

FREQUENCY DOMAIN SURROUND SOUND PRODUCTION FROM COINCIDENT MICROPHONE ARRAY WITH DIRECTIONAL ENHANCEMENT

SAMUEL SAMSUDIN NG AND SAPNA GEORGE

STMicroelectronics Asia Pacific Pte Ltd
samsudin.ng@st.com

Coincident sound recording and matrix encoding techniques represent the sound stage by a set of intermediate signals, which encode the directional information. A passive decoder employs *virtual microphone* signal mixing to combine these intermediate signals into multiple of signals pointing to various directions corresponding to the surround sound playback speaker configuration. Such passive decoding technique generally suffers from high channel crosstalk, which narrows the reproduced soundstage and creates directional confusion. This paper provides an efficient frequency-domain method to decode the intermediate signals into speaker playback signals with reduced channel crosstalk. Such method can be used, for example, to convert coincident sound recording signals or to decode matrix-encoded stereo signals into multichannel surround sound playback signals. The scheme has been tested and implemented on embedded platform, giving enhanced surround sound playback experience on standard ITU 5.0 surround sound playback setup.

INTRODUCTION

In [1], a method of surround sound production by means of signal mixing (virtual microphone matrixing) from coincident microphone signals was described. Virtual microphone signals with any first order directivity and pointing to any direction can be produced by weighted combination of intermediate signals. These signals possess specific properties, which encode the spatial information. To produce multichannel surround sound, one virtual microphone signal is derived for each surround playback channel. Virtual microphone matrixing effectively decodes the spatial information, enabling approximation of the actual sound stage during playback.

A well-known spatial-encoded signal representation is the coincident B-format signals [2] which consist of W (omni-directional), X (front-facing figure-of-eight) and Y (left-facing figure-of-eight) signals. The directional response of B-format signals for a sound source coming from direction θ° can be modelled as

$$W(\theta) = \frac{1}{\sqrt{2}} \quad (1)$$

$$X(\theta) = \cos(\theta) \quad (2)$$

$$Y(\theta) = \sin(\theta). \quad (3)$$

In practice, these signals can be acquired by special microphone array such as Soundfield microphone or horizontally-coincident arrangement of an omni-directional and two figure-of-eight microphones.

Another well-known signal representation is matrix-encoded stereo signals [3], which encodes multichannel surround sound signals as stereo signals. The stereo signals are commonly referred to as left-total (Lt) and right-total (Rt) signals. In this case the directional response can be modelled as

$$Lt(\theta) = \cos\left(\frac{\theta - 90}{2}\right) \quad (4)$$

$$Rt(\theta) = \sin\left(\frac{\theta - 90}{2}\right). \quad (5)$$

The directional response of B-format and matrix-encoded stereo signals are illustrated in Figure 1(a) and 1(b), respectively.

The virtual microphone matrixing process is described by Equation 6, where $S_i(n)$ ($i = 1, 2, \dots, M$) is the M intermediate signals, $C_j(n)$ ($j = 1, 2, \dots, P$) is the virtual microphone signals corresponding to the P playback channels, n is the sample index, and γ_{S_i, C_j} is the

channel-coefficient for intermediate signal $S_i(n)$ and playback channel signal $C_j(n)$. By designing the channel-coefficient matrix properly, the intermediate signals can be combined to produce the desired multichannel surround sound signals.

$$\begin{bmatrix} C_1(n) \\ C_2(n) \\ \vdots \\ C_P(n) \end{bmatrix} = \begin{bmatrix} \gamma_{S_1, C_1} & \cdots & \gamma_{S_M, C_1} \\ \gamma_{S_1, C_2} & \cdots & \gamma_{S_M, C_2} \\ \vdots & \ddots & \vdots \\ \gamma_{S_1, C_P} & \cdots & \gamma_{S_M, C_P} \end{bmatrix} \cdot \begin{bmatrix} S_1(n) \\ \vdots \\ S_M(n) \end{bmatrix} \quad (6)$$

