Music Perception
Winter 1998, Vol. 16, No. 2, 223–241

Timbre Similarity: Convergence of Neural, Behavioral, and Computational Approaches

PETRI TOIVIAINEN
University of Jyväskylä

MARI TERVANIEMI
University of Helsinki

JUKKA LOUHIVUORI
University of Jyväskylä

MARIEKE SAHER, MINNA HUOTILAINEN, &
RISTO NÄÄTÄNEN
University of Helsinki

The present study compared the degree of similarity of timbre representations as observed with brain recordings, behavioral studies, and computer simulations. To this end, the electrical brain activity of subjects was recorded while they were repetitively presented with five sounds differing in timbre. Subjects read simultaneously so that their attention was not focused on the sounds. The brain activity was quantified in terms of a change-specific mismatch negativity component. Thereafter, the subjects were asked to judge the similarity of all pairs along a five-step scale. A computer simulation was made by first training a Kohonen self-organizing map with a large set of instrumental sounds. The map was then tested with the experimental stimuli, and the distance between the most active artificial neurons was measured. The results of these methods were highly similar, suggesting that timbre representations reflected in behavioral measures correspond to neural activity, both as measured directly and as simulated in self-organizing neural network models.

EXPERIMENTAL research on music perception and cognition has relied mainly on behavioral studies. Recent neurophysiological studies have indicated that relevant information about musical processing can also be

Address correspondence to Petri Toiviainen, University of Jyväskylä, Department of Music, P. O. Box 35, FIN-40351 Jyväskylä, Finland. (e-mail: petri.toiviainen@jyu.fi)

obtained by means of brain measurements (Besson & Faïta, 1995; Besson, Faïta, & Requin, 1994; Crummer, Walton, Wayman, Hantz, & Frisina, 1994; Tervaniemi, Ilvonen, Karma, Alho, & Näätänen, 1997; Zatorre, Evans, & Meyer, 1994). Although studies based on these methods may elucidate some cognitive aspects beyond a given musical task, using a single method only may account for just a part of the processes underlying it. Therefore, use of several methods in parallel may help to obtain a more complete picture of the processes under examination. Moreover, each method may be prone to extraneous factors characteristic of that particular method; the effect of these factors may be reduced if several methods are used concurrently. Behavioral studies, for instance, provide a means of measuring subjective experience, but the cognitive processes underlying this experience may not necessarily be unraveled with these experiments. In addition, the results obtained may be influenced by either the instruction provided or by the answering strategies of the subjects. Brain studies, on the other hand, may not measure aspects that correspond to subjective experience. Further, in these studies it may be difficult to isolate the response evoked by the actual stimulus from those evoked by other aspects of the environment. Given these facts, it is obvious that any theory about a cognitive process can be better justified if there are converging results from experiments that use different measuring methods (see, e.g., Garner, Hake, & Eriksen, 1956; Leman, 1997).

Besides behavioral experiments and brain measurements, perceptual and cognitive processes of music have been studied by using computational approaches (see, e.g., Balaban, Ebcioglu, & Laske, 1992; Bharucha, 1987; Miranda, in press; Todd & Loy, 1991). A problem with some of these simulations is that it is not always clear how the models relate to neural mechanisms and subjective experience. Accordingly, for any such model to be adequate, it would be essential to demonstrate that the results obtained from simulations with the model converge with results from both behavioral and brain studies. This kind of convergence is necessary for the model to correspond to the cognitive process involved both at the output level and at the internal levels.

Some computational models of cognition, especially symbol-based ones, have received serious criticism because of their weak relationship to brain functions (Oaksford & Chater, 1991; Valentine, 1997; Wason & Johnson-Laird, 1972). This has led to a growing interest in another kind of computational model, specifically, artificial neural networks. Although these models are inspired by the assumed affinity to the neural mechanism of the brain, their relevance is also debatable. The limitation most often mentioned is that the majority of artificial neural network algorithms do not provide a convincing model of learning. This applies in the first place to the so-called supervised learning networks. In order to provide an adequate model of learning, an artificial neural network should adapt to external stimuli without any direct intervention of the programmer. This holds true for the self-organizing neural networks. In these networks, the role of the programmer is restricted to the planning of the model's architecture. Further, the learning algorithm should, at least on some abstract level, conform to current knowledge about the adaptation mechanisms of the brain. In our view, the Kohonen self-organizing map (SOM) meets these preconditions (Kohonen, 1997). It provides a simple, yet effective, learning algorithm for simulating the formation of topographical feature maps in the neural system.

In the present study, the similarity of responses evoked by sound stimuli differing in timbre was measured by applying three different methods, namely, behavioral tests, recordings of brain electric potentials, and computer simulations. The aim was to examine the mutual similarity of the results obtained by these methods. The present strategy required that these results could be meaningfully compared. In the behavioral test, the perceptual similarity of stimuli was quantified from subjects' numerical assessments of tone similarity. A method was needed for measuring the degree of similarity of the brain responses. For this, we recorded the mismatch negativity (MMN), which is a successful method for determining the degree of similarity of auditory stimuli in neural and perceptual terms (Näätänen & Alho, 1997). The MMN reflects a discrepancy between sound parameters represented by a neural memory trace and a new sound; the degree of similarity between two successive sounds can be quantified as the MMN amplitude and latency. The neural response to the sound stimuli was simulated by means of a model consisting of a computational auditory model and the SOM. After the SOM was trained with a large set of instrumental sounds, its response to the sound stimuli used in the other two experiments was measured. The distances between these responses were then compared with the behavioral measures and the MMN data.

Timbre

Perceptually, a musical sound is often described in terms of four attributes: volume, pitch, perceived duration, and timbre. The physical counterparts of the first three perceptual attributes are intensity, frequency, and physical duration, respectively. Timbre is considered to arise from all the remaining physical attributes and is, thus, a multidimensional attribute of sound. It is associated with the time-varying spectrum of sound.

