Coding of Elevation in Acoustic Image of Space

Rudolf Susnik, Jaka Sodnik and Saso Tomazic

Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

ABSTRACT

Spectral features of sound are believed to be the primary cues for the human perception of spatial sound elevation. It has also been observed that people connect higher frequencies of sound with a higher elevation of the sound source and lower frequencies with lower elevations. The most common approach to creating an acoustic image delivered by headphones is to use Head Related Transfer Functions (HRTFs). Unfortunately, satisfactory perception of elevation can only be achieved with personalized HRTFs which are impractical to measure. This paper describes an alternative method of sound coding for representation of the virtual sound source elevation in an acoustic image. Our method is based on coding particular elevations with sound stimuli which differ in spectral content. Sound stimuli were created by various signal processing techniques (e.g. filtration, modulation). Experiments show that, in certain cases, test subjects were able to perceive up to 60 different elevations in the range of –40° to 90°.

INTRODUCTION

In localizing sound sources, a listener uses various acoustic cues present in the perceived sound. These cues are the result of the travel of sound waves through the space between a sound source and the eardrum. The environment primarily affects sound by reflections from objects located around the listener and the sound source. These reflections are normally not perceptible when direct sound is played; otherwise, they act as a distance cues. The most important are reflections from the body, head and pinna, which impact direction-dependent Head Related Transfer Functions (HRTF). HRTFs are complex-valued free field transfer functions from a sound source in a certain direction to the eardrum. They include all information about direction-dependent cues, such as inter-aural time difference (ITD), inter-aural level difference (ILD) and spectral features. Since every person has a different shape and dimension of the body, head and pinna, HRTFs depend on the individual characteristics of the person.

It is well established that the binaural inter-aural time difference and the inter-aural level difference provide the primary cues for the horizontal localization of a sound source (Blauert 2001; Sodnik et al. 2004), whereas monaural spectral modifications caused by reflections and diffractions provide the primary cues for vertical localization (Algazi, Avendano, Duda 2001; Bloom 1977). Spectral shape features affected by pinna become apparent at frequencies above 4 kHz (Hofman, Van Opstal 1998; Bronkhorst 1995), where the wavelength becomes comparable to the pinna size. Algazi, Avendano, Duda (2001) have pointed out that one of the most characteristic pinna effects is the so-called "pinna notch", which appears within the octave from 6 kHz to 12 kHz. This supports the general belief that a sound source must have high-frequency energy for accurate perception of elevation (Hofman, Van Opstal 1998). Nevertheless, sound sources containing only low frequencies can also be well located by elevation (Algazi, Avendano, Duda 2001). The fact is that besides pinna effects, disturbances caused by the body influence perception of elevation and that these elevation cues, which are primarily due to torso reflection and head diffraction, manifest as spectral features below 3 kHz (Algazi, Avendano, Duda 2001; Hofman, Van Opstal 1998). These cues are significant away from the median plane (Algazi, Avendano, Duda 2001).

In previous studies on elevation localization, monaural spectral cues and source elevation manipulation, different sound sources such as narrowband, passband, broadband, etc. have been used (Bloom 1977; Hofman, Van Opstal 1998; Jin et al. 2004; Langedijk, Bronkhorst 2002; Meijer 1992; Sodnik et al. 2004; Sodnik et al. 2005; Watkins 1978; Ziberna, Zazula 2003). Studies that used narrowband sounds (tones or narrowband noise) demonstrated that perceived elevation varies with the centre frequency of a sound, whether the same sound is typically perceived at fixed elevation regardless of its actual elevation (Hofman, Van Opstal 1998). In the study where bandpassed Gaussian noise was used (Langedijk, Bronkhorst 2002), it was observed that spectral cues in the 4 – 16 kHz frequency band are essential for correct localization of broadband sounds and that the most important elevation cues are present in the 5.7 – 11.3 kHz band. Another study with broadband noise filtered with a notch filter at different central frequencies demonstrated that a sharp dip in the high frequency spectrum of certain signals heard monaurally is sufficient information to evoke a sensation of source elevation (Bloom 1977).

