Acoustic Communication

Hearing and Speech

Torben Poulsen

tp@elektro.dtu.dk

Version 2.1.1
March 2008

Lecture Note 31230-05

DTU Electrical Engineering - Acoustic Technology

DTU Electrical Engineering, Acoustic Technology
Building 352, DTU
Technical University of Denmark
DK-2800 Lyngby
Denmark

E-mail: info-at@elektro.dtu.dk Web: www.elektro.dtu.dk

(+45) 45 25 39 30

Contents

1. Introduction

1.1 Background

The aim of the present lecture note is to give the student an understanding of the function of the ear, the perception of sound (psychoacoustics), of speech and speech intelligibility.

Comments and suggestions for improvement of the lecture note are very welcome.

1.2 Overview

The hearing sense can be illustrated schematically as in Figure 1-1 below.

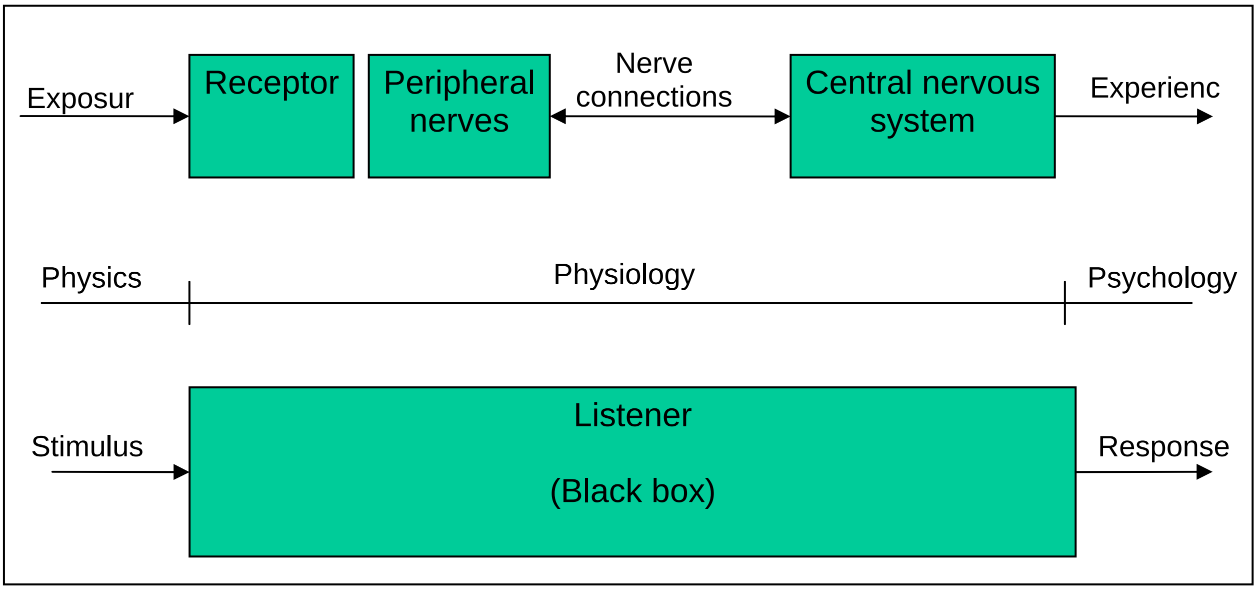


Figure 1-1 Schematic description of the hearing sense

In the upper part of the figure, the receptor receives the exposure and from the peripheral nerves, the signal is transmitted to the central nervous system via the nerve connections. This gives rise to an experience, the perception. The situation involves Physics (sound), Physiology (cochlea, nerves) and Psychology (the perception of the sound). In the lower part of the figure, the same situation is shown in a purely psychoacoustic way. Here we have a listener listening to a sound and the result is a response from the listener. The listener is considered a black box. The relation between stimulus and response is the main objective in psychoacoustics. The results from psychoacoustic investigations are used to investigate the details of the black box, i.e. the details of the upper part of the figure.

In Figure 1-2 an overview is given for various signals and their corresponding spectra. The figure is from (Zwicker and Fastl 1999).

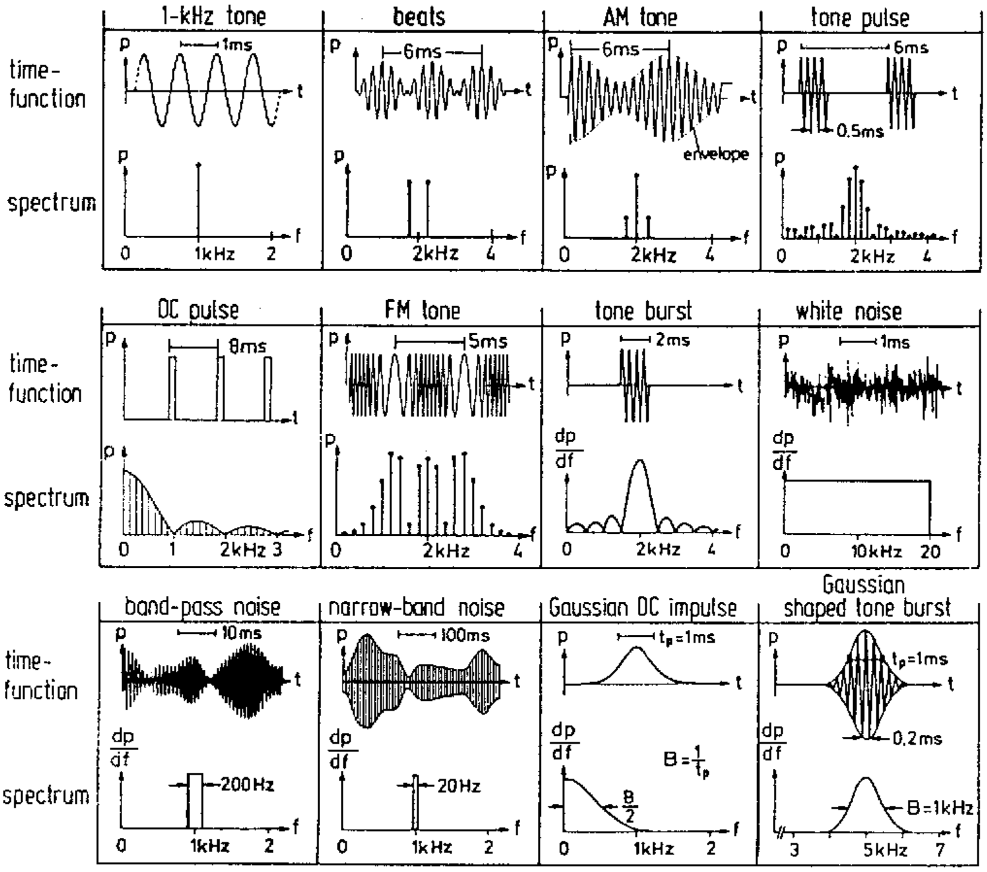


Figure 1-2. Overview of signals and corresponding spectra that are used in hearing research

2. The Ear

The ear can be divided into four main parts: The outer ear, the middle ear, the inner ear and the nerve connection to the brain. The first three parts (the peripheral parts) are shown in Figure 2-1.

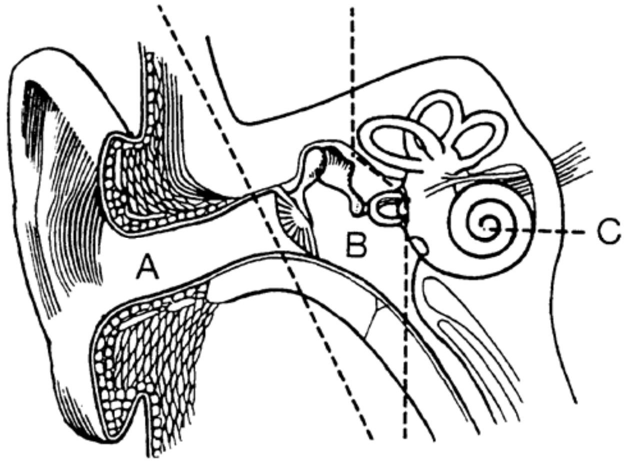


Figure 2-1. Drawing of the ear. A is the outer ear. B is the middle ear. C is the inner ear. From (Hougaard et al. 1995)

Part A being the outer ear, B is the middle ear and C is the inner ear. The sound will reach the outer ear, progress through the outer ear canal, reach the tympanic membrane (the eardrum), transmit the movements to the bones in the middle ear, and further transmit the movements to the fluid in the inner ear. The fluid movements will then be transformed to nerve impulses in the inner ear and the impulses are transmitted to the brain through the auditory nerve.

2.1 The outer ear

The outer ear consists of the pinna (or the auricle) and the ear canal. The Pinna plays an important role for our localization of sounds sources. The special shape of pinna produces reflections and diffraction so that the signal that reaches the ear will be dependent on the direction to the sound. The pinna has common features from person to person but there are big individual differences in the details. Localisation of sound sources is difficult if a hearing protector or a crash helmet covers the pinna. The outer part of the ear canal is relatively soft whereas the inner part is stiff and bony. At the end of the ear canal the tympanic membrane is situated. The length of the ear canal is approximately 25 mm and the diameter is approx. 7 mm. The area is approx. 1 cm2. These numbers are approximate and vary from person to person. (Compare the area 1 cm2 with the calculation: pi*3.5^2 = 38.5 mm^2).

The ear canal may be looked upon as a tube that is closed in one end and open in the other. This will give resonances for frequencies where the length of the ear canal corresponds to 1/4 of the wavelength of the sound. With a length of 25 mm and a speed of sound of 340 m/s the resonance frequency will be

$$f_{res} = 340 / (4 * 0.025) = 3.4 \text{ kHz}$$

This calculation is correct if the ear canal was a cylindrical tube with rigid walls. The tympanic membrane is not perpendicular on the axis of the canal and most ear canals will have one or two bends. The bends implies that it is usually not possible to see the tympanic membrane at the end of the ear canal. It is necessary to make the canal straighter, which may be done by pulling pinna upward and backwards.

The tympanic membrane is found at the end of the canal. The membrane is not perpendicular to the axis of the ear canal but tilted approx. 30 degrees. The tympanic membrane is shaped like a cone with the top of the cone pointing inwards into the middle ear. The thickness is approx. 0.1 mm.

2.2 The middle ear

The middle ear consists of three small bones: hammer, anvil and stirrup. Often, the Latin names are also used: Malleus, Incus and Stapes. These bones are the smallest bones in the human body. Figure 2-2 shows a drawing of the middle ear. The function of the middle ear is to transmit the vibrations of the tympanic membrane to the fluid in the inner ear. From Figure 2-2 it is seen that the hammer, M, is fixed to the tympanic membrane (1) from the edge and into the centre of the membrane (the top of the cone). The anvil, (Incus, I, 2) connects the hammer and the stirrup (Stapes, S) and the footplate of the stirrup makes the connection into the inner ear. Sometimes this connection is called the oval window. The footplate rotates around the point marked (3). The middle ear is filled with air and is connected to the nose cavity (and thus the atmospheric pressure) through The Eustachian tube (ET, 4). The fluid in the inner ear is incompressible and an inwards movement of the stirrup will be equalised by a corresponding outward movement by the round window (RW).

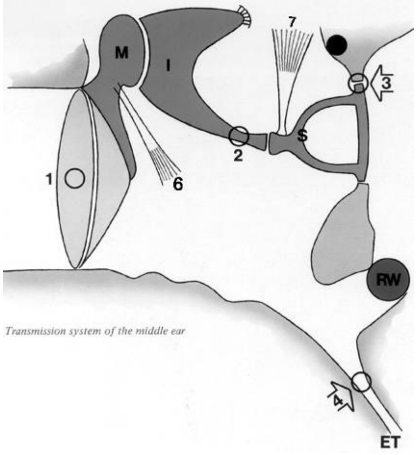


Figure 2-2. Drawing of the middle ear. See text for details. From (Engström and Engström 1979)

Usually the Eustachian tube is closed but opens up when you swallow or yawn. When the tube is open, the pressure at the two sides of the tympanic membrane is equalised. If the Eustachian tube becomes blocked (which is typically the case when you catch a cold) the equalisation will not take place and after some time the oxygen in the middle ear will be assimilated by the tissue and an under-pressure will build up in the middle ear. This causes the tympanic membrane to be pressed inwards and the sensitivity of the hearing is reduced.

The chain of middle ear bones forms a lever function that - together with the area ratio between the tympanic membrane and the footplate of stapes - makes an impedance match between the air in the outer ear and the liquid in the inner ear. The lever ratio is approx. 1.3 and the area ratio is approx. 14. The total ratio is thus 18, which corresponds to approx. 25 dB.

Two small muscles, tensor tympani (6) and stapedius (7), see Figure 2-2, are attached to the bones and will be activated by the so-called middle ear reflex. The reflex is elicited when the ear is exposed to sounds above approx. 70 dB SPL whereby the transmission through the middle ear is reduced. The reduction is about 20 dB at 125 Hz, 10 dB at 1000 Hz and less than 5 dB at frequencies above 2000 Hz. To a limited extent, the middle ear reflex can protect the inner ear from excessive exposure. Because the reflex is activated by a signal from the brain, there will be a delay of about 25 to 150 ms before the effect is active. The reflex has therefore no protective effect on impulsive sounds.

The transmission through the middle ear looks like a band pass filter with a maximum around 800 Hz and declines towards higher and lower frequencies.

2.3 The inner ear

The inner ear consists of a snail-shell shaped structure in the temporal bone called Cochlea. The cochlea is filled with lymph and is closely connected to the balance organ that contains the three semicircular canals. There are 2.75 turns in the snail shell and the total length from the base to the top is 32 mm. A cross section of one of the turns is shown in Figure 2-3. This figure shows that the cochlea is divided into three channels (Latin: Scala) called scala vestibuli (1), scala tympani (3) and scala media (2).

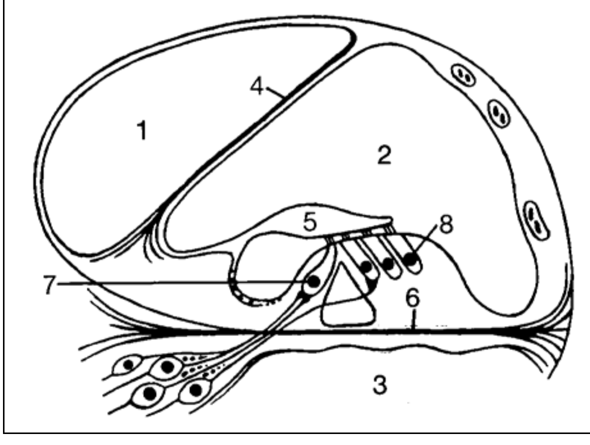


Figure 2-3. Cross section of a cochlea turn. See text for details. From (Hougaard et al. 1995)

There are two connections (windows) from cochlea to the middle ear cavity. The oval window is the footplate of the stirrup and is connected to Scala Vestibuli (1). The round window is connected to Scala Tympani (3). The round window prevents an over-pressure to build up when the oval window moves inwards. Scala Vestibuli and Scala Tympani are connected at the top of the cochlea with a hole called Helicotrema.

The Basilar membrane (6 in Figure 2-3) divides scala tympani from scala media. The width of the basilar membrane (BM) changes from about 0.1 mm at the base of the cochlea to about 0.5 mm at the top of the cochlea (at helicotrema). The change of the BM-width is thus the opposite of the width of the snail shell. The function of the BM is very important for the understanding of the function of the ear.

A structure - called the organ of Corti - is positioned on top of the Basilar Membrane in Scala Media. The organ of Corti consists of one row of inner hair cells (7 in Figure 2-3) and three rows of outer hair cells (8 in Figure 2-3). The designations 'inner' and 'outer' refer to the centre axis of the snail shell which is to the left in Figure 2-3. The hair cells are special nerve cells where small hairs protrude from the top of the cells. There are approx. 3000 inner hair cells and about 12000 outer hair cells. A soft membrane (5 in Figure 2-3) covers the top of the hair cells. The organ of Corti transforms the movements of the Basilar membrane to nerve impulses that are transmitted to the hearing centre in the brain.

The inner hair cells are the main sensory cells. Most of the nerve fibres are connected to the inner hair cells. When sound is applied to the ear, the basilar membrane and the organ of Corti will vibrate and the hairs on the top of the hair cells will bend back and forth. This will trigger the (inner) hair cells to produce nerve impulses.

The outer hair cells contain muscle tissue and these cells will amplify the vibration of the basilar membrane when the ear is exposed to weak sounds so that the vibrations are big enough for the inner hair cells to react. The amplification function of the outer hair cells is nonlinear which means that they have an important effect at low sound levels whereas they are of almost no importance at high sound levels. The amplifier function - sometimes called the cochlear amplifier - is destroyed if the ear is exposed to loud sounds such as gunshots or heavy industrial noise. This is called a noise induced hearing loss. The amplifier function also deteriorates with age. This is called an age related hearing loss.

2.4 The central auditory pathway

From the cochlea the nerve signals are transmitted via the auditory nerve (the eight nerve) through the central auditory pathway to the auditory cortex. In principle, the signals from the right cochlea are transmitted to the left auditory cortex and the signals from the left cochlea are transmitted to the right auditory cortex. However, several connections exist between the right and left sides where 'crosstalk' can take place. The pathway consists of several 'relay stations', nuclei, where the signal is transmitted from one nerve cell to the next. The nuclei are situated in the brain stem. This is illustrated in Figure 2-4.

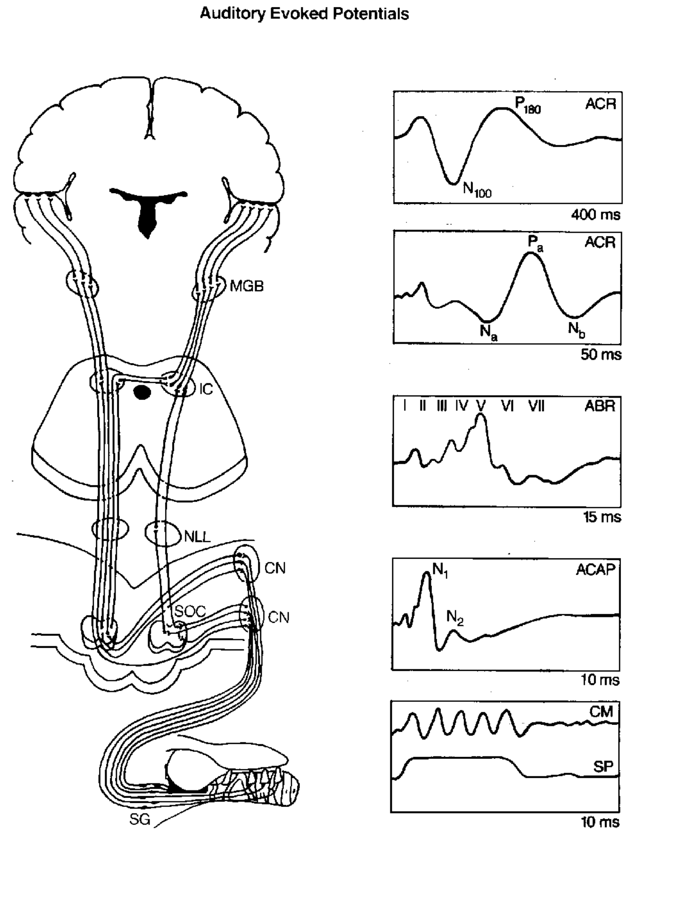


Figure 2-4. The central auditory pathway (Elberling & Osterhammel, 1989). Note that only the connections from the right cochlea are shown.

The relay stations are: CN: Cochlear Nucleus (ventral and dorsal), SOC: Superior Olivary Complex, NLL: Lateral Lemniscus, IC: Inferior Colliculus, MGB: Medial Giniculate Body. The right side of the figure shows the various signals that can be recorded by means of electric response audiometry, ERA. The abbreviations here are SP: Summating potential, CM: Cochlear microphonic, ACAP: Auditory compound action potential, ABR: Auditory brain stem response, ACR: Auditory cortical response.

Figure 2-4 shows only half of the system, i.e. the connections coming from the right cochlea. There are similar connections from the left cochlea. Furthermore, the figure only shows the ascending system where the information is transmitted from the cochlea to the brain. This is called the afferent system. There is also another system where the information goes in the opposite direction, from the brain to the cochlea. This is called the efferent system.

2.5 The BM frequency analyser

The basilar membrane acts like a frequency analyser. When the ear is exposed to a pure tone, the movement of the basilar membrane will show a certain pattern and the pattern is connected to a certain position on the basilar membrane. If the frequency is changed, the pattern will *not* change but the position of the pattern will move along the basilar membrane. This is illustrated in Figure 2-5 for the frequencies 400 Hz, 1600 Hz and 6400 Hz. The 400 Hz component produces BM-movement close to the top of the cochlea. 6400 Hz produces a similar pattern but close to the base of the cochlea. Note that a single frequency produces movements of the basilar membrane over a broad area. This means that even for a single frequency many hair cells are active at the same time. Note also that the deflection of the BM is asymmetrical. The envelope of the deflection (shown dotted in Figure 2-5) has a steep slope towards the low frequency side and a much less steep slope towards the high frequency side. The same different slopes are also found in masking thresholds and it can be shown that masking is closely related to the basilar membrane movements.

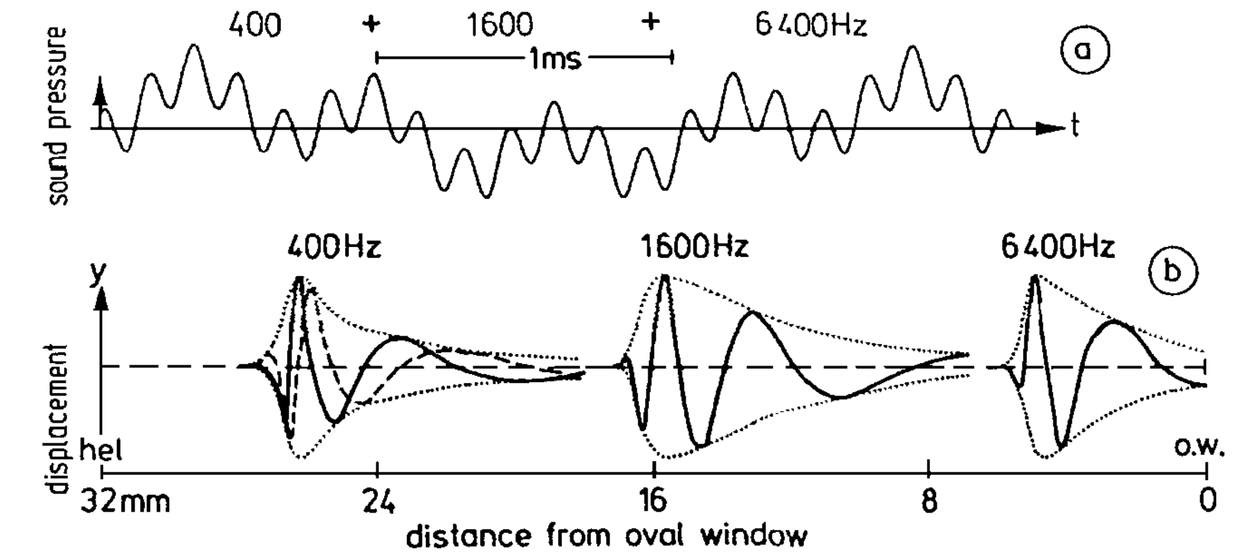


Figure 2-5. Movement of the basilar membrane (b) when the ear is exposed to a combination of 400 Hz, 1600 Hz and 6400 Hz (a). O.W.: Oval window (base of cochlea). Hel: Helicotrema (top of cochlea). From (Zwicker and Fastl 1999)

Hint: 20 mm from the Oval Window corresponds roughly to the frequency 1 kHz.

The non-linear behaviour of the outer hair cells and their influence on the BM movement is illustrated in Figure 2-6. This figure shows the BM-amplitude at a certain position of the basilar membrane as a function of the stimulus frequency. (Note that this is different from Figure 2-5 where the amplitude is shown as a function of basilar membrane position for different frequencies). There are at least three nonlinear phenomena illustrated in the figure.

1) At low exposure levels (20 dB) the amplitude is very selective and the 'high' amplitude is achieved only in a very narrow frequency band. For high exposure levels (80 dB) the 'high' amplitude is achieved at a much wider frequency range. Thus, the filter bandwidth of the auditory analyser changes with the level of the incoming sound.

2) The frequency where the maximum amplitude is found, change with level. At a high level, the frequency is almost one octave below the max-amplitude frequency at low levels.

3) The maximum amplitude grows non-linearly with level. At low levels (20 dB) the maximum BM-amplitude is about 60 dB (with some arbitrary reference). At an input level of 80 dB the maximum BM amplitude is about 85 dB. In other words the change in the outside level from 20 dB to 80 dB, i.e., 60 dB, is reduced (compressed) to a change in the maximum BM-amplitude of

only 25 dB.

These non-linear phenomena are caused by the function of the outer hair cells. The increase of amplitude at low levels is sometimes called ‘the cochlear amplifier’. In a typical cochlear hearing loss, the outer hair cells do not function correctly or they are destroyed. In other words: The cochlear amplifier does not work. This will be seen as an elevated hearing threshold and is called a hearing loss.

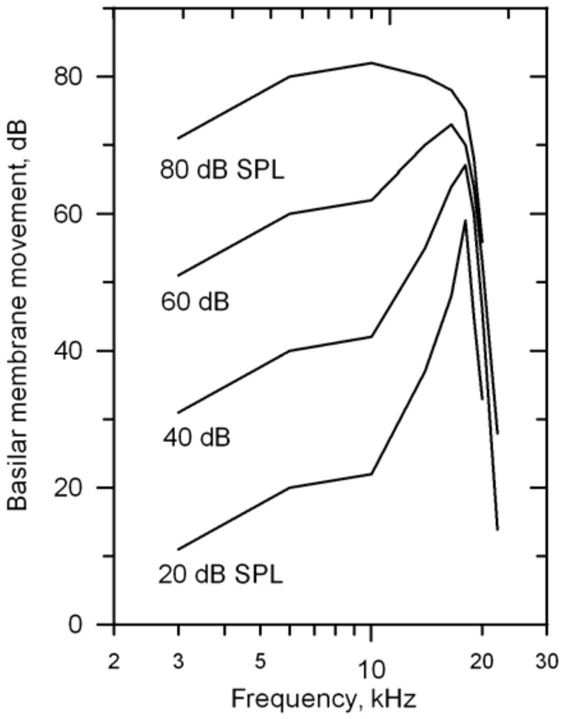


Figure 2-6. Movement of the Basilar membrane at a fixed point for stimulus levels from 20 dB SPL to 80 dB SPL. Redrawn from (Kemp 1988)

The function of the cochlear amplifier is illustrated in Figure 2-7. The acoustic stimulus makes the Basilar membrane vibrate and the inner hair cells will be activated and send neural impulses towards the brain. The basilar membrane motion will cause the hairs (stereocilia) on the outer hair cells to bend and thus modulate the current flow in the outer hair cells. This current flow makes the outer hair cell vibrate and thereby enhance the motion of the basilar membrane.

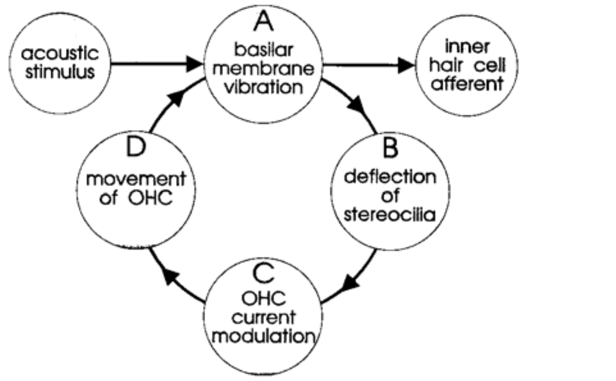


Figure 2-7 Schematic representation of the cochlear amplifier. If the loop is broken at any point, the Basilar membrane is driven only by the input acoustic stimulus. The inner hair cells play no part in the amplification as they are passive motion detectors.

From Moore: Hearing, p. 46.

3. Human hearing

The human hearing can handle a wide range of frequencies and sound pressure levels. The weakest audible sound level is called the hearing threshold and the sound level of the loudest sound is called the threshold of discomfort or the threshold of pain.

3.1 The hearing threshold

The hearing threshold is usually given as a certain level. Below that level, nothing is heard and above the level, everything is heard. In Figure 3-1 this classical interpretation is illustrated by the solid curve for an example where the threshold is 12,5 dB. However, this is not the situation in real life. For presentation levels close to the threshold the probability of getting a ‘yes I heard the tone’ response will vary according to the S-shaped dashed curve. This curve is called a psychometric function. For high levels, all presentations will be heard (100% yes-responses). For very low level, no presentations will be heard (0% yes-responses). In the transition-range around the threshold, the number of yes-responses will be within 100% and 0%. The S-curve can be modelled as a cumulative normal distribution or (more simple) as a logistic function.

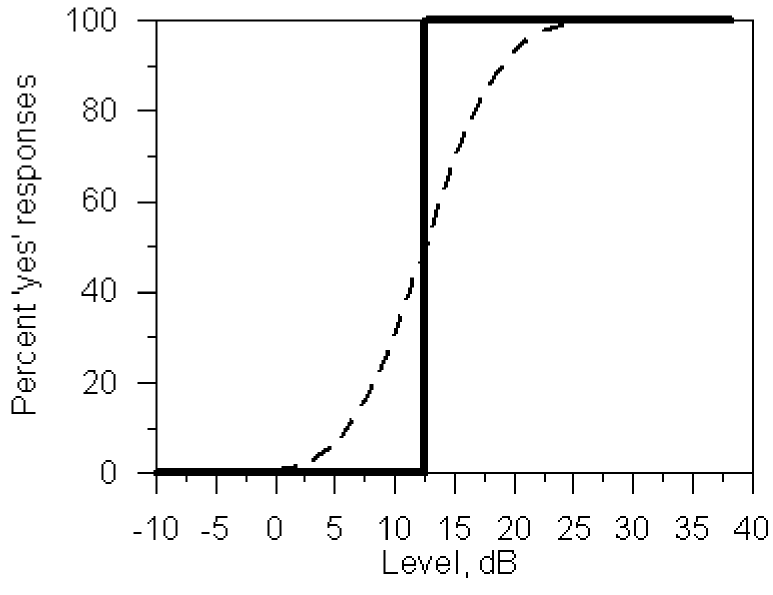


Figure 3-1. Heavy, solid line describes the classical interpretation of the threshold. The S-shaped curve indicates the real situation. This curve is called a psychometric function.

A simple approximation to a cumulative normal distribution is given by

$$P(x) = 1 - ½ (1 + C_1x + C_2x^2 + C_3x^3 + C_4x^4)^{-4}$$

where $C_1 = 0{,}196854$; $C_2 = 0{,}115194$; $C_3 = 0{,}000344$; $C_4 = 0{,}019527$

The error of this approximation will be less then $2{,}5 \bullet 10^{-4}$

The logistic function is given by

$$P(x) = 1 - \frac{1}{1 + e^{x/a}}$$

If the slope parameter 'a' is set to 1, a simple expression is obtained that can be used to describe experimental data. A good approximation to the cumulative normal distribution is obtained if the 'a' parameter is set to 0,569.

The result of a threshold determination is shown in Figure 3-2. The percentage of positive responses from repeated presentations (at the same level) has been calculated and plotted in the figure.

