Frequency and Pitch Representation Using Self-Organized Maps

Christos Zarras*1, Konstantinos Pastiadis#2, George Papanikolaou*3, George Papadelis#4

*Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki, Greece
#Department of Music Studies, School of Fine Arts, Aristotle University of Thessaloniki, Greece
[1]chzarras@auth.gr, [2]pastiadi@mus.auth.gr, [3]pap@eng.auth.gr, [4]papadeli@mus.auth.gr

ABSTRACT

Previous works on computational approaches for the description of pitch phenomena have employed various methodologies, deterministic and probabilistic, which are based on psychophysiological auditory stimuli modeling, representations and transformations (e.g. spatial, temporal, spatiotemporal), both at peripheral and more central stages of the auditory chain. Then, a confirmatory phase, utilizing data from behavioral (or even imaging) studies, is usually followed to assess the validity of the computational methods.

The human auditory perception relies on interconnected neuronal networks, which have been shown to demonstrate multi-directional activity and dynamical, adaptive, and self-organizing properties, together with strong tonotopical organization along the auditory pathway up to the primary auditory cortex.

This paper focuses on the exploration of properties and effectiveness of a certain type of computational approaches, namely self-organized networks, for the description of frequency and pitch related phenomena. A Self-Organized connectionist model is presented and tested.

We explore the ability of Kohonen type neural networks (Self-Organizing Feature Maps, SOFMs or SOMs) to organize based on frequency information conveyed by sound signals. Various types of artificially generated sound signals (ordered along a frequency/pitch axis) are employed in our simulations, including single tones, harmonic series, missing fundamental series, band limited noises, and harmonics with formants. Simple Fourier representations and their physiologically plausible frequency-to-pitch mappings (e.g. tonotopy in the cochlea) are used as network inputs. The networks' efficiency is investigated, according to various structural parameters of the network and the organizing procedure, together with aspects of the obtained tonotopical organization.

Our results, using different types of input spectra and various SOM implementations, demonstrate a clear ability for self-organizing according to (fundamental) frequency or pitch. However, when certain test configurations were used, the networks showed observable inability to organize, revealing limitations in the resolving ability of the network related to the required number (density) of neurons compared to the dataset size. Some more difficulties were also observed, relating to the type of signals for which an organized network can identify pitch.

The results of this work indicate that, under some provisions, such a model could be effective in frequency and pitch indication, within certain limitations upon training parameters and types of signals employed. Further work will compare the efficiency of the proposed representation with classical computational approaches upon various aspects of pitch perception, together with examination of feasibility and possible advantages of employing SOMs in the description of pitch perception in various types of auditory dysfunction.

Keywords: pitch perception, tonotopical organization, self-organized maps, Kohonen maps

Topic areas: Pitch and tonal perception, Computational Modeling of Music, Acoustics and psychoacoustics

I. INTRODUCTION

Pitch perception is a common but still intriguing field in music perception. Various computational models have been proposed for the explanation of pitch perceptual effects. The complexity of human auditory system, that is constituted by functionally different interconnected subsystems, compound with the amount of pitch phenomena that has to be illustrated makes the development of a unified model challenging.

In literature three main approaches exist, with which most classical models can be associated, temporal, spatial and spatio-temporal. Temporal approach models have their roots in Licklider's (1954) initial model. They exploit features extracted from time domain representations of the acoustic signal, such as peak to peak distances, basilar membrane neural firing timings and autocorrelation functions. On the other hand, spatial approaches are based on spectral characteristics of the input signal and the way these are displayed along an implied dimension reflecting representations of tonotopic ordering (e.g. external ear). In this category, Terhardt's model (1979) (1982) utilize common sub-harmonics, Wightman's (1973) model employ the autocorrelation of spectra, while Houtsma and Goldstein (1972) match spectral excitation pattern to harmonic templates. Finally, spectro-temporal models (Meddis & Hewitt 1991) combine a spatial model of the basilar membrane that filters the input signal, with a subsequent temporal model such as autocorrelation.

Most of the above described modeling approaches towards the description and/or prediction of psychophysical performance in relation to tonality (rather in terms of physical or perceptual units and less of musical meaning) rely in a big part to evidence from physiology, but in addition employ several assumptions on structures and functions of audition. Based on such assumptions, several tonality phenomena have been addressed such as frequency discrimination or pitch scaling. However, these assumptions are not always possible to be supported by physiological or behavioral evidence (Hawkins et all, 1995). Additionally, and most importantly, their predictions refer mainly to normal hearing, which could pose weakness or severe deflections from behavioral data when other modes of hearing are considered, e.g. impaired hearing.

A possible route to overcome such problems of incompatibility (as for example in impaired hearing) could be to attempt to make special adaptations of the above models (e.g. by adoption of proper arithmetic functions) in order to comply with findings and data from such cases.

