8.6 Loudness & timbre

Compared to an acoustic guitar, the electric guitar sounds different even if both play the same note. Pitch and tone duration may be identical, but given "species-appropriate" play, sound (timbre) and loudness will differ. This is due to the different way the partials evolve during the respectively sounding note, i.e. due to the individual, instrument-typical attack- and decay-behavior of the particular partials.

In simple models, each partial is assigned a frequency that is more or less an integer multiple relative to the fundamental (dispersion, Chapter 1.3). Per partial an initial level is specified, and also a decay time-constant that has the effect of an exponential decay of the amplitude over time, or of a linear decay of the level. However, more exact analyses show that most levels of the partials decrease according to more complicated functions; thus we have the basis for developing more complex models that define every (primary) partial as a sum of secondary partials. Other approaches are also possible – let's remind ourselves here that the quantity of mathematically equivalent models is in fact not limited.

As we reduce the sound pressure level (SPL) of all partials by the same dB-amount, the volume drops; as we turn down the level of the higher partials by turning the treble control counter-clockwise, the sound gets dull – that's well-known. It is much more difficult to answer the question *how* volume and sound depend on physical sound-parameters, and what would be the characteristics of a good or a bad sound to begin with. The volume of a tone (termed **loudness** in the following) is a monotonous function of the SPL-level (termed merely **level** in the following). Since the level is dependent on the power, "more power = more loudness" holds. Of course, we need to define this simple dependency in much more detail, because otherwise there's the danger that the result of the consideration would quickly read: a 100-W-amp is louder than a 50-W-amp … however this cannot be the general statement.

First, we need to distinguish between amplifier power and sound power (or acoustic power). The amplifier power (that would strictly speaking have to be sub-classified into effective – or active, or wattful – power, reactive – or wattless – power, and apparent power) is the power that the amplifier delivers to the loudspeaker: e.g. 10 Watts (10 W). The largest part of this power is converted into heat by the loudspeaker (Chapter 11); only about 1 – 10% are converted into sound. For example, a highly efficient guitar speaker would convert 9 W of the 10 W electrical power fed to it into heat, and radiate 1 W as sound. In the immediate vicinity of the speaker, this acoustic power is concentrated onto a small spherical surface and generates a high intensity (the intensity is the power per area [3]). Assuming that the loudspeaker generates a short sound impulse, this results in an imagined spherical wave that propagates around the speaker and increases its radius (and thus its surface) with increasing time. Since the surface grows with the square of the radius, the intensity drops with the square of the distance. This is in the free, unperturbed sound field that we now focus on first. Because the intensity is in a square-relationship with the sound pressure, the simple $1/r$-law (**one-over-*r*-law**) is applicable: doubling the distance to the loudspeaker reduces the sound pressure by half, or as equivalent: the SPL drops by 6dB (more details in [3]). As an example: an efficient guitar loudspeaker generates an SPL of 110 dB at 1 m distance given an input of 10 W amplifier power. At a distance of 2 m the SPL is therefore 104 dB, and at 10 m distance it is 90 dB. If the objective is to generate not 90 dB at 10 m distance but 100 dB, the amplifier power needs to be upped to 100 W, and for 110 db it would have to be 1000 W. So, already here we notice the limits of this model that may remain linear only with regard to the sound wave – for the loudspeaker, load-limits need to be respected, the efficiency is of course power-dependent, and the speaker will die on us when overloaded.

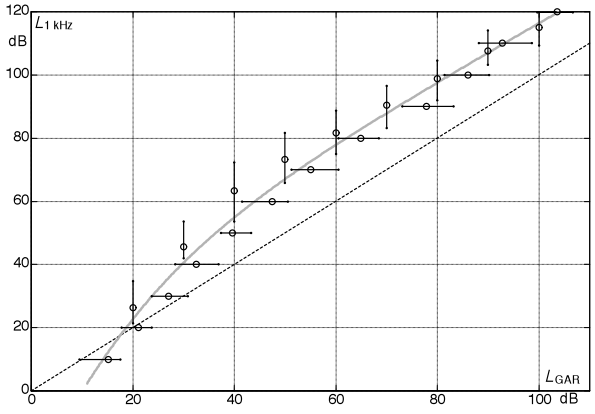
In the open, given unperturbed sound propagation, the level decreases by 6 dB per doubling of the distance. This fact is usually noticed with horror by the guitarist playing an open-air concert for the first time: that amp that was way too loud every time back in the club now is hopelessly drowned out all of a sudden. In the open, the reflections from the walls and the ceiling are missing – they lead to the sound reaching the listener not just once but (as echo) repeatedly. In the room, a superposition of free sound field and diffuse sound field is generated, with the free sound field dominating close to the loudspeaker, and the diffuse sound field dominating further away. The border between the two sound fields is represented by the diffuse-field-distance (also called reverberation radius). It amounts to a few meters in regular rooms (more precise information is found in [3]). Beyond the reverberation radius, the SPL stays independent of the location **within the room**; or so says simple theory. For the above example, this would imply: if the reverberation radius were 5 m, we would get (for 10 W input, and calculated starting from the speaker) at this distance a decrease in SPL down to 96 dB. In the remaining room ($r > 5$ m) the SPL would be 96 dB independent of the location. Of course, additional factors such as beaming effects, the actual geometry of the room, and the distribution of reflectors and absorbers would have to be considered – but this would go beyond the scope intended here. This example is to show that – before we start thinking about sound volume – sound source and room need to be looked into: which electrical power do we have, what is the efficiency of the loudspeaker, into what kind of room does the speaker radiate, and at last: where is the listener located? The SPL developing at the ear of the listener is the result of all these parameters, and from it – not just from the power of the amplifier – we can obtain indications for the generated loudness.

Psychoacoustics investigates the connection between SPL and **loudness**. Nowadays there is a standard for that – which is not undisputed. How loud you perceive a sound to be is a highly personal matter that is still interesting to science. And so we inquire with test persons (subjects) about their impression of loudness, we have them give categorical assessments (soft, loud, very loud), we make them perform magnitude estimates (double as loud as the reference sound), and let them determine thresholds (now the sound becomes audible). It is to be expected that not all human beings hear exactly the same thing, and neither that one and the same person will give the exact same response when asked again. This insight, however, will not be of much help – the psychoacoustician will want to know by how many dBs the level needs to be increased in order to make the subject perceive double the loudness. It is right here where the problems start: in fact, there is a multitude of experiments targeted to find out exactly that – but unfortunately there is also a multitude of answers or resulting models, not all of which generally correspond. Estimating the doubling or halving of loudness is a frequently practiced experiment from which the whole scale from *inaudible* up to *too loud* is assembled. Hellbrück [1993] has addressed this topic extensively and describes both the pros and the cons of the standardized loudness model of Stevens/Zwicker: power law, or exponential function? Stevens and his sidekicks had the subjects judge loudness relationships, and therefrom derived the **loudness power law** – it teaches that loudness depends on SPL according to a power law. In order to **double** the loudness of a 1-kHz-tone (in the level range > 40 dB), the **level needs to be increased by 10 dB** according to this law. Accordingly, upping the level by 20 dB corresponds to quadrupling the loudness, and +30 dB will match eight-fold the loudness. Recalculating this in terms of amplifier power: to double the loudness (and given linearity), the amplifier power needs to be increased by factor of 10 ten! Thus, compared to a 10-W-amp, only a 100-W-amp will be double as loud, and not a 20-W-amp. Still, a lot needs to be added here. To start with, the above law is applicable a priori only to a 1-kHz-tone. Then we find in Hellbrück's book the lovely but unsettling citation: *the possibility should be considered that the whole of the sone-scale is a pure artifact from psychometric methods that have been applied inappropriately and mindlessly.*

Sone, that's the unit for loudness. Mindlessly investigated? Let's not go there – psychologists and engineers will probably continue to bandy that ball for further decades. If we don't want to abort everything with the quite unsatisfactory insight that, due to the individual scatter, establishing an exact functional correspondence will not be possible, then what remains is forming statistical **mean values**. The difficulty is shown by an example from the beginnings of calculating loudness: during some auditory experiments it was noticed that broadband noise is much louder than a 1-kHz-tone although both have the same SPL value. Apparently, the SPL-value is unsuitable as a measure for the perceived loudness, leading to this question: by how many dB the two sounds will be different if both are adjusted to the same loudness? For the experiment described in [12], a special noise is used, the so-called *uniform exciting noise* (UEN) that may be imagined approximately as pink noise (kind of similar to a spoken long, slightly dark "sh"). One possibility to estimate the loudness is to present the 1-kHz-sinetone (e.g. at 80 dB) and ask the subject to adjust the level of the noise such that both sounds (presented alternately, not concurrently) are perceived equally loud. The reverse approach would also be possible; the noise is presented and the 1-kHz-tone is adjusted to the same loudness. Surprisingly, different values result from the two approaches even if the unavoidable small scatter is averaged out. There clearly is a **systematic deviation** (on top of the stochastic one): the adjustable magnitude is adjusted too high. For a presented 79-dB-noise, an adjustable tone is set to 90 dB, but for a presented 90-dB-tone, the noise is adjusted to 78 dB to be equally loud.

The measurements shown in **Fig. 8.36** give three results:
- For the two sounds to be subjectively of the same loudness, the level of the 1-kHz-tone needs to be in part more than 20 dB above the level of the noise.
- The results are dependent on the measurement procedures.
- The scatter is considerable.

