Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Modelling of the cochlea response as a versatile tool for acoustic signal processing

Marinus Boone[a], Diemer De Vries[a], Tjeerd Andringa[b], Anton Schlesinger[a], Jasper Van Dorp Schuitman[a], Bea Valkenier[b] and Hedde Van De Vooren[b]

[a]University of Technology Delft, Lorentzweg 1, 2628 CJ Delft, Netherlands
[b]University of Groningen, Dept. Artificial Intelligence, P.O. Box 407, 9700 AK Groningen, Netherlands
m.m.boone@tudelft.nl

The inner ear or cochlea processes the acoustic signals that enter the oval window into a specific time-frequency pattern. Many acoustic signal processing methods are based on this behaviour. A fundamental method is to calculate this time-frequency response by solving the differential equation of the movement of the basilar membrane, followed by a visualisation of the excitation patterns in a time-frequency plot. For that purpose Continuity Preserving Signal Processing (CPSP) is a promising method.
An overview is given of a project that is carried out by TUD (University of Technology Delft) together with RUG (University of Groningen) being sponsored by STW (Dutch Technology Foundation). The project divides into four sub-projects which are closely related: Automatic Keyword Spotting, Machine Analysis and Diagnostics, Speech Intelligibility Enhancement for Hearing Aids, and Quality Assessment of Room Acoustics. Results that have been obtained in the project will be summarised. Detailed results of the sub-projects are presented in separate papeers of this conference.

1 Introduction

In this paper an overview is presented of a project that is presently being carried out by Delft University of Technology (Acoustical Imaging and Sound Control) together with University of Groningen (Artificial Intelligence). The origin of the project stems from the fact that much overlap in interest has been found on the essence of the human hearing which relates back in a great extend to the signal processing of the cochlea. The cochlea can be seen as the acoustic-neural transducer of humans and other natural species.

Research on the acoustic-mechanical principles of the cochlea goes back to the fundamental work of Von Békésy [1], [2], [3]. He made anatomical preparations of the human cochlea and observed the movements of the basilar membrane due to sinusoidal exitation with different frequencies, using microscopic techniques. From his work it was concluded that the basilar membrane can be seen as a distributed second order mechanical system that is activated by the vibrations transferred from the ear drum by the middle ear ossicles to the oval window of the cochlea. The excitation results in a transverse traveling wave along the basilar membrane, which resonates at positions along the membrane, depending on the frequency content of the excitation. This so-called tonotopical behavior forms the basis for the way in which the vibations on the basilar membrane give rise to neural impulses that are transferred to the brain by the auditory nerve and are observed as sound.

Following Von Békésy, the mechanical properties of the cochlea have been studied by many researchers. We mention here Zwislocki [4], Viergever [5], and Netten and Duifhuis [6]. The tonotopical behavior of the inner ear has also been used by several researchers to explain the psychoacoustic behavior of the human ear. The positions and widths of the excitation pattern on the basilar membrane are strongly correlated with the critical bands of Zwicker et al. [6] that are related to loudness perception.

The combination of the tonotopical spread of frequencies along the basilar membrane, in combination with the temporal character of the neural activity, has led to a common practice to visualize sounds as time-frequency plots of amplitudes. Such representations are generally known as spectrograms. They can give a very good insight in the time-frequency content of all kinds of sounds, and are well suited to characterize sound for different purposes.

It is interesting to compare the time-frequency response of the cochlea with different mathematical approaches of time-frequency analysis, where it is usually searched for the minimization of the product of temporal and frequency uncertainty, sometimes called the Heisenberg uncertainty of signal analysis. See Hut el al. [8].

By inspection of sound spectrograms it can be observed that in many cases there are harmonic patterns that last for a certain amount of time. We see this for instance in music, but also in speech and certain kinds of traffic noise. Therefore it makes sense to characterise such sounds by looking at the temporal development of these areas of maximum amplitude response as a function of time and frequency. To be able to do so an efficient calculation scheme of the time-frequency response of the cochlea is needed together with an algorithm to detect and analyse the time frequency maxima as so-called ridges. Much effort on this topic has been given by Van Hengel [9] and Andringa [10].

2 Theory

In this project we use different methods to obtain a time-frequency representation of audio signals.

In its most fundamental form, time-frequency analysis consists of the application of a set of filters $w_i(t)$ where each filter acts as a narrow band filter on the input signal $g(t)$.

