SURROUND RECORDING USING COINCIDENT AND SPACED MICROPHONES COMBINED WITH 2-TO-3 UPMIXING

Benedict Slotte

Nokia Corp.
Technologies & Platforms, Product Technologies
Joensuunkatu 7E, 24100 Salo, Finland
benedict.slotte@nokia.com

ABSTRACT

This paper will discuss a few possible methods in surround recording that aims to combine the advantages of both spaced and coincident microphones. The text will concentrate on conventional microphone techniques and leave out more advanced methods such as beamforming and wave field synthesis. Specifically as a means of recording the front channels using conventional stereo microphones or coincident pairs, an upmixing method will be considered in combination with ambience recording using conventional spaced pairs. The results of using this method will be compared with a few other established methods.

1. INTRODUCTION

The ideas described in this paper were originally based on the author's work in developing a surround recording technique that:

(a) provides accurate and detailed imaging (i.e. high directional resolution, low blurring of phantom images) for the front sector (left, center, right, i.e. L, C, R),
(b) achieves a spacious and cohesive representation of the hall ambience,
(c) uses a minimum number of channels, and
(d) uses conventional microphone patterns and technology (1st-order patterns, single microphones or microphone pairs, no large multimicrophone arrays).

This work is not part of any research project and thus no formal subjective testing to assess its performance has been carried out, except for the author's personal listening evaluations.

The front channel recording technique will be discussed first, and the addition of ambience will be treated as a separate matter. From the definition of the goals of the methods described here ((a)-(d) above), it is obvious that in this case a typical concert hall recording is the norm: the front channels reproduce the actual ensemble on stage as well as a part of the early reflections, whereas the back channels reproduce further early reflections, hall ambience, and audience noise if wanted.

In the entire text it is assumed that the by now well-established 5.0/5.1 loudspeaker layout (ITU-R BS 775-1, where the L and R loudspeakers are at ±30° from the C (center) loudspeaker, and the Ls and Rs (left and right surround) are at ±110°) is used.

2. RECORDING THE FRONT CHANNELS

2.1. Some things learned from stereo

The history of stereo recording provides a good starting point for achieving goals (a) and (d) above. Although personal taste and opinions differ quite a lot among recording engineers and researchers, one can generally say that conventional stereo recording techniques can be pictured along an axis between two extremes:

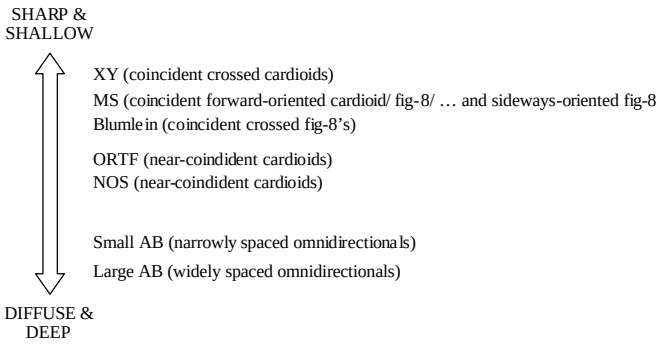

Figure 1. *Subjective characteristics of stereo image of a few well-known stereo microphone techniques*.

In this picture, "sharp" means that the reproduced soundstage has a high directional resolution so that individual sound sources can be accurately pinpointed, and "deep" means that the phantom images of sound sources localize not only on the arc between the loudspeakers but also further away from the listener (well behind the plane defined by the loudspeakers).

In other words, coincident techniques tend to provide the sharpest localization along the whole stereo arc, while they lack depth, whereas the opposite is true for spaced pairs. This is due to the fact that pure amplitude differences between stereo channels, as when coincident techniques are used, generally cause sharper localization than pure time-of-arrival differences as in spaced techniques. Conversely, the relative delays present in spaced techniques cause phase differences that tend to widen and diffuse the stereo image. In the Blumlein setup much of the reverberation is picked up out-of-phase and thus also gives an impression of the sound extending a bit beyond the loudspeaker base (±30°). The Blumlein technique of course also picks up a greater number of early reflections, meaning an emphasized sense of depth. Blumlein generally also offers sharp localization, which is why it is very favoured among some recording engineers. Also the MS (mid-side) technique has similar properties as XY and Blumlein, depending somewhat on the relative gains assigned to the M and S signals. It should be noted that one particular near-coincident setup, namely the "sphere microphone" manufactured by Schoeps, is also claimed to have accurate localization [1]. The author does not own such a microphone and has not compared it to the other microphone setups described here, but is inclined not to believe that the imaging sharpness could be as high as for coincident recording, particularly at high frequencies and for sounds relatively close to the center of the soundstage. Also the Stereo-180 microphone setup [2] uses near-coincident placement (the distance is as small as 4.6 cm), and although it is claimed to have better optimized localization at least at low and midrange frequencies, the stereo image disintegrates somewhat at high frequencies.

Some of the differences between recording techniques as regards subjective stereo image depth are of course due to the differing amount of early reflections picked up by the different microphone patterns. In AB recording (especially wide AB) the inherent time delays act as further early reflections, which is likely to be another reason why there is more depth in the soundstage (it is well known that early reflections play a crucial role in providing such depth). However, even when only one particular polar pattern is employed (e.g. cardioid), it is easy to demonstrate that spacing the microphones apart acts to increase subjective depth. The ORTF and NOS techniques (which use angled cardioids with spacings of 17-30 cm) act as good compromises between these two extremes, which is the reason why many recording engineers favour them most of all.