In Fig. 8.36, the scatter is indicated as **interquartile ranges**; these represent 50% of the measurement values, with the values "above" and "below" discarded. As an example: 50 % of the subjects (the "middle" half) adjust the level of a 1-kHz-tone to an SPL of 83 … 97 dB for equal loudness with a 70-dB-noise, 25% of the subjects set the level to smaller than 83 dB, and the remaining 25% adjust the level to more than 97 dB. Additionally, the median value is given as a dot. We can unequivocally take from this experiment that noise is perceived louder that a 1-kHz-tone of the same level; however, the quantitative evaluation is subject to considerable scatter, and the latter moreover is dependent on the adjustment method. Psychoacoustics factors this in by defining two different loudnesses: a standard loudness level, and an object loudness level (that of the test sound). N.B.: the loudness comparison with the 1-kHz-tone historically was the first method to determine the loudness of any sounds, i.e. objects, via using a standard, i.e. the 1-kHz-tone).



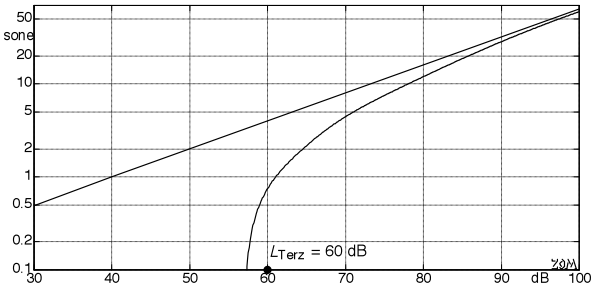


Fig. 8.36: Loudness comparison noise/sinetone (left), loudness with/without background noise (above), [12]. "GAR" = UEN, "Terz" = third octave band.

Keeping constant the level of the standard and making variable the level of the object (i.e. in this case the noise-level) yields the **object loudness**. Conversely, making variable the level of the 1-kHz-tone we obtain the **standard loudness**. The value interpolated between the two curves (the grey line in the figure) is called **interpolated loudness level** in older literature. The tem loudness level was introduced in order not to always have to talk about the "level of the equally-loud 1-kHz-tone" – rather, the **loudness level** with the unit **phon** is specified; the numeric value is that of the level of the 1-kHz-tone of the same loudness. Thus, if noise is perceived as equally loud compared to a 90-dB-tone (of 1 kHz), then this noise has a loudness level of 90 phon. This now makes the loudness sensation quantifiable – with numeric values that are difficult to interpret, though: 80 phon are not double as loud as 40 phon but 16 times that loudness. For this reason, additionally the **loudness** (measured in **sone**) was introduced. A 1-kHz-tone of 40 dB level serves as a reference point delivering the loudness of N = 1 sone. Since a level increase of 10 dB has the effect of doubling the loudness, 2 sone match 50 dB, 4 sone match 60 dB, 8 sone match 70 dB, and so on. Below the level of 40 dB, this correspondence is not valid anymore: in this range already smaller level changes have the effect to doubling the loudness.

The upper line in the right-hand section of Fig. 8.36 shows the relation between the level of the 1-kHz-tone (abscissa) and loudness (ordinate). Again: this is only for 1-kHz-tones – other spectral compositions necessitate other curves. Another prerequisite is that the 1-kHz-tone is presented by itself i.e. without other sounds being present. If the latter are presented concurrently, the loudness of the 1-kHz-tone may be **partially masked** i.e. reduced. The lower line in the right hand graph of Fig. 8.36 shows such a scenario: besides the 1-kHz-tone, a pink noise with the third-octave level of 60 dB is presented at the same time. If the 1-kHz tone has a high level (e.g. 90 dB), the two curves barely differ – the noise has little influence on the loudness of the tone. However, as the level of the tone is reduced (e.g. to levels below 57 dB), the tone becomes altogether inaudible because it is "masked" by the noise. Thus, when there is a masking sound present, the loudness grows more strongly with the level compared to the situation without masking noise.

For the practical musical performance situations we can learn from these relations that small variations in the sound power (e.g. +10%) are insignificant for the loudness perception. If the power of an amplifier is increased from 40 W to 44 W (and given a proportional change in sound power), we will – as a rule – not perceive a change in loudness. According to common practice, the just noticeable difference for amplifier power is estimated at about +50%. The difference between a 40-W-amp and a 60-W-amp is just about noticed – while doubling the power is clearly perceivable. Any musician deliberating whether to buy a 50-W-amp, or "for good measure" rather a 60-W-amp should be particularly weary of the efficiency of the loudspeaker. That is because, for example, a Celestion G-12-M is rated in the datasheet at 100 dB/1m while the G-12-M Greenback is rated at 97 dB/1m. Purely in terms of figures, the greenback requires double the power in order to generate the same SPL as the G-12-H. How these datasheets were established, is of course an entirely different story, and that (besides the loudness) the color of the sound (the timbre) plays a pivotal role – well, that opens yet another can of worms. It would go too far here to elaborate on all parameters that weigh in when determining loudness and timbre; those interested are recommended to read up in Fastl's book "Psychoacoustics" [12] – on 462 pages, it represents a comprehensive overview of the most important basics and models. The literature list in the appendix gives further info on related books.

The **color of sound** (timbre, sound- or tone-color) is the last sound parameter that we visit here. For many readers, it will be the most important one – but unfortunately it is also the most complex one. The sound-color – "the sound" – is being evaluated according to highly individual criteria, and trying to establish a model to calculate it always leads to failure. Of course, the sound-color depends on the sound spectrum, but already the metrological determination of the latter will be unsuccessful unless very simple sounds are analyzed. Harmonically complex tones are one thing, but a guitar solo played against a full accompaniment is another. Seeking to attribute roughness or fluctuation strength (based on modulation-indices and -frequencies) to a sound is futile because this cannot be determined in the guitar solo. Every spectral analysis may optionally be interpreted as a spectral weighing with the complex transmission function of a bank of band filters, or a convolution in time with the impulse responses of these filters. Bandwidth and impulse response cannot both be limited to a rectangular range, though, and thus every spectral analysis will lead to spectral and time-related **leakage**. The term spectral leakage intends to express that even the spectrum of a sine-tone is not measured discretely at on point of the frequency scale but as a continuously distributed spectral density. A Fourier series expansion is only possible in special cases (e.g. when the signal period is known), but this is meaningless in practice. Because the spectrum of the pure tone is presented in a broadened ('smeared') fashion, it is difficult to separate closely adjacent notes. Since spectral and time-related blur are reciprocal to each other, it would be possible to extend the duration of the analysis and thus to decrease the spectral leakage – but then the time-related leakage (describing the broadening – 'smearing' – along the time axis) increases. In concrete terms: if 1 Hz separation is desired in the frequency domain, the blur in the time domain is 1 s. The exact relation between the two quantities does not need to be deduced here[♣], for orientation $\Delta t \cdot \Delta f = 1$ suffices. If the analysis-blur along the time-axis is to be reduced to 10 ms, the spectral blur increases to 100 Hz. If we seek to, for example, extract from a musical piece the partials of the lead guitar, and therefore subject the wav-file to a DFT-analysis, it will be very difficult to decide which of the lines belong to the guitar, and which should be traced to other instruments. It may be possible in some cases, but fail in others.

Particular significance needs to be assigned to the "**attack**" (the onset of the tone). Many instruments can correctly be identified only via the structure of their attack; suppressing the first 100 ms tampers greatly with the sound. A good time- and frequency-resolution is desirable in this time range if the structure of the partials is to be meaningfully detected. The spectral and time-related leakage effects cannot be seen as errors per se; rather, they are kind of analysis-immanent artifacts. A Blackman-Harris-window is not more wrong or more right than a Kaiser-Bessel-window – it is just different. That, however, also means that one window modifies the structure of the partials differently compared to another window. If guitar tones were composed of harmonic partials of infinite duration, the analysis would be relatively simple. But they're not: the frequency relations of the partials are not integer multiples but they are spread out, and in addition they are slightly shifted (due to the frequency-dependent bearing impedances, Chapter 2). The amplitudes of the partials are not constant over time, and they do not decay according to simple functions, either. Moreover, the almost always present other instruments weigh in, as well, because pure solo-playing of any length of time does not occur much. Spectral analyses can certainly help to establish orienting impressions: are only odd-numbered partials dominant, how strong is the fundamental, do strong partials already stop at 1 kHz or do they extend up to 5 kHz? However, already with the evolution with time, with the fluctuations of the partials, it does get complicated, and the results of the analyses become dependent on the parameters of the analysis filters to a large extent.

♣ See e.g.: Zollner, M., *Frequenzanalyse*, Hochschule Regensburg, 2009;
or: Zollner M., *Signalverarbeitung*, Hochschule Regensburg, 2009.

Starting not with the spectral analysis of a whole ensemble, but recording and analyzing the sound of a single instrument played in the anechoic chamber, will usually result in spectra like those depicted in **Fig. 8.37**. They give the insight that e.g. a clarinet generates predominantly odd-numbered partials – this even being in good agreement with the wave mechanics of this aero-acoustic resonator (open on one side, i.e. "gedackt" pipe). The graphs on the left and in the middle stem from different books – both are supposed to show the spectrum of a clarinet. The graph on the right shows the spectrum of a cello. That the two clarinet spectra differ so much is not necessarily the result of grave measurement errors but easily due to the variability of this sound. Indeed, there is not "the" tone of a clarinet, and just as little is there "the" spectrum of a clarinet. We may be able to recognize characteristic differences in the cello-spectrum in Fig. 8.37 compared to the clarinet-spectrum, but these become meaningless in view of the spectral differences between the clarinets. Conclusion: **single spectra** hold little validity.