Such a filter $w_i(t)$ can be characterized by a time and frequency center and a time and frequency spread, being defined by:

$$\omega_0 = \frac{1}{2\pi\left\|g\right\|^2}\int_{-\infty}^{\infty}\omega\left|\tilde{g}(\omega)\right|^2\mathrm{d}\omega\,, \tag{1}$$

$$t_0 = \frac{1}{\left\|g\right\|^2}\int_{-\infty}^{\infty}t\left|g(t)\right|^2\mathrm{d}t\,, \tag{2}$$

$$,\sigma_\omega^2 = \frac{1}{2\pi\left\|g^2\right\|}\int_{-\infty}^{\infty}\left(\omega-\omega_0\right)^2\left|\tilde{g}(\omega)\right|^2\mathrm{d}\omega \tag{3}$$

$$\sigma_t^2 = \frac{1}{\left\|g\right\|^2}\int_{-\infty}^{\infty}\left(t-t_0\right)^2\left|g(t)\right|^2\mathrm{d}t\,, \tag{4}$$

in which $\tilde{g}(\omega)$ is the Fourier transform of $g(t)$ and $\left\|g\right\|$ is the norm of $g(t)$, given by

$$\left\|g\right\|^2 = \int g(t)g^*(t)\mathrm{d}t \tag{5}$$

or, using Parsevals theorem:

$$\|g\|^2 = \frac{1}{2\pi}\int g^*(\omega)g(\omega)\mathrm{d}\omega \,, \tag{6}$$

where it is noticed that $g(t)$ can be a complex function.

It was derived by Gabor [11] that the following unequality holds:

$$\sigma_t\sigma_\omega \geq \tfrac{1}{2} \,. \tag{7}$$

For a good time-frequency resolution this product should be as small as possible.

The so-called Gabor filters fulfill the property $\sigma_t\sigma_\omega = \frac{1}{2}$ and have the form:

$$g_G(t) = Ae^{-b(t-t_0)^2}e^{i\omega_0(t-t_0)},\ A,b\in\mathbb{C} \,. \tag{8}$$

It was found by Hut et al. [8] that the cochlear response has less resolution than the Gabor filters and can therefore not replace the Gabor filters for general purpose high resolution time-frequency filtering.

There are many ways of obtaining time-frequency response functions of auditory signals. A well known method is the short term Fourier method where the signal is divided into small overlapping time slices that are Fourier transformed. Another often used method is by using so-called gammatone filters, as introduced by De Boer [12]. The mathematical formulation is given by

$$h_n(t,f_c) = at^{n-1}e^{-2\pi bt}\cos\left(2\pi f_c t+\phi\right), \tag{9}$$

$(t \geq 0,\ n \geq 1)$, where f_c is the center frequency of the filter and n the order of the filter. *a, b* ϕ are parameters that determine the amplitude, the duration and the phase of the filter. Gammatone filters can be implemented in a very efficient way.

To let the time-frequency response of the applied filters correspond as good as possible with human perception it is preferred to mimic the time-frequency response of the basilar membrane. Although the basilar membrane response includes a non-linear behaviour, a good approximation is obtained by linear modeling. An analytic model can be obtained with the following differential equations:

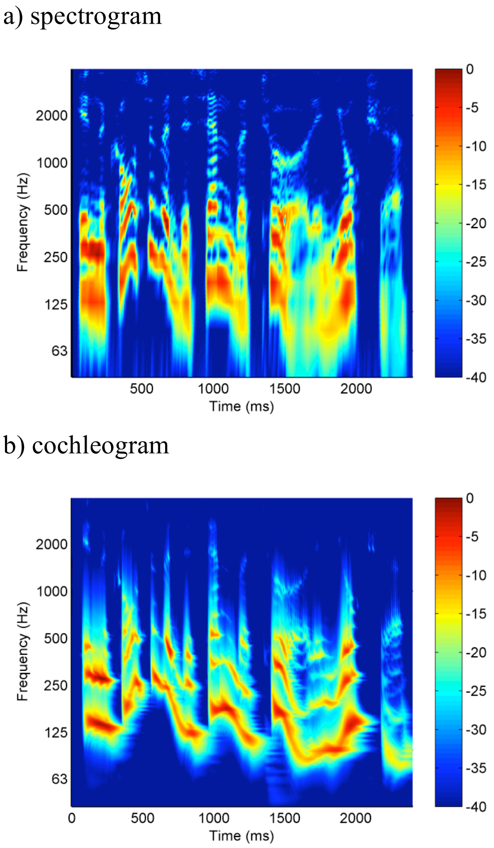
$$m\frac{\partial^2 y(x,t)}{\partial t^2} + d(x,y)\frac{\partial y(x,t)}{\partial t} + s(x)y(x,t) = p(x,t), \tag{10}$$

$$\frac{\partial^2 p(x,t)}{\partial x^2} - \Gamma\frac{\partial^2 y(x,t)}{\partial t^2} = 0 \,. \tag{11}$$

