THE PSYCHOACOUSTICS OF MULTICHANNEL AUDIO

J. ROBERT STUART

Meridian Audio Ltd
Stonehill, Huntingdon, PE18 6ED
England

ABSTRACT

This is a tutorial paper giving an introduction to the perception of multichannel sound reproduction. The important underlying psychoacoustic phenomena are reviewed – starting with the behaviour of the auditory periphery and moving on through binaural perception, central binaural phenomena and cognition.

The author highlights the way the perception of a recording can be changed according to the number of replay channels used. The paper opens the question of relating perceptual and cognitive responses to directional sound or to sound fields.

1. INTRODUCTION

Multichannel systems are normally intended to present a three-dimensional sound to the listener. In general, the more loudspeakers that can be applied, the more accurately the sound field can be reproduced. Since all multichannel systems do not have a 1:1 relationship between transmitted channels and loudspeaker feeds, a deep understanding of the human binaural system is necessary to avoid spatial, loudness or timbral discrepancies.

This paper reviews some of the basic psychoacoustic mechanisms that are relevant to this topic.

2. PERCEPTION

For the purpose of this paper, we are going to break the listening process down, following the signal path into the following bottom-up hierarchy.

- Auditory periphery: taking each ear as an independent device.
- Binaural perception: seeing how the basic behaviour is modified by two-eared listening.
- Spatial perception: reviewing the low-level inter-aural interactions that give instinctive spatial perception.
- Cognition: how the useful percept depends on spatial and multichannel factors.

3. PERIPHERAL AUDITORY FUNCTION

Sounds are encoded in the auditory periphery on a loudness-pitch basis.

3.1 Pitch

The cochlea aids frequency selectivity by dispersing excitation on the basilar membrane on a frequency-dependent basis. Exciting frequencies are mapped on a pitch scale roughly according to the dispersion (position) and the integral of auditory filtering function. Several scales of pitch have been used; the most common being *mel*, *bark* and *E.*. Fig. 1 shows the inter-relationship between these scales.

The mel scale derives from subjective pitch testing and was defined so that 1000mel ≡ 1kHz. The other scales are derived from measures of the auditory filter shape. Fig. 4 shows the relationship between the now dominant measure (E) and frequency (Hz). The E scale plays an important role in understanding frequency-dependent perceptual phenomena, including selectivity, masking and loudness.

The frequency-selectivity of the periphery can be determined in psychoacoustic tests. The selectivity varies with frequency and intensity. Fig. 2 shows the frequency dependence of the Equivalent Rectangular Bandwidth (Erb) at different applied intensities.

Fig. 3 shows the selectivity – or frequency shape – of the auditory filter at 1kHz. It can be seen that as the applied intensity increases, the filter broadens.

This is thought to be due to the equivalent of agc effects combined with an active process that is effective at low intensities – near threshold. Obviously the auditory selection bandwidth is a compromise between time and frequency discrimination.

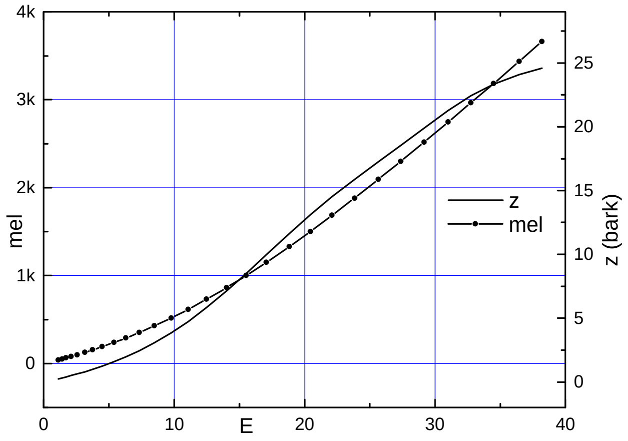

Figure 1 Sowing the inter-relationship between the three pitch scales – E, mel, bark.

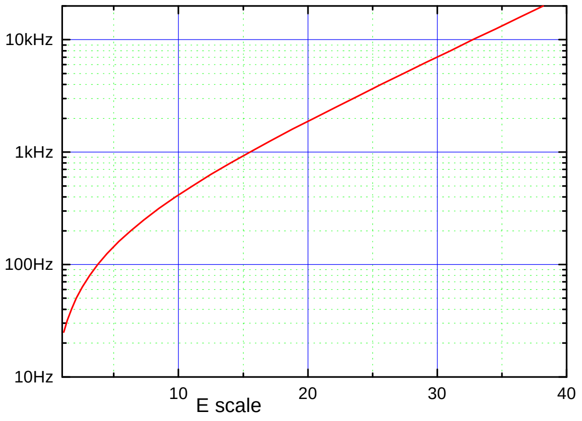

Figure 4 Showing the relationship between the E scale and frequency.