In previous studies it was also established that loudness can affect perception of elevation (Hofman, Van Opstal 1998). This term is closely linked to the sensitivity of the human ear, which does not directly reflect sound intensity in dB. Human hearing has a maximum sensitivity in the region between 3 and 4 kHz (Blauert 2001).

The aim of our research is to find appropriate methods for coding the elevation of a virtual sound source in an acoustic image of space delivered by headphones. Acoustic image is an approach to help blind people orient in space. It is created with spatial sound synthesis, done in a way that most efficiently describes the visual image. An image is first captured (camera, radar, sonar) and then converted to sound. One of the practical systems uses temporal scanning of the image from left to right. Azimuth (horizontal direction) is encoded with time delay, elevation (vertical direction) with sound frequency, and distance with sound amplitude. This is quite efficient, but unnatural to the brain. We propose an approach based on the division of visual space into several subspaces. The acoustic image is created by measuring the distances to obstacles in subspaces, randomly selected in accordance with an appropriate space dispersion function.

The synthesis of spatial sound can be made by the use of Head Related Transfer Functions (HRTFs).

Using HRTFs, satisfactory perception of elevation can be achieved only with personalized HRTFs measured for each individual. Since HRTF measurements are impractical, an alternative method of special sound coding, based on the fact that human brains connect higher frequencies of sound with a higher elevation of the sound source and lower frequencies with a lower elevation, to represent the elevation of a virtual sound source in an acoustic image should be used. In an attempt to determine appropriate sound samples for better elevation coding in a virtual acoustic image, i.e. creating elevation illusions, we prepared a few sound stimuli considering findings from the studies mentioned before. Different sound stimuli were processed for listening through headphones. All experiments were performed in an anechoic chamber.

METHODS

Psychoacoustics

Loudness – the sensitivity of the human ear varies with frequency. The ear is most sensitive in the frequency range between 3 and 4 kHz, which means that sound pressure levels that are just noticeable at these frequencies are not detectable at other frequencies. In other words, two tones of equal power but different frequencies would not sound equally loud. The unit measure for perceived loudness of sound is called a sone. A sone is defined as the loudness of a 4 kHz tone with a 40 dB sound pressure level (Tsutsui et al. 1992; Blauert 2001).

Masking is a phenomenon by which a sound called the "masked signal" is rendered inaudible by the sound called the "masker". This happens when sounds of similar loudness get closer both in time and frequency or when the masker is much louder than the masked signal. In the latter case, masking also becomes stronger when sounds get closer both in time and frequency. Two types of masking are known: backward masking occurs when the masked signal ends before the masker begins and forward masking occurs when the masked signal begins after the masker has ended (Tsutsui et al. 1992).

Critical bands arose from the idea that the human hearing system analyses the frequency range of incident sound using a set of sub-band filters whose ranges are ranges of critical bands as well (Zwicker, Flottorp & Stevens 1957; Tsutsui et al. 1992). The frequencies within a certain critical band are similar in terms of perception and are therefore processed separately from other critical bands. The width of critical bands increases with frequency, and therefore critical bands are much narrower at low frequencies than at high frequencies. The ear can only make sense of one signal per critical band, but there is another mechanism at work that fine tunes the ability to discern frequency. If two tones are within the same critical band, the listener hears some sort of unresolved sound. Two distinct tones would be only heard when the two tones exist in separate critical bands. On the other hand, the loudness of a broadband signal with flat amplitude does not depend on its bandwidth in the case where the bandwidth is equal to or lower than the critical band, as stated by Zwicker, Flottorp & Stevens (1957). The loudness of such a signal depends on the number of occupied critical bands.

Just Noticeable Differences (JND) – when the difference in frequency between the two tones, presented one after another, is too small, listener perceive both tones as having the same pitch (Just Noticeable Difference (JND) for Three Frequencies). Whenever the variation of an original physical stimulus lies within a certain difference limen or just noticeable difference (JND) the associated sensation is judged as remaining the same. As soon as the variation exceeds the JND, a change in sensation is detected. This is due a natural limit of an ability to establish a relative order of pitch when two sine tones of the same intensity are presented.

Test signals

According to the facts given above, ten different test sounds signals were generated for experiments with headphones. In every case, at least one of the factors that affect elevation perception as described above is present.

