nederlands akoestisch genootschap
NAG

journaal

nr. 168
nov. 2003

Virtual acoustics for consumer electronics

dr Ronald M. Aarts
Philips Research Labs, Prof Holstlaan 4 (WO 02), 5656 AA Eindhoven

Abstract

Today and tomorrow's audio and video, portable audio and multi-media applications put increasing demands on sound reproduction techniques. On one hand there is a need for reductions in both cost and size, on the other hand we wish to enhance the experience of the user beyond today's possibilities. A good sound reproduction system is in general in conflict with the boundary conditions for consumer products both by size as well as by price requirements. A possible way to ease these conflicts is to enhance the reproduction and perception of sound for listeners by exploiting the combination of psycho-acoustics, loudspeaker configurations and digital signal processing. Various examples will be given, such as increasing the perceived bass response of loudspeakers, and increasing the number of loudspeaker channels (converting stereo to multichannel sound). When multichannel reproduction through loudspeakers is not a viable option, the same percept can be simulated over headphones. This method, using active noise control principles, is discussed as well.

Part 1: Two-to-Five Channel Sound Processing

1 INTRODUCTION-Part 1

Since the introduction of digital versatile disk (DVD) and super audio CD (SACD), a revival of multichannel audio has appeared in sound systems for consumer use today. It is, however, desirable to maintain the compatibility with the existing two-channel stereo recordings and/or broadcasting. Therefore the conversion of two-channel stereo to multichannel format has been studied extensively over the decades, and a considerable number of publications exists, see [1] and references there in. In this part of the paper we focus on an algorithm for format conversion from two-channel stereo to five-channels (two-to-five). The desired setup is shown in Fig. 1, in which the channels are labeled L (left), C (center), R (right), S_L (left surround), S_R (right surround) according to convention.

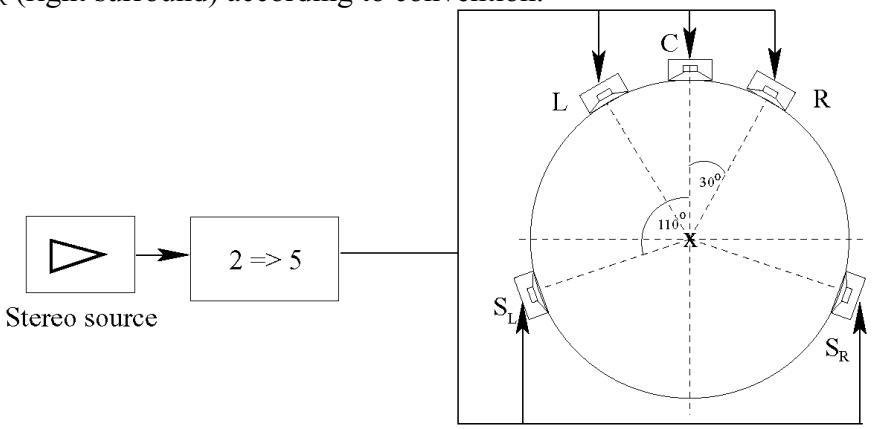

Figure 1: ITU reference configuration [2]. The reference listening position (sweet spot) is indicated by **x**. Left and right channels are placed at angles ±30° from C, and the two surround channels are placed at angles ±110° from C.

This setting is adopted from the ITU multichannel configuration [2], with three loudspeakers placed in front of the listener, and the other two at the back. The front channels are used to provide a high degree of directional accuracy over a wide listening area for front stage sounds, particularly dialogues, and the rear channels produce diffuse surround sounds providing ambience and environment effects. An additional loudspeaker (subwoofer) may be used to augment bass reproduction, which is often called 5.1 system with .1 referring to the low frequency enhancement (LFE) channel. In this paper, however, we do not use a subwoofer, since it can easily be added when necessary without affecting the algorithm.

2 THE CENTER LOUDSPEAKER

We consider the three-channel approach first. It is known that the sound quality of stereo sound reproduction can be improved by adding an additional loudspeaker between the loudspeakers. We propose an algorithm to derive the center channel without these drawbacks using Principal Component Analysis (PCA) [3] which produces two vectors indicating the direction of both dominant signal y and remaining signal q as shown in Fig. 2 by dashed lines. Note that these two directions are perpendicular to each other, creating a new coordinate system. These two signals are then used as basis signals in the matrix decoding, a point that is different from other existing two-to-five sound systems.

