Spatial Sound
Principles and Applications
Bosun Xie

Multichannel spatial surround sound

Publication details
Bosun Xie
Published online on: 17 Oct 2022

How to cite: Bosun Xie. 17 Oct 2022, Multichannel spatial surround sound from: Spatial Sound, Principles and Applications CRC Press
Accessed on: 21 Aug 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Multichannel horizontal surround sounds are discussed in Chapters 3 to 5. A real space possesses three dimensionalities. Multichannel three-dimensional spatial surround sounds, shorten for multichannel spatial surround sound or multichannel 3D (surround) sound, should be developed to recreate the three-dimensional spatial information of sound. As an extension of multichannel horizontal surround sounds, multichannel spatial surround sounds are considered the new generation of spatial sound techniques and are addressed in this chapter. In Section 6.1, the summing localization theorems of a horizontal virtual source are extended to a three-dimensional space to provide a basis for the succeeding analyses. The principle of summing localization with two loudspeakers in the median and sagittal planes is analyzed in Section 6.2. The vector base amplitude panning, a typical signal mixing method for spatial surround sound, is examined in Section 6.3. The principle of spatial Ambisonics, another typical spatial surround sound technique and signal mixing method, is discussed in Section 6.4. Some examples of spatial Ambisonic reproduction are also given. Some advanced spatial surround sound systems and related problems are addressed in Section 6.5.

6.1 SUMMING LOCALIZATION IN MULTICHANNEL SPATIAL SURROUND SOUND

6.1.1 Summing localization equations for spatial multiple loudspeaker configurations

Similar to the cases of horizontal surround sound, virtual source localization is an important aspect of spatial surround sound. Summing localization equations for multiple horizontal loudspeakers in Section 3.2.1 should be extended to spatial multiple loudspeaker configurations to analyze the virtual source localization in a spatial surround sound (Xie X.F., 1988; Rao and Xie, 2005; Xie et al., 2019). The foundation of this extension is Wallach’s (1940) hypothesis, discussed in Section 1.6.3.

The coordinate shown in Figure 1.1 is used. At low frequencies, the head shadow is neglected, and the two ears are approximated by two points in a free space and separated by a distance of $2a$ (head diameter). For a point source in the direction of $(r_s, \theta_s, \phi_s)$ and a source distance of $r_s \gg a$, the incident wave can be approximated as a plane wave, and the pressures in the two ears are given by

$$P_L = P_A \exp[-jk(r_s - a \sin \theta_s \cos \phi_s)] \quad P_R = P_A \exp[-jk(r_s + a \sin \theta_s \cos \phi_s)], \quad (6.1.1)$$
where $P_A$ is the amplitude, and $k$ is the wave number. The interaural phase difference in pressures is calculated from Equation (6.1.1) as

$$\Delta \psi = \psi_L - \psi_R = 2ka \sin \theta_s \cos \phi_s,$$  \hspace{1cm} (6.1.2)

or the interaural phase delay difference $ITD_p$ is given by

$$ITD_p(\theta_s, \phi_s) = \frac{\Delta \psi}{2\pi f} = \frac{2a}{c} \sin \theta_s \cos \phi_s.$$ \hspace{1cm} (6.1.3)

When the sound source is located in the cone of confusion in Figure 1.21, $\sin \theta_s \cos \phi_s$ and $ITD_p$ are constant. Therefore, $ITD_p$ only is inadequate for determining the unique position of a sound source.

In Figure 6.1(a), if the head rotates around the vertical ($z$) axis anticlockwise with a small azimuth $\delta \theta$ in the horizontal ($x$–$y$) plane, $ITD_p$ becomes

$$ITD_p(\theta_s - \delta \theta, \phi_s) = \frac{2a}{c} \sin(\theta_s - \delta \theta) \cos \phi_s.$$ \hspace{1cm} (6.1.4)

Letting $\delta \theta \to 0$, the variation rate or derivative of $ITD_p$ with respect to $\delta \theta$ is expressed as

$$\frac{dITD_p(\theta_s, \phi_s)}{d(\delta \theta)} = -\frac{2a}{c} \cos \theta_s \cos \phi_s.$$ \hspace{1cm} (6.1.5)

Equation (6.1.5) indicates that the variation in $ITD_p$ with head rotation is relevant to the elevation $\phi_s$. For a given $\theta_s \neq 90^\circ$, the magnitude of $ITD_p$ variation maximizes in the horizontal plane with $\phi_s = 0^\circ$. As the source elevation departs from the horizontal plane to a high or low elevation, the magnitude of variation decreases. At the top or bottom with $\phi_s = \pm 90^\circ$, $ITD_p$ is invariant against head rotation. Therefore, $ITD_p$ variation caused by head rotation provides information on vertical displacement from the horizontal plane. However, the variation in $ITD_p$ with head rotation alone does not provide enough information on up-down discrimination because of the even function characteristic of $\cos \phi_s$.

As shown in Figure 6.1(b), if the head turns around the front-back ($x$) axis to the left with a small angle $\delta \gamma$ (tilting), the positions of the left and right ears become $(0, a \cos \delta \gamma, -a \sin \delta \gamma)$ and $(0, -a \cos \delta \gamma, a \sin \delta \gamma)$ in Cartesian coordinates, respectively. Similar to the above derivation, $ITD_p$ becomes

$$ITD_p(\theta_s, \phi_s, \delta \gamma) = \frac{2a}{c} (\sin \theta_s \cos \phi_s \cos \delta \gamma - \sin \phi_s \sin \delta \gamma).$$ \hspace{1cm} (6.1.6)
Letting $\delta \gamma \rightarrow 0$, the variation rate or derivative of $ITD_p$ with respect to $\delta \gamma$ is given by

$$
\frac{dITD_p(\theta_s, \phi_s)}{d(\delta \gamma)} = -\frac{2a}{c} \sin\phi_s. 
$$

(6.1.7)

In a horizontal plane with $\phi_s = 0^\circ$, $ITD_p$ is invariant against head tilting. As the source departs from the horizontal plane to a high or low elevation, the magnitude of $ITD_p$ variation with head tilting increases. Head tilting provides supplementary information on up-down discrimination because of the odd function characteristic of $\sin \phi_s$. This analysis is the mathematical expression of Wallach’s hypothesis and was experimentally validated by Perrett and Noble (1997).

In the case of summing localization with multiple loudspeakers, if $M$ loudspeakers are arranged on a spherical surface with a large radius $r_0 > a$, then the incident wave near the origin can be approximated as a plane wave. Let $(\theta_i, \phi_i)$ be the direction of the $i$th loudspeaker and $A_i$ be the normalized amplitude or gain of the corresponding loudspeaker signal. According to Section 3.2.1, for a unit signal waveform in the frequency domain, binaural sound pressures are a linear superposition of the plane wave (far-field) pressures caused by all loudspeakers; for the arbitrary signal waveform $E_A(f)$, the following equation should be multiplied by $E_A(f)$:

$$
P_L = \sum_{i=0}^{M-1} A_i \exp\left[-jk(r_0 - a \sin \theta_i \cos \phi_i)\right] 
$$

$$
P_R' = \sum_{i=0}^{M-1} A_i \exp\left[-jk(r_0 + a \sin \theta_i \cos \phi_i)\right] 
$$

(6.1.8)

Similar to the case of two-channel stereophonic sound in Section 2.1.1, the interaural phase delay difference is evaluated using

$$
ITD_{p, SUM} = \frac{\Delta \Psi_{SUM}}{2\pi f} = \frac{1}{\pi f} \arctan \left[ \frac{\sum_{i=0}^{M-1} A_i \sin(k a \sin \theta_i \cos \phi_i)}{\sum_{i=0}^{M-1} A_i \cos(k a \sin \theta_i \cos \phi_i)} \right]. 
$$

(6.1.9)

$ITD_p$ is the dominant cue for lateral localization at low frequencies; as such, the lateral position of the virtual source for a fixed head is found by comparing Equations (6.1.9) and (6.1.3):

$$
\sin \theta_i \cos \phi_i = \frac{1}{ka} \arctan \left[ \frac{\sum_{i=0}^{M-1} A_i \sin(k a \sin \theta_i \cos \phi_i)}{\sum_{i=0}^{M-1} A_i \cos(k a \sin \theta_i \cos \phi_i)} \right]. 
$$

(6.1.10)
Equation (6.1.10) indicates that the virtual source direction generally depends on $ka$ or frequency. At very low frequencies with $ka \ll 1$, Equation (6.1.10) can be expanded as a Taylor series of $ka$. If the first expansion term is retained, the equation is simplified as

$$\sin \theta_i \cos \phi_i = \frac{\sum_{i=0}^{M-1} A_i \sin \theta_i \cos \phi_i}{\sum_{i=0}^{M-1} A_i}. \quad (6.1.11)$$

In this case, the virtual source direction is independent of $ka$ or frequency.

If the head rotates around the vertical ($z$) axis anticlockwise with a small azimuth $\delta \theta$, the variation rate $ITD_{p, SUM}$ with $\delta \theta$ can also be evaluated. At very low frequencies with $ka \ll 1$, the result is

$$\frac{dITD_{p, SUM}}{d(\delta \theta)} = -\frac{2a}{c} \sum_{i=0}^{M-1} A_i \cos \theta_i \cos \phi_i \left(\sum_{i=0}^{M-1} A_i\right)^{-1}. \quad (6.1.12)$$

If $ITD_p$ variation caused by head rotation provides information on vertical displacement from a horizontal plane, then comparing Equation (6.1.12) with Equation (6.1.5) yields

$$\cos \theta_i' \cos \phi_i' = \frac{\sum_{i=0}^{M-1} A_i \cos \theta_i \cos \phi_i}{\sum_{i=0}^{M-1} A_i}. \quad (6.1.13)$$

Similarly, if the head tilts around the front-back axis with a small angle $\delta \gamma$, the variation rate $ITD_{p, SUM}$ with $\delta \gamma$ can also be evaluated. If head tilting provides supplementary information on up-down discrimination, then comparing the result of multiple loudspeakers with Equation (6.1.7) at very low frequencies with $ka \ll 1$ yields

$$\sin \phi_i' = \frac{\sum_{i=0}^{M-1} A_i \sin \phi_i}{\sum_{i=0}^{M-1} A_i}. \quad (6.1.14)$$

Equations (6.1.11), (6.1.13), and (6.1.14) are a set of summing localization equations for multiple loudspeakers in a three-dimensional space.

In addition, if the head rotates around the vertical axis with an angle $\delta \theta$, $ITD_{p, SUM}$ in Equation (6.1.9) becomes
The perceived virtual source direction at low frequencies of $ka << 1$ is expressed in the following equation by selecting the rotation azimuth $\delta \theta$, so that $ITD_{p, \text{SUM}}$ given in Equation (6.1.15) vanishes:

$$
\tan \hat{\theta}_I = \frac{\sum_{i=0}^{M-1} A_i \sin \theta_i \cos \phi_i}{\sum_{i=0}^{M-1} A_i \cos \theta_i \cos \phi_i}.
$$

Equation (6.1.16) only determines that the virtual source is located in the left-right symmetrical vertical plane with respect to the new orientation of the head and then identifies the azimuth $\hat{\theta}_I$ of the virtual source with respect to the fixed coordinate. The virtual source is not certainly located in the horizontal plane with $\phi_I = 0^\circ$, though it may be located at an arbitrary elevation in the left-right symmetrical vertical plane with respect to the new orientation of the head. If Equations (6.1.11), (6.1.13), and (6.1.14) yield values are nearly consistent with $\theta_I \approx \theta'_I, \phi_I = \phi''_I = \phi''_{\text{ho}}$ and the result of Equation (6.1.13) is nonzero, then Equation (6.1.16) can also be derived by dividing Equation (6.1.11) with Equation (6.1.13).

### 6.1.2 Velocity and energy localization vector analysis for multichannel spatial surround sound

The velocity and energy localization vector analysis of horizontal sound reproduction in Section 3.2.2 can be extended to the cases of spatial sound reproduction (Gerzon, 1992a). The unit vector of an arbitrary three-dimensional source direction $(\theta_S, \phi_S)$ with respect to the origin of a coordinate can be written as $\hat{r} = [\cos \theta_S \cos \phi_S, \sin \theta_S \cos \phi_S, \sin \phi_S]^T$. If $M$ loudspeakers are arranged on a spherical surface, the direction of the $i$th loudspeaker is represented by the unit vector $\hat{r}_i = [\cos \theta_i \cos \phi_i, \sin \theta_i \cos \phi_i, \sin \phi_i]^T$, and the normalized signal amplitude of the $i$th loudspeaker is $A_i$, then the velocity localization vector is evaluated with the following equation as a three-dimensional extension of Equation (3.2.26):

$$
r_v = \sum_{i=0}^{M-1} A_i' \hat{r}_i,

A_i' = \frac{A_i}{\sum_{i'=0}^{M-1} A_i'}.
$$

If the perceived virtual source direction is consistent with the direction of a velocity localization vector (the inner normal direction of the superposed wavefront), then a set of summing localization equations for multiple loudspeakers in a three-dimensional space is derived by decomposing Equation (6.1.17) into three Cartesian components:
In the case of the unit velocity vector magnitude \( r_v = 1 \), Equation (6.1.18) is equivalent to Equations (6.1.13), (6.1.11), and (6.1.14). A correct interaural phase delay difference in reproduction requires that \( r_v = 1 \). The closer \( r_v \) to a unit, the more stable the virtual source against head rotation. Generally, the velocity vector magnitude is evaluated with the following equation:

\[
\begin{align*}
    r_v & = \sqrt{\left( \sum_{i=0}^{M-1} A_i \cos \theta_i \cos \phi_i \right)^2 + \left( \sum_{i=0}^{M-1} A_i \sin \theta_i \cos \phi_i \right)^2 + \left( \sum_{i=0}^{M-1} A_i \sin \phi_i \right)^2} \\
    & \sqrt{\sum_{i=0}^{M-1} A_i} \tag{6.1.19}
\end{align*}
\]

Similarly, the perceived virtual source direction at high frequencies is assumed to be just opposite to the direction of sound intensity. Equation (3.2.24) for horizontal summing localization can be extended to the case of a three-dimensional space:

\[
\begin{align*}
    r_E \cos \theta_E \cos \phi_E & = \frac{\sum_{i=0}^{M-1} A_i^2 \cos \theta_i \cos \phi_i}{\sum_{i=0}^{M-1} A_i^2} \\
    r_E \sin \theta_E \cos \phi_E & = \frac{\sum_{i=0}^{M-1} A_i^2 \sin \theta_i \cos \phi_i}{\sum_{i=0}^{M-1} A_i^2} \\
    r_E \sin \phi_E & = \frac{\sum_{i=0}^{M-1} A_i^2 \sin \phi_i}{\sum_{i=0}^{M-1} A_i^2} \tag{6.1.20}
\end{align*}
\]

The energy velocity magnitude is evaluated as follows:

\[
\begin{align*}
    r_E & = \sqrt{\left( \sum_{i=0}^{M-1} A_i^2 \cos \theta_i \cos \phi_i \right)^2 + \left( \sum_{i=0}^{M-1} A_i^2 \sin \theta_i \cos \phi_i \right)^2 + \left( \sum_{i=0}^{M-1} A_i^2 \sin \phi_i \right)^2} \sqrt{\sum_{i=0}^{M-1} A_i^2} \tag{6.1.21}
\end{align*}
\]
A single source always yields $r_e = 1$. In the design of mid- and high-frequency signal mixing for a spatial surround sound, $r_e$ may be optimized to as close to a unit as possible.