Timbre is defined by the American Standards Association (1960) as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." This definition is, however, inadequate, because it actually defines what timbre is not rather than what it is. A number of attempts have been made to extract the most salient acoustic attributes affecting the perception of timbre. By examining maps obtained by applying the technique of multidimensional scaling (Kruskal, 1964a, 1964b; Shepard, 1962a, 1962b) to similarity ratings, it has been found that the spectral energy distribution in the steady-state portion of a tone is one of the main contributors to the perception of timbre (Grey, 1977; Iverson & Krumhansl, 1993; Plomp, 1976; Wedin & Goude, 1972; Wessel, 1979). In perceptual terms, this dimension in the timbre space corresponds to the brightness of tones. The second dimension found in most studies is the rise time, that is, the time between the onset and the instant of maximal amplitude.

Mismatch Negativity and Timbre

Auditory perception is based on a chainlike activation of several neural populations, starting from the cochlea, continuing into the primary auditory cortex (located in the temporal lobe), and extending thereafter to include neuronal networks from frontal to parietal and even to occipital areas (for a review, see Buser & Imbert, 1992). The perception of timbre results from the intact functioning of the right auditory cortex areas. This was the conclusion of a study (Samson & Zatorre, 1994) in which the effect of right versus left temporal lobe lesions on processing spectrally versus temporally complex information was determined.

Neural mechanisms activated by timbre changes, independent of the focus of the subject's attention, can be investigated by recording MMN (Näätänen, Gaillard, & Mäntysalo, 1978). MMN is a component in the auditory event-related potential (ERP), which mainly originates in the auditory cortex (for a review, see Alho, 1995; see also Giard, Perrin, Pernier, & Bouchet, 1990; Hari et al., 1984). The MMN is typically elicited by the rare deviants in a so-called oddball paradigm, in which a frequent standard stimulus is occasionally replaced by a deviant stimulus differing from standard tones in, for instance, intensity or duration. The MMN reflects change detection in a process in which the memory traces representing the constant standard stimulus and the incoming deviant stimulus are compared (Näätänen, 1992). In other words, the MMN reflects a memory-based process rather than a reaction to the presence of a stimulus; it is never elicited by the first stimulus in a sequence, nor is it elicited when the deviant stimulus is presented alone without intervening standards (Näätänen, 1990, 1992). MMN occurs independently of attention; it is elicited even when the subject's attention is directed to something completely irrelevant to the task, like reading or playing a computer game (e.g., Alho, Woods, Algazi, & Näätänen, 1992).

Recently Tervaniemi, Winkler, and Näätänen (1997) performed an MMN study aimed at determining whether changes in the spectral component of sound timbre, as reflected by MMN elicitation, would be preattentively encoded. This was established by intermixing nine spectrally rich missing-fundamental sounds with a relatively rare sinusoidal tone, all having a pitch of 300 Hz. Although the exact timbre of the nine standard sounds was different because their frequency compositions were varied, an MMN was elicited by the rare sinusoidal tones. This suggests that the preattentive processes underlying MMN grouped the spectrally rich sounds together, contrasting them with the qualitatively different deviant timbre. This result confirms that the MMN paradigm is appropriate for studying timbre perception.

Kohonen Self-Organizing Map and Timbre

We receive a constant stream of information from the environment through our senses. This information is highly dimensional and complex. In order to deal with this complexity, the central nervous system tends to reduce the dimensionality of the incoming information. Various kinds of ordered feature maps can be identified at least in the somatosensory, auditory, and visual modalities. The feature maps are compressed representations of the observed signals, containing information about the most relevant features and their interrelationships. It is commonly believed that these maps originate from self-organization of neural connectivity structure. There exists a well-developed computational theory of self-organization (Kohonen, 1997), which is based on the assumption that lateral inhibition and changes in neural connectivity are responsible for self-organization in the central nervous system. This theory of self-organization has been formalized into a simple, yet effective, numerical algorithm. Given a set of input vectors in a multidimensional vector space, the Kohonen SOM identifies the most salient features of the input set, that is, the dimensions with greatest variance, and maps those features onto a two-dimensional grid of artificial neurons,[1] while retaining the topological relationships between the input vectors. A mathematical description of the SOM and the self-organization algorithm is provided in Appendix 1. The SOM has been used successfully for the classification of timbre by a number of researchers (Cosi, De Poli & Lauzzana, 1994; De Poli, Prandoni, & Tonella, 1993; Feiten & Günzel, 1994; Toiviainen, 1996, 1997; Toiviainen, Kaipainen, & Louhivuori, 1995).

[1] Subsequently, the word "artificial" is omitted in this context whenever there is no danger of confusing artificial neurons with biological ones.

Any timbre classification model has to extract the most significant parameters of the incoming sound signal by means of a preprocessing stage. In timbre and speech research, this has traditionally been based on methods like short-time Fourier transform, cepstrum, or linear predictive coding (Rabiner & Shafer, 1978). All these methods rely on the analysis of a series of successive frames, and a quasi-periodic model of the signal, that is, on the assumption that the properties of the signal do not change significantly within an analysis frame. This assumption may cause subtle dynamic phenomena to be discarded. More recent multiresolution analysis approaches, such as the wavelet transform (Kronland-Martinet & Grossmann, 1991), seem to alleviate this shortcoming, but they are still based on a mathematical transformation of the signal without taking into account the principles of human auditory processing.

Recently, knowledge about the neuromechanical properties of the human auditory periphery has increased significantly. On the basis of this knowledge, a number of computational models of the auditory periphery have been developed (Cohen, 1989; Ghitza, 1986; Meddis, 1986; Van Immerseel & Martens, 1992). A great deal of evidence in the literature supports the use of these auditory modeling techniques in systems aimed at classifying and recognizing sounds. For instance, Van Immerseel and Martens (1992) found that their model for phonetic classification and segmentation of speech utterances clearly performed better when the preprocessing of sound was based on the properties of the auditory periphery, compared with traditional preprocessing strategies. In a timbre classification experiment with the SOM (Toiviainen et al., 1995), it was found that using an auditory model for preprocessing the sound stimuli led to significantly better performance; the correlation between the responses on the SOM and the respective similarity ratings by human subjects was significantly higher than when short-time Fourier transform was used. Here we use the same auditory model in the preprocessing stage.