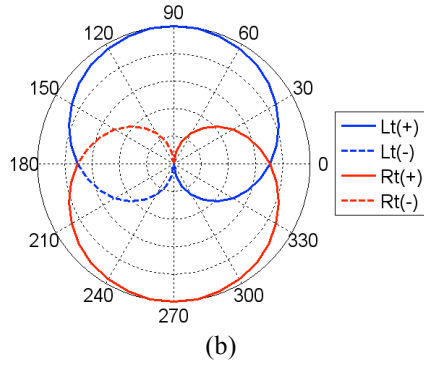
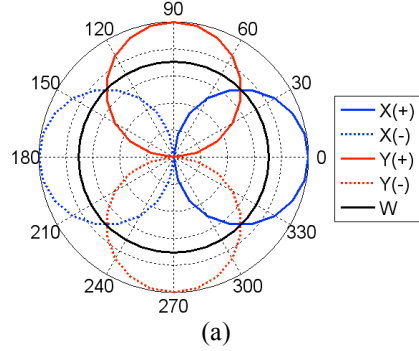


Figure 1: Directional response of (a) B-format and (b) matrix-encoded signals.

As an illustration, a B-format channel-coefficient design solution to derive a virtual microphone signal with directivity d_{C_j} and pointing to the direction α° is

$$C_j^\alpha(n) = \gamma_{W, C_j^\alpha} \cdot W + \gamma_{X, C_j^\alpha} \cdot X + \gamma_{Y, C_j^\alpha} \cdot Y \quad (7)$$

$$\gamma_{W, C_j^\alpha} = \frac{1}{\sqrt{2}} \cdot (2 - d_{C_j}) \quad (8)$$

$$\gamma_{X, C_j^\alpha} = \frac{1}{2} \cdot d_{C_j} \cdot \cos(\alpha) \quad (9)$$

$$\gamma_{Y, C_j^\alpha} = \frac{1}{2} \cdot d_{C_j} \cdot \sin(\alpha). \quad (10)$$

For matrix-encoded signals, the solution could be

$$C_j^\alpha(n) = \gamma_{L, C_j^\alpha} \cdot Lt(n) + \gamma_{R, C_j^\alpha} \cdot Rt(n) \quad (11)$$

$$\gamma_{L, C_j^\alpha} = \cos\left(\frac{\alpha - \pi/2}{2}\right) \quad (12)$$

$$\gamma_{R, C_j^\alpha} = \sin\left(\frac{\alpha - \pi/2}{2}\right). \quad (13)$$

Figure 2(a) shows an example of the directional response of virtual microphone signals for standard ITU surround sound playback, derived from B-format signals. The B-format signals are mixed into five virtual cardioid signals pointing to the direction of 30° , 330° , 0° , 110° and 250° . These directions correspond to the ITU 5.0 loudspeaker location for playback channels left-front (L), right-front (R), center (C), left-surround (Ls) and right-surround (Rs) respectively. Figure 2(b) shows similar directional response of the playback channel signals derived from matrix-encoded signals, with different virtual microphone orientation.

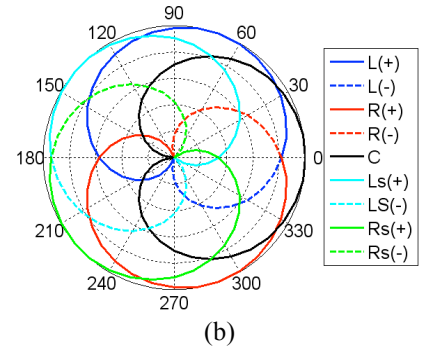
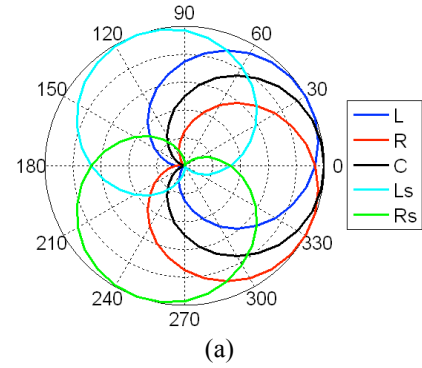


Figure 2: Directional response of virtual microphone signals for standard ITU 5.0 playback configuration derived from (a) B-format and (b) matrix-encoded signals.

