19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

MODELS OF TIMBRE USING SPECTRO-TEMPORAL RECEPTIVE FIELDS: INVESTIGATION OF CODING STRATEGIES

Elhilali, Mounya[1]; Shamma, Shihab[1]; Thorpe, Simon J.[2]; Pressnitzer, Daniel[3]
[1] Institute for Systems Research; University of Maryland; A.V.Williams Bldg, College Park, MD 20742; USA; {mounya,sas}@umd.edu
[2] Centre de Recherche Cerveau Cognition, CNRS-Université Paul Sabatier Toulouse 3 Faculté de Médecine de Rangueil, 31062 Toulouse Cedex 9 simon.thorpe@cerco.ups-tlse.fr
[3] Laboratoire Psychologie de la Perception, CNRS-Université Paris Descartes & DEC, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France; Daniel.Pressnitzer@ens.fr

ABSTRACT
Timbre designates all of the perceptual characteristics of sounds that cannot be described as pitch, loudness or duration. Behavioral experiments combined with multidimensional scaling techniques have proposed that a few main acoustic dimensions subserve the perception timbre for homogeneous ensembles of sounds (e.g., Western musical instrument sounds). It is unclear however whether these dimensions can describe all aspects of timbre, and, most importantly, that they can capture the subtle differences that allow for sound source recognition. Here we investigate a computational model of timbre that has in principle a large number of dimensions, but of which only a small subset are used to describe each individual sounds. These dimensions are based on spectro-temporal receptive fields (STRFs) obtained from physiological recordings in the primary auditory cortex, and capture a multiscale cortical representation of dynamic sound spectra. The STRF model is used here to predict listener's perception of differences between musical instruments, and between musical and vocal sounds. This technique presents a biologically plausible approach to reproduce perceptual results, and offers an alternative view to understand timbre perception

INTRODUCTION
Natural sounds carry a wealth of information about the environment. Our brains are able to effortlessly integrate a multitude of acoustic cues constantly arriving at our ears, and to derive coherent percepts and judgments about the varied attributes of the sound. This facility to analyze an auditory scene is based on a multi-stage process in which sound is analyzed in a myriad of neural circuits, populated with neurons which encode a multitude of features and levels of abstractions of the acoustic information. At the peripheral stages, auditory nuclei exhibit well-defined tuning characteristics. In contrast, later stages exhibit a more complex and labile behaviour, presumably to be able to express context-sensitive rules that can influence auditory scene analysis and sound identification [1]. Decades of physiological and psychoacoustical studies [7,10] have revealed elegant strategies at various stages of the auditory system for the representation of the signal cues underlying auditory perception.

Of all attributes that have been defined to describe the perception of acoustic signals, *timbre* remains the most mysterious and least amenable to a simple mathematical abstraction [7]. Unlike pitch and loudness, it has resisted descriptions along an ordered scale, and instead has fallen along impressionistic juxtapositions such as sharpness-dullness and transient-sustained. The multifaceted complexity of the timbre percept essentially stems from its sensitivity to spectral *shape* and temporal *dynamics,* cues that cannot be readily captured along simple dimensions. Furthermore, they often co-vary with other cues such as intensity (loudness) and fundamental frequency (pitch) when produced by real instruments. Thus, the timbre of a musical note changes substantially when it is “plucked” instead of “bowed”, muted instead of amplified, or when played on an open string versus a fingered placement [6,8].

We report in this article on recent efforts to develop a model of auditory processing that can potentially account for major aspects of the timbre percept. The physiologically inspired model is

assessed by its ability to generate a relational structure of musical timbre [13]. The model consists of two stages: a *representational* stage in which sound is transformed into a pattern that embodies its perceptually significant spectro-temporal features. It is followed by a *classification* stage that organizes the perceptual patterns due to different musical instruments into a pair-wise confusion-matrix and aims to predict the "perceptual-distances" among them. In a first set of simulations, we show that the model can qualitatively capture interesting differences between various types of instruments, such as their manner of production. This property is dependent on the spectro-temporal analysis used, as selecting only spectral or only temporal cues fail to produce such a classification. In the second part of the study, we compare the model simulations to behavioural data. We collected timbre dissimilarity judgments between musical instruments, using the same instruments for both psychophysics and the model simulations. A high, significant correlation was observed between the model predictions and the listeners' judgments. Finally, we examine biologically plausible coding strategies involving spike-timing information and show that this could reduce the computational cost of the classification stage.

