IEEE-INNS Intern. Joint Conf. on Neural Networks (IJCNN 2001), Washington, 2001, pp. 1132-1137, IEEE/Omnipress, 2001

Model and Application of a Binaural 360° Sound Localization System

C. Schauer, H.-M. Gross
Dept. of Neuroinformatics, Ilmenau Technical University, D-98684 Ilmenau, Germany

Abstract

We introduce a biologically inspired sound localization system, based on an open two-microphone configuration. Its purpose is to perform a robust, 360-degree detection of objects, in particular humans, in the horizontal plane. In our approach, we consider neurophysiological findings to discuss the biological plausibility of the coding and extraction of spatial features, but also meet the demands and constraints of a practical application in the field of human-robot interaction. We are processing Interaural Time Differences for the left/right detection and spectrum-based features to discriminate between in front and behind. Tests in real environments are used to demonstrate the function and the application of the system. We will briefly touch the point of handling different acoustic room characteristics and give reasons why the proposed model can show advantages over conventional correlation methods.

1 Introduction

In recent years a lot of promising work on the problem of spatial hearing has been published – many investigations and models of auditory perception exist from neurobiology to psychoacoustics [5, 11, 2]. However, although numerous applications in robotics and human-machine interaction are imaginable, only a few working examples are known. There could be different reasons for that: on the one hand, the models normally can include only a few details of the complex neural coding and processing mechanisms in the real auditory system. On the other hand, when targeting at localization systems working in everyday environments, many acoustic effects arising from very different acoustic characteristics must be faced. Without using dummy-head recordings, our own sense of hearing is often insufficient, when we try to evaluate spatial information in the playback via headphones or stereo speaker systems. Hence, our approach is to model a certain biologically inspired architecture, which is well understood and based on reliable findings, and discuss its behavior and capabilities in varying real–world scenarios, e.g. the problem of human-robot interaction.

2 Biological Background & Model Concept

In biology sound localization means the evaluation of the horizontal and vertical directions and calculation of the distance of a sonic event. Thereby the nature of the sound is just as important as who is listening: sometimes a rough idea about the direction is sufficient to be alarmed or attempt to escape whereas another *listener* is interested in the exact point of origin. Since the categories azimuth, elongation and distance are more or less important to an animal, it is not surprising that different species solve the localization task with varying expense and precision. We have to take this into account, when we interpret morphological findings and compare, e.g., the limited sound localization of the chicken with a highly specialized hunter like the barn owl.

2.1 Acoustic Effects & Auditory Features

To evaluate spatial information, the auditory system utilizes acoustic effects caused by a varying distance between the sound source and the two ears and the shape of the head and body. We can categorize these effects in intensity differences and time delays. Since the processing in the central auditory system is distributed in parallel frequency bands, a varying sound level is represented by spectral differences, interaural time delays by phase differences. Beside those binaural cues, monaural mechanisms contribute to localization, especially if the sound source is located in the medial sagittal plane where interaural effects can hardly occur. We use the terminology currently found in literature and specify the following spatial cues as candidates for our modeling: (i) Interaural Time Differences (ITD), (ii) Overall Interaural Intensity Differences (IID) and (iii) Sound Color and Tuning Frequencies, according to direction–depending Head-related Transfer Functions (HTF).
As mentioned, the success or precision of sound localization also depends on the type of a sound event, its dynamic and spectral shape. Low frequencies, whose wavelengths are large compared to the size of the head and the pinnae, reach the two ears undamped with almost the same intensity. The only reliable spatial information to be found in low–frequent sounds is an ITD

(for humans up to 0.65ms), if the sound is heard from the left or the right. The problem with ITD processing is, that it is limited to the lower frequency range. If the wavelength is shorter than the ear distance, the ITD representation by phase differences becomes ambiguous. According to a maximum ITD of 0.65ms this problem occurs above approximately 1.5 kHz.

Nevertheless, we can estimate the direction of those sounds because of another effect: the higher the frequency, the less the sound waves can bend around the head – IIDs arise and become the major spatial cue. In [1] a comprehensive study of sound localization based on different types of binaural and monaural information is presented, including findings about the localization blur: The achieved precision in the horizontal plane corresponds conspicuously to the relation of azimuth angle variation and ITDs.

