# MODELING OF COLORATION OF VIRTUAL SOUND SOURCES IN LISTENING ROOMS

*Sharaf Hameed and Ville Pulkki*

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Otakaari 5, 02150 Espoo, Finland

sharaf.hameed@hut.fi

## ABSTRACT

Coloration of a virtual source is defined as the difference between the timbre of a virtual source and that of a reference real source. The coloration of virtual sources created by amplitude panning techniques has been previously studied and modeled in anechoic conditions using a simple auditory model implementing binaural loudness summation. In this study, the same model has been tested in reverberant conditions. The performance of the model was tested by conducting listening tests that analyzed the coloration of virtual sources perceived by listeners. The results reveal that the model is able to predict coloration very well for the optimum listening area in listening rooms.

## 1. INTRODUCTION

Localization of amplitude panned virtual sources has been extensively studied in the past. Yet, other perceptual properties of virtual sources like coloration have not been studied as extensively until recently. Such studies are essential in the evaluation of the quality of reproduced sound. A generic perceptual model that measures the quality of different sound reproduction systems would certainly need to be able to predict how colored a virtual source produced by a system would sound to an average listener.

A virtual source can be created in many ways. These include amplitude panning methods, time panning methods, Ambisonics, etc. The aim of these methods is to produce point-sized virtual sources in a specific direction. A virtual source is defined as an auditory object that is perceived in a location that does not correspond to any physical sound source.

The main factors in spatial sound perception are direction perception of the virtual source and coloration effects. In amplitude panning, a variation in timbre between a real source and a virtual source can be observed. This difference is termed as coloration. Timbre is defined as that attribute of auditory sensation in terms of which a listener can judge two sounds as being dissimilar, under equal loudness and without using the criteria of pitch or duration.

This work has been carried out as an extension to studies in which coloration of amplitude panned virtual sources is investigated in different setups [1][2]. Earlier studies have also modeled it using a binaural auditory model in anechoic conditions. In this study, the model is tested in a listening room.

The next section briefly discusses virtual sources, timbre and other key concepts. In section 3, the listening test procedure and results are described. The results are discussed in section 4, and finally, conclusions are drawn in section 5.

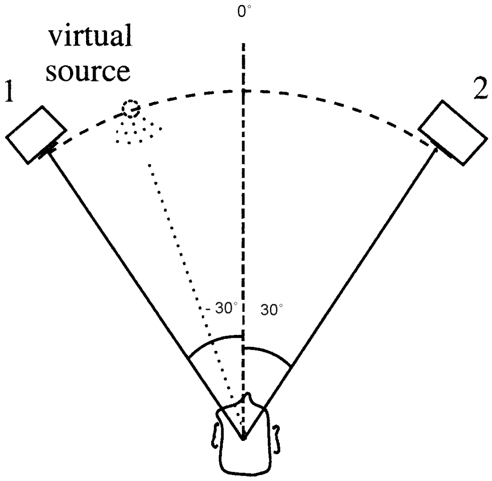

Figure 1. *Standard stereophonic listening setup. Depending on the amplitude gains for each loudspeaker, the virtual source can be positioned at any point between the speakers. If the amplitudes are equal, the virtual source is perceived by the listener as being directly in front of him, at the centre point between the two loudspeakers. (adapted from [1])*

## 2. PERCEPTUAL ATTRIBUTES

### 2.1. Virtual source

A virtual source denotes an auditory object that is perceived in a location that does not correspond to any physical sound source [1]. Typically the auditory cues of virtual sources do not correspond to cues of any real sources. In auditory research they are of interest because the way listeners perceive such distorted cues reflects human mechanisms for spatial sound perception. If the auditory cues of a virtual source were equal to a small-sized real source, the virtual source could then be described as "pointlike" or "sharp". Usually there exist some deviations in virtual source cues in different frequency bands and between different cues. A virtual source may then be perceived to be "spread", i.e., it is no longer point-like, and the perceived size of a virtual source is bigger. In some cases the virtual source can be perceived to be "diffuse", the direction of which is then undefined. In some cases it appears inside the listener's head. Fig. 1 shows an example of a virtual source created using the stereophonic setup.