I. THE COMPUTATIONAL MODEL

Stage1: Cortical Representation Space

Sound in its journey from the eardrum to the cortex undergoes a profound transformation from a simple one dimensional temporal pressure waveform to an elaborate multidimensional representation. In models of the early auditory stages [5,16], the acoustic signal is transformed into an "auditory spectrogram" - a frequency-time representation that is the end-result of frequency analysis in the cochlea, followed by edge detection and temporal smoothing. Figure 1(A) depicts the auditory spectrogram of a violin note, whose spectral and temporal cross-sections are depicted by the bold lines in Fig.1(C). The spectral structure of the sound has many harmonic peaks (due the pitch of the note) and an overall envelope (dashed bold line) reflecting the resonances of the body of the violin. It also highlights the temporal modulations in some channels that reflect the soft onset, sustained bowing, and the vibrato (bold line). By contrast, the temporal and spectral modulations of a piano (playing the same note) are quite different (thin lines in Fig.1(C)). Temporally, the onset of a piano rises and falls much faster, and its spectral envelope (dashed thin line) is smoother.

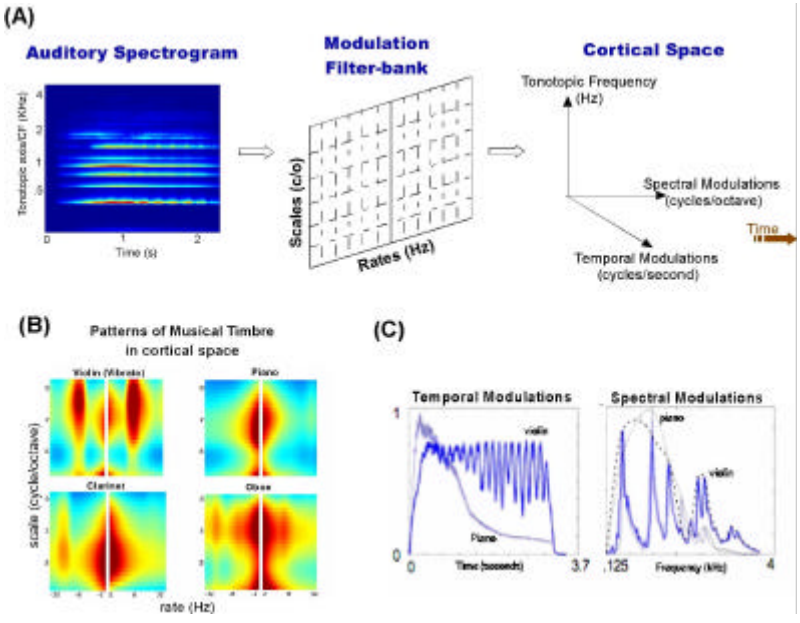

Figure 1: Representation of sound in a higher-dimensional cortical space. **(A)** Sound is first transformed into an auditory spectrogram in peripheral stages. Its modulation content is then analyzed in the cortex by an array of modulation-selective "filters", (STRFs). (**B)** Viewed in the cortical space, each instrument has a distinctive signature pattern that captures its spectro-temporal activation, and hence reflects its timbre. Red indicates strong responses, Blue is weak. **(C)** A contrastive view of the temporal modulations (left) and spectral modulations (right) of a piano and violin playing the same note.

These spectro-temporal modulations are analyzed in subsequent stages of the model, which can be used to predict the varied percepts produced by the sound [2]. The analysis mimics aspects of cortical auditory processing. It consists of a bank of "modulation selective filters" that detect various spectro-temporal features ultimately created by the instrument and player. Each cortical filter (usually referred to as the spectro-temporal receptive field, or STRF) is "tuned" or best activated by a particular patterns of spectral peaks and temporal rates. The auditory cortex

contains a large variety of such STRFs with different spectral bandwidths, asymmetries, dynamics, and directional preferences [12,14], hence giving rise to the description of such an analysis as a *multi-scale and multi-rate analysis* along the spectral and temporal dimensions of the spectrogram, respectively. Therefore, an ordered bank of such multi-resolution filters, tuned to a range of bandwidths and dynamic rates, provides by its responses a unique characterization of the spectrogram, one that is sensitive to the spectral shape and temporal dynamics over the entire stimulus. Mathematically, this "cortical" model performs a two-dimensional wavelet analysis of the spectrogram (with all details provided in [2]). Figure 1B illustrates the responses generated by four instruments (violin, piano, clarinet and oboe) playing the same note (G3), and averaged over their spectral dimension and durations. Each of the panels provides an estimate of the distribution of energy in the various spectral and temporal modulations of the sound. For instance, the vibrato of the violin concentrates its peak energy near 6 Hz, while by contrast the rapid onset of the piano distributes its energy. Similarly, the unique pattern of peaks and valleys in the spectral envelopes of each instrument produces a broad distribution along the bandwidth axis, whereas the piano's smooth profile activates broad bandwidths. Each instrument, therefore, produces a correspondingly unique spectrotemporal activation that could be used to recognize it or distinguish it from others.

