Course: Digital Audio Signal Processing

The Listener's Model 1

The organ of hearing

Aurelio Uncini

Dept. of Information Engineering, Electronincs and Telecommunications

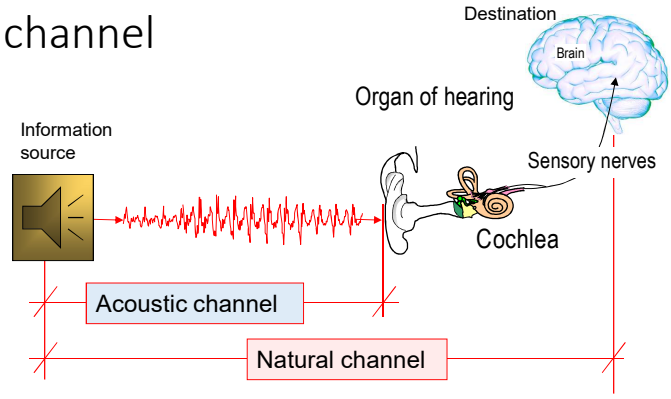
SAPIENZA
UNIVERSITÀ DI ROMA

intelligent signal processing
and multimedia lab

Premise

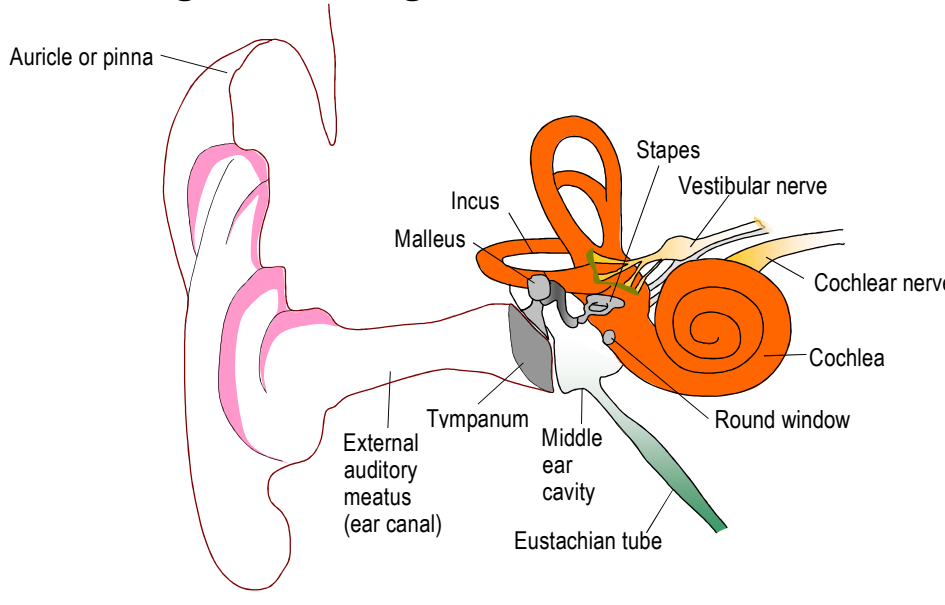
- The ***organ of hearing*** is one of the most sophisticated of the human organism and the phenomenon of acoustic perception is very complex.
- In the communications theory, for the optimal source coding it is necessary to specify the receiver model, in the high-definition audio a precise receiver model does not exist.
- In this case, in fact, the end user is the *organ of hearing* and to realize an optimal coding of the audio signal it is necessary to know the way in which the hearing perceives the sounds.
- The relationships that exist between the physically measurable acoustic and subjective subjective quantities, linked to both physiological and psychological mechanisms, are the subject of the discipline called **psychoacoustic**.

The complexity of natural communication channel



- While the acoustics deals with measurable physical quantities, the perception of sound by the listener is always subjective.
- Thus, a question of primary importance, both in acoustics and in the manipulation of the audio signal, concerns the ***listener's model***

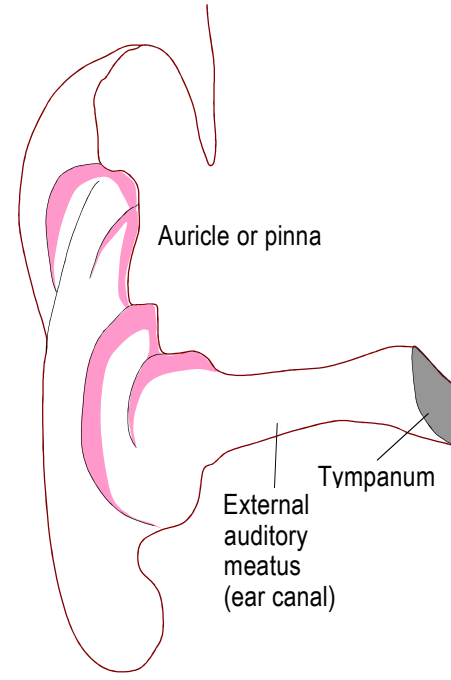
Human Organ of Earing



Outer Ear

The **outer or external ear** is formed by the auricle or pinna and the external **auditory canal** or **ear canal**.

The **ear canal** can increase the sound pressure level by 10-12 dB for frequencies around 3000 Hz (corresponding to speech), while it has little resonator effect for frequencies below 1000 Hz and above 7000 Hz.



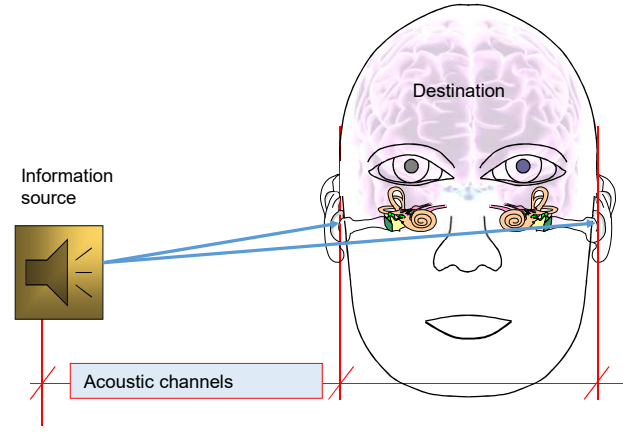
Outer Ear

The **outer ear** makes it possible to identify the spatial origin of the sound.

The waves reflected from the various points of the pinna and towards the mouth of the ear canal, are each other not in phase due to the different paths.

From these phase displacements and the consequent spectral variations, even if minimal, the **nervous system is able to obtain directional information of the incoming sound**.

Although experimental evidence has shown that with only one ear there is some sense of the direction of arrival, **the precise localization of the sound source requires the presence of two ears**.



