PAN
ARCHIVES OF ACOUSTICS
Vol. **44**, No. 3, pp. 561–573 (2019)
DOI: 10.24425/aoa.2019.129271

Prediction of Psychoacoustic Metrics Using Combination of Wavelet Packet Transform and an Optimized Artificial Neural Network

Mehdi POURSEIEDREZAEI(1), Ali LOGHMANI(2)*, Mehdi KESHMIRI(2)

(1) *Mechanical Engineering Group, Pardis College*
Isfahan University of Technology
Isfahan 84156-83111, Iran; e-mail: s.pourseiedrezaei@pa.iut.ac.ir

(2) *Department of Mechanical Engineering*
Isfahan University of Technology
Isfahan 84156-83111, Iran; e-mail: mehdik@cc.iut.ac.ir
*Corresponding Author e-mail: a.loghmani@cc.iut.ac.ir

(*received December 10, 2018; accepted April 23, 2019*)

In this paper, a modified sound quality evaluation (SQE) model is developed based on combination of an optimized artificial neural network (ANN) and the wavelet packet transform (WPT). The presented SQE model is a signal processing technique, which can be implemented in current microphones for predicting the sound quality. The proposed method extracts objective psychoacoustic metrics including loudness, sharpness, roughness, and tonality from sound samples, by using a special selection of multi-level nodes of the WPT combined with a trained ANN. The model is optimized using the particle swarm optimization (PSO) and the back propagation (BP) algorithms. The obtained results reveal that the proposed model shows the lowest mean square error and the highest correlation with human perception while it has the lowest computational cost compared to those of the other models and software.

Keywords: sound quality measurement; psychoacoustic metrics; wavelet packet transform; optimized artificial neural network.

1. Introduction

The sound quality (SQ) is a general interpretation of human feelings from a received sound (Blauert, Jekosch, 1998; Pleban, 2014). Since physical characteristics of a sound cannot express the human audial stimulation, human perception of the sound should be assessed. A common approach to reach this purpose is to undertake auditory tests, which are complex and time-consuming. In the past decades, many researchers paid much attention to the SQ of radiated sound from medical equipment, home appliances, vehicles, airplanes and trains to evaluate both the quality of the equipment and the pleasantness of the perceived radiated sound.

Several psychoacoustic indices, including A-, B-, C-, D-weighted sound pressure level, loudness, sharpness, roughness, fluctuation strength, tonality, annoyance, and pleasantness, have been presented to quantitatively illustrate subjective feelings of perceived sounds (Carletti, 2013; Fastl, Zwicker, 2007). Failure to consider masking effects results in a poor correlation between weighting functions and the perceived feelings (Parmanen, 2007; Pleban, 2010). This has given rise to develop several psychoacoustic metrics including loudness (Fastl, Zwicker, 2007; Klonari *et al.*, 2011; de Oliveira *et al.*, 2009; Wang *et al.*, 2014), sharpness (Leite *et al.*, 2008; Wang *et al.*, 2007), roughness (Aures, 1985b; Miskiewicz *et al.*, 2007; Szczepańska-Antosik, 2008; Vencovský, 2016), and tonality (Aures, 1985a; Hasting, Davies, 2002; Kim *et al.*, 2012) for describing the perceived feelings. Each of these metrics represents one or more particular aspect of the sound. The SQ combines these indices for predicting the pleasantness or annoyance (Kaczmarek, Preis, 2010).

Human will experience different perceptions from various sound samples with different frequency content (Silva, 2002). Thus, selecting a signal processing approach for extracting the sound features based on

hearing characteristics is important. The fast Fourier transform (FFT) transforms a signal from time domain to frequency domain; however, it is suitable just for stationary signals. For time-frequency analysis, there are some other approaches including the short-time Fourier transform (STFT), wavelet transform (WT), Wigner-Ville distribution (WVD), and the Hilbert-Huang transform (HHT), which are usually employed for feature extraction of non-stationary signals (Błazejewski *et al.*, 2014; Huang *et al.*, 2015). The continuous wavelet transform (CWT) is commonly used for data analysis, while the discrete wavelet transform (DWT) is applied for image compression and pattern recognition (Mallat, 2009; Qin, Sun, 2015). The Wigner-Ville distribution is a popular approach thanks to its good time-frequency resolution, however, generates results with coarser granularity than those of the wavelet transform methods (Xing *et al.*, 2016). In (Wang *et al.*, 2007), DWT-based approaches were developed for SQ evaluation (SQE) of non-stationary vehicle noises. As an extension to the DWT, wavelet packet transform (WPT) provides a specific filter bank with identical features to the center frequencies and critical bandwidths (Majeed *et al.*, 2015; Parfieniuk *et al.*, 2006).

Given the complex and nonlinear relationships in the human auditory system, intelligent approaches have been investigated in the calculation of psychoacoustic indices. Two intelligent methods, namely artificial neural network (ANN) and support vