A reprint from

American Scientist

the magazine of Sigma Xi, The Scientific Research Honor Society

The Science of Hi-Fi Audio

Despite great advances in quantifying sound quality, engineers are still struggling to satisfy the subjective ways listeners respond to music.

John G. Beerends and Richard Van Everdingen

The swell of the orchestra reaches a crescendo, all of the instruments together creating a swirling field of sound that fills the concert hall and surrounds the listener. Anyone who has ever attended a classical music concert has probably encountered that joyous feeling of being completely immersed in sound. But most of us don't have an orchestra at home, and a large orchestra probably would not fit in there, anyway.

Engineers have been seeking ways to re-create the immersive experience of a live music performance ever since 1877, when Thomas Edison made the first crude recording of himself reciting "Mary Had a Little Lamb." The ultimate goal has been high-fidelity audio, or hi-fi: the reproduction of sound without audible noise and distortion, based on a flat frequency response within the human hearing range. In terms of technology, that goal might now seem easily attainable. Even moderately priced consumer equipment can process sound accurately; given that humans only have two ears, a simple stereo setup with two speakers would seem sufficient for the job. Yet modern designers of hi-fi audio systems keep adding more speakers with more audio channels without ever quite managing to recapture the sensation of musical immersion.

We have spent our careers pursuing a scientific, perception-based approach for assessing audio devices, so we are keenly aware of the obstacles to attaining hi-fi sound. Above all, every person has different ears, a different brain, and unique, personal preferences. It is therefore difficult to separate facts from opinions and fake claims when discussing the quality of recording and playback.

One of us (Beerends) memorably experienced the subjectivity of sound while attending a hi-fi trade show, where a small company demonstrated a very expensive audiophile amplifier. During that demo, a soft hum was audible to me in the silent intervals of the music. At first, the man running the equipment could not perceive the hum. Only after I suggested that he listen to the loudspeakers at a closer distance could he, too, perceive the hum. Nobody appreciates a humming amplifier, so presumably multiple engineers at the company failed to notice the sound that was obvious to me.

For recordings of speech, at least, test subjects largely tend to agree in their assessments of reproduction quality, especially when they are listening to familiar voices. But for music, individual preferences tend to dominate, greatly complicating the situation. Whether people are listening through headphones, earbuds, Bluetooth speakers, home stereo, automotive audio, or any audio system you can dream of, their judgements of musical sound quality show large differences from individual to individual.

The upshot is that audio engineers can achieve high quality rather easily for the recording and playback of speech, but the recording and playback of hi-fi music remains elusive. Even multi-channel systems cannot consistently and satisfactorily re-create most listeners' experiences of, say, the rich, diffusive sound of a large classical orchestra. In fact, such complex audio setups miss the most important subjective aspect of listening to music: being immersed in the sound. We argue that there is a better and simpler solution.

A Search for Transparency

An essential quality of hi-fi audio is what's called *transparency*. For a well-designed audio device—regardless of whether it is for music recording, compression, storage, streaming, or playback—there should be no discernible difference between the input and the output, as if the device itself were transparent and invisible. Using that audiophile amplifier as an example, we could take a sample of the input signal and compare it with a sample of the output signal. If we then subtract the output from the input (after aligning the amplitude and compensating for a possible delay), we should get an overall zero signal.

If the subtracted signal is not exactly zero, the difference between the input and output might still be so small that it is not audible, making the device transparent from a perceptual point of view. But if the device is not perceptually transparent, we then want to have an interpretation algorithm that can quantify the extent to which the system falls short of the transparency

QUICK TAKE

The goal of high-fidelity audio is to capture the feeling of a live musical event. Doing so requires more than just reproducing sound accurately, without audible distortion or noise.

Perceptual measurement techniques provide an effective way to evaluate sound quality for spoken words. But the techniques cannot fully capture subjective impressions of music.

A sense of immersion is crucial for a satisfying musical experience. Most commercial systems fail in that regard; the authors propose a new solution, using both direct and diffuse sound.

Keith Jefferies/Stockimo/Alamy Stock Photo

A live musical experience depends on many factors. The instruments and the acoustics of the performance space affect the sounds that reach the listener. But the ways that listeners respond also depend heavily on each individual's unique characteristics, both physiological and psychological. A satisfying hi-fi system should do more than reproduce sounds accurately; it should also re-create the feeling of immersion produced by a live event.