### 2.2. Amplitude panning

Amplitude panning is a very widely used technique. In this method, a coherent sound signal is applied to two or more loudspeakers with different amplitudes. The loudspeakers are located at different directions and equal distances from the listener. Depending on the amplitude proportions, the listener perceives a virtual source in a particular direction. Amplitude panning is also sometimes called intensity panning.

A sound signal x(t) is applied to each loudspeaker with different amplitudes, which can be formulated as

$$x_i(t) = g_i\, x(t),\ i = 1, 2, ...,N, \tag{1}$$

where $x_i(t)$ is the signal to be applied to loudspeaker i, $g_i$ is the gain factor of the corresponding channel, t is the time variable and N is the number of loudspeakers [3].

The basic principle is that when two or more sound sources radiate coherent or partially coherent signals, only one auditory event may appear (virtual source), with its spatial attributes depending on the positions of all of the contributing sound sources and on the signals they radiate. This is called summing localization [4]. The signals from each loudspeaker are summed at the ear canal forming new signals. The attributes of these signals decide the perceived localization and timbre.

### 2.3. Stereophonic listening

Stereophonic reproduction is the simplest case of amplitude panning. Because of its simplicity, the setup is widely used. Two loudspeakers are positioned in front of the listener in such a way that they subtend an angle of 60° at the listening spot, as illustrated in Fig. 1. The perceived direction of an amplitude panned virtual source is not unequivocal. It depends on the frequency and temporal structure of the sound signal. The azimuthal angle of the generated virtual source can be estimated from the gain factors of the loudspeakers using a panning law. Several panning laws exist that give an estimate of this panning angle, such as the sine law, tangent law, etc.

### 2.4. Direction perception of a virtual source

The ability to resolve the direction of a sound source is possible largely due to the existence of two cues present in natural sound: Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs) [4]. Due to the physical nature of sounds, these cues are not equally effective at all frequencies. Interaural level differences are more prominent at high frequencies because the low frequency signals diffract around the head due to their wavelength being much larger than the size of the human head. Interaural time differences are effective at all frequencies. In normal listening conditions, the sound from a source reaches our ears via a number of different paths. Some of the sound arrives by the direct path, while others reach our ears after subsequent reflections from different surfaces in the room. In such a scenario, humans are able to accurately locate the position of the loudspeaker because when two brief sounds (time difference between 1 to 5-30 ms) are fused into a single sound, the location of the total sound is determined largely by the location of the first sound. This is known as the precedence effect [4]. This effect is strongly shown for sounds of a transient character, and typically requires the sound to have a wide bandwidth. It plays a very important role in our perception of everyday sounds, and in particular, listening to sounds in a room, where, had it not been for the precedence effect, we would have been aware of the reflected sounds and we would be unable to perceive sound source directions.

### 2.5. Timbre

Timbre denotes tone quality or tone color of sound [5]. It has been defined by the American National Standards Institute as "that attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar". It relates to the quality of a sound. It depends

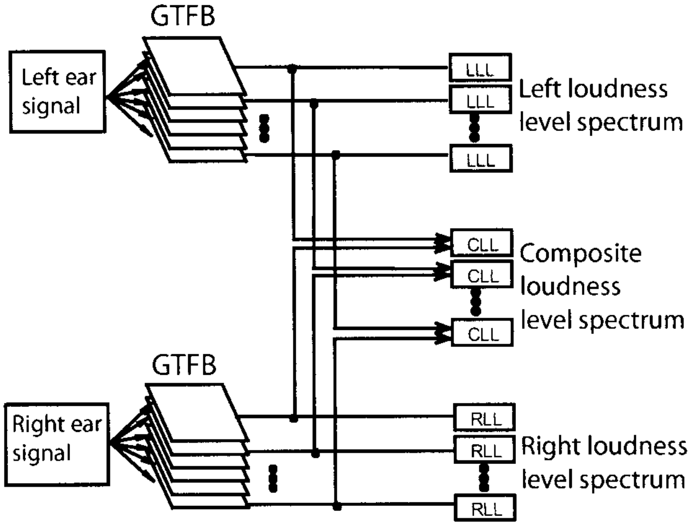
GTFB
Left ear signal
LLL
LLL
LLL
Left loudness level spectrum
CLL
CLL
CLL
Composite loudness level spectrum
GTFB
Right ear signal
RLL
RLL
RLL
Right loudness level spectrum