The paths between the source and the auditory channels are modeled with transfer functions indicated as **head related transfer functions** (HRTF).

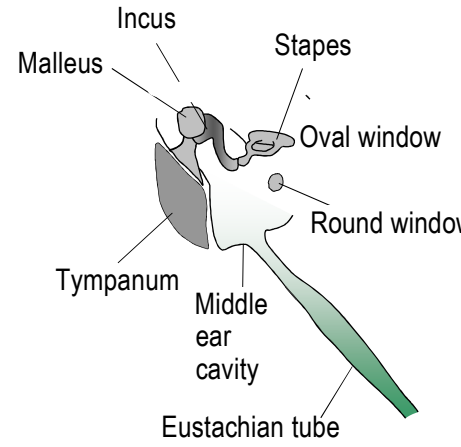
Middle Ear

The **outer ear** and the middle ear must transmit changes in air pressure to the inner ear.

The force coming from the tympanum vibration acts on the **malleus** (i.e the hammer) is transmitted through the **incus** (i.e the anvil) to the base of the **stapes** (i.e. the stirrup), which exerts a mechanical pressure on the *oval window* (or *fenestra vestibuli*).

Therefore, these ossicles form a system of levers, able to triple the transmitted force.

Furthermore, the surface of the oval window is about 30 times lower than that of the eardrum.



The ossicles act as an impedance matching device

Middle Ear

The acoustic pressure acting on the oval window is about 2 orders of magnitude higher than that of the incoming sound on the eardrum. The middle ear is able to ``**defend itself" from very intense sounds**: the tympanic muscle (or tympanic tensor) stiffens and prevents the eardrum from deforming too much.

At the same time, the stapedius muscle, which acts on the stapes, removes it from the oval window and reduces the vibration transfer (Stapedio reflex).

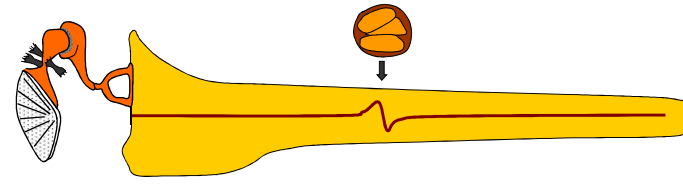
This mechanism takes a few moments to become operational, **so it is not effective in the presence of sudden and very intense sounds, such as explosions**.

The middle ear communicates with the outer ear through a small channel (**Eustachian tube** or trumpet) that connects the tympanic box with the pharynx; this makes it possible to balance the external static pressure acting on the eardrum membrane, which is stressed only by rapid pressure variations due to sound waves.

Inner Ear

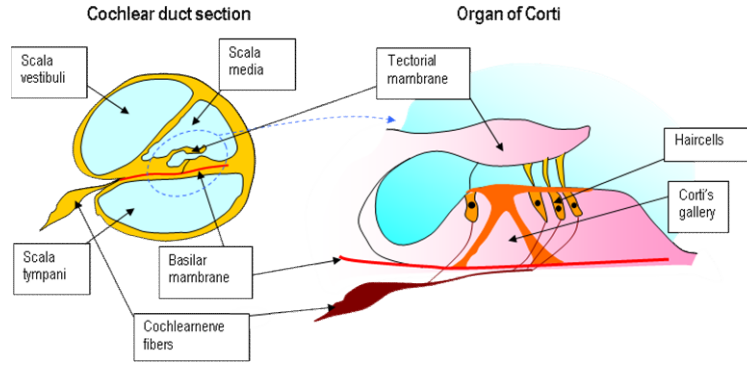
The inner ear is a very complex structure, which also includes the organ of equilibrium, where the processes of transforming the sounds take place in electrochemical stimuli conveyed to the brain via the acoustic nerve.

Schematic operation of the cochlea. The basilar mambrana constitutes the platform of the organ of Corti. It is not a linear membrane but a sheet of cellular fibers (collagen). It is thick and narrow at the base, thin and wide at the apex. The different rigidity causes the mechanical "frequency tuning" of the cochlea.



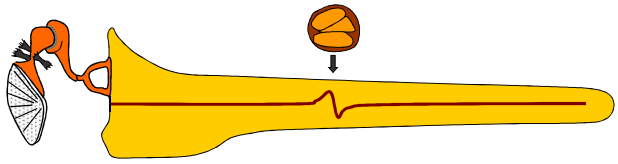
The inner ear includes the *cochlea* (the real organ of hearing) and the labyrinth, which regulates the equilibrium. The cochlea, shown in Figure *snail-shaped* tube divided into two channels, called the *vestibular* and *tympanic canal*, separated by the basement membrane. The signal is transmitted from the oval window, runs through the cochlea to its center from one side to the other of the vestibular canal and continues its cycle through the tympanic canal. During the passage of the signal, the basal membrane is under stress due to the difference in pressure present between the two channels. Inside the cochlea is the *Organ of Corti* that is formed by a series of *ciliated cells* and nerve fibers through which the sound is perceived and transmitted the action potential to the brain.

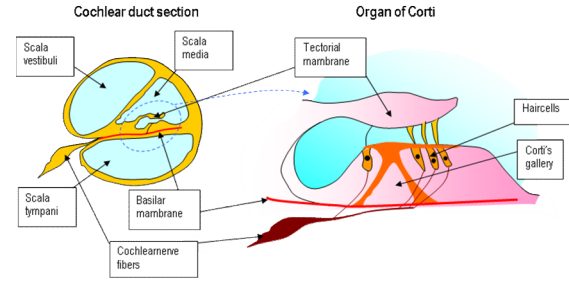
Inner Ear



The organ of Corti represents the sensory structure able to transduce the sound stimulus, is located on the surface of the basilar membrane. The organ contains ciliated cells, supporting cells and a membrane overlying the hair cells, called the tectoria membrane. The receptor cells are particular in that they have at one end of the membrane the cilia called stereociglia; the ends of such cilia are inserted into the tectorial membrane. The anatomy of the organ of Corti causes the sound waves to determine the mechanical folding of the stereociglia that generates potential receptorial cells in ciliated cells