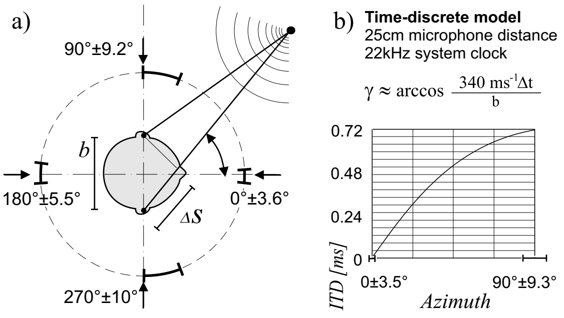

Figure 1: Localization blur in the horizontal plane:
a) noise-burst experiments (according to[1]).
b) Approximation of ITD–based localization in a time-discrete model.

2.2 Model Architecture

Although we are not able to simulate the tricky but advantageous merging of multiple spatial cues, we expect that already the analysis of ITDs will provide sufficient information to solve many localization tasks. Our assumption is based on the following ideas: (i) ITDs seem to be the most important cue to horizontal localization (except for front/behind discrimination) which we are interested in. (ii) ITD analysis provides a high azimuthal resolution and works best with onsets and low–frequent or broadband sounds like the human voice. (iii) In contrast to the processing of IIDs, sound color and tuning frequencies, an ITD–based system requires no dummy–head technique. (iv) Especially for ITD-processing we can account for numerous neurobiological findings.

Our work on real–world–capable ITD processing is similar to Lazzaro's neuromorphic auditory localization system [9], but follows a more pragmatic approach. In our simulations, we use digital algorithms for the preprocessing and coincidence detection within spike patterns, as well as an uniform spike–response neuron in the other parts of the model. The motivation to use spikes is, that ITD calculation requires a timing that is more precise than the description of neural activities by firing-rates. The applied spike response neuron model is inspired by Gerstner's work [6] and takes up fundamental properties of biological cells: the spatial and temporal integration of stimuli via postsynaptic potentials (PSP) in the dendritic tree, the generation of an action potential when reaching a threshold, and the effect of diminished sensitivity during a period of refraction. An absolute refractory period and axonal delays are not modeled. To describe the impulse response of a synapse (PSP), we chose the so called α–function $f_\alpha(t) = \frac{t}{\tau} e^{1-\frac{t}{\tau}}$, the afterhyperpolarization (AHP) follows a simple exponential fading function. The combination of these potentials results in neural dynamics, which are more complex than that of leaky integrate-and-fire models.

3 Components of the system

The ITD-processing system was extended by a simple in front/behind discrimination, so that we can describe the overall architecture by four stages:

Filtering and spike coding: The analog signals from two microphones are filtered by a cochlear model (all–pole–gammatone filter) [13] and coded into spikes.

ITD calculation: For every frequency channel the spike patterns from left and right are cross–correlated. The time–code of binaural delay is transformed into a place code representing interaural phase differences.

Mapping and selection: The resulting pattern is projected onto a non–tonotopic representation of the azimuthal locations of sound sources. As the result of a winner-take-all process, only one direction is dominant in the final representation.

In front/behind discrimination: With the help of a special microphone configuration an estimation of interaural spectral differences determines the in front or behind orientation. A 360°-map is formed.

3.1 Filtering and spike coding

The frequency analysis in the cochlea as the basis for the tonotopic organization of the auditory pathways is realized by an all–pole gammatone (APG) filter cascade [13]. With respect to the broadband tuning in the auditory nuclei involved in ITD processing [5] we calculate 16 logarithmically arranged channels in the relevant frequency range from 100 Hz to 2.5 kHz from the digitized microphone signals. The output of the filter corresponds to the mechanical properties of the cochlear basilar membrane and must be transformed into a neural response, the specific timing of spike trains in the auditory nerve. For this we use a receptor model, simulating the interaction of inner hair cells and ganglion cells [1]. Since their firing is connected to the movement cycles of the basilar membrane, the resulting spike pattern shows the effect of *phase locking* on the acoustic stimulus [12].

3.2 Cross-Correlation

The majority of the known ITD–detectors is based on Jeffress' coincidence model [7], whose basic idea is the cross-correlation of corresponding frequency-bands in a highly specialized neural structure. The time-window, necessary for the correlation function, is realized by counterpropagating neural delay lines. This thesis written about 50 years ago is still up to date and corresponds to findings in the auditory brainstem: the evidence of binaural delay line–structures seem to be clearly connected to ITD-processing.