2.2. Can it be extended to three channels ("LCR stereo")?

The fact that a higher degree of inter-channel delay increases subjective depth as well as localization blur is equally valid in the 3-channel case. This can be easily demonstrated by real (or computer-simulated!) recording of sound sources even in the absence of early reflections and reverberation. Similarly, coincident 3-channel recording tends to increase localization sharpness. Thus a coincident method best fulfils goal (a)

above – but how to extend the XY technique to 3 channels? Intuitively, one could simply derive a center channel (C) from the L and R signals as:

$$C_o = k_C\left(L_i + R_i\right) \tag{1}$$

where k_C is a gain coefficient (which would normally be of the order of 1). The "*i*" index mean "input", "*o*" means "output". However, doing this leads to a very center-heavy soundstage unless the L and R signals are also expanded as follows:

$$\begin{aligned} L_o &= L_i + k_S\left(L_i - R_i\right) \\ R_o &= R_i + k_S\left(R_i - L_i\right) \end{aligned} \tag{2}$$

where k_S is a gain coefficient for the difference signal, i.e. a stereo expansion coefficient (which would also normally be of the order of 1). This stereo expansion by amplifying the difference signal is well known and forms the idea behind e.g. M-S processing used in mastering to alter the stereo width of recorded material. $k_C = 0$ and $k_S = 1$ means that the original signals are preserved. $k_S = -0.5$ means that the stereo width is reduced to zero. Equations (2) could be further simplified by adding just the opposite channel (not the difference of the channels), but acting explicitly on the sum and difference signals maintains a greater degree of intuitivity for the parameter user interface if the transformation is implemented as a VST plugin etc. (the author has implemented it in the Reaktor 4.0 program by Native Instruments [3]).

2.3. A practical implementation of the upmixing/ width control

To keep the total overall loudness approximately constant in typical cases when the coefficients are varied, a normalization factor (the denominator in equations (3) below, part of which was found empirically) can be added. Other user-friendliness enhancements include mapping of suitable, convenient parameters to k_C and k_S above. For example, it would be desirable to have an L and R "width" value of 0 correspond to mono and 1 to no change. For reasons that shall be briefly discussed below, a "center level vs left and right level" control should also be provided. All this can be done as follows:

$$\begin{aligned} L_o &= \frac{L_i + \frac{k_{\exp} - 1}{2}\left(L_i - R_i\right)}{\sqrt{1 + \frac{10^{\frac{g_C}{10}}}{2} + \frac{{k_{\exp}}^2}{4}}} \\ C_o &= \frac{\frac{10^{\frac{g_C}{20}}}{2}\left(L_i + R_i\right)}{\sqrt{1 + \frac{10^{\frac{g_C}{10}}}{2} + \frac{{k_{\exp}}^2}{4}}} \\ R_o &= \frac{R_i + \frac{k_{\exp} - 1}{2}\left(R_i - L_i\right)}{\sqrt{1 + \frac{10^{\frac{g_C}{10}}}{2} + \frac{{k_{\exp}}^2}{4}}} \end{aligned} \tag{3}$$

or, in matrix notation,

$$\begin{pmatrix} L_o \\ C_o \\ R_o \end{pmatrix} = \frac{\begin{pmatrix} 1+\frac{k_{\text{exp}}-1}{2} & -\frac{k_{\text{exp}}-1}{2} \\ \frac{10^{\frac{g_C}{20}}}{2} & \frac{10^{\frac{g_C}{20}}}{2} \\ -\frac{k_{\text{exp}}-1}{2} & 1+\frac{k_{\text{exp}}-1}{2} \end{pmatrix} \begin{pmatrix} L_i \\ R_i \end{pmatrix}}{\sqrt{1+\frac{10^{\frac{g_C}{10}}}{2}+\frac{k_{\text{exp}}{}^2}{4}}} \tag{4}$$

In other words,

$$\begin{aligned} k_C &= \frac{10^{\frac{g_C}{20}}}{2} \\ k_S &= \frac{k_{\text{exp}}-1}{2} \end{aligned} \tag{5}$$

so k_{exp} is an L and R expansion coefficient (≥0) and g_C determines the level of the C channel (in dB) compared to L and R for sounds arriving from the center of the soundstage (0°).

This transformation is similar in principle to the 2-to-3-channel upmixing proposed by M. A. Gerzon in [4]. Gerzon further requires the matrix to be energy preserving, which would eliminate either k_S or k_C by making one dependent on the other. Furthermore, the normalization factor would be different. The author prefers instead the above approach since it has one more degree of freedom, and tuning by ear is better in practical work. Although the normalization may not be the optimum one for all signals, it preserves the overall subjective loudness well for typical practical values of g_C and k_{exp} (which are usually in the range 0 to 10 dB and 2 to 9, respectively) and is thus useful in practical mixing. To further improve the localization sharpness, high frequencies (above about 5 kHz) can be slightly emphasized in the L_o and R_o signals, and damped in C_o, as is also done in [4]. A couple of dB is usually enough. This could be handled by two additional parameters g_{Cc} and g_{LRc}, both given in dB, and used so that the former damps the C channel by the given amount (above about 5 kHz only), and the latter amplifies the L and R channels correspondingly (the exact frequency above which this correction is applied is not so critical, as is also noted by Gerzon, but it should be no less than about 5 kHz).

2.4. Does upmixing really make sense?

Why derive a center channel instead of just recording it using a third microphone? If the number of channels is not an issue, then not very much is gained by doing it this way. But the fact is that when high-quality coincident microphones are used, the “derived” (i.e. obtained through upmixing) center channel works very well, and it is justified as soon as there is any reason to save one channel for something else (spot microphones, ambience microphones). This can be an issue especially when recording using e.g. ADAT bit-split mode at high sample rates, since the number of channels is then halved, or if storage space is to be saved.

Other potential reasons to use the upmixing method are e.g. ease of microphone setup (existing coincident stereo microphones can be used), maximally small size, and cost.

"High-quality microphones" in this case means that the microphones have a very stable polar pattern, a smooth frequency response, and small diaphragms so that they can be placed very close to each other. The author uses 2 Schoeps MK4Vg micropones for this purpose. In addition to being of high quality in general, these microphones can be easily mounted one above the other (with the upper one turned upside down), since the capsules have their maximum pickup in the radial direction, to ensure very good time alignment between the L and R signals when picking up sounds in the horizontal plane. Using the miniaturized versions of the microphones instead of capsules with CMC5 preamplifier bodies (see picture below) would of course make the total size smaller. (It would also be possible to use Schoeps's dedicated CMXY microphone, but the fact that its capsules are not coincident in the horizontal plane may create problems at very high frequencies.) Furthermore, the preamplifier has to have very good phase matching between channels, and it goes without saying that the microphone pair itself also has to be matched.