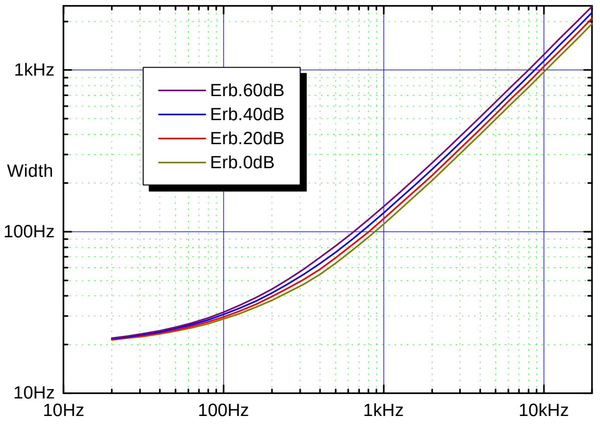

Figure 2 Showing the way the ERB noise-bandwidth varies with frequency and level. The bandwidth is plotted for applied intensities of 60, 40, 20 and 0dB spl.

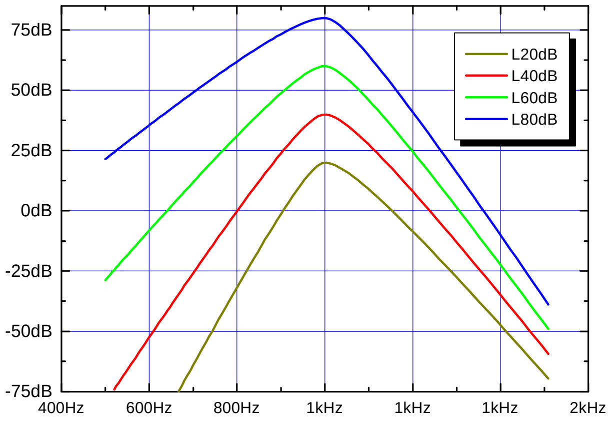

Figure 3 Showing peripheral selectivity at 1kHz and for 20, 40, 60 and 80dB spl.

Obviously this frequency selectivity describes our fundamental ability to discriminate sounds in the frequency domain. It also defines – through the way excitation spreads to adjacent frequencies – the way in which one sound may mask another. When the excitation region of two stimuli overlap, each is masked by the other and the total loudness is less than the sum of the loudness of each taken alone.

3.2 Threshold

The auditory periphery also exhibits a sensitivity that varies with frequency. Two commonly referenced curves are shown in Fig. 5; Minimum Audible Field (MAF) and Minimum Audible Pressure (MAP).

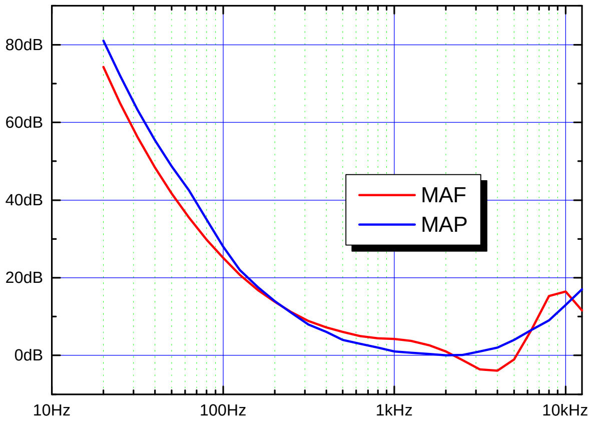

Figure 5 Showing the two hearing threshold curves; Minimum Audible Field (MAF) and Minimum Audible Pressure (MAP).

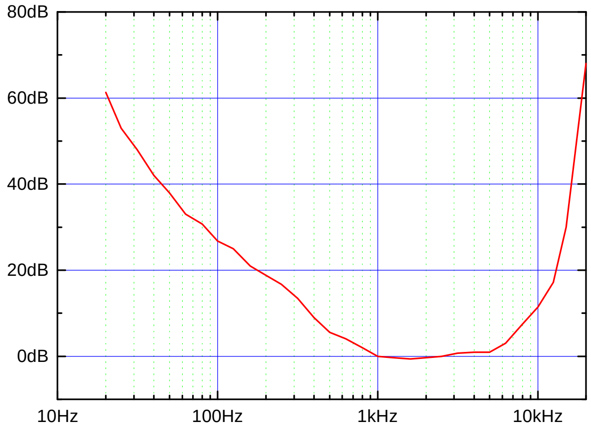

Figure 6 Showing the equivalent internal auditory system noise which is partially responsible for the MAP threshold.

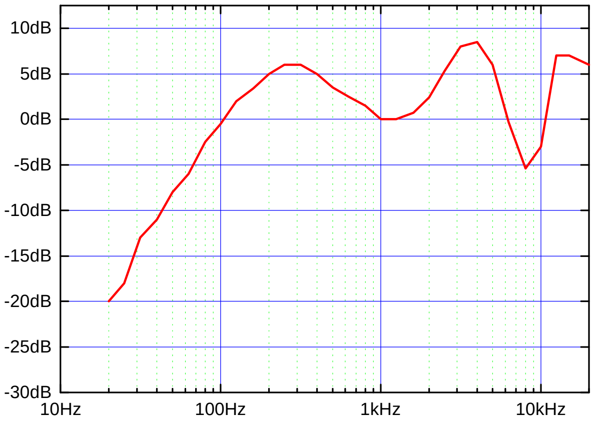

Figure 7 Showing the form of the average diffuse-field frequency response effects of the external auditory system.