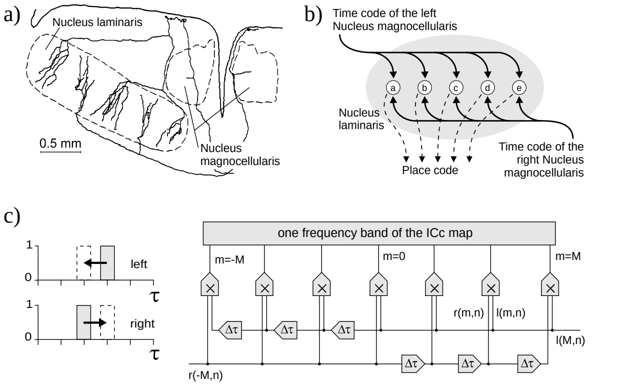

Figure 2: a) Dual delay–line structure in the brainstem of the barn owl [3]. b) Model of the coincidence mechanism. c) Asynchronous shifting is used to prevent the *missing of spikes* in the time-discrete delay-lines. This model is more similar to single delay line structures of mammals

According to Jeffress' coincidence model, the calculation of ITDs is realized by counterpropagating axonal delay lines. Coincidence cells, located at different positions along the axons, generate spikes if they receive a simultaneous stimulation from the left and the right hemispheres. Because of the different time delays, depending on the length of the propagating fibers, each cell becomes sensitive to a certain ITD. In this way, the temporal information of ITD is transformed into a place code, represented in the spatial distribution of the activity in a neural structure.

3.3 Mapping and selection

Even under advantageous conditions, e.g., an outdoor experiment with just one sound source, no noise or reverberation sound, the ITD representation is never a perfect single peak but contains local maxima in a flashy symmetrical arrangement. Since the coincidence detection is similar to the calculation of the cross-correlation of periodical signals, its result is just as periodical. Therefore it is easy to associate the displacement of the local maxima to periodical components of the acoustic stimulus [2]. In the context of localization it is the feature of tonotopy to distinguish ITDs from ambiguous phase differences by a recombination of frequency bands. Phase differences are located at different positions in the ITD map depending on the characteristic frequencies. In a convergent projection from many frequency bands they produce a diffuse activation. The position of the detected ITD is independent of the tonotopic organization and gives rise to a less ambiguous feature (figure 3). The idea of a summation of the tonotopic response is strongly supported by findings in the Inferior Colliculus (IC) of the barn owl, where ambiguous activations of single high frequency bands of the central IC, but a definite response in the non-tonotopic extern IC could be observed [9].

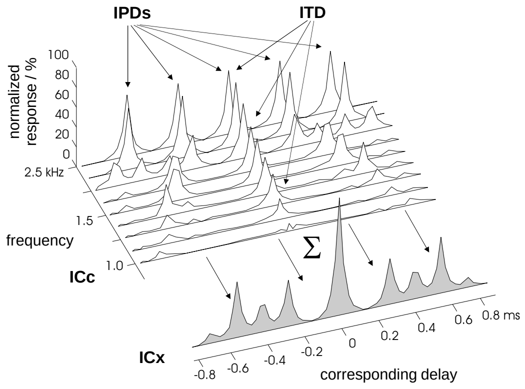

Figure 3: The combination of the tonotopically distributed response of the coincidence detector in an one-dimensional IC-model enables a distingtion between ITD and phase differences (IPDs).

The one-dimensional response is often disarranged by interferences with other sources, echos, or ambiguities which could not be suppressed yet. To obtain reliable results, we need to simulate a focusing mechanism, that selects a dominant ITD. Our model uses a structure

[1] The receptor model is similar to the spike response model with, the exception, that the output of the cochlear filter is used as a generator potential instead of PSPs at the dentrites.

[2] In simulations one actually has to face two kinds of periodical inputs: (i) the characteristic frequency and (ii) the maximum spike frequency of the receptor cells. Because a delay line of the coincidence detector is activated by just one neuron, the response of the coincidence cells codes also the periodicity within spike bursts from the receptors. In contrast to biology, this yields additional local maxima in the ITD map.

containing lateral and self excitation and an interneuron which integrates the instantaneous activity of the net and generates recurrent inhibition to all cells. In the resulting winner-take-all (WTA) process only a single region of dominant feature representation can maintain activity [8]. For the application to dynamic acoustic scenes, the network is capable of moving the focus of attention to moving or new sound sources.