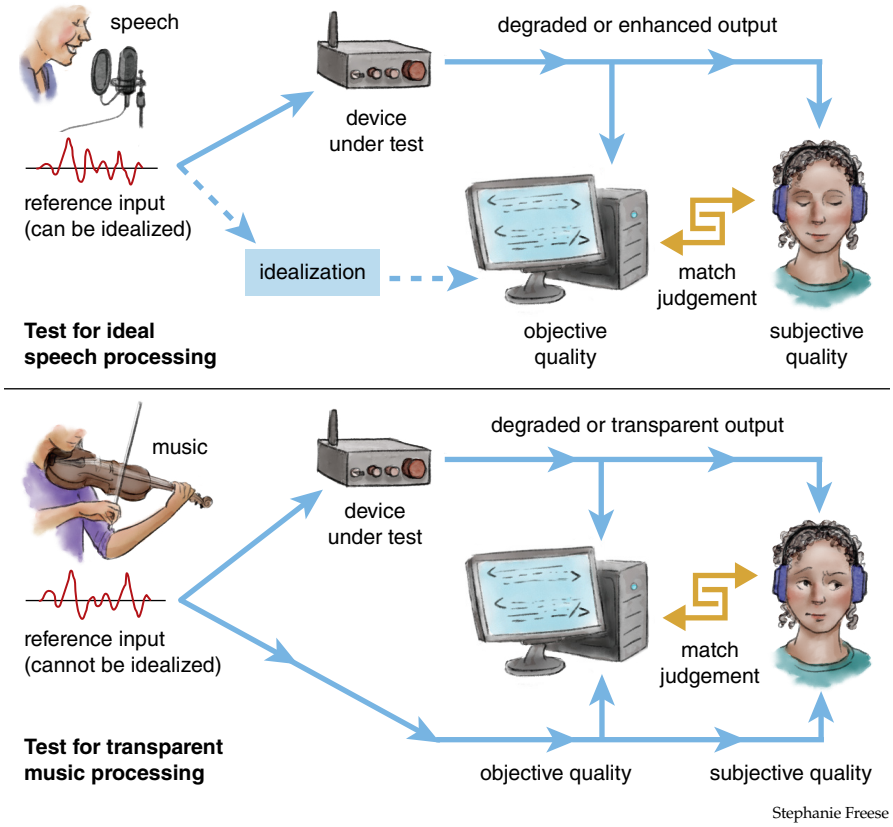
ideal. Following this approach, audio engineers have designed perceptual measurement systems that assess audible degradations of perceived audio quality (*see illustration on page 34*).

An effective perceptual measurement method was developed in the early 1990s by one of us (Beerends) in collaboration with Jan Stemerdink at KPN Research NL, the research arm of the biggest Dutch telecom company. The initial version of this method, called Perceptual Speech Quality Measure or PSQM, could assess the software used to code and decode *narrowband* speech, the kind commonly used for telephone communications; PSQM demonstrated high correlations between subjective evaluations and objective measurements of speech quality. In 1996, the International Telecommunication Union (ITU) endorsed PSQM as a worldwide standard ("Recommendation P.861 PSQM"). An improved version of PSQM, which also allowed for the assessment of wideband speech (used for high-definition communication), was developed in 2001 by KPN Research and British Telecom and accepted by the ITU as "Recommendation P.862 PESQ."

In 1998 the ITU adopted a similar perceptual measurement technique for assessing the quality of music encoded using common digital formats, such as MP3, AAC, WMA, and OGG ("Recommendation BS.1387"). However, assessing the quality of coding-decoding systems, or *codecs*, is far more difficult when dealing with music than it is with speech—especially assessing how much the sound quality has been degraded when the codecs behave nontransparently.

Listeners have more widely divergent opinions on the effect of degradations on music than they do on speech. Furthermore, the varied ways that people perceive and process sound (due to both innate physiological differences and subjective, psychological ones) are far more important when listening to music than they are when listening to speech. Even simple differences in *perceptual threshold*, the level at which certain frequencies become audible, can lead to large differences in listeners' quality assessments. In particular, degradation that occurs at high frequencies, above roughly 8 kilohertz, has limited impact on how people perceive speech but can have a large impact in the way they perceive music. Because of these complicating factors, perception-based measurements of audio quality show significantly poorer correlations with subjective evaluations when the experiments use music rather than speech.

A fundamental obstacle to developing a more accurate objective perceptual quality assessment method is that typical listeners, who simply want to enjoy their audio system, generally do not have access to an ideal reference signal. Instead, they judge the sound quality of their system against their own subjective, internal ideal.



Stephanie Freese

Perceptual measurement techniques are used to assess devices that code, decode, store, or transmit sounds. A reference input signal is fed into the device being tested, such as an audio amplifier. The reference signal and the output signal from the device are then played back for listeners, who evaluate the subjective quality of the resulting sounds. Objective computer models attempt to simulate how the listeners will respond. For speech (*top*), engineers can idealize the sound for comprehensibility. For music (*bottom*), engineers can test only for transparency: how closely the output matches the input.

