Music Perception and Cognition

0. Introduction

What's the purpose of?

- **Hearing**: overcome the limitations of our vision.
- **Listening:** gain information that becomes the basis for taking a decision on any topic.
- **Understanding**: matching what has been heard and listened to pre-existing knowledge (to make sense of it).

1. The physiology of audition

1.1. The problem of perception

*"The problem of perception is initially a **problem of taxonomy** in which the individual animal must "classify" the things of its world (…) The internal taxonomy of perception is **adaptive** but is not necessarily veridical in the sense that it is concordant with the descriptions of physics."*

Edelman, Neural Darwinism, p.26

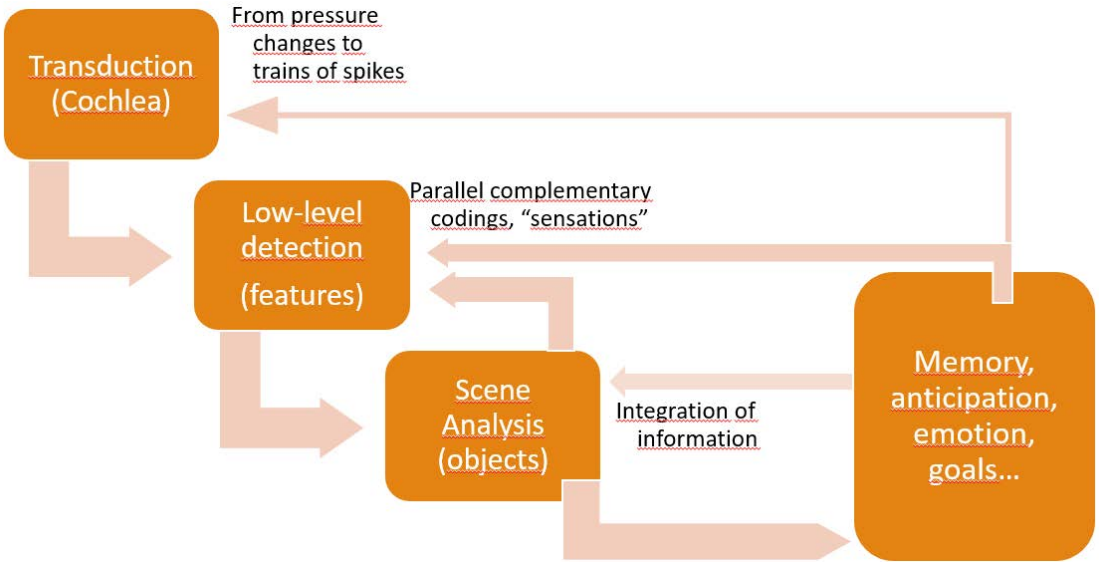
1.2. Physical properties / Sensations (perceptual properties) / Musical properties

Perception does not necessarily reflect reality.

3 types of properties:

1. **Physical:** frequency (periodicity), amplitude, waveform, duration
2. **Perceptual:** pitch, loudness, timbre, length
3. **Musical:** melodies, dynamics, voices/instruments/texture/chords, figures/rhythm

1.3. Stages in "perception"



1.4. Perception versus Cognition

- **Bottom-up processing**: data-driven approach stating that perception directs cognition (from ear to brain). Influences decisions and behavior (**perception directs cognition**).
- **Top-down processing**: behavior influenced by CONCEPTUAL DATA (from brain to ear). The expectations influence perception and behavior (**perception is constructed by cognition**).

PERCEPTION	COGNITION
Pressure converted to **electrochemical** patterns	Electrochemical patterns becoming **meaningful** and **adaptive**
Transduction and **basic sensations** (pitch, loudness, timbre, localization…)	Musicality, musical knowledge (tension, movement, emotion, preference…)
Bottom-up	Top-down
Automatic	Consciousness-mediated
Encapsulated	Distributed

1.5. Sensory systems

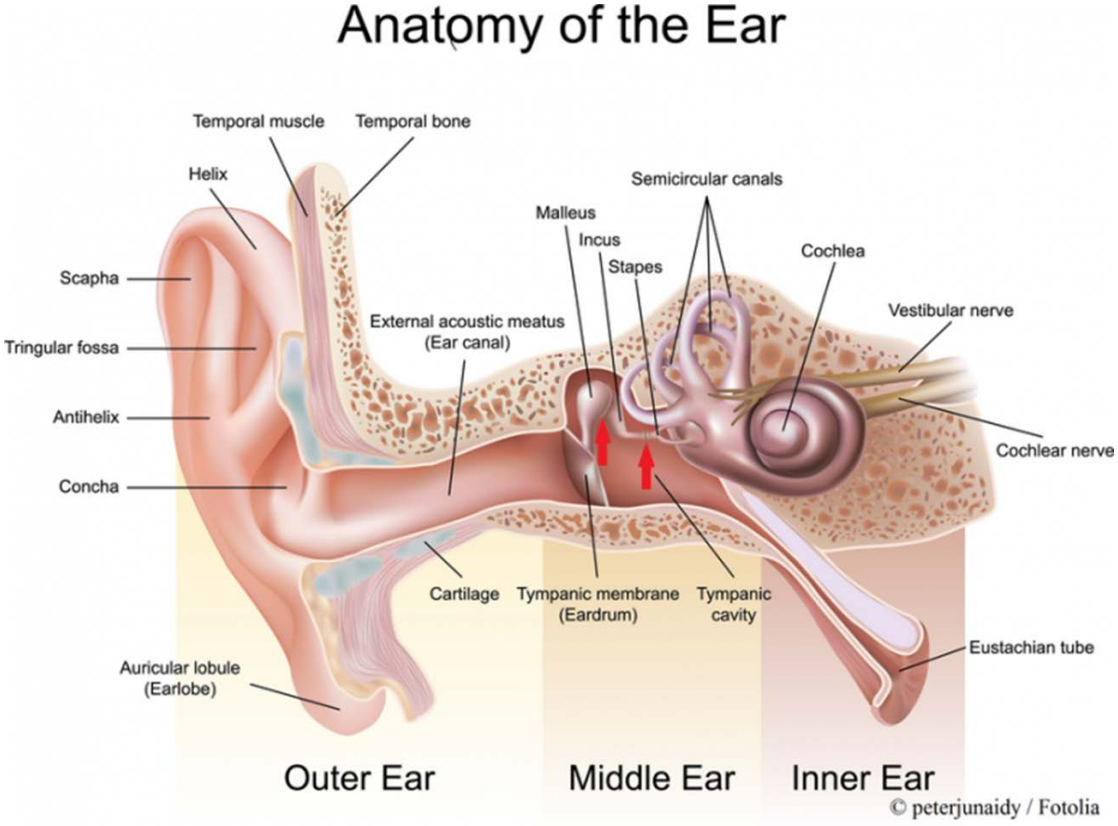
- **Interoceptors:** keep the brain informed about the **status of internal organs** (stomach, heart, etc.)
- **Proprioceptors:** keep the brain informed about the **body position** (equilibrium - vestibular system, limb movements - kinesthetic system)
- **Exteroceptors:** keep the brain informed about the **external world** (traditional 5 senses)

Stages followed by the information:

1. **Transduction**
2. **Transmission to the cortex**

3. **Integration with other senses and info**

1.6. The Ear



1.7. The Outer Ear

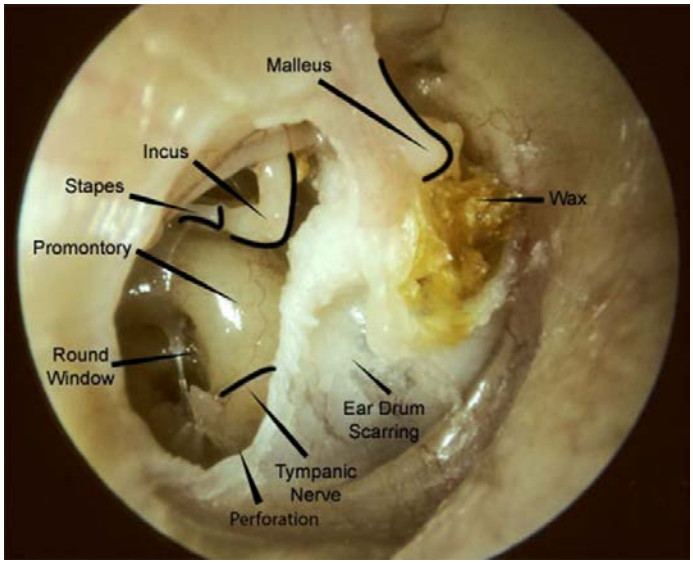
Functions:

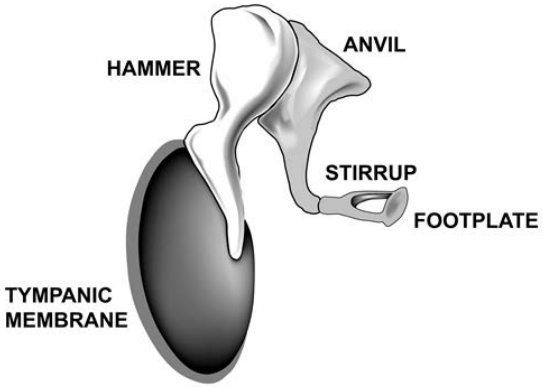
- **Sound collection** (pinna macro-structure)
- **Localization** (pinna micro-structure)
- **Protection** (inner canal, wax, inner hair)

1.8. The Middle Ear

Functions:

- **Amplification** (ossicles - lever principle)
- **Pressure compensation** (Eustachian tube)
- **Protection** (acoustic reflex)

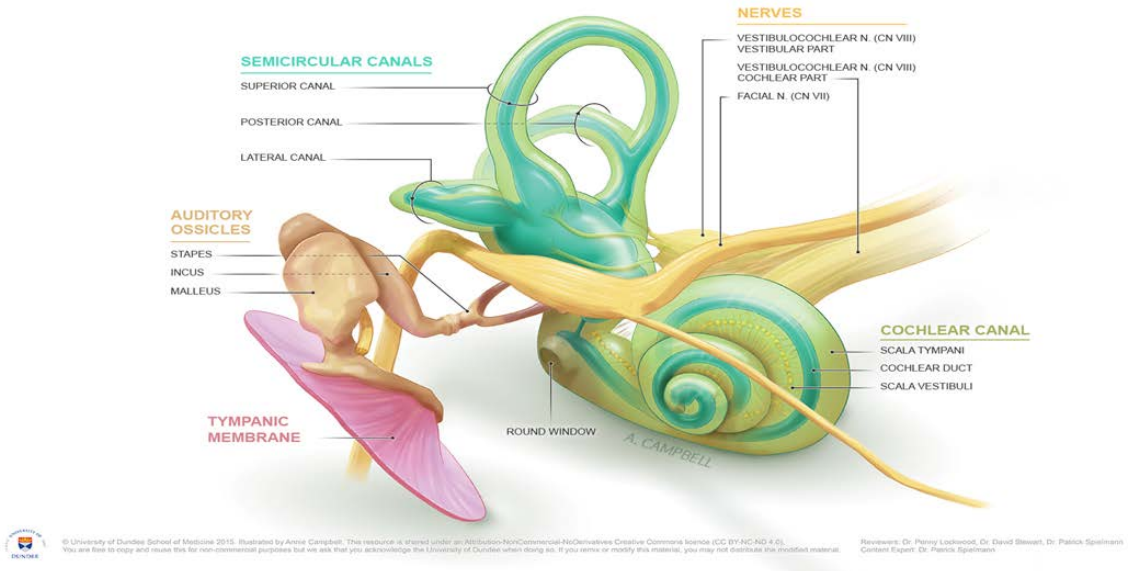


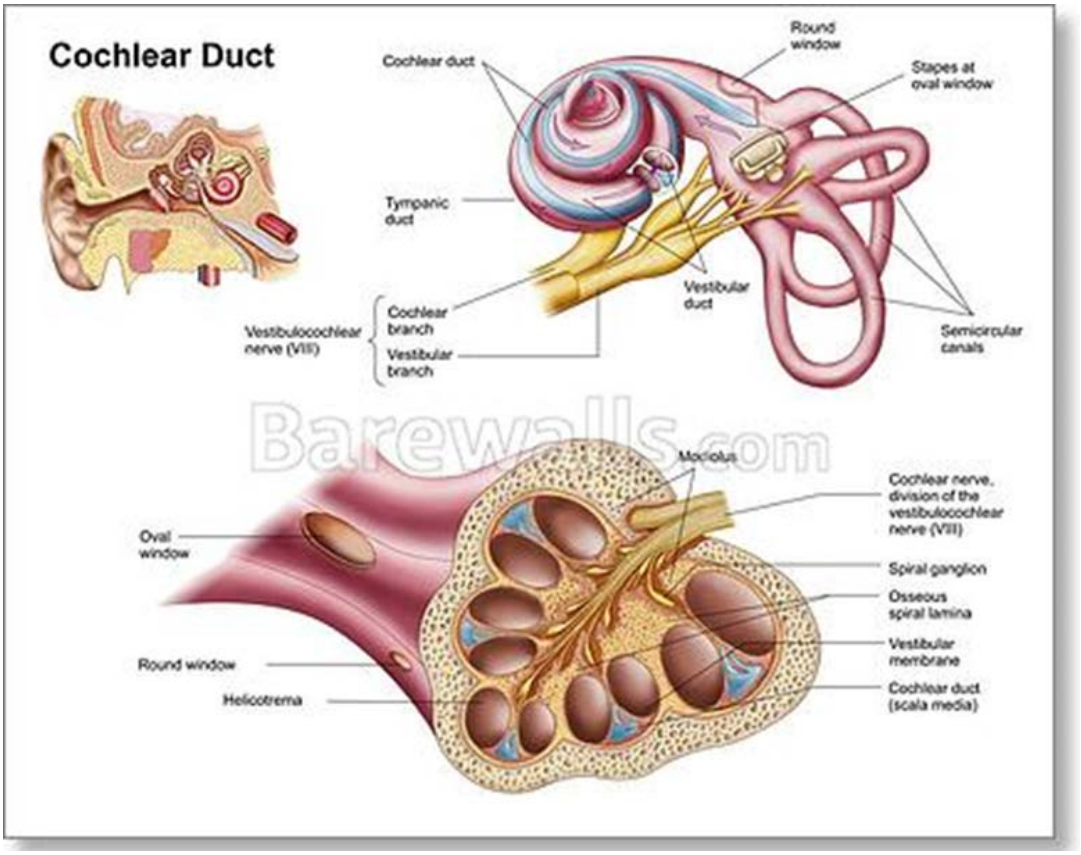


1.9. The Inner Ear

Functions:

- **Transduction** (cochlea)
- **Equilibrium** (semicircular canals)





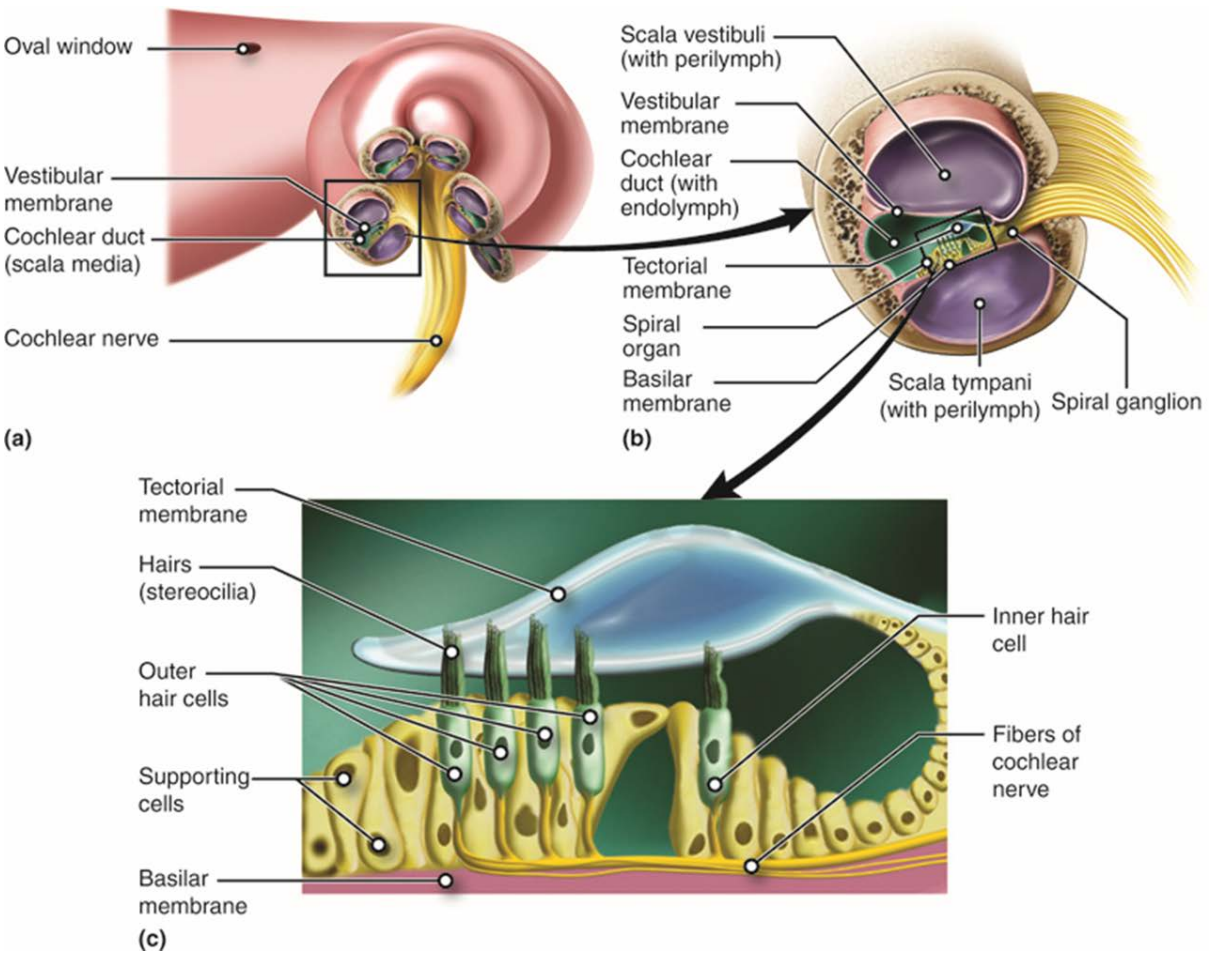
When the pressure waves reach the ear, the ear transduces this mechanical stimulus (pressure wave) into a nerve impulse (electrical signal) that the brain perceives as sound. The pressure waves strike the **tympanum**, causing it to vibrate. The mechanical energy from the moving tympanum transmits the vibrations to the **3 bones of the middle ear**. The **stapes** transmits the vibrations to a thin diaphragm called the **oval window,** which is the outermost structure of the **inner ear**. The structures of the inner ear are found in the **labyrinth**, a bony, hollow structure that is the most interior portion of the ear. Here, the energy from the sound wave is transferred **from the stapes through the flexible oval window and to the fluid of the cochlea**. The vibrations of the oval window create pressure waves in the fluid (perilymph) inside the cochlea. The **cochlea** is a whorled structure, like the shell of a snail, and it contains receptors for **transduction of the mechanical wave into an electrical signal.** Inside the cochlea, the **basilar membrane** is a mechanical analyzer that runs the length of the cochlea, curling toward the cochlea's center.

- Sound wave represents alternating areas of high and low pressure
- **Tympanic membrane vibrates** in response to sound wave
- **Vibrations are amplified across ossicles**
- Vibrations against oval window set up **standing wave in fluid of vestibuli** (frequency of standing wave is the same as sound wave)
- **Pressure bends the membrane of the cochlear duct** at a point of maximum vibration for a given frequency, **causing hair cells in the basilar membrane to vibrate**.

1.10. The Cochlea

- Coiled tube (2.5 turns) and 3 chambers

- **Basilar membrane and tectorial membrane (c)**
- Organ of Corti, transduction (c)
- Outer and inner hair cells (c)
- Efferent and afferent neurons: depending on the direction in which information travels across the nervous system. **Afferent neurons carry information from sensory receptors of the skin and other organs to the central nervous system (i.e. brain and spinal cord), whereas efferent neurons carry motor information away from the central nervous system to the muscles and glands of the body.**



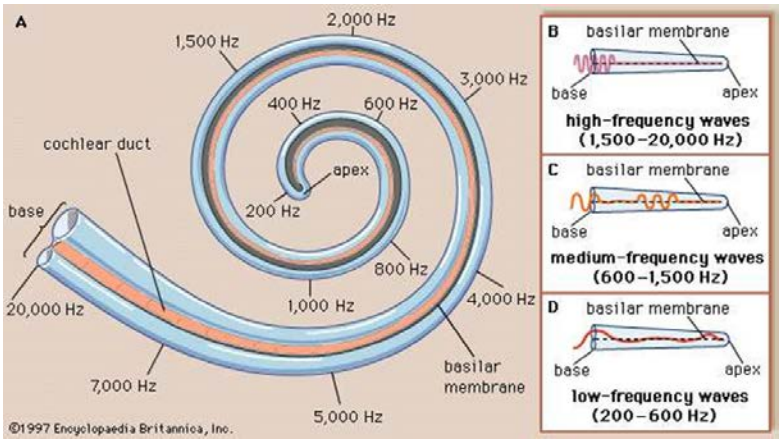
- **The mechanical properties of the basilar membrane change along its length**, such that it is thicker, tauter, and narrower at the outside of the whorl (where the cochlea is largest), and thinner, floppier, and broader toward the apex, or center, of the whorl (where the cochlea is smallest). Different regions of the basilar membrane vibrate according to the frequency of the sound wave conducted through the fluid in the cochlea. For these reasons, the **fluid-filled cochlea detects different wave frequencies (pitches) at different regions of the membrane**. When the sound waves in the cochlear fluid contact the basilar membrane, it flexes back and forth in a wave-like fashion. Above the basilar membrane is the **tectorial membrane**.
- The site of transduction is in the **organ of Corti** (spiral organ). It is composed of **hair cells** held in place above the basilar membrane like flowers projecting up from soil, with their exposed short, hair-like **stereocilia contacting or embedded in the tectorial membrane above them**. **The inner hair cells are the primary auditory receptors and exist in a single row, numbering**

approximately 3,500. The stereocilia from inner hair cells extend into small dimples on the tectorial membrane's lower surface. **The outer hair cells are arranged in three or four rows**. **They number approximately 12,000, and they function to fine tune incoming sound waves**. **The longer stereocilia that project from the outer hair cells actually attach to the tectorial membrane**. All of the stereocilia are mechanoreceptors, and when bent by vibrations they respond by opening a gated ion channel. As a result, the hair cell membrane is depolarized, and a signal is transmitted to the cochlear nerve. **Intensity (volume) of sound is determined by how many hair cells at a particular location are stimulated.**

- The hair cells are arranged on the basilar membrane in an orderly way. The basilar membrane vibrates in different regions, according to the frequency of the sound waves impinging on it. Likewise, the hair cells that lay above it are most sensitive to a specific frequency of sound waves. Hair cells can respond to a small range of similar frequencies, but they require stimulation of greater intensity to fire at frequencies outside of their optimal range.
- **Place theory, which is the model for how biologists think pitch detection works in the human ear, states that high frequency sounds selectively vibrate the basilar membrane of the inner ear near the entrance port (the oval window)**. **Lower frequencies travel further along the membrane before causing appreciable excitation of the membrane**. **The basic pitch-determining mechanism is based on the location along the membrane where the hair cells are stimulated**. The place theory is the first step toward an understanding of pitch perception. Considering the extreme pitch sensitivity of the human ear, it is thought that there must be some auditory "sharpening" mechanism to enhance the pitch resolution.
- When sound waves produce fluid waves inside the cochlea, the basilar membrane flexes, bending the stereocilia that attach to the tectorial membrane. **Their bending results in action potentials in the hair cells, and auditory information travels along the neural endings of the bipolar neurons of the hair cells** (collectively, the auditory nerve) **to the brain**. **When the hairs bend, they release an excitatory neurotransmitter at a synapse with a sensory neuron, which then conducts action potentials to the central nervous system**. The cochlear branch of the vestibulocochlear cranial nerve sends information on hearing. The auditory system is very refined, and there is some modulation or "sharpening" built in. **The brain can send signals back to the cochlea, resulting in a change of length in the outer hair cells, sharpening or dampening the hair cells' response to certain frequencies**.
- **The inner hair cells are most important for conveying auditory information to the brain**. **About 90 percent of the afferent neurons carry information from inner hair cells**, with **each hair cell synapsing with 10 or so neurons**. Outer hair cells connect to only 10 percent of the afferent neurons, and each afferent neuron innervates many hair cells. The afferent, bipolar neurons that convey auditory information travel from the cochlea to the medulla, through the pons and midbrain in the brainstem, **finally reaching the primary auditory cortex in the temporal lobe**.

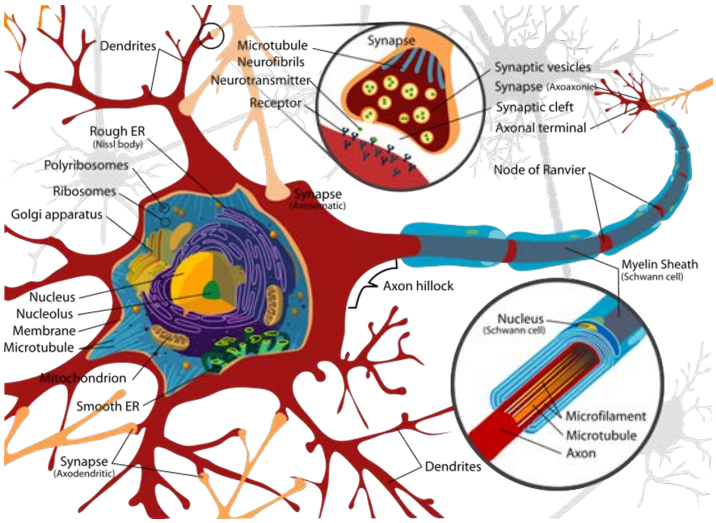
1.11. Traveling waves in the cochlea

- Maximum vertical displacement of basilar membrane depending on sound frequency:



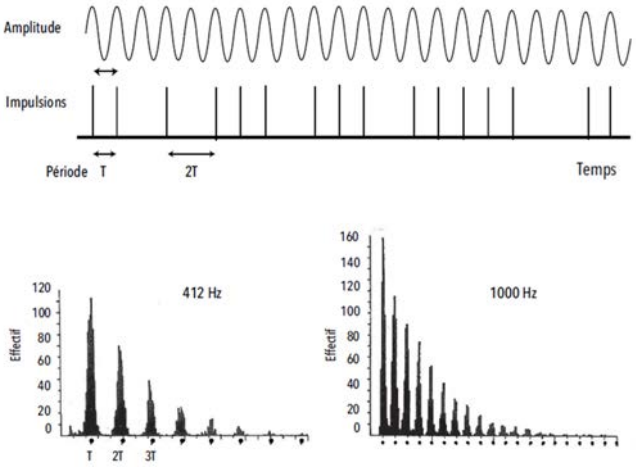
1.12. Neurons

- Spontaneous stochastic activity
- Binary devices (on: +50 mV, off: -60 mV)
- Electrochemical transmission
- Speed limitation (<4000 spikes/s, <<120m/s)
- **Refractory period**: time in which a nerve cell is unable to fire an action potential (nerve impulse)



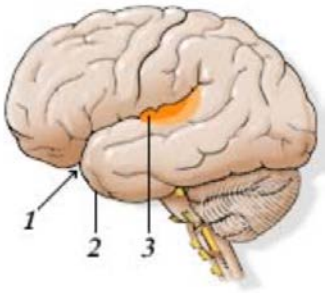
1.13. Coding of acoustic information

- **Phase-locking**: **tendency of a neuron to fire action potentials at particular phases of an ongoing periodic sound waveform, such as the sinusoidal waveforms** that are typically used in physiological studies of the auditory system. **Phase-locking as a means of frequency coding.**
- Information about sound intensity is coded in 2 ways: **rate (firing rate of neurons)** and **place (number of neurons active) → neural correlates of perceived loudness**
- Statistical behavior of neuron ensembles



1.14. Auditory pathways

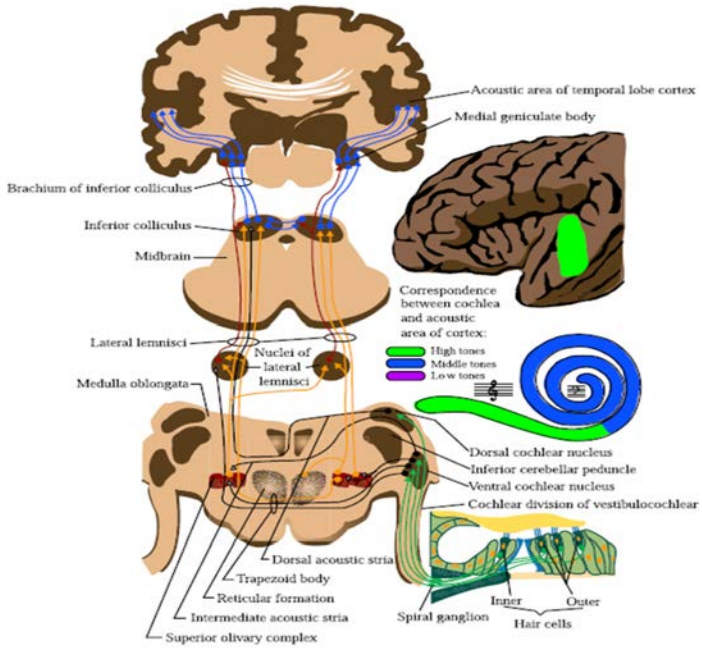
Auditory messages are conveyed to the brain via two types of pathway: the **primary auditory pathway** which exclusively carries messages from the cochlea, and the **non-primary pathway** (also called the reticular sensory pathway) which carries all types of sensory messages.
The primary pathway is short and it ends in the **primary auditory cortex** (3), located in the temporal area (2) within the lateral sulcus (1).



The tonotopic organization of the auditory nerve is also preserved throughout the auditory pathway.
Another categorization:

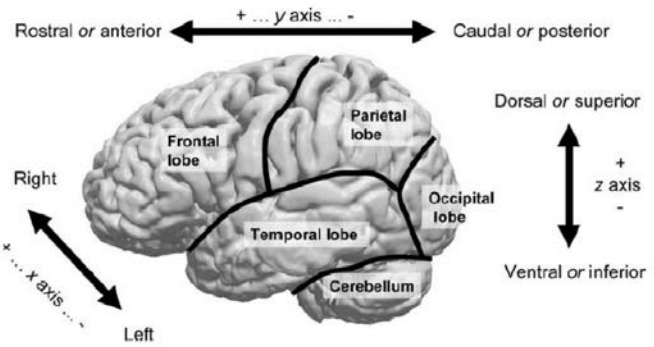
- **The "what" pathway**: monaural and receives information from only one ear. Concerned with: spectral (frequency) and temporal (time) features and hardly concerned with the spatial aspects. Focused on identifying and classifying different types of sounds.
- **The "where" pathway**: binaural and receives information from both ears. Involved in the localization of a sound stimulus.

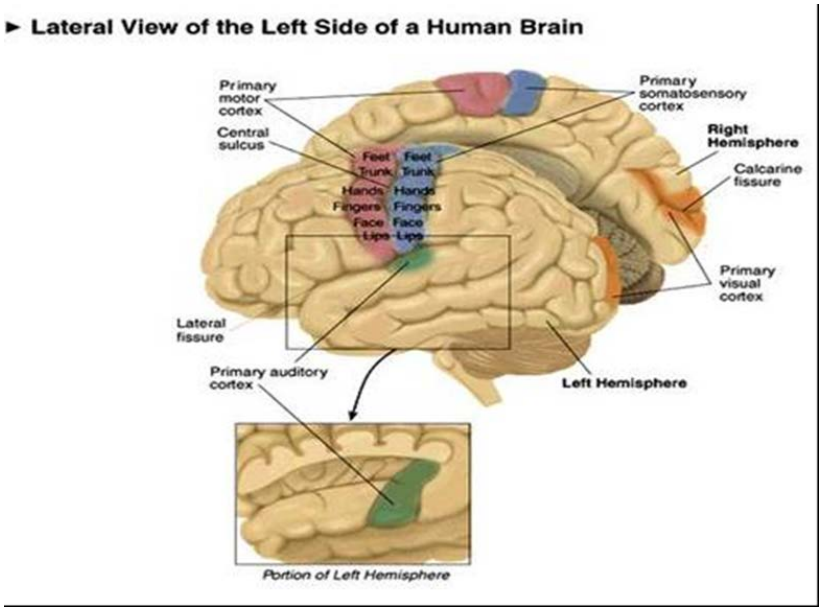
- Cross-lateralization
- Rough analyses (localization, loudness, noisiness, pitch…)



1.15. The brain

- **Hemispheres & lobes**:
  - 2 hemispheres
  - Each hemisphere contains 4 lobes: **frontal** lobe (cognitive functions, control of voluntary movement or activity), **parietal** lobe (processing information about temperature, taste, touch and movement), **occipital** lobe (vision) and **temporal** lobe (memories, taste, sound, sight and touch).
- **Lateralization**: functions are performed by distinct regions of the brain.
- **Neural plasticity**: ability of neural networks in the brain to change through growth and reorganization. These changes range from individual neuron pathways making new connections, to systematic adjustments (e.g. circuit and network changes from learning something new or from practice).





[PERFECTO QUIZ CARD] **Sympathetic vs. parasympathetic nervous system:**

Sympathetic (fight or flight):

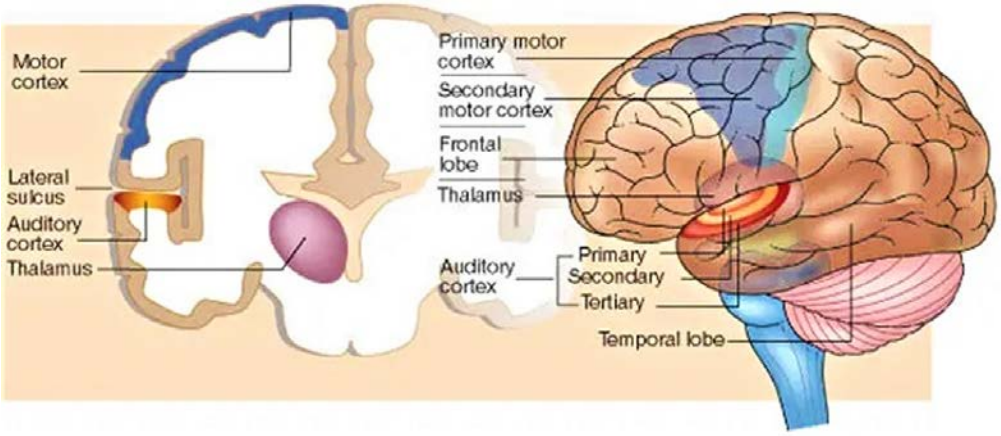
- controls unconscious fast actions (e.g. heart rate, breathing, adrenaline)

Parasympathetic (rest and digest):

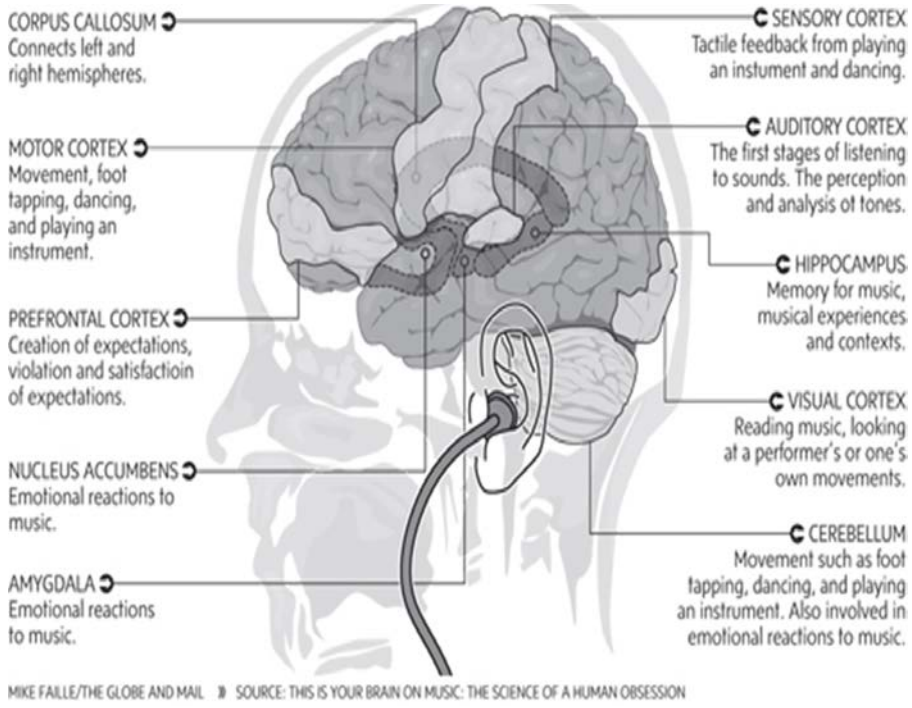
- Slower, regular processes
- Sleep
- Could be activated by music

1.16. Auditory cortex

- Temporal lobe
- Primary, secondary and associative regions



1.17. The musical brain



1.18. Timings in the musical brain (according to ERP studies)

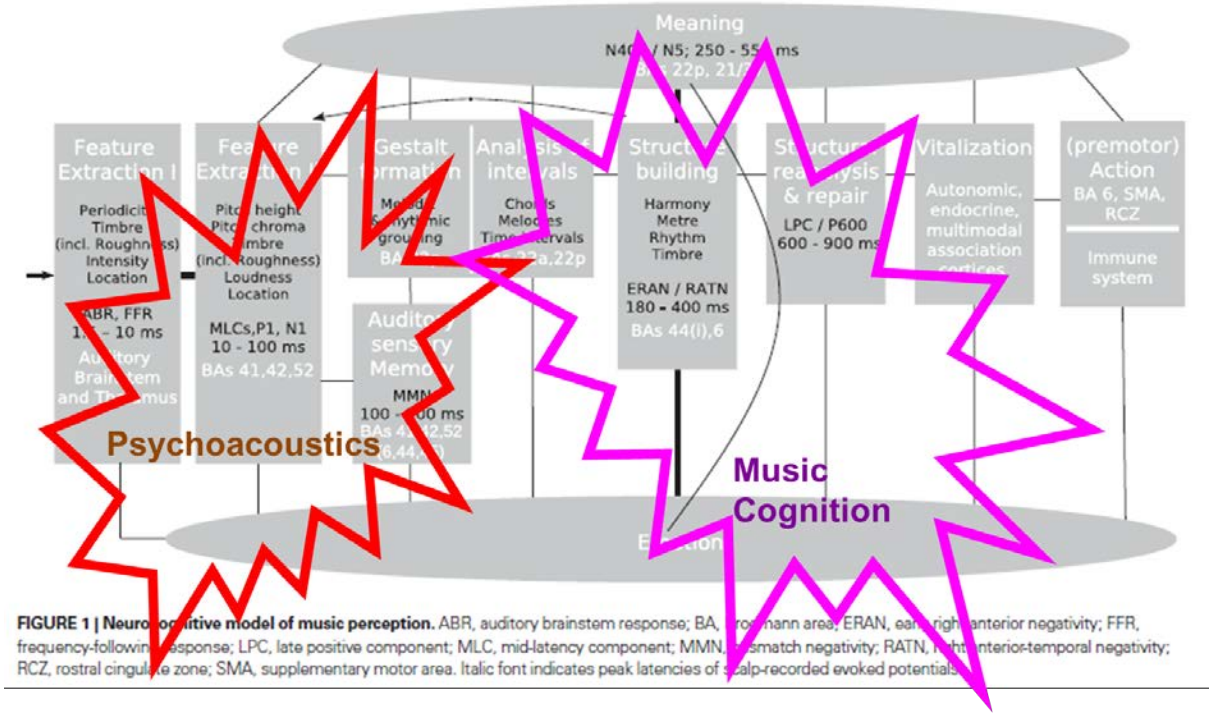


FIGURE 1 | Neuro[illegible]nitive model of music perception. ABR, auditory brainstem response; BA, [illegible]mann area; ERAN, ear[illegible]right anterior negativity; FFR, frequency-following [illegible]esponse; LPC, late positive component; MLC, mid-latency component; MMN, [illegible]smatch negativity; RATN, r[illegible]ht anterior-temporal negativity; RCZ, rostral cingul[illegible]e zone; SMA, supplementary motor area. Italic font indicates peak latencies of [illegible]alp-recorded evoked potentials.

1.19. Music Cognition

Perceiving, learning, remembering, producing and performing.

(PERFECTO QUIZ CARDS) **Categories of cognitive processing:**

1. **Automatic Processing**: unintentional/involuntary/effortless/with limited processing
2. **Controlled Processing**: flexible, intentional control of the individual, with conscious awareness.

2. Research methods and techniques

2.1. How to "measure" (music) perception and cognition?

- **Surveys** (not an experimental procedure)
- **Physiological measures**
- **Behavioral measures**
- **Electro-Encephalography (EEG)**
  - **Event-Related Potentials** (**ERP**)
- **Magneto-Electro-Encephalography (MEG)**
  - **Positron Emission Tomography (PET)**
  - **Functional Magnetic Resonance Imaging (fMRI)**

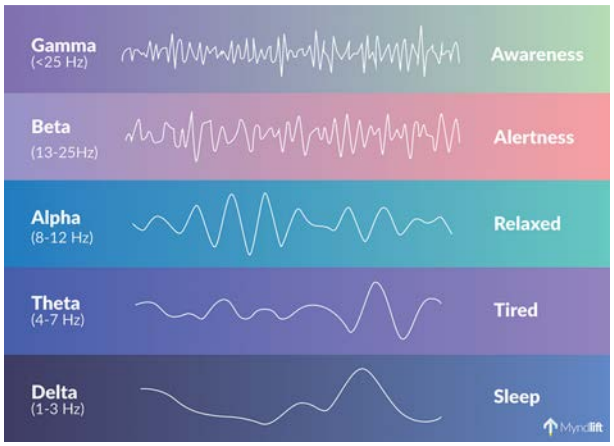
2.2. Physiological measures

Heart rate, skin conductance, blood pressure, temperature, muscle tension.

2.3. Behavioral measures

Errors solving a task, reaction time, relatedness or similarity (5 or 7 point Likert-type scale), choice (forced, not forced, 2/3/4 choices), eye-fixation, motion capture, Quantified Self.