The ten test sound signals can be divided into four groups: in first group, signals were processed using modulation; in second, lowpass filtering; in third, bandstop filtering; and the fourth group represents a harmonic sine signal. The frequency range for all four groups of signals is between 1 kHz and 19 kHz, and the basic frequency step is 250 kHz. The basic frequency step represents the frequency distance between two neighbouring signals, depending on processing techniques. Where modulation is used, the frequency step is the difference between the frequencies of two carriers; where bandpass filtering is used, the frequency step means the difference between the central frequencies of the two filters; and for lowpass filtering, the step is the difference in the filters' edge frequencies.

From each of these sounds, sets of 73 different sounds were created using a certain processing technique. In the experiment, the suitability of a particular set of sounds for elevation coding was observed. Therefore, each sound from a set would code a certain elevation. Test signals were processed in Matlab 6 and then converted to Windows Audio Video format (WAV) files.

Modulated pink noise. Pink noise is frequently used as a signal in acoustical experiments. Its main characteristic is its distinctive frequency-dependent amplitude density. The power spectral density (P) of pink noise is inversely proportional to the frequency f, giving a negative slope of 3 decibels per octave (Equation 1).

$$P \propto \frac{1}{f} \tag{1}$$

Modulation of pink noise, i.e., multiplying pink noise with a harmonic signal (carrier) of arbitrary frequency f_c, produces a signal that does not have the property described by Equation 1. In Figure 1, the amplitude spectrum of pink noise modulated by a 10 kHz harmonic signal is depicted.

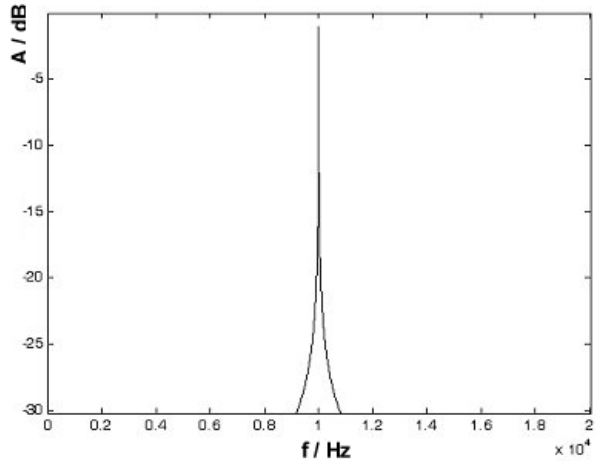

Figure 1. Pink noise modulated by a 10 kHz harmonic signal

Since multiplying shifts pink noise to a frequency equal to the frequency of the carrier (harmonic signal) f_c, the spectrum of the modulated signal has a peak at a frequency equal to the frequency of carrier. Above f_c, the spectrum of the modulated signal has a negative slope, which is not 3 dB per octave. Under f_c, the spectrum is its own reflection from above f_c.

Modulated pink noise, as we call it, has so far very little in common with pink noise. Anyway, considering the assumption that the central frequency of the signal resembles the elevation of the sound source, multiplying pink noise with a harmonic signal of arbitrary frequency could manifest in a varying perception of elevation.

White noise filtered by bandstop filters with different bandwidths. Since the "pinna notch" is one of the primary localization cues, six experiments using white noise filtered with bandstop filters were performed with a different bandwidth in each. Bandwidth remained constant in each case, and illusions of sound source elevation were therefore created by variation of the filter's central frequency. Bandwidths used were: 30 Hz, 600 Hz, 1 kHz, 2 kHz, 3 kHz and 4 kHz.

Pink noise filtered by lowpass filter. Sound stimuli were processed by lowpass filtering. A filter with a certain cut-off frequency was used for each sound stimulus, so information about elevation is contained in the bandwidth. In the acoustic image of the space, sounds with narrower bandwidth would represent low elevations and sounds with broader bandwidth would represent higher elevations.

White noise, filtered by lowpass filter. Reasons for including this sound source into the experiment were the same as for lowpass-filtered pink noise. The main difference between the two signals is in their spectral content; therefore, we may further observe the impact of low frequencies on elevation perception.

