Loudness, Pitch and Timbre

John G. Beerends

May 2026

QUICK TAKE

Loudness, pitch and timbre are the best-known perceptual characterizations of stationary sounds. However, mapping real-world, physical measurements to loudness, pitch and timbre is quite complex. The main challenge is to obtain consistent and experimentally reproducible bridges between quantities that are objective—based on instrumental measurements of physical quantities—and those that appear subjective, which incorporate biomechanical, neurological, and psychological factors.

People have different perceptions and widely accepted measurement standards are thus based on the average behavior of large sets of test subjects. For certain applications age is an important factor that must be considered in the prediction of the perceived loudness, pitch or timbre of a sound. The core problem in psychoacoustics is to derive facts from opinions.

A better science of sound perception is developed by studying the neural firing patterns found in the auditory pathway and central processing in the brain, allowing to construct more accurate models. A major success of the advanced modelling is the ever more efficient coding of audio signals and measurement methods that allow us to predict the perceived quality of speech and music.

In journalism it is important to separate facts from opinions, in science one can ask the question whether it is possible to derive facts from opinions. In fields like physics and mathematics personal opinions play no role but in psychoacoustics opinions are the core of the research that has been and is carried out. Modern psychoacoustic research is evolving towards the study of increasingly more complex signals in a wide variety of experimental contexts. It's clear that the more complex signals are getting, the more personalized our opinions become. At a certain point, e.g. when dealing with speech and music signals, the acquired personal talents and taste may start to leak into our research, but even with simple stationary short duration sounds one can run into the fact versus opinion discussion.

A famous fact versus opinion controversy already popped up around the 1840s. Georg Ohm claimed that the mathematical theory of Fourier could explain the perceived pitch in complex tones while August Seebeck performed a series of experiments which led him to conclude that the perceived pitch cannot be explained by this theory. Seebeck claimed that the repetition pattern determines the pitch because he could hear a pitch in a signal that did not have a matching spectral component in the Fourier analysis. Ohm then dismissed his personal perception as an "illusion of the ear" while Seebeck argued that our ear is the sole judge of auditory sensations. The question how to derive scientific facts from personal opinions is still debated in psychoacoustics, only think about the many claims made in the world of HiFi audio that is driven by personal opinions.

A first step in the process of deriving psychoacoustic facts from personal opinions is agreement on the definition of properties like pitch and loudness. In 1951 the American Standards Association (ASA; currently known as ANSI) published a document that defined the pitch of stationary sounds as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high, such as a musical scale" and loudness as "the intensive attribute of an auditory sensation, in terms of which sounds may be ordered on a scale extending from soft to loud". They also defined timbre, another important characterization of stationary sounds, as "that attribute of auditory sensation in terms of which a listener can judge

that two sounds similarly presented and having the same loudness and pitch are dissimilar". Although these definitions have slightly changed over time, they provide an excellent starting point for our discussion on how to derive scientific psychoacoustic facts from personal opinions.

Most of the early psychoacoustic research was carried out with short duration, stationary signals, which are roughly defined as signals that show no variation, or mathematically well-defined small variations, over time. These types of artificial signals allow us to investigate auditory perception under well-defined circumstances. Typical stimulus durations in these studies are between 0.2 and 2 seconds and stimuli are mostly presented over headphones to allow for precise control of the acoustic stimulus. When we listen to such a stationary sound a first fundamental characterization is its loudness. The loudness control of your audio equipment is by far the most important audio control allowing you to adapt the loudness to your personal preference. For a large class of signals we will also perceive a clear pitch, allowing us to appreciate music. Unfortunately, even with simple stationary signals the relationship between physical parameters such as sound pressure level (SPL), that is determined by the amplitude of the signal; the frequencies of which it is composed or the repetition time; and psychoacoustic parameters like loudness and pitch are not simple, they are subjective quantities that can only be derived by asking subjects their opinion. Can these subjective quantities nevertheless be scientifically (objectively) measured?

A question that was not answered in the ASA document is whether subjective test procedures can be developed that allow for objective measurements of loudness, pitch and timbre using the ASA definitions. For loudness and pitch simple operational procedures have been developed. One can use listening test subjects and ask them to judge the loudness and pitch in comparison to a predefined stationary "ideal" sound and use an averaging procedure over individual opinions. In these experiments subjects are asked to rate the loudness and pitch in comparison to an ideal reference pure tone (sinusoid) of 1000 Hz played at a level of 40 dB SPL for around 1 second. This soft reference tone lies somewhere in the middle between the absolute hearing threshold in humans (around 0 dB SPL) and the loudest comfortable tone (around 80 dB SPL) on the logarithmic decibel scale. The frequency of the tone lies in the middle between the

lowest and highest audible frequency (20 Hz - 20,000 Hz, thus spanning around 10 octaves) on a logarithmic frequency scale. For loudness the Sone scale is used, and the loudness of the 1000 Hz 40 dB SPL reference tone is defined as one Sone. Sounds that are twice as loud get a loudness score of two Sone and sounds that are twice as soft get a loudness score of half a Sone. The loudest comfortable level of a 1000 Hz tone gets an average score of around 16 Sone. For pitch, the Mel scale is used, and the 1000 Hz 40 dB SPL tone is defined as 1000 Mel. Sounds that are twice as high get a pitch score of 2000 Mel and sounds that are twice as low get a pitch score of 500 Mel. The highest audible frequency of around 20 kHz gets an average score of around 3200 Mel. The most important result from these experiments is that we perceive sound on logarithmic scales, doubling the power results in less than doubling of the loudness and doubling the frequency results in less than doubling of the pitch.

For timbre the situation is more complicated than for loudness and pitch as it is defined as everything that is not loudness and pitch, a too wide definition. Although timbre is sometimes referred to as the color of a sound, we do not have words or a scale that characterizes the timbre of a sound. In vision we have agreed upon terms and definitions for color, and the complete perceptual color space is largely defined by three colors. In the case of hearing there are only more or less standardized "sound color" terms for noise, “white” for a flat frequency distribution where the amplitude of all audible frequencies is the same (power per Hertz is constant), and “pink” for an equal power per octave which more closely matches the logarithmic characteristic of human perception. In the course of history more noise color definitions have been established, brown (or Brownian), red, green, blue, violet, velvet, grey, black, all with their specific frequency distributions. In some situations, noise stimuli with different timbres can be ordered from low to high, e.g. when we have three band limited white or pink noise signals, 100-500 Hz, 500-2,000 Hz and 2000-10,000 Hz, subjects have no problem in ordering them from low to high, fulfilling the pitch definition. However, matching the noise bands to a single pure tone to determine the pitch will give a wide variety of responses and our intuition would not identify these stimuli as having “pitch”. Furthermore, it is not possible to play all musical notes within one octave using only three

bandlimited noise signals. It is clear that the narrower the bandwidth of the noise the more “pitchy” it will get.

Musical instruments can in most cases be identified easily by their onset, playing the same note on a piano, a violin or a trumpet will reveal huge differences in the onset. The stationary part of the notes will still allow us, in most cases, to identify the instrument by their timbre differences. And the differences between two musical instruments of the same type will be largely determined by their timbre, also coined the “tone color” of an instrument. People use words like “bright” and “dark” in the characterization of the timbre of a musical instrument but there is no agreed “standard” terminology and thus no “standard” measurement approach.