Inner Ear

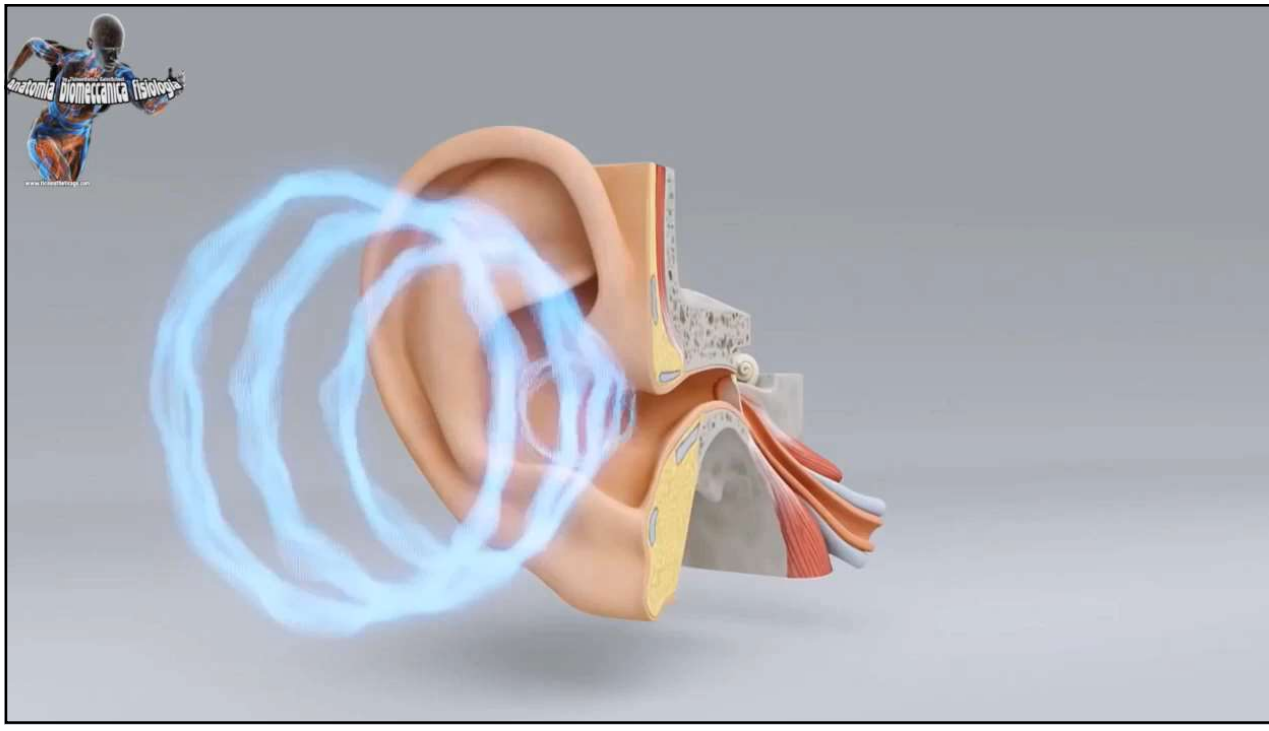




The basilar membrane response to an input sound produces a small oscillation at the position where the resonance occurs. Assuming the tension negligible along the whole longitudinal direction and a linear behavior, the membrane is similar to a bank of resonators (or filters) tuned to different frequencies.

The cochlea behaves therefore as a filter bank that performs a spectral analysis and that provides a frequency position map of the input signal.

The mechanical behavior of the cochlea is non-linear and very important psychoacoustic effects are produced. In normal listening conditions (30-70 dB SPL) the cochlea operates in two different damping regimes: for levels below 30-40 dB SPL there is a sort of amplification and sometimes a delay; for a level above 60-70 dB SPL there is an attenuation.

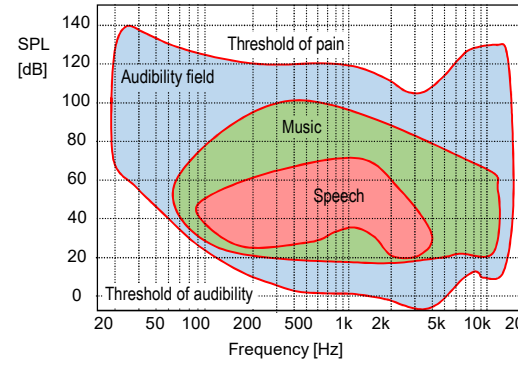


Principles of Psychoacoustics

It is known that the human ear is a sophisticated transducer that is able to pick up sounds of even many different levels.

Figure shows the perception areas of the most common acoustic sound perceivable by the hearing organ.

In fact, we also know that at lower sound levels, the sounds in vocal-frequency range are better perceived and that raising the level of acoustic intensity, this phenomenon is less pronounced.



The human auditory system is not only formed by the ear. The way in which the information transmitted by the acoustic nerve is processed by the brain is more important than the information itself.

Fletcher and Munson curve

In general terms, called $s(t)$ the function that characterizes the sound (i.e the objective sound), and called $p(t)$ the function of the *perceived sound*, we can postulate the existence of a nonlinear operator Ψ defined as a psychoacoustic operator such that $p(t) = \Psi[s(t)]$.

Failed to attempt to represent the operator in an analytical way, Fletcher and Munson proposed an experimental evaluation methodology

Fletcher-Munson sperimental curves represent the equal loudness (or isophonic) contours for the human ear denote as *normal audiogram*, for which a listener perceives a constant loudness when presented with pure steady tones.

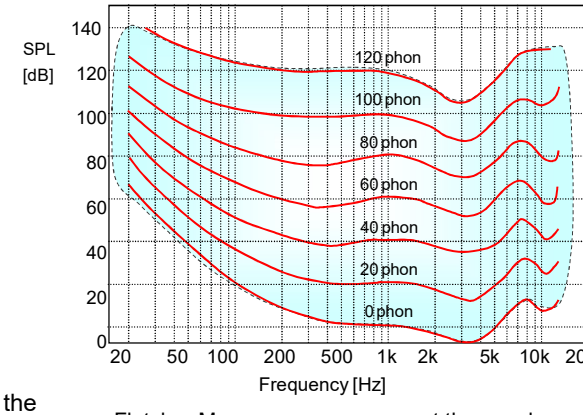
Note that human ears are most sensitive to frequencies around 1kHz to 5kHz. That's the most useful range for hearing human speech.

Fletcher and Munson curve

In the interval between the audibility and the pain thresholds, level curves are experimentally obtained, corresponding to an identical perceived sound level.

The ear can perceive with equivalent levels sounds of different frequencies having very different intensities. Listening and comparison tests are performed between two sounds of different frequency, one of which is a reference sound at a fixed frequency and equal to 1 [kHz] produced with a certain intensity (in [dB]). The other sound is emitted at a certain frequency and its intensity is varied.

A group of listeners are asked to evaluate at which intensity of the sound under examination the sound stimuli are perceived as equal. In this way, isofonic curves are constructed, such as those shown in Fig. 2.5, corresponding to the same sound level perceptions of sine-wave sounds as the frequency varies.



