EURASIP Journal on Applied Signal Processing 2005:9, 1323–1333

Analysis of the IHC Adaptation for the Anthropomorphic Speech Processing Systems

Alexei V. Ivanov

Computer Engineering Department, the Belarusian State University of Informatics and Radioelectronics, 220013 Minsk, Belarus
Email: alexei_v_ivanov@ieee.org

Alexander A. Petrovsky

Real-Time Systems Department, the Bialystok Technical University, 15351 Bialystok, Poland
Email: palex@it.org.by

Received 1 November 2003; Revised 5 September 2004

We analyse the properties of the physiological model of the adaptive behaviour of the chemical synapse between inner hair cells (IHC) and auditory neurons. On the basis of the performed analysis, we propose equivalent structures of the model for implementation in the digital domain. The main conclusion of the analysis is that the synapse reservoir model is equivalent in its properties to the signal-dependent automatic gain-control mechanism. We plot guidelines for creation of artificial anthropomorphic algorithms, which exploit properties of the original synapse model. This paper also presents a concise description of the experiments, which prove the presence of the positive effect from the introduction of the depicted anthropomorphic algorithm into feature extraction of the automated speech recognition engine.

Keywords and phrases: inner hair cell (IHC), Meddis IHC model, IHC adaptation, auditory models, modulation spectrum filtering.

1. INTRODUCTION

1.1. Anthropomorphism, psychoacoustics, and auditory physiology

Many contemporary speech processing techniques tend to reflect properties of the human auditory apparatus. As a rule, most of the information about the way human beings process acoustic data comes into artificial applications from the field of psychoacoustics (for classical psychoacoustics work, refer to [1]).

Apart from the experiments with subjects that have reliably diagnosed and anatomically localised auditory pathology, psychoacoustics treats the whole human auditory system as a "black box" and tries to infer its properties without particular interest to its internal structure. Most of the psychoacoustical experiments include analysis of the responses to "simple" sounds, like pure tones, wideband noise, coloured noises, clicks, and so forth. But a lot of evidence (simultaneous and nonsimultaneous masking, pitch perception, etc.) points to the fact that the auditory system is essentially a nonlinear system.

From the system identification theory, it is known that the response of the linear system to an arbitrary excitation can be derived from the study of responses of such system to simple sounds, for example, tones, noises, and clicks. There is no need to study the internal structure of the linear black box as far as responses to the simple input signals are known. Strictly speaking, for the case of nonlinear systems, this black box approach is not applicable. There are mainly two possibilities to model a nonlinear system: either to construct a semiparametric statistical learning machine, a "neural-network-like" structure, and let it adapt through a kind of learning algorithm, or follow the parametric approach and somehow infer the internal structure of the nonlinear system to be modelled, parse it into smaller and, hopefully, simpler building blocks, then tune parameters of those blocks, so that model response matches that of the original system.

The first alternative suffers from the problems in creating the representative training set, as well as from the absence of a priori information regarding the required model complexity. The mentioned difficulties virtually prohibit application of this approach to the auditory modelling. The second of the mentioned approaches corresponds to the physiologically grounded studies of the auditory apparatus.

Among the solutions, which could benefit most from the employment of the physiological models, one can name the development of cochlear implants, the objective and quantitative quality assessment of the coded audio reconstruction, anthropomorphic audio coding, and automated

speech recognition applications. While the first two mentioned branches are concentrated on the closest possible literal reproduction of the auditory apparatus properties in the artificial device, the latter imply a computationally efficient way to implement the "biological" audio processing algorithm with a certain predefined precision.

In spite of being precise and objective, the physiological hearing models neither provide a clear signal processing interpretation of those phenomena, nor give a ready answer regarding the relevance of the modelled phenomena to the hearing process in general. Thus, straightforward application of the physiological models to the fields of audio coding and speech recognition may not easily gain advantage over the conventional algorithms [2]. Before the employment of a certain physiological model into the mentioned applications, one should answer the questions of why it is important (i.e., what result is expected from it) and what is the most efficient way of its implementation. This reasoning leads to a conclusion that the further analysis of the available physiological models with the aim of finding their algorithmical interpretation is needed. This paper is further devoted to such kind of analysis.

Particularly, we are aiming at analysing the adaptation of the chemical "inner-hair-cell auditory nerve" (IHC-AN) synapse, and trying to infer its importance to the artificial anthropomorphic audio signal (and particularly speech) processing systems in adverse environments. Indeed, strong onset responses of the auditory nerve (AN) fibers to the presented stimulus are followed by the "adaptation", that is, gradual decrease of the response amplitude over time while the stimulus amplitude remains constant. This "adaptive strategy" at first glance seems to be advantageous since it allows an emphasis of nonstationarities within the incoming signal.

2. RESERVOIR MODEL OF IHC-AN CHEMICAL SYNAPSE

Physiological research into the way the inner ear converts an acoustical stimulation into a response of the auditory nerve fibers (for a brief summary and review, refer to [3]) among many other findings led to the conclusions that

(i) inner hair cells are mechanical vibrations sensory cells;
(ii) each IHC makes chemical synapses with approximately 10–30 peripheral axons of primary bipolar neurons which cell bodies contained in the spiral ganglion and modiolar axons forming the auditory (VIIIth) nerve;
(iii) one can distinguish three groups of afferent neurons based on the level of their spontaneous activity: low-spontaneous rate, medium-spontaneous rate, and high-spontaneous rate fibers. The level of spontaneous activity of the fiber is closely related to the form and the size of the synapse it formed with IHC;
(iv) chemical nature determines the following properties of IHC-AN synapses: adaptive responses, synaptic delays, quantised response amplitudes.

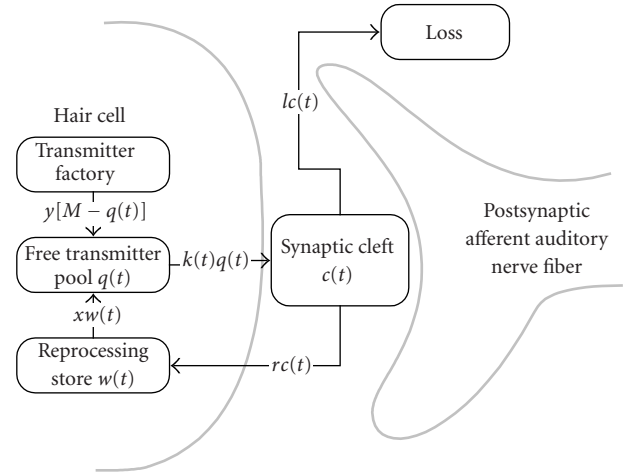

Figure 1: Schematic representation of the Meddis reservoir model.