Harmonic sine signal. Sound source elevation illusions in this experiment were created by varying the frequency of the sound signal. This case represents the most direct implementation of elevation coding by frequency. We expect that such coding should be very effective in the vertical dimension, but also ineffective in the horizontal direction.

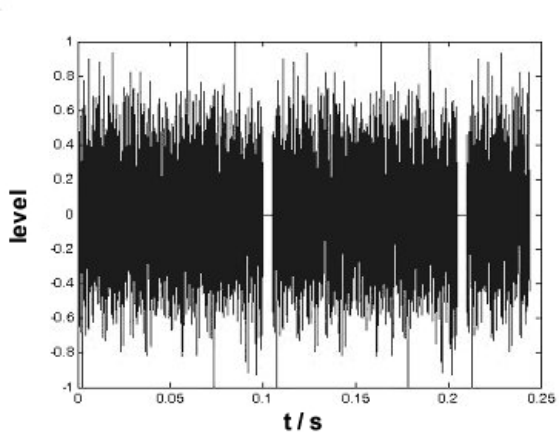

Figure 2. Test sequence – test signal followed by a 5 ms period of silence

The duration of all test signals was 100ms. Test signals were played in sequences where the signal was followed by a 5 ms period of silence as depicted in Figure 1.

Test persons, environment and apparatus

24 volunteers between 17 and 30 years of age participated in each experiment. None of the subjects involved in the experiment had previous experience with spatial sound created by HRTF and played through headphones. They reported no problems with hearing or seeing.

All experiments were conducted in an anechoic room with dimensions 4m × 6m × 3m (W×L×H). The room had an A-weighted ambient background noise level of 27 dB SPL, measured with a Lutron SL-4012. Sound signals were generated using an Acer TravelMate 4000 notebook computer with a Digigram VXpocket 440 sound card. We used Sennheiser Control HD270 headphones.

Experiment

The aim of the experiment was to determine noticeable differences between sounds in each set of sounds. A custom-made computer application was made in Microsoft Visual Basic 6. The application was used to play sound signals processed in Matlab 6. Since only changes in elevation were being considered in the present experiment, the test environment in our research was limited in the vertical dimension (elevation) from –40° to 90°. Azimuth (horizontal dimension) was set to 0°, which is straight in front of the listener.

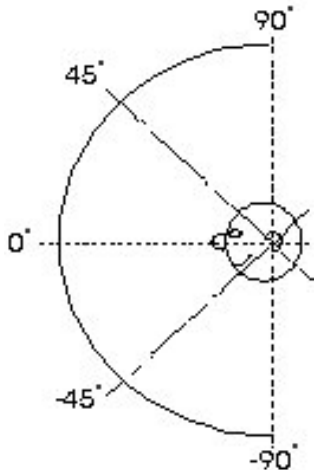

Figure 3. Vertical dimension – elevation

During the experiment, test subjects were seated in an anechoic room facing a wall. First, every test subject was introduced to spatial sound played through headphones. An example with sound moving around subject's head, changing its elevation and azimuth, was played. After one minute, the subject was told to pay particular attention to sound source position in the vertical dimension. The whole introduction procedure lasted 2 minutes. Sound intensity was adjusted for each subject separately to fit each subject best. Therefore, sound played through the headphones had an A-weighted level between 65 dB SPL and 70 dB SPL, measured at the headphones' membrane.

Sound signals imitating different sound source elevations were played successively upward, beginning with sound simulating the lowest elevation. Every virtual position of the sound source was played for no more than 3 seconds. It was randomly determined at every elevation whether the next elevation would be the same, lower or higher. The subject's task was to raise a thumb whenever he or she heard a sound source position raised in elevation or to point the thumb downwards whenever he or she heard the sound source position drop in elevation. If subject was unable to hear the difference between two sounds (let's call them $x_0(t)$ and $x_1(t)$) representing neighbouring elevations (h_0 and h_1), sound $x_0(t)$ was played again and then the step in elevation was doubled, i.e., the next sound played was $x_2(t)$. If the subject heard a difference between $x_1(t)$ and $x_0(t)$, which could happen when $x_0(t)$ was repeated, it was considered that the basic step difference was heard. The algorithm for the test procedure is depicted in Figure 3.