The actual mathematical properties of upmixed 2-channel recording vs "real" 3-channel recording will be investigated in some more detail below, but it suffices to say in advance that the difference as regards performance is not as big as one might expect. After all, MS recording also works well in practice although it is based on similarly "synthesized" polar patterns.

Figure 2. *Mounting of 2 Schoeps MK4Vg capsules (with CMC5 bodies) above one another to achieve perfect time alignment for horizontal pickup.*

It is also worth noting that the panning of spot microphones becomes particularly straightforward when the upmixing method is used: conventional 2-channel amplitude panning can be utilized (it is then applied before the upmixing, of course). When the upmixing to 3 channels is then carried out, both the coincident main microphone signals and the panned contributions from the spot microphones go through exactly the same transformation, which ensures perfectly consistent images. Recording the front channels using 3 spaced or coincident microphones, and then panning spot microphones using e.g. conventional pairwise amplitude panning, would not yield consistent results, at least not for off-center listening. Of course, this depends also on *how* the signal from the spot microphones is used – to solidify the location of the particular sound, or merely to bring up its level to make it better heard. Although such panning of monophonic spot microphone signals is generally considered bad practice as it does not sound as natural as when using stereo spot microphones, the author maintains that the required missing ambience can be separately provided by other means that do not interfere with the accurate localization provided by the main coincident and spot microphones.

It is of course possible to go even further along the way of reducing the number of channels, and record 5 channels (L, C, R, Ls, Rs) using only 3 microphones (e.g. two figure-8's and one omnidirectional) in a so-called double-MS configuration so that the "M" channel is shared, as described by Schoeps in [5]. Similarly, the Soundfield microphone allows the user to synthesize various directional patterns pointing in different directions (from 4 internal microphone capsules). However, as stated above, coincident recording of ambience is not recommended since it easily leads to subjectively impaired depth and envelopment (the correlation between channels becomes too high especially at low frequencies).

2.5. Does coincident recording really make sense?

As stated above, the coincident technique causes a severe lack of image depth, but experience shows that this depth can be very well restored by the ambience provided by the Ls and Rs (surround) channels when properly recorded – especially if part of this ambience is also recorded closer to the stage and routed to L and R, as will be described later in this text.

What about the artistic considerations? The author likes to favour accurate, pinpoint imaging for the simple reason that since the visual image is lacking anyway (assuming that the final medium is audio only), the directional resolution of the auditory one should be as high as possible. (If on the other hand the visual component *is* available, it may perhaps be more beneficial *not* to provide too sharp imaging unless the visual and auditory images can be guaranteed to coincide fairly well.) At the end of this text, a few more aspects of coincident vs spaced recording will be discussed.

2.6. 2-to-3 upmixing: vector analysis

Using *energy vector* and *velocity vector* analysis, it is possible to easily find the optimum values for the parameters g_C, $k_{\exp}$, g_{Cc} and g_{LRc} when the recording angle, microphone angle and microphone patterns are known. The actual calculation of the vectors will not be considered here (refer instead to e.g. [4], [6]), but it is useful here to investigate the *behaviour* of the energy and velocity vectors, and especially the "virtual microphone polar patterns" generated by the 2-to-3 upmixing. It is assumed here, as in the rest of this paper, that the loudspeakers are at 0° and ±30°.

For this, suppose first that the starting point is a 2-channel "XY" setup consisting of coincident cardioids at a 90° angle:

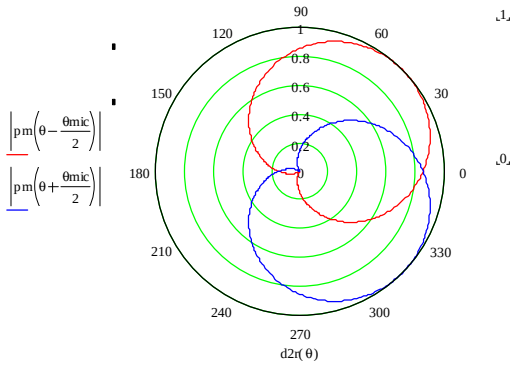

Figure 3. *Polar patterns of XY microphone setup.*

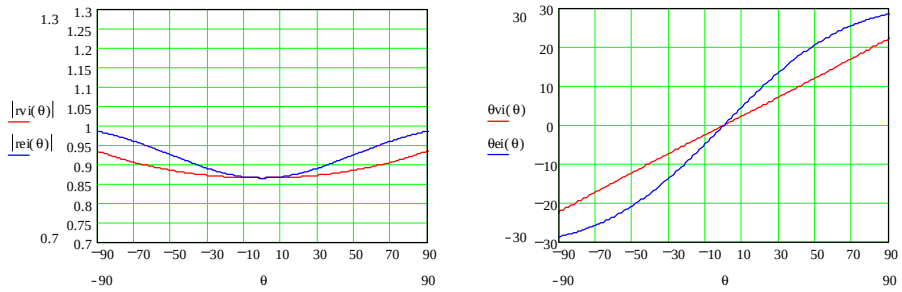

Figure 4. *Vector magnitudes (left,* $\mathbf{r}_v$ *= velocity vector,* $\mathbf{r}_e$ *= energy vector) and directions (right,* $?_v$ *= velocity vector angle,* $?_e$ *= energy vector angle) for XY microphone setup. The indices "i" mean "input", i.e. original unprocessed L and R signals.q is the sound incidence angle.*

These graphs were generated by the program Mathcad 6.0, and the same program was also used to find optimum parameters for the rest of the cases considered in this section. Generally, velocity and energy vector graphs (as in fig. 4 above) can be interpreted as follows [4]:

1. The velocity vector direction provides an indication of where the reproduced phantom image will localize *at frequencies below about 700 Hz*.
2. The energy vector direction correspondingly provides an indication of where the reproduced signal will localize *at frequencies between about 700 Hz and 3.5 kHz* (and also about the behaviour of phantom images *for off-center listeners*).
3. *At frequencies above some 5 kHz*, the phantom image tends to get pulled more towards the louder of the loudspeakers than predicted by the energy vector direction.
4. The energy vector magnitude should be as close to 1 as possible for good phantom image stability when the listener moves sideways or rotates his/her head. The degree of instability is proportional to the difference between this value and 1, i.e. $1-\left|\mathbf{r}_e\right|$.
5. The velocity and energy vector directions should be as close to each other as possible for sharp imaging (i.e. various parts of the frequency spectrum should have their phantom images in the same direction).
6. The angle between the points at which the velocity and energy vector directions equal ±30° is the effective *recording angle*.

In other words, the XY setup described above has a recording angle greater than 180°, and recording of an ensemble occupying an angle of e.g. 80° would result in a reproduced angle of only 1/3 to 2/3 of the ±30° stereo arc (as seen in the vector direction graphs in fig. 4) This is why usually the angle between the cardioids has to be correspondingly increased (110° is common) to get a better channel separation and thus a wider stereo image. However, as will be seen now, a small microphone angle such as 90° can work well when combined with the 2-to-3-channel upmixing method described above.

Let us now assume that the desired recording angle is 100° and that the original XY recording (using a microphone angle of 90°) is to be thus processed using the upmixing technique. First of all, it is worth noting that the C channel should preferably be some 6-9 dB louder than L and R for direct sound arriving from straight ahead. If instead C has roughly the same amplitude as L and R, colouration and/or unnecessarily pronounced center image instability will result. Thus the "C level vs L and R" (g_C) control is fundamental and should preferably be set first, after which the "expansion" ($k_{\exp}$) control can be set to achieve the desired width. While these adjustments are being done, it is worth noting also that g_C will alter the overall balance of

center of stage vs edges. Assuming that g_C = 8 dB, a value of $k_{\exp}$ = 5.3, and high frequency corrections of $g_{Cc} = g_{LRc}$ = 1 dB gives the following vectors and virtual polar patterns:

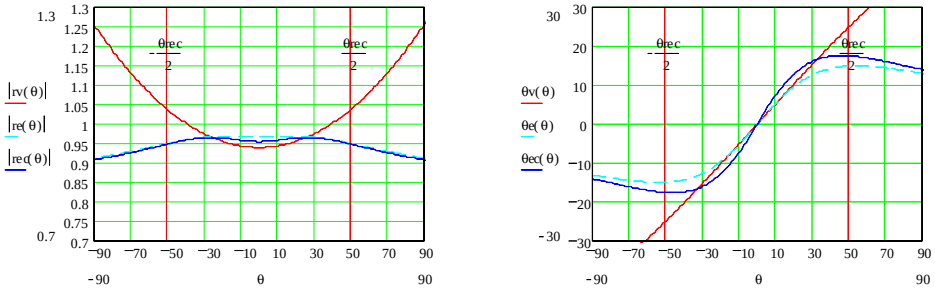

Figure 5. *Velocity and energy vector magnitudes (left) and velocity and energy vector directions (right) for upmixed XY recording. Dashed curve = original energy vector, solid curve = correction above 5 kHz.*

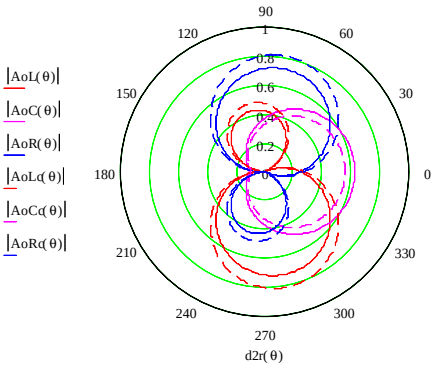

Figure 6. *Polar patterns generated from the original XY microphone setup (see fig. 4) by the upmixing. Dashed lines = correction above 5 kHz.*

Many things can be noted in figs. 5 and 6. First of all, in fig. 6 it is seen that the upmixing in effect generates two (approximate) hypercardioid patterns and one (approximate) cardioid pattern from the two original cardioid patterns. Secondly, as seen in fig. 5, both the velocity and energy vector amplitudes are considerably closer to 1 than in the original XY setup (fig. 4). What this means, as noted above, is that the stereo image will be much more stable as a result of adding the "virtual center microphone". At the edges of the recording sector (and beyond), the velocity vector amplitude rises higher and higher above 1. Subjectively this is heard as a more and more phasey sound, since (as is also seen in fig. 6) there is a considerable out-of-phase lobe in the equivalent polar pattern of the opposite channel's microphone. This is generally unpleasant and one of the reasons why this upmixing should always be tuned by ear. Whether this effect is objectionably strong or not will of course also depend on how much sound (in this case: usually reverberation) is entering from such directions.

Furthermore, the energy vector (fig. 5, right side) does not go all the way to ±25° (within the recording angle q_{rec} shown by the markers) as the velocity vector does. According to the above this would mean that higher frequencies would not localize as far out on the sides as they should – but in practice this effect does not seem to be as noticeable as the graph would indicate, since it is probably cancelled out by the effect of high frequencies generally being "pulled towards" the closest (loudest) loudspeaker. (The angle ±25°, instead of ±30°, has been chosen here since to avoid the slightly annoying phasiness, one may need to sacrifice a little bit of stereo width, and this happens to be one of the disadvantages of this upmixing process.) However, it still

pays to slightly enhance the width at high frequencies, and what this means in terms of the polar patterns is shown by the dashed lines in fig. 6)

Finally, one can see from the polar patterns that the center of the soundstage will be slightly damped. This might be desirable sometimes (e.g. if the main XY pair is placed quite close to the ensemble), but in general the g_C parameter can be used to tune the center level to taste (with a corresponding need to readjust k_{exp}). Anyway, 8 dB is generally a good starting point.