Fletcher-Munson curves represent the equal-loudness (or isophonic) contours for the human ear denote as normal audiogram, for which a listener perceives a constant loudness when presented with pure steady tones [2]. Note that human ears are most sensitive to frequencies around 1kHz to 5kHz. That's the most useful range for hearing human speech.

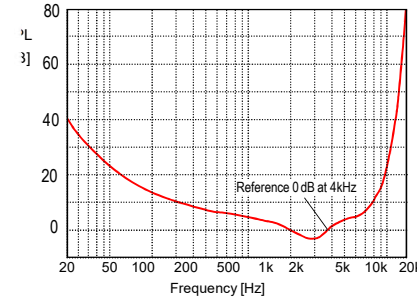
Approximate Model of Absolute Audibility Threshold

An approximate model of the absolute threshold of audibility is given by the following formula

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4, \quad \text{[dB SPL]} \quad (2.1)$$

This model, whose performance is shown in Fig. 2.6, has been calculated on the basis of averages made on young listeners with a "good hearing" [7].

The quantity $T_q(f)$ calculated with Eqn. (2.1) is used in the audio coders for the compression based on perceptual models (e.g. MPEG1) in order to check that the quantization noise of the encoder is always below this threshold.

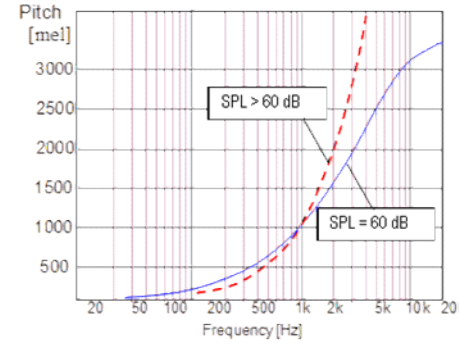


Remark 2.1. Note that, when using the $Tq(f)$ for the audio codec design, should be taken into account thta the listening level is expressed in decibel, assuming that the listening level is such that the lowest level of the signal is about 0 [dB] (i.e. the amplitude of ± 1 bit at 4kHz).

Complex Sounds Perception

The pitch of a sound is a subjective feeling. The American Standards Association for example defines pitch as "*an attribute of the auditory sensation in which a sound can be ordered in a musical scale*". It is known that the perceived pitch does not always correspond exactly to the objective frequency of the signal.

The acoustic intensity has a significant effect on the perceived pitch for very high and very low frequencies. There is a particularly noticeable discordance above 1 kHz, between the real frequency of pure sound and the average pitcht perceived by the listener. As the sound pressure increases, the pitch of a low frequency sound decreases, while the pitch of a high frequency sound increases.



In order to have a frequency scale consistent with the perceived pitch, the **mel scale** has been introduced that measures the perceived pitch by the listener vs objective sound frequency (Fig. 2.7 continuous tract). The *mel scale*, named by Stevens, Volkmann, and Newman in 1937 [3], is a perceptual scale of pitches judged by listeners. It is clear from the graph in Fig. 2.7, where is shown the curve that describes the relationship between Hertz and mel, as the relationship between the frequency (in Hertz) and the pitch (in mels) is non-linear. Moreover, the scale is not valid for the frequencies of a sound with a complex spectrum, but rather for the fundamental frequency of the sound itself.

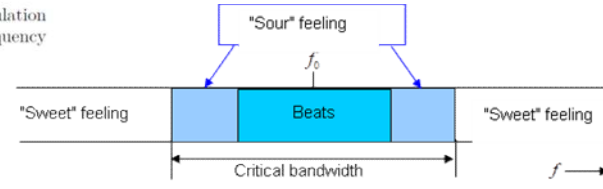
Critical Bands

The concept of *critical bandwidth* is of central importance in the modeling of auditory perception and widely used in modern signal encoders based on perceptive models.

To understand the concept of critical band we consider a simple experiment. If we add two sinusoidal signals of frequencies close to each other we have the *beats phenomenon* expressed by the relation[1]

$$\sin\omega t + \sin(\omega + \Delta\omega)t = 2\cos(\Delta\omega/2)t\cos(\omega + \Delta\omega/2)t$$

The previous relationship can be easily interpreted in terms of amplitude modulation in which the carrier signal $2\cos(\omega + \Delta\omega/2)t$ is modulated with a much lower frequency $\cos(\Delta\omega/2)$ signal.



2.3.4.1 Consonance and Dissonance

From the perceptual point of view, as shown in Fig. 2.8, if the frequencies of the two pure sounds are close to each other, the beat will be heard, that is the perception of a single sound modulated in amplitude (or with a tremolo). However, in the event that the frequencies are more distanced, a *sound sour* or dissonant sound will be heard. By further increasing the distance you will hear two distinct and not unpleasant sounds or *consonant sounds*.

Critical Bands

By "consonance" we mean the quality of an interval or a chord (more complex sound made up of several notes), which produces in the listener a sense of satisfaction and stability (chord or consonant interval) or a tension effect (chord or dissonant interval).

According to a theory elaborated in the nineteenth century by H. Helmholtz, an interval[2] is consonant when the two notes that form it have one or more harmonics in common. The more harmonics in common the more the interval is consonant. However, Western musical theory considers deterministic intervals as consonants while others as dissonant ones.

From the psychoacoustic point of view, the evaluation of consonance and dissonance in reality concerns the single person and the difference between the frequencies of the

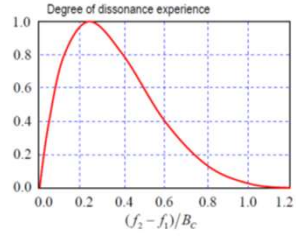


Fig. 2.9 Dissonance curve of two pure sounds. Quantification of the dissonance experienced when two pure tones of frequencies f_1 and f_2 are sounded simultaneously [27].

two pure tones such that the sound begins to be perceived again as "unpleasant" or "'dissonant', is defined as *critical band* as indicated in Fig. 2.8.

Critical Bands Width

A relevant aspect of psychoacoustics concerns the modality for determining the average bandwidth, of the critical bands valid for the whole population.

