



Audio Engineering Society

Conference Paper 26

Presented at the International Conference on Spatial & Immersive Audio, 2023 August
23-25, Huddersfield, UK

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Immersive Audio Test Signals for Musical Applications

Arthur Moelants¹

¹PXL University of Applied Sciences and Arts

Correspondence should be addressed to Arthur Moelants (arthur.moelants@gmail.com)

ABSTRACT

This paper describes a collection of test signals that are specifically tailored to current and future challenges in immersive audio listening formats on both speakers and headphones. As the human auditory system uses a combination of different techniques to localize sound, immersive audio reproduction systems should take these factors into account when their goal is to achieve accurate localization. The elements in these test signals were specifically selected to challenge the various aspects of sound localization, especially in musical applications, and can thus highlight any possible flaws in immersive sound reproduction systems. On top of that, some more practically oriented test signals are included, created to evaluate and calibrate the channel order and loudness of multichannel playback systems.

The test signals are available for download at <https://www.immersive.pxl.be/downloads> or through request via the email address listed above.

1 Introduction

These test signals were conceived in demand of qualitatively comparing different immersive music playback systems, as well as for the development of immersive technology. On top of the increased channel count, immersive sound reproduction systems also face new challenges as sound localization is expanded into three dimensions.

The first part of this paper is a summary of most relevant localization techniques, the properties of sound reproduction systems and sound itself, regarding clarification on the reasoning behind the included elements of each specific test signal. In the second half, the test signals and included elements and sounds will be described and linked to the relevant localization aspects.

Four types of tests were created: Musical Test, Localization Test, Layout Test and Calibration Test. The first two are referred to as 'subjective' tests, as they are supposed to aid in evaluation the localization quality of immersive technology components through subjective listening tests. The latter two are 'objective' tests, as they aid in correctly setting up any immersive playback system.

2 Localization principles

Before the test signals can be discussed, a basic understanding of localization principles is necessary, as the content of these tests are derived directly from them. The following section provides an overview of all relevant methods that humans use to localize sound in space, and some relevant consequent acoustic phenomena.

2.1 Human auditory system

While this paper does not aim to give a complete explanation of localization principles, there are still a few key elements that should be pointed out.

For a more thorough guide on how human localization works, Agnieszka Roginska & Paul Geluso's book "Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio" (2017) would be highly recommended [2].

The human auditory system uses several techniques to localize the position of a sound source. These techniques can be divided into three categories: interpreting information that requires only one ear (monaural cues), interpreting information that requires the combination of two ears (binaural cues) and, lastly, using movement.

2.1.1 Binaural cues

Humans have two ears. Even though we can already localize fairly well using only one ear (see 2.1.2 Monaural Cues), being able to compare stimuli from both ears to each other helps tremendously with localizing sound.

To locate a sound source on the horizontal plane, our brain relies heavily on these binaural cues, utilizing the interaural level difference (ILD) and interaural time difference (ITD) between our ears[1]. For example: when a sound comes from the left, it will reach the left ear earlier and will usually be louder too. However, ILD and ITD do not work equally across the entire audible frequency range [2].

ILD works better for higher frequencies, whereas we rely more on ITD for lower frequencies. Without going into too much detail, this difference is largely due to our head being an obstacle between our ears, as well as the physical distance between them. When a sound's wavelength is longer than the distance between our ears, it can diffract ("bend") around the head and won't have a significant level difference (ILD), yet we are able to detect time and phase differences (ITD). When the wavelength is shorter than the distance between our ears, our head will form an obstacle, reflecting and absorbing sound, creating a noticeable level difference (ILD).

Depending on the exact distance between a person's ears, the general crossover frequency can be found around 1500Hz (22-23cm) [3] to 2000Hz (17.5cm)

[4]. However, there is a significant overlap between the two frequencies. Furthermore, Brughera, Dunai and Hartmann have demonstrated that the frequency range of a source has an influence on these ILD and ITD threshold frequencies. ITD can be detected for sine waves down to 1400Hz, whereas for wide, broadband signals that limit can be up to 4000Hz. [5] Especially in music (where sound sources are rarely pure sine waves) the importance of the frequency range of the source becomes apparent when trying to evaluate the localizability of an immersive playback systems.

2.1.2 Monaural cues

For monaural cues, stimuli are only presented to one ear (either the other ear is completely blocked, or the sound is presented via headphones). However, information, like depth or elevation, can still be extracted [2] [6].

2.1.2.1 Depth perception

To estimate the distance of a sound source to the subject, there are four possible principles that can be used [6] [7].

1. The intensity of a sound source will be lower at greater distances. When we take a familiar sound (see Section 2.2.3), like a calmly talking voice, we have a certain expectation of how loud it should be at a certain distance.
2. The ratio of direct and indirect (reflected, reverberant) sound changes dependent on the distance to subject.
3. Higher frequencies are attenuated over greater distances, as these find more resistance in the air, compared to the large wavelengths of lower sounds (imagine the sound of a music festival from a large distance).
4. The correlation between the sound arriving at the left and right ear also reveals how close by a source presumably is. The more correlated the signals at both ears are, the closer it is.

The first three phenomena can solely rely on monaural cues. The fourth one, however, does utilize interaural (binaural) cues.