Voices also have their own characteristic timbre; a child voice sounds “brighter” than an adult male voice and when a single voice pronounces the steady-state vowels "aaaaaaaa" and "eeeeeeee" with the same pitch and loudness, the timbre determines the audible difference. This shows the complexity of timbre, the overall distribution of a child voice contains relatively more high frequencies and has an overall timbre that is significantly brighter than the timbre of an adult male, but the characteristic patterns of the steady-state aaaaaaaa’s are roughly the same for both voices and allow us to recognize an “a” in a wide variety of situations. We can even say that the “a” and “e” vowels are characterized by their timbre. In acoustic phonology timbres are described by the envelopes of the spectral frequency distributions and the first two maxima in the envelope largely define the vowel.

Another problem in discussing the timbre of a musical instrument or a voice is the impact of the room in which we are listening. We are all familiar with the timbre changes when listening to a sound in a bathroom, or a room with strong resonances that enhance certain low frequencies. Reflections interfere with the direct sound and the timbre changes that are introduced by the room can be described by the reflection pattern, or in a first approximation by the reverberation time, the time it takes a sound to decrease 60 dB in level (RT60). If we are listening in an anechoic room, where there are no resonances and reflections, and the RT60 is zero, and thus the room does not color the sound, the directional properties of the instrument or voice will cause a strong dependency of the timbre on the spatial positioning. If a voice is recorded from upfront

the timbre will be brighter than recorded from the back. For many musical instruments the timbre is strongly dependent on the direction, and if we make a recording of a Stradivarius in an anechoic room the sound will have an unnatural timbre that is determined by the chosen positions of the Stradivarius and the microphone. This shows that a room that does not color the sound, i.e. that has little or no reflections, does not necessarily reveal the “true” timbre of a sound source. For sound sources with strong varying radiation patterns, like e.g. violins, “*the*” timbre is difficult to define. Only in the case that a source has a constant and smooth directivity pattern the “true” timbre can be defined easily.

A room which is judged to be of high acoustic quality will have a reflection pattern that smoothly integrates the sound over all directions without disturbing resonances and that will thus make a Stradivarius sound excellent. Unfortunately, the RT60 is only a first rough indicator in the characterization of a room and two rooms with the same RT60 can sound completely different. In room acoustics many additional parameters have been developed to quantify the impact on the perceived timbre of a sound played or recorded in that room. The best-known ones are early decay time (EDT), clarity (C80), and definition (D50).

We see that describing timbre as “that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar” is a too general definition, too many aspects of sound are included in it. The same sound source recorded in two different rooms and played back over headphones can have completely different timbres. As a first step we could split the timbre definition into a description of the timbre of the sound source versus the description of the timbre of the room, thus quantifying the impact of the transmission path between the sound source and the ear.

If we want to assess the timbre of a sound itself, the best experimental approach is thus to use high quality headphones. Conveniently, when we generate the signals on a computer, the source characteristics can be manipulated exactly, allowing us to investigate the perceptual properties of timbre more precisely. These experiments have shown that the timbre of a sound itself is a multi-dimensional property that is difficult to characterize. In a first approximation the

sound analyses carried out by the ear can be described by a filtering operation using a set of around twenty to thirty different, overlapping, frequency band filters, referred to as critical bands. The power distribution over these frequency bands characterizes the perceived timbre of a sound to a large extent. This results in a timbre characterization that uses a twenty to thirty dimensional space, where the power in each critical frequency band is interpreted as an independent dimension. We see that in hearing the description of timbre, sometimes referred to as tone-color, requires significantly more dimensions than the three required to describe color in vision.

The width of the critical band filters can be assessed in masking experiments where a sharp single loud frequency peak is found to mask softer single frequency peaks that lie in the vicinity of the loud peak. If two frequency peaks are separated by more than about one critical band the softer peak is no longer masked by the louder peak. If there is a single sharp maximum, only one critical band is stimulated, and we perceive a clear pitch of which the strength gets stronger as the maximum gets more peaked and a timbre that can best be described as “pure”. If we use noisy sounds that have no local maxima, and that stimulate all critical bands, we can use the noise colors to describe the timbre. When there are a few maxima, then sometimes it is possible to use vowel-like characterizations that either have a “noisy” or a “pitchy” character.

The same multi-dimensional approach can be taken in quantifying “the timbre of a room”, the RT60 and EDT can be assessed separately for each critical band. This frequency dependent measurement of the RT60 and EDT allows a significantly better characterization of the impact a room has on the timbre of a sound listened to, or recorded, in that room. We should also be aware that once a room is characterized, the impact on the perceived timbre is dependent on the distance between the sound source and the ear, or the microphone in the case of an audio recording. The larger the distance the bigger the impact on the perceived timbre. A musician playing an acoustic instrument will hear his instrument at close distance, minimizing the impact of “the timbre of the room”, but also maximizing the timbre distortion caused by irregularities in the radiation pattern of the instrument. An acoustic instrument with an irregular radiation pattern needs a high acoustic quality room that integrates the sound over all directions smoothly in order to sound good. Note that when a recording is made of a musical instrument in a room with little or

no reflections the timbre of the instrument cannot be restored by using an artificial reverberation which introduces the same reflections as a high-quality room because only a limited part of the irregular radiation pattern is processed.

Subjectivity in the Domain of Psychoacoustic Research

The subjective nature of loudness, pitch and timbre entails that we all have different experiences when we hear a certain sound. If we ask a single person to judge the loudness, pitch or timbre in an experiment we will get a personal opinion and big differences are found if we carry out subjective experiments with a large, heterogeneous, set of subjects. Age is an important factor in psychoacoustics. As we get older, we lose our ability to hear high frequencies, so for example, my current upper limit in the frequency domain (age 72) lies in the order of 7 kHz, while a young reader (< 20 years old) may have an upper limit of 20 kHz. In the time domain, extreme short clicks are inaudible for me but can be clearly audible for a young listener.

Decades of psychoacoustic experiments have shown that the ear analyzes sounds in short duration intervals, or time windows, of about 20 to 40 ms, largely independent from age. Our cognition integrates these observations over several consecutive time windows, somewhere around seven for loudness. Loudness thus increases with duration until about 200 ms after which a stable judgment is reached for stationary sounds. In psychoacoustic research many experiments have been carried out to be able to standardize objective loudness measurements. In the most widely used type of loudness measurement subjects are asked to adjust the loudness of a reference pure tone of 1000 Hz to the loudness of a test pure tone over a wide range of frequencies (40 Hz – 15 kHz) and intensities (inaudible – 120 dB SPL), leading to the so-called equal loudness contours. The first large study on the loudness of pure tones was carried out in 1933 by Fletcher and Munson using one second duration stimuli but their study does not give a clear indication of the age dependency. In 1957 Robinson and Dadson used the same experimental approach with a clear age categorization. Their results show that for a 15 kHz pure tone the absolute threshold of audibility for the 60-year age category is about 60 dB higher than for the 20-year category. They also carried out measurements with headphones as well as with

loudspeakers to take into account the diffraction effect around the head. These measurements have been reproduced in many investigations and have resulted in a worldwide technical standard for equal loudness contours, ISO 226. From these measurements one can derive frequency weighing curves with which the loudness of stationary sounds can be measured. The most widely used one, called dB(A), was already standardized by the ASA in 1936 using a rather fast integration time of 125 ms, allowing to track rather fast loudness changes. Later standards (see IEC 61672) added new weighing characteristics (B and C) and standardized two different integration times, 35 ms (fast), allowing to track very fast loudness changes, and 1000 ms (slow). The fast integration time roughly matches the analysis window of the ear but overestimates the loudness of short duration stimuli while the longest one provides a good match to the perceived loudness for longer duration stationary sounds.