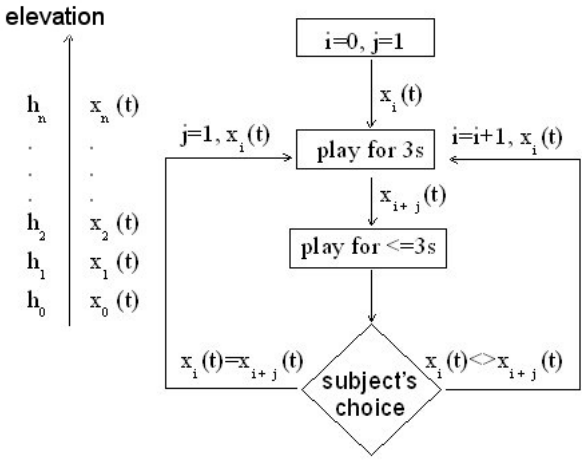

Figure 4. Test procedure algorithm

The whole test procedure for each subject lasted approximately 20 minutes.

RESULTS

Approximately 2300 samples were obtained during the experiment. Each sample is an answer to the question of whether the difference in sound stimuli was heard or not. All cases where test subjects signalled changes with raised or lowered thumb as well as cases where the test subject was unable to signal anything, i.e., when the sound source remained the same or the listener did not hear any difference, are included in the total number of answers.

Results in the following sections are represented with a mean value $\bar{x}$ calculated from N samples x_i , where i is index, by Equation 2.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (2)$$

Results for particular signals, except for white noise filtered by bandstop filter with a bandwidth 30 Hz, are presented graphically in Figures 5 to 13. Results for white noise filtered by a bandstop filter with bandwidth of 30 Hz are not presented because no one of the test subjects was evidently able to observe any difference between different sounds of that type.

Results in Figures 5 to 13 can be interpreted as a representation of "just noticeable differences" (JND) in frequency for specific signals. In contrast to JND defined in Zwicker, Flottorp & Stevens (1957) where signals are harmonic, the signals presented here have a certain bandwidth. Since difference in frequency, i.e., differences between central frequencies of filters, bandwidth of filter, etc., was constant, the results show frequency ranges where JND is lower or greater than 250 Hz.

For each set of signals, another piece of information in the figures is a representation of elevation positions dependent on the central frequency of the filter, i.e. each elevation position is represented by sound signal with certain central frequency. Let us denote the sound signal filtered by a filter with central frequency 1 kHz to have elevation position N=1. Sound signals filtered by filters with higher central frequencies should therefore be denoted by N>1 respectively. For lower frequencies where JND is lower than our basic step of 250 Hz, it is clear that a 250 Hz step will be sufficient for increasing elevation up to certain frequency. For higher frequencies, where JND is greater than 250 Hz, greater frequency steps are demanded. Results can be read from the figures. The figures also show the number of elevation positions if such coding were used in an acoustic image of space. The corresponding elevation resolution can be found in Table 1.

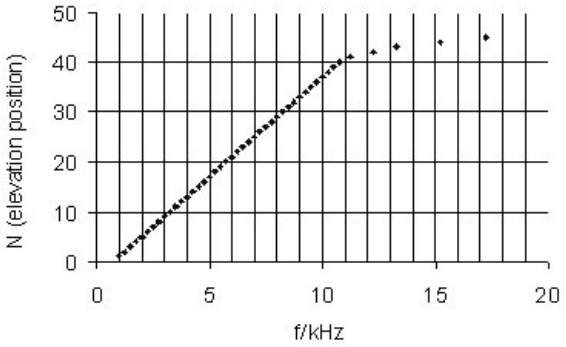

Figure 5. Modulated pink noise, up to 45 different elevation positions can be achieved

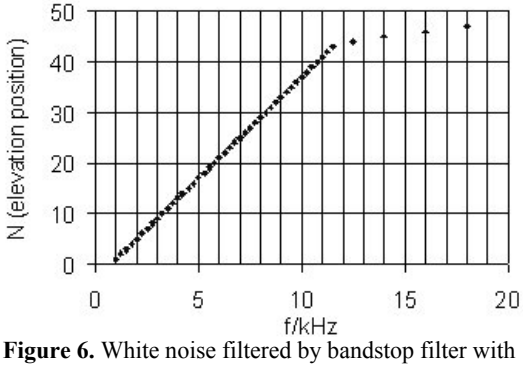

Figure 6. White noise filtered by bandstop filter with bandwidth 600 Hz, up to 47 different elev. pos. can be achieved