2.1.2.2 Spectral cues

Through direction-dependent filtering, a person can localize sound, both in the horizontal, as well as in the median (vertical) plane.

For this, the detailed geometry of the outer ear, or pinna, is used [1] [8]. The shape of our outer ear influences phenomena such as reflection, shadowing, dispersion, diffraction, interference, and resonance differently for specific frequency bands, depending on the relative position of the sound source to the head.

As interaural cues (ILD and ITD) are absent in the median plane, elevation is determined using only spectral cues (loudness based on frequency), utilizing the detailed geometry of the pinnae [1] [8]. High frequency components are emphasized for elevated source, while being attenuated for sound coming from the back [1] [2].

2.1.3 Movement based localization

A listener's movement can aid in localization abilities and provide clarity. Head movement shifts both binaural and monaural variables. It is limited to a fixed body frame with three axes (pitch, yaw and roll).

When we move our head, the perceived timbre changes and the direction-dependent filtering shifts in a way that further informs us where a source is located.

Aside from that, the general principles of ILD still remain. Due to the "head shadow" effect, a source will still be attenuated when the sound source is at the opposite side of the head [7].

Head movements improve localization performance and can all but eliminate reversals between, for instance, front and back [2]. Neglecting head movement in immersive playback systems, however, will not just reduce the localization accuracy, but will completely confuse the localization system and therefore break any intended illusion.

On speakers, head movement correspondence is 'built in' as speakers are stationary within a room. When using (classic, non-headtracked) headphones, however, the sound follows the movement of the head and will result in 'inside-the-head-locatedness' (see Section 2.4.2) [2].

Aside from moving our head, physically moving our position towards or away from a source can shift monaural cues and improve localisation.

Furthermore, the sound source itself can also move through space.

2.2 Important factors

2.2.1 Frequency range

The frequency range of a sound is the spectral width it covers and can be either wide (or broadband), for

instance 20Hz to 2000Hz (± 10 octaves) - e.g., pink noise - or very narrow, like a sine wave.

Broadband sources are easier to localize, compared to narrowband sounds [2] [4].

2.2.2 First wavefront

The first part of a sound that arrives at our ears is called the 'first wavefront', also referred to as 'transients', or 'attack'. The stronger the first wavefront, the easier the sound is located. [2]

2.2.3 Familiarity and predictability

The more familiar we are with a certain sound (e.g., human voice), the easier we can locate the position of the source [1] [2] [4]. For spectral cues, it can be important to be familiar with the spectral contents of a source.

A good example is the famous "The Virtual Barber Shop" video on YouTube [9]. As it is only a static binaural experience, it uses only binaural cues (ILD and ILD, see Section 2.1.1.1) and monaural cues (spectral filtering and depth through space, see Section 2.1.1.2), while ignoring movement (see Section 2.1.1.3). However, it still gives a relatively convincing result. This is in large part due to the familiarity, (absent) complexity and predictability of the sound sources. The most prominent sound source is a human voice, which has very familiar spectral content. When dealing with less familiar sounds (e.g., an electric guitar hooked up to a specific amp), it becomes harder to reverse-engineer what the original source sounds like without spectral localization cues. On top of that, the narrative structure of "The Virtual Barber Shop" and easy to imagine context, make it easy to predict or follow how a sound source moves through the room.

2.3 Psychoacoustic phenomena

Aside from physical events in the auditory system, there are also several noteworthy psychoacoustical phenomena that impact localization.

2.3.1 The "Pitch height" effect

The perceived position of a sound source can vary depending on the spectral information. This is especially true for narrowband sounds and is described as the "pitch height" effect [4] [10] [11].

In experiments, a narrowband source (pink noise bandpassed at 1/3 octave) with a center frequency of 8kHz was perceived to come from above, while in reality the loudspeaker was at ear level. [1] Blauert referred to the frequency band by which the direction

of sound image is determined as the ‘directional band’ [10], which were researched by Hyunkook Lee and Rory Wallis [13][14].

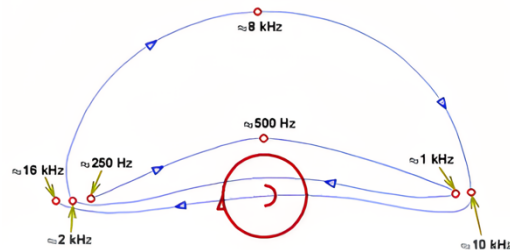


Fig. 1: Directional bands.

Courtesy of J. Blauert (1997) [1], edited by Audiologieboek.nl [4]

2.3.2 Inside-the-head-locatedness

When a sound source follows our exact head movements (e.g., when we listen through a normal pair of stereo headphones), a sound won’t appear to be outside the head. It seems to be somewhere on an imaginary axis between our ears. Even when a sound is positioned in the center, it should – in real life – be perceived on the left when we turn our head right. As long as the position of the source follows the head movement, instead of keeping its position in space, we can only assume it’s in our head.

2.3.3 Masking and BMLD

Masking occurs when two similar sounds play at the same time, or from the same general location. One sound masks the other, confusing the perception of either sound. Masking can thus be prevented by having the sound sources come from different directions [6].

Masking can also occur when sound comes from two different discrete sources. The sound that arrives the earliest will mask the later one [14].