Another possibility could be to exploit well documented properties of the human neural and cognitive systems such as learning and adaptation through the flow of information, avoiding specific and possibly restrictive assumptions of certain types of processing. Artificial neural networks (or connectionist models) that exhibit properties of self-organization could be a promising candidate.

Actually, such an approach would be in accordance with significant amount of evidence on the type and functionality of representations and organization of pitch towards more central areas of the auditory system, where tonotopy and abstraction along with feature detection have been detected (Hawkins et all, 1995), (Bharucha,1998).

In addition, there exist several conceptualized and computationally realized approaches of neural association and learning for pitch tasks (Bharucha,1998). For example, in the spirit of spatial models for pitch perception, Taylor and Greenhough (1994) used the ART NN employing supervised learning for training. The network acquires the ability to perform the required operation gradually, as more combinations of input-output data are presented. Cohen and Grossberg's (1995) SPINET model uses a model of hearing periphery and produces a distribution of strength values across a spatial representation of pitch as output, rather than the frequency of most likely pitch. Additionally, it is generally accepted that neural networks which form the human auditory system demonstrate tonotopical organization properties. In analogy to these, artificial self-organizing networks require no a priori knowledge of the relationship between input and output. Consequently, no set of patterns is inserted into the system, in order to be compared with frequency representations of the input signals. Training is conducted through continuously feeding the network only with input data. The procedure clearly resembles the function of human brain. Bharucha (2009) suggests that the principles of self-organization may extend across domains, such as auditory and music cognition, even though different representations are used as input.

This work is a preliminary step in the investigation of the value of artificial Self-Organization in describing and modeling aspects of pitch perception, within the general framework of connectionist models.

The main objective of this work is to investigate the main factors that affect the ability of a Kohonen's Self Organizing Feature Map (SOFM) network to be organized when a rich (regarding pitch) dataset of spectral representations is presented as input. The resulting organized maps are compared between various executions of a self-organization experiment, when different organization parameters (such as network size and shape, frequency distance between successive pitches, organization epochs, and initial neighborhood of neurons) and input datasets are used. The input datasets, in these experiments, are Fourier transforms of artificially generated sound signals. Mel transform is also employed in some experiments and its effect, as an intermediate to quantifications of the fundamental into pitch, is recorded.

II. Method

In our paper, we study aspects of the efficiency of a SOFM in frequency or pitch representation by making extensive use of test experiments with various combinations of signals and parameters. The information fed to the SOFM consists of spectral representations of the examined signals. These representations are further organized and conditioned using various types of parameters, thus constituting the experimental datasets. Finally, the SOFM efficiency is compared along combinations of the experimental datasets and additional configurations of parameters regarding the SOFM itself.

A. Description of the testing/experimental procedure

In this study, we investigated the competency of Kohonen's Self Organizing Feature Maps to be self-organized according to the harmonic content of periodic signals. That is, the self-organization is examined against orderings of the root (fundamental) of harmonic series spectra or their implied pitches. Initially, each network was presented and organized with a set of frequency representations of artificially generated sound signals. The members of each dataset were Fourier transforms of signals whose pitch varied in a predefined range. A detailed description of the datasets follows in B.

After the network organizing phase has completed, the same dataset was presented to the network as a test input. A different dataset was then used as input in order to test the ability of the network to generalize between different types of signals, by identifying the same pitch.

B. Datasets

The input datasets were Fourier Transforms of artificially generated sound signals. Each dataset consisted of signals of the same type. The types of signals and their respective spectra were: pure tones (only the fundamental present), complex tones composed as harmonic series, complex tones composed as harmonic series but with the fundamental missing, harmonic series filtered by a formant filter, Mel transformed pure and complex tones. An additional factor tested, was the logarithmic scale of frequency amplitude.

The sampling frequency was 44100Hz, the FFT size was 1024 samples. The frequency range for pure tones and the fundamental of complex harmonic tones was from 100Hz to 4000Hz in steps of 10Hz. This leaves us with 391 signals (spectral representations) within each dataset. In experiments with Mel transformed spectra, the O'Shaugnessy's equation was used for Hz-to-Mel transformation:

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right)$$

C. Network configurations

The SOFM network, initially, consisted of 400 neurons, arranged in a 20x20 square in gridtop topology. The training epochs were 200. The ordering phase during which weigh vectors of neighbouring neurons, besides the winning neuron, alter their values, lasted for 100 epochs. The initial neighbourhood, which includes the neurons that alter their weigh vectors at the first epoch was 15 neurons. The training was conducted in batch mode.

D. Tools

For all experiments, the Matlab software was used for the signals generation, network training and results presentation. All parameters that are not included in C but can be configured in Matlab, were set to the default values of train function.

III. Results

The simulations conducted in this work can be classified in three main categories. The first category consists of experiments that test the ability of a self-organizing map to identify pitch in a dataset that includes different type of signals than that used in the organizing phase.