Loudness estimation may be used to quantify the risk of noise-induced hearing loss resulting from extensive exposure to energetic sounds. For loudness above 80 to 85 dB(A) sounds may cause hearing loss, and the harmfulness of loud sounds is determined by the energy, i.e. duration times level. For 85 dB(A) around 8 hours of exposure is the limit and every 3 dB increase in level, thus doubling the power, halves the safe listening time. For 100 dB(A) this results in a maximum acceptable duration of about 15 minutes. Unfortunately, loudness has a drug-like property, people tend to get hooked on loudness and the preferred setting that is found for speech and music is mostly too loud. Even for simple speech communication we tend to set the loudness very high, in telephony the preferred loudness at the entrance of the ear measures up to 20 dB(A) higher than found in natural communication. In terms of power this means that the ear gets a 100 times more powerful signal than in the natural situation. The standard level for telephony as recommended by the International Telecommunication Union lies close to this high level. For music the situation is even worse, ear buds and modern loudspeaker technology make it too easy to listen to music on a too high level, way above the safe loudness limit. In pop music events the performers, but also to a large extent the listeners, may therefore need hearing protection to prevent noise-induced hearing loss later in life.

Another issue in loudness measurements we have to consider is that when a technical measurement predicts that two sounds are equally loud, they can elicit a very different perceived annoyance. If I play my music loudly it will please me, but my neighbors will be disturbed and complain that I play too loud. Loudness is thus not necessarily a good predictor of annoyance, making it difficult to agree on standards that define the maximum acceptable loudness in situations where people have a conflict of interest. The counterintuitive result of 90 years of psychoacoustic research is the fact that the dB(A), in its first form standardized in 1936 and only marginally updated, is still widely used (see IEC 61672), while advanced measurements that use more accurate perceptual modeling, and that are based on the perceptual Sone scale, are seldom used. The reason for this is simple: loudness is a subjective quantity, the loudness of an 80 dB(A) 20 kHz tone for me (age 72) is zero Sone, while for a young listener (< 20 years old) it may be a few Sones. Therefore, to apply the Sone scale we would have to develop models that factor in age, hearing threshold, sound masking characterizations and some kind of averaging over individual differences, to get an objective, scientifically accepted measurement. Even if we take these factors into account correctly, our measurement procedure may end up with a problem that is often encountered with technical measurements that use a perceptual approach; they are roughly correct only within a limited context. There is thus no advantage in using Sone over dB(A) in the measurement of loudness. Even the advanced perceptual models as standardized by the International Organization for Standardization (ISO) that take into account different aspects of masking and loudness growth (e.g. Zwicker, Stevens and Moore-Glasberg-Schlittenlacher model, see ISO 532) are seldom used. The annoyance of your vacuum cleaner, refrigerator, heat-pump, etc, is still being quantified in dB(A)'s.

For narrow-context situations the International Telecommunication Union standardized dedicated loudness models, for speech the ITU-T (Telecom sector) standardized the Loudness Rating (LR, Recommendation P.79), for audio the ITU-R (Radio sector) standardized the Loudness Units relative to Full Scale (LUFS, Recommendation BS.1770). The LUFS scale is focused on the modeling of loudness as perceived over large durations in digitally represented audio signals.

For pitch the situation is even more complex than for loudness. Pitch can be measured on the physical Hz scale by asking subjects to adjust the frequency of a reference pure tone of varying frequency, using a predefined level, to the pitch of the sound that is tested. When we present a pure tone of 1000 Hz we perceive a single pitch. However, when we simultaneously present two pure tones, say 1000 and 1200 Hz, we hear three pitches, two related to the spectral frequencies of the pure tones and one related to the repetition rate of 200 Hz, the waveform repeats itself exactly every 5 ms. This example illustrates the root cause of the Ohm Seebeck controversy as to what dominates the pitch, the time or frequency domain. A first step towards solving this controversy was already introduced by Helmholtz in 1863 by his observation that one can perceive two different kinds of pitches in complex tones, depending on whether the attention is focused on the harmonics of the complex tone, analytic listening, or on the complex as a whole, synthetic listening. In the early 1970s Terhardt introduced the terms spectral and virtual pitch for the perceived pitch in respectively the analytic and synthetic mode of perception, where “virtual” indicates that one can perceive a pitch that has no equivalent spectral frequency component. Decades of psychoacoustic research has shown that both viewpoints are correct, pitch can be dominated by the repetition rate or by the loudest spectral component(s) in the signal. The theory of pitch perception is still being developed, but the fact that we perceive three pitches in a stationary signal composed of a 1000 Hz and a 1200 Hz tone shows that we have to introduce some kind of probability function which in this example should allow to predict the relative strength of the repetition pitch and the two spectral pitches, depending on the experimental context.

The perceptual one-dimensional definitions for loudness and pitch can be easily interpreted if we use ideal signals such as pure tones where the amplitude determines the loudness and the repetition time T coincides with the period, and hence with the frequency F (F=1/T, repetition rate). For complex signals the loudness can still be measured by a one-dimensional quantity (as is the amplitude). However, when it comes to pitch, it must be defined as a two-dimensional quantity, where one dimension is focused on the frequencies of the loudest spectral components and one dimension on the most dominant repetition rates.

Another point in our assessment of loudness, pitch and timbre is that we should realize that when we make a high-quality recording of a sound and play it back, the pitch should remain the same, but the loudness and timbre should be adaptable to one's personal preference. For loudness this is simply carried out by the volume control, the most widely used psychoacoustic control in audio equipment. For timbre, that is also impacted by the listening room, this is more difficult. The most widely used approach was developed by Baxandall in the early fifties using a simple passive timbre control circuit that only uses capacitors and resistors to balance the low (20-200 Hz), mid (500-2,000 Hz) and high frequencies (4000-20,000 Hz). This approach is still widely used and can be found as a bass and treble knob on most HiFi amplifiers today. With the introduction of digital processing more advanced adaptations, like parametric equalization and room correction/simulation have become possible. With modern signal processing techniques it is possible to compensate timbre distortions by the play back room. Although timbre is difficult to quantify, it is rather easy to adapt it to one's own preference and often it is also possible to compensate timbre degradations in the recording / playback chain.

For pitch we seldom need control in playing back music. In fact, although pitch variations are extremely important in music, pitch variations introduced by the audio system are almost always perceived as a severe degradation in audio quality. Old analog audio formats like the LP, compact cassette and open reel tape, suffer from pitch variations which are referred to as "wow and flutter", and are always perceived as unwanted degradations. In the digital domain, the recording and reproduction of pitch without audible degradations is less of a problem and with the introduction of the compact disc (CD) in 1982, wow and flutter are no longer an issue in high quality audio reproduction in home audio systems.

Advanced Modeling of Auditory Perception Based on Neural Coding

A more advanced modeling of the auditory system starts with the description of how sound enters the ear and is transported to the inner ear, where the transduction to neural pulses is carried out. This step is already quite complex and can be modeled with a biomechanical approach that can

roughly be described as a filtering and amplitude compression process. When sound reaches the cochlea in the inner ear, a traveling wave will arise and cause a vibration pattern on the basilar membrane where different sections respond to different frequencies. This results in a tonotopic organization, whereby frequency is translated to a place of mechanical activity. The basilar membrane's vibration is mechanically linked to hair cells causing stereocilia on these hair cells to deflect, which in turn triggers the release of neurotransmitters that initiate nerve impulses. This process thus carries out the transduction from sound to neural pulses that are processed by our brain where the loudness, pitch and timbre of the sound are determined. The first one to describe this process was von Békésy (1960) and later research showed that the filtering that is carried out by the basilar membrane is further enhanced by an additional active feedback mechanism via outer hair cells.