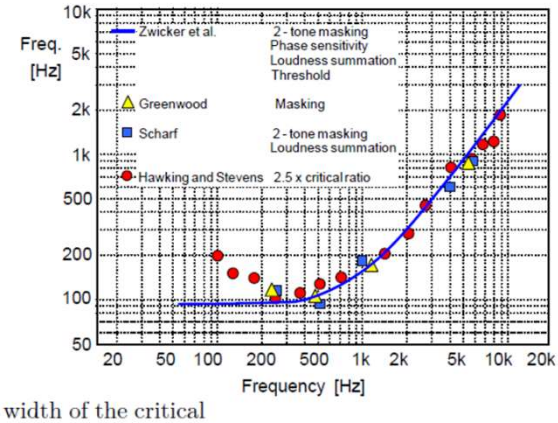
The critical bandwidth also influences the sensation of the intensity of a sound, the perception of a sound like noise, the *masking* or covering of a sound with another sound.

The width of the critical band depends on the way in which the hearing organ resolves the frequencies. The critical band is that band of width such that the subjective response changes very abruptly (empirical definition, reported by Scharf in [9], [13]).

According to this definition, a simple model of the hearing apparatus, also supported by the mechanical model of the cochlea, very useful in many practical applications, consists of a **passband filter bank with a width equal to the critical bands** followed by an energy detector.

Therefore the definition of the critical bands is strictly connected to the physical **modeling of the cochlea with a filter bank**.

More recently, Moore and Glasberg [11] have shown that the value of critical bands depends not only on the frequency but also on the sensitivity of the hearing system. For frequencies lower than 1kHz, where the hearing apparatus is less efficient (see loudness curves) the width from the critical band is narrower than those shown in Fig. 2.10.



For an average population of listeners, as indicated in [10], the width of the critical bands BW_c, can be approximated with the following function

$$BW_c(f) = 25 + 75[1 + 1.4(f/1000)^2]^{0.69}, \quad [\text{Hz}] \tag{2.2}$$

Although the previous expression is continuous with respect to frequency, in practical cases it is usual to think of the critical bands as band-pass filter bank conforming to the Eqn. (2.2). Table 2.1 shows the value of the central frequencies and the BW_c related to them [9].

Table 2.1 Central frequencies and relative bandwidths of the filter bank ideally representing the critical bands.

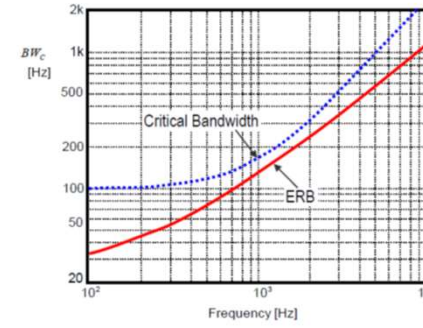
Band No.	Center Freq. [Hz]	Bandwidth [Hz]	Band No.	Center Freq. [Hz]	Bandwidth [Hz]	Band No.	Center Freq. [Hz]	Bandwidth [Hz]
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10500	9500-12000-
6	570	510-630	15	2500	2320-2700	24	13500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

2.3.4.3 Equivalent Rectangular Bandwidth (ERB)

Although the critical band model in Eqn. 2.2 is widely used, it is often referred to as an alternative expression in which the frequency responses of the filter bank are assumed to be perfectly rectangular. This approximation, which emerged from research in the field of psychoacoustics, led to the definition of the *equivalent rectangular bandwidth* (ERB), which can be approximated by the following expression

$$\mathrm{ERB}(f) = 24.7[4.37(f/1000)+1]. \tag{2.3}$$

The ERBs defined in this way, as can be seen in Fig. 2.10, differ from the expression (2.2). The use of Eqn. (2.3) is of primary importance in the physical modeling of auditory perception and in the design of the filter banks used in the coding of the audio signal (discussed later in the book).



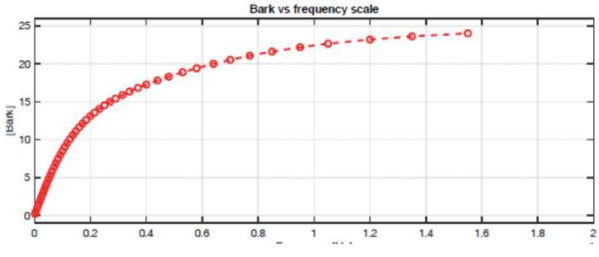
2.3.4.4 Bark Scale

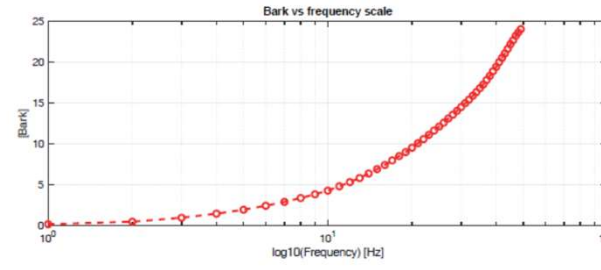
Proposed by Zwicker in 1961 [46], Bark's scale is a psychoacoustic scale for subjective measures of loudness, that is, a frequency scale in which the distance between to two channel of the filter-bank is defined on a perceptive metric, i.e. defined as one [Bark]. In fact, based on the results of many experiments, the Bark scale is defined so that the critical bands of human hearing have a width of one [Bark]. The range of perceivable frequencies (i.e. [20, 20kHz]) is covered by 25 [Bark] [46]-[10].

An expression to convert the distance in frequency to [Hz] in distance in [Bark] turns out to be

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan[(f/7500)^2], \quad [\mathrm{Bark}] \tag{2.4}$$

Fig. 2.12 shows the relationship between the bandwidth in [Bark] and the frequency in [Hz].





2.3.5 Masking

Listening to a sound composed of two pure sounds, the listener does not always perceive the components distinctly. When one of the two sounds has a lower intensity than the other, the latter is inaudible or *masked*. The *masking phenomenon* consists, in practice, in raising the audibility threshold in a interval of frequencies centered at a higher intensity sound, called *masking tone*.

Acoustic masking occurs when two or more frequencies near each other are simultaneously present. In this case the hair cells corresponding to the frequency of the lower level tone could already be excited by he presence of high level (closed in frequency) tone. Thus, the brain has no way of knowing that it is present. In practice, if a tone is only slightly stronger than the other, we can not hear the lower level tone. However, the masking has enormous advantages for audio compression. If we can not hear a sound, it does not need to be coded.