In principle, if we had access to a listener's ideal sound, we could design a processing method that delivers a personalized perfect audio quality. For speech, we can do something quite close to that, because test subjects largely agree about what ideal spoken-word audio should sound like. Such consensus means that it's possible to create a perceptual measurement technique to assess the end-to-end quality of any voice connection, such as a video meeting or a cell-phone call. One of us (Beerends) was the main developer of yet another speech-quality standard known as P.863 POLQA, adopted by the ITU, which compares such connections against an average, ideal speech representation derived from a large database of speech-quality assessments. No such standard exists for music.

Another obstacle to objectively assessing the quality of music processing is that our ears hear sounds, not digital signals. Subjective audio tests therefore require a transduction device—headphones, a loudspeaker, or set of loudspeakers—to assess an audio signal. The device that we use has to be of superior quality for a listener to hear small degradations in the audio output, especially if we are evaluating high-quality devices that are designed to come close to perceptual transparency. When we are testing such devices, we will allow subjects to directly compare the reference input with the output, making it easier for them to detect small degradations in the output signal. For instance, we might let them hear what the audio sounds like before and after it passes through an amplifier or through a Bluetooth streaming system.

The situation becomes trickier still if we want to assess the quality of headphones and loudspeakers using perceptual modeling, because we run into the problem that the output is an acoustic wave that we need to transform back to data that we can feed into a perceptual measurement model. Accurately recording the output of a loudspeaker or a headphone is difficult, and it can be carried out in a variety of ways that each lead to different assessments of the device under test.

This is the core problem in the science of hi-fi audio quality assessment: Subjective tests of microphones, headphones and loudspeakers are all based on judgments that use an unknown internal ideal. Developing an objective perceptual measurement of listeners' subjective and diverse responses is exceedingly difficult.

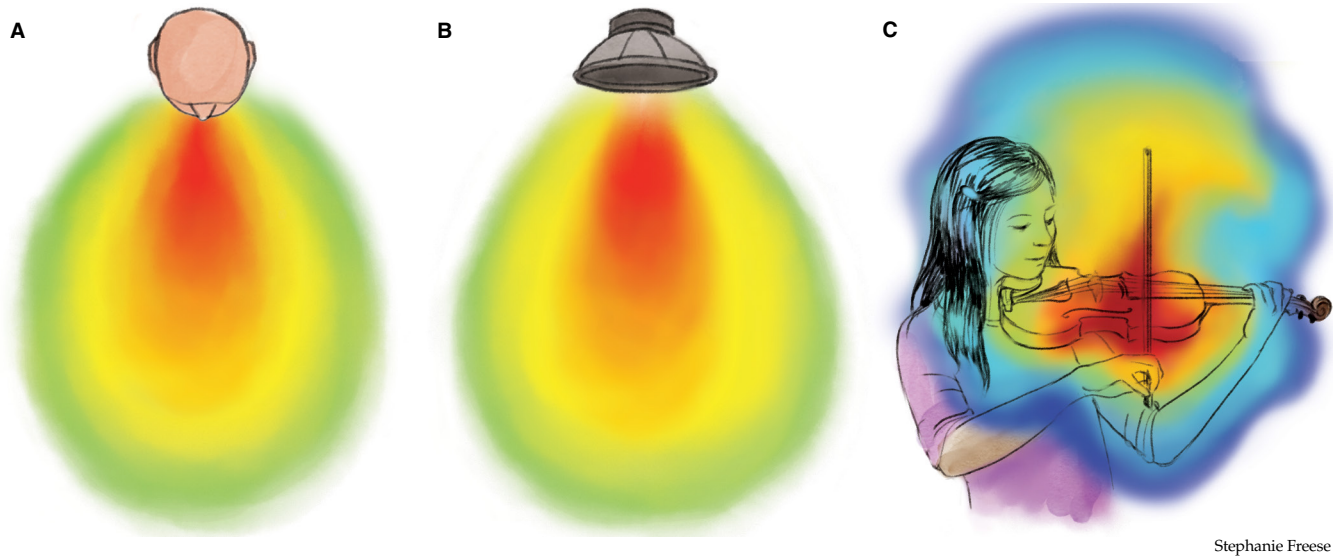
From Recording to Playback

What we really want to do is create objective perceptual measurements that can assess the complete life of a piece of music from recording to playback. That process includes everything from transduction, in which recording microphones convert sound into electronic signals, to reproduction, in which headphones or loudspeakers convert the final versions of those signals back into sound that the listener can hear.

At this point, the acoustic environments in which the recording and the playback occur become important. When you listen to a recorded sound, the room where the recording was made has a significant effect on the audio quality. Listen, for instance, to a voice recorded in a bathroom and you will hear that acoustic reflections from the room dominate the audio quality. The way we reproduce the recording also has a significant impact on the audio quality.