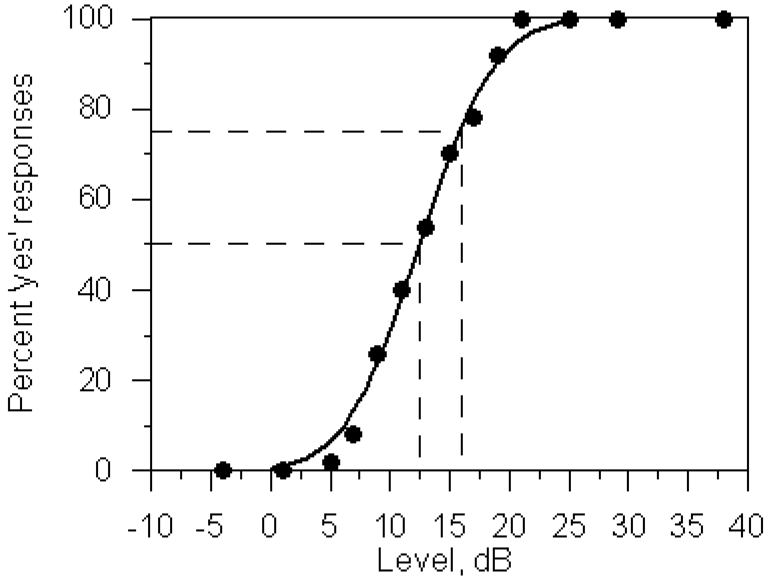


Figure 3-2. The result of a threshold determination.

For numerical reasons it is impractical to use the curve directly as the result. The threshold must therefore be defined as one of the points on the curve. Often the 50% point is used as the threshold (12,5 dB in this example) where half of the presentations will be heard. Remember that this is a matter of definition; any other point on the psychometric function could be just as good. In some cases the 75% point is used in order to obtain a more 'certain' threshold.

The hearing threshold is frequency dependent, see Figure 3-3. At 1000 Hz the threshold is about 2 dB SPL whereas it is about 25 dB SPL at 100 Hz and about 15 dB at 10000 Hz.

The threshold curve in Figure 3-3 is measured under the following (somewhat special) conditions:

- Frontally incoming sound (called frontal incidence)
- Signals are single pure tones (one frequency at a time)
- Binaural listening (i.e. listening with both ears)
- No reflections (i.e. an anechoic room)
- No background noise (i.e. an anechoic room)
- Subjects between 18 and 25 years of age with normal hearing
- The threshold determined by means of either the ascending or the bracketing method.

The curve is the median value (not the mean) over the subject's data. The sound pressure level, which is shown in the figure, is the level in the room at the position of the test subject's head but measured *without* the presence of the test subject. This curve is also called the absolute threshold (in a free field) and data for the curve may be found in ISO 389-7 (ISO-389-7 1996) and in ISO 226 (ISO-226 2003).

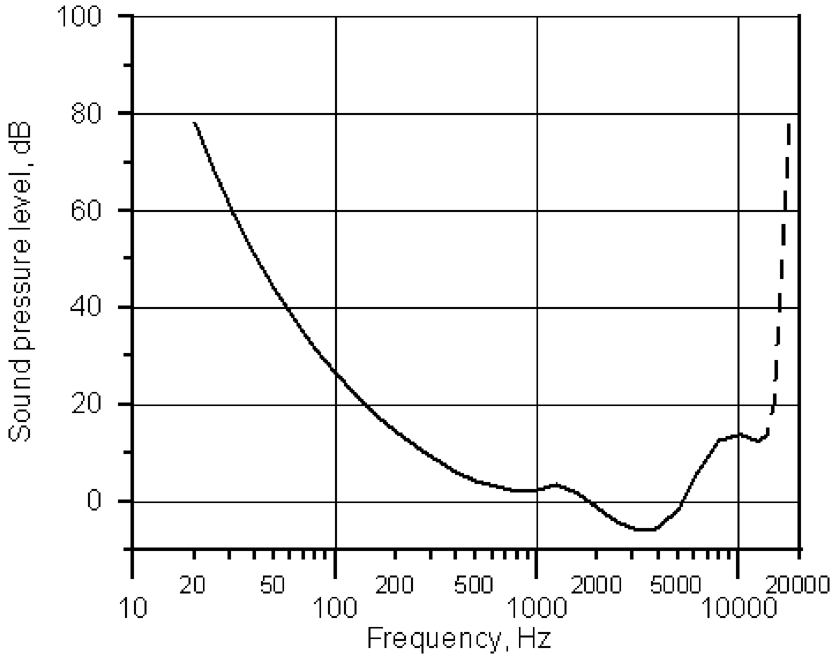


Figure 3-3. The binaural hearing threshold in a free field. Redrawn from (ISO-226 2003)

In ISO 389-7 also threshold data for narrow band noise in a diffuse sound field are found. This threshold curve is similar to the curve in Figure 3-3 and deviates from the pure tone curve only by a few dB (-2 to +6) in the frequency range 500 Hz to 16 kHz.

3.2 Audiogram

For practical use it is not convenient to measure the hearing threshold in a free or a diffuse sound field in the way described in the previous section. For practical and clinical purposes, usually only the deviation from normal hearing is of interest. Such deviations are determined by means of headphones connected to a calibrated audiometer. The measurement is called audiometry and result of the measurement is called an audiogram.

Figure 3-4 shows an audiogram for a person in the frequency range 125 Hz to 8000 Hz. The zero line indicates the average threshold for young persons and a normal audiogram will give data points within 10 to 15 dB from the zero line. An elevated hearing threshold (i.e. a hearing loss) is indicated downwards in an audiogram and the values are given in dB HL. The term 'HL' (hearing level) is used to emphasize that the curve show the *deviation* from the average normal hearing threshold.

For many years the Telephonics TDH 39 earphone has been (almost) the only earphone used for audiometry. This is because standardized calibration values for the earphone are necessary in order to make correct measurements. Such calibration values do not exist for ordinary earphones used for e.g. music listening. Around year 2000, calibration values for a new earphone, Sennheiser HDA 200, became available. The calibration values for HDA 200 cover the frequency range from 125 Hz to 16 kHz (see ISO 389-5, ISO 389-8)

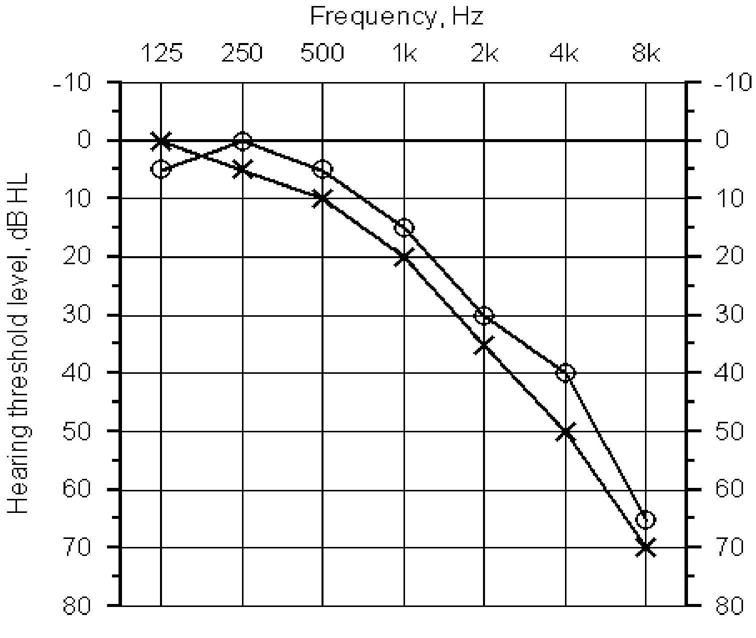


Figure 3-4. Audiogram for a typical age-related hearing loss.

(Note that the aspect ratio is not correct in this figure. Usually the following is asked for: the size of one octave is equal to the size of 20 dB)

Audiometry is performed with headphones for each ear separately. The results from the left ear are shown with 'X' and the results from the right ear are shown with 'o'.

In order for the audiometry to give correct results, the audiometer must be calibrated according to the ISO 389 series of standards. These standards specify the sound pressure level in dB (SPL) that shall be measured in a specific coupler (an artificial ear) when the audiometer is set to 0 dB HL. The values in the standards are earphone specific, which means that the audiometer must be recalibrated if the earphone is exchanged with another earphone.
Table 3-1 shows reference values for two earphones commonly used in audiometry.

F, Hz	125	250	500	1k	2k	3k	4k	6k	8k	10k	12,5k	14k	16k
TDH 39	45,0	25,5	11,5	7,0	9,0	10,0	9,5	15,5	13,0	-	-	-	-
HDA 200	30,5	18,0	11,0	5,5	4,5	2,5	9,5	17,0	17,5	22,0	28,0	36,0	56,0

Table 3-1. Calibration values in dB SPL for a Telephonics TDH 39 earphone and a Sennheiser HDA 200 earphone. The TDH 39 earphone can not be used above 8 kHz. The TDH 39 data are from ISO 389-1 (ISO-389-1 1991). The HDA 200 data are from ISO 389-5 (ISO-389-5 2006) and ISO 389-8 (ISO-389-8 2001).

Note that sound pressure level, dB SPL, and hearing level, dB HL, is not the same. An example: From Figure 3-3 it can be seen that the hearing threshold at 125 Hz is 22 dB SPL (measured in the way described previously). If a person has a hearing loss of 5 dB HL at this frequency the threshold would be 27 dB SPL. In an audiogram the 5 dB hearing loss will be shown as a point 5 dB below the zero line (as seen for the right ear, Figure 3-4). Another example: At 4000 Hz the free field threshold is −6 dB (see Figure 3-3). A hearing loss of 50 dB HL (e.g. left ear, Figure 3-4) will give a threshold of 44 dB SPL.

The test-retest uncertainty in an audiogram is about 5 dB when a clinical threshold determination method is used. This means that a repetition of the threshold determination could give results that may go up or down by 5 dB compared to the first measurement. This should be kept in mind e.g. when the audiogram is used to set the amplification of a hearing aid, called ‘hearing aid fitting’.

Screening is a simplified hearing test where pure tones at the audiometric frequencies are presented at e.g. 20 dB above the normal hearing threshold. If the test person can hear these tones no hearing loss seems to be present. If one or more of the tones are not detected, the hearing threshold will be determined at these frequencies. Screening is a fast method to detect a hearing impaired person, but because of the uncertainty of the test real hearing losses may not be detected – and normal hearing persons may be called hearing impaired. This is illustrated in Table 3-2

No hearing loss		Hearing loss	
RESULT OF SCREENING TEST			
Negative result	Positive	Positive	Negative
True	False positive	True	False negative

Table 3-2 Result of a screening test. The top row shows the real distribution (in a population) of hearing losses. Only a few have hearing loss. The screening test may be negative (all frequencies were heard, the test for hearing loss was negative) or positive (some frequencies were not head, the test for hearing loss was positive). The riskof the test lies in the false positive and the false negative.

3.3 Loudness Level

The definition of loudness level is as follows: For a given sound, A, the loudness level is defined as the sound pressure level (SPL) of a 1000-Hz tone which is perceived equally loud as sound A. The unit for loudness level is Phon (or Phone). In order to measure loudness level a 1 kHz tone is needed and this tone should then be adjusted up and down in level until it is perceived just as loud as the other sound. When this situation is achieved, the sound pressure level of the 1 kHz tone is per definition equal to the loudness level in phone. For a 1000-Hz tone the value in dB SPL and in Phone will be the same.

The loudness level for pure tones has been measured for a great number of persons with normal hearing. The result is shown in Figure 3-5.

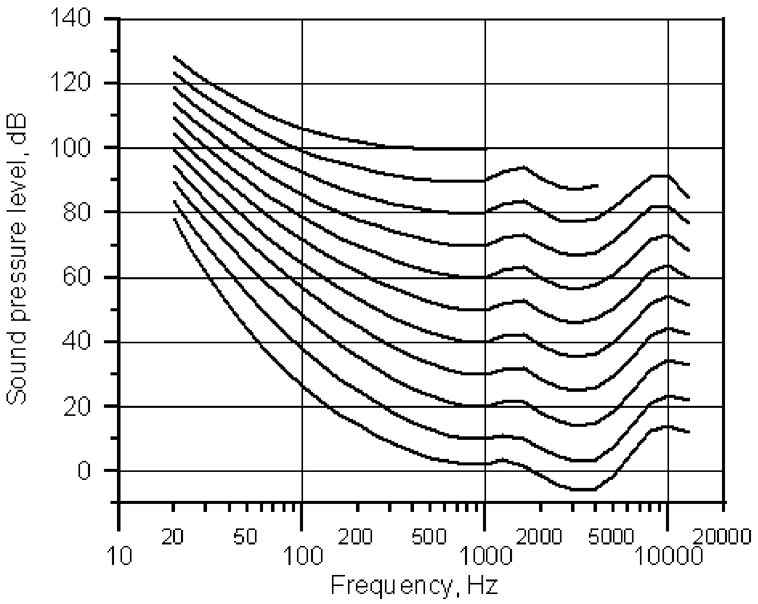


Figure 3-5. Equal loudness level contours. Redrawn from (ISO-226 2003)

Some examples - see Figure 3-5: A 4000-Hz tone at 26 dB SPL will be perceived with the same loudness as a 1000-Hz tone at 30 dB SPL and thus the loudness level of the 4000 Hz tone is 30 Phone. A 125-Hz tone at 90 dB SPL will have a loudness level of 80 Phone.

The curves in Figure 3-5 are − in principle − valid only for the special measurement situation where the tones are presented one at a time. They should not be used directly to predict the perception of more complicated signals such as music and speech because the curves do not reflect the effect of masking and temporal matters. Reflections in a room are not taken into account either.

Translations of Loudness Level:

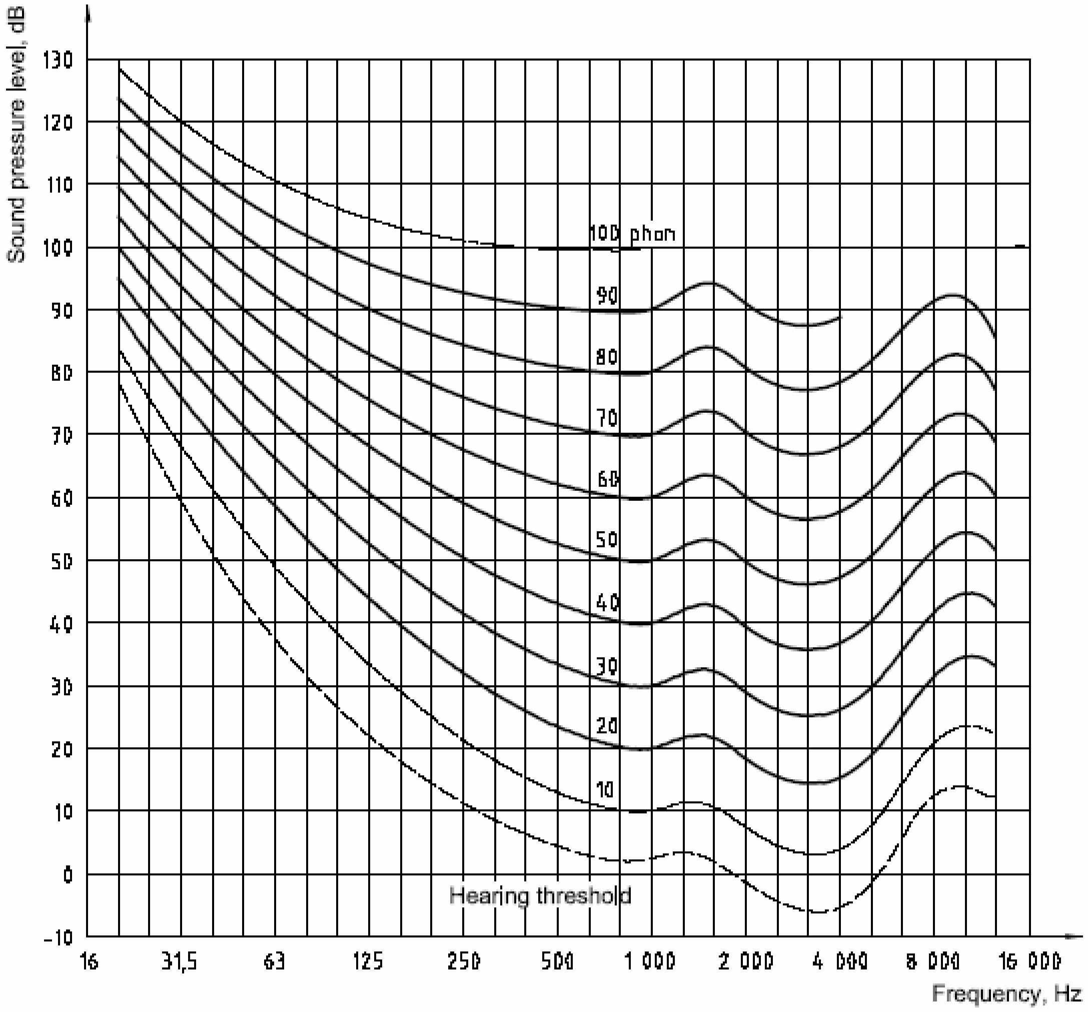
Danish:	Hørestyrkeniveau (enhed: Phon)
German:	Lautstärkepegel (Einheit: Phon)
French:	Niveau de Sonie.

3.4 ISO 226

The ISO standard ISO 226: ’Acoustics- Normal equal-loudness-level contours’ have a formula for deriving sound pressure level from loudness level and also a formula for deriving loudness level from sound pressure level. The data are also given in tables and in graphical form. See Figure 3-6. The data in the standard are based on twelve independent experimental investigations performed under the same experimental conditions. Annex C of the standard gives a short description of the measurement details and the psychoacoustical principles used in the investigations.

Annex A
(normative)

Normal equal-loudness-level contours for pure tones under free-field listening conditions



NOTE 1 The hearing threshold under free-field listening condition, T_f, is indicated by a dashed line.

NOTE 2 The contour at 10 phon is drawn by dotted lines because of the lack of experimental data between 20 phon and the hearing thresholds. Moreover, the 100-phon contour is also described by a dotted line because data from only one institute are available at this loudness level.

Figure A.1 — Normal equal-loudness-level contours for pure tones
(binaural free-field listening, frontal incidence)

Figure 3-6. The original figure from ISO 226 (2003)

4. Masking

The term ‘masking’ is used about the phenomenon that the presence of a given sound (sound A) can make another sound (sound B) inaudible. In other words A masks B or B is masked by A. Masking is a very common phenomenon which is experienced almost every day, e.g. when you need to turn down the radio in order to be able to use the telephone.

The situation described above is also called simultaneous masking because both the masking signal and the masked signal are present at the same time. This is not the case in backward and forward masking. (Backward and forward refer to time). Simultaneous masking is best described in the frequency domain and is closely related to the movements of the Basilar membrane in the inner ear.

The masking phenomenon is usually investigated by determining the hearing threshold for a pure tone when various masking signals are present. The threshold determined in this situation is called the masked threshold contrary to the absolute threshold.

4.1 Complete masking

If the ear is exposed to white noise, the hearing threshold (i.e. masked threshold) will be as shown in Figure 4-1 where also the absolute threshold is shown. The masked threshold is shown for different levels of the white noise.

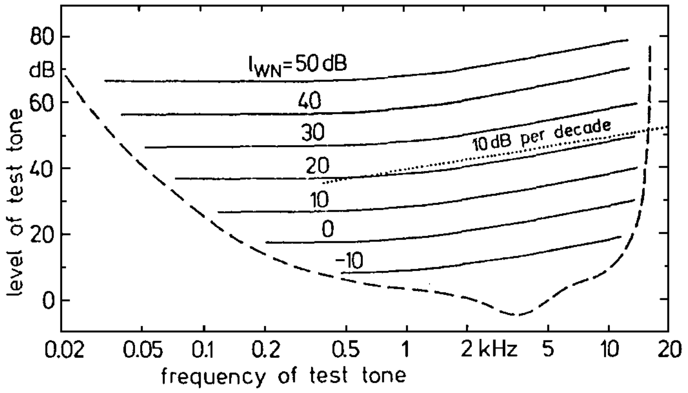


Figure 4-1. Masking from white noise. The curves show the masked threshold for different spectrum levels of white noise. From (Zwicker and Fastl 1999)

The masked thresholds are almost independent of frequency up to about 500 Hz. Above 500 Hz the threshold increases by about 10 dB per decade (= 3 dB/octave). A 10-dB change in the level of the noise will also change the masked threshold by 10 dB.

If a narrow band signal is used instead of the white noise, the masked threshold will be as shown in Figure 4-2. Here the masked threshold is shown for a narrow band signal centred at 250 Hz, 1 kHz and 4 kHz respectively. Generally, the masking curves have steep slopes (about 100 dB/octave) towards the low frequency side and less steep slopes (about 60 dB/octave) towards the high frequency side.

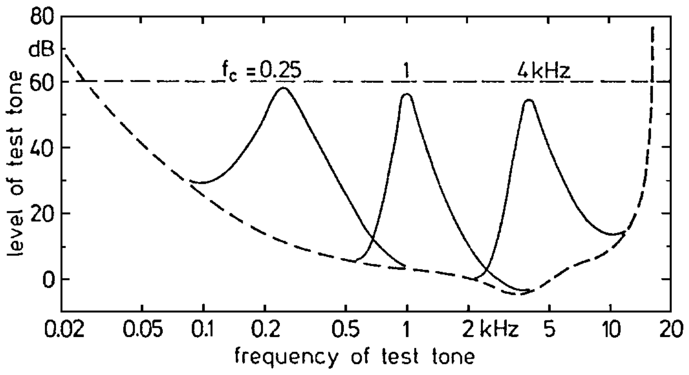


Figure 4-2. Masking from narrow band noise. The curves show the masked threshold when the ear is exposed to narrow band noise (critical band noise) at 250 Hz, 1 kHz and 4 kHz respectively. From (Zwicker and Fastl 1999)

The masking curves for narrow band noise are very level dependent. This is illustrated in Figure 4-3. The slope at the low frequency side is almost independent of level but the slope at the high frequency side depends strongly on the level of the narrow band noise. The dotted lines near the top of the curves indicate experimental difficulties due to interference between the noise itself and the pure tone used to determine the masked threshold.

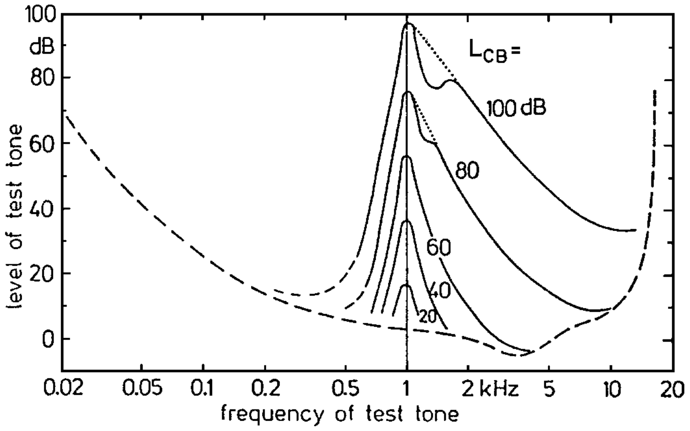


Figure 4-3. The influence of level on the masked threshold. The slope towards higher frequencies decreases with increasing level, i.e. masking increases non-linearly with level. From (Zwicker and Fastl 1999)

The masked threshold for narrow band noise is mainly caused by the basilar membrane motion. The different slopes towards the low and the high frequency side are also seen here and also the nonlinear level dependency is seen. Compare with Figure 2-5.

A model of simultaneous masking is found in Buus (1997)

4.2 Partial masking

The term 'Complete masking' is used when the presence of a given sound (sound A) can make another sound (sound B) inaudible. Partial masking is a situation where sound A influences the perception of sound B even though sound B is still audible. The influence is mainly seen in the loudness of sound B.

An example: When you listen to a standard car-radio while you are driving at, e.g. 100 km/h, you will adjust the level of the radio to a comfortable level. There will be some background noise from the engine, the tires, and the wind around the car (at least in ordinary cars). Then, when you come to a crossing or a traffic light and have to stop you will hear that the radio-volume is much too high. This is an example of partial masking where the background noise masks part of the radio signal and when the background noise disappears the masking disappears too and the radio signal becomes louder than before. (Some modern car radios are equipped with a speed dependent automatic level control. The example above is not fully convincing for such cars.)

4.3 Forward masking

It has been shown that a strong sound signal can mask another (weak) signal which is presented after the strong signal. This kind of masking goes forward in time and is therefore called forward masking.

Forward masking is also called post-masking (post = after). Figure 4-4 shows the results of a forward masking experiment. The effect lasts for about 200 ms after the end of the strong signal.

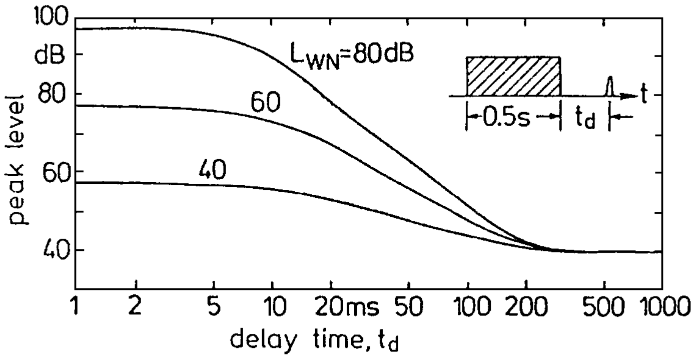


Figure 4-4. Threshold of a 20μs Gaussian pressure impulse as a function of the delay, td after the end of a white noise signal. Redrawn from (Zwicker and Fastl 1999)

4.4 Backward masking

It has been shown that a strong sound signal can mask another (weak) signal which appears before the strong signal. This kind of masking goes back in time and is therefore called backward masking.

Backward masking is also called pre-masking (pre = before). Figure 4-5 shows the result of a backward masking experiment. The effect is restricted to about 20 ms before the start of the strong signal.

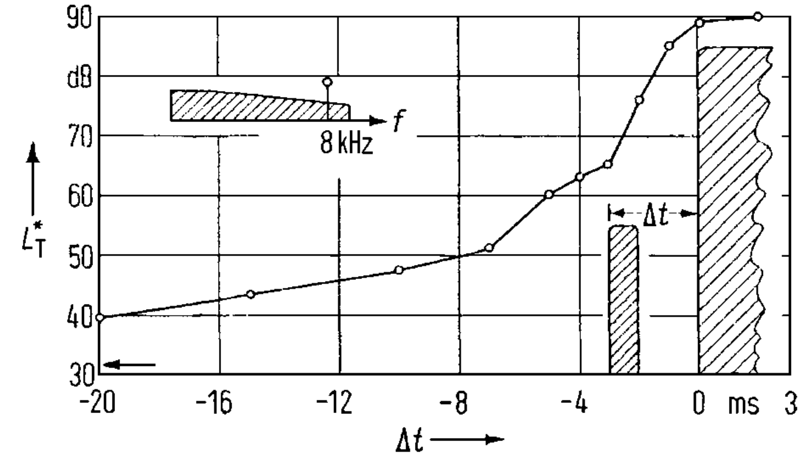


Figure 4-5. Threshold of a 1 ms impulse as a function of the time, Δt, between the impulse and the start of a long lasting signal. (Zwicker and Fastl 1999)

4.5 Interaural masking, Central masking

Interaural masking or Central masking is very different from the other masking phenomena especially from simultaneous masking. Central masking means masking in the central nervous system. Due to the connection between the two sides of the central auditory pathway, sound presented to one ear can influence the perception of a sound presented to the other ear. Thus, the two ears are not independent.

The monaural hearing threshold (measured in quiet) may change by as much as 15 dB when a noise is presented in the other ear. This effect must be taken into account in audiometry when a masking noise is presented to the non-test ear (contralateral masking). Contralateral masking is recommended if the hearing level differ by more than 40 dB in the two ears.

4.6 Comodulation masking release, CMR

The comodulation masking release, CMR, is defined as: The reduction in signal threshold that occurs for an amplitude modulated masker when additional frequency components - with the same pattern of modulation as the masker - are added to the stimulus (Plack, 2005). The release of masking is called comodulation masking release (CMR) because it occurs primarily when the modulation at remote frequencies is similar to that around the signal frequency, that is, when the on frequency and the remote masker bands are comodulated (Buus, 1997).

In simultaneous masking, it is assumed that detection of the signal (typically a pure tone) depends only on the signal-to-noise in the relevant frequency band around the signal frequency. The relevant band is here the auditory filter around the pure tone signal (see chapter 6). In other words, the noise power in the band determines the threshold. From Figure 4-6 it is seen that the threshold increase when the masker bandwidth increase up to a certain bandwidth which is 400 Hz in this case. When the bandwidth increases further the threshold does not increase further. This is an example of a measurement of the bandwidth of an auditory filter.

This way of looking at masking is adequate in most cases, but fails if the masker is amplitude modulated. In such cases a dramatic decrease of the threshold is seen. The decrease is believed to depend on modulations in bands away from the signal band.

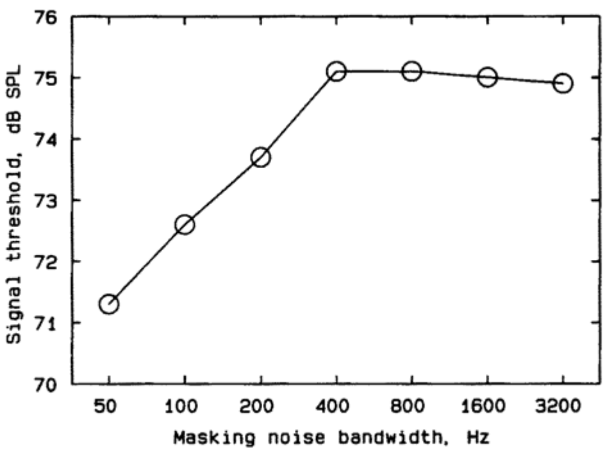


Figure 4-6. The threshold of a 2 kHz pure tone as a function of the bandwidth of a noise masker centred around 2 kHz (From Moore, 1995)

The circle points in Figure 4-7 constitute a curve that is similar to the curve in Figure 4-6 that shows the dependence of bandwidth for a stationary masker. If the masker is amplitude modulated the threshold for the pure tone signal decrease (Figure 4-7). When the bandwidth of the masker becomes 1000 Hz the threshold is lowered by as much as 10 dB.

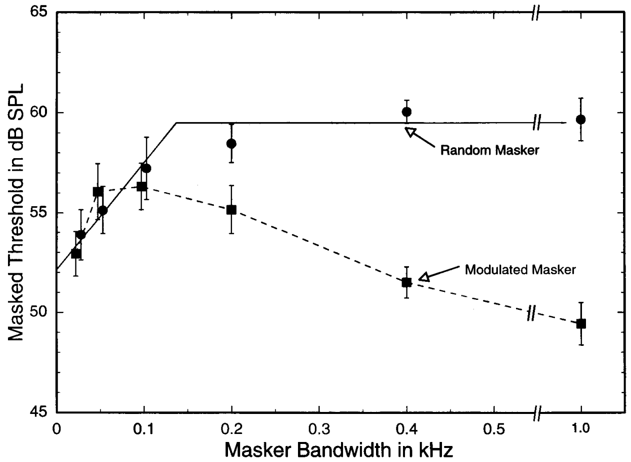


Figure 4-7. Threshold of a 350 ms tone at 1 kHz masked by bands of random noise (circles) or by amplitude modulated noise (squares). The maskers were centred around 1 kHz. (From Buus, 1997)

Buus (1997) gives the following explanations of comodulation masking release: “One possible explanation for CMR is that detection may be mediated by a simultaneous comparison of envelopes in different critical bands. …. The envelope comparison may be described by an equalization-cancellation (EC) model, which extracts envelopes in different critical bands, adjusts their root-mean-square (rms) amplitudes to be equal (equalization), and subtracts envelopes in channels remote from the signal from the envelope in the signal channel (cancellation). In the comodulation condition, the envelopes are similar when the signal is absent but dissimilar when it is present. Thus, the presence of the signal is indicated by a large remainder after cancellation.

Detection of decorrelation between the envelopes in the signal and the cue bands[1] may also explain CMR. The predictions by this model and the EC model may be identical. When the signal is absent, comodulation causes the correlation to be high; when the signal is present, it alters the envelope in the signal band, which in turn decreases the correlation. The decorrelation caused by a signal close to threshold in a typical CMR experiment is sometimes comparable to the just-detectable decrease in correlation of two bands of noise, but the effects of masker bandwidth are opposite in CMR and decorrelation detection."