Figure 2. *Modeling of the left-ear, right-ear and composite loudness level spectra (LLL, RLL and CLL spectra respectively) in the binaural auditory model. The cochlear filtering of the inner ear is modeled using 84-band gammatone filter bank (GTFB) (from [2]).*

primarily on the spectrum of the stimulus, but it also depends on the waveform, the sound pressure, the frequency location of the spectrum and the temporal characteristics of the stimulus [6].

The five major acoustic parameters of timbre are: (i) the range between tonal and noiselike character, (ii) the spectral envelope, (iii) the time envelope in terms of rise, duration, and decay, (iv) the changes in spectral envelope (e.g. formant-glide) and fundamental frequency (micro intonation), and (v) the prefix (the onset of a sound is quite dissimilar to the ensuing lasting vibration) [7]. Thus, some method is necessary to describe the spectrum of a sound that takes this multidimensional aspect into consideration. In [8], a quantitative approach has been described by Plomp, where the bandwidth is divided into eighteen 1/3-octave frequency bands and the variations of the levels in each band convey information about the sound. A bandwidth of 1/3-octave is slightly greater than the critical bandwidth ($\mathrm{ERB}_\mathrm{N}$). Thus, in Plomp's model, timbre is related to the relative level produced by a sound in each critical band.

Such a method is useful to distinguish between two sounds that are represented successively, but is not enough to identify sounds, say, to identify which musical instrument produced the sound. This is because the transmission path may markedly color the frequency spectrum. Since the sound field is usually asymmetric to the human head, binaural loudness must be considered in the binaural modeling of coloration. It has been found that the perceived timbre of a sound source is dependent on the sum of decoded loudness of ear canal signals [9]. However, it also depends on the interaural level difference of the sound source.

In this study, the coloration (timbre variation) of amplitude panned virtual sources is being modeled. Amplitude panning is essentially a time-invariant reproduction method and does not produce any time-varying modifications to sound objects. This is fortunate because temporal patterning can be of crucial importance to timbre perception. The scope of this study can, thus, be narrowed down to static coloration of perceived timbre.

### 2.6. The comb filter effect

Combining an audio signal with a slightly delayed (delay time less than 10 ms) version of itself will give periodic cancellations in the frequency domain. These cancellations will occur for every frequency that is an integer multiple of the delay plus a half wavelength. This leads to a comb filter effect. Increasing the delay between the signals lowers the frequency of the first notch. At small delay values, therefore, low frequencies are unaffected. The depth of the notch in the comb filter is also affected if the delayed signal has an amplitude that is smaller than the direct signal. In stereophonic listening in anechoic conditions (Fig. 1), when the listener's head is pointing between the loudspeakers, the loudspeaker signals from the two loudspeakers arrive at any one ear with different arrival times. This difference in arrival time depends on the size of the human head, which leads to the distance from one ear to the two loudspeakers differing by approximately 7 cm. This corresponds to the value of the arrival time difference being approximately 0.2 ms. Because of this time difference, a comb filter effect is observed in the ear canal signals. If the amplitudes of the signals at the two loudspeakers are equal, the virtual source is perceived at the center point between the two loudspeakers, right in front of the listener. The timbre of this virtual source is, thus, prominently colored in anechoic conditions due to the comb filter effect.

### 2.7. Auditory model

The auditory model used in this study takes as input the ear canal signals and computes the frequency dependent loudness level spectra. The middle ear, cochlea, hair cells and auditory nerve have been modeled in the HUTear 2.0 software package [10]. The middle ear is modeled using a filter that approximates a response function derived from the minimum audible pressure (MAP) curve. The cochlear filtering of the inner ear is modeled using an 84-band gammatone filter bank, with center frequencies following the ERB scale [11]. The loudness in each frequency band in each ear is calculated using Zwicker's formulae [12]. The loudness of each frequency band at each ear are then summed together to yield a set of loudness levels in each band, termed as composite loudness level spectrum (CLL spectrum), also termed as specific loudness pattern. This is based on Zurek's model [13]. Fig. 2 shows the part of the model where the right ear, left ear and composite loudness patterns (LLL, RLL and CLL patterns respectively) are computed.