Audio engineers often use reference-standard headphones when asking test subjects to make audio quality judgments. Unfortunately, headphones produce an unnatural auditory effect: They make it seem as if sound is localized in the center of your head, whereas in real life the sound will be localized at some external source. When you move your head, your perception of that source will change; when you move your head while wearing headphones, everything stays the same. To make headphone playback more realistic, we therefore add a set of personalized corrections called *head-related transfer functions*. With the proper corrections applied, the sound localization will seem to move along with the listener's head movements.

Listening to audio playback over loudspeakers presents its own challenges, because the setup of the reproduction room has a significant effect



Stephanie Freese

Simple, directional sounds are relatively easy to reproduce accurately. A human voice (*a*) and an individual loudspeaker (*b*) have similar, directional properties and so produce similar sound fields. We also normally listen to spoken voices one at a time. In contrast, a single musical instrument such as a violin (*c*) produces a sound field with wildly varying directional properties. Combining multiple instruments makes the situation even more complex.

on the perceived audio quality. The advantage of the loudspeaker approach is that the room degrades the playback in the same way that it would have degraded the live source in that room. We can therefore make a monophonic recording of an acoustic source in an anechoic room (which prevents sound from reflecting) and play it back through a single loudspeaker with the same directional properties as the source, such that there is a transparent relationship between recording and playback. In contrast with the headphone experience, there is no need for a head-related transfer function correction. You could go into that listening room, rotate your head in any direction, and move around freely while maintaining full transparency between the recording and the original live sound.

The drawback of using a loudspeaker for playback is that it requires that the loudspeaker's directional properties match those of the original acoustic source. For a single voice, made by one person and coming from one direction, we can easily do that. But if we try to make a loudspeaker match up with the sound field radiating from a musical instrument, we run into trouble (*see figure above*).

What we really want to do is create objective perceptual measurements that can assess the complete life of a piece of music, from recording to playback.

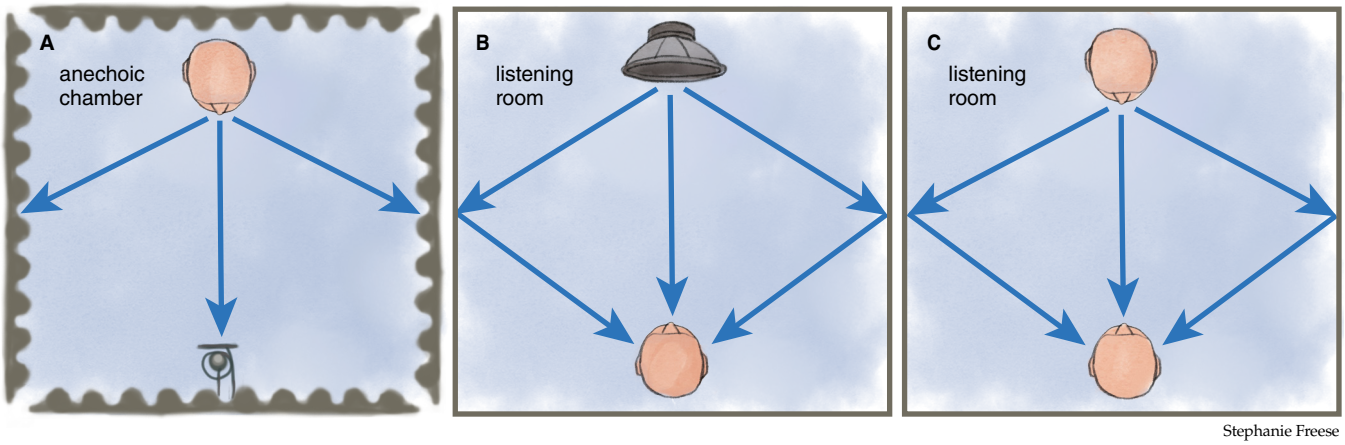
Musical instruments can have complex directivity patterns, with some frequencies more likely to reach the listener directly but others more likely to arrive via reflection, so recording them in an anechoic room will result in an unbalanced sound. Many modern recordings use electronic instruments that lack a natural reverberation, which introduces another issue. Audio engineers often add artificial reverberation to electronic instruments and to recordings made in sound-dampened studios; such reverb will also become imbalanced when applied to an anechoic room recording.

The situation becomes even more complicated if we apply a "dry" recording approach, with no added reflection or reverb, to a performance with multiple acoustic sources, such as an orchestra. To reproduce those sound locations, we would need a large (possibly very large) number of anechoic mono recordings played back over at least the same number of correctly placed loudspeakers. It's a rather impractical approach for a large orchestra that cannot be contained within a recording studio or a living room.