Figure 4-8 shows the result of measurements where the threshold of a 2 kHz tone is determined as a function of the bandwidth of the masker which is centred around the masker. The signal is a 2 kHz pure tone and the masker is random noise around the pure tone. The upper panels are individual results from four individual listeners. The lower left panel shows the average data and the lower right panel shows the output from a mathematical model. The individual data show some differences between test subjects, but the general picture of the curves is the same. This intersubject variability is typical for psychoacoustic experiments. The mathematical model is Torsten Dau's modulation filterbank model. Similar results are found at 1 kHz and 4 kHz.

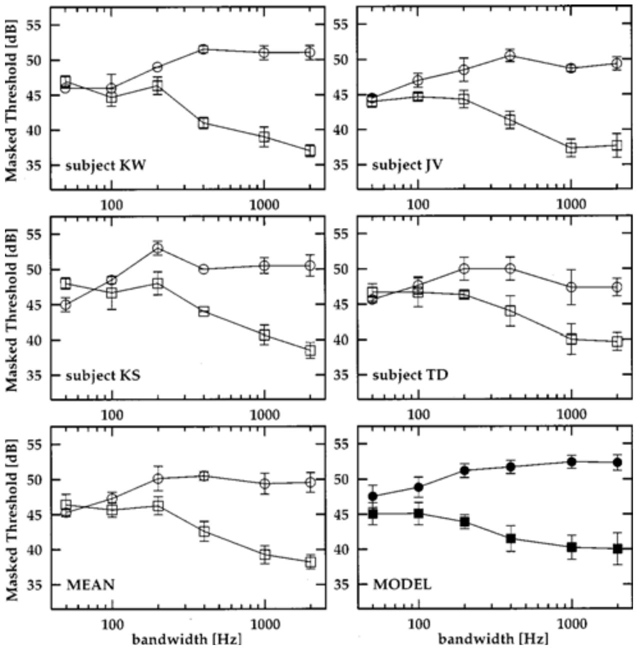


Figure 4-8. Signal threshold as a function of the masker bandwidth in random noise (circles) and modulated noise (squares) .The modulator bandwidth was 50 Hz; the signal frequency was 2 kHz. The upper four panels show individual data. The error bars show plus and minus one standard deviation from three repetitions. The lower left panel show the average data. In this panel, the vertical bars show the variability across subjects. The lower left panel shows the output from a model calculation (From Verhey et al. 1999).

The CMR concept is believed to have a great importance for speech perception and speech intelligibility, localization, the 'cocktail party effect', and other delicate detection tasks.

[1] The masker bands that are added away from the signal frequency are usually called cue bands

5. Loudness

The term ‘loudness’ denotes the subjective perception of strength or powerfulness of a sound. The unit for loudness is Son or Sone. Note that ‘loudness’ and ‘loudness level’ are different concepts. Translation of terms:

	Loudness Unit: *Sone*	**Loudness Level** Unit: *Phone*
Danish	Hørestyrke	Hørestyrkeniveau
German	Lautheit	Lautstärkepegel
French	Sonie	Niveau de Sonie

5.1 The loudness curve

The Sone scale was established in order to avoid the mental confusion between sound pressure levels (in dB) and the subjective perception of loudness: A 1 kHz tone at 80 dB SPL is not perceived double as loud as the same tone at 40 dB SPL. Figure 5-1 shows the relation between the Sone and the Phone scales. (Hint: for a 1 kHz tone, phone and dB SPL is the same number). Arbitrarily it has been decided that one sone should correspond to 40 phones. The curve is based on a great number of loudness comparisons. The curve is called a loudness curve.

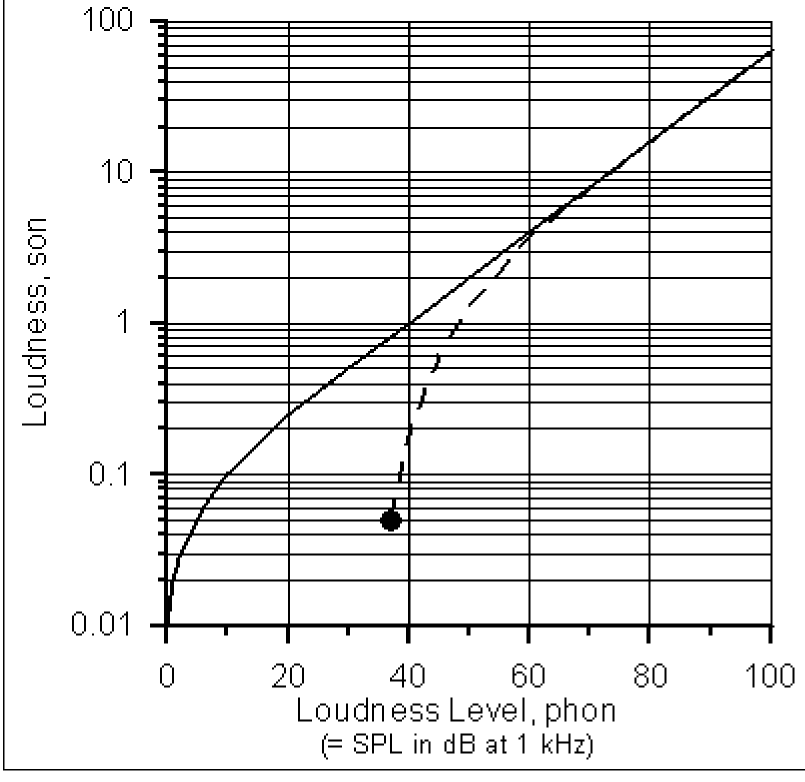


Figure 5-1. The loudness curve for a normal hearing test subject (solid line) and for a person with a cochlear hearing loss (dashed)

The ordinate in Figure 5-1 is logarithmic. The abscissa is linear in Phone (or dB for 1 kHz). As dB is in itself a logarithmic quantity, the figure shows a double logarithmic relation. The straight part of the solid line in Figure 5-1 corresponds to Stevens' power law:

$$N = kp^{0.6}$$

where N is the loudness in sones, k = 0.01, and p is sound pressure in micropascals (µPa). Near the hearing threshold, the curve becomes steeper. This can be expressed in the equation by introducing the threshold sound pressure, p_0 ,:

$$N = k(p - p_0)^{0.6}$$

This is called a modified power function. Other ways of modifying the power function to account for the bended curve near threshold is seen in the literature.

The curve in Figure 5-1 shows the relation between loudness (sones) and loudness level (phones). The straight part of the curve is given by

$$N = 2^{\frac{L-40}{10}}$$

where N is the loudness (in sones) and L is the loudness level (in phones). The curve shows that a doubling of the loudness corresponds to a 10-phone increase in loudness level (or a 10-dB increase in SPL if we are dealing with a 1 kHz tone). For many daily life sounds a rule of thumb says that a 10-dB increase is needed in order to perceive a doubling of the loudness.

The loudness curve becomes steeper near the hearing threshold. This is also the case for a person with a cochlear hearing loss (e.g., the very common hearing impairment caused by age). An example of such a hearing loss is shown by the dashed curve in Figure 5-1 where the threshold (1 kHz) is a little less than 40 dB SPL. The steeper slope means that - near the threshold - the loudness increases rapidly for small changes in the sound level. This effect is called *loudness recruitment*. Recent research have shown that – for this kind of hearing loss – the loudness at threshold has a value significantly different from nil[2] as indicated in the figure (Florentine and Buus 2001). In other words, listeners with cochlear hearing loss suffer from *softness imperception*, rather than they exhibit loudness recruitment.

Note that at higher sound levels the perception of loudness is the *same* for both normal and impaired listeners. This observation is one of the major problems for hearing impaired persons and is very important for the construction and fitting of hearing aids.

Figure 5-2 shows the loudness function for a 1 kHz pure tone (same as in Figure 5-1) and the loudness function for a white noise signal (i.e. a signal with a wide spectrum, a broad band signal). This loudness function is determined by means of the procedures magnitude estimation and magnitude production.

[2] 'nil' means no loudness perception at all

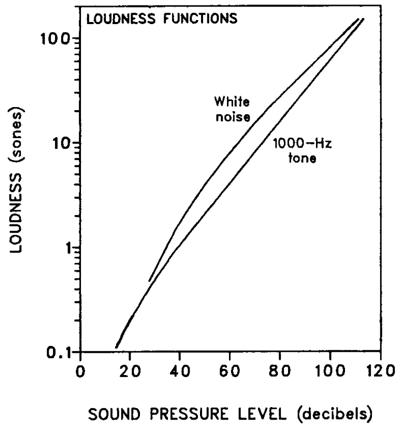


Figure 5-2. Loudness function for white noise compared to the loudness function for a 1000 Hz pure tone. The white noise function is determined by magnitude estimations and magnitude production (average of fifteen subjects). From Scharf & Houtsma (1995).

In magnitude estimation the test subject reports a number that represent the test subject's perception of loudness (how loud the noise is perceived). In magnitude production the test subject adjusts the loudness of the noise to match the magnitude of a number given by the experimenter. Above about 30 dB the pure tone curve is a straight line (a power function), whereas the noise curve is bowed in the mid-level range. Near threshold, the noise curve grows more rapidly (steeper curve) than the tone curve. Above about 60 dB it grows more slowly.

5.2 Temporal integration

The perception of loudness needs some time to build up. This means that short duration sounds (less than one second) are perceived as less loud than the same sound with longer duration. The growth of loudness as a function of duration is called temporal integration. The growth resembles the exponential growth of a time constant. It has been shown that the time constant is about 100 ms. The loudness will grow towards the final value according to this curve. It can be seen that a 50 ms sound will only reach about 40% of the loudness that an 800 ms sound would give.

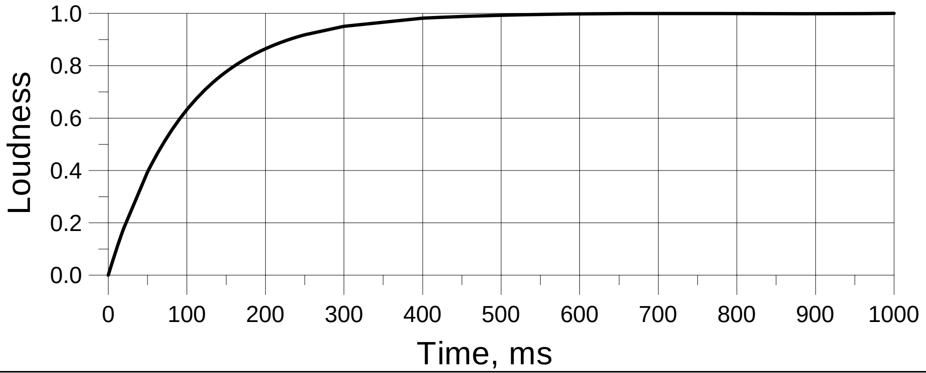


Figure 5-3. Growth of loudness (arbitrarily set to 1). Corresponding to the voltage growth of a RC-circuit with a time-constant 100 ms.

The loudness growth can be described by an exponential relation

$$I(t) = \frac{I_\infty}{1 - e^{-t/\tau}}$$

where I(t) is the intensity of a signal of duration t, I_∞ is the intensity for a long duration, and τ is the time constant. Based on an extensive number of investigations, the best fit of the time constant is about 80 to 100 ms.

Temporal integration has been measured by means of loudness comparisons of tone pulses of different durations. Figure 5-4 show an example of such measurements. It can be seen that a 5 ms tone pulse need a 17 dB higher level than a 640 ms tone pulse (presented at 35 dB SPL) in order for the two pulses to have the same loudness. It can also be seen that the steepness of the curve is less at 95 dB compared to 35 dB.

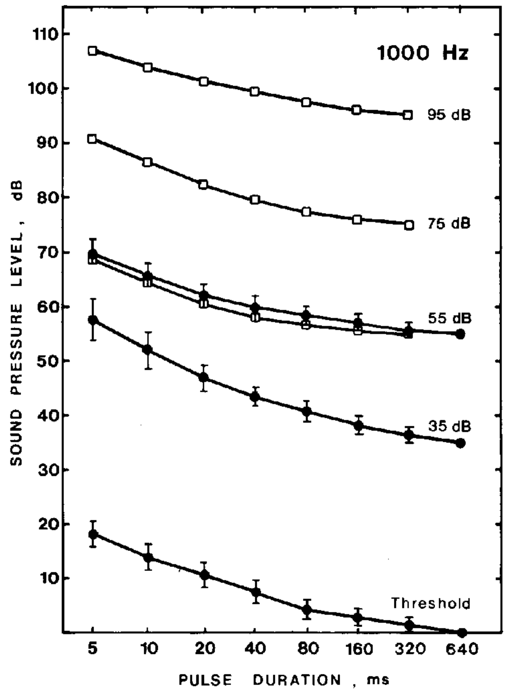


Figure 5-4. Temporal integration. Based on loudness comparisons of tone pulses of different durations. The lowest curve shows the threshold of the tone pulses. (Poulsen 1981)

Short sounds - like a pistol shot, fireworks, handclap, etc. - are perceived as weak sounds although their peak sound pressure levels may be well above 150 dB SPL. This is one of the reasons why impulsive sounds are more dangerous than other sounds. The nerve signal generated by such short sounds will need only about 10 ms to reach the brain (like any other sounds) but this does not mean that the listeners can make a loudness decision within this short time frame.

Temporal integration measurements have been used to derive the shape of the loudness function. Loudness comparisons between 5 ms and 200 ms tone pulses show a considerable change with the overall level, see Figure 5-5. At low levels and at high levels the level difference (for equal loudness) is about 15 dB. At medium levels, the difference is about 25 dB.

This change in temporal integration with level can be used to derive the shape of the loudness function. This is shown in Figure 5-6. It is seen that the loudness function is not a straight line at levels well above threshold (as indicated in Figure 5-1).

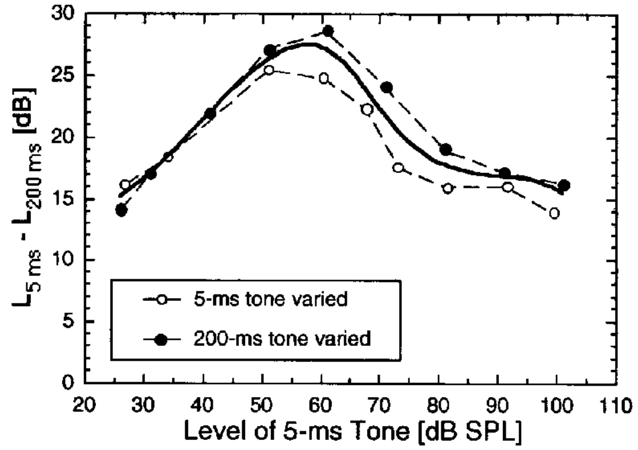


Figure 5-5. Temporal integration as a function of level. Loudness comparison between 5 ms and 200 ms tone pulses. (Buus et al. 1997)

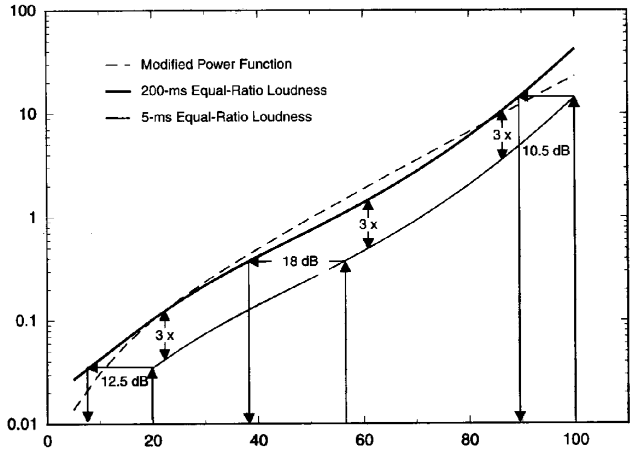


Figure 5-6. Loudness function derived from the temporal integration data in Figure 5-5 (Buus et al. 1997)

The abscissa shows sensation level

5.3 Measurement of loudness

Many years ago it was thought that a sound level meter with filters corresponding to the ears' sensitivity (described by the equal loudness level contours (Figure 3-5) could be used to measure loudness. This is not the case.

Figure 5-7 show the characteristics for the commonly used A- and C- filters, but due to masking and other phenomena, these filters will not give a result that corresponds to loudness. For the determination of loudness, special calculation software is needed. For stationary sounds two procedures can be found in (ISO-532 1975). For non-stationary sound, loudness calculations are found in professional Sound Quality calculation software. For research purposes loudness models (software) can be found on the Internet (e.g. at http://hearing.psychol.cam.ac.uk/Demos/demos.html)

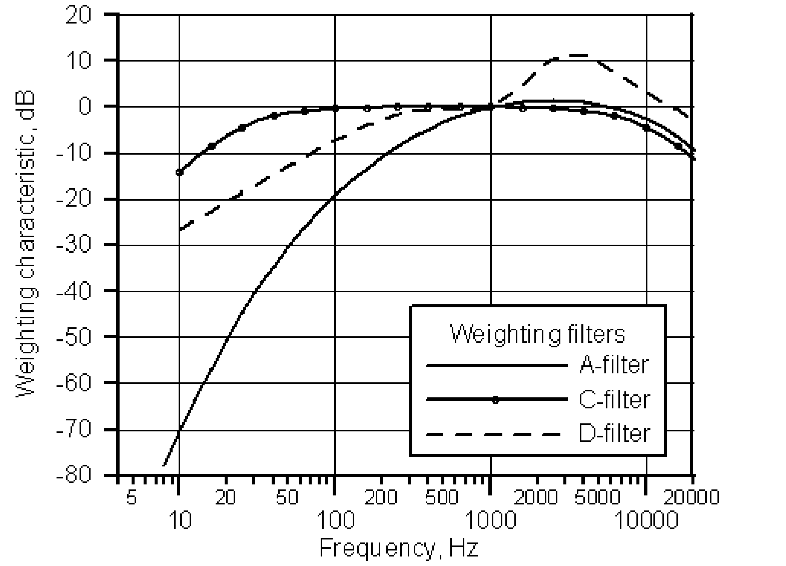


Figure 5-7. Filter characteristics for the A, C and D filter. The data for the A and the C filter are from (IEC-651 1979). The data for the D filter is from (IEC-537 1976).

The main effect of the A-filter is that it attenuates the low frequency part of the signal. The attenuation is e.g. 20 dB at 100 Hz and 30 dB at 50 Hz. Wind noise and other low frequency components are attenuated by the A-filter and is therefore very practical for many noise measurement situations.

The C-filter is 'flat' in the major part of the audible frequency range. It may me used as an approximation to a measurement with linear characteristic.

The D-filter is mainly used in connection with evaluation of aircraft noise. The frequency range around 3 kHz is known to be annoying and therefore this frequency range is given a higher weight.

5.4 Loudness scaling

A procedure for categorical loudness scaling can be found in Brand & Hohmann, JASA, vol 112, October 2002, p 1597-1604.

5.5 Level discrimination

The concept of the smallest detectable change in a sound level is called 'Level discrimination' or 'Intensity discrimination' and these terms are widely used in the literature. The just noticeable difference, JND, is one of the classical concepts in psychophysics. The JND for loudness was determined by Riesz as early as in 1928 (Riesz, 1928), and has been investigated by many researchers since then. Sometimes the JND is called the 'difference limen', DL, but this abbreviation should be reserved for the 'discrimination loss' that is used in clinical speech audiometry (DL = the deviation from 100% word intelligibility).

Level discrimination can be measured by different principles. The most common are

- Detection of an amplitude modulation in a continuous signal
- Detection of an amplitude difference (level difference) between two separate stimuli

The JND will depend on the principle of the stimulus presentation, the psychometric method used, the criterion for calling the difference just noticeable (i.e. which point on the psychometric function), and a number of other variables.

The JND is expressed in decibels. Figure 5-8 shows the JND in dB for a pure tone and for white noise as a function of the overall level of the tone, $L_{T,}$ and the white noise, L_{WN} , respectively. For the pure tone, it is seen that a difference of about 1 dB is needed at a level of 30 dB, whereas only 0.2 dB is needed at a level of about 90 dB.

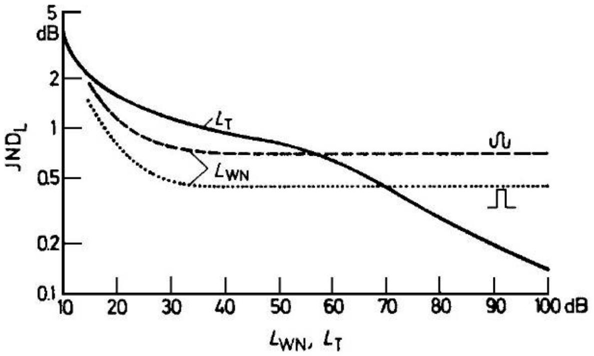


Figure 5-8. JND as a function of the level for a1 kHz tone (solid line) and for white noise. Dashed: sinusoidal modulation,4 Hz; dotted: rectangular modulation (from Zwicker, 197x)

For the white noise, the situation is different. At 30 dB, a level difference of about 0.5 dB is necessary (depending on modulation type) but this is also the case at 90 dB.

For such wideband noises, it is usually found that the smallest detectable change is approximately proportional to the magnitude of the stimulus. This means that

$$\Delta p = kp$$

Where p is the sound pressure and k is a constant called the Weber fraction. The relation is Weber's law:

- The just noticeable difference between stimuli is proportional to the magnitude of the stimuli.

If Δp is the pressure difference between two stimuli that are just detectably different in loudness, we can calculate the level difference

$$\Delta L = 20\log\frac{p+\Delta p}{p} \qquad \text{or} \qquad \Delta L = 10\log\frac{I+\Delta I}{I}$$

If Weber’s law is correct (*k* is a constant), then the level difference, Δ*L*, is also constant. From Figure 5-8 (and not too close to the hearing threshold) it is seen that the level difference is constant (around 0.5 dB) for white noise, but it is not constant for the pure tone.

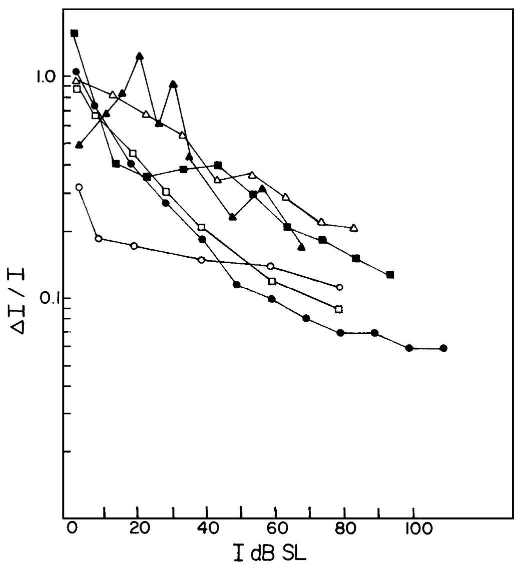


Figure 5-9. Results of level discrimination measurements show as Weber fractions. If Weber’s law is fulfilled, the curves should be ‘horizontal’. (From Green 1976).

Original figure legend:

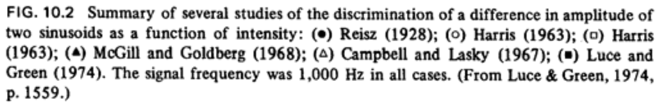


FIG. 10.2 **Summary of several studies of the discrimination of a difference in amplitude of two sinusoids as a function of intensity: (●) Reisz (1928); (○) Harris (1963); (□) Harris (1963); (▲) McGill and Goldberg (1968); (△) Campbell and Lasky (1967); (■) Luce and Green (1974). The signal frequency was 1,000 Hz in all cases. (From Luce & Green, 1974, p. 1559.)**

Figure 5-9 shows some of the older data on level discrimination of a 1 kHz pure tone. The data shows the same trends but do not agree entirely. This is caused by different measurement techniques, different training of the tests subjects, etc. The data show a general decline of ΔI/I as a function of I. If Weber’s law were fulfilled, the curves would show no slope, but this is obviously not the case. Even though the deviation from a no-slope line is obvious it has become common practice to talk about the ‘near miss to Weber’s law’. Green (1976) has a comprehensive discussion of the implications for models of intensity discrimination.

If the level difference ΔL (in dB) is plotted against the overall level L (in dB), an almost straight line is obtained with a slope of 0.9, but if Weber’s law was fulfilled, the slope would be 1.0. Shown in this way the ‘near miss’ expression becomes more reasonable.

Figure 5-10 shows an example of level discriminations measured at a number of frequencies. The near-miss to Weber’s law is clearly seen. A frequency effect is present showing that the discrimination is worse at the higher frequencies.

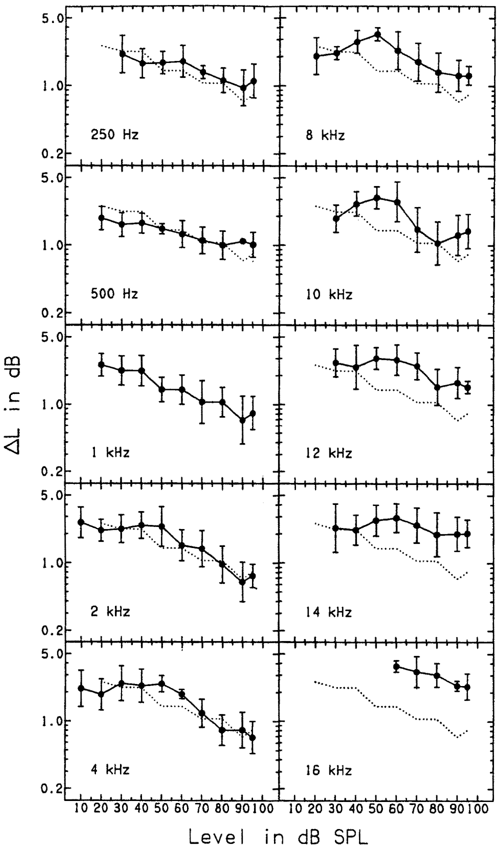


Figure 5-10. Just noticeable level difference for pure tones as a function of the presentation level. (From Florentine et al. 1987)

Original figure legend:

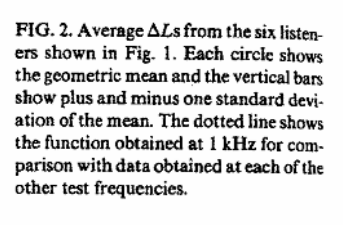


FIG. 2. Average ΔLs from the six listeners shown in Fig. 1. Each circle shows the geometric mean and the vertical bars show plus and minus one standard deviation of the mean. The dotted line shows the function obtained at 1 kHz for comparison with data obtained at each of the other test frequencies.

6. The auditory filters

The movements of the basilar membrane in the inner ear constitute a frequency analyser where the peak of the envelope moves along the basilar membrane as a function of frequency. See Figure 2-5. The width of the envelope peak may be seen as an indication of the selectivity of the analyser filter and it has been common practice to describe the frequency selectivity of the ear as a set of filters, a filter bank, which cover the audible frequency range. It should be noted though that the concept of a filter bank is a very coarse description and should be seen as a typical engineering approximation to the real situation.

Frequency selectivity is important for the perception of the different frequencies in complex sound signals such as speech and music. We rely on our frequency selectivity when we distinguish different vowels from each other.

The concept of frequency discrimination is different from frequency selectivity. Frequency discrimination is the ability to hear the difference between two tones that are close in frequency (one frequency at a time).

6.1 Critical bands

The bandwidth of the filters in the filter bank can be determined by means of various psychoacoustic experiments. Many of these are masking experiments that led to the formulation of the critical band model.

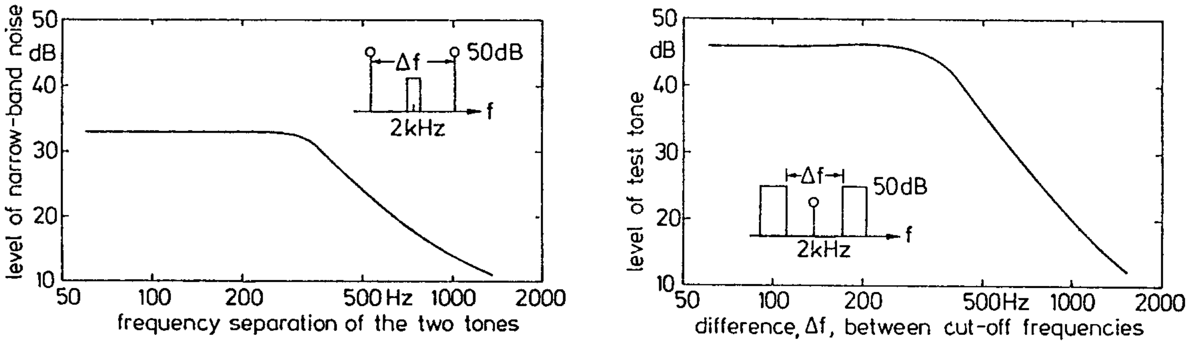


Figure 6-1. Example of masking experiments used to determine critical bandwidth. Left part: threshold of narrow band noise as a function of the frequency distance between two pure tones. Right part: threshold of a pure tone as a function of the frequency distance between two narrow band noises. (Zwicker and Fastl 1999)

An example of a classical masking experiment for determining the critical bandwidth is given in Figure 6-1. In the left part of the figure, the threshold of a narrow band noise (bandwidth less than a critical band) has been determined as a function of the frequency separation between two pure tones. Starting from a small separation it is seen that the threshold is constant up to a separation of about 300 Hz. When the separation is further increased, the threshold decreases. This knee-point on the threshold curve is defined as the critical bandwidth.

In the right part of Figure 6-1 the same experiment is made but now the threshold of a pure tone between two narrow band noises is determined. Also in this situation a threshold curve is obtained that is constant up to a certain bandwidth and then decreases. Again, the critical bandwidth is defined as the frequency separation where the curve starts to decrease.

The result of such the investigations is shown in Figure 6-2. It is seen that the bandwidth (Critical Bands) is almost constant at 100 Hz up to a centre frequency of about 500 Hz and above this frequency, the bandwidth increases. The increase in bandwidth above 500 Hz is similar to the increase in bandwidth for one-third-octave filters.

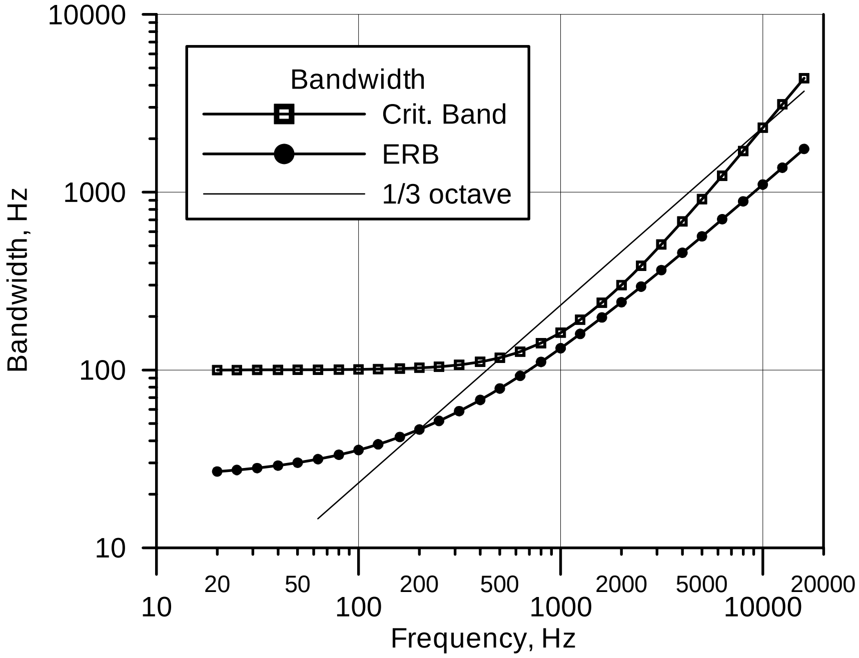


Figure 6-2. Bandwidth of critical bands and Equivalent Rectangular bandwidth, ERB. The bandwidth of 1/3-octave filters (straight line) is shown for comparison. The curves are computed from the formulas given in the text.