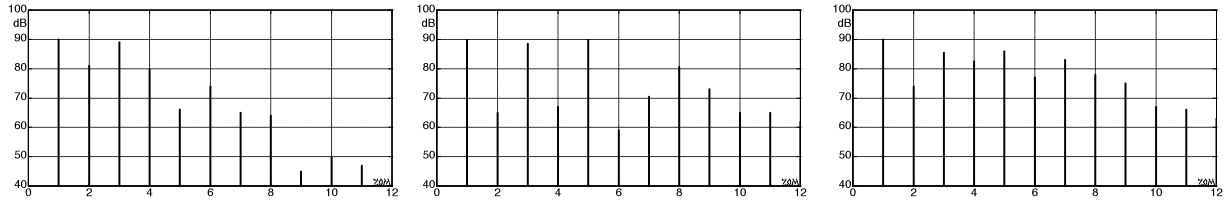


Fig. 8.37: Sound-spectra of some instruments: clarinet, clarinet, cello.

Regarding the sound of the violin, Dickreiter[♣] elucidates: *the build of the partials of the violin is relatively irregular, i.e. it changes from note to note. The reason is found in the complicated resonance properties of the resonance body that strongly influence the material characteristics and the construction.* Thus, the spectrum of a d^1 may look entirely different from that of a g^1, and of course the relative position of violin and microphone plays a role since the radiation happens with a frequency dependent directionality. The first "electronic organs" sought to imitate the sound of specific instruments by generating periodic tones with a spectrum that had a supposedly instrument-typical envelope – such as e.g. the cello-spectrum from Fig. 8.37. It was more or less accepted that the resulting sound was only very remotely reminiscent of a cello; to sound "kind-of-electronic" was probably o.k. The main criticism was: the sound of simple organs is too "sterile"; it does not live – the instrument-typical beats are missing. The latter then were subsequently included via amplitude- and frequency-modulators (vibrato, tremolo), but again the result sounded artificial again because the effect was not relative to the partials but global. Only with the emergence of the sampling keyboards and the availability of huge solid-state memories could instrument sounds with an acceptable degree of naturalness be synthesized.

It's not that the spectral representation would be entirely unsuitable to visualize instrument-typical characteristics – spectra can fully describe signals. It's just that the information included in a single spectrum is too limited to already extract the instrument-typical from it. Typical is e.g. the presence of accompanying sounds that nevertheless contribute to the recognition of an instrument. The hammer-noise of the piano ("plock"), the blowing-noise of the flute, the scraping noise of the violin bow, the "squeak" in the horn attack, the impact of the strings of a bass on the fingerboard (drastically emphasized in the slap-bass style) – these are examples for such additional sounds, and there are many more. Typical are spectral maxima (**formants**) that are at a fixed frequency, or move along dependent on the fundamental; typical are time-related fluctuations of partials.

[♣] Dickreiter M.: Handbuch der Tonstudiotechnik, Saur 1979.

All these characteristics aid the hearing system to categorize sound-colors, and to eventually allocate them to specific instruments. This then is done on the basis of learned knowledge – those who never have consciously heard an oboe will not recognize it, and only hear a strange nasal tone. Even those who in fact know how an oboe sounds will find recognizing the instrument difficult if one period is cut out from the oboe-tone and periodically repeated (*looped*). An oboe-typical spectrum is created – but it's out of typical context. In the auditory signal analysis (i.e. when we listen) the arriving sounds are automatically compared with known patterns stored in our memory. If the presently heard sound and the memorized one more or less match, the decision is made: sounds like an oboe, and/or like a musical instrument, and/or nasal, and/or dangerous, or whatever else could be found in the match. We can imagine the sound-color identification as a multi-stage process: in a first hierarchical stage, the inner ear determines the time-variant spectrum of the non-masked partials, i.e. the momentary sound-color – customarily described by *one single* spectrum. However, since (as taught be signal theory) a spectrum cannot be ascertained for a point in time but only for a time-range, the term *momentary* must not be taken too narrow a view on. The speech analytic evaluates sections of about 10 – 30 ms length, and it indeed is a powerful tool; as it is applied, it is often underlined that for the evolution of the hearing system, analyzing speech was even more important than analyzing music. That does sound convincing – but it does not mean that each and every musical analysis has to comply. For percussive sound, shorter durations of analysis may be purposeful, and for very low bass-notes longer ones, as well (because it allows for a finer frequency resolution). Still, an analysis-duration of 20 ms is quite workable as an orientation value; this means 50 spectra per second. These of course are not all identical but time-variant. On the basis of this spectral ensemble, the next-higher sound-color determination can happen which already yields more than just a "sound kinda like aaa". It could e.g. yield "sounds like a trumpet". In order for this already rather complex analysis to be successful, typical patterns about tone-onset, fluctuations, duration and decay need to be memorized. If the deviations are too big, the recognition algorithm fails. Cutting off the first 100 ms of a note will substantially lower the recognition rate; apparently already this short section includes important instrument-specific information that is not available in the later parts of the evolution of the note. Alternatively (and this is something we must not overlook), the cut sound will not be matched to the correct instrument because nothing about it has been learned yet (i.e. no corresponding patterns have been memorized).

In the processing stage still higher up, the evaluation steps can start that lead to the verdict: "sounds like Josh Redman", or "That be Hendrix on the Strat". Such judgments are, however, not part of the present reflections … so let us return to the color of sound, the timbre, and its signal-theoretical basis. We have already known for some time what the color of sound is NOT, and from this the following exclusion-definition originated: color of sound is that which remains if loudness and pitch are abstracted from. Alternatively, according to an old Acoustical-Society-of-America-definition: *color of sound is the perception attribute that still distinguishes two sounds although loudness and pitch are equal*. Somehow that feels like a trash-can-esque definition into which we can throw everything that cannot be defined precisely. Borrowing from optics helps to move along a bit: like we can objectively define visually perceived colors on the basis of spectral intensity distributions, the color of sound in auditory perception can be ascribed to the envelope of the sound spectrum. Like a picture consists of strung-together locally distributed color spots, the tone of an instrument consists of momentary timbres strung-together sequentially in time. We need to allow for the fact that this comparison will arrive rather quickly at its maximum load and hit a wall – the two sensory channels do, after all, exhibit strong differences besides some similarities.

In order to explain the possibilities and limits of the spectrum-based analysis of tone color, a dyad shall serve: two added-up sine-tones (300 Hz, 312 Hz) of equal level that are abruptly switched on at t = 100 ms (**Fig. 8.39**). The time-function would therefore be:

$$x(t) = \sin(\omega_1 t) + \sin(\omega_2 t) \;=\; 2 \cdot \sin\left(\frac{\omega_1 + \omega_2}{2} \cdot t\right) \cdot \cos\left(\frac{\omega_1 - \omega_2}{2} \cdot t\right) \qquad \text{Beating}$$

Already this simple example exemplifies that there is more than one possibility of representation for every signal: the **dyad** may either be seen as the **sum** of two tones, or as the **product** of two other (!) tones. Instead of adding a 300-Hz-tone and a 312-Hz-tone, it is also possible to multiply a 306-Hz-tone by a 6-Hz-tone. A spectral analysis merely and always disassembles the signal into its additive components, and not into its multiplicative components, showing one 300-Hz-line and one 312-Hz-line in the spectrum. The 6-Hz-envelope that is so nicely revealed in the time function (Fig. 8.39, upper left) remains hidden in the spectral analysis. Even the 300/312-Hz-pair-of-lines will only be represented as two separate lines for suitable analysis parameters – and since there is an infinite number of parameter-variants, there will be an infinite number of spectra.

The long-term spectrum identified for $-\infty < t < \infty$ is pointless; rather, the **spectrogram** obtained by shifting a short window-section is required (**Fig. 8.38**). In the left-hand graph, a rectangular evaluation-window is shown; it is slid across the signal as a multiplicative weighing (over time). From the signal weighed this way (shown at ***b***), the DFT-spectrum is calculated as a function of the time-shift. Since undesirable jumps occur at the window-borders for this type of window, the rectangular window is not applied in practice; windows with a rounded-off shape are customary.

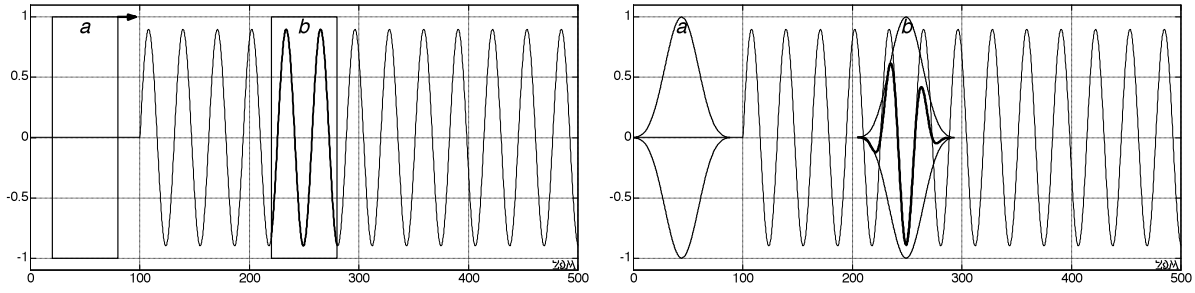
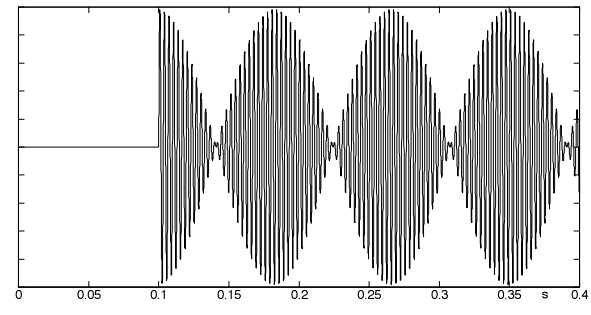
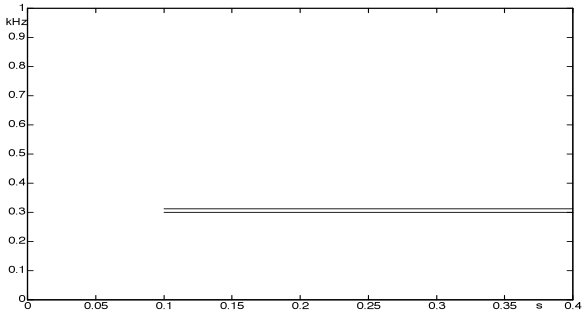
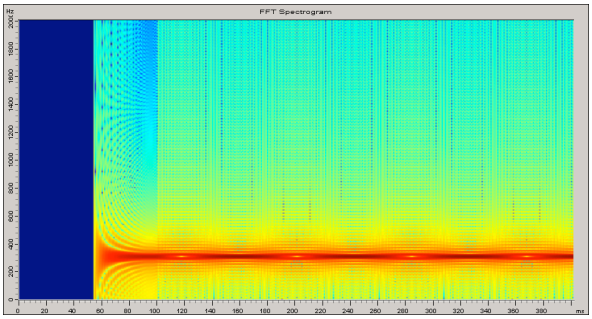


Fig. 8.38: Time-function of a sine-tone; with two different weighing windows.