In equations (10) and (11) *m, d(x,y)* and *s(x)* are mass surface density, damping, and stiffness, respectively, while *x* represents distance along the membrane, and *y(x,t)* is the displacement of the membrane. *p(x,t)* is the pressure in the cochlear fluid and *Γ* is a constant determined by the density of the cochlear fluid, the cross-sectional area of the cochlear channels and the width of the partition, which are all assumed to be constant and are hence not shown explicitly in the model. More details of this model are given in Netten en Duifhuis [6] and Diependaal et al. [13].

The numerical values of the different parameters in these equations have been well established by extensive research and the filter set has been implemented in a very efficient way by Van Hengel [9] and Andringa [10]. They also included so-called ridge detection as an analysis tool for continuity preserved signal processing (CPSP).

Figure 1 shows a comparison between time-frequency analysis based on the short term Fourier transform (spectrogram) and a numerical simulation of the cochlear response (cochleogram) for a small speech segment of sound. Notice the difference in spectral resolution and continuity of the harmonic components.


a) spectrogram

b) cochleogram

Figure 1: Comparison of the spectrogram and the cochleogram of a short example of male speech.

Depending on the application, a time-frequency analysis method can be chosen at will. For some applications there is also a synthesis step to return from a time-frequency respresentation of the signal back to the time domain. More than one audio channel may be involved, for instance separate channels respresenting the audio signal at the left and right ear, or even a larger set of audio signals from an array of microphones. A general processing scheme including analysis and synthesis of one audio channel is shown in figure 2.

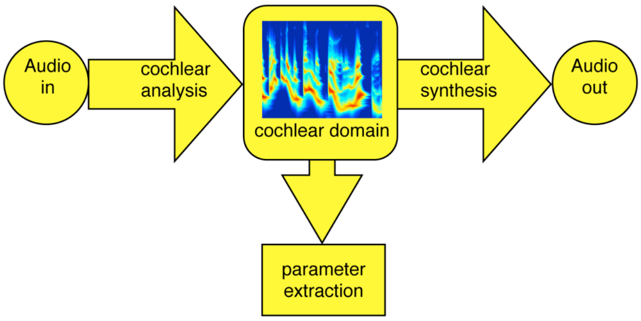


Figure 2: Flow chart of cochlear analysis and cochlear synthesis of an audio channel. Also indicated is the extraction of audio parameters from the cochlear domain respresentation of the signal.

3 Overview of the subprojects

In this chapter we present a short overview of the different subprojects that are involved. A common nomer in all subprojects is the time-frequency representation and processing of audio. Some subprojects have a separate paper at the Acoustics-08 conference, where more detailed information is presented.

3.1 Automatic Keyword Spotting (AKS)

In the AKS project we concentrate on the recognition of keywords in a wide variety of (degraded) acoustical conditions which is an unresolved problem in modern automatic speech recognition. We will illustrate our approach with a description of recognizing vowels.

Vowels are a subset of the voiced parts of speech which is very robust to noise. The informative parts of voiced speech are the harmonic complexes and these can be extracted relatively easy from a distorted signal. In order to extract the harmonic complexes, the signal components that exhibit relatively high energy levels are selected first. And second, the selection is reduced to the signal components that can be related to harmonic complexes. The harmonic complexes are then resolved by adding less reliable parts of the signal.

From ten different males, five vowels were selected from a spoken sentence. These vowels were used as an input to select the harmonic complexes as described above. Because harmonics do not cover all frequencies of the spectrum we focus on estimating the information underlying the surface structure in order to categorize the vowels. In order to approximate the real frequencies of the possible formants spectral peaks of the harmonic complex were extrapolated. For 12 vowels minimal distances were calculated of the estimated formant frequencies to the formant frequencies for Dutch vowels as described in the literature [18]. Based on these minimal distances the vowels were ordered on probability. The vowels that are uttered turn out to be categorized in the three-best group of hypothesized vowels (figure 3).