Methods

STIMULI AND SUBJECTS

The stimuli were produced by additive synthesis with a Power Macintosh 7600 computer. All stimuli had a fundamental frequency of 440 Hz and consisted of 16 partials. The partials formed a harmonic overtone series, meaning that each consecutive partial had a frequency 440 Hz higher than that of the preceding partial (880 Hz, 1320 Hz, etc.). All stimuli had a duration of 500 ms and attack and decay times of 50 ms.

The distribution of spectral energy, perceived as brightness, was the only parameter varied in the set of stimuli. Brightness was defined as the ratio of maximum amplitudes of successive partials. For instance, if brightness was set to 0.5, then the second partial of the stimulus had a maximal amplitude of 0.5 of that of the first partial, the third had a maximum of 0.5 times of that of the second, and so forth. Five stimuli, having brightness values of 0.1, 0.3, 0.5, 0.7, and 0.9, were used in the experiments.[2] Subsequently, the stimuli will be referred to as very dark, dark, medium dark, medium bright, and bright, respectively. The very dark stimulus resembled a horn sound, whereas the bright stimulus resembled a bagpipe. The stimuli were equalized for perceived loudness. Figure 1 displays graphically the amplitudes of the partials of the stimuli.

A commonly used quantity for expressing the brightness of a sound in physical terms is the spectral centroid. This is the weighted average of the spectral energy across frequency (see Appendix 2). The higher the spectral centroid of a tone, the brighter it is perceived. Table 1 presents the spectral centroids of each stimulus in hertz and in critical band rate (Bark scale, see Appendix 2).

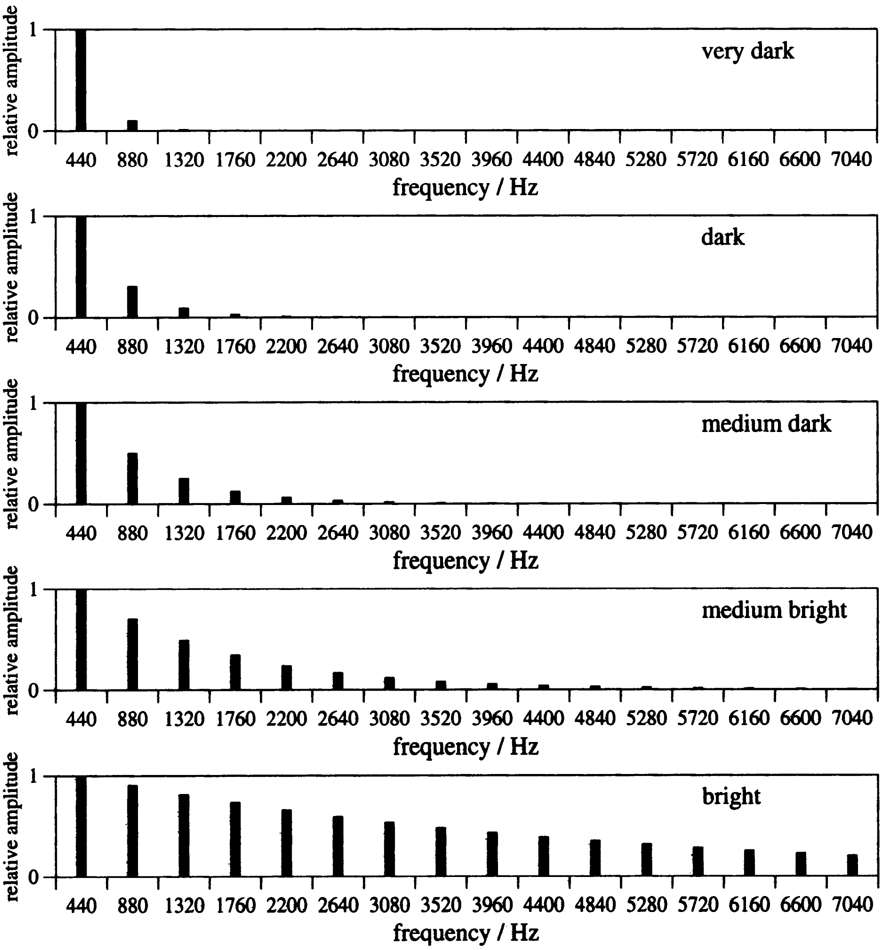

Fig. 1. Maximum amplitudes of partials of each of the five tones.

[2] Sound files containing the stimuli used in the experiments are available on the World-Wide Web via the URL "http://www.jyu.fi/~ptoiviai/convergence.html".

Table 1
Spectral Centroids of the Stimuli

	Standard (very dark)	Deviant 1 (dark)	Deviant 2 (medium dark)	Deviant 3 (medium bright)	Deviant 4 (bright)
$\langle f \rangle$/Hz	489	629	880	1443	2799
$\langle f \rangle$/Bark	4.82	5.98	7.78	10.84	15.24

Nine subjects (mean age 26, range 20–31 years, four females) were employed in the present MMN and similarity rating experiments. All subjects were right-handed and reported no history of neurological and/or auditory problems. After being informed about the test procedure, all subjects gave informed consent. Each subject participated in the MMN experiment before performing the similarity rating test.

PROCEDURE

Similarity Rating

The subjects participated in a short similarity rating task for assessment of the individual stimulus discrimination abilities. They were presented (binaurally, through headphones) with tone pairs, each pair preceded by a warning sound. Each possible pair of the five stimuli, including the pairs consisting of identical stimuli, was presented three times. The total number of tone pairs presented was thus 75. The pairs were presented in random order. After each presentation of a tone pair, subjects indicated on a response form to what extent the two tones differed from each other, with answers falling in five categories (identical, very similar, quite similar, quite different, and very different). A test trial of five tone pairs was presented to acquaint the subjects with the procedure before the experiment. The duration of the similarity rating experiment was approximately 10 min.

MMN Experiment

In the MMN experiment, the standard stimulus (p = .80) was the very dark tone, and the four remaining tones served as deviants. Subjects were presented with six blocks of 600 stimuli each, in which each deviant had a probability of .05 (total deviance probability was therefore .20). The stimuli were presented via headphones at a loudness of 70 dB SPL and were separated by a silent interstimulus interval of 400 ms. The sounds were transferred to PC-based NeuroStim software, which was used while the sounds were presented to the subjects during the experiment.