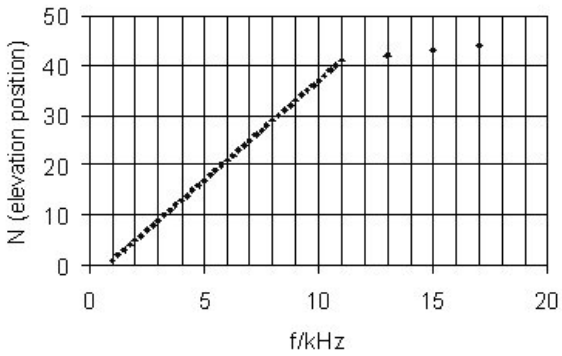

Figure 7. White noise filtered by bandstop filter with bandwidth 2 kHz, up to 44 different elev. pos. can be achieved

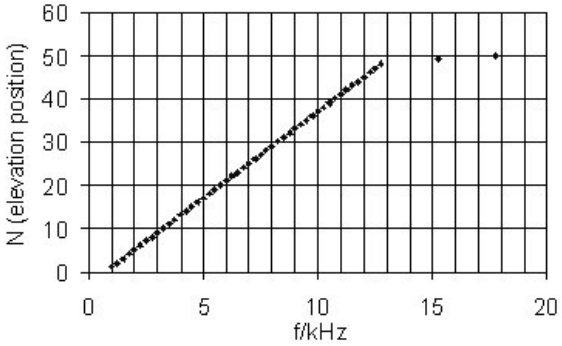

Figure 8. White noise filtered by bandstop filter with bandwidth 2 kHz, up to 50 different elev. pos. can be achieved

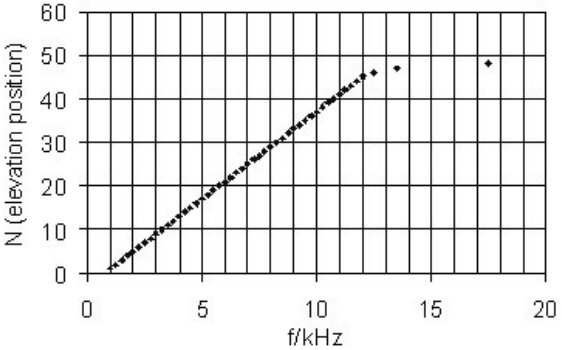

Figure 9. White noise filtered by bandstop filter with bandwidth 3 kHz, up to 48 different elev. pos. can be achieved

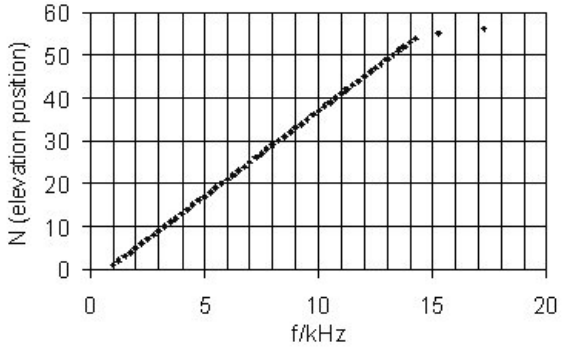

Figure 10. White noise filtered by bandstop filter with bandwidth 4 kHz, up to 56 different elev. pos. can be achieved

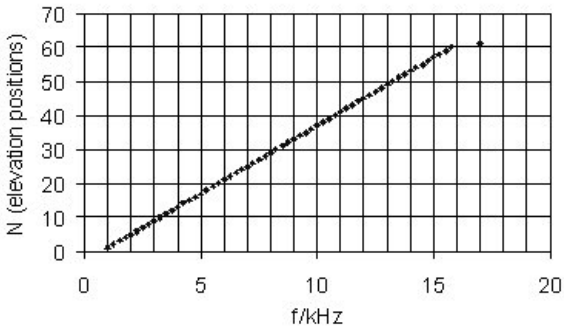

Figure 11. Pink noise filtered by lowpass filter, up to 61 different elevation positions can be achieved