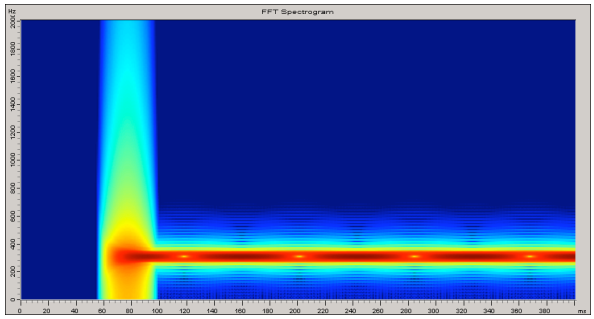
However, a fundamental problem still remains with the window-weighing as shown on the right (here: Kaiser-Bessel window): the spectrum is determined based on the windowed (i.e. modified) signal. **Fig. 8.39** shows – for the two-tone signal mentioned above – spectrograms derived with different windows. The signal was identical for each spectrum; the differences stem exclusively from the different analysis-parameters. The window-length is specified by the point-number N, a frame-length of 46 ms belongs to N = 2048. The time specified as abscissa in the color-spectrum marks the beginning of the window. Since the width of the latter is not 0 but e.g. 46 ms, we understand why the analysis pushes the start of the dyad ahead e.g. to the 54-ms-point – although both sine-tones are switched on only at the 100-ms-point! At exactly this time shift, the start of the signal falls into the rectangular window, and therefore the corresponding spectrum also starts from 54 ms. Increasing the number of points to 4096, the window-length grows to 92 ms, and the spectrogram (linked to the rectangular window) starts at 8 ms.



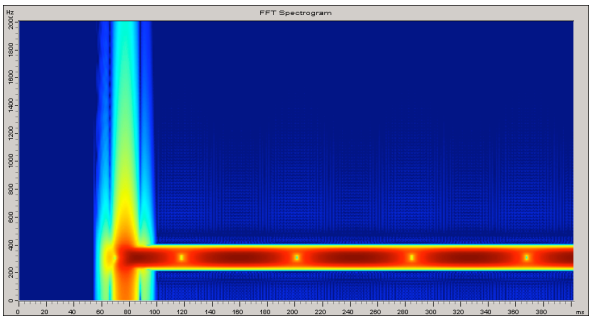




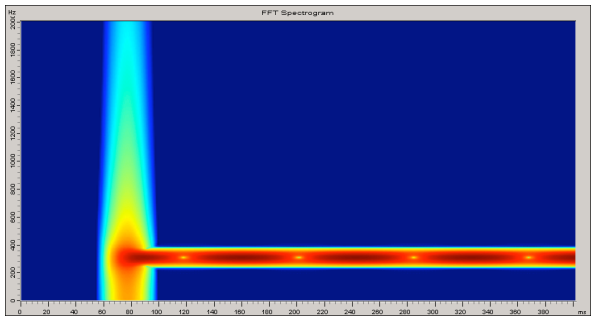
Rectangular window, $N = 2048$



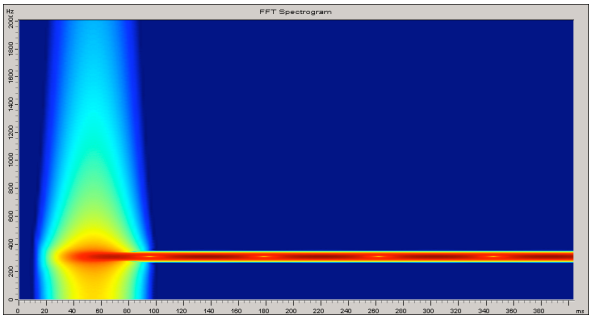
Hanning-window, $N = 2048$



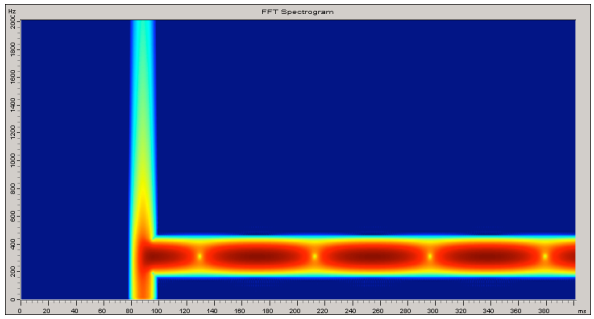
Flat-Top-window, $N = 2048$



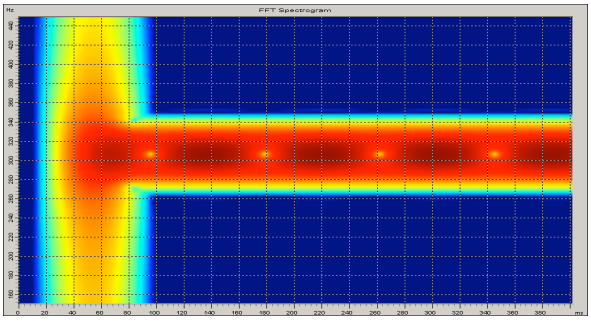
Kaiser-Bessel-window, $N = 2048$



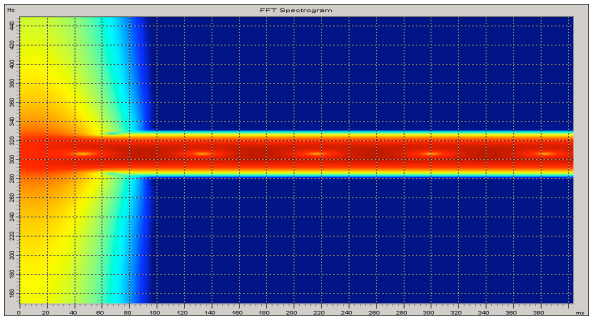
Kaiser-Bessel-window, $N = 4096$



Kaiser-Bessel-window, $N = 1024$

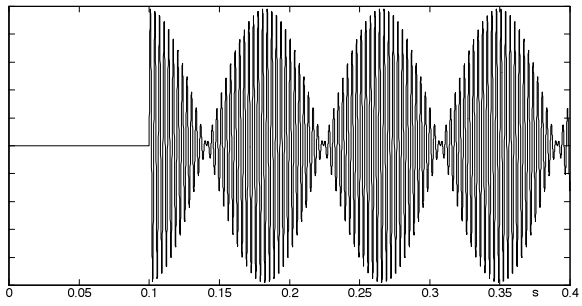


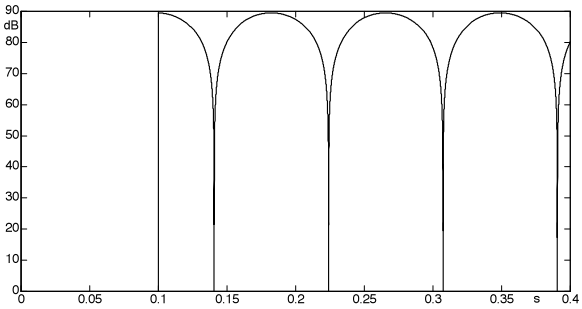
Kaiser-Bessel-window, $N = 4096$

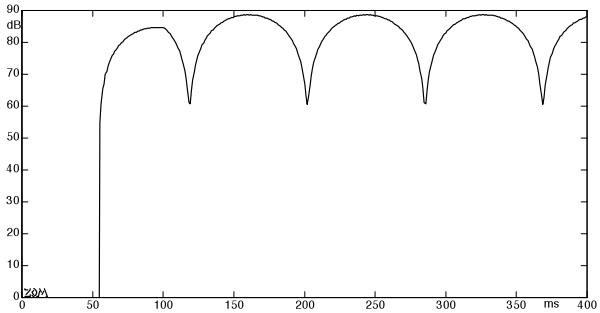


Kaiser-Bessel-window, $N = 8192$

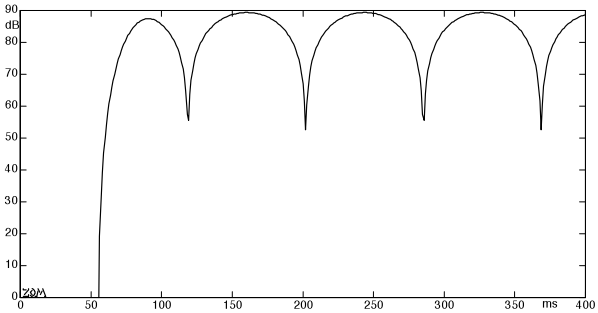
Fig. 8.39a: DFT-Spectrograms of an abruptly switched-on beating (300 Hz / 312 Hz), $\Delta L = 90$ dB.