## 3. LISTENING TESTS

### 3.1. Setup

The test setup is shown in Fig. 3. The loudspeakers were placed in the listening room in front of the listener at 30˚ and –30˚. The test subject was equidistant from the two loudspeakers. The test method implements simultaneous recording of the ear canal signals of the test subject, so that the actual ear canal signals of the test subject can be recorded and processed with the model. This is accomplished by two electret microphones that are set at the entrance of the open ear canals of the test subject. The microphones were located about 5 mm outside the ear canal entrance. They minimally changed the perceived sound within the range of frequencies being tested and corresponded well with the ear canal signals.

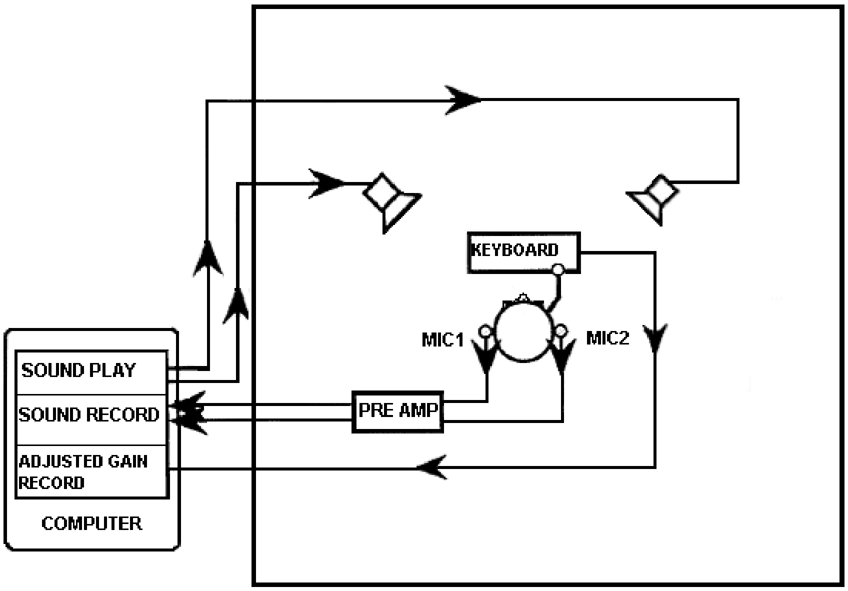

Figure 3. *Listening test setup. The test signals were recorded at the listening simultaneous with when the test subject had made his adjustments.*

The test signals used in these tests were 6-Barks wide. Tests were conducted for both 0 ms and 2 ms delay between the two loudspeakers. The two stimuli used in the 0 ms case together had bandwidths covering Bark channels of 3 to 14 (center frequencies between 250 Hz and 2150 Hz). The first test used a test sample covering Bark channels 3-8 and the second test covered channels 9-14. The magnitude spectrum of the reference test sound was flat over the 6-Bark bandwidth. The test subject could adjust the timbre of the broadband virtual source using a computer keyboard as a graphic equalizer. The keys on the keyboard were arranged in such a way that it could be used as an equalizer with 6 bands, one for each Bark bandwidth in the sound. Each band had four keys, with four corresponding controls: +3 dB, +0.5 dB, -0.5 dB and -3 dB, each representing an increase or decrease in level of the broadband virtual source.