Stage 2: Classifying Cortical Activation Patterns

The hypothesis behind this work is that the activation patterns generated by the cortical model reflect closely the perceptually-significant features of the timbre of the signals. Consequently, it should be possible to organize the patterns according to their similarity or differences, and thus predict our perception of their timbre distances. In the following simulation, sets of instrument sounds were processed through the cortical model and the full output was then further analyzed to compute a matrix of the pair-wise distances among the instruments. This matrix measures the distances between the representations of any pair of instruments as viewed by the cortical model. It reflects how similar the instruments are for the model, and, we hypothesize, how 'perceptually similar' these instruments would sound to human listeners. To construct this matrix, we compared the scale-rate plots of all instruments (similar to those shown in Fig.1(B)) using the L2-norm as a measure of distance (or dissimilarity).

II. USING THE MODEL TO EXPLORE MUSICAL INSTRUMENTS SPACE

Method

For a first test of the model, we extracted musical sounds from the well-known RWC Music Database [3]. The set tested included: Piano (PF), Classical Guitar (CG), Harp (HP), Vibraphone (VI), Marimba (MB), Cello (VC), Violin (VN), Flute (FL), Oboe (OB), Clarinet (CL), Harmonica (HM), Trumpet (TR), Trombone (TB), and Bassoon (FG). Each instrument was tested for 3 different notes (B2, Bb3, Csharp3), as well as up to 3 different instrument manufacturers. On average, 3-5 tokens were recorded from each instrument and used in the analysis described here. The full model was run, and the similarity matrix constructed. Each entry in the matrix is the *average* distance between the corresponding two instruments, measured by first computing the distances between the samples for the two instruments played at the same note, and then averaging the results across all notes and all samples.

Results

The results of the cortical patterns comparisons are shown in Figure 2. The broad organization that emerges is that instruments that share a *manner* of production (e.g., bowed or plucked), and/or important *physical attributes* (e.g., wood box or strings) tend to cluster together in the matrix. For instance, two classes of instruments are roughly distinguished in Fig.2(left): transient *versus* sustained, e.g., percussive (plucked, and struck) instruments *versus* wind and bowed instruments. Finer distinctions within each group are also apparent. For instance, within the transient group, the "string" instruments (Piano, Classical Guitar, and Harp) cluster together apart from the two non-string instruments (marimba and Vibraphone). Within the sustained group, the string instruments (Violin and Cello) cluster apart from most wind instruments. The latter group further subdivides into two types: Flute/Oboe/Clarinet/Harmonica, and the brass instruments Trombone and Tuba. The Bassoon remains somewhat apart.

The features that give rise to these classifications in the cortical model are varied but can be roughly divided into temporal and spectral. Thus, the primary division of all instruments into sustained and transient clearly reflects the temporal dynamics of the sound, which in turn

primarily correlates with the *manner* of sound production. However, within each of these two classes of sounds, spectral cues take over, giving rise to the stringed *versus* non-stringed distinction, and the further subdivisions within each which are likely related to the *physical attributes* of the instruments. To examine explicitly the contributions of these temporal and spectral factors, we re-computed the distance matrices based *solely* on either the rate (dynamics) or the scale (spectral) dimensions by effectively collapsing one of the dimensions in the scale-rate plots (Fig.1(B)). Therefore, each instrument is now characterized by either a one-dimensional scale or rate profile. The results of the purely-temporal matrix are shown in Fig.2(middle), which primarily replicates the transient/sustained distinction described earlier. Note that within these two groups, the clustering of the instruments is now more cohesive since all share very similar dynamics. Relatively small distinctions among the instruments remain, however, presumably due to vibrato, shimmer, and other dynamic features that are unique to some instruments (e.g., the violin). The matrix due to the purely-spectral cues (Fig.2(right)) shares some of the major classifications of the temporal matrix, e.g., the transient/sustained grouping, indicating the consistency of dynamic and spectral cues. However, spectral cues on their own are quite effective in distinguishing many of the instruments, e.g., Vibraphone/Marimba, Piano/Guitar, Clarinet/Harmonica, and Trombone/Tuba. Therefore, the overall conclusion of this qualitative analysis is that relying exclusively on temporal or spectral cues is not as reliable as making use of both types of cues.