Properties of the chemical IHC-AN synapse are successfully captured by the so-called "reservoir models," in which neurotransmitter is produced and stored in the IHC to be released in accordance with IHC transmitter release probability that changes with mechanical vibrations in the inner ear. First reservoir models for IHC-AN synapses were proposed as early as [4, 5].

Meddis has put forward [6] and further developed [7, 8, 9, 10, 11] a model of IHC, which includes a version of reservoir model of chemical synapse. The latest model [10, 11] allows for a nice fit between experimental and model data for all thee groups of IHC-AN synapses (low-, medium-, and high-spontaneous rate fibers) with only calcium conductance parameters being changed.

It must be noted here that in reality neurotransmitter release into synaptic cleft is a probabilistic and quantal process. However, to a certain degree, the dynamical properties of the synapse may be reflected by the model that assumes that neurotransmitter flow is deterministic and continuous. From the practical point of view, this assumption corresponds to the averaging of the synapse response over many identical stimulations. Latest Meddis models [9, 10, 11] depart from this assumption offering better correspondence to the data recordings of individual experiments. For the purpose of the analysis of the core properties of IHC-AN synapse and construction of the anthropomorphic artificial algorithms, we further narrow our consideration to the deterministic and continuous case.

Meddis version of the reservoir model is represented by schematic drawing in Figure 1, and is described by the set of (1). "Free transmitter pool" is the main storage facility for the transmitter that is immediately ready to be released from the cell to the "synaptic cleft." It is filled with neurotransmitter coming from the "transmitter factory" as well as that recycled at the "reprocessing store." Neurotransmitter is being released into "synaptic cleft" with a certain rate, dependent upon IHC stimulation, as well as instantaneous quantity of the stored transmitter. From the "synaptic cleft," transmitter is either being returned to the cell for reprocessing or lost by diffusion.

We assume that the pool capacity equals M. The quantity of the transmitter stored in the pool at a certain time instant will be denoted by $q(t)$. The rate, at which the factory produces new transmitter, is proportional to the free volume of the pool $y[M - q(t)]$, here operation $[\cdots]$ constitutes the choice of the biggest value between zero and the value inside square brackets. Alternatively we may put that coefficient y becomes zero at the moment the pool is filled to the limit. We denote the instantaneous amount of the transmitter in the reprocessing by $w(t)$. The recirculation rate is proportional to the amount of the transmitter in the reprocessing $xw(t)$. The rate, at which transmitter is sent to the cleft, is equal to the product of membrane permeability $k(t)$ and the quantity of the transmitter in the pool $q(t)$. The quantity of the neurotransmitter in the cleft at certain instant will be denoted by $c(t)$. Rates of neurotransmitter loss and return for reprocessing are proportional to the amount of the transmitter in the clefts $lc(t)$ and $rc(t)$, respectively.

As it follows from the above-presented description, Meddis version of the reservoir model is described by the following set of differential equations:

$$\begin{aligned} \frac{dq(t)}{dt} &= xw(t) + y[M - q(t)] - k(t)q(t), \\ \frac{dc}{dt} &= k(t)q(t) - (l + r)c(t), \\ \frac{dw}{dt} &= rc(t) - xw(t). \end{aligned} \tag{1}$$

Initial conditions of the model are taken in accordance with the assumption that at a certain instant t_0 the system is in an equilibrium state:

$$\begin{aligned} xw(t_0) + y[M - q(t_0)] &= k(t_0)q(t_0), \\ k(t_0)q(t_0) &= (l + r)c(t_0), \\ rc(t_0) &= xw(t_0). \end{aligned} \tag{2}$$

3. ADAPTATION PROPERTY OF THE RESERVOIR MODEL OF IHC-AN CHEMICAL SYNAPSE

Figure 2 presents a typical response of the Meddis model to the excitation. Signal $k(t)$ is an input to the reservoir model and is computed by earlier stages of cochlear model (the cochlear filter bank [12] in combination with the first part of IHC model [10]) when the test tone of 6 kHz is presented. IHC medium-spontaneous rate fiber model gets its input from the cochlear filter bank section with the closest to 6 kHz centre frequency. It is running at the sampling frequency of 16 kHz.

Typical values of the model coefficients were taken from the works of Meddis [6, 7, 8, 9] and are as follows:

$$\begin{aligned} M = 10, \quad x = 66.3, \quad y = 10, \\ l = 2580, \quad r = 6580. \end{aligned} \tag{3}$$

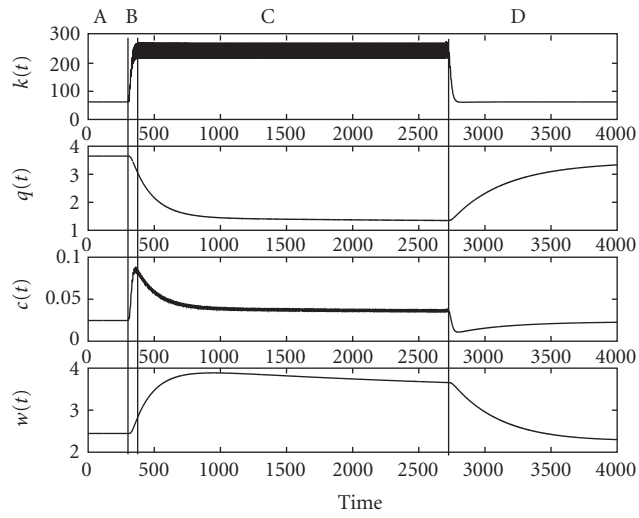

Figure 2: Reservoir model response to the excitation with the 6 kHz tone, CF ~ 6 kHz, F_s = 48 kHz, medium-spontaneous rate fiber. A: steady state, B: onset, C: adaptation, D: offset.

In order to perform this digital simulation (depicted in Figure 2) of the synapse model, the forward difference approximation of the set of differential equations (1) was used, as it is advised in [8].

As it can be seen from the above figure, there are four distinct regions in the model response signal $c(t)$: steady-state response to a long-term absence of stimulation (denoted as region A); onset response (region B)—brief rise of the response level to higher values; subsequent adaptation of the response level to a much lower activity (region C); offset region (region D), when synapse recovers from the stimulation and response level slowly converges to a steady-state level.

For a detailed review of adaptation properties of IHC, please refer to [11].