For recording live music, we strive to capture an immersive feeling similar to the experience of the original event. Ideally, the acoustics of the recording room would provide proper acoustic integration of all the instruments, including their directional patterns. In the room where we play back the recording, we want to reproduce the sound field as it would have been experienced live, taking into account the crucial feeling of immersion.

Re-creating the Immersive Experience

We now run into a dilemma, because we have arrived at two distinctly different approaches to the recording and playback of hi-fi sound. One is focused on transparency in the "here and now," optimizing the sound from a single, simple directional source. The other is focused on transparency in the "there and then," attempting to re-create the experience of a complex, multi-source, diffuse live event. The two approaches require completely different, incompatible recording and playback techniques.



Stephanie Freese

Recording and playback of a spoken human voice can be carried out effectively in an anechoic recording room (*a*), where the sound-damped walls mean that the microphone picks up only the direct sounds. On playback, a loudspeaker (*b*) that re-creates the voice produces the same direct and reflected sounds as does a human speaker (*c*) at the same location; in audio terms, there is a transparent relationship between recording and playback.

If we are aiming for the illusion of "there and then," we need to figure out the minimum number of audio channels required for hi-fi quality loudspeaker reproduction. We've known for a long time that one is not enough. The invention of stereophonic sound by British electronics engineer Alan Blumlein in the 1930s significantly improved the perceived loudspeaker reproduction quality of music events compared with mono. In stereo recordings, we can use time and intensity differences between the two channels to allow the listener to hear different musical instruments in different locations.

For headphone reproduction, two channels are sufficient, although they require meticulous, personalized head-related transfer function corrections. For loudspeaker reproduction, the sound quality is determined by a number of characteristics, of which the acoustics of the listening room is a dominating factor. For a high-quality audio experience, the acoustic resonances in the listening room should be damped and we should aim for a low reverberation time, preferably less than 0.5 seconds, allowing the listener to hear the longer echoes that were captured in the original recording environment. On the other hand, if the reverberation time of the listening room is too low, such as in an anechoic room, people lose the feeling of being immersed by the sound. To compensate for that effect, an extremely large number of audio channels would be required.

The reproduction of music is seldom improved by adding more playback channels beyond the typical two.

In a typical listening location, such as a living room, the number of audio channels needed for hi-fi quality loudspeaker reproduction of music events is not clear. Although expanding the number of recording-playback channels from one to two (from mono to stereo, that is) was a great improvement, extending that principle to four-channel "quadraphonic" sound was a commercial failure in the 1970s. The likely reason for the lack of public acceptance is that musical events seldom require localization behind the listener. In a concert hall, you seldom hear musical instruments behind you; the immersive experience of a concert performance is influenced instead by the more subtle, diffuse sound field that reaches your ears from all directions. To replicate that experience, a multi-channel system should reproduce only the diffuse field over the back channels.

The recent development of elaborate home theater surround systems with more than a dozen channels seems inconsistent with the characteristics that improve music reproduction. Commercial systems such as Dolby Atmos and DTS-X are useful mainly for watching films and playing games, media in which the sound effects require a more exact localization. The reproduction of music is seldom improved by adding more playback channels beyond the typical two. While the number of audio channels has been growing in home theater systems and high-end audio systems in vehicles, recording of music has remained mainly in stereo. In general, multi-channel systems introduce complexity in the setup and often introduce sound-localization errors that diverge from the live experience.

For music, the feeling of being immersed in a natural-sounding diffuse field is much more important than an improved sense of localization. Adding more reproduction channels can even lead to undesirable, uncontrolled degradations that people describe as "hearing things jumping around." Creating a high-quality, immersive diffuse field turns out to be quite difficult, however. Engineers have developed many complex algorithms for achieving such immersion, often using four speakers possibly with an added center one. But such so-called *two-to-five up mixing algorithms*, which extend stereo reproduction to five channels, tend to provide a poorer front sound-image quality along with only a marginal improvement in immersion. In most cases, listeners report that they prefer the original stereo reproductions, even though stereo audio cannot fully cap-