### 6.1.3 Discussion on spatial summing localization equations

The summing localization equations in a three-dimensional space are derived in Sections 6.1.2 and 6.1.3. Similar to the case of horizontal summing localization in Section 3.2.3, these equations are based on different physical and psychoacoustic hypotheses and restrictions. They are related to but different from each other; therefore, they are valid under different conditions.

As indicated in the succeeding sections, experimental results proved that appropriate three-dimensional loudspeaker configurations and signal mixing can recreate a summing virtual source in vertical directions. Previous experiments indicated that the high-frequency spectral cue caused by the diffraction of the pinna is important to front-back and vertical localization. However, the analyses in Section 12.2.2 reveal that a high-frequency spectral cue in multiple loudspeaker reproduction does not match with that of the target source. Therefore, the high-frequency spectral cue does not account for vertical summing localization. As stated in Section 1.6.5, auditory localization is a comprehensive consequence of multiple localization cues. The cooperative effects of multiple localization cues enhance accuracy in localization. However, the information provided by various localization cues may be somewhat redundant. When some cues are unavailable, the auditory system may still be able to localize sound to some extent. As stated in Section 1.6.3, Wallach hypothesized that the variations in $ITD_p$ caused by head rotation and tilting provide information on vertical localization, and this hypothesis has been experimentally validated. Equations (6.1.11), (6.1.13), and (6.1.14) are derived on the basis of Wallach’s hypothesis; therefore, the dynamic $ITD_p$ variation caused by head turning accounts for the vertical summing localization at low frequencies in multiple loudspeaker reproduction. The examples in Sections 6.2 and 6.4 can be regarded as the experimental validation of Equations (6.1.11), (6.1.13), and (6.1.14). Some experiments have demonstrated that subjects can discriminate the sound source in up-and-down directions even for a fixed head (Perrett and Noble, 1997). The scattering and diffraction effects of the torso and shoulders may provide supplementary information on up-down discrimination at low and mid frequencies. The same experiments have also shown that increasing the dynamic cues further improves the up-down discrimination.

Similar to the cases of horizontal surround sound in Section 3.2.3, if Equations (6.1.11), (6.1.13), and (6.1.14) yield consistent results, the perceived virtual source direction can be identified uniquely, and the perceived direction is stable as the head turns. If the three equations reveal inconsistent results, Equation (6.1.11) based on $ITD_p$ can be reasonably used to identify the lateral displacement of the virtual source (cone of confusion). Equation (6.1.13) based on head rotation and Equation (6.1.14) based on head tilting can be comprehensively utilized to identify the virtual source position in the cone of confusion. If the difference in the results of the three equations is obvious, the virtual source is unstable during head turning. This situation should be avoided in a practical signal mixing design. The analyses in Sections 6.2 and 6.4 are based on the logic and consideration presented here.

Equations (6.1.11), (6.1.13), and (6.1.14) are only valid at low frequencies. For wideband stimuli, vertical localization is the comprehensive consequence of dynamic cues at low frequencies, spectral cues caused by the head and pinna at high frequencies, and even spectral cues caused by the torso and shoulders at mid and low frequencies. In practical multichannel spatial sound reproduction, if dynamic and spectral cues provide consistent information, a summing virtual source is accurate and stable. If two cues or a cue in different frequency bands provide inconsistent or even conflicting information, the auditory system may utilize...
dominant or some partly consistent cues to localization, but the virtual source becomes blurry. Moreover, excessive conflicting information may lead to splitting virtual sources in different directions at different frequencies or auditory events with an uncertain position, depending on the power spectra of the stimuli. Strict analysis should comprehensively consider various localization cues and should be based on the mechanisms of auditory information processing by the high-level neural system. Unfortunately, this analysis is currently not feasible. Therefore, the aforementioned localization theorem and equations for spatial surround sound are incomplete. In some instances, they may qualitatively interpret some experimental results. In other instances, they fail to do so. However, the succeeding discussions in this chapter indicate that the results of these localization equations are approximately or qualitatively consistent with the experimental findings in most cases. Therefore, in the absence of other strict means of analysis, the aforementioned localization theorem and equations are still helpful to the practical analysis and design of multichannel spatial surround sound.

Equation (6.1.16) is derived by considering head rotation to an orientation so that ITD_{p, SUM} vanishes. It is related but not completely equivalent to the ITD_p variation caused by head rotation.

The velocity localization vector-based Equation (6.1.18) and energy localization vector-based Equation (6.1.20) are similar to the cases of horizontal reproduction in Section 3.2.3. In the case of \( rv = 1 \), Equation (6.1.18) is equivalent to Equations (6.1.11), (6.1.13), and (6.1.14). In general cases, the velocity localization vector-based direction (Makita hypothesis) is not always the practical perceived direction. The mid- and high-frequency virtual source direction given by Equation (6.1.20) is also a rough approximation. Other features of the aforementioned virtual source localization theorems for spatial surround sound (such as the applicable frequency range) are similar to those for horizontal surround sound in Section 3.2.3 and are omitted here.

6.1.4 Relationship with the horizontal summing localization equations

The relationship between the spatial and horizontal summing localization equations is analyzed in this section. A special summing spatial auditory phenomenon, i.e., the virtual source-elevated effect in horizontal loudspeaker arrangement, is addressed.

If all loudspeakers are arranged in the horizontal plane with \( \phi_i = 0^\circ \), Equations (6.1.11), (6.1.13), and (6.1.14) yield the following results.

In terms of ITD_p for a fixed head orientation,

\[
\sin \theta_i \cos \phi_i' = \frac{\sum_{i=0}^{M-1} A_i \sin \theta_i}{\sum_{i=0}^{M-1} A_i}. \quad (6.1.22)
\]

In terms of the variation rate of ITD_p caused by head rotation around the vertical axis,

\[
\cos \theta_i' \cos \phi_i' = \frac{\sum_{i=0}^{M-1} A_i \cos \theta_i}{\sum_{i=0}^{M-1} A_i}. \quad (6.1.23)
\]
In terms of the variation rate of $ITD_p$, caused by head tilting around the front-back axis,

$$\sin \phi_I^r = 0^\circ. \quad (6.1.24)$$

If the three equations above yield consistent results, i.e., a fixed head orientation, head rotation, and head tilting provide consistent low-frequency localization information, the solutions of the three equations satisfy $\phi_I = \phi_I' = \phi_I'' = 0^\circ$ and $\theta_I = \theta_I'$. In this case, Equation (6.1.22) is identical to summing localization Equation (3.2.7) in the horizontal plane. In addition, dividing Equation (6.1.22) by Equation (6.1.23) yields Equation (3.2.9) and leads to $\theta_I = \theta_I'$. In this case, as stated in Section 3.2.3, combining Equations (3.2.7) and (3.2.9) is appropriate to examine the horizontal summing localization. The analysis of horizontal Ambisonics in Chapter 4 is based on this consideration.

If Equations (6.1.22), (6.1.23), and (6.1.24) yield inconsistent results, i.e., a fixed head orientation, head rotation, and head tilting provide inconsistent low-frequency localization information, an appropriate solution that is approximately consistent with experimental results should be obtained by considering the psychoacoustic principle of auditory localization. Low-frequency $ITD_p$ is a dominant cue for lateral localization; as such, Equation (6.1.22) should be initially used to identify the cone of confusion in which the virtual source is located. However, the two situations or hypotheses may occur, depending on which vertical or elevation localization information among the inconsistent information provided by Equations (6.1.23) and (6.1.24) is dominant.

The first situation or hypothesis considers the result of Equation (6.1.24). According to this equation, the dynamic $ITD_p$ variation caused by head tilting around the front-back axis provides information that the virtual source is located in the horizontal plane. If this information is basically consistent with other vertical localization information (such as spectral information), the auditory system may identify the vertical position based on the more consistent information. Therefore, the perceived virtual source is located in the horizontal plane. In this case, $\phi_I = \phi_I' = \phi_I'' = 0^\circ$, and Equations (6.1.22) and (6.1.23) become

$$\sin \theta_I = \frac{\sum_{i=0}^{M-1} A_i \sin \theta_i}{\sum_{i=0}^{M-1} A_i}, \quad (6.1.25)$$

and

$$\cos \theta_I' = \frac{\sum_{i=0}^{M-1} A_i \cos \theta_i}{\sum_{i=0}^{M-1} A_i}. \quad (6.1.26)$$

If the qualitative results of Equations (6.1.25) and (6.1.26) are consistent, the horizontal azimuth of the virtual source is determined using Equation (6.1.25), which is based on $ITD_p$. The front-back ambiguity is resolved with Equation (6.1.26), which is based on $ITD_p$ variation caused by the head rotation around the vertical axis.
In the second situation or hypothesis, if the comprehensive information of the dynamic \( ITD_p \) variation caused by the head rotation around the vertical axis and other cues (such as spectral cues) is considered, the perceived virtual source is not always located in the horizontal plane. In this case, the result of Equation (6.1.24) should be discarded, and the perceived virtual direction is determined via a combination of Equations (6.1.22) and (6.1.23).

The first hypothesis is used in the analysis of the pair-wise amplitude panning for a horizontal loudspeaker configuration in Section 3.2.3 and in Chapters 4 and 5. Actual horizontal loudspeaker configurations and signal mixing are unable to recreate the exact spectral cue of a horizontal target source at high frequencies, but they may not obviously conflict with that of the horizontal target source. As such, Equation (3.2.9), based on the head oriented to the virtual source, is used to replace Equation (6.1.26) for the analysis of summing localization with a horizontal loudspeaker configuration in Chapters 4 and 5 and to evade the second situation. Consequently, the results basically match with experimental findings in most cases, but they may exhibit some limitations.

In practical horizontal loudspeaker reproduction, the two aforementioned situations may occur depending on the competition and coordination among different localization cues. The final results may depend on various related conditions, such as the head-turning mechanism and stimulus characteristics (such as spectra).

The virtual source-elevated effect in a horizontal loudspeaker arrangement is analyzed to illustrate the second situation. For a two-channel stereophonic loudspeaker arrangement with a large span angle and an identical signal amplitude, the perceived virtual source is located at a high elevation, especially as the head rotates around the vertical axis. A similar phenomenon occurs in reproduction with two rear loudspeakers with a large span angle (such as two surround loudspeakers in a 5.1-channel configuration). This virtual source-elevated effect was observed in early studies on stereophonic sound (Boer, 1947). By the end of the 1950s, Leakey (1959) interpreted this effect via \( ITD \) variation caused by head rotation. Some studies have provided experimental results with additional details (Lee, 2017).

A pair of stereophonic loudspeakers are arranged in a horizontal plane with elevations of \( \phi_L = \phi_R = 0° \) and azimuths of \( \theta_L = \theta_0 \) and \( \theta_R = -\theta_0 \). The amplitudes of two loudspeaker signals are identical to \( A_L = A_R \). For a fixed head, the law of sine in Equation (2.1.6) predicts that the low-frequency virtual source is located at the horizontal front direction of \( \theta_I = 0° \) and \( \phi_I = 0° \).

However, Equations (6.1.22), (6.1.23), and (6.1.24) yield

\[
\sin \theta_I \cos \phi_I = 0, \tag{6.1.27}
\]

\[
\cos \theta_I' \cos \phi_I' = \cos \theta_0, \tag{6.1.28}
\]

and

\[
\sin \phi_I' = 0. \tag{6.1.29}
\]

The three equations indicate the following results:

1. Equation (6.1.27) yields \( \sin \theta_I = 0 \) or \( \cos \phi_I = 0 \), \( \theta_I = 0° \) or \( 180° \), or \( \phi_I = \pm90° \). Therefore, for a fixed head oriented to the front, \( ITD_p = 0 \), and the virtual is located in the median plane.

2. If the virtual source is located at the frontal-median plane, then \( \cos \theta_I' = 1 \), Equation (6.1.28) yields \( \cos \phi_I' = \cos \theta_0 \). Therefore, during the head rotation around the vertical
axis, the dynamic $ITD_p$ variation provides information that the virtual source deviates from the direct front to a high or low elevation in the median plane. This deviation increases as the half-span angle $\theta_0$ of the loudspeaker pair increases. For a pair of horizontal loudspeakers arranged at the two sides with $\pm \theta_0 = \pm 90^\circ$, Equation (6.1.28) yields $\cos \phi' = 0$, and the perceived virtual source is located at the top (or bottom) direction.

3. The result of Equation (6.1.29) is inconsistent with that of Equation (6.1.28), indicating that the head tilting around the front-back axis fails to provide consistent information for up-down discrimination.

