The Basics of High Fidelity
Part 7: What Do We Really Want? Immersion!

In the previous six parts we have deepened our understanding of HiFi and came to the conclusion that HiFi stands for "naturalness" in the sense that we are striving for a certain amount of transparency between recording and reproduction. We have also seen that there are two different kinds of transparency, "here and now" (augmented reality) versus "there and then" (virtual reality), which require non-compatible optimizations on the recording/reproduction chain. This leads us towards the big question "What Do We Really Want?" The simple quick answer is Quality, which in the ideal case should be optimized towards individual preferences using simple control mechanisms.

I have carried out many subjective tests, especially regarding the assessment of speech quality, and have seen strange personal preferences regarding what we like. We also know that people tend to adjust to a certain sound. If you listen a long time to your own loudspeakers, you may develop a personalized preference bias. Musicians also suffer from this and it is not a good strategy to ask a musician about his favorite loudspeaker. In general, musicians listen to their own instrument at close distance which automatically results in a distorted view on the reality of listening to music. And to be clear, we are allowed to develop our own preference bias, but we should realize that this leads to a shift from science to art.

At this point, it is wise to define the terms audio and sound quality more precisely. We will use the term audio quality whenever it is related to the transparency goal and sound quality whenever it is related to a personalized preference. This implies that for audio quality, in either the electric or acoustic domain, we need a predefined ideal allowing us to force subjects towards a unified opinion. For sound quality, which is only defined in the acoustic domain, we have to deal with personal preferences which are sometimes difficult to average over large sets of subjects.

So, let's go back to science and try to develop characterizations which can be used in both the audio and sound quality domain. In psychoacoustics, the four most fundamental perceptual characteristics that can be controlled in HiFi audio are:

- Loudness (amplitude/volume control)
- Repetition Pitch (time/playback speed control)
  Note: Pitch has 2 flavors, Repetition Pitch, dominating perception, and Spectral Pitch dominated by the sharp frequency peaks of the signal.
- Timbre (frequency distribution/bass-treble tone color control)
- Localization (spatial distribution/immersion control)

Now, let's use these characterizations to find out "What Do We Really Want!" in HiFi audio.

The first characterization brings trouble. Why? Because most people love Loudness too much! Almost all modern HiFi equipment is capable of producing more than natural high Loudness and if we leave it up to the users, they will increase the volume to a level where the Loudness will damage their ears. And if you compare two HiFi loudspeakers, the louder one will almost always be preferred. In the world of telephony, many loudness experiments have been carried out and they show that the preferred Loudness of a telephone is about 20 dB higher (i.e. 100 time more powerful) than a natural voice at 1 meter distance. In fact, this high Loudness level is standardized by the ITU (International Telecommunication Union) as the optimal Loudness level. Loudness is like a drug, you adapt and need more and more to be

satisfied, in the end always leading to a damaging high Loudness level. The extreme difference of 20 dB in the voice example is for a large part caused by the fact that standard telecommunication voice experiments are carried out with band-limited speech presented in one ear. In general, shortcomings in audio reproduction are partly compensated by unnatural loud playback, e.g. subjects tend to listen to music with headphones/earbuds with a too high loudness setting. High levels of voice playback can lead to a decrease in intelligibility due to an increase in upward spread of masking. From a technical perspective, Loudness is no challenge at all and the volume knob on your HiFi system is by far the most important control knob that allows you to adapt the reproduction Loudness towards your personal preference. But it should be used with care.