The signal was first played only through the left loudspeaker. This was the reference real source. The virtual source was then created by playing the same sound through both the left and right loudspeakers with equal gains. The level of the virtual source was 60 dB SPL(A), measured at the listening position with a sound pressure level meter. The starting level of the virtual source was always randomized prior to each trial. The test subject's task was to match the virtual source and the reference real source maximally in timbre as well as loudness, using the keyboard as a virtual equalizer. When a match was found, the subject hit the enter key. Following this, two additional sets of test samples were played. During this period the subject was instructed to be as immobile and silent as possible. The latter sample was recorded. The signal was then bandpass filtered to remove any unwanted microphone noise outside the range under test and used as inputs to the auditory model. This recorded signal now contains the ear canal signals of the test subject corresponding to the reference real source and the adjusted virtual source. The auditory model then calculated the Central Loudness Level spectrum for the real source ($CLL_{Real}$) and that for the adjusted virtual source ($CLL_{Virtual}$) for the matched signals.

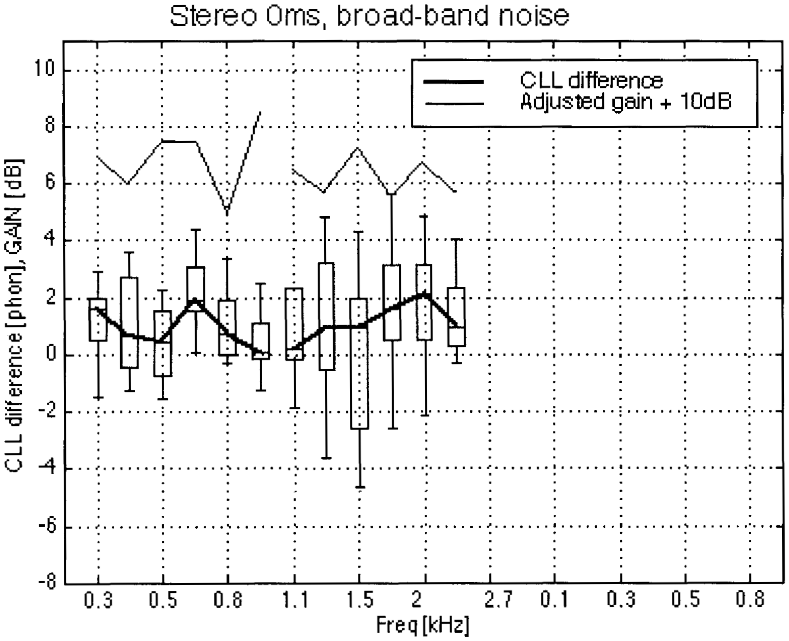

Figure 4. *Coloration in stereophonic listening in reverberant conditions with no delay between the loudspeakers: Barring a positive offset of about 1 phon, the plot remains near zero across the tested frequencies. Variations never exceed ±1 phon.*

The test was performed by five listeners, with three sets of trials each. They were all male researchers at this laboratory and had interest in musical acoustics and some experience in performing listening tests. The virtual source was multiplied by a gain coefficient, whose value was changed by 3 dB or 0.5 dB every time the subject pressed a button. Thus, a set of gain adjustment coefficients was obtained, one coefficient for each frequency band tested. These coefficient values were cascaded together to generate a pattern of gain adjustment values.

The aim of the experiment is to test if the model is able to predict the coloration of virtual sources, i.e. if it can compute the perceived loudness pattern from the reference real source. If the output pattern of the auditory model ($CLL_{Real}$) matches $CLL_{Virtual}$, it can be concluded that the model can predict coloration. Coloration is defined as the difference between the loudness pattern (CLL spectrum) of a single reference real source and that of the virtual source. Since the listener has adjusted the virtual source to be maximally similar to the real source, if the model is accurate, there should now be no coloration, i.e. the CLL difference plot should remain at zero.

The experiment was also repeated with a 2 ms delay between the two channels. This was done to test whether the model also works if the listener is outside the optimum listening position or sweet spot. If the listener is outside the sweet spot, there will be a delay between the arrival of the sound from the nearer loudspeaker and from the loudspeaker further away. This causes the signals to be interaurally decorrelated, unlike the case with 0 ms delay, in which the ear canal signals were interaurally correlated at all times. For the 2 ms case, only the low frequencies were tested, since in earlier studies, a discrepancy was observed at those frequencies. The failure was in the form of a dip of about 4 phons in the CLL difference plot at 300 Hz. Thus the signal used for the 2 ms tests had a bandwidth covering Bark channels 1-6.