Minimum Audible Pressure (MAP) refers to sounds applied to the ear-canal – for example by headphones – in which the outer ear mechanisms of head-diffraction and pinna effects are negated.

Minimum Audible Field (MAF), refers to sounds presented to the listener in a diffuse external field. In other words it combines the diffuse-field external auditory frequency response – shown in Fig. 7 – with the MAP threshold.

It is thought that the general shape of the MAP threshold is not exclusively to do with transmission efficiency (i.e. mechanical response). Rather, the shape of the threshold also reflects internal noise which masks signals according to the indication of Fig. 6. Note however, that MAF indicates a higher sensitivity at low frequencies and this is thought to be due to an effective increase in internal noise when the ear-canal is excluded – as it is when wearing headphones.

3.3 Loudness

The second important parameter encoded in the auditory periphery is loudness. Loudness is a subjective measure normally expressed in sones, where one sone is the loudness of a 1kHz tone presented at 40dBspl. Fig. 8 shows the growth of loudness in sone for a pure 1kHz tone and for a wide-band white noise as a function of intensity.

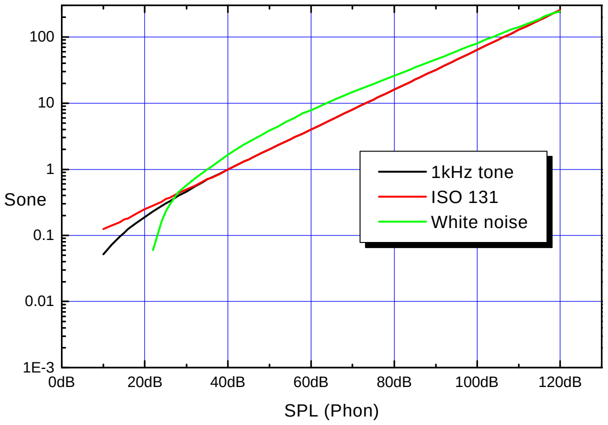

Figure 8 Showing the relationship between the loudness (in Sone) of a 1kHz tone or a white noise and the applied intensity in spl. The straight-line labelled ISO 131 shows the standardised definition.

It can be seen that, above 30dB spl, the loudness grows as a power law of intensity and reasonably uniformly. The noise behaves differently due to the shape of the auditory threshold, auditory filtering and non-linear effects.

It was mentioned earlier that when two stimuli are applied there can be a degree of mutual masking if there is a region of filter overlap. Fig. 9 shows how the loudness progression of a 1kHz tone varies when it is masked by a broadband white noise of the intensity shown.

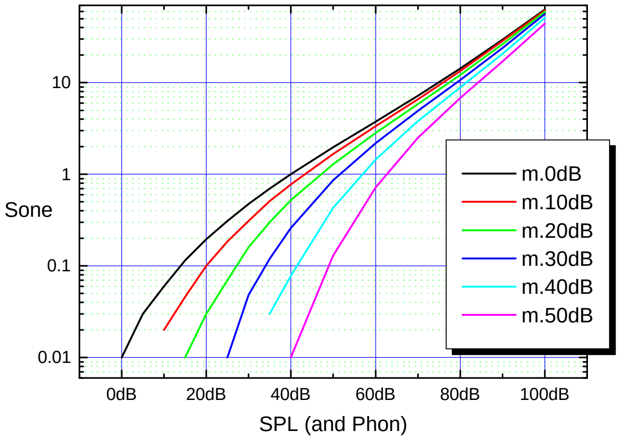

Figure 9 Showing how the loudness of a 1kHz tone is effected by the presence of a broad-band white-noise masker. The white noise intensity is varied between 0dB and 50dBspl.

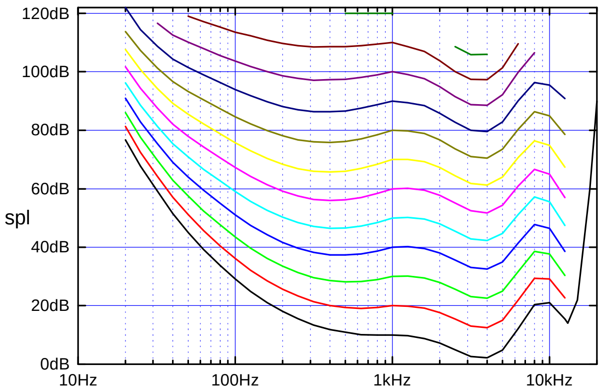

Figure 10 Showing the equal-loudness contours for diffuse-field presentation from ISO 226.

Fig. 9 illustrates some useful points. It is significant that a fixed intensity 1kHz tone appears to get quieter as the white-noise masker is increased in level. Also, it will be seen that for significantly-masked sounds, the growth of loudness with intensity is very rapid. A circumstance where loudness grows rapidly with intensity indicates a masking phenomenon.

Fig. 10 shows the familiar data of equal-loudness contours from ISO 226. It can be seen that the low frequency threshold is combined with rapid loudness growth – substantiating the assertion that an internal noise like that of Fig. 6 is partially responsible for the threshold.