The binaural masking level difference (BMLD) is a phenomenon that expands on this. It was discovered by presenting a sound that is identical at both ears, while masking it by a noise that is also identical at both ears. When flipping the polarity of either left or right side of only one of the signals (and consequently creating a different perceived direction), the masked sound can be perceived more easily.

2.3.4 Phantom sourcing

When identical sound information is played through more than one speaker, the position of sound will be perceived in between those speakers, creating a ‘phantom source’ [2]. Changing the level or delay of one of the channels, relative to the other(s), will shift the position towards the loudest or earliest speaker, as it affects the ILD or ITD.

3 Sound reproduction qualities

This section provides an overview of all relevant variables that can be found within the sound reproduction system and the sound source itself, as they both have an influence on the localization parameters discussed in Section 2 and thus have an influence on the localization accuracy.

In theory, any immersive sound reproduction system should comply with all localization principles at once. A binaural system, for example, shouldn’t only aim to reproduce binaural cues (Section 2.1.1), but should also focus on head movement (Section 2.1.3), as well as monaural cues (Section 2.1.2), so that when one earphone gets taken off, localization can still be done using only the present monaural cues, as one would when covering up one ear in daily life.

3.1 Playback system properties

Immersive sound reproduction systems, like multichannel speaker layouts, binaural headphone systems and immersive soundbars, aim to mimic the localization facets, if not naturally present. There are several factors that can either improve or degrade the quality of good localization on any playback system.

3.1.1 Spatial resolution

The number of (virtual) channels in an immersive playback system can influence its location accuracy. The more discrete channels, the clearer a sound can be pinpointed in space. Fewer channels, on the other hand, will result in phantom sourcing (Section 2.3.4).

3.2.2 Positioning

The physical position of the (virtual) speakers, and their relation to each other, needs to be considered too. Most immersive systems work with speakers all around the listener, yet some systems work with only a row of front speakers. To get a proper translation, the listener must have their speakers set up in the way the manufacturer of the format intended.

3.2.3 Room acoustics

The need for an acoustically well-treated room is not unique to immersive formats, yet the increased number of speakers will highlight poor room acoustics even more.

On headphones, virtual room acoustics can have a major influence on localization ability and overall ‘immersiveness’ or ‘spaciousness’. Adding additional reflections aids in verifying binaural (Section 2.1.1) and monaural (Section 2.1.2) cues.

3.2.4 Filtering and frequency response

Immersive playback systems may be programmed to filter specific channels. On headphones, the use of an HRTF will not only impact the overall sound, but also the localization ability through binaural and monaural cues, especially spectral cues (Section 2.1.2.2).

Additionally, the frequency response of the drivers of the speakers or headphones might also have a meaningful impact on the localization quality.

3.2.5 Data compression and codecs

Finally, possible data compression should also be considered, as some important localization information might be lost, depending on the codec used. Note needs to be taken that some systems compress beds and objects differently.

3.3 Properties of the sound source

After discussing several techniques and challenges relevant to localization, it is equally important to analyze the properties of audio signals.

3.3.2 Acoustic properties

Sound can be broken down into four elements. These elements can be described as:

1. Frequency
2. Amplitude
3. Waveform
4. First wavefront

3.3.3 Processing properties

Additionally, sound can be processed, resulting in a change in any of the above properties. This processing might however also impact the localization ability.

1. Filtering (EQ)
 - Spectral cues (2.1.2.2)
2. Dynamic processing (Compression, Distortion...)
 - First wavefront (2.2.2), ILD (2.1.1)
3. Modulation (Vibrato, Chorus, Flanging, Phasing)
 - Doppler effect [2], Spectral cues (2.1.2.2)
4. Reflections (Reverb, Delay)
 - ITD, ILD (2.1.1), monaural cues (2.1.2)
5. Level in relation to other sounds
 - Masking (2.3.3)

3.3.4 Locational properties

An additional relevant property that can be attributed to a single sound source, is its location relative to the listener, defined utilizing:

1. Azimuth (angle on horizontal plane)
2. Elevation (angle on median plane)
3. Distance

Additionally, a sound source can be either stationary or moving. Movement can be described through:

1. Direction
2. Speed (constant or changing)
3. Predictability

4 Relevance

The combination of several localization techniques is what makes our auditory system capable of localizing sound in a (virtual) space. By combining factors, we not only have several independent localization methods, but also create a system that can use one factor's output to verify another's. However, if there are discrepancies between these factors, confusion will occur.

Example 1: A sound seems to be coming from the left, according to ILD, ITD and spectral information, yet when the listener turns their head, the sound follows that exact movement.

This results in 'inside-the-head-locatedness' 2.3.2 and usually happens with headphones that don't provide headtracking.

Example 2: A sound was positioned at a certain height in the median plane during mixing, yet the playback system of the listener does not provide the necessary spectral cues to localize that sound as such. As a result, the listener will perceive the sound to be at ear level.

Some immersive playback systems manipulate sound to achieve a desired perceived position without the need for a physical sound source at that specific position. On headphones HRTF filtering can be used to emulate the spectral cues our pinnae would give to create the illusion of sound being above or behind. Even on speakers phantom sources are used as a way of tricking the auditory system.