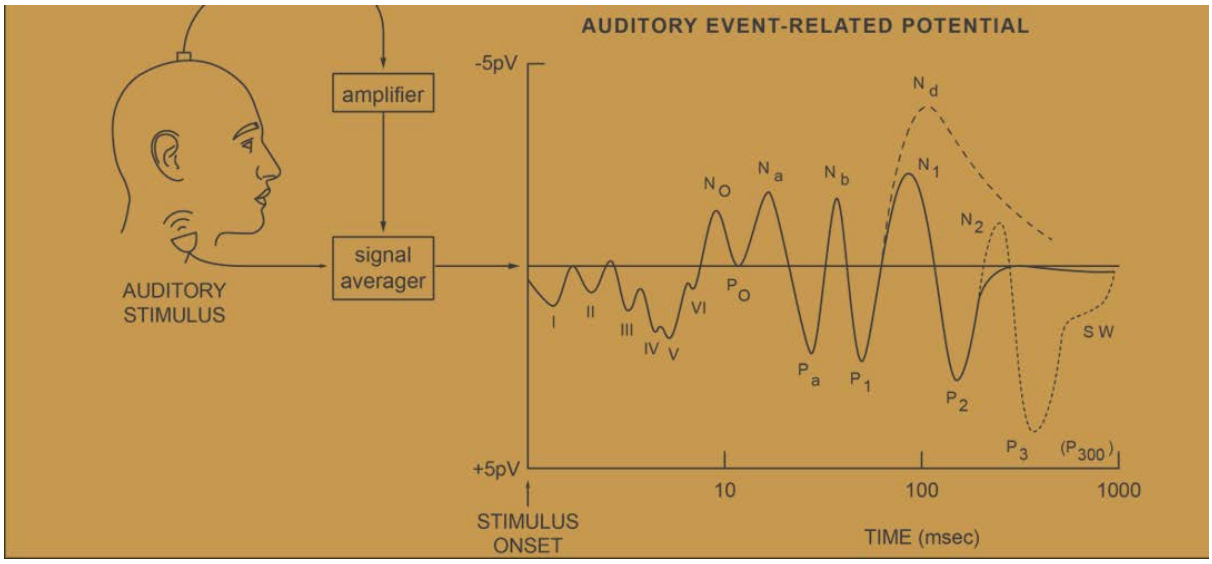
2.4. Electro-encephalography: spontaneous EEG



2.5. Electro-encephalography: Event-Related Potentials (ERP)

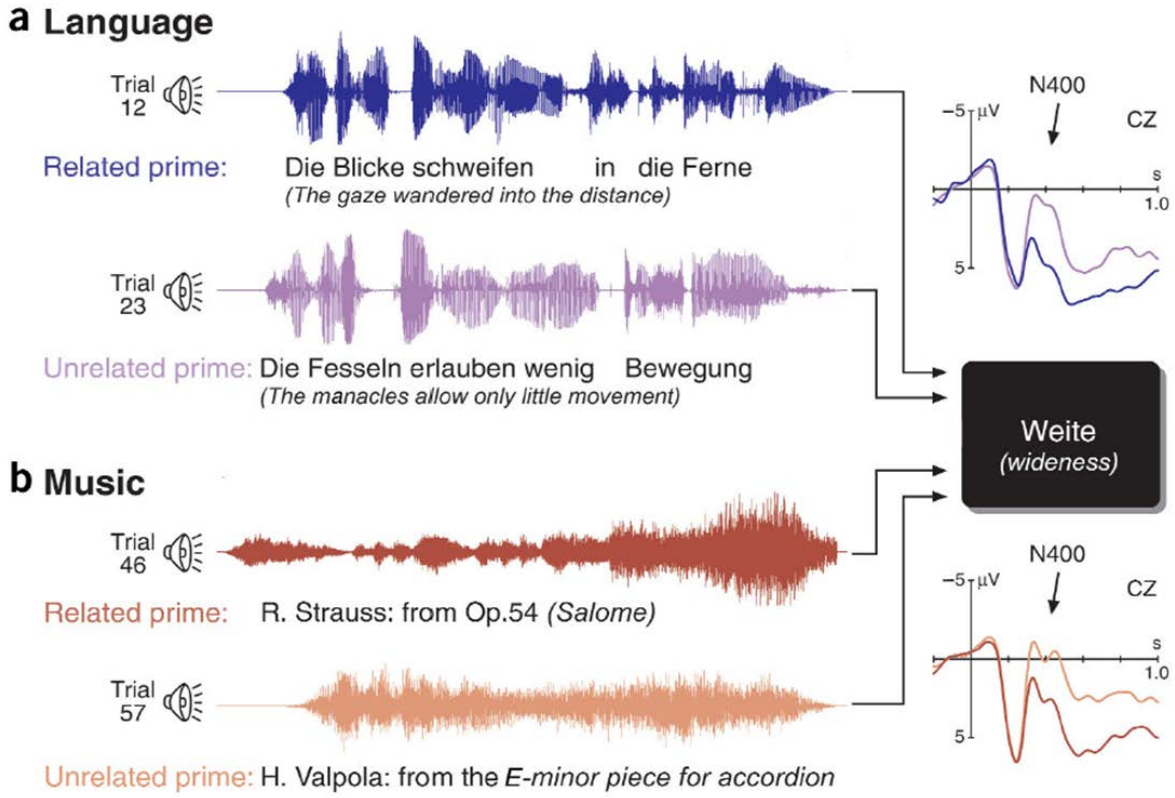
ERP: neural signal that reflects **coordinated activity of an ensemble of neurons**, observed "after" certain events or stimuli have been processed.

When a sufficiently large number of neurons having a **similar anatomical position and orientation** are **synchronously activated**, their summed fields may be strong enough to be detectable as ERPs or ER Fields at the **surface of the head**.



- **Good temporal resolution** but **poor spatial resolution**.
- **Pxxx**: positive peaks, **Nxxx**: negative peaks
- **MMN**: **Mismatch Negativity (80-200 ms after the event)**. It reflects the **changes in the content of our auditory short-term memory** (when something "new" is presented, a peak can be observed), even though we are not aware of that.
- **N400 (200-500 ms after the event)**: **lack of semantic associations** between terms or expressions.
- **P600 (around 600 ms)**: **syntactic violations** in language and in music.

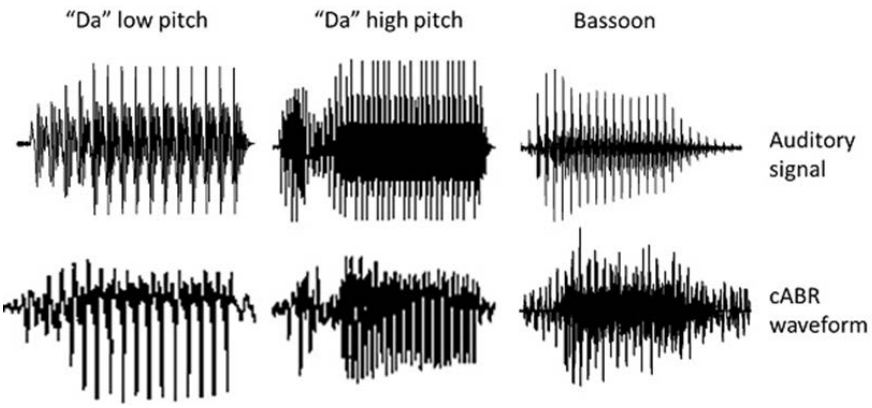
A **semantic context** is set either by speech (a) or by music (b). Hearing **unrelated words** generate a **negativity peak** in centro-parietal electrodes around 400 ms. But when there is a **semantic relationship** between the prime and the word, **the negativity is not observed**. The same **N400** is observed when the context is set using music and the words have no semantic relationship with it:



Koelsch et al. (2004). Music, language and meaning: brain signatures of semantic processing, Nature Neuroscience, 7(3).

2.6. Frequency-Following Response (FFR, a.k.a. cABR)

The FFR is an **evoked potential generated by periodic or nearly-periodic auditory stimuli**. Part of the **auditory brainstem response (ABR)**, the FFR reflects sustained neural activity integrated over a population of neural elements. It is often **phase-locked** to the individual cycles of the stimulus waveform and/or the envelope of the periodic stimuli.
cABR: auditory brainstem response to complex sounds.



2.7. fMRI: functional Magnetic Resonance Imaging

fMRI measures brain activity by detecting changes associated with blood flow. This technique relies on the fact that cerebral blood flow and neuronal activation are coupled. When an area of the brain is in use, blood flow to that region also increases.

- **Changes in oxygen consumption** in particular areas of the brain when performing a given task. Oxygen nuclei have nuclear magnetic resonance, i.e. their spins act as small magnetic fields and can be aligned according to a magnetic field. Depending on the metabolism (high vs low oxygenated blood), the spins misalign and then this process can be "plotted".
- **Excellent spatial resolution** but **bad temporal resolution**
- **Non-invasive and innocuous**

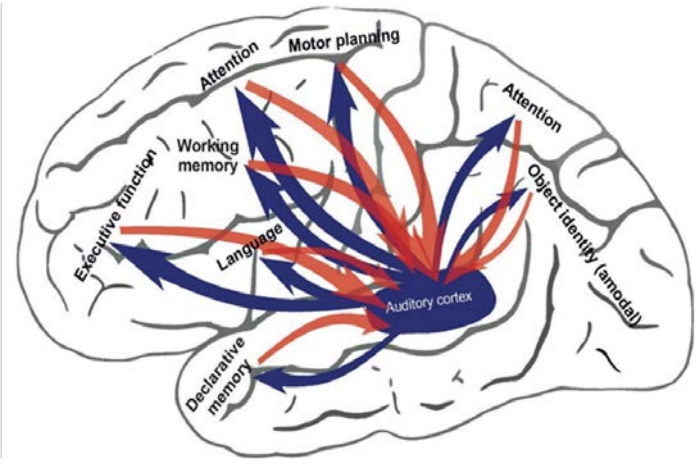
2.8. PET: Positron Emission Tomography

Radioactive contrast substance injected.
Measurements of local blood flow, metabolism, neuro-receptor bindings…

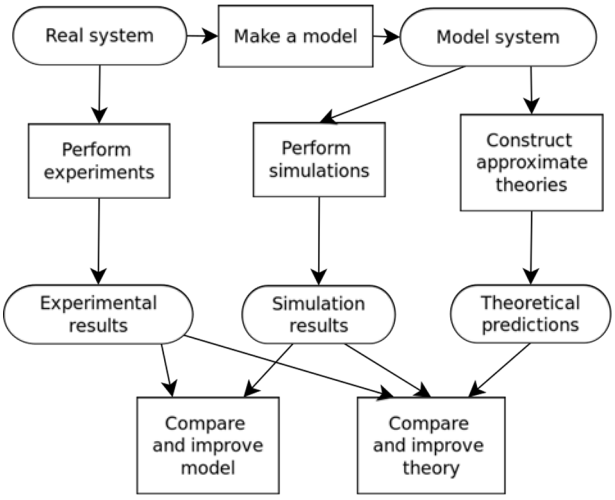
- **Good spatial resolution** but **poor temporal resolution**
- **Small health risk involved**

2.9. Diffusion Spectrum Imaging

Mapping the gradient of water (instead of blood) molecule diffusion. Estimation of the trajectories of **neuron firing** (instead of activity) in that area.
Connectome revealed: a connectome is a comprehensive map of neural connections in the brain ("wiring diagram"). A connectome is constructed by tracing the neuron in a nervous system and mapping where neurons are connected through synapses.



2.10. Computer simulation



3. Psychophysics of basic sound dimensions: Frequency resolution

3.1. Frequency resolution of our hearing system

Frequency resolution ability can be interpreted as the **filtering capability of the auditory system**. It can be studied by the shape of the underlying filter – called the **auditory filter** – that is centered at the frequency component of interest.

The ability to hear frequencies separately is known as **frequency resolution** or **frequency selectivity**.

- When signals are perceived as a combination tone, they are said to reside in the same **critical bandwidth**.
- A complex sound is split into different frequency components and these components cause a peak in the pattern of vibration at a specific place on the cilia inside the basilar membrane within the cochlea. These components are then coded independently on the auditory nerve which transmits sound information to the brain. This individual coding only occurs if the frequency components are different enough in frequency, otherwise they are in the same critical band and are coded at the same place, therefore they are perceived as one sound instead of many.
- The filters that distinguish one sound from another are called auditory filters, listening channels or critical bandwidths. Frequency resolution occurs on the basilar membrane due to the listener choosing a filter which is centered over the frequency they expect to hear, the signal frequency.

- Masking illustrates the limits of frequency selectivity. If a signal is masked by a masker with a different frequency, then the auditory system was unable to distinguish between the two frequencies.

Different methods can be used to estimate the shape of this auditory filter. In general these methods use a pure tone signal (mostly just above hearing threshold) and they are based on the assumption that a subject can detect the tone if the signal-to-noise ratio (SNR) within the auditory filter, exceeds a critical threshold. The reduction of the SNR is the reason that problems arise in a noisy environment. Noise can be interpreted as all sounds different from the signal pure tone. Most methods use the principle of changing the level and/or frequency content of the noise in order to obtain the shape of the filter. The noise added on purpose in the method is called **'the masker'** and the differences between methods are mainly related to the maskers applied. **The auditory filter is nonlinear and its shape depends on the method and signal level used**, e.g. **the filter broadens with increasing signal level**.

- Which is the precision of our hearing system to resolve (separate) different components (partials) of a sound stimulus? Most of our knowledge about that comes from studies about **masking**. This has also implications for: **loudness**, **pitch** and **timbre**. The sensitivity of human hearing depends on the frequency range where the sound is present. We are more sensitive in the "low" regions and less sensitive at "high" frequencies.
    - In the cochlea of the human ear, sounds are processed in spectral bands, which are independent such that sounds in separate bands do not interfere with each other, but **sounds within the same band do interfere with the perception of each other (auditory masking)**. The width of these bands is frequency dependent and increases with increasing frequency. This has been approximated with several models, including the Bark and ERB scales.
    - The distance between pitches are perceived differently depending on their frequencies. A perceptually small step in pitch (measured in frequencies) is much larger at higher frequencies than at low frequencies. This can be approximated with the Mel scale.

3.1.1. Range of hearing

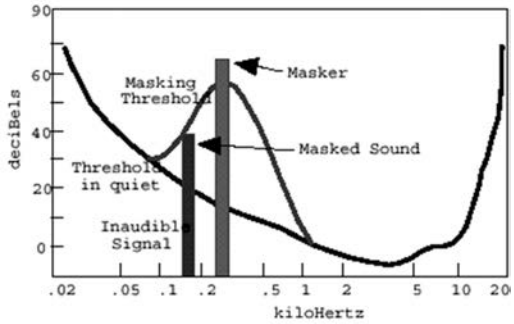
Only one tone:

- **Hearing threshold**: our hearing is limited by the hearing threshold (a.k.a threshold in quiet): faintest sounds that are just audible without any other sounds being present. Measured by playing tones and someone detecting the presence of the tone. Threshold of hearing as a function of frequency (**ISO 389-7 hearing threshold**). Tones beyond that threshold are audible.
- **Risk of damage threshold**: risk damaging the auditory system.
- **Threshold of pain**: so painful that not only damages the ear but also creates pain.
- **Music**: from 10 Hz to 10 kHz
- **Speech**: from 80/90 Hz up to 5 to 7 kHz

3.2. Masking

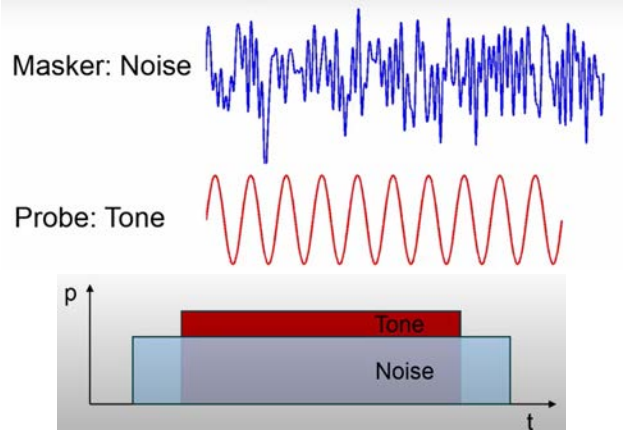
The **unmasked threshold** is the quietest level of the signal which can be perceived without a masking signal present (hearing threshold). The **masked threshold** is the quietest level of the signal perceived when combined with a specific masking noise. The **amount of masking** is the difference between the masked and unmasked thresholds.

- Simultaneous masking occurs when a sound is made inaudible by a noise or unwanted sound of the same duration as the original sound.



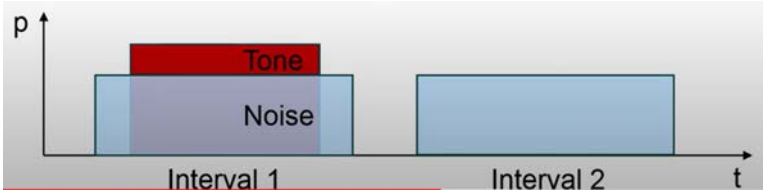
3.2.1. The Masking Experiment

- **Detection in a single interval**: Bias → somebody might be very sensitive when it is just something audible, someone else might want to hear it very clearly.



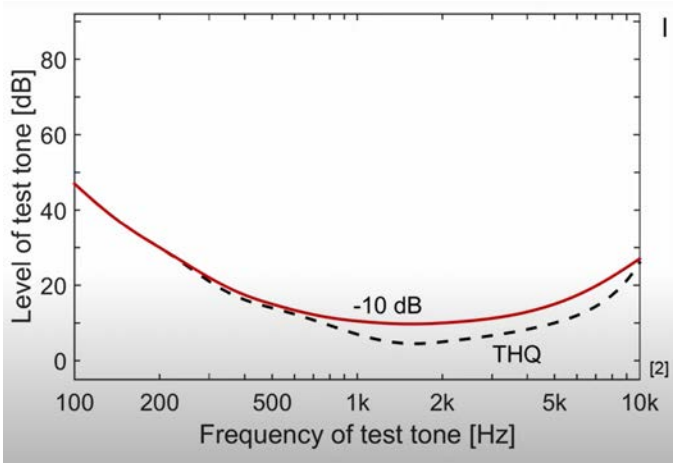
Find tone's amplitude to be just detectable (threshold).

- **Forced-choice detection of the interval with the tone**: 2, 3 intervals and hide the tone in some interval. Respondents should detect which interval contains the signal. **75% is the threshold. This is the standard procedure used for masking experiments**.

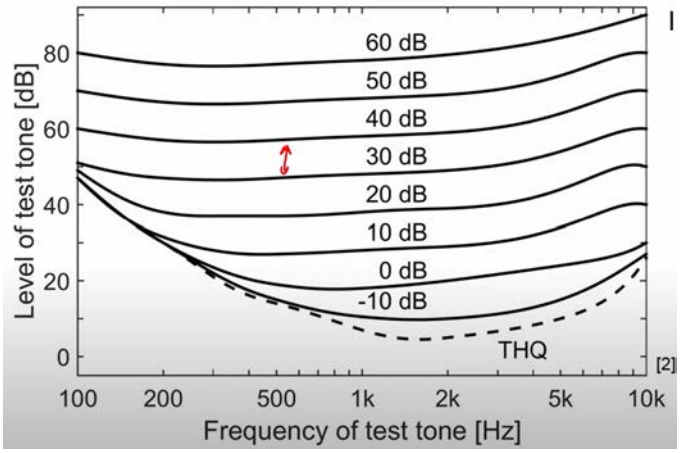


3.2.2. Simultaneous Masking of a Tone by White Noise

White Noise has constant spectral energy per frequency band (constant spectral density level l in dB/Hz). If we play into that noise a tone we get the masking pattern for that tone (when l is -10 dB):



If we increase the level of the noise in 10 dB steps, the tone level has to be increased in equal steps of 10 dBs, basically a linear process.

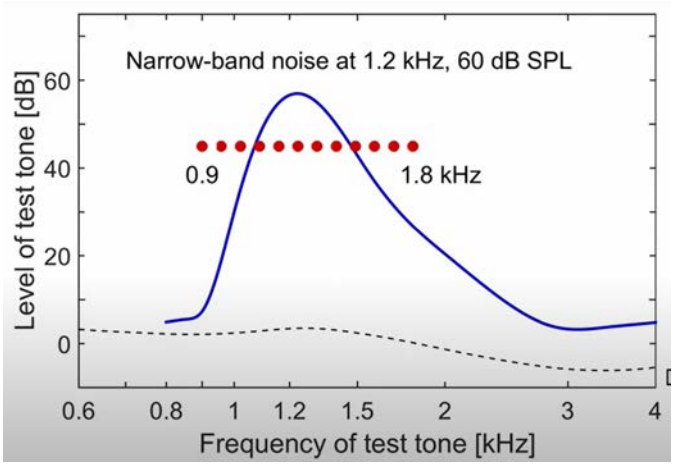


Towards higher frequencies there is a **slight rise**: **our auditory filters get wider towards high frequencies**.

What happens if we reduce the bandwidth of the noise, so that the noise is more specific to certain frequencies **(narrowband noise)**?

3.2.3. Simultaneous Masking of a Tone by Narrow-Band Noise

If we shift the center frequency, we see that masking shifts with it, it is specific with the center frequency, there are areas with no energy of the noise but masking is still present.



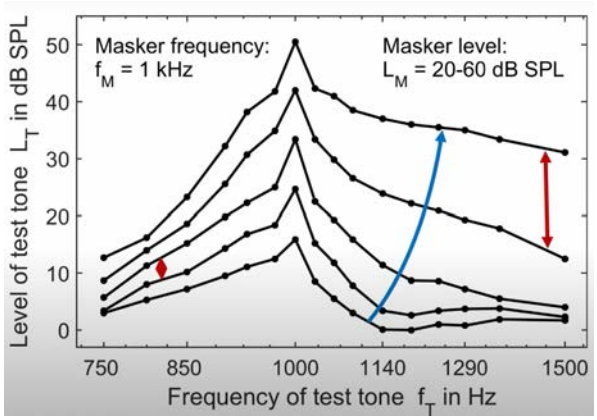
Summary (Masking part 1):

- **Hearing range**:
  - Hearing threshold as lower level of audibility, rises sharply towards low and high frequencies.
- **Simultaneous masking by noise**
  - Noise raises threshold of audibility for tones
  - 10 dB more noise → 10 dB threshold increase
  - Masking is **specific to frequencies of the masker**, **decays to higher and lower frequencies**.

3.2.4. Simultaneous Masking of a Tone by a Tone

Masking of a narrowband noise is specific and close to the center frequency of the noise. Similar effects happen with masking of a tone by a tone. The masking pattern changes, most strongly masking near the center frequency of the tone and decays towards higher and lower frequencies, however **at higher levels and higher frequencies masking is more pronounced that for lower frequencies**. Strong excessive masking, **masking increases more than proportionally when we increase the level of the tone masker**. **Masking is overly proportional, upward spread of masking**, **with high masker levels one can mask sounds at faraway frequencies even very pronouncedly**.

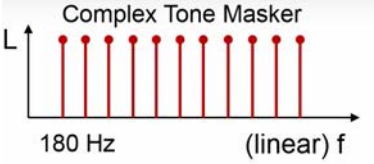
- **At high levels more masking towards higher frequencies**.
- **Upward spread of masking**: masking increases more than proportionally with masker level at high frequencies.

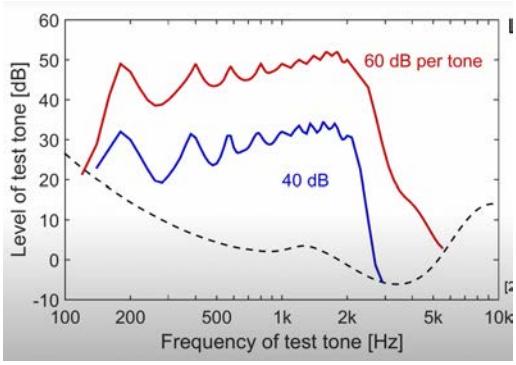


We now speak of the masked detection threshold: **steep curve for low frequencies** (the basilar membrane is not much affected by the sound) and **shallow slope for high frequencies**.

3.2.5. Simultaneous Masking of a Tone by a Complex Tone

Compose the masking patterns as the **sum of the individual masking patterns for each tone in the complex tone**. **Spectral gap: a bit more masking in those valleys (between tones)**. But overall, masking is really **determined by individual components**. Many natural sounds are made of complex sounds, more realistic.

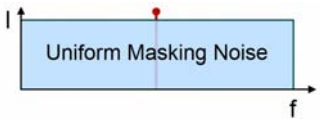


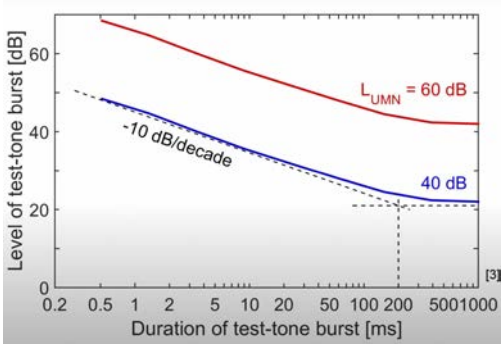


3.2.6. Simultaneous Masking: Effect of Tone Duration

Irrespective of the noise level, **the energy of the test tone needs to be turned up, if the test tone gets shorter than about 200 ms**. Beyond 200 ms, we have steady-state conditions. **Shorter than 200 ms we need to increase the level of the tone, and that increase is 10 dB per decade**.

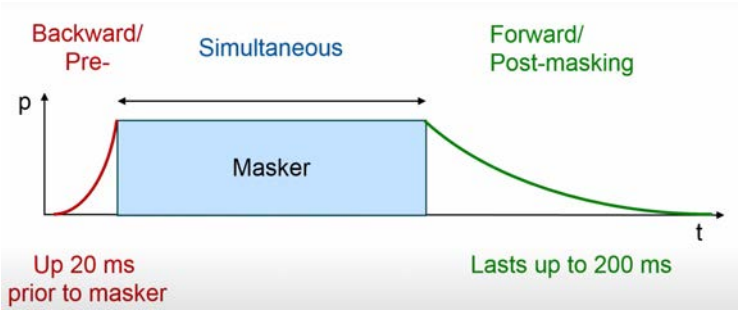
- Masking increases by around 10 dB/decade when test tones are shorter than 200 ms, **independent of masker level**.
- **The auditory system is a long-term integrator over 200 ms that looks for the energy of the tone to detect the tone in the mask**.
- If you have short impulsive sounds, to hear them, for them to have a certain loudness, their level needs to be very high and it can be dangerous.





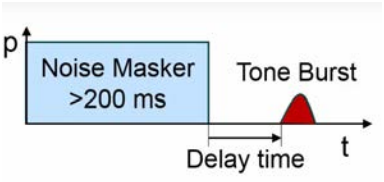
3.2.7. Temporal Masking Overview

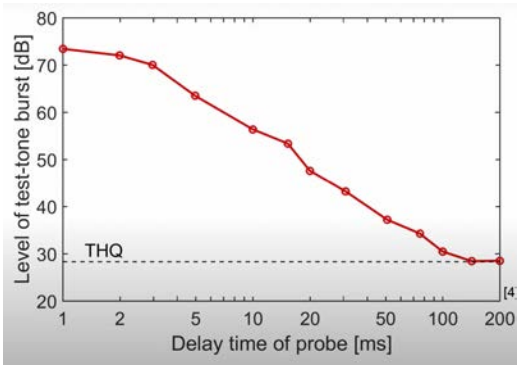
We can see masking as a relative timing issue of the tone relative to the masker. There is also a situation where the probe tone is present before a masking sound and there is even masking (pre masking or backward masking). You have to raise the level of the probe, about 20 ms, or can be 50 ms, with hearing impairment these durations are more. The other masking, more pronounced, forward or post-masking, up to 200 ms.



3.2.8. Forward Masking (Post-Masking)

The masker has to be longer than 200 ms. On the logarithmic scale of time, you can see almost linear decay of masking over time, with time after the mask lasting out to beyond a hundred ms.





Summary (masking part 2):

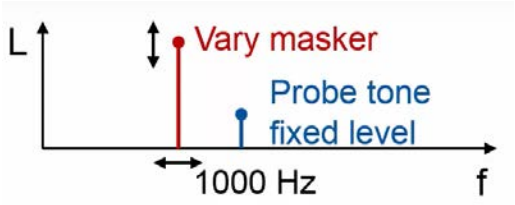
- **Simultaneous masking by tones**:

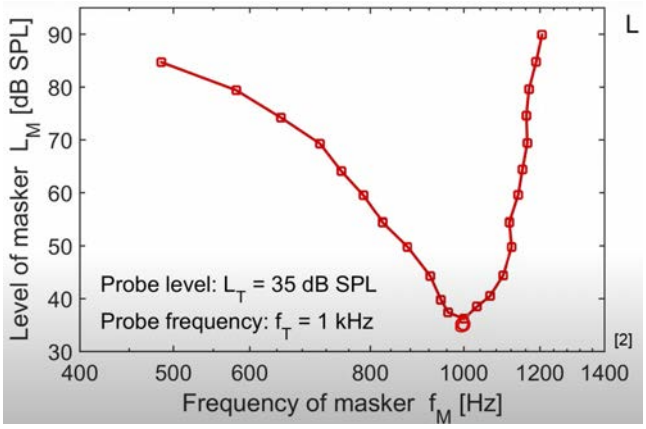
- **Masking strongest near masker frequency**
- **Upward spread of masking**: masking increases strongly at high frequencies if masker level >50 dB.

- **Temporal masking effects**
  - Raise probe level by 10 dB for duration < 200 ms
  - **Backward masking**: masking before sound onset
  - **Forward masking**: masking decays over 200 ms after the end of masker.

3.2.9. Simultaneous Masking of a Tone by a Tone: Tuning Curve

We keep the probe tone level fixed





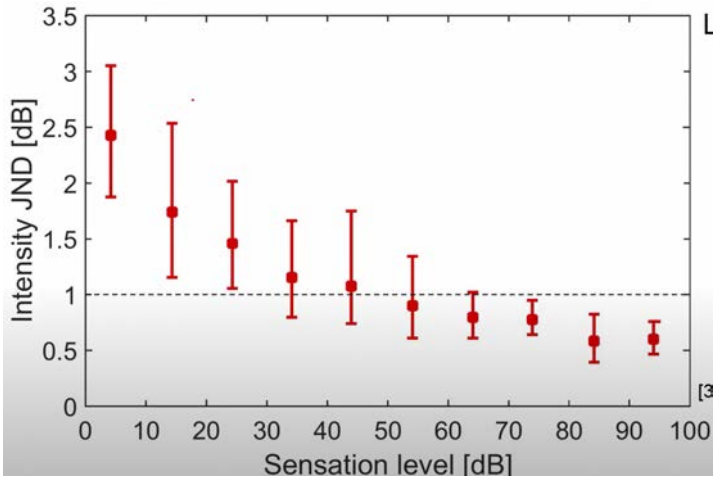
Masker level necessary to mask the probe tone.

If the masker is of a higher frequency it will be very hard to mask a low frequency probe. **The tuning curve increases sharply towards the higher frequencies** → filter function in the auditory system. When energy from the masker falls into the filter that also holds the probe → energetic masking, critical band concept.

3.2.10. Just-Noticeable Differences in Intensity

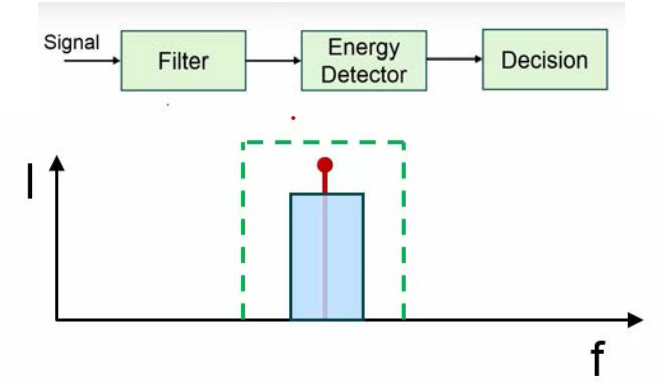
Certain threshold to find differences in intensity. The sensation level is not the DB SPL, it is relative to the threshold of hearing. 10 dB sensation level → 10 dB above the threshold of hearing.

- Just above the threshold → sensitivity is quite poor (about 2 dB of difference)
- From 30 to 70 dB → sensitivity of 1 dB , we cannot detect changes below that.
- Masking at the output of an auditory filter will also be produced by a change in the probes energy at the output of the filter that has to overcome this 1 dB threshold.



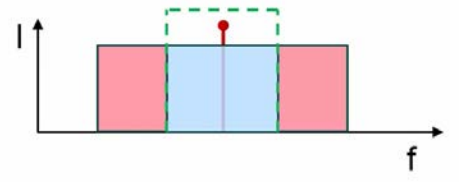
3.2.11. Critical Band Concept

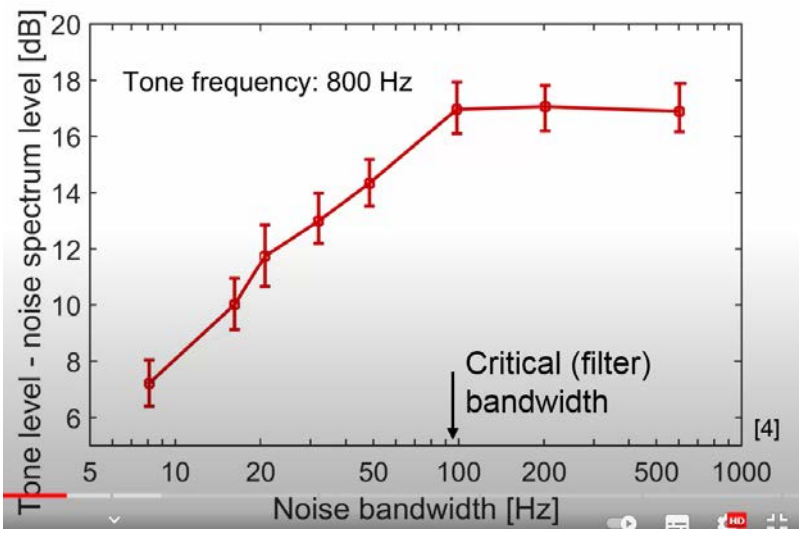
Auditory filter is composed of a bank of filters. Look at the **output of the filter for changes of intensity or energy to make a decision where the probe tone was heard or not**, signal comes into the auditory system and it is filtered. **If we place a masker noise, then the auditory system can detect the presence of the tone only by an energetic change at the output of the filter that overcomes the energy of the noise (around 1 dB, increase at the output of the filter)**.



Noise and tone energy in the filter: all noise energy determines the tone's detection threshold.

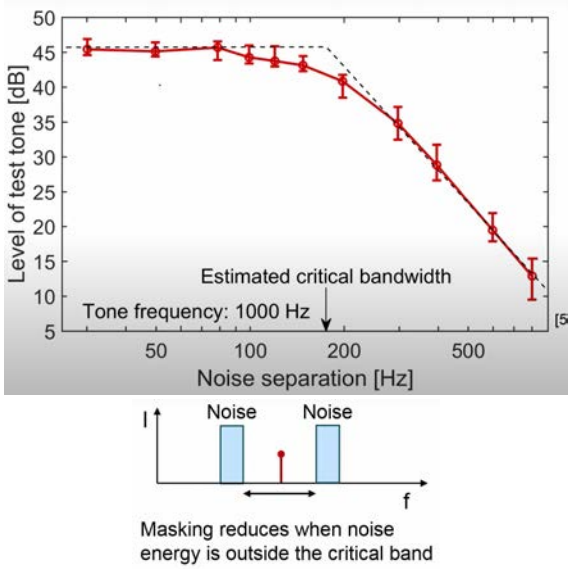
Nice **linear increase until certain bandwidth**. Outside that area, the energy of the noise does not contribute to the masking. **Noise energy outside the filter is not relevant to tone detection: energy comparison is done at filter output**.





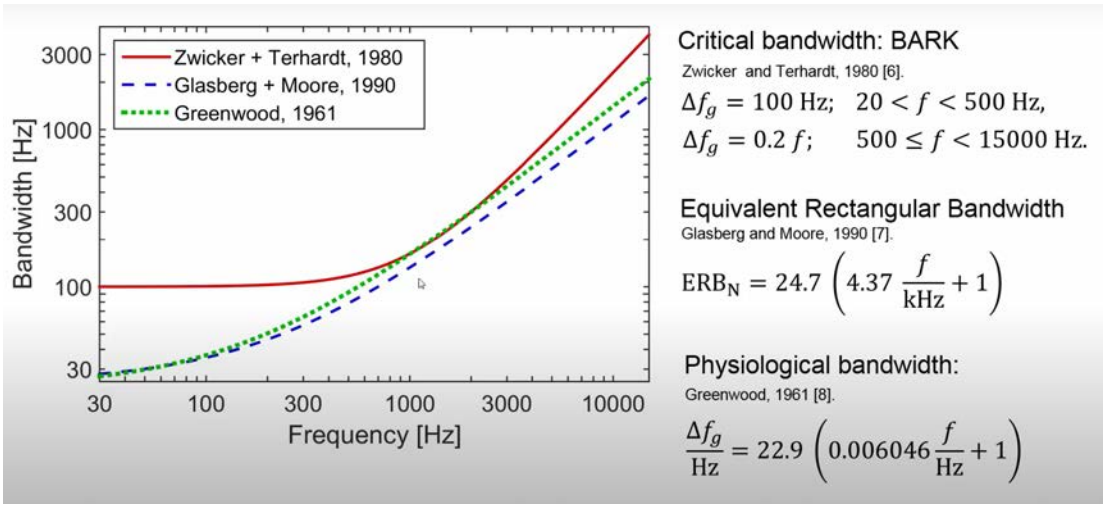
3.2.12. Critical Band in Masking: Notched-Noise Experiment

Tone with fixed level and frequency, two flanking noises, when further apart the noise energy will not fall within the filter, so the probe tone can be detected easily. Noise separation small → raise the tone level to detect it. Corner point → estimate of critical bandwidth.

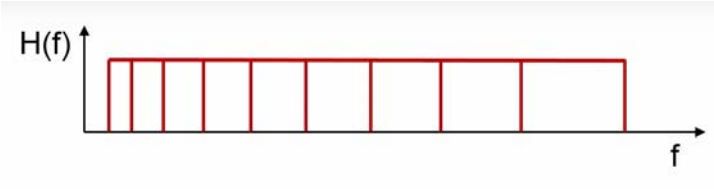


3.2.13. Critical Bandwidth of the Auditory Filter

- **Zwicker + Terhardt (1980)** → **BARK**: about 100 Hz wide at low frequencies (about constant), and about 20% of the center frequency of the filter at high frequencies (above 500 Hz).
- **Glasberg + Moore (1990)** → **Equivalent Rectangular Bandwidth**: different at low and high frequencies. Much narrower at low frequencies, more narrow at high frequencies compared to Zwicker. General idea: high frequencies constantly relative to the frequency.
- **Greenwood (1961)** → **physiological bandwidth**: physiological measurements. Pretty close to Zwicker at mid frequencies, but narrow at low and high frequencies.

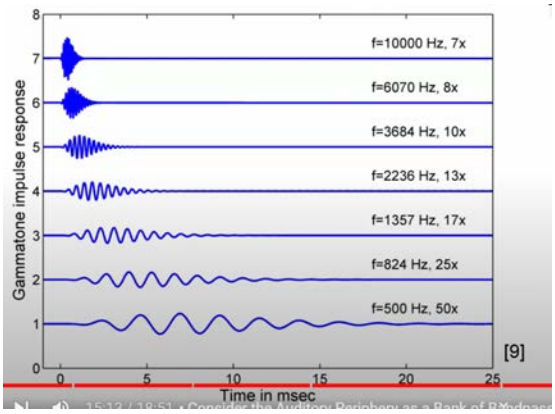


3.2.14. Consider the Auditory Periphery as a Bank of Bandpass-Filters:



Critical bands wider at high frequencies:

- Shorter group delay
- Less temporal dispersion
- 1/f noise to compensate wider filters
- Bandwidth is 1 ERB or 1 BARK



Gammatone filters widely used for a linear filterbank. **For low frequencies, the ringing of the filter is very long** (several ms for the filter to ring out after you excite it with a click). **At high frequencies**, energy decays much more quickly, so temporal resolution is better, because the **ringing of the filters is much shorter**.

When the filters are wider, if you have wide noise there will be more energy in the high frequency filter, we often use **1/F noise or pink noise** which will approximate to compensate for that increase in the bandwidth.

Summary (auditory frequency selectivity and critical bands):

- **Auditory system behaves like a filterbank**:
    - **Energetic masking, loudness**: frequency selective
    - **Phase sensitivity limited across filters**: the sensitivity of phase changes across filters is limited, but if they are in the same filters, the phase sensitivity is fairly high.
- **Critical bandwidth**:
    - Narrow filters at low frequencies, wide at high
    - Bandwidth roughly proportional to frequency: increases with center frequency.
    - Considered as a filterbank with overlapping bandpass-filters.

Summary II (masking) + additional info:

- How effective the masker is at raising the threshold of the signal depends on the frequency of the signal and the frequency of the masker.
- The greatest masking is when the masker and the signal are the same frequency and this decreases as the signal frequency moves further away from the masker frequency
- **On-frequency masking**: the masker and the signal are within the same auditory filter.
- **Off-frequency masking:** The amount the masker raises the threshold of the signal is much less, but it does have some masking effect because some of the masker overlaps into the auditory filter of the signal. Off-frequency masking requires the level of the masker to be greater in order to have a masking effect.
- The masker masks high frequency signals much better than low frequency signals.
- As the masker frequency increases, the masking patterns become increasingly compressed: high frequency maskers are only effective over a narrow range of frequencies, close to the masker frequency.
- MP3: parts of the signals which are outside the critical bandwidth are represented with reduced precision. The parts of the signals which are perceived by the listener are reproduced with higher fidelity.
- Varying intensity levels can also have an effect on masking. The lower end of the filter becomes flatter with increasing decibel level, whereas the higher end becomes slightly steeper.

3.3. Basilar membrane behavior

- A sine sound does not only excite the part of the basilar membrane that corresponds to its frequency but also other frequencies as well (**traveling waves**). The traveling wave moves from the **base** (high frequencies) to the **apex** (low frequencies), and declines after passing the resonance frequency. Therefore, we expect that a sound at a given frequency also affects the detection of a sound at another frequency.
- **Traveling wave with multiple “peaks”. Peaks corresponding to loud “partials” of the sound**.
- Tonotopic distribution (**high frequencies close to oval window**): tonotopy is the **spatial arrangement of where sounds of different frequencies are processed in the brain**. **Tonotopy in the auditory system begins at the cochlea**.

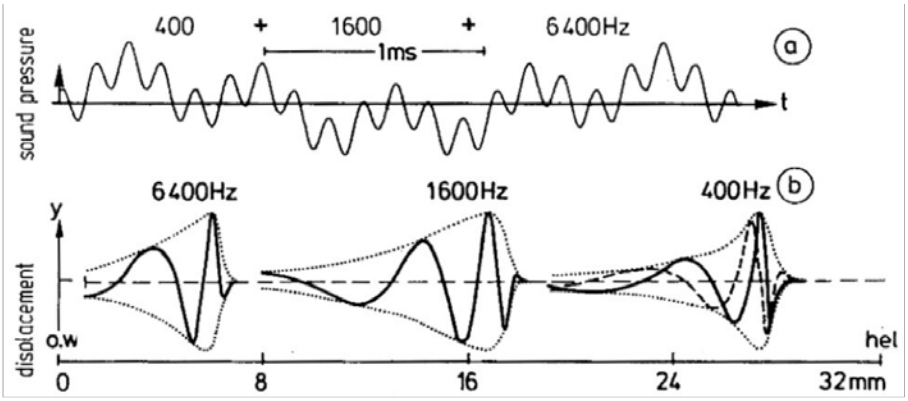


Fig. 3.5a, b. Schematic drawing of the transformation of frequency into place along the basilar membrane. In (a) three simultaneously presented tones of different frequencies expressed as compound time function produce travelling waves (b), that reach their maximum at three different places corresponding to the characteristic frequencies

3.4. Causal mechanisms of masking: basilar membrane

- **Asymmetry of masking**: upwards spreading
- The louder, the more masked range.

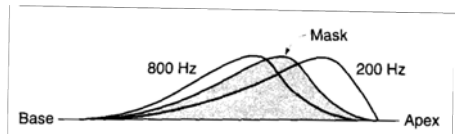
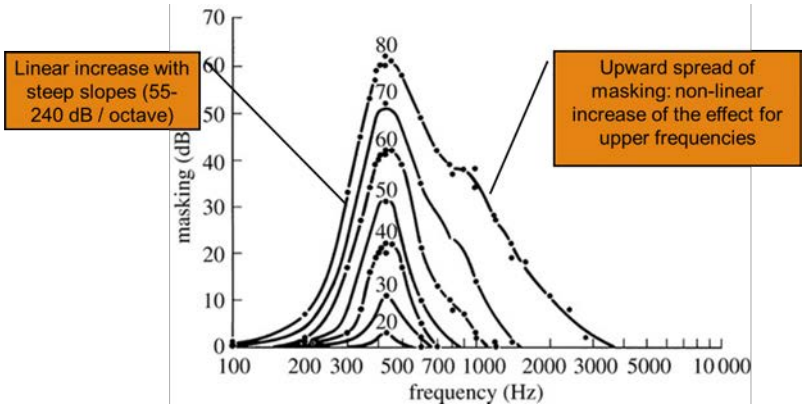


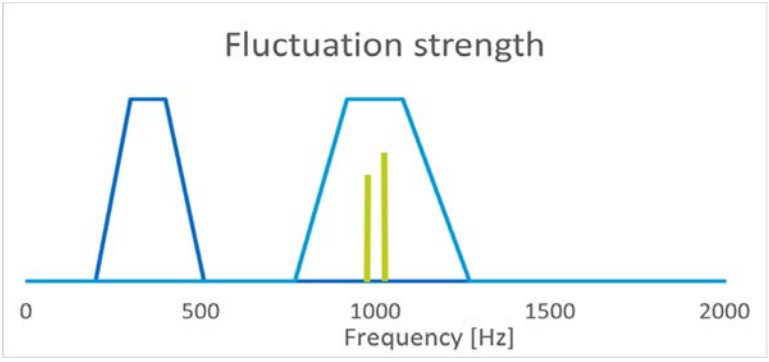
Figure 10.35
Vibration patterns caused by 200- and 800-Hz test tones and the 400-Hz mask, taken from basilar membrane vibration patterns in Figure 10.29. The pattern for a 400-Hz tone is used for the masking pattern (shaded). Since the mask actually contains a band of frequencies, the actual pattern would be wider than is shown here. It would, however, still be asymmetrical and would overlap the 800-Hz vibration more than the 200-Hz vibration.