Initially, a 20x20 neurons SOFM was organized by pure tones ranging from 100Hz to 4000Hz. The set of figures below presents the input dataset, the number of hits for each neuron during the organizing phase, and the distribution of input signals-pitches in the network when the organized network is presented with the same dataset.

In Figure 1 the amplitude of FFT coefficients is in linear scale, resulting in narrow areas of non-zero values. Figure 2 displays the number of dataset signals used during organization, for which each neuron was the winning one. The signals were allocated to neurons throughout the network, but in some cases, a single neuron was the winning for up to four signals. In figure 3, a route of successive pitches through the network is obvious revealing a tonotopic organization.

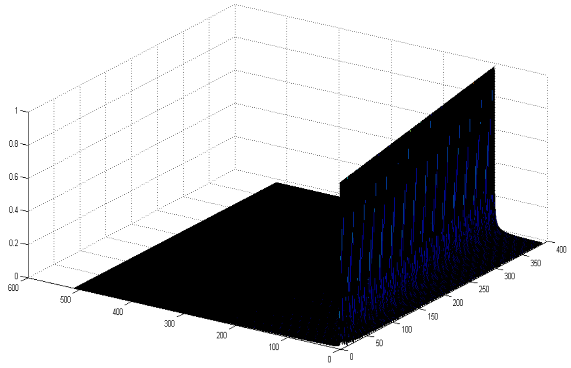

Figure 1. Input dataset consisting of FFT transformed pure tones. The three axes represent signal index (391 spectra), frequency bins (512 frequency points of one-sided FFT) and spectral amplitude respectively.

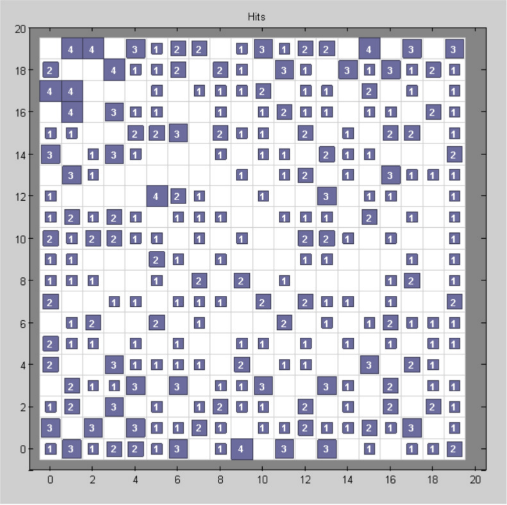

Figure 2. Sample hits during training for a 20x20 SOM network. Axes represent indices of the neurons' within the SOM's structure (for example a grid of 20x20 neurons, without any implied a-priori ordering). Each number within the grid (appending to the respective cell, namely indexed neuron) represents the number of dataset's members, namely spectra of the test signals, that were assigned to the respective neuron during training.

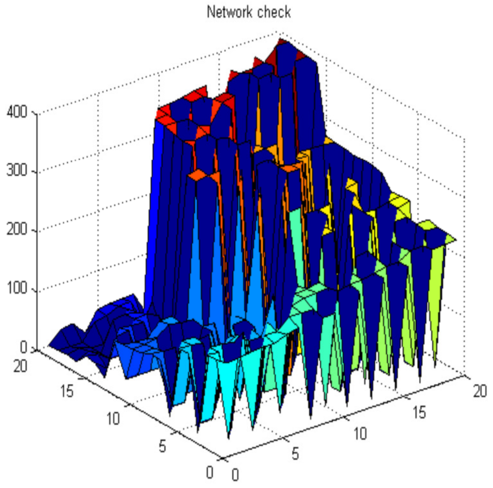

Figure 3. Pitch distribution of test signals in the organized network (20x20 neurons network, 400 test signals with increasing pitch). Pitch increases with height.

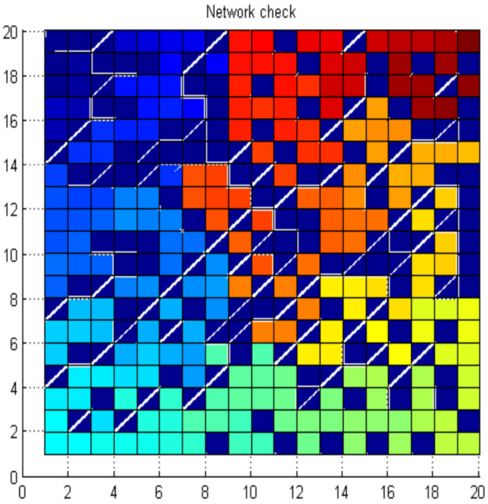

Figure 4. Pitch distribution of test signals in the organized network (20x20 neurons network), in a 2D representation. Pitches range from blue (low) to red (high).

In an attempt to use the above organized network to identify pitches in other datasets, it was presented with complex tones consisting of harmonics with and without the fundamental, as well as formant filtered or not. Figures 4, 5 and 6 display the network response to complex tones.