Subjects were comfortably seated in a recliner in an electrically shielded and sound-attenuated room. Through headphones, subjects were binaurally presented with each of the six blocks of stimuli. The order of the blocks was randomized among subjects. All subjects were instructed to concentrate on reading a book of their own choice and to pay no attention to the stimuli. If needed, they were allowed breaks between the blocks in order to maintain a high level of attention to reading. The duration of the MMN experiment was approximately 75 minutes, including the breaks between the blocks.

The electrical signal was measured from the scalp with silver–silver chloride electrodes at 10 sites: Fpz, Fz, Cz, Pz, L1, L2, R1, R2, and left and right mastoids Lm and Rm. The lateral electrodes were placed on the axes drawn from Fz to each of the mastoids, at one third (L1, R1) and two thirds (L2, R2) of the distance between Fz and the mastoids (see Figure 3 for the schematically drawn positions of the electrodes). Eye movements were monitored with Fpz and HEOG electrodes (horizontal electro-oculogram, attached to the right outer canthus). The reference electrode was attached to the nose. The recording band pass was –0.1 to 100 Hz, and the sampling rate was 500 Hz. The data were amplified with SynAmps EEG amplifiers and stored on a computer disk for off-line averaging.

SOM Simulation

A timbre map was constructed by training a SOM on a set of real instrument sounds. The training set consisted of all sounds of the first three volumes of McGill University Master Samples (Opolko & Wapnick, 1989), making a total of 45 sounds. All these sounds had a fundamental frequency of 440 Hz. To obtain a vector representation of the tones, the stimuli were preprocessed using the peripheral part of an auditory model by Van Immerseel and Martens (1992), modified by Leman (1995) for musical purposes. The output of the auditory model is a 20-component vector, updated every 0.4 ms; each component represents the probability of neural firing in the respective auditory channel since the last update of the output vector. This output was low-pass filtered to smooth amplitude modulations and sampled 50 times during the first 500 ms of each tone. Each tone was thus represented by a vector of 1000 components (for details, see Toiviainen, 1996). The set of 45 vectors thus obtained was then used to train an SOM of 10 x 10 neurons. The training consisted of 100,000 cycles, during which the neighborhood radius was linearly decreased from 5 to 0 and the learning rate from 0.5 to 0.

After the SOM was trained, its response to the five stimuli was measured one at a time. For each stimulus, the focus of response was determined (for the definition of the response focus in the SOM, see Appendix 1). The distance between the responses evoked by any two stimuli was defined to be the distance between the respective foci of response (see Figure 2).

DATA ANALYSIS

Similarity Ratings

The similarity ratings for each subject were stored as a 5 x 5 matrix of data, where each matrix component was obtained by averaging the ratings for the three presentations of the respective tone pair. Because the order in which tones of a pair are presented in timbre similarity rating experiments has been found to have only little effect (Grey, 1977; Iverson & Krumhansl, 1993), the original response matrices were transformed into triangular ma-

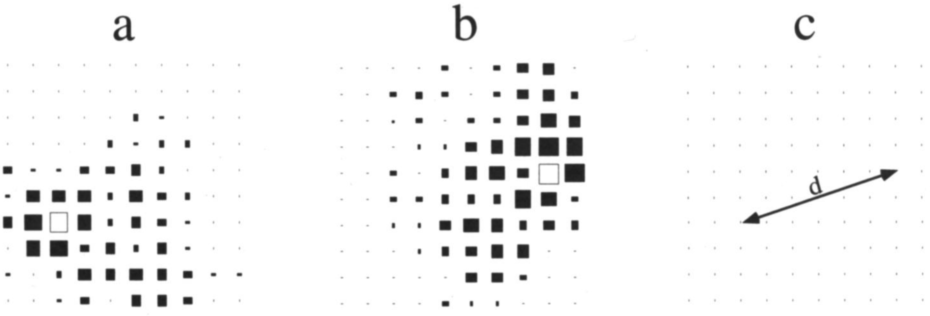

Fig. 2. Definition of the distance of two responses on the self-organizing map (SOM). Each black square represents a neuron; a large square denotes high activation of the respective neuron (i.e., small difference between the input vector and the synaptic vector of the neuron). White squares represent foci of response. (a) response to (a fictitious) stimulus a; (b) response to stimulus b; (c) d = distance between the responses to the two stimuli.

trices by averaging the ratings of pairs of stimuli presented in different orders. The ratings of all pairs consisting of identical stimuli, that is, the diagonal elements of the matrices, were ignored. Therefore, each matrix consisted finally of 10 elements.

Consistency of ratings between the subjects was examined by calculating intersubject correlations. A one-way repeated measures analysis of variance (ANOVA) and a subsequent post-hoc Newman-Keuls test were used to verify the statistical significance of the similarity ratings. The half-matrices were averaged across subjects to obtain mean similarity ratings of each of the tone pairs. In this analysis, only the ratings for the pairs consisting of the standard (very dark) and each of the four deviants were considered.

MMN Experiment

The data were averaged individually for every subject separately for each type of deviant as well as for the standard. Epochs with a voltage change exceeding 100 μV were automatically rejected. The analysis period was 1000 ms including a 100-ms prestimulus baseline. Only those standards that were not directly preceded or followed by a deviant were analyzed. Then, the data were re-referenced by subtracting the average value between the two mastoids and band-pass filtered at 1–30 Hz.

Next, the standard-tone grand-average ERP was subtracted from that of each of the deviants individually for each subject; the resulting waves are called ***subtraction waves***. Following the common procedure, the MMN peak latency and amplitude were determined from the subtraction waves, and the MMN was defined as the most negative deflection in the subtraction wave between 100 and 300 ms after stimulus onset at the Fz electrode. The amplitudes were calculated as the mean during the 20-ms time window centered on the most negative peak, measured separately for each subject and for each deviant.

A one-tailed *t* test was used to determine whether MMN amplitudes differed significantly from zero. A one-way repeated measures ANOVA and a subsequent post-hoc Newman-Keuls test were used to verify the statistical significance of the amplitude and latency differences between the four deviant-timbre tones.