The critical bandwidth may be calculated from the empirical formula:

$$CB = 25 + 75(1 + 1{,}4F^2)^{0{,}69}$$

where *CB* is the bandwidth in Hz of the critical band and *F* is the frequency in kHz (not in Hz).

If the audible frequency range is ‘filled up’ with consecutive critical bands from the lowest frequency to the highest frequency, it is seen that 24 critical bands will cover the whole frequency range. Each of these ‘critical band filters’ has been given a number called Bark. Bark number one is the band from zero to 100 Hz; Bark number two is the band from 100 Hz to 200 Hz, etc. Band no. 8 has a centre frequency of 1000 Hz and goes from 920 Hz to 1080 Hz. The band around 4000 Hz is no. 17 and has a bandwidth of 700 Hz. A formula for the calculation of the number of critical bands (the Bark-number) that are needed to cover the range from ‘zero’ Hz up to a given frequency is given by

$$z = \frac{26{,}8}{1 + {}^{1{,}96}\!/_{F_c}} - 0{,}53$$

where F_c is the frequency in kHz (not in Hz).

The critical bands are not fixed filters similar to the filters in a physical filter bank as the numbers given above may indicate. The critical bands are a result of the incoming sound signal and as such much more ‘flexible’ than physical filters would be.

6.2 Equivalent Rectangular Bands

The auditory filters have also been determined by means of notched noise measurements where the threshold of a pure tone is determined in the notch of a broadband noise as a function of the width of the notch (Patterson 1976). This leads to the concept of equivalent rectangular bandwidth, i.e. the bandwidth of a rectangular filter that transmits the same amount of energy as the auditory filter. The bandwidth of such rectangular filters is shown in Figure 6-2 as a function of centre frequency.

The notched-noise principle is illustrated in Figure 6-3.

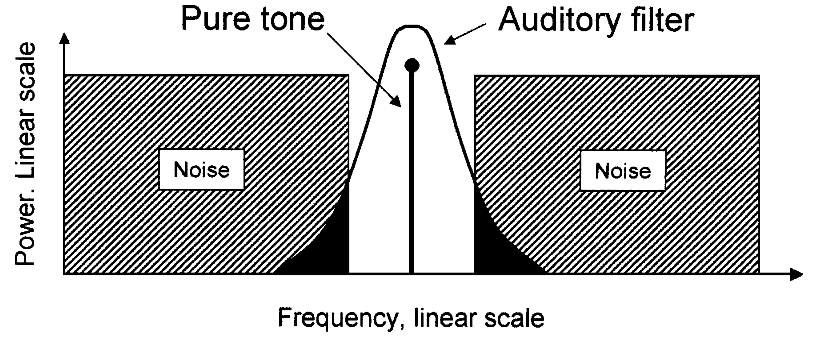


Figure 6-3. Principle of determination of the auditory filter bandwidth by means of notched noise.

A pure tone is positioned centrally in the spectral notch of a broadband noise signal. The frequency of the pure tone (the signal) is kept constant and the threshold of the pure tone is determined as a function of the width of the notch. It is believed that this detection task is made by means of an auditory filter centred around the signal. The black area is proportional to (represents) that part of the noise that passes through the auditory filter. If the notch is narrow, much of the noise will be within the filter and thus more masking of the pure tone takes place (higher pure tone threshold). When the spectral notch becomes wider, less noise will be within the filter, the signal will be easier to detect and the signal threshold decrease.

Determined in this way, Patterson suggested that the shape of the auditory filter could be described (approximately) by the formula:

$$W(g) = (1 + pg)\exp(-pg)$$

$$g = \frac{|f - f_c|}{f_c}$$

where *g* is the normalized frequency, i.e. the deviation from the centre of the filter. From the threshold detection data, the parameter *p* can be calculated. This parameter determines the sharpness of the filter and may be different for the upper and lower slopes. In order to avoid the difficult description of the auditory filter by means of the -3 dB bandwidth (or -10 dB bandwidth), the concept of equivalent rectangular bandwidth is introduced.

The Equivalent Rectangular Bandwidth, ERB, is defined as the bandwidth of a theoretical rectangular filter that passes the same power of white noise as the (auditory) filter in question. The equivalent rectangular bandwidth is equal to

$$ERB = \frac{2f_c}{p_l} + \frac{2f_c}{p_u}$$

where p_l and p_u are the lower and upper part of the filter.

The equivalent rectangular bandwidth may be calculated from the empirical formula:

$$ERB = 24{,}7(4{,}37F_c + 1)$$

where *ERB* is the bandwidth in Hz and F_c is the centre frequency in kHz (not in Hz).

Similar to calculating the number of critical bands, *z,* used to cover the frequency range from ‘zero’ to a given frequency, the number of ERB’s needed to cover that range is approximately given by

$$E = 21{,}4\log_{10}(4{,}37F_c + 1)$$

where F_c is the centre frequency in kHz (not in Hz). *E* may be called the ERB number.

Description of the frequency selectivity by means of Equivalent Rectangular Bandwidth, ERB, is very useful in auditory models. In (Moore 2001) many more details are given about the determination and use of the ERB.

6.3 Critical ratio

The idea behind the critical band and the ERB concept is that a number of adjacent filters constitute the audible frequency range. This means that - in a pure tone masking experiment - only the noise in the frequency range within a filter can contribute to the masking of the tone.

This model was made by Fletcher in 1940 when he investigated the masking effect. He described the results by means of a critical ratio.

Critical ratio is defined as the difference between the level of a pure tone and the density level of a white noise that just masks the tone. CR is determined by

$$CR = L_M - L_W$$

where L_M is the masking threshold in dB SPL and L_W is the density level of the noise in dB/Hz.

The density level can be calculated from the sound pressure level of the noise. If the sound pressure level is determined with a bandwidth of 10 kHz the level for a bandwidth of 1 Hz is

$$L_W = SPL – \log(10000/1) = SPL – 40$$

The result is given in dB/Hz. The same calculation procedure can be used to determine the density level for other bandwidths, e.g. when the sound pressure is measured in 1/3-octave bands.

6.4 Psychoacoustic tuning curve

The frequency selectivity of the ear can be described by the so-called psychoacoustic tuning curve, sometimes abbreviated to PTC.

The PTC is determined in the following way. A pure tone at a fixed level is presented to the ear. The level of this fixed tone shall be close to the hearing threshold, e.g. 20 dB above the threshold. Then it is determined how loud another tone shall be in order to just mask the fixed tone. This masked threshold is determined as a function of frequency. An example is given in Figure 6-4.

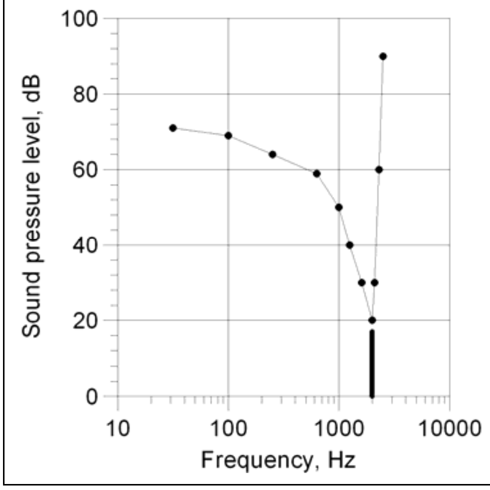


Figure 6-4. Psychoacoustic tuning curve. The fixed tone is at 2 kHz and is shown as the vertical bar.The level of the tone is about 18 dB in this example.

The shape of the curve resembles a real tuning curve measured e.g. in the auditory nerve. The width of the dip is an expression of the frequency selectivity.

6.5 Frequency discrimination

Frequency discrimination is – in opposition to frequency selectivity – the ability to distinguish a slight change in frequency. Usually two pure tone stimuli are presented with frequencies that are slightly different. The task is to detect when the frequency has changed, in other words: determination of the difference limen for frequency. This method uses discrete frequencies. Another method is to use a slow (4 Hz) frequency modulation of the frequency and then determine the just noticeable change in the frequency. Results are shown in Figure 6-5 and Figure 6-6.

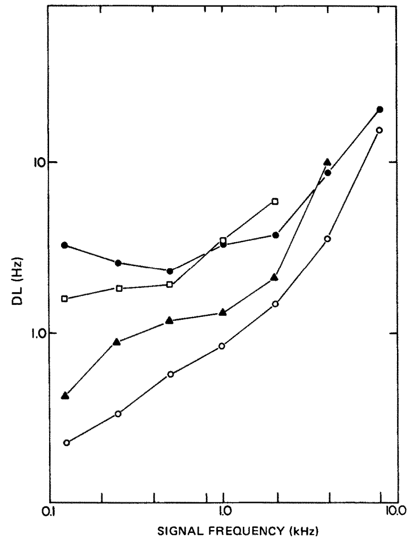


Figure 6-5. Old measurements of the difference limen for frequency. The filled circles and the open squares are measurements from 1931. The filled triangles are from 1952, the open circles are from 1968. (From Green 1976).

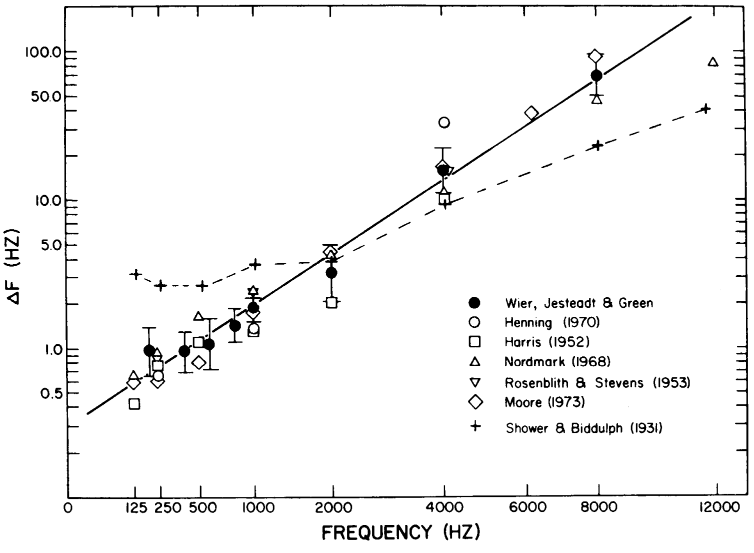


Figure 6-6. Results from a number of studies of frequency discrimination. All measurements are measured with discrete frequencies except the '+' curve from 1931. (From Moore1997)

A general result is that it is possible to detect a frequency change of about 1 to 2 Hz for a 1 kHz pure tone, i.e. approximately 0,1%. At 4000 Hz the JND is about 20 Hz, i.e. 0,5%

7. The speech organs

A speech signal is produced in the following way. Air is pressed from the lungs up through the vocal tract, through the mouth cavities and/or the nose cavities.

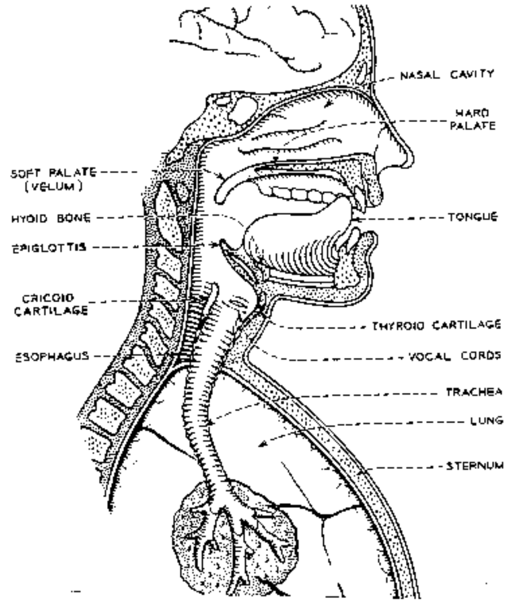


Figure 7-1. Drawing of the speech organs. From (Flanagan 1972)

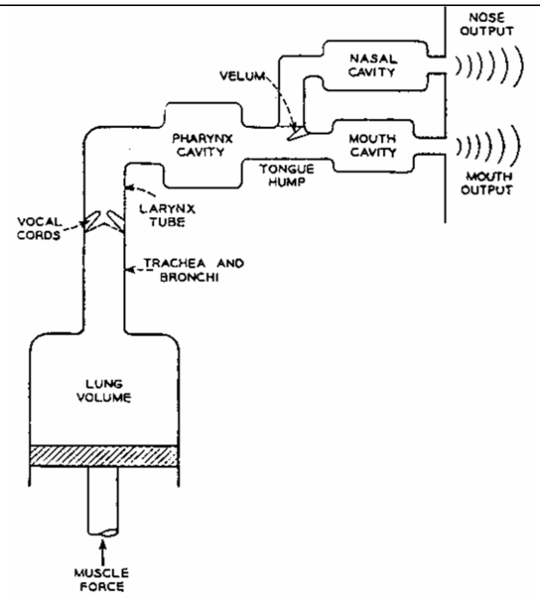


Figure 7-2. Schematic illustration of the speech organs. From (Flanagan 1972)

Speech sounds are produced by means of an air flow that is pressed up from the lungs and out in the free air through the vocal tract (i.e. windpipe (trachea) and throat (larynx and pharynx)), the mouth (oral cavity), the nose cavity, the teeth and the lips. The sound is radiated from the mouth and the nose.

Figure 7-1 and Figure 7-2 shows the speech organs and a schematic model of the speech organs. The lungs may be modelled by a piston that presses the air up through the vocal cords. For voiced sounds (e.g. vowels), the vocal cords are vibrating and generate an acoustic signal. The acoustic signal from the vocal cords is then formed by the cavities in the throat, mouth and nose before it is emitted from the mouth/nose. For unvoiced sounds (e.g. the consonants p, t, k, s, f (but not l, m, n, r, v)) the sound is not generated at the vocal cords but is produced when the air is passing the teeth or at a sudden opening of the lips.

7.1 Breathing

Normal breathing is usually performed through the nose. A normal in- and exhalation period is composed by 40% inhalation and 60% exhalation. During speech the breathing shall deliver the necessary air for the speech production and the breathing rhythm is therefore changed. The inhalation is fast (covers only 10% of the time) and is usually performed through the mouth. The exhalation (about 90% of the time) is used to produce the speech and last therefore considerably longer. The strength of the airflow is used to control the speech level.

Speech may also be produced during inhalation. This is the situation when one is breathless or when you wipes or cry. It is also common practice to say 'yes' during inhalation when you are listening to another person in order to indicate that you are still listening.

7.2 Larynx

A drawing of larynx is shown in Figure 7-3. In Larynx both the false vocal cords (or folds) and the true vocal cords are found. The vocal cords constitute a narrowing of trachea and the opening between the vocal cords is called Glottis. Around the vocal cords the thyroid cartilage is found which is squeezed a little in the frontal part. This pinching of the thyroid is about 90º for men and about 120º for women. The male 'Adams apple' can therefore easily bee seen. In the upper part of the larynx the Epiglottis is found. It can be bended and close for the larynx. This happens during swallowing.

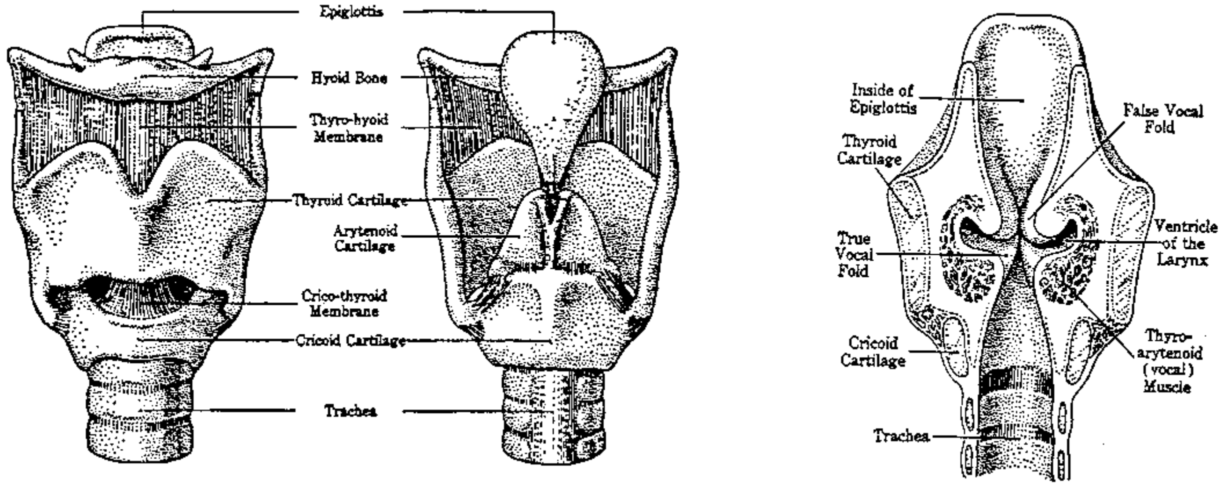


Figure 7-3. Larynx seen from the front and from the back. To the right a vertical cut is shown. From (Borden and Harris 1980)

The vocal cords are separated (i.e. open) when no speech is produced and the air can pass freely during breathing. During speech the vocal cords are open in unvoiced sounds (e.g. s,t) but are pressed together in voiced sounds (e.g. vowels) and voiced consonants (e.g. z,d). Muscles in Larynx perform this static contraction, but the muscles do not vibrate the vocal cords.

The vibration of the vocal cords is caused by the overpressure made by the lungs under the contracted vocal cords that then will open up. When the air then pass through glottis, the pressure will decrease (Bernoulli effect) and the static contraction of the vocal cords will close the opening again. Hereafter a new overpressure will build up and the process will be repeated.

Figure 7-4 shows the different phases of the vocal cord open/close period. The concentration of dots indicates the amount of pressure.

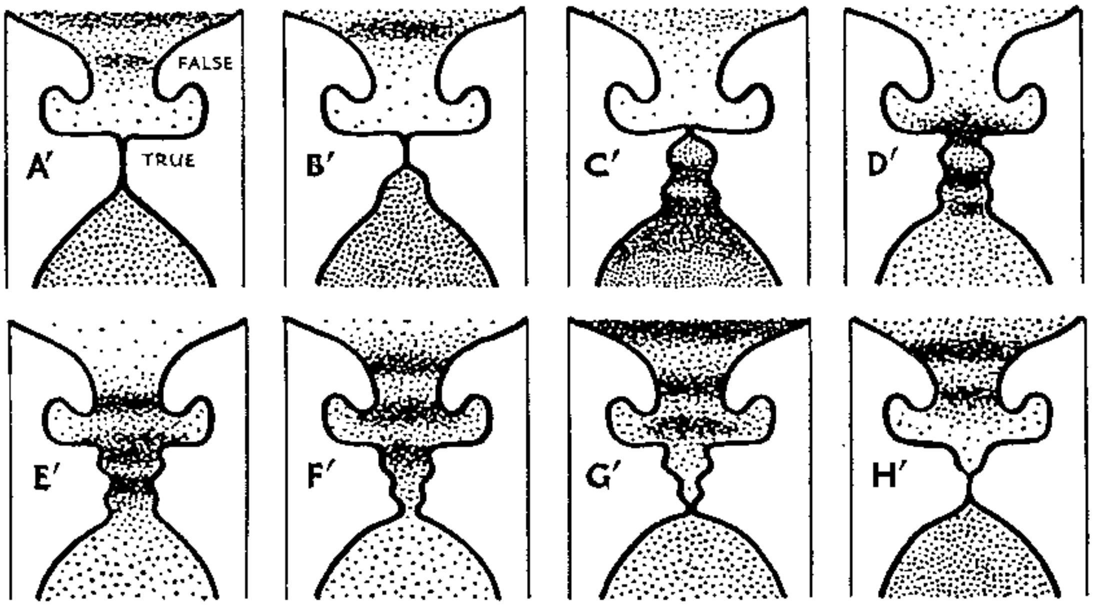


Figure 7-4. Position of the vocal folds during a period of vibration. The movement of the vocal cords will produce a periodic acoustic signal. The fundamental frequency of the sound will typically be 150 Hz for men and 300 Hz for women. The fundamental frequency is perceived by the listener as the speakers pitch. During normal running speech the pitch will vary. This is called intonation. From (Borden and Harris 1980)

The area of glottis during a vibration period is shown in Figure 7-5. The volume velocity is also shown in the figure.

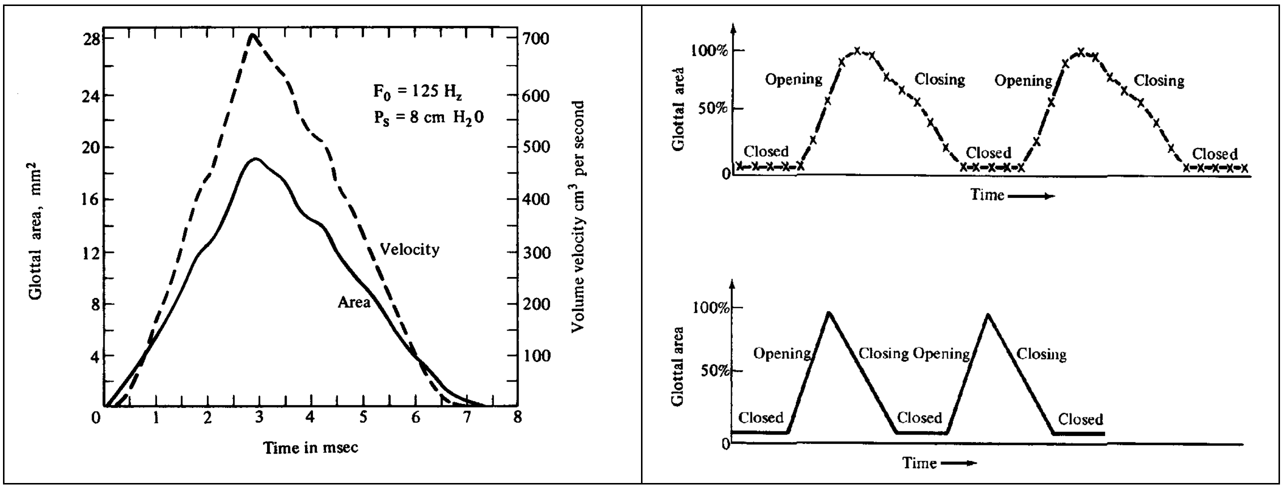


Figure 7-5. Area of the Glottis opening and the volume velocity through Glottis during one period of a vocal cord vibration. From (Minifie et al. 1973).

The area and the volume velocity show a triangular course during the vibration period. The period lasts about 8 ms corresponding to a 125 Hz fundamental frequency (f_0). The vocal cord signal is thus a periodic triangular signal. The spectrum of this signal is a line spectrum where the distance between the lines is the fundamental frequency (= 125 Hz in this example). The amplitude of the lines in a triangular signal decrease by 12 dB per octave, i.e. inversely proportional to the frequency squared (f^2). Figure 7-6 show two examples of the spectrum of the vocal cord signal. The upper part is for a male (f_0 = 150 Hz), the lower part is for a child (f_0 = 350 Hz).

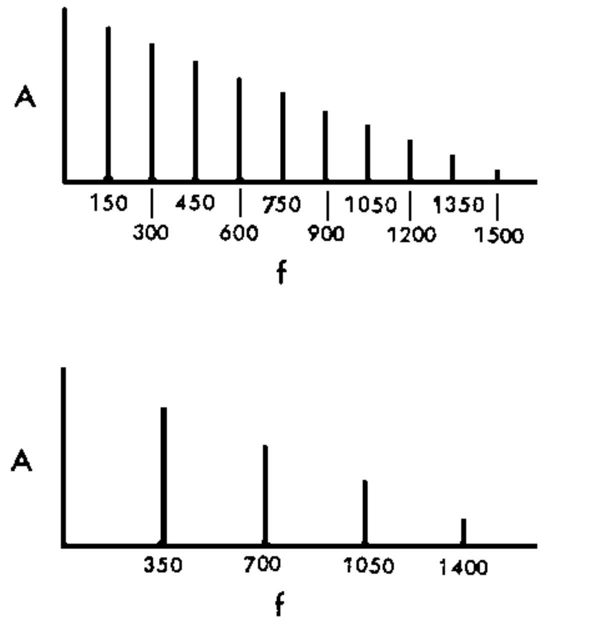


Figure 7-6. Source spectrum for an adult male with fundamental frequency 150 Hz (upper part) and for a child with fundamental frequency 350 Hz (lower part). From (Borden and Harris 1980)

7.3 Pharynx, mouth and nose

Pharynx is found above the Larynx. Pharynx is a cavity that continues upwards into the nose and forward into the mouth. The tongue is found at the bottom of the mouth cavity and the palate is found at top of the mouth cavity. The frontal part of the palate is hard, whereas the inner part is soft (also called velum). The uvula is found as the inner part of the soft palate. In the frontal part of the mouth cavity, the teeth and the lips are found.

These speech organs are used to produce different speech sounds. The airflow from the lungs can be influenced by narrow passages e.g. between the teeth and/or the lips. The cavities in pharynx and the mouth/nose constitute an acoustic transmission line where the properties can be changed by the physical dimensions of the cavities by means of the palate, the tongue, lips, etc.

The tongue is especially movable and can change the size of the mouth cavity very dramatically. The tip of the tongue is used e.g. during the pronunciation of ‘t’ and ‘d’ where the tip is pressed against the backside of the teeth. The rearmost part of the tongue is pressed against the soft palate when ‘k’ and ‘g’ is pronounced.

During normal breathing, the epiglottis is lowered in order for the air to pass freely in and out of the nose. Epiglottis will typically be raised during normal speech to prevent the sound to pass through the nose. Only for the nasal sounds (e.g. m, n) the epiglottis will be open.

The uvula is used to produce a rowing ‘r’. This is not very common in Danish but usual in e.g. German and Dutch.

The jaw is usually moved up and down during speech, but it is possible to talk with a fixed jaw.

8. The speech signal

8.1 Overview of speech production

A schematic illustration of the production of voiced sounds is given in Figure 8-1. The source spectrum is a line spectrum where the distance between the lines corresponds to the fundamental frequency. The fundamental frequency is around 125 Hz for men, around 250 Hz for woman and around 300 for children, but there are big individual variations. There are thus more lines in a male spectrum compared to a female. The source spectrum decreases with the square of the frequency ($1/f^2$). The source spectrum is formed by the 'tube' consisting of trachea, throat (pharynx) and the mouth. This structure is simulated in Figure 8-1 by a cylindrical tube of length 17 cm.

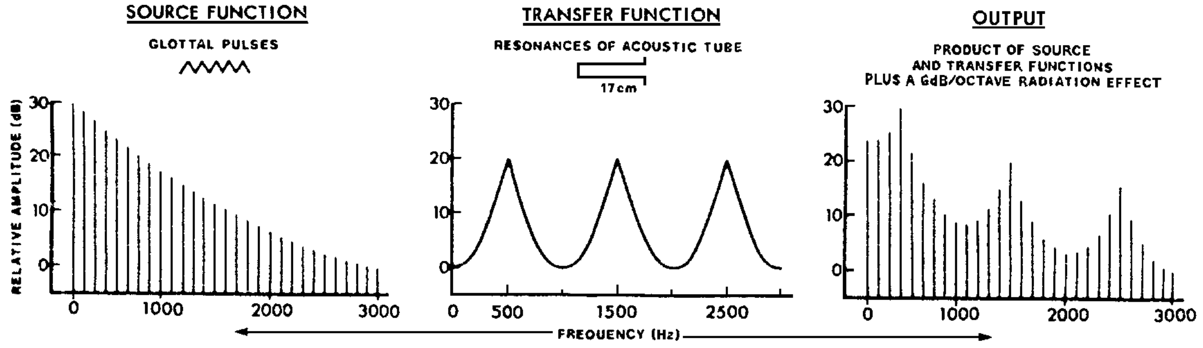


Figure 8-1. The principle of vowel generation. From (Borden and Harris 1980)

The tube has pronounced resonances (where the length of the tube corresponds to the odd multiples of 1/4 wavelength) indicated by the peaks at 500, 1500 and 2500 Hz. The final spectrum radiated from the mouth is then the product of the two spectra. The final spectrum is a line spectrum with characteristic peaks caused by the transfer function. The peaks are called *formants* and the formants are positioned differently for each vowel. Table 8-1 shows the formants frequencies (in round numbers) for the three most different vowels. The sounds are /i/: as in eve, /a/ as in father, /u/ as in moon. There are individual differences from person to person.

	/i/	/a/	/u/
1. formant	225	700	250
2. formant	2200	1200	700
3. formant	3000	2500	2200

Table 8-1. Formant frequencies in Hz of the vowels /i/, /a/ and /u/.

The *un*voiced sounds are produced in many different ways, e.g. by pressing air out through the teeth /s/, out between the lips and the teeth /f/, by sudden opening of the lips /p/, sudden opening between tongue and teeth /t/ and between tongue and palate /k/. These sounds are called unvoiced because the vocal folds do not vibrate but stays open in order for the air to pass.

8.2 Voiced sounds

The signal from the vocal cords is a periodic signal (triangular signal) with a spectrum where the components decrease in level with increasing frequency. This fall off is inversely proportional to f-squared, f^2 .

During pronunciation of (voiced) vowels the transmission channel can be modelled by a tube closed in one end (at the vocal cords) and open in the other end (at the lips). The lowest resonance frequency for such a tube occurs at ¼ of the wavelength. The distance from the vocal cords and up through the throat and the mouth to the lips is about 17 cm for an adult male. The resonance frequency is thus (340 m/s) / (4 x 0.17 m) = 500 Hz. The next resonances exist at the odd harmonics of the first and are thus 1500 Hz, 2500 Hz, 3500 Hz and so on. Figure 8-1 show such a simple description of the spectrum for a voiced sound: source spectrum, transfer function and resulting spectrum (including a 6 dB/octave radiation increase towards higher frequencies).

The peaks in the resulting spectrum are called formants. The size and the position of the formants are characteristic for the individual vowel.

The throat and the mouth cavities do not constitute a simple cylindrical tube, but the shape is more complicated. Figure 8-2 shows the cross sectional area as a function of the distance from the vocal cords when the vowels /i/, /a/ or /u/ are pronounced.

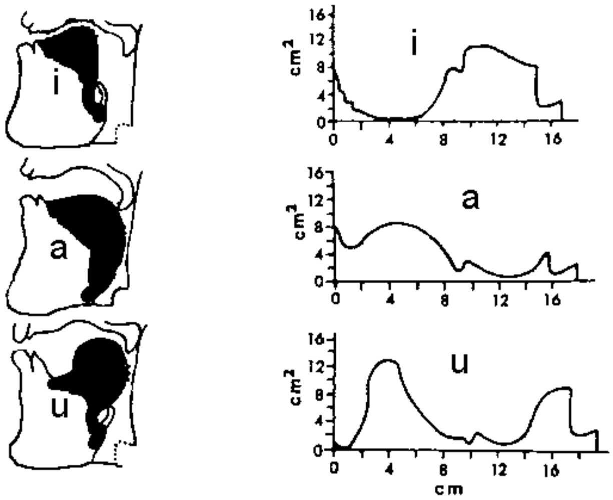


Figure 8-2. Position of the tongue (shown black) and the cross sectional area of the vocal tract. From (xx).

The transmission path may be looked upon as a combination of tubes with different length and cross section areas. The formant frequencies for these three vowels are:

	1. formant	2. formant	3. formant
/i/	225 Hz	2200 Hz	3000 Hz
/a/	700 Hz	1200 Hz	2500 Hz
/u/	250 Hz	700 Hz	2200 Hz

The formant frequencies are given here in round numbers. For male speakers the frequencies are typically somewhat lower and for females and children they are typically a little higher. Correspondingly there are similar differences between different speakers (of the same gender). In Figure 8-3, the spectrum of /a/ and /i/ is shown for two different speakers (i.e. different fundamental frequencies). It is seen that the shape of the spectrum is the same even though the fundamental frequencies are different.