However, not all immersive playback systems succeed equally well in creating these illusions. Being able to compare them in their ability to meet the various localization factors becomes increasingly important.

For that, a specifically designed test signal, that contains all necessary elements to evaluate whether an immersive playback system fails in any of the earlier described localization factors, is necessary.

The proposed test signal that will be described later in this paper, can be useful for evaluating the quality of existing immersive playback systems and their components, or for qualitative, controlled testing when creating new technology.

More often than not, immersive technology is tested using a variety of pre-existing songs or audio samples. The drawback of this approach is that most sound excerpts do not readily give a clear readout of all possible factors that should be taken into account. While a good, thorough test using a multitude of sources material will always be key to making sure technology works the way it was intended to, there aren't always enough time and resources for thorough testing.

A good test signal, specifically designed for immersive listening formats, should provide a way to test as many factors as possible using only a few organized tests, which should point out the flaws in a system quickly and clearly.

Through repeatable test setups it is possible to check how accurate a playback system is, compared to another. As it's usually rather hard to properly do an "AB comparison" with immersive test setups, a subjective listening test, with a consistent questionnaire, can make qualitative comparison possible.

The proposed test signal consists of elements that can clearly be pointed out. As an example, a possible way to question a listener could be through a 'radar map', on which they can indicate a point, representing the perceived location of a specific sound.

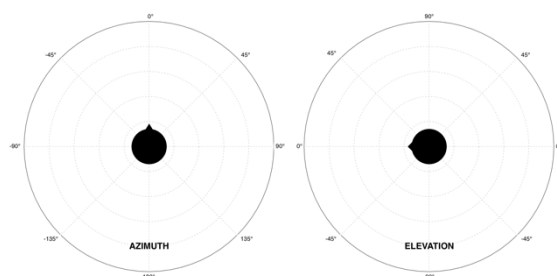


Fig. 2: localization of a specific sound can be questioned using a radar map. The left circle is used for indicating the azimuth, the right one for the elevation.

5 State of the art

At present, each immersive sound reproduction system comes with its own demos, which are often made to impress, rather than assist to a critical view on its ability to correctly localize sound.

Pre-existing songs can also be used to test these sound reproduction systems. However, different immersive playback systems require different inputs. The number of channels and the placement of speakers is rarely entirely the same across different systems. A

song is usually mixed for one immersive format and can thus not be used to properly compare the different playback systems.

While test signals exist, it is presently still hard to find them in multichannel formats that comply with today's standards, ranging anywhere from six to sixteen channels and beyond.

Another problem can be that a test signal is purposely made exclusive for their own specific playback system. Apple Inc. has a qualitative test signal – which even served as part of the inspiration for our musical test signal – yet it can exclusively be listened to through the Settings app on an iPhone, while wearing Apple's AirPods earphones. Consequently, it's rather impossible to critically compare Apple's Spatial Audio system to other immersive systems, without an external test signal that can be used across different immersive sound reproduction systems.

6 The test signals

Based on the earlier described criteria, four groups of test signals have been created.

- 1) Musical test (subjective) ^{6.2}
- 2) Localisation test (subjective) ^{6.3}
- 3) Layout test (objective) ^{6.4}
- 4) Calibration test (objective) ⁰

6.2 Musical Test

The Musical Test signal was designed to measure the localization accuracy of an immersive sound reproduction system within a musical context.

The Musical Test signal is an eight-bar loop in C major at 120BPM, containing nine different sound elements. Some immersive reverb (LiquidSonics' 'Cinematic Rooms' (Section 7.13)) was added, to represent mixing decisions in most immersive music mixes.

The Musical Test signal consists of the following elements:

1. Bass
2. Piano
3. Pad
4. Celesta
5. Kick
6. 'Machines'
7. Pluck
8. Sticks
9. Shaker

These elements were carefully chosen during the composition process as they are built up of varying sound characteristics (Section 3.2.1). As a result, several localization principles, must be used by the listener's hearing in order to locate all sound sources. As discussed in (Section 2.1), a listener will use binaural, monaural, spectral cues (and additionally head or physical movement), each of which works better for a certain sound characteristic. The Musical Test is therefore designed to include varying elements that each challenge different aspects of an immersive sound reproduction system's capability to adhere to these hearing principles.

There are four melodic elements: Bass, Piano, Pad & Celesta. That span over different frequency ranges and each have different transient behavior, ranging from very prominent transients (the Celesta) to none at all (the Pad).

The other five elements are percussive, however they too span a wide frequency range altogether, ranging from only containing high frequencies (the Shaker) to low ones (the low thumping 'Machines' sound).

6.2.2 Movement

In the different versions of the Musical Test, varying levels of movement are present. These variants are divided into five categories that each have a certain degree of movement:

1. **Static:** no moving elements
2. **Simple movement:** one element, moving horizontally or vertically
3. **Medium simple:** Two elements moving simultaneously in different directions
4. **Medium complex:** more than three elements moving in several directions
5. **Complex:** all elements moving in any possible direction

6.2.3 Technical

In a Dolby Atmos or AuroMax configuration, the Bass, Piano and Pad sounds are assigned as beds, whereas the other, percussive, elements are all objects.

This allows for fast and instinctive discovery of a routing problem in either the objects or the beds. If only melodic sounds are heard, there is likely a routing error in the objects. If only rhythmic elements are heard, there will be a problem with the bed.