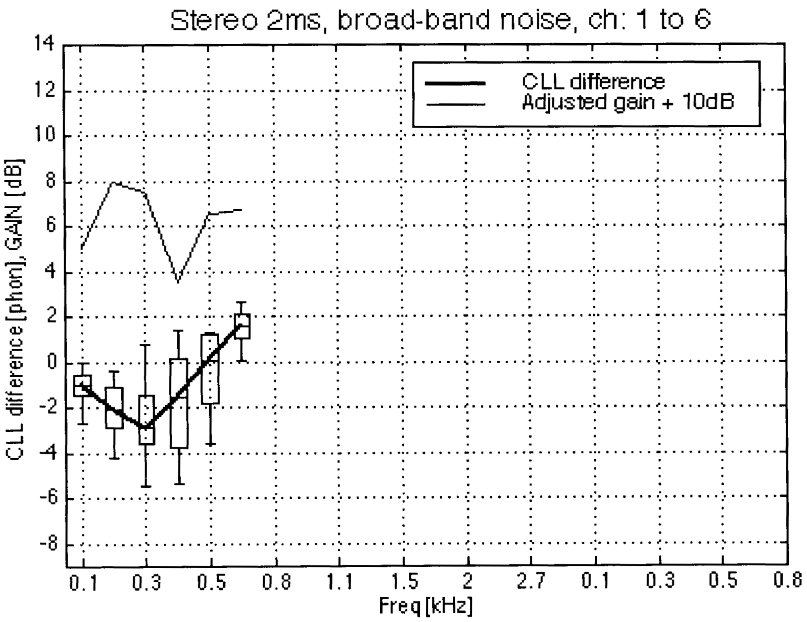

Figure 5. *Coloration in stereophonic listening with a 2 ms delay between the left and right loudspeaker channels: The prominent dip at 300 Hz, which was present in the anechoic tests with 2 ms delay, persist even in reverberant conditions.*

### 3.2. Results

The results of the experiments with 0 ms delay for frequency bands 3-8 Barks and 9-14 Barks are shown together in Fig. 4. A slight positive bias is noticed, which is possibly because the reference signal was fixed and only the virtual source was adjusted by the subjects in all tests. Such a testing procedure is likely to introduce a bias in the results. Similar biasing effects have been observed in [14] and [15]. Repeating the experiments, this time also keeping the virtual source fixed and matching the timbre and loudness of the real source to it, is likely to eliminate the slight positive bias. Barring the bias, the plot behaves well with frequency. It can be seen that the CLL difference plot variations never exceed ±1 phon. This indicates that the model is successful.

Although care was taken to ensure that the listening spot was exactly the same for all the test subjects, slight differences in the listening position and the exact position of the ears are unavoidable. This may also explain the variations in adjustments by different test subjects. The test was also repeated with the 2 ms delay between the loudspeakers. The results of this experiment, shown in Fig. 5, show that there is a 3-phon dip at 300 Hz. This is discussed further in the next section.

## 4. DISCUSSION

In earlier studies [2], a binaural auditory model was shown to be able to predict the perceived coloration reasonably well in anechoic conditions for the 0 ms case, which corresponds to the optimum listening position. The model has no additional binaural mechanism implemented other than simple summation of loudness from both ears. In this study, the model has also been validated in a listening room. It appears that for this condition (0 ms delay) in stereophonic listening, the model is adequate for coloration of virtual sound sources.

Amplitude panning is essentially a time invariant method. The signal is fed with different gains to the two (or three) loudspeakers to create a virtual source between them. The signals are always applied simultaneously; there is no time delay between the channels. Also, the only effect of head rotation is that the first comb filter notch moves higher in frequency [1]. A rotated head is the same as the virtual source being positioned to the side of the listener. Thus, the model can be reliably used as a general coloration model for the best listening position in amplitude panning.