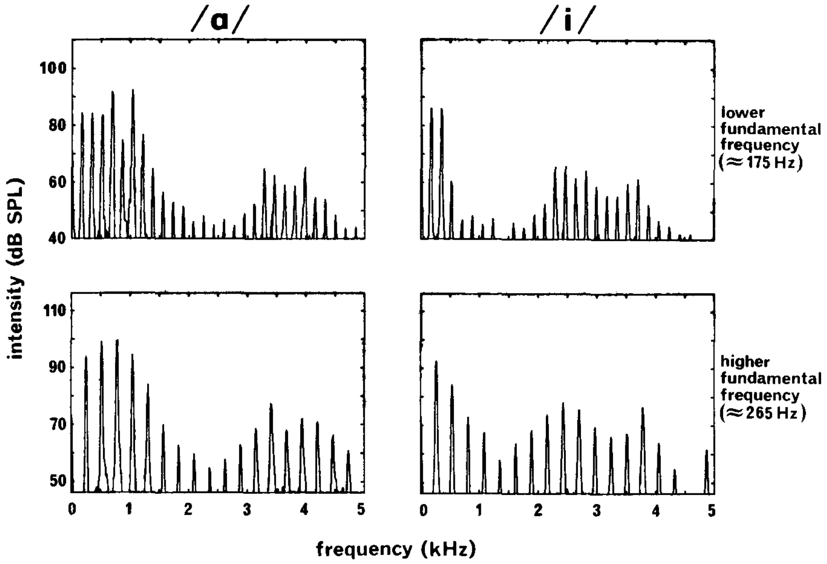
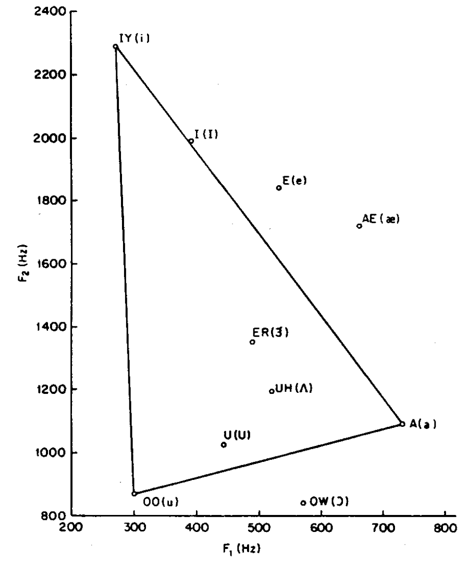


Figure 8-3. Spectrum of /a/ and /i/ for two voices with different fundamental frequency. From (Moore 1986).

The three vowels /i/, /a/ and /u/ belong to the so-called cardinal vowels. They represent three extremes in relation to tongue position and formant distribution. This is shown in Figure 8-4 where F2 is plotted as a function of F1 and where the three vowels represent a triangle.

All vowels may be characterised by the relative position and strength of the formants. This is shown in Figure 8-5 where 76 persons have pronounced 10 different vowels. It is clearly seen that each vowel may be grouped by means of the position of F2 and F1.



FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	ɪ	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	a	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

Figure 8-4. F2 as a function of F1. /i/, /a/ and /u/ constitute a triangle. From (Moore 1986).

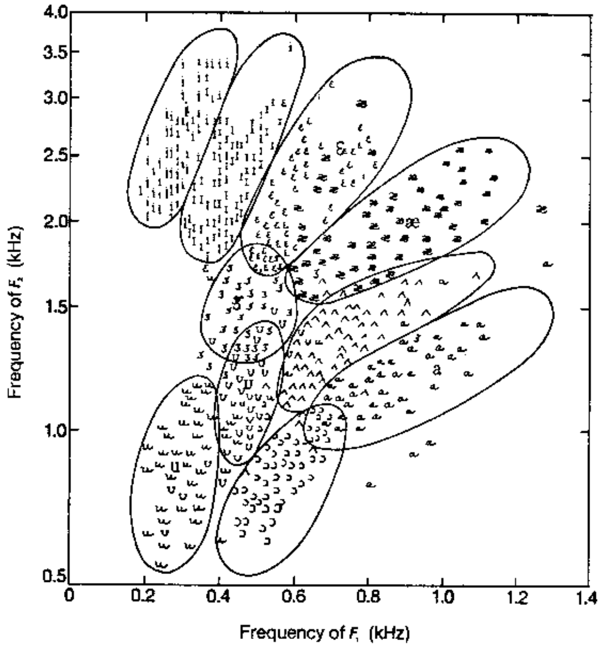


Figure 8-5. F2 as a function of F1 measured for 76 different speakers. From (Moore 1986).

The structure of the formants in the vowels has been investigated by means of a special spectrum analyzer, a sonagraph, which was developed around 1940 and which is now substituted by computer programs (giving the same graphical output). The result of the analysis – a sonagram – is shown in Figure 8-6.

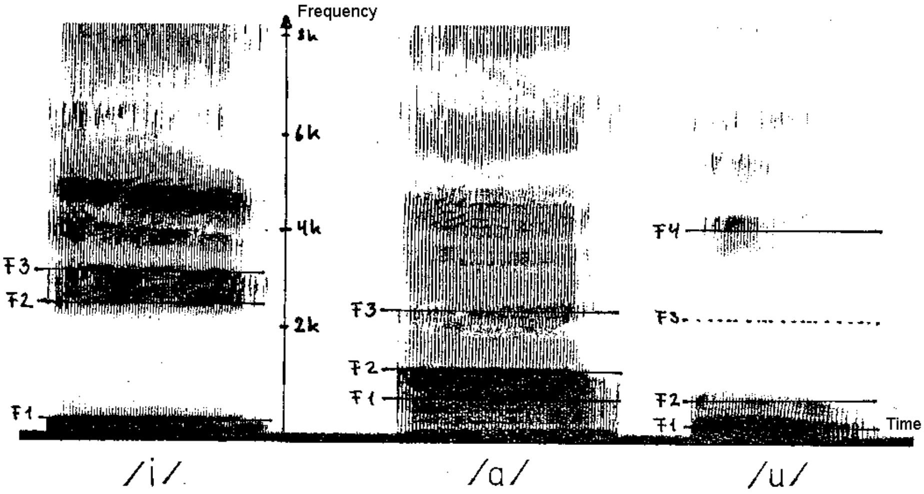


Figure 8-6. Example of sonagram recordings

The sonagram has the time as abscissa and the frequency (linear) as the ordinate. The density of the print corresponds to the intensity of the frequency component. The formants are seen as horizontal bars.

Examples of time functions and corresponding sonagrams are shown in Figure 8-7 and Figure 8-8. For the vowel /i/ there are approximately five periods within the 50 ms shown in the figure. This means that the fundamental frequency is around 110 Hz. It is possible to see the slowly varying F1 (225 Hz) overlaid by the much more rapidly varying F2 (2200 Hz). The vowel /u/ is almost a pure sine wave.

The structure of the formants is here described in a stationary situation where the vowels are pronounced one at a time. In running speech the formants will be influenced by the speech sounds that precede the vowel and the speech sound that follows the vowel. This is illustrated in Figure 8-9 where the formants of the sounds 'babb', 'dadd' and 'gagg' are shown. It is seen that the position of the formants change during the pronunciation.

Reference (Minifie et al. 1973) has a comprehensive description of acoustical and electrical models for the generation of the formants.

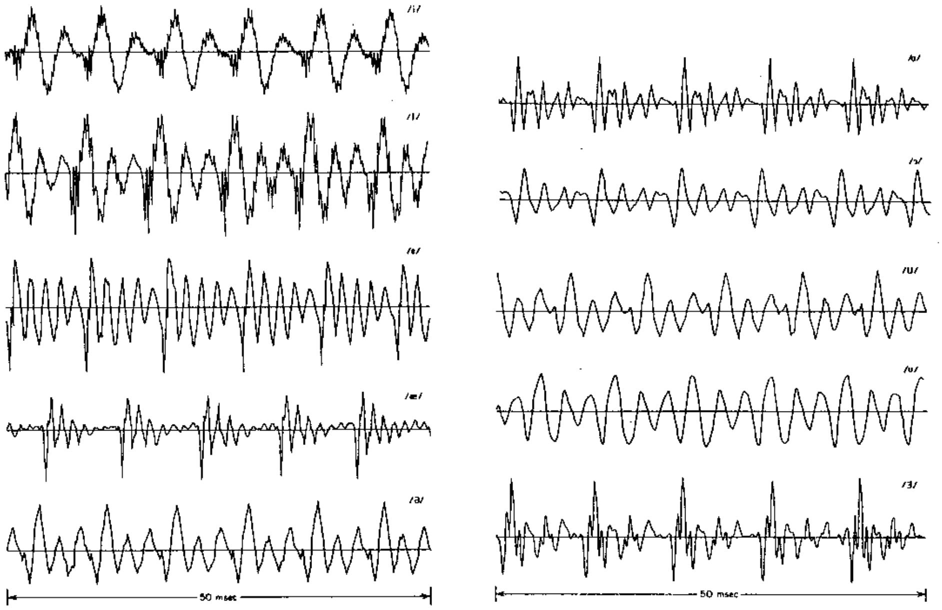


Figure 8-7. Time functions for different vowels From Rabiner and Shafer (1978).

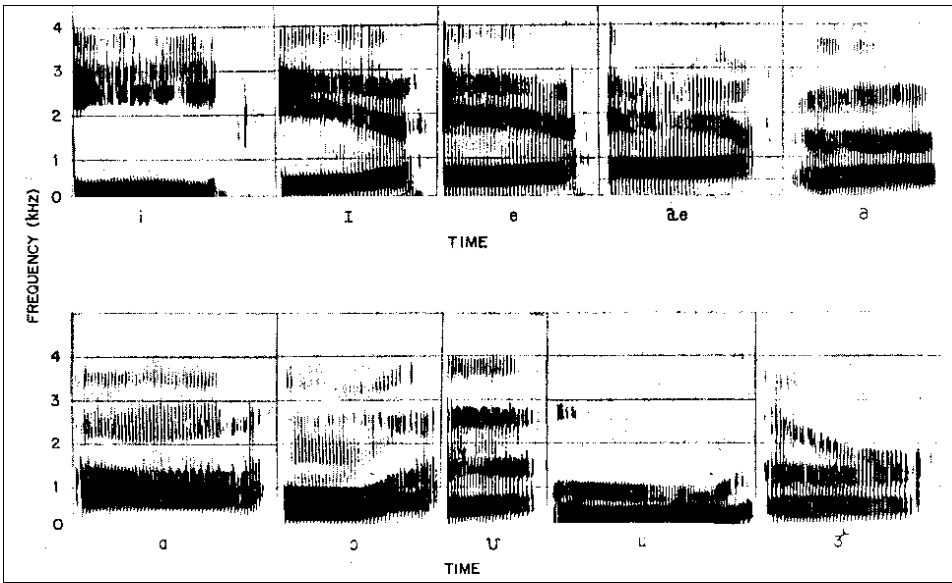


Figure 8-8. Sonagrams of the time functions in Figure 8-7

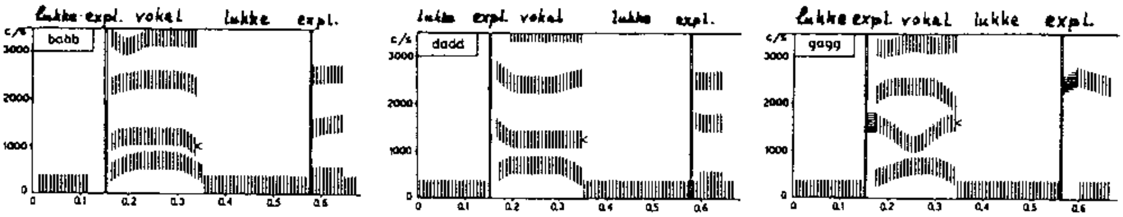


Figure 8-9. Sonagrams for 'babb', 'dadd' and 'gagg'. From (Jørgensen 1962)

8.3 Unvoiced sounds

Unvoiced sounds are defined as speech sounds where the vocal cords do not vibrate. The unvoiced sounds can e.g. be produced when air is pressed between the teeth and the lips (i.e. /f/) or between the teeth (i.e. /s/). Other examples are the sound produced when the air flow is suddenly opened: /p/ produced between the lips, /t/ produced between tongue and teeth, /k/ produced between tongue and palate. The spectrum of the unvoiced consonants is different from the formant structure. The vocal cords are usually open in this situation and the sound source is not situated in the 'bottom' of the throat but closer to the lips. The source is thus shunted by the cavities in the mouth and the throat.

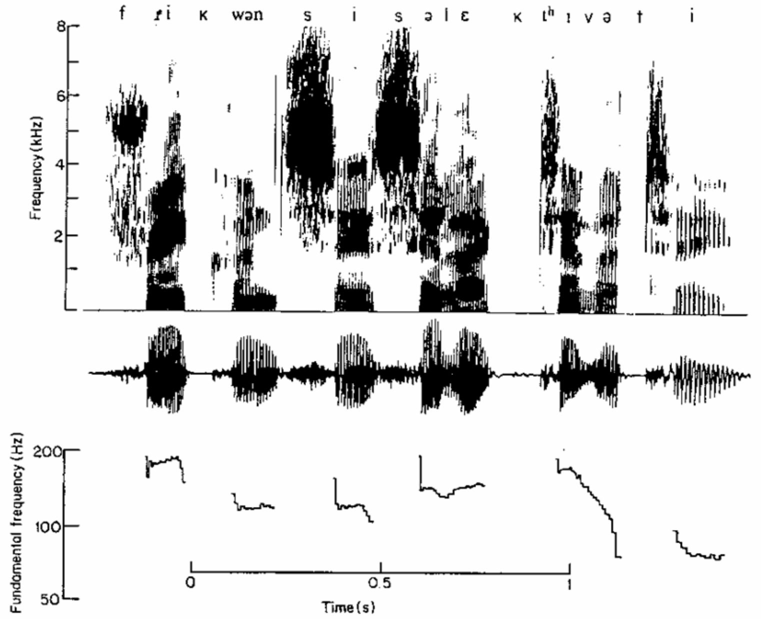


Figure 8-10. Spectrogram (Sonagram), time function and fundamental frequency course for the utterance 'Frequency selectivity'. From (Moore 1986)

Rosen & Fourcin, 1986

Figure 8-10 show the sonogram, the time function and the variation in the fundamental frequency, of the utterance 'frequency selectivity'. The high-frequency components /s/ and /t/ are easily recognised and it is also seen that these high-frequency parts have relative low amplitude.

8.4 Whispering

If the vocal cords are pressed (slightly) against each other – without closing totally – the air flow will be turbulent and produce acoustical noise. This source signal is used in whispering.

8.5 Level and dynamic range

The sound pressure level for male speech is around 65 dB measured in an anechoic room 1 m in front of the speakers mouth. This speech level (65 dB SPL) is based on a measurements of the long-term average for a great number of speakers. For women the level is on average 3 dB lower, i.e. 62 dB SPL (compare the number of lines in the spectrum). During normal speech the level will vary ±15 dB around the mean value. For a raised voice the level will increase by 10 to 15 dB and the spectrum will change.

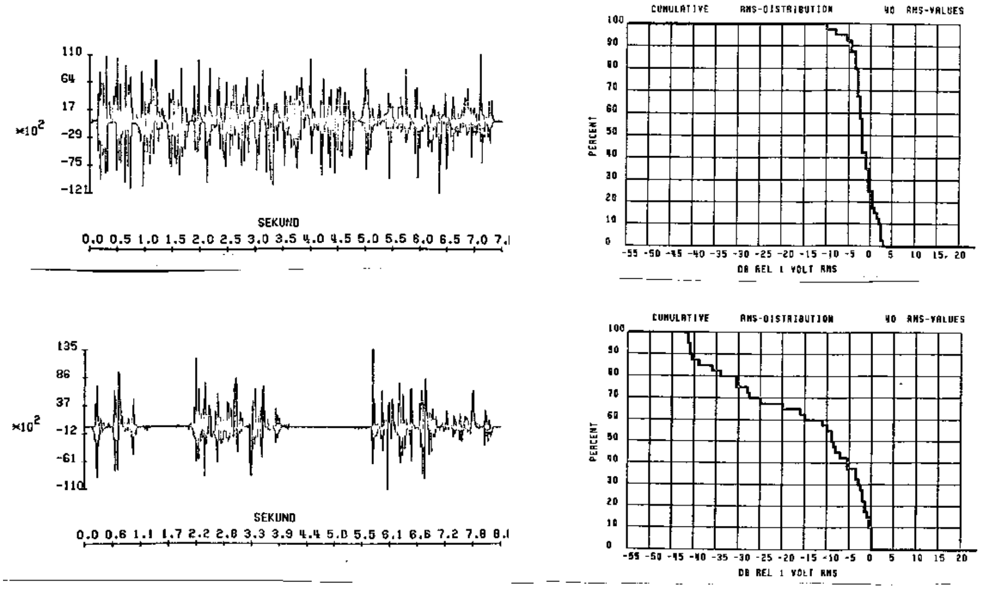


Figure 8-11. Time function and cumulative distribution of fast and slow speech. Measured with 125 ms RMS values.

The level of running speech is often given as an equivalent level, Leq, measured over a reasonable long time period (30 seconds or more). Due to the non-stationary character of the speech signal, short integration times will not give a representative result.

For the level of word lists (e.g. Dantale CD) it should be remembered that the (silent) intervals between the words constitute the major part of the duration of the list. It is therefore necessary to make a correction for the silent intervals. Usually measurements are based on a list where the intervals have been removed. See also (Ludvigsen 1987).

Figure 8-11 shows the time function and the cumulative distribution of 128 ms RMS values for two different male speech signals. One of the signals is a recording of the news from the radio (fast speech) and the other signal is a slow reading of the official weather forecast. In the first situation the dynamic range is only 15 dB whereas it is about 40 dB in the second situation. In normal running speech the dynamic range is ± 15 dB around the mean value, i.e. a dynamic range of 30 dB. (This ± 15 dB dynamic range is used in the Speech Transmission Index calculation (p.81). In the Articulation Index calculation procedure (p. 74) the range is set to +12 dB and –18 dB).

Speech as a function of time is not fully symmetrical around the zero line. It has been shown that amplitude distribution for running speech can be approximated by a gamma distribution, see

Figure 8-12.

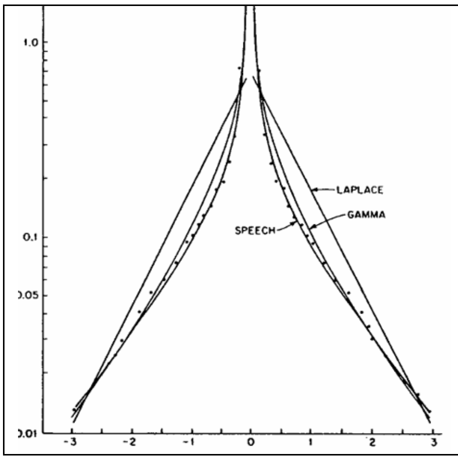


Figure 8-12. Distribution of amplitude density for speech. (Rabiner and Schafer 1978)

8.6 Long term spectrum

The long term spectrum for Danish speech is shown in Figure 8-13 for male and female voices respectively. The spectrum is shown as a density spectrum (dB/Hz). The spectrum falls off by approximately 9 dB/octave above 500 Hz and by approximately 3 dB/octave below 500 Hz.

Figure 8-14 show also the long term spectrum but in this figure it is a 1/3 octave spectrum. In this situation the spectrum falls off by 6 dB/octave above 500 Hz and the spectrum falls off by 6 dB/octave below 500 Hz. It is important to distinguish between these two spectrum illustrations. The shape of the spectrum depends also on the recording conditions (the acoustics of the room, microphone distance and microphone type).

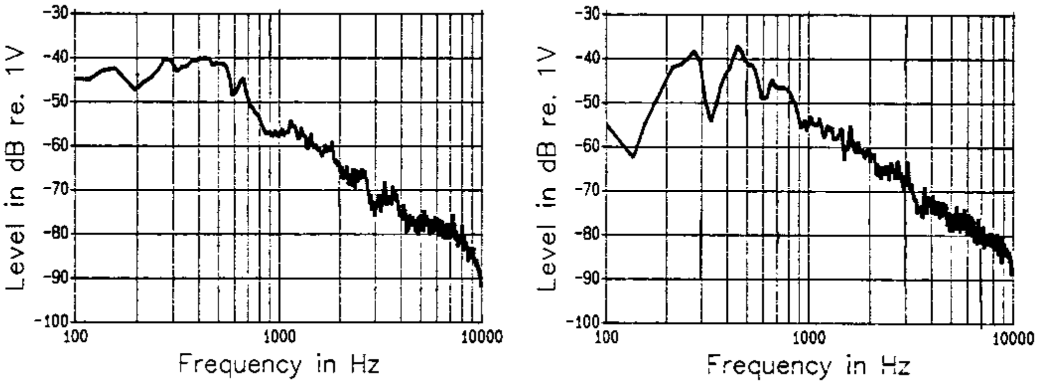


Figure 8-13. Long-term ***density spectrum*** *of Danish speech. Left: average of four male voices. Right average of four female voices.*

For measurement purposes, a speech shaped noise signal is often used. The frequency components in speech shaped noise are weighted so that the spectrum is the same as the long term spectrum of speech.

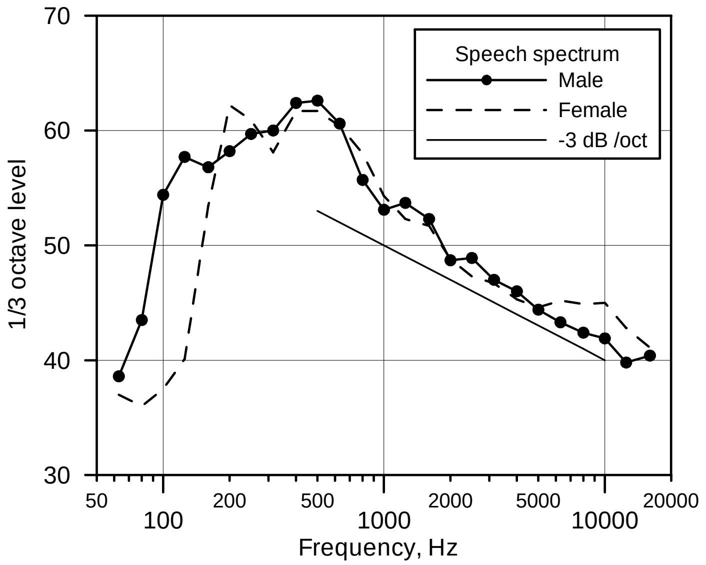


Figure 8-14. The long-term speech spectrum for male and female speech shown as a ***1/3-octave spectrum****. For comparison a line with slope –3 dB per octave (= –10 dB per decade) is shown. Redrawn from (Byrne et al. 1994)*

The general long-term speech spectrum is shown in Figure 8-14. This general spectrum is calculated from as the average of 18 speech samples from 12 languages. The spectrum is based on English (several dialects), Swedish, Danish, German, French (Canadian), Japanese, Cantonese, Mandarin, Russian, Welsh, Singhalese and Vietnamese. Almost 400 talkers participated in the investigation. Note that the spectrum is a one-third octave spectrum. This means that the curves are tilted 3 dB/octave compared to the result of a FFT-calculation. (The result of a FFT is a density spectrum).

The long term spectrum of speech is almost independent of the language. This is not surprising when the speech production mechanism is taken into account. From a European point of view, even ‘strange’ languages as Chinese and Japanese have long term spectra which resemble Danish. This is very practical for the specification of speech transmission channels.

The spectrum for women falls off below 200 Hz because their fundamental frequency typically is around 250 Hz. The maximum is found around 500 Hz for both gender and above 500 Hz the two curves are almost identical. The slope above 500 Hz is approximately minus 10 dB per decade (or -3 dB/octave).

When the voice is raised the overall level will increase, but at the same time also the (long term) spectrum will change. The same is true for loud voices and for shouting. This is shown in Figure 8-15. The overall speech levels are approximately

Normal:	63 dB SPL (for a female speaker)
Raised:	68 dB SPL
Loud:	75 dB SPL
Shout:	82 dB SPL

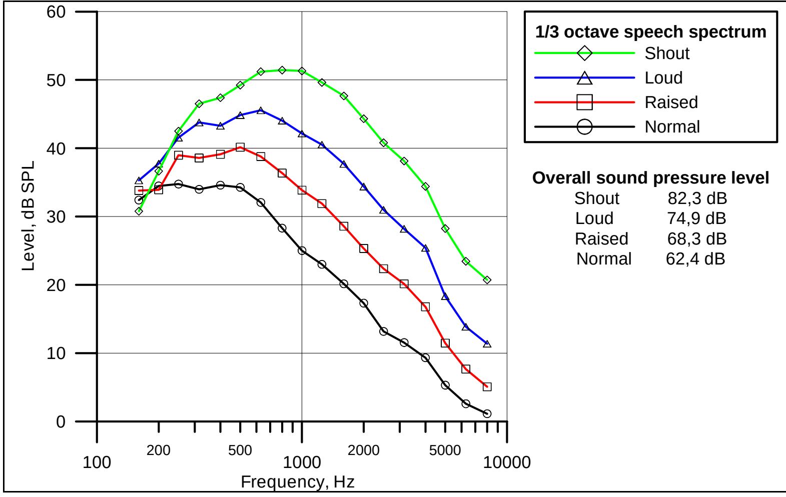


Figure 8-15. The long-term speech spectrum for normal speech level, raised and loud voices, and for shouting. From ANSI-S3.5 (1997)

8.7 Envelope spectrum

The envelope function of speech can be determined by squaring and subsequent low pass filtering (corner frequency approximately 30 Hz) of the speech signal. The envelope function describes the slow variation of the squared signal as a function of time. The spectrum of the envelope function is called the envelope spectrum or the modulation spectrum. The envelope function can also be determined by means of a Hilbert transformation.

Figure 8-16 show an example of 1/3 octave *envelope* spectra. Measurements were made both on the wideband speech signal (lin) and on the octave bands from 125 Hz to 8 kHz.

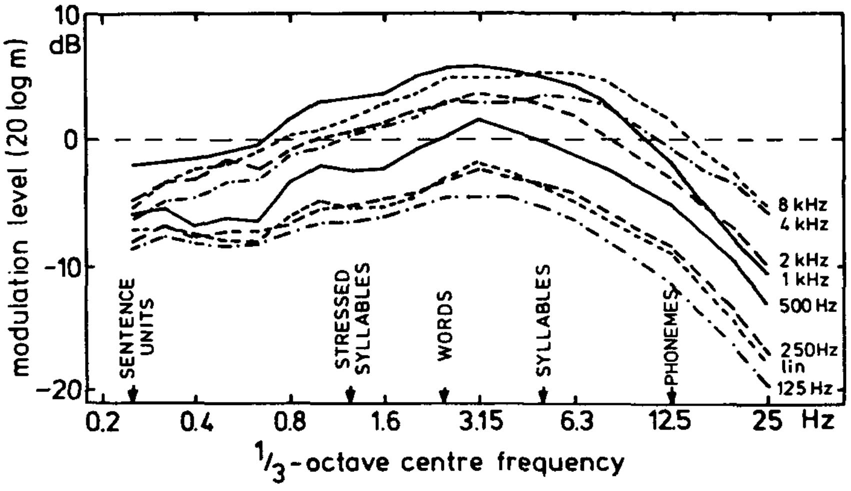


Figure 8-16. Envelope spectrum for one minute speech. Unfiltered (lin) and octave filtered (centerfrequency from 125 to 8 kHz). From (Steeneken and Houtgast 1983).

It is seen that the envelope spectrum has components between 0.3 Hz and 25 Hz. There is a maximum at (1/3 octave level) around 3 Hz and there is a rapid fall-off towards higher frequencies. The individual octave bands have almost the same envelope spectrum as the unfiltered speech signal. Only for the 4 kHz and the 8 kHz octave the maximum is shifted to around 6 Hz.

The spectral values in Figure 8-16 are calculated from RMS measurements in 1/3 octave bands. They are multiplied by √2 as this is the ratio between the RMS and the peak value (for a pure tone). The result is normalised by the mean intensity of the signal. In this way the modulation index, m, is calculated. Finally, the modulation index is transformed to a modulation level by calculating 20 log(m). Thus a modulation level of 0 dB corresponds to a modulation index of m = 1.

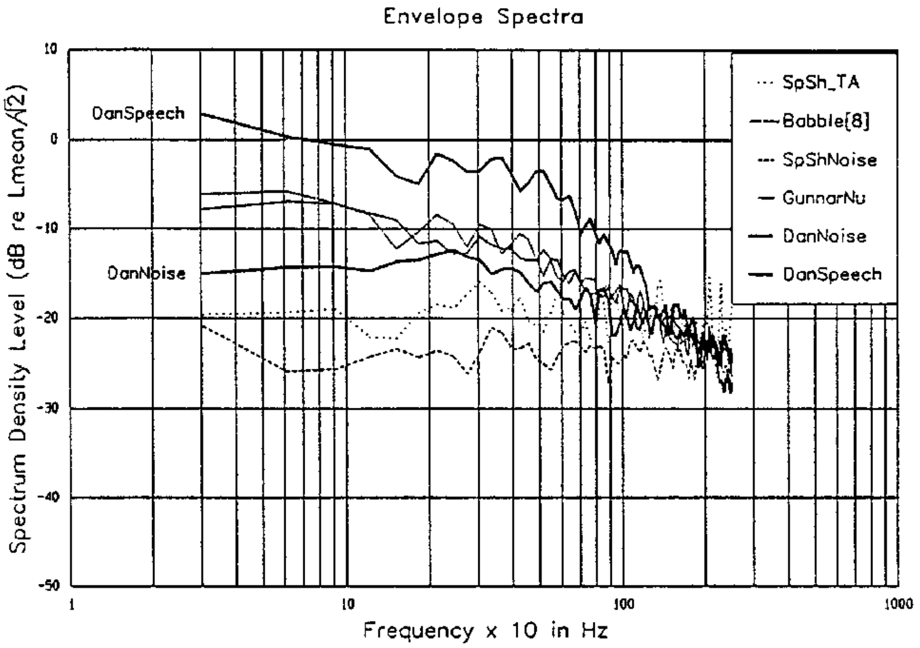


Figure 8-17. Envelope spectra for speech from simultaneous multiple speakers.

Figure 8-17 shows envelope spectra for speech signals with a different number of simultaneous speakers. The upper curve is for a single speaker. When more (simultaneous) speakers are present, the more flat the envelope spectrum will be because the low frequency modulation is reduced.

Note the difference between the 'lin' curve in Figure 8-16 and the 'DanSpeech' curve in Figure 8-17. Both curves show the envelope spectrum for a single speaker. The difference is caused by the fact that Figure 8-16 is based on 1/3 octave analysis whereas Figure 8-17 show the spectral density. A conversion between the two methods shows that the two curves are the same.

8.8 Directivity

The level of speech around the human head change with the direction. For the low frequency parts of the speech signal only little variation is seen. For the higher frequency components, a pronounced directivity is seen. This is illustrated in Figure 8-18.

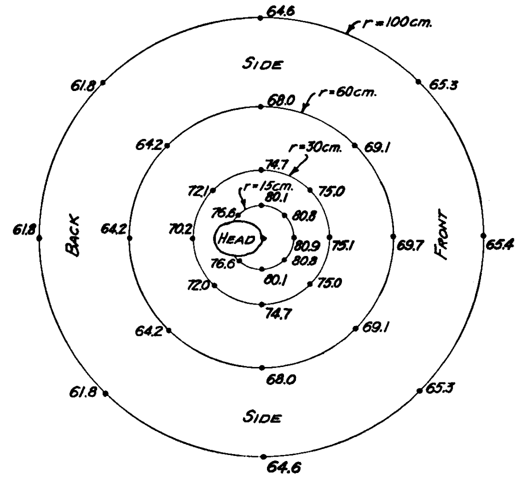


Fig. 59.—Relative Speech Intensity Levels Around the Head of the Speaker—Whole Speech.