To have a more fundamental understanding of loudness, pitch and timbre we can look at the firing patterns as found in the auditory nerve. A first observation is that the firing rate of a neuron, the number of impulses per second, largely corresponds to the loudness of a sound but the absence of sound is not characterized by a zero-firing rate, but some level of spontaneous activity. This has been a dangerous path in evolution, especially in combination with the behavior of our active outer hair cells that pump energy back into the mechanical auditory system allowing to hear extremely low-level sounds. Unfortunately, this advanced processing can lead to the situation where complete silence in the world is mapped towards a loud internal sound caused by some form of incorrect spontaneous activity. This form of activity can lead to a very loud disturbing tinnitus (“ringing in the ears”). Experiments have also shown that the number of neurons that are firing increases with the level of the sound and that the high level of activity, associated with the relatively intense frequency components in the signal, suppress (inhibit) the activity in the vicinity of these neurons. Inhibition is an important feature of the processing of sound in our hearing system, and it plays an even more dominant role in vision.

Another important observation is that the timing of neural firings shows a phase locking behavior, the neurons fire in roughly the same manner within each repetition of the audio signal, sometimes skipping a single or more signal repetitions. This phase locking can be used to derive

the onset, offset and repetition pitch in an audio signal. The process of skipping one repetition in the firing results in a timing difference that is doubled and can be interpreted as the generation of a subharmonic, a virtual pitch with half the frequency as found in the audio signal. This generation of subharmonics forms the basis of what is known as octave ambiguity, where two pitches that differ by a factor two are perceived to be similar.

. The development of neural models has led to an improved understanding of our auditory system, but the mapping from the physical domain towards the main psychoacoustic sound characterizations; loudness, pitch and timbre, are still being developed. For loudness an important aspect in the modeling is found in the masking properties of the ear. Masking occurs in both the time and frequency domains and is already partially introduced by the mechanical properties of the ear. A short duration pulse will be smeared out by the inertia of the mass that is distributed alongside the basilar membrane, resulting in a smeared out tonotopic representation along the basilar membrane. On the auditory nerve level, a short duration pulse inhibits a neuron for a short duration, the neural dead time, also known as the refractory period, and thus masks later short duration pulses. This period also limits the maximum firing rate of a neuron. Furthermore, a hair cell neuron combination that fires on a certain spot along the cochlea--representing the tonotopic decomposition as carried out by the inner ear---also inhibits neurons in the vicinity that represent slightly higher and lower frequency components. Advanced loudness models like the Moore-Glasberg-Schlittenlacher model as specified in ISO 532, incorporate an approximation of the masking as found in the auditory system. The core idea in these models is that soft time-frequency components that are masked by loud time-frequency components have a loudness of zero and thus do not contribute to the overall loudness of the signal. However, these models only show a small improvement in correlation between subjective and objective measurements of loudness and the use of these models is limited to special situations. Also, the fact that the individual loudness characteristics are so different between listeners, especially showing large differences between age categories, makes it hard to develop generally applicable loudness models. Also, the fact that even a simple quantity like the dB(A) is a good predictor of

the average perceived loudness of stationary sounds makes the practical use of advanced loudness models limited.

The improved understanding of the masking properties has, however, led to the development of audio compression techniques that allow the bit rate of digitally coded sound to be reduced with only a marginal loss in quality. It turns out that for high quality audio many details in the signal are masked and can be removed without significant loss in quality. The main problem is that high-quality compression requires complicated masking models. Current standards like the ones developed within MPEG use direct modeling of masking. First, they transform the audio signal into a time-frequency representation. Then they calculate which soft time-frequency components that lie in the vicinity of loud time-frequency components can be omitted in the coding of the audio signal (see illustrations on page 16-18, Figures 1, 2, 3). Even this direct modeling approach in the signal encoding is already quite complex and is still being improved using more advanced modeling approaches. Especially the loudness growth of masked components that pop up just above the threshold is difficult to model. In high-quality audio compression only the time-frequency components that have zero loudness can be omitted without loss in audio quality. In contrast, the decoding process, i.e. the reconstruction of the time signal from the compressed time-frequency representation, only requires a relatively simple model that can also be standardized, thus allowing simple algorithms for use in audio playback. Therefore, most audio compression standards can focus on the standardization of the playback algorithm, allowing for further quality improvements in the encoding process without the need to update the standard.

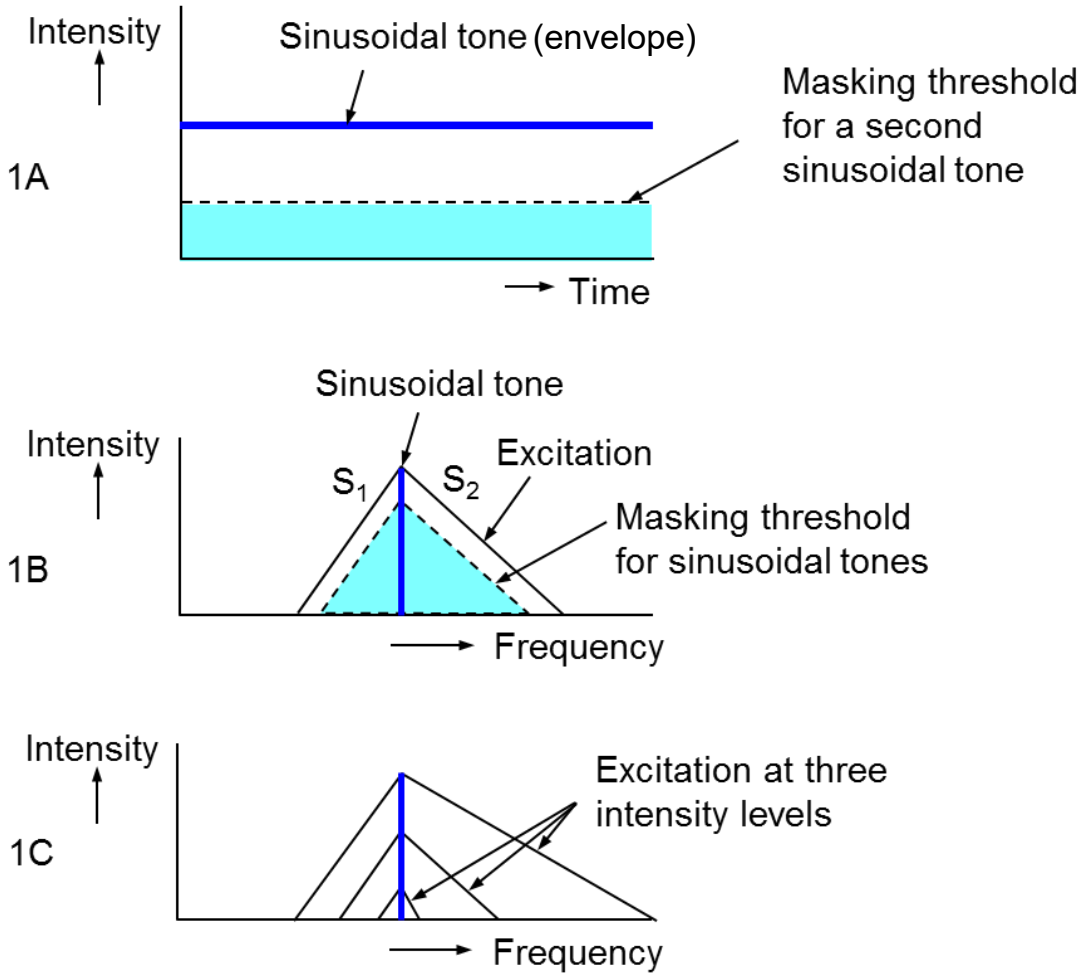


Figure 1. Masking of a soft sinusoidal tone by a second loud one in time- (1A) and frequency domain (1B). All tones within the blue triangle in 1B are inaudible, they have a zero partial loudness. For loud sinusoidal tones the right slope S_2 is steeper for low loudness values (1C). Interference that leads to audible beats is not taken into account.