1 DIRECTIONAL ENHANCEMENT

As can be seen in Figure 2, the resulting surround sound signals from virtual microphone matrixing have considerably high channel crosstalk. For example, assume a sound is intended to come from the direction

of 30° (left-front direction). The signal generated in the left-front output channel is at the maximum level. Simultaneously, there is also considerably high signal output level from the adjacent channels: center, right-front and left-rear channels. This may pull the reproduced sound image towards the center (narrowing the stereo width), or towards the rear-left direction (creating front-back confusion or directional ambiguity). However, these are the limitation of first order B-format signals, as it is only possible to derive first order virtual microphone signals. Similarly for decoding of matrix-encoded signal into surround sound, the so-called *passive* decoding method results in undesirably high channel crosstalk.

To prevent narrowing of the sound image and enhance the overall directional clarity of the surround sound, channel crosstalk has to be minimized. It is achieved by forming a complimentary virtual microphone signal (cancellation signal) for each of the virtual microphone signal (main signal). The main and cancellation signal pair has to fulfil a certain directional property: the null of the cancellation signal should point to the main signal direction.

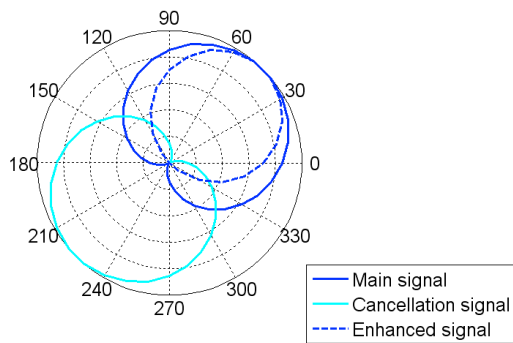


Figure 3: Directional response of main, cancellation, and enhanced signals.

One possible arrangement is a pair of back-to-back virtual cardioid microphone signals. The cancellation signal is then subtracted from the main signal to produce a narrower lobe, as illustrated in Figure 3. This is performed in the frequency domain similar to the spectral subtraction technique [4]. With the specific arrangement of the main and cancellation virtual microphones, the subtraction attenuates the crosstalk signal components while preserving the desired main signal component.

The following section presents the detail of the frequency domain surround sound production scheme with directional enhancement. More details can be found in the patent publication document of the scheme [5].

2 FREQUENCY-DOMAIN SURROUND SOUND PRODUCTION WITH DIRECTIONAL ENHANCEMENT

The surround sound mixing and directional enhancement are performed in the frequency domain by first transforming the time-domain intermediate signals using fast fourier transform (FFT). Overlap-add method can be implemented to segment the signals prior to the transform. The power spectral density (*psd*) of the main and cancellation signal for each surround sound output channel are calculated at each Bark frequency band. This is motivated by psychoacoustic principles while also reducing computational complexity. The directional enhancement gain at each Bark frequency band is calculated by taking the *psd* ratio of the main and cancellation signals. Subsequently, the intermediate signals are mixed and the directional enhancement gain matrix is applied.

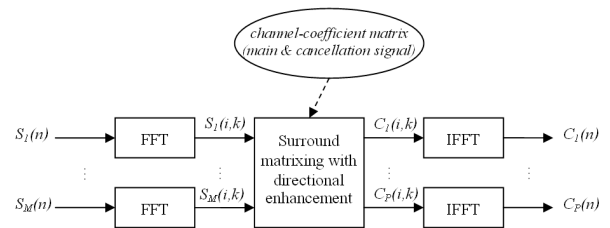


Figure 4: Top-level processing diagram of frequency domain surround sound production method

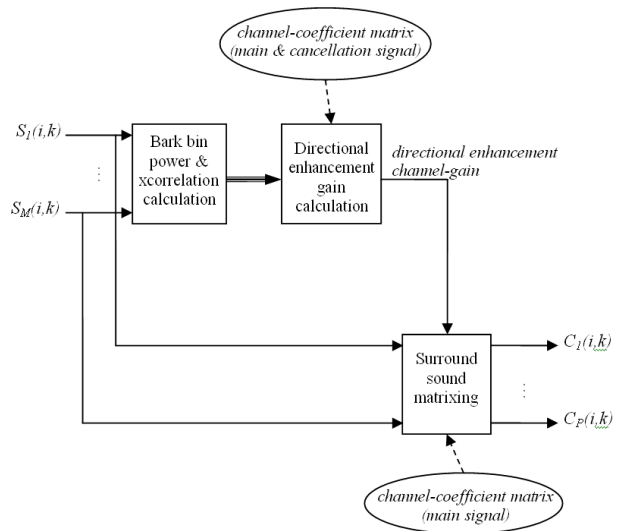


Figure 5: Diagram of surround sound matrixing with directional enhancement process

By simple algebraic expansion, *psd* of any virtual microphone signal can be derived from the power and cross-correlation of the intermediate signals. Hence for any surround sound configuration and number of channels, the computational complexity is more or less

constant and dependent only on the number of the intermediate signals. In most cases the number of the intermediate signals required is always less than the number of the surround sound playback channel.