Exact reproduction of Repetition Pitch, the second characterization, used to be a huge problem in the analog world. Play a single piano note and reproduce it over a classical HiFi system using a vinyl recording, or even worse, using a cassette recording, and you will in most cases be able to clearly perceive audible wow and flutter in the Repetition Pitch. In the old analogue world, you needed an expensive studio quality tape recorder to hear no wow and flutter. But fortunately, if you run the same experiment with a CD or a streaming platform, the wow and flutter will be inaudible. And although a tremolo represents a wanted flutter, there is seldom a post-processing in our HiFi system that adds flutter. The only control that is useful in the manipulation of Repetition Pitch is playback speed control, which in its most advanced form is used to playback speech faster than live while maintaining the correct Repetition Pitch using advanced PSOLA (Pitch Synchronous Over Lap Add) algorithms. Note that Repetition Pitch is related to frequency content, but is dominated by repetition time, e.g. an harmonic sound with frequencies of 1000, 1200, 1400, 1600 and 1800 Hz has a dominant Repetition Pitch of 200 Hz (inverse repetition time, sometimes referred to as Virtual Pitch) as perceived in the synthetic mode of perception. When we force ourselves in an analytic mode of perception, we can perceive the four Spectral Pitches that are less dominant. Our second conclusion is that we are happy with our perfect digital flutterless reproduction, while playback speed control with pitch preservation opens up a new world of post-processing possibilities.

Timbre, the third characterization, is predominantly determined by the frequency response of the complete recording/reproduction chain. In the recording part we have the problem that musical instruments often have a wild varying radiation pattern that will lead to an incorrect Timbre when a single close microphone recording technique is used. In a high quality recording room, such as a concert hall, the room acoustics take care of the integration over all directions leading to a well-balanced Timbre. If we make a recording at a place where we perceive a high acoustic quality the play back of such a recording will not result in a high quality sound due to the fact that our ears cannot de-colorize the sound as would have been carried out in the live situation. Recording engineers often use close microphone techniques that will thus suffer from an unbalanced Timbre, especially for musical instruments that have a wild varying radiation pattern. This unbalance is often accentuated by an artificial reverberation that is added to the dry recording. This makes the recording of an acoustical event more art than science.

If the Timbre problem in the recording room is solved we run into the problem of the response of the reproduction room which is in most cases poor due to (room) resonances, leading to the conclusion that the recording and the reproduction rooms are the dominating factors in the final Timbre. With headphone reproduction, one can bypass the reproduction room degradation and when using individualized recording play back HRTF’s with head tracking one can achieve high quality although the lack of bass feeling will always be noticeable. Due

to the fact that the timbre problem is so difficult to optimize audio engineers have searched for means to adapt the Timbre to one's own preference. The most widely used approach was developed by Baxandall who designed 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.

Localization, the fourth characterization, is determined by the spatial distribution of the acoustic events, the Localization of the microphones, the spaciousness of the recording room, the artificially controlled spatial distribution (including added spaciousness), the spatial distribution of the loudspeakers and the spaciousness of the reproduction room. If we use headphones, the spaciousness of the reproduction room plays no role, but we need individualized binaural recordings with headtracking where the playback response of the headphone is adapted towards the position of the head. But even in the case that the spatial distribution is of acceptable quality, i.e. good Localization, the feeling of immersion is seldom acceptable with headphone reproduction. Furthermore, the optimal spatial distribution is strongly dependent on the goal we strive for, "illusion here and now" (augmented reality) versus "illusion there and then" (virtual reality), requiring non-compatible optimizations of the recording/playback chain.

In modern recordings, we see more and more artificially created spaciousness where there is no reality to strive for. Some people like to add artificially generated spaciousness in their home HiFi system on top of the spaciousness as found in the recording, especially if the reproduction room sounds too dry. Optimal spaciousness may thus require post-processing and/or recording/playback techniques that require a multi-channel approach. We should be aware of the fact that multi-channel approaches tend to produce more problems than they solve due to the problem of Localization stress. This is especially true for full six-degrees-of-freedom audio reproduction. For music reproduction, the feeling of Immersion, restricted to small head movements and focused on three-degrees-of-freedom, is more important than the accurate Localization of musical instruments. Advanced systems like Dolby ATMOS, high order Ambisonics, wave field synthesis or object based audio coding ([1], [2], [3], [4], [5]) are only useful in films and games where sound effects require a more exact Localization and where binaural and monaural decolorization requires a special recording/playback approach that often forces one towards the use of a foley artist. In contrast, music reproduction requires a well-balanced diffuse field which by definition cannot be localized. And although lateral reflections may contribute to an increased spatial impression, they may cause degradations in case they arrive within a 30 ms window from the direct sound. Early strong reflections are better suppressed by using diffusion at the source. One could even say that diffusion at the source is the main reason why the Leslie Hammond sound is world famous.