Nevertheless, when the training was conducted with the complex tones dataset consisted of harmonics, the training sample hits (figure 7), as well as the network response to the same dataset (figure 8) present a fine monotonic organization.

From these figures it is obvious that pitches of signals belonging to a specific dataset cannot be identified by a SOFM organized by a different dataset. In the case of pure tones training and harmonics testing, the rich frequency content at high frequencies, even for low pitch signals, results in the activation of neurons that where assigned to high pitches during training. The same applies to formant filtered harmonic sequences, where the estimated pitch of almost all signals, concentrates at neurons assigned to formant frequencies.

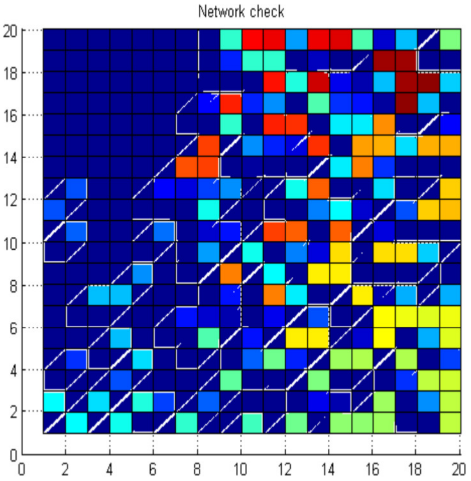

Figure 4. Network response to complex tones (fundamental + harmonics) (Blue corresponds to lower pitches while red to higher)

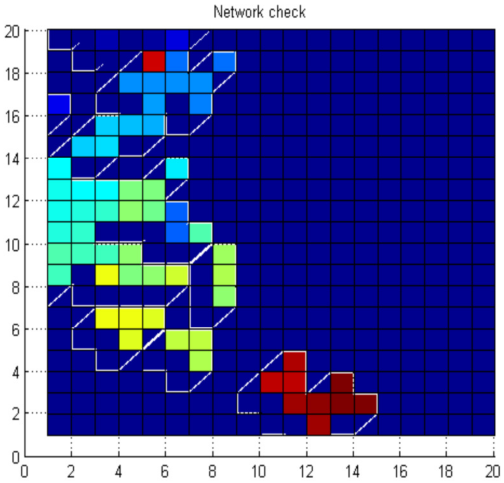

Figure 5. Network response to formant filtered complex tones (Blue corresponds to lower pitches while red to higher)

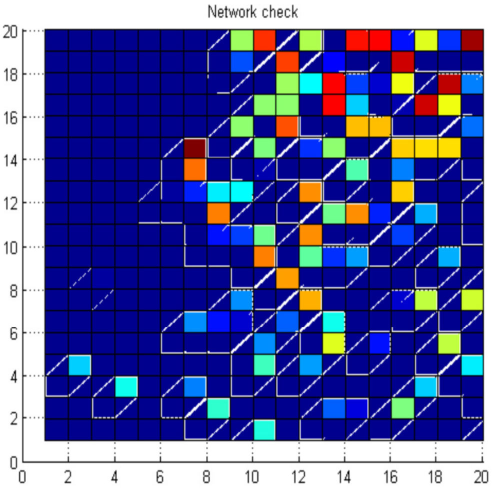

Figure 6. Network response to complex tones with missing fundamental (Blue corresponds to lower pitches while red to higher)

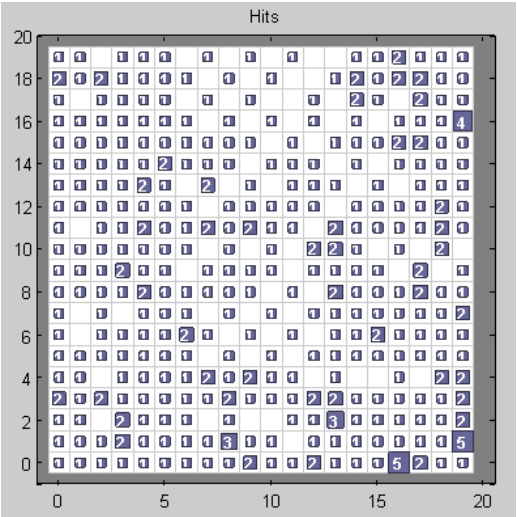

Figure 7. Number of sample hits during training when the harmonics dataset was used

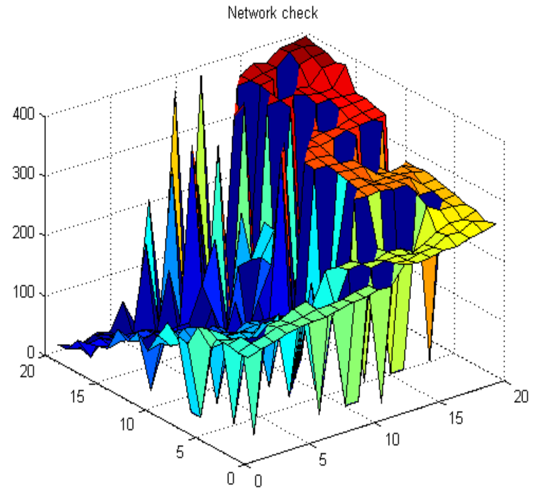

Figure 8. Neuron activation surface when harmonic complex tones were used both for training and testing.