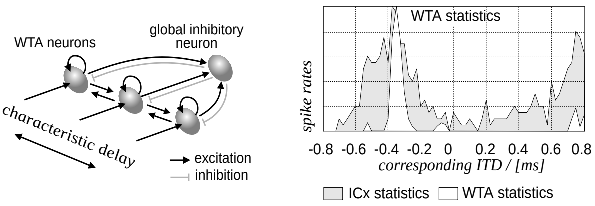

Figure 4: Left: Structure of the WTA network.
Right: Result of the selection process.

3.4 In Front or Behind

We believe the description of the ITD–processing of our model to be plausible and conclusive. However, at this point, we usually had to face one quite unpleasing question: What's the use of this expensive, biologically inspired ITD–model, if we cannot solve the very simple but important problem: Is what I hear right in front or behind me? As mentioned, this task requires the processing of HTFs, including cues in the monaural spectrum, which is perhaps the most complicated part of spatial hearing. Although multi-cue approaches are becoming more and more relevant for artificial hearing systems [14], no concrete applications are known. Especially if no dummy–head recordings are available, the standard solution is to calculate multiple time–delays in larger microphone arrays.

But, do we really have to model HTF-analysis to achieve a horizontal 360°–localization? The precision of the pure ITD–system is already in the range of the general human capabilities in spatial hearing.The only thing missing is simply to discriminate between front and behind. But what means simply - we would need to know the unaltered sound without head–related transfer as a reference, expert knowledge about our very own HTFs and tuning frequencies or the memory of how typical acoustic events sound in front or behind us. Since we are not able to model this amazing capability, we state the question in a slightly different way: Is it possible to generate a reference signal with a sound color typical for the opposite direction, that means the biologically little plausible idea to have one ear directed forward and the other one backward. By comparing the left and right spectrum of such a stereo–signal we could answer the front/behind question – without knowledge about the sound, memory of a reference sound and even without detailed HTF-analysis. If we utilize the acoustic bending effect, that is limited to low frequencies and build a frequency–dependent directional microphone characteristic, we can specify that the more light sound color belongs to the microphone pointing forward, the more hollow sound comes from the backward microphone.

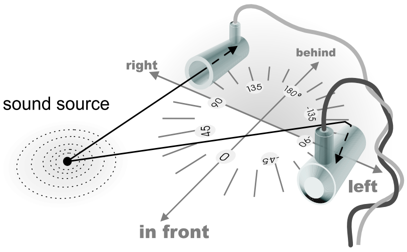

Figure 5: Simple microphone configuration for the spectrum-based front/behind discrimination

We decided to use little tubes, like artificial ear channels, in a configuration shown in figure 5. To compare the sound color of the binaural signals we determine the actual spectrum in the left and right channel, e.g. by a short-term FFT in time windows of n samples. For each FFT–vector $[f_1 \dots f_n]$ we calculate a value $C = \frac{1}{n}\sum_{i=1}^{n} i \cdot f_i$, the *center of mass* of the spectral shape. It is level-independent and does not correspond to a real frequency in the signal, but, a simple difference of the left and the right C-value is realizing the comparison of the spectral shapes: $C_{left} < C_{right} \hookrightarrow infront$; $C_{left} > C_{right} \hookrightarrow behind$. If the values are very similar, a front/behind discrimination is not necessary, because the source is about 90° left or right and the ITD-calculation alone is providing a definite result. The short–term spectral comparison can be unstable due to window effects of the FFT–calculation. By using samples up to 250 or 500ms long, we can prevent this but the disadvantage is a delayed response of the whole system.

4 Simulations and results

4.1 Offline simulation

Firstly, the localization system was tested offline with data recorded in an open environment including background noise but only little echo effects. Narrow and broadband sounds, including numerous speech signals, were recorded by 2 microphones (omni-directional characteristic, base distance b=0.25 m).