3.4 Temporal encoding

The neural code from the auditory periphery partially represents specific loudness – that is a two-dimensional representation of loudness vs. pitch. Real-world sounds are rarely as uniform as the simple objective stimuli of tone and noise in the preceding examples, and indeed contain important cues in their time-structure.

Sounds are also encoded through

- onset and offset (overall envelope and transients)
- synchronously for waveforms or envelopes < 800Hz
- loudness dependency through temporal pre- and post-masking effects

4. PERIPHERAL BINAURAL AUDITORY FUNCTION

The previous section reviewed the important parameters of the auditory periphery from the psycho-acoustics of a single ear.

Multichannel sound reproduction is naturally about spatial aspects of sound – or stereo[1] – and so this section looks at the relevant aspects of two-eared listening.

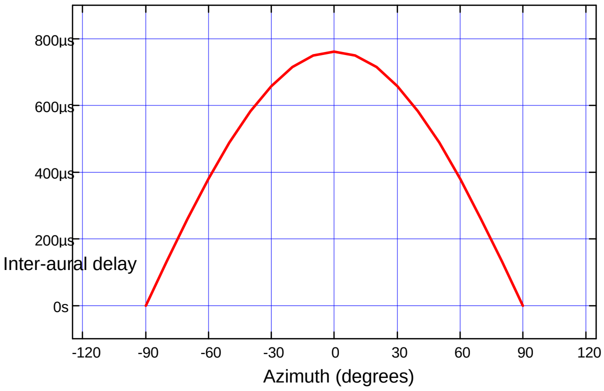

Figure 11 Showing the inter-aural time-delay for a point-source signal at different azimuths. In this diagram zero azimuth is fully to one side.

4.1 Head and Pinna effects

Humans listen with two ears. Two spaced ears give a mean time arrival difference for sounds in different locations of up to 0.7ms – and intensity difference due to head 'shadow'. These basic phenomena are at the root of the mechanisms that allow us to determine the direction of an external sound.

The time-delay difference due to path-length and diffraction effects is shown in Fig. 11.

Pinna effects also make important spectral modification according to angle of incidence and this filtering action combined with head diffraction is used to encode direction.

Fig. 12 shows an example of measurements giving the frequency response variation for single-tone point-sources at different azimuths for the near ear and in the horizontal plane. There are a number of features in the response that vary considerably with azimuth; note especially, the sharp notches that vary position with azimuth and probably provide a significant cue to position.

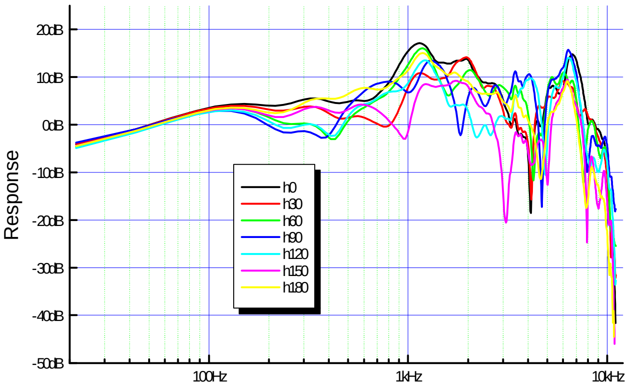

Figure 12 Showing the variation in frequency response for single tones, measured in the ear-canal. The responses shown cover from full ahead (azimuth 0°) to full behind and are for the near ear. The responses for the shadowed ear are obviously different again.

[1] *Stereo* (from Greek), means 'solid'. Current abuse of the term takes it to mean 'two-channel'. This is not the case, stereo i.e. solid sound, can be conveyed or reproduced by many loudspeakers.

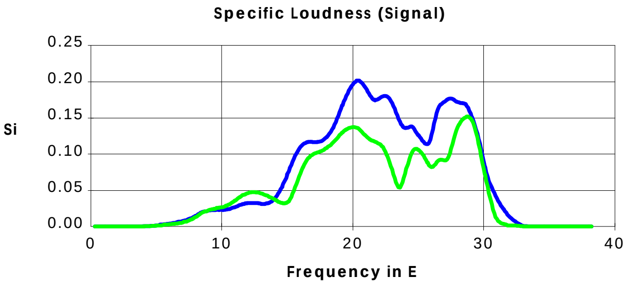

Figure 13 Showing the difference in internal loudness representation of a white noise source at 30° (upper) and 150° (lower). The graph plots specific loudness against the E frequency scale.

Fig. 12 illustrates the response variations for pure tones; most real-world sounds are more complex and so important cues can be obtained in the way the harmonic (or multi-frequency) content of the sound of an object changes with azimuth – either with object or head movement.

Fig. 13 is a representation from auditory modelling of the peripheral excitation (basilar membrane) resulting from an external white-noise source. The graph plots specific loudness on the E frequency scale.

4.2 Masking effects

With one-eared listening, the masking provided by a masker can be readily determined by experiment. Generally speaking, except for stimuli with particular envelopes, the masking can be predicted from the spread of excitation each component produces on the basilar membrane. Fig. 14 illustrates the basic concept of masking in a multi-tone stimulus (in this case a violin note). The hearing threshold is modified by the stimulus, and some components of the original sound are effectively masked[2].