4. ANALYSIS OF THE RESERVOIR MODEL OF IHC-AN CHEMICAL SYNAPSE

Looking at the equation set (1), one can easily notice that functions $c(t)$ and $f(t) = k(t)q(t)$ are linked with the linear constant-coefficient differential equation of the first order with zero-free member:

$$\frac{dc(t)}{dt} + (l + r)c(t) = f(t). \tag{4}$$

Thus, (4) describes a linear time invariant system, which performs transformation of $f(t)$ into $c(t)$. Taking forward difference approximation of the differential problem and assuming that both functions take discrete values at discrete-time instances, it is possible to approximate this system with

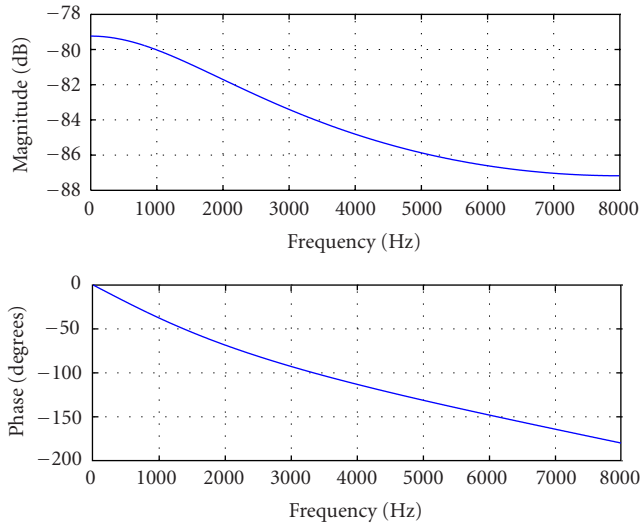

Figure 3: Frequency characteristic of filter A ($F_s = 16\,\text{kHz}$).

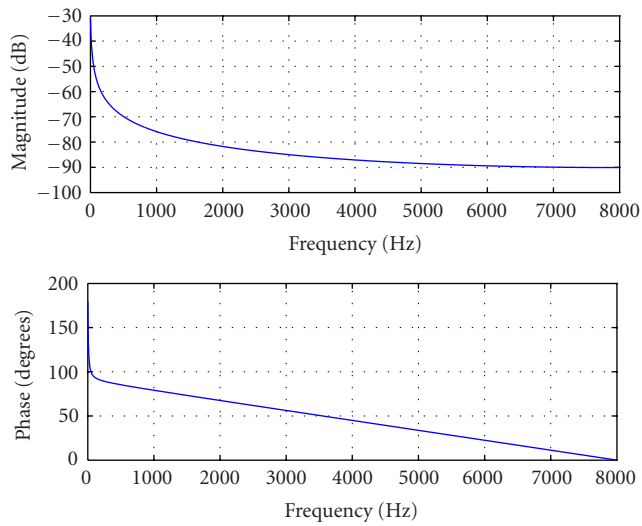

Figure 4: Frequency characteristic of filter B ($F_s = 16\,\text{kHz}$).

a digital filter:

$$c(n) = \frac{1}{F_s} f(n-1) - \frac{(l+r-F_s)}{F_s} c(n-1), \tag{5}$$

$$H_A = \frac{(1/F_s)z^{-1}}{1-\left((1-(l+r)/F_s)\right)z^{-1}}. \tag{6}$$

Here F_s denotes the sampling frequency. We will further refer to this filter as "filter A." With the typical values of parameters l and r, this filter is a lowpass filter, which has rather smooth slope response characteristic that is presented in Figure 3.

Further analysis of the equation set (1) leads to a conclusion that functions $s(t) = M - q(t)$ and $f(t) = k(t)q(t)$ are also linked with the linear constant-coefficient differential equation of the first order with zero-free member:

$$\frac{d^3s(t)}{dt^3} + (x+y+l+r)\frac{d^2s(t)}{dt^2} + \left((x+y)(l+r)+xy\right)\frac{ds(t)}{dt} + xy(l+r)s(t) = \frac{d^2f(t)}{dt^2} + (x+l+r)\frac{df(t)}{dt} + xlf(t). \tag{7}$$

We note that this equation is valid for all such values $s(t) = M - q(t) \geq 0$. If $s(t) = M - q(t) \leq 0$, then it must be substituted with the following equation, which is obtained from (7) by letting $y = 0$:

$$\frac{d^3s(t)}{dt^3} + (x+l+r)\frac{d^2s(t)}{dt^2} + x(l+r)\frac{ds(t)}{dt} = \frac{d^2f(t)}{dt^2} + (x+l+r)\frac{df(t)}{dt} + xlf(t). \tag{8}$$

The performed digital simulations show that for realistic input signals and reasonably high sampling frequency, it is enough to use (7) only.

Again it is possible to approximate the system described by (7) with a digital filter:

$$\begin{aligned} s(n) &= \frac{b_1}{a_0} f(n-1) + \frac{b_2}{a_0} f(n-2) + \frac{b_3}{a_0} f(n-3) \\ &\quad - \frac{a_1}{a_0} s(n-1) - \frac{a_2}{a_0} s(n-2) - \frac{a_3}{a_0} s(n-3), \\ a_0 &= F_s^3, \\ a_1 &= -3F_s^3 + F_s^2(x+y+l+r), \\ a_2 &= 3F_s^3 - 2F_s^2(x+y+l+r) + F_s\left((x+y)(l+r)+xy\right), \\ a_3 &= -F_s^3 + F_s^2(x+y+l+r) \\ &\quad - F_s\left((x+y)(l+r)+xy\right) + xy(l+r), \\ b_1 &= F_s^2, \\ b_2 &= -2F_s^2 + F_s(x+l+r), \\ b_3 &= F_s^2 - F_s(x+l+r) + xl. \end{aligned} \tag{9}$$

We denote this filter as "filter B." This is a lowpass filter with rather sharp frequency response characteristic (see Figure 4) for typical values of parameters x, y, l, and r.

Filter B has two real zeros and three real poles:

$$n_{B1,2} = 1 - \frac{1}{2F_s}\left((x+l+r) \pm \sqrt{(x+l+r)^2 - 4xl}\right), \tag{10}$$

$$p_{B1} = 1 - \frac{l+r}{F_s}, \qquad p_{B2} = 1 - \frac{x}{F_s}, \qquad p_{B3} = 1 - \frac{y}{F_s}. \tag{11}$$

The above conclusions imply that there must be a link between functions $c(t)$ and $s(t) = M - q(t)$ in the form

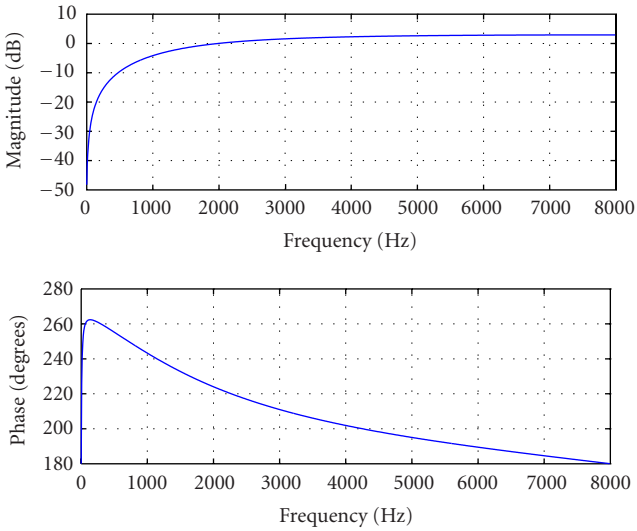

Figure 5: Frequency characteristic of filter C ($F_s = 16$ kHz).

of the linear constant-coefficient differential equation of the first order with zero-free member. Indeed, it is the case

$$\frac{d^2c(t)}{dt^2} + (x+l+r)\frac{ds(t)}{dt} + xlc(t) = -\frac{d^2s(t)}{dt^2} - (x+y)\frac{ds(t)}{dt} + xys(t). \tag{12}$$

As in the case of (7), this equation is valid for $s(t) = M - q(t) \geq 0$, if it is less than zero, then y in (12) should be put to zero.