The top-level processing diagram is shown in Figure 4. The surround sound matrixing and directional enhancement block is further shown in detail in Figure 5. For simplicity of explaining the concept as well as signal labeling, the core algorithm is illustrated using B-format as input signals and standard ITU 5.0 surround sound as output signals. The input signals are labeled as W , X and Y , while the output channels are labeled as L , R , C , Ls and Rs .

2.1 Windowing and frequency transform

Using overlap-add technique, the time-domain intermediate signals are first segmented into block of samples. Windowing function is applied, followed by frequency transform by fast fourier transform (FFT). These FFT signals are input to the surround sound production block. For a typical 44.1 or 48 kHz sampling rate, block size of 1024 samples with 50% window overlap is adequate.

2.2 Bark-band power spectral density calculation

One of the main features of this scheme is the direct calculation of the psd of the main and cancellation signals via the intermediate signals. For example with B-format input signals, the psd of output channel signal ch can be calculated according to Equation 14. In this equation, i refers to the frame (time) index, k_b refers to the FFT-bin border corresponding to bark band index b and $[\gamma_{W,ch}, \gamma_{X,ch}, \gamma_{Y,ch}]$ refer to the channel-coefficient corresponding to the virtual microphone signals for output channel ch . Essentially, to calculate the psd for any virtual microphone signal, only the signal power $PW(i,b)$, $PX(i,b)$ and $PY(i,b)$, and cross-correlation $CWX(i,b)$, $CWY(i,b)$ and $CXY(i,b)$ are required. These are calculated following Equation 15 to Equation 20 where $*$ denotes complex conjugate.

$$PSD_{ch}(i,b) = \sqrt{\left(\sum_{j=k_b}^{k_{b+1}-1} \left| \gamma_{W,ch} \cdot W(i,j) + \gamma_{X,ch} \cdot X(i,j) + \gamma_{Y,ch} \cdot Y(i,j) \right|^2 \right)} \quad (14)$$

$$= \sqrt{\begin{pmatrix} \gamma_{W,ch}^2 \cdot PW(i,b) \\ + \gamma_{X,ch}^2 \cdot PX(i,b) \\ + \gamma_{Y,ch}^2 \cdot PY(i,b) \\ + 2 \cdot \begin{pmatrix} \gamma_{W,ch} \gamma_{X,ch} CWX(i,b) \\ + \gamma_{W,ch} \gamma_{Y,ch} CWY(i,b) \\ + \gamma_{X,ch} \gamma_{Y,ch} CXY(i,b) \end{pmatrix} \end{pmatrix}}$$

$$PW(i,b) = \sum_{j=k_b}^{k_{b+1}-1} |W(i,j)|^2 \quad (15)$$

$$PX(i,b) = \sum_{j=k_b}^{k_{b+1}-1} |X(i,j)|^2 \quad (16)$$

$$PY(i,b) = \sum_{j=k_b}^{k_{b+1}-1} |Y(i,j)|^2 \quad (17)$$

$$CWX(i,b) = \sum_{j=k_b}^{k_{b+1}-1} (W(i,j) \cdot X^*(i,j)) \quad (18)$$

$$CWY(i,b) = \sum_{j=k_b}^{k_{b+1}-1} (W(i,j) \cdot Y^*(i,j)) \quad (19)$$

$$CXY(i,b) = \sum_{j=k_b}^{k_{b+1}-1} (X(i,j) \cdot Y^*(i,j)) \quad (20)$$

2.3 Directional enhancement gain calculation and surround sound signal generation

First, the main and cancellation signals' channel-coefficient are designed. As an example, for front left channel, the main signal has cardioid directivity pointing to 30° (location of front left speaker in the ITU 5.0 playback configuration) while the cancellation signal has cardioid directivity pointing to 210° .