The next set of experiments focuses on network and training parameters. The purpose of this investigation is to determine whether better training can be achieved by altering the training procedure. The first parameter explored was the network size. The same dataset used in the previous experiment was used to train an 8 by 8 and a 30 by 30 SOFM. In the experiment described above, the number of training signals was almost equal to the number of network neurons. Here, this number was decreased and increased significantly. The number of network hits and the pitch distribution surface are presented in figures 9, 10, 11 and 12.

These figures indicate that different pitch signals are activating neurons throughout the network, even if the last has fewer or more neurons compared to the number of signals in the dataset. The choice of fewer neurons introduces a reduction in pitch resolution for the network. On the other hand, employing a large network affects significantly the duration of training.

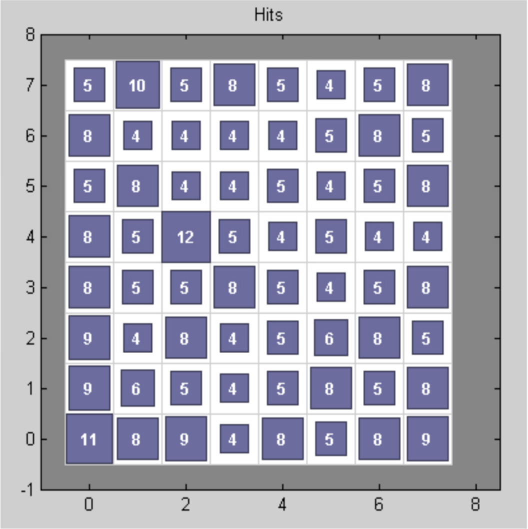

Figure 9. Number of sample hits during training for the 8 by 8 network when trained with a 400 signals dataset

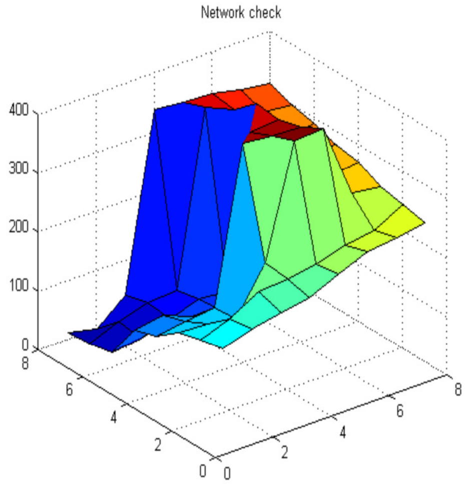

Figure 10. Neuron activation surface for the 8 by 8 network when tested with the 400 signals dataset

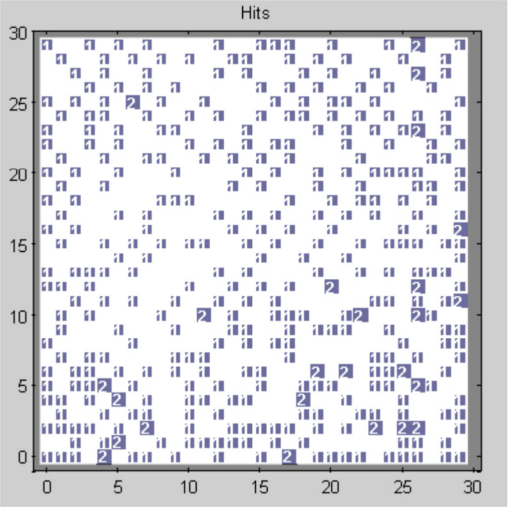

Figure 11. Number of sample hits during training for the 30 by 30 network when trained with a 400 signals dataset

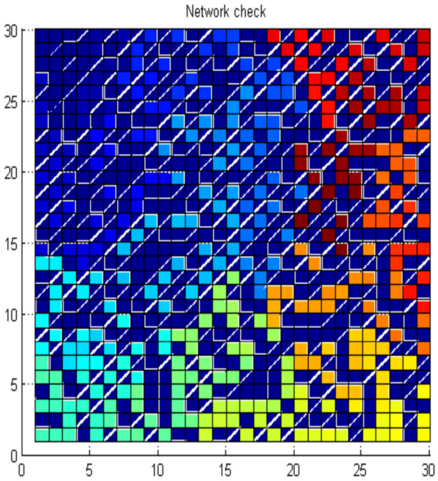

Figure 12. Neuron activation surface for the 30 by 30 network when tested with the 400 signals dataset (Blue corresponds to lower pitches while red to higher)