SOM Simulation

The distance between the response on the SOM evoked by the standard (very dark) stimulus and that evoked by each of the four deviants was determined.

Results

SIMILARITY RATINGS

The average of the intersubject correlations was 0.90 ($df = 8$), each being significant at the $p < .05$ level. The similarity ratings were thus consistent between subjects.

When the similarity ratings were compared with the respective brain recording and computer simulation data, only the ratings for the pairs consisting of the standard (very dark) stimulus and one of the deviants were compared with the respective data obtained from the brain recordings and the computer simulation. The mean similarity ratings for these pairs are shown in the first row of Table 2. Each mean rating is

TABLE 2

Mean Mismatch Negativity (MMN) Latencies and Amplitudes, Mean Similarity Ratings, and Distances of the Self-Organizing Map (SOM) Response

	Deviant 1 **(dark)**	**Deviant 2** **(medium dark)**	**Deviant 3** **(medium bright)**	**Deviant 4** **(bright)**
Mean similarity rating	1.96 (0.43)	3.02 (0.68)	4.20 (0.56)	4.76 (0.43)
SOM response distance	1.41	3.16	5.00	6.71
Mean MMN amplitude (µV)	–2.04 (0.9)	–3.20 (1.3)	–4.76 (1.0)	–6.39 (1.6)
Mean MMN latency (ms)	172 (31)	115 (10)	112 (9)	116 (10)

Note—Standard deviations are given in parentheses.

obtained by averaging across all ratings of all subjects for the respective pair. As can be seen, the ratings are in line with the brightness values of the deviants. In other words, the brighter is the deviant, the higher is the respective rating, $F(3,24) = 277.12$; $p < .0001$. The ratings differed between all deviant tones ($p < .001$, according to a Newman-Keuls post-hoc test).

MMN EXPERIMENT

The mean MMN amplitudes and latencies elicited by the deviants are displayed in the last two rows of Table 2. The MMN was elicited by all four timbre deviants. The values were significantly different from zero for all four deviants [t values ranged from $t(8) = 6.74$ to 14.4, all $p < .0001$]. Figure 3 shows the MMN amplitudes and latencies as a function of stimulus deviation. The more the deviant differed from the standard, the larger was the MMN amplitude, $F(3,24) = 23.83$; $p < .00001$. The MMN amplitude differed between all deviant-timbre tones (p values between .04 and .0001, according to a Newman-Keuls post-hoc test). Also the MMN latency differed between deviant-timbre tones, $F(3,24) = 28.45$; $p < .00001$. However, only the latency of the dark deviant was significantly longer than that of the other deviants ($p < .0001$, by a Newman-Keuls post-hoc test).

SOM SIMULATION

Figure 4 displays the foci of response on the SOM for the five stimuli. The distances between the SOM responses to the standard stimulus and each of the deviants are presented in the second row of Table 2. Again, these distances correspond to the brightness values of the deviants.

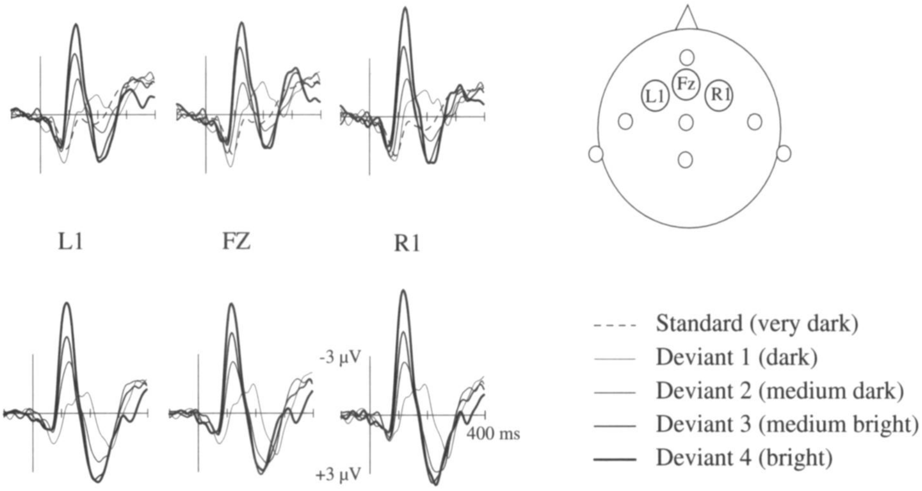

Fig. 3. Upper panel: Grand-average event-related potentials (ERPs) of nine subjects to standard timbre (the dashed line) and to four deviant-timbre tones (the line thickness denotes the magnitude of deviance) recorded over the frontocentral scalp areas. Bottom panel: difference curves in which ERPs to standard timbre have been subtracted from ERPs to deviant-timbre tones (the line thickness denotes the magnitude of deviance). The schematic illustration in the upper right corner denotes the electrode montage used.

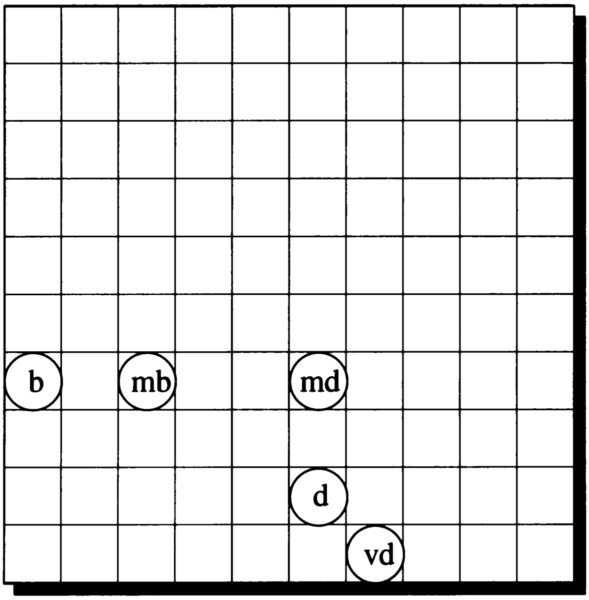

Fig. 4. The foci of response on the self-organizing map (SOM) for the stimuli used (vd = very dark, d = dark, md = medium dark, mb = medium bright, b = bright).