The Celesta has both very rhythmic, as well as melodic properties. Due to its 'Left Wide' position however, it is assigned as an object, but more on that later (Section 7.4)

The position of each element in the Musical Test can be seen in figure 3.

Bed elements are squares, whereas objects are represented as spheres.

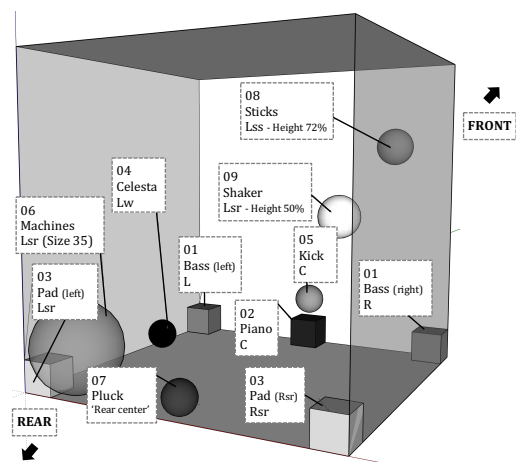


Fig. 3: Position of the elements in the Musical Test. Channel based sources represented by cubes, objects by spheres.

Furthermore, there can be one or more moving elements within the loop, depending on the specific test chosen. If a person is not able to correctly identify the movement direction of the source, it usually implies one or more discrepancies between the localization factors.

The Musical Test has a loudness of -18 LUFS and a True Peak of -3dB.

6.3 Localization Test

The second test, the Localization Test is designed to evaluate the quality of the localization cues in an immersive sound reproduction system through subjective-listening testing.

6.3.2 Predictability

The test comprises two components. On the one hand, it involves testing through a predictable pattern in which the listener can anticipate the location of the sound source in the space, and on the other hand, through an unpredictable pattern where the listener does not know in advance where a source will be positioned.

6.3.2.1 'Predictable pattern' test

Through the predictable pattern, we can determine whether an anticipated movement corresponds to its perceived path. For example, a circular movement in which the elevation is not changed should not give the

impression that the source is moving through the median plane.

In one of the ‘predictable pattern tests’, a voice describes its current and upcoming location while moving through the virtual space in a simple, continuous (predictable) motion.

An additional test with an instrument (the Shaker) is also available. This takes away the psychological effect of being informed of where the sound should be perceived.

An interesting study can be performed when the spoken position is uncorrelated to its actual position. For instance, when the voice says, “I’m going left”, yet in reality it’s going up. A proper immersive playback system should allow the listener to easily identify this discrepancy.

The ‘predictable pattern’ test consists of the Voice and the Shaker elements.

6.3.2.2 ‘Unpredictable pattern’ test

Alongside the ‘predictable pattern’ test, an evaluation through an unpredictable pattern can be made too. This allows verification on whether an immersive listening system is accurate enough to localize a sound source without prior knowledge about its upcoming position. This is done by making the sound appear at random positions. During a subjective listening test, the listener can indicate the perceived position on a radar map (see fig. 2), or any alternative.

The ‘Unpredictable Pattern’ test consists of seven elements that feature varying localization difficulties: Voice, Shaker, Live Recorded Piano, Cello, Celesta, Sticks and Pad.

6.4 Layout Test

The Layout Test is, in contrast to the two-prior subjective-listening tests, a technical, objective test, designed to evaluate the routing of an immersive playback system.

One channel at a time, a voice indicates the expected position of that channel, e.g. “One, left”, or “Ten, right top front”.

If an indicated position is not correlated to the actual, physical position in space, it uncovers an error in the patching. If a channel number is skipped, it implies a defective component in the signal chain.

As there are many immersive formats and channel orders, mistakes can easily be made and aren’t always easily discoverable through listening content (e.g. music or a movie) due to phenomena like masking. Through the use of e.g. a DAW, panning tests with a simple signal (e.g. pink noise) can easily be done,

however, not all listening situations offer the ability to listen through a DAW. Furthermore, using the Layout Test from this collection of test signals insures that the channel layout and channel order are the same as for the other tests in the collection.

The Layout test is thus a good first starting point when setting up an immersive playback system and makes it possible to accurately verify that all channels are connected and working as expected.

6.4 Calibration Test

The fourth and final test is designed to measure or calibrate the loudness of an immersive listening system per channel. As with the Layout Test, this test is intended for conditions where the playback system itself does not include a test signal, whilst also ensuring that the levels match with the other test signals from this collection.

While only pink noise is needed for this test, there doesn’t currently seem to be a test signal available as ADM or any other type of Wave file that goes up to sixteen channels (9.1.6) or higher.

The Calibration Test consists of a pink noise signal, measured at -20dB RMS, with -6dB True Peak.

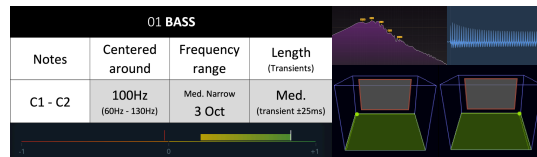
7 Elements

In this section, each element that is part of at least one test signal will be described, including its properties and the possible localization principles it can help to evaluate.