Additional training parameters that affect training include pitch resolution, i.e. the distance of successive pitches in the training set as well as training epochs, initial neighborhood and ordering phase duration. The cases included in the second stage, were increased successive pitch distance from 10Hz to 50Hz, decreased number of epochs from 200 to 50, reduced initial neighborhood size from 15 to 3 and diminishing the ordering phase duration from 100 to 0. The training results for all four cases are presented in figures 13, 14, 15 and 16. The results reveal that while training epochs and initial neighborhood decrease affects slightly the training process, the remaining parameters have impacted training significantly. Pitch resolution decrease has led to an undesired activation of a single neuron by many signals whose pitch extends to a wide frequency range. Apparently such a network cannot be considered as successfully trained. Furthermore, ordering phase value decrease, results in a reduction of neurons activated.

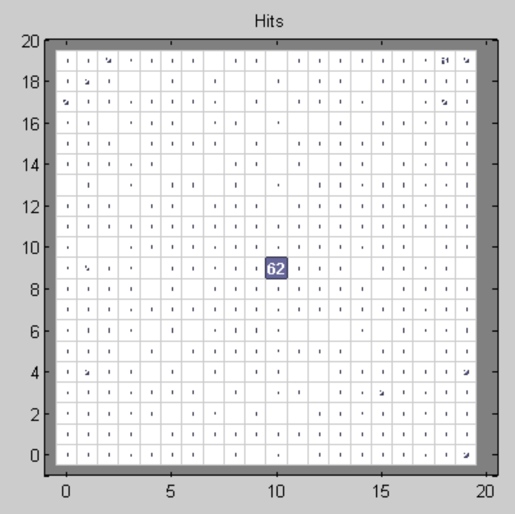

Figure 13. Sample hits during training for the 50Hz pitch distance dataset

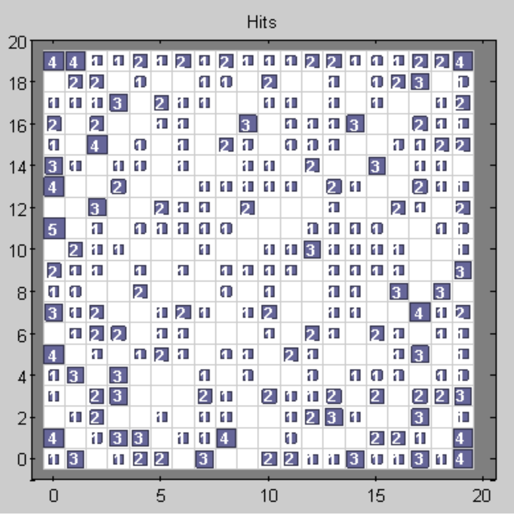

Figure 14. Sample hits during training for the 50 epochs case

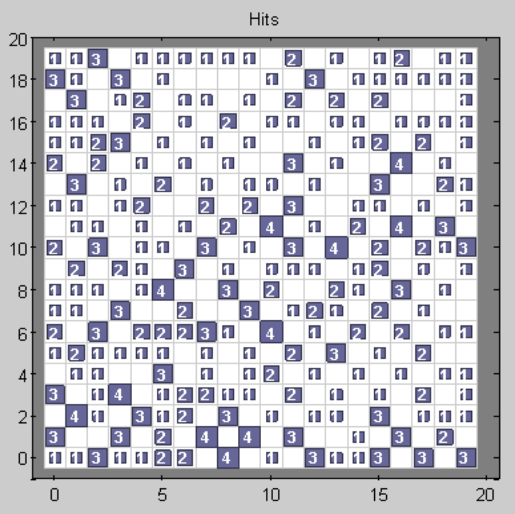

Figure 15. Sample hits during training for the 3 neurons initial neighborhood case

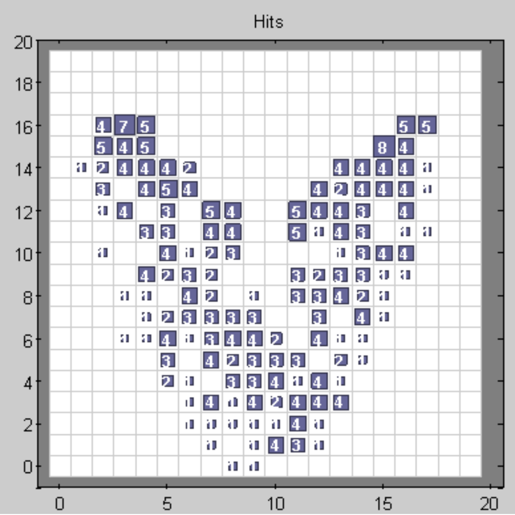

Figure 16. Sample hits during training for the zero epochs ordering phase case

In the third set of experiments, the above mentioned dataset of pure tones spectra was further transformed. The motivation for these experiments was the fact that low pitch resolution of the training dataset prevents network training success. Thus transforming the spectrum amplitude to a logarithmic scale, produces more common, non-zero, coefficients for each pair of successive spectra. This increases the amount of information available to the network, in order to sort spectra according to pitch. Furthermore, Mel scale is widely used in pitch perception. The SOM's ability to be self-organized when Mel transformed spectra are presented is examined here as well.