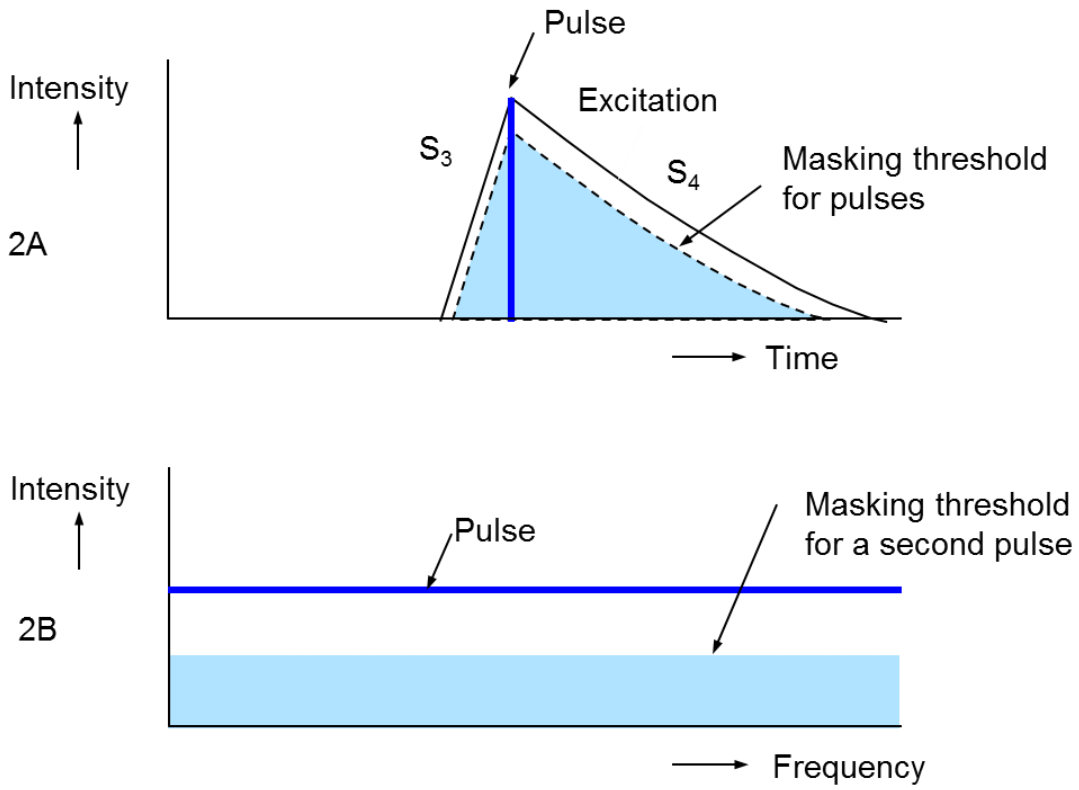


Figure 2. Masking of a soft pulse by a loud pulse in time (2A) and frequency domain (2B). All pulses within the blue triangle in 2A are inaudible, they have a zero partial loudness. For pulses the left slope S_3 is significantly steeper than the right slope S_3.

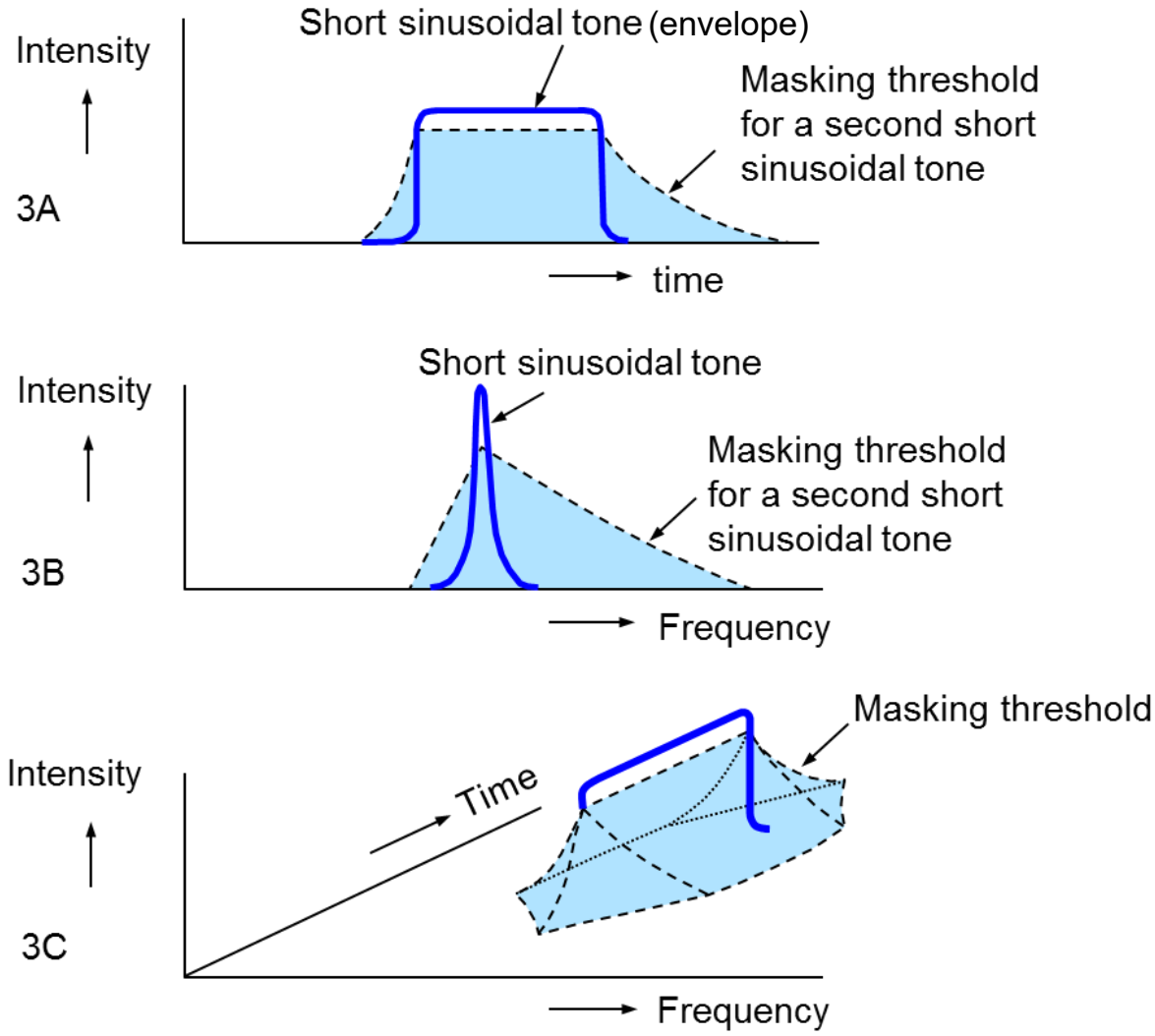


Figure 3. Masking of soft short sinusoidal tones by a loud short duration sinusoid in time (3A), frequency (3B) and time-frequency domain representation (3C). All tones with the blue "tent" shaped figure are inaudible, they have a zero partial loudness.

