A BALANCED STEREO WIDENING NETWORK FOR HEADPHONES

OLE KIRKEBY

Speech and Audio Systems Laboratory, Nokia Research Center, Helsinki, Finland
ole.kirkeby@nokia.com

The purpose of a stereo widening network for headphones is to compensate for the non-ideal listening conditions experienced when listening to music mixed for playback over two widely spaced loudspeakers. Our network is balanced in the sense that there is a constraint on the sum of the magnitude responses of the cross-talk (left input to right output, and right input to left output) and the direct paths (left input to left output, and right input to right output). In order to ensure that only a minimum of spectral colouration is added to the original sound only frequencies below 2kHz are processed. Consequently, a very natural characteristic of the reproduced sound is maintained, and it makes the stereo widening algorithm well suited for applications to high-quality digital source material. A structurally balanced implementation makes it possible to run the network very efficiently in fixed-point precision.

INTRODUCTION

The music that has been recorded over the last four decades is almost exclusively made in the two-channel stereo format which consists of two independent tracks, commonly denoted left (L) and right (R). The two tracks are intended for playback over two loudspeakers, and they are mixed to provide a desired spatial impression to a listener positioned centrally in front of two loudspeakers that ideally span 60 degrees. Portable devices use headphones rather than loudspeakers for the reproduction. Headphone reproduction potentially offers very good control of what the listener hears. When music is played back over loudspeakers, the perceived sound depends to a large extent on the characteristics of the acoustical environment, and on the position of the listener in that environment. The acoustics of a car cabin is very different from a living room, for example, and the listener’s position relative to the loudspeakers is also different in the two situations.

It has long been known that headphone listening can cause in-head localisation [1] and fatigue [2]. The exact reasons why this happens are still not clear but it is widely accepted that the auditory system struggles with inputs that do not contain the interaural time- and level differences generated by physical sound sources [3], [4]. Thus, a processing scheme for converting material mixed for playback over loudspeakers to a format suitable for headphone listening ought to insert some binaural cues [5]. However, modification of high-fidelity music recordings must be done with great care. There is evidence to suggest that certain types of post-processing for the purpose of spatial enhancement has a detrimental effect on the sound quality [6]. The advantages of a spatial enhancer must be weighed against the blurring and spectral colouration it adds to a high-fidelity music recording. See [6] for the results of a listening test performed to assess the performance of a number of spatial enhancers, including the one described in the following, with high-fidelity music recordings.

1 STEREO WIDENING FOR HEADPHONES

The block diagram in Figure 1 shows a common way to implement spatial enhancement of two-channel stereo material. The direct paths are implemented by two identical filters \( H_d \), and the cross-talk paths are implemented by two identical filters \( H_x \). This kind of symmetric structure is very general, and many technical papers and patents are based on it even if they are not presented explicitly in this form. It is the derivation and the properties of \( H_d \) and \( H_x \) that are different.

![Figure 1: An implementation of a spatial processor.](image)

Weak decorrelation [7], [8] of the source signals guarantees that the two signals that are input to the widening network always differ to some extent even if the two signals from the digital source are identical. The spatial effect of decorrelation is that the component common to both L and R is not heard as being localised
at a single point, but rather it is spread out slightly so that it is perceived as having a finite size. This helps to give the impression that the central image is produced by a physical sound source, and it also ensures that the central image contains some binaural cues after being passed through the widening network. In addition, it reduces the attenuation of the monophonic component caused by the interference between the direct path and the cross-talk path.

Equalization, such as a bass boost that can compensate for a poor transducer response, is conveniently implemented by post-processing so that it does not affect the dynamic properties of the stereo widening network itself.

2 FILTER DESIGN AND IMPLEMENTATION

2.1 Stereo played back over loudspeakers

Fig. 2 illustrates what a listener hears when positioned centrally in front of two loudspeakers.

![Diagram showing four paths from two loudspeakers to two ears](image)

Sound coming from the left loudspeaker is heard at both ears, and, similarly, sound coming from the right loudspeaker is also heard at both ears. Consequently, there are four paths from the two loudspeakers to the two ears. When the loudspeakers are positioned symmetrically with respect to the listener the direct path from the left speaker to the left ear is the same as the direct path from the right speaker to the right ear, and, similarly, the cross-talk from left speaker to the right ear is the same as the cross-talk from the right speaker to the left ear. As in Figure 1 we denote the direct path by subscript d and the cross-talk path by subscript x. The direct path and the cross-talk path each has a frequency-dependent gain, $G_d$ and $G_x$ respectively, and a frequency-dependent delay, $t$ and $t+\text{itd}$ respectively, associated with it. The difference between the two delays is the interaural time difference $\text{itd}$. We can derive approximate values for $G_d$ and $G_x$ by considering the physics of sound propagation. It is well known that when an object, such as the head of a human listener, is positioned in an incident sound field, such as that produced by a loudspeaker, the sound field is not disturbed when the wavelength is long compared to the size of the object [9]. Given the size of the human head, this means that $G_d$ and $G_x$ are constant and substantially equal below 1kHz. At high frequencies, where the wavelength is short compared to the size of the object, there is a pressure build-up on the object’s near side, and a pressure attenuation, also referred to as shadowing, on the object’s far side. If the object has a relatively simple shape, so that it does not focus the sound, and furthermore it is rigid, there will be a pressure doubling on it’s near side at very high frequencies, and no sound will reach the shadow zone. Based on this, we can set $G_d$ and $G_x$ to one at frequencies below 1kHz, and $G_d$ to two and $G_x$ to zero at high frequencies. The interaural time difference $\text{itd}$ is strictly speaking also dependent on the frequency but we will assume that it is constant. It is widely accepted that the value of $\text{itd}$ is the most important cue for determining the location of a source in the horizontal plane [10]. Thus, for frontal sources $\text{itd}$ is zero whereas $\text{itd}$ is about 0.7ms for sources directly to the side of the listener.