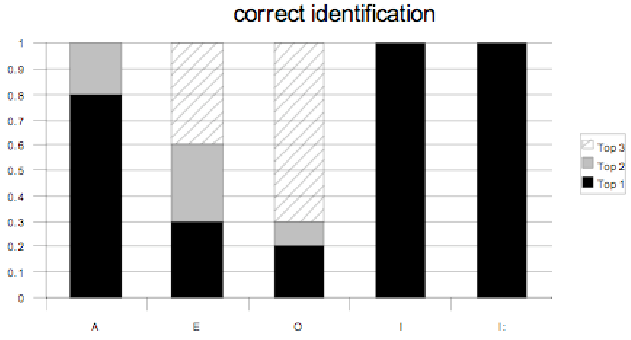


Figure 3: Proportion of correctly recognized vowels in the one-best, two-best and three-best group.

Two aspects suppress these results. Firstly all local maxima, without making a selection, are used as possible formants. An even better formant estimation might be obtained if time integration of the formant hypotheses limit or expand the formant hypotheses space. A second aspect that might suppress the results is the fact that only the formant frequencies and no other vowel characteristics are used. In the future these two aspects will be added to serve the task of vowel recognition in arbitrary acoustical environments.

3.2 Machine Analysis and Diagnostics (MAD)

Acoustical signals are a useful source of information about the functioning of machines; human operators can often detect and diagnose machine failure by simply listening to the sound they produce. The goal of the MAD project is to investigate the possibilities to approach or even surpass this remarkable ability of humans by an automatic system. In general, the performances of sound recognition systems are limited, amongst others due to unexpected input. However, since machines often contain rotating parts producing sounds with a highly tonal and relatively simple structure (see for instance Figure 4), we expect that the sound of specific machines can be modelled accurately.

We model the sound in terms of signal components (SCs). SCs are physically coherent trajectories in the time-frequency plane with a positive local signal-to-noise ratio. The (relative) positions of SCs together with their energetic development are supposed to contain all information necessary to obtain the machine's status. However, since extracted SCs usually stem from different sound processes, this higher-level information can only be revealed after grouping.

In Figure 4 we show such a mixture of sounds, produced by the compressor of a gas turbine (*Solar Centaur*) during startup. After 2.6 seconds, a new sound source enters the scene (harmonic sweeps). In order to classify this sound, its SCs has to be extracted (bottom panel) and grouped together to segregate it from its pulse-like background. After grouping (result not shown here), the new sound source can be analyzed separately to determine whether or not it is related to machine failure.

At this moment we have an effective SC-extraction algorithm and a grouping algorithm based on harmonicity. Future work should lead to the modeling of (other) machine behavior and expected defects and the application of other grouping principles.

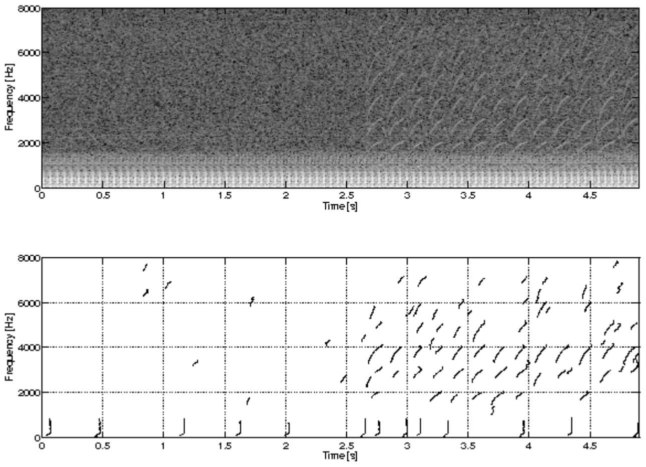


Figure 4: Spectrogram (top panel) shows the relatively simple structure of a machine sound. It can be modelled effectively by extracting the signal components (bottom panel).

3.3 Speech Intelligibility Enhancement of Hearing Aids (SEHA)

For the design of a successful hearing aid, the preservation of continuity on a phoneme-scale is of great importance. Considering the grade of rehabilitation by a hearing aid, the algorithmic reduction of interference has to be performed alongside with sustained speech quality and listening comfort to enhance speech intelligibility [14].

The challenges that hearing aids have to face are real-world situations, such as situations in which many talkers compete and in which reverberation as well as additive noise complicate the intelligibility of a target-speaker. A second issue for the design of a hearing-aid algorithm is its computational complexity and its real-time performance. Although automatic speech recognition systems perform more and more successfully, their application is generally dependent on a training phase, previous knowledge as well as a high degree of computational effort. These systems are generally referred to as top-down approaches. In the SEHA-project we focus on bottom-up approaches, which enable an instantaneous speech processing, whereas the computational complexity and load are comparatively low.