For pitch we see two main mechanisms that encode the pitch of a signal. Firstly, the tonotopic mapping that is introduced by the inner ear supports the idea of Ohm that the perceived pitch is derived from by the power distribution of the sound in the frequency domain. However, the fact that subjects hear a 200 Hz dominant pitch in a sound that is composed of a set of only upper harmonics, e.g. 400, 600, 800, 1000 and 1200 Hz, shows that in many cases the perceived pitch is not dominated by the frequency components of the sound but by the repetition pitch. One could expect that the repetition pattern in the waveform is directly coded and interpreted, however in an experiment carried out by Houtsma and Goldstein (1972) it was shown that listeners can also perceive a missing fundamental with a harmonic two-tone complex with one pure tone in each ear. Their results show that the formation of pitch is a process that is dominated by a central cognitive process that allows the derivation of a repetition pitch even in the case that each ear only hears a single pure tone, in which case the auditory system has no direct access to the repetition time of the envelope. The fact that we can hear a repetition pitch even if there is no repetition in the signal makes the term “virtual pitch” as introduced by Terhardt a correct description of our perception. In my PhD research, carried out in the 80’s, I reformulated the stochastic harmonic approach, as formulated by Goldstein, towards a stochastic subharmonic approach thus allowing to calculate a pitch power density function from which the perceived strength of both the spectral and the virtual pitches can be derived.

In the many psychoacoustic experiments that have been carried it is shown that pitch is derived from three different kinds of cues depending on whether the individual harmonics are resolved, processed by different critical bands, or unresolved, processed within the same critical band. The first cue is derived from the place along the basilar membrane where there are significant mechanical activities for both resolved and unresolved components, the tonotopic mapping. The second cue is derived from timing cues of the resolved components. The third cue is derived from timing cues that result from the envelope of unresolved components. The extent to which one perceives the pitches of a complex tone in analytic mode, focused on the dominant frequency spectrum components, or in synthetic mode, focused on the dominant repetition rates, largely depends on the experimental context, individual preferences, training etc. In general, one

can say that when a signal contains many frequencies, subjects are inclined to listen in the synthetic mode of pitch perception. The current accepted vision is that pitch is a pattern-recognition process that is dominated by a central processor that tries to find a (missing) fundamental frequency F0 which fits best to the frequencies of the spectrum using the tonotopic information from the resolved harmonics as well as the repetition rate derived from the phase locked neural pulses of the resolved harmonics. To a lesser extent the phase locked neural pulses of the time envelope of unresolved harmonics are used (see illustration on page 22).

Most current pitch algorithms use an autocorrelation approach where correlations are calculated between the signal and delayed versions of that signal, allowing to find dominant spectral and virtual pitches. This approach works well for most “natural” signals which are dominated by harmonically related frequencies but fails on large sets of (inharmonic) “test” signals that are used in psychoacoustics. The core of the “pitch problem” as found with many test signals is the fact that one can perceive several pitches that are sometimes equally loud, and the dominant pitch is difficult to predict and often dependent on the experimental context. One approach to deal with those ambiguous signals is to calculate a probability density function from which the probability that a certain pitch is perceived can be calculated. Such an algorithm can be designed on the basis of the phase locking found in the neural firings where sometimes a single or more periods are skipped, which can be interpreted as a mechanism that generates subharmonic pitches. Whenever the subharmonic pitches of individual frequency components coincide, they strengthen the perceived virtual repetition pitch.

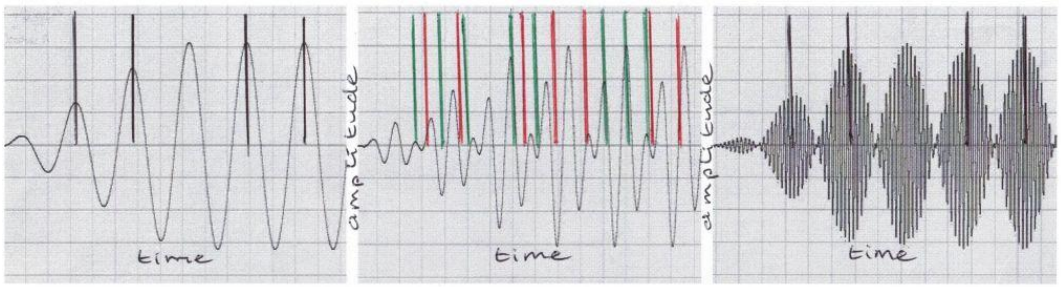
An example of the three different mechanisms of pitch perception, including the subharmonic analysis approach, is given in the illustration on page 22. Three different stimuli are analyzed:

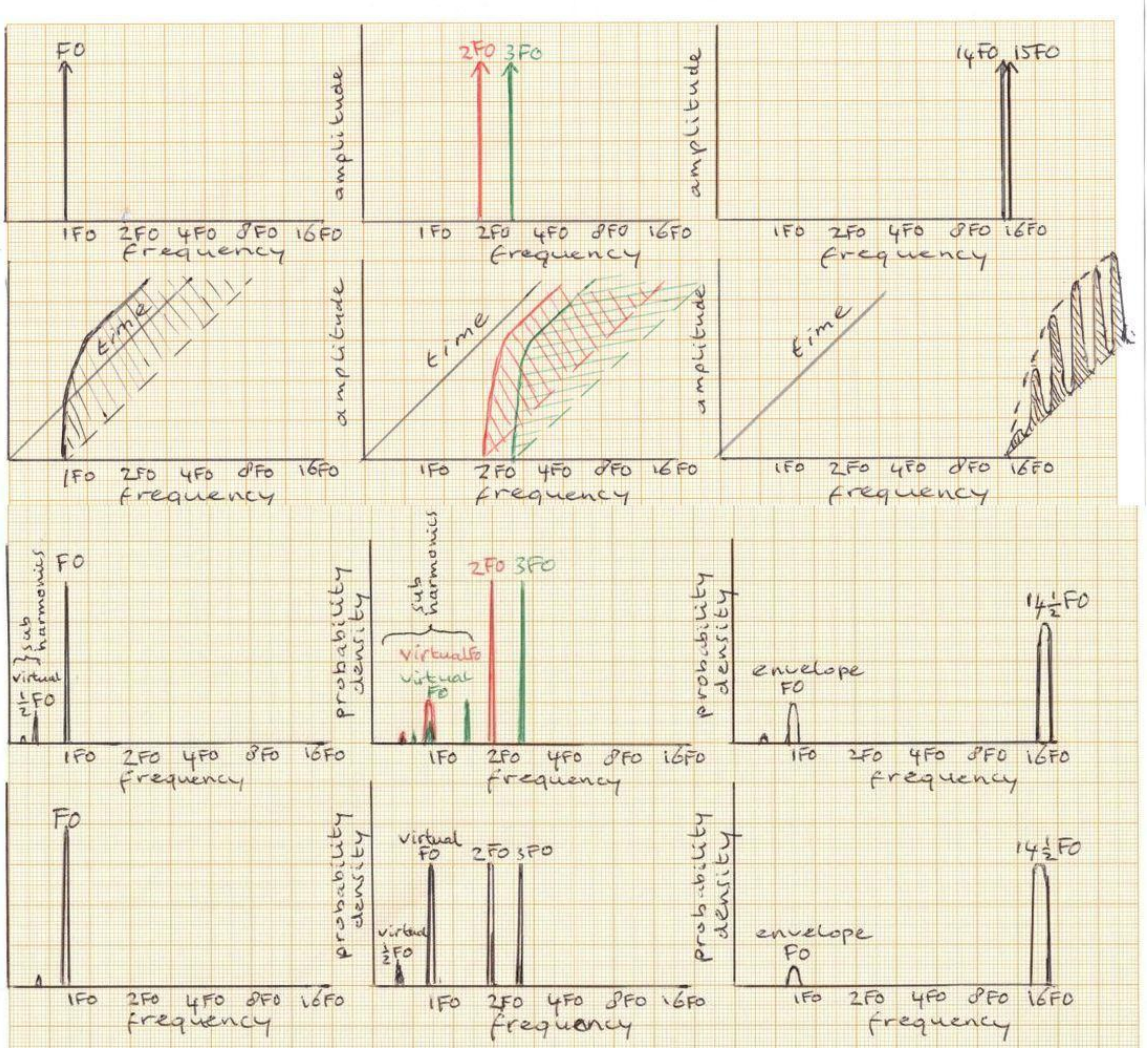
1. a pure tone that is associated with an unambiguous pitch