CORRELATIONS

Figure 5 displays the X-Y scatter plots of the physical attributes of the stimuli and the data obtained from MMN experiment, similarity ratings, and SOM simulation.

Table 3 presents the matrix of correlations between the physical attributes of the stimuli (difference of spectral centroids on the Bark scale) and the data obtained from the MMN measurements, similarity ratings, and SOM simulation.

As shown in Table 3, the correlations between the different measures are all significant with the exception of the MMN latency, which failed to cor-

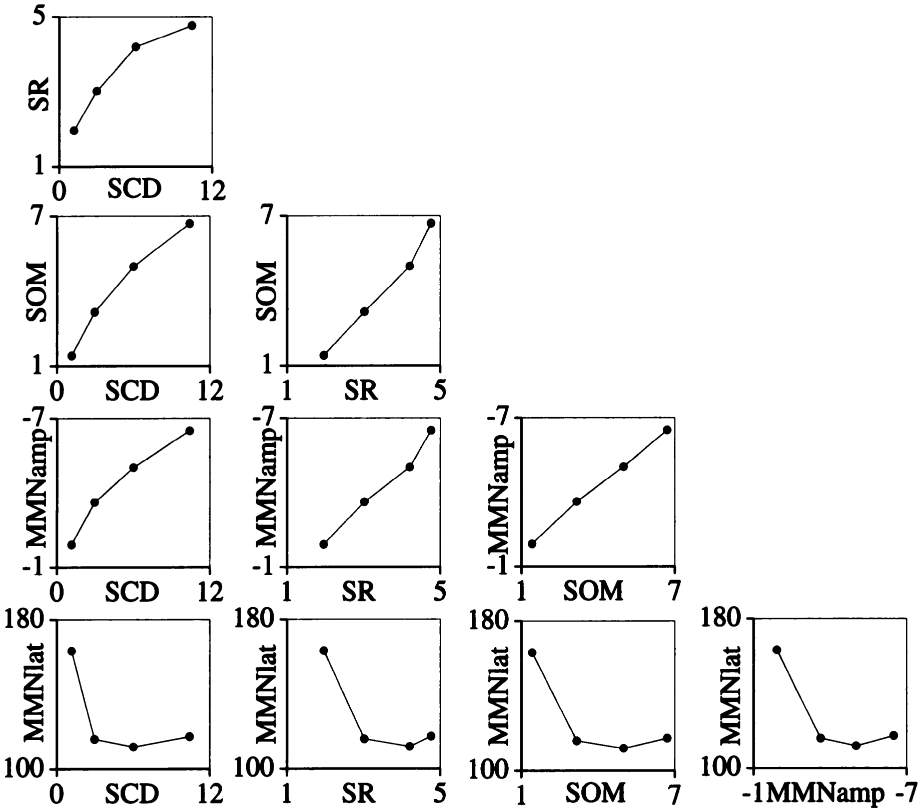

Fig. 5. X-Y scatter plots of the physical attributes of the stimuli and the data obtained from mismatch negativity (MMN) measurements, similarity ratings, and the self-organizing map (SOM) simulation. The plots are in the same order as the respective correlations of Table 3. SCD = difference of spectral centroids (on the Bark scale); SR = mean similarity ratings; MMNamp = mean MMN amplitude; MMNlat = mean MMN latency; SOM = distance of responses on the self-organizing map.

TABLE 3
Correlations (*df* = 2) Between the Physical Attributes of the Stimuli and the Data Obtained from Similarity Ratings, Mismatch Negativity (MMN) Measurements, and Self-Organizing Map (SOM) Simulations

	Difference of		SOM	
	Spectral Centroids	Similarity	Response	MMN
	(Barks)	Rating	Distance	Amplitude
Similarity rating	0.951*			
SOM response distance	0.982**	0.992***		
MMN amplitude	−0.975 *	−0.991***	−0.999***	
MMN latency	−0.628 (ns)	−0.808 (ns)	−0.758 (ns)	0.783 (ns)

Note—*$p < .05$, **$p < .01$, ***$p < .005$, ns = not significant.

relate significantly with any other measure. Similarity ratings, MMN amplitudes, and SOM response distances converge strongly, all the three respective correlations being significant at the $p < .005$ level.

Discussion

The present findings demonstrate that the obtained behavioral, neural, and computational measures of timbre similarity converge strongly. The more dissimilar the subjects rated a given pair of tones, the more dissimilar were the respective neural responses, both as measured directly and as simulated with the SOM. In the data set obtained, the mutual dependence between the similarity ratings, the MMN amplitudes, and the response distances on the SOM was almost linear; only the MMN latency failed to correlate significantly with the other measures.

Some studies on music perception and cognition have demonstrated the similarity between behavioral and brain measurement data (Crummer et al., 1994; Janata, 1995; Lang et al., 1990; Tervaniemi, Ilvonen, et al., 1997). Further, a few studies have established the similarity between behavioral and computational measures (e.g., Leman, 1995, 1997; Leman & Carreras, 1997; Parncutt, 1994; Toiviainen, 1996, 1997; Toiviainen et al., 1995). The present study is, however, the first one in the domain of music that has demonstrated the mutual convergence of all three of these measures within a single experimental design.

The high correlation between the results of the behavioral test and the brain measurements implies that automatic (unconscious) processes of the brain measured with MMN have a close connection to the behavioral (conscious) level. Previous support for this hypothesis has been obtained from studies on discrimination tasks (Lang et al., 1990; Tervaniemi, Ilvonen, et al., 1997; Tiitinen, May, Reinikainen, & Näätänen, 1994). The task of similarity rating is, however, more demanding than that of discrimination, because the subjects have to assess the stimuli along a scale instead of just judging whether a stimulus differs from another. Therefore, in comparison with the discrimination task studies, the present findings can be seen as providing even stronger evidence for the aforementioned hypothesis.