If it is desired that $?_v$ = ±25° at the edges of the recording angle, if g_C = 8 dB, and if the microphone setup is as described above (coincident cardioids at 90°), it is possible to derive the following table of required values for k_{exp}:

q_{rec}	k_{exp}
40°	13.6
60°	9
80°	6.7
100°	5.3
120°	4.4
140°	3.7
160°	3.2
180°	2.8

Table 1. k_{exp} vs q_{rec} when g_C = 8 dB, the desired maximum $?_v$ = ±25°, and the microphones are coincident cardioids at 90°.

This text will not go further into the details of deriving upmixing parameter values for other microphone angles (or polar patterns) and other values of g_C, since these things must be mostly tuned by ear in real life. But as a summary, it can be said that the best distribution of the ensemble at the actual recording is in a quite wide arc around the XY microphone pair. A wide arc is better than a narrow one since the latter requires more extreme artificial widening, which also means a higher degree of antiphase signals from sound picked up outside the recording angle. Also, since widening is about amplifying the difference between channels, any mismatches in the microphone frequency responses or polar patterns will then also be amplified. Yet another disadvantage of extreme widening is the fact that for phantom images intended to be close to the edge (e.g. close to the left loudspeaker), the antiphase component in the loudspeaker on the opposite side (i.e. the *right* loudspeaker) will be so strong that when the listener is further outside the sweet spot (in this case, far to the *right* of the sweet spot), the whole stereo image can be "folded over" so that everything that should have been at the extreme left is heard instead from the right half of the soundstage.

Another point worth noting is that if, as is usually the case, pickup of reverberation from the back of the hall is not wanted in the "dry" LCR stereo signal derived from the XY pair (since it is going to be added separately from other microphones in the mixing), a microphone angle of 90° is a good choice. If instead the microphone angle is widened to between 90° and 180°, it means that smaller values of k_{exp} will be sufficient, and also that the virtual center channel microphone will approach a subcardioid or omnidirectional polar pattern, and more pickup from the rear of the hall will result.

2.7. Further notes about increasing the width

The slight reduction of apparent width as a result of the upmixing was mentioned above. Although this might seem bad enough, in the author's opinion this disadvantage is well outweighed by the increased localization sharpness and the increased width and envelopment of the early reflections and reverberation that can be brought by conventional use of the surround channels.

It might seem tempting to add some width by using a spaced pair (one microphone some tens of centimetres to the left of the XY pair, and one to the right, mixed into the L and R channel, respectively). This is not recommended since (unless the spaced pair has instead a much wider spacing and is very carefully mixed in) it will then also blur the sharp imaging produced by the XY pair. The author has used a spaced pair which is gently low-pass filtered (e.g. starting from 80 Hz, to compensate for possible lack of bass in the cardioids) with good results, but the damping of higher frequencies has then been high enough not to blur the imaging.

2.8. Comparison to three microphones without upmixing

As a comparison point to the above, real 3-channel recording using three coincident hypercardioid microphones will be considered next. (The patterns of the microphones can of course be changed depending on the recording situation, and the C microphone need not have the same pattern as the L and R microphones, but in this case only hypercardioid will be considered. Further examples can be found in [7].)

Assuming that the same recording angle and angular spread of phantom sources as in the above upmixing example is desired (i.e. 100° and ±25°), the required angle between the L and R microphone can be found to be 165°, and the vectors and polar patterns will then be as follows:

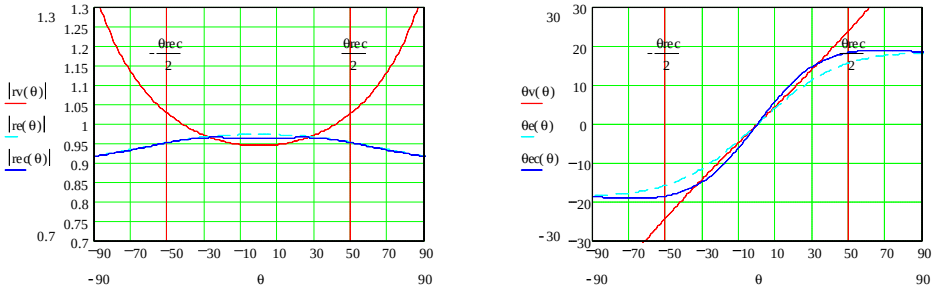

Figure 7. *Velocity and energy vector magnitudes (left) and velocity and energy vector directions (right) for 3-channel coincident hypercardioid recording. Dashed curve = original energy vector, solid curve = correction above 5 kHz.*

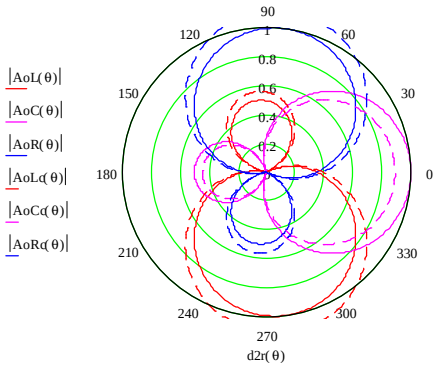

Figure 8. *Polar patterns of 3 coincident hypercardioid microphones. Dashed lines = correction above 5 kHz.*

Comparing figs. 7 and 5, and 8 and 6, it is easy to see that the difference is not dramatic, apart from the pattern of the center channel microphone (the difference in the overall gain, which comes from the fact that the upmixed signals were normalized as in eq. 3 whereas the 3-channel signals were not, is arbitrary and can be ignored). Especially the vectors behave very much the same way inside the recording sector. What this means is that as regards localization, the difference between the real and upmixed 3-channel recordings would be very small. Thus the point made at the beginning of this paper is valid. (Note also that a high-frequency correction of ±1 dB as in the upmixed case has been applied here as well.)