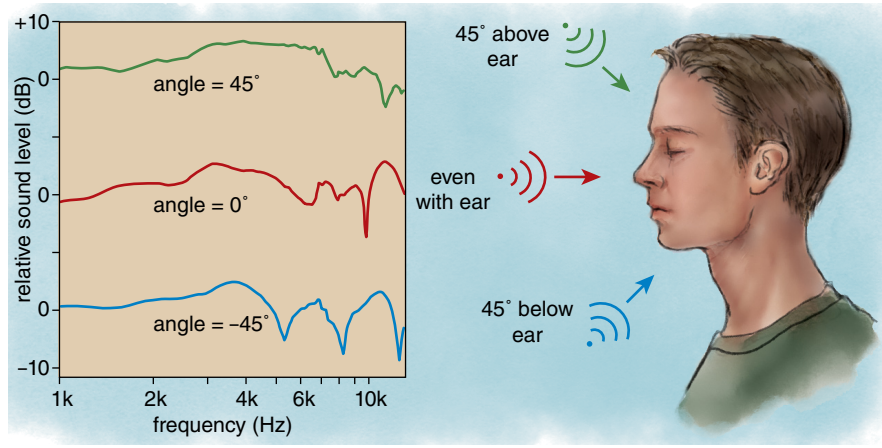
ture the feeling of immersion from a live music event.

The major reason why immersion is so difficult to attain is that it is a highly cognitive concept, one that was only recently introduced in the world of sound reproduction over loudspeakers. The feeling of being immersed is related to the perceived sound quality and is therefore difficult to define and measure. In general, when engineers discuss *quality* they are referring to two different dimensions: function and beauty. Quality optimization usually starts with the former. An excellent car, for example, should never fail in its function of transportation; it must fulfill that role with high reliability. Once function is achieved, the focus shifts toward beauty. But because sonic beauty lies in the ear of the beholder, it is difficult to quantify and optimize.

In sound-quality research, we have therefore focused more on functional quality aspects, such as localization, and less on beauty aspects, such as immersion. The first studies related to immersion were carried out in the context of speech perception, addressing familiar problems such as the functional difficulty of understanding a single voice when you are immersed in a loud party. The goal here is to improve functional localization in order to optimize speech intelligibility. The same basic motivation inspires home theater systems that prioritize localization accuracy over auditory beauty.

In recent years, audio researchers have begun to focus more intently on the beauty aspect of immersion. In a 2019 study, Callum Eaton and Hyunkook Lee at the University of Huddersfield in the U.K. asked a group of consumers and audio professionals to rate 10 aspects of sound quality in relation to immersion. Eaton and Lee found that horizontal sound perception was more important than vertical, but they could not determine to what extent subjects prefer to be immersed by a sound. If we take a single-voice recording and play it over a standard stereo setup or over four loudspeakers, increasing the number of loudspeakers will improve the feeling of immersion but will not improve the perceived sound quality. For this reason, a single direct-radiating loudspeaker is preferable for reproducing a single-voice recording.

Many audio designers have recognized the importance of widespread



Stephanie Freese

Perception of sound depends on the location and orientation of the listener relative to the source. For instance, the ear responds differently to sounds above, at, and below the horizontal plane. When the listener moves, therefore, the perceived sounds change. To make music seem more realistic when heard through headphones, audio engineers add a set of corrections (called *head-related transfer functions*) that restore some of the sense of location.

directivity of loudspeakers for high-quality music reproduction. At the same time, we know that multiple direct-radiating loudspeakers are not well suited to creating an immersive diffuse field for music. To improve the feeling of immersion, those designers have used additional sound drivers that do not radiate directly toward the listener.