3.5. Masking patterns

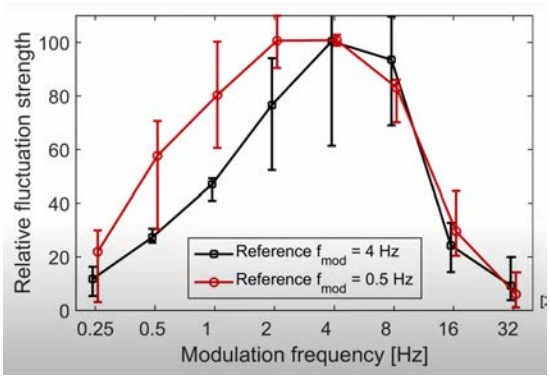
Example for a narrowband noise masker centered at 410 Hz. Each curve shows the elevation in threshold of a pure-tone signal as a function of signal frequency. The overall noise level in dB SPL (signal-noise difference) for each curve is indicated in the figure:



3.6. Beating (or *fluctuation strength*)



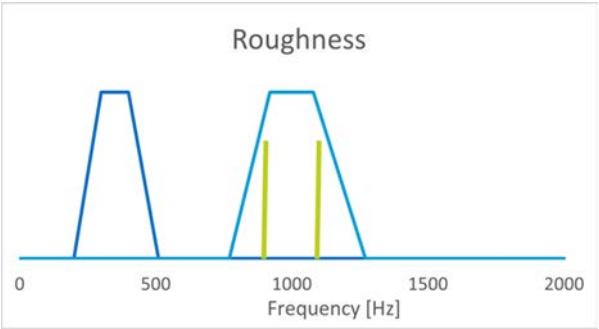
- Sound fluctuating.
- If two tones are heard simultaneously, and are in the same critical band, they will be perceived as a single tone with a "beat frequency" modulating the loudness of the tone. The beat frequency is the difference in Hz between the two tones' frequencies. If the tones are in different critical bands, they will be perceived as two distinct tones. 24 critical bands as defined by Bark scale.
- Peaks at around 4 Hz modulation frequency



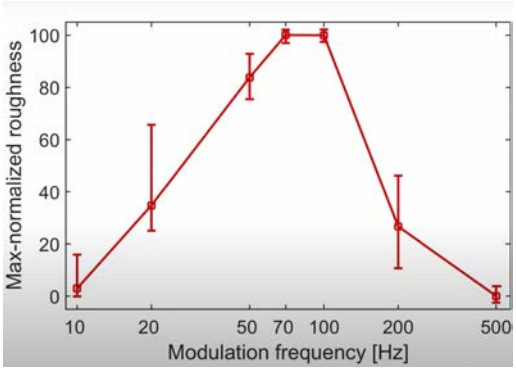
Fluctuation strength indicates the perception of fluctuation which peaks for 4 Hz AM or FM.

3.7. Roughness

Expresses how rough a sound is. Musically, could be equivalent to "dissonance"



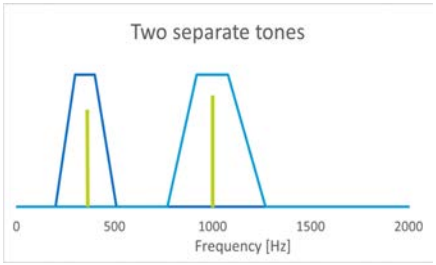
- Roughness indicates the rough-sounding perception of stimuli with AM or FM rates around 70 Hz.



Summary (sound quality measures):

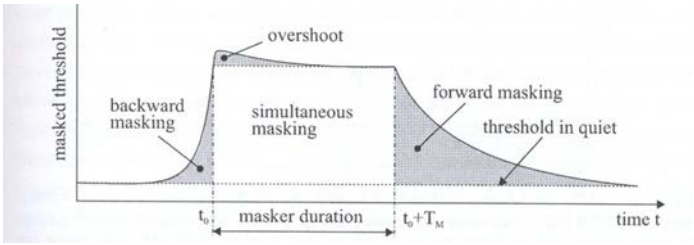
- Sharpness:
  - Perception of “sharp” sounding stimuli
  - High-frequency energy contributes dominantly
- Fluctuation strength and roughness:
  - Related to AM or FM modulations with perception of fluctuating (around 4 Hz) or rough (around 70 Hz).

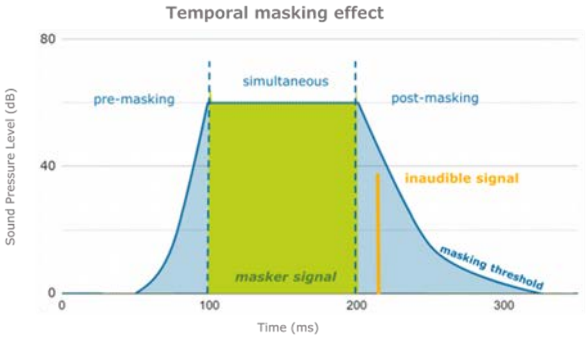
3.8. Full resolution



3.9. Temporal masking

Forward: 30-300 ms
Backward: 5 ms





Herre, Jürgen & Dick, Sascha. (2019). Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review. Applied Sciences. 9. 2854. 10.3390/app9142854.

3.10. Causal mechanisms of masking: **neural functioning**

Forward masking:

- Response of the basilar membrane continues after the end of the masker ("**ringing**").
- Masker produces short-term adaptation or fatigue in the auditory nerve or higher centers in the auditory system.
- Neural activity persists at some level in the auditory system (blocking that of the second tone)

Backward masking:

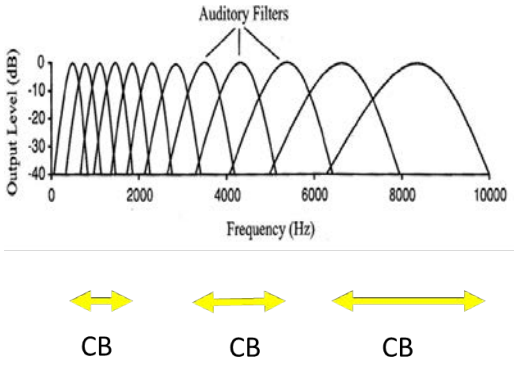
- **Sensory memory of the first tone not properly formed**.

3.11. Auditory Filters

- Fletcher (1940) postulated that the auditory system behaves like a bank of pass-band filters with overlapping passband.
- Helmholtz (1865) already had similar ideas.
- Auditory filters can be measured in **amplitude** and **phase** as functions of **frequency**.
- Frequency selectivity can be modeled with a bank of bandpass filters with overlapping bands. Each different point (around 1mm) of basilar membrane corresponds to a filter with a different center frequency.

Measurement of auditory filters:

- Present two sinusoids with same level and same frequency to the listener. The level was adjusted just above sensation level.
- Next, vary the frequency of the sinusoids in the opposite direction.
- The two sinusoids become inaudible at the point where they do not fall into one auditory filter anymore. In this case, the energy within each of the two auditory filters becomes too small to be detected).



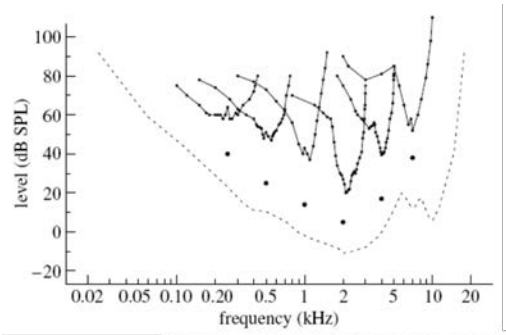


Figure 1. Psychophysical tuning curves (PTCs) determined in simultaneous masking, using sinusoidal signals at 10 dB SL. For each curve, the solid circle below it indicates the frequency and level of the signal. The masker was a sinusoid which had a fixed starting phase relationship with the 50 ms signal. The masker level required for threshold is plotted as a function of masker frequency on a logarithmic scale. The dashed line shows the absolute threshold for the signal. Data from Vogten (1978).

3.12. Critical Bands

[PERFECTO QUIZ CARD] **Critical bands in 4 steps:**

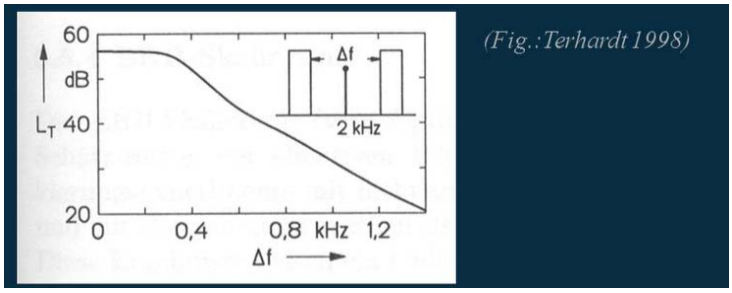
1. an individual's pure tone threshold (in quiet) is at 100 Hz

2. with low-intensity WBN, the threshold increases

3. as WBN continues to increase the masked threshold continues increase until the masker and tone are at the top of the critical band centered on 1000 Hz

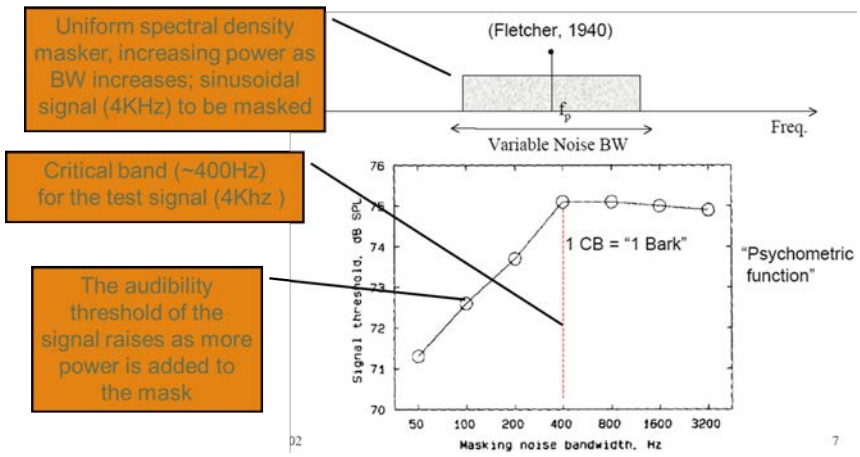
4. once critical band is full, fletcher argues that pure tone threshold will not change with increase in WBN

- Zwicker (1961) measured the critical bandwidth using **two narrow-band maskers** which masked a sine target at the center of the critical band. He recorded the detection threshold of the sine tone (varied in level) as a function of the **frequency gap between both maskers**. Unfortunately, the interference between the lower frequency noise masker and the sine target led to interference effects, and combination tones at different frequencies become audible, while the signal remains undetected. This leads to the abrupt decrease in the detection threshold at 0.3kHz.



- **The distance in the basilar membrane whereby two tones of different frequencies do not interfere themselves anymore**.
- A frequency region that is **"critical" to masking**, **tone interactions** and **loudness summation**.
- **A gammatone filterbank is the most accurate model of the auditory filters** and their critical bands.

3.13. Measuring Critical Bands (Fletcher's experiment)



3.14. Critical Bands: Bark estimation

- Zwicker et al.
- Barkhausen (proper name) → Bark

$$B_c = 52548 / (z^2 - 52.56\ z + 690.39),\ \text{with } z \text{ in bark}$$

Critical band partition

Z/Bark	f_u/Hz	f_o/Hz	Δf_G/Hz	f_m/Hz
0	0	100	100	50
1	100	200	100	150
2	200	300	100	250
3	300	400	100	350
4	400	510	110	450
5	510	630	120	570
6	630	770	140	700
7	770	920	150	840
8	920	1080	160	1000
9	1080	1270	190	1170
10	1270	1480	210	1370
11	1480	1720	240	1600
12	1720	2000	280	1850
13	2000	2320	320	2150
14	2320	2700	380	2500
15	2700	3150	450	2900
16	3150	3700+	550	3400

Critical Band (Bark)	Center Frequency (Hz)	Bandwidth (Hz)
1	50	100
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	1170	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

3.15. Critical Bands: ERB estimation

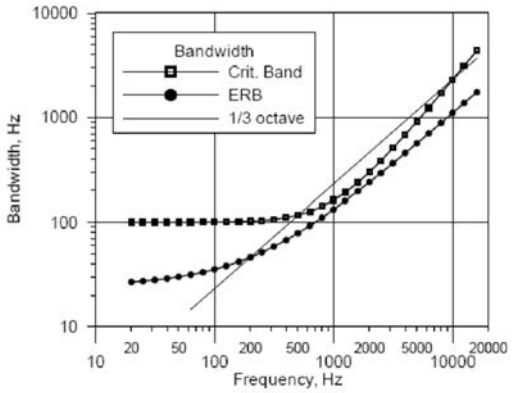
- Moore et al.
- ERB (Equivalent Rectangular Bandwidth)

$$ERB(f) = 24.7 * (4.37\, f / 1000 + 1)$$

CF (f0)	ERB	CF	ERB
27.5	31.1	880	115.5
55	33.7	1760	212.2

110	38.9	3520	434.4
220	49.4	7040	994.8
440	70.8		

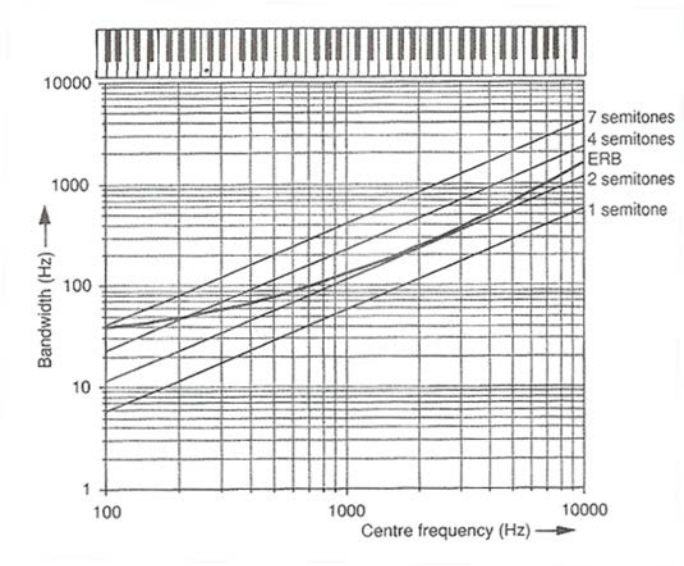
3.16. Critical Bandwidth comparison



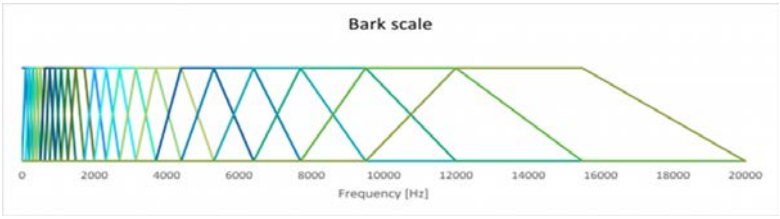
3.17. Critical Bandwidth: application

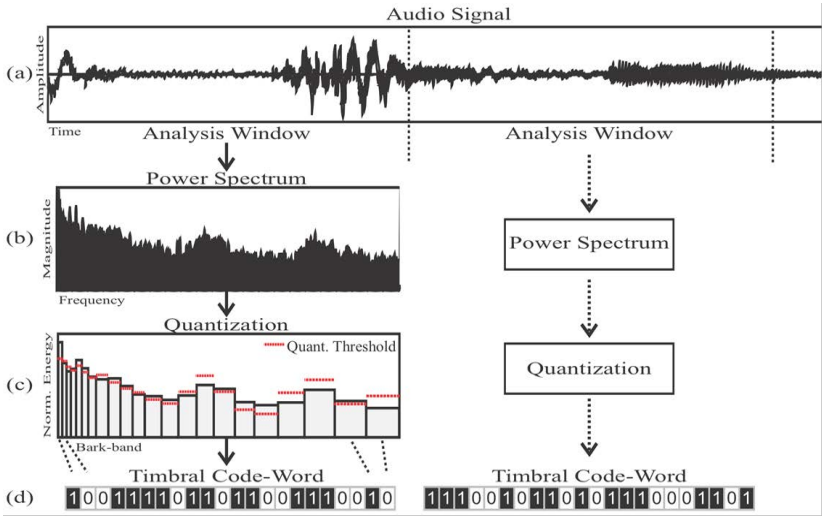
- What is the relation between the figure and third-octave equalizers?

An octave band is a frequency band that spans one octave. Equalization is the process of adjusting the volume of different frequency bands within an audio signal. The use of ⅓ octave bands is more convenient mathematically than the use of critical bands, so for analysis and equalization, ⅓ octave bands have become the defacto standard.



- How can we encode the perceived spectral shape of sounds in a “compact” way? Using 24 CBs of Bark scale.





Haro M, Serrà J, Herrera P, Corral Á (2012) Zipf's Law in Short-Time Timbral Codings of Speech, Music, and Environmental Sound Signals. PLOS ONE 7(3): e33993

3.18. Perceptual coding models: what do they eliminate?

Perceptual coding operates by analyzing the whole audio signal and **deleting all those parts of it which are deemed to prove inaudible** because of their quietness, or closeness in time or pitch to some other

louder signal component present at the same time. Perceptual models try to approximate or predict the judgment of auditory quality perceived by human listeners.
In speech and audio coding applications, practically all distortions caused by the algorithms are due to quantization of the signal. The objective of perceptual modeling is then to choose the quantization accuracy such that the perceptually degrading effect of quantization is minimized. Signal components which are more important to a human listener are quantized with a higher accuracy than those which are less important.

https://wiki.aalto.fi/display/ITSP/Perceptual+modelling+in+speech+and+audio+coding

[LAB 1: AUDIOGRAMS]

- Decibels:
    - dBHL: Hearing Level
    - dBMasking: Masking level
    - dBSPL: Sound level
    - dBSize: Size
    - dBV: Voltage
    - dBW: Electrical power
    - 0 dBHL means a normal average healthy hearing level
    - -20 dBHL means an outstanding or amazing hearing level: in fact it means that the person showing it would have a threshold 20 dB below (better hearing) than the average listener. He/she could hear sounds with 20 dB less amplification than the average listener.
- The frequencies that deteriorate earlier are the high frequencies
- The range of frequencies perceived by the human ear is 20 Hz - 20 kHz for a young and healthy ear.
- Why do the audiograms show only frequencies up to 8 kHz? Because most of the significant sounds for humans have most of their energy up to that value.
- The 2 thick black lines surrounding the letters in the audiograms indicate the range of intensities for most of the speech sounds.
- Crosses and circles in the audiograms indicate left ear and right ear, respectively.
- Why was “dog barking”, “jack hammer” or “lawnmower” included in the audiogram? To indicate their approximate range of frequencies and loudness and offer some term for comparison.

[LAB 2: MASKING]

- Asymmetry of Masking by Pulsed Tones:
    - Masking will be more effective when the mask is 1200 Hz and the signal is 2kHz than the other way around.
    - A masking tone tends to mask more effectively tones of higher frequencies than tones of lower frequencies.

    - A masking tone tends to mask more effectively tones that are coded between the oval window and the place of the mask than tones between the place for the mask and the helicotrema.
    - As the critical bandwidth for 150 is around 60 Hz whereas for 6.3 kHz it is around 700 Hz, masking should be stronger in the latter than in the former case, because both tones are separated less than a CB.
- Determination of Critical Bands by Masking:
    - The CB of 200 Hz (measured using Barks) is 100 Hz
    - The CB of 200 Hz (measured using ERB) is less than 50 Hz (around 47 Hz)
    - The CB of 400 Hz (measured using Bark) is a bit more than 100 Hz
    - The CB of 400 Hz (measured using ERB) is more than 50 Hz but less than 100 Hz (around 66 Hz)
    - The CB of 2 kHz (measured using Barks) is 300 Hz approx.
    - The CB of 2 kHz (measured using ERB) is 240 Hz approx.
    - The CB of 4 kHz (measured using Bark) is 700 Hz approx.
    - The CB of 4 kHz (measured using ERB) is between 456 and 502 Hz approx., depending on the formula.
    - The CB of 7 kHz (measured using Bark) is 1300 Hz approx.
    - The CB of 8 kHz (measured using ERB) is between 888 and 1175 Hz approx., depending on the formula.
    - The CB of 16 kHz (measured using ERB) is more than 1700 Hz

CB is 240 Hz:

- A narrow noise bandwidth (e.g. 10 Hz) would make it possible to hear more tones than a 240 noise bandwidth.
- As the bandwidth of the noise increases from 10 Hz to 240, the amount of tones that can be heard decreases.
- If we compare the 500 Hz and the 1000 Hz bandwidth cases, there will be no difference in the number of tones perceived.

- Critical Bands by Loudness Comparison:
    - When the bandwidth is greater than a critical band, the subjective loudness increases because: many more independent neurons are spiking and the energy is passing through more than one auditory filter.
- Backward and Forward Masking
    - Forward masking refers to the masking of a tone by a sound that ends a short time (up to about 20 or 30 ms) before the tone begins. This effect suggests that recently stimulated sensors are not as sensitive as fully-rested ones.
    - Backward masking refers to the masking of a tone by a sound that begins after the tone has ended (up to 10 ms later, but the amount of masking decreases as the time interval increases). This effect apparently occurs at **higher centers of processing in the nervous system** where the neural correlates of the later-occurring stimulus of greater intensity overtake and interfere with those of the weker, earlier stimulus.
    - Forward masking will be more effective than backward masking when the gap signal-mask is long (100 ms).

- When the masker is presented at the same time than the signal (0 ms delay) we will hear the signal fewer times than when t > 0 irrespectively of the forward or backwards masking condition.

4. Psychophysics of basic sound dimensions: Loudness

4.1. Psychophysics

The following scientists gave shape to the field of psychophysics:

- Hermann von **Helmholtz** (1821-1894)
- Ernst Heinrich **Weber** (1795-1878)
- Gustav Theodor **Fechner** (1801-1887)

4.2. Physical and perceptual features of sounds

There are basic relationships between physical and perceptual features of sounds. Each feature has a specific role.

- Waveform amplitude → loudness (the larger, the louder)
- Waveform period → pitch (the longer, the lower)
- Waveform shape → timbre (the more "rippled" i.e. far from sinusoidal,the richer)

4.3. Psychophysical Laws

Psychophysical laws are **mathematical expressions relating a physical property with a perceptual sensation**. The researchers elaborated these laws based on experimental procedures: they collected subjects and once they gathered enough data, they tried to come up with a function of the best fit.

- Response = f (sensory stimulation)
- Is f linear, potential, exponential?
- Are sensations totally independent?
- Collect "subjective" judgements when presenting different intensities, pitches, along a single dimension.
- Find the best fit between the physical magnitudes and the perceptual estimations

Important psychophysical laws:

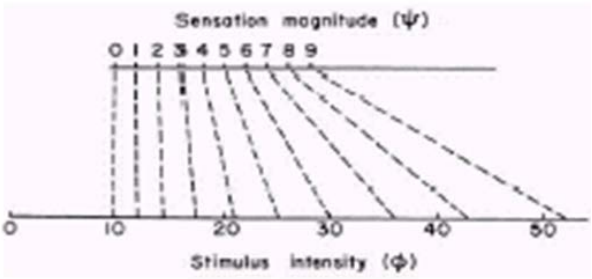
- **(1831) Weber's Law**: it is not very faithful, it was the first attempt to find the just-noticeable change. It states that **the just-detectable change in stimulus intensity (JND or DL) is proportional to the intensity**:

$$\Delta l \,/\, l \;=\; k$$

- **(1860) Fechner's Law:** he proposed a logarithmic relationship between physical property and sensation **(stimulation and sensation are not linearly linked**):

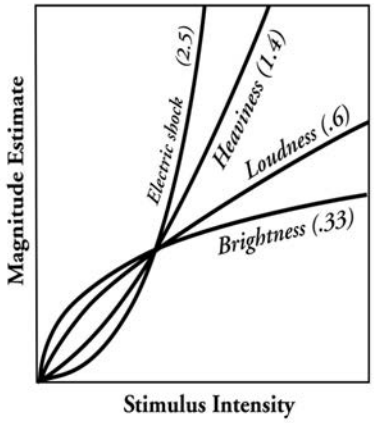
  $$R \;=\; k\, log(l)$$

  R is the sensation, l is the physical property, k is a cte. to be found or adjusted from data.



- **(1957) Stevens' Power Law**: he studied different sensations, and for each one, a power law could fit quite well. It is a better approximation (for today):

  $$R \;=\; k\, l^{p}$$

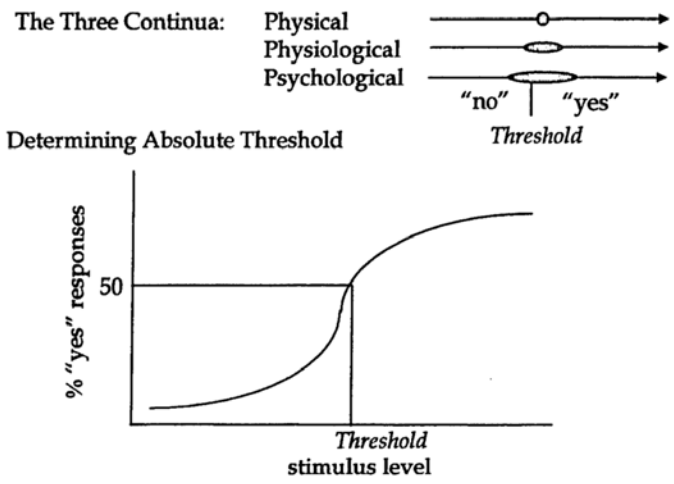


4.4. Absolute thresholds

Can you hear a sound? Can you hear a pitch? Can you feel pain? Can you see a light? In the case of loudness, **an absolute threshold refers to the first sound pressure level which can be heard by a subject**. In other words, the change of pressure has to reach a certain amount of change for us to notice it.

- **Absolute threshold**: **minimum amount of stimulation** needed to feel a given sensation.

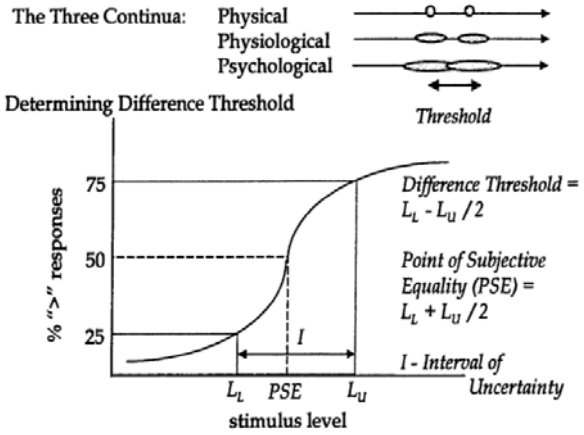
However, subjectivity creates an area of variation, it can be seen when we compare various responses of different individuals, so a **statistical** approach is needed. For example, if more than 50% answer X when… It is likely that some subjects cannot hear it but on average we can say that most will hear something.



4.5. Differential thresholds

- **Differential threshold: minimum difference** that can be perceived with relation to a given sensation. How much one sound should change compared to another to be considered different (with respect to some sensation). It has many names: "Just noticeable difference" (JND), "Differential threshold" or "Differential Limen" (DL).

Criteria is much more conservative: the first difference or jump that is noticed **by more than 75 %** of the listeners if the JND for that sensation.



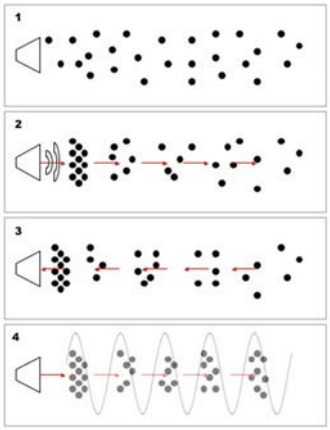
4.6. Loudness

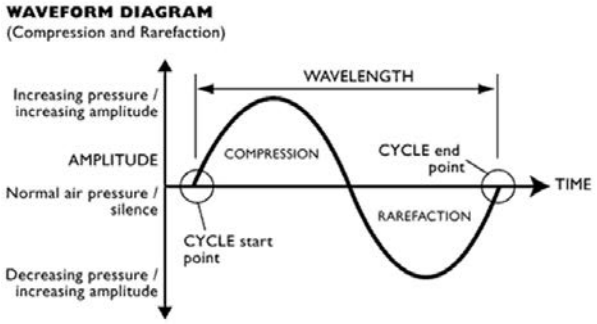
Loudness is the **subjective sensation** generated by the **intensity of the air pressure**.

Amplitude (physical) → Loudness (sensation)

4.6.1. Compression and rarefaction

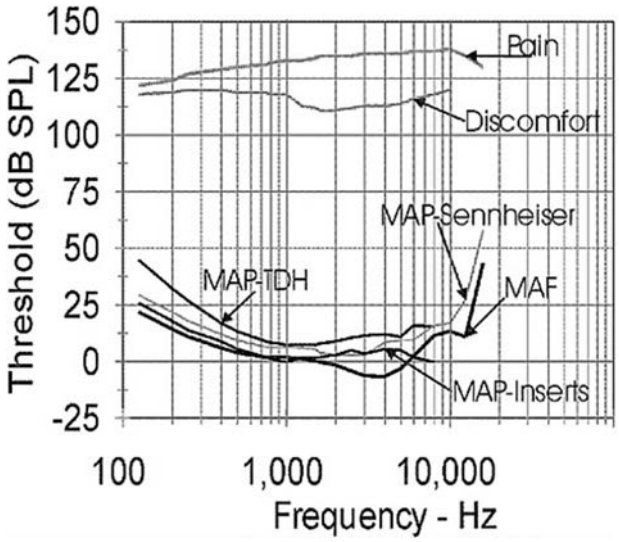
- **Compression raises air pressure, which raises the loudness sensation**.
- Rarefaction is the reduction of density, the opposite of compression, also raises loudness sensation (second figure below).





4.6.2. Absolute thresholds

Different measurement conditions yield slightly different curves



4.6.3. The dB SPL

dB SPL is the unit preferred to measure the **Sound Pressure Level**. Mathematical transformations:

$$dB_{SPL} = 20\ log(P/P_0)$$

- It usually ranges from 0 to 130, and uses the minimum audible pressure for 1 kHz as reference value (P_0)
- What does 0 dB SPL mean? No pressure? Means that the pressure is the same as the reference value. And negative dB SPL? Below the average hearing threshold (below the reference value).

- dBs are not additive: first you should convert dB SPL to Pressure, and then sum them and convert again to dB:

$$P = P_0 * 10^{(dB/20)}$$

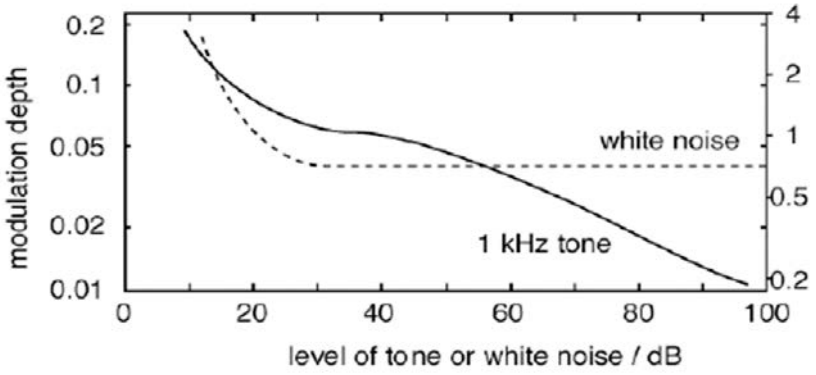
- An increase of 20 dB → sound pressure * 10

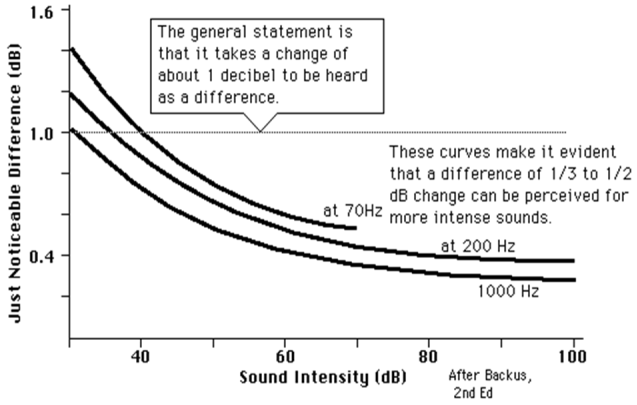
Table of sound levels L (loudness) and corresponding sound pressure and sound intensity			
Sound Sources Examples with distance	Sound Pressure Level L_p dBSPL	Sound Pressure p N/m^2 = Pa	Sound Intensity I W/m^2
Jet aircraft, 50 m away	140	200	100
Threshold of pain	130	63.2	10
Threshold of discomfort	120	20	1
Chainsaw, 1 m distance	110	6.3	0.1
Disco, 1 m from speaker	100	2	0.01
Diesel truck, 10 m away	90	0.63	0.001
Kerbside of busy road, 5 m	80	0.2	0.0001
Vacuum cleaner, distance 1 m	70	0.063	0.00001
Conversational speech, 1 m	60	0.02	0.000001
Average home	50	0.0063	0.0000001
Quiet library	40	0.002	0.00000001
Quiet bedroom at night	30	0.00063	0.000000001
Background in TV studio	20	0.0002	0.0000000001
Rustling leaves in the distance	10	0.000063	0.00000000001
Threshold of hearing	0	0.00002	0.000000000001

- Avoid recurrent events generating **>90 dB SPL** (avoid physical damage of hair cells)

4.6.4. Differential thresholds

- **1 dB of change in loudness is the smallest difference perceptible by normal human hearing.** Less than 1 dB is very difficult (you have to be trained).
- **A change of 3 dB** is accepted as the **smallest difference** in level that is **easily heard** by most listeners (double power in W).
- **A change of 6 dB** is accepted as a significant difference in level for any listener (power by four).
- **A change of 10 dB** is accepted as the difference in level that is perceived by most listeners as **"twice as loud"** or **"half as loud"** (power by 10).

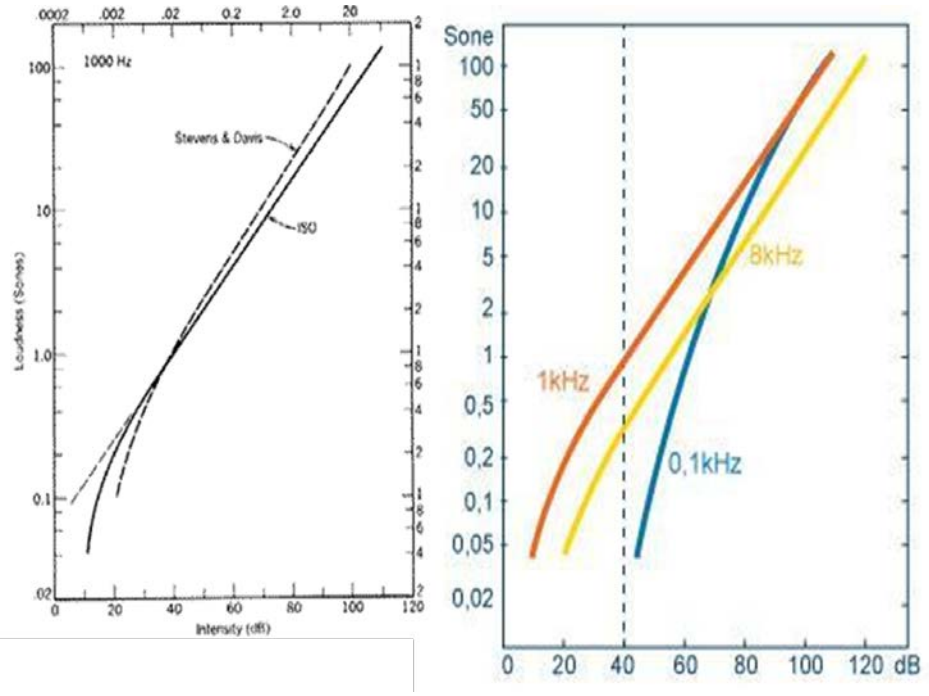




4.6.5. Sones

Sones is a **subjective scale for loudness**, it has been accepted by ISO. Ask people how much louder a sound is than another one. (Subjectively):

- 1 Sone = loudness sensation for 1 kHz @ 40 dB SPL = 40 Phone (see next section)
- **A 10 dB increase in level gives a doubling in loudness (twice the loudness sensation).**
- Which is the loudness of 1kHz @ 60 dB SPL? 4 sones.



4.6.6. Phones

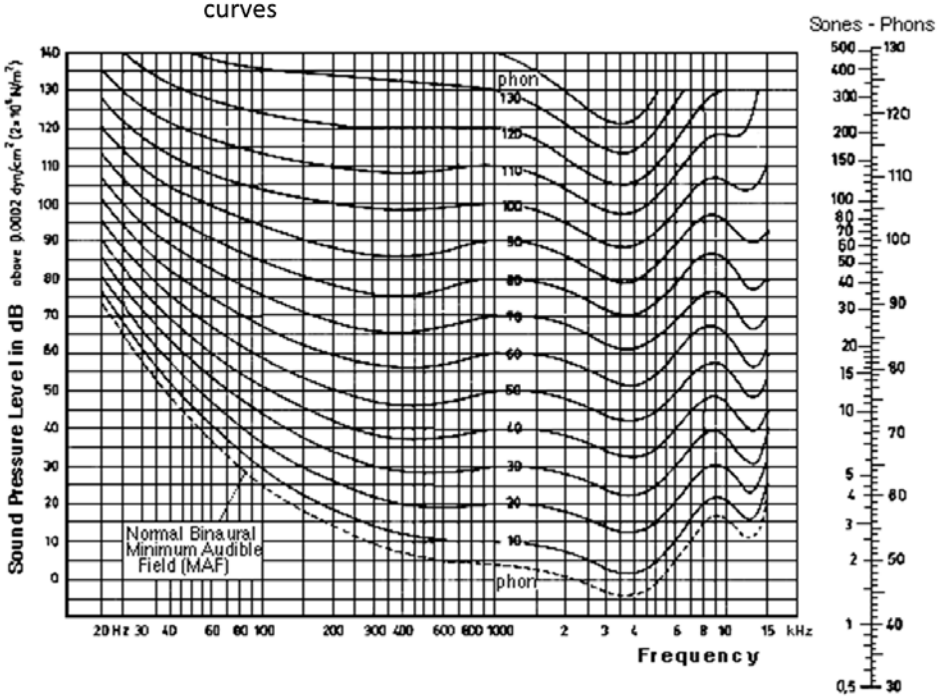
Fletcher and Munson did studies motivated by the telephone, posing the question: tones of different frequencies at the same dB SPL level, are they perceived equally loud? This led to the **equal loudness curves** (also known as **isophonic**). Our hearing system is not linear. **High pressure curves are a bit flatter than the low pressure curves**.

How much do we need to amplify/attenuate a tone to get “equal loudness” between both?

- Tones with equal loudness have the same phone level.

- The “Loudness” function in hi-fi amplifiers and in music player presets: “Loudness” EQ **enhances low frequencies for a better balance**, **especially for low sound pressures**.

Equal loudness, isophonic, or *Fletcher & Munson* curves

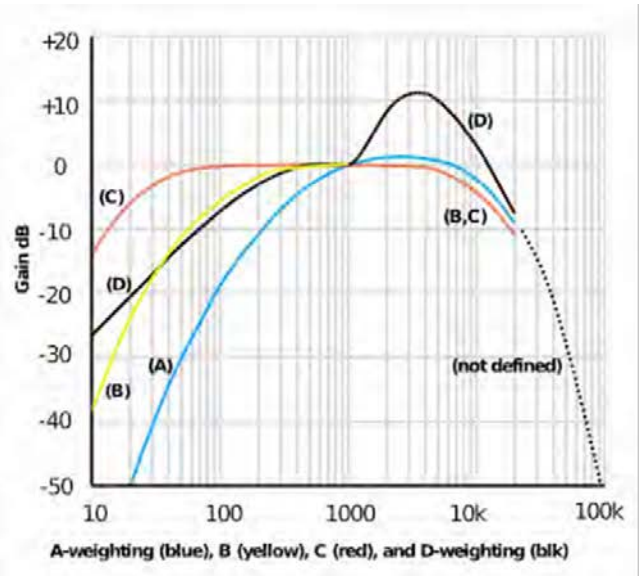


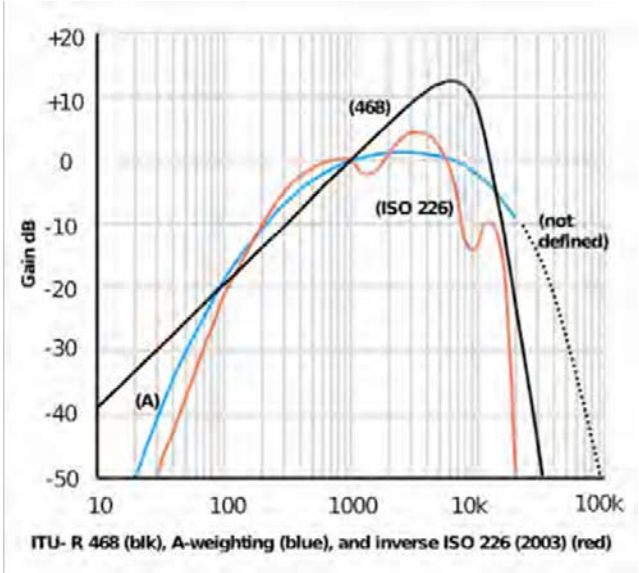
4.6.7. Loudness weighting scales

A machine just produces certain dB SPL, but this does not take into account the human sensations. In order to approximate human sensations (certain applications need it), **loudness weighting scales** were created:

- dB(A) = 40 phon curve (approx) (SOFT SOUNDS): in an effort to account for the relative loudness perceived by the human ear, as the ear is less sensitive to low audio frequencies. Originally defined for the measurement of low-level sounds (around 40 phon), is now commonly used for the measurement of environmental noise and industrial noise. However, if applied to noisy situations, it is not so good (it makes a wrong weighting), but it is very much used everywhere.
- dB(B) = 70 phon curve
- dB(C) = essentially flat, for high pressure levels

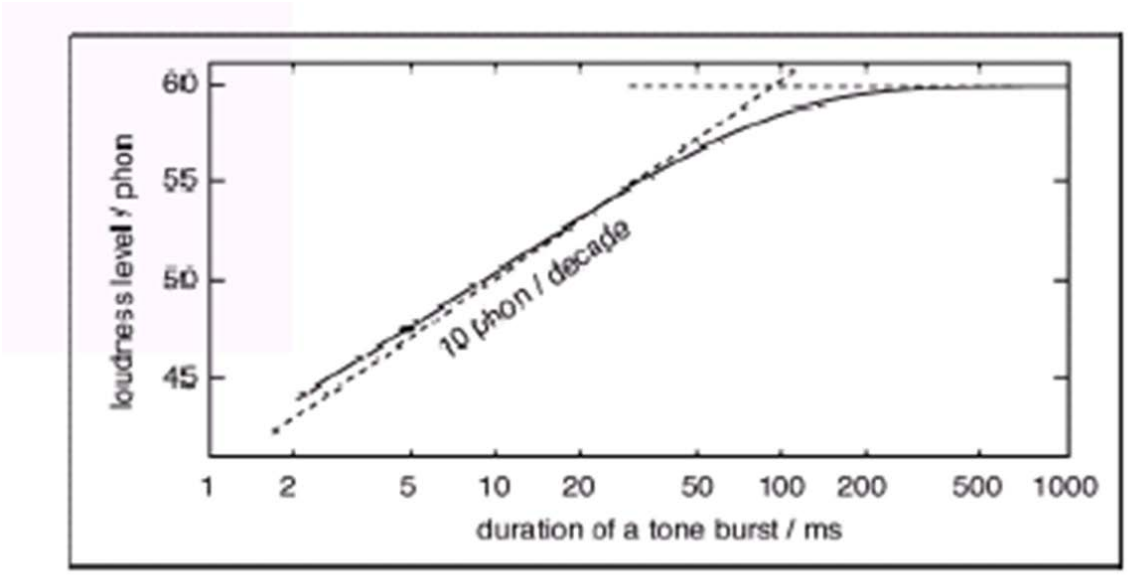
Low frequencies have to be weighted in a negative way (weight down).





4.6.8. Loudness and duration

- Does duration affect our perception of loudness? Yes, if we go to the extremes.
- Energy integration time < 150 ms
- 100 ms for the loudness perception to start working, if < 100 ms then the integrator cannot work properly (tricky sensations).
- The longer the tone, the louder, up to 150 ms.