Once the channel-coefficient $[\gamma_{W,ch}, \gamma_{X,ch}, \gamma_{Y,ch}]_{main}$ and $[\gamma_{W,ch}, \gamma_{X,ch}, \gamma_{Y,ch}]_{cancel}$ are designed, the psd of the main and cancellation signal $PSD_{ch,main}(i,b)$ and $PSD_{ch,cancel}(i,b)$ are calculated according to the equation presented in the previous section. The cancellation gain at each bark bin, which is the amount of attenuation applied to the frequency region to reduce the channel crosstalk, is calculated according to

$$gain_{ch}(i,b) = \frac{PSD_{ch,main}(i,b) - cFac \cdot PSD_{ch,cancel}(i,b)}{PSD_{ch,main}(i,b)} \quad (21)$$

where $cFac$ is a parameter to control the amount of cancellation. To prevent unwanted artifacts such as spectral hole or musical noise, the gain may be limited as well as smoothed over time.

The bark-bin gain values are subsequently mapped to the corresponding FFT-bin according to

$$gain_{FFT_{ch}}(i,k) = gain_{ch}(i,b_k) \quad (22)$$

where b_k is the bark-bin index b which corresponds to FFT-bin index k . The value of gain might change abruptly between adjacent FFT bins and may cause undesired artifacts. The gain may be smoothed over the frequency bin to minimize such artifacts. Finally, the surround sound output signals are mixed from the intermediate signals according to Equation 23.

$$\begin{bmatrix} C_1(i,k) \\ C_2(i,k) \\ \vdots \\ C_p(i,k) \end{bmatrix} = \begin{bmatrix} \text{gainFFT}_{C_1}(i,k) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \text{gainFFT}_{C_p}(i,k) \end{bmatrix} \quad (23)$$

$$\begin{bmatrix} \gamma_{S_1,C_1} & \cdots & \gamma_{S_M,C_1} \\ \gamma_{S_1,C_2} & \cdots & \gamma_{S_M,C_2} \\ \vdots & \ddots & \vdots \\ \gamma_{S_1,C_p} & \cdots & \gamma_{S_M,C_p} \end{bmatrix} \begin{bmatrix} S_1(i,k) \\ \vdots \\ S_M(i,k) \end{bmatrix}$$

3 RESULTS

Figure 7 shows the polar pattern of a specific design of virtual microphones for ITU 5.0 surround sound playback configuration. Directional enhancement was applied to the design, resulting in narrower lobe. This means reduced playback channel crosstalk.

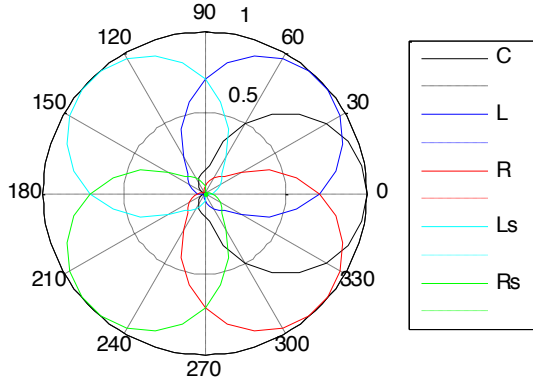


Figure 6: Directional enhanced polar pattern of virtual microphone signals for ITU 5.0 surround sound playback

Actual B-format sound recording has been conducted, with a person walking around and across the microphone while announcing his location. The signals was mixed into 5.0 surround sound signals with the method presented in this paper. During playback, the sound scene was recreated with improved directional clarity over simple surround sound mixing.

REFERENCES

- [1] Samsudin, S. George, "Low Complexity MPEG Encoding For Surround Sound Recordings", *United States Patent #8,332,229*
- [2] R. K. Furness, "Ambisonics – An Overview", *Proceedings of the AES 8th International*

Conference: The Sound Of Audio, Washington, D.C., USA (1990).

- [3] K. Gundry, "A New Active Matrix Decoder for Surround Sound", *Proceedings of the AES 19th International Conference on Surround Sound*, Schloss Elmau, Germany (2001).
- [4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 208-211 (1979).
- [5] Samsudin, S. George, "System and Method for Efficient Sound Production using Directional Enhancement", *United States Patent Application Publication US20130044894 A1*.