And what is the fundamental reason for the liking of immersive diffuse fields? It's evolution. Close-by objects have low levels of diffuse field and can be dangerous, objects that are further away have higher levels of diffuse field and are perceived as less dangerous, leading to a higher feeling of comfort.

Our fourth conclusion is that creating an optimal immersion is important in the perceived quality, but that optimization is complicated. In years of trying to get control over this issue, my conclusion is that the best processing in home systems is strongly dependent on the recorded material and should be focused on allowing to control the diffuse field in such a manner that it can easily be adapted to the content that is played.

A trivial post processing strategy that improves the feeling of immersion with stereo recordings is to just add two extra surround loudspeakers that simply reproduce a slightly lower volume of the left and right front loudspeaker [6]. A relative level of −3 to −10 dB is suggested. These speakers should be designed in such a way that they only contribute to the diffuse field, thus allowing for a simple control over the amount of Immersion, without the introduction of Localization errors. A simple but effective way of creating a diffuse surround speaker is to use a cone shaped diffuser that creates a 360 degrees horizontal radiation pattern. In the optimal construction, this cone is designed in such a way that there is minimal contribution to the direct field, e.g. by limiting the cone radiation to 300 degrees (see Figure 1).

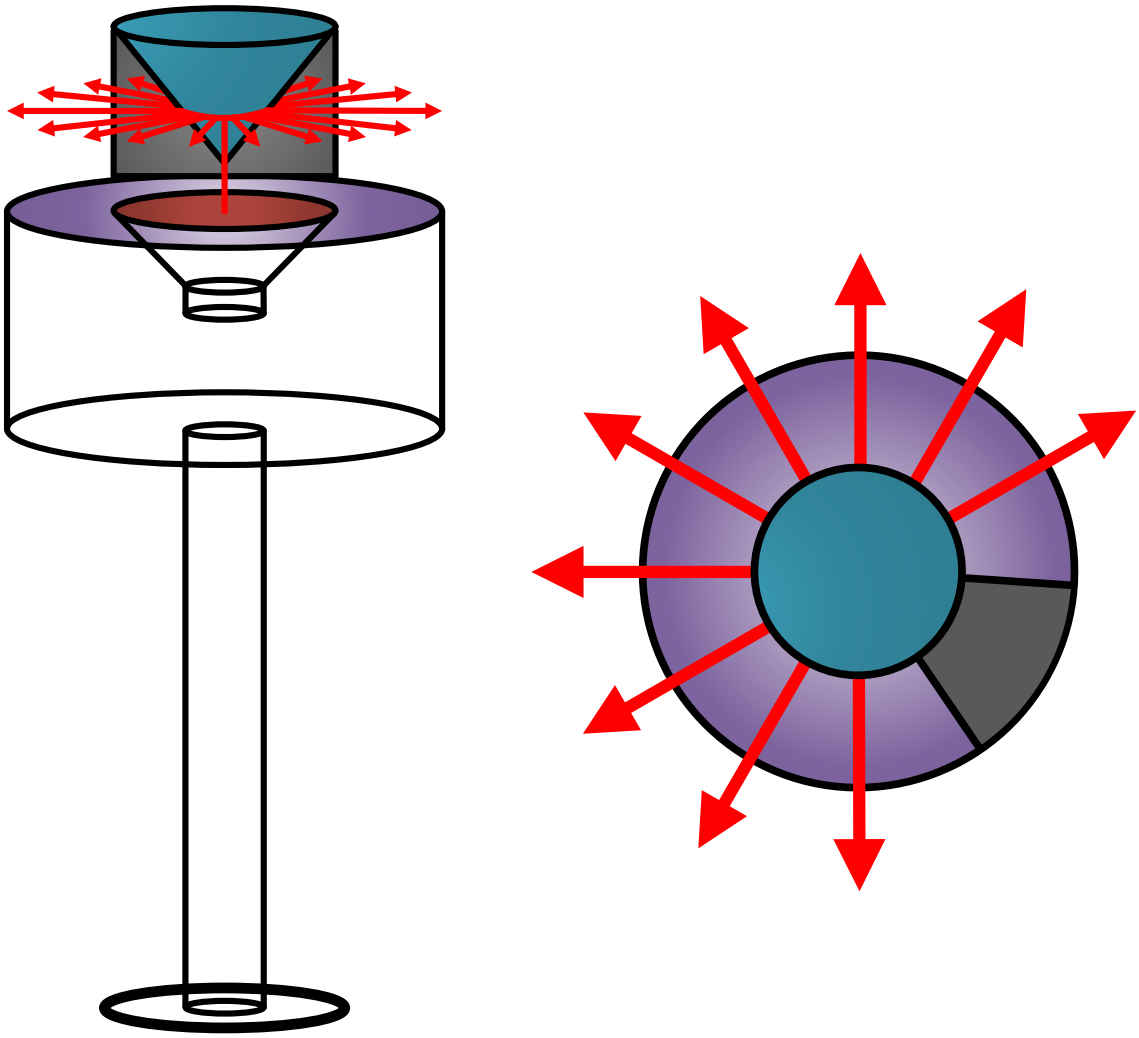

Figure 1. Side and top view of the suggested surround loudspeaker used to create a diffuse field behind the listener. The grey block is attached to the mounting of the cone and is covered with absorbent material that provides shielding of the direct sound, so that dispersion is mainly limited to around 300 degrees omnidirectional.