The digital filter, which is equivalent to the system (12), is defined as follows:

$$\begin{aligned} s(n) &= \frac{d_0}{c_0}f(n) + \frac{d_1}{c_0}f(n-1) + \frac{d_2}{c_0}f(n-2) \\ &\quad - \frac{c_1}{c_0}s(n-1) - \frac{c_2}{c_0}s(n-2), \\ c_0 &= F_s^2, \\ c_1 &= -2F_s^2 + F_s(x+l+r), \\ c_2 &= F_s^2 - F_s(x+l+r) + xl, \\ d_0 &= -F_s^2, \\ d_1 &= 2F_s^2 - F_s(x+y), \\ d_2 &= -F_s^2 + F_s(x+y) - xy. \end{aligned} \tag{13}$$

We will further denote this filter as "filter C." It is a high-pass filter with rather sharp frequency response characteristic (see Figure 5) for typical values of its parameters.

We also note that a cascade connection of filters B and C should be equivalent to filter A. This is true and can be immediately proved by looking at (9), (13), and (5).

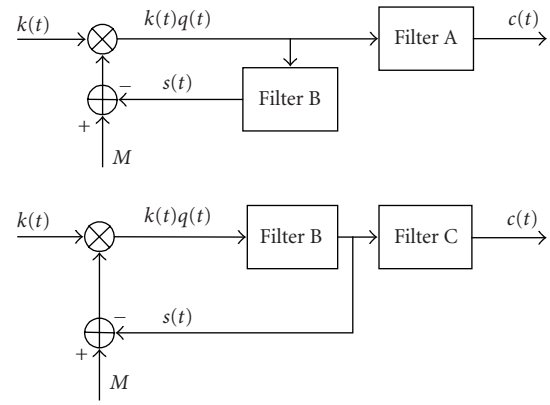

Figure 6: Reservoir model equivalent structures.

5. EQUIVALENT DIGITAL STRUCTURES FOR THE RESERVOIR MODEL

Analysis of the Meddis reservoir model allows us to plot its equivalent structures for realisation in the digital form (see Figure 6). The realisation with the help of filter A is more preferable since it is more computationally efficient.

Apart from the linear digital filters, the developed equivalent representations include an operation of multiplication of the signals in the time domain. It should be noted that, in general, multiplication of time-varying signals does not comply with the superposition principle, thus the reservoir model equivalent structure performs a nonlinear signal transformation. The signal $q(t) = M - s(t)$, which is multiplied by $k(t)$ is confined in the interval $[0, M]$ in accordance with the reservoir model definition. It consists mainly of low-frequency components of signal $k(t)q(t)$ in accordance with the properties of filter B.

Operation of the multiplication in the equivalent structure may be viewed as an automatic gain-control (AGC) operation. The gain $q(t)$ is a parameter, which slowly varies through time between M in the case of weak input signal and zero in the case of strong one.

Our equivalent structure of the Meddis reservoir model has similarities with that plotted in the works of Perdigao [13, 14].

6. LINEAR APPROXIMATION OF THE SIGNAL MULTIPLICATION OPERATION IN THE EQUIVALENT STRUCTURE OF THE RESERVOIR MODEL

It is possible to build a linear digital filter, which approximates the effect of the AGC mechanism for the case of small deviations of the system from the equilibrium state. Particular form of such filter is dependent on initial conditions, namely, the steady-state input signal value k_0.

A method we are going to use is thoroughly investigated in [15]. Similar methods of differential equation linearisation (which lead to the identical results) are widely known and used in the classical literature on theoretical mechanics.

Indeed, we assume that the system depicted in Figure 6, at a certain time instant, resides in the equilibrium. For such case, we may write

$$\begin{aligned} f_0 &= k_0 q_0, \\ q_0 &= M - s_0, \\ y(l+r)s_0 &= l f_0. \end{aligned} \tag{14}$$

Any deviations from the steady state are sufficiently small:

$$\begin{aligned} k(n) &= k_0 + \delta k(n), \\ f(n) &= f_0 + \delta f(n), \\ q(n) &= q_0 + \delta q(n), \\ s(n) &= s_0 + \delta s(n). \end{aligned} \tag{15}$$

Thus, for such system at an arbitrary time instant, we may write the following set of equations (see Figure 6):

$$\begin{aligned} f_0 + \delta f(n) &= (k_0 + \delta k(n))(q_0 + \delta q(n)), \\ q_0 + \delta q(n) &= M - (s_0 + \delta s(n)), \\ (a_0 + a_1 + a_2 + a_3)s_0 + a_0\delta s(n) + a_1\delta s(n-1) + a_2\delta s(n-2) & \\ + a_3\delta s(n-3) &= (b_1 + b_2 + b_3)f_0 + b_1\delta f(n-1) \\ &\quad + b_2\delta f(n-2) + b_3\delta f(n-3). \end{aligned} \tag{16}$$

Coefficients in the third equation of the set are those of filter B. Comparing sets (15) and (16), we may conclude that the following set of equations holds for deviations:

$$\begin{aligned} \delta f(n) &= k_0\delta q(n) + q_0\delta k(n), \\ \delta q(n) &= -\delta s(n), \\ a_0\delta s(n) + a_1\delta s(n-1) + a_2\delta s(n-2) + a_3\delta s(n-3) & \\ = b_1\delta f(n-1) + b_2\delta f(n-2) + b_3\delta f(n-3). & \end{aligned} \tag{17}$$

A solution of the equation set (17) with respect to variables $\delta k(n)$ and $\delta f(n)$ is represented as

$$\begin{aligned} q_0 &= \frac{My(l+r)}{y(l+r) + lk_0}, \\ q_0(a_0\delta k(n) + a_1\delta k(n-1) + a_2\delta k(n-2) + a_3\delta k(n-3)) & \\ = a_0\delta f(n) + (b_1k_0 + a_1)\delta f(n-1) & \\ + (b_2k_0 + a_2)\delta f(n-2) + (b_3k_0 + a_3)\delta f(n-3). & \end{aligned} \tag{18}$$

This equation represents a desired linear digital filter, which linearly approximates AGC of the equivalent structure. This filter is capable of transforming the signal $\delta k(t) = k(t) - k_0$ into $\delta f(t) = \delta(k(t)q(t)) = f(t) - f_0$ under the condition that these deviations are sufficiently small. The transfer function of this filter is expressed as

$$\begin{aligned} H_D(z, k_0) & \\ = \frac{My(l+r)}{(y(l+r) + lk_0)} & \\ \cdot \frac{a_0 + a_1z^{-1} + a_2z^{-2} + a_3z^{-3}}{a_0 + (b_1k_0 + a_1)z^{-1} + (b_2k_0 + a_2)z^{-2} + (b_3k_0 + a_3)z^{-3}}. & \end{aligned} \tag{19}$$

Note the explicit dependency of the form of this transfer function on the value of k_0. We will further denote this filter as "filter D."