2. two resolved pure tones of low harmonic order that provide two spectral pitches and a competing virtual repetition pitch constructed by a central auditory processor

3. two unresolved pure tones of high harmonic order that provide a rough sounding spectral pitch and a competing repetition pitch derived from the envelope of the unresolved harmonics

The virtual repetition pitch estimation by the central processor has a natural ambiguity built into it, whenever a certain F0 fits to a harmonic complex tone F0/2, a pitch that is one-octave lower, will fit as well. An extreme case of this ambiguity can be constructed by using only octave related pure tones, F0, 2F0, 4F0, 8F0, … etc, that are smoothly introduced, and which are slowly increased or decreased in frequency (Shepherd, 1964). This leads to an extremely ambiguous pitch perception that corresponds to frequencies of F0 and 2F0. When F0 is below the auditory threshold we hear a pitch of 2F0, but the slow introduction of F0 at a certain level introduces a faint pitch with frequency F0 that gets continuously stronger. The pitch switching point between F0 and 2F0 is a bifurcation point (tipping point) but since the spectral frequencies are continuously rising or falling this bifurcation switch is masked and we create the illusion of an ever rising or falling pitch. Note that if pitch were a true one-dimensional percept, then this illusion could not have been created.


amplitude
time
time
amplitude
time


F0
2F0 3F0
14F0 15F0
amplitude
1F0 2F0 4F0 8F0 16F0
frequency
time
sub harmonics
virtual ½F0
probability density
virtual F0
envelope F0
14½F0

Three different types of "idealized" pitch processing are depicted, including the non-stationary onset, which entails a slow increase in amplitude. The left column shows the processing of a single pure tone with frequency F0 (the first harmonic) that is smoothly switched on. The first row gives the time signal including several phase-locked pulses (vertical lines) with some repetition cycles skipped. The second row gives the frequency domain representation, which is a single frequency F0 in this case. The third row gives the time-frequency representation that results after filtering in the first stage of the auditory processing. Along the time axis only the envelope of the signal is drawn. The fourth row gives the frequency representation of the statistical information that is available for the central processor. The skipping of phase locked pulses, that occurs in some repetition cycles, results in the subharmonic virtual pitches F0/2, F0/3, ... The fifth row provides the final pitch probability density, which in this case shows a clear dominant repetition and spectral pitch with frequency F0. Houtgast (1976) showed that a faint subharmonic pitch (F0/2) is even audible in a single pure tone.

The middle column shows the processing of two resolved pure tones, the 2nd (red) and 3rd (green) harmonics of F0, that are smoothly switched on. These tones are processed in two separate auditory filter bands with central frequencies that are twice and three times as high as F0. The first row gives the time signal including several phase locked pulses that are generated separately for both tones in the two separate filter bands (in red and green). The second row gives the frequency domain representation, in this case two spectral frequencies 2F0 (red) and 3F0 (green). The third row gives the time-frequency representation that results after filtering by two separate critical band filters. Along the time axis only the two resulting envelopes are drawn (red and green). The fourth row gives the frequency representation of the statistical information that is available for the central processor. The skipping of phase locked pulses, that occurs in some repetition cycles,

results in the subharmonic virtual pitches 2F0/2, 2F0/3, ..(red) and 3F0/2, 3F0/3, (green). The fifth row provides the final pitch probability density where the coincidence of the 2nd and 3rd subharmonic results in a clear virtual repetition pitch with frequency F0. Note that also F0/2 shows up as a soft repetition pitch resulting from the 4th and 6th subharmonic. If the number of harmonics is increased, the virtual pitch will start to dominate the spectral pitches.

The right column shows the processing of two unresolved pure tones, the 14th and 15th harmonics of F0, that are smoothly switched on. These tones are processed within the same auditory filter with a central frequency that is 14.5 times as high as F0. The first row gives the time signal including a number of phase locked pulses that are generated by the envelope in this filter. The second row gives the frequency domain representation, in this case two frequencies 14F0 and 15F0. The third row gives the time-frequency representation that results after filtering in a single critical band. Along the time axis only the envelope of the signal is drawn. The fourth row gives the frequency representation of the statistical information that is available for the central processor. The skipping of the envelope phase locked pulses, that occurs in some repetition cycles, results in a soft subharmonic pitch (F0/2). The fifth row provides the final pitch probability density, which in this case corresponds to a soft virtual envelope pitch with frequency F0 and a spectral pitch of 14.5F0. The virtual repetition pitch that results from the processing of unresolved harmonics is sometimes referred to as residue pitch and is less prominent than the virtual repetition pitch from resolved harmonics (compare the F0 peaks of the middle and right column of the fifth row).

Concluding Remarks and Further Research Topics

We discussed how the modeling of sound perception can be improved by looking more into the details of how it is processed by our hearing system where central cognitive processing plays an important role in the perception of loudness, pitch and timbre. Roughly speaking we can say that sounds in the world are represented by a neural triggering pattern where the loudness is determined by the number of neural pulses, while the pitch is determined by the tonotopic mapping as found in the inner ear in combination with the repetition patterns found in the neural representation. For loudness, the advanced models have not led to a significant improvement in widely accepted technical measurements, the dB(A) is still the most widely accepted measurement for loudness. However, these models have led to significantly improved audio codecs that exploit the fact that some signal parts have zero partial loudness and need not to be coded, allowing other details in the audio signal to be coded more exactly.

For pitch, we have shown that we need two dimensions, time and frequency, to describe the perception. Although they are strongly related, as shown by Fourier, they can be manipulated in such a way that we perceive two different types of pitches, one related to the loudest frequency components, and one related to the dominant repetition rates. The repetition pitch has two different flavors, a dominant one that is found by a central processor operating on resolved harmonics, and a less dominant one that is directly available in the time envelopes of unresolved harmonics. The two-dimensional character of pitch can lead to conflicting perceptions that allow us to create an auditory illusion of an ever rising or falling pitch (Shepard tone). In natural sounds a straightforward pitch algorithm that uses an autocorrelation approach, where correlations are calculated between the signal and delayed versions of that signal, will give good predictions of the dominant pitch.