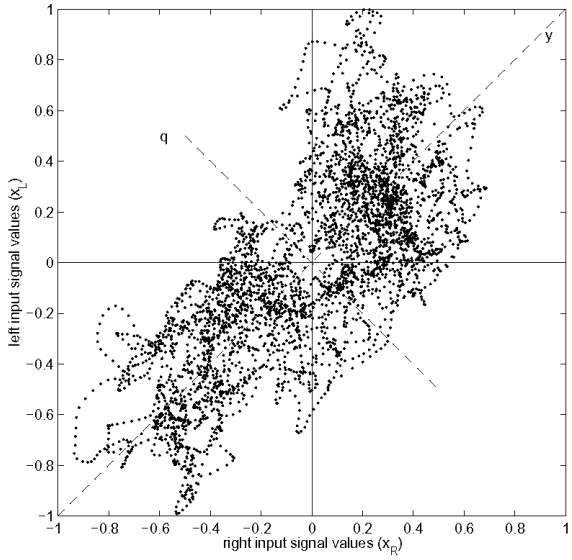

Figure 2: A Lissajous plot of a stereo signal recorded from the fragment "The great pretender" by Freddy Mercury. Dashed lines represent new coordinate system based on both the dominant signal y and remaining signal q forming a direction of a stereo image α.

To derive the center channel's gain using the direction of a stereo image, we process the audio signal coming from a CD (sampling frequency F_s = 44.1 kHz) on a sample basis. The direction of a stereo image in terms of angle α being the angle between the y and the positive x_R axis (see Fig. 2), which can be efficiently tracked [1]. Recalling Fig. 2 with the left channel corresponding to $\alpha = \pi/2$, and right channel to $\alpha = 0$, α fluctuates around $\pi/4$ creating a phantom source almost equidistant between left and right channels. To map this stereo vector onto a three-channel vector, we double the angle α producing a new mapping. We can then find the projections of the vector onto the LR-axis, and C-axis using sine and cosine rules. It should be pointed out that this mapping works only for nonnegative 2α. This is because for negative 2α, a multiplication by a factor two results the vector to be in a lower

quadrant, and therefore no gain can be derived for the center channel. To overcome this problem, extra information should be used which is described in the next section.

3 THE SURROUND LOUDSPEAKERS

The surround channels are generally used to create ambience effects for music, while for applications in the film industry the surround channels are used for sound effects. Environmental and ambience effects can be computed by considering left and right channel variation ($x_L - x_R$) in the original signals. This variation is usually referred to as the *anti-phase* components, the amount of which can be represented by the remaining signal q (Fig. 2). However, it can be expected that when the amount of the dominant signal equals or almost equals to that of the remaining signal, an ambiguity appears since there is no way to determine the direction vector uniquely. In this situation the distribution in Fig. 2 is no longer an ellipse but has a circle-like form ($|y| \approx |q|$), causing α to be not well defined. In [4] it is shown that the correlation coefficient $\rho_0(k)$ can be computed recursively by using only a few arithmetic operations [4].

3.1 Three-dimensional mapping

To avoid ambiguity when the amount of the dominant signal approaches that of the remaining signal, the use of both the direction of the stereo image and the correlation coefficient is necessary. The latter is included in the mapping by, for example, placing the surround channels in the vertical plane, as shown in Fig. 3. The angle β can be defined to represent the actual surround information by means of the adaptive correlation coefficient, for example, by using Eq.(1), where $\rho_0(k)$ is the (with time index k) correlation coefficient.

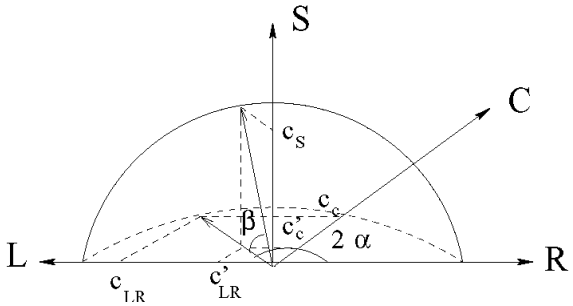

Figure 3: Three-dimensional mapping showing front (horizontal plane) and surround channels (vertical plane). Parameter β determines the level of surround information with respect to the front channel sounds.