The steady-state output $f_0(k_0)$ of the system is derived from the equilibrium set of (14) and is expressed as

$$f_0(k_0) = q_0k_0 = \frac{My(l+r)}{y(l+r) + lk_0}k_0. \tag{20}$$

Filter D is a highpass filter with quite sharp frequency response characteristic (see Figure 7) for a typical value of $k_0 = 10$.

In order to illustrate the dependence of the properties of filter D upon the value of k_0, Figure 8 depicts frequency characteristic of that filter with $k_0 = 1000$. As it can be seen from the comparison of Figures 7 and 8, apart from the change of the gain, the cut-off frequency of the filter is getting bigger with the increase of k_0.

From the digital filter theory it is known that the linear digital filter is "bounded-input bounded-output" (BIBO) stable if all of its poles lay inside the unit circle in z-plane. Filter D has three real poles. Analytical derivation of their values is rather complex in general. To perform such derivation, one could take advantage of Cardano formula for the roots of cubic equation.

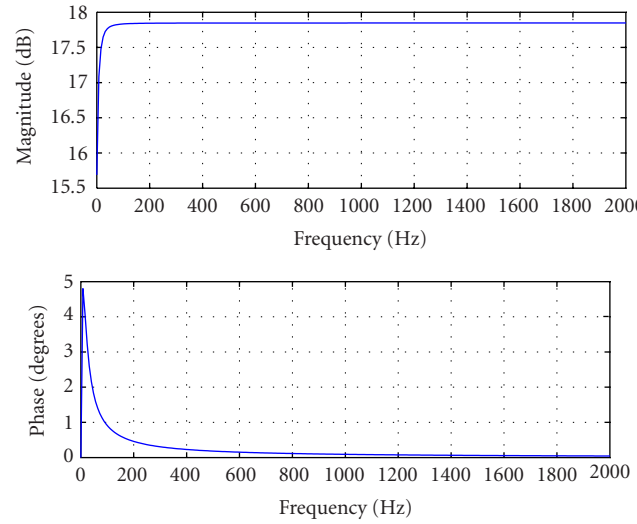

FIGURE 7: Frequency characteristic of filter D ($F_s = 16$ kHz, $k_0 = 10$).

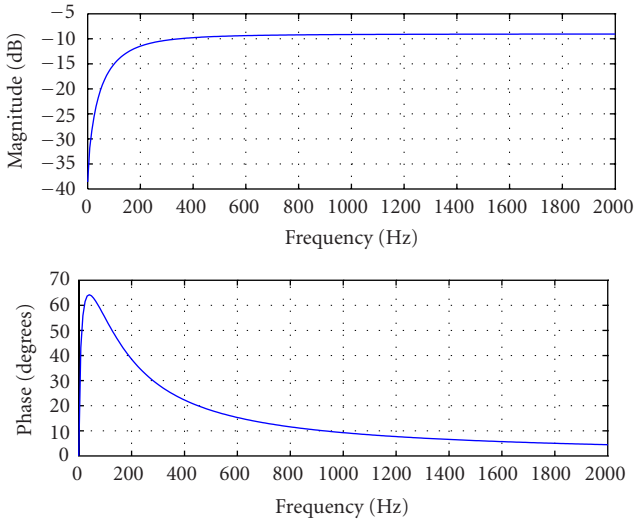

FIGURE 8: Frequency characteristic of filter D ($F_s = 16\,\text{kHz}$, $k_0 = 1000$).

An alternative way is to estimate positions of the filter poles. Indeed, for realistic values of $k_0 \sim 10^1$–10^2 with quite good precision, filter D poles lay in the vicinity of its zeros. Zeros of filter D coincide with poles of filter B (11), and approximately we may put

$$p_{D1} \approx n_{D1} = 1 - \frac{l+r}{F_s}, \qquad p_{D2} \approx n_{D2} = 1 - \frac{x}{F_s}, \qquad p_{D3} \approx n_{D3} = 1 - \frac{y}{F_s}. \tag{21}$$

It must be noted also that if $k_0 \to 0$, then $p_{DN} \to n_{DN}$.

Pole p_{D1} first leaves the unit circle while sampling frequency is being decreased, indeed the realistic values of $l + r$ are significantly larger than the values of x and y. Consequently, approximation of the position of the first pole gives us a condition of filter D stability, while $k_0 \to 0$:

$$F_s > \frac{l+r}{2}. \tag{22}$$

Pole p_{D1} moves to the right on the real axis if the value of k_0 is being increased. This allows for filter D to become stable with increased k_0 even if it was unstable with the smaller values of k_0. This leads us to a conclusion that (22) represents sufficient condition for filter D to be stable with arbitrary realistic values of k_0.

In the work [8], it is required that the sampling frequency must be sufficiently large for a successful digital implementation of the reservoir model. Our finding of stability condition (22) puts a quantitative restriction on the sampling frequency for the linearised approximation.

Under the same assumption of small deviations from the equilibrium state, it is possible to construct an equivalent linear filter, which would serve as linear approximation of relation of the signals $\delta k(t)$ and $\delta c(t) = c(t) - c_0$, that is, the input and the output signals of the reservoir model measured relatively to their corresponding equilibrium values.

Such filter (further denoted as filter E) corresponds to the cascade of the filters D and A. Its frequency response characteristic is presented in Figure 9. Filter E transfer function is defined as

$$H_E(z, k_0) = \frac{1}{F_S} \cdot \frac{My(l+r)}{(y(l+r) + lk_0)} \cdot \frac{(1 - n_{D2}z^{-1})(1 - n_{D3}z^{-1})z^{-1}}{1 + ((b_1k_0 + a_1)/a_0)z^{-1} + ((b_2k_0 + a_2)/a_0)z^{-2} + ((b_3k_0 + a_3)/a_0)z^{-3}}. \tag{23}$$

It should be noted that the pole of filter A and the first zero of filter D are equal, thus they are removed from (23).