The virtual source-elevated effect also occurs in the case of regular horizontal loudspeaker configurations with more loudspeakers (such as four) and identical signal amplitudes. In this case, $ITD_p$ and its dynamic variation with head rotation on a vertical axis vanish, and this observation only matches with a real source in the top or bottom direction. However, Equation (6.1.24) provides an inconsistent or conflicting information on $\sin \phi'$ and fails to yield consistent information on up-down discrimination. This inconsistency reveals the limitation of the analysis in this section. On the other hand, the virtual source-elevated effect is applicable to recreate the perception of “flying over” with a horizontal loudspeaker configuration (Jot et al., 1999). A similar method is also applicable to an irregular loudspeaker configuration in a horizontal plane.

The above example indicates that the experimental results of a virtual source elevated in a horizontal loudspeaker arrangement can be partly and qualitatively interpreted by the dynamic cue caused by head turning. Some quantitative differences may be observed between the theoretical and experimental results, but the above analysis is at least qualitatively consistent with Wallach’s hypothesis, that is, the head rotation around the vertical axis provides information on the vertical displacement of a sound source from the horizontal plane. The virtual source elevated in a horizontal loudspeaker arrangement may be a comprehensive consequence of multiple cues, especially for wideband stimuli. For example, the scattering and diffraction of the torso do not provide information that the virtual source is located below the horizontal plane. The auditory system has adapted to a natural environment with sound sources on the top rather than below directions. However, the effect of multiple localization cues may cause a quantitative difference between the practical perceived directions and those predicted by Equations (6.1.27), (6.1.28), and (6.1.29). Moreover, other studies have interpreted the virtual source-elevated effect with a spectral cue or an interchannel crosstalk (Lee, 2017). Therefore, the analysis in this section is partly based on some hypotheses and thus incomplete; more psychoacoustic validations and corrections are required.

### 6.2 SIGNAL MIXING METHODS FOR A PAIR OF VERTICAL LOUDSPEAKERS IN THE MEDIAN AND SAGITTAL PLANE

The summing localization and other spatial auditory perceptions with a pair of loudspeakers in the horizontal plane are analyzed in Section 1.7.1 and in Chapters 2 to 5. They include recreating a virtual source between loudspeakers with pair-wise amplitude and time panning and recreating spatial auditory perceptions with decorrelated loudspeaker signals. These methods are applicable to two-channel stereophonic and horizontal surround sound reproduction. In multichannel spatial surround sound, a virtual source should also be recreated in the median plane and other sagittal planes with a pair of adjacent loudspeakers in different elevations. The feasibility of this method is analyzed in the following section. A comparison between the analysis and experiment also validates or reveals the limitations of the localization theorem in Section 6.1.
The virtual source localization equations in Section 6.1.1 are used to analyze the summing localization in the median plane at low frequencies (Rao and Xie, 2005; Xie and Rao, 2015; Xie et al., 2019). If all loudspeakers are arranged in the median plane with \( \theta_i = 0^\circ \) or \( 180^\circ \), then \( \sin \theta_i = 0 \) and \( i = 0, 1 \ldots (M - 1) \). Equation (6.1.9) or (6.1.11) yields the interaural phase delay difference \( ITD_{p,\text{SUM}} = 0 \), and the summing virtual source (if it exists) is located in the median plane with its perceived azimuth satisfying \( \theta_i = 0^\circ \) or \( 180^\circ \). Substituting this result into Equation (6.1.13) and letting \( \theta_i = \theta_i' \) yield

\[
\cos \phi_i' = \pm \frac{\sum_{i=0}^{M-1} A_i \cos \theta_i \cos \phi_i}{\sum_{i=0}^{M-1} A_i}. \tag{6.2.1}
\]

When the virtual source is located in the frontal-median plane with \( \theta_i = 0^\circ \) or when the variation rate of \( ITD_{p,\text{SUM}} \) with head rotation calculated from Equation (6.1.12) is negative, the positive sign should be chosen in Equation (6.2.1). Otherwise, the negative sign should be selected in Equation (6.2.1). According to Wallach’s hypothesis, Equation (6.2.1) is used to evaluate the elevation angle of the summing virtual source in the median plane quantitatively, and Equation (6.1.14) is applied to resolve the up-down ambiguity.

Equation (6.2.1) is used to analyze the examples of summing localization with two loudspeakers in the median plane. As shown in Figure 6.2, five loudspeakers numbered 0, 1, 2, 3, and 4 are arranged in the median plane. Loudspeakers 0, 1, and 2 are arranged in the frontal-median plane with an azimuth of \( \theta_i = 0^\circ \) and elevations of \( \phi_0 = -\phi_2, \phi_1 = 0^\circ \), and \( 0^\circ < \phi_2 < 90^\circ \), respectively. Loudspeaker 3 is arranged on the top with an azimuth of \( \theta_3 = 0^\circ \) and an elevation of \( \phi_3 = 90^\circ \). Loudspeaker 4 is arranged in the rear-median plane with an azimuth of \( \theta_4 = 180^\circ \) and an elevation of \( \phi_4 = \phi_2 \). The normalized amplitudes of loudspeaker signals are denoted by \( A_0, A_1, A_2, A_3, \) and \( A_4 \). For pair-wise amplitude panning between loudspeakers 1 and 2, \( A_0, A_3, \) and \( A_4 \) vanish. In this case, Equation (6.1.14) yields \( \sin \phi''_i \geq 0 \), and the virtual source is located in the upper half of the median plane. The variation rate of \( ITD_{p,\text{SUM}} \) with

![Figure 6.2 Five loudspeakers arranged in the median plane.](image-url)
According to Equation (6.2.2), when the interchannel level difference (ICLD) $d_{21} = 20 \log_{10}(A_2/A_1)$ changes from $-\infty$ dB to $+\infty$ dB, the elevation of the virtual source varies from $0^\circ$ to $\phi_2$. In this case, the panning can recreate a virtual source between the two adjacent loudspeakers. Figure 6.3 illustrates the result calculated from Equation (6.2.2) with $\phi_2 = 45^\circ$.

Similar analyses indicate that pair-wise amplitude panning is also able to recreate a virtual source between two adjacent loudspeakers 0 and 1 or 2 and 3. By contrast, for pair-wise amplitude panning between loudspeakers 0 and 2, Equation (6.2.1) yields

$$\cos \phi'_1 = \cos \phi_2 = \cos \phi_0. \quad (6.2.3)$$

Therefore, for a pair of up–down symmetrical loudspeakers, the virtual source is located in the direction of either loudspeaker 0 or 2 in spite of the interchannel level difference $d_{20} = 20 \log_{10}(A_2/A_0)$, i.e., pair-wise amplitude panning fails to recreate a virtual source between two loudspeakers. This result is similar to the case when a pair of lateral loudspeakers with a front-back symmetrical arrangement fails to recreate a stable lateral virtual source (Section 4.1.2).

To analyze the pair-wise amplitude panning between loudspeakers 2 and 4, which have a front-back symmetrical arrangement, selecting a new elevation angle of $-180^\circ < \varphi \leq 180^\circ$ in the median plane is convenient. As shown in Figure 6.4, the new elevation angle is related to the default angle when $\theta = 0^\circ$ and $\varphi = 90^\circ - \phi$ and when $\theta = 180^\circ$ and $\varphi = -90^\circ + \phi$. Therefore, the top, front, and back directions are denoted by $\varphi = 0^\circ$, $90^\circ$, and $-90^\circ$, respectively. As shown in Figure 6.4, the span angle between two loudspeakers is $2\varphi_2 = 2(90^\circ - \phi_2)$, where the elevation angles of the front and back loudspeakers are $\varphi_2$ and $\varphi_4 = -\varphi_2$, respectively. For the normalized loudspeaker signal amplitudes $A_2$ and $A_4$, Equation (6.2.1) yields

$$\cos \phi'_1 = \cos \phi_2 = \cos \phi_0. \quad (6.2.3)$$
The form of Equation (6.2.4) is similar to the law of sine for stereophonic reproduction in the front horizontal plane described in Equation (2.1.6), resulting in a similar variation pattern of a virtual source angle with an interchannel level difference. Therefore, pair-wise amplitude panning can recreate a virtual source between a pair of loudspeakers with a front-back symmetrical arrangement in the median plane. For a half-span angle of $\varphi_2 = 45^\circ$ between two loudspeakers, Figure 6.5 illustrates the variation in the perceived elevation angle $\varphi'_I$ with $\text{ICLD} = d_{24} = 20 \log_{10}(A_2 / A_4)$ (dB). For $d_{24} \rightarrow +\infty$, $d_{24} = 0$ and $d_{24} \rightarrow -\infty$, $\varphi'_I$ are 45°, 0°, and −45°, respectively.

The physical origins of Equation (6.2.4) are different from those of law of sine for stereophonic reproduction. The law of sine for stereophonic reproduction is derived from $\text{ITD}_p$,

$$\sin \varphi'_I = \frac{A_2 - A_4}{A_2 + A_4} \sin \varphi_2 = \frac{A_2 / A_4 - 1}{A_2 / A_4 + 1} \sin \varphi_2. \quad (6.2.4)$$

Figure 6.4 Pair of loudspeakers with a front-back symmetrical arrangement in the median plane.

Figure 6.5 Elevation of a virtual source from the calculation and experiment for pair-wise amplitude panning between a pair of front-back symmetrical loudspeakers 2 and 4. In the figure, solid circle and square represent the results of full-bandwidth and low-pass filtered pink noise, respectively.
which is a dominant cue for azimuthal localization at low frequencies. Equation (6.2.4) is derived from the variation in \( ITD_p \), caused by head rotation, which is supposed to be a cue of elevation localization at low frequencies.

Figures 6.3 and 6.5 also illustrate the results of virtual source localization experiments, including the means and standard deviations from eight subjects. Pink noise with a full-audible bandwidth and a 500 Hz low-pass filtered bandwidth are used as stimuli. The experimental results exhibit a tendency similar to that of the analysis. However, in Figure 6.3, for pink noise with a full-audible bandwidth and ICLD \( d_{21} = -6 \) dB, the elevation angle from the experimental result is closer to the direction of the front loudspeaker 1 than that from the analysis. This inconsistency may be attributed to the following reasons:

1. The accuracy of elevation localization is limited compared with that of azimuthal localization.
2. The approximate (shadowless) head model is used in the analysis.
3. The spectral cue in amplitude panning deviates from that of the target (real) source. The mismatched spectral cue may influence the localizations.
4. The localization information provided by head rotation is inconsistent with that given by tilting for pair-wise amplitude panning.

These reasons should be further examined experimentally.

For a pair of loudspeakers 0 and 2 with an up-down symmetrical arrangement in the median plane, the virtual source jumps from the direction of one loudspeaker to another when the ICLD changes. This experimental result is also consistent with that of theoretical analysis. In addition, the standard deviation of the experimental results in Figures 6.3 and 6.5 is larger than those of usual localization experiments in the horizontal plane. These standard deviations are larger probably because pair-wise amplitude panning with two loudspeakers in the median plane only provides information on dynamic \( ITD_p \) variation for vertical localization. The perceived virtual source becomes blurry (for stimuli with a full-audible bandwidth) because of incorrect spectral information at high frequencies.

Some studies have explored the feasibility of recreating a virtual source between two loudspeakers in the median plane to develop spatial surround sound. The results vary across works. Pulkki (2001a) experimentally demonstrated that pair-wise amplitude panning fails to recreate virtual sources between two loudspeakers at \( \phi = -15^\circ \) and \( 30^\circ \) in the median plane for pink-noise or narrow-band stimuli. This finding is similar to the case of pair-wise amplitude panning between loudspeakers 0 and 2 shown in Figure 6.2, although the loudspeaker configuration in Pulkki’s experiment was not completely symmetrical in the up-and-down direction. An analysis similar to that in Equation (6.2.3) can interpret Pulkki’s experimental results.

Wendt et al. (2014) conducted a virtual source localization experiment with a pair of loudspeakers at an elevation of \( \phi = \pm 20^\circ \) in the median plane. They used pink noise as a stimulus and set ICLD to 0, ±3, and ±6 dB. During the experiment, they instructed their subjects to keep their head immobile to the frontal orientation. The results indicated that the perceived elevation of a virtual source can be controlled by ICLD, but the quantitative results are obviously subject-dependent. Therefore, reproduction fails to create consistent localization cues.

In the development of advanced multichannel spatial sound, a series of virtual source localization experiments is also conducted to investigate the summing localization in the median plane (ITU-R Report, BS 2159-7, 2015). The experimental results from the ITU-R Report indicate that pair-wise amplitude panning for white noise stimuli can recreate a virtual source between two loudspeakers at \( \phi = 0^\circ \) and \( 30^\circ \) or \( 0^\circ \) and \( -30^\circ \) in the median plane; however, it
fails to recreate a virtual source between two loudspeakers at \( \phi = -30^\circ \) and \( 30^\circ \). The results of the aforementioned analysis are also consistent with those in the ITU-R Report.

Lee (2014) investigated the perceived virtual source movement caused by ICLD in the median plane. They arranged two loudspeakers at elevations of \( 0^\circ \) and \( 30^\circ \) in the frontal-median plane and used the signals of a cello and a bongo as stimuli. The results indicated that an ICLD of 6–7 dB is enough to make the virtual source fully move in the direction of either loudspeaker. An ICLD of 9–10 dB is sufficient to completely mask the crosstalk from the loudspeaker with a weaker signal so that the crosstalk is inaudible. The thresholds of the full movement and masking are lower than those of the horizontal stereophonic loudspeaker configuration (Section 1.7.1).