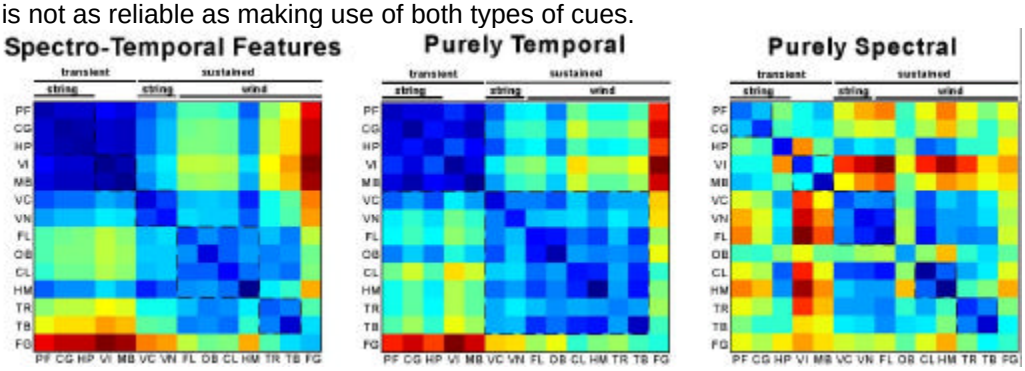

Figure 2: Confusion matrices representing the timbre-distance between any pair of instruments. The instruments featured are the Piano (**PF**), Classical Guitar (**CG**), Harp (**HP**), Vibraphone (**VI**), Marimba (**MB**), Cello (**VC**), Violin (**VN**), Flute (**FL**), Oboe (**OB**), Clarinet (**CL**), Harmonica (**HM**), Trumpet (**TR**), Tuba (**TB**), and Bassoon (**FG**). Red (Blue) indicates largest (smallest) distances between any pair of instruments. **(Left)** The confusion matrix computed from the complete rate-scale plots of all indicated instruments (as in Fig.1(C)). **(Middle)** The confusion matrix computed from *only* the temporal dynamics of the sounds (roughly, collapsing the scale axes in the rate-scale plots of Fig.1(C). **(Right)** The confusion matrix computed from *only* the spectral modulations of the sounds (roughly, collapsing the rate axes in the rate-scale plots of Fig.1(C).

III. PREDICTING PERCEPTUAL SIMILARITY WITH THE TIMBRE MODEL

Method

To further investigate the relation between model dissimilarity and perceptual dissimilarity, we ran a psychoacoustical test of timbre judgments. Normal-hearing listeners had to judge the dissimilarity between two timbres that were played with the same pitch. The instruments tested behaviorally were also extracted from the RCW musical database [3], and consisted of: Piano (PF), Vibraphone (VI), Marimba (MB), Cello (VC), Violin (VN), Oboe (OB), Clarinet (CL), Trumpet (TR), Bassoon (FG), Trombone (TB), and Saxophone (Sx); in addition to a voice singing the vowel ‘A’ (Av) and ‘I’ (Iv). The note played was D4 for all sounds. The listening test was performed on 7 subjects. Listeners made subjective judgments of dissimilarity between each pair of sounds, in both orders. Prior to the experiment, subjects were acquainted with the range of timbre differences. Listeners were allowed to listen to the sounds as many times as desired before they made their judgment on a continuous subjective scale. For the model simulations, we replicated the analysis described before but using the single tokens of the set of sounds tested behaviorally.

Results

In Figure 2B(left), we show the psychoacoustical distance matrix obtained from the human listening tests. The behavioral results reveal a good agreement among subjects, as indicated by the relatively small values of the standard deviation (middle panel). In addition, the dissimilarity

matrix is fairly symmetrical around the diagonal (left panel), allowing us to combine both half off-diagonals into the upper half matrix for comparison with the simulation results (Fig.2B rightmost panel). The model simulations with the behaviorally-tested set of sounds are shown in Figure 3(A,left). Both the original analysis using average measures (Fig. 2(left) and this analysis with single tokens (Fig. 3(A,left)) show qualitatively similar patterns at the corresponding instruments. We then compared the model's classification with the behavioral panel results. Visual inspection indicates that the major features of the dissimilarity judgments are captured by the model. Quantitatively, the model yields a good match to the behavioral data; with a statistically-significant correlation coefficient r=0.7 ($p < 10^{-5}$).