$$\beta(k) = \arcsin(1-\rho_0(k)), \qquad (1)$$

3.2 Matrixing

The system as described so far reproduces four channel signals as L, C, R and S from two input signals. Therefore, we have a 4_2 reproduction matrix, see Eq.(2), where y and q are the rotated input signals (see Fig.2), the c's are the various (time variant) coefficients, based on the projections shown in Fig. (3), and $g = \cos^2 \beta$ (see [1] for more details).

$$\begin{bmatrix} u_L(k) \\ u_R(k) \\ u_C(k) \\ u_S(k) \end{bmatrix} = \begin{bmatrix} c_L(k) & gw_L(k) \\ c_R(k) & gw_R(k) \\ c_C(k) & 0 \\ 0 & c_S(k) \end{bmatrix} \begin{bmatrix} y(k) \\ q(k) \end{bmatrix} \quad (2)$$

The components of the left-hand side of Eq. 2 denote the signals for Left, Right, Center loudspeakers, and u_S the mono surround signal. The basis signals are obtained by rotating the coordinate system of x_L and x_R , to y and q as shown in Fig. 1. Finally, a decorrelator is used to obtain stereo surround because of its simplicity. This decorrelator can be viewed as two FIR comb filters (h_L and h_R) with two taps each for surround left and surround right. A time delay of δ (=440 samples) $\approx$ 10 ms is used between the taps, which is determined experimentally. The choice of the time delay δ is a subtle compromise between the amount of widening and the sound diffuseness. The greater δ is, the more diffuse the sounds will be, and at some point it will lead to confusion.

4 CONCLUSIONS-Part 1

A method to convert two-channel stereo to multichannel sound has been presented. A three-dimensional representation has been used to produce each channel's gain, which is time varying. PCA is proven to be a powerful tool to detect the direction of a stereo image, which is then used to derive the center channel's gain. Furthermore, a robust tracking algorithm for computing the cross correlation between left and right channel has been used to improve the sound quality of the surround channels.

Part 2: Bandwidth extension of band-limited signals in particular for reproducing low pitched signals through small loudspeakers

5 INTRODUCTION-Part 2

In many sound reproduction applications, it is not possible to use large loudspeakers, due to size and/or cost constraints. Typical applications are portable audio, multimedia, TV, and public address systems, to name just a few. These devices are often small in size, and therefore the transducers are inherently small as well. However, probably the most well-known characteristic of small loudspeakers is a poor low-frequency (bass) response. In practice this means that a significant portion of the audio signal may not be reproduced (sufficiently) by the loudspeaker. For loudspeakers used in such applications reproduction below 100 Hz is usually negligible, whereas in some applications this lower limit can easily be as high as several hundred hertz. The bass portion of an audio signal contributes significantly to the sound ‘impact’, and depending on the bass quality, the overall sound quality will shift up or down. Therefore a good low-frequency reproduction is essential.

6 VIRTUAL PITCH

Pitch is a subjective, psychophysical quantity. According to the American Standards Association pitch is ’that attribute of an auditory sensation in terms of which sounds may be ordered on a (musical) scale extending from high to low’. For a pure tone, where the fundamental frequency corresponds to the frequency of the tone, the pitch is unambiguous and- if we neglect the influence of sound level on pitch- one can identify pitch with the frequency of the pure tone. For a complex tone, consisting of more than one frequency, the

situation is more complicated. Pitch should then be measured by psychophysical experiments. A pitch that is produced by a set of frequency components, see Fig. 4 -b, rather than by a single sinusoid, is called a *residue*.

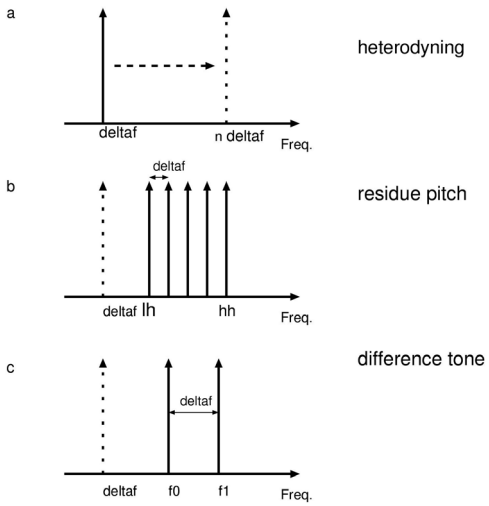

Figure 4: Possible options for psychoacoustic bass enhancement. The dotted frequency component denotes the perceived pitch (but is not necessarily acoustically radiated). (a) Frequency doubling. (b) Residue pitch. (c) Difference tone.