However, Barbour (2003) experimentally revealed that pair-wise amplitude panning for pink noise and speech stimuli fails to recreate a virtual source between two loudspeakers with a front-back symmetrical arrangement in the median plane with \((\theta = 0^\circ, \phi_2 = 60^\circ)\) and \((\theta = 180^\circ, \phi_3 = 60^\circ)\), or \(\phi_{2,4} = \pm30^\circ\) for the new elevation angle in Figure 6.4. It is also unable to recreate a virtual source between two loudspeakers arranged in the median plane with one at \(\phi= 0^\circ\) and another at \(\phi = 45^\circ, 60^\circ, \) or \(90^\circ\). Therefore, Barbour’s results differ from those of the analysis of Equation (6.2.4) and those of the aforementioned experiments possibly because the dynamic cue caused by head-turning might not be fully utilized in Barbour’s experiment. Actually, as stated in Section 1.6, for wideband stimuli, dynamic and spectral cues contribute to vertical localization. However, a pair of loudspeakers in the median plane fails to synthesize the correct or desired spectral cues above the frequency of 5–6 kHz (see the discussion in Section 12.2.2). Therefore, summing localization is impossible if a dynamic cue is not fully utilized. In other words, summing localization in the median plane is caused by the dynamic cue rather than the spectral cue. However, a mismatched spectral cue may degrade the perceived quality of a virtual source, making the virtual source blurry and unstable.

Equation (6.2.4) is based on Wallach’s hypothesis that head rotation provides dynamic information on vertical localization. As stated in Section 1.6.3, since Wallach (1940) proposed the hypotheses on front-back and vertical localization, the contribution of head rotation to front-back discrimination has been verified via numerous experiments. However, few appropriate experiments have been performed to validate the contribution of head-turning to vertical localization because completely and experimentally excluding contributions from other vertical localization cues is difficult. Therefore, the analysis and experiments presented in this section can be regarded as a quantitative validation of Wallach’s hypothesis on vertical localization (Rao and Xie, 2005; Xie et al., 2019), e.g., head rotation provides information on source vertical displacement from the horizontal plane, and head tilting yields additional information on up-down discrimination. The analysis in this section aims to validate Wallach’s hypothesis on vertical localization. The analysis and experiment in this section are also a validation of the summing localization equations derived in Section 6.1.1.

In summary, the equations derived in Section 6.1.1 are applicable to the analysis of the summing localization in the median plane. The analysis and experiments indicate that a pair-wise amplitude panning for appropriate two-loudspeaker configurations in the median plane can recreate a virtual source between loudspeakers, but the position of a virtual source may be blurry. However, for some other up-down symmetrical loudspeaker configuration, a pair-wise amplitude panning fails to recreate a virtual source between loudspeakers. These results are applicable to designs of loudspeaker configurations in the median plane. Notably, the results are appropriate for stimuli with dominant low-frequency components. For stimuli with dominant high-frequency components, the results are different.

This analysis can be extended to the vertical summing localization in sagittal planes (cone of confusion) other than the median plane (Xie et al., 2017b). For example, by using pair-wise amplitude panning, a pair of loudspeakers with an up-down symmetrical arrangement
at \((\theta, \phi)\) and \((\theta, -\phi)\) fails to recreate a virtual source between them. By contrast, a pair of loudspeakers arranged asymmetrically at \((\theta, \phi)\) and \((\theta, 0^\circ)\) can recreate a virtual source between them.

The vertical summing localization by interchannel time difference (ICTD) or pair-wise time panning has also been explored in some studies. Lee (2014) conducted an experiment with cello and bongo stimuli and indicated that pair-wise time panning fails to recreate a virtual source between two vertical loudspeakers, especially in the median plane. Tregonning and Martin (2015) experimentally demonstrated that ICTD between the signals of horizontal and elevated loudspeakers changes the perceived elevation, but it never makes the virtual source fully move in the direction of either loudspeaker. For speech and conga stimuli, an ICTD of 5 ms leads to a maximal vertical movement of the virtual source. As ICTD further increases, the width and vertical spread of the virtual source increase.

Litovsky et al. (1999) experimentally investigated the precedence effect of two sound sources (loudspeakers) in the median plane and obtained results that vary from other studies. They indicated that the delay for localization dominance in the median plane is similar to that in the horizontal plane, but the precedence effect is weaker in the median plane than in the horizontal plane. Conversely, Wallis and Lee (2014) observed that the precedence effect does not operate in the median plane. In addition, decorrelated signals are less effective for creating an auditory event with a vertical spread than for creating an auditory event with a horizontal extension (Gribben and Lee, 2014).

Overall, cues for horizontal and vertical auditory perceptions differ; as such, rules for summing spatial auditory events and perceptions in horizontal and vertical directions also differ. Moreover, different studies may yield various results, which may depend on the characteristics of stimuli. Hence, further studies are needed.

6.3 VECTOR BASE AMPLITUDE PANNING

The reproduced sound field and perceived effect of a multichannel sound depend on loudspeaker configuration and signal mixing. Pair-wise amplitude panning is a common signal mixing method and is applicable to a pair of adjacent loudspeakers in the horizontal, median, and sagittal planes to recreate a virtual source between loudspeakers. Other signal mixing methods are used to recreate a virtual source in a three-dimensional space. Pulkki (1997) proposed a vector base amplitude panning (VBAP) method. It is another typical and general signal mixing method that can be regarded as an extension of pair-wise amplitude panning in the case of three-dimensional loudspeaker configurations.

VBAP can be conveniently analyzed in terms of the velocity localization vector discussed in Section 6.1.2. The origin of coordinates is located at a spherical center and coincident with the head center. A spherical surface is divided by spherical-triangular grids. \(M\) loudspeakers are arranged in the vertices of the grid. The direction \(\Omega_i = (\theta_i, \phi_i)\) of the \(i\)th loudspeaker is represented by a unit vector \(\mathbf{r}_i = [\cos \theta_i \cos \phi_i, \sin \theta_i \cos \phi_i, \sin \phi_i]^T\) pointing from the origin to the loudspeaker. The target source direction \(\Omega_S = (\theta_S, \phi_S)\) is represented by a unit vector \(\mathbf{r}_S = [\cos \theta_S \cos \phi_S, \sin \theta_S \cos \phi_S, \sin \phi_S]^T\). The virtual source within a spherical triangle is recreated by the loudspeakers forming the triangle. Therefore, a virtual source at different directions may be recreated by loudspeakers in different spherical triangles. This feature is similar to that of pair-wise amplitude panning in a horizontal plane.

Figure 6.6 illustrates the principle of VBAP. The unit vectors of three active loudspeakers are \(\mathbf{r}_1, \mathbf{r}_2,\) and \(\mathbf{r}_3\). The amplitudes \(A_1, A_2,\) and \(A_3\) of three loudspeaker signals are real and non-negative values (in phase). According to the discussion in Sections 3.2.2 and 6.1.2 and Equation (6.1.17), the linear combination of \(\mathbf{r}_1, \mathbf{r}_2,\) and \(\mathbf{r}_3\) with weights of \(A_1, A_2,\) and \(A_3\) is
the velocity localization vector \( \mathbf{r}_v \) in reproduction, which should be matched with the target virtual source direction (Makita’s hypothesis), then

\[
\mathbf{r}_s = A_1 \mathbf{r}_1 + A_2 \mathbf{r}_2 + A_3 \mathbf{r}_3.
\]  

(6.3.1)

Equation (6.3.1) can be written in a matrix form as follows:

\[
\mathbf{r}_s = \begin{bmatrix} \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \end{bmatrix} A,
\]  

(6.3.2)

where

\[
\mathbf{r}_s = \begin{bmatrix} \cos \theta_s \cos \phi_s, \sin \theta_s \cos \phi_s, \sin \phi_s \end{bmatrix}^T \quad A = \begin{bmatrix} A_1, A_2, A_3 \end{bmatrix}^T
\]  

(6.3.3)

\[
\begin{bmatrix} \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \end{bmatrix} = \begin{bmatrix} \cos \theta_1 \cos \phi_1 & \cos \theta_2 \cos \phi_2 & \cos \theta_3 \cos \phi_3 \\ \\ \sin \theta_1 \cos \phi_1 & \sin \theta_2 \cos \phi_2 & \sin \theta_3 \cos \phi_3 \\ \\ \sin \phi_1 & \sin \phi_2 & \sin \phi_3 \end{bmatrix}.
\]

The loudspeaker signal amplitudes can be solved from Equation (6.3.2) as

\[
A = \begin{bmatrix} \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \end{bmatrix}^{-1} \mathbf{r}_s.
\]  

(6.3.4)

The directions of the three loudspeakers are unparallel; as such, \( \mathbf{r}_1 \), \( \mathbf{r}_2 \), and \( \mathbf{r}_3 \) are not linearly correlated, and the inverse matrix in Equation (6.3.4) certainly exists. As expected, Equation (6.3.4) proves that a virtual source is recreated only by the corresponding loudspeakers if a target source is located in the same direction as any of the loudspeakers. If the target source is located in an arc between two adjacent loudspeakers, the virtual source is
recreated by the two loudspeakers, and the signal for the other loudspeaker vanishes. If the target source is located between two adjacent loudspeakers in a horizontal plane, VBAP is simplified into horizontal pair-wise amplitude panning.

The resultant signal amplitudes $A_1$, $A_2$, and $A_3$ should be normalized further. The normalized amplitudes of actual loudspeaker signals are $A_1$, $A_2$, and $A_3$ multiplying $A_{\text{total}}$. For constant-power normalization, the following equation is obtained:

$$A_{\text{total}} = \frac{1}{\sqrt{A_1^2 + A_2^2 + A_3^2}},$$

(6.3.5)

The aforementioned discussion is the basic principle of VBAP. VBAP satisfies the low-frequency optimized criterion of the direction of the velocity localization vector. However, it does not satisfy the optimized criterion of the unit velocity vector magnitude $r_v = 1$ except the target virtual source located in the loudspeaker direction. If the head rotates to an orientation so that $\text{ITD}_p$ vanishes, the lateral displacement of the perceived virtual source with respect to the fixed coordinate is identical to that of the target source. However, for a fixed head with a frontal orientation, the lateral displacement of the perceived virtual source may differ from that of the target source because of the mismatched $\text{ITD}_p$ in reproduction. Pulkki (2001a) also validated this result by using a binaural auditory model. In addition, the vertical displacement of the perceived virtual source may differ from that of the target source because of the mismatched variation in $\text{ITD}_p$ caused by head rotation. The accuracy of the perceived virtual source direction in reproduction depends on the positions of three active loudspeakers on a spherical surface and can be analyzed using the localization equations in Section 6.1.1. The situation here is similar to the cases of pair-wise amplitude panning in a horizontal plane (Section 4.1.2) or in a median plane (Section 6.2).

Four active loudspeakers with at least one out-of-phase loudspeaker signal are needed to further satisfy the optimized criterion of $r_v = 1$. The function of an out-of-phase loudspeaker signal is similar to the cases of horizontal reproduction described in Sections 3.2.2 and 4.1.3. Out-of-phase loudspeaker signals are used in global or local Ambisonic signal mixing for horizontal and spatial reproduction, as discussed in Sections 4.3, 5.2.3, 5.2.4, and Section 6.4. The analysis here also reveals the difference between VBAP and Ambisonic signal mixing. VBAP employs three active loudspeakers to recreate a virtual source within a spherical triangle formed by loudspeakers. Therefore, the error of the perceived virtual source direction is within the region bounded by active loudspeakers despite the position of a listener. The crosstalk from loudspeakers in opposite directions and a serious localization error in off-central listening position are averted. As the total number of loudspeakers increases, the active loudspeaker triangle grid becomes dense, and the localization error is further reduced. VBAP is simple and easily implemented in program production. Therefore, it possesses some advantages in practical uses. Some experiments on VBAP with various loudspeaker configurations have validated the aforementioned analysis (Pulkki, 2001a; Wendt et al., 2014).

### 6.4 Spatial Ambisonic Signal Mixing and Reproduction

#### 6.4.1 Principle of Spatial Ambisonics

Spatial Ambisonics is another typical multichannel spatial sound and signal mixing method (Gerzon, 1973). It was developed in the 1970s and originally termed Periphony. Currently, it is a hot topic in spatial sound. As an extension of horizontal Ambisonics, spatial Ambisonics involves the decomposition and reconstruction of a sound field by a series
of directional harmonics (spherical harmonic functions). Similar to horizontal Ambisonics, spatial Ambisonics can be analyzed with various mathematical and physical methods. A traditional analysis based on a virtual source localization theorem is addressed in this section, and a stricter analysis based on sound field reconstruction is presented in Chapter 9.

The first-order spatial Ambisonics involves four independent signals. If the maximal amplitude of independent signals is normalized to a unit, a set of normalized independent signals (amplitudes) can be chosen as

\[
W = 1, \quad X = \cos \theta \cos \phi, \quad Y = \sin \theta \cos \phi, \quad Z = \sin \phi,
\]

where \((\theta, \phi)\) represent the azimuth and elevation of the target source in the original sound field. Equation (6.4.1) represents the signals captured by four coincident microphones in the original sound field. \(W\) is the signal from an omnidirectional microphone (spherical symmetry); \(X, Y,\) and \(Z\) are signals from three bidirectional microphones (symmetrical directivity around the polar axis) with their main axes pointing to the front, left, and top directions, respectively. The directional patterns of these four microphones are illustrated in Figure A.1 in Appendix A. \(W, X, Y,\) and \(Z\) also represent the pressure and the \(x\) (front-back), \(y\) (left-right), and \(z\) (up-down) components of the velocity of the medium in the original sound field, respectively.

In reproduction, \(M\) loudspeakers are arranged on a spherical surface around a listener. The distance from each loudspeaker to the origin (center of the sphere) satisfies the condition of a far-field distance. The direction of the \(i\)th loudspeaker is \((\theta_i, \phi_i)\) with \(i = 0, 1, 2\ldots M – 1,\) and loudspeaker signals can be written as a form of sound field signal mixing, i.e., as a linear combination of the independent signals \(W, X, Y,\) and \(Z\) given in Equation (6.4.1):

\[
A_i(\theta_s, \phi_s) = A_{\text{total}} \left[ D_{00}^{(i)}(\theta_s, \phi_s) W + D_{11}^{(i)}(\theta_s, \phi_s) X + D_{12}^{(i)}(\theta_s, \phi_s) Y + D_{10}^{(i)}(\theta_s, \phi_s) Z \right] \\
= A_{\text{total}} \left[ D_{00}^{(i)}(\theta_s, \phi_s) + D_{11}^{(i)}(\theta_s, \phi_s) \cos \theta_s \cos \phi_s + D_{12}^{(i)}(\theta_s, \phi_s) \sin \theta_s \cos \phi_s \right] + D_{10}^{(i)}(\theta_s, \phi_s) \sin \phi_s \right]
\]

where the decoding coefficients \(D_{00}^{(i)}(\theta, \phi), D_{11}^{(i)}(\theta, \phi), D_{12}^{(i)}(\theta, \phi),\) and \(D_{10}^{(i)}(\theta, \phi)\) can be chosen by using various methods depending on the loudspeaker configuration and optimized criteria for a reproduced sound field.