Response magnitude in the equilibrium state is derived from (2) and (20) and it looks like

$$c_0(k_0) = \frac{1}{(l+r)} f_0(k_0) = \frac{My}{y(l+r) + lk_0} k_0. \tag{24}$$

The notion that poles of filter E coincide with those of filter D leads to a conclusion that condition of the stability of the filter is identical to that of filter D.

7. PRACTICAL OUTCOME OF THE PRESENTED RESERVOIR MODEL ANALYSIS

As it can be seen from Figure 6, the reservoir model is equivalent to a kind of signal-dependent gain-control mechanism. The presented equivalent structure may be perceived as the interpretation of the IHC adaptation mechanism from the algorithmical signal processing point of view. In the equivalent structure, filters A, B, and C are all linear time-invariant structures, the only nonlinear element here is the multiplication of the signals. Implementation of the equivalent structure via a combination of filters A and B seems more preferable among the alternatives, presented in Figure 6, since it requires less computational effort.

A brief look at the poles of filter A (6) and B (11) gives an indication that their stability conditions are identical to that of filter D (22). This fact is a direct result of employment of forward difference approximation of the differential problem in the filter synthesis. All known digital implementations of the IHC reservoir model [16, 17, 18] share this method of differential approximation. However, this limitation seems impractical from the technological point of view, since it prevents implementation of the described equivalent structure, as well as other implementations mentioned above, for signals with sampling frequency below $\sim 4,6\,\text{kHz}$ using the

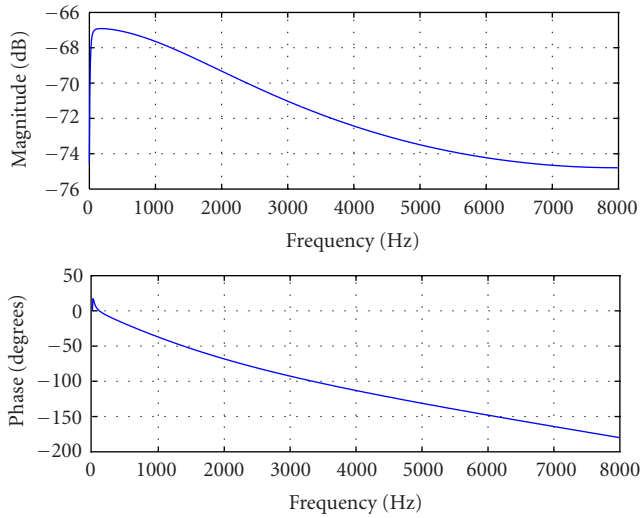

Figure 9: Frequency characteristic of filter E ($F_s = 16\,\text{kHz}$, $k_0 = 50$).

realistic values of the model parameters. Indeed, such limitation of the lowest possible sampling frequency makes efficient combination of the model with multirate cochlear filter banks impossible.

Fortunately, there exist other methods of approximation of the differential problem in the digital domain, for example, bilinear transformation. In accordance with its properties, any stable analog linear time-invariant filter, described by the corresponding differential equation, is converted into a stable digital filter. With the help of bilinear transformation, it is possible to construct universally stable digital filters A and B regardless of the sampling frequency. This procedure as well as its combination with computationally efficient implementation of the multirate cochlear filter bank is described in detail in [19].

However, in the case of bilinear transformation, unlike the situation with difference approximation, the coefficient b_0 of the filter B is not equal to zero:

$$\begin{aligned}
H_B(z) &= \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3}}{a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}},\\
a_0 &= 8F_s^3 + 4F_s^2(x + y + l + r)\\
&\quad + 2F_s\big((x + y)(l + r) + xy\big) + xy(l + r),\\
a_1 &= -24F_s^3 - 4F_s^2(x + y + l + r)\\
&\quad + 2F_s\big((x + y)(l + r) + xy\big) + 3xy(l + r),\\
a_2 &= 24F_s^3 - 4F_s^2(x + y + l + r)\\
&\quad - 2F_s\big((x + y)(l + r) + xy\big) + 3xy(l + r),\\
a_3 &= -8F_s^3 + 4F_s^2(x + y + l + r)\\
&\quad - 2F_s\big((x + y)(l + r) + xy\big) + xy(l + r),\\
b_0 &= 4F_s^2 + 2F_s(x + l + r) + xl,\\
b_1 &= -4F_s^2 + 2F_s(x + l + r) + 3xl,\\
b_2 &= -4F_s^2 - 2F_s(x + l + r) + 3xl,\\
b_3 &= 4F_s^2 - 2F_s(x + l + r) + xl.
\end{aligned} \tag{25}$$

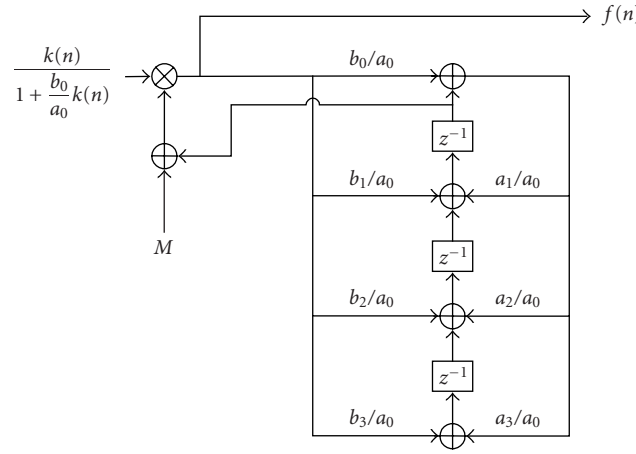

Figure 10: Transposed direct form II realization.

This fact leads to the additional operations at the implementation of the signal flow of Figure 6. Indeed, writing a set of equations describing the signal flow over the feedback loop of Figure 6 results in the following relations:

$$\begin{aligned}
f(n) &= k(n)q(n) = k(n)\big(M - s(n)\big),\\
\sum_{i=0}^{3} b_i f(n-i) &= \sum_{i=0}^{3} a_i s(n-i).
\end{aligned} \tag{26}$$

It is evident that simple substitution of the second equation into the first does not lead to the expression of the output signal $f(n)$ through the current value of the input signal $k(n)$ and previous values of the signals f and s. The current value of the output is present on the both sides of the equation. Separation of the variables leads to the following expression for the output signal:

$$f(n) = \frac{k(n)}{1 + (b_0/a_0)k(n)} \cdot \left(M - \sum_{i=1}^{3} \frac{b_i}{a_0} f(n-i) + \sum_{i=1}^{3} \frac{a_i}{a_0} s(n-i) \right). \tag{27}$$

It appears that the most computationally effective way to implement filter B with its signal feedback is a transposed direct form II structure (Figure 10). This realisation minimises the number of delay units.