To keep the front image stable, the audio signal for these speakers must be delayed on top of the compensation required to achieve the same propagation time for the surround units as for the front speakers. The amount of additional delay to create the diffuse field can, in combination with the level settings, be used to adapt the feeling of immersion towards the content that is played. In a co-operation with a number of small HiFi companies in The

Netherlands a series of experiments were carried out where experts could adjust the delay and level of the diffuse surround speakers. For the delay the optimal value was between 10 and 20 ms depending on the acoustic properties of the room where the recording was made. Roughly speaking, more delay could be allowed for recordings that are made in large concert halls than for dry pop recordings. The optimal level of the additional surround speakers also dependent on the properties of the recording but also differed largely between experts. When the system was evaluated with naïve listeners the preferred level of the extra diffuse field speakers showed an even larger variation. Some subjects set the level close to the just noticeable difference, about 20 dB below the level of the direct field loudspeaker, while others choose to set it above the level of the direct field loudspeakers. From the 24 subjects that were used 23 preferred to switch on the extra diffuse field speakers for the majority of the used music fragments while 16 subjects always switched it on. One subject only switched it on in 43% of the fragments.

The most interesting conclusion from the immersion experiments was the significant increase in perceived overall sound quality when the diffuse surround speakers were switched on. Using a five point scale ranging from 1, a very small improvement to 5, a very big improvement, the experts judged the overall sound quality improvement around 3 while the naïve listeners judged the quality improvement even bigger with scores around 4. For dry recordings the optimum is around 10 ms while for large concert hall recordings it is around 20 ms. In the ideal postprocessing, this immersion control is combined with diffuse field processing of the Left and Right front loudspeakers to correct directivity timbre problems as explained in Part 3 (see Left and Right Diffuse Field Fillers in Figure 2).