Generally, for left-right symmetrical loudspeaker configurations, decoding coefficients satisfy symmetrical relationships similar to those in Equations (5.2.16)–(5.2.18). If loudspeaker configurations are also up-down symmetric, for any pair of up-down symmetrical loudspeakers \(i\) and \(i',\) their azimuths and elevations satisfy \(\theta_i = \theta_{i'}\) and \(\phi_i = -\phi_{i'},\) and

\[
D_{00}^{(i)}(\theta, \phi) = D_{00}^{(i)}(\theta, \phi), \quad D_{11}^{(i)}(\theta, \phi) = D_{11}^{(i)}(\theta, \phi), \quad D_{12}^{(i)}(\theta, \phi) = D_{12}^{(i)}(\theta, \phi), \quad D_{10}^{(i)}(\theta, \phi) = -D_{10}^{(i)}(\theta, \phi).
\]

Symmetry greatly simplifies the decoding coefficients.

The decoding coefficients in Equation (6.4.2) are solved in accordance with various optimized criteria. Similar to the case of horizontal Ambisonics with an irregular loudspeaker configuration in Section 5.2.3, the optimized criterion of the perceived virtual source direction...
matching that of a target source can be used when the head is fixed and when the head is turning at low frequencies; as such, \( \theta_i = \theta_i' = \theta_i \) and \( \phi_i = \phi_i' = \phi_i \) in Equations (6.1.11), (6.1.13), and (6.1.14). With these parameters, a set of linear equations for decoding coefficients can be obtained. Furthermore, \( \theta_i = \theta_i, \phi_i = \phi_i \), and \( r_i = 1 \) can be set in Equation (6.1.18) by using the optimized criterion of the velocity localization vector; the normalized pressure amplitude in the origin is set to a unit, so the following equation is obtained:

\[
\sum_{i=0}^{M-1} A_i(\theta_i, \phi_i) \cos \theta_i \cos \phi_i = \cos \theta \cos \phi
\]

\[
\sum_{i=0}^{M-1} A_i(\theta_i, \phi_i) \sin \theta \cos \phi_i = \sin \theta \cos \phi
\]

Substituting Equation (6.4.2) into Equation (6.4.4) leads to a set of linear equations or a matrix equation for decoding coefficients:

\[
S'_{3D} = A_{total} [Y'_{3D}] [D'_{3D}] S'_{3D},
\]

where \( S'_{3D} = [W, X, Y, Z]^T \) is a 4 × 1 column matrix or vector composed of independent signals in Equation (6.4.1), the subscript “3D” denotes the case of three-dimensional spatial Ambisonics; and \( [D'_{3D}] \) is an \( M \times 4 \) decoding matrix to be solved:

\[
[D'_{3D}] =
\begin{bmatrix}
D^{(1)}_{00}(\theta, \phi) & D^{(1)}_{11}(\theta, \phi) & D^{(2)}_{11}(\theta, \phi) & D^{(1)}_{10}(\theta, \phi) \\
D^{(1)}_{00}(\theta, \phi) & D^{(1)}_{11}(\theta, \phi) & D^{(2)}_{11}(\theta, \phi) & D^{(1)}_{10}(\theta, \phi) \\
\vdots & \vdots & \vdots & \vdots \\
D^{(1)}_{00}(\theta_{M-1}, \phi_{M-1}) & D^{(1)}_{11}(\theta_{M-1}, \phi_{M-1}) & D^{(2)}_{11}(\theta_{M-1}, \phi_{M-1}) & D^{(1)}_{10}(\theta_{M-1}, \phi_{M-1})
\end{bmatrix}.
\]

\( [Y'_{3D}] \) is a 4 × \( M \) matrix with its entries related to the directions of loudspeakers:

\[
[Y'_{3D}] =
\begin{bmatrix}
1 & 1 & \ldots & 1 \\
\cos \theta_0 \cos \phi_0 & \cos \theta_1 \cos \phi_1 & \ldots & \cos \theta_{M-1} \cos \phi_{M-1} \\
\sin \theta_0 \cos \phi_0 & \sin \theta_1 \cos \phi_1 & \ldots & \sin \theta_{M-1} \cos \phi_{M-1} \\
\sin \phi_0 & \sin \phi_1 & \ldots & \sin \phi_{M-1}
\end{bmatrix}.
\]

Similar to the case of Equation (5.2.24), for \( M \gg 4 \), the decoding coefficients in Equation (6.4.5) can be solved from the following pseudo-inverse methods if the loudspeaker configuration is chosen appropriately so that the matrix \( [Y'_{3D}] \) is well conditioned (Section 9.4.1):

\[
A_{total} [D'_{3D}] = \text{pinv} [Y'_{3D}] = [Y'_{3D}]^T \left\{ [Y'_{3D}] [Y'_{3D}]^T \right\}^{-1}.
\]

The solution given in Equation (6.4.8) satisfies the condition of constant-amplitude normalization. For \( M = 4 \), the decoding coefficients can be directly solved from the inverse of
the matrix \([Y'_{3D}]\). Except for some regular loudspeaker configurations, the solution shown in Equation (6.4.8) usually does not satisfy the criterion of constant power in reproduction:

\[
P_{w'} = \sum_{i=0}^{M-1} A_i^2 \neq \text{const.} \tag{6.4.9}
\]

For the solution expressed in Equation (6.4.8), the direction of the energy localization vector evaluated using Equation (6.1.20) generally does not coincide with that of the velocity localization vector, i.e., \((\theta_{E}, \phi_{E}) \neq (\theta_{v}, \phi_{v})\). However, as indicated by Equation (6.4.4), the first-order spatial Ambisonics with a decoding matrix given by Equation (6.4.8) is an example in which Equations (6.1.11), (6.1.13), and (6.1.14) are consistent. In other words, the first-order spatial Ambisonics creates ITD, and its dynamic variation with head rotation and tilting that match with those of the target source at very low frequencies.

Similar to the case of horizontal Ambisonics (Section 4.3.1), spatial Ambisonics has various forms of independent signals. In principle, four independent linear combinations of \(W\), \(X\), \(Y\), and \(Z\) can serve as a set of independent signals of spatial Ambisonics. Different sets of independent signals can be converted mutually. The decoding equations for different sets of independent signals are related via linear matrix transformation. In practice, four independent signals can be captured by four cardioid or subcardioid microphones with symmetrical directivity around the polar axis. The main axes of four microphones are pointed to the four non-parallel directions: left-front-up (LFU), left-back-down (LBD), right-front-down (RFD), and right-back-up (RBU) (Figure 6.7). The normalized amplitudes of microphone signals are

\[
A_i' (\Delta \Omega_i') = A_{mic} \left[ 1 + b \cos (\Delta \Omega_i') \right] \quad i' = 0, 1, 2, 3. \tag{6.4.10}
\]

where \(0.5 \leq b \leq 1\), \(\Delta \Omega_i'\) is the angle between the target source direction and the main axis of the \(i\)th microphone, and \(A_{mic}\) is a coefficient associated with the acoustic-electric efficiency or gain of microphones. The independent signals expressed in Equation (6.4.10) are termed A-format spatial Ambisonic signals.

Figure 6.7 Main axis directions of the four microphones for recording the A-format first-order spatial Ambisonic signals.
W, X, Y, and Z given in Equation (6.4.1) can be conveniently chosen as the independent signals of the first-order spatial Ambisonics. W, X, Y, and Z can be derived from the signals presented in Equation (6.4.10) through linear transformation. Let

\[ A'_{LFU} = A'_0 (\Delta \Omega'_0) \quad A'_{LBD} = A'_1 (\Delta \Omega'_1) \quad A'_{RFD} = A'_2 (\Delta \Omega'_2) \quad A'_{RBU} = A'_3 (\Delta \Omega'_3). \]  

(6.4.11)

The linear combinations of the four signals in Equation (6.4.1) yield

\[
\begin{align*}
W^* &= A'_{LFU} + A'_{LBD} + A'_{RFD} + A'_{RBU} = 4A_{mic} \\
X^* &= A'_{LFU} - A'_{LBD} + A'_{RFD} - A'_{RBU} = \frac{4b}{\sqrt{3}} A_{mic} \cos \theta_S \cos \phi_N \\
Y^* &= A'_{LFU} + A'_{LBD} - A'_{RFD} - A'_{RBU} = \frac{4b}{\sqrt{3}} A_{mic} \sin \theta_S \cos \phi_N \\
Z^* &= A'_{LFU} - A'_{LBD} - A'_{RFD} + A'_{RBU} = \frac{4b}{\sqrt{3}} A_{mic} \sin \phi_N.
\end{align*}
\]  

(6.4.12)

The normalized independent signals W, X, Y, and Z are obtained by equalizing the signals in Equation (6.4.12) with an appropriate gain. In practice, the levels of X, Y, and Z are increased by 3 dB to ensure a diffused-field power response identical to that of the M signal. The independent signals are written as

\[
\begin{align*}
W' &= W = 1 \quad X' = \sqrt{2} X = \sqrt{2} \cos \theta_S \cos \phi_S \quad Y' = \sqrt{2} Y = \sqrt{2} \sin \theta_S \cos \phi_S \\
Z' &= \sqrt{2} Z = \sqrt{2} \sin \phi_S.
\end{align*}
\]  

(6.4.13)

The decoding equation should be appropriately changed to accommodate the independent signals in Equation (6.4.13). The spatial Ambisonic system with independent signals given by Equation (6.4.13) is termed the first-order B-format spatial Ambisonics (Gerzon, 1985).

The second- and higher-order directional harmonics (appendix A) are supplemented as new independent signals to improve the performance in reproduction, and the first-order spatial Ambisonics can be extended to the higher-order spatial Ambisonics. Under the optimized criterion of matching each order approximation of reproduced sound pressures with those of the target source within a given listening region and below a certain frequency limit, the decoding coefficients or matrix and loudspeaker signals are solved via a pseudo-inverse method similar to Equation (6.4.8). This problem is addressed in Section 9.3.2. For the second- and higher-order spatial Ambisonics with decoding coefficients given in Equation (6.4.4), localization Equations (6.1.11), (6.1.13), and (6.1.14) yield consistent results. Equation (6.4.4) only specifies the zero and first-order decoding coefficients, flexibly leaving room for choosing the second- and higher-decoding coefficients.

Moreover, the equivalence between A-format and B-format spatial Ambisonics allows a post-equalization of recorded source signal at a given direction (Favrot and Faller, 2020). By using an appropriate matrix, the recorded B-format signals can be converted into a set of A-format signals with the main axis of polar pattern of one of A-format signal \( A'_0(\Delta \Omega'_0) \) points to the concerned source direction. The resultant signal \( A'_0(\Delta \Omega'_0) \) is equalized and, and the equalized signal along with the other un-equalized A-format signals are converted back into B-format signals by an invert matrix.
Similar to the case of horizontal Ambisonics, the decoding coefficients at high frequencies can be derived according to various optimized physical and psychoacoustic criteria (Arteaga, 2013), such as optimized direction of the energy localization vector in Equation (6.1.20), maximized energy vector magnitude in Equation (6.1.21), and constant overall power; thus,

$$\theta_E = \theta_e, \quad \phi_E = \phi_e, \quad \max(r_E) \quad \text{Pow}' = \text{const.} \quad (6.4.14)$$

However, if the energy localization vector is considered, the optimized criteria yield a set of nonlinear equations of decoding coefficients, and nonlinear optimization methods are required. For higher-order spatial Ambisonics with irregular loudspeaker configurations, nonlinear optimization is complicated and can be usually solved with numerical methods (Scaini and Arteaga, 2014). Moreover, it may satisfy various optimized criteria partly or approximately, rather than completely and exactly. For example, Zotter et al. (2012) provided a constant power solution and Epain et al. (2014) provided a constant angular spread solution for irregular loudspeaker configuration. As the order of spatial Ambisonics increases, the upper-frequency limit of the exact reconstruction of a sound field increases. In this case, optimizing high-frequency decoding may be unnecessary. All these aspects are similar to the horizontal Ambisonics with an irregular loudspeaker configuration in Section 5.2.3, so they are no longer described here. For reproduction within a large-sized listening region, the in-phase solution of decoding coefficients can be derived in a manner similar to the horizontal case in Section 4.3.2 (Malham and Myatt, 1995; Daniel, 2000; Neukom, 2007).

The features and limitations of spatial Ambisonics are similar to those of horizontal Ambisonics (discussed in Section 4.3.4, so they are no longer presented in this section).

Virtual source localization experiments for spatial Ambisonics with different orders and loudspeaker configurations have been conducted (Capra et al., 2007; Morrell and Reiss, 2009; Power et al., 2012; Xie et al., 2017a, 2019). Overall, the results indicate that spatial Ambisonics can recreate a virtual source at three-dimensional directions (including vertical directions). The localization performance improves as the order increases. The third-order spatial Ambisonics usually exhibits a satisfactory localization performance. In Section 12.2.2, the upper frequency limit for Ambisonics to reconstruct target binaural pressures accurately increases as the order increases. The third-order spatial Ambisonics can reconstruct target binaural pressures approximately up to 1.87 kHz and can reconstruct ITDp and its dynamic variation with head turning below 1.5 kHz. In other words, the third-order spatial Ambisonics can recreate correct localization information below 1.5 kHz. Therefore, the results of localization experiments on spatial Ambisonics can be regarded as experimental validations of the localization hypothesis and theorem presented in Section 6.1.