CPSP gives a means to adhere to the trade-off between noise-reduction and signal-distortion. With respect to the preservation of the continuity of speech, SEHA is geared to biologically motivated models of speech-processing.

Research in that field (prominently in the field of research named CASA, - computational auditory scene analysis) revealed that the gap between the computational applications of biologically inspired models and the processing of a healthy human hearing is still far from bridged. A full understanding of the human auditory processing is supposed to be required for an equivalent computational operation. However, also combinations of different processing schemes of noise reduction appear promising to achieve a successful enhancement of speech intelligibility [14], [15].

In the SEHA-project we concentrate on combinations of prosperous approaches that enhance speech intelligibility. In a first design, we combined optimized spatial beam-forming [16] with a biological model of modulation perception and binaural interaction [14]. The results were assessed with the speech intelligibility index. In a variety of adverse acoustical situations, the combined processing scheme shows an improvement comparable to either of the underlying processing schemes alone [17].

Future work of SEHA is devoted to an optimization of complementary processing schemes as well as a further incorporation of other successful models of speech perception.

3.4 Quality Assessment of Room Acoustics (QARA)

Acoustical parameters, which describe the acoustical qualities of a room, are generally determined from impulse responses. These responses can be determined from measurements on single positions in the room or from measurements along an array with multiple closely spaced microphone positions. However, it turns out that these parameters as determined from measurements suffer from large spatial fluctuations, something which does not match with human perception.

In this project it is tried to reach more reliable results by applying mechanisms of the human auditory processing. These include the absolute threshold of hearing, overlap of the auditory filters and forward and simultaneous masking.

The resulting signal processing algorithm was tested on an impulse response set, measured along a line array in the "Concertgebouw" in Amsterdam, the Netherlands. Various acoustical parameters were determined from the impulse responses before and after applying the algorithm. Results for reverberation time, clarity index and early decay time are shown in figures 3, 4 and 5.

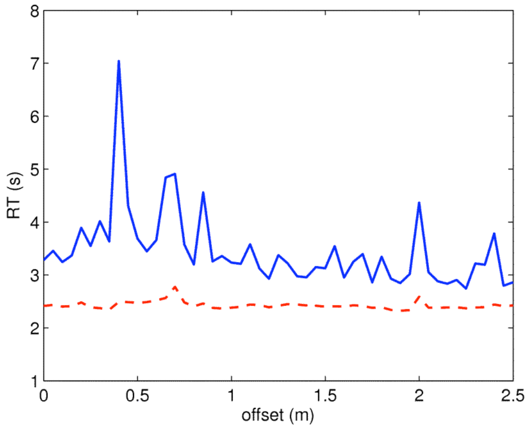


Figure 3: Reverberation time RT in the 125 – 2000 Hz frequency band as a function of offset. The parameter was determined both with (red dashed line) and without (solid blue line) simulation of auditory processing.

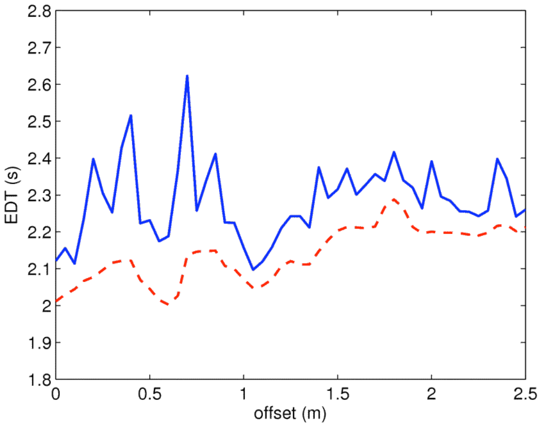


Figure 4: Early decay time EDT in the 125 – 2000 Hz frequency band as a function of offset. The parameter was determined both with (red dashed line) and without (solid blue line) simulation of auditory processing.

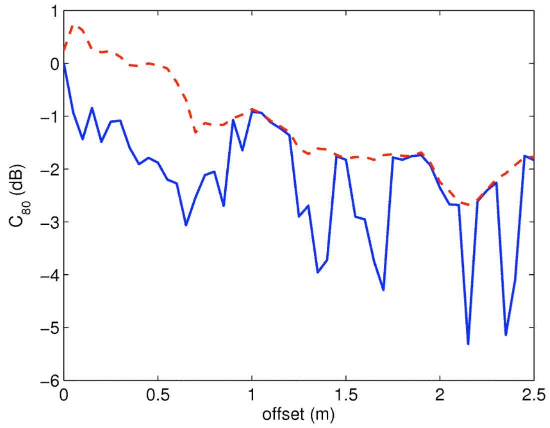


Figure 5: Clarity index *C80* in the 125 – 2000 Hz frequency band as a function of offset. The parameter was determined both with (red dashed line) and without (solid blue line) simulation of auditory processing.