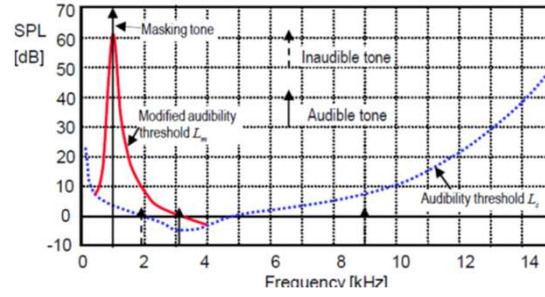


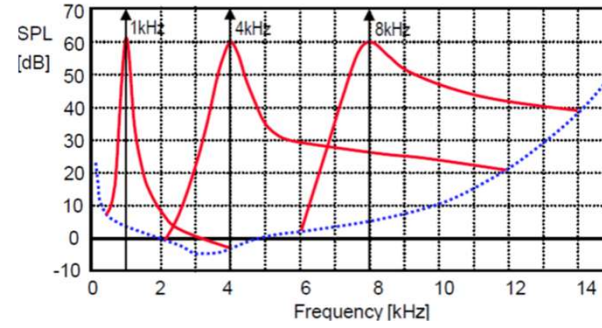
Fig. 2.13 Masking for pure sounds. a) Modified hearing threshold for a 1 kHz masking tone; b)

Definition 2.1. *Masking coefficient* - We define $M(f,p)$ masking coefficient for pure tones, depending on the frequency f and the acoustic intensity p, as the difference expressed in [dB] between the level of normal audibility threshold L_s (or reference threshold) and the increase of this threshold in presence of masking sound

$$M(f,p) = L_m(f,p) - L_s(f,p). \tag{2.5}$$

In Fig. 2.13, the qualitative values of $L_m(f,p)$ are shown for different frequency values and the level of the masking tone [14]. From the figure we can see how the shape of the threshold depends on both the level and the frequency. For low signal levels in general the surrounding masked frequencies shrinks. In particular, Figs 2.13-b) shows the approximate values of the *modified audibility threshold* by the presence of the masking tone at frequencies of 1, 4 and 8 kHz. Observe that as the frequency increases, the modified audibility threshold widens and its shape is not symmetrical with respect to the masking tone.

In Fig. 2.14, the average values of measurements taken on different subjects of the modified audibility threshold at various frequencies are reported. It should be noted that the presence of "holes" in the progression of the masking coefficient is due to



average measured trend of the modified audible thresholds for masking tones of 1, 4 and 8 kHz.

In Fig. 2.14, the average values of measurements taken on different subjects of the modified audibility threshold at various frequencies are reported. It should be noted that the presence of "holes" in the progression of the masking coefficient is due to the beats between the masking and the masked tone. In the measurements, for a more realistic estimation of the masking thresholds, instead of the pure tones, are used narrow band noise sources (notch noise). In general it can be said that using narrowband noise the following phenomena are observed:

- the coefficient M increases;
- the frequency range decreases;
- the trend of the L_m curve is smoother.

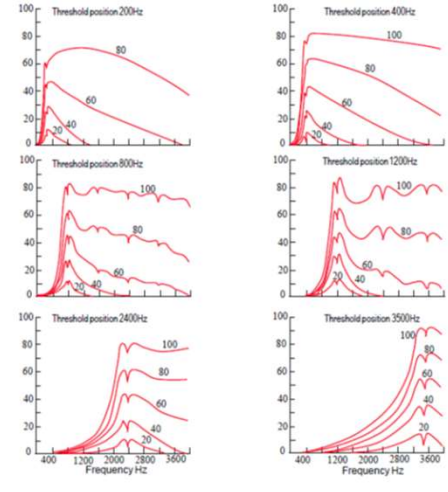


Fig. 2.14 Trend of the masking coefficient $M(f,p)$ for sinusoidal masking sounds of different intensity and frequency [14].

For example, in Fig. 2.15 two experiments are reported. In the first, 2.15-a), the audibility threshold is measured for a 400 Hz masking tone, with various intensities reported on the curves; in the second experiment, limited white noise is used in the band around 400 Hz (between the frequencies of 365 Hz and 455 Hz), Fig. 2.15-b).

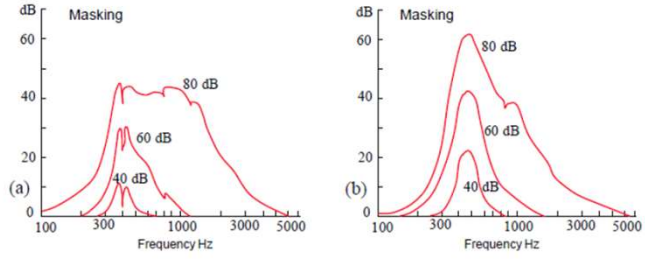


Fig. 2.15 Masking effect. a) For pure masking sound at 400Hz; b) white noise between 365Hz and 455Hz.

Perceptive and Physical Model of the Auditory System

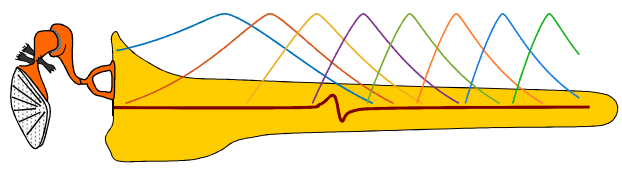
The energy of the acoustic wave, affects the tympanic membrane, and is transmitted through the ossicles (which act as an impedance adjustment device) to the oval window of the cochlea.

The basilar membrane, through the organ of Corti, excite small hair cells along.

Due to the simple mechanical resonance, a simple tone excites the hair cells at a particular point in the basilar membrane.

High frequency tones excite the hair cells near the oval window, while the low frequency tones excite the hair cells at the opposite end of the membrane.

So, one common method used by engineers to model the basilar membrane mechanism is to use a *critical band filter banks* (CBFB).



The concept of critical band is strongly related to the phenomenon of masking. The modification of the audibility threshold due to sinusoidal or narrow band noise, is closely linked to the way in which the hearing organ performs the spectral analysis of the acoustic signal.

The physical modeling of the hearing organ is a very complex matter; and various philosophies and models are available for this purpose in literature. In general terms, however, the models are realized in several stages. The first performs a spectral decomposition of the signal; the second, generally non-linear, represents a dynamic model of the ciliated cell.

The roex(*p*) Function Filter Bank

The determination of a filter-bank model of the auditory system, is the central point is to specify the physical model of the auditory apparatus and to simulate it numerically.