For the sake of completeness of the picture, the following formula presents a version of the digital filter A, which is obtained with the help of bilinear transformation:

$$H_A(z) = \frac{1}{l + r + 2F_s} \cdot \frac{1 + z^{-1}}{1 + \big((l + r - 2F_s)/(l + r + 2F_s)\big) z^{-1}}. \tag{28}$$

The formulae (25), (27), and (28), as well as the Figures 6 and 10, contain exact instructions for the implementation of the reservoir IHC model, which remains stable at any sampling frequency. As it was noted above, this property saves computational load and is desirable for efficient incorporation of the model into multirate cochlear filter bank.

Linear approximation (23) of the reservoir model might be viewed as a computationally effective way to implement the model when input signal does not significantly deviate from a certain fixed stationary value. It might also serve as the linear time-variant filter, which simulates the reservoir model, when the slowly varying stationary value of the signal k_0 is known in advance or is estimated through a long-term moving average procedure.

This linear approximation is also important because of its link to the RASTA filtering technique [20, 21], a well-established channel normalisation and speech augmentation means in ASR. Although the nature of this link needs further investigation, both techniques represent low-passband filters, running in separate frequency channels, which are converted with the help of nonlinearity. In the case of RASTA, each frequency channel is decimated to represent one frequency bin of the short time Fourier transform spectrogram and converted into modulation-frequency domain by Jah-log transformation [16]. In the case of reservoir model there is no explicit decimation and the passband signal is transformed by "BM vibration—membrane permeability" transformation [6], which somewhat resembles Jah-log transform.

8. EXPERIMENTS

Several experiments were run in order to validate the original assumption that the anthropomorphic auditory modelling in general and IHC adaptation model in particular may indeed augment performance of the ASR systems. A comparison involved three experimental setups, which are described in more detailed fashion in [22].

(i) BASELINE: an ASR feature extraction (FE) algorithm, which is based on linear time-invariant perceptually aligned filters.

(ii) A-MORPHIC: anthropomorphic feature extraction algorithm [22], which combined linear time-variant cochlear filters to model auditory suppression and the above-described IHC reservoir model implementation. However, results mainly reflect effect of the IHC reservoir model since speech recordings in the experiment had approximately the same loudness level ($\sim$ 40 dB SPL).

(iii) RASTA: the conventional RASTA algorithm-based feature extraction [16].

In order to be effective, ASR FE algorithms should convey as much information about the speech source as possible. The measure of the amount of conveyed information, that is, the mutual information between a speech source S, which at any instant of time resides in one of the possible states C_i, $i = 1, 2, \ldots, N$, and a measured feature vector component X is defined as follows:

$$\begin{aligned} I(S,X) &= H(S) - H(S \mid X) \\ &= -\sum_{\forall C_i \in \{c\}} P(C_i) \log_2 P(C_i) \\ &\quad + \sum_{\forall C_i \in \{c\}} \int_{G(X)} P(C_i, X) \log_2 P(C_i \mid X)\, dX. \end{aligned} \tag{29}$$

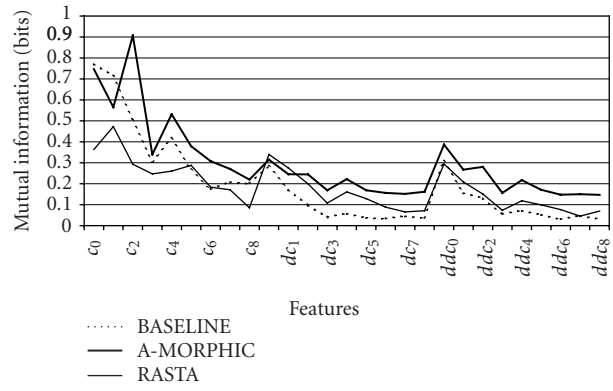

Figure 11: Mutual information of feature components ($_{\Delta}X = 0.01$).

Estimation of the mutual information has been performed with the help of the following procedure [22]:

$$\begin{aligned} I_{{}_{\Delta}X}(S,X) \approx \log_2 N + \frac{1}{N}\Bigg(& \sum_{\forall i}\sum_{\forall j} N(C_i, {}_{\Delta}X_j) \log_2 N(C_i, {}_{\Delta}X_j) \\ & - \sum_{\forall i} N(C_i) \log_2 N(C_i) \\ & - \sum_{\forall j} N({}_{\Delta}X_j) \log_2 N({}_{\Delta}X_j) \Bigg). \end{aligned} \tag{30}$$

Here N denotes a total number of feature frames in the measurement; $N({}_{\Delta}X_j)$—a number of frames when the feature value falls into the interval $[\min(X) + (j-1){}_{\Delta}X, \min(X) + j{}_{\Delta}X]$; $N(C_i)$—a number of frames which were generated in the state C_i; $N(C_i, {}_{\Delta}X_j)$—a number of frames, belonging to the certain feature interval, which were generated by the source in the state C_i.

Phonetically labelled TIMIT speech corpus was used in this experiment. Probability distributions were approximated with histograms that had a step size $_{\Delta}X = 0.01$. Results, which are presented in Figure 11, show that A-MORPHIC features are generally the most informative.

Another experiment was performed to estimate a degree of invariance of the feature vectors to different kinds of adverse interference. To provide estimates of the feature invariance degree a simple Euclidian distance between feature vectors was used. Exact experiment description may be found in [22]. Results of the experiment, which are presented in Table 1, reflect a mean distance of the feature vectors in adverse conditions to those perceived in a "clean" environment. As it can be seen from the table, A-MORPHIC features are less invariant to the adverse interference than RASTA. Anyway, a distance between "clean" and severely noisy (SNR 0 dB) features in the case of A-MORPHIC FE matches that between "clean" and mildly-noisy (SNR 30 dB) features in the BASELINE case.

Results of the depicted experiments are also supported by the reported in [22] comparison of the speech recogniser

Table 1: Expected mean distance between the feature vectors in adverse conditions and clean environment.

Feature extraction algorithm	Interference			
	Noise 30 dB	Noise 10 dB	Noise 0 dB	Convol. channel
BASELINE	0.41597	0.78894	1.05047	0.49298
RASTA FE	0.09842	0.17563	0.22338	0.05300
A-MORPHIC	0.26853	0.44951	0.42615	0.16665

performances (refer to [23] for a description of the recogniser). It's main result is that in adverse environments the recogniser with A-MORPHIC FE performs at least as good as the one with RASTA FE. These facts support the conjecture of the present paper that application of the anthropomorphic algorithms in technical devices, namely, ASR engines, is fruitful.

9. CONCLUSIONS

Analysis of the physiological model of the chemical IHC-AN synapse creates an opportunity to implement it in the form of the anthropomorphic algorithm, which is computationally efficient and thus may be used in technical devices. The equivalent digital and linearised equivalent representations create alternatives for a traditional direct difference approximation of the original set of differential equations. These representations allow for a multiple "accuracy versus computational load" tradeoffs at the implementation stage. Within the described framework, it is possible to create implementations, which remain stable regardless to the signal sampling frequency.