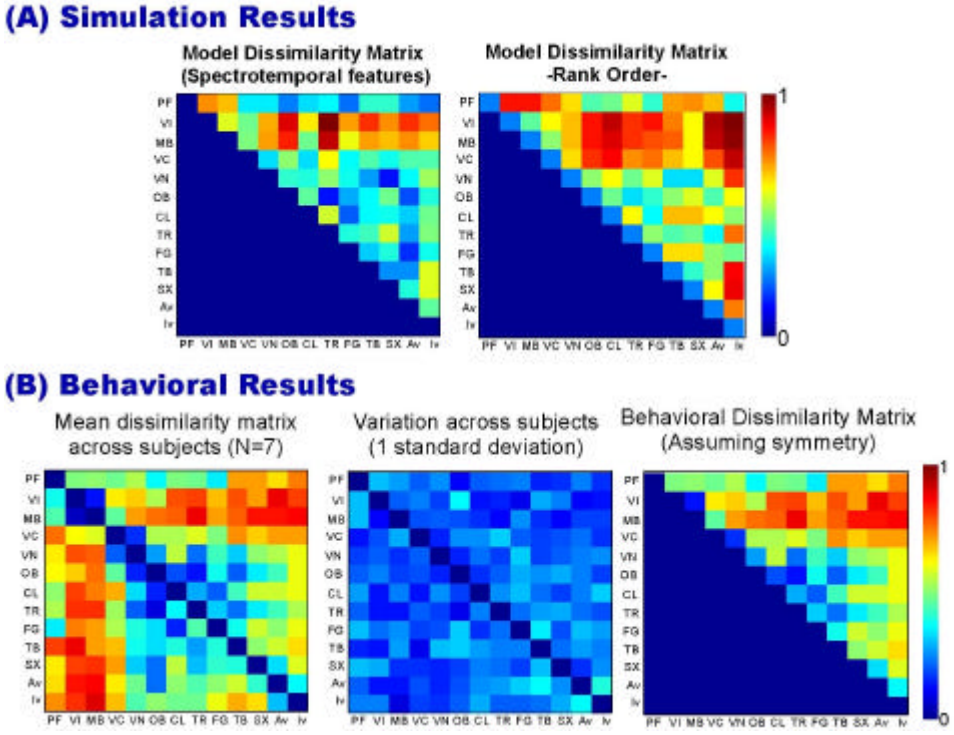

Figure 3: Simulation and Behavioral Results of Musical Timbre Judgments. **(A)** The *left* matrix shows the pair-wise distances between every instrument pair, based on their corresponding cortical activation patterns. The *right* panel shows a similar dissimilarity matrix obtained from the spike-timing information of the cortical activation of each sound. (**B)** Behavioral judgments of human subjects listening to the same sets of sounds yield a nearly symmetrical matrix, with a small variance across subjects (right panel).

IV. EXPLORING TIME-BASED CODES FOR TIMBRE

Timing-based Coding strategies

The L2-norm that we used to estimate the distance between model outputs is conceptually simple but would in practice require a comparison between the rates of activation of all possible feature detectors. This is a very expensive strategy in terms of possible neural computation. We thus investigated whether a sparser code could be found by using the information contained in the time course of activity of the STRF-based model. The use of spike-timing as a source of information in auditory processing has been hypothesized for pitch [11], spatial localization [9], and loudness [4] before. Here, we explore aspects of spike-timing to extract information about the *timbre* of these natural sounds.

We chose to investigate the rank-order principle as an efficient coding strategy that has been successfully implemented in the identification of visual natural objects [15]. A rank order code is based on the principle that neural feature detectors that are well matched to the sensory input will tend to produce spikes rapidly, whereas a poorer match will produce late spikes or no spikes at all. The sequence of spikes (which neuron fired first, which neuron fired second, etc.) is the code that is used to characterize neural activity. To estimate the distance between activity patterns produced by two inputs, it is thus not necessary to compute the rate of activation in all feature-detectors: the first few active feature-detectors are identified for each given input, and these sparse sequences of activation are then compared to produce the distance estimate [15].