For each element a table is included that sums up the range of notes being played, the frequency the sound is centered around, it’s audible frequency range (the range that is louder than -20dB), and the total length of each sound. Additionally, a rough representation of the frequency spectrum is included, as well as a waveform that can give an indication of the transient behavior. For the stereo sounds, a phase correlation meter is also included. Lastly, the position of the sound is represented as a small green dot in the cube. The images are intentionally kept small as they would otherwise take up too much space in this paper.

Alongside the properties of the source itself, for each element there is a list of localization properties it can point out and why that is.

7.1 Bass



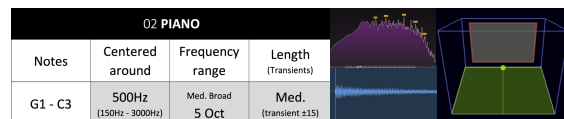
Origin: Logic Pro X, Retro Synth synthesizer.

The Bass is included in the Musical Test 6.2 as a low, sustained element with a relatively short (2.2.2) first wavefront. It's a stereo source, routed to the front left and right (L, R) as well as the LFE channel.

This element can point out:

1. **Masking** (2.3.3) due to overlapping frequencies with Kick, Piano, Pad and 'Machines' at lower spatial resolutions.
2. Correlation for frequencies below 500Hz.
3. **Phantom sourcing** (2.3.4) as the sound is present in L, R and LFE channels.
4. **LFE control** (objective).
 - The functioning and level of the LFE channel of a playback system can be evaluated, knowing that:
 - Bass main level: -9dB
 - Bass LFE level: -6dB

7.2 Piano



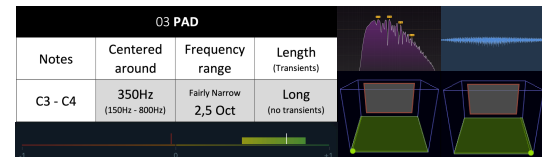
Origin: Logic Pro X, Sampler, 'Grand Piano'.

The sampled Piano was added to the Musical Test 6.2 as an easy to localize, center positioned element (C).

This element can be useful for its:

1. **Ease of localization** due to its wide frequency range (2.2.1), short first wavefront (2.2.2) and fairly familiar (2.2.3) sound.
 - As it's positioned exactly at the center, it can be a clear **anchor point** to compare the position of other elements to during a subjective-listening test.
2. **Masking** (2.3.3) with Bass, Kick, Pad and 'Machines' at lower spatial resolutions.

7.3 Pad



Origin: Logic Pro X, ES E synthesizer.

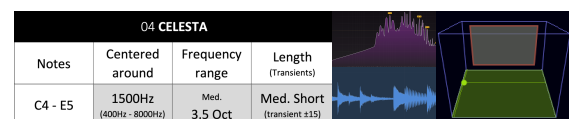
The Pad is a difficult to locate element and is featured in the Musical 6.2 and Localization 6.3 tests.

It's a stereo source, positioned at the rear (Lsr, Rsr).

This element can point out:

1. **Localization difficulty** as it's fairly narrow-band (2.2.1) and has no transients.
2. **Directional bands** (2.3.1) around 200-500Hz, possibly resulting in a more frontal perceived location.
3. **Correlation** (ITD, 2.1.1) for frequencies below 500Hz.
4. **Masking** (2.3.3) with Bass, Piano, Kick and 'Machines' at lower spatial resolutions.

7.4 Celesta



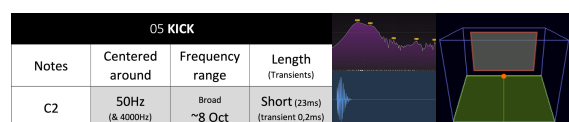
Origin: Logic Pro X, Sampler, 'Celesta'.

The Celesta is the Musical Test's 6.2 main melody and was added as a midrange element with short first wavefront (2.2.2). It's placed at the 'left wide' (Lw) position.

This element can point out:

1. **Spatial resolution** due to the less common 'Lw' position. Absence of a discrete channel may result in phantom sourcing (2.3.4).
2. **Directional bands** (2.3.1) due to the narrowband sinusoids of whom the Celesta is made up of. A higher up, or further back position may be perceived.
3. **Data compression** (3.2.5) on objects with a melodic, continuous character.

7.5 Kick



Origin: Logic Pro X, Ultrabeat, 'Advanced Machines' library.

The Kick is an element within the Musical Test 6.2, as a broadband, low-end emphasized, percussive sound, and is positioned at the front center (C). This object-based element is the only one utilizing the ‘size’ parameter (set at 35).

This element can point out:

1. **Ease of localization** due to the fast first transient (2.2.2) and relatively familiar (2.2.3) character.
2. **Masking** (2.3.3) with Bass, Piano, Pad and ‘Machines’ at lower spatial resolutions.
3. **Data compression** (3.2.5) on objects, as lower frequencies may be sacrificed by compression codecs.

7.6 ‘Machines’

06 MACHINE			
Notes	Centered around	Frequency range	Length (Transients)
G2	150Hz (90Hz - 500Hz)	Med. 3,5 Oct	Short (20ms) (transient 5ms)

Origin: Logic Pro X, Sampler, ‘Advanced Machines’.
Panning: Size parameter set at ‘35’

The ‘Machines’ was added to the Musical Test 6.2 as an unfamiliar, percussive, low sound at the left rear position (L_{sr}). This object-based element is the only one featuring the ‘size’ parameter (set at 35).