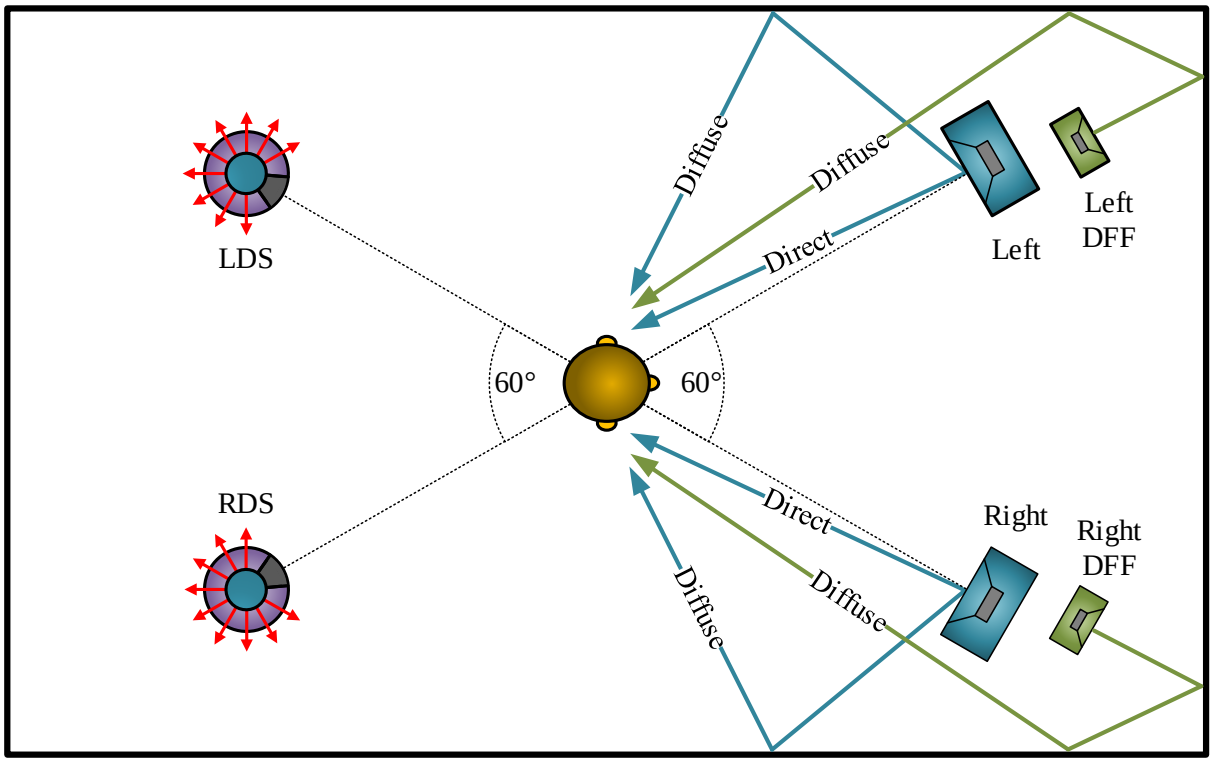

Figure 2. Loudspeaker setup that allows for Immersion control by use of the cone shaped diffuser of Figure 1 that only contributes to the surround diffuse field. The Left and Right Diffuse Surround (LDS and RDS) loudspeakers are delayed by about 15 ms in order to keep the front image stable. Depending on the characteristics of the recording, the level setting and delay of LDS and RDS can be used to optimize the feeling of immersion. Note that the Left and Right loudspeaker create a flat direct field and a non-flat diffuse field that is equalized with the Left and Right Diffuse Field Fillers (DFFs) as explained in Part 3.

More complex algorithms, that may also provide a center channel, can be formulated (see e.g. [7]) but in general, two-to-five up mixing algorithms provide a poorer front image quality and only a marginal improvement in Immersion. In most cases, the original stereo reproduction is preferred [8]. Note that the Left and Right Diffuse Surround speaker in Figure 2 are radiating towards the walls of the listening room instead of directly radiating towards the listener as used in standard surround set-ups [8], [9]. This prevents Localization degradations that can be characterized as ”hearing things jumping around”.

With this advanced diffuse field reproduction approach, we can achieve levels of immersion that sound better than advanced multichannel recording/playback systems. An extra strong point in the proposed diffuse field approach is that it allows to have control over the feeling of immersion by the diffuse field in such a manner that it can easily be adapted to the content that is played.

Summarizing:

1. We like Loudness, but we get more than is good for us. The volume control is the most important control knob on our HiFi system that allows to set the volume to our personal preference that is dependent on the characteristics of our ears.
2. We don't want unnatural Pitch variations (wow and flutter in the old analogue world). All modern HiFi systems can provide this. Sometimes, with speech, we may want to speed up the play back without degrading the pitch, modern signal processing techniques allow us to do so (PSOLA).
3. We want sound with a natural Timbre without disturbing resonances. The reproduction room dominates this. Global Timbre optimization is possible with the Baxandall approach. The Baxandall knobs "bass" and "treble" are important control knobs that allow to set the timbre to our personal preference.
4. We have to choose between the Localization "here and now" (augmented reality), "there and then" (virtual reality) or "anything goes" (extended reality) and have to adapt our recording/reproduction chain accordingly. The amount of immersion of stereo recordings can be optimized with a diffuse field volume control using the setup of Figure 2. This immersion control knob should replace all the difficult surround choices with their control knobs that are currently used in multi-channel home systems and that often suffer from Localization errors. The immersion control knob is the second most important control knob that allows to adapt the feeling of immersion to our personal preference.
5. We could add a final point that we want the reproduced sound to be free of unwanted disturbing (background) noises, signal interruptions and nonlinear distortions. From a technical point of view this requirement is esay to fulfill.