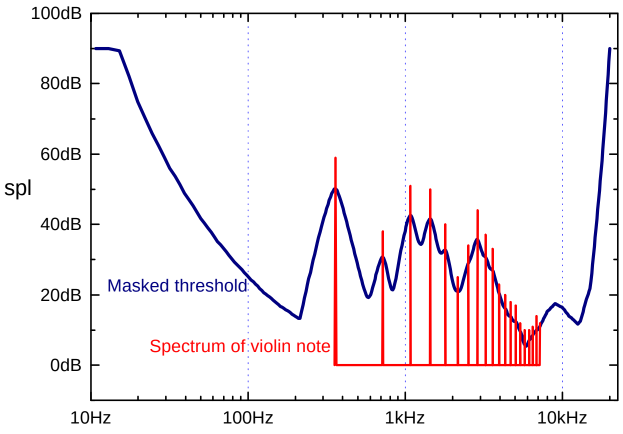

Figure 14 Showing the monaural masked threshold for a multi-tone stimulus (in this case a bowed violin note).

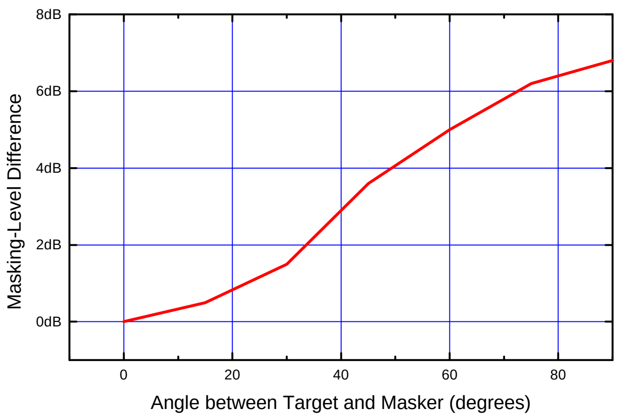

Figure 15 Showing the form of masking difference according to the angular spacing between masker and target.

As Fig 14 shows, the masked threshold for each component is dependent on the position of the probe frequency with respect to the masker. For two-eared listening, the masked threshold also varies with position. Sounds are more effectively masked when the masker and target have the same location.

Fig. 15 shows the way in which the masked threshold produced by a white noise varies as the angle between the target and masker is changed. Overall, by placing sounds in different locations, the degree of masking can be reduced by up to 7dB. This difference is very important in multichannel systems for several reasons.

1. The design of multichannel lossy compression systems needs to account for the reduced masking available for spaced sounds.
2. Matrix decoders or spatial synthesis schemes may reveal components in lossy-compressed materials that were not intended to be heard.
3. On a more positive note, if multichannel systems can spatially separate sounds, then they can be clearer or more individual to the benefit of realism. It should be obvious that the fewer loudspeakers used to render a performance, the more components of that performance will mask each other.

4.3 Localisation: Temporal cues

The previous sections reviewed the mechanisms by which amplitude differences could provide cues to the location of an external acoustic object.

Another important source of information on externalisation is in the time-structure of arriving sounds, and the relevant parameters are:

- onset and offset (overall envelope and transients)
- synchronously for waveforms or envelopes < 800Hz

So, in addition to intensity cues, data arises in time and phase differences between the signals from both ears.

[2] This mechanism is exploited in the design of lossy perceptual coders.

It is an important requirement for a natural-sounding multichannel system that these different mechanisms are exploited in a co-ordinated way. Listener fatigue or confusion rapidly occurs when the location cues are contradictory.

4.4 Localisation: Precedence effects

It is well known that sounds often appear to come from the direction of first arrival, somewhat independently of amplitude. This is entirely reasonable – especially since most naturally occurring sonic events will tend to make the first-arriving sound also the loudest.

There is a trade-off between time-arrival difference and loudness effects.

4.5 Localisation: Sound-field effects

The normal two-eared listener will make head movements. Apart from small movements, which can rapidly aid the confirmation of a direction hypothesis, by far the most powerful direction-determining behaviour is to turn to face the direction of the apparent sound.

Normally, an external sound will grab attention and the combination of cues from time-arrival and spectral changes, set up an initial listener-hypothesis of its location. If the sound continues, the listener can get a very accurate 'fix' by turning the head to set up a similar sound in both ears. When the sound source is dead-ahead, each ear produces a similar response and the listener is facing perpendicular to the wavefront.

Some stereo and pseudo-stereo systems do not achieve good agreement between the first hypothesis and the net wavefront. In particular, some methods of spatial encoding rely on equalisation to fool the pinna and head effects and may even require the listener to remain fixed – thereby introducing a significant 'unreal' quality to the percept.

Sound-field replay methods look at the apparent direction of a source *in the absence of a listener*. Localisation can be confirmed by moving around or head-turning.