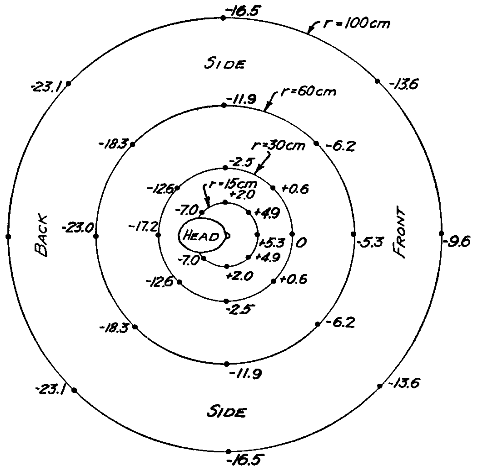


Fig. 60.—Relative Speech Intensity Levels around the Head of the Speaker. Band of Speech 2800 cps to 4000 cps.

Figure 8-18. Measurements of speech levels around the human head. The left part of the figure shows speech levels for the full speech spectrum (Whole Speech). The right part shows relative levels for the frequency band 2800 Hz to 4000 Hz. Reference position is 30 cm in front of the speaker. (From Fletcher 1953)

From the figure it is seen that the important high-frequency part of the speech is mainly radiated in the frontal direction. This effect can be demonstrated in an anechoic chamber. The consonants of speech are not heard if the talker turns around and talks away from the listener.

9. Speech intelligibility

9.1 Some phonetics

We use the written langue for written communication (as in these lecture notes) and from the spelling and sequence of the words we get the meaning of the text. When we speak to each other, on the contrary, there are a number of other factors such as intonation, pauses etc that determines how the sentences are understood. The listener has no opportunity to see the 'spelling' and there is therefore a risk that words may be confused with other words. If not all details are heard by the listener, the listener may very well obtain another meaning than the one the talker tried to express.

Statements are expressed in sentences that comprise words. The words are taken from the vocabulary of the talker and if the communication shall be successful the words shall also be in the vocabulary of the listener. The words are put together after certain rules, syntax (word order etc.). The same word may have different conjugations. E.g. 'love', 'loves', loved'. These words have a slightly different meaning. They are put together from smaller parts, which change the meaning: love, love+s, love+d. The part '+s' is also found in hate+s, live+s and '+d' is found in hate+d and live+d. These parts are called *morphemes* and are defined as the smallest parts that can change the meaning of a word.

The spoken language is made in such a way that morphemes, words and sentences can be made from a relatively little number of 'building blocks' (approximately 30-40) that are called *phonemes*. The phonemes are defined as the smallest meaning-distinctive parts of the speech signal.

In Danish there are 17 consonant phonemes and 9 short vowel phonemes. From these 26 phonemes it is possible to make 43.740 single-syllable words with a short vowel (in Danish e.g. er, at, det, skal, skrap, må, der, majs, ønsk).

	Number of combinations *before* the vowel	Example (Danish)	Number of combinations *after* the vowel	Example (Danish)
1 consonant	15	Ris	14	Av
2 consonants	36	Stå	48	Ulv
3 consonants	8	Skrap	18	Ønsk
Total	59		80	

In some words there are no consonant before the vowel or no consonants after the vowel. As there are 9 short vowel phonemes the calculation will be: (59+1) x (80+1) x 9 = 43.740 possible combinations.

In Danish only about 2000 of these combinations are real words. There is e.g. nothing in Danish whish is called 'basp', jøb', 'stung', 'tyn', 'ræst' or 'pad'. This is utilised in some speech intelligibility measurements methods where such nonsense words are used.

In the calculation above no distinction is made between words with and without glottal stop (Danish: stød), i.e. the Danish ‘man‘ and ‘mand’ is only counted once. If glottal stop, long vowels and multi syllable words are taken into account, it will result in an enormous number of words which all are made from the same (small) number of phonemes. The phonemes can be illustrated in the following (simplified) Figure 9-1.

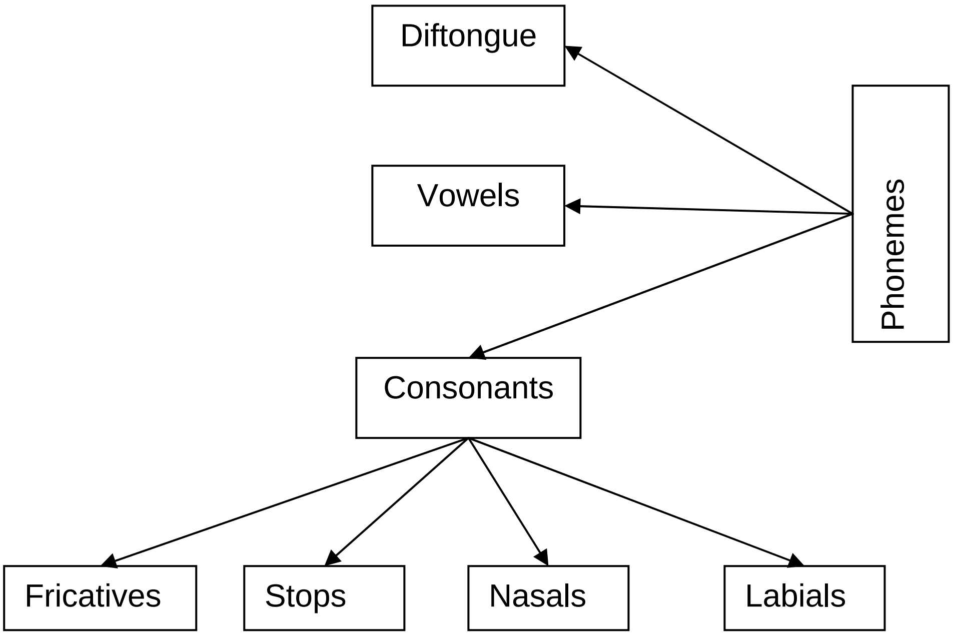


Figure 9-1. Overview of phonemes

Both the fricatives and the stop sounds can be subdivided into voiced and unvoiced sounds. Acoustically it can be said that the consonants only exists because of the vowels they are connected to. If one takes the Danish word SALE and cuts from the beginning of the start consonant the word will be heard as SALE, TALE, DALE, ALE. If the words MÅDE and NÅDE are recorded in a computer and we then swap the parts of the signals that belong to ‘M’ and ‘N’ (i.e. MÅDE becomes N+ÅDE and NÅDE becomes M+ÅDE), then N+ÅDE will still sound like MÅDE and M+ÅDE still sound like NÅDE.

The language contains a certain amount of excess information. This redundancy makes it possible to understand the meaning even though not all words or details are being heard. It is also the redundancy which makes it possible to find spelling errors in a text.

The following table gives some terms used to describe speech sounds:

Dentals	Tooth sounds, t, d

Diftongue	Double sound, vowel plus j, v or r *(English examples??)*
Fricative	Friction sound, f, v, s, j, h
Intonation	Sentence melody. Questions will have an increasing pitch towards the end of a sentence. 'He comes?' compared to 'he comes'
Labials	Lip sounds. P, b
Laterals	The sound comes from the side of the mouth when the tongue is put against the palate. L-sound.
Stops	Stop sound, p, t, k (with aspiration), b, d, g without aspiration. All unvoiced
Nasals	Voiced consonants, m, n.

9.2 Definition of speech intelligibility

The elements in a speech intelligibility test is shown in Figure 9-2

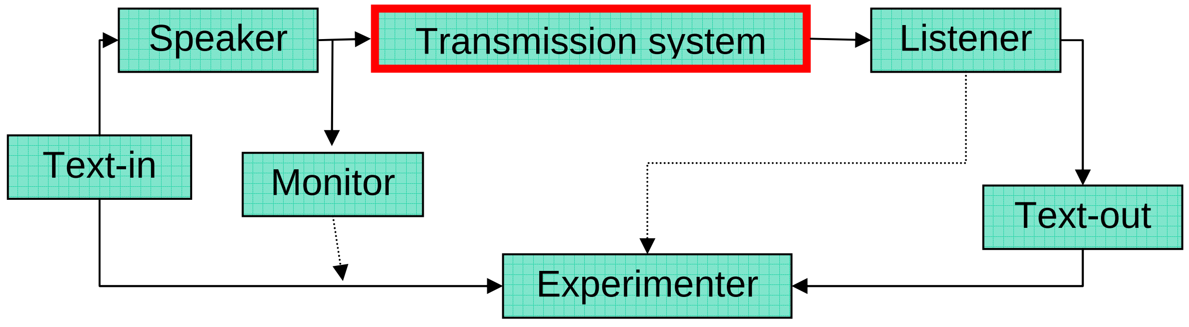


Figure 9-2. Illustration of the elements that are involved in a speech intelligibility test.

A speaker read aloud the word material which reaches the listener through the transmission channel. The listeners recording of what he/she hears is then compared with the speaker's word material. Speech intelligibility is defined as the percentage of the words that are understood correctly. More precisely it can be expresses as:

Speech intelligibility (%) = (100/T) *R

where T is the number of words in the test and R is the number of correct words.

In the clinic the speech intelligibility is often given as the deviation from 100% intelligibility and is called Discrimination Score (DS) or Discrimination Loss (DL). In English literature also the designations Articulation Score and Percent Correct are seen.

Speech intelligibility is an important parameter in speech communication systems and it is thus important to be able to measure this property. From Figure 9-2 it is seen that the transmission channel alone constitute only a minor part of the whole system and unfortunately the speech intelligibility depends also on a number of other factors. Some of these are:

- Word material (sentences, words, numbers, logatomes)
- Presentation method (single words, in a text)
- Open/closed response test (‘infinite’ or limited number of words)
- Speaker (pronunciation, talking speed, speech level, dialect)
- Listener (hearing ability, training, vocabulary, dialect)
- Scoring method (word, phoneme, oral, written)

Figure 9-3 show a typical course of the speech intelligibility as a function of the signal level. The curves show an S-shaped course where the intelligibility changes from 0% to 100%. For persons with a cochlear hearing loss the intelligibility may not reach 100% but decrease if the level is increased too much. The full lines in Figure 9-3 are for single syllable words and the dashed lines are for two-syllable words (also called spondees).

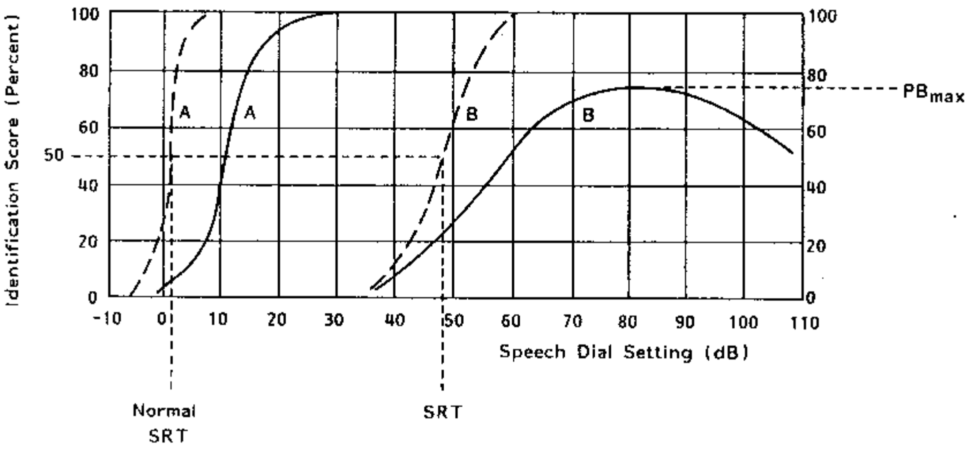


Figure 9-3. Speech intelligibility in % as a function of the signal level (Speech Dial Setting). A: normal hearing. B: hearing impairment. SRT: Speech Recognition Threshold defined as the level where 50% intelligibility is obtained.

9.3 Word material

The intelligibility of a word depends on whether it is heard in isolation or it is part of a meaningful sentence. This is illustrated in Figure 9-4. In a meaningful sentence it is often possible from the redundancy of the language to guess the word even though the word is not heard or part of the word is misinterpreted.

The language redundancy plays also a role in multi-syllable words. For a given S/N the intelligibility will increase with increasing number of syllables.

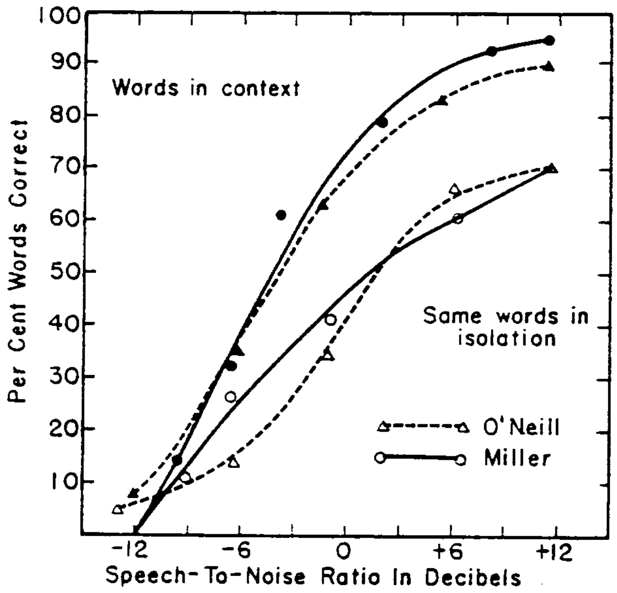


Figure 9-4. Intelligibility of words heard in isolation or in a context. (Malmberg 1970)

The size of the word material has also an influence on the intelligibility. If the listener knows that only the single-syllable numbers from 0 to 10 are included in the test a higher intelligibility will be measured compared to a test where all possible single-syllable words are included. This is illustrated in Figure 9-5.

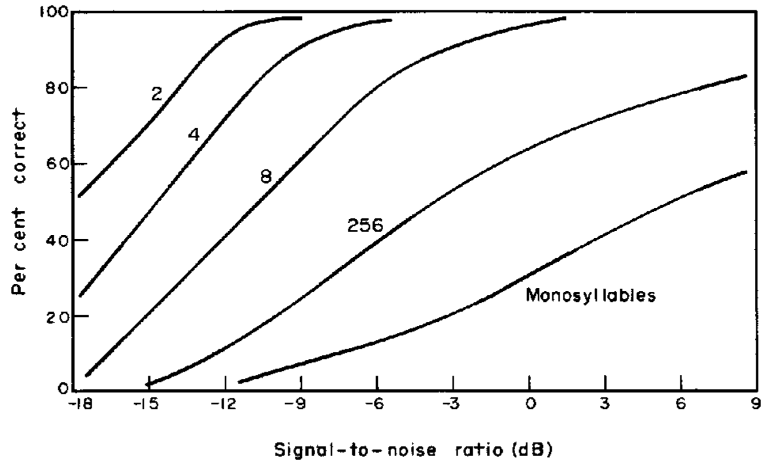


Figure 9-5. Intelligibility as a function of the size of the vocabulary (Ainsworth 1976)

In speech intelligibility tests, it is necessary to use a well defined word material. Either the words shall be part of the listener's active vocabulary or the words shall not be known by the listener. The unknown word situation can be obtained by means of constructed artificial (single-syllable) words called logatomes. They are constructed from a consonant-vowel-consonant paradigm, CVC. The first and/or the last consonant may be a double consonant. Examples of such constructed (Danish) logatomes are shown in the list below:

PØV	SKASK	NUS	GONG
NØST	DOSP	JØN	LASP
TYV	SKUL	REG	KÆV
HAS	STET	FIK	VIM

Among such constructed CVC words some will be meaningful words (e.g. the Danish words TYV and FIK in the list above) and will increase the intelligibility. It is common practice to remove these meaningful words from the CVC word material and only use the remaining nonsense syllables or nonsense words. If the words furthermore are selected so that the frequency of the various phonemes in the material is the same as in the normal language the material is said to be phonetically balanced, PB.

In certain applications, the nonsense word is included in a carrier sentence that is otherwise meaningful. Some Danish examples are given below:

- Vi tager hinanden MOP I hånden og går
- En skov kan BASP jo også have sin virkning
- Ellers er her JØB næsten som du havde ventet det
- Det er en STUNG temmelig vanskelig patient
- Nu vil jeg TYN huske ham altid
- Som de ser JØD er det en ganske almindelig mand
- Vi har et PAD temmelig stort hus
- Du kan jo RÆST nok forstå de hader ham

In these examples the nonsense word is included as the 4th word in the sentence. Carrier sentences are well suited for the determination of speech intelligibility in room with a long reverberation time. The sound field is built up by the first words in the sentence and the nonsense word is masked in the same way as in running speech. The method also has the advantage that the listener will know when the nonsense word is presented. The carrier sentence principle may also be used in test of speech-controlled systems, in automatic gain control systems (AGC), adaptive noise reduction and the like. The carrier sentence principle is not limited to nonsense words.

See ANSI-S3.2 (1989) for different speech materials. The standard specifies test materials that have been thoroughly validated. The standard also specifies methods for selecting and training the talkers and listeners; for designing, controlling, and reporting the test conditions; and for analyzing and reporting the test results.

9.4 Dantale I

For use in speech intelligibility measurements (e.g. speech audiometry), a CD has been produced with Danish words. The CD, called Dantale I, contains eight word lists (each with 25 words) for adults, three lists with numbers, four lists with words for children and one list with words for small children. There are furthermore three tracks with running speech. The Dantale CD is described in (Elberling et al. 1989)

For the adult lists 200 words are selected among common single-syllable words (nouns, adjectives and verbs). The words are distributed in such a way that all lists contain (almost) the same amount of double consonants, labials, stops (plosives), /s/ or /t/. The word in the children lists are selected as word pairs where only one phoneme differ in the two words (minimal pairs). Danish examples are: HÅR FÅR, KÆLK MÆLK, VAND MAND, STEN BEN. For masking purposes the CD contain a noise signal with (almost) the same spectrum as the speech. The noise is signal is amplitude modulated corresponding to the envelope of four speakers talking simultaneously.

The CD also contains a number of calibration signals which are necessary for a well controlled measurement. See also reference (Elberling et al. 1989; Keidser 1993).

Often the intelligibility score is given as a function of the signal-to-noise ratio (SNR). An example of this is shown in Figure 9-6 for the word-material on the Dantale CD. This CD contains eight tracks of 25 words each. The words are common Danish single-syllable words that are distributed phonetically balanced over the eight lists so that the lists can be regarded as equivalent. The words are recorded on the left channel of the CD and on the right channel a noise signal is recorded with (almost) the same spectrum as the words. The noise signal is amplitude modulated in order to make it resemble normal speech for four concurrent speakers.

The result in Figure 9-6 is obtained with the words and the noise on the Dantale CD with untrained Danish normal hearing listeners. It is seen that even at a signal-to-noise ratio of 0 dB almost all words are understood. It is also seen that an increase of just 10 dB in SNR can change the situation from impossible to reasonable, e.g. from -15 dB (10%) to - 5 dB (70%). It is a general finding that such a relatively small improvement of the signal-to-noise ratio can improve the intelligibility situation dramatically. In other words, if the background noise in a room is a problem for the understanding of speech in the room, then just a small reduction of the background noise will be beneficial.

9.5 Dantale II

Another Danish speech test , called Dantale II, has been developed and is available on a CD. Sentences consisting of five words are used in this test. Each sentence is a combination of a name, a verb, a number, an adjective, a substantive. The test subject repeat the words that is understood and the experimenter counts the number of correct words. An adaptive procedure is used to determine the speech reception threshold, SRT. The noise signal is non-modulated speech shaped noise. See Wagener, Josvassen and Ardenkjær (2003) for details.

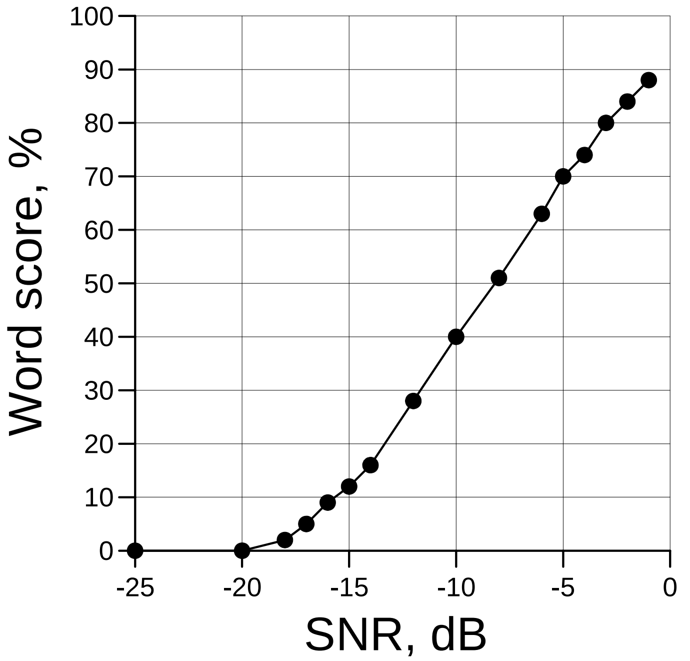


Figure 9-6. Word score for the speech material DANTALE as a function of speech-to-noise ratio (SNR). Redrawn from (Keidser 1993)

9.6 Measuring methods

The hearing threshold for speech is defined as the level where 50% of the words are understood correctly. This is called the Speech Recognition Threshold, SRT, (IEC 645-2, 1993). The threshold may be measured by means of an up-down procedure where the intelligibility is measured over 10 presentations and where the level is increased or decreased depending on whether the intelligibility is above or below 50%. The SRT may also be determined by a Bekesy procedure with running speech where the listener (the test subject) shall try to keep the intelligibility at 50%.

The intelligibility at a given level (or S/N) is determined by presentation of a list of words (e.g. 25 words) and then count the number of incorrect responses. This number given in percent is the discrimination loss (DL). This scoring method is called word score and a word can be either right or wrong.

Instead of counting (whole) words it may be advantageous to count the number of correct (or wrong) phonemes. This is called phoneme score. For single-syllable words, a so-called triple-score method can be used. In this case, the scoring is based on the initial consonant(s), the vowel, and the final consonant(s).

A method where the test subject in principle can answer with all possible words is called an open test. Contrary to this, a test is called closed if the test subject must choose between a limited number of words. The responses may e.g. be limited to the numbers between 1 and 10.

Confusion of different phonemes may be investigated by a closed test where the possible words rhymes, a rhyme test. In English this could be CAT, SAT, FAT, THAT. Some Danish examples with four alternatives are given below:

BUS	BUSK	HUS	MUS
BLÅ	GRÅ	GÅ	NÅ
NÆB	VÆG	ÆG	KNÆ
BIL	HVID	ILD	PIL
HEST	FEST	GÆST	RIST
NAT	VAT	HAT	KAT

Experience shows that consonants of the same type (fricatives, stops, etc.) are often confused under difficult listening conditions, i.e. in a bad S/N. With a rhyme test it is possible to study the detailed confusions between the initial consonants and between the final consonants.

Typical confusions that occur in all speech intelligibility measurements because the words are pronounced in the same way:

High	hi
See	sea
Way	weigh
No	know

9.7 Statistical issues

When speech intelligibility test are used to evaluate the quality of a transmission system, it is important to use the results with some cautions. The test could be an evaluation of two loudspeaker systems in a room. Or it could be an evaluation of two different hearing aid fittings. In such cases a change in the intelligibility from e.g. 50% to 60% may not be a real improvement, but could just be a matter of statistical uncertainty in the test. This is illustrated in Figure 9-7.

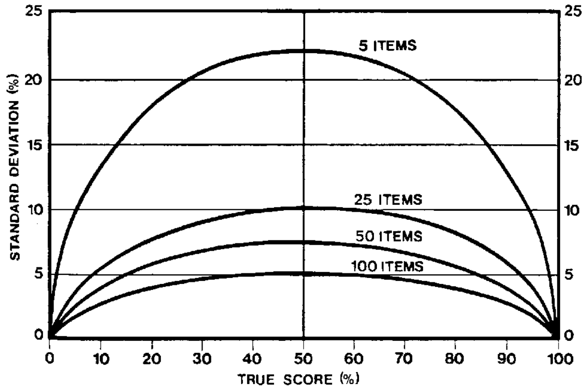


Figure 9-7 Standard deviation of the score as a function of the score.

When the score is very low (i.e. almost nothing is understood), the variance in the data is small. The same is true at a very high intelligibility (almost everything is understood). In the range around 50%, which is the interesting one for speech intelligibility tests, the variance is much bigger. The standard deviation is about 10% for a word list with 25 words (as in DANTALE). This means that in order to see a real difference, the change must be greater than two standard deviations, i.e. 20%.

In Figure 9-8 an example is shown where number of units (e.g. words) is 50. The different distributions are clearly seen.

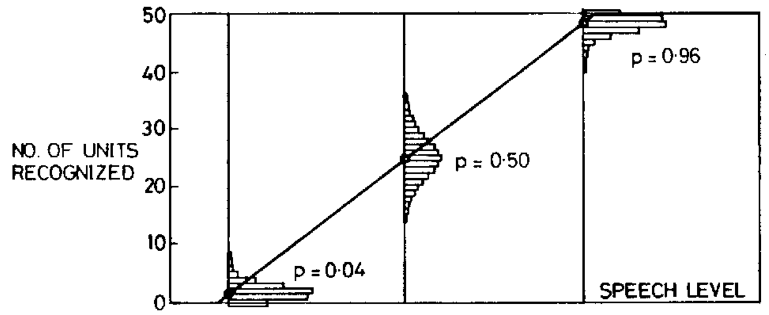


Figure 9-8. Illustration of the different width of the distribution for an example with 50 units

For more discussion of the statistical matters, see chapter 2 in reference (Martin 1987)

9.8 Distortion

The speech intelligibility may be reduced if the speech signal is distorted. Figure 9-9 show the intelligibility of nonsense syllables under low-pass and high-pass conditions. From the figure it is seen that high-pass filtering from about 1.000 Hz will only reduce the intelligibility to about 80%. If the high-pass filter is moved to 3.000 Hz the intelligibility will be reduced to 20%. A low-pass filtering at 3.000 Hz will have almost no influence on the intelligibility whereas a low-pass filtering at 1.000 Hz will reduce the intelligibility to about 30%.

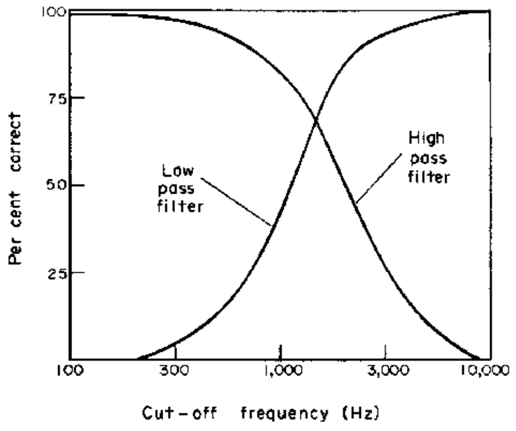


Figure 9-9. Intelligibility for low- and high-pass filtered speech.

Peak clipping of a speech signal will have a minimal influence on the intelligibility. Figure 9-10 (left) shows a speech signal that is peak clipped and even in the situation where the signal becomes a square wave the intelligibility is good (but the sound quality is of course very bad). It is the correct position of the zero crossings that secures the intelligibility. Measurements with signals where the zero crossings have been moved away from their original positions (Figure 9-10, right) show a very bad intelligibility.

The speech intelligibility of a transmission system is usually measured by means of a list of words (or sentences) where the percentage of correctly understood words gives the intelligibility score. The transmission system could be almost anything, e.g. a telephone line or a room. The intelligibility depends on the word material (sentences, single words, numbers, etc.), the speaker, the listener, the scoring method and the quality of the transmission system.

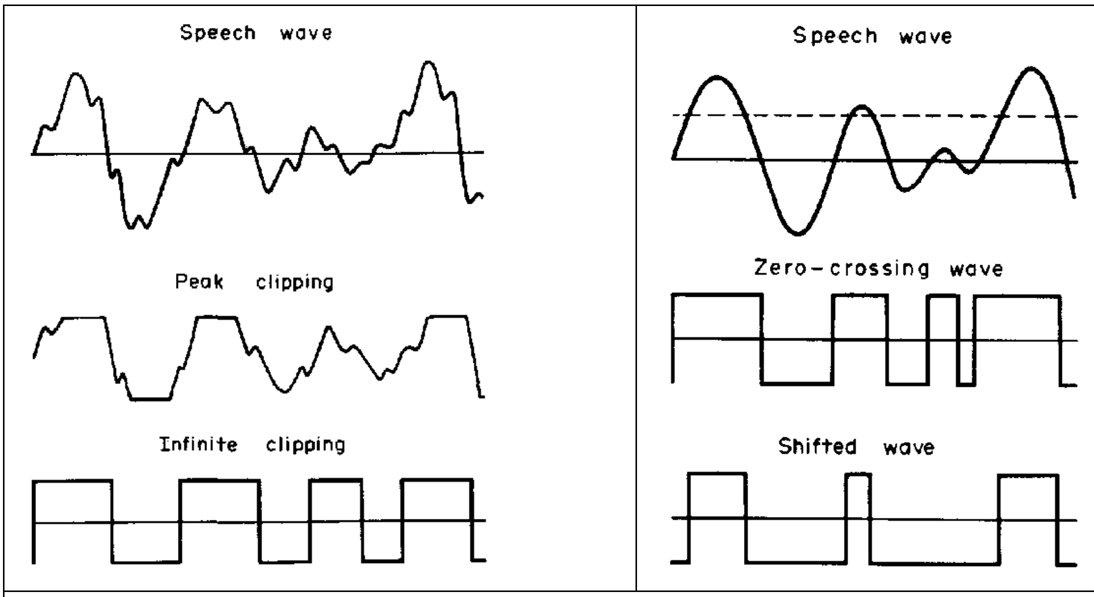


Figure 9-10. Left: Peak clipping of a speech signal. Right: Peak clipping and shift of zero crossings. (Ainsworth 1976)

It is time consuming and complicated to measure speech intelligibility with test subjects. Therefore measurement and calculation methods have been developed for the estimation of the expected speech intelligibility in a room or on a transmission line.

Articulation Index, AI (ANSI-S3.5 1969): Determination of the signal-to-noise ratio in frequency bands (usually one octave or one-third octave). The SNR values are weighted according to the importance of the frequency band. The weighted values are added and the result normalised to give an index between zero and one. The index can then be translated to an expected intelligibility score for different speech materials.

Speech Intelligibility Index, SII (ANSI-S3.5 1997): This method is based on the AI principle, but the weighting functions are changed and a number of ‘corrections’ to the AI-method are implemented. One of these is the correction for the change in speech spectrum according to the vocal effort (shouting, raised voice, low voice).

Speech Transmission Index, STI (Steeneken and Houtgast 1980): In this method the modulation transfer function, MTF, from the source (the speaker) to the receiver (the listener) is determined. The MTF is determined for octave bands of noise (125 Hz to 8 kHz) and for a number of modulation frequencies (0,63 Hz to 12,5 Hz). The reduction in modulation is transformed to an equivalent signal-to-noise ratio and as in the AI method these values are added and normalised in order to yield an index between zero and one. The index can then be translated to an expected intelligibility score for different speech materials.

Rapid Speech Transmission Index, RASTI (IEC-268-16 1988): This is an abbreviated version of STI. Only the frequency bands 500 Hz and 2 kHz and only nine different modulation frequencies are used. The result is an index which is used in the same way as in STI.