It was found that effect of the IHC adaptation model is equivalent to the action of signal-dependent automatic gain control mechanism. It is also conjectured that effect of the linearised equivalent representation resembles that of RASTA, an algorithm engineered with the aim of alleviating the influence of additive and convolutive noises. This interpretation of the IHC-AN synapse model gives us reasons to believe that it is important as a mean of increasing ASR robustness to the real-world environments (e.g., "too slow" and "too fast" varying additive and convolutive noises) and also as a mean of enhancement of the useful signal in the speech coding applications. Presented and referenced experiments confirm viability of the application of the discussed anthropomorphic algorithm to the ASR field. However, the exact form of the relation between the IHC-AN synapse model and RASTA should be investigated further.

ACKNOWLEDGMENTS

The authors would like to thank G. Kubin and the anonymous reviewers for the valuable insights they provided as this article was developed. This work is supported in part by the Bialystok Technical University under the Grant no. W/WI/02/03.

REFERENCES

[1] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer, Berlin, Germany, 1990.
[2] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, no. 1, pp. 3–27, 1998.
[3] D. C. Geisler, *From Sound to Synapse: Physiology of the Mammalian Ear*, Oxford University Press, New York, NY, USA, 1998.
[4] M. R. Schroeder and J. L. Hall, "Model for mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America*, vol. 55, no. 5, pp. 1055–1060, 1974.
[5] Y. Oono and Y. Sujaku, "A model for automatic gain control observed in the firings of primary auditory neurons," *Abstracts of IECE Transactions*, vol. 58, no. 6, pp. 61–62, 1975.
[6] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, 1986.
[7] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *Journal of the Acoustical Society of America*, vol. 83, no. 3, pp. 1056–1063, 1988.
[8] R. Meddis, M. J. Hewitt, and T. M. Shackleton, "Implementation details of a computation model of the inner hair-cell auditory-nerve synapse," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1813–1816, 1990.
[9] E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, "A revised computational inner hair cell model," in *Proc. 11th International Symposium on Hearing*, pp. 112–121, Grantham, UK, August 1997.
[10] C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, "A revised model of the inner-hair cell and auditory-nerve complex," *Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2178–2188, 2002.
[11] C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, "Adaptation in a revised inner-hair cell model," *Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 893–901, 2003.
[12] A. Ivanov and A. Petrovsky, "Auditory models for robust feature extraction: suppression," in *Proc. IEEE Signal Processing Workshop*, pp. 23–28, Poznan, Poland, October 2003.
[13] F. S. Perdigao and L. V. Sa, "Properties of auditory model representations," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH '97)*, vol. 5, pp. 2499–2502, Rhodes, Greece, September 1997.
[14] F. S. Perdigao and L. V. Sa, "Auditory models as front-ends for speech recognition," in *Proc. NATO Advanced Study Institute on Computational Hearing*, Il Ciocco, Italy, July 1988.
[15] D. N. Morgan, "On discrete-time AGC amplifiers," *IEEE Trans. Circuits Syst.*, vol. 22, no. 2, pp. 135–146, 1975.
[16] A. Härmä and K. Palomäki, "HUTear—a free Matlab toolbox for modeling of human auditory system," in *Proc. 1999 MATLAB DSP Conference*, pp. 96–99, Espoo, Finland, November 1999.
[17] M. Slaney, "Auditory toolbox, version 2," Tec. Rep. 1998-10, Interval Research Corporation, Palo Alto, Calif, USA, 1998.
[18] R. D. Patterson and M. H. Allerhand, "Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform," *Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.
[19] A. V. Ivanov and A. A. Petrovsky, "A composite physiological model of the inner ear for audio coding," in *Proc. 116th AES Convention*, Berlin, Germany, May 2004, preprint 6082.
[20] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[21] J. Baszun and A. Petrovsky, "Enhancement of speech as a preprocessing for hearing prosthesis by time-varying tunable modulation filters," in *Proc. 17th International Congress on Acoustics*, Rome, Italy, September 2001.
[22] A. V. Ivanov and A. A. Petrovsky, "Anthropomorphic feature extraction algorithm for speech recognition in adverse environments," in *Proc. 9th International Conference "Speech and Computer" (SPECOM '04)*, St. Petersburg, Russia, September 2004.
[23] A. V. Ivanov and A. A. Petrovsky, "Speech recognition based on hybrid neural network/hidden Markov model approach," *Neurocomputers Design and Applications*, no. 12, pp. 27–36, 2002.

Alexei V. Ivanov received the M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 1995. He received the Ph.D. degree from the Computer Engineering Department, the Belarusian State University of Informatics and Radioelectronics, in 2004. His Ph.D. thesis is entitled "Feature space construction based on anthropomorphic information processing for speech recognisers in adverse environments." In 2000, he joined Lernout & Hauspie Speech Products NV Research Laboratory, Wemmel, Belgium, as a Research Engineer. Currently he is with the Computer Engineering Department, the Belarusian State University of Informatics and Radioelectronics, working as a Researcher in the field of automated speech recognition. His research interests include application of the detailed hearing models to artificial speech processing systems and, in particular, construction of the anthropomorphic feature extraction algorithms for speech recognition, with the aim to increase its robustness towards adverse interference. Dr. Ivanov is a Member of the Institute of Electrical and Electronics Engineers (IEEE); the IEEE Signal Processing & Information Theory Societies; the Association for Computing Machinery (ACM); the International Speech Communication Association (ISCA); the Acoustic Engineering Society (AES); and the Acoustical Society of America (ASA).

Alexander A. Petrovsky received the Dipl.-Ing. degree in computer engineering in 1975 and the Ph.D. degree in 1980 both from the Minsk Radio-Engineering Institute, Belarus. In 1989, he received the Doctor of Science degree from The Institute of Simulation Problems in Power Engineering, Academy of Science, Kiev, Ukraine. In 1975, he joined Minsk Radio-Engineering Institute. He became a Research Worker and Assistant Professor and since 1980 he has been an Associate Professor at the Computer Science Department. From 1983 to 1984, he was a Research Worker at the Royal Holloway College and the Imperial College of Science and Technology, University of London, UK. Since May 1990, he has been a Professor and Head of the Computer Engineering Department, the Belarusian State University of Informatics and Radioelectronics, and he is with the Real-Time Systems Department, Faculty of Computer Science, Bialystok Technical University, Poland. Recently his main research interests are in acoustic signal processing, such as speech and audio coding, noise reduction and acoustic echo cancellation, robust speech recognition, and real-time signal processing. A. A. Petrovsky is a Member of the Russian A. S. Popov Society for Radioengineering, Electronics and Communications, and an Editorial Staff Member of the Russian journal Digital Signal Processing, AES, IEEE, and IIAV.