The localization of single sources was comparatively simple and robust – the directions of all tested broadband sounds were determined correctly [3]. Figure 6

[3] A successful localization means the emergence of an unequivocal winner–neuron in the WTA–layer at the theoretically cor-

illustrates the dynamical focusing on a moving source, emitting pink noise. While the coincidence pattern displays a diffuse activation and disturbances, the capability of the WTA network to detect a dominant ITD leads to a clear feature representation. The focus is stably locked to the correct direction, even if the sound source is moving, which is an important feature of the strong WTA dynamics (figure 6).

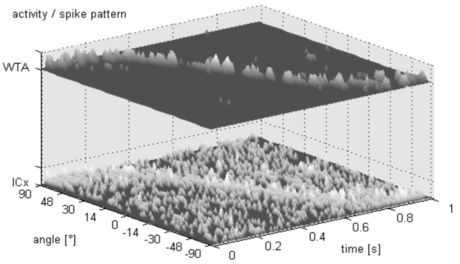

Figure 6: Model behavior for a moving source emitting pink noise. Visualization of the IC output (bottom) and the WTA-layer (top).

If multiple sources are present in the acoustical scene, the requirements to the localization system change considerably. Because of interferences between periodical sound components, the dominance of a certain source has to be caused by its intensity or broader spectral constitution. The experiment shown in figure 7 demonstrates, how the focus of attention is shifted from a narrowband sound toward a voice stimulus.

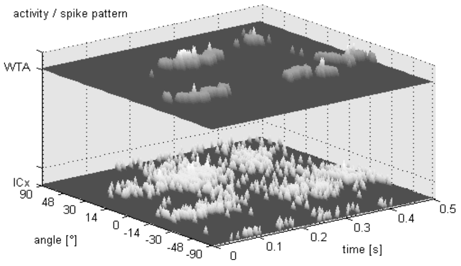

Figure 7: Localization of a continual narrow band sound and a human call setting in about 100 ms later.

4.2 Online simulation

Because the synchronous simulation is still not real-time capable and acoustic problems occur in smaller rooms, we use only a short block of samples from the signal's onset for online processing. The simulation is fast enough to trigger a turn-reflex of the active–vision head of a robot and showed robust results on hand-claps and similar signals. In large rooms even human speech can be localized reliably.

Several experiments were run in a shopping center with only little reverberations but a high background noise and in an empty, relatively echoic lecture hall. We noticed that the WTA process is able to focus on a sound source in about 10ms – often unaffected by the first echos reaching the microphones. For most broadband signals, this time is longer than the arrival of a first wavefront, which has been considered as the longest part of reverberate signals we can localize. But only if a voiced sound hits the resonance frequency of a small room (in our recordings resonances build up after 30ms or later), the focus of the WTA layer may be shifted to an apparently random position of an interference. This way we can model major aspects of the precedence effect – the dominance of the original sound event over its echos.

The successful processing of voiced, reverbereated sounds demonstrates the superiority of the proposed model over conventional correlation methods. The result of a simple cross-correlation of more or less sinusoidal signals is a sine itself, that means ambiguous with a smooth, hard to derive maximum. The maximum search in a mean response strongly depends on the size and position of the applied time window. Our model, too, suffers partly from the periodicity of sine-components of a sound but benefits from some special properties: (i) The spike-pattern mainly codes temporal information like phasing - amplitude is coded indirectly by the spike rate - the effect is a sharp peak as the result of the coincidence detection instead of a smooth sine–response. (ii) The broadband response of the cochlea filter together with the tonotopically distributed processing and the recombination of frequency bands effectively prevents ambiguous responses. (iii) Because competing sound sources, noise or the disturbance by interferences need energy and time to shift the focus in the WTA–layer, the system shows a hysteresis property and thus prefers the signals onsets. Based on these qualities the system integrates mechanisms of onset, transient, and ongoing sound processing, and realizes the localization beyond the first wavefront. Due to the hysteresis of the WTA-layer, we become adaptive to the overall loudness and sensitive to onsets or sudden changes. In practice, we gain the advantage of achieving the correct result without having to determine the exact time of onsets and echos.

4.3 Building a 360°–map of horizontal angles

No matter how precise and reliable the ITD–based localization might determine angles from -90° to +90°, a universal application, e.g. on a mobile robot, would fail

rect position. Corresponding to the inaccuracy of the angle–measurement and the localization blur, also adjacent cells are considered to be correct winners.

without the ability to discriminate between in front and behind. Because we could not answer conclusively the question of how to map the different spatial information, especially since the front/behind detector has less biological background, we decided to run the separated algorithms in parallel. The front/behind prediction is simply used to interpret the final WTA-output of the ITD system.