For timbre, we have seen that the definition is too wide, everything that is not loudness or pitch includes a world of possible auditory cues, even for short duration stationary stimuli. Timbre is a multi-dimensional property of a sound that is difficult to quantify. As a start we must distinguish the timbre of a sound source from timbral changes that are introduced by the acoustic path. In the assessment of the timbre of acoustic sources the radiation pattern of the source, in

combination with the reflection patterns created by the listening room, determines the perceived timbre. When a recording of a source is made in a room with little or no reflections the timbre of the source cannot be restored by using an artificial reverberation which introduces the same reflections as a high-quality room because only a limited part of the irregular radiation pattern is processed. Furthermore, the broad definition of timbre would also include spatial cues, which clearly does not match our intuition of timbre as the color of a sound. In general, spatial cues are used to attribute sounds to sources that are treated as separate auditory objects. The research into timbre has therefore shifted towards parameters that can be more precisely defined, such as sharpness, brightness, clarity, tonality, roughness, fluctuation strength, etc., all of which are investigated in more precise experimental situations with headphone presentation, thus ruling out the impact of the listening room.

In modern psychoacoustics more natural signals are used and the focus has shifted towards an analysis in terms of objects or scenes. While in vision it is clear what is meant by an object or a scene, this is not as clear in the case of hearing. A first idea in auditory scene analysis is to define objects in the context of hearing as separable, identifiable *streams*. In his seminal book "auditory scene analysis" (1990) Albert Bregman starts his first chapter with the “Historical difference between auditory and visual perception” and he argues that in evolution the focus of both systems is on scene analysis and the forming of auditory streams is a first step in hearing that allows us to identify an acoustic object. A simple example of such a streaming experiment is to generate a mono signal that consists of two (harmonic) stationary tones that are perceived as a single unified sound, or object, but if one (harmonic) tone is modulated, causing audible changes in either the temporal or frequency domain envelope, we would hear two objects. The extent to which we hear two objects will be dependent on the type of modulation, but also on the spectral differences between the harmonic tones. If we perform this experiment in stereo, with one (harmonic) tone played in the left and one played in the right ear, then the effect is even stronger, demonstrating the role of spatial cues in the formation of auditory objects.

In natural scenes there is a correspondence between the visual and auditory objects, but if we set up “unnatural” audiovisual experiments, the visual domain dominates the formation of

objects. When there are no visual cues the formation of objects is largely dominated by correspondences in the temporal and/or frequency domain envelopes in the signals and by differences in the signals between the left and right ear that generate localization cues that allow us to localize objects in space. Two main types of difference are used in sound localization, timing differences and timbre differences that result from the shadowing effect of the head. Our head also introduces diffraction effects causing spectral peaks and dips that can be used in localization. These diffraction effects also lead to significant loudness and timbre differences between experiments that use headphones versus free-field loudspeaker presentation.

The identification of arbitrary physical objects is largely dominated by visual cues and the extent to which auditory cues are used is an important topic in current psychoacoustics. The main contribution of the auditory system towards the identification of objects is found in situations where the object is outside the visual field and where the auditory image thus provides a first localization cue. In this case the weakest point of auditory imaging, front back confusion, is solved by the visual system. An important research question is how to build an internal auditory representation of the object, the auditory image, from the neural representations that are found in the auditory system. In his treatise on hearing “The temporal auditory imaging theory inspired by optics and communication”, Adam Weisser (2021) states: “If the visual image occurs in space, then a hypothetical auditory image must occur in time. Then, we should expect that just as the optical object-image pair in vision can be represented as a spatial envelope [the contour of the geometrical shape of the object, JGB] that propagates between points in space (e.g., between the object origin and the image origin), so should the acoustical object-image pair of hearing be representable by a temporal envelope”. The extent to which this idea can be used to develop new approaches in auditory scene analysis is not clear as the world of non-stationary long duration signals is infinite and the modeling required for understanding our behavior is consequently getting more complex and will therefore be dominated by the experimental context.

If we extend our auditory experiments towards the use of natural speech and music, the behavior of subjects will get even more complex and be determined by the experimental context. One of the interesting questions in this domain is whether one can predict the perceived quality of

recorded speech and music. To do so we will need a psychoacoustic model that allows us to model all important aspects of our hearing system, including cognitive factors that often dominate our perception of sound quality. In addition to accurately modeling masking, important cognitive issues to be modeled are the separation of linear (timbre) and non-linear distortions, the impact of noise in silent and very soft intervals and the impact of esoteric distortions as introduced by modern speech and audio codecs. The modeling of “never seen” degradations is particularly difficult, and modern artificial intelligence approaches are often “overtrained” on known sets of degradations. It also turns out that audio quality can only be predicted to a certain extent as the quality domain is dominated by personal taste. In the accurate prediction of audio quality we need a predefined “ideal” speech or music signal and only in the case of recorded speech there is a more or less agreed upon “ideal”. Translating individual perceptions and opinions into scientific facts remains a fascinating quest.

Bibliography

American Standard Association. 1951. Acoustic Terminology.

Beerends, J. G. 1989. A stochastic subharmonic pitch model. Chapter 5 PhD. University Eindhoven. https://alexandria.tue.nl/repository/books/304393.pdf

Beerends, J. G., et al. 2013, Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II – Perceptual Model. *Journal of the Audio engineering society* 61:385-402.

Bosi, M., et al. 1997. ISO/IEC MPEG-2 advanced audio coding. *Journal of the Audio engineering society* 45:789-814.

Bregman, A. S., 1994. Auditory scene analysis: The perceptual organization of sound. MIT press.

Cheveigné, A. de, and H. Kawahara, 2002. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America* 111:1917-1930.

Goldstein, J. L. 1973. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America* 64:1496-1516.

Houtgast, T. 1976. Subharmonic pitches of a pure tone at low S/N ratio. *Journal of the Acoustical Society of America* 60:405-409.

Houtsma, A. J. M., and J. L. Goldstein. 1972. The central origin of the pitch of complex tones: Evidence from musical interval recognition. *Journal of the Acoustical Society of America* 51:520-529.

Langner, G. 2015. The Neural Code of Pitch and Harmony. Cambridge University Press.

Moore, B. C. J., B. R. Glasberg, and T Baer. 1997. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society* 45:224-239.

Moore, B. C. J. 2012. An Introduction to the Psychology of Hearing, Sixth Edition. Brill.

Plack, C. J., A. J. Oxenham, R. R. Fay, and A. N. Popper Editors. 2005. Pitch: Neural Coding and Perception. Springer.

Robinson, D. W., and R. S. Dadson. 1957. Threshold of hearing and equal-loudness relations for pure tones, and the loudness function. *Journal of the Acoustical Society of America* 29:1284-1288.

Siedenburg, K. C., Saitis, S. McAdams, A. N. Popper, and R. R. Fay Editors. 2019. Timbre: Acoustics, Perception, and Cognition. Springer.

Terhardt, E. 1979. Calculating virtual pitch. *Hearing Research* 1:155-182.

Thiede, T., et al. 2000. PEAQ - The ITU-Standard for Objective Measurement of Perceived Audio Quality. *Journal of the Audio Engineering Society* 48:3-29.

Weisser, A. 2021. Treatise on hearing: The temporal auditory imaging theory inspired by optics and communication. https://doi.org/10.48550/arXiv.2111.04338

Yost, W. A. 2009. Pitch perception. *Attention, Perception & Psychophysics* 71:1701–1715.