The high correlation between the results of the behavioral test and the computer simulation provides support for the view that artificial neural networks are appropriate for studying some aspects of human behavior. By means of properly designed artificial neural network models, one could, for instance, simulate a given behavioral task. The results thus obtained could provide new working hypotheses that could then be tested by actual behavioral experiments.

The high correlation between the results of the brain measurements and the computer simulation suggests that the SOM, after being trained, responds to sound stimuli in a way similar to how the brain responds. Moreover, this finding supports the view that the SOM provides a convincing model of how the response structure of the brain arises through self-organization and exposure to environmental stimuli. Therefore, SOM-based computer models can be useful in the design of brain measurements; simulations carried out with such models may help to formulate new hypotheses for these studies.

The stimuli used in the present experiment were fairly simple compared with those found in a natural sound environment. One should, however, bear in mind that ERP (thus also MMN) research has traditionally used pure sinusoidal tones to ensure sufficient control over stimulus material and that only recently have temporally and/or spectrally complex sounds been used. So the present stimulation should be seen as a successful compromise between the tradition of MMN research and the natural sound environment. In the future, the present experiment could be repeated with more complicated, possibly natural, sound stimuli.

In summary, the study reported here provides the first demonstration of the convergence of behavioral, neural, and computational measures of a musical task. The present findings strongly support the view that relevant information about perceptual and cognitive processes of music can be obtained by concurrently using these three research paradigms.[3]

References

Alho, K. (1995). Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. *Ear and Hearing, 16,* 38–50.

[3] We are grateful to Carol Krumhansl for valuable comments and suggestions. This work was supported by the Academy of Finland.

Alho, K., Woods, D. L., Algazi, A., & Näätänen, R. (1992). Intermodal selective attention. II. Effects of attentional load on processing of auditory and visual stimuli in central space. *Electroencephalography and Clinical Neurophysiology, 82,* 356–368.

American Standards Association (1960). *Acoustical terminology.* New York: American Standards Association.

Balaban, M., Ebcioglu, K., & Laske, O. (1992). *Understanding music with AI.* Cambridge, MA: MIT Press.

Besson, M., & Faïta, F. (1995). An event-related potential (ERP) study of musical expectancy: Comparison between musicians and non-musicians. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1278–1296.

Besson, M., Faïta, F., & Requin, J. (1994). Brain waves associated with musical incongruities differ for musicians and non-musicians. *Neuroscience Letters, 168,* 101–105.

Bharucha, J. J. (1987). Music cognition and perceptual facilitation: a connectionist framework. *Music Perception, 5,* 1–30.

Buser, P., & Imbert, M. (1992). *Audition: A Bradford book.* Cambridge, MA: MIT Press.

Cohen, J. (1989). Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America, 85,* 2623–2629.

Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research, 23,* 71–98.

Crummer, G. C., Walton, J. P., Wayman, J. W., Hantz, E. C., & Frisina, R. D. (1994). Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *Journal of the Acoustical Society of America, 95,* 2720–2727.

De Poli, G., Prandoni, P., & Tonella, P. (1993). Timbre clustering by self-organizing neural networks. In L. Finarelli & F. Regazzi (Eds.), *Proceedings of X Colloquium on Musical Informatics.* Milan: University of Milan.

Feiten, B., & Günzel, S. (1994). Automatic Indexing of a sound database using self-organizing neural nets. *Computer Music Journal, 18(3),* 53–65.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review, 63,* 149–159.

Ghitza, O. (1986). Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech & Language, 1,* 109–130.

Giard, M.-H., Perrin, F., Pernier, J., & Bouchet, P. (1990). Brain generators implicated in the processing of auditory stimulus deviance: A topographic event-related potential study. *Psychophysiology, 27,* 627–640.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America, 61,* 1270–1277.

Hari, R., Hämäläinen, M., Ilmoniemi, R., Kaukoranta, E., Reinikainen, K., Salminen, J., Alho, K., Näätänen, R., & Sams, M. (1984). Responses of the primary auditory cortex to pitch changes in a sequence of tone pips: Neuromagnetic recordings in man. *Neuroscience Letters, 50,* 127–132.

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America, 94,* 2595–2603.

Janata P. (1995). ERP measures assay the degree of expectancy violation of harmonic contexts in music. *Journal of Cognitive Neuroscience, 7,* 153–164.

Kohonen, T. (1997). *The self-organizing map.* 2nd ed. Berlin, Heidelberg: Springer-Verlag.

Kronland-Martinet, R., & Grossman, A. (1991). Application of time-frequency and time-scale methods (wavelet transform) to the analysis synthesis and transformation of natural sounds. In G. De Poli, A. Piccialli, & C. Roads (Eds.), *Representations of musical signals.* Cambridge: MIT Press.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrica, 29,* 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrica, 29,* 115–129.

Lang, A. H., Nyrke, T., Ek, M., Aaltonen, O., Raimo, I., & Näätänen, R. (1990). Pitch discrimination performance and auditive event-related potentials. In C. H. M. Brunia,

A. W. K. Gaillard, & A. Kok (Eds.), *Psychophysiological Brain Research* (Vol. 1, pp. 294–298). Tilburg: Tilburg University Press.

Leman, M. (1995). *Music and schema theory: Cognitive foundations of systematic musicology.* Berlin: Springer-Verlag.

Leman, M. (1997). The convergence paradigm, ecological modelling, and context-dependent pitch perception. *Journal of New Music Research, 26,* 133–153.

Leman, M., & Carreras, F. (1997). Schema and Gestalt: Testing the hypothesis of psychoneural isomorphism by computer simulation. In M. Leman (Ed.), *Music, Gestalt, and computing: Studies in cognitive and systematic musicology* (pp. 144–168). Berlin: Springer-Verlag.

Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America, 79,* 702–711.

Miranda, E. R. (in press). *Readings in music and artificial intelligence.* Amsterdam: Gordon and Breach.

Näätänen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences, 13,* 201–288.

Näätänen, R. (1992). *Attention and brain function.* Hillsdale, NJ: Erlbaum.

Näätänen, R. & Alho, K. (1997).Mismatch negativity: The measure for central sound representation accuracy. *Audiology & Neuro-otology, 2,* 341–353.