The tests included all sorts of common sounds (clicks, hand claps, voices, pink noise) and were performed in an empty lecture hall. As a good example for a moving sound source, the processing of a 12 word long sentence is shown (figures 8), where the tracked speaker position is traveling once around the microphones.

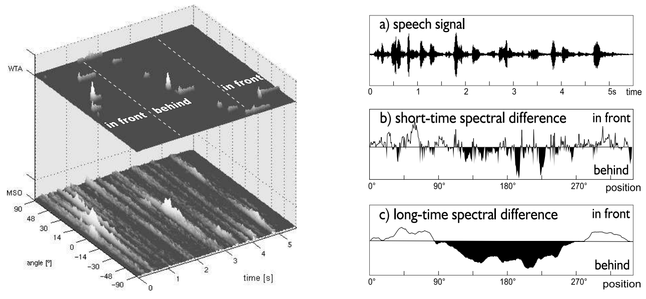

Figure 8: Left: 360-degree localization of a speaker moving around the microphones. Right: a) Speech–signal, b) short–term (12.5 ms) and c) long–term (0.5 s) spectral difference.

Also the front/behind discrimination was additionally testet in the shopping center. In comperatively quiet situations the obtained ratio between the speech signal of a user (1-4 meter distance) and background noise or other voices was at about 5-6dB. The command–words and hand–claps of the user were detected with a propability of 81% and a precission of +/- 10° (90% within +/- 20°).

5 Conclusion

Summarizing our experiments, the presented model yields convincing results in open environments. Ambiguities and disturbances in the ITD representation at the level of the IC are successfully suppressed by the WTA process. Thereby the simulation of the spike-based selection model proved to be uncomplicated – the limited dynamic range of a spike coded WTA input enables a robust operation of the network. While testing the system in complex acoustic situations, it turned out that an ITD-dependent differentiation between multiple sources, as a typical application of the model, has to be based on a sequential calculation and selection of ITDs (see figure 7). Because the ITD feature is locked to the phase of the signal, the localization fails if interferences occur between voiced sounds. Thus, even a realtime–capable implementation will require an onset detector to deliver reasonable results, and the one restriction which remains, is that our system is limited to relatively quiet situations and can not perform the so-called cocktail party–effect.

With the extension of the front/behind detector we have, for the first time, a suitable tool to perform a horizontal 360° sound localization on a low-cost two-microphone platform.

References

[1] Jens Blauert. *Spatial Hearing : The Psychophysics of Human Sound Localization.* MIT Press, 1996.

[2] Jens Blauert. Process and Trends since 1982. In [1]. pages 394–409.

[3] C.E. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience*, 10:3227–3246, 1990.

[4] G. Ehret and R.Romand The Central Auditory System. Oxford University Press, New York 1997.

[5] G. Ehret *The Auditory Midbrain, a "Shunting Yard" of Acoustical Information Processing.* In [4], 259-316.

[6] Wulfram Gerstner, Richard Kempter, J. Leo van Hemmen, and Hermann Wagner. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383:76–78, September 1996.

[7] L.A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol.*, 41:35–39, 1948.

[8] Samuel Kaski and Teuvo Kohonen. Winner-Take-All Networks for Physiological Models of Competitive Learning. *Neural Network*, 7:973–984, 1994.

[9] John Lazzaro and Carver Mead. A silicon model of auditory localization. *Neural Computation*, 1(1):41–70, 1989.

[10] A. Popper, R. Fay. The Mammalian Auditory Pathway: Neurophysiology. Springer, New York 1992.

[11] Eric M. Rouiller. *Functional Organization of the Auditory Pathways* In [4], 3-96.

[12] Mario A. Ruggero. *Physiology and Coding of Sound in the Auditory Nerve.* In [10], 40.

[13] Malcolm Slaney. Lyon's cochlea model. Technical Report 13, Apple, Advanced Technology Group, 1988.

[14] A. van Schaik, C. Jin & S. Carlile. Human Localisation of Band–Pass Filtered Noise. Int. Journal of Neural Systems, vol. 9, number 5: 441–446, October 1999.