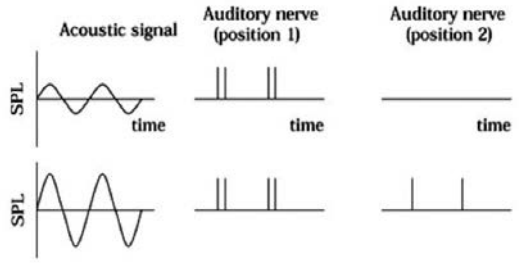


4.6.9. Neural coding of intensity

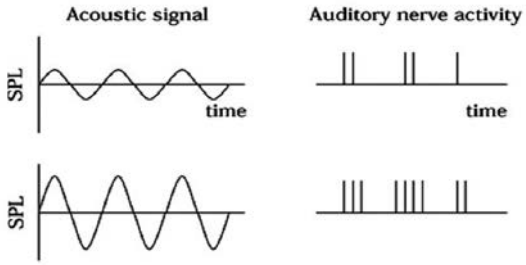
There is a limitation of speed of neurons sending spikes. **If there are many neurons sending information, then the sound will be perceived as louder**. **Also, if the spikes are generated in many places, the sound will be perceived as louder.**

- **Firing rate works for low intensities**
- **Number of neurons works for mid-loud intensities.**

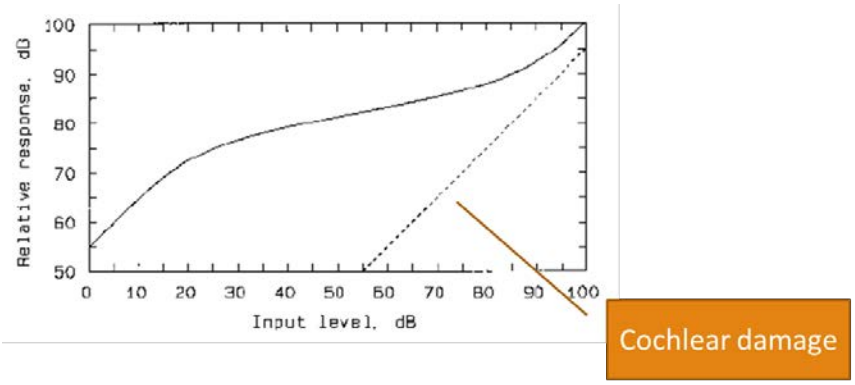
Number of neurons hypothesis

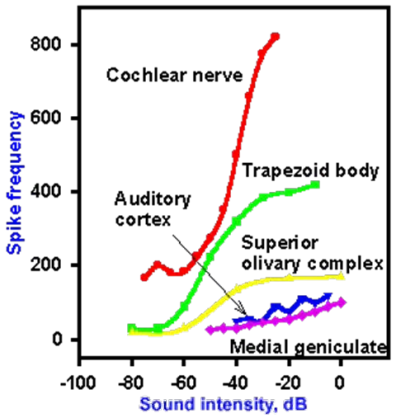


Firing rate hypothesis



- Compression caused by the basilar membrane.
- In the first stages of transmission there is high spike frequency, the intensity is mostly encoded there. In fact, this is one of the earliest sensations encoded (**Intensity is mostly encoded in the early transmission through the auditory pathways**)
- Within the critical region (critical band) sounds of equal total energy have equal loudness → firing rate of each neuron and the number of neurons trade-off against one another perfectly. As soon as the sounds are spread out over a larger frequency range, the sound with the larger bandwidth sounds louder.

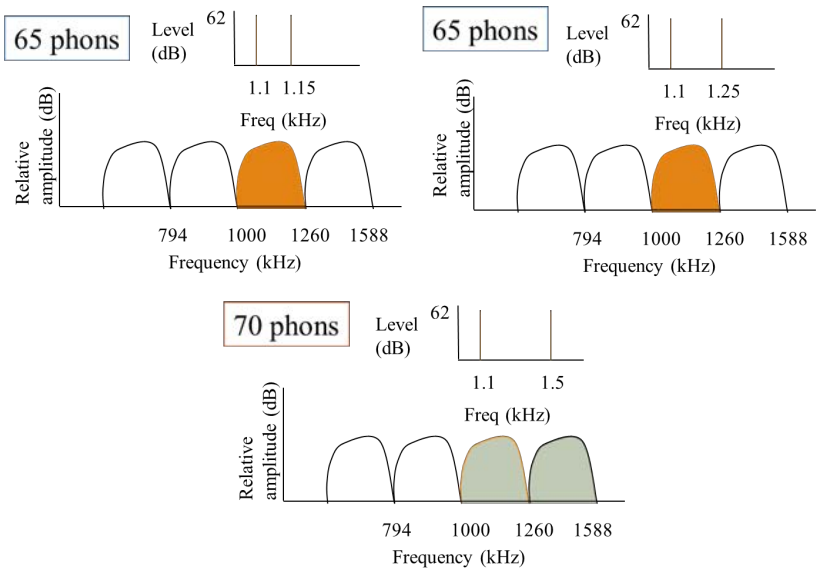




4.6.10. Loudness of complex sounds

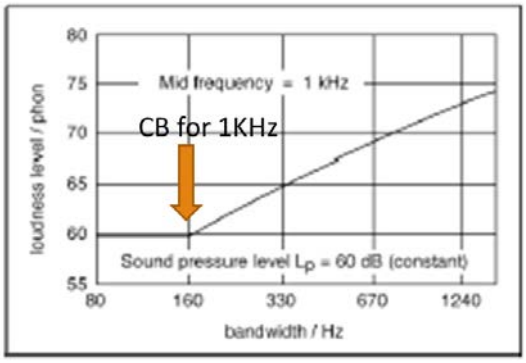
Examples of complex sounds can be sounds with many harmonics or frequency components. Here, Critical Bandwidth comes into play.
e.g. 500, 520, 540 Hz share the same CB, but 500, 720 and 940 Hz go through different filters.

- Does loudness increase by adding energy at any place in the spectrum? **We must consider the frequency resolution of our hearing system. If the energy is placed outside the critical band, loudness will increase.**
- For a complex sound of fixed intensity and bandwidth W, if W is less than the critical bandwidth, then loudness is independent of W. If W is increased beyond the critical bandwidth, the loudness begins to increase.



Processing complex sounds:

1. **Peripheral processing** (filtering according to the outer and middle ear specificities)
2. **Computation of the excitation pattern considering the masking effects** (cochlea + neural firing approximation)
3. **Conversion of the excitation pattern into band-specific loudness computation**
4. **Summation of the specific band-loudness into the final loudness value**.



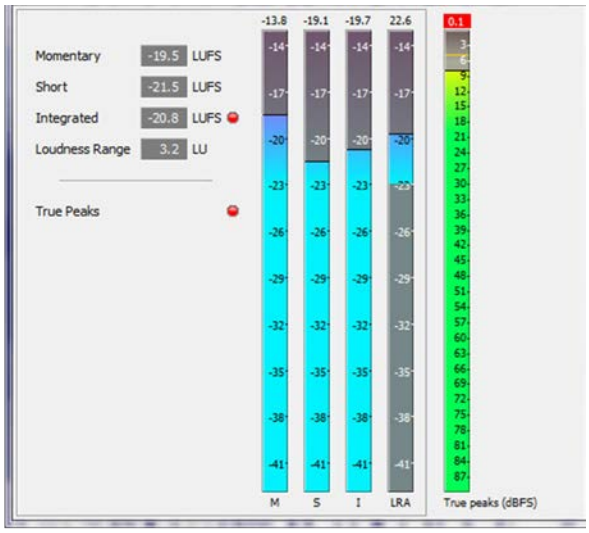
Loudness increases (additively) only when there is energy beyond the critical band

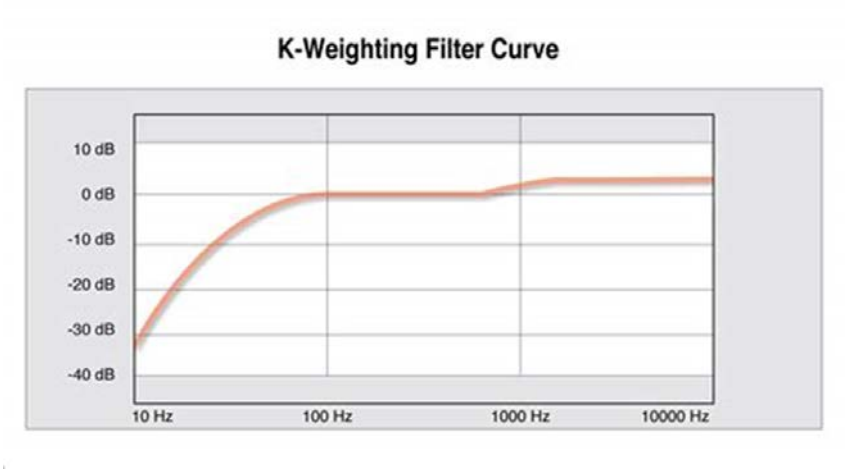
4.6.11. Approximating loudness: LUFS (EBU-128)

Loudness Units relative to Full Scale or Loudness Units Full Scale (LUFS). Standardized measure of audio sonority **taking into account human perception and intensity of electric signal**. There are two principal ways of measuring sonority: **peaks** and **RMS**. Peaks measures the highest level of the audio (songs with higher levels sound louder than the rest), while RMS measures the mean level (can cause distortion).

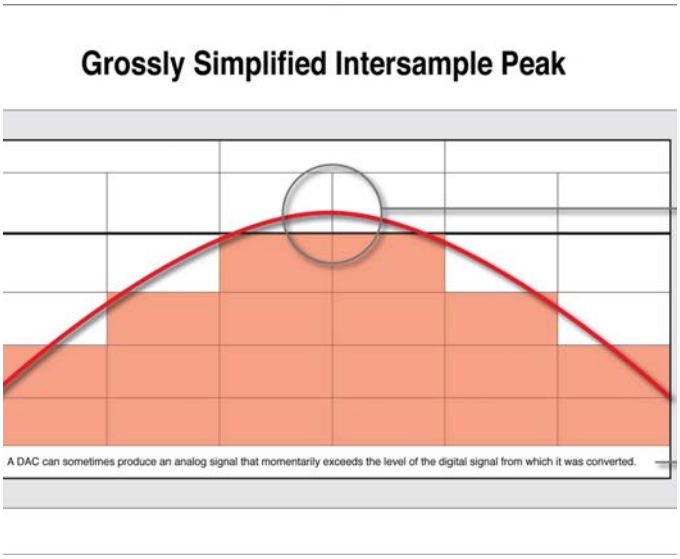
- LUFS: better option than dBFS (Peaks) or RMS values.

- Weighting curve roughly based on our hearing. **Negative value**, the further from 0 the quieter the sound.
- In music production, you have to learn how to use it, because it is the most close to human perception. All music distributors use LUFS (e.g. Spotify).



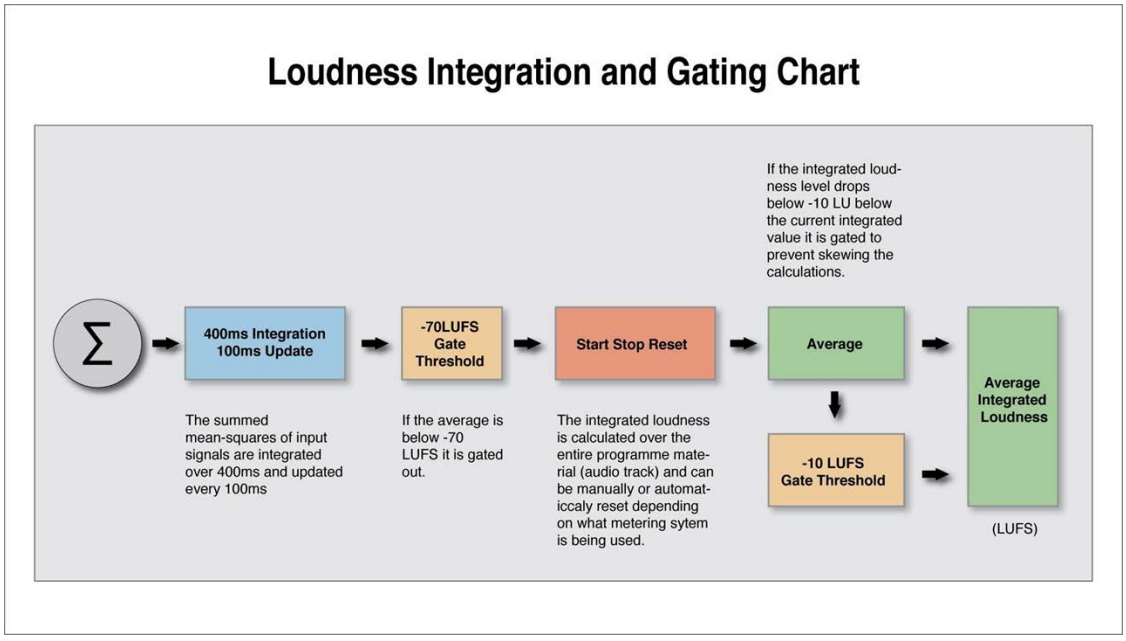


True Peak: oversamples the data to interpolate if the reconstructed analog waveform could go beyond 0 dBFS.



LUFS:

- **Momentary loudness**: K-weighted RMS in 400 ms
- **Short-term loudness**: integrated loudness in 3 seconds' windows
- **Integrated loudness**: average loudness for a whole track
- **Loudness range (LRA)**: statistical distribution of short-term loudness within a track



Targeted values in streaming platforms:

Service	Codecs Streamed	Normalization enabled by default	Normalization type	Boost Y/N	Loudness Target	Peak Headroom Target	Is limiting used?	Bitrates/Resolutions
Apple Music	AAC	No	Both	Yes	-16LUFS	-1.0dBTP	No	256kbps (VBR)
Spotify	AAC & Ogg Vorbis	Yes	Both	Yes	-14LUFS	-1.0dBTP	Yes	160kpbs (free) 320kbps (premium)
Tidal	AAC & FLAC	Yes	Album	No	-14LUFS	-1.0dBTP	No	320kHz 44.1kHz/16bit 96kHz/24bit
Amazon Music	MP3	Yes	Track	No	-14LUFS	-2.0dBTP	No	320kbps
YouTube	AAC	Mandatory	Track	No	-14LUFS	-1.0dBTP	No	256kbps
Pandora	AAC+ (low & standard) MP3 (high)	Mandatory	Track	Yes	-14LUFS	-1.0dBTP*	No	64kbps (free) 192kbps (subscription)
Deezer	MP3 & FLAC (HiFi)	Mandatory	Track	No	-15LUFS	-1.0dBTP	No	128kbps (free) 320kbps (premium) 44.1kHz/16bit
Qobuz	FLAC	No	N/A	N/A	N/A	N/A	N/A	Up to 192kHz/24bit
SoundCloud	Opus/Ogg Vorbis (free & go) AAC (go+)	Yes	Track*	Yes	-14LUFS	-1.0dBTP	Yes	64kbps (free & go) 256kbps (go+)
Facebook	AAC							128kbps
Instagram	AAC							128kbps

[LAB 4: LOUDNESS]

- Decibels:
  - The dBSPL is a measure of sound pressure level
  - dB scales are useful when measured values span across a wide range: the log helps to make the values to be more constrained.
  - A decibel is a relative measurement that always works by comparison to a reference measure.
  - dBSPL values can be negative when the sound pressure level is below the hearing threshold.
  - A 1dB decrease can be barely perceivable when loudspeakers are used.
  - In order to get a clear “half loudness” sensation, 10 dBSPL jumps should be used.
  - A 1dB decrease is close to the JND for loudness for mid and high-frequency tones or broadband noises.
  - When intensity decreases by 6 dBSPL, loudness decreases a bit less than a half.
- Intensity, loudness and distance
  - The sound pressure level decreases about 6 dB each time the distance to the source is doubled (in a normal room this will not be the case, because of reflections from walls, ceiling, floor and objects in the room).
  - Everytime we double the distance it decreases about less than a half
- Loudness scaling
  - On the sone scale, the loudness in sones S is proportional to sound pressure p raised to the 0.6 power: S = Cp^0.6, where C depends on the frequency.
  - Loudness doubles for about a 10 dB increase in sound pressure level → but it is not true for every frequency. For low frequencies the curves are closer than for 1 kHz. This means that you achieve double or half loudness sensations with less than 10 dB changes.

    - Some investigators have found that the exponent varies with tone frequency, increasing at low frequency and low level to approach a value of 1.0 → an exponent of 1.0 would mean that loudness doubled for a 6 dB increase in sound pressure level instead of the usual 10 dB.
- Equal loudness curves (Frequency response of the ear)
    - The sensitivity of the ear varies with the frequency and the quality of the sound.
    - In their famous experiments of 1933, Fletcher and Munson determined curves of equal loudness for pure tones, demonstrating the relative insensitivity of the ear to sounds of low frequency at moderate to low intensity levels.
    - Hearing sensitivity reaches a maximum around 4000 Hz, which is near the first resonance frequency of the outer ear canal, and again peaks around 12 kHz, the frequency of the second resonance.
- Temporal integration
    - Numerous experiments have pretty well established that the "ear" averages sound energy over 200 ms, so loudness grows with duration up to this value, loudness level increasing by 10 dB when the duration is increased by a factor of 10.
    - The loudness level of broadband noise seems to depend somewhat more strongly on stimulus duration than the loudness level of pure tones.
    - Loudness of sounds shorter than 150 ms is reduced because the brain needs more time to integrate the energy.
    - Loudness progressively decreases only for very short durations.

5. Psychophysics of basic sound dimensions: Pitch

5.1. Pitch

- ANSI (1973): "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high."
- Plack (2005): "The aspect of auditory sensation whose variation is associated with musical melodies."

Pitch is a subjective sensation. Our brain builds it in response to some properties (e.g more prominent peaks in the spectrum). From complex sounds it is difficult.

- **F0 frequency** -> pitch
- **periodicity** -> pitch

Pitch helps us to compare melodies. So useful depending on the culture.

5.2. Pitch from a noisy source?

It seems that **periodicity is very important for the pitch sensation**.

5.3. Pitch: different waves, different spectra

Different waves, different spectra but the same pitch sensation. Even without a fundamental in the spectrum. The envelope is the same though.

5.4. Pitch ascending endlessly

Shephard tone: sound consisting of a superposition of sine waves separated by octaves. It creates the auditory illusion of a tone that seems to continually ascend or descend in pitch, yet which ultimately gets no higher or lower.

5.5. Frequency perception: absolute thresholds

- 20Hz - 20Khz (in young, intact hearing systems, 16Khz in adults)
- HF threshold permanent decrease due to:
    - Ototoxic substances: aspirin, quinine
    - Exposure to high levels of sound pressure (>85 dBSPL) for long periods of time.

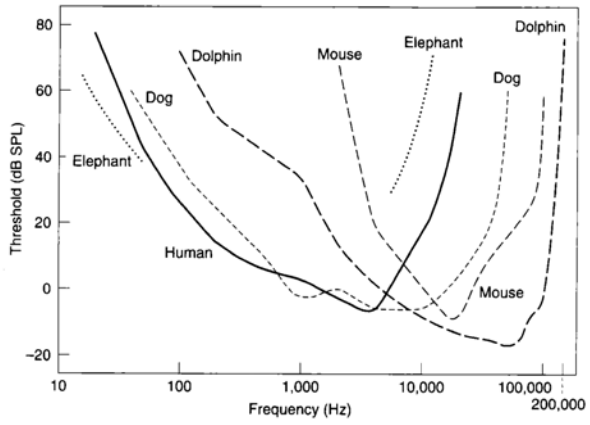
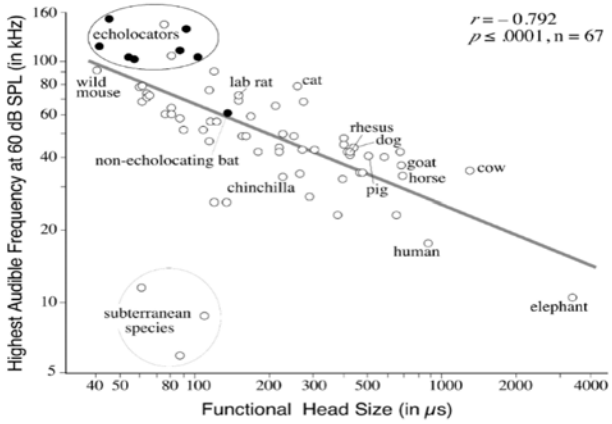


Figure 10.10
Audibility curves for a few animals. Notice that the frequency scale is logarithmic, so high frequencies are allotted less distance than low frequencies. Only the low and high ends of the elephant curve are shown to prevent overlap with the other curves. (Based on data from Au, 1993; Heffner, 1983; Heffner & Heffner, 1980, 1985; Heffner & Masterton, 1980.)

5.6. Head size (cochlear volume) and threshold for frequency

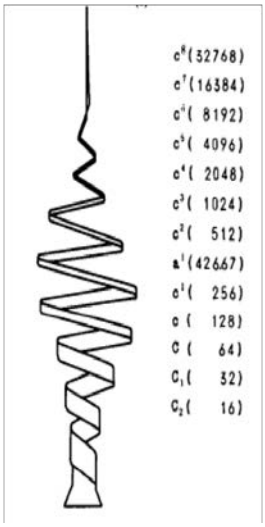
- The bigger the head size, the lower the highest audible frequency.

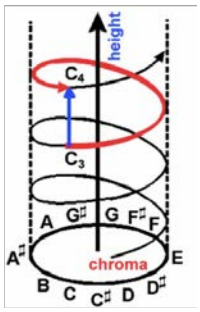


5.7. Dimensions of the pitch sensation: chroma and height

Chroma: assuming the equal-tempered scale, one considers 12 chroma values represented by the set: {C, C#, D, D#, E, F, F#, G, G#, A, A#, B} Tones are judged to be more similar when they share "chroma" (just up to 5 Khz)
Tones are ordered in a spatial way (height) from low to high.

Shepard's helix model: pitch height and chroma:





5.8. Dimensions of the pitch sensation: strength

Pitch strength is determined by:

- The **microstructure** of the waveform (the simpler, the stronger pitch)
- The **amplitude envelope** (random phase -> weak pitch)

5.9. Pitch sensation thresholds

Region of pitch: 30 - 5000Hz
Highest piccolo note: 4500Hz

- Is there any chance for high-frequency music?
    - Not much harmonics up to 5KHz. Not really rich sounds.

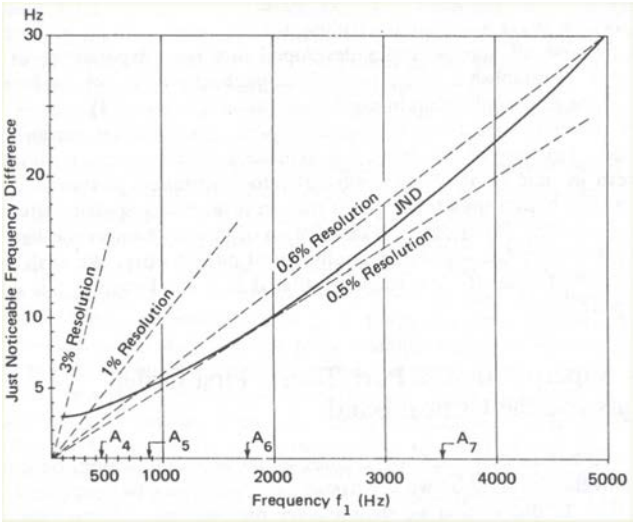
5.10. Pitch differential threshold (JND)

$$Cents = 1200 * log2\,(f2/f1)$$

- High frequency tones: 0.5% of the base frequency; **<10 cents**
- Low frequency tones: better than 3% and around 2 or 3 Hz; **around 40 cents**
- Mid frequency tones: **around 4 cents**

Why not using more "notes" in music? (as we can hear > 100 "notes" between A440 and A880)

Discrimination is best for frequencies near 2 kHz and degrades rapidly above 4 kHz. Discrimination is better for longer durations.



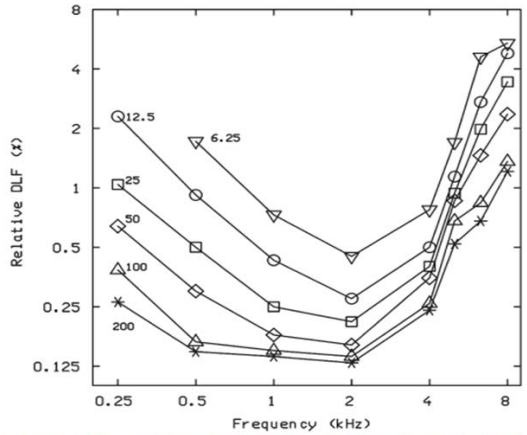


Fig. 4.2 Frequency difference limens (smallest detectable relative frequency difference) for pure tones. Each curve is for a different stimulus duration (in ms). Discrimination is best for frequencies near 2 kHz and degrades rapidly above 4 kHz. Discrimination is better for longer durations. From Moore (1973) with permission.

5.11. Perceivable “Quanta”

How many different “sounds” can we perceive (considering loudness JND and pitch JND). We can consider that **our perception is quantized**. Using the threshold for audibility and pain, and looking at the different JND for loudness and pitch, we can approximate how many different sounds we can perceive. This can be useful for **audio coding**.

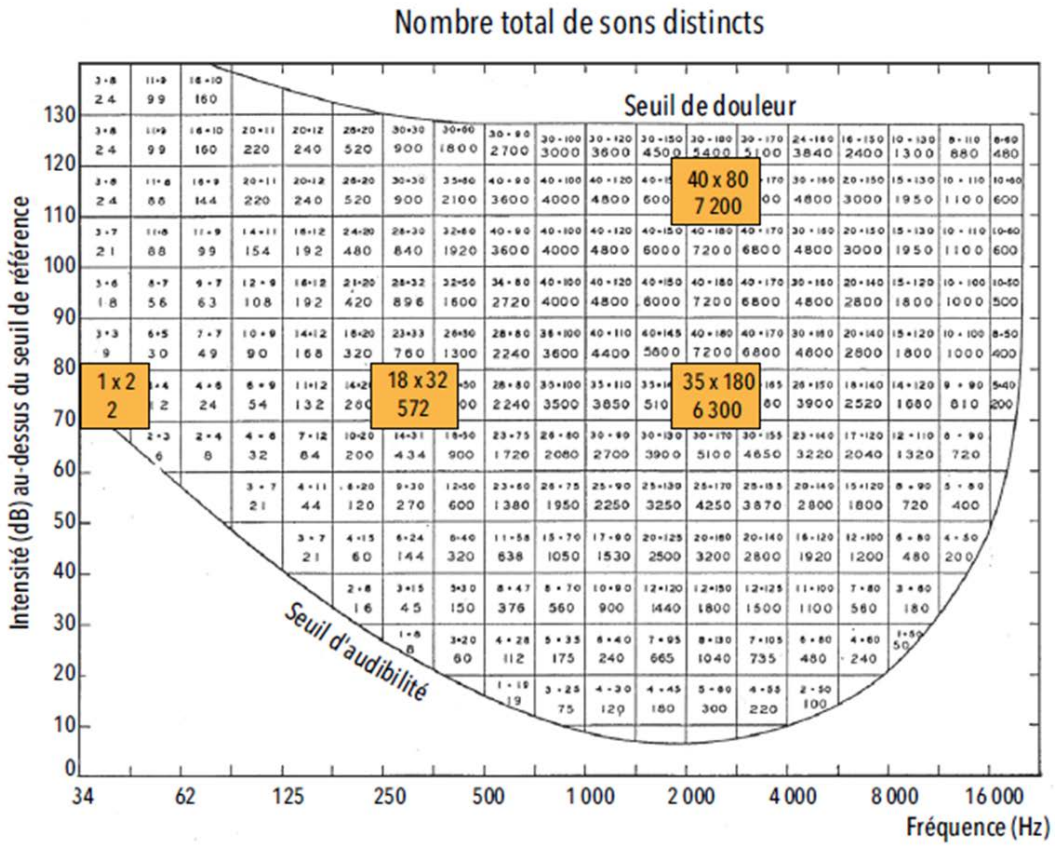
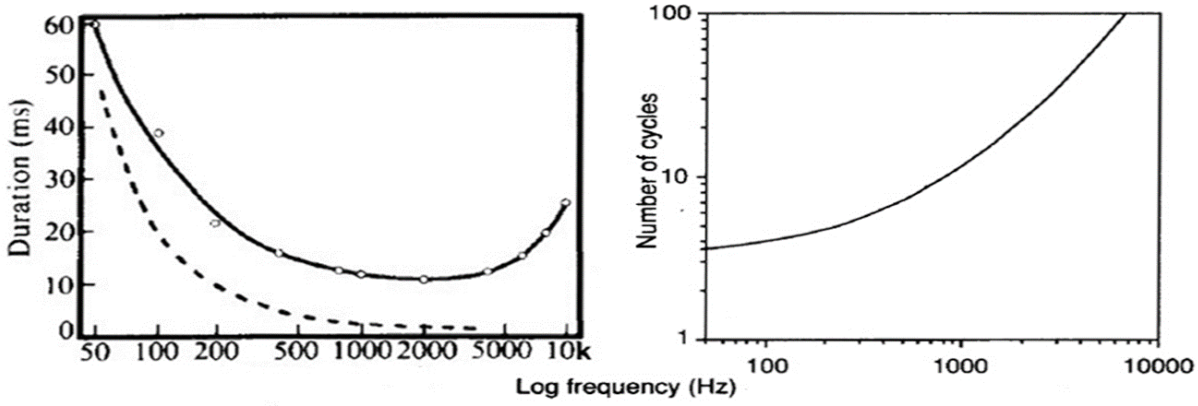


Figure 3.24 Calcul des « quantas » acoustiques sur la base des seuils différentiels.

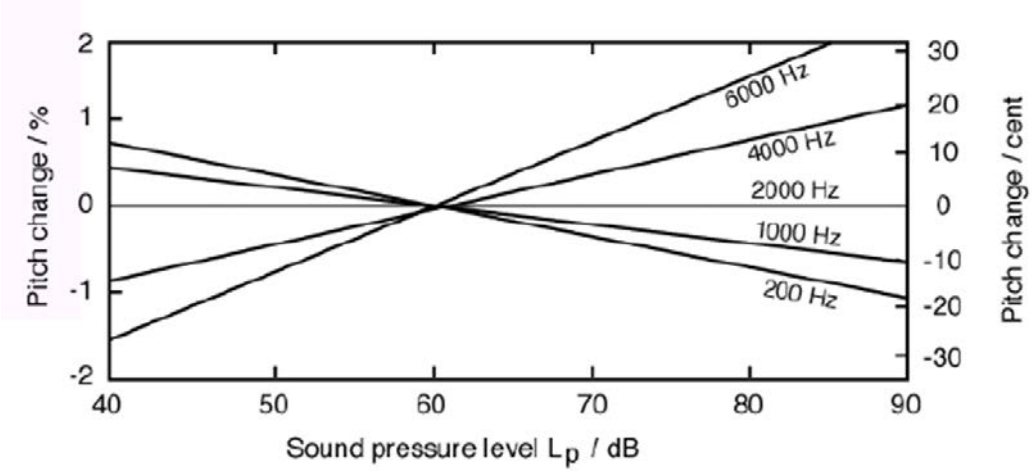
D'après Stevens, S., & Davis, H., 1938, p. 153.

5.12. Duration effects on pitch

Minimum cycles or **minimum duration**? Both.



5.13. Intensity effects on pitch



5.14. Subjective scale for pitch: Mel

How far in frequency do we have to be in order to feel a tone as doubled in pitch?

$$m = 1127\, ln(1 + f/700)$$

8 kHz = 2840 mel

500 Hz = 607 mel

$$f = 700\,[exp(m/1127) - 1]$$

Mel-scaling is used in signal processing to build filters that approximate our frequency resolution (MFCC).

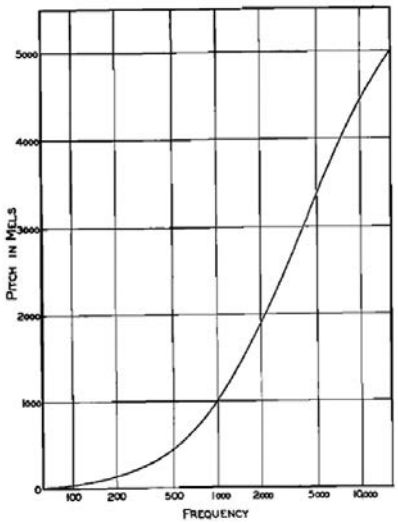
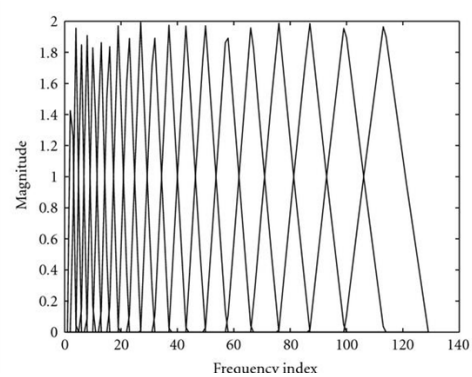


FIG. 2. The pitch function, showing the relation of perceived pitch (in mels) to the frequency of the stimulus. The values in mels are derived from the curve of Fig. 1; the 1000-cycle tone is arbitrarily assigned the value of 1000

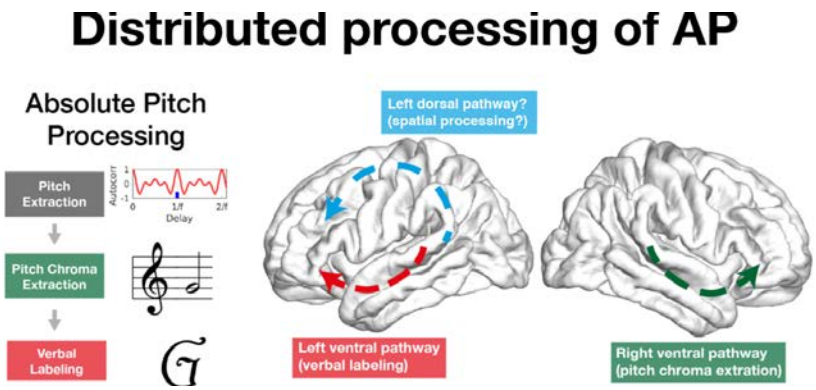
Mel filters: to characterize the spectral content of a sound.



5.15. Absolute pitch

Absolute pitch is the ability to name a sounded pitch or to produce a pitch in response to a note name. Possessors of AP name pitches rapidly and without effort or conscious strategy. It is contrasted with relative pitch (RP), the ability to identify notes relative to sounded reference pitch(es).

- Pitch memory is equal for possessors and non-possessors of AP for delays up to one minute, but **only AP possessors perform above chance for longer delays**.
- Various findings implicate **verbal codes**.
- AP can be learned most easily during a limited period of development, possibly comparable to a critical period for language learning.
- AP involves **several neurally separate subprocesses** (pitch perception, classification, labeling, storage in long-term memory, retrieval from memory).



5.16. Mechanisms for the pitch sensation of pure tones

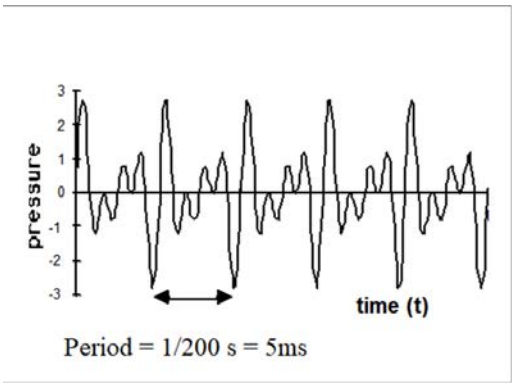
- **Place theories**: place of maximum in basilar membrane excitation (**excitation pattern**): which fibers are excited?
- **Timing theories**: temporal pattern of firing: how are the fibers firing? needs phase-locking.

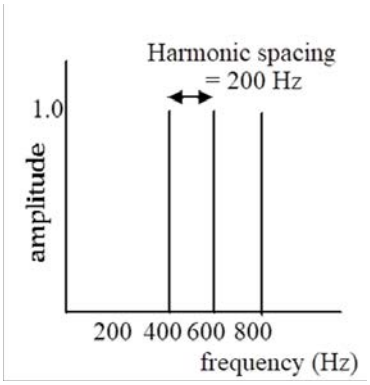
5.17. Place theory and the missing fundamental

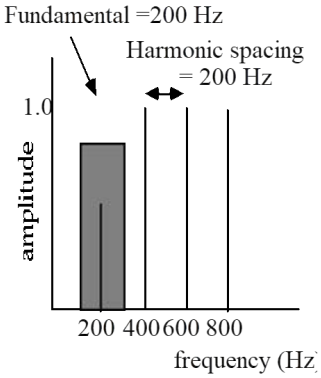
Place theory states that high frequency sounds selectively vibrate the basilar membrane of the inner ear near the entrance port (the oval window). Lower frequencies travel further along the membrane before causing appreciable excitation of the membrane. The basic pitch-determining mechanism is

based on the **location** along the membrane where the hair cells are stimulated. The place theory is the first step toward an understanding of pitch perception. Considering the extreme pitch sensitivity of the human ear, it is thought that there must be some auditory “sharpening” mechanism to enhance the pitch resolution.

The pitch of a tone having partials at 300, 400, 500 and 600 Hz will be 100 Hz, even though there is no energy present there.
Neurons are activated at the same time. **For high frequencies we need to include place theory.**
How do we properly track the bass F0 from small radio tuners, small headphones and small loudspeakers?

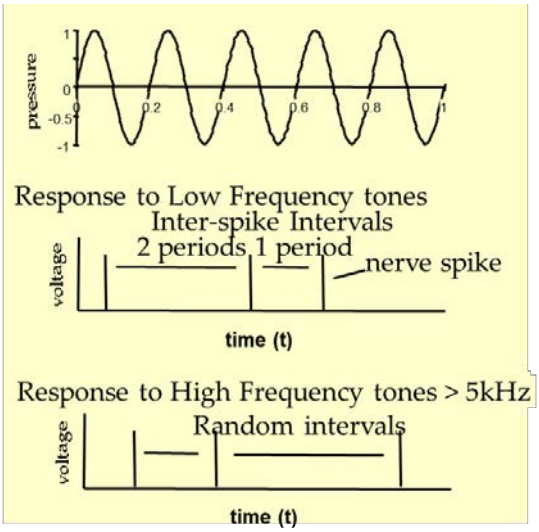


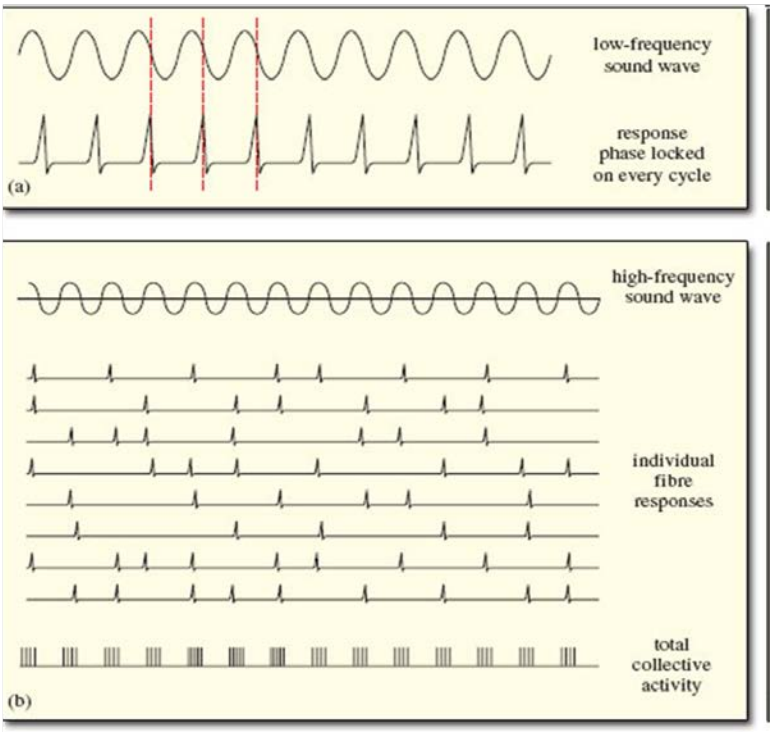




5.18. Timing theory: Phase-locking

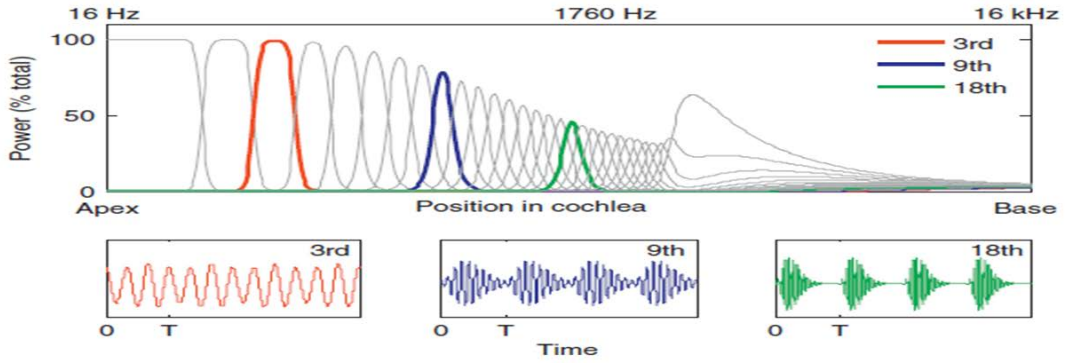
- Inter-spike interval for low frequency tones: 2 periods 1 period
- Inter-spike interval for high frequency tones (>5kHz): random intervals

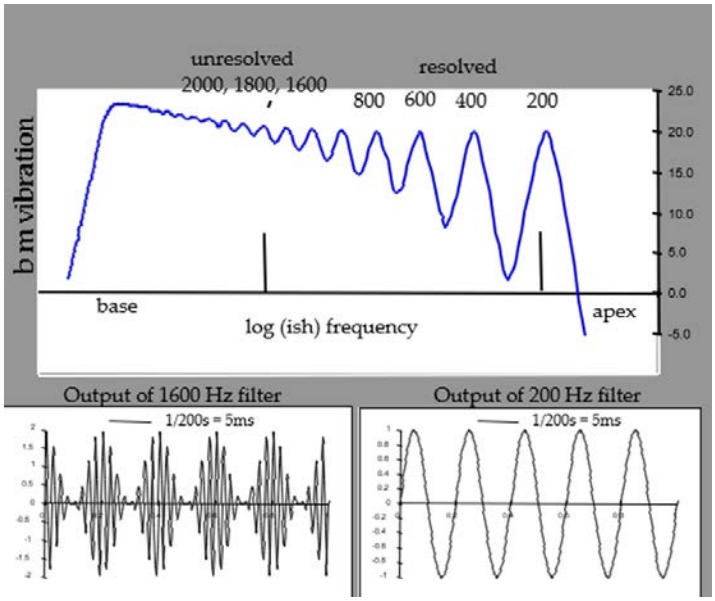




5.19. Timing theory: Unresolved harmonics in critical bands

Beating in HF auditory filters follows the periodicity of the “fundamental”.





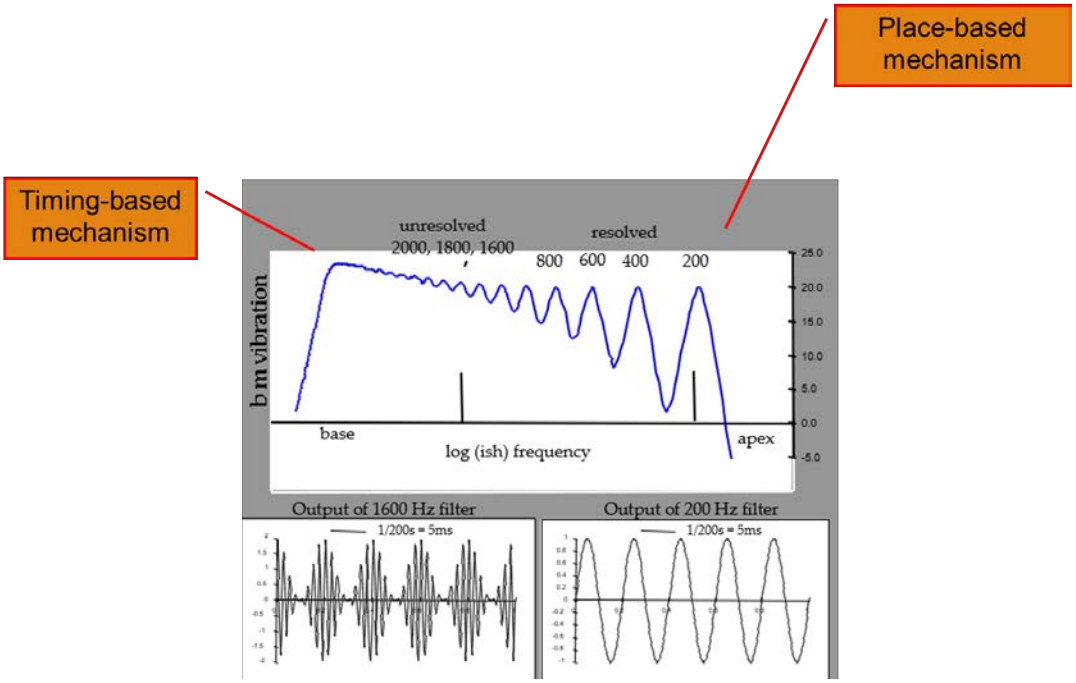
5.20. Two operating mechanisms

Place theory (Helmholtz) and timing theory (Seebeck) are closely linked with the volley principle (volley theory), a mechanism by which groups of neurons can encode the timing of a sound waveform. The combination known as the place-volley theory considers coding low pitches by timing pattern and high pitches by rate-place pattern.