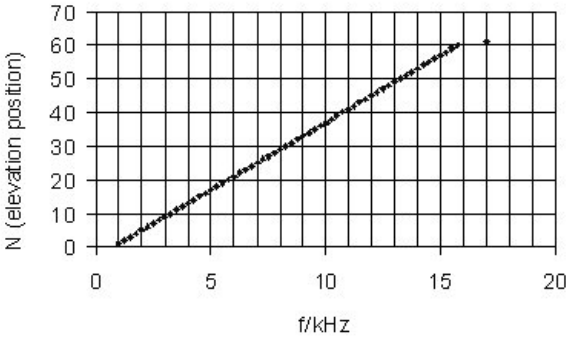

Figure12. White noise filtered by lowpass filter, up to 61 different elev. pos. can be achieved

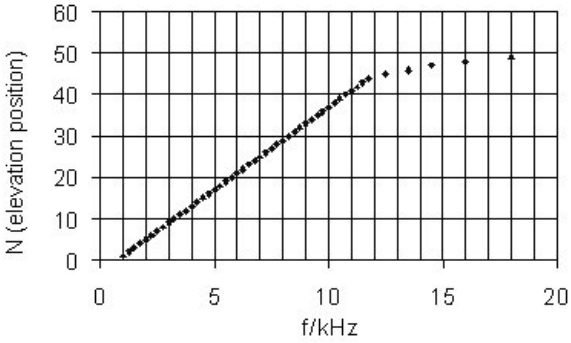

Figure 13. Harmonic sine signals, up to 46 different elev. positions can be achieved

As observed by Nandy & Ben-Arie (1996), only coding with personalized HRTFs demonstrates absolute relations in the vertical dimension; using alternative coding would require the user to learn it. If it is possible to learn such coding where the only restriction is JND in frequency and the elevation range is –40° to 90°, then Table 1 shows elevation resolution.

Table 1. Elevation resolution for; N - number of perceivable elevation positions in range from –40° to 90°

Sound	N	Resolution/°
Modulated pink noise	45	2,95
White noise, bandstop (B=30Hz)	0	-
White noise, bandstop (B=600Hz)	47	2,83
White noise, bandstop (B=1kHz)	44	3,02
White noise, bandstop (B=2kHz)	50	2,65
White noise, bandstop (B=3kHz)	48	2,77
White noise, bandstop (B=4kHz)	56	2,36
Pink noise, lowpass	61	2,17
White noise, lowpass	61	2,17
Sine	46	2,89

DISCUSSION

In this experiment we proved that the spectral content of sound affects perception of sound source elevation and is therefore applicable to be used as elevation coding in acoustic imaging of space.

From the data in Figure 1 and Table 1 we can extract the most suitable candidates for elevation coding, which should be treated as a combination of a certain sound signal and signal processing method; therefore, we can evaluate the suitability of each signal processing technique and sound signal for elevation coding. In our experiment it turned out that simulating "pinna notch" with properly designed bandstop filters, lowpass filtering and modulation of signals with non-uniform spectral content may satisfy our goals.

As previously mentioned, pink noise has non-uniform frequency-dependent amplitude density. The peak of amplitude density varies with frequency when using modulation. Therefore, the central frequency carries information about the elevation of the sound source. A 250 Hz frequency step is suitable up to 11 kHz; above that, it should be increased to 1 kHz and 2 kHz. 45 perceptible positions in elevation are achievable with modulated pink noise.

White noise filtered with bandstop filters confirms "pinna notch" elevation dependency in consideration of stopband bandwidth importance. A notch filter is suitable in individualized HRTFs, but in other cases the effect of the "pinna notch" should be emphasized.

Lowpass-filtered pink and white noise appeared as the most separable sound signals. They were both separable up to nearly 16 kHz, which equals 61 elevation positions. If we take into consideration the elevation range from –40° to 90° then this would mean separability by 2.17°, much better than the separability in a real environment, which was recognized to be approximately 6° (Sodnik et al. 2005).