Figures 17 and 18 display the sample hits and neuron activation surface for the 50Hz resolution dataset. While organization failed for the linear scaled spectrum amplitude, transforming spectra to logarithmic scale, improved training.

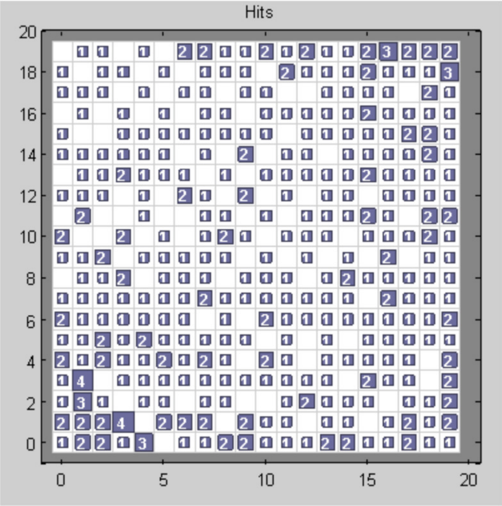

Figure 17. Sample hits during training with logarithmic scaled spectra amplitude, for the case of 50Hz resolution.

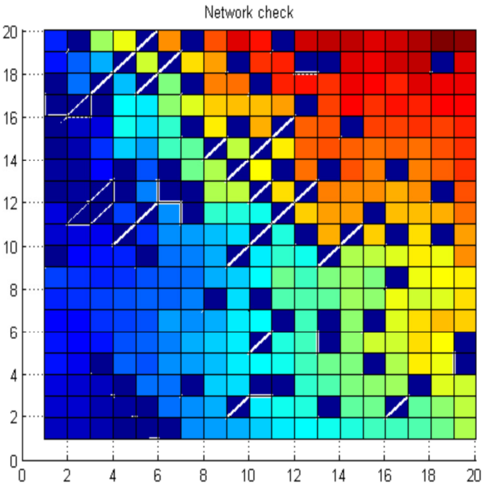

Figure 18. Neuron activation surface for the 50Hz pitch resolution dataset (Blue corresponds to lower pitches while red to higher)

This last figure reveals that training of a self-organizing map can be more efficient with this type of input patterns compared to linear spectra, while pertaining the ordering properties of the input series pitches (Figure 21).

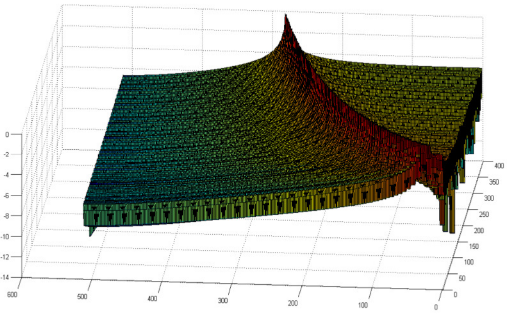

Figure 19. Mel transformed input spectra with logarithmic scaled amplitudes

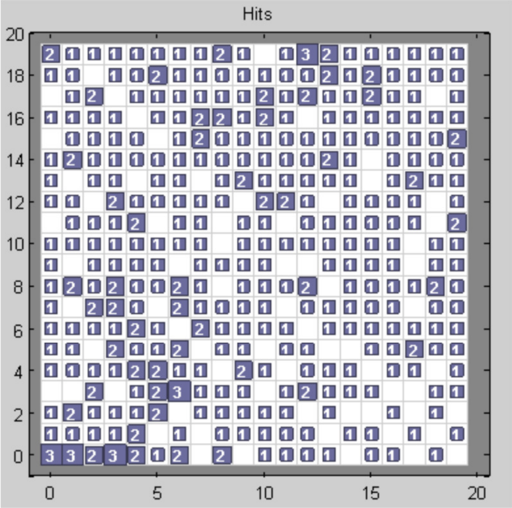

Figure 20. Sample hits during training with Mel transformed spectra dataset

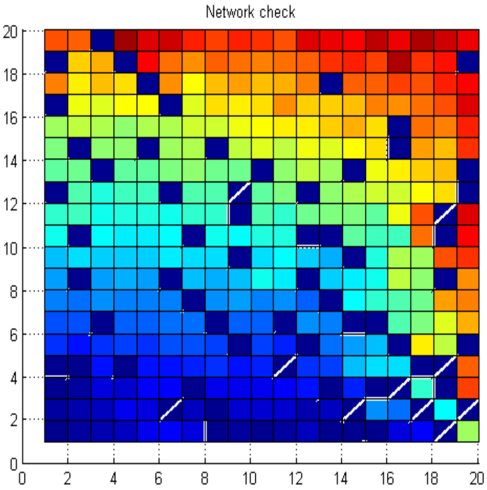

Figure 21. Neuron activation surface for the Mel transformed dataset (Blue corresponds to lower pitches while red to higher)

Figures 19 and 20 present the Mel transformed dataset of pure tones spectra, with logarithmic scaled amplitudes and the sample hits of the SOM network during training.