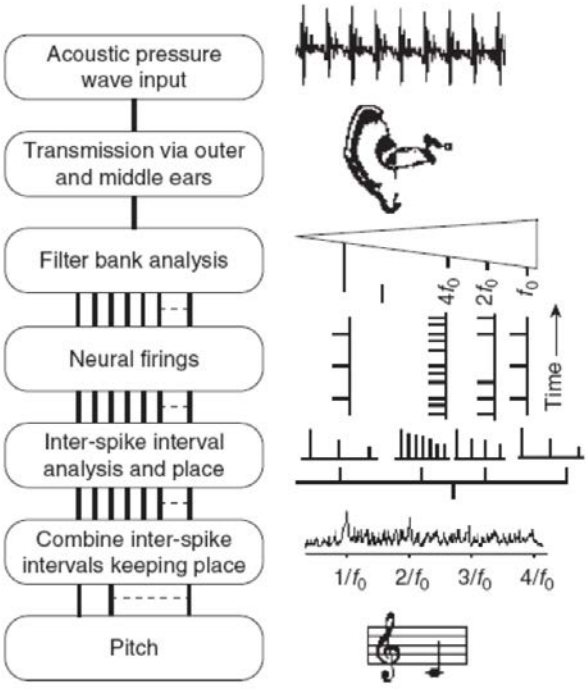
- Large vibrations with low rate are produced at the apical end of the basilar membrane
- Large vibrations with high rate are produced at the basal end.

Summary:

- Timing theory: the pitch of a pure tone is determined by the period of neuron firing patterns, either of single neurons, or groups as described by the volley theory. The more intense the vibration is, the more the hair cells are deflected and the more likely they are to cause cochlear nerve firings.
- Place theory: pitch is determined according to the locations of vibrations along the basilar membrane.



2 complementary mechanisms: place + periodicity



[LAB 4: PITCH]

- Pitch salience and tone duration:
    - Very brief tones are described as “clicks”, but as the tones lenghen, the clicks take on a sense of pitch which increases upon further lengthening.
    - The dependence of pitch salience on duration follows an “acoustic uncertainty principle”, there is a tradeoff between the uncertainty in frequency and the duration

of a tone burst (delta-f · delta-t = K). K, which can be as short as 0.1, depends upon intensity and amplitude envelope.
    - The duration needed to establish a clear pitch sensation for 300 Hz is longer than the duration required for higher frequencies.
    - Pitch salience depends on both the duration and the number of periods. For a given frequency, a tone must last for more than a minimum amount of time in order to evoke some clear pitch rather than just a "click". The minimum duration is inversely proportional to the frequency until high values of frequency (above 4000 kHz approx.). That is, for low frequencies we need more time to evoke a clear pitch sensation than for medium frequencies, then for frequencies higher than approx. 4000 kHz the minimum duration rises again (more slightly). In the case of the number of cycles, there is a direct proportional relationship with respect to frequency.
- Frequency Difference Limen or JND:
    - 100 Hz → Approx. 3Hz JND and 51 cents
    - 200 Hz → approx. 3Hz JND and 25 cents
    - 400 Hz → approx. 3Hz JND and 13 cents
    - 1000 Hz → 5 Hz JND and approx. 8 cents
    - 2000 Hz → 10 Hz JND and approx. 8 cents
    - 4000 Hz → more than 20 Hz JND and approx 9 cents
    - The JND for pitch has been found to depend on the frequency, the sound level, the duration of the tone, and the suddenness of the frequency change. Typically, it is found to be about 1/30 of the critical bandwidth at the same frequency.
- Virtual pitch:
    - One of the most remarkable properties of the auditory system is its ability to extract pitch from complex tones. Missing fundamental → virtual pitch.
    - When the partials are not exactly harmonics of a missing fundamental, we arrive at a "virtual pitch" by some strategy that may weigh several possibilities, and when the choice is difficult the pitch may be ambiguous (e.g. bass notes we hear from loudspeakers of very small size that radiate negligible power at low frequencies, subjective strike note of carillon bells, tuned church bells and orchestral chimes).
    - The pitch of the complex tone is determined by the **spectral structure**, which makes possible the "missing fundamental" phenomenon.
    - Masking and virtual pitch: we perceive the sinusoidal tone but not the complex tone as its harmonics have been masked. A place theory of pitch sensations cannot explain why the masking of the fundamental does not impede listening to it. A pitch perception theory has to explain why the masking of the fundamental does not impede perceiving it.
- Analytic vs synthetic pitch
    - Our auditory system has the ability to listen to complex sounds in different modes. When we listen analytically, we hear the different frequency components separately; when we listen synthetically or holistically, we focus on the whole sound and pay little attention to its components.

6. Psychophysics of basic sound dimensions: Timbre

6.1. Timbre

- ANSI (1960) + Plomp (1970): “that attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch and duration are dissimilar.”
- Hostmsa (1989): These properties include the sound’s spectral power distribution; its temporal envelope… rate and depth of amplitude or frequency modulation, and the degree of inharmonicity of its partials. The timbre of a sound therefore depends on many physical variables.
- Tone color, texture, quality… Highly multidimensional sensation.

Spectrum -> timbre

Time -> timbre

Waveform shape -> timbre

- The most direct way is modifying the shape of the signal (waveform), then we are having a different timbre sensation.
- 2 Listening modes:
    - **Default**: Source-based
    - **Reduced**: Acousmatic, sound qualities.

[PERFECTO QUIZ CARD] **Timbre**

- ASA (Auditory Scene Analysis): differences in sonic quality not captured by loudness, pitch or duration.
- By characteristics: sensory attributes (sharpness, brightness, nasality, richness), some of which are continuous, some discrete, some categorical.
- Important for classification of genre

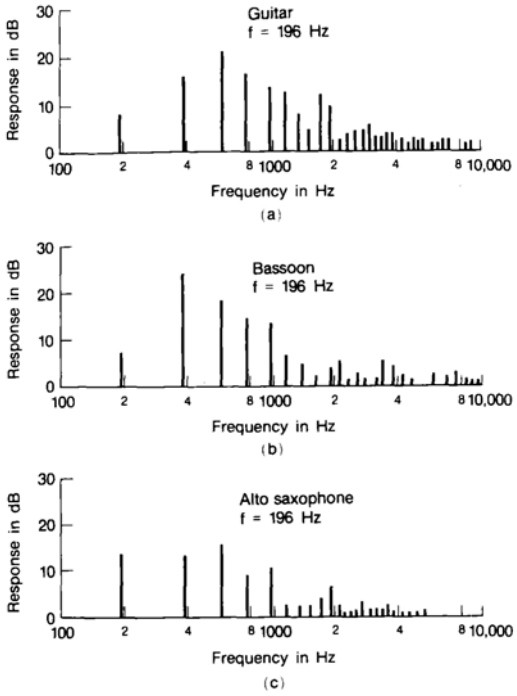
[PERFECTO QUIZ CARD] **Perception of timbre**

- Set of abstract sensory attributes, some of which are continuously varying (e.g. attack sharpness, brightness, nasality, richness), others of which are discrete or categorical (e.g. blatt at the beginning of a sforzando trombone).
- One of the primary perceptual vehicles for the recognition, identification, and tracking over time of a sound source (singer’s voice, clarinet, set of carillon bells).
- Techniques: multidim scaling

6.2. Defining attributes of timbre

- Range between **tonal** and **noiselike** character
- The **spectral envelope** (frequency)
- The **amplitude envelope** in terms of rise, duration and decay (time)
- The change both of spectral envelope (formant glide) or fundamental frequency micro intonation
- The **prefix**: an onset of a sound quite dissimilar to ensuing lasting vibration.

6.3. Timbre = Spectrum ?



6.4. Envelopes: the temporal factor

- Envelopes are functions representing the temporal evolution of partials, overall amplitude, etc..

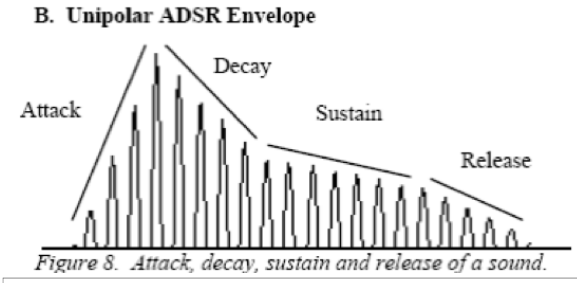
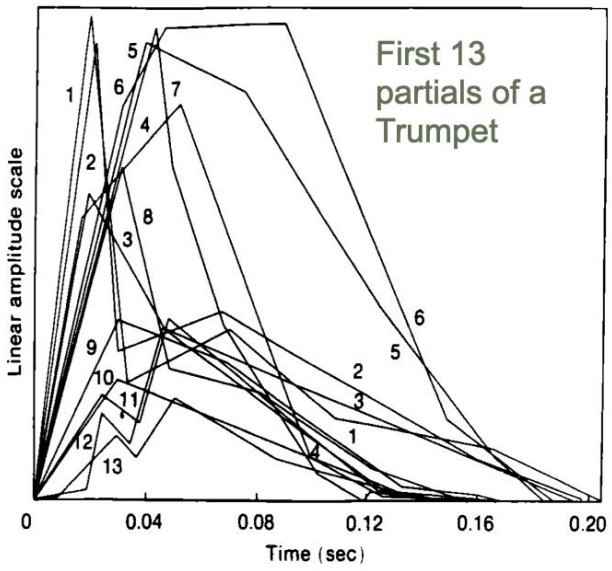
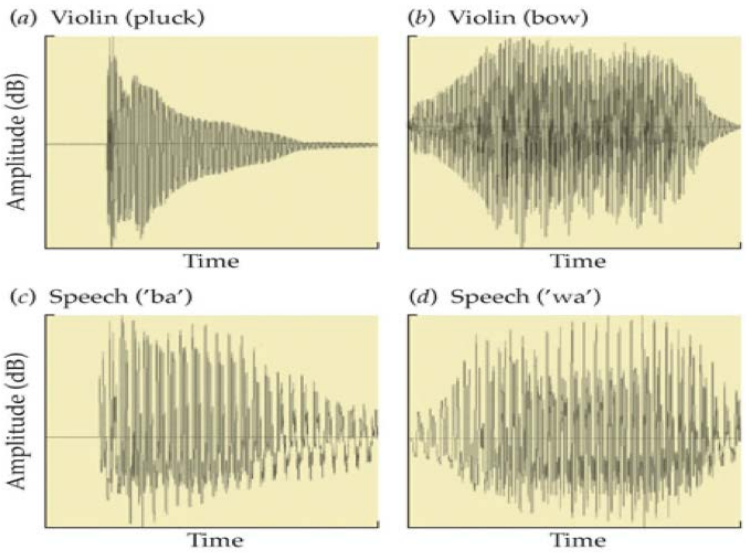


Figure 8. Attack, decay, sustain and release of a sound.



6.5. The attack

- Really important when it comes to **source characterization. Believed to convey most of the source information.**
- Noisy character in some instruments
- Different attacks change the timbre sensation…. But not necessarily the source name -> invariant perception of source.



6.6. Formants

- **Spectral regions of high-energy**, that impart distinctive characteristics to the sounds. Allows us to distinguish the vowels. Prominent harmonics with prominent energy.

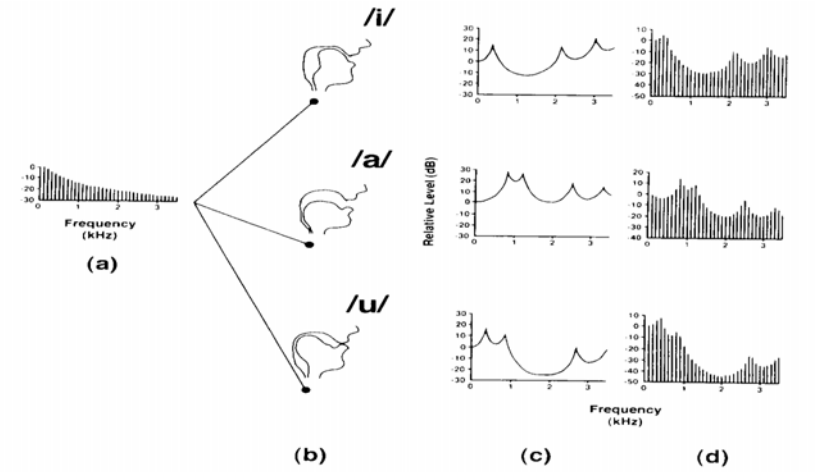
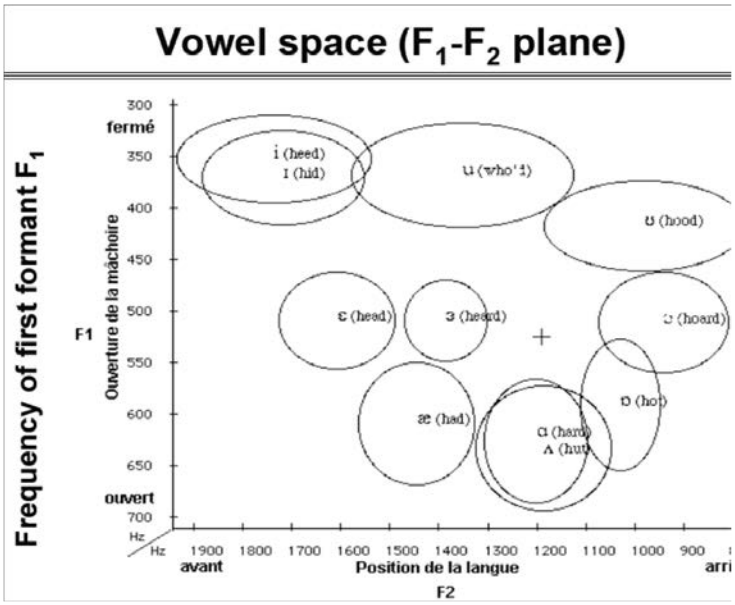
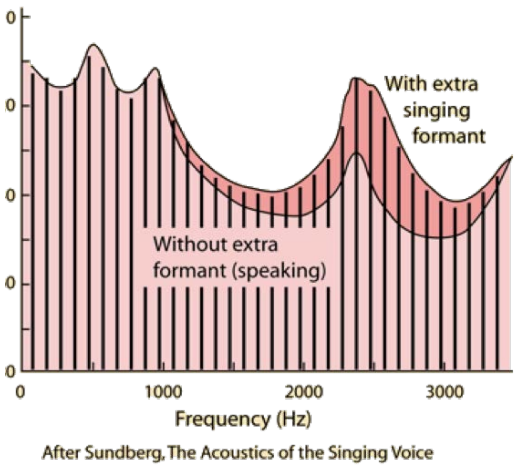


FIG. 8.1 Illustration of how three different vowel sounds are produced. Part (a) shows the spectrum of the sound produced by vibration of the vocal folds. It consists of a series of harmonics whose levels decline with increasing frequency. Part (b) shows schematic cross-sections of the vocal tract in the positions appropriate for the three vowels. Part (c) shows the filter functions or transfer functions associated with those positions of the vocal tract. Part (d) shows the spectra of the vowels resulting from passing the glottal source in panel (a) through the filter functions in panel (c). Adapted from Bailey (1983) by permission of the author.

- Two formants can be enough to identify different classes of sounds (e.g vowels)



- The **singer's formant** is a skill developer by male singers to stand in front of the orchestra



After Sundberg, The Acoustics of the Singing Voice

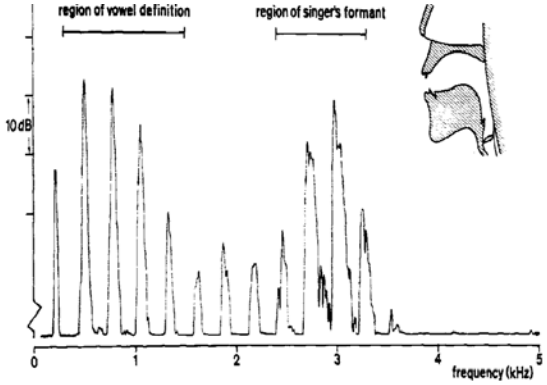
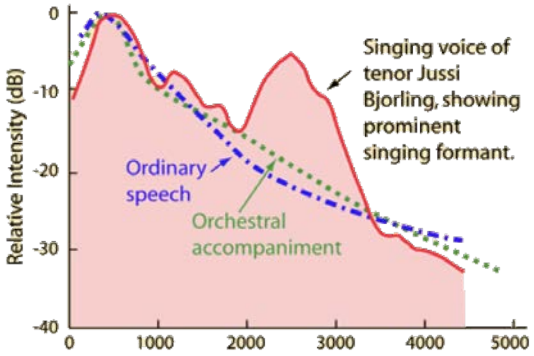


Figure 4.6. The vowel [ɔ] (as in "hawed") sung at approximately 262 Hz (C_4). The spectral envelope indicates desirable vowel definition and singer's formant. Note the favorable balance in sound energy between the region of vowel definition and that of the singer's formant. (From Richard Miller and Harm Kornelis Schutte, "The Effect of Tongue Position on Spectra in Singing." *The NATS Bulletin*, January/February, 1981, Vol. 37, No. 3. By per-



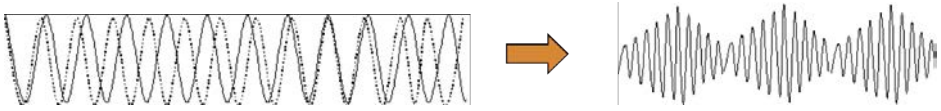
If you need to capture in a simplified way the spectral shape of a sound, which way would you use to encode the shape? Bark only needs 24.

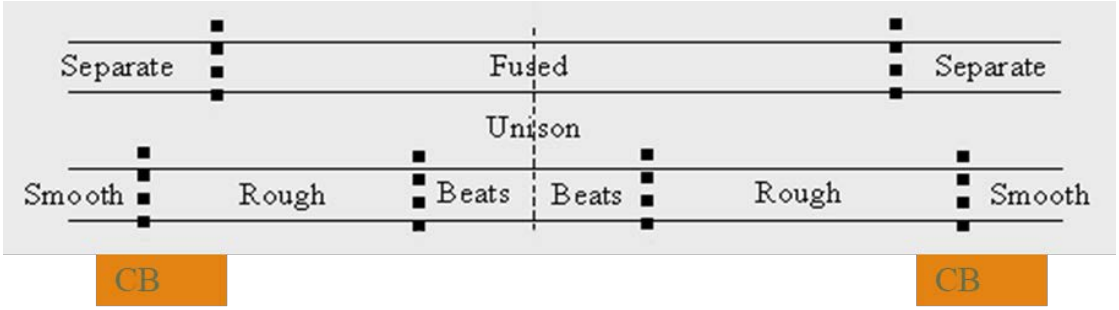
6.7. Combination of tones: Tartini tones

- At **loud and very loud levels**, 2 tones that are **close in frequency** may create a **"ghost" tone** corresponding to the **sum** and to the **difference (Tartini Tone)** of the frequencies.
- Other extra tones may also appear, all of them obbeying to a **non-linear distortion** process **happening in the middle-ear or/and in the early auditory processing**.

6.8. Combination of tones: Beating-Roughness

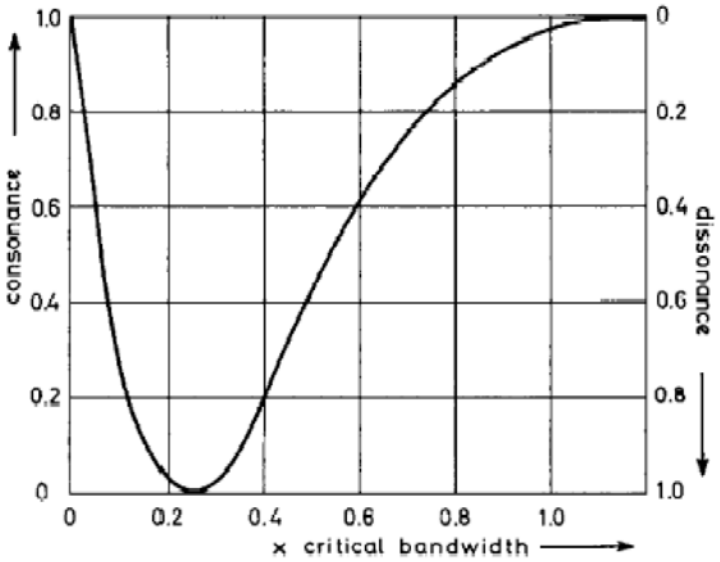
- f0-f1 < 6Hz -> **beating single tone**
- f0-f1 > 24Hz & f0/f1 < 0.1 -> **roughness**





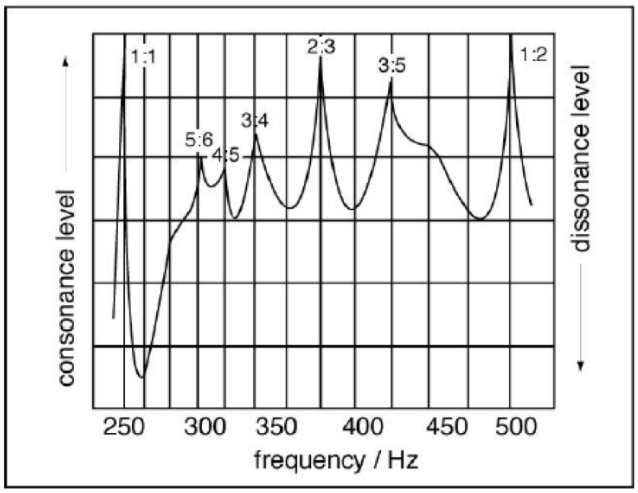
6.9. Combination of tones: Sensory consonance-dissonance

- **Consonance**: ratio among their frequencies is a **natural number or a simple fraction**.
- **Dissonance**: Ratios with **non-natural numbers**, e.g 16:15 - minor second, 45:32 - tritone

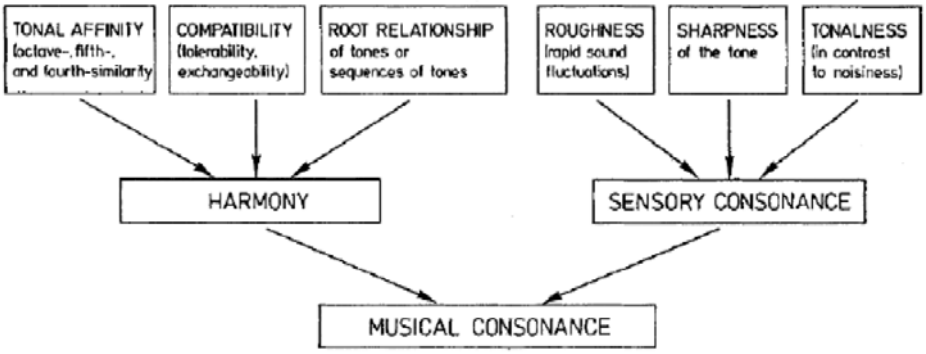


- **Highest consonances**: obtained for **unison** and **octave** relationships (1:1, 1:2), **fifth** (2:3), **Major sixth** (3:5)
- **Lower consonances**: are obtained for **fourth** (3:4), **Major third** (4:5), and **minor third** (5:6)
- **Minor second yields the lowest consonance**
- Other **low-consonant combinations** correspond to the **minor ninth**, **Major and minor sevenths**, **tritone**, and **Major second.**

Sensory consonance as a function of the distance between partials. The frequency of the base tone is 250 Hz:

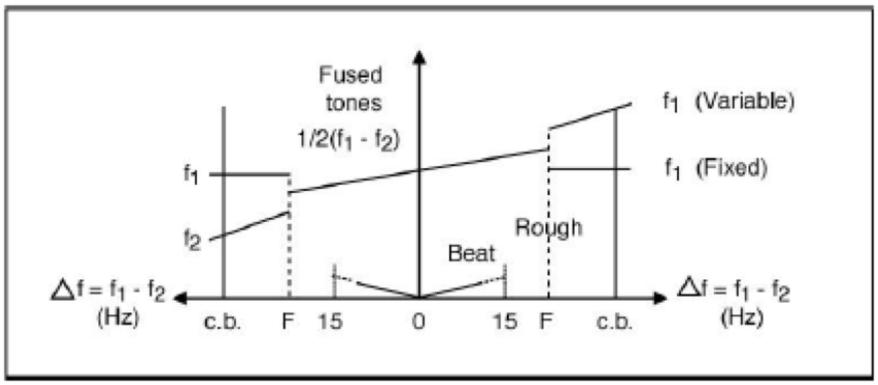


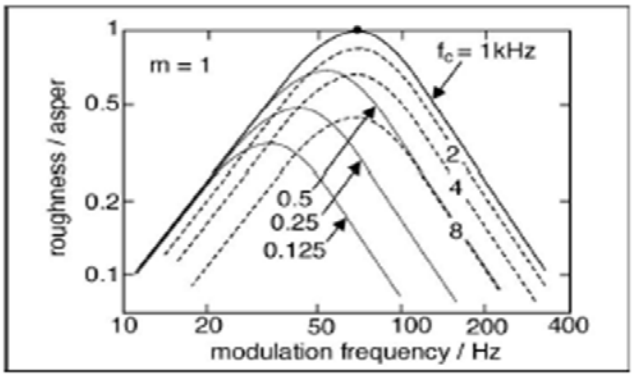
6.10. Sensory Consonance and Musical Consonance

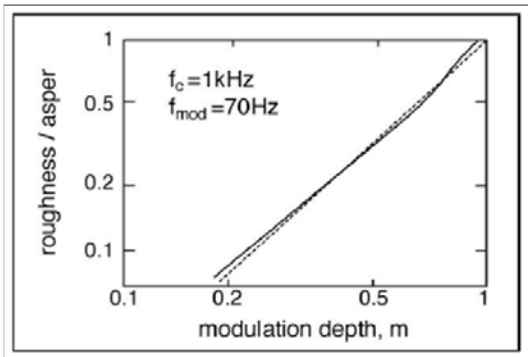


6.11. Roughness

- **Partials sharing a critical band but separated more than 15Hz** create a **deep AM** -> **roughness**.
- Measured in "**asper**"
- 1 asper is the roughness generated by a 1 Khz tone at 60 phone, modulated at 70 Hz, full depth.







6.12. Fluctuation Strength

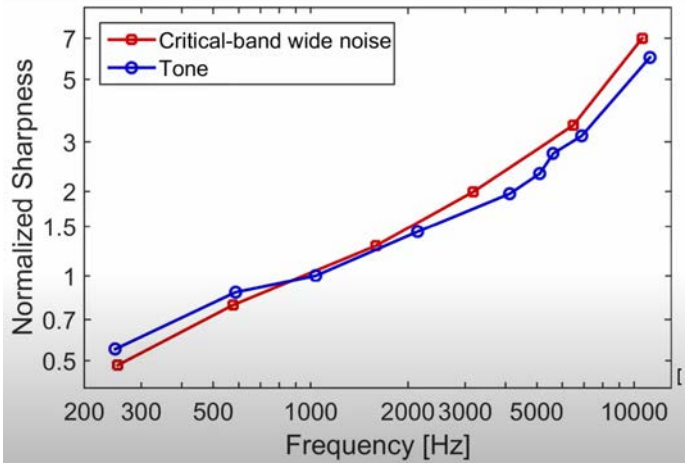
Sometimes used to characterize timbre. Sensation produced by **slow beating**. For fast beating the quality of the sensation changes (roughness). Used in several studies, for example, on the annoyance of sounds.

- FS ~ Slow beating measured in "**vacils**"
- 1 vacil = 1Khz tone, fully modulated at 4 Hz at a level of 60 phons
- Fast Beating (>20Hz) -> Roughness

6.13. Sharpness

Expresses the quality of a sound to sound sharp, strongly related to how pleasant we feel with the sound. Dominated by **energy and high frequencies**, it is a question of the **spectral envelope of a sound**. If there is strong high frequency energy in a sound then sounds will be sharper than if there are not strong high frequency energy present.
Normalized sharpness as a function of center frequency of noise that is wide as a critical band. Sharpness increases the higher up we go up in center frequency of the noise. The high frequencies are much more strong and dominating than low frequencies (x axis is logarithmic).

- Sharpness increases with high-frequency energy.
- **Sharpness is one measure of sound quality**, also fluctuation strength or temporal modulations are another important aspect.

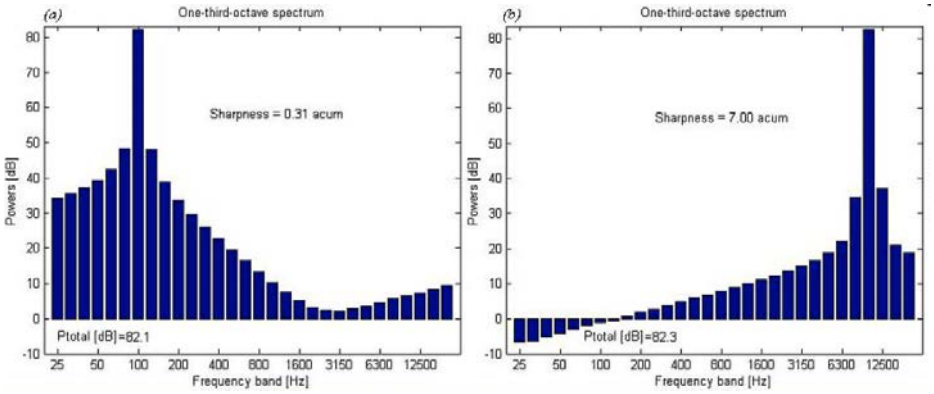


- Sharpness = Brightness

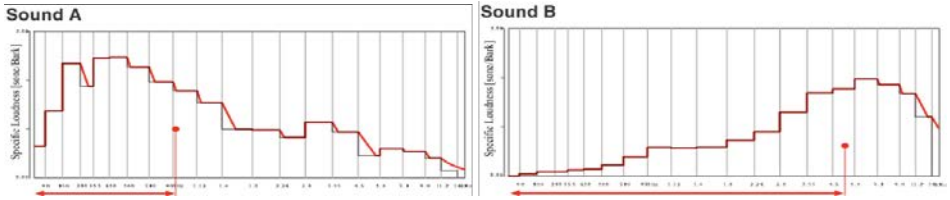
- Sensation related to the position of the **spectral centroid**, the gravity center of the spectrum (point that divides energy into 2 halves):

$$\sum_{i=1}^{i=N} a_i f_i \Big/ \sum_{i=1}^{i=N} a_i$$

- Measured in "**acums**"
- 1 ACUM is the sensation generated by 1 Bark bandwidth noise, centered at 1 Khz, at 60 phon.
- **Sharpness increases with frequency and intensity**



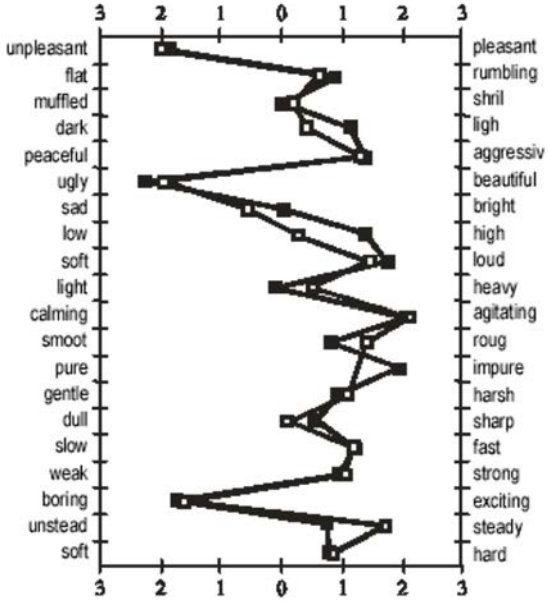
Trumpet dark (left) vs. bright (right) sound:



6.14. Sound description

Sound description is a **way to describe timbre sensations**. Techniques:

- **Rating scales/Likert scales**: How much does the sound have some listed properties?
- **Semantic differential**: Based on pairs of opposed adjectives; see figure.



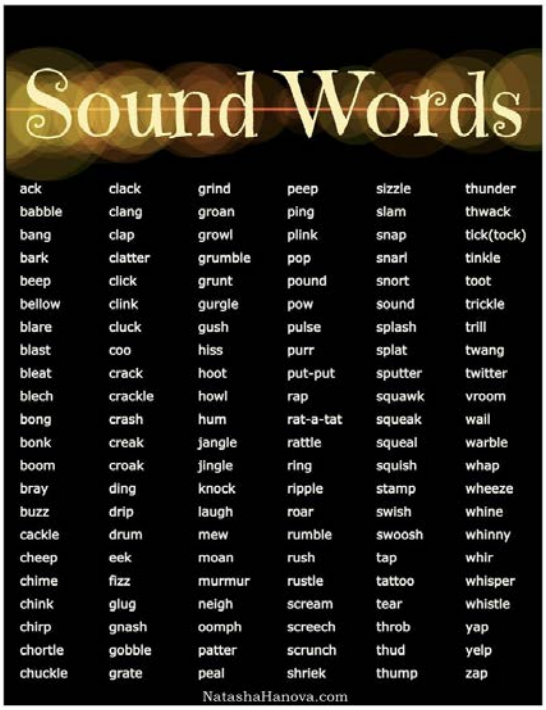
The most frequent adjectives are bright-sharp/dull-mellow, fast/slow, soft-rough.

6.15. Use of timbre descriptors in applied contexts

- **Sound branding**: growing discipline. Giving a product a given quality that tells about a given status, power, functionality, etc. Audio branding (especially in Northern Europe) has a strong presence, with many researchers working on conveying messages/images on the content of the music. Take advantage of some cultural associations, personal or historical associations to transmit values of a product/company.

Table 4 Correlation between the sound metric and seven major sounds

Driving condition	Sound metric	Correlation	Sound metric	Correlation
60 km/h	Roughness	−88.4	Articulation index	91.3
100 km/h	Zwicker loudness	−94.0	Roughness	−82.8
Sun roof	Sharpness	−90.6	Roughness	−75.2
Turn signal	Zwicker loudness	−92.2	Fluctuation strength	−95.4
Door lock	Sharpness	−89.8	Articulation index	88.0
Second gear, WOT, group 1	Zwicker loudness	−78.7	Roughness	82.5
Second gear, WOT, group 2	Zwicker loudness (0–2 Bark)	92.6	Sharpness	−93.7
Third gear, WOT, group 1	CEO	−84.0	Zwicker loudness (0–2 Bark)	−97.5
Third gear, WOT, group 2	Sharpness (8 Bark low)	−87.2	Fluctuation strength	−73.3

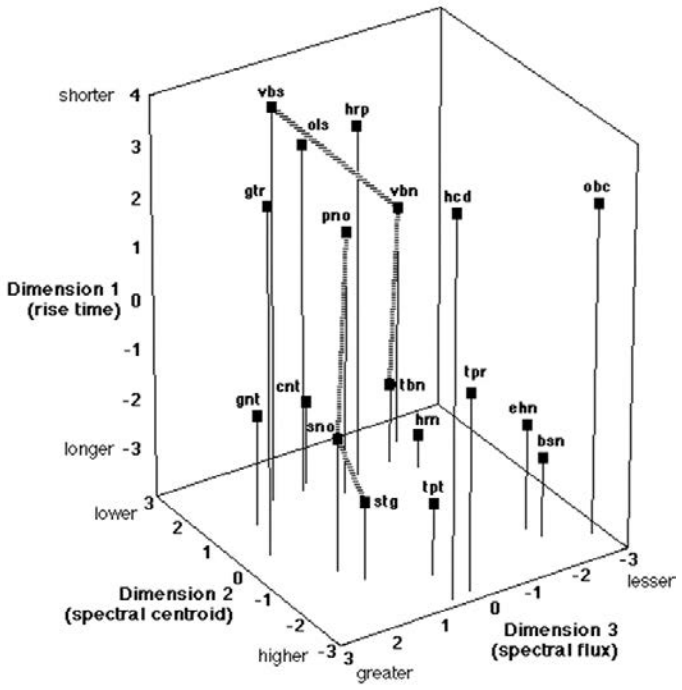


6.16. Timbre spaces

Timbre Spaces: emerged in the 70s in California. Developed a lot by MacAdams in Paris in the 90s. The idea would be good to get a **graphical representation** of timbre sensations, to decide if two sounds are close in our mind. Researchers presented pairs of sounds and subjects were asked to rate the similarity. Some techniques called **multidimensional scaling**, generate a **map of similarities**. This allows us to **interpolate** between two sounds in order to **create hybrid sounds**.

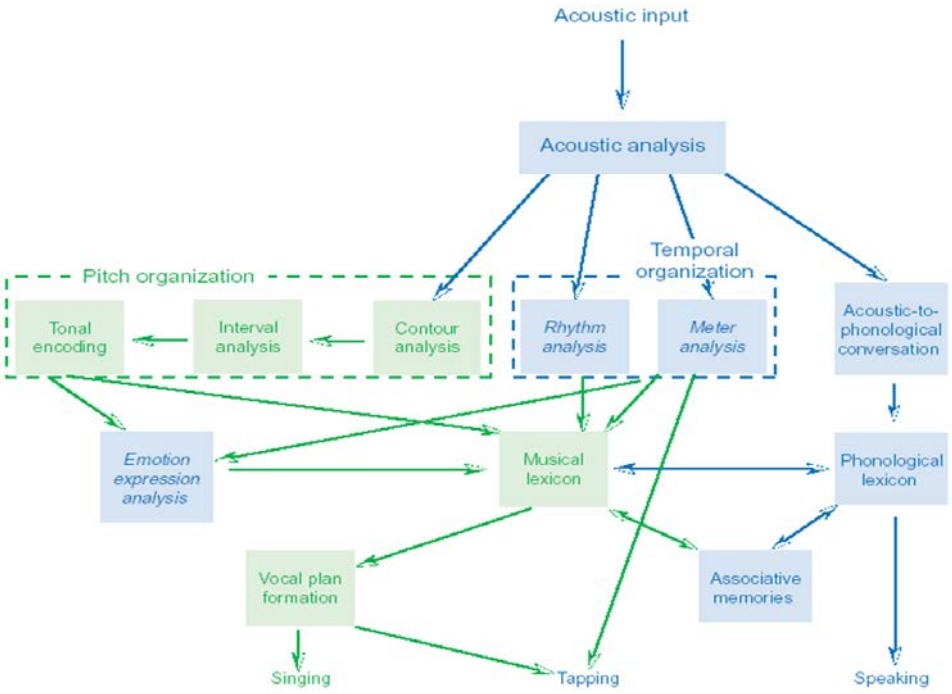
Methodology:

- Similarity judgments between pairs of sounds
- Multidimensional scaling techniques → **perceptual space**
- Audio analysis of the presented sounds → **audio descriptors. Dimensions**:
    - **Rise time**: time to reach maximum amplitude
    - **Spectral centroid**: gravity center of the spectrum
    - **Spectral flux**: difference between the spectral energy in consecutive frames



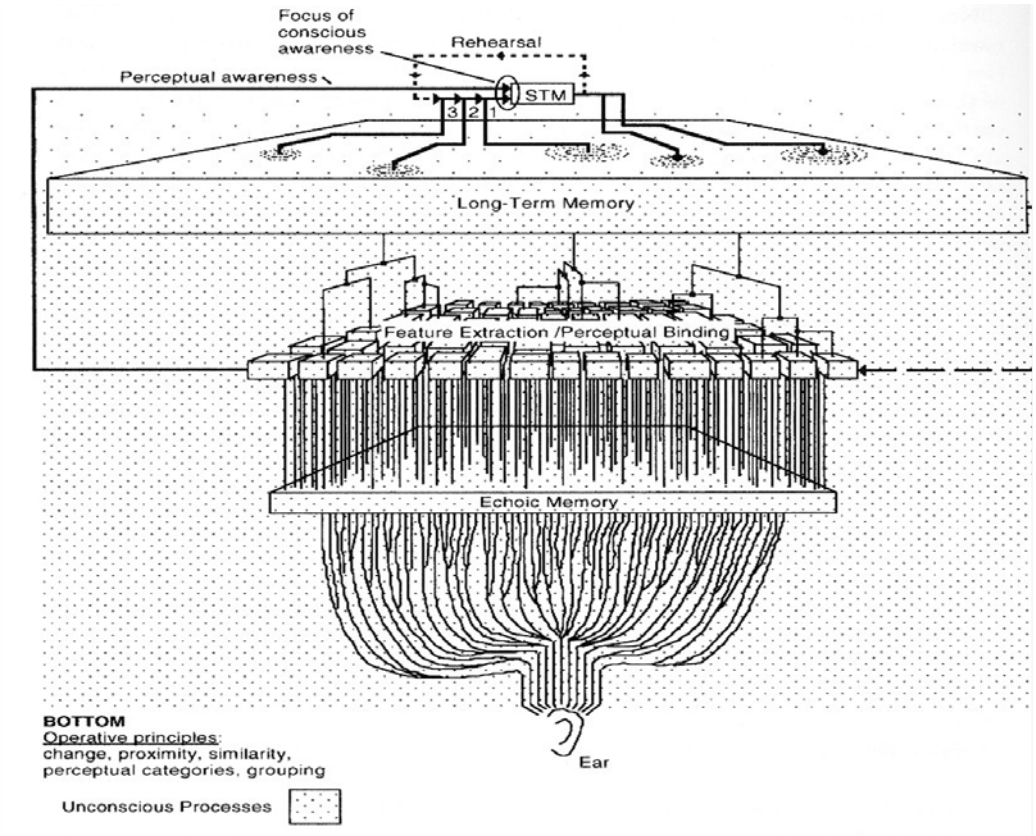
7. Perceptual organization (making sense of sonic patterns)

7.1. A graphical model of some involved processes



7.2. Perceptual Binding

Our sensory systems are constantly inundated with information from the outside world. The human nervous system has evolved to be able to take that sensory information and convert it into neural representations that are then used in our everyday cognitive functions and interactions with the world. One critical step is referred to as **perceptual binding: the process of merging individual bits of sensory information into coherent representations**.



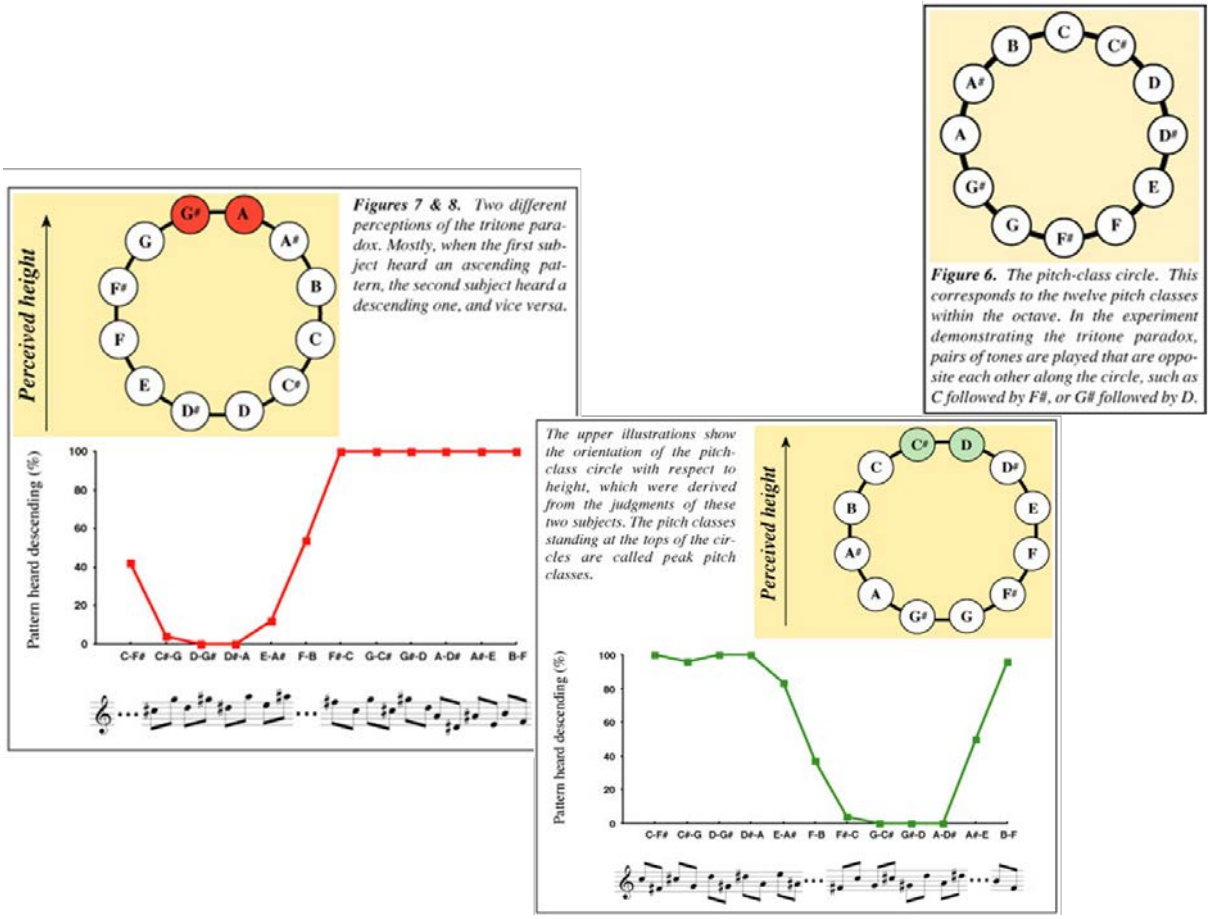
7.3. Perceptual Organization

Pomerantz & Kubovy, 1986:

- "... Perceptual Organization is central to the key question of perception: how do we **make the leap** from information detected by our sensory receptors… to our perceptions of the world? This requires not just the detection of information but also the organization of that information into **veridical percepts.**"
- "...Perceptual Organization is the process by which particular **relationships among potentially separate elements** (including parts, features, and dimensions) are perceived (**selected from alternative relationships**) and guide the interpretations of those elements… in sum, how we process sensory information in **context**."