### 6.4.2 Some examples of the first-order spatial Ambisonics

The first-order spatial Ambisonics can be implemented via various loudspeaker configurations. Some examples are given here. These examples are used to illustrate the principle of first-order spatial Ambisonics. They can theoretically optimize the localization performance at low frequencies (velocity localization vector) but may fail to optimize the overall perceived performance. As a consequence, they may not completely satisfy the practical requirements.

A classic example is reproduction with a loudspeaker configuration, as shown in Figure 6.8 (Xie X.F., 1988; Gerzon, 1992a). Eight loudspeakers are arranged in the vertices of a hexahedron (cuboid), and the listener’s head is located at the center of the hexahedron with an equal distance to all loudspeakers. This loudspeaker configuration is left-right and up-down symmetric. The position of the ith loudspeaker is denoted by $(\theta_i, \phi_i)$ with $i = 0, 1, \ldots, 7$. From Equation (6.4.8), the normalized loudspeaker signals (amplitudes) for low-frequency optimization are given by
where $\gamma_{x,i}$, $\gamma_{y,i}$, and $\gamma_{z,i}$ are the angles between the directional vector of the $i$th loudspeaker (with respect to the center of the cuboid or origin) and the $x$, $y$, and $z$ axes of the coordinate; $\cos \gamma_{x,i}$, $\cos \gamma_{y,i}$, and $\cos \gamma_{z,i}$ are the direction cosines of the $i$th loudspeakers. This hexahedral configuration is accommodated for most practical room shapes in reproduction.

In a special case of a regular hexahedral configuration, the positions of eight loudspeakers are

$\begin{align*}
\text{LFU} & : \theta_{\text{LFU}} = 45^\circ \quad \phi_{\text{LFU}} = 35.3^\circ \\
\text{LBU} & : \theta_{\text{LBU}} = 135^\circ \quad \phi_{\text{LBU}} = 35.3^\circ \\
\text{RFU} & : \theta_{\text{RFU}} = -45^\circ \quad \phi_{\text{RFU}} = 35.3^\circ \\
\text{RBU} & : \theta_{\text{RBU}} = -135^\circ \quad \phi_{\text{RBU}} = 35.3^\circ
\end{align*}$

where the notations L, R, F, B, U, and D in the subscripts denote left, right, front, back, up, and down, respectively. Substituting Equation (6.4.16) into Equation (6.4.15) yields the (matrix) decoding equation for loudspeaker signals:

$$A_i(\theta_j, \phi_j) = \frac{1}{8} \left[ W + \frac{1}{\cos \gamma_{x,i}} X + \frac{1}{\cos \gamma_{y,i}} Y + \frac{1}{\cos \gamma_{z,i}} Z \right]$$

$$= \frac{1}{8} \left[ W + \frac{1}{\cos \theta_j \cos \phi_j} X + \frac{1}{\sin \theta_j \cos \phi_j} Y + \frac{1}{\sin \phi_j} Z \right] \quad i = 0, 1...7,$$

(6.4.16)
Equation (6.4.17) is equivalent to the normalized signals captured by eight coincident hypercardioid microphones with a symmetrical directivity around the polar axis, and the main axes of eight microphones are pointed to the directions of eight vertices of the regular hexahedron:

\[
A_i \left( \Delta \Omega_i^r \right) = \frac{1}{8} \left[ 1 + 3 \cos \left( \Delta \Omega_i^r \right) \right], \tag{6.4.18}
\]

where \( \Delta \Omega_i^r \) is the angle between the target source direction and the main axis of the \( i \)th microphone (which coincides with the direction of the \( i \)th loudspeaker). For a loudspeaker configuration in Equation (6.4.16) and the signals in Equation (6.4.17), the velocity localization vector in reproduction satisfies the optimized criteria of \( \theta_v = \theta_S, \phi_v = \phi_S, \) and \( r_v = 1. \) At the same time, the overall power of loudspeaker signals is a constant. The energy localization vector satisfies the optimized criteria of \( \theta_E = \theta_S \) and \( \phi_E = \phi_S, \) but the energy vector magnitude is \( r_E = 0.5. \) The regular hexahedral configuration is one of a few cases in which the direction of a virtual source satisfies the optimized criteria of velocity localization vector and energy localization.

Another example is the first-order spatial Ambisonics with a regular tetrahedral loudspeaker configuration, as shown in Figure 6.9 (Xie X.F., 1988). The listener’s head is located at the center of the regular hexahedron with an equal distance to all loudspeakers. Four loudspeakers are arranged in the four vertices of the regular tetrahedron, i.e., in left-front-down (LFD), right-front-down (RFD), back-down (BD), and up (U) directions with respect to the listener. This loudspeaker configuration is the simplest among configurations for first-order spatial Ambisonics because it involves four independent signals and requires at least four loudspeakers in reproduction. The positions of the eight loudspeakers are:

\[
\begin{align*}
LFD: & \quad \theta_{LFD} = 60^\circ \quad \phi_{LFD} = -19.47^\circ \quad \bcancel{RFD:} \quad \theta_{RFD} = -60^\circ \quad \phi_{RFD} = -19.47^\circ \\
BD: & \quad \theta_{BD} = 180^\circ \quad \phi_{BD} = -19.47^\circ \quad \bcancel{U:} \quad \theta_{U} = 0^\circ \quad \phi_{U} = 90^\circ 
\end{align*}
\tag{6.4.19}
\]

Substituting Equation (6.4.19) into Equation (6.4.8) yields the (matrix) decoding equation for loudspeaker signals:

\[
\begin{bmatrix}
A_{LFD} \\
A_{RFD} \\
A_{BD} \\
A_{U}
\end{bmatrix}
= \begin{bmatrix}
1 & 1.414 & 2.449 & -1 \\
1 & 1.414 & -2.449 & -1 \\
4 & 1 & -2.828 & 0 \\
1 & 0 & 0 & 3
\end{bmatrix}
\begin{bmatrix}
W \\
X \\
Y \\
Z
\end{bmatrix}.
\tag{6.4.20}
\]

For this loudspeaker configuration and signals, the overall power of loudspeaker signals is a constant, but the direction of a virtual source does not satisfy the optimized criterion of the energy localization vector.

Similar to the case of horizontal Ambisonics, the first-order spatial Ambisonic reproduction has more loudspeaker configurations. The decoding equation and loudspeaker signals for these configurations can also be derived. For example, for the conical configuration with \( M \) loudspeakers shown in Figure 6.10 (a), \( (M - 1) \) loudspeakers are arranged in the circle of the bottom of the cone, and a loudspeaker is arranged in the vertex of the cone (up direction). The listener’s head center is located at the center of the cone with an equal distance to
all loudspeakers. The folding loudspeaker configuration with four loudspeakers shown in Figure 6.10 (b) is also theoretically available for the first-order spatial Ambisonic reproduction (Xie and Xie, 1992, 1992b). However, the details are not presented here because of page limitations.

### 6.4.3 Local Ambisonic-like signal mixing for vertical loudspeaker configuration

The spatial Ambisonic signal mixing described in the preceding sections is a global signal mixing method through which signals are fed to all spatial loudspeakers to recreate a virtual source. Similar to the case of horizontal reproduction, local Ambisonic-like signal mixing – by which signals are fed to loudspeakers arranged in the median or sagittal plane – is available to recreate a vertical virtual source in a given plane (Yi and Xie, 2020).

In the first-order local Ambisonic-like signal mixing, three loudspeakers, i.e., loudspeakers 0, 1, and 2 in Figure 6.2, are used to recreate a virtual source in the frontal median plane. Similar to the derivation of Equation (5.2.32) for three frontal horizontal loudspeakers, the
normalized signal amplitudes for three vertical loudspeakers are a linear combination of elevation harmonics up to the first order and given by

\[
A_0 = A_{total} \left( 1 - \cos \phi_S - \frac{1 - \cos \phi_2}{\sin \phi_2} \sin \phi_S \right)
\]

\[
A_1 = A_{total} \left( -2 \cos \phi_2 + 2 \cos \phi_S \right)
\]

\[
A_2 = A_{total} \left( 1 - \cos \phi_S + \frac{1 - \cos \phi_2}{\sin \phi_2} \sin \phi_S \right),
\]

where \( \phi_S \) is the target source elevation in the median plane, and \( \phi_2 \) is the elevation of loudspeaker 2. For constant-power normalization, the normalized factor \( A_{total} \) is expressed as

\[
A_{total} = \frac{1}{\sqrt{2(1 - \cos \phi_S)^2 + 2 \left( \frac{1 - \cos \phi_2}{\sin \phi_2} \sin \phi_S \right)^2 + 4 \left( \cos \phi_S - \cos \phi_2 \right)^2}}.
\]

Figure 6.11 presents the local Ambisonic-like panning curves of Equations (6.4.21) and (6.4.22) with \( \phi_2 = 45^\circ \). For target source elevation in each loudspeaker direction (\( \phi_S = 0^\circ \) or \( \pm 45^\circ \)), the signals for two other loudspeakers vanish, and a virtual source is recreated by a single loudspeaker. For target source elevation between two loudspeakers, the virtual source is recreated by three loudspeakers, and the signal for the third loudspeaker is out of phase. The panning curves in Figure 6.11 are extended to create a target source outside the boundary of a three-loudspeaker array.

The three localization equations can be proved to yield consistent results by substituting Equation (6.4.21) into Equations (6.1.11), (6.1.13), and (6.1.14):

\[
\sin \theta_1 \cos \phi_1 = 0 \quad \cos \theta'_1 \cos \phi'_1 = \cos \phi_S \quad \sin \phi'_1 = \sin \phi_S
\]

or

\[
\theta_1 = \theta'_1 = 0^\circ \quad \phi_1 = \phi'_1 = \phi'_2 = \phi_S.
\]
In this case, $ITD_p$ and its dynamic variation with head rotation and tilting in reproduction are consistent with those of the target source, and the perceived virtual source direction matches that of the target source at low frequencies. This feature is superior to that of vertical pair-wise amplitude panning discussed in Section 6.2. Further analysis using head-related transfer functions (HRTFs) (Section 12.1.3) yields similar results.

Another feature of local Ambisonic-like signal mixing given by Equation (6.4.21) is that it can theoretically recreate a vertical virtual source at the elevation outside the boundary of a three-loudspeaker array. A virtual source localization experiment for a three-loudspeaker configuration with elevations of $\phi_0 = -45^\circ$, $\phi_1 = 0^\circ$, and $\phi_2 = 45^\circ$ indicates that the local Ambisonic-like signal mixing can recreate a stable virtual source within the elevation region of $\pm 60^\circ$. The local Ambisonic-like signal mixing can also be extended to vertical loudspeaker configurations in other sagittal planes.

### 6.4.4 Recreating a top virtual source with a horizontal loudspeaker arrangement and Ambisonic signal mixing

The first-order spatial Ambisonics encodes spatial information into four independent signals, namely, $W$, $X$, $Y$, and $Z$, and requires at least four loudspeakers arranged in space (such as a regular tetrahedral configuration) in reproduction. The first-order horizontal Ambisonics excludes $Z$ and loses the vertical localization information. However, if the three independent signals for horizontal Ambisonics are revised as per (Jot et al., 1999), we have:

$$W_1 = \sqrt{1 + \sin^2 \phi_s} \quad X_1 = \cos \theta_s \cos \phi_s \quad Y_1 = \sin \theta_s \cos \phi_s.$$  \tag{6.4.25}

Then, a perceived virtual source can be recreated in the top direction by using horizontal loudspeaker configurations. The nature of this method is the virtual source-elevated effect discussed in Section 6.1.4.

For a regular configuration with $M \geq 3$ loudspeakers in the horizontal plane, the independent signals $W$, $X$, and $Y$ of the first-order horizontal Ambisonics are replaced with $W_1$, $X_1$, and $Y_1$ expressed in Equation (6.4.25). For a target source in a horizontal plane with $\phi_s = 0^\circ$, the independent signals in Equation (6.4.25) are identical to those of the first-order horizontal Ambisonics in Equation (4.3.3), thereby causing an identical horizontal localization effect. For a target source in the top direction $\phi_s = 90^\circ$, $X_1 = Y_1 = 0$ and the amplitudes of all loudspeaker signals are equal:

$$A_i(\phi_s = 90^\circ) = \sqrt{2}A_{total} \quad i = 0, 1, \ldots (M - 1).$$  \tag{6.4.26}

According to the analysis in Section 6.1.4, the perceived virtual source is located in the top direction.

### 6.5 ADVANCED MULTICHANNEL SPATIAL SURROUND SOUNDS AND PROBLEMS

#### 6.5.1 Some advanced multichannel spatial surround sound techniques and systems

For practical uses, especially for ultra-high-definition video/audio and new cinema sound requirements, multichannel spatial surround sounds have emerged as new-generation sound reproduction techniques after multichannel horizontal surround sounds with and without
accompanying pictures. Since 2000, multichannel spatial surround sound has been con-
ered an important trend in the audio field (Rumsey, 2013).

Multichannel spatial surround sound aims to improve the capability of reproducing the
three-dimensional spatial information of sounds in all or some of the following aspects
depending on the applications and contents of programs:

1. Improve the sound effects that match with pictures, i.e., recreate frontal virtual sources
   in left-right and vertical dimensions in the entire screen area because an ultra-high-
definition video with a large screen provides larger horizontal and vertical view angles.
2. Enhance the three-dimensional virtual source effect, including lateral, rear, and top vir-
tual sources, and improve the resolution in the reproduction of the spatial information
   of sound.
3. Increase the reproduction of the three-dimensional spatial information of discrete
   early reflections and late diffused reverberation to produce a good auditory spatial
   impression.
4. Improve the reproduction of non-reflective ambience components to obtain a good
   immersive sense.
5. Enlarge the listening region.
6. Ensure compatibility with existing and common techniques and systems.

Similar to the case of horizontal surround sound, the main considerations and core problems
in a multichannel spatial surround sound are the appropriate selection of the number of chan-
nels, loudspeaker configurations, and signal mixing based on the psychoacoustic principle
and available system resources to achieve the desired spatial auditory perceptions. However,
for multichannel spatial surround sounds, more choices are available in the number of chan-
nels and loudspeaker configurations. For practical uses and commercial competition, various
multichannel spatial surround sound techniques with different loudspeaker configurations
have been developed since 2000. These techniques are also termed immersive sound or audio.
Some techniques and systems were originally developed for cinema sound reproduction with
a large-sized listening region. Subsequently, they were simplified for domestic or consumer
uses. Other techniques and systems were directly established for domestic reproduction with
a small-sized listening region. For domestic uses, these techniques involve an irregular loud-
speaker configuration in a three-dimensional space and can be applied to reproduction with
and without accompanying picture.