When three microphones are used, there is of course more freedom to choose the pattern of the center microphone according to the amount of rear attenuation needed. The upmixing process leaves no freedom to change the virtual center pattern when g_C and k_{exp} have once been set. However, polar pattern choices other than cardioid are equally valid for the upmixed 2-channel XY microphones, although such patterns will generally lead to stronger antiphase signal components, which might be objectionable.

2.9. Does upmixing work for noncoincident microphones?

There is evidence that the upmixing method clearly improves the reproduction even when using AB (spaced pair) microphone setups [8], even though the imaging accuracy cannot reach that of coincident recording. For this reason the upmixing method is of course also suitable as a general 2-to-3-channel process e.g. for domestic listening to 2-channel recordings (of any kind), and it has also been applied commercially. It should be noted that aspects of these processing methods are the subjects of patents assigned to M. A. Gerzon or Trifield Productions Ltd. [4].

However, the (generalized) method was presented here in conjunction with a specific recording technique in order to demonstrate how it can transform a 2-channel coincident recording into a virtual 3-channel counterpart that in many respects has clearly better performance than the original 2-channel recording.

3. RECORDING THE AMBIENCE

Since the goal of the method(s) presented here is to favour the front channels as regards localization accuracy, and make the ambience subjectively spacious and diffuse (without causing too strong localization that could interfere with the frontal localization), it is clear that the left and right surround channels benefit from being recorded using a spaced pair. Reducing the interaction with the imaging provided in the front sector also means minimizing the direct sound from the stage.

3.1. Polar pattern

The simplest way of recording hall ambience is to use two spaced microphones, one for the left surround (Ls) and one for the right surround (Rs). These microphones can be placed either further back from the main microphones, or at the sides [9]. This would be in line with the goals set at the beginning of this paper about making surround recordings that exhibit both spaciousness and imaging sharpness while using a minimum number of channels (in this case 4). (Actually, the method proposed by Schoeps in [5] would use only 3 channels, but at the expense of clearly reduced spaciousness.)

The spaced omnidirectional pair is a well-known method of recording ambience. However, omnidirectional microphones also pick up considerable amounts of direct sound, and if this is to be reduced, the microphones have to be placed far away from the stage (or close to the rear of the hall), which in turn might cause the ambience to be less cohesive (even when the signals from the main microphone are properly delayed). Another possibility is to use figure-of-8's with their null planes oriented towards the stage (four such microphones in a square arrangement form a so-called *Hamasaki square*). However, the author would like to stress that a problem with figure-of-8's is that their "null" region is much narrower than that of cardioids. Thus, a sideways-oriented figure-8 (as often used in the Hamasaki square) is bound to pick up much more direct sound than a backwards- (or slightly sideways-) oriented cardioid, since for the figure-8's null to be really effective (i.e. damp most of the direct sound), the ensemble has to be quite small. As an illustration of this, consider the following comparison of polar patterns and damping (in dB) as a function of pickup angle for both a cardioid and a figure-of-8:

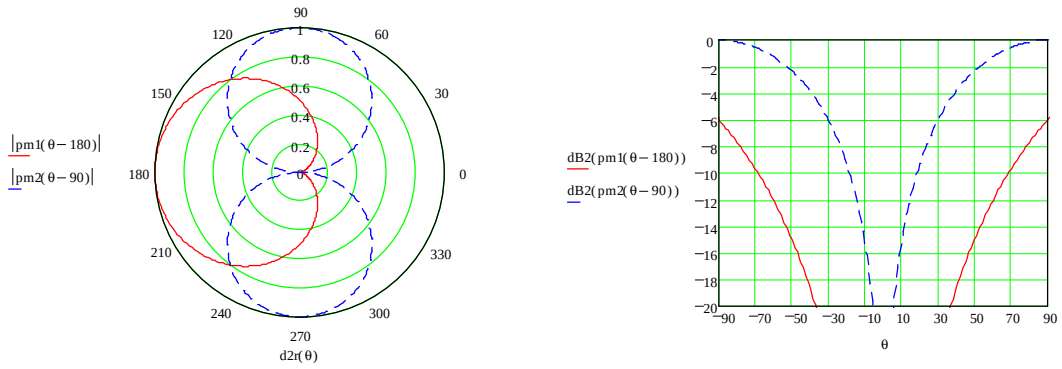

Figure 9. *Polar patterns (left) and attenuation in dB (right) of backwards-facing cardioid (solid line) and sideways-facing figure-8 (dashed line). In the left picture, the stage would be at angle 0°.*

As the graphs show, the region of efficient attenuation is several times broader for the cardioid than the figure-of-8. For example, if the desired angle to be suppressed is 100°, the cardioid will attenuate by 15 dB or more whereas the figure-of-8 will have a lowest attenuation of only 2 dB.

3.2. Microphone distance

The question of the ideal distance between the microphones used for ambience pickup has no universal answer. There are a number of conflicting requirements. First of all, according to research, low (ideally: zero) correlation between the Ls and Rs channels (as well as any other channels reproducing ambience) is favourable since it causes a subjectively higher degree of envelopment [9]. This is important also in the low frequency region. What this means is that the microphones would have to be spaced apart by at least about one wavelength or the diffuse-field distance (several metres in a typical concert hall), whichever is smaller.

However, such wide spacing has a very annoying effect as regards direct sound (typically, audience applause) picked up by the ambience microphones: the sound localizes quite strongly to the Ls and Rs loudspeaker instead of being detached from them and "floating" between and around them. It is thus clear that

the ideal spacing depends on frequency. A narrower spacing would eliminate this “hole in the middle” effect for the applause. This would not occur as easily for low frequencies due to their wavelength being greater in relation to the microphone spacing (which causes a higher correlation).