This element can point out:

1. **Size parameter support:** The element can be used to verify whether a playback format supports the ‘size’ function.
2. **Localization difficulty** for low, unfamiliar (2.2.3) sounds. Although ‘Machines’ is also a short (2.2.2), low, percussive sound, which, on the other hand, may aid localization through ITD (2.1.1).
3. **Directional bands** (2.3.1) that might occur at 200Hz and, 500Hz, resulting in more frontal perceived position, as opposed to its actual rear placement.
4. **Data compression** (3.2.5) on object-based elements in the rear position. As they are both objects featuring low frequency content, ‘Machines’ can be evaluated against the Kick element, which is positioned in the front and might be compressed differently, depending on the codec used during (re)export or streaming.
5. **Masking** (2.3.3) with Bass, Piano, Pad and Kick at lower spatial resolutions.

7.7 Pluck

07 PLUCK			
Notes	Centered around	Frequency range	Length (Transients)
(C2)* G3	450Hz (190Hz - 600Hz)	Med. Narrow 3 Oct	Short (23ms) (transient 1ms)

Origin: Logic Pro X, EFM1 synthesizer.

The Pluck is part of the Musical Test 6.2 as a way of testing spatial resolution due to its uncommon placement at the ‘rear center’ (L_{sr}, R_{sr}). The element is object based.

Note needs to be taken that the Pluck is a very short sound, that only occurs once every four bars, at the “4 &” count.

This element can point out:

1. **Spatial resolution** (3.1.1), as common surround formats don’t have a discrete speaker at the ‘center rear’ position. As a result, a phantom source (2.3.4) will likely be created. However, some immersive formats, such as object-based headphone systems, may overcome this and create a discrete channel, reducing localization confusion.
2. **Directional bands** (2.3.1) due to the sound being built up of a relatively simple combination of narrowband sinusoids. While the element is located at the rear position, a directional bands (2.3.1) of 500Hz may cause it to be perceived at the middle.
3. **Localisation difficulty** due to the short (2.2.2) and sporadic nature of this element, which reduces the ability to localize using head movements as the sound is short and the interval between two sounds is too great to anticipate sequential events, unless when paying active attention.

7.8 Sticks

08 STICKS			
Notes	Centered around	Frequency range	Length (Transients)
A5	3500Hz (& 1000Hz)	Med. Broad 4 Oct	V. Short (5ms) (transient 0,7ms)

Origin: Logic Pro X, Ultrabeat, ‘Adept Machine’ library.

The Sticks element was added to the Musical Test 6.2 and the Localization Test 6.3 as the percussive element with the shortest first transients (0,7ms) and duration (5ms) (2.2.2).

It’s a mono, object-based element positioned at the right rear side, almost completely in the height layer (72%) (between R_{sr} and R_{tr}).

This element can be useful for its:

1. **Ease of localization** due to its broadband spectrum (2.2.1), short transient (2.2.2), emphasis at higher frequencies and it being a relatively familiar (2.2.3) sound.
2. **Elevation:** The Sticks element is positioned 72% towards the height layer and can thus be used to evaluate whether spectral cues (2.2.1) allow the element to be localized at its correct position.

7.9 Shaker



Origin: Logic Pro X, Drummer, 'Heavy Kit' library.

The Shaker was added to the Musical Test 6.2 and the Localization Test 6.3 as the easiest to localize element. It's an object-based element that moves between left and right rear sides, while also moving between the lower and height layer (Lsr, Rsr, Rtr, Ltr) in a continuous movement.

This element can be useful for its:

1. **Ease of localization** due to its prominent transients (2.2.2). Being played continuously, head movements (Section 2.1.3) are able to detect a shift in the spectral cues (2.2.1).
2. **Elevation:** The Shaker moves through the height layer, creating spectral cues (2.2.1).
3. **Directional banding** as a narrowband (2.2.1) sound with focus around 10kHz.

7.10 Live Recorded Piano

10 LIVE RECORDED PIANO			
Notes	Centered around	Frequency range	Length (Transients)
G#2 - G#6	600Hz (250Hz - 1500Hz)	Med. 3,5 Oct	Med (transient 0,2ms)

This Piano element, captured at a recording studio in Mechelen, Belgium, was added to the 'unpredictable' part of the Localization Test 6.3.2.2 as a broadband melodic instrument with a short first transient (2.2.2) on each note.

Unlike the sampled and synthesized elements in the Musical Test 6.2, this element contains the natural room ambience present in the microphones. This can create a sense of depth through monaural cues 2.1.2.

7.11 Cello

11 CELLO			
Notes	Centered around	Frequency range	Length (Transients)
G2 - C4	900Hz (& 150Hz)	Fairly Broad. 4,5 Oct	Long (transient 0,2ms)

The cello was added to the 'unpredictable pattern' part of the Localization Test 6.3.2.2 as a close relative to the Live Recorded Piano element, that differs in transient behavior (2.2.2), as it features a slow attack.