In Fig. 4 -b the fundamental frequency is missing, yet it will still be perceived as a residue pitch, which in this case is also called *virtual pitch*. The psychoacoustic phenomenon responsible for this effect is the 'missing fundamental' effect. There is a vast amount of literature on this topic; see [6]. As the frequency of a pure tone decreases to very low values, say less than 100 Hz, the pitch becomes more difficult to determine. This is also true for the missing fundamental effect, and because the proposed algorithm is aimed at this very low frequency range, we need psychoacoustic data regarding the perception of virtual pitch for this range. Unfortunately, only sparse data is available. The work of Ritsma [7,8] investigates the existence region of the tonal residue, for frequencies above 200 Hz.

7 PROCESSING SCHEME

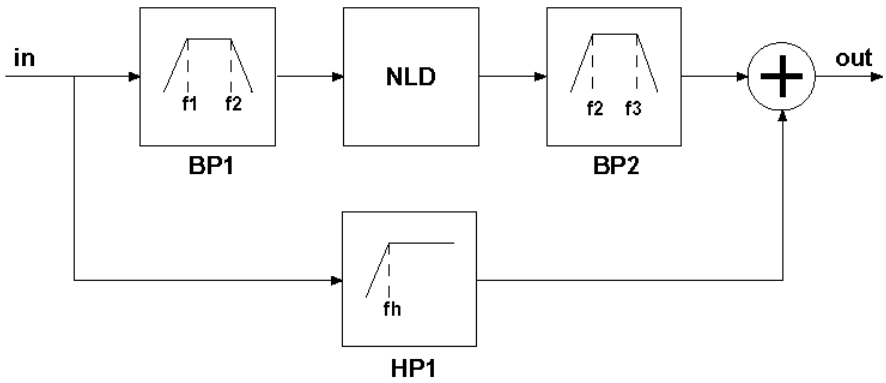

Figure 5: Signal processing for psychoacoustic bass enhancement. The input signal is summed and filtered to obtain the bass portion. Then harmonics are created by the non-linear device (NLD) and added to left and right output signals. In the direct path a high-pass filter is implemented.

Fig. 5 presents the general processing scheme that we propose for psychoacoustic bass enhancement [5]. As the system is 'merely' based on a psychoacoustic model of pitch perception, and uses loudspeaker characteristics in a very general sense (it is only assumed

that reproducing lower frequencies is less efficient than reproducing higher frequencies), the method can be employed for any kind and/or size of loudspeaker.

7.1 NON-LINEAR DEVICE - NLD

The non-linear device, or harmonics generator, 'shifts' signal components in a low frequency range to a higher frequency range. The pitch of the input signal is preserved, because the components in the higher frequency range are harmonics of the original components. The preservation of the original low pitch is due to the virtual pitch of the harmonics signal. Because this element is a non-linear device, any single output component depends on all input components. Moreover, at the output, frequency components will be generated, which are not present at the input. This is a desired effect, since this is how the harmonics are obtained. However, it also leads to sum and difference components, which are not desired, for they are not harmonically related to the input signals.

8 CONCLUSIONS-Part 2

In this part of the paper we have proposed a psychoacoustically based signal processing system to enhance the perceived bass response of a loudspeaker below its cut-off frequency. The main concept of this system is to replace very low frequency components by their harmonics, through controlled non-linear processing. The resulting harmonics yield the same (virtual) pitch as the original signal, due to the missing fundamental effect. The added harmonics interfere with frequency components in the original audio, which in some cases may alter the timbre of the signal to some extent.