2.2 Binaural synthesis

If sound is emitted from a loudspeaker that is placed to one side of the listener, say straight to the left, it is possible to measure the sound pressures it produces at the two ears of the listener. The same sound pressures can be reproduced at the listener’s ears through headphones by processing the signal that is input to the loudspeaker with a pair of digital filters. Ideally, what the listener then hears over the headphones is the sound of a loudspeaker positioned straight to the left. This approach is called binaural synthesis [11], and it is a scientifically credible way to emulate natural listening conditions. When it is applied separately to a stereo recording’s left and right channel, the listener should hear the equivalent of two widely spaced loudspeakers. However, an attempt to model natural listening with very good accuracy introduces noticeable spectral colouration of the reproduced sound, particularly at frequencies above 3kHz, and this colouration is unacceptable for high-fidelity music material.

Binaural synthesis has been used extensively in the research community during the last decade, mainly for the purpose of understanding the localisation mechanism of the human auditory system.

2.3 A balanced stereo widening network

Our network implements binaural synthesis with a careful choice of the head-related transfer functions that are realised by the direct gain $G_d$, the cross-talk gain $G_x$, and the interaural time difference $\text{itd}$. The values of
those parameters are based on the physics of sound propagation, as discussed in section 2.1. We set both $G_d$ and $G_x$ to one at low frequencies, and at high frequencies we set $G_d$ to two and $G_x$ to zero. Thus, if neither $G_d$ or $G_x$ vary too rapidly in the transition band, the sum of $G_d$ and $G_x$ is always very close to two. We can ensure that the network does not amplify its input by scaling the direct path and the cross-talk path by a factor of 0.5. Consequently, when a pure sine is input in one channel, the sum of the amplitudes of the two outputs is the same as the amplitude of the input. For this reason, we say that the network is balanced (the technical term for a pair of filters with this property is 'amplitude complementary'). At low frequencies, the input is split equally between the two outputs; at high frequencies it is passed straight through from the input to the output.

In order to minimise processing artifacts, in particular comb-filtering of the monophonic component at high frequencies, it is advantageous to make the low-pass characteristic of $G_x$ more dramatic than the effect it emulates in real life. Consequently, frequencies above 2kHz are considered 'high'. The resulting magnitude responses of $H_d$ and $H_x$ are shown in Fig. 3.

![Figure 3: The magnitude responses of the direct- and cross-talk paths of the network](image)

The value of the interaural time difference $\text{itd}$ affects the amount of widening perceived by the listener. The highest value encountered when listening to real sound sources is around 0.7ms, which corresponds to about 30 samples at a sampling frequency of 44.1kHz. A slightly higher value, 0.8ms for example, is good for a high degree of widening, but if $\text{itd}$ is made too large (>1ms) the reproduced sound becomes very unnatural and uncomfortable to listen to.

### 2.4 Implementation

An efficient implementation of the balanced stereo widening network is based on the simple digital filter structure shown in Figure 4.

![Figure 4: A simple filter structure and its magnitude response. $N_{\text{group}}$ is the group delay of the lowpass filter](image)

This filter structure takes advantage of the fact that it is very easy to modify the output from a linear phase lowpass filter so that the result corresponds to the output of another linear phase filter that also passes low frequencies straight through but which has a different magnitude response at high frequencies. Such a magnitude response is sometimes referred to as a shelving filter. Thus, a magnitude response of the type shown in Figure 4 can be realised from the output of a lowpass filter at very little extra expense. The extra processing requires a separate delay line whose length $N_{\text{group}}$ in samples corresponds to the group delay of the lowpass filter, plus two multiplications, by $g$ and $1-g$ where $g$ is the high-frequency magnitude response, and one addition of the outputs from the two branches. In practice, the group delay of the lowpass filter is of the order of 0.25ms which corresponds roughly to $N_{\text{group}}$ having a value of 10 samples at 44.1kHz.

Fig. 5 demonstrates how the filter structure shown in Fig. 4 can be used to achieve a computational saving by using the outputs from a single lowpass filter and a delay line to implement both the direct path and the cross-talk path.

![Figure 5: Implementation of a single virtual loudspeaker](image)
Fig. 6 shows a block diagram of the balanced stereo widening network in the case when $g_d$ is 2 and $g_x$ is 0, and when the overall levels in both channels have been scaled by a factor of 0.5. This network is very robust numerically, and it is suitable for an implementation in fixed point arithmetic. The linear phase low-pass filtering in the cross-talk path can be realized in many ways. The different techniques are well documented in the literature. A distinct advantage of the network shown in Figure 6 is that it is 'structurally balanced' in the sense that when a pure sine is input in one channel, the sum of the amplitudes of the two outputs is guaranteed to be the same as the amplitude of the input. This is so because the part of the signal not taken out by the lowpass filter is 'leaked' to the other channel, so loosely stated no part of the signal is lost. For this reason the design of a good lowpass filter for this structure is not as critical as one might expect.

3 CONCLUDING REMARKS
A balanced stereo widening network can be used to convert conventional stereo music recordings to a format suitable for headphone listening. Good externalisation is achieved while maintaining a natural quality of the reproduced sound, and as an added benefit listening fatigue is reduced. The network can be run efficiently in real-time on a fixed-point platform with modest computational resources. Parts of the technology described are the subject of a patent application filed by Nokia Corporation.

REFERENCES