7.12 Voice

12 VOICE			
Notes	Centered around	Frequency range	Length (Transients)
F2 - E3	1500Hz (700z - 2500Hz)	Broad. 8 Oct	~ (transient 5ms)

The last element is a human voice. As this is by far the most familiar (2.2.3) sound to the auditory system. The Voice is used in the Localization Test 6.3 and is used for introductions and additional information in the Musical 6.2, Layout 6.4, and Calibration 0 tests.

In the 'predictable' part of the Localization Test 6.3.2.1, it first 'walks' around in a circle while informing the listener of its current and upcoming position, after which it returns to the center before 'flying' overhead from front to back, informing, again, its position continuously, creating predictability (2.2.3).

In the 'unpredictable' part of the Localization Test 6.3.2.2, it indicates either the referral number of its own location, or that of another element (Shaker, Live Recorded Piano, Cello, Celesta, Pad or Sticks), as the element appears at a new, unpredictable position for each sequential numbered location point.

7.13 Reverb

All instruments, except for the Pluck, have a bit of immersive reverb added to 'glue' to instruments together, as would be the case in most music mixes. The reverb was created using LiquidSonics' Cinematic Rooms plugin. The reverb was sent to the 7.1.2 bed. The level sent to the input of the input varies between $-\infty$ to -14dB, with most instruments sitting between -24 and -18dB.

The added reverb should not be confused with 'room simulation', which would have an impact on the ILD and ITD cues. (2.1.1)

It's not up to the test file to aid in localization by adding realistic early reflections. In this sense, an immersive sound reproduction system that adds their own room emulation at the output can still be tested.

8 Output

The test signals were created in Logic Pro X and mixed in Pro Tools, in a Dolby Atmos environment on a 9.1.6 system.

Aside from being suitable for several channel and object-based speaker systems, these test signals were also created specifically with immersive sound reproduction on headphones in mind.

At this moment, the test signals are available as a Dolby Atmos ADM-file, with a maximal resolution of 9.1.6.

9 Practical use

The complete collection of test signals is ideal for fully evaluation the calibration and localization accuracy of any immersive audio sound reproduction system and certain types of immersive audio technology.

The order in which the tests are labeled was chosen carefully. The Musical Test is the first one, the Calibration Test the last. The Localization and Layout tests sit respectively between them.

When a quick evaluation needs to be performed, solely utilizing the first test (Musical Test) should suffice as the elements it contains are designed to easily point out any flaws in an immersive audio playback system. However, when deeper evaluation is necessary, the other tests can be addressed.

The Localization Test allows for further evaluation and comparison of different immersive sound reproduction systems. It is mainly designed to be combined with any form of a questionnaire, allowing for objective data collection of perceived (subjective) experiences.

When setting up and completely evaluating an immersive audio playback system, or when testing new immersive audio technology, the suggested workflow of using these test signals would be to work from the fourth test towards the first one. This implies that all channels of the immersive audio playback system are first calibrated to be at the expected loudness level, utilizing the Calibration Test. Then, evaluating the channel order through the Layout Test. Next, testing and evaluating the localization accuracy of the system or technology utilizing the Localization Test and, finally, making sure that the system works effectively for music reproduction, utilizing the Musical Test.

10 Future work

The presented test signal is a first version and will likely be improved upon through further research and practical testing.

A few short-term expansions may include additional output formats support, the inclusion of ambisonic elements and adding additional movement parameters for the elements within the test.

11 Conclusion

These immersive tests signals will hopefully allow engineers and developers to have an objective ground on which to base their findings, through the Musical and Localization tests, while also providing some practical aid in configuring immersive sound reproduction systems.

References

- [1] J. Blauert, "Spatial Hearing: The Psychophysics of Human Sound Localization", (MIT Press, Cambridge, MA. 1997), 1997.
- [2] A. Roginska and P. Geluso, "Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio", 2017.
- [3] J. Zwislöcki and R. S. Feldman, "Just Noticeable Differences in Dichotic Phase", 1956.
- [4] Audiologieboek, 2.7.1(2) richtingshoren, URL: <https://audiologieboek.nl/content/2-7-12-richtingshoren..>
- [5] F. Wightman and D. Kistler, "Monaural sound localization revisited", 1997.
- [6] D. Breebaart, "Modeling binaural signal detection", 2001.
- [7] N. Franssen, "Stereofonica", 1962.
- [8] A. Brughera, L. Dunai and W. Hartmann, Human interaural time difference thresholds for sine tones: The high-frequency limit, 2013.
- [9] D. Webb, "The Virtual Barber Shop", URL: <https://www.youtube.com/watch?v=IUDTHIUDTlv>, 2008.
- [10] Motokuni Itoh, Kazuhiro Iida and Masayuki Morimo, "Individual differences in directional bands in median plane localization", 2007.
- [11] P. Pedley and R. Harper, "Pitch and the Vertical Localization of Sound", 1959.
- [12] R. Wallis en H. Lee, "Directional bands revisited", 2015.

- [13] R. Wallis en H. Lee, “The Effect of Interchannel Time Difference on Localization in Vertical Stereophony”, 2015.
- [14] H. Lee, The Relationship between Interchannel Time and Level Differences in Vertical Sound Localisation and Masking, 2011.
- [15] D. Cohen, “Music: Its Language, History, and Culture”, 2015.