In particular in [20], for the determination of CBFB a family of functions called **rounded exponentials** called roex(*p*), turns out to be

$$W(g,p) = (1+pg)e^{-pg} \tag{2.6}$$

where p and g are parameters that depend on the frequency and are determined according to the following relationships. Said f the frequency and f_0 the central frequency in [Hz] of the filter, g represents the normalized frequency

$$g = \frac{|f - f_0|}{f} \tag{2.7}$$

while the p parameter depends on the ERB according to the relation

$$p(f) = \frac{4f}{\mathrm{ERB}(f)} = \frac{4f}{24.7 + 0.108f}.$$

Thus, p determines the bandwidth and slope from the filter curve. For a filter with an asymmetrical shape, the p parameter assumes different values: p_l for the low part of the frequency response and p_u for the high part. For example, in [24] the following values are proposed

$$p_u(f_0) = p(f_0)$$
$$p_l(f_0) = p(f_0)\left(1 - \frac{0.38}{p(1\text{kHz})}(L - 51\text{dB})\right) \tag{2.8}$$

where $\text{ERB}(f_0)$ represents the equivalent rectangular band (defined in (2.8)) and L represents the equivalent level of the input noise expressed in [dB/ERB].

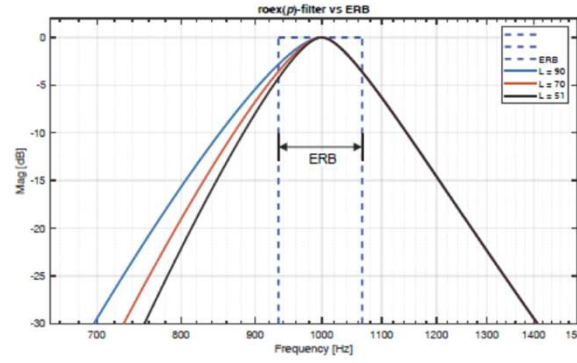


Fig. 2.16 Frequency responses of the roex(p) for filters L=51, 70, 90; and equivalent rectangular bandwidth (ERB). The dotted line represents the equivalent rectangular band.

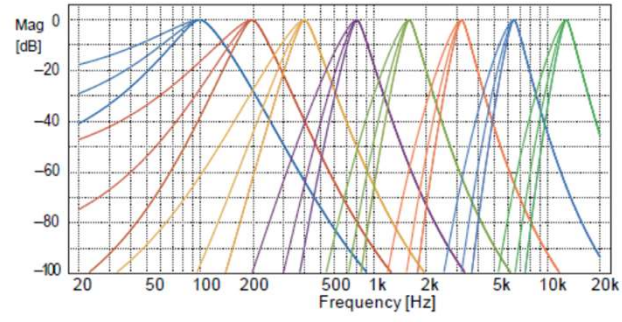


Fig. 2.17 Replies in frequency of the roex(p) filters calculated with the Eqn. (2.8) with f_0 = 100, 200, 400, 800, 1.6 kHz, 3.2 kHz, 6.4 kHz, 12.8 kHz; L = 50, 70, 90 [dB/ERB]

This type of filter has a symmetric amplitude response on a linear frequency scale for a level $L = 51$ dB/ERB at 1 kHz; for Eqn. (2.8) results, in fact, $p_l(f_0) = p_u(f_0 0) = p(f_0)$ (see Fig. 2.16).

For $L > 51$ and $L < 51$ [dB/ERB] we have, respectively, an increase or decrease of the slope from the curve to the low frequencies. On the other hand, the frequency response at the high frequency filter for $(f > f_0)$ is independent from the L level. Fig. 2.17 shows some amplitude response of roex(p) filters with different parameters.

Gammatone Filter Banks

In [19]-[22] Patterson et. al., proposed a model of psychoacoustic filter bank always based on the cochlear physical model, based on a particular filter bank called gammatone filter bank (GFB).
The GFB, is characterized by an impulsive response of the type

$$g(t) = at^{n-1}e^{-bt}\cos(2\pi f_0 t + \phi), \quad t > 0 \tag{2.9}$$

where a control the gain that is typically used to normalize the filter gain to unit, f_0 is the central frequency of the filter, the order n determines its slope, ϕ is the phase which is usually set to be 0, the term b represents a decay factor that determines the duration of the impulse response and thus the filter's equivalent rectangular bandwidth (see Eqn. (2.3)) that, for a given f_0, usually is computed as

$$b = 1.019 \cdot 24.7 \cdot (4.37 \cdot f_0/1000 + 1).$$

A fourth-order $n = 4$ gives the best fit with a wide range of human masking data [19]. Most auditory models use fourth-order GTFs, but lower or higher orders might be used in the sound-processing strategy in cochlear implants to optimize speech reception sco

Observe (see Fig. 2.19) that the amplitude response of the GFB for $n = 4$ is very similar to the amplitude response of the bank determined by the roex(p)function for $L = 51$ [dB/ERB].

Remark 2.3. Note that the gammatone filter bank model works well for auditory applications because the frequency response is "pseudo-resonant" similar to the dissonance curve of two pure sounds (see Fig. 2.9), and is relatively easy to match with measured responses.

The gammatone filters are particularly long, some efficient implementations of discrete-time gammatone filters can be found in [25] and [26].

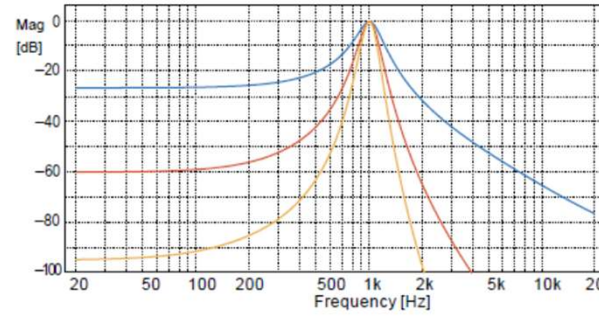


Fig. 2.18 Frequency responses of a gammatone filters for orders=2,4,6; for $F_s = 44.1$ kHz.

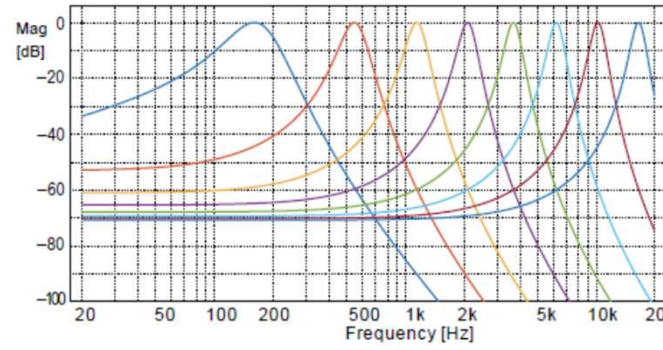


Fig. 2.19 Frequency responses of a typical gammatone filter-bank of 8-filter for $f_{min} = 80$ Hz, $f_{max} = 12$ kHz, order $= 4$, $\text{pad}_{bw} = 4$, $F_s = 44.1$ kHz.