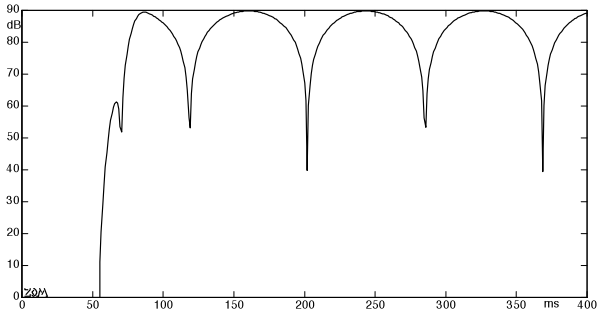




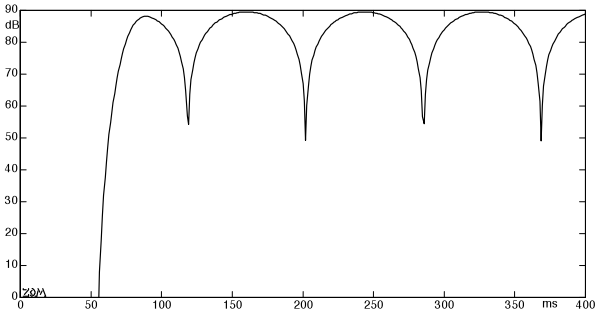
Rectangular window, $N = 2048$



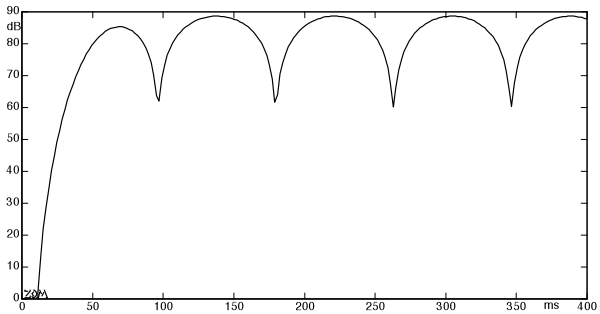
Hanning-window, $N = 2048$



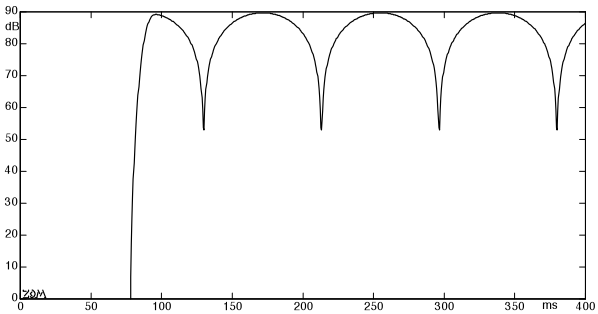
Flat-Top-window, $N = 2048$



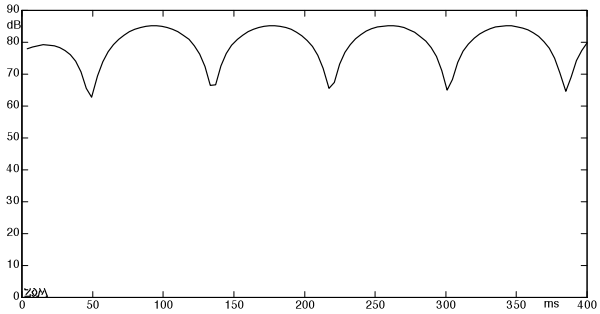
Kaiser-Bessel-window, $N = 2048$



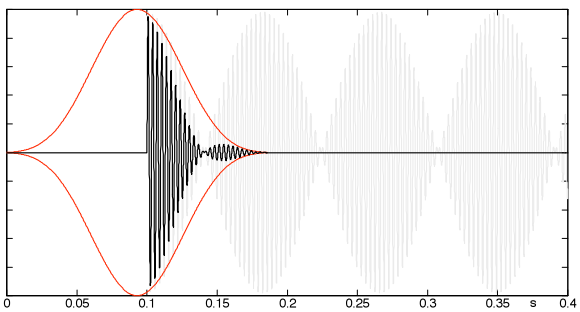
Kaiser-Bessel-window, $N = 4096$



Kaiser-Bessel-window, $N = 1024$



Kaiser-Bessel-window, $N = 8192$



Kaiser-Bessel-window, $N = 8192$, time-domain

Fig. 8.39b: DFT-level graphs (at 306 Hz) of an abruptly switched-on beating (300 Hz / 312 Hz).

It would be possible to scale the time axis such that $t = 0$ specifies the *end* of the window; in that case the corresponding shifts would show up at the end of the signal. Just to be clear: it again needs to be emphasized that this is not a software error of the analysis program, but a system-immanent artifact of all spectral analyses. Depending on the window-length (= on the impulse response of the filter), the analyzed signal becomes longer. Moreover, changes result in the direction of the ordinate, as well: the switching-on click as vertical streak, and the spectral leakage as vertical broadening of the spectral lines. In fact, from 100 ms there should be two lines running in parallel towards the right, as shown in the top-right graph; instead *one single* streak is shown. The simple reason: for $N = 2048$, the analysis bandwidth is too small, and the two lines cannot be represented separately. If we take the bandwidth as the reciprocal of the window-width, we obtain the bandwidth of $\Delta f = 22$ Hz – that is too broad for a line distance of only 12 Hz. For the Kaiser-Bessel-window (**Fig. 8.40**) used in the following, we moreover need to consider that the effective duration is only about $1/4^{th}$ of the frame length; and that the effective bandwidth therefore will be about four times that of the rectangular window♣.

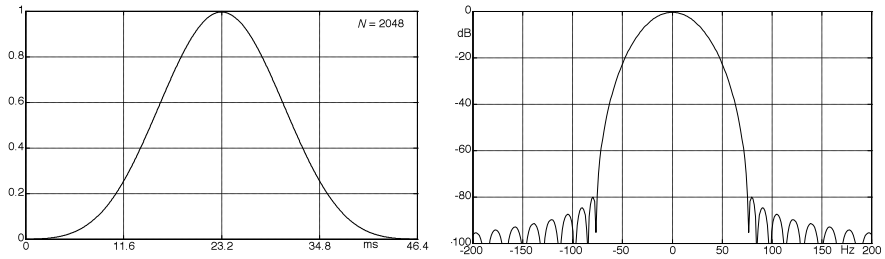


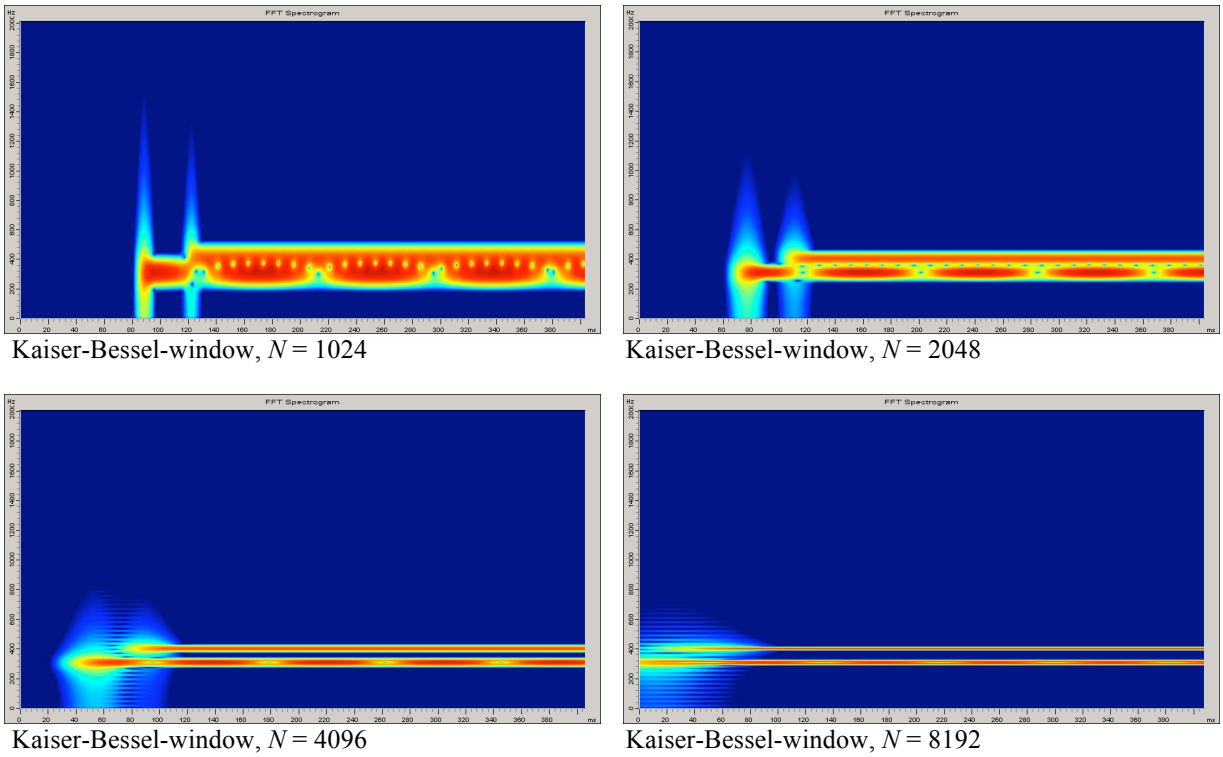
Fig. 8.40: Time-function (left) and spectral function of the Kaiser-Bessel-window, $N = 2048$, sampling frequency: fa = 44.1 kHz.