Part 3: 3D headphones based on active noise cancellation

9 INTRODUCTION-Part 3

Headphone virtualizers are systems that aim at giving the user the illusion that the sound is coming from loudspeakers rather then from the headphones themselves [9-14]. Systems that are commercially available today are not optimized for the individual listener. This results in large localization errors for most listeners. The system at hand is personalized in that it requires a calibration procedure, which can be carried out conveniently by the listener. This system consists of conventional headphones into which miniature microphones have been mounted. The sound reproduction using headphones gives the same listening experience to the user as the reference (multichannel) loudspeaker system. This is achieved by taking all contributions into account: the room impulse responses, the loudspeaker characteristics, the headphone characteristics and the properties of the listener's head and torso. Besides the usual computational requirement for a headphone virtualizer, this system needs in addition two low-cost microphones and two analog to digital converters to convert the microphone signals.

9.1 Technology background

The way in which sound propagates from the loudspeaker towards the ear-drums of the listener depends on the loudspeaker, the room and the physical properties of the listener (e.g. the shape of the head, ears, and torso). If loudspeaker reproduction is emulated using headphones, these sound characteristics have to be taken into account and compensation for the sound reproduction characteristics of the headphones is required.
The physical properties of the head and outer ears of the listener modify the sound as it travels from the source to the ear-drums. The transfer functions describing this sound

propagation from multiple sound sources to both ears are known as head-related transfer functions (HRTFs). Multichannel audio can be filtered with the HRTFs of the listener and the inverses of the headphone to ear transfer functions prior to headphone sound reproduction. In this way the multichannel loudspeaker system can be emulated very accurately. Note that only one loudspeaker driver is required at each side of the head in order to make multichannel virtual sound. Sounds add in a linear way in the air so that headphone signals of the virtual left loudspeaker can be added to those of the virtual right loudspeaker to obtain virtual stereo for example. When audio is filtered with HRTFs that are measured from another person, there are large errors in the vertical and front/back localization. Therefore the sound reproduction system should be personalized.

9.2 Configuration

The headphones are equipped with integrated microphones [14] and are connected to a digital signal processing unit (DSP). During the calibration, the DSP is connected to the multichannel loudspeaker setup. A noise signal is played through each of the loudspeakers consecutively and is picked up by the microphones. The DSP then computes how the sounds should be processed prior to headphone reproduction, such that exactly the same sound is generated at the position of the microphones, which are very close to the ears. The algorithm that is used is described in the next section. When the calibration is completed, the listener can manually choose between loudspeaker or headphone sound reproduction, showing the capabilities of the system. A variant is that the calibration is carried out using only one loudspeaker. The subject needs to change his/her orientation after each measurement such that this loudspeaker corresponds to the left front, right front, left rear, right rear and center loudspeaker position.

10 Active Noise Cancellation

The algorithm that is used during the calibration is essentially an active noise cancellation algorithm. An introduction to active noise cancellation can be found in [15]. Its application to headphone listening will be explained below. In sound reproduction systems, sound signals can be filtered prior to reproduction by loudspeakers to ideally obtain perfect sound reproduction at a finite number of positions in space. The filters can be found by first placing microphones at the relevant positions and using the difference between the ideal sound and the reproduced sounds at these positions as errors signals for an adaptive algorithm. In classic adaptive-filter theory these error signals are obtained by comparing the desired signals with the adaptive filter outputs. In active noise cancellation the error signals are obtained by comparing the desired signals with adaptive filter outputs that are filtered by acoustic transfer functions. The classical adaptive filter is depicted in Fig. 6 (top) where it is used to equalize the acoustic transfer function from the loudspeaker to the microphone H(z). Here, the update uses the input signal of the adaptive filter W(z) and the difference between the reference signal d[n] which resembles x[n] and the adaptive-filter output. The reference signal d[n] can be a delayed version of x[n] for example, so that W(z) can converge to a stable solution with W(z) H(z) equal to this delay. Instead of equalizing a signal x[n] that is filtered by an acoustic transfer function H(z), sound reproduction systems need to filter this signal prior to playback, as depicted in Fig. 6 (bottom). Both systems are equivalent if the adaptive filter W(z) is constant. In practical applications it suffices to demand that the adaptive filter is slowly varying. The latter system is termed the filtered-x algorithm in [15] which indicates that a filtered version of the signal x[n] is used in the update. This filter _(z), which corresponds to the acoustic transfer function H(z), is not exactly known in practical situations. An estimate

of it can be used however, and the filtered-x algorithm is known to be robust to estimation errors herein.