2.3.7 Temporal Aspects of Masking

Real acoustic signals, as in speech and in music, are generally strongly time variations and for a more in-depth study we must also consider the dynamic effects of masking. This can be done by considering the cases in which masking and masked signals are not applied simultaneously. In fact, it is known that by applying an impulsive signal of a certain intensity at the end of the impulse, the ear remains "deaf" for a certain time interval.

Dynamic masking phenomena are generally detected by exciting the auditory apparatus with short reference signals dispensed in instants close to the beginning or end of a long-lasting masking sound. As a masking signal, is generally used a broadband noise of suitable duration and level, while a short suitably shaped sinusoidal tone (burst) is used as reference. When the masked (or reference) signal is applied later to the masking tone, the effect is defined as *forward masking*. On the contrary, when the reference tone preceding the masking tone is the effect is defined as *backward masking*[15]-[25]. Other masking effects are detected when the reference is placed at the beginning or end of a long-lasting pulse. These effects are referred to as *forward fringe masking* and *backward fringe masking* respectively. The various methods of detecting the phenomenon of dynamic masking are illustrated in Fig. 2.20.

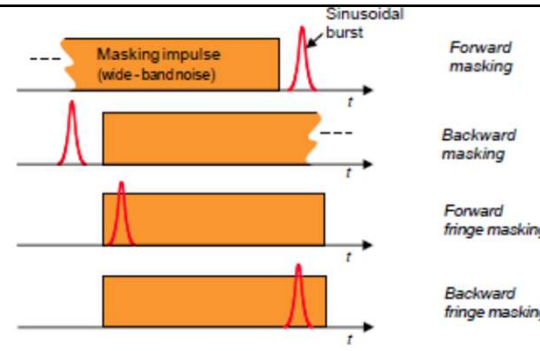


Fig. 2.20 Temporal succession of the signals to be dispensed to the subjects for the dynamic study of the temporel masking phenomenon.

The effects of the various types of masking are illustrated in Fig. 2.21. In this case the masking signal is a 70 dB SPL wideband noise, while the reference is a 1.9 kHz burst. The masking tone is administered continuously except for a certain interval or silence window lasting 25, 50, 200 or 500 msec [16]. The beginning of the silence window is considered as a reference to the time scale shown in the abscissa of Fig. 2.21. The four sets of points shown on the graph represent: as abscissa the instant of presentation of the reference burst; as ordered the threshold of audibility of the burst presented in the various instants.

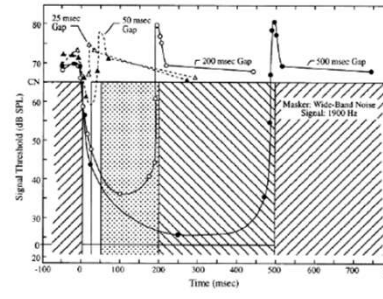


Fig. 2.21 Dynamic masking phenomenon with silence window of 25, 50, 200, 500 msec. The figure simultaneously illustrates the: forward, backward, forward fringe and backward fringe masking. (Courtesy of [16]).

The hearing threshold of the reference in the case of silence is shown with the symbol "0" in the figure (at about 23 dB SPL). In the case of continuous masking, i.e. without any window of silence, the threshold of audibility of the reference signal is indicated in the figure with the symbol "CN". The forward masking manifests itself as a gradual decrease in the level of the reference threshold that begins with the silence window and continues just beyond the first 100 msec. The effect disappears almost completely for later times. The effect of backward masking consists in a sudden increase in the threshold of audibility of the reference. This phenomenon precedes by a few tens of msec the reintroduction of the masking tone. The forward fringe masking is evidenced by the presence of a peak, of more than 10dB in the audibility threshold, which manifests itself immediately after the reintroduction of the masking signal. The backward fringe masking manifests itself as a small amplitude peak that precedes the cessation of the masking signal by a few tens of msec.

Remark 2.4. Note that, it is evident that the backward masking and the backward fringe masking are non-causal phenomena; i.e. the *effect*, that is, precedes the *cause* that generated it. These phenomena detected experimentally are probably related to the physiological characteristics of transmission of stimuli from the auditory apparatus to the brain. Moreover, a certain instability of these phenomena has been noted: in a subject in which the experiment is repeated several times the phenomenon of backward masking decreases [17].

The phenomenon of forward masking even if more robust (object of many past and present studies) consists in a very complex function of the stimulus parameters. There is not yet an accurate model that is able to correctly predict the masking phenomenon as a function of the parameters (frequency, intensity, shape, etc.) of masking and masked signals. Some studies on these aspects can be examined in more detail in [17]-[25]. Fig. 2.22 shows a time-frequency mask indicative of the phenomenon of masking.

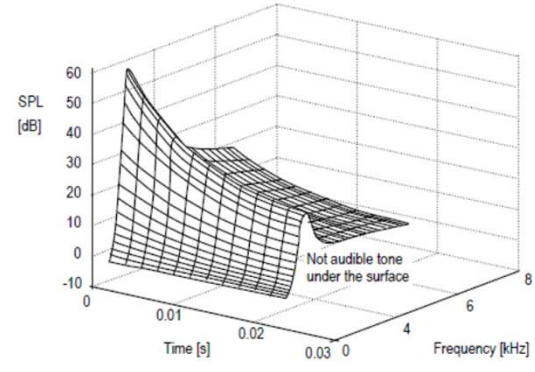


Fig. 2.22 Time-frequency surface of the masking phenomenon. The tone under the surface are not audible.

Remark 2.5. Note that, while currently the main psychoacoustic models used in the encoding of the audio signal are mainly based on energy considerations related to the signal in the critical bands. In recent years, studies of perceptive phenomena of the hearing apparatus, in particular of dynamic masking phenomena, have considerably increased.

As already mentioned above, the audio encoders used in modern acoustic transmission and diffusion channels are based on filter bank model and the space for further improvements is still considerable. There are, therefore, some areas of study and research aimed at the possible improvement of the coding of acoustic signals. Aspects such as the asymmetry of masking phenomena are not yet properly considered. In fact, it is known that noise produces a more significant masking effect compared to pure sounds: some encoders, for example, first make an empirical estimate of the characteristics of the signal and then use this measure to adjust the mask. Such methodologies have produced satisfactory results for a broad category of signals and it is possible that less empirical approaches based on more consistent psychoacoustic models produce further improvements and greater robustness.

A further area of improvement in making audio encoders is to consider the temporal aspects of masking. Modeling such phenomena could greatly improve the quality of compressed signals. In fact, in the scientific literature on the subject there are numerous data on this subject, but a consistent model is not yet available that can be introduced into an audio codec in a systematic way.