The only major discrepancy in reverberant conditions is the persistent dip at 300 Hz in the 2 ms delay case. In [2], monaural tests were performed for 2 ms case. The results suggest that some binaural mechanism exists, other than the loudness summation implemented in the model. It was shown by testing with narrowband pulse trains that this unknown mechanism is unlikely to be the same as the auditory mechanism responsible for the precedence effect. The CLL difference spectra show a dip, which implies that the virtual source was perceived as less loud than was predicted by the model. The auditory model is unable to predict coloration correctly because it has no such binaural computational units implemented in it. The results are speculative; a more precise explanation would necessitate more evidence about the exact nature of this binaural mechanism. In any case, the auditory model based on binaural loudness summation performs reasonably well when there is no inter-channel delay in anechoic conditions, and very well in reverberant conditions as shown in this study. It can be concluded that simple binaural summation of loudness is sufficient for modeling of coloration of virtual sources.

## 5. CONCLUSIONS

In this study, coloration was modeled using a simple auditory model that only implements one binaural mechanism - neural summation of signals from the left and right ear. Listening tests were conducted using broadband signals in the stereophonic setup in a listening room. It was found that when the setup was symmetric, the binaural auditory model performed very well in reverberant conditions. However, if there is a time delay between the channels, the model showed consistent deviations from ideal performance. This (and results from monaural tests from earlier studies) implies that there are possibly some binaural effects other than simple summation that need to be considered in such cases.

Since, in amplitude panning, the signals are applied to the loudspeakers simultaneously, there is no time delay between the channels. It may be concluded that, for the optimum listening position, the binaural auditory model discussed in this paper can be used in the modeling of coloration of virtual sources that have been created using the amplitude panning technique in reverberant conditions.

## 6. REFERENCES

[1] V. Pulkki. Coloration of amplitude-panned virtual sources. In *Proc. AES 110th Convention*, Amsterdam, The Netherlands, 2001.

[2] K. Ono, V. Pulkki, and M. Karjalainen. Binaural modeling of multiple sound source perception: coloration of wideband sound. In *Proc. AES 112th Convention*, Munich, Germany, 2002.

[3] V. Pulkki. Spatial Sound Generation and Perception by Amplitude Panning Techniques. PhD thesis, Helsinki University of Technology, 2001.

[4] J. Blauert. Spatial hearing - The Psychophysics of Human Sound Localization (Revised Edition). MIT Press, Mass., USA, 1997.

[5] R.H. Gilkey and T.R. Anderson (ed.). Binaural and Spatial Hearing in Real and Virtual Environments. Lawrence Erlbaum Assoc., NJ, USA, 1997.

[6] American Standards Association. Acoustical Terminology. American Standards Association, New York, 1960.

[7] J.F. Schouten. The perception of timbre. Reports 6th International Congress in Acoustics, Tokyo, Japan, 1(GP-6-2), 1968.

[8] R. Plomp. Aspects of Tone Sensation. Academic Press, London, 1976.

[9] U.T. Zwicker and E. Zwicker. Effects of binaural loudness summation and their approximation in objective loudness summation. In Proc. Inter-Noise90, 1990.

[10] A. Härmä and K Palomäki. Hutear - a free matlab toolbox for modeling of the auditory system. In Proc. Matlab DSP Conference, 1999. http://www.acoustics.hut.fi/software/HUTear/.

[11] B.C.J. Moore, R.W. Peters, and B.R. Glasberg. Auditory filter shapes at low center frequencies. J. of the Acoustical Society of America, 88:132–140, July 1990.

[12] E. Zwicker and H. Fastl. Psychoacoustics: Facts and Models. Springer-Verlag, Heidelberg, Germany, 1990.

[13] P.M. Zurek. Measurements of binaural echo suppression. J. of the Acoustical Society of America, 66:1750–1757, 1979.

[14] B.C.J. Moore, S. Launer, D. Vickers, and T. Baer. Loudness of modulated sounds as a function of modulation rate, modulation depth, modulation waveform and overall level. In A.R. Palmer, A.Rees, A.Q Summerfield, and R. Meddis, editors, Psychophysical and Physiological Advances in Hearing, pages 465–471. Whurr, London, 1998.

[15] H. Gockel, B.C.J. Moore, and R.D. Patterson. Influence of component phase on the loudness of complex tones. Acustica - Acta Acustica, 88:369–377, 2002.