7.4. The tritone paradox

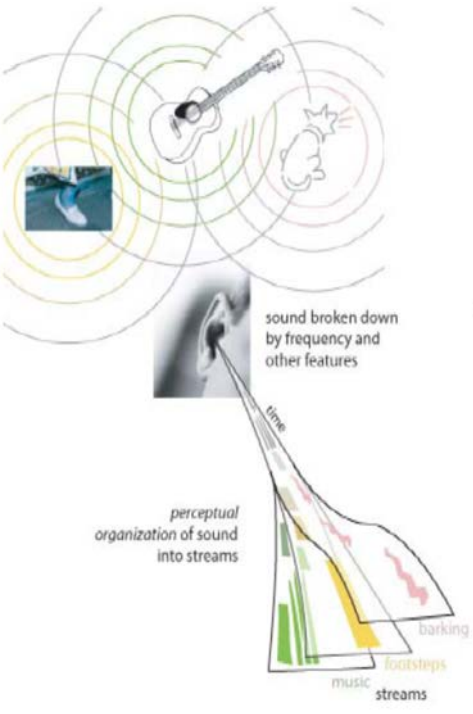
The tritone paradox is an **auditory illusion** in which a **sequentially played pair of Shepard tones** separated by an interval of a tritone is heard as ascending by some people and as descending by others. The way these tone pairs are perceived varies depending on the listener's language or dialect.

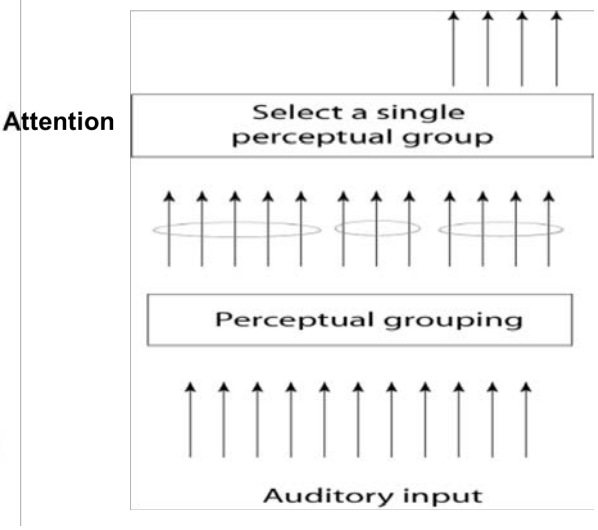


7.5. Audio-visual interactions: the McGurck effect

The McGurk effect is a perceptual phenomenon that demonstrates an **interaction between hearing and vision in speech perception**. The illusion occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the **perception of a third sound**.
If a person is getting poor quality auditory information but good quality visual information, they may be more likely to experience the McGurk effect.

7.6. Perceptual organization (II)





7.7. Some terms

- **Source**: physical entity that gives rise to the sound pressure waves (e.g. a violin)
- **Stream**: percept of a group of successive and/or simultaneous sounds as a **coherent whole** appearing to come from a **single source** (e.g. brass section)
- The sounds we hear at any one time usually come from a number of different sources.
- In most cases we can hear and identify each of the different sound sources as having its own pitch, timbre, loudness and location (stream and source). In other cases several sources are processed as a single stream as their features do not qualify for being considered as "distinct" (e.g. string section). In other (exotic) cases, a single source may yield different streams.

7.8. Auditory Scene Analysis

Auditory Scene Analysis (ASA) can be conceptualized as a **two-stage process**:

1. The mixture of sounds is **decomposed** into a **collection of sensory elements** (onsets, pitch trajectories, modulations, spectral tracks, etc.)
2. Elements that are likely to have arisen from the same event are **grouped to form a perceptual structure (stream)** which can be interpreted by higher centers in the brain.

For example, when listening to a violin performance, it is the task scene analysis to group the acoustic events emitted from the physical source (violin) into a perceptual stream (mental experience of a violin being played).

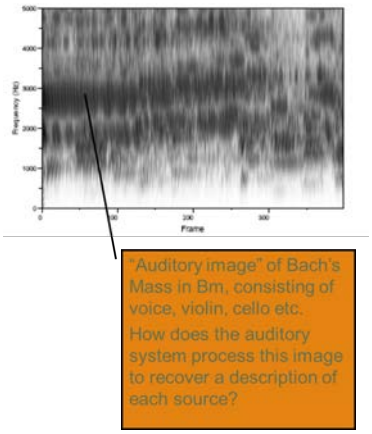
Another way of listening: **reduced listening**.

In most listening situations, a mixture of sounds reaches the ears. However we can:

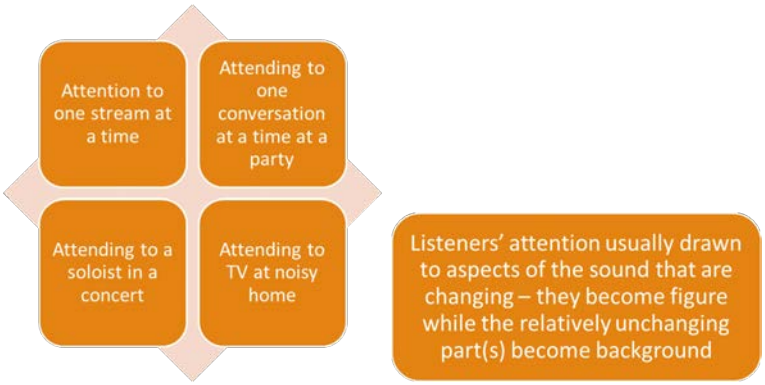
- **Attend one conversation** among many competing voices and other background sounds (cocktail party).

- **Follow the melodic line** played by the violins in an orchestral recording.

This problem is of great scientific interest, and a solution also has engineering applications.



7.9. The Figure-Ground Phenomenon

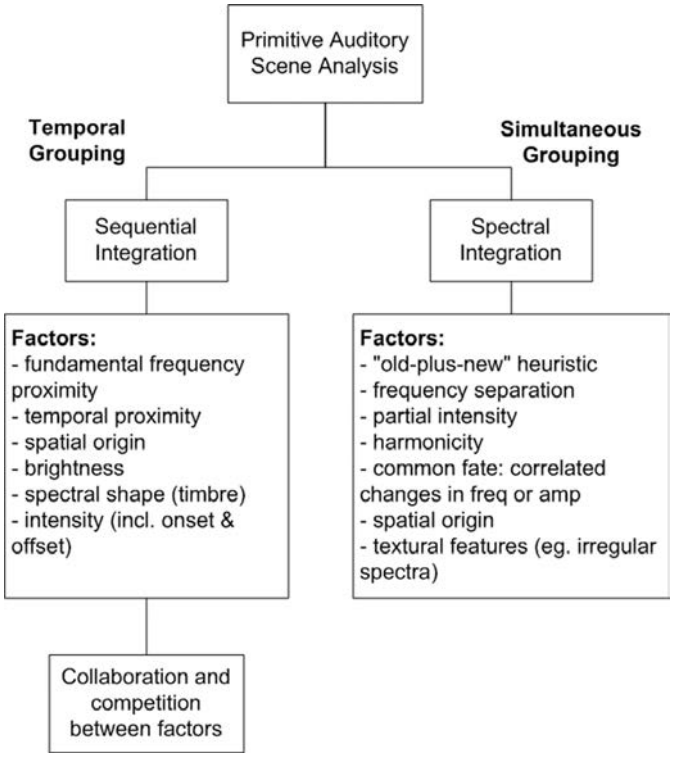


7.10. Figure-Ground: Scrambled melodies

- **Frequency separation** progressively improves recognition of the melodies
- **Intensity** first, then **timbre** are used to favor groupings of notes, improving the identification of the melodies.

7.11. Grouping heuristics

- **Simultaneous grouping**: the grouping together of the **simultaneous frequency components** that come from a single source.
- **Sequential grouping**: the **connecting over time of the changing frequencies** that a single source produces from one moment to the next.



Gestalt Laws: principles by which humans tend to group visual or auditory information:

A description of auditory and visual Gestalt Laws.

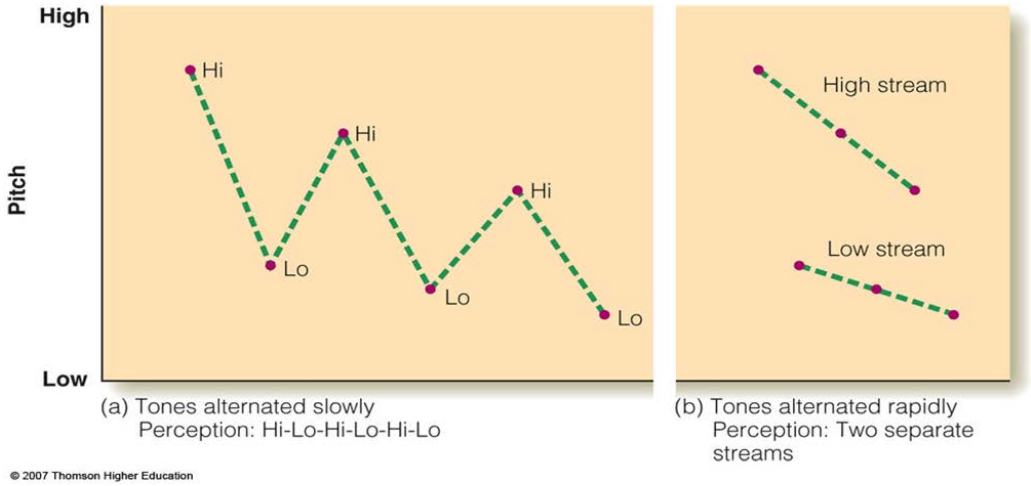
Name	Audition	Vision
Proximity (Belongingness)	Sounds arriving from places *close in space* tend to be grouped	Elements *close together* in space tend to be grouped
Similarity	Sounds with *similar timbre and pitch* tend to be grouped	Elements *shaped alike* tend to be grouped
Good Continuation	Sounds that follow a *regular pitch contour* tend to be grouped	Elements that follow a *regular spatial contour* tend to be grouped
Closure	Interrupted auditory stimuli tend to be perceived as *continuous* when plausible	*Borders are interpreted/completed* to specify shapes
Simplicity (Pragnanz)	Frequencies with *simple harmonic ratios* tend to be grouped	*Prototypical* shapes tend to be regular, simple, symmetric
Common Fate	Sounds with *synchronous rhythm patterns* tend to be grouped	Elements that *move together* tend to be grouped

7.12. Proximity

Stream segregation depends on proximities (temporal, pitch, etc.)

When stream segregation occurs, we are unable to attend fully to the events in both streams at the same time.

- Figure-background phenomena
- Rhythmic confusions
- Bi-stability

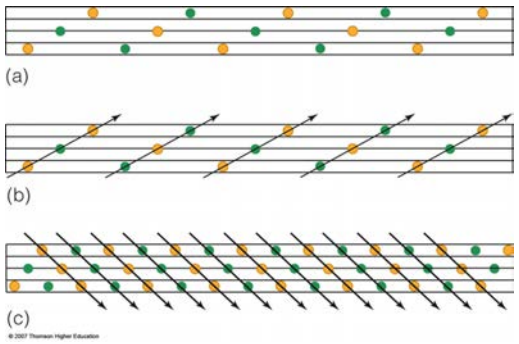


7.13. Wessel's illusion

The repeating series of 3 notes presented by Wessel (1979). The yellow circles stand for a tone with one timbre, and the green circles for a tone with a different timbre.

When the tones are presented **slowly**, they are perceived as **ascending sequences** of notes that alternate in timbre.

When the tones are presented **rapidly**, they are perceived as **descending sequences** of notes with the same timbre. This is Wessel's timbre illusion.

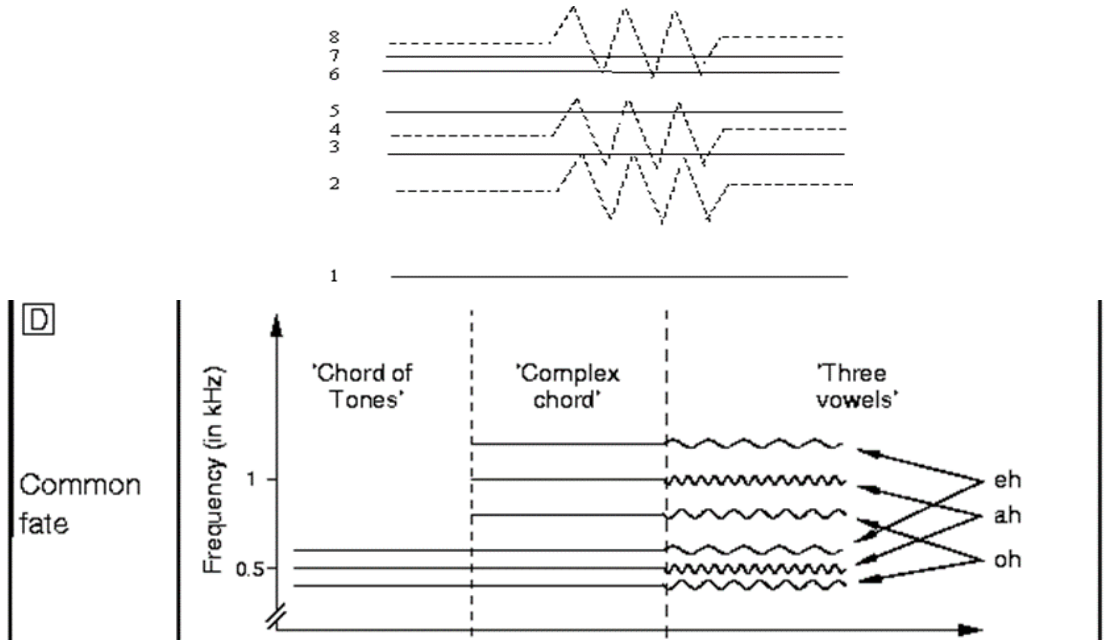


7.14. Common Fate

Components in sound act together. They tend to start and finish together. They tend to change in pitch or intensity together.

Therefore, if we have a complex sound and the components are coordinated then they are fused, e.g. onset disparities, and AM and FM (tremolo & vibrato).

- **Common fate: elements that change in the same way are perceptually linked together. Correlated changes in frequency or amplitude.** For example, if harmonics 2, 4 and 8's frequency is modulated (FM) they separate from harmonics 3, 5, 6 and 7.



7.15. Closure

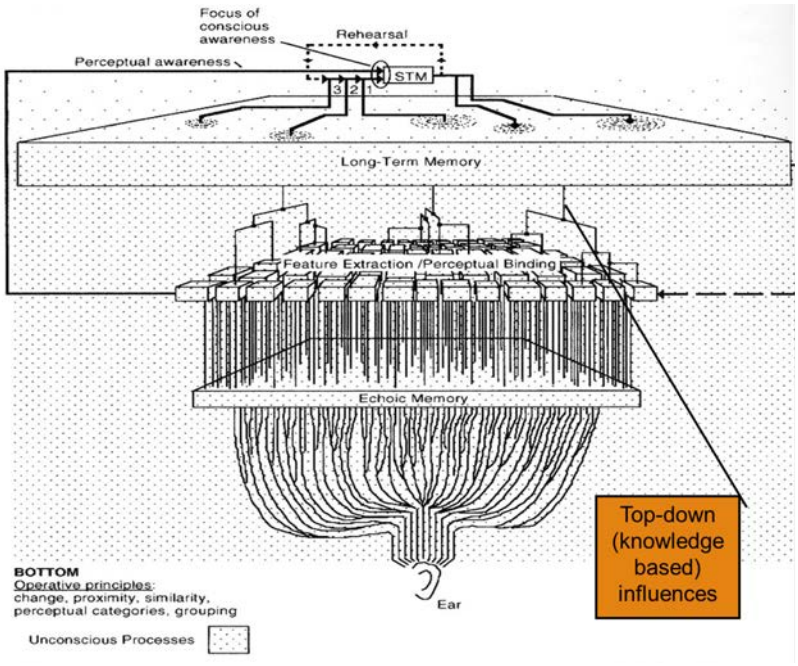
- A source maybe obscured or absent -but its percept continues.
- Drums occlude long notes, but we do not worry about that.
- In speech this is known as "**phoneme restoration**".

8. Music and memory

Music is a technology of memory: music helps/improves our memory ("music is not just cheese cake").

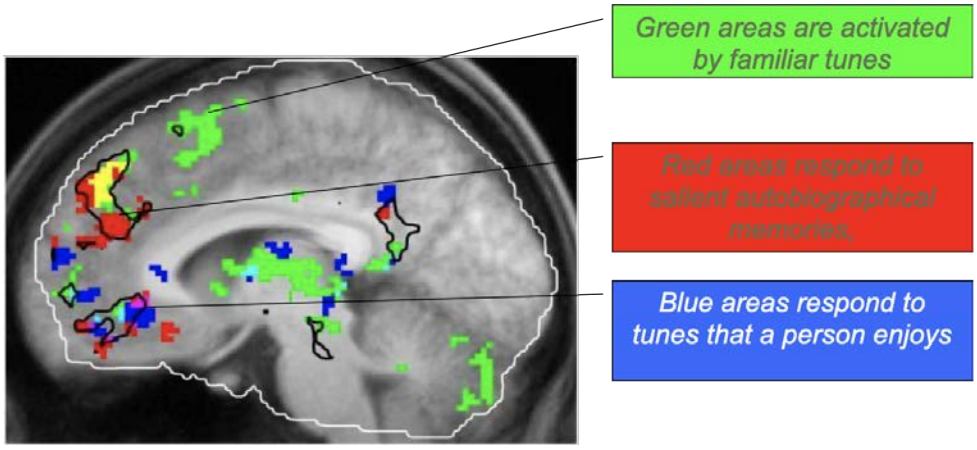
8.1. Memories

- **Echoic memory**: sensory memory that registers **specific to auditory information**. Once an auditory stimulus is heard, it is stored in memory so that it can be processed and understood. Stored for **slightly longer periods of time than iconic memories** (visual memories).
- **Short term memory (STM) (working memory)**
- **Long term memory (LTM)**

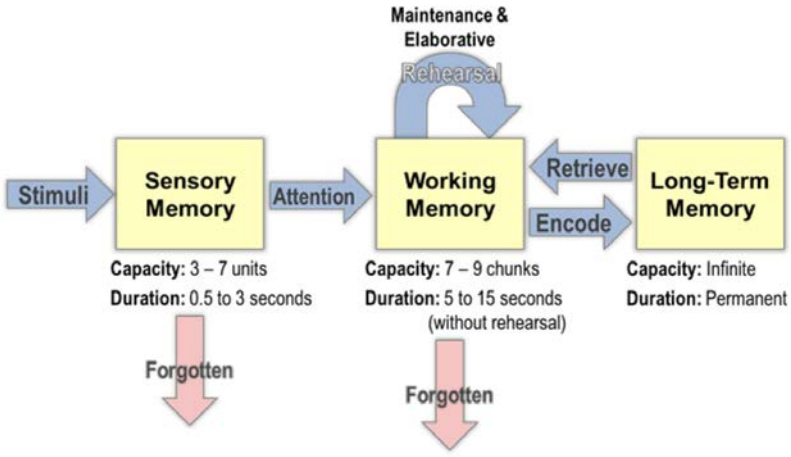


8.2. Simultaneous activation of different “memories” when listening to music

Memories are distributed across our cortex: some traces in **auditory cortex**, some in **visual cortex**, some in **motor cortex**, some in **frontal cortex.**



8.3. Multi-storage view of memory



8.4. A functional view of memory systems

- **Echoic Memory**: Feature extraction & perceptual binding. <-> fast decay
- **STM**:
    - **Segmentation and chunking**.
    - Forgetting information from STM can be explained using the theories of **trace decay** and displacement: **memories leave a trace in the brain** (some form of physical and/or chemical change in the nervous system).
    - <-> Decay of traces: trace decay theory states that forgetting occurs as a result of the **automatic decay or fading of the memory trace**. This theory suggests STM can only hold information for between 15 and 30 seconds unless it is rehearsed.
    - <-> Interference of other traces
    - **Displacement theory**: because of its limited capacity, STM can only hold small amounts of information. When STM is "full", new information displaces or "pushes out" old information and takes its place.
    - **Interference theory**: memory can be disrupted or interfered with by what we have previously learned or by what we will learn in the future
- LTM:
    - Explicit learning rehearsal using verbal codes (inner voice)
    - mnemonics: any learning technique that aids information retention or retrieval in the human memory for better understanding.
    - **Implicit learning** (learning by exposure)
    - **Consolidation** (hippocampus and sleep effect are fundamental)
    - Forgetting from LTM can be explained using the theories of interference, retrieval failure and lack of consolidation.
    - <-> Interference and Reconstruction

The encoding process is **interactive** (different memories from different types of memory can interact between them). Memory operations with music:

- Reaction to familiar tones
- Evaluative judgment: i like it or not (areas in cortex activate)
- Belonging to my autobiographical memories (frontal cortex)

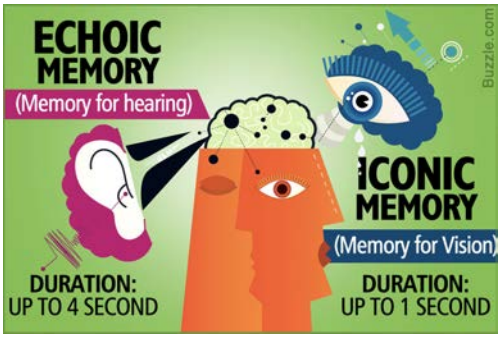
Categories are the most valuable mechanisms for our memory.

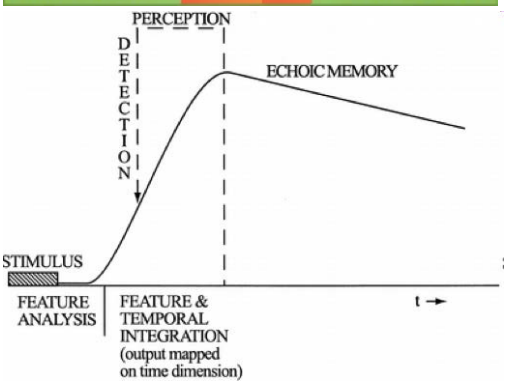
8.5. Iconic memory

- Iconic memory involves the **memory of visual stimuli**. The word iconic refers to an icon, which is a pictorial representation or image. Icon memory is how the brain remembers an image you have seen in the world around you. ... Iconic memory is a type of sensory memory that lasts just milliseconds before fading.
- Experiment: You just have to stare at the little plus sign in the middle for about a minute, and then quickly look at the white surface. The effect will be stronger if you blink your eyes repeatedly.

8.6. Echoic or Sensory Memory

- Trace that stimuli leave just **after the transduction**
- Probably it is accounted for by the connections up to the thalamic centers.
- It accounts for the **basic feature extraction** (pitch, intensity, onset, noise, harmonic pattern, spatial position…
- **Extracted features are "bounded"** (what goes with what, e. g. a given partial with another one and with a fundamental coming all of them from the same direction)
- **Active up to 4 seconds, very fast degradation after 50ms**
- **Sensory coding** (no conceptual or categorical coding available)





8.7. Short-term Memory

- **Temporal storage** for helping the permanent encoding (aka "**Working Memory**")
- Ground of our sensation of "present", a shifting focus of awareness selects what/how much is processed.
- Has limitations: **capacity (7+/-2 items) and time (around 10-20 seconds).**

- **Chunking** and **rehearsal** help to overcome these limitations.
- Has "**sensory-specific**" **sub-blocks** (for dealing with visual-spatial, phonological, pitch information –parallel processing of them).
- **Attentional mechanisms** "modulate" what is stored in short-term and which type of processing is devoted to that.

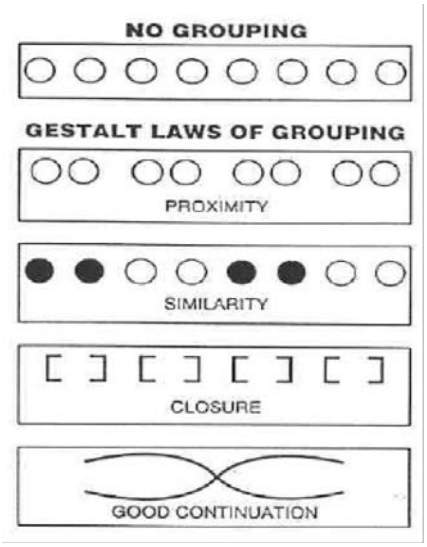
8.8. Operations in STM: Segmenting

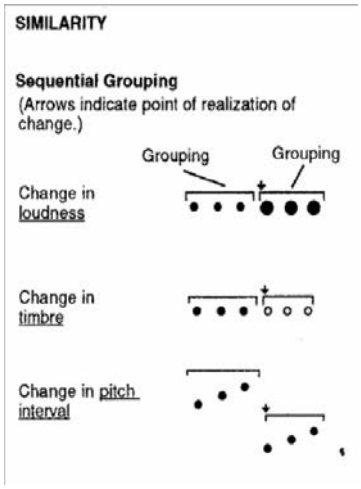
The continuous musical texture is broken into shorter sequences using "**segmentation cues**":

- Closure and change detection
- Pauses, silences, stretching of notes
- Instrument changes
- Cadences
- Accents and other metrical elements
- Tendency changes, contrasts (up-down melody, long-short notes, etc)

8.9. Operations in STM: Grouping

The **Gestalt laws of grouping** are a set of principles to account for the observation that humans naturally perceive objects as organized patterns and objects. Five categories: Proximity, Similarity,Continuity, Closure and Connectedness.





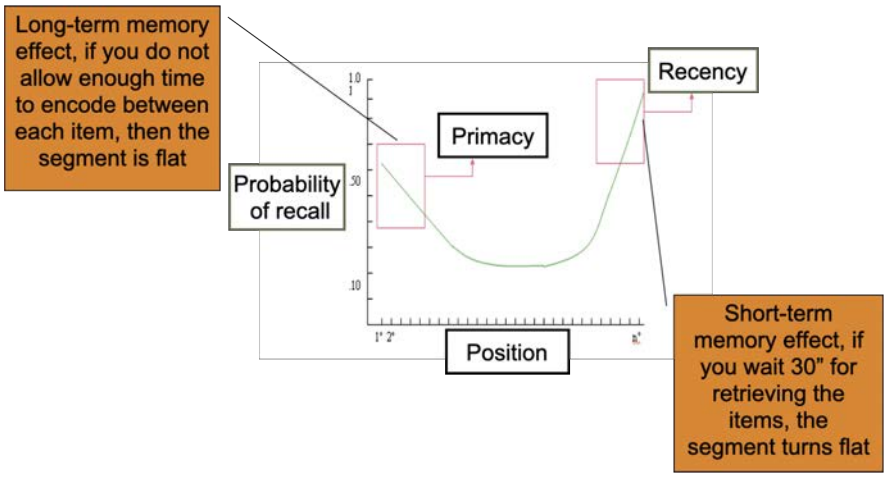
- Lerdahl & Jackendoff **"grouping well-formedness rules"** and **"grouping preference rules"** (Generative Theory of Tonal Music)
- *Well-formedness rules* define **(abstract) structural descriptions** that can be **derived from the surface structure** (a series of hypotheses for organizing the musical structure).
- ***Grouping preference rules*** define the conditions that allow a listener to choose the preferred interpretation of the structure from all of the possible ones that conform to the well-formedness rules (connected to real perceptual events).

8.10. Operations in STM: Chunking

- **Encoding or consolidation of small groups of elements into a compact larger or more abstract element, which is then encoded, recognized or remembered.**
- Example: Memorize this series of letters:
    - F-B-I-C-I-A-U-S-A-C-N-N-I-B-M
    - You create chunks of 3 letters that "go together" in an existing memory
    - **Musical scales and chords can be used as elements that facilitate chunking** (you usually do not code the individual notes)
- C-E-G-B -> C major 7
- Cadences
- **Similarity and relatedness between the sequential elements facilitates chunking**; not only a "top-down" (knowledge-based) effect

8.11. Serial position effect

Tendency of a person to recall the first (**primacy effect**: initial terms presented are most effectively stored in LTM) and last items (**recency effect**: items are still present in STM) in a series best, and the middle items worst.



8.12. Pitch-specific short-term memory

STM plays an important role when we listen to music. Without such memory, we would not be able to tell how the different tones in a phrase are related to each other. **Memory for the pitch of a tone is retained in a specialized memory system.** Deutsch experiments:

- Remarkable dissociation between tones and spoken words in memory. Play a tone, then another tone which is either the same in pitch or differs by a semitone → most people find it very easy to decide whether the 2 tones are the same or different, even when the tones are separated by a 5s delay.
- If there are extra notes in between: comparing test tones is very difficult: we retain pitches in a specialized memory store, that memory loss is caused by interference between pitches that

are held in this store (if instead of notes in between, a spoken sentence is in between, it is again much easier to decide).

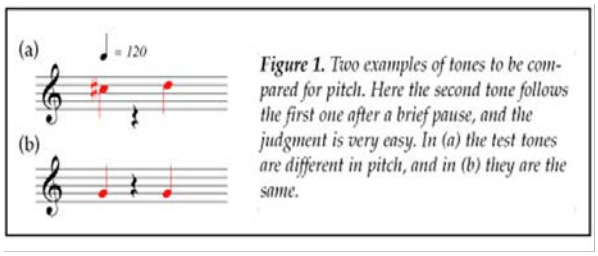


Figure 1. Two examples of tones to be compared for pitch. Here the second tone follows the first one after a brief pause, and the judgment is very easy. In (a) the test tones are different in pitch, and in (b) they are the same.

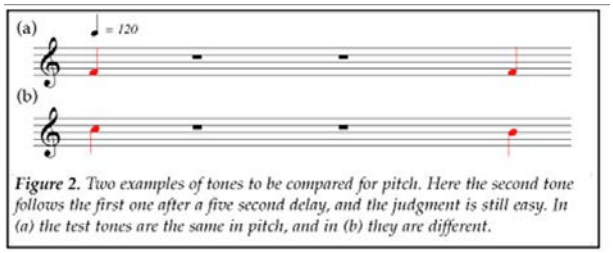
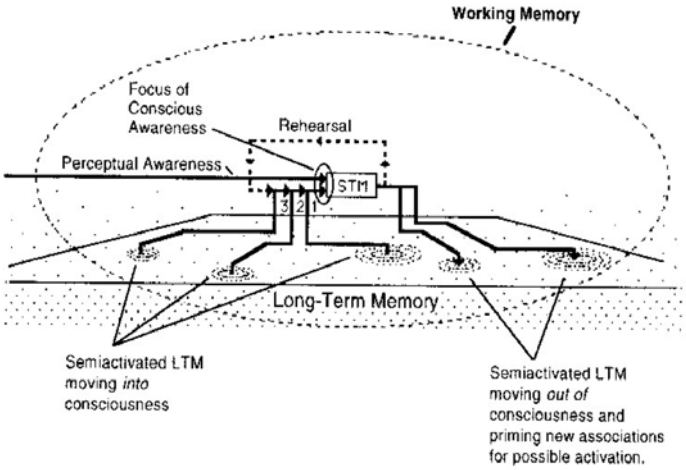
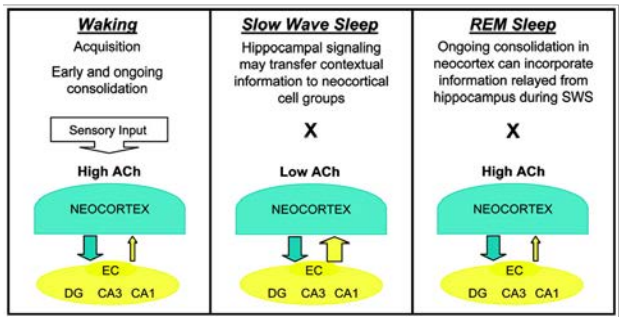


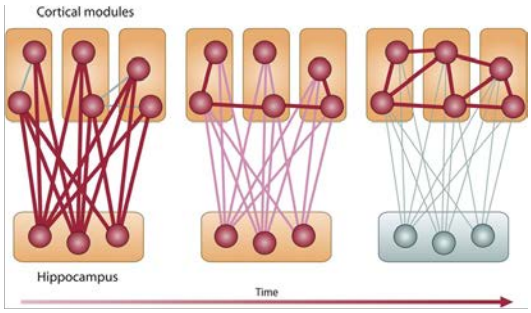
Figure 2. Two examples of tones to be compared for pitch. Here the second tone follows the first one after a five second delay, and the judgment is still easy. In (a) the test tones are the same in pitch, and in (b) they are different.

8.13. STM-LTM

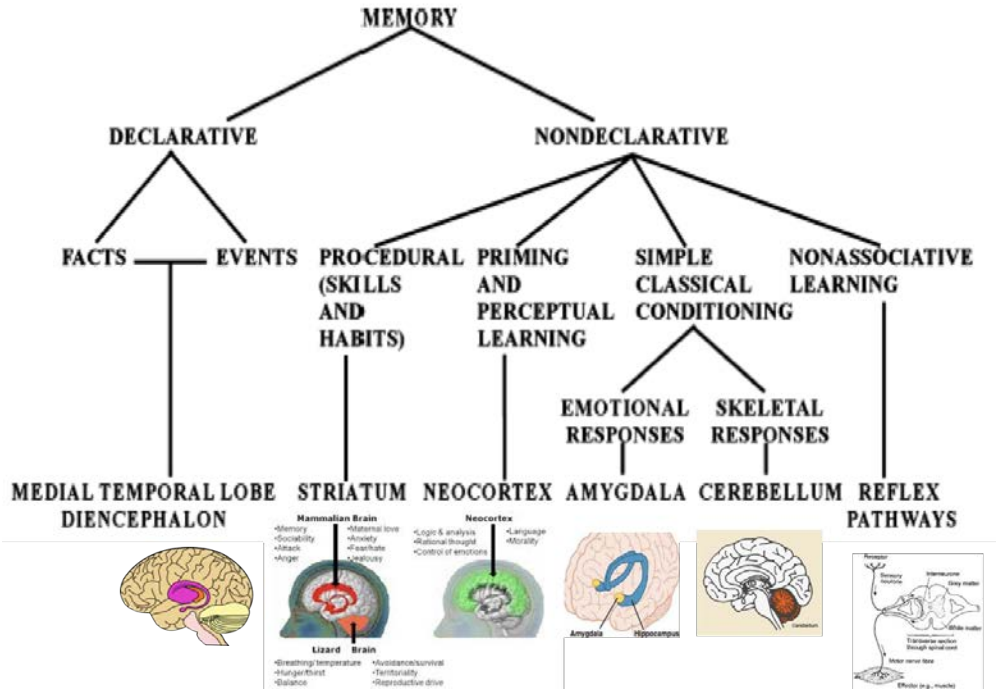


8.14. Memory consolidation





8.15. (Long-term) Memory systems of the brain

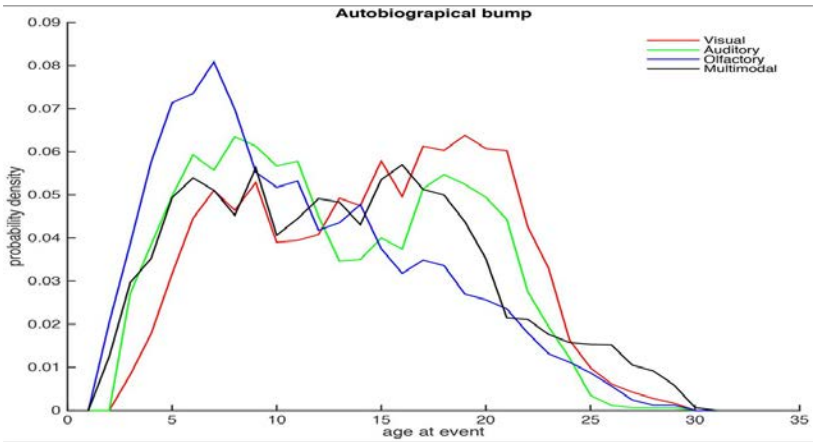


8.16. Declarative vs Non-Declarative Memories

- **Declarative** (**explicit** memory):
  - Consciously available
  - Fast learning, even "single-trial"
  - **Events (autobiographical memory) + Facts**
  - *What is this note name? What's the name of the piece?*
    - **Episodic memory**: memory for specific events in time (e.g. "that" performance of Don Giovanni that we saw in Vienna) → **autobiographical memory**
    - **Semantic memory:** memory about things of the world, common-sense (e.g. Don Giovanni is an Opera by Mozart)
  - Both are **associative** and **distributed**
  - Both are **reconstructive ("extra info",** not encoded in the original event, can be generated at retrieval time)
- **Non-Declarative** (**implicit memory or procedural memory**):
  - Contents consciously "unavailable" or interfering its development when trying to make them conscious
  - Slow learning
  - Automatic once learning has happened ("compiled knowledge")
  - Often modality-specific
  - *How does the melody go? How should I play this phrase with this instrument?*
  - **Procedural knowledge is slowly acquired, usually "by doing", it is very long-lasting.**
  - Usually related to "sequences" of sensory/motor representations (e.g. finger movements, notes, etc.) → **grammars** (they define "correct sequences")
  - Difficult to retrieve verbally, often causing interference

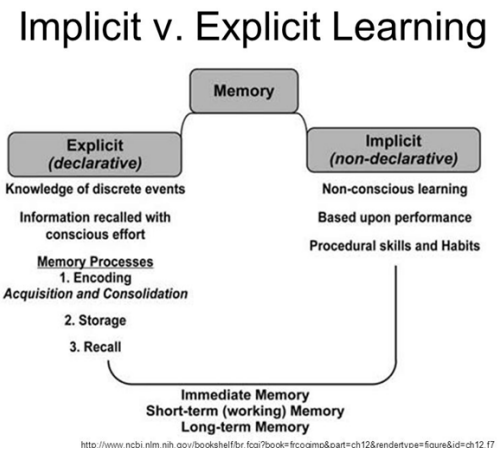
- After some amount of practice knowledge is “compiled” into big chunks that are not accessible to introspection.

8.17. Autobiographical memories



8.18. Implicit musical knowledge

- Western listeners, even those without musical instruction, have extensive knowledge of typical tonal and harmonic patterns (and melodic, rhythmic and metric patterns too): **tonal hierarchies (i.e. “key”)**.
- **Statistical regularities are extracted from musical input by the learning brain provided proper and frequent exposure** (the cues for the tonal hierarchy -the key- are present in the surface details of a melody -duration and frequency of occurrence of pitches--)
- Knowledgeable listeners, even from another culture, can pick up on those cues, provided a certain amount of exposure.
- 30% explicit knowledge (verbalized) and 70% tacit knowledge (non verbalized).



8.19. Types of retrieval

Retrieval operations in decoding depth order (from less to more):

- **Recognition**: acknowledgement that a pattern in STM is stored in LTM (did you hear this song before? Does it sound familiar?) Recognition may happen without further recall or reminding. For music we have some “absolute memory”: we are able to recognize familiar songs just by listening to a few seconds or even milliseconds.
- **Recall or Recollection**: activation of a LTM encoded pattern by a “diffuse” effort of will (what is the title of this song? Who was the composer?)
- **Reminding**: activation of a LTM encoded pattern by a pattern in STM (which memories are activated by listening to this song?)

8.20. Phrasing

The way notes are joined/separated/played into a meaningful “unit”, which separates from the previous one and from the next one. It is a musical resource to **help segmentation in STM**.

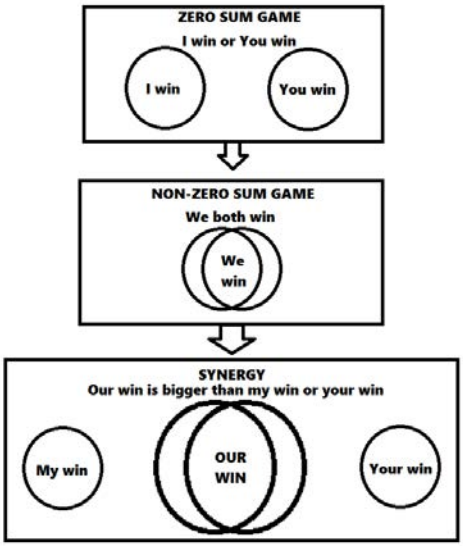
8.21. Themes, variations and motives

Musical strategies to **increase the memorability and association** of musical materials.

9. The when/what hierarchies (making sense of sonic patterns)

9.1. The game of music

- **Music is Non-zero sum infinite game**: All players may win
- The players are the listeners vs the composer/performer
- **Goal**: Anticipating what will happen, achieve a good prediction. The user is able to make some “errors” in the prediction. In a positive timestamp they will get a **reward** (game), but **listeners are not winning all the time**. This keeps the game interesting.



Pierre Boulez: Serialism

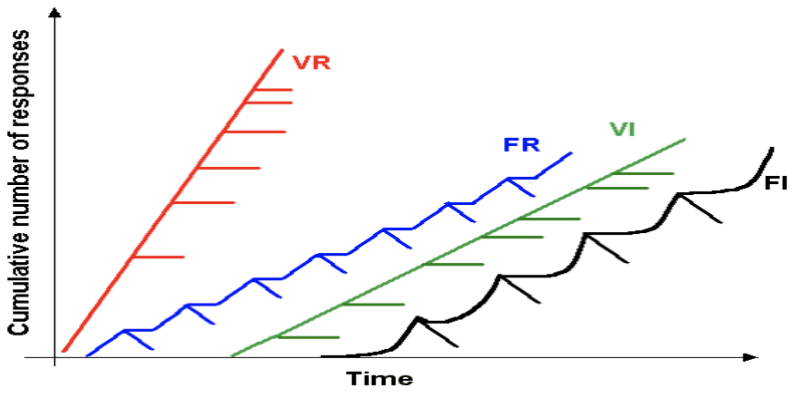
- Lack from typical harmonies
- We can't tap along
- No intention to convey a method

The when/what:

- Making sense of sonic patterns.
- Meter requires that our brain constructs a **hierarchy of weights**.

Music as a variable-ratio reinforcement system.

Curves obtained using different rewards:



VR-> Variable ratio. Type of **reward** (cannot be totally anticipated)
Coin Machines used that.