We have determined the contributing factors that dominate the perceived quality in music reproduction and we could ask ourselves whether it is possible to find an objective measurement approach that correlates well with the subjectively perceived quality? In Part 6 we have discussed the problem of assessing reproduction quality with subjects and have seen that the idea of an idealization dominates the procedure in the subjective assessment. This idea of idealization can be used in objective quality assessment, similar to the procedure used in the objective assessment of speech quality [10], [11], [12]. A first implementation of this idea is given in [13].

So, have we reached our destination in the HiFi story? No, one final point has to be discussed: Telephony. Although a classic telephone connection is considered to be the rock bottom in HiFi, there are some interesting observations to be made regarding the conversational speech quality of a voice link. It starts with the observation that with a telephone we are not only dealing with listening, but also with talking and interacting. When you talk, you can hear your own voice and when you hear your own voice in the wrong manner, the conversational quality is terrible, even if the listening quality is perfect. This will be discussed in the final paper, Part 8 on Telephony.

[1] M. A. Gerzon, "Periphony: With-Height Sound Reproduction", J. Audio Eng. Soc., vol. 21, 2-10 (1973 Feb.).

[2] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video", J. Audio Eng. Soc., vol. 33, 859-871 (1985 Nov.).

[3] Furness, Roger K. "Ambisonics-an overview," Audio Engineering Society Conference: 8th International Conference: The Sound of Audio, pp. 181-190 (1990).

[4] A. J. Berkhout, D. de Vries and P. Vogel, “Acoustic control by wave field synthesis,” J. Acoust. Soc. Am., vol. 93, pp. 2764-2778 (1993 May).

[5] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, & F. Zotter, “Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State,” Proceedings of the IEEE, 101(9), pp. 1920-1938 (2013 Sep.).

[6] M. Tohyama and A Suzuki, “Interaural cross-correlation coefficients in stereo-reproduced sound fields,” J. Acoust. Soc. Am., vol. 85, pp. 780-786 (1989 Feb.).

[7] R. Irwan and R. M. Aarts, “Two-to-Five Channel Sound Processing," J. Audio Eng. Soc., vol. 50, pp. 914-926 (2002 Nov.).

[8] F. Rumsey, “Controlled Subjective Assessment of Two-to-Five Channel Surround Sound Processing Algorithms," J. Audio Eng. Soc., vol. 47, pp. 563-582 (1999 July/Aug.).

[9] ITU-R BS.1116, “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”, International Telecommunication Union, Geneva, Switzerland (1997).

[10] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I – Temporal Alignment,” J. Audio Eng. Soc., vol. 61, pp. 366-384 (2013 June).

[11] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II – Perceptual Model,” J. Audio Eng. Soc., vol. 61, pp. 385-402 (2013 June).

[12] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment,” Geneva, Switzerland (2011 Jan.).

[13] J. G. Beerends, K. van Nieuwenhuizen, and E. vd Broek, “Quantifying Sound Quality in Loudspeaker Reproduction”, J. Audio Eng. Soc., vol. 64, pp. 784-799 (2016 Oct.).

John G. Beerends, send comments to johnbeerends@hotmail.com
Published in Hifi Video Test 6/2008 (in Dutch), translated and updated over the period 2012-2023 in co-operation with Richard van Everdingen.

Part 1: Transparency and Perceptual Measurement Techniques

Part 2: Reproduction Philosophy “Here and Now” versus “There and Then”

Part 3: The Ideal Loudspeaker, Diffuse Field Equalization

Part 4: The Ideal Loudspeaker, Reflections and Resonances