Sine signals represent the most straightforward method for coding with frequency. Results are as expected from Zwicker, Flottorp & Stevens (1957). Such signals are very annoying to listen to, and there is another problem of localizing such signals in the horizontal dimension, which should also be tested with the other signals used in the experiments described here.

Elevation coding on the basis of signal processing without considering HRTFs could be described as artificial elevation coding, as opposed to natural elevation coding, which resembles the reflections and diffractions from body, pinna etc. described with HRTFs. The problem with artificial elevation coding is in perception of absolute elevation. It is claimed (Nandy & Ben-Arie 1996) that only relative differences in elevation can be perceived by artificial coding. On the other hand, in his project "The vOICe – Seeing with Sound", Meijer (1992) has shown that the brain is a very flexible organ that can even learn absolute perception of elevation coded by frequency.

ACKNOWLEDGMENTS

This work was supported by Ministry of High School and Science of Slovenia within a project Algorithms and optimization methods in telecommunications.

REFERENCES

Algazi, VR, Avendano, C, Duda, RO 2001, 'Elevation localization and head-related transfer function analysis at low frequency', *Journal of Acoustical Soc. of America*, vol. 109, no. 3, pp. 1110 – 1122.

Blauert, J 2001, *Spatial Hearing, The Psychophysics of Human Sound Localization*, MIT Press.

Bloom, PJ 1977, 'Creating Source Elevation Illusions by Spectral Manipulation', *Journal of the Audio Engineering Soc.*, vol. 25, no. 9, pp. 560 – 565.

Bronkhorst, AW 1995, 'Localization of real and virtual sound sources', *Journal of Acoustical Soc. of America*, vol. 98, no. 5, pp. 2542 – 2553.

Hofman, PM, Van Opstal, AJ 1998, 'Bayesian reconstruction of sound localization cues from responses to random spectra', *Biological Cybernetics*, vol. 84, no. 4, pp. 305 – 316.

Jin, C, Corderoy, A, Carlile, S, van Schaik, A 2004, 'Contrasting monaural and interaural spectral cues for human sound localization', *Journal of Acoustical Soc. of America*, vol. 115, no. 6, pp. 3124 – 3141.

Just Noticeable Difference (JND) for Three Frequencies, Retrieved: June 9, 2005, from http://webphysics.davidson.edu/faculty/dmb/soundRM/jnd/jnd.html

Langedijk, EHA, Bronkhorst, AW 2002, 'Contribution of spectral cues to human sound localization', *Journal of Acoustical Soc. of America*, vol. 112, no. 4, pp. 1583 – 1595.

Meijer, PBL 1992, 'An Experimental System for Auditory Image Representations', *IEEE Trans. on Biomedical Engineering*, vol. 39, no. 2, pp. 112 – 121.

Nandy, D & Ben-Arie, J 1996, 'An auditory localization model based on high frequency spectral cues', *Annals of Biomedical Engineering*, vol. 24, no. 6, pp. 621 – 638.

Sodnik, J, Susnik, R, Bobojevic, G & Tomazic, S 2004, 'Smerna locljivost navideznih izvorov zvoka pri cloveku', *Electrotechnical Review*, vol. 71, no. 3, pp. 121 – 127.

Sodnik, J, Susnik, R, Stular, M & Tomazic, S 2005, 'Spatial sound resolution of an interpolated HRIR library', *Applied Acoustics*, vol. 66, no. 11, pp. 1219 – 1234.

Tsutsui, K, Suzuki, H, Shimoyoshi, O, Sonohara, M, Akagiri, K, Heddle, RM 1992, 'ATRAC: AdaptiveTransform Acoustic Coding for MiniDisc', *Proceedings of 93rd Audio Engineering Society Convention*.

Watkins, A 1978, 'Psychoacoustical aspects of synthesized vertical local cues', *Journal of Acoustical Soc. of America*, vol. 63, no. 4, pp. 1152 – 1165.

Ziberna, G, Zazula, D 2003, 'Racunalniska tvorba 3D zvoka v virtualnih prostorih', *Electrotechnical Review*, vol. 70, no. 3, pp. 96 – 102.

Zwicker, E, Flottorp, G & Stevens, SS 1957, 'Critical bandwidth in loudness summation', *Journal of Acoustical Soc. Of America,* vol. 29, no. 5, pp. 548-557.