10. Articulation Index, AI

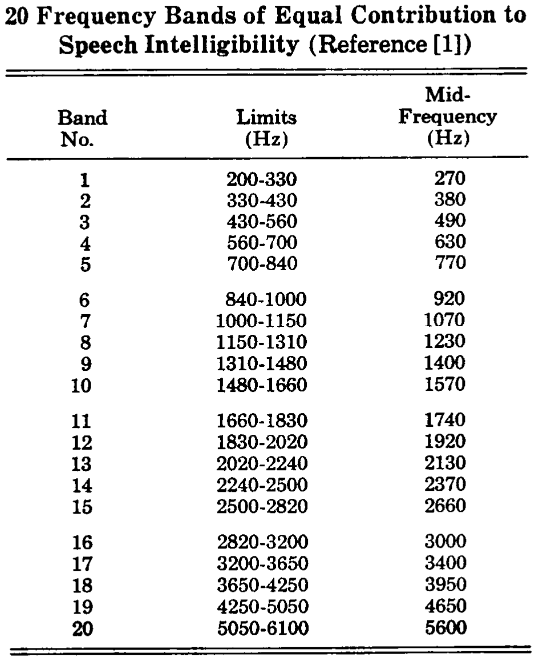
10.1 Background

The American standard, ANSI S3.5 – 1969, describes a method for calculation of the expected speech intelligibility in a transmission system. The idea behind the method is 1) that background noise can influence the intelligibility due to masking and 2) that not all frequency components in the speech signal are equally important for the intelligibility. Based on knowledge of the background noise and knowledge about the masking phenomena it is possible to make an estimate of the speech intelligibility. The method was originally developed at the Bell Laboratories in the USA with the aim of evaluate the intelligibility in telephone lines.

In the ANSI standard the method is described in various versions. A choice must be made depending on which objective measurement data that are available.

10.2 The 20 band method

This version is based on the assumption that the speech signal can be divided into 20 bands who each contribute equally to the speech intelligibility. A requirement is that the peak value of the speech signal (i.e. the 1% fractal for 125 ms RMS values) in each band is at least 30 dB above the hearing threshold in the band. In other words the speech shall be clearly audible. Figure 10-1 show the band limits of the 20 bands.



20 Frequency Bands of Equal Contribution to Speech Intelligibility (Reference [1])

Band No.	Limits (Hz)	Mid-Frequency (Hz)
1	200-330	270
2	330-430	380
3	430-560	490
4	560-700	630
5	700-840	770
6	840-1000	920
7	1000-1150	1070
8	1150-1310	1230
9	1310-1480	1400
10	1480-1660	1570
11	1660-1830	1740
12	1830-2020	1920
13	2020-2240	2130
14	2240-2500	2370
15	2500-2820	2660
16	2820-3200	3000
17	3200-3650	3400
18	3650-4250	3950
19	4250-5050	4650
20	5050-6100	5600

Figure 10-1. Frequency bands that contribute equally to the intelligibility

The principle of the method is now to determine which part of the speech (i.e. how many of the 20 bands) that are masked by a background noise. Figure 10-2 is used for this.

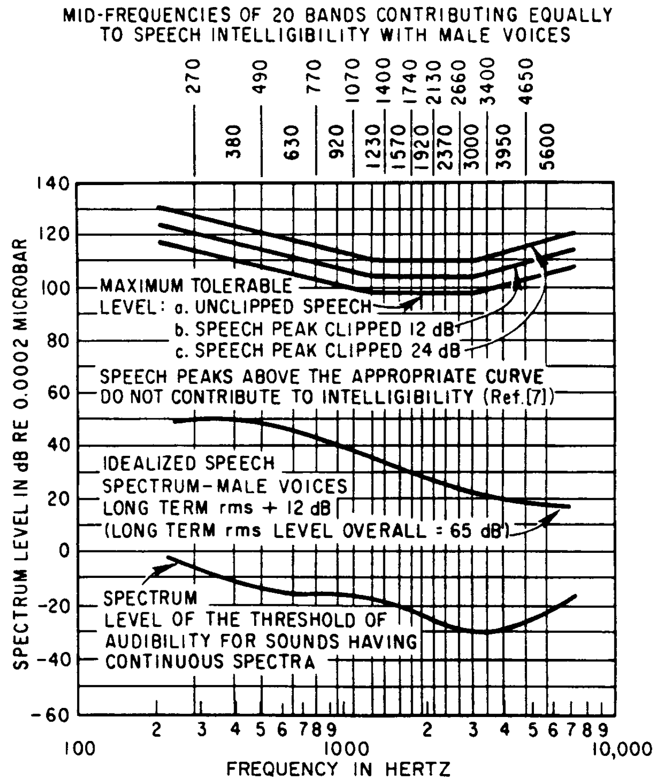


Figure 10-2. Work sheet for the AI calculation. 20 band method.

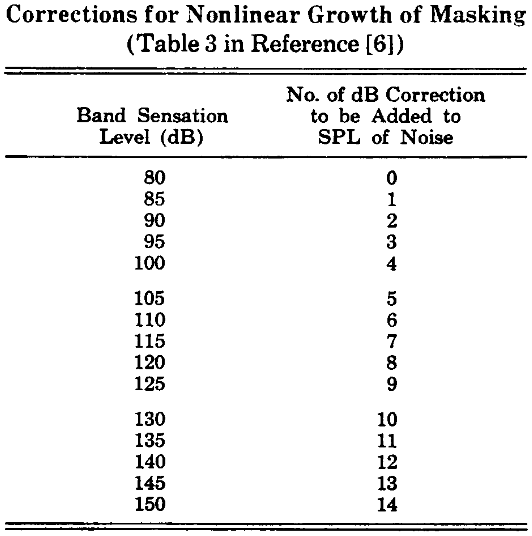
The work sheet, Figure 10-2, is a graph where two curves must be plotted:

1. The long term RMS level of the speech signal + 12 dB, measured (or estimated) at the ear of the listener in each of the 20 bands.

2. The long term RMS level of the background noise at the ear of the listener in each of the 20 bands

The speech level may be estimated from the middle curve in figure 10.2. This curve show a an idealized speech spectrum with a mean level of 65 dB SPL measured 1 meter in front of a male speaker. The speech spectrum here is the 1% fractal of 125 ms RMS values (here called the peak level) which is the long term mean value + 12 dB. If the speech is transmitted through a channel with a given frequency characteristic and amplification, the speech spectrum must be corrected accordingly.

The level of the background noise must be corrected for the nonlinear course of the masking at levels above 80 dB, see Figure 10-3.



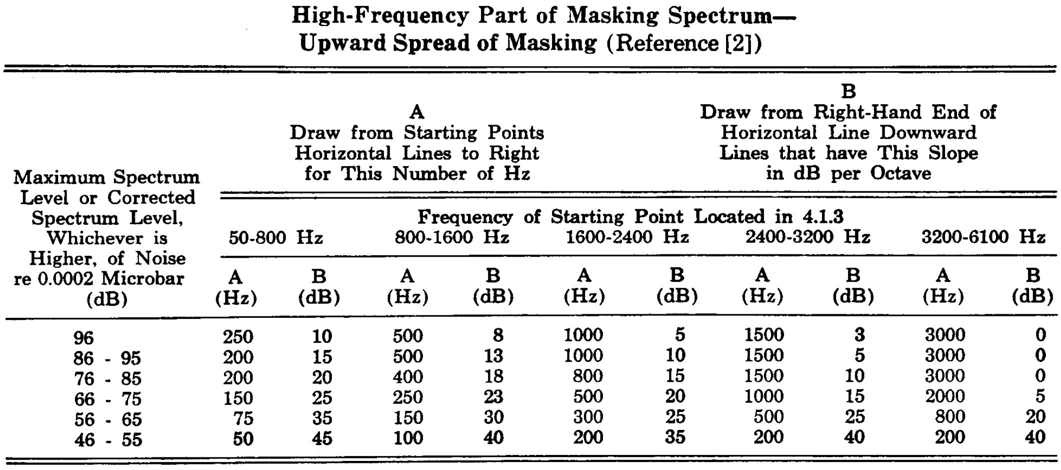
Corrections for Nonlinear Growth of Masking (Table 3 in Reference [6])

Band Sensation Level (dB)	No. of dB Correction to be Added to SPL of Noise
80	0
85	1
90	2
95	3
100	4
105	5
110	6
115	7
120	8
125	9
130	10
135	11
140	12
145	13
150	14

Figure 10-3. Corrections to account for the nonlinear growth of masking at high levels

On the basis of the (possibly corrected) noise spectrum a new spectrum is constructed which results in essentially a curve of the masking threshold. (This is a little difficult to explain but easier to do in practise):

> Draw a straight line corresponding to a level 3 dB below the maximum of the noise spectrum. This line intersects the noise spectrum at two points. The point at the higher frequency is called the starting point.
>
> From the starting point the upward spread of masking is drawn (i.e. towards higher frequencies) first as a horizontal line and thereafter as a downward sloping line. The length of the horizontal part and the slope of the falling line must be read from column A and B in Figure 10-4.



High-Frequency Part of Masking Spectrum— Upward Spread of Masking (Reference [2])

Maximum Spectrum Level or Corrected Spectrum Level, Whichever is Higher, of Noise re 0.0002 Microbar (dB)	A Draw from Starting Points Horizontal Lines to Right for This Number of Hz						B Draw from Right-Hand End of Horizontal Line Downward Lines that have This Slope in dB per Octave			
	Frequency of Starting Point Located in 4.1.3									
	50-800 Hz		800-1600 Hz		1600-2400 Hz		2400-3200 Hz		3200-6100 Hz	
	A (Hz)	B (dB)	A (Hz)	B (dB)	A (Hz)	B (dB)	A (Hz)	B (dB)	A (Hz)	B (dB)
96	250	10	500	8	1000	5	1500	3	3000	0
86 - 95	200	15	500	13	1000	10	1500	5	3000	0
76 - 85	200	20	400	18	800	15	1500	10	3000	0
66 - 75	150	25	250	23	500	20	1000	15	2000	5
56 - 65	75	35	150	30	300	25	500	25	800	20
46 - 55	50	45	100	40	200	35	200	40	200	40

Figure 10-4. Values to be used in order to take the upward spread of masking into account

The masking towards lower frequencies is drawn as a line which starts 57 dB under the starting point and has a slope of -10 dB/octave.

The masking from the noise is described by these lines together with the (corrected) spectrum. Figure 10-5 shows an example of the use of the above rules.

Example of the Calculation of an AI by the 20-Band Method

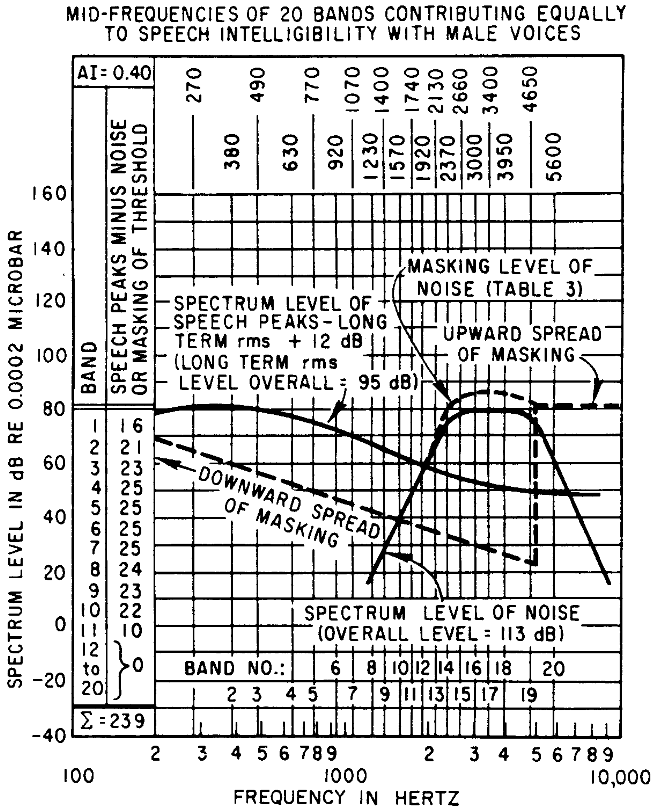


Figure 10-5. Example of the use of the work sheet

If the noise has more than one maximum, the procedure with finding the starting point and drawing of masking lines must be repeated for each maximum. If there is no background noise this drawing exercise can be omitted and the absolute hearing threshold shown in figure 10.2 is used instead.

We now have two curves in the figure: the speech level and the masking threshold. The difference (or distance) between these two curves are used for the estimation of the speech intelligibility. This is done in the following way:

For each of the 20 bands the difference in dB between speech level and masking level is determined. If the difference is 0 or negative it is set to 0. If the difference is greater than 30 dB it is set to 30 dB.

The 20 differences are then added and the result is normalised to a value between 0 and 1 by dividing by 600 (= 20 bands * 30 dB). The value is called Articulation Index, often abbreviated to AI.

The number 600 is caused by the fact that an ideal system with a 30 dB dynamic range in all of the 20 bands must give a value of 1. An articulation index of 1 corresponds to the best possible speech intelligibility. An index of 0 corresponds to no intelligibility. Figure 10-6 show the connection between the Articulation Index and a number of word materials and measuring methods.

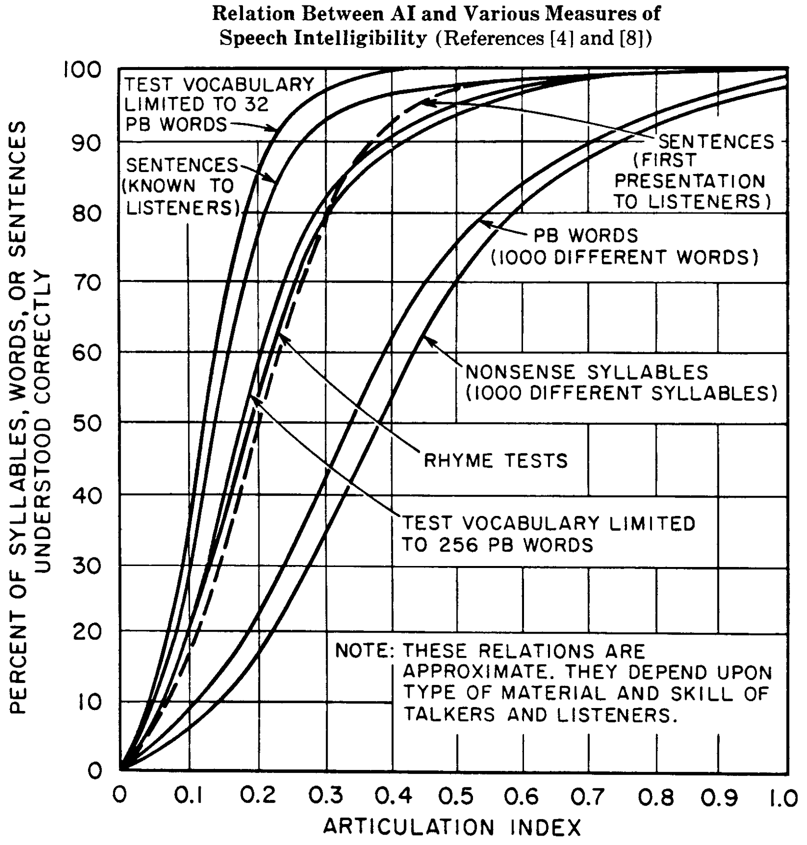


Figure 10-6. Relation between speech intelligibility (in %) and articulation index, AI

10.3 1/1-octave and 1/3-octave method

The 1/1-octave and the 1/3-octave versions of the AI method are just simple modifications of the 20-band method. These versions are included in the ANSI standard because filters corresponding to the 20 bands are not available in practise, whereas 1/1- and 1/3-octave filter are easily accessible. The graphs used for these filters are shown in Figure 10-8 and Figure 10-7 .

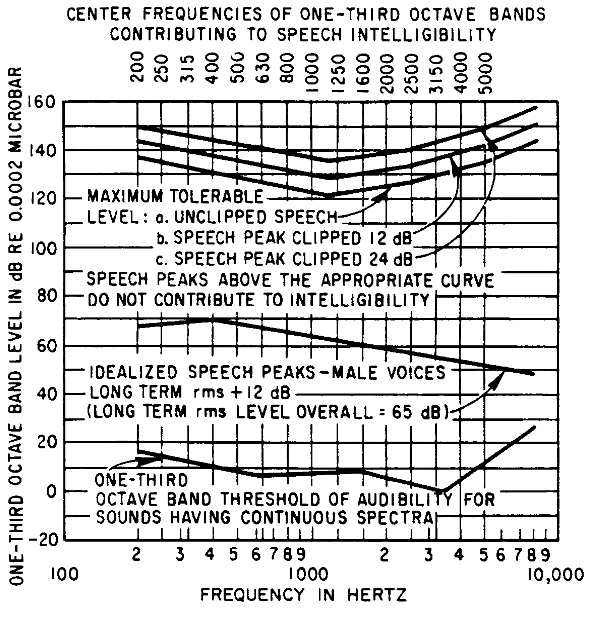


Figure 10-7. Work sheet for the 1/3-octave method

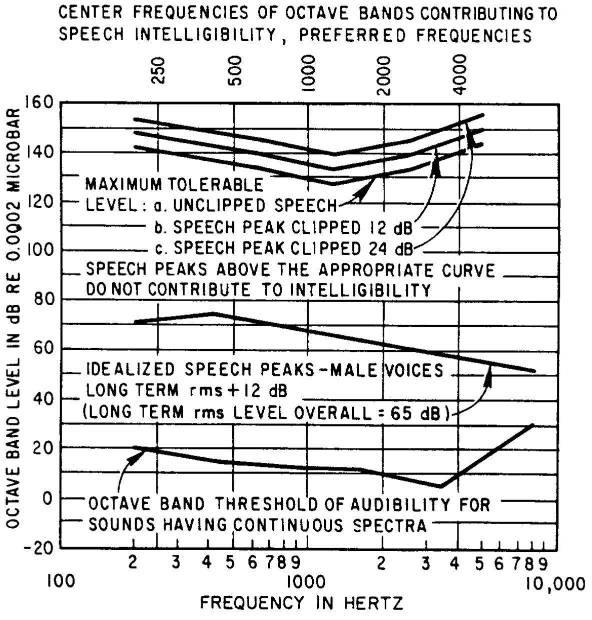


Figure 10-8. Work sheet for the 1/1-octave method

The graphs are used in the same way as the graph for the 20-band method. In the 20-band method each of the bands contributed equally to the intelligibility. This assumption is not valid for 1/1 and 1/3 octave bands and therefore weighting factors are introduced. These weighting factors must be applied before the summation of the speech level - masking threshold differences. The weighting factors are shown in Table 10-1.

Center Frequency, Hz	1/3 oct weight	1/1 oct weight
200	0,0004	
250	0,0010	0,0024
315	0,0010	
400	0,0014	
500	0,0014	0,0048
630	0,0020	
800	0,0020	
1000	0,0024	0,0074
1250	0,0030	
1600	0,0037	
2000	0,0038	0,0109
2500	0,0034	
3150	0,0034	
4000	0,0024	0,0078
5000	0,0020	

Table 10-1. Weight factors for 1/3 octave bands and 1/1 octave bands

10.4 Corrections to AI

The articulation index method is based on normal hearing speakers and listeners in usual situations where the background noise is constant. The ANSI standard contains a number of corrections to the 'normal' situation where more special situations are taken into account. These situations comprise non stationary noise, peak clipping of the speech, reverberation, unusual speech level, possibility of lip reading. Figure 10-9 show the reduction in AI as a function of reverberation.

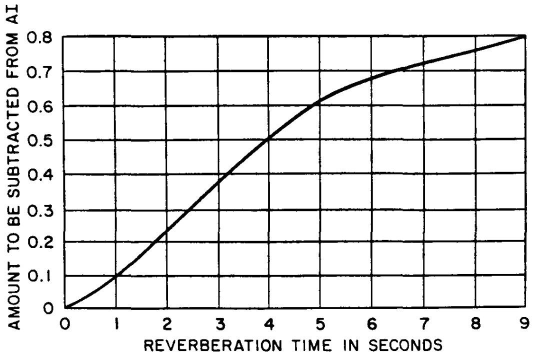


Figure 10-9. Reduction in AI as a function of reverberation time in seconds.

Figure 10-10 show the improvement in AI which can be achieved if the listener is able to see the mouth of the talker.

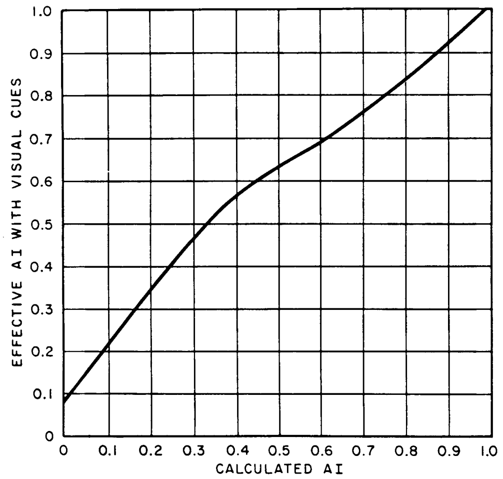


Figure 10-10. Correction (improvement) of AI if visual cues (e.g. lip-reading) are available

The AI method has been used in many other connections than for its original purpose, e.g. the prediction of the speech intelligibility for hearing impaired persons, Ludvigsen (1987).

The 1969 version of the ANSI standard has been revised and a new version, called Speech Intelligibility Index (SII), became available in 1997.

11. Speech Transmission Index, STI

An objective method for determination of speech intelligibility has been developed at Institute for Perception, The Netherlands. The method is based on an assumption that the envelope of the speech signal must be perceived correctly in order for the speech to be understood correctly. The result of the method is the *Speech Transmission Index*, STI. The method is described in a number of papers e.g. Houtgast & Steeneken (1973, 1985), Steeneken & Houtgast (1980). It is relatively complicated to measure/calculate the STI and therefore a simpler version of the method has been suggested. This version is called RASTI (Rapid STI) and is described in IEC publication 268: Sound system equipment, part 16, (IEC 268-16, 1988).

11.1 Speech Transmission Index, STI

As in the articulation index method (AI) the aim is to estimate the effective Signal-to-Noise ratio (SNR) between speech and background noise. The speech is looked upon as a combination of modulated signals. The principle in the STI method is to determine how much of the modulation that is preserved at the receiver's position (at the listener). This is measured as the *Modulation Transfer function*, MTF.

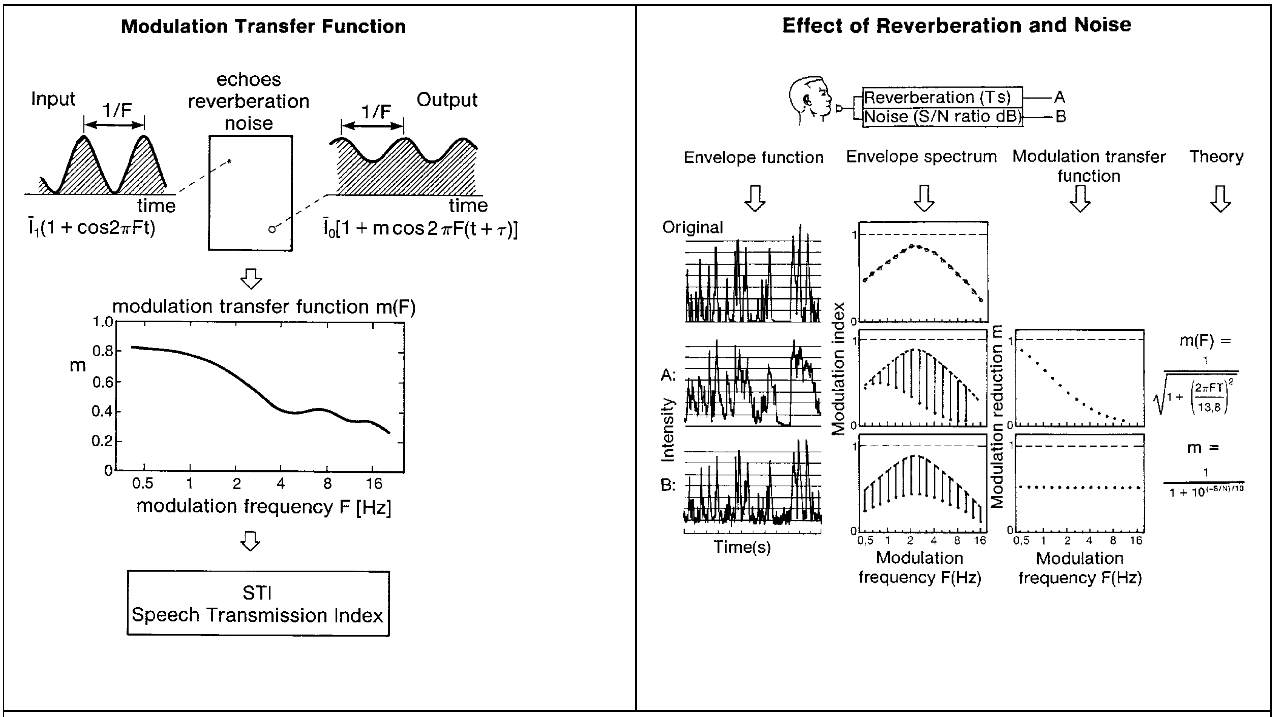
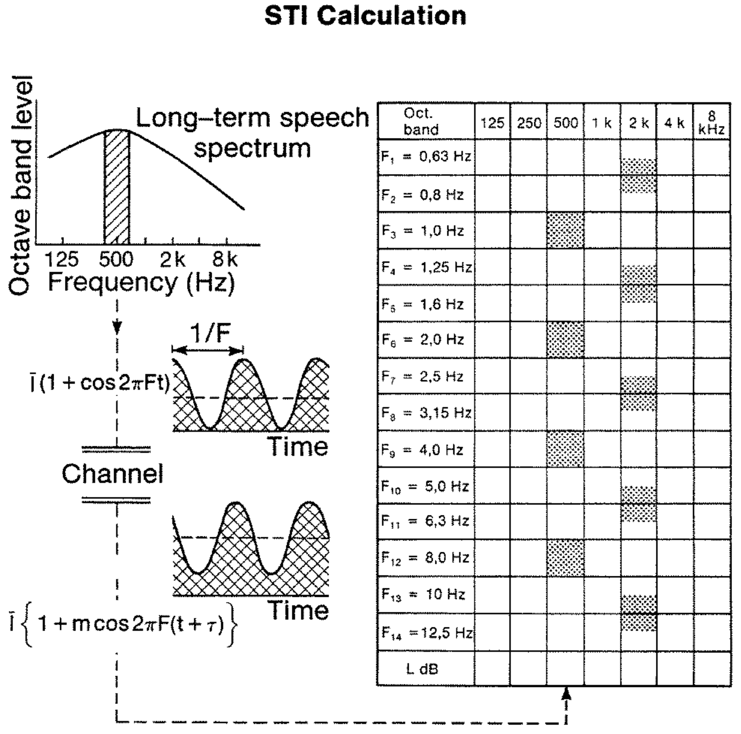

Modulation Transfer Function
Input
1/F
echoes
reverberation
noise
1/F
Output
time
time
$\bar{I}_1(1 + \cos 2\pi Ft)$
$\bar{I}_0[1 + m\cos 2\pi F(t + \tau)]$
modulation transfer function m(F)
m
1.0
0.8
0.6
0.4
0.2
0
0.5
1
2
4
8
16
modulation frequency F [Hz]
STI
Speech Transmission Index
Effect of Reverberation and Noise
Reverberation (T s)
Noise (S/N ratio dB)
A
B
Envelope function
Envelope spectrum
Modulation transfer function
Theory
Original
A:
B:
Intensity
Time(s)
Modulation index
Modulation reduction m
1
0
0,5
1
2
4
8
16
Modulation frequency F(Hz)
Modulation frequency F(Hz)
$m(F) = \frac{1}{\sqrt{1 + \left(\frac{2\pi FT}{13,8}\right)^2}}$
$m = \frac{1}{1 + 10^{(-S/N)/10}}$

Figure 11-1. Principle of Modulation Transfer Function. Left: Principle of modulation reduction. Right: Examples of modulation reduction: A, reduction caused by reverberation. B, reduction caused by noise.

Figure 11-1 shows an input signal (representing the speech) which is sinusoidal modulated. The modulation is 100%. When this signal is sent into a transmission channel or out in a room, the modulation depth will be reduced at the receiving point if noise or reverberation has been present in the course from sender to receiver. The reduction of the modulation is described by the MTF, m(F), which is a function of the modulation frequency, F. MTF is the basis for the calculation of STI.

A noise signal with the same long term spectrum as speech (speech shaped noise) is used in the measurement of STI. The noise signal is divided into 7 octave bands from 125 Hz to 8 kHz, see figure 11.2. Each octave band is separately modulated with 14 modulation frequencies (one at a time) from 0.63 Hz to 12.5 Hz. The modulation is sinusoidal and 100% and the signal is emitted from the sender (the speaker).


STI Calculation
Octave band level
Long–term speech spectrum
125 500 2k 8k
Frequency (Hz)
1/F
$\bar{I}(1+\cos 2\pi F t)$
Time
Channel
Time
$\bar{I}\left\{1+m\cos 2\pi F(t+\tau)\right\}$

Oct. band	125	250	500	1 k	2 k	4 k	8 kHz
F_1 = 0,63 Hz					▨		
F_2 = 0,8 Hz					▨		
F_3 = 1,0 Hz			▨				
F_4 = 1,25 Hz					▨		
F_5 = 1,6 Hz					▨		
F_6 = 2,0 Hz			▨				
F_7 = 2,5 Hz					▨		
F_8 = 3,15 Hz					▨		
F_9 = 4,0 Hz			▨				
F_{10} = 5,0 Hz					▨		
F_{11} = 6,3 Hz					▨		
F_{12} = 8,0 Hz			▨				
F_{13} = 10 Hz					▨		
F_{14} =12,5 Hz					▨		
L dB							

Figure 11-2. Overview of frequency bands and modulation frequencies in the STI calculation.

For each of the 98 combinations of octave band and modulation frequency the reduction in modulation is determined. This is the MTF as a function of octave band centre frequency, k, and modulation frequency, F. The reduction in modulation, m(F,k), is transformed to a signal-to-noise ratio by means of

$$SNR_{F,k} = 10 \log (m / (1-m)) \text{ dB}$$

As in the AI method the SNR is truncated to a dynamic range of 30 dB but here the limits are ±15 dB (in AI the limits are +12 dB and –18 dB). If SNR is >15 dB then SNR is set to 15 dB and if SNR is < –15 dB then SNR is set to –15 dB.

The 14 SNR values corresponding to the 14 modulation frequencies are then averaged

$$\text{Average(SNR)}_k = 1/14 \bullet \sum \text{SNR}_{F,k}$$

We now have 7 SNRs corresponding to the seven octave bands. These SNRs are also averaged after some weighting factors have been applied. The weighting factors are shown in the following table:

Freq. Hz	125	250	500	1k	2k	4k	8k
Weight, w	0,1129	0,143	0,114	0,114	0,186	0,171	0,143

The average SNR over the 7 bands (k) is calculated as

$$\text{Average(SNR)} = \sum w_k \bullet \text{Average(SNR}_k)$$

Finally the SNR is normalised to a number between 0 and 1 giving the Speech Transmission Index, STI:

$$\text{STI} = [\text{Average(SNR)} + 15] / 30$$

The STI represents an average signal-to-noise ratio deduced from the modulation transfer function, MTF.

The number of modulation frequencies used in the calculations is not very important. In one of the descriptions, Houtgast et al. (1980), only 18 modulation frequencies from 0,4 Hz to 20 Hz is used in 1/3 octave steps, but at the same time it is mentioned that as a first approximation the STI can be obtained by using 6 modulation frequencies from 0,5 Hz to 16 Hz (in octave steps). In the same reference no weighting factors are used, i.e. all octave bands have the same weight.