In 2002, Dabringhaus published the first music program by using the 2 + 2 + 2 channel
 technique (www.mde.de/frame2.htm). This technique is based on the ITU-5.1-channel con-
figuration, but the center loudspeaker and subwoofer are omitted. Instead, a pair of loud-
speakers arranged above the L and R loudspeakers are supplemented (Ehret et al., 2007).

As stated in Section 5.3, some other loudspeaker configurations in DTS-HD are for
7.1-channel spatial surround sound (DTS Inc., 2006), which can partly recreate vertical
sound information.

1. In configuration 2, a pair of loudspeakers on the two sides of $\theta = \pm 90^\circ$ above the
   horizontal plane are supplemented to the standard 5.1-channel configuration. This con-
figuration improves the reproduction of vertical information at the lateral-up and top
directions; it is also compatible with 5.1-channel reproduction. However, it does not
improve the reproduction of lateral and rear information in the horizontal plane.
2. In configuration 3, a loudspeaker in the horizontal rear direction and another loud-
speaker in the top direction are supplemented to the standard 5.1-channel configura-
tion to improve the reproduction of rear and top information.
3. In configuration 4, a pair of loudspeakers is supplemented to the positions above the left and right loudspeakers in a standard 5.1-channel configuration (Figure 6.12, where the subwoofer is omitted).

4. In configuration 7, a loudspeaker is supplemented to the position above the center loudspeaker, and another loudspeaker is supplemented to the rear in standard 5.1-channel configuration.

Configurations 4 and 7 are beneficial to recreating a vertical localization effect that matches with a large-screen video.

Layer-wise loudspeaker configurations are often used in multichannel spatial surround sound reproduction. For instance, Auro Technologies introduced a series of multichannel spatial surround sound systems and loudspeaker configurations (termed Auro-3D). These loudspeaker configurations are constructed with supplemented upper-layer loudspeakers into a 5.1 or 7.1-channel horizontal configuration and are compatible with horizontal reproduction. The Auro 9.1 channel system, which was introduced in 2016, is intended for domestic reproduction and is composed of a two-layer loudspeaker configuration (Figure 6.13, where the subwoofer is omitted; Theile and Wittek, 2011). The middle or horizontal layer at an elevation of $\phi = 0^\circ$ involves five loudspeakers identical to that of an ITU-5.1-channel configuration, i.e., located at azimuths of $\theta = 0^\circ$, $\pm 30^\circ$, and $\pm 110^\circ$ respectively. The upper layer at an elevation of $\phi = 30^\circ$ involves four loudspeakers located above the horizontal left, right, left-surround, and right-surround loudspeakers, i.e., located at azimuths $\pm 30^\circ$ and $\pm 110^\circ$, respectively. This loudspeaker configuration may fail to recreate a vertical virtual source in the frontal-median plane.

Figure 6.12 Loudspeaker configuration 4 in a DTS-HD 7.1-channel system.

Figure 6.13 Two-layer loudspeaker configuration in an Auro 9.1 channel system.
plane, but it can recreate vertical localization information in the directions of upper-layer loudspeakers. On the basis of the Auro 9.1 channel system, the Auro 10.1-channel system is constructed by adding a top channel; the Auro 11.1-channel system is developed by adding a top channel and an upper-center channel in the upper layer; and the Auro 13.1-channel system is created by adding a top channel, an upper-center channel in the upper layer, and left-back and right-back surround channels in the middle layer (i.e., 7.1-channels are present in the middle layer). The Auro 11.1 and 13.1-channel systems were originally developed for cinema sound, but they are applied to domestic reproduction after simplification.

The 10.2-channel sound, which was proposed by Holman in cooperation with the Integrated Media System Center at the University of California, was developed for domestic and cinema reproduction; it is shortened for USC 10.2-channel system (Holman, 1996, 2001; ITU-R Report, BS 2159-7, 2015). For domestic reproduction, it consists of a two-layer loudspeaker configuration (Figure 6.14). The middle layer at an elevation of $\phi = 0^\circ$ involves eight loudspeakers at azimuths of $\theta = 0^\circ$, $\pm 30^\circ$, $\pm 60^\circ$, $\pm 110^\circ$, and $180^\circ$. In comparison with the ITU-5.1-channel sound, the USC 10.2-channel sound can be used to extend the virtual source distribution in the frontal horizontal plane and improve the reproduction of early lateral reflections by adding a pair of left/right-wide loudspeakers at larger azimuths of $\theta = \pm 60^\circ$. Adding a back (rear) surround loudspeaker at $\theta = 180^\circ$ enhances the reproduction of a virtual source and reflection from the rear. The upper layer at an elevation of $\phi = 45^\circ$ consists of a pair of left/right-height loudspeakers at an azimuth of $\phi = 45^\circ$ to improve the reproduction of a virtual source and reflection from vertical directions. In practice, two types of loudspeakers with different radiation patterns are arranged at an azimuth of $\pm 110^\circ$ in the middle layer. One type is a loudspeaker with a traditional direct radiation pattern for recreating localization effects. Another type is a loudspeaker with a dipolar radiation pattern, which is elevated above the direct radiation loudspeaker. Combined with the reflections from the lateral and rear walls of a listening room, dipolar-type loudspeakers enhance ambience reproduction. Therefore, the USC 10.2-channel system has 12 loudspeakers with a full-audible bandwidth. In addition, two low-frequency effect channels are available for decorrelated low-frequency contents. The system employs bass management for all full-audible channels (Section 14.3.3). All low-frequency components below 120 Hz are reproduced by a pair of subwoofers arranged on two sides to enhance the auditory spatial impression at low frequencies.

Another 10.2-channel sound was introduced by Samsung Electronics Co., Ltd. in the Republic of Korea for ultra-high-definition digital television (Kim et al., 2010; ITU-R Report, BS 2159-7, 2015). It is shortened for the Samsung 10.2-channel system. The two-layer configuration of the main loudspeakers is illustrated in Figure 6.15. The loudspeaker configuration

![Figure 6.14](image) Two-layer loudspeaker configuration for a USC 10.2-channel sound.
in the middle layer at an elevation of $\phi = 0^\circ$ is similar to that of the optional 7.1-channel configuration in the ITU standard (ITU-R BS 775-1, 1994), which involves seven loudspeakers located at azimuths of $\theta = 0^\circ$, $\pm 30^\circ$, $\pm 90^\circ$, and $\pm 135^\circ$, respectively. The four horizontal surround loudspeakers can also be arranged within the azimuthal region of $\pm 60^\circ$ to $\pm 150^\circ$. The upper (top) layer at $\phi = 45^\circ$ (or $30^\circ$ to $45^\circ$) involves three loudspeakers arranged in the Left-front-up (LFU), right-front-up (RFU), and back-up (BU) directions with azimuths $\theta = \pm 45^\circ$ (or $\pm 30^\circ$ to $\pm 45^\circ$) and $180^\circ$, respectively. The back-up loudspeaker can also be arranged within the range of $\theta = 180^\circ$ and $\phi = 45^\circ$ to $90^\circ$. Two low-frequency effect channels are also present, and two subwoofers are arranged below the middle layer. The loudspeaker configuration in the Samsung 10.2-channel system improves the localization and reflection reproduction in rear and vertical directions.

The 22.2-channel sound by NHK (Japan Broadcasting Corporation) was developed for sound reproduction of ultra-high-definition video system called Super High Vision (Hamasaki et al., 2004, 2007; ITU-R Report, BS 2159-7, 2015). The Super High Vision comprises 100° horizontal–frontal viewing angle and about 4,000 scanning lines ($7680 \times 4320$-pixel video image), the definition is 16 times that of HDTV and twice that of a 70 mm film. The 22.2-channel sound improves the ability to recreate three-dimensional spatial information, including localization in a three-dimensional space, envelopment, and ambience in consumer reproduction. It consists of a three-layer loudspeaker configuration (Figure 6.16). Some international

![Figure 6.15 Two-layer loudspeaker configuration for Samsung 10.2-channel sound.](image)

![Figure 6.16 Three-layer loudspeaker configuration for the 22.2-channel sound.](image)
standards specify the channel labeling and loudspeaker positions in the 22.2-channel system. However, channel labeling may vary among these standards. The labeling specified by ITU is illustrated in Figure 6.16 [ITU-R, BS 2051-2 (2018)], which is identical to those of the Society of Motion Picture and Television Engineers (SMPTE ST 2036-2, 2008) but different from those of ISO/IEC (ISO/IEC 23001-8, 2015). In addition, the loudspeaker positions of the ITU and ISO/IEC standards slightly differ, but both standards allow variations in loudspeaker positions within a certain region. Loudspeaker positions may also be slightly different in other studies and even in studies from the same group in different periods. The loudspeaker positions in the ITU standard are specified as follows:

1. Middle (or ear) layer
   Five frontal loudspeakers at an elevation of \( \phi = 0^\circ \text{–} 5^\circ \) and azimuths of \( \theta = 0^\circ \), \( \pm22.5^\circ \text{–} \pm30^\circ \), and \( \pm45^\circ \text{–} \pm60^\circ \)
   Five lateral and back (rear) surround loudspeakers at an elevation \( \phi = 0^\circ \text{–} 15^\circ \) and azimuths of \( \theta = \pm90^\circ \), \( \pm110^\circ \text{–} \pm135^\circ \), and 180°

2. The top (or upper) layer
   Eight loudspeakers in elevation \( \phi = 30^\circ \text{–} 45^\circ \) and azimuths \( \theta = 0^\circ \), \( \pm45^\circ \text{–} \pm60^\circ \), \( \pm90^\circ \), \( \pm110^\circ \text{–} \pm135^\circ \), and 180°
   A loudspeaker directly in the top direction with an elevation of \( \phi = 90^\circ \)

3. Bottom layer
   Three loudspeakers at an elevation of \( \phi = -15^\circ \text{–} -30^\circ \) and azimuths of \( \theta = 0^\circ \) and \( \pm45^\circ \text{–} \pm60^\circ \)

4. Subwoofers
   Two subwoofers at a low elevation of \( \phi = -15^\circ \text{–} -30^\circ \) and an azimuth of \( \theta = \pm30^\circ \text{–} \pm90^\circ \)

Here, all loudspeakers are assumed to be arranged on a spherical surface with an equal distance to the center. If loudspeakers are arranged at different distances to the center, appropriate delay and magnitude correction for loudspeaker signals are needed to compensate for the unequal loudspeaker distances.

### 6.5.2 Object-based spatial sound

The preceding discussions focus on channel-based spatial sounds. The spatial information of sound is represented by channel (loudspeaker) signals. In a channel-based method, the configuration of loudspeakers in reproduction is predefined, and loudspeakers signals are prepared at the stage of program production according to the desired perceived effects and the physical or psychoacoustic rules of signal mixing. The resultant signals are inappropriate for reproduction with different loudspeaker configurations unless some signal conversion procedures are supplemented. As the number and options of loudspeakers increase, especially for multichannel spatial surround sound, this problem becomes serious. Ambisonics is a scene-based spatial sound. The spatial information of sound is encoded into a set of independent signals. Independent signals, which are independent of the number and configuration of loudspeakers, are decoded into loudspeaker signals in reproduction according to the practical number and configuration of loudspeakers. In comparison with channel-based methods, the scene-based method is relatively flexible.

Since 2000, object-based spatial sound has been considered a remarkable development. In an object-based spatial sound, audio contents with identical spatial and other properties are grouped into an audio object to be transmitted. A set of metadata, which describe the temporary spatial and other properties of audio objects (such as the temporary position of a target source), are transmitted with audio objects. The metadata can be regarded as the
parameters and side information of audio objects or data about data. Audio objects and metadata are independent from the number and configuration of loudspeakers or even the manners or techniques in reproduction. During reproduction, audio objects are distributed to loudspeakers according to the information of metadata, practical loudspeaker configuration, and certain signal mixing rules. Object-based methods can recreate virtual sources and other spatial auditory perceptions (such as envelopment similar to that caused by diffused reverberation). In an object-based method, the signal mixing or signal rendering is moved from the stage of program production to the stage of reproduction. This method is flexible and able to cater to different numbers and configurations of loudspeakers. By using different signal rendering methods, object-based methods can also be used for spatial sound reproductions with different principles and techniques, such as wave field synthesis (Chapter 10) and binaural technique (Chapter 11).

Various object-based spatial sound techniques or combinations of channel- and object-based techniques have been developed. Examples include the ISO/IEC MPEG-H 3D Audio standard (ISO/IEC 23008-3, 2015; Herre et al., 2014, 2015), the Dolby Atmos by Dolby Laboratories (2012, 2015, 2016), the AuroMax by Auro Technologies and Bacro Audio Technologies (2015), and the Multi-Dimensional Audio (MDA) by DTS Inc. (ETSI TS 103 223, V1.1.1, 2015). The basic considerations of these examples are similar.

Dolby Atmos was first introduced by Dolby Laboratories in 2012 as a spatial sound technique for cinema and then modified for domestic (consumers) use. Currently, it is widely used in many domestic theaters and cinemas. In addition to audio objects, Dolby Atmos supports the transmission of parts of audio contents called "beds" with traditional channel-based methods. In other words, the Dolby Atmos involves two types of audio contents, namely, audio objects and beds. Audio objects are appropriate for directional localization contents, and beds are designed for ambient contents. In reproduction, loudspeaker signals are rendered by an appropriate mixing of audio objects and beds according to practical loudspeaker configurations and by using a special processor. Dolby Atmos for cinema uses supports up to 118 audio objects, 5.1, 7.1, or even 9.1-channel beds and reproduction of up to 64 loudspeakers. The 5.1 and 7.1-channel beds are similar to previous Dolby 5.1- and 7.1-channel systems. The 9.1-channel beds involve three frontal channels, four horizontal surround channels on the sides and rear, and two (left and right) top channels. Figure 6.17 illustrates the block diagram of the Dolby Atmos for cinema uses (the B-chain processing in the figure is discussed in Section 14.7).