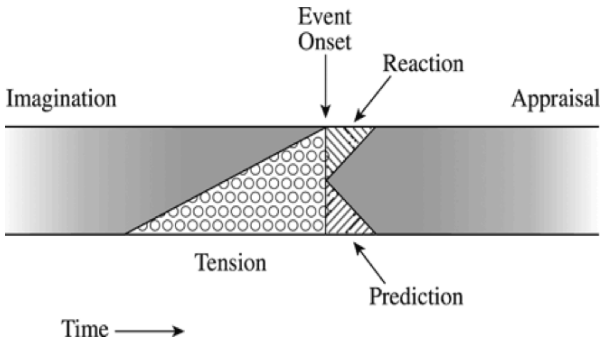
9.2. David Huron's ITPRA

He proposed very abstract theories that can provide useful insights.
When we are reacting to music, there are **4 main stages**:

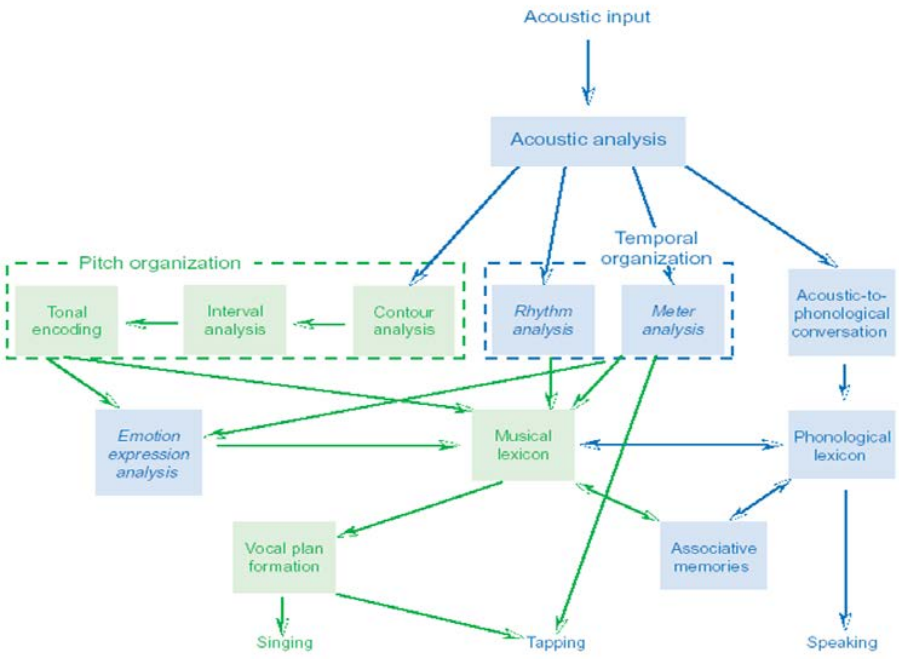
1. Imagining what is going to happen…
2. Creation of a certain amount of tension

3. Predicting what is going to happen and..
4. Reacting and appraisal (evaluate, assess)

Illustration of the time course for Huron's **ITPRA theory of expectation**. Feeling states are first activated by imagining different outcomes (I). As an anticipated event approaches, physiological arousal typically increases, often leading to a feeling of increasing tension (T). Once the event has happened, some feelings are immediately evoked related to whether one's predictions were borne out (P). In addition, a fast reaction response is activated based on a very cursory and conservative assessment of the situation (R). Finally, feeling states are evoked that represent a less hasty appraisal of the outcome (A):



9.3. A graphical model of some involved processes:

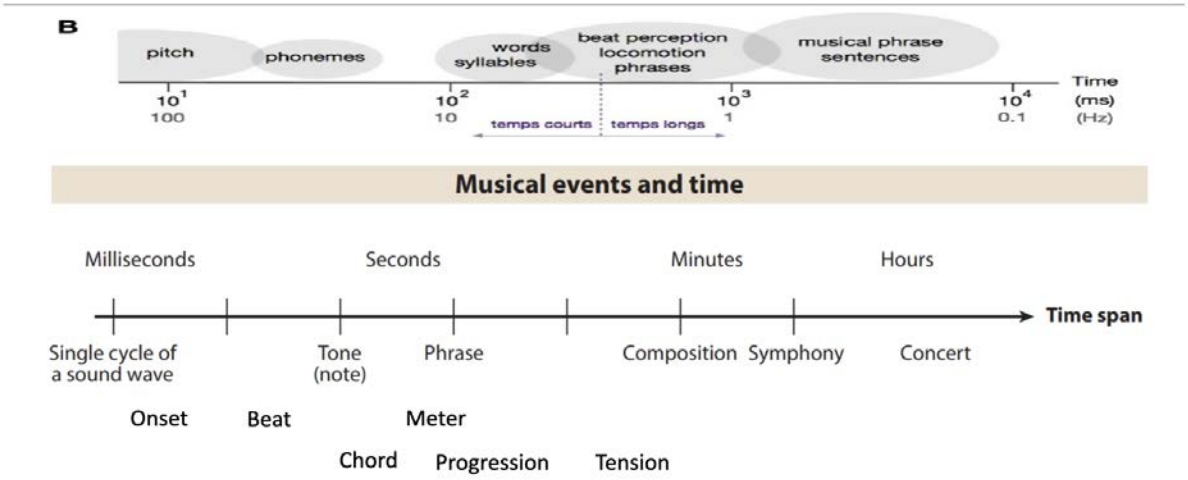


Contour analysis: Done by our brain. Once we have the analysis done, the brain goes up and down, in order to do an **interval analysis** and then it encodes the **melody**. **Encoding of tonal content and linguistic aspects** is necessary so that we can interpret music emotionally.

9.4. What/when frameworks:

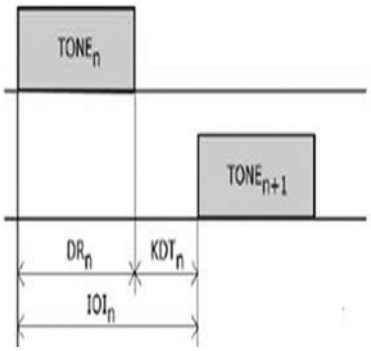
- They "frame" our **listening experience** (**"when" can we expect something**, **"what" to be expected**). How the NOW relates to the previous moments.
- **Rhythm hierarchies (when)**: pulse, tempo, meter make certain "moments" more important than others. "When" is related to: rhythm, temporal organization...
- **Tonal hierarchies (what)**: scales, keys, modes make certain "notes" more important than others. "What" is related to: tonalities, pitches, harmonies...

9.5. Temporal scales for music events:



9.6. IOI: Inter Onset Interval

- IOI: **time between the onset of successive events.**
- What is an onset? Changes in loudness, pitch or timbre
- What is an event? Whatever happens between two onsets or between an onset and offset
- **Perceived IOI differs from physical IOI** (e.g. lengthened when playing staccato –the sensed event lasts longer than the physical note-, visual info can bias our sensation, reverb…)



- **Simple relationships** (1:1, 1:2, 1:3) are usually found in music.

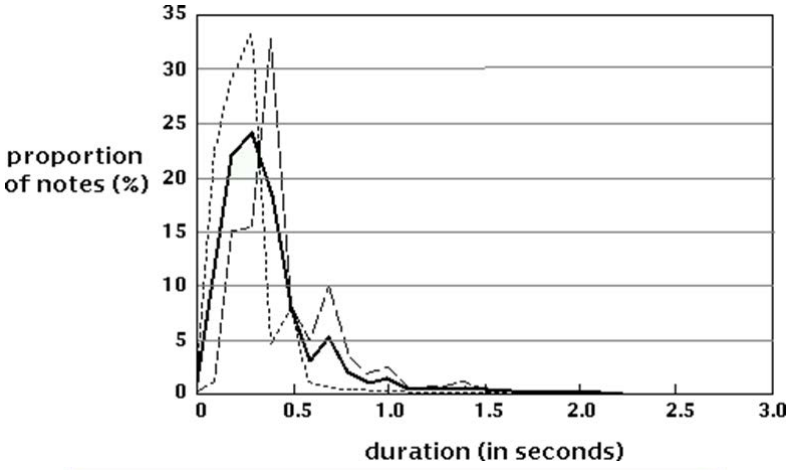
- **Simple ratios help our memory and production processes**. When asked to produce arrhythmic patterns, subjects produce near-equal IOI (isochrony: is the postulated rhythmic division of time into equal portions by a language).
- **Rhythmic sequences usually combine short times** (200-300 ms) and **long times** (450-800 ms).
- **Short-term Memory constraints**

9.7. Musical Speed Limits

- A Melody can be **too fast**, **too slow** or **understandable**

9.8. Distribution of note durations

Distribution of note durations in 52 instrumental and vocal works. Dotted line: note durations for the combined upper and lower voices from Bach's two-part Inventions. Dashed line: note durations in 38 songs (vocal lines only) by Stephen Foster. Solid line: mean distribution for both samples (equally weighted):



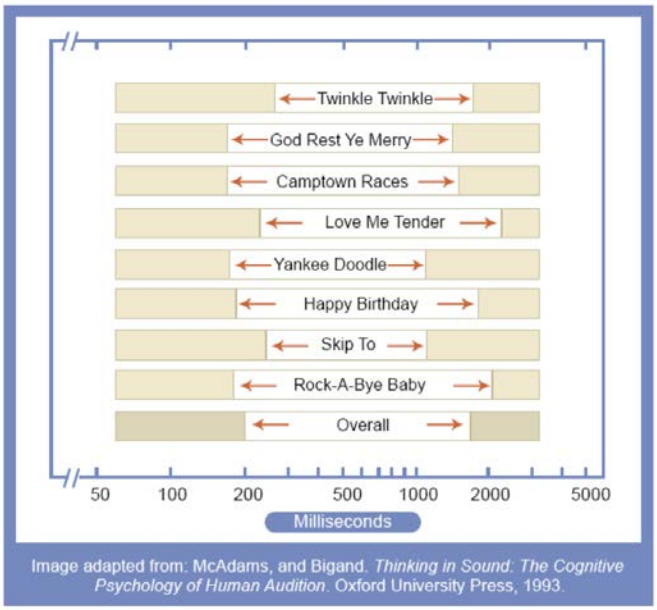


Image adapted from: McAdams, and Bigand. *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford University Press, 1993.

9.9. Beat/Tempo:

- **Beat**: a series of **regularly recurring, precisely equivalent stimuli** (Cooper and Meyer, 1960). A **chain of events, roughly equally spaced in time** (Parncutt, 1987).

- Many basic activities with periods between 500ms and 1s (baby sucking, heart beating, walking...)
- Beat induction is “Universal” and favors “entertainment” (**being in-synch**)
- Is beat to be found in the IOI’s ?

- **Tempo**: count of **beats** per time unit.
  - It can range from 30 to 600 bpm (equivalently: IOI between 100ms to 2 seconds)
  - *Preferred tempo*: listeners adjustment to make music more “natural” (average: 100bpm). Most popular music is set around this value and range.

9.10. Beat induction:

- Beat induction is favored when beat:
  - **coincides with notes onsets**
  - **coincides with longer notes**
  - is **regular**
  - **aligns with beginning of musical phrases**
  - **aligns with points of harmonic change**
  - **aligns with onsets of repeating melodic patterns**.

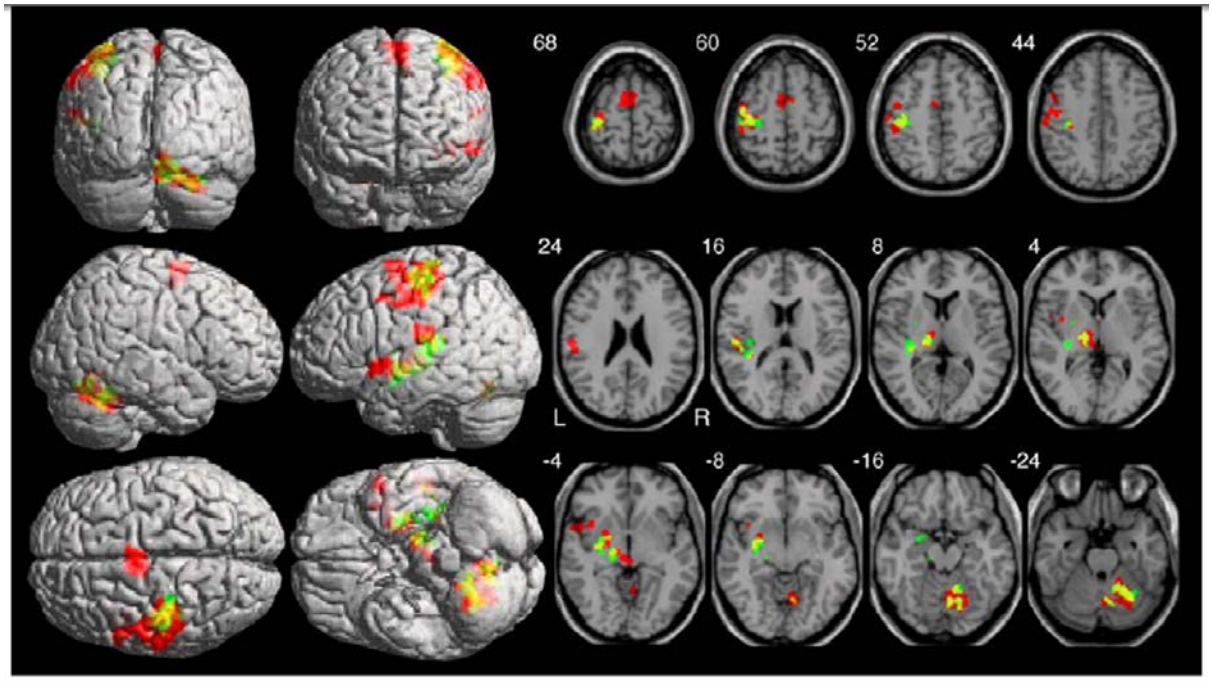


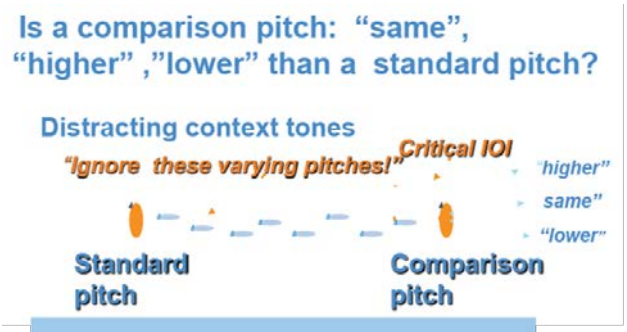
Figure 3. Brain activation during paced tapping. Foci of activation for isorhythmic-only tapping (green label), polyrhythmic-only tapping (red), and overlap (yellow) depicted on rendered projection images (left) and selected axial slice (right). Isorhythmic tapping activated M1, S1, and PMA, and the temporal operculum in contralateral neocortex as well as the basal ganglia and cerebellum. Polyrhythmic tapping activated the same structures, as well as the supramarginal gyrus, SMA, preSMA, cingulate cortex, and the middle and superior temporal gyri. Note greater extent of activation for polyrhythmic then isorhythmic tapping in commonly activated areas, and new areas of activation mostly for polyrhythmic tapping. **Numbers next to upper left of each horizontal brain slice refer z-axis in MNI-Talairach space. Additional details in text; full reporting of the activated areas appears in Table 1.**
doi:10.1371/journal.pone.0002312.g003

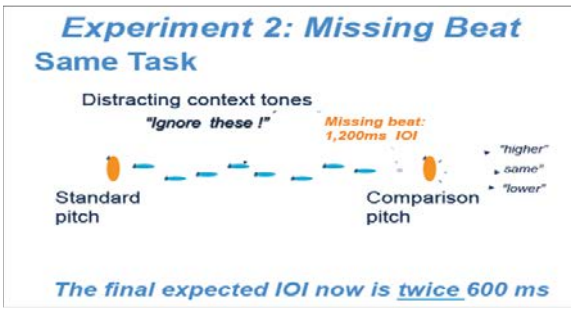
9.11. Rhythm hierarchies / beat induction:

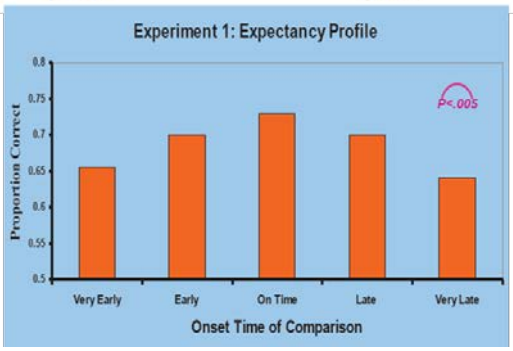
- Experiment in class about tapping a melody without the drum:
  - General agreement about most appropriate times to tap or clap

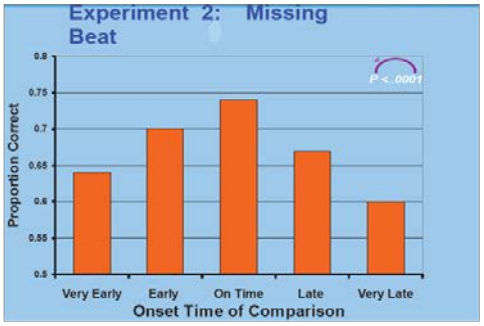
- Different subjects give different responses, even though there is a “central tendency”, so different solutions are “good!”.
- Different patterns => different degree of difficulty
- Unconscious mental processes are involved in deciding when to tap

9.12. Beat as a patterning device framing our attention and expectations:

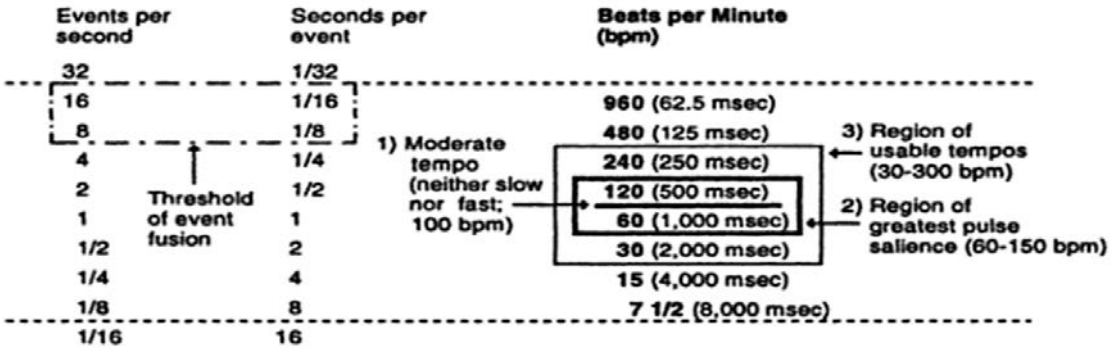




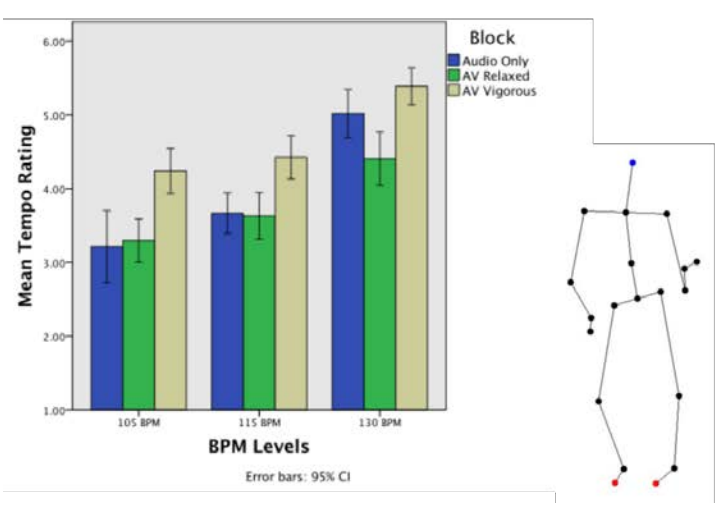


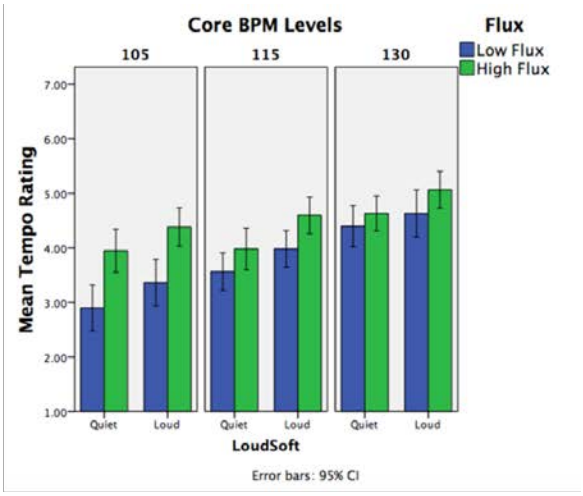


9.13. Time(s) and tempo:



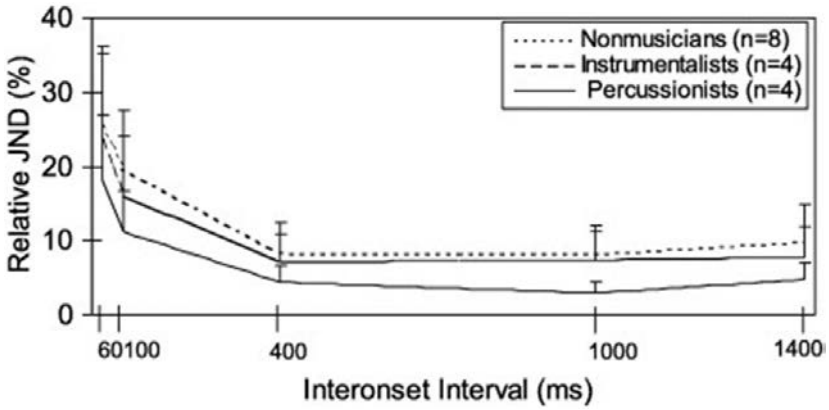
9.14. Factors affecting tempo estimation:





9.15. Tempo JND:

- At 100bpm a listener can discriminate a change of about 10% in spacing (60 ms or ~10bpm).

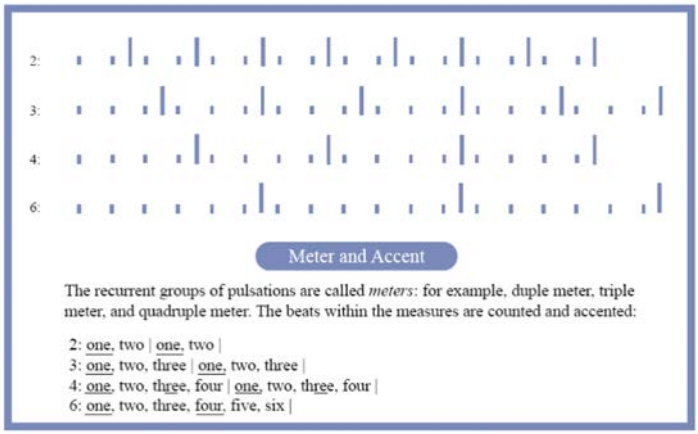


9.16. Accent:

- A stimulus (in a series of stimulus) which is marked for consciousness in some way (Cooper and Meyer, 1960). A perceivable element that increases the perceptual salience of an event.
- How can we add accent? Accent == contrast
    - Lengthening IOIs
    - Increasing intensity (dynamics)
    - Changing articulation
    - Timbre variation
    - Changing direction of melodic contour (from ASC to DESC)
    - Big Leaps (leaps == saltos)
    - Modulating (to dissonant)
    - Deviating from regularity (in timing, in pitch)

9.17. Accent and meter:

Beats can be weak or strong (accented), this leads to meter.



9.18. Influences of native language on segmentation/meter strategies:

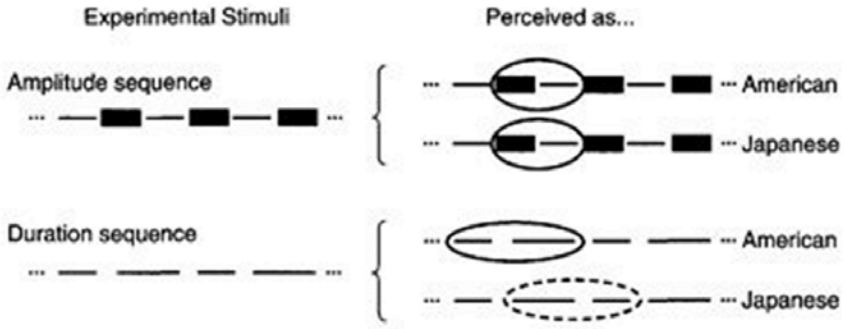
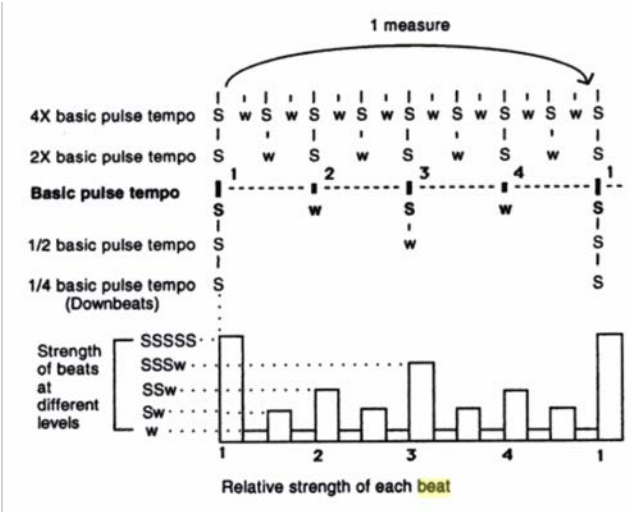


Figure 3.19 *Left side:* Schematic of sound sequences used in the perception experiment. These sequences consist of tones alternating in loudness ("amplitude sequence," top), or duration ("duration sequence," bottom). In the amplitude sequence, thin bars correspond to softer sounds and thick bars correspond to louder sounds. In the duration sequence, short bars correspond to briefer sounds and long bars correspond to longer sounds. The dots before and after the sequences indicate that only an excerpt of a longer sequence of alternating tones is shown. *Right side:* Perceived rhythmic grouping by American and Japanese listeners, indicated by ovals. Solid black ovals indicate prefer-

9.19. Metrical hierarchy:

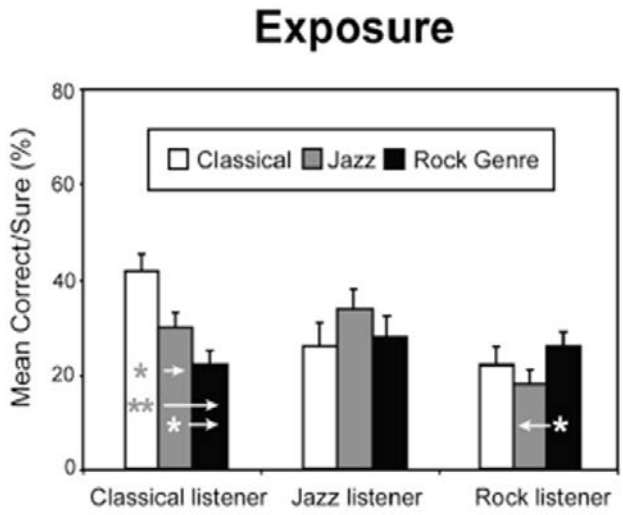
- **The weaker a beat, the greater the metrical tension when placing an accented event on it**. (We expect it to be placed in the strong one).



9.20. Expressive timing & expectation:

- **Timing**: temporal microstructure characteristic of a **music performance** (is the outcome of the performer's interpretation.)
- **Deviations from a strictly uniform pulse that occur in live performance**. These deviations most **commonly occur near the ends of phrases and other grouping units.**
- At **phrase boundaries listeners are most sensitive to timing alterations.**

9.21. Exposure-dependent sensitivity to timing alterations:

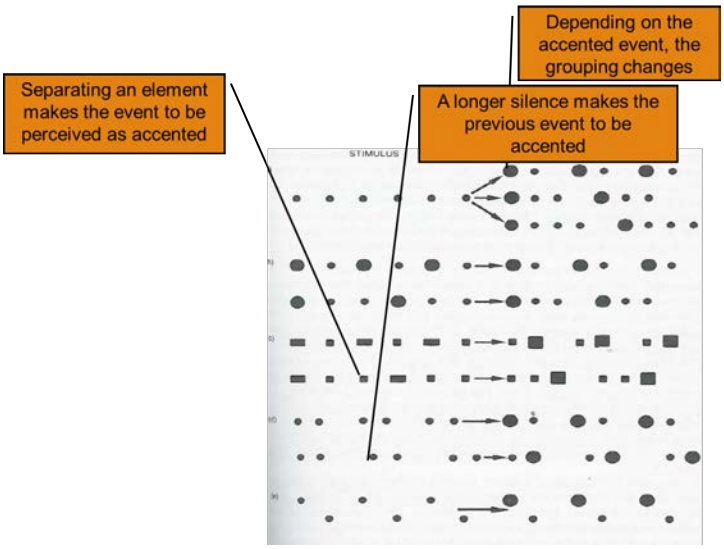


Honing, H. & Ladinig, O. (in press, 2008). Exposure influences timing judgments in music. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5).

Honing, H. (2008). Musical competence and the role of exposure. *Proceedings of the Music and Language II Conference*, Boston: Tufts University.

9.22. Rhythmic pattern:

- **Accent causes grouping which determines periodicities and then a perceived rhythmic pattern**
- **Rhythm is a perceptual attribute** (it "emerges" as a combination of bottom-up & top-down processes)
- **Rhythm=grouping+meter**



9.23. 3 dimensions of rhythmic semantic experience:

3 dimensions of rhythmic semantic experience:

- **Cognitive-structural**: rhythms have meter, accents, clarity and complexity…
- **Movement**: rhythms can be fast-slow, dense or sparse (number of events per time unit), and can be related to human motoric activities (walking, knocking, dancing, jumping…)
- **Emotional**: rhythms can be playful, solemn, rigid, excited, calm…

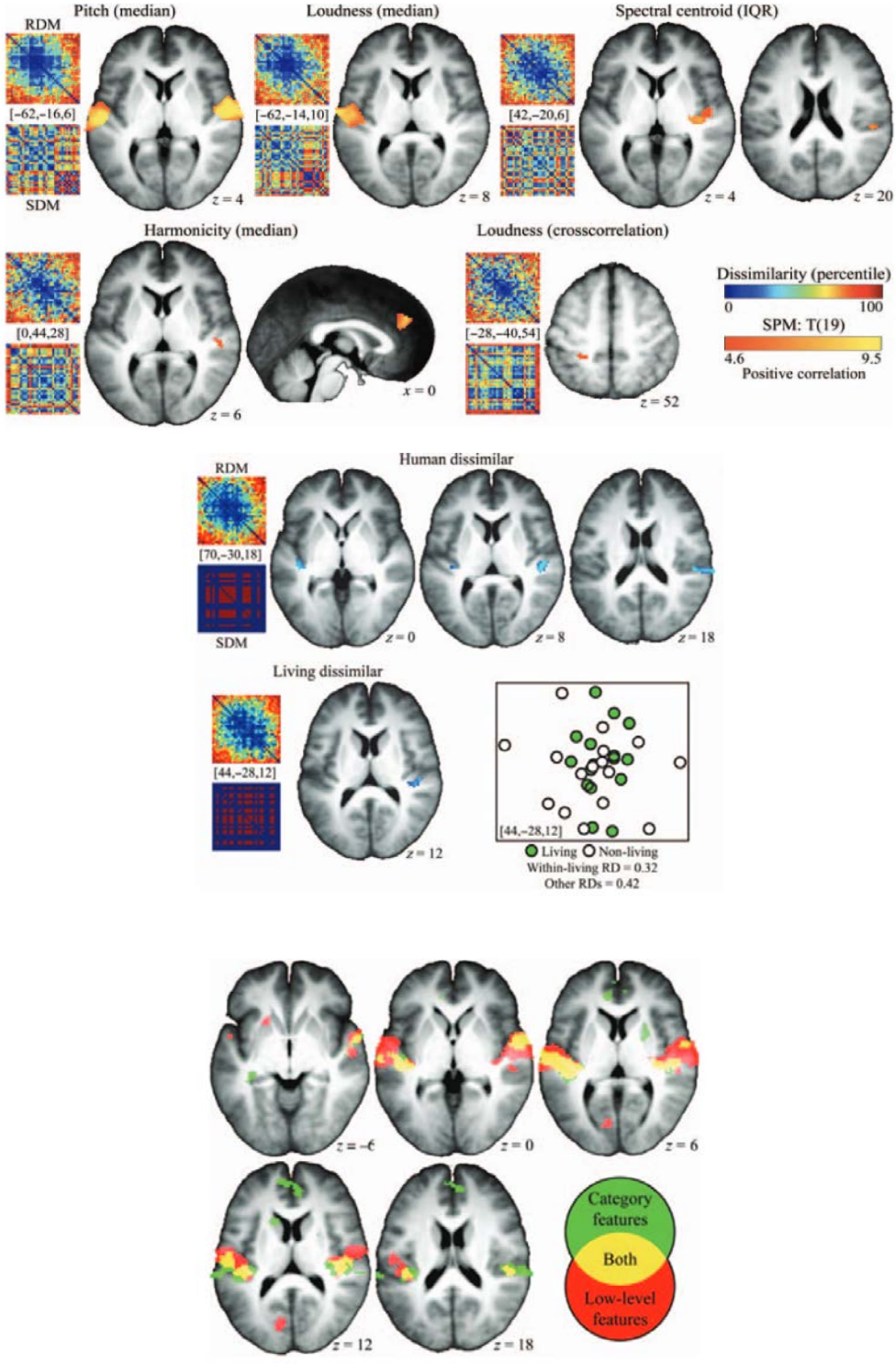
9.24. The "what":

Knowledge structures: **Categories**

- **A group of nonidentical objects or events that an individual treats as equivalent**.
- Equivalent could mean:
    - Their internal representations are similar or close
    - They generate similar behaviors (avoidance / approach)
    - They generate similar internal states (e.g. emotions) (pleasure/pain)
- **Categorization reduces the overwhelming complexity of the natural world** (frequencies → notes, spectro-temporal profiles → instruments)
- "Static" knowledge structure for representing facts and hierarchical relationships between them.

9.25. Sensory qualities and categories in the brain:

Living vs non-living, human vs non-human, also music vs speech:



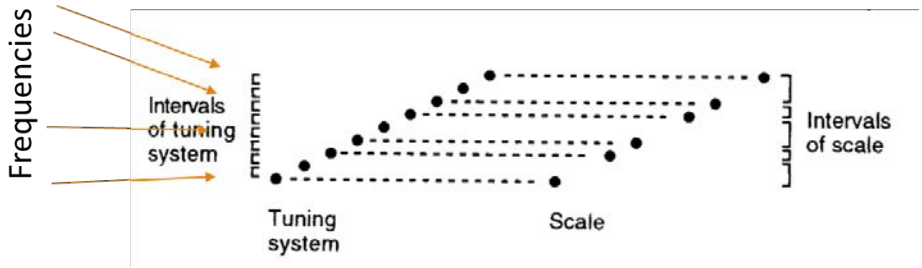
9.26. Explicit musical categories ?

- Note (vs silence)
- Note (C, C#,...)
- Dynamics (pianissimo,...)
- Instrument (Flute, string section, brass)
- Rhythm pattern(reggae, swing,...)
- Duration (black, quaver)

- Key (A, atonal, ….)
- Mode (Major, minor, ….)
- Chord (C Maj, DMin7th)
- Genre (Be-bop, Gregorian, Italian Opera, Thrash-metal...)
- Expression (Frullato, sforzando)

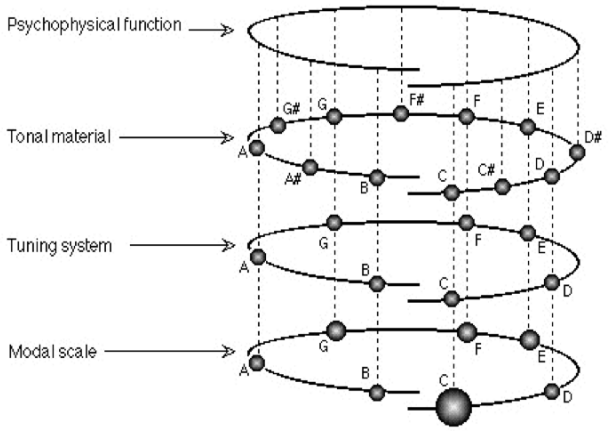
9.27. Melodic categories: tuning systems, scales, pitches

- **Tuning systems**: reduce the variety of audible frequencies into a small number of classes to be discriminated (preference of simple ratios: ⅔, 3/4, ⅘,...)
- **Scales**: subpopulations of a tuning system that **allow pitch information to be managed under the constraints of our memories** (<=7 pitches considered).
- Scales provide a framework to admit "pitch nuances" and small mistunings as instances of the learned categories.



9.28. Representation of melodic knowledge:

- **Psychophysical function** (frequencies → pitches)
- **Pitches** (all pitches usable in a musical culture)
- **Tuning system** ("Available" pitches for building melodies)
- **Modal Scale** (some pitches are assigned roles, weighted with different importance, etc.)
- **Melodic schema** (in a given context, we expect or anticipate the most likely melodic evolution)



9.29. From harmonics to notes and intervals in Western music:

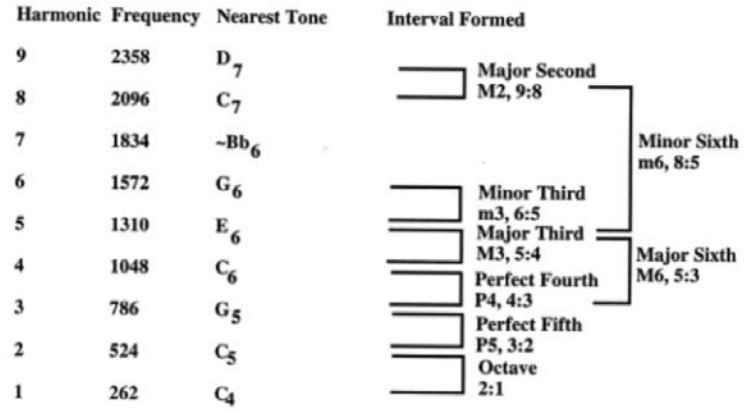
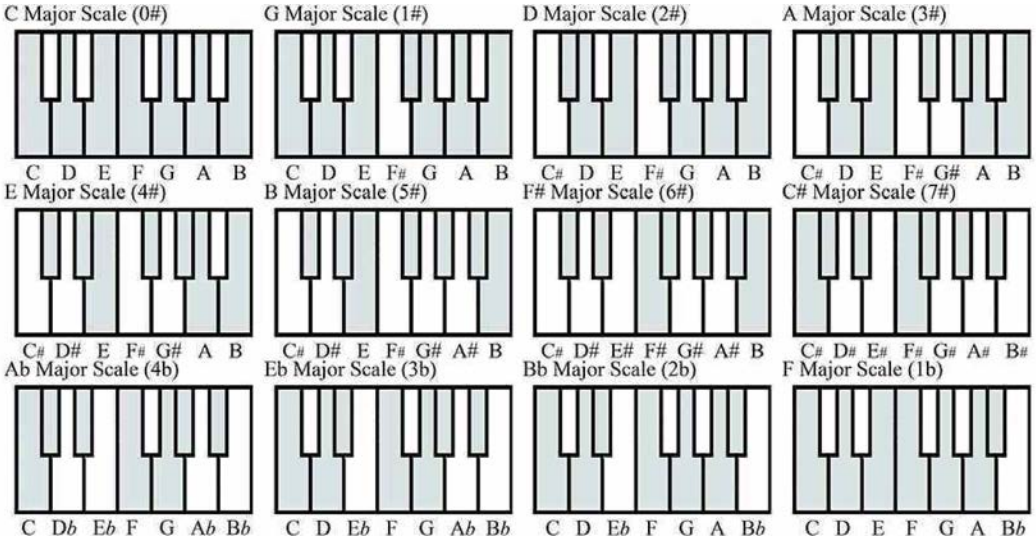


Fig. 2. The first nine harmonics of middle C. The frequencies and nearest tones are indicated, as well as intervals described as elements of Western tonal-harmonic music.

9.30. Tuning:

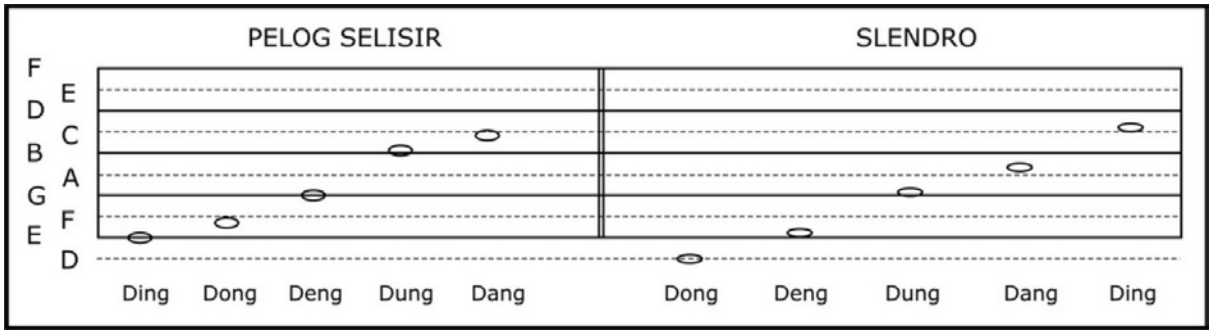
	EQUAL TEMPERAMENT		PYTHAGOREAN		JUST		MEAN–TONE	
	Ratio	Cents	Ratio	Cents	Ratio	Cents	Ratio	Cents
Unison	1:1	0	1:1	0	1:1	0	1:1	0
Aug unis					1:1.055	92	1:1.045	76
Mi 2nd	1:1.059	100	1:1.053	90	1:1.067	112	1:1.070	117
Maj 2nd	1:1.122	200	1:1.125	204	1:1.125	204	1:1.118	193
Mi 3rd	1:1.189	300	1:1.185	294	1:1.200	316	1:1.196	310
Maj 3rd	1:1.260	400	1:1.265	408	1:1.250	386	1:1.250	386
P 4th	1:1.335	500	1:1.333	498	1:1.333	498	1:1.337	503
Tritone	1:1.414	600						
Aug 4th			1:1.404	588	1:1.406	590	1:1.398	580
Dim 5th			1:1.424	612	1:1.422	610		
P 5th	1:1.498	700	1:1.500	702	1:1.500	702	1:1.496	697
Mi 6th	1:1.587	800	1:1.580	792	1:1.600	814	1:1.600	814
Maj 6th	1:1.682	900	1:1.687	906	1:1.667	884	1:1.672	890
Aug 6th					1:1.778	996	1:1.747	966
Mi 7th	1:1.782	1000	1:1.778	996	1:1.800	1018	1:1.789	1007
Maj 7th	1:1.888	1100	1:1.898	1109	1:1.875	1088	1:1.869	1083
Octave	1:2	1200	1:2	1200	1:2	1200	1:2	1200

9.31. Some scales in Western music:

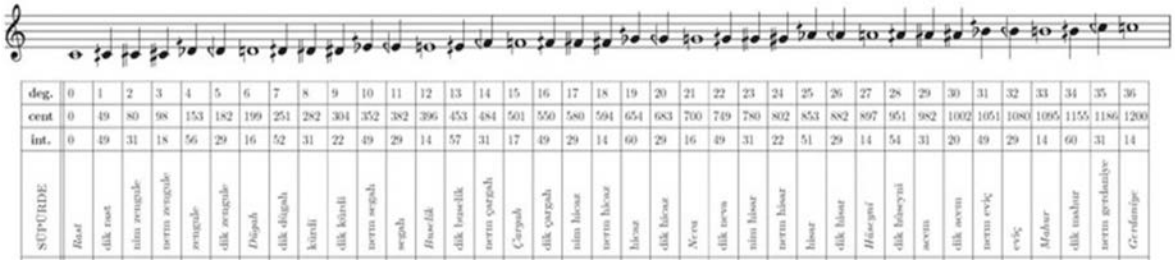


9.32. Scales in Gamelan music:

Scales in **gamelan** music (sounds from Indonesia, predominantly percussion, but also some xylophones, bamboo flutes amongst others):



9.33. Scales (Makams) in Turkish music:



9.34. Scale structure:

- **Scale**: 7 tones (degrees) per octave with **asymmetric pattern of pitch spacing** (intervals) between them.
- The different tones take on different roles in the fabric of music, with the tone being the most central and stable (the tonic).
- Diverse musical traditions make use of a tonic:
    - Utility of psychological reference points in organizing mental categories

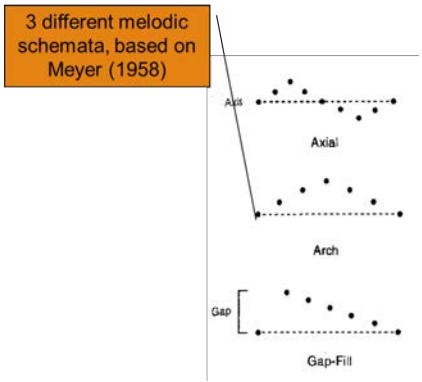
- **Listeners are very sensitive to scale structure**
- **Sour notes are very salient**.
- Trainor & Trehub (1992) experiment:
    - Non-musician adults were sensitive to notes in a melody that were changed by one semitone (outside of the scale) compared to four semitones (inside the scale), but 8-month old infants were not.
- Trainor & Trehub (1994):
    - 5 year-old children with no formal training in music were sensitive to out-of-scale alterations.

Contour: the pattern of successive pitch changes within a melody defines its contour (what is important is the **direction of pitch changes**, rather than the extent of the change)
Transposition: Frequency scaling.