3.3. Trapezoid arrangement and other tricks

The above idea logically leads to an alternate arrangement of 4 ambience microphones: to achieve both cohesive reproduction of audience noise (e.g. applause) and good envelopment through low correlation all the way down to bass frequencies, it is suggested to have one of the two pairs (whose signals are routed to Ls and Rs) quite narrowly spaced (of the order of 1 m or less) and the other (whose signals are routed to L and R) spaced apart by several metres. It would also be possible to partly mix the left channel of the widely spaced pair into Ls, and the right channel into Rs, and/or to use suitable frequency-dependent weighting (low frequencies taken mostly from the widely spaced pair). If a cohesive reproduction of audience noise is wanted, one might ask why not use another “main pair” for this. Actually, the narrowly spaced pair can be seen as a secondary “main pair”. However, it is not even possible to achieve such accurate localization in the rear sector as has been required above for the front sector (partly due to the properties of human hearing, and partly due to the very large angle between the Ls and Rs loudspeakers). Thus it is enough to just make sure that sounds do not localize too strongly to the Ls and Rs loudspeakers themselves.

To minimize the pickup of direct sound, cardioids should be used throughout (unless something prevents this, e.g. too strong echoes from the back of the hall). Although in principle the wide spacing of the front pair of cardioids would lead to a considerable “hole in the middle” if the sound picked up by it were reproduced on its own, this is only desirable since the more central parts of the frontal soundstage are to be filled in by the solid stereo image generated by the coincident pair.

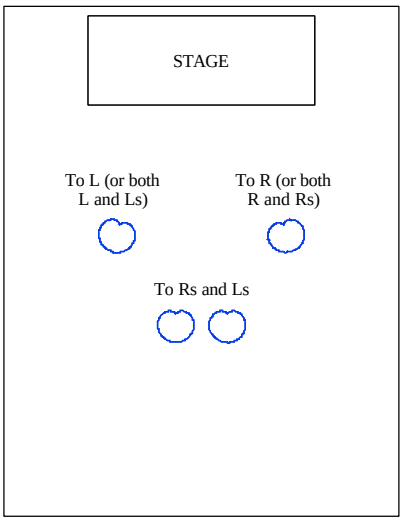

Figure 10. *Possible arrangement of ambience microphones to achieve better cohesion for high-frequency sounds combined with desired low correlation at low frequencies.*

Another valid way of making the reproduction of applause more homogeneous is to create a virtual “rear center” channel by mixing the sum of Ls and Rs equally into both Ls and Rs (with a very moderate gain so as not to make the ambience seem narrower). Yet another way is to generate artificial early reflections from the Ls and Rs channels and feed these early reflections to L and R. This might seem slightly counterintuitive and surprising, but the method works well and it not only makes the ambience more homogeneous but also greatly increases its depth. This is easy to understand since just as artificial rear early reflections can make the frontal

soundstage subjectively "deeper" by simulating lateral reflections in a real hall, frontal early reflections generated from the rear channels can work the same way.

4. PUTTING IT TOGETHER

By now it should be clear that surround recording done using the methods described in this text would be a matter of combining 3-4 different "layers":

1. L, C and R signals from (a) upmixed coincident pair or (b) three coincident microphones, consisting mostly of direct sound and some early reflections,
2. "side ambience", consisting mostly of early lateral reflections and reverberation, from the widely spaced ambience pair, mixed into L, R, Ls and Rs in suitable proportions,
3. "rear ambience", consisting mostly of reverberation and early reflections from the middle and rear parts of the hall, mixed into Ls and Rs
4. optional spot microphones, and flanking microphones for the extreme left and right parts of the ensemble.

Note that if the reproduction of low bass is particularly important, then omnidirectional microphones could be used in the flanking microphone pair or, even better, a fairly narrowly spaced omnidirectional pair could be added to the coincident cardioid setup used as the main microphone. In the latter case the omnidirectional pair should have a smooth roll-off above some 50-100 Hz. A few possible combinations of the "layers" described above are shown here:

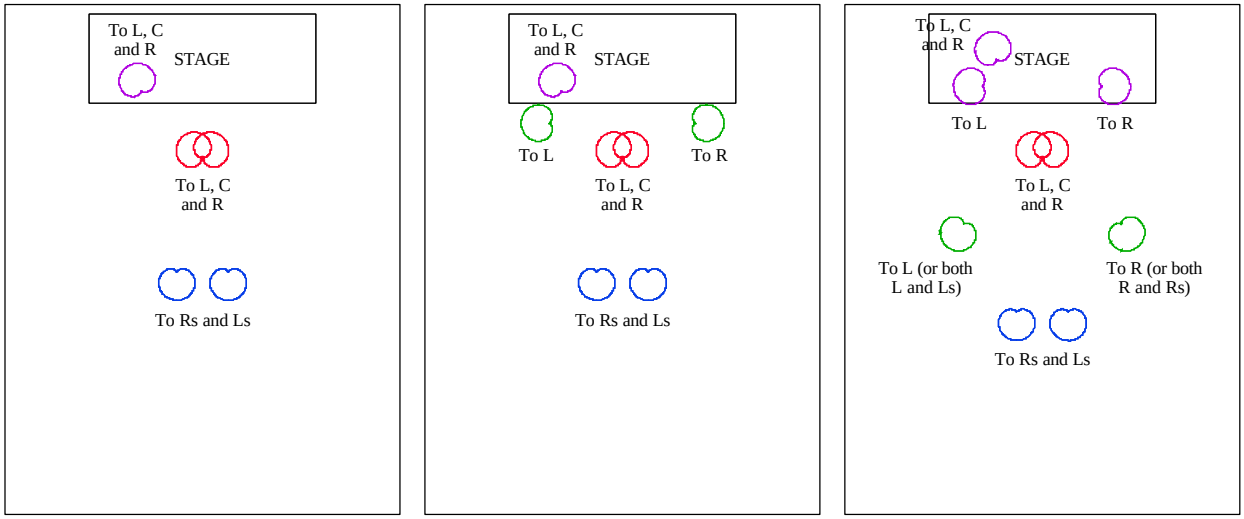

Figure 11. *Possible combinations of microphone "layers" according to the above (left: "absolute minimum" setup (including optional spot microphone), center: combined flanking/"side ambience" microphones added, right: separate flanking microphones and "side ambience" microphones).*