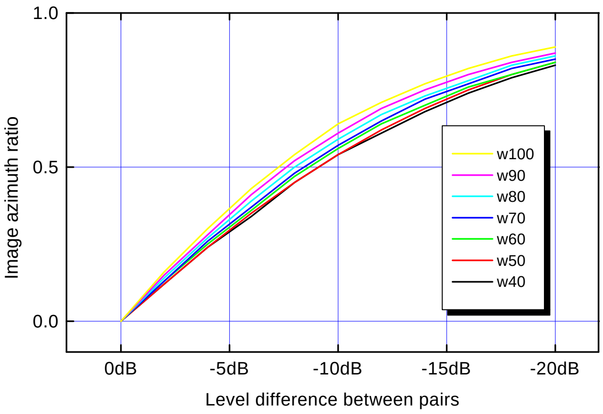

Figure 16 Showing the apparent wavefront in intensity stereo. Two speakers subtend angles between 40° and 100°. The apparent position is the azimuth ratio, where 0 is mid-way and 1 is in line with the louder speaker.

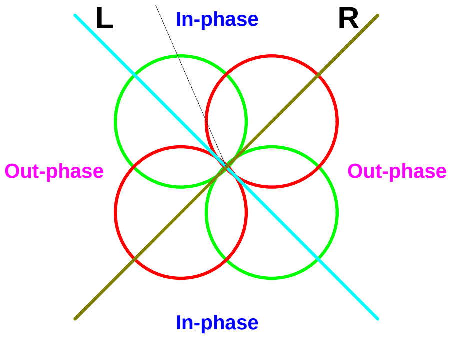

Figure 17 Showing the polar diagram of a common stereo microphone – the crossed-pair of velocity capsules. The left and right polar diagrams are sinusoidal.

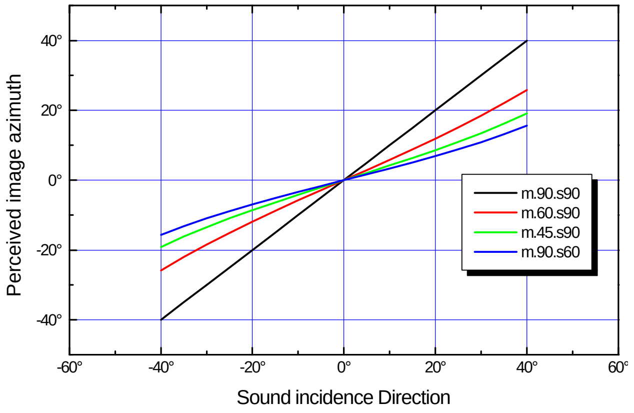

Figure 18 Showing how the azimuth position of a source sampled by a microphone with the polar-response shown in Fig. 17, is represented when replayed over two loudspeakers. The parameters are the angle between the microphones and between the speakers (from the listeners perspective).

Fig. 16 shows how the apparent wavefront direction can be imputed for intensity stereo. Two loudspeakers present the same signal at different amplitudes; the two signal vectors combine to produce a wavefront whose apparent direction places the image between the loudspeakers. Fig. 18 shows the way azimuth can be mapped from an angular position with respect to a crossed figure-of-eight microphone (see Fig. 17), to an apparent position between two loudspeakers.

5. Central Binaural Processing

The central binaural processor is extremely sophisticated. By combining the signals from two ears, many of the thresholds seen in one-eared listening become significantly modified. In almost all cases the binaural listener is more acute.

For example binaural temporal acuity is significantly higher than in the monaural case. Arrival-time differences of the order of 30us at 50 phon can be perceived.

5.1 Binaural thresholds

In binaural listening there are also significantly modified detectability thresholds due to binaural interaction. Some examples include:

- lower hearing threshold with two ears
- sub mono-threshold interpolation
- binaural masking and release
- binaural masking-level differences (≅ 12dB)
- binaural beats (interaction between separate sounds in each ear)
- subliminal perception: (see e.g. Groen)

Each of the mechanisms listed is a full subject – sufficient for many papers – the interested reader should consult the reading list at the end of this paper.

5.2 Binaural post-processing

The binaural perception process also significantly modifies the perceived sound. For example, external sounds may suffer comb-filtering, yet the binaural processor removes this effect.

This could be better explained with reference to the changing amplitude-with-azimuth data shown in Figs. 12 and 13. It is a remarkable feature of the binaural processor (and cognition) that the marked difference in internal excitation seen in Fig. 13 can be used to determine the location of the sound; yet, were the source to move between the two positions, the percept would be of continuity – to the extent that the timbre of the noise would not change.

The perceptual process at this point begins to separate the timbre of the sound according to the hypothesis on direction and range.

This raises another important issue in multispeaker replay: there will inevitably be a timbre mismatch between phantom sources and 'hard' loudspeaker sources. Fig. 19 shows the correction one should apply to a centre speaker which is used to contribute to a normally phantom central image.

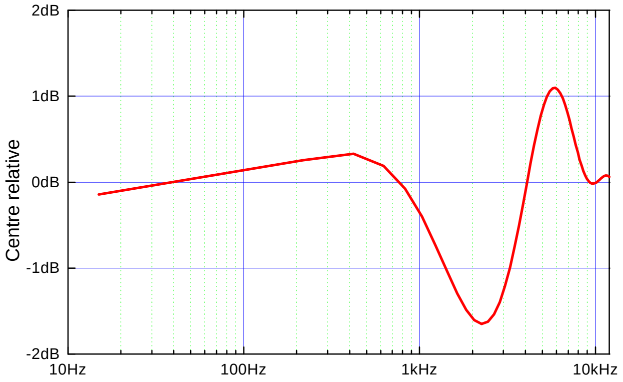

Figure 19 Showing the form of timbre correction to apply to a centre speaker reproducing a normally-phantom source.