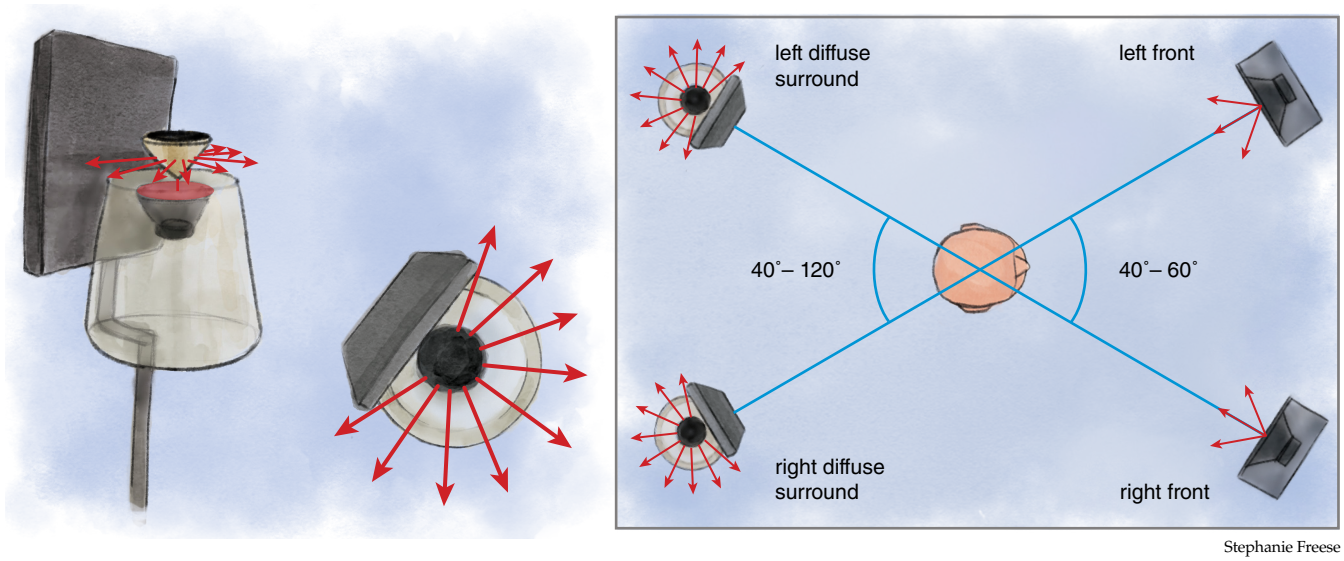
The best known of the people pursuing this direction is probably Amar Bose, founder of Bose Corporation, who in the 1960s designed a loudspeaker enclosure that has additional drivers in the back panel to produce reflections against the walls, thereby improving the balance between the direct and diffuse fields. In the 1980s, Kenneth Kantor and Alexander de Koster from Teledyne Acoustic Research in Cambridge, Massachusetts, extended the idea and developed an enclosure that uses extra backward radiating drivers to equalize the diffuse field room response independently from the direct field. One of us (Beerends) demonstrated the quality improvement from widespread directivity in 1988 for Dutch loudspeaker manufacturer BNS, using an extra set of back-radiating loudspeakers, which can be added to any regular stereo setup, to equalize the diffuse field response.

The weakness of all these setups is that they primarily create a frontally localized diffuse field. That distribution of sound does not closely replicate the diffuse field that a listener experiences during a concert-hall performance.

Once function is achieved, the focus shifts toward beauty. But because sonic beauty lies in the ear of the beholder, it is difficult to quantify and optimize.

A Simple Loudspeaker Solution

Today's home audio listening experience often falls into one of two extremes. At one end, we have a simple, tabletop Bluetooth speaker or a mono radio/television loudspeaker producing a single-source sound with one exact location, allowing excellent speech reproduction. At the other end, we have elaborate, multi-channel home theater setups producing highly detailed but mostly exaggerated localizations. In the middle of these extremes,



Stephanie Freese

The authors' experimental loudspeaker setup can produce a realistic mix of direct and diffuse sound. A cone-shaped diffuser (*left*) radiates sound in all horizontal directions (*arrows*), while a sound-absorbent block (*gray*) shields the listener from sounds that would arrive directly. Two front loudspeakers (*right*) create direct sound while two rear loudspeakers create an adjustable level of diffuse sound, mimicking the immersive experience of a live concert.

we have the traditional stereo setup that many people still use for listening to their favorite music. However, none of these designs does much to recreate a diffuse sound field that allows for a rich, immersive music listening experience.

We see a big missed opportunity, because excellent quality of immersion can be achieved using ordinary stereo recordings reproduced by a regular stereo loudspeaker setup, complemented only by two additional omnidirectional loudspeakers that project most of their sound energy toward the walls. In our experiments, we have shown that the two additional speakers can be designed to contribute only to the diffuse field, so the degree of immersion can be easily controlled without introducing localization errors. This setup also reduces undesired *comb filtering effects*, the sharp frequency peaks and dips that arise when sound waves interfere between the front and rear loudspeakers.

We have devised a simple but effective way to create a diffuse-radiating surround speaker using a cone-shaped diffuser that produces, for a substantial part, a 360-degree pattern of sound that radiates horizontally. Optimally, the speaker is designed to minimize its contribution to the direct field, for example, by limiting the actual radiation to about 300 degrees.

The basic layout of a complete loudspeaker configuration designed for an optimal sense of immersion can be adapted to one's personal preferences (*see figure above*). In our setup, the left and right diffuse speakers mainly radiate toward the walls of the listening room, as opposed to the standard surround setups in which the surround speakers radiate directly toward the listener. This approach prevents the "things jumping around" effect. Our setup can't create a full three-dimensional diffuse field because it is designed mostly to spread out sound along the horizontal plane, but the feeling of immersion is dominated by horizontal sound anyway.

The proof of the playback is in the listening, so over the past few years we have carried out a series of experiments in cooperation with a number of small hi-fi companies in The Netherlands. These experiments were conducted in four locations: three in a professional listening room, and one in a home environment. Both professional audio engineers and nonexpert listeners were asked to set the optimal playback level of the front loudspeakers, after which they were asked to adjust the level of the diffuse surround speakers for maximum perceived overall audio quality. We also adjusted the time delay between the surround speakers and the front ones, to keep the main stereo image (the sense of sound location) stable and prevent the rear speakers from creating unwanted localization from behind.