IV. Discussion

In this work, a connectionist realization for pitch and frequency representation was presented and tested. The model is based on Kohonen's Self Organizing (Feature) Maps. Their ability for self-organization according to frequency or pitch was explored.

The results reveal that SOMs can effectively self-organize and identify pitch, assigning different complex or pure tones to distinct neurons monotonically. There is a specific route through neurons, in every trained network, on which the assigned pitch gradually increases.

Nevertheless, a SOM would exhibit failure in self-organizing in a meaningful and monotonic pattern, when the training dataset demonstrates low pitch resolution. In this case, transforming spectral amplitudes in logarithmic units would alleviate a significant portion of the issue. Additionally, and in accordance with clues from previous works (Bharucha 2009) an organized network may not identify pitch, when the training and test datasets demonstrate different timbral characteristics. This fact indicates that a pre-processing stage, where the input signal may be classified according to specific timbral properties, would be necessary. Thereafter, the appropriate SOM that corresponds to the specific timbral category could be employed to determine pitch. Finally, pitch perception related transformations of spectral frequency scale (such as Mel transformation) not only do they not affect self-organization negatively, but seem to suggest less ambiguous pitch representations.

A point that has to be stressed is that in this work the input datasets to the examined network consist of Fourier spectra of the test signals. However, this is just an initial choice in order to step into the investigation of properties and capabilities of the SOM representations. Actually, a more prudent choice would be to obtain spatiotemporal representations of the test signals much like the type of neural activation patterns answered in the auditory periphery. This could be accomplished by incorporating a precursory analysis stage in the form of a computational auditory model. And indeed, most of the existing models are structured in this way, namely by employing modules simulating the auditory periphery's function (Hawkins 1995). In our work, even an elementary incorporation of logarithmic intensity treatment and a pitch-like scaling of frequency axis in spectral representations showed to have a beneficial effect, as already mentioned. Moreover, this would be mandatory in cases of examining aspects of the application of our modelling approach in impaired hearing, which is one of the future targets of our efforts. Additionally, some more choices, which are inspired from auditory physiology and are already taken in existing approaches, would be advisable. For example, limiting the number of harmonics used for pitch representation is one rational selection stemming from facts related to the spectral resolving capabilities of the auditory periphery.

In the next steps of the investigation, in addition to the above improvements, we shall proceed with a study of the pitch scaling which is provided by the current representation (assigning a unique pitch value to each neuron) and further warping to pitch scaling data from behavioural experiments (e.g. with normal hearing listeners). Accordingly, performance in related pitch tasks (such as difference limens determination) could be attempted and compared against other previously proposed models in terms of predictive power. As already mentioned, the current type of investigation is prospectively sought to be employed in a modelling framework of auditory performance for hearing impairments.

REFERENCES

Bharucha, J.J. (1998*). Neural nets, temporal composites and tonality. In D. Deutsch (Ed.), The Psychology of Music (2d Ed.).* New York: Academic Press.

Bharucha, J.J. (2009). From frequency to pitch, and from pitch class to musical key: Shared principles of learning and perception. *Connection Science*, 21, 177-192.

Cohen, M.A., Grossberg, S., & Wyse, L. (1995). A spectral network model of pitch perception. *The Journal of the Acoustical Society of America,* 98, 862-879.

Hawkins H.l., Mcmullen T.a., Popper A.n., Fay R.r. (1995*). Auditory computation.* 1st edition, Springer.

Houtsma, A.J.M., & Goldstein, J.L. (1972). The central Origin of the pitch of complex tones: Evidence from music interval recognition. *Journal of the Acoustical Society of America*, 51, 520-529.

Kohonen, T. (2001). *Self-organizing maps*. Third, extended edition. Springer.

Licklider, J.C.R. (1954). Periodicity pitch and place pitch. *Journal of the Acoustical Society of America,* 26, 945.

Meddis, R., & Hewitt, M.J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America,* 89, 2866-2882.

Taylor, I., & Greenhough, M. (1994). *Modeling pitch perception with adaptive resonance theory artificial neural networks.* Cambridge: MIT Press.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, 1, 155-182.

Terhardt, E. (1982). Pitch of complex signals according to virtual pitch theory: Tests, Examples, and Predictions. *Journal of the Acoustical Society of America,* 71, 671-678.

Wightman, F.L. (1973). The pattern transfromation model of pitch. *Journal of the Acoustical Society of America*, 54, 407-416.