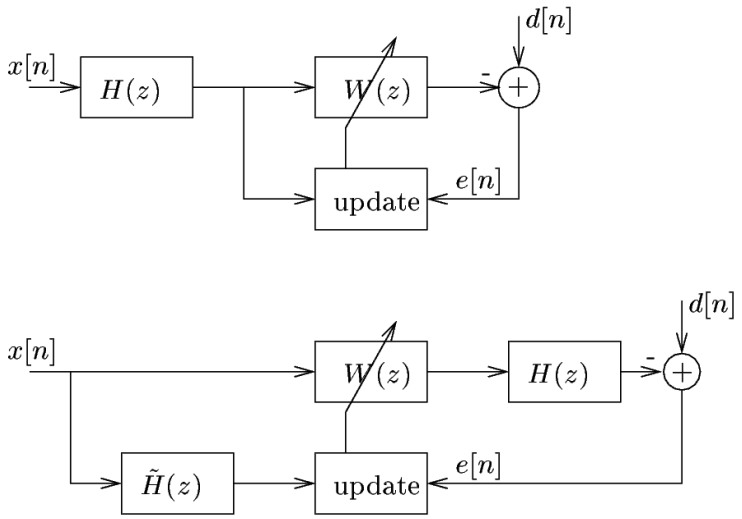

Figure 6: Conventional adaptive filter (top) and filtered-x equivalent (bottom).

Using the above outlined filtered-x algorithm, we apply this as follows. A loudspeaker, which we want to simulate, is playing a noise signal x(n), which is received by the microphone, depicted as the adder in the Fig. 6 (bottom). (The system is shown for only one side of the headphones. The processing for the other side is identical and works independently.) This noise signal is also filtered, by filter W, before it is fed to the headphones. The filtering is done in such a way that the microphone signal is minimized. In this way the adaptive filter W will adjust such that (apart from a minus sign) the filtered headphone signal received at the microphone will become approximately equal to the signal due to the loudspeaker directly.

11 CONCLUSIONS-Part 3

The performance of a system that delivers multichannel sound using headphones is analyzed. The system is calibrated using active noise cancellation techniques.

Acknowledgments

The author would like to thank his colleagues Roy Irwan, Erik Larsen and Daniël Schobben for their parts in the work presented in this paper.

References

[1] R. Irwan and R.M. Aarts. Two-to-five channel sound processing. *J. Audio Eng. Soc.*, 50(11):914-926, November 2002.

[2] ITU-R Recommendation BS.1116, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, International Telecommunication Union, Geneva, Switzerland, (1994).

[3] S. Haykin. *Neural Networks*. Prentice-Hall, N.J., 1999. Second Edition.

[4] R.M. Aarts, R. Irwan, and A.J.E.M. Janssen. Efficient tracking of the cross-correlation coefficient. *IEEE Trans. on Speech and Audio Proc.*, 10(6):391-402, September 2002.

[5] E. Larsen and R.M. Aarts. Reproducing low-pitched signals through small loudspeakers. *J. Audio Eng. Soc.*, 50(3):147-164, March 2002.

[6] A.J.M. Houtsma and J.L. Goldstein. The central origin of the pitch of complex tones: Evidence from musical interval recognition. *J. Acoust. Soc. Am.*, 51(2 (Part 2)):520-529, 1972.

[7] R.J. Ritsma. Existence region of the tonal residue I. *J. Acoust. Soc. Am.*, 34(9):1224-1229, September 1962.

[8] R.J. Ritsma. Existence region of the tonal residue II. *J. Acoust. Soc. Am.*, 35(8):1241-1245, August 1963.

[9] D. Schobben and R.M. Aarts. 3D headphones based on active noise cancellation. *Convention Paper 5713 Presented at the AES 113th Convention 2002 Oct. 5-8, Los Angeles, CA, USA*, 2002.

[10] J. Blauert. *Spatial hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1983.

[11] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4):171-218, 1992. Special issue on auditory environment and telepresence.

[12] F.L. Wightman and D. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoust. Soc. Am.*, 85(2):858-867, February 1989.

[13] F.L. Wightman and D. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *J. Acoust. Soc. Am.*, 85(2):868-878, February 1989.

[14] R.M. Aarts. Headphones with integrated microphones, 2000. European patent EP1201101.

[15] P.A. Nelson and S.J. Elliott. *Active control of sound*. Academic Press, 1992.