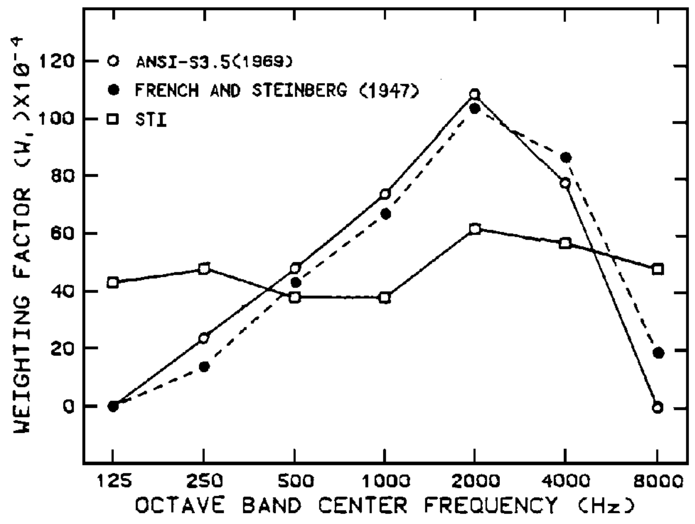


Figure 11-3. Weight factors used in AI and STI

Figure 11-3 shows the difference in weight factors used in the AI procedure (1/1 octave version) and in the STI procedure. It is seen that especially around 2 kHz the weights are different.

The STI in itself is not enough to predict the speech intelligibility. In order to use the STI it is necessary to know the relation between speech intelligibility and STI. This is illustrated in figure 11.4 for some typical word materials.

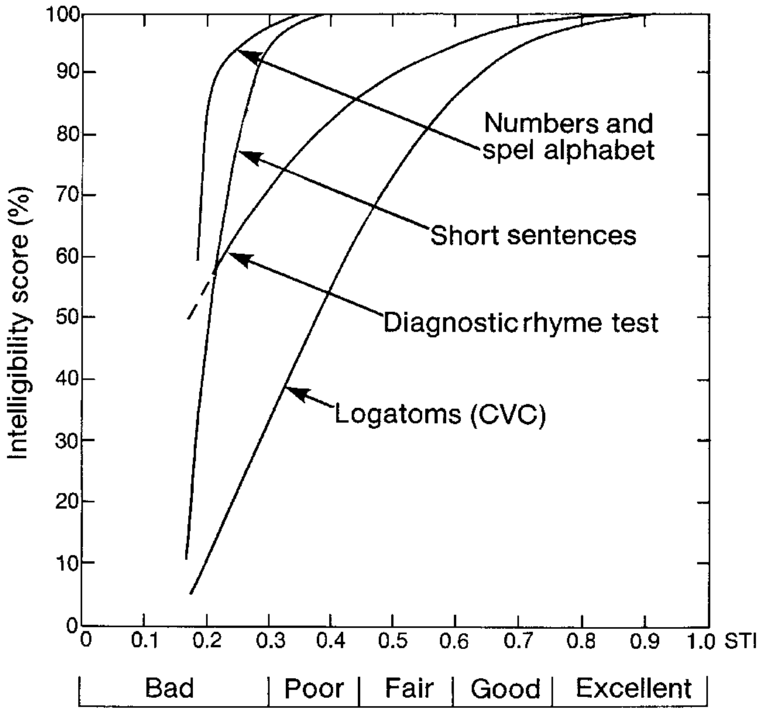


Figure 11-4. Relation between intelligibility and STI for different word materials. Compare with Figure 10-6.

11.2 RASTI

As seen in the STI section it is a somewhat complicated matter to calculate STI. This is not satisfying for an objective method that should replace the cumbersome subjective measuring methods. Therefore a simpler method, RASTI (= RApid STI) has been developed based on the same principles as in STI. In RASTI the number of combinations of octave bands and modulation frequencies are reduced from 98 to 9. In Figure 11-2 the combinations of frequency bands and modulation frequencies used in RASTI are shown in grey. Only the octave bands 500 Hz and 2 kHz are used and the modulation frequencies are selected so that they cover the most important range.

Figure 11.5 show the envelopes curves for the 500 Hz and the 2 kHz noise bands. Figure 11.6 shows the calculation procedure. There are no weighting factors in the RASTI calculation but because of the 5 modulation frequencies used in the 2 kHz band and only 4 modulation frequencies in the 500 Hz band, there will be a slightly higher weight for the 2 kHz band in the average calculation.

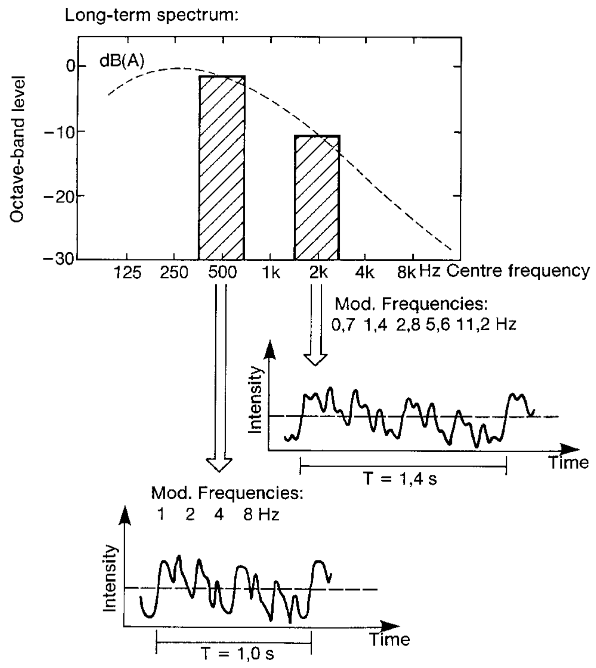


Figure 11-5. The RASTI signal. All modulation frequencies are present at the same time. At 500 Hz the period is 1 s, at 2 kHz the period is 1,4 s.

Modulation reduction factor: $m(F)_k$

F: modulation frequency
k: 500 Hz or 2000 Hz

1: $x = 10\log\left\{\frac{m}{1-m}\right\}$ dB

2: $-15 \leqslant x \leqslant 15$ dB

3: $\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i$

4: $\mathrm{RASTI}_{\mathrm{index}} = \frac{\bar{x}+15}{30}$

Figure 11-6. RASTI calculation based on the modulation transfer function described by the modulation reduction factor $m(F)_k$

The company Brüel & Kjær has produced a system that works according to the RASTI principle (B&K 3361). The system consists of a sound generator and an analyzer. The sound generator emits all the modulated noise signals as one signal and the analyzer measures the degree of modulation and calculates the RASTI value. All modulation frequencies and both noise bands are emitted simultaneously and it is therefore possible to make a RASTI measurement within a very short time (8 to 32 seconds).

RASTI measurements have been shown to be in reasonable accordance with the results of real subjective speech intelligibility measurements in auditoria, churches etc where the problem is reverberation or stationary background noise. If the background noise is non stationary and thus may contribute to the modulation the RASTI measurements will show a too optimistic result.

12. Binaural hearing

We are listening with two ears- usually. The information that we receive from the two ears is used to tell us where the sound source is located. The neural paths from the two ears are connected in several ways and thus a kind of cross correlation can be made in the brain.

12.1 Horizontal plane

The situation in the horizontal plane can be described by Figure 12-1.

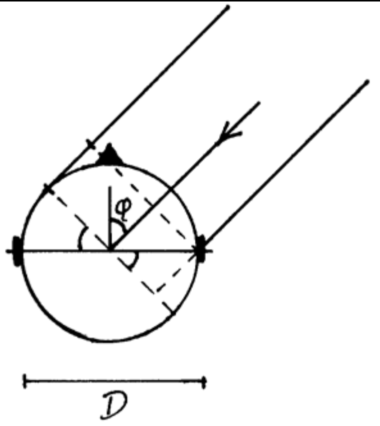


Figure 12-1. Simple model for sound incidence in the horizontal plane

Sound coming from a certain direction will reach first the nearest ear and after some time, the interaural time delay (ITD), the other ear. If the wavelength of the sound is small compared to the size of the head (i.e. high frequencies), a shadow effect, the interaural level difference (ILD), will occur.

If the sound source is far away (i.e. parallel incidence to the two ears) and the wavelength is small compared to the size of the head, the travel distance can be approximated by

$$\Delta s = \frac{D}{2}(\varphi + \sin\varphi)$$

and thus the interaural time delay will be

$$\Delta t = \frac{\Delta s}{c}$$

where c is the speed of sound. In the transition area between high and low frequencies (0,5 – 2 kHz) it is necessary to distinguish between phase delay and group delay.

For low frequencies, the interaural time delay can be calculated from the approximate expression

$$\Delta t_{LF} = \frac{D}{2c}(3\sin\varphi)$$

The time delay for a sphere model of the head is shown is shown in Figure 12-2. Note that the maximum delay is about 0,65 ms (= 650 µs).

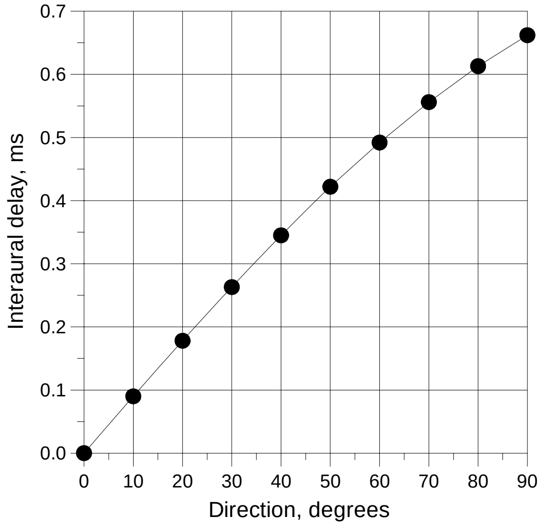


Figure 12-2. Interaural time delay, ITD, for a sphere model. Measurements on a human head gives the same result. The result is shown for only one side of the sphere/head.

The interaural level difference is shown in Figure 12-3. The level difference is small at low frequencies (long wavelength) and as much as 20 dB at high frequencies (5 kHz).

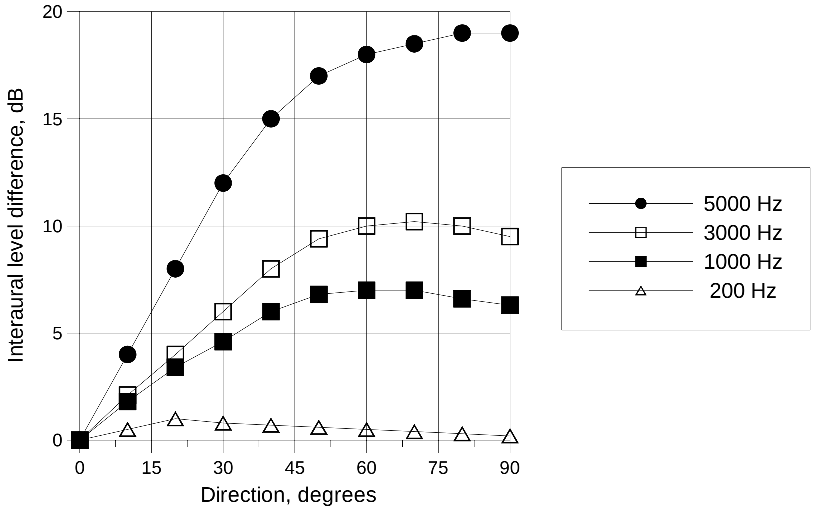


Figure 12-3. Interaural level difference, ILD. The result is shown for only one side of the sphere/head.

The ability to distinguish between the directions to two sound sources is called the minimum audible angle, MAA. This has been measured in the horizontal plane by Mills (Mills 1958). The famous result is shown in Figure 12-4.

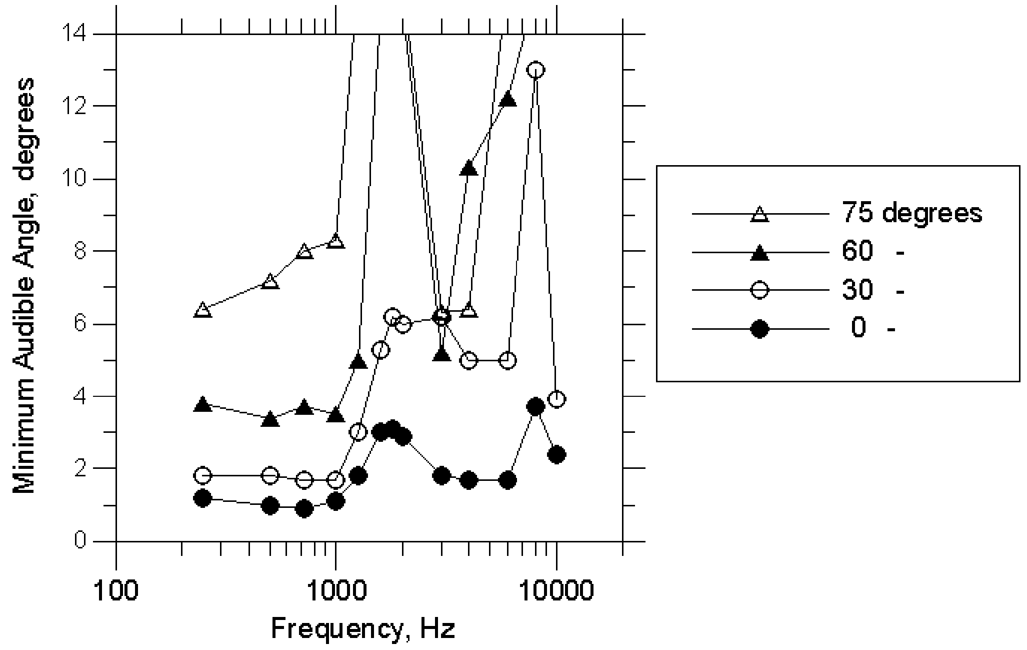


Figure 12-4. The minimum audible angle in the horizontal plane. (Mills 1958).

In the frontal direction, 0°, it is possible to distinguish within about 1° at low frequencies. Around 1500 Hz the MAA increase to 3 degrees and just at 8 kHz the MAA show a peak of about 4 degrees.

12.2 Head related transfer function

The head related transfer function, HRTF, is defined by

$$\mathrm{HRTF}(\varphi, \theta) = p_{\mathrm{ear}} / p_{\mathrm{center}}$$

where p_{ear} is the sound pressure at the ear and p_{center} is the sound pressure at a position corresponding to the centre of the head.

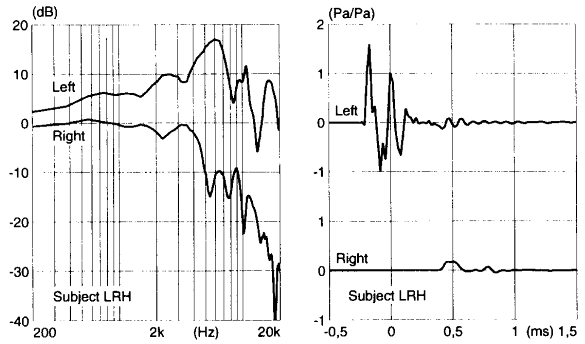
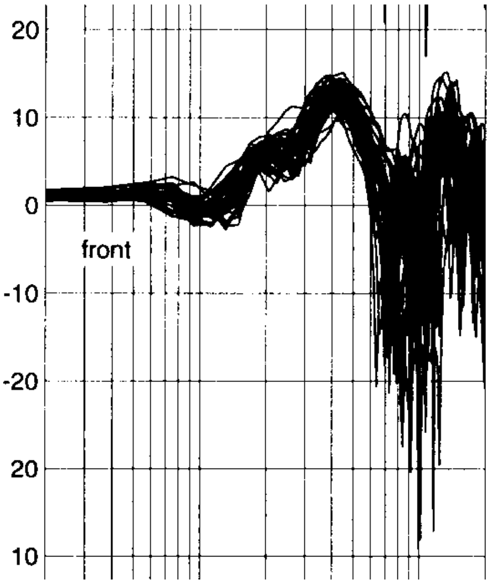


Figure 12-5. HRTF for the left and right ear for sound incidence directly from the left (Hammershøi 1995).

The HRTF differ considerably between persons. This is illustrated in Figure 12-6 and Figure 12-7 where results from 40 ears are shown.



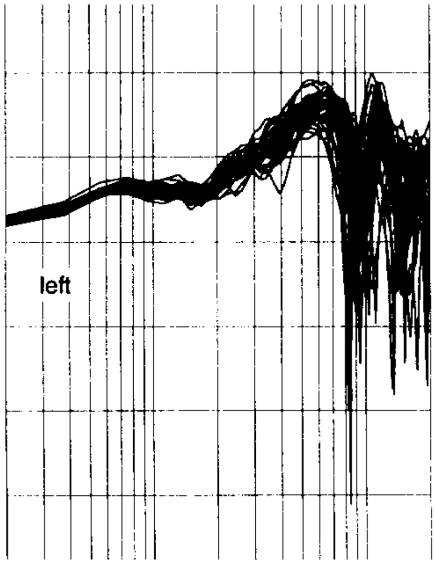
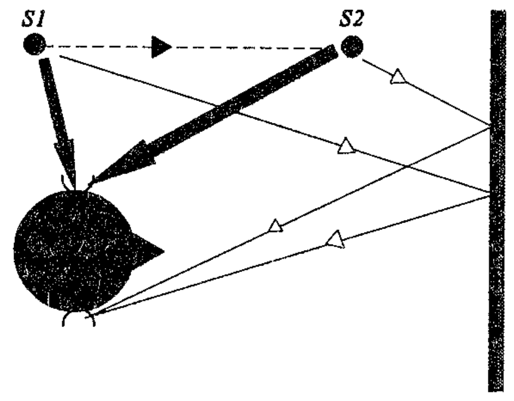


Figure 12-6. HRTF for 40 persons. Frontal sound incidence. (Hammershøi 1995).

Figure 12-7. HRTF for 40 persons. Sound incidence from the left. (Hammershøi 1995).

12.3 Moving sound sources

The previous sections in this chapter are all about fixed positions of source and receiver. If the sound source moves from one position to another the situation becomes more complicated. This is illustrated with an example in Figure 12-8. The path for direct sound (shown with thick arrows) will become longer. The path for the reflections from the wall (shown for the right ear only) will become shorter.

	*Figure 12-8 Schematic illustration of a sound source that moves from S1 to S2.*

13. Hearing loss

Hearing losses can be divided into two main groups

- Conductive losses, i.e. losses where the ‘problem’ is found in the ear canal or in the middle ear. The name comes from the effect that the sound cannot be conducted into the sensory organ, the inner ear.
- Sensorineural losses, i.e. losses where the ‘problem’ is found in the sensory organ or in the nerve connections to the brain. The sound cannot be perceived in a normal way.

The conductive losses can often be alleviated by means of an operation in the ear canal or the middle ear. Combinations of conductive and sensorineural losses are also seen. A distinction between a conductive and a sensorineural hearing loss can be made by means of the so-called air-bone gap in an audiogram.

The sensorineural losses can further be divided into

- Cochlear losses, i.e. some defect in the cochlea, the inner ear.
- Retro-cochlear losses, i.e. a defect in the transmission path from the cochlea to the brain.

The most common type of a cochlear hearing loss is the age-related hearing loss. The clinical term for this is Presbyacusis. The cause of this type of hearing loss is that the outer hair cells have deteriorated over time and do not function as well as previously. The cochlear amplifier does not amplify as much as it did at a younger age. This means that the hearing threshold will be elevated.

Another common type of a cochlear hearing loss is the noise-induced hearing loss. In this case the outer hair cells have been destroyed because the ear has been exposed to too much noise (e.g. at a work place).

Some cochlear hearing losses are congenital, i.e. present from the time of birth. Previously hearing losses caused by German measles during pregnancy were often seen. This is now prevented by a vaccination before pregnancy. Some cochlear hearing losses are hereditary and thus related to genetic factors in the family.

The above characterization of hearing losses is very coarse. There are big individual differences and these individual differences must be taken into account when a hearing loss is treated with a hearing aid.

The effects of a cochlear hearing loss can be summarized as follows

- Reduced sensitivity. This is measured by means of the audiogram. The reduced sensitivity will give an elevated threshold that is seen as a downwards deviation from the zero line in the audiogram.
- Abnormal loudness. The loudness perception will be different from normal hearing

people's loudness perception. This effect is pronounced at low levels i.e. near, but above the hearing threshold. Traditionally this has been called loudness recruitment, but it may be more correct to characterize the effect as loudness imperception. At levels well above hearing threshold the loudness perception is the same for cochlear hearing impaired persons and normal hearing persons.

- Reduced frequency selectivity. This means that the analysing filters (critical bands) are broader and thus fewer are needed to cover the audible frequency range. The broader bands make it more difficult to distinguish the different frequency components in a complex sound signal like speech.
- More upward spread of masking. This means that some frequency components may be masked even though the components are above the absolute threshold of the hearing impaired.
- Reduced temporal resolution. The distinction between short and longer signals is deteriorated and forward masking is prolonged. Both effects will have a negative influence on speech intelligibility.
- Reduced binaural processing. The ability to analyse and utilize the signals from the two ears is reduced. This will make it more difficult to localize sounds and to suppress the influence from background noise.

13.1 ISO 1999

ISO 1999 (1990) has formulas and tables showing estimations of a hearing loss as a function of level and duration of *noise* exposure. See Figure 13-1.

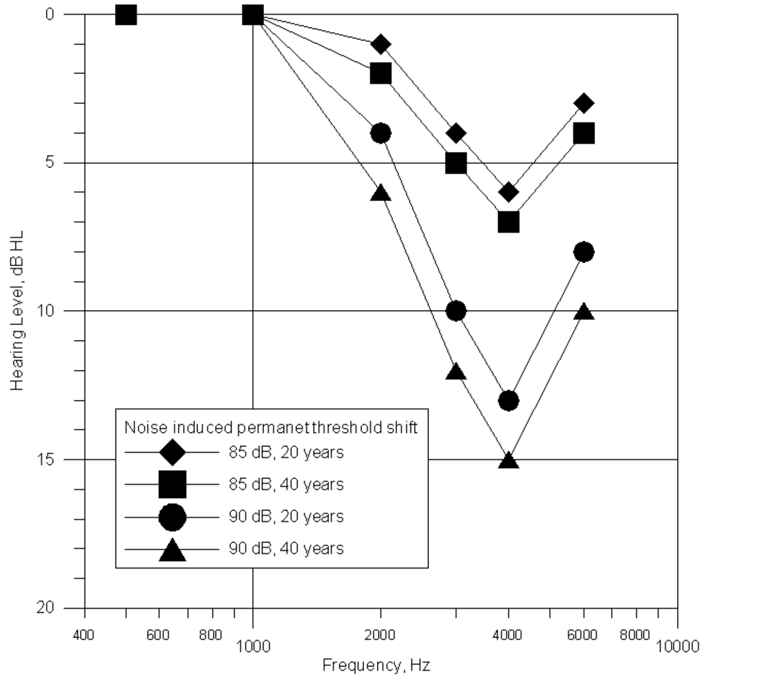


Figure 13-1. Estimated hearing loss caused by noise. From ISO 1999 (1990)

13.2 ISO 7029

ISO 7029 (2000) has formulas and tables showing estimations of a hearing loss as a function of *age* for otologically normal persons. Distinction is made between males and females. See Figure 13-2.

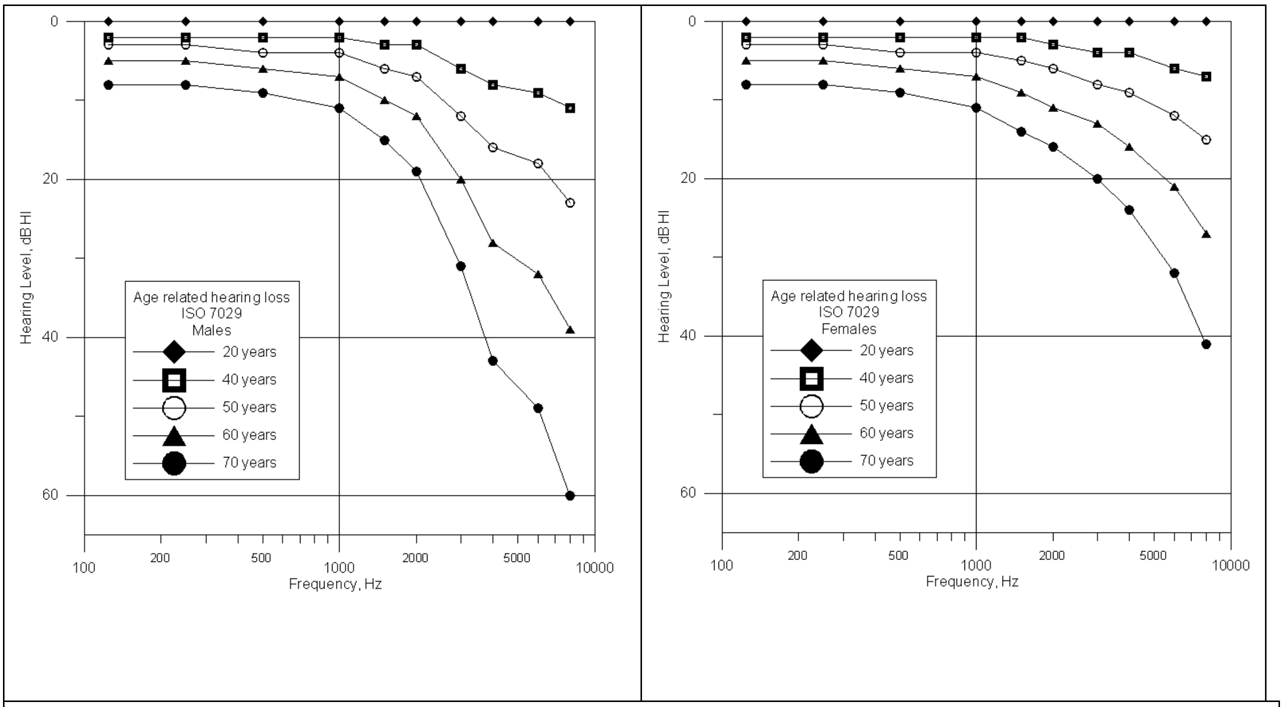


Figure 13-2. Age related hearing loss. Left panel. Males. Right panel: Females. From ISO 7029 (2000)

14. References

Ainsworth, W. A. (1976) Mechanisms of speech recognition, Pergamon Press.

ANSI-S3.5 (1969). American National Standard methods for the calculation of the Articulation Index. New York, American National Standards Institute, Inc.

ANSI-S3.5 (1997). American National Standard methods for the calculation of the Speech Intelligibility Index. New York, American National Standards Institute, Inc.

Borden, G. and K. Harris (1980) Speech science primer, Williams & Wilkins.

Buus, S., M. Florentine, et al. (1997) "Temporal integration of loudness, loudness discrimination and the form of the loudness function." Journal of the Acoustical Society of America **101**(2): 669-680.

Byrne, D., C. Ludvigsen, et al. (1994) "Long-term average speech spectra ..." J. Acoust. Soc. Am. **96**(no. 4): 2110?-2120?

Elberling, C., C. Ludvigsen, et al. (1989) "DANTALE, a new Danish Speech material." Scandinavian Audiology **18**: 169-175.

Engström, H. and B. Engström (1979) A short survey of some common or important ear diseases, Widex.

Flanagan, J. L. (1972) Speech analysis, Synthesis and Perception, Springer.

Florentine, M. and S. Buus (2001) Evidence for normal loudness growth near threshold in cochlear hearing loss. 19 Danavox Symposium, Kolding, Denmark.

Hammershøi, D. (1995). Binaural technique - a method of true 3D sound reproduction. Department of Acoustics, Aalborg University.

Hougaard, S., O. Jensen, et al. (1995) Sound and Hearing, Widex.

IEC-268-16 (1988). Sound system equipment - Part 16: The objective rating of speech intelligibility in auditoria by the RASTI method, International Electrotechnical Commission.

IEC-537 (1976). Frequency weighting for the measurement of aircraft noise (D-weighting). Geneva, Switzerland, International Electrotechnical Commission.

IEC-651 (1979). Sound level meters. Geneva, Switzerland, International Electrotecnical Commission.

ISO-226 (2003). Acoustics - Normal equal-loudness-level contours. Geneva, International Standardization Organization.

ISO-389-1 (1991). Acoustics - Reference zero for the calibration of audiometric equipment - Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones. Geneva, Switzerland, International Organization for Standardisation.

ISO-389-5 (2006). Acoustics - Reference zero for the calibration of audiometric equipment - Part 5: Reference equivalent threshold sound pressure levels for pure tones in the frequency range 8 kHz to 16 kHz“. Geneva, Switzerland, International Organization for Standardization.

ISO-389-7 (1996). Acoustics - Reference zero for the calibration of audiometric equipment- Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions. Geneva, Switzerland, International Organization for Standardization.

ISO-389-8 (2001). Acoustics - Reference zero for the calibration of audiometric equipment - Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones (ISO/DIS). Geneva, Switzerland, International Organization for Standardization.

ISO-532 (1975). Acoustics - Method for calculating loudness level. Geneva, Switzerland, International Organisation for Standardisation.

Jørgensen, E. F. (1962) Almen Fonetik, Rosenkilde og Bagger.

Keidser, G. (1993) "Normative data in quiet and in noise for DANTALE - a Danish speech material." Scandinavian Audiology **22**: 231-236.

Kemp, D. T. (1988) "Developments in cochlear mechanics and techniques for noninvasive evaluation." Adv Audiol **5**: 27-45.

Ludvigsen, C. (1987) "Predictin of speech intelligibility for normal-hearing and cochlearly hearing-impaired listeners." J. Acoust. Soc. Am. **82**: 1162-1171.

Malmberg, B. (1970) Manual of Phonetics, North-Holland.

Martin, M. (1987) Speech audiometry, Taylor & Francis.

Mills, A. W. (1958) "On the minimum audible angle." J. Acoust. Soc. Am. **30**: 237-246.

Minifie, F. D., T. J. Hixon, et al. (1973) Normal aspects of speech, hearing and language, Prentice-Hall.

Moore, B. C. J. (1986) Frequency selectivity in hearing, Academic Press.

Moore, B. C. J. (2001) Frequency Resolution. 19th Danavox Symposium: Genetics and the function of the auditory system, Kolding, Denmark.

Poulsen, T. (1981) "Loudness of Tone Pulses in a Free Field." Journal of the Acoustical Society of America **69**(6): 1786-1790.

Rabiner, L. R. and R. W. Schafer (1978) Digital processing of speech signals., Prentice-Hall.

Steeneken, H. and T. Houtgast (1980) "A physical method for measuring speech-transmission quality." J. Acoust. Soc. Am. **67**: 318-326.

Steeneken, H. J. M. and T. Houtgast (1983) The temporal envelope of speech and its significance in room acoustics. ICA, Paris.

Zwicker, E. and H. Fastl (1999) Psychoacoustics. Facts and models., Springer.

Zwicker, E. (197x). Scaling. In Keidel & Neff (eds.), Handbook of sensory physiology, Vol IV.

14.1 Further reading:

Moore, B. C. J. (2003). An introduction to the psychology of hearing. 5th Edition. Academic press ISBN: 0-12-505627-3

Yost, W. A. (2000). Fundamentals of hearing. An introduction. 4th Edition. Academic press. ISBN: 0-12-775695-7

Plack, C. J. (2005). The sense of hearing. Lawrence Earlbaum Associates. ISBN: 0-8058-4884-3

Ludvigsen, C.: Comparison of certain measures of speech and noise level. Scand. Audiol. Vol 21, p 23, 1992

Quackenbush, S.R.; Barnwell, T.P. & Clements, M.A. (1988): Objective measures of speech quality. Prentice-Hall, New Jersey. (Contains good descriptions of Articulation Index (p.37-41) and of Speech Transmission Index (s. 41-44))

Bistafa & Bradley, JAES, vol 48, 2000, p. 531

Bradley, JAES, vol 46, 1998, p. 396

Hansen, M. & Kollmeier, B.: JAES, vol 48, 2000, p. 395

About standardized audiological, clinical tests see: http://www.phon.ucl.ac.uk/home/andyf/natasha/

15. Index