It can be seen from the figures that the spatial fluctuations reduce considerably when auditory processing is simulated. The graphs become much smoother, which is in agreement with human perception.

In the near future the algorithms will be tuned further, to make them more robust. Also, the current algorithm does not yield much improvement for spatial parameters like lateral fraction *LF* and inter-aural cross-correlation *IACC*. It will be investigated if unwanted spatial fluctuations in these parameters can also be reduced.

5 Conclusions and Outlook

In this paper an overview is presented of work that is currently in progress on a project based on time-frequency analysis and synthesis of audio signals for different applications. The principle of time-frequency analysis is used here in relation to the working principle of the human ear which is essentially a time-frequency acoustic to neural signal transducer. By following the same principles with signal processing tools, results can be obtained that are close to human perception, which is important for our different subprojects.

Acknowledgments

This project is co-financed by the Dutch Technology Foundation STW, including contributions from Gasunie Research, Sound Intelligence, Varibel, Philips Research, Televic and B&K.

References

[1] Georg von Békésy, “The Variation of Phase Along the Basilar Membrane with Sinusoidal Vibrations”, J. Acoust. Soc. Am., **19** (1947) 452 – 460.

[2] Georg von Békésy, “The Vibration of the Cochlear Partition in Anatomical Preparations and in Models of the Inner Ear”, J. Acoust. Soc. Am., **21** (1949) 233 – 245.

[3] Georg von Békésy, “On the Resonance Curve and the Decay Period at Various Points om the Cochlear Partition”, J. Acoust. Soc. Am., **21** (1949) 245 – 254.

[4] J. Zwislocki, “Theory of the acoustical action of the cochlea”, J. Acoust. Soc. Am. **22** (1950) 778 – 784.

[5] M. A. Viergever, “Mechanics of the inner ear”, PhD-thesis, Delft University of Technology, 1980.

[6] S.M. Netten and H. Duifhuis, “Modelling an active, nonlinear cochlea – In: Mechanics of Hearing, E. de Boer and M. A. viergever (Eds), Delft University Press, Delft, 1983, 143 – 151.

[7] E. Zwicker and E. Terhardt, “Analytical expression for critical band rate and critical bandwidth as a function of frequency”, J. Acoust. Soc. Am. **68** (1987) 1523 – 1525.

[8] R. Hut, M.M. Boone and A. Gisolf, “Cochlear Modeling as Time-Frequency Analysis Tool”, Acta Acustica united with Acustica **92** (2006) 629 – 636.

[9] P.W.J. van Hengel, “Emissions from cochlear modelling”, PhD-thesis, Rijksuniversiteit Groningen (1996).

[10] T. Andringa, “Continuity Preserving Signal Processing”, PhD-thesis, Rijksuniversiteit Groningen (2002).

[11] D. Gabor, “Theory of communication”, J. IEE **93** (1946) 429 – 457.

[12] E. de Boer, “Synthetic whole-nerve action potentials for the cat”, J. Acoust. Soc. Am. **58** (1975) 1030 – 1045.

[13] R.J. Diependaal, H. Duifhuis, H.W. Hoogstraaten and M.A. Viergever, “Numerical methods for solving one-dimensional cochlear models in the time domain”, J. Acoust. Soc. Am. **82** (1987) 1655 – 1666.

[14] B. Kollmeier and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction”, J. Acoust. Soc. Am., **95** (1994) 1593 – 1602.

[15] T. Wittkop et al. “Speech processing for hearing aids: noise reduction motivated by models of binaural interaction”, Acta Acustica, **83**1(997) 684 – 699.

[16] M.M. Boone, “Directivity measurement of a highly directive hearing aid: the hearing glasses”, AES 120th Convention Paper 6829, 2006.

[17] A. Schlesinger and M. M Boone, “Dual Noise Suppression in Hearing Aids”, AES 124th Convention Paper 7386, 2008.

[18] P. Adank, P., R. van Hout, and R. Smits), “An acoustic description of the vowels of northern and southern standard Dutch”, .J. Acoust. Soc. Am., **116** (2004) 1729-1738.