9.35. Melodic schemata:

Melodic Schemata: (3 types: **Axial**, **Arch**, **Gap-fill**), based on Meyer (1958)

- **Schemas are dynamic knowledge structures.**
- **Some of them related to physical or visual concepts**
- Help to recognise and to code a series of events or objects
- Help to predict possible next musical notes, directions of contour, etc…
- **Tuning systems and scales do not include temporal information, schemata does.**
- **Tonality is one of the most powerful music schemas as it includes, implicitly, temporal dependencies derived from the sequential pattern of notes.**



9.36. Tonal hierarchies:

- By being exposed to organized musical input **our brain "implicitly" learns the statistics of pitch occurrences (either in melodies as in chords)**
- The cues for the **tonal hierarchy (the key)** are present in the surface details of a melody – **duration and frequency of occurrence of pitches**

9.37. Statistical learning:

- Is it enough with transition matrices of order 1 (i.e., considering the previous note)?

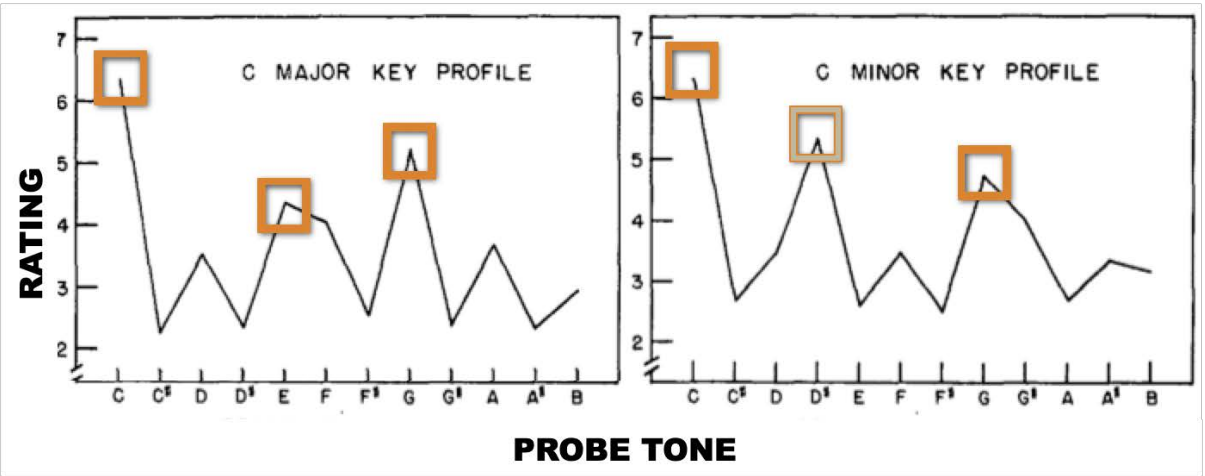
  Are we "markovian learners"?

	do	do#	re	re#	mi	fa	fa#	so	so#	la	la#	ti
do	26.42	0.06	21.70	0.17	15.27	1.62	0.10	10.22	0.00	6.49	0.00	17.95
do#	5.88	11.76	61.76	8.82	0.00	0.00	0.00	0.00	0.00	5.88	0.00	5.88
re	33.53	0.26	21.06	0.55	26.26	5.43	0.00	6.60	0.00	1.61	0.00	4.69
re#	12.50	0.00	45.00	0.00	8.33	30.00	0.00	4.17	0.00	0.00	0.00	0.00
mi	10.38	0.03	32.47	0.03	20.97	17.65	0.59	15.79	0.02	1.88	0.00	0.19
fa	0.58	0.02	13.51	0.90	44.26	16.15	0.04	18.36	0.03	4.73	0.06	1.34
fa#	1.57	0.00	8.38	0.00	19.37	5.24	20.94	20.94	1.05	20.94	0.00	1.57
so	14.87	0.01	3.08	0.22	16.60	21.25	1.20	28.12	0.11	12.08	0.31	2.15
so#	2.70	0.00	0.00	0.00	2.70	5.41	2.70	29.73	5.41	37.84	8.11	5.41
la	3.55	0.06	2.51	0.00	0.97	5.11	0.55	54.37	0.25	18.83	1.05	12.75
la#	21.60	0.00	1.05	2.79	0.35	1.05	0.00	14.98	0.00	41.46	16.72	0.00
ti	33.59	0.00	8.46	0.00	0.58	0.48	0.17	5.36	21.91	22.01	0.02	7.43

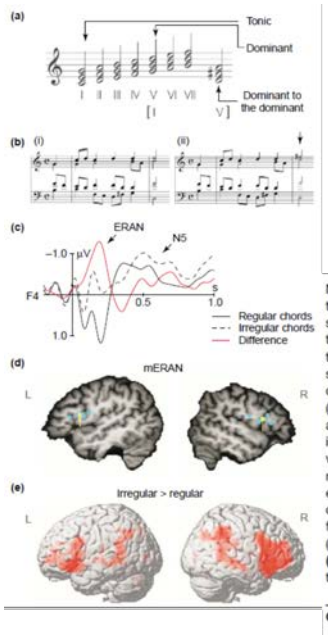
Likelihood that a pitch on the y axis is followed by a pitch on the x axis in a sample of 50,000 notes of German folk music.

9.38. Western Tonal Hierarchies:

- Krumhansl & Kessler (1982) experiment:
  - Certain notes are sort of more important than others (first level: Tonic; second level: Dominant & Mediant; etc.)
  - Key profiles: play a sequence setting a tonal context, play a probe tone (does it fit or not?)
  - Listeners may notice “in-scale” vs. “out-of-scale” pitches
  - Connection with Harmonic Pitch Class Profiles (HPCP) in MIR



9.39. Music Syntax processing in the brain:



Neural correlates of music-syntactic processing. **(a)** In major–minor tonal music, chord functions are arranged within harmonic sequences according to certain regularities. Chord functions are the chords built on the tones of a scale. The chord on the first scale tone, for example, is denoted as the tonic and the chord on the fifth scale tone as the dominant. The major chord on the second tone of a major scale can be interpreted as the dominant to the dominant (square brackets). **(b)** One example for a regularity-based arrangement of chord functions is that the dominant–tonic progression is a prominent marker for the end of a harmonic sequence, whereas a tonic–dominant progression is unacceptable as a marker of the end of a harmonic sequence. **(i)** The sequence shown ends on a regular dominant–tonic progression, **(ii)** the final chord of this sequence is a dominant to the dominant (arrow). This chord function is irregular, especially at the end of a harmonic progression (sound examples are available at www.stefan-koelsch.de/TC_DD). **(c)** Electric brain potentials (in μV) elicited by the final chords of the two sequence types presented in b (recorded from a right-frontal electrode site [F4] from twelve subjects). Both sequence types were presented in pseudorandom order with equal probability in all twelve major keys. Brain responses to irregular chords clearly differ from those to regular chords. The first difference between the two black waveforms is maximal at about 0.2 s after the onset of the chord (this is best seen in the red difference wave, which represents regular subtracted from irregular chords) and has a right-frontal preponderance. This early right anterior negativity (ERAN) is usually followed by a later negativity, the N5 (short arrow). **(d)** With MEG, the magnetic equivalent of the ERAN was localized to the inferior frontolateral cortex (adapted from Maess *et al.* with permission of Nature Publishing Group [http://www.nature.com/] [8]; single-subject dipole solutions are indicated by blue disks, yellow dipoles indicate the grand-average of these source reconstructions). **(e)** fMRI data obtained from 20 subjects using a similar chord-sequence paradigm (the statistical parametric maps show areas that are more strongly activated during the processing of irregular than during the processing of regular chords). Corroborating the MEG data, the fMRI data indicate activations of IFLC. Additionally, the fMRI data indicate activations of the ventrolateral premotor cortex, the anterior portion of the STG, and posterior temporal lobe structures.

Current Opinion in Neurobiology 2005, **15**:1–6

10. Music and Emotion

10.1. What is an emotion (I)?

"everyone knows what an emotion is, until asked to give a definition" (Fehr & Russell, 1984)

- Emotion is both an everyday concept and a scientific construct.
- Emotion as opposed to reason: Descartes' error (Damasio 1994)
- Emotions have **survival value**, they are not "bells and whistles" of behavior
- They are linked to survival issues: danger, competition, loss and cooperation.
- [Wikipedia]: emotions are **psychological states** brought on by **neurophysiological changes**, variously associated with **thoughts**, **feelings**, **behavioral responses**, and a degree of pleasure or displeasure. There is currently no scientific consensus on a definition. Emotions are often intertwined with mood, temperament, personality, disposition or creativity. PET scans and fMRI scans help study the affective picture processes in the brain. Emotions produce different physiological, behavioral and cognitive changes.

[PERFECTO QUIZ CARD] **What is an emotion?**

Hard to define and explain, no precise agreed upon definition, agreed upon characteristics.

- Relatively brief, intense, rapidly changing responses to potentially important events in the external or internal environment.
- Usually of social nature

- Response to internal or external events
- Related "synchronized" subcomponents: cognitive changes, subjective feelings, expressive behavior, action tendencies.

[PERFECTO QUIZ CARD] **Perceived vs. Induced Emotion**
In music, perceived = inferred, induced = evoked.

[PERFECTO QUIZ CARD] **Five Basic Emotions**
Anger, Fear, Happiness, Sadness, Disgust

[PERFECTO QUIZ CARD] **Complex Emotions**
Combinations of basic, also more abstract things like nostalgia.

[PERFECTO QUIZ CARD] **Mood vs. Emotion**
Mood:

- Less intense
- Lasts longer & is more stable
- Less likely to be provoked
- Influenced by environment, physiology and attention

Emotion:

- More intense (strong and fast)
- Shorter in duration, clear boundaries
- Response to particular stimulus
- Usually creates action impulse

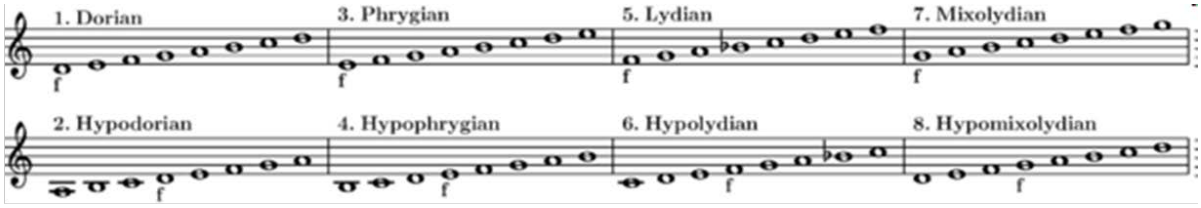
[PERFECTO QUIZ CARD] **How to measure emotion**

- Self report (explicit)
- Indirect measures (implicit): reaction time test, etc.
- Expressive behavior, e.g.: facial expression, posture, movement
- Physiological response: GSR (Galvanic Skin Response), heart rate, etc.

10.2. Communicating vs. generating emotions

- Music is a **communicative device** (as speech): is this music happy/sad/aggressive/solemn?
- Music is a **brain changer** (as drugs): do you feel happy/sad/aggressive/tender when listening to X music?
- Do you always feel sad when you listen to sad music? Feel aggressive when listening to such kind of music?

10.3. Modes and Emotions



Name	Mode	D'Arezzo	Fulda	Espinoza
Dorian	I	serious	any feeling	happy, taming the passions
Hypodorian	II	sad	sad	serious and tearful
Phrygian	III	mystic	vehement	inciting anger
Hypophrygian	IV	harmonious	tender	inciting delights, tempering fierceness
Lydian	V	happy	happy	happy
Hypolydian	VI	devout	pious	tearful and pious
Mixolydian	VII	angelical	of youth	uniting pleasure and sadness
Hypomixolydian	VIII	perfect	of knowledge	very happy

10.4. Affective phenomena

Preferences: evaluative judgments of stimuli in the sense of liking or disliking, or preferring or not over another stimulus (*like, dislike, positive, negative*)	**Interpersonal stances:** affective stance taken toward another person in a specific interaction, colouring the interpersonal exchange in that situation (*distant, cold, warm, supportive, contemptuous*)
Emotions: relatively brief episodes of synchronized response of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance (*angry, sad, joyful, fearful, ashamed, proud, elated, desperate*)	**Attitudes:** relatively enduring, affectively coloured beliefs and predispositions towards objects or persons (*liking, loving, hating, valuing, desiring*)
Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (*cheerful, gloomy, irritable, listless, depressed, buoyant*)	**Personality traits:** emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (*nervous, anxious, reckless, morose, hostile, envious, jealous*)

10.5. What is an emotion (II)?

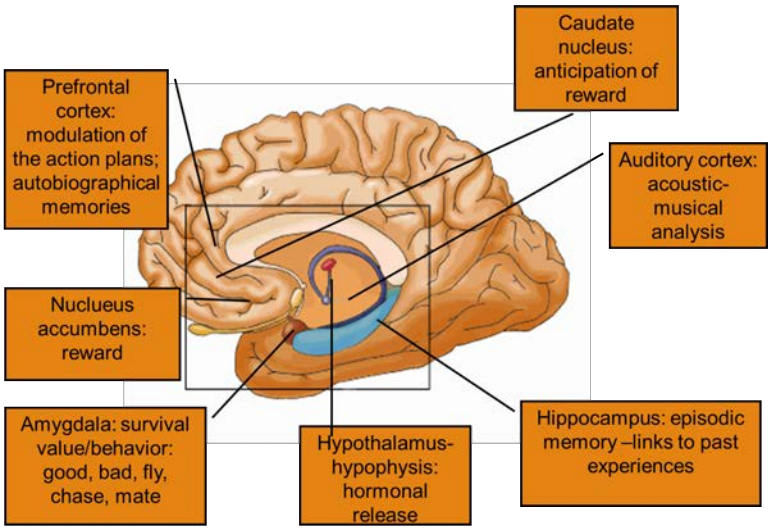
Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can:

- Give rise to **affective experiences** such as feelings of arousal (activation at different behavior levels), pleasure/displeasure, etc.

- Generate **cognitive processes** (e.g. increasing attention, appraisals, labeling processes, social bonding)
- Activate widespread **physiological adjustments** (e.g. increasing heart-rate, sweating, crying…)
- Lead to **behavior** that is often, but not always, expressive, goal-directed, and adaptive (e.g. running away, reiterating exposure…)

10.6. The limbic system

The limbic system is a set of brain structures located on both sides of the thalamus, immediately beneath the medial temporal lobe of the cerebrum primarily in the forebrain. It supports a variety of functions including **emotion**, **behavior**, **long-term memory and olfaction**. Emotional life is largely housed in the limbic system, and it critically aids the formation of memories.



10.7. Why does music convey (express) emotion?

Hearing **resemblance** between the music and the natural expression of the emotion (similarity to speech):

- Loudness and spectral dissonance found in an angry voice and in angry music.
- **Minor scale resembling spectra of subdued (depressed) speech**
- **Melodic** contours may resemble **questions**, **severe statements,** etc.
- Big and fast melodic leaps analog of happy jumps
- Descending contours analog of body movements (arms,head)

Accumulated connotations a certain musical phenomena acquire in a culture: we learn in our culture which musical cues correspond to which feeling:

- Brass instrumentation and slow tempo meaning “solemnity”
- Drums meaning “dance”
- Atonal music meaning “mystery”

10.8. Musical Features used to express emotions

Table 1. Summary of musical features correlated with discrete emotions in musical expression.

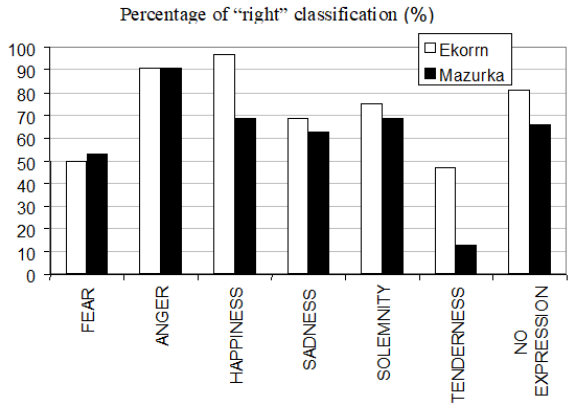
Emotion	Musical features
Happiness	Fast tempo, small tempo variability, major mode, simple and consonant harmony, medium-high sound level, small sound level variability, high pitch, much pitch variability, wide pitch range, ascending pitch, perfect 4th and 5th intervals, rising micro intonation, raised singer's formant, staccato articulation, large articulation variability, smooth and fluent rhythm, bright timbre, fast tone attacks, small timing varibility, sharp contrasts between "long" and "short" notes, medium-fast vibrato rate, medium vibrato extent, micro-structural regularity
Sadness	Slow tempo, minor mode, dissonance, low sound level, moderate sound level variability, low pitch, narrow pitch range, descending pitch, "flat" (or falling) intonation, small intervals (e.g., minor 2nd), lowered singer's formant, legato articulation, small articulation variability, dull timbre, slow tone attacks, large timing variability (e.g., rubato), soft contrasts between "long" and "short" notes, pauses, slow vibrato, small vibrato extent, ritardando, micro-structural irregularity
Anger	Fast tempo, small tempo variability, minor mode, atonality, dissonance, high sound level, small loudness variability, high pitch, small pitch variability, ascending pitch, major 7th and augmented 4th intervals, raised singer's formant, staccato articulation, moderate articulation variability, complex rhythm, sudden rhythmic changes (e.g., syncopations), sharp timbre, spectral noise, fast tone attacks/decays, small timing variability, accents on tonally unstable notes, sharp contrasts between "long" and "short" notes, accelerando, medium-fast vibrato rate, large vibrato extent, micro-structural irregularity
Fear	Fast tempo, large tempo variability, minor mode, dissonance, low sound level, large sound level variability, rapid changes in sound level, high pitch, ascending pitch, wide pitch range, large pitch contrasts, staccato articulation, large articulation variability, jerky rhythms, soft timbre, very large timing variability, pauses, soft tone attacks, fast vibrato rate, small vibrato extent, micro-structural irregularity
Tenderness	Slow tempo, major mode, consonance, medium-low sound level, small sound level variability, low pitch, fairly narrow pitch range, lowered singer's formant, legato articulation, small articulation variability, slow tone attacks, soft timbre, moderate timing variability, soft contrasts between long and short notes, accents on tonally stable notes, medium fast vibrato, small vibrato extent, micro-structural regularity

Note. Shown are the most common findings in the literature. For a more detailed treatment of studies, see Gabrielsson and Juslin (2003), Juslin (2001a), Juslin and Laukka (2003), and Juslin and Lindström (2003).

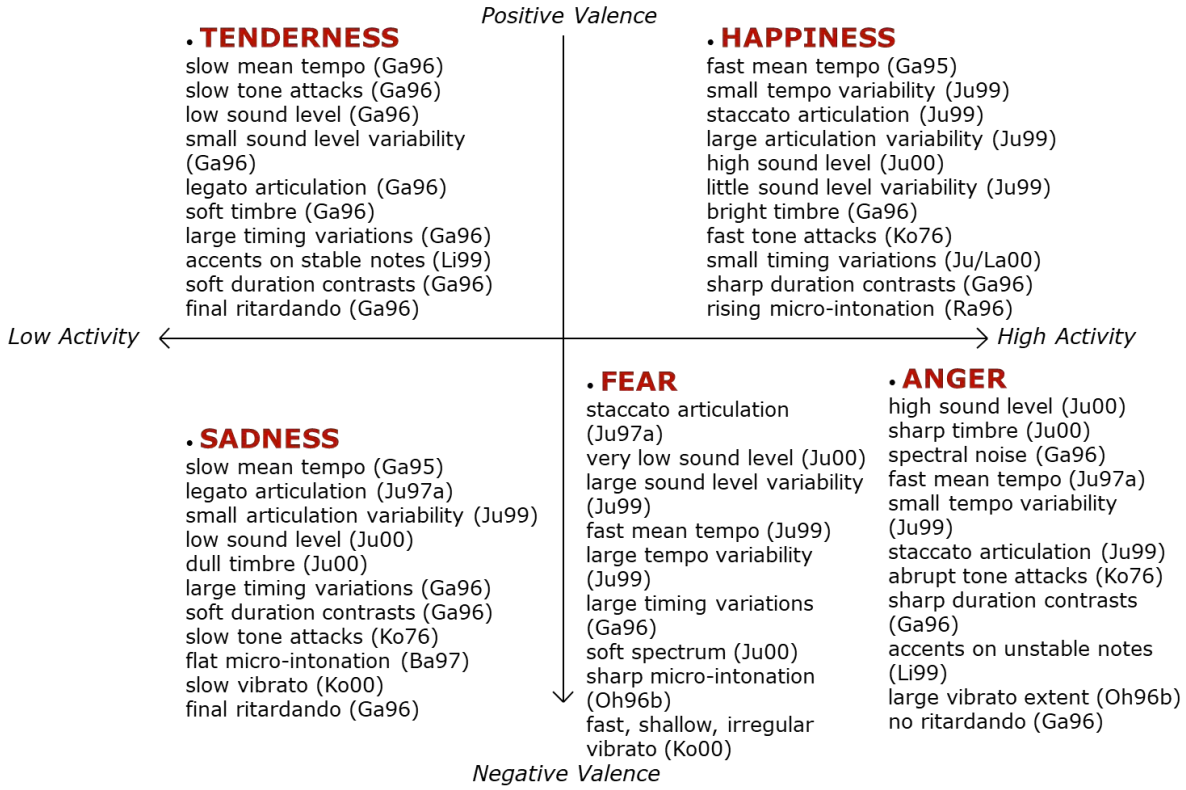
10.9. Synthesis of Emotion

An example of the "analysis by synthesis" strategy:

- Different renditions of the same piece are synthesized by changing musical parameters, then we study the effect of the parameter-tuning on the perceived emotions (anger, happiness, tenderness, neutral…)

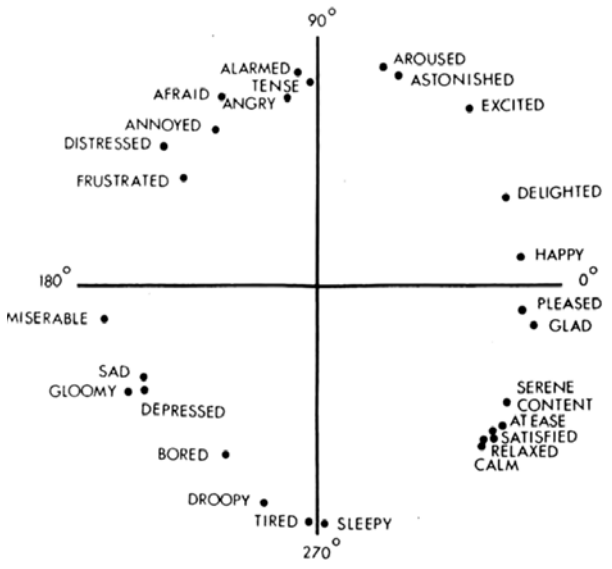


10.10. Dimensional representation with determining features



10.11. Russell's (1980) circumplex model

The circumplex model of emotion suggests that emotions are distributed in a two-dimensional circular space, containing **arousal** and **valence** dimensions. **Arousal represents the vertical axis and valence represents the horizontal axis**, while the center of the circle represents a neutral valence and a medium level of arousal. In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both of these factors. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions and affective states.



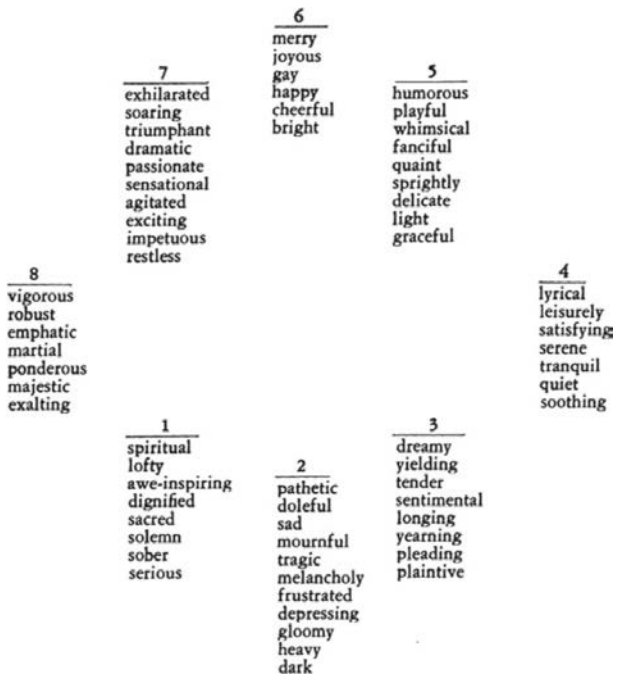
[PERFECTO QUIZ CARD] **Circumplex Model of Emotion:**
Arousal and Valence

[PERFECTO QUIZ CARD] **Arousal:**
Degree of alertness/energy, or readiness/attention.

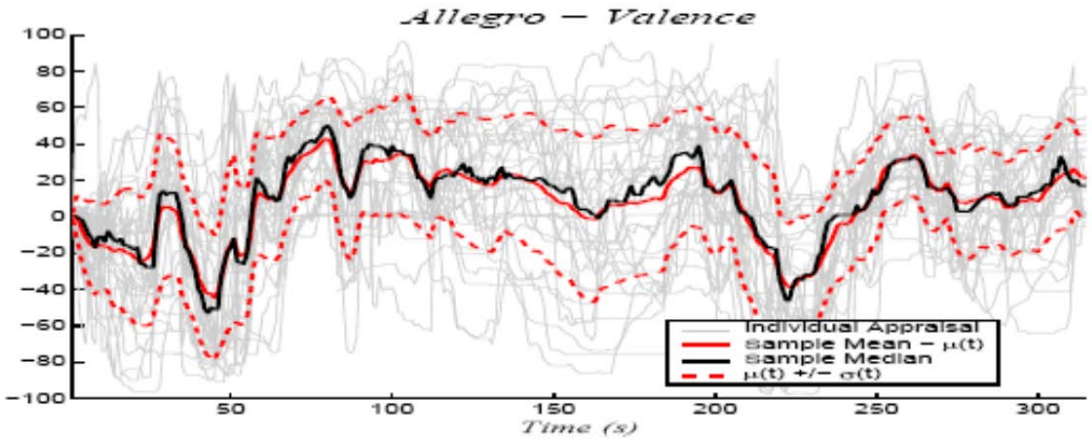
[PERFECTO QUIZ CARD] **Valence:**
Degree of aversiveness or sensitivity.

10.12. Hevner's (1936) categorical model

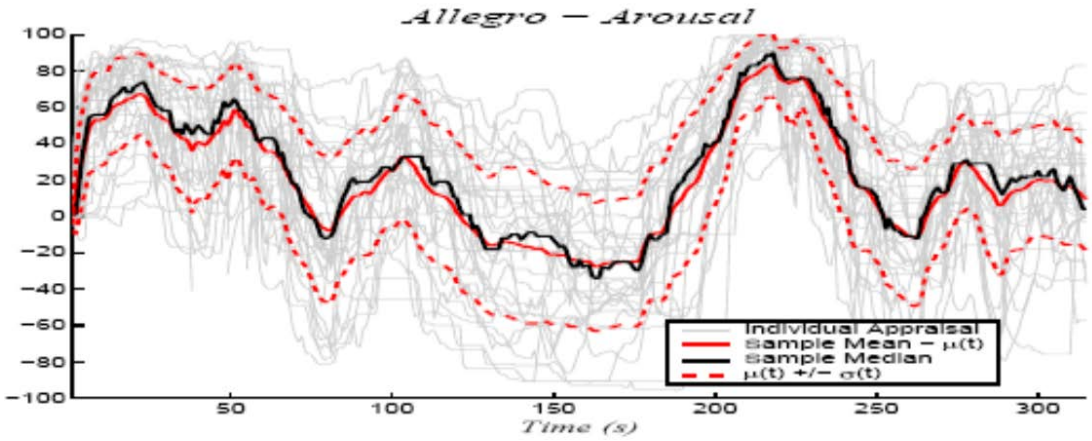
Eight clusters of affective terms:



10.13. Emotion in time



(a) Valence



(b) Arousal

10.14. Why does music induce emotions? Juslin & Västfjäll (2008)

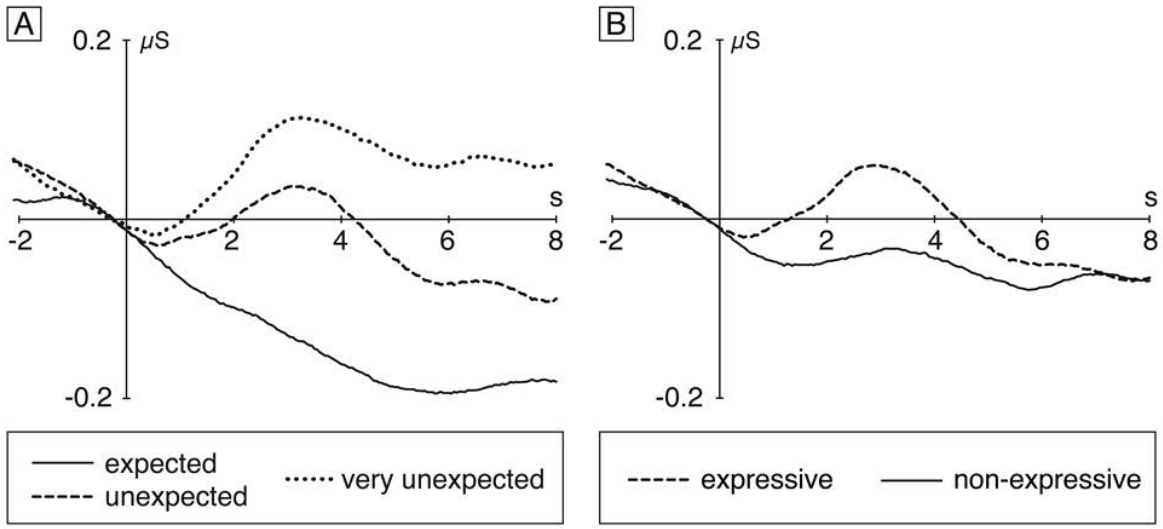
- **Musical expectation**: the anticipation infinite game
- **Arousal**: the activation of our systems
- **Mood contagion**: “monkey see, monkey do”
- **Associations**: “they are playing our song, darling”, can be unconscious
- **Imagery**: multimodality

10.15. Violations of musical regularities elicit emotional responses

First, the original version of a piano sonata was played by a pianist. This original version contained an unexpected chord as arranged by the composer. After the recording, the MIDI file with the unexpected (original) chord was modified offline using MIDI software so that the unexpected chord became expected, or very unexpected chord. From each of these 3 versions, another version without musical expression was created by eliminating variations in tempo and key-stroke velocities (excerpts were modified offline using MIDI software). Thus, there were 6 versions of each piano sonata: version with expected, unexpected, and very unexpected chords, and each of these versions played with and without musical expression.

10.16. Skin conductance responses (SCRs)

The **skin conductance response**, also known as the electrodermal response, is the phenomenon that the **skin momentarily becomes a better conductor of electricity when either external or internal stimuli occur that are physiologically arousing**. **Arousal** is a broad term referring to **overall activation**, and is widely considered to be **one of the two main dimensions of an emotional response**. Measuring arousal is therefore not the same as measuring emotion, but is an important component of it. **Arousal** has been found to be a **strong predictor of attention and memory**.



A: grand-average of SCRs elicited by expected, unexpected (original) and very unexpected chords (averaged across expressive and non-expressive conditions). Compared to expected chords, **unexpected and very unexpected chords elicited clear SCRs**. Notably, the SCR elicited by very unexpected chords was larger than the SCR to unexpected (original) chords, showing that **the magnitude of SCRs is related to the degree of harmonic expectancy violation**.

B: grand-average of SCRs elicited by expressive and non-expressive chords (averaged across expected, unexpected and very unexpected conditions). **Compared to non-expressive chords, chords played with musical expression elicited a clear SCR**.

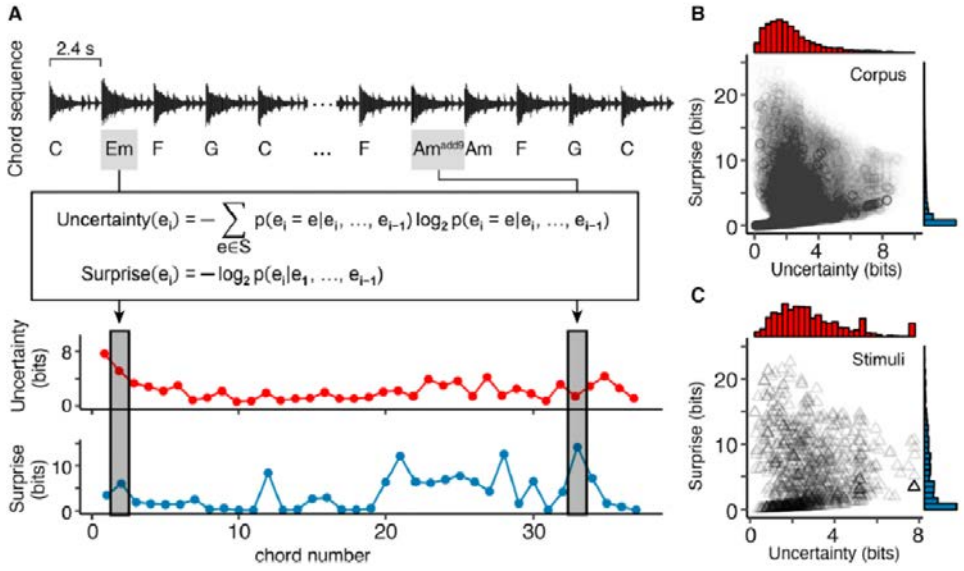


Figure 1. Quantifying Uncertainty and Surprise of a Chord

(A) An unsupervised statistical-learning model was trained on a corpus of 745 US Billboard "Hot 100" pop songs to derive the uncertainty (red) and surprise (blue) of chords (here, "Knowing Me, Knowing You" by ABBA; refer to Audio S1). Uncertainty is the lack of a clear expectation when anticipating an event *before* it is heard, while surprise occurs when what is *actually* heard deviates from expectations. Uncertainty of chord e_i is quantified by its entropy, or expected negative log-probability, taken across the set of all chords S in the corpus and conditional on the previous context of chords $\{e_1, \ldots, e_{i-1}\}$ in the progression. Surprise of chord e_i is quantified by its information content, and is the negative log-probability of the actual chord conditional on the context. Gray bars indicate points of high uncertainty but low surprise, and low uncertainty but high surprise. Subjects (n = 79) were asked to either rate the pleasantness of each chord (2.4 s) from 30 pop song chord progressions behaviorally or listen attentively and focus on how they fitted together in the context while undergoing fMRI scanning.

(B and C) Scatterplot and marginal densities of the uncertainty and surprise for all chords in the McGill Billboard corpus [21] (circles, n = 80,943) and in our chord stimuli (triangles, n = 1,039; Table S1).

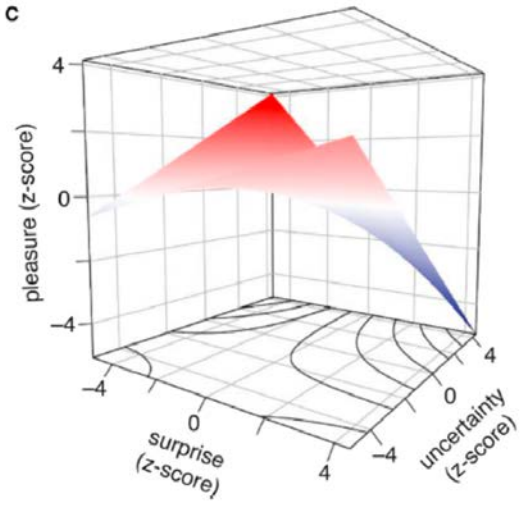


Figure 2. Uncertainty and Surprise Jointly Shape the Pleasure Rating of a Chord

(A) Standardized pleasure ratings to a chord progression taken from "Knowing Me, Knowing You" by ABBA (Audio S1). Diamonds indicate mean pleasantness ratings for each chord. Filled circles indicate fitted values from a linear mixed model with chord uncertainty, surprise, and their interaction as predictors. Error bars indicate 95% confidence intervals (95% CI). Low-level acoustic parameters were also included as covariates to control for sensory confounds.

(B) Contour plot demonstrating how pleasantness ratings jointly depend on uncertainty and surprise. When the tonal harmonic context does not allow for a prediction with high precision (i.e., when uncertainty is high), the pleasantness of a surprising chord is low. However, when the uncertainty is low, surprising chords are highly pleasurable.

(C) Data from (B) replotted in 3D. Although reminiscent of the characteristic inverted-U response from empirical aesthetics, the regression surface is in fact a saddle for which pleasantness varies nonlinearly across different levels of uncertainty and surprise.

together in the progression while undergoing fMRI scanning in Experiment 2. As before, we confirmed that our stimuli were unfamiliar to the subjects. Despite the sluggishness of the blood-oxygen-level-dependent (BOLD) response, the long duration of each chord (2.4 s) meant that metabolic changes could still be measured on a chord-to-chord level. We also used multiband echo-planar imaging (EPI) [37, 38] to allow for a sub-second temporal resolution while maintaining good spatial coverage. We focused our analysis on brain regions previously shown to be implicated in music-evoked emotions across multiple studies [1]: the bilateral amygdala and adjacent anterior hippocampus, bilateral auditory cortex, right nucleus accumbens, left caudate nucleus, and the pre-supplementary motor area. Given that musical pleasure depends on joint effects of uncertainty and surprise, we hypothesized that the underlying brain regions would also show the same interaction.

10.17. Chills

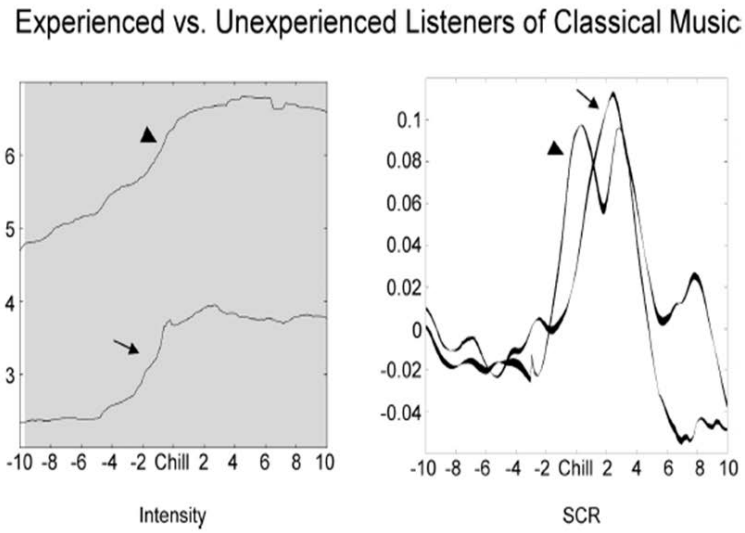
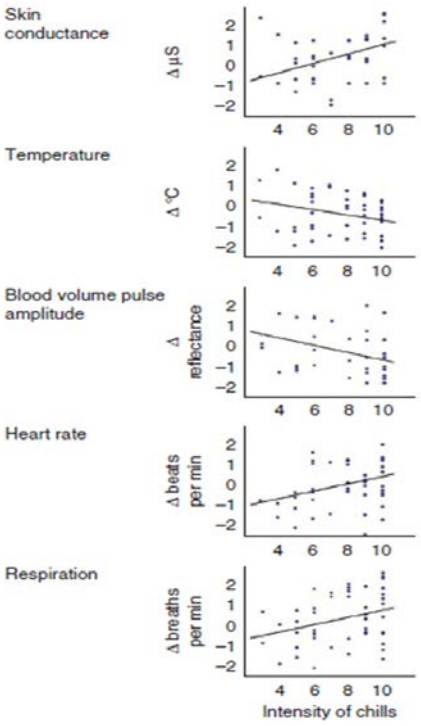
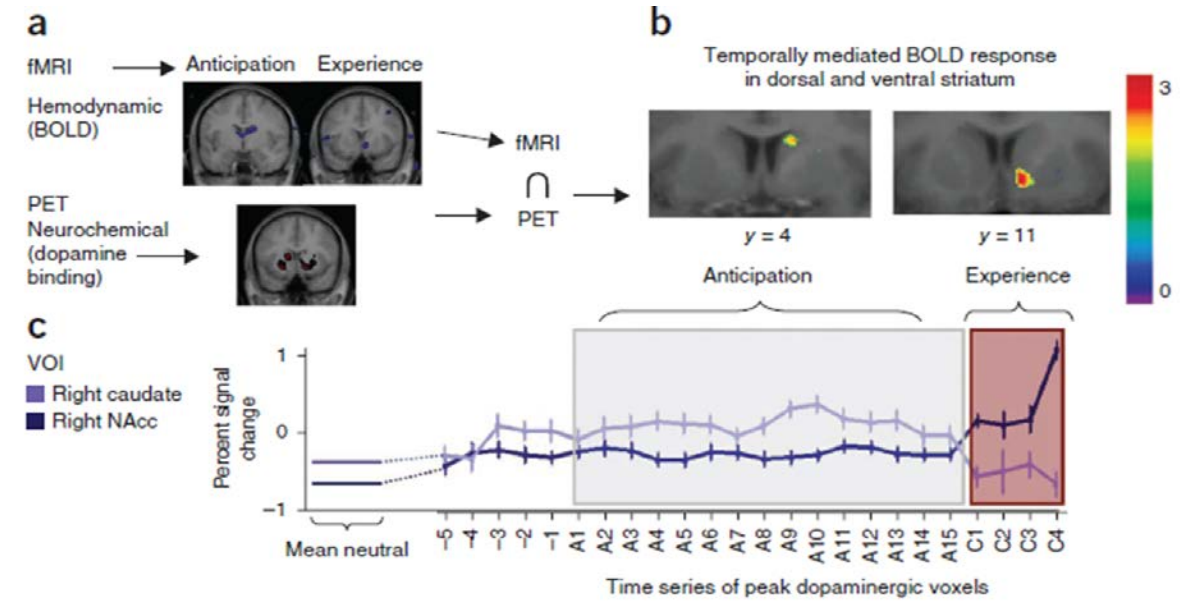


FIGURE 5. Comparison of chill samples (20 s) of participants highly experienced (arrowhead) and inexperienced (arrow) in classical music. Left intensity of feeling ratings, right skin conductance response (SCR). Significant differences (permutation test, $p < .05$) are shaded grey.





Tractography is a 3D modeling technique used to visually represent nerve tracts using data collected by diffusion MRI. It uses special techniques of MRI and computer-based diffusion MRI.

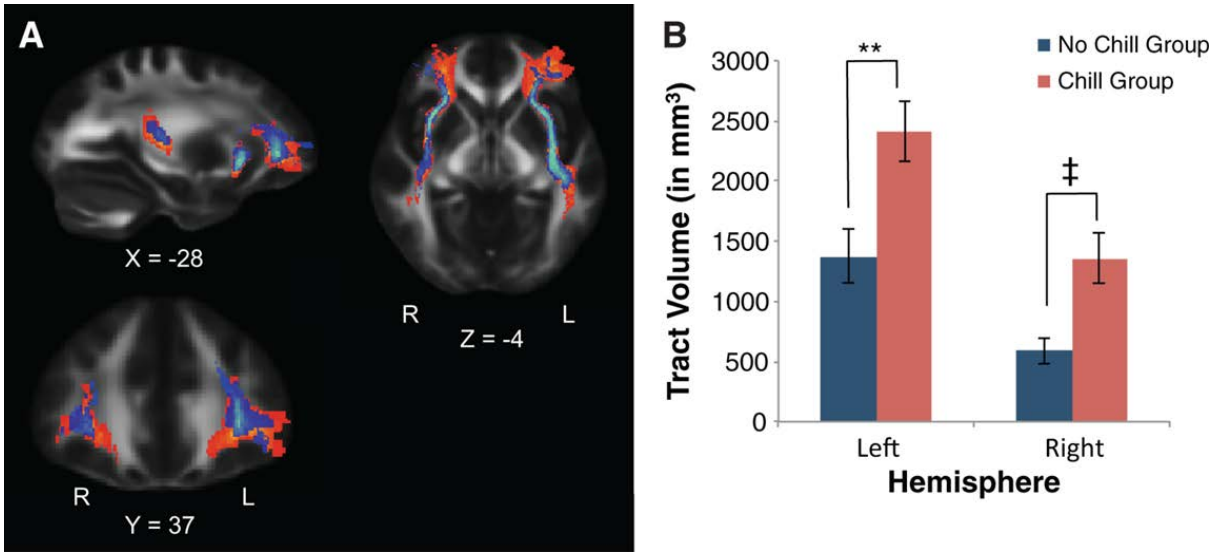


Fig. 3. Larger tract volume from pSTG to aIns and mPFC in Chill responders: (A) Diffusion tractography showed increased tract volume between auditory perception regions in the STG and emotional and social processing regions in the aIns and mPFC. (B) Tract volume between the STG, aIns and mPFC was significantly larger in individuals who frequently experience chills in response to music compared to matched controls. **P < 0.01 uncorrected. ‡ P < 0.05 after Bonferroni correction. Error bars denote standard error.