In 2014, Dolby Laboratories introduced the Dolby Atmos for domestic use, which is similar to but is modified or simplified from the Dolby Atmos for cinema use (Dolby Laboratories, 2016). For domestic use, the number and configuration of loudspeakers are flexible, and Dolby Atmos supports the reproduction of up to 24 horizontal loudspeakers and 10 top loudspeakers. Typical configurations are constructed by adding four top (overhead) loudspeakers to the horizontal 5.1- or 7.1-channel configurations, leading to the 9.1- or 11.1-channel configuration (Figure 6.18). The horizontal left and right loudspeakers are arranged at azimuths

![Figure 6.17 Block diagram of Dolby Atmos for cinema uses. (adapted from Dolby Laboratories, 2012.)](image-url)
of ±22°–±30°, and the four horizontal surround loudspeakers are arranged at azimuths of ±90°–±110° and ±135°–±150°, respectively. A pair of top loudspeakers are arranged in front of the listening position, and another pair of top loudspeakers are arranged behind the listening position. The subwoofer is arranged on the floor between the horizontal center and left loudspeakers. If two top loudspeakers are used in reproduction, they are arranged slightly in front of the listening position. This loudspeaker configuration favors the reproduction of beds in Dolby Atmos. However, the details of the signal rendering in Dolby Atmos have yet to be published until 2021.

The MPEG-H 3D Audio is a standard for new-generation spatial sound and signal coding techniques by the International Organization for Standardization (ISO) and the Moving Pictures Expert Group (MPEG) under the International Electrotechnical Commission (IEC). It supports many existing spatial sound techniques and formats, including object- and channel-based methods, Ambisonics, VBAP, binaural reproduction, and audio coding techniques. The details are addressed in Section 13.5.6.

The principle and structures of other object-based techniques and methods are similar to MPEG-H 3D Audio. The Advanced Television Systems Committee (ATSC) in the USA specified the standard of ATSC Audio 3.0, an immersive audio standard for television (ATSC Standard Doc A/342-1, 2017). The ATSC Audio 3.0 involves an 11.1-channel loudspeaker configuration similar to that in Figure 6.18 and supports channel-based, object-based, and Ambisonic reproduction. AuroMax was developed for cinema reproduction, but it can be modified for domestic uses. AuroMax involves 11.1- or 13.1-channel beds, so it is compatible with Auro 11.1- or 13.1-channel sound, as described in Section 6.5.1. It also has more audio objects. AuroTechnologies recommended examples of 20.1-, 22.1-, and 26.1-channel reproduction. Another feature of AuroMax is that it supports the wave field synthesis technique to recreate virtual sources at distances beyond the boundaries of a loudspeaker array (Chapter 10). The Multi-Dimensional Audio (MDA) group, which consists of five corporations, including DTS, also developed an object-based spatial sound technique. The principle is similar to the above. The rendering stage also involves VBAP and Ambisonic techniques. Channel-based beds are modeled as a special MDA group of objects.

6.5.3 Some problems related to multichannel spatial surround sound

Commercial spatial sound is developed from horizontal to three-dimensional reproduction and from channel-based methods to object-based methods. A combination of channel- and object-based spatial sound is a new-generation technique. Since 2010, the Audio Engineering
Multichannel spatial surround sound

Society (AES) has organized many conferences and sessions on this special issue (Rumsey, 2013; and The AES 40th Conference; The AES 138th Convention). The perceived performance of multichannel spatial surround sound is greatly prior to traditional multichannel horizontal surround sound but at the cost of increasing the complexity. Multichannel spatial surround sounds are being developed quickly, and various techniques have been established, but they may cause some confusion.

To standardize multichannel spatial surround sound, some international organizations and societies have formulated some standards and specifications, which are being constantly revised and updated. Early in 2006, the Committee on Digital Cinema Technology under SMPTE established the standard of distribution master audio channel mapping and channel labeling for 16-channel digital cinema sound (SMPTE 428-3, 2006). The SMPTE ST 2036-2 (2008) specified the audio channel mapping and labeling for program production in ultra-high-definition television. The SMPTE ST 2098-5 (2018) further defines names and abbreviations for immersive audio channels and immersive sound field groups associated with digital cinema immersive audio presentation. This standard also provides informative guidance on typical locations of cinema loudspeakers used for immersive audio reproduction.

Among the works and standards of ITU, ITU-R BS 1909 (2012) specified the performance requirements for an advanced multichannel stereophonic (and spatial surround) sound system for use with or without an accompanying picture, including a virtual source in all directions around a listener, the sensation of three-dimensional spatial impression, a stable virtual source in the entire area of high-resolution large-screen digital imagery, and excellent sound quality over a wide region. The ITU-R Report BS 2159-7 (2015) compared multichannel sounds for domestic and broadcasting uses and gave the results of the subjective assessment of various loudspeaker configurations. The ITU-R BS 2051-2 (2018) recommended possible loudspeaker configurations, channel labeling, and metadata requirements for multichannel spatial surround sound. Among the standards of ISO/IEC, in addition to the MPEG-H 3D Audio mentioned in Section 6.5.2 (ISO/IEC 23008-3, 2015), the IEC 62574 (2011) and ISO/IEC 23001-8 (2015) specified various loudspeaker configurations, channel labels, and program reproduction levels for multichannel spatial surround sound. Notably, loudspeaker positioning and channel labeling in ISO/IEC standards are different from those in ITU standards.

Similar to the case of 5.1-channel sound, various standards of multichannel spatial surround sound for domestic reproduction only recommend the loudspeaker configurations, but these standards do not restrict signal mixing for loudspeakers. Various signal mixing methods, such as traditional pair-wise amplitude panning, VBAP, Ambisonics, and some spatial microphone techniques (Section 7.3), are applicable to multichannel spatial surround sound with different loudspeaker configurations. Theoretically, the virtual source localization theorems in Section 6.1 can be applied to optimize loudspeaker configuration and signal mixing, but the results are usually valid in the central listening position rather than the off-central listening position even for reproduction within a small-sized region. However, these analyses are beneficial to the design of multichannel spatial surround sound for a small-sized listening region. They also help reveal and validate the limitation of systems and techniques. Object-based methods allow more flexible and diverse loudspeaker configurations and signal mixing, which should also be optimized in accordance with the psychoacoustic principles of spatial hearing.

As an example of flexible signal mixing, a signal mixing for loudspeaker configuration similar to the DTS 7.1-channel configuration 2 described in Section 6.5.1 is outlined (Rao and Xie, 2004). In this configuration, a pair of left-high and right-high loudspeakers are arranged at \( \theta = \pm 90^\circ \) and \( \phi = 30^\circ - 45^\circ \). A pair of horizontal surround loudspeakers is moved...
backward to $\theta = \pm 135^\circ$. Only the case of a virtual source in the left-half space is outlined because of the left-right symmetry.

1. In the left-front space ($0^\circ \leq \theta \leq 30^\circ$, $0^\circ \leq \phi \leq 90^\circ$), a virtual source is recreated by four loudspeakers, i.e., the left-front, center, left-high, and right-high loudspeakers. For a target source in the median plane with $\theta = 0^\circ$, the signal for the left-front loudspeaker vanishes, and a virtual source is recreated by the center, left-high, and right-high loudspeakers.

2. In the front-lateral space ($30^\circ \leq \theta \leq 90^\circ$, $0^\circ \leq \phi \leq 90^\circ$), a virtual source is recreated by three loudspeakers, i.e., the left-front, left-high, and right-high loudspeakers. In the lateral plane, the signals of left-front loudspeakers vanish, and the virtual source is recreated by the left-high and right-high loudspeakers.

3. In the rear-lateral space ($90^\circ < \theta \leq 135^\circ$, $0^\circ \leq \phi \leq 90^\circ$), a virtual source is recreated by three loudspeakers, i.e., the left-surround, left-high, and right-high loudspeakers.

4. In the rear space ($135^\circ \leq \theta \leq 180^\circ$, $0^\circ \leq \phi \leq 90^\circ$), a virtual source is recreated by four loudspeakers, i.e., the left-surround, right-surround, left-high, and right-high loudspeakers. In the horizontal plane, the signals of left-high and right-high loudspeakers vanish, and the virtual source is recreated by the left-surround and right-surround loudspeakers.

The analysis via the localization equations in Section 6.1.1 and experiments indicated that this loudspeaker configuration and signal mixing can recreate a virtual source in the upper-half space. In addition to recreating a stable virtual source in the front and rear region, the stability of a lateral virtual source is improved in comparison with 5.1-channel reproduction.

The differences in the distances from each loudspeaker to an off-central listening position lead to variations in the propagating time. In this case, the analysis of summing localization and other spatial auditory perception becomes complicated and may deal with the mechanisms of the high-level nervous system to processing binaural information comprehensively (Sections 1.7.1 and 1.7.2). This issue has been explored in some studies. However, complete models for analysis and optimized reproduction performance at the off-central listening position are still unavailable. As such, future studies should develop such models. In the current stage, the analysis, optimization, and evaluation of off-central performance are mainly based on psychoacoustic experiments.

The results of psychoacoustic or subjective assessment experiments on multichannel spatial surround sound generally depend on loudspeaker configurations, signal types, and mixing methods. As the number of reproduction channels increases, virtual sources in many directions can be recreated by the corresponding single loudspeaker, thereby reducing the dependence of the summing virtual source with two or more loudspeakers. In this case, the listening region is enlarged, and the perceived localization quality at the off-central position is improved. This conclusion is valid for both channel- and object-based methods. For example, subjective assessments (by using a method similar to that in Section 15.4) on two-channel stereophonic sound, 5.1-channel sound, and 22.2-channel sound indicate that 22.2-channel sound is obviously superior to the two other types in terms of various perceived attributes (ITU-R Report, BS 2159-7, 2015). Subjective experiments also reveal that the 22.2-channel sound in domestic reproduction demonstrates a good subjective effect within a larger listening region, e.g., off-central to the left or right about 1–2 m and to the front or back about 1 m (Hamasaki et al., 2007). Conversely, using the NHK 22.2-channel sound as a reference, Kim et al. (2010) assessed the subjective directional quality and overall perceived quality of multichannel spatial sound reproduction with various numbers of loudspeakers in the upper (top) layer. They used the stimuli with VBAP signal mixing, including a moving virtual source.
soundtrack. The results indicated that the Samsung 10.2-channel configuration with three loudspeakers in the upper layer yields grading scales similar to those of the 22.2-channel sound. Howie et al. (2017) conducted a double-blind listening test to evaluate a listener’s discrimination among four common channel-based spatial surround sound reproduction with music stimuli, namely, the NHK 22.2-channel sound, ATSC 11.1-channel sound, Samsung 10.2-channel sound, and Auro 9.1-channel sound. The results revealed that listeners can easily discriminate between NHK 22.2-channel sound and the three other reproductions. Listeners can also discriminate between the three other reproductions with a significantly lower success rate.

In channel- and object-based methods, multichannel spatial surround sound requires more independent signals, which demand an increase in the bandwidth of transmission or capacity of storage. The current digital transmission and storage techniques meet these demands (Chapter 13).

Overall, compared with multichannel horizontal surround sound, multichannel spatial surround sound improves the ability to recreate spatial information and perceive performance. However, it requires more reproduction channels and is more complicated. Therefore, the problem of how to make a compromise between complexity and perceptual performance in reproduction remains. In addition, multichannel spatial surround sound opens up a new dimensionality, and new problems should be further investigated, including loudspeaker configuration and signal mixing/rendering, microphone technique, up/down mixing, signal transmission, and storage, related psychoacoustics, and subjective assessments. The prospect of multichannel spatial surround sound depends on practical demands.

6.6 SUMMARY

Multichannel spatial surround sound aims to recreate the three-dimensional spatial information of sound. It is an extension of multichannel horizontal surround sound and a new-generation sound reproduction technique.

Horizontal summing localization theorems are extended to a three-dimensional space. However, the mechanisms of summing localization and other auditory events in a vertical direction are different from those in a horizontal plane. In addition to spectral cues at high frequencies, the dynamic variation in $ITD_p$ caused by head-turning is a cue for vertical localization, as proposed by Wallach. The summing localization equations for a multichannel spatial surround sound at low frequencies are derived by considering $ITD_p$ and its dynamic variation caused by head rotation and tilting. The analysis of velocity and energy localization vectors in horizontal reproduction is also extended to three-dimensional space reproduction.

The summing localization with pair-wise amplitude panning in the median plane can be interpreted and predicted by the summing localization theorems at low frequencies. Conversely, corresponding summing localization experiments are regarded as experimental validation of the theorems. For some appropriate loudspeaker configurations in the median or sagittal plane, the pair-wise amplitude panning can recreate a virtual source between a pair of adjacent loudspeakers but the virtual source may be blurry. However, for up-down symmetrical loudspeaker configurations, the pair-wise amplitude panning fails to recreate the virtual source between the loudspeakers.

Some optimized loudspeaker configuration and signal mixing methods can be derived from the summing localization theorem, such as the VBAP and Ambisonics methods, although the derivations are suitable for central listening positions rather than the off-central listening positions. Even so, the related analyses and experiments are beneficial to the design of
Multichannel spatial surround sound because these analyses and experiments reveal and validate the limitation of the systems and techniques at least.

Multichannel spatial surround sounds are currently a frontier field in sound reproduction, especially the object-based spatial sound is now a trend of development. Various advanced multichannel spatial surround sounds have been developed. As compared with multichannel horizontal surround sound, multichannel spatial surround sound improves abilities to recreate spatial information and improves perceived performances greatly. However, it requires more reproduction channels and is more complicated. Some international organizations and societies have formulated some standards and specifications for multichannel spatial surround sounds. These standards are constantly revised and updated. Numerous new problems related to multichannel spatial surround sound need to be further investigated. The prospect of multichannel spatial surround sound depends on practical demands.