5.3 Binaural Loudness

Loudness for binaurally presented sound is not simply related to the mono equivalent. Lateral inhibition causes the loudness in each ear to grow 'as masked', and as sounds are located in space, the stimulus magnitude will be interpreted in the manner illustrated in the previous section.

Binaural listening also changes the form of the loudness function. Switching from presenting a sound to one ear (mono), to binaural presentation results in:

- near threshold: an approximate doubling in Sone i.e. 10 Phon,
- mid-loudness (say around 50 Phon): we see a 4 Phon increase,
- at high level (say 80 Phon): a 3 Phon increase.

An important observation is that if multichannel reproduction succeeds in exploiting direction cues to give a better (wider) physical separation of sounds – then not only will those sounds be more separated (less masking), but the loudness balance between the sounds will be different. Assuming successful design of the encode/decode, the possibility exists for sounds to be separated naturally.

6. COGNITION

6.1 Perception of objects

The perception of music or speech in surround depends on our ability to 'externalise' perceptions into 'acoustic objects'.

We do not hear tones and noise. Rather, the arriving sound elements are separated into various hypotheses of real sources: head-turning or continuity in the evolution of the sound will then confirm or deny the hypothesis.

Without direct visual cues, instruments will stream into a number of separated items; with more or less success depending on the quality and design of the system.

The process by which a percept is resolved as a real external acoustic object is known as 'cognition'.

Some factors that effect the grouping of components feeding this 'cognitive' process are their:

- amplitude
- fundamental frequency
- timbre
- envelope patterns
- onset disparities
- correlated changes
- contrast with earlier and later sounds
- spatial location.

So, initially a hypothesis is formed about probable external acoustic sources based on the components of the arriving sound.

Internal contributions to the cognition process seem to use an iterative process based on the external hypothesis. Other perceptual attributes of acoustic object formation may be:

- constancy/correlated changes
- similarity/ contrast
- auditory streaming
- continuation
- common fate
- onset/offset disparities
- timbre/envelope correlation
- language
- rhythm
- closure (replacement of missing sounds)
- attention.

6.2 Cognitive elements in sound

Regarding general object cognition, the following elements contribute to the overall process:

- Monaural elements of sound: pitch, loudness, timbre; auditory object formation; 'object' grouping.
- Binaural additions: auditory object location and separation, 'object' externalisation.
- Spatial characteristics: spaciousness, ambience recognition, distance perception.

6.3 Cognitive elements of Music

Multichannel sound systems are normally aimed at reproducing music or speech performances.

For speech, the cognitive process obviously involves many complex interactions, as cues from the loudspeakers confirm or deny hypotheses about persons in the surrounding acoustic space. Language plays a very important part in differentiating sounds.

So far as music is concerned, there are a number of additional levels of cognition including:

- cognition of the 'sound object' itself
- cognition of the music
- cognition of the music's structure
- cognition of the content, or meaning of the music.

Obviously, music normally combines elements of theme, melody, harmony, rhythm. It also arises from instruments, whose segregation in the listening process may rely on very small cues.

Continuity applies, in that it is not normal experience for instruments to change character or position suddenly; although in the music flow – on a context-dependent basis – the instruments may 'come and go' i.e. start and stop playing.

6.4 Multichannel object separation

The binaural cognitive process allows the listener to separate sounds in the environment and from each other. In many circumstances, each object component will be presented in very poor signal/noise conditions, and subtle cues radically alter the perception.

For this reason, the benefits brought to sound reproduction by moving from the essentially 2-D presentation of mono or stereo to the 3-D of multichannel are highly significant. Not only is spatial separation important in object formation, but by presenting different wavefront options, the generally lower masking allows clearer segregation.

Fig. 20 illustrates a hypothetical cognitive process as the notional signal/noise ratio is changed from –20dB to +20dB on a piano stream.

Compared with two-speaker stereo, multichannel brings:

- easier and more emphasised auditory object externalisation
- simpler instrument streaming
- changed loudness balance through the binaural process
- changed timbre perception through location-correction
- markedly different ambient perception
- increased speaker directivity
- Increased acuity for channel or processing errors

7. SUMMARY

This paper has examined the perceptual and cognitive processes in a 'bottom-up' hierarchy starting with the auditory periphery.

Although it is common to consider that we hear the externally-applied noises in a passive way, this paper has taken pains to illustrate that this is in fact a poor model

Rather, the cognition of the material routinely transmitted on multichannel systems, relies on the presentation of *cues* in the auditory space. These cues are interpreted by the listener, using a considerable amount of internal learned data, as an overall collection of external objects from which streams of content arise.

So, the design of multichannel systems requires a good understanding of both perception and cognition.