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective attention effect on evoked potential reinterpreted. *Acta Psychologica, 42,* 313–329.

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language, 6,* 1–38.

Opolko, F., & Wapnick, J. (1989). *McGill University Master Samples User's Manual.* Montreal: McGill University, Faculty of Music.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception, 11,* 409–464.

Plomp, R. (1976). *Aspects of tone sensation.* London: Academic Press.

Rabiner, L. R., & Shafer, R. W. (1978). *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice-Hall.

Samson, S., & Zatorre, R. J. (1994). Contribution of the right temporal lobe to musical timbre discrimination. *Neuropsychologica, 32,* 231–240.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika, 27,* 219–246.

Tervaniemi, M., Ilvonen, T., Karma, K., Alho, K., & Näätänen, R. (1997). The musical brain: Brain waves reveal the neurophysiological basis of musicality in human subjects. *Neuroscience Letters, 226,* 1–4.

Tervaniemi, M., Winkler, I., & Näätänen, R. (1997). Pre-attentive categorization of sounds by timbre as revealed by event-related potentials. *NeuroReport, 8,* 2571–2574.

Tiitinen, H., May, P., Reinikainen, K., & Näätänen, R. (1994).Attentive novelty detection in humans is governed by pre-attentive sensory memory, *Nature, 372,* 90–92.

Todd, P. M., & Loy, D. G. (1991). *Music and connectionism.* Cambridge, MA: MIT Press.

Toiviainen, P. (1996). Optimizing auditory images and distance metrics for self-organizing timbre maps. *Journal of New Music Research, 25,* 1–30.

Toiviainen, P. (1997). Optimizing self-organizing timbre maps: two approaches. In M. Leman (Ed.), *Music, Gestalt, and computing: studies in cognitive and systematic musicology* (pp. 337-350). Berlin: Springer-Verlag.

Toiviainen, P., Kaipainen, M., & Louhivuori, J. (1995). Musical timbre: Similarity ratings correlate with computational feature space distances. *Journal of New Music Research, 24,* 282–298.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88,* 97–100.

Valentine, E. (1997). Deconstructing cognition: towards a framework for exploring non-conceptualised experience. In P. Pylkkänen, P. Pylkkö, & A. Hautamäki (Eds.), *Brain, mind and physics* (pp. 3–12). Amsterdam: IOS Press.

Van Immerseel, L. M., & Martens, J.-P. (1992). Pitch and voiced/unvoiced determination with an auditory model. *Journal of the Acoustical Society of America, 91,* 3511–3526.

Wason, P. C., & Johnson-Laird, P. N. (1972). *The psychology of reasoning: Structure and content.* London: Batsford.

Wedin, L., & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology, 13,* 228–240.

Wessel, D. (1979). Timbre space as a musical control structure. *Computer Music Journal, 3,* 45–52.

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *The Journal of Neuroscience, 14,* 1908–1919.

Appendix 1
Description of the SOM

The SOM has n input neurons, each having a specified activation level a_i. The input to the network is, thus, an n-dimensional vector $\mathbf{a} = (a_1, a_2, ..., a_n)$. The SOM also has m output neurons receiving activation from the input neurons. The output neurons usually form a two-dimensional planar array. Each input neuron is connected to each output neuron. A weight w_{ij} is associated to the connection from input neuron i to output neuron j. The connections to output neuron j can thus be represented by an n-dimensional vector $\mathbf{w}_j = (w_{1j}, w_{2j}, ..., w_{nj})$.

There are several variants of the self-organization algorithm; the one used in this study is as follows:

1. The weights are initially given random values, the radius of the topological neighborhood ρ and the learning rate η are chosen.
2. An input vector a is chosen randomly from the set of all possible input vectors and is presented to the network.
3. The Euclidean distance d_j of the input vector a from every weight vector $\mathbf{w}_j$, $j = 1, ..., m$, is calculated according to the formula

$$d_j = \sqrt{\sum_i (a_i - w_{ij})^2}. \qquad (1)$$

4. The output neuron with the least distance d_j is chosen as the winner.
5. For all output neurons lying inside the topological neighborhood of the winner, the weight vectors are moved toward the input vector according to the formula

$$\mathbf{w}_j \leftarrow \mathbf{w}_j + \eta(\mathbf{a} - \mathbf{w}_j). \qquad (2)$$

   After the modifications, the topological neighborhood becomes more sensitive to input vector a and similar input vectors.
6. The radius of the topological neighborhood ρ and the learning rate η are gradually decreased in order to achieve a more precise mapping after the gross topology has found its shape.
7. The training cycle, that is, steps 2–6, is repeated for a predefined number of times. This is typically of the order 10,000 ... 100,000.

After the SOM has completed the training, test vectors can be presented to it. The response evoked by the vector can be visualized as an activation pattern (see Figure 2). This can be carried out by defining an activation function, whose value depends on the error $\mathbf{a} - \mathbf{w}_j$, so that a small error leads to a high activation value, and vice versa. The response focus evoked by the vector **a** is defined as the neuron whose weight vector is closest to vector **a**. In other words, neuron *r* is the response focus, if

$$d_r = \min_j d_j, \tag{3}$$

where d_j is defined as in Equation 1.

Appendix 2
Spectral Centroid and Critical Band Rate

The spectral centroid $\langle f \rangle$ of each stimulus in hertz can be calculated by using the formula

$$\langle f \rangle = \frac{\sum_i a_i f_i}{\sum_i a_i} \tag{4}$$

where a_i is the maximal amplitude and f_i the frequency of partial *i*, and the summation is carried out across all partials.

The centroid can also be expressed in critical band rate (Bark scale). When exposed to a stimulus, the auditory system performs a spectrographic analysis. The cochlea can be regarded as a bank of band-pass filters whose center frequencies are ordered tonotopically. Critical band rate is a measure of tonotopic position in the auditory system and is thus useful in producing tonotopic spectra of sounds. The critical band rate (on the Bark scale), can be calculated according to

$$z = \frac{26.81}{1 + 1960/f} - 0.53 \tag{5}$$

where frequency *f* is expressed in hertz (Traunmüller, 1990).