t + 1 + e^{-2t}(A' \cos t + B' \sin t).$$

2. As a more interesting example, to solve

$$\ddot{y} + 4\dot{y} + 5y = \sin 2t, \tag{1.10.3}$$

we look for a particular integral of the form

$$y = a \cos 2t + b \sin 2t.$$

Equating coefficients of $\cos 2t$ and $\sin 2t$ we get two equations:

$$\begin{aligned} -8a + b &= 1 \\ a + 8b &= 0. \end{aligned}$$

Solving these equations, we get $a = -\frac{8}{65}$, $b = \frac{1}{65}$. So the general solution to (1.10.3) is

$$y = \frac{\sin 2t - 8 \cos 2t}{65} + e^{-2t}(A' \cos t + B' \sin t).$$

The case of forced harmonic motion of interest to us is the equation

$$m\ddot{y} + \mu\dot{y} + ky = R\cos(\omega t + \phi). \tag{1.10.4}$$

This represents a damped harmonic motion (see §1.9) with forcing term of amplitude R and angular velocity ω.

We could proceed as above to look for a particular integral of the form

$$y = a\cos\omega t + b\sin\omega t$$

and proceed as in the second example above. However, we can simplify the calculation by using complex numbers (see Appendix C). Since this differential equation is linear, and since

$$Re^{i(\omega t+\phi)} = R(\cos(\omega t + \phi) + i\sin(\omega t + \phi))$$

it will be enough to find a particular integral for the equation

$$m\ddot{y} + \mu\dot{y} + ky = Re^{i(\omega t+\phi)}, \tag{1.10.5}$$

which represents a complex forcing term with unit amplitude and angular velocity ω. Then we take the real part to get a solution to equation (1.10.4).

We look for solutions of equation (1.10.5) of the form $y = Ae^{i(\omega t+\phi)}$, with A to be determined. We have $\dot{y} = Ai\omega e^{i(\omega t+\phi)}$ and $\ddot{y} = -A\omega^2 e^{i(\omega t+\phi)}$. So plugging into equation (1.10.5) and dividing by $e^{i(\omega t+\phi)}$, we get

$$A(-m\omega^2 + i\mu\omega + k) = R$$

or

$$A = \frac{R}{-m\omega^2 + i\mu\omega + k}.$$

So the particular integral, which actually represents the eventual "steady state" solution to the equation since the complementary function is decaying, is given by

$$y = \frac{Re^{i(\omega t+\phi)}}{-m\omega^2 + i\mu\omega + k}.$$

The bottom of this expression is a complex constant, and so this solution moves around a circle in the complex plane. The real part is then a sine wave with the radius of the circle as amplitude and with a phase determined by the argument of the bottom.

The amplitude of the resulting vibration, and therefore the degree of resonance (since we started with a forcing term of unit amplitude) is given by taking the absolute value of this solution,

$$|y| = \frac{R}{\sqrt{(k - m\omega^2)^2 + \mu^2\omega^2}}.$$

This amplitude magnification reaches its maximum when the derivative of $(k - m\omega^2)^2 + \mu^2\omega^2$ vanishes, namely when

$$\omega = \sqrt{\frac{k}{m} + \frac{\mu^2}{2m^2}},$$

when we have amplitude $mR/(\mu\sqrt{km+3\mu^2/4})$. The above value of ω is called the *resonant frequency* of the system. Note that this value of ω is slightly greater than the value which one may expect from Equation (1.9.4) for the complementary function:

$$\omega = \frac{\sqrt{-\Delta}}{2m} = \sqrt{\frac{k}{m} - \frac{\mu^2}{4m^2}},$$

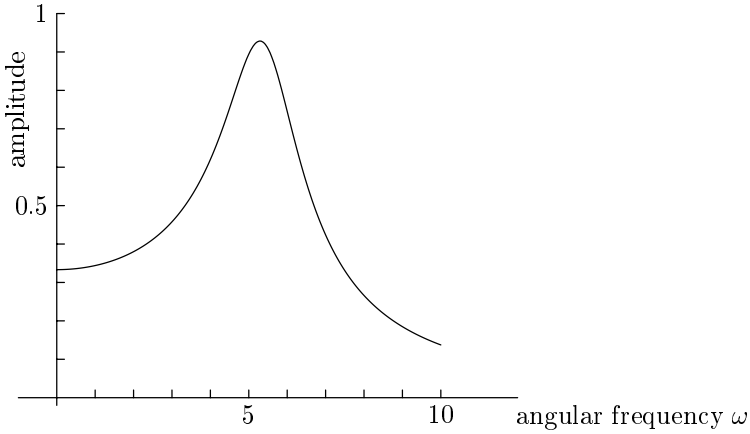
or even than the value of ω for the corresponding undamped system:

$$\omega = \sqrt{\frac{k}{m}}.$$

Example. Consider the forced, underdamped equation

$$\ddot{y} + 2\dot{y} + 30y = 10\sin\omega t.$$

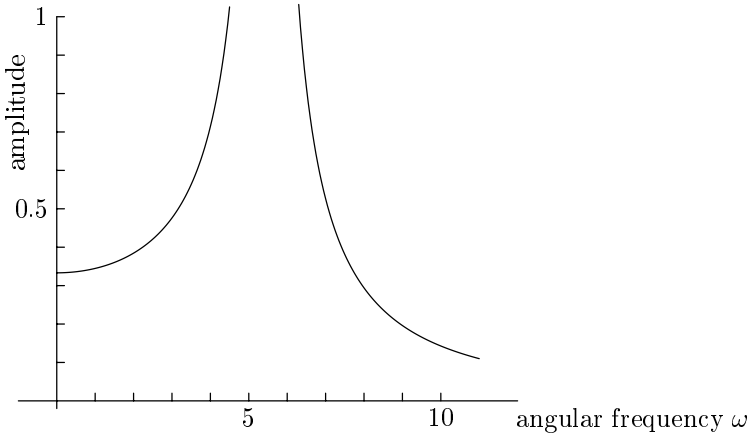
The above formula tells us that the amplitude of the resulting steady state sine wave solution is $10/\sqrt{900-56\omega^2+\omega^4}$, which has its maximum value at $\omega = \sqrt{31}$.



Without the damping term, the amplitude of the steady state solution to the equation

$$\ddot{y} + 30y = 10\sin\omega t,$$

is equal to $10/|30-\omega^2|$. It has an "infinitely sharp" peak at $\omega = \sqrt{30}$.



At this stage, it seems appropriate to introduce the terms *resonant frequency* and *bandwidth* for a resonant system. The resonant frequency is the frequency for which the amplitude of the steady state solution is maximal. Bandwidth is a vague term, used to describe the width of the peak in the above graphs. So in the damped example above, we might want to describe the bandwidth as being from roughly $4\frac{1}{2}$ to $6\frac{1}{2}$, while for the undamped example it would be somewhat wider. Sometimes, the term is made precise by taking the interval between the two points either side of the peak where the amplitude is $1/\sqrt{2}$ times that of the peak. Since power is proportional to square of amplitude, this corresponds to a factor of two in the power, or a difference of about 3 dB.[7]

[7]The exact value is $10\log_{10}(2)$ dB.