For the delay, we found that the optimal value was between 10 and 20 milliseconds, 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-music recordings. The optimal volume level for the front speakers depended marginally on the preference of the test subject and not on the properties of the recordings, as they were equalized in loudness. The optimal level for the surround speakers depended significantly on both the test subjects and on the properties of the recording.

We were interested to learn that listeners' preferred levels for the diffuse field varied significantly. Some subjects set the level very low, close to the minimum noticeable volume, about 20 decibels below the level of the direct field loudspeaker. Others choose to set the diffuse sound level very high, even above the volume of the direct field loudspeakers. We also gave our testers the option to turn off the surround speakers entirely. Among the 24 test subjects, 23 chose to switch on the extra diffuse field speakers for most of the music samples, and 16 subjects chose to keep the speakers on the whole time. Even our least enthusiastic subject switched on the diffuse speakers for 43 percent of the samples.

Overall, our testers reported a significant increase in perceived overall sound quality when the diffuse surround speakers were switched on. Using a 5-point scale, ranging from 1 (a very small improvement) to 5 (a very big improvement), the audio experts judged the overall sound quality im-

Piano Piano!/flickr.com/photos/hansthijs/3585941755/

Quadraphonic sound was an attempt by the audio industry to create a home hi-fi experience that was more immersive than conventional stereo. Despite the wide availability of quadraphonic recordings and equipment in the 1970s, the technology flopped—probably because it failed to capture the way the people actually experience immersion and the locations of sounds.

provement around 3 on average. The nonexpert listeners judged the quality improvement even bigger, with average scores around 4.

The most encouraging aspect of these experiments is that only two small additional surround speakers were needed to produce a significant increase in overall perceived sound quality. Our diffuse field approach did not introduce the degrading localization errors that occur in many surround-sound systems. The setup we created allows for a simple "immersion control": Listeners can easily adapt the main volume, diffuse volume, and time delay characteristics of any standard stereo recording to their personal immersion preferences.

Hi-Fi in Your Life

The long quest for high-quality, widely accessible hi-fi audio is far from over. The extreme dependence of the optimal audio experience, especially perceived immersion, on personal preferences makes it difficult to design an objective system for assessing the overall sound quality of a system. For mono speech and music, and to some extent stereo music, audio engineers largely have conquered the basics. Perceptual models have been developed that show good correlation between objective measurement and subjectively perceived speech and music quality. There are also useful models for spatial audio quality, although they do not take into account personalized immersion optimization.

The major shortfalls of currently available commercial audio systems are that most of them provide only limited or ineffective amounts of immersion, and that none of them allow easy adaptation of immersion to personal preferences. Those preferences also vary strongly depending on the room in which the sound reproduction takes place. One of us (Beerends) has been experimenting with home theater systems that can generate artificial sound reflections, using algorithms to simulate the acoustics of concert halls. This approach allows listeners to optimize the feeling of immersion in rooms that sound too dry, lacking enough acoustical reflections. However, such systems do not make it easy to dial in an optimal level of immersion, and they can lead to sound localization errors.

The diffuse sound setup that we developed offers a simpler yet effective way to optimize the feeling of immersion, but for now it exists only as a prototype. Currently no company manufactures such a system. We hope that this article will encourage manufacturers to commercialize a system that can be hooked up to any standard hi-fi set, allowing for easily controlled and optimized immersion into the music.

Bibliography

Beerends, J. G., and J. A. Stemerdink. 1992. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society* 40:963–978.

Beerends, J. G., and J. A. Stemerdink. 1994. A perceptual speech quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society* 42:115–123.

Beerends, J. G., A. P. Hekstra, A. W. Rix, and M. P. Hollier. 2002. PESQ, the new ITU standard for objective measurement of perceived speech quality, part II—perceptual model. *Journal of the Audio Engineering Society* 50:765–778.

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.

Eaton, C., and H. Lee. 2019. Quantifying factors of auditory immersion in virtual reality. In: *2019 AES International Conference on Immersive and Interactive Audio*, eds. T. Tew and D. Williams. Audio Engineering Society.

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.

John G. Beerends researched speech, audio, and video quality assessment for Dutch telecom operator KPN. For the past 22 years he worked for TNO (Dutch Organization for Applied Scientific Research), focusing on the idealization approach in speech quality assessment. Richard van Everdingen has worked in audio and video distribution, with an emphasis on audio quality optimization. In 2010 he founded Delta Sigma Consultancy, where he has conducted research on audio and video-related issues. Email: johnbeerends@hotmail.com