If indeed time-related and spectral leakage have effects in every spectral analysis, it stands to reason to ask whether the like would not appear also within the hearing process – after all, the signal is broken up into its spectral components there, as well. And sure, leakage will of course be present, too. However, because the **auditory filters** are adaptive and non-linear, we cannot specify *one* bandwidth and *one* attack time – things are more complicated. Too complicated for the present explanations that are merely intended as an overview, and therefore reference is made to specialist literature, e.g. Fastl's "Psychoakustik" [12, available also in English language]. The hearing system processes two tones of large frequency distance in separate channels, while tones close in frequency are jointly processed. The two-tone signal mentioned above cannot be separated into its two components by the auditory system, and one tone of quickly fluctuating loudness is heard – i.e. as product, not as sum. We hear something that does not actually exist in the spectrum: a 306-Hz-tone! Already this simple example proves how difficult it can be to extrapolate from a spectrum to the auditory perception. It is not entirely impossible; the parameters of the analysis can be adapted, after all. Therefore **Fig. 8.39** includes different analyses, with varying window-types and -lengths. All show the **switching-click**, to start with. The longer the window, the longer the switching click. It has to be that way: if, during the shifting of the window, the signal-start just about falls into the window, it is only an impulse of very short duration that is analyzed – the spectrum of which is necessarily broad-band. The more the window is shifted beyond the signal-start, the longer the signal to be analyzed (windowed), and the more narrow-band the spectrum. Is the switching click audible? No! In any case not as the figures would let us assume. It therefore is purposeful not to show the color-spectrum with a dynamic of 90 dB (as is the case in Fig. 8,39) but with only 40 dB: visual and auditory impressions are a better match that way.

♣ We will not investigate in detail here what is to be understood by the term „effective".
More details may be obtained from: M. Zollner, *Signalverarbeitung*, Hochschule Regensburg, 2009,
as well as from: M. Zollner, *Frequenzanalyse*, Hochschule Regensburg, 2009.

We now take a look at the fast fluctuations that can be clearly seen in the time-function. They also appear in a time-section of the spectrum, in the so-called **slice** (level over time with fixed frequency, Fig. 8,39b). Forming the logarithm of the envelope yields the curve shown in the graph at the upper right, and the evaluation of the DT-analysis yields the graphs below. Again it is clear that the time-related leakage has the effect of very differently shaped level curves – depending on the window-type and -duration. Thus we retain: the **DFT-analysis** delivers a multitude of different spectra that – to begin with – allow for only few conclusions regarding the perception of the sound. Supplementary algorithms enable modeling of hearing-typical assessments (auditory critical-band filters, contouring algorithms, spectral and time-related masking), but the scientific investigations have yet to arrive at a true breakthrough.

The two-tone signal analyzed in Fig. 8.39 already revealed the fundamental issues found in any spectral analysis. Yet, it is a very simple signal – instrument tones are of considerably more complex build, not to mention chords or tutti-sections. Compared to the latter, the **triad** analyzed in **Fig. 8.41** is still rather simple: three added-up sine-tones of equal level but switched on at different times. The 300-Hz-tone and the 312-Hz-tone are switched on at t = 100 ms, and the 400-Hz-tone comes in at t = 134 ms. Analysis is again done using the Kaiser-Bessel-window, the level dynamic in the figure is, however, reduced from 90 dB to 50 dB (compared to Fig. 8.39).



Kaiser-Bessel-window, N = 1024 Kaiser-Bessel-window, N = 2048

Kaiser-Bessel-window, N = 4096 Kaiser-Bessel-window, N = 8192

Fig. 8.41: DFT-spectrograms of a triad (300 Hz / 312 Hz / 400 Hz). ΔL = 50 dB.

The tone onset blurs as the window length increases, but the spectral separation improves in turn. The latter does not need be all that great, though – because with this triad, again only one single tone is heard. Not a sine-tone but rather a lively bubbling tone-mixture – with *one single* pitch. Only when listening repeatedly, one could also tend to hear an oscillation between two pitches … but certainly not anything like what the analysis done with N = 8192 would suggest.

As powerful PCs became available, the desire developed for sound analyses to – at last – depict "the correct" spectrum, meaning not just 512 lines but 4096, or even 16384, for good measure. The latter number implies, however, that the sampling window (at a sampling frequency of 44.1 kHz) has a length of 372 ms, which is too long compared to the hearing system even when applying the (shorter) effective length. For sound analysis, N = 4096 represents a tried and tested compromise that offers the basis for supplementary DFT-analyses and post-processing. The latter is urgently required: the 2k-analysis shown in Fig. 8.41 gives the impression of two sound-parts starting at different times. Objectively seen, this is indeed correct: a beat from 100 ms, and a sine-tone from 134 ms. Our hearing system, however, does not care: it perceives one single tone-onset and not two. Even when the two partial sounds start with a delay of 70 ms between them, they are not heard separately in time. The simple reason is that the beating in the dyad impedes the recognition of the time-structure. Only from an offset of about 100 ms, the additional tone coming in with the delay in this example (!) is recognized as such (compare to Chapter 8.5, though).

Not to stick exclusively to synthetic tones, let us now turn to a real guitar tone: **Clapton's intro to "Stepping Out"**. The guitar plays by itself a number of times – this facilitates the spectral analysis a lot. **Fig. 8.42** shows spectra and time-functions: in the upper two lines of graphs those of a G_3, and below for a C_4. That's four times that "same" G_3, but with considerable differences! Clapton's sound may not be described with one single spectrum, after all – and that is the same for J.H., R.B., G.M. and all the other big names: virtuosity implies change, and that holds for the spectra, as well.

Still, we of course can wring a few commonalities from the G_3-spectra: they all feature a gap between 1 and 1.5 kHz, and a spectral maximum between 1.5 und 2 kHz. This is the range where the (second) formants of the vowels "ø" and "y" (using the definitions of the international phonetic alphabet, IPA) reside, so these tones can be attested an ø- and y-like timbre. Moreover, the strength of the low partials is notable: there are neither exclusively even-numbered, nor exclusively odd-numbered partials. And finally: the brilliance of a single-coil-guitar (which would feature a resonance of 3 – 4 kHz) is not achieved; rather we have a strong, mid-range-y, trumpet-y sound … or a saxophone sound, or a cello-sound with flute-like harmonics? Journal-literature – (rightfully) praising this phase of Clapton's as pure genius – has found, and still finds, many comparisons. It seems strange that to describe a guitar sound, one would have to borrow from the realm of wind instruments, or strings – but maybe in the far distant future, a trumpet instructor will shout at his pupil: *blow with more emphasis on the mids; more like Clapton's guitar sound*!

Irrespective of whether trumpet- or cello-like, what does determine that sound and its variance that appears even for the same notes? First, let's look at the second part of that question which is easier to answer: even when fretting the same string at the same fret, the sound depends on the location of picking, and on the movement of the plectrum. And on the plectrum itself – although that was certainly not swapped during one take of the recording. The angle of the plectrum (parallel or slanted relative to the string), the basic movement (up- or down-stroke), the angle of the movement (relative to the fretboard), place of picking (closer to or further from the bridge – these are all sound-determining parameters. Then there is how the left hand is at work: even slight bends can make partials vanish into interference-gaps. That is why the four analyzed G_3's are not identical, and that is why there is no "one" G_3-spectrum, and not "the" C_4-spectrum, either, and least of all "the" Clapton-spectrum. Not to forget: guitar, cable, amp, room, and recording technique of course also influence the sound – but these would be time-invariant per recording … presumably, EC will not have jumped back and forth between amp and mike. But then, come to think of … one could surmise that some musicians ……

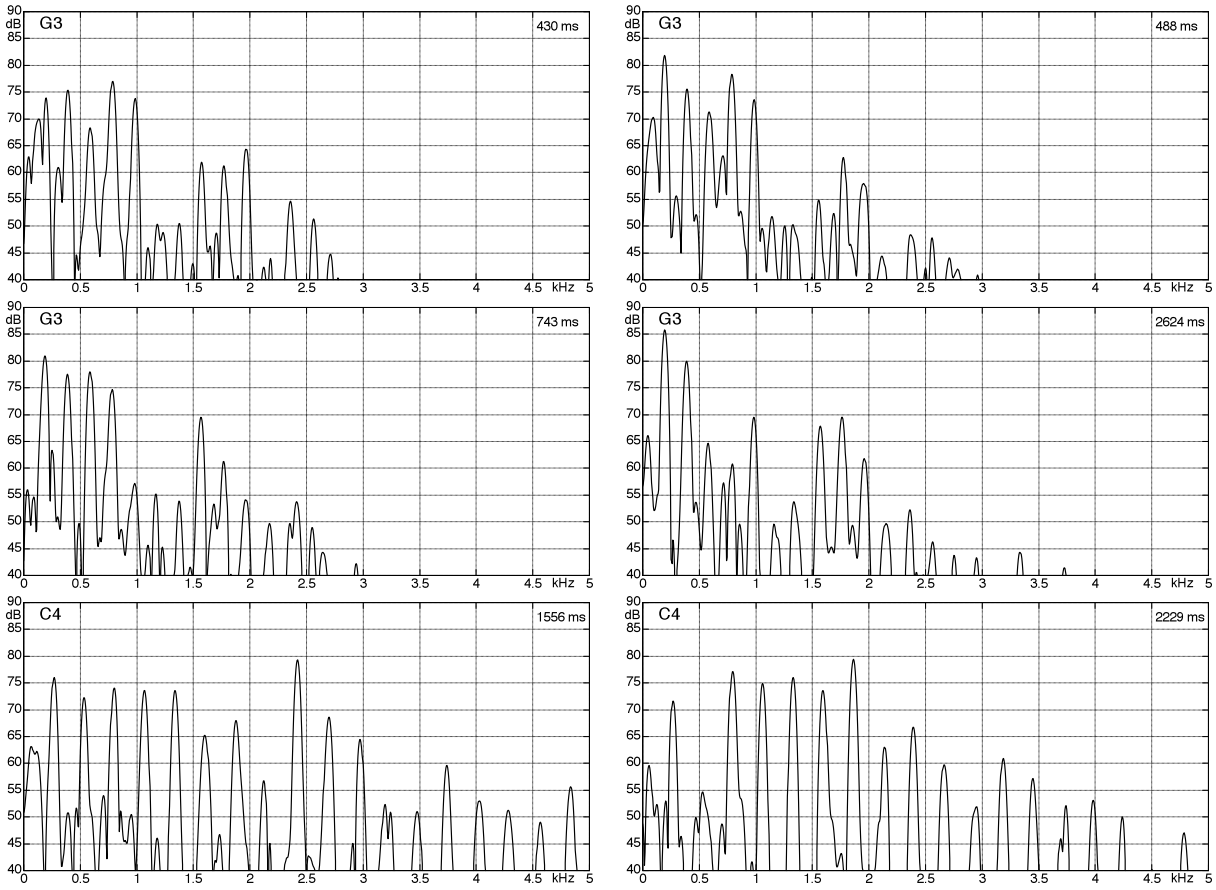


Fig. 8.42a: Individual spectra for the spectrograms in Fig. 8.43a. Kaiser-Bessel-window, $N = 2048$.