In general, the important target for the designer of multichannel systems, is to achieve stability and continuity. The overall percept will not be realistic if:

- the sound space appears to move, or
- contradictory binaural cues result from the encode/decode process, or
- head-turning does not tend to confirm the location of sound objects.

For the interested student, a list of reading is appended.

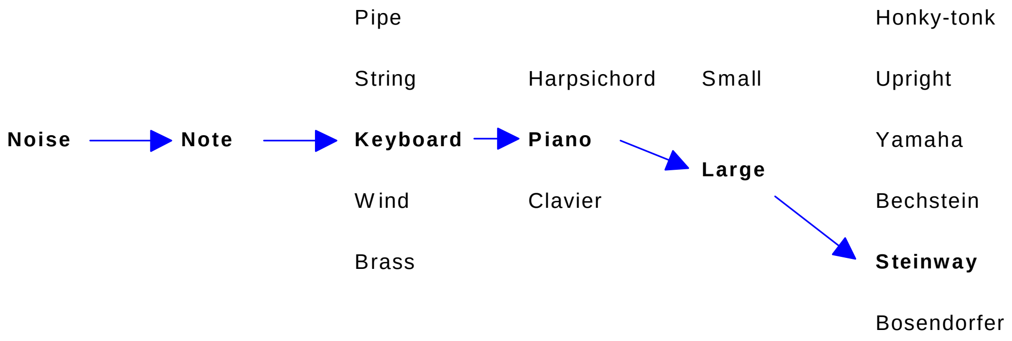

Figure 20 Giving an illustrative example of the change in cognition of a piano stream as the signal/noise ratio moves from –20dB to +20dB (left to right).

8. Further Reading

Bibliography

1 Blauert, J. *Spatial Hearing* (MIT Press, 1983)

2 Bregman *Auditory Scene Analysis*

3 Carterrette, E.P. and Friedman, M.C. *Handbook of Perception*, **IV**, 'Hearing' (Academic Press, 1978)

4 Deutsch, D. *The Psychology of Music* (Academic Press, 1982)

5 Moore, B.C.J. *An Introduction to the Psychology of Hearing* (Academic Press, 1991)

6 Tobias, J.V. *Foundations of Modern Auditory* Theory (Academic Press, 1970)

Perception

7 Buus, S. 'Release from masking caused by envelope fluctuations' *J. Acoust. Soc. Amer.*, **78**, 1958–1965 (1985)

8 Groen, J J. 'Super and subliminal binaural beats' *Acta Oto-Lar*, **57**, p224

9 Hall, J.W. 'Experiments on Comodulation Masking Release', in *Auditory processing of complex sounds*, Eds Yost, W.A. and Watson, C.S., Erlbaum and Assoc. (1987)

10 Irwin, R.J. 'Binaural summation of thermal noises of equal and unequal power in each ear' *American Journal of Psychology*, **78**, 57–65 (1965)

11 Lochner, J.P.A. and Burger, J.F. 'Form of the loudness function in the presence of masking noise' *J. Acoust. Soc. Amer.*, **33**, 1705–1708 (1961)

12 Scharf, B. 'Loudness summation between tones from two loudspeakers' *J. Acoust. Soc. Amer.*, **56**, 589–593 (1974)

13 Scharf, B. and Fishken, D. 'Binaural summation of loudness' *J. Exp. Psychology*, **86**, 374–379 (1970)

14 Stuart, J.R. 'Predicting the audibility, detectability and loudness of errors in audio systems' *AES 91st convention*, New York, preprint 3209 (1991)

15 Stuart, J.R. 'Estimating the significance of errors in audio systems' *AES 91st convention*, New York, preprint 3208 (1991)

16 Stuart, J.R. 'Psychoacoustic models for evaluating errors in audio systems' *PIA*, **13**, part 7, 11–33 (1991)

17 Yost, W.A. and Watson, C.S., (eds) of *'Auditory processing of complex sounds'*, Eds Erlbaum and Assoc., section VI (1987)

Cognition

18 Deutsch, D. 'The octave illusion and auditory perceptual integration' in *Hearing Research and Theory*, Eds Tobias, J.V. and Schubert, E.D., 99–142 (Academic Press 1981)

19 Terhardt, E., 'Music perception and sensory information acquisition: relationships and low-level analogies', *Music Perception* **8** no 3, 217-239, (Spring 1991)

20 Umemoto, T. 'The Psychological Structure of Music' *Music perception* **8** No 2, 115–128 (Winter 1990)

Surround sound

21 Acoustic Renaissance for Audio, Technical Subcommittee. 'A Proposal for the High-Quality Audio Application of High-Density CD Carriers' *Privately published document*, (1995)

22 Perrott, D. R. 'Auditory and Visual Localisation: Two modalities One world', *Proceedings of AES 12th International Conference "The Perception of Reproduced Sound"*, 221–231 (June 1993)

23 Schroeder, M. R. 'Listening with Two Ears' *Music perception* **10** No 3, 255–280 (Spring 1993)

24 Snow, W.B. 'Basic principles of Stereophonic Sound' *Journal of SMPTE*, **61** 567–589 (1953)

25 Steinberg, J.C. and Snow, W.B. 'Physical factors in Auditory Perspective' *Journal of SMPTE*, **61** 420–430 (1953)