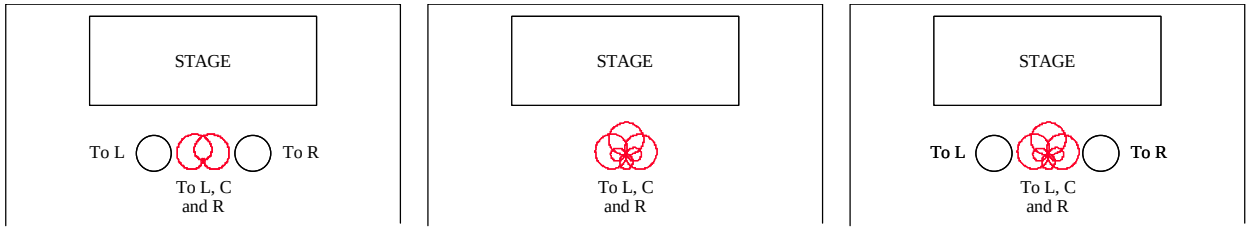

Figure 12. *Alternate implementations of the main microphone setup (left: "small AB" omnidirectional pair added for low frequency enhancement, center: three coincident microphones, right: three coincident microphones with "small AB" pair).*

It is important to note that unless it is acceptable to sacrifice some (or even most of!) the imaging sharpness provided by the coincident main XY pair, the "flanking" microphones have to be mixed in with utmost care. The flanking microphones should pick up as little sound as possible from the more central parts of the stage, which means also that sideways-oriented cardioids (as in fig. 11, right) or supercardioids are the best choice.

How does this "coincident + spaced" setup perform in general? First of all, as is the case with all recording methods relying on "main microphones", it may be difficult to properly record a larger ensemble from such a single spot. Some recording engineers favour instead multiple spot microphones, and some go so far as to get rid of the "main microphones" altogether. In fact, all coincident or near-coincident main microphone techniques fail at some point when the listener moves out of the "sweet spot" (or "sweet area") and the dimensions of the loudspeaker setup is enlarged (also discussed by Griesinger in [9]). This is because of the following simple fact: any localization of phantom sources based on either time delay or amplitude differences (or both) between two or more loudspeakers will break down when the additional delays and/or pressure amplitude differences (caused by the skewed distances to the loudspeakers when the listener is outside the sweet area) are of the same order of magnitude as the delays and/or amplitude differences present in the original signals fed to the loudspeakers. The *only* 5.1 front channel recording method that would consistently work at any possible listener location would be a setup consisting of three widely spaced microphones, whose signals would each be mixed to only one loudspeaker (L, C or R). However, although this is a valid recording method that is favoured by many, it completely lacks the sharp localization between loudspeakers that can be achieved using the coincident methods described earlier in this paper.

The coincident recording method, on the other hand, suffers from a smaller sweet area (which is, however, not nearly as big a problem as it might seem since so much information is present in the ambience reproduced more or less equally by all five loudspeakers). It even so happens that the proper recording of ambience can stabilize the stereo image to the point that a phantom image that would otherwise have wandered to the wrong side of the L-C-R soundstage (as mentioned earlier, phantom sources can "fold over" into the wrong half of the soundstage if the L and R microphones (real or virtual!) have strong rear lobes) can still appear to be roughly at its correct position – or at least the correct half of the soundstage – even though the listener is far outside the sweet area.

In 5.1 surround, just as in stereo, there are several versions of more or less successful compromises between the two extremes of coincident and widely spaced microphones. The 3-channel equivalents of near-coincident 2-channel recording methods (such as ORTF and NOS) are such proposals as INA3 and OCT [10]. These microphone setups will not be further described here, but it suffices to say that they lack some of the localization sharpness provided by the coincident method (especially at high frequencies). However, in return they (especially OCT) can provide more stable reproduction of sound sources at the extreme left or right of the soundstage when the listener is far outside the sweet spot. (The demonstration accompanying this paper will compare the performance of the coincident method(s) with near-coincident methods such as OCT and INA3, and various spaced setups.)

5. CONCLUSIONS

This paper has described how working surround recordings, combining both sharp phantom imaging and depth, can be made using an absolute minimum of 4 microphones, and how additional microphones can be used to advantage. The aspects of coincident vs spaced recording have been discussed, and the proposed method has been put into perspective by comparing its performance with a few other known methods. Surround recording has many degrees of freedom, and most methods are simply various combinations of basic principles known from 2-channel stereo recording. The method proposed in this paper can find use when stereo image sharpness is of primary concern.

6. REFERENCES

[1] Theile, G. *Further developments of loudspeaker stereophony*, preprint 2947, AES 89th convention, Sept. 1990.

[2] Olson, L. T. *The Stereo-180 microphone system*, JAES vol. 27, no. 3, Mar. 1979.

[3] Native Instruments website, http://www.nativeinstruments.de.

[4] Gerzon, M. A. *Optimum reproduction matrices for multispeaker stereo*, JAES vol. 40 no. 7/8, Jul./Aug. 1992.

[5] Schoeps newsletter no. 6, *Surround recording techniques*, http://www.schoeps.de/PDFs/newsl+cmc6xt-E.pdf.

[6] Gerzon, M. A. *Panpot laws for multispeaker stereo*, preprint 3309, AES 92nd convention, Mar. 1992.

[7] McKinnie, D. and Rumsey, F. *Coincident microphone techniques for three channel stereophonic reproduction*, preprint 4429, AES 102nd convention, Mar. 1997.

[8] Gerzon, M. A. *Microphone techniques for 3-channel stereo*, preprint 3450, AES 93rd convention, Oct. 1992.

[9] Griesinger, D. *The psychoacoustics of listening area, depth, and envelopment in surround recordings, and their relationship to microphone technique*, AES 19th international conference.

[10] Theile, G. *Natural 5.1 music recording based on psychoacoustic principles*, AES 19th international conference.

[11] Gerzon, M. A. *General metatheory of auditory localisation*, preprint 3306, AES 92nd convention, Mar. 1992.