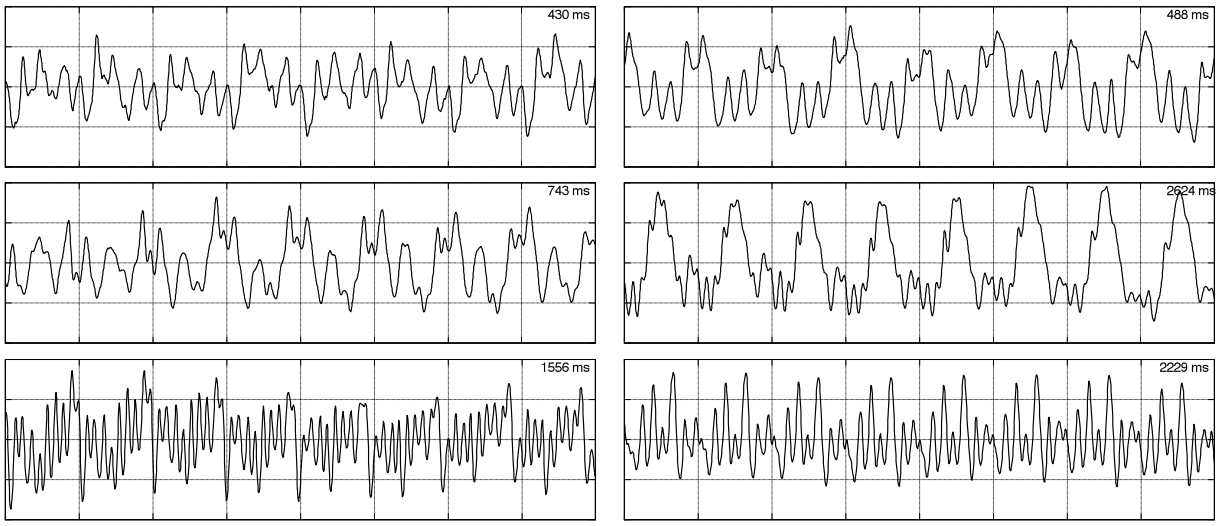


Fig. 8.42b: Time-functions for the spectra in Fig. 8.41a. Time grid = period of fundamental.

But where exactly do we now have the analytical proof for Clapton's "unique" or at least "groundbreaking" Bluesbreaker-sound; what is so special about these notes and their spectra? Ultimately: nothing at all! Listening to them in isolation, cut out of the intro, they sound plenty unspectacular. Maybe like a trumpet, or like a cello, or even synthetic. It gets interesting only as a group of notes is sounded, as soon as every note is presented with its attack- and decay-processes only existing in full context. However, it is exactly those processes that elude any spectral analysis that could reasonably be interpreted.

Somewhat easy to detect is the "lingering" of individual (in fact already terminated) notes which is due to the strong amplification. The **color-spectrogram** in **Fig. 8.43a** shows this, and it aurally creates the impression of a mighty, fat, powerful sound that can be reigned in only with difficulty. However, the short attack-noises limited to 20 – 40 ms duration yield all those problems that have already been described in the context of the dyad- and triad-signals. Of course it is possible to calculate corresponding spectra, but they will be highly parameter-dependent. Dear PhD-students who are just now in the process trying to cut another facet into the diamond that is psychoacoustics: don't let yourselves be discouraged by that! EC needed 21 years to produce these sounds – you don't have to have them analyzed within 2 days. Sure, it is not impossible, but simply applying a bank of Gammatone filters with contouring-algorithm – that ain't enough. Here's a hot tip: do *synthesize* the sound using the supposed partials, and listen. This approach very quickly reveals, which formation-rules are verifiable but not relevant to the auditory system, and what might constitute a "groundbreaking" sound. And speaking to the gear-heads: you won't do anything wrong bringing out that original '58 (or was it a '59, after all?), but absolutely necessary it is not. Required are the right fingers, the right micro-timing, the right bends. "Clapton is God" was the writing, not "Paula is goddess". This is illustrated by many EC-epigones appearing on Youtube, covering *Stepping Out* with at times remarkable equipment (but at times showing dismal timing, too). It becomes quite clear that the finger-work is much more essential than the question of "R8 or R9?".

It is time to come back to the starting point of this chapter: to the timbre (or tone color). The latter may without doubt be determined on the basis of a spectrogram – but in infinite variations, because there are infinite possibilities to parametrization of spectrograms. If we do not want to test all of then, then an overlapping 4k-DFT with Kaiser-Bessel-window for the steady-state part of the guitar-tone will deliver some first orientation values. The onset of tone (attack) is more difficult to analyze because here the spectrum can change as much as 20 dB within 10 ms – a typical case of conflict between time-domain-resolution and frequency-domain-resolution. If several instruments sound at the same time, the analysis becomes particularly difficult. For the graph in **Fig. 8.43a**, only a single guitar plays, and the behavior of individual partials can clearly be observed. This behavior is, however, difficult to measure since these partials rarely maintain their frequency, not even in the seemingly steady-state part of a note. We find subtle up-bends (at around 1000 ms), down-bends (also called pre-bends, around 1900 ms), and half-step bends (around 1600 ms). Thus, it is not sufficient to set the cursor on the 180th DFT-line and to analyze how its level evolves. This would again be merely the behavior of the level of this DFT-line but not that of a special partial – the frequency of the latter is changing, after all (e.g. from the 180th line to the 191st DFT-line). Contouring- and pitch-follower-algorithms (which one indeed is that "closest neighbor"?) are applied to assist in this scenario, which is another reason for the multitude of parameters. Once these problems have been solved (it is, after all, not impossible to track partials), new challenges present themselves: the partials not only change their frequency but also their level! And not too little or too slowly, at that: we see e.g. 6 dB / 10 ms. Mind you, the attack- and decay processes of the DFT-analysis may run with the same speeds. Thus, if we change the DFT-parameters, the level fluctuations also may change. This multi-variant analysis (or optimization) would go far beyond the scope intended here, and so what can remain is merely the qualitative statement: the partials change their amplitude and frequency even within one single played note. At least the frequency shift is a global one (all partials change their frequency by the same percentage), but the amplitude shifts are partial-specific. Not all frequency- and amplitude changes are audible; there are absolute thresholds, masked thresholds depending on neighboring tones, and pre- and post-masking in time. Only that which is above threshold is fed to the final post-processor that then forms – among other things – the timbre, the tone color.

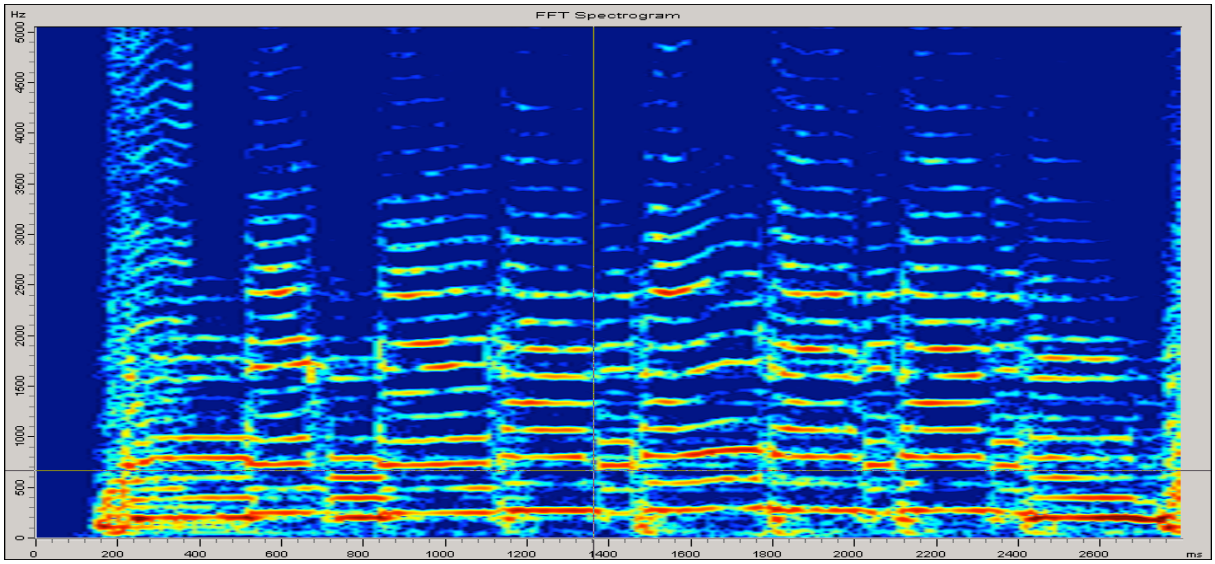


Fig. 8.43a: Excerpt from *Stepping Out* (Mayall / Clapton), guitar-intro. ΔL = 40dB.

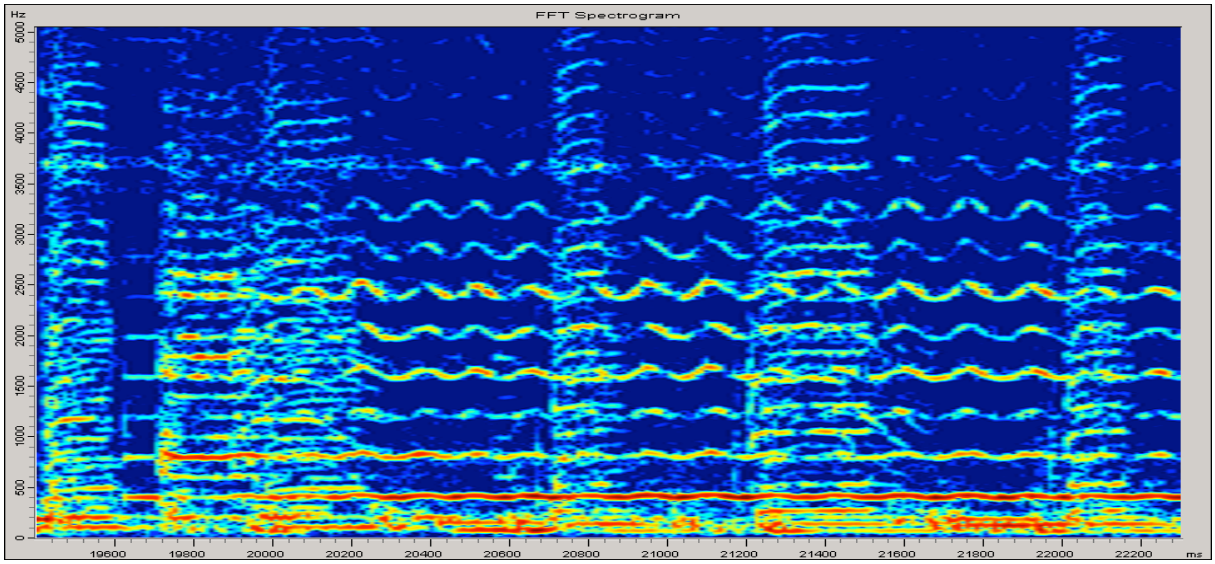


Fig. 8.43b: Excerpt from *Stepping Out* (Mayall / Clapton), guitar note with finger vibrato (7 Hz). ΔL = 40dB.

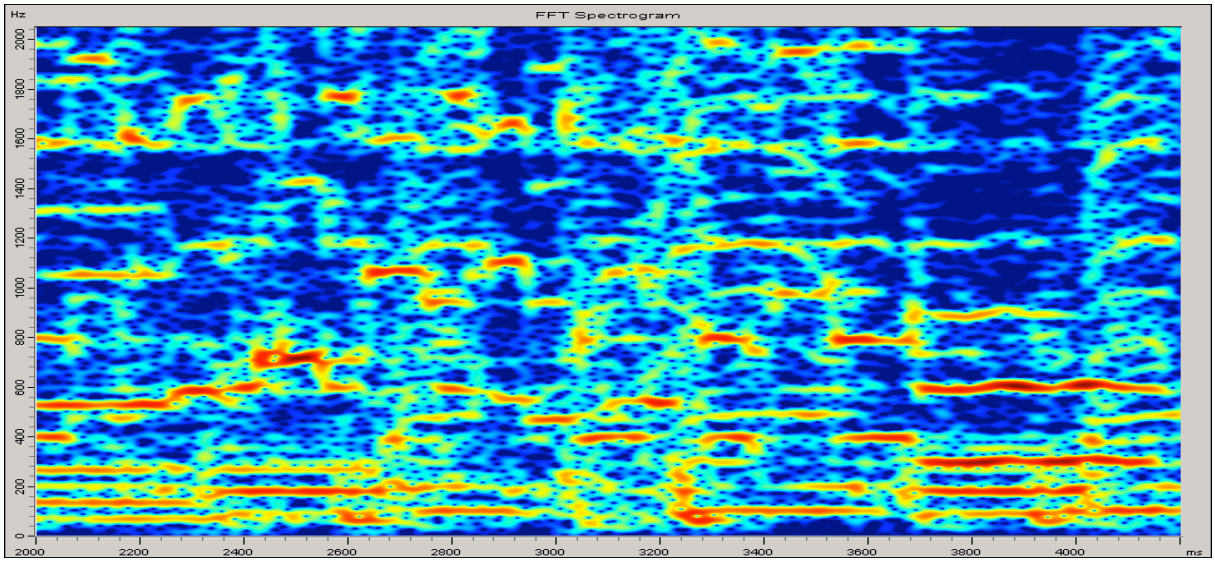


Fig. 8.43c: Excerpt from *Stepping Out* (Mayall / Clapton), fast eight-note triplets. ΔL = 40dB.

All in all: not a trivial analysis. This is not supposed to sound too discouragingly, therefore let's quickly look at **Fig. 8.43b**. In this graph, the visual analysis is facilitated by a gestalt-law that helps in auditory tone-recognition, as well: the low of common fate (see also Chapter 8.2.4). All lines that move back and forth in synchrony are partials of a guitar note, in between the horns provide (vertical) accents, and the electric bass lays the foundation below. It may be added for the Strat-purists: no, you do not need a whammy bar for that; this is done with a left-hand finger. To bend a note by ± ¼-step with a modulation frequency of 7 Hz – that is Clapton at his best. In **Fig. 8.43c**, things get more hairy again. This is one of the passages with faster playing, and vibrato is not really possible with note-durations of as small as 100 ms. In this section, already the pitch-tracking is a true challenge, not to mention an automatic timbre-analysis.

(Translator's note: the following paragraph only makes sense and works for German speech sounds <u>and</u> words. It was impossible to find suitable correspondences in English without a complete re-write/re-draw. I have tried to make sense nonetheless, using again the International Phonetic Alphabet – IPA – where necessary ...)
If we do not want to wait until research offers reliable algorithm, we can only resort to onomatopoeia as it has been practice for centuries. This is an effort of pattern matching between the spectral maxima of the guitar tone to be described, and those of a speech sound (formant = frequency of a spectral envelope-maximum). From this, it suddenly becomes understandable that a "flute-like" *("flöten-artig" in German)* guitar sound does not need to unconditionally sound like a flute. Maybe that guitar sounds merely like a spoken "ø" *(as in the German "flöte")*, it "fløøøøøtes" without being that instrument. The corresponding (second) ø-formant is at 1500 Hz. It may be a bit higher up, if a female speaker is assumed (N.B.: it's *she the Paula*, after all). It wouldn't be counterproductive, either, that the famed blue Cøløstiøn-speakers have a maximum in their transmission curve around that frequency.

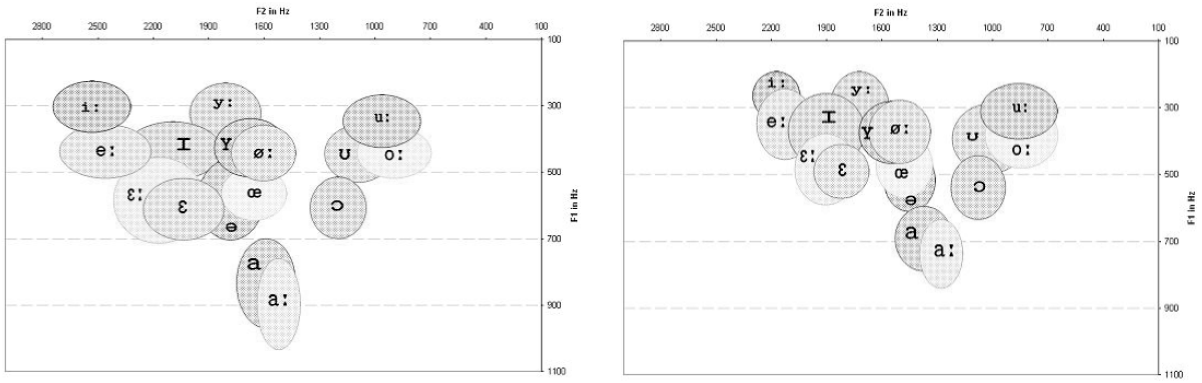


Fig. 8.44: Formant-frequencies of the German language, f/m speaker; (Sendlmeier/Seebode, TU Berlin).

In short: timbre (the tone color) depends on everything involved in tone generation. Not only that: the subjective assessment criteria of the listener play a role. To objectively visualize the sound that generates a timbre, the SPL-time-function is a complete but rather unsuitable and abstract quantity. The hearing system does not directly process the time-function, but a short-term spectrum determined according to complex rules. The phase is of secondary importance in this short-term spectrum; the behavior over time of the spectral amplitudes yields the primary hearing-relevant data-set. From the latter, and with suppression of masked (inaudible) ranges, a secondary above-threshold data-set is derived. Contouring-algorithms (maximum-detection), curve-following- and grouping-algorithms join what belongs together, and enable – on the basis of memorized knowledge – recognition of instrument-typical characteristics: timbre, pitch, and loudness, among others. There is a good deal of arbitrariness involved here: whether strongly modulated tones are attested a fixed pitch with a special modulation timbre, or a variable pitch with a fixed timbre: that is under the sovereignty of the listening "subject" .