Applying the rank-order classification strategy to timbre dissimilarity estimation

We used the firing pattern obtained from each STRF in the cortical model as a time-evolving template of the activation driven by every given sound. The spiking time for a given STRF was computed by integrating its energy output and producing a spike when an arbitrary threshold was reached. This firing-rank template was then compared to the STRF firing ranks of any other arbitrary sound hence building a new dissimilarity matrix between every pair. In our current analysis, we encoded the initial 50 spikes from a population of about 400 feature-selective STRFs. This yields a powerful reduction in the information encoded. The dissimilarity matrix obtained is shown in Figure 3 (A, right). Despite the information compression, rank order coding is as good as the full L2-norm measure, it even results in an apparent improvement in predicting the human results with a correlation coefficient of r=0.76 ($p < 10^{-5}$).

CONCLUSIONS

The hypothesis investigated in this work is that timbre may be quantified in terms of the sound-induced activity in a model of auditory cortical processing. Using a set of natural instruments and introducing some variability in the sound tokens used for each instrument, we found that the model provided a classification that was consistent with intuition. Using single tokens and a restricted instrument set, we observed a significant correlation between model-predicted dissimilarity and perceptually-measured dissimilarity. More experiments need to be performed to confirm these preliminary results, but a few speculations might be drawn from the current study. First, on the nature of the appropriate representation to predict timbre dissimilarity: according to our first set of simulations, purely temporal or purely spectral analyses are not as versatile in segregating instruments nor in providing insights into their hierarchical relationships as spectro-temporal representations. In fact, it seems that a simple "superposition" of the results is not sufficient either. Instead, it is essential to perform the classification on the full spectro-temporal representation since collapsing one or the other axes likely distorts or destroys the *joint* spectrotemporal features in the representation. Second, even though a strategy based on distance measurement between a very large set of feature detectors might seem inefficient, biologically-plausible coding strategies can be proposed to reduce the complexity of the problem while avoiding any significant loss in predictive power.

References: [1] A. S. Bregman: Auditory scene analysis: The perceptual organization of sound. Cambridge, Massachusetts: MIT Press (1990).
[2] T. Chi, P. Ru, S. A. Shamma: Multiresolution spectrotemporal analysis of complex sounds, J of Acoust Soc of America, **118, No.** 2 (2005) 887-906.
[3] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka: RWC music database: Music genre database and musical instrument sound database, Proc of ISMIR (2003).
[4] P. Heil, D. R. Irvine: First-spike timing of auditory-nerve fibers and comparison with auditory cortex. Journal of Neurophysiology, **78, No. 5** (1997) 2438-54.
[5] R. Lyon, S. A. Shamma: Auditory Computation, volume 6 of Springer Handbook of Auditory Research, chapter: Auditory representations of timbre and pitch, Springer-Verlag New York, Inc., (1996) 221-270.
[6] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, J. Krimphoff: Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes, Psychol. Research **58** (1995) 177-192.
[7] B. Moore: An Introduction of the Psychology of Hearing, Academic Press, London (1989).
[8] J. Marozeau, A. de Cheveigne, S. McAdams, S. Winsberg: The dependency of timbre on fundamental frequency, Journal of the Acoustical Society of America, 114, No. 5 (2003) 2946-2957.
[9] I. Nelken, G. Chechik, T. D. Mrsic-Flogel, A. J. King, J. W. Schnupp: Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. Journal of Computational Neuroscience, **19, No. 2** (2005) 199-221.
[10] J. O. Pickles: An Introduction to the Physiology of Hearing Academic Press, Auditory Computations, Ed. by H. Hawkins and E T. McMullen and A. Popper and R. Fay, 221—270, Springer Verlag (1988).
[11] D. Pressnitzer, A. de Cheveigne, I. M. Winter: Physiological correlates of the perceptual pitch shift for sounds with similar waveform autocorrelation, Acoustic Research Letters Online, **5, No. 1** (2004) 1-6.
[12] Read HL, Winer, JA, Schreiner, CE: Functional architecture of auditory cortex. Current Opinion in Neurobiology, **12** (2002) 433-440.
[13] P. Ru, S. A. Shamma: Representation of Musical Timbre in the Auditory Cortex, J of New Music Research, **26 No. 2** (1997) 154-169.
[14] Shamma S., J. Fleshman, P. Wiser: Response Area Organization in the Ferret Primary Auditory Cortex, Journal of Neurophysiology, **69, No. 2** (1993) 367-383.
[15] S. Thorpe, A. Delorme, R. Van Rullen: Spike-based strategies for rapid processing. Neural networks, **14, No. 6-7**(2001)715-725.
[16] K. Wang, S. A. Shamma: Self-normalization and noise-robustness in early auditory representations. IEEE transactions on speech and audio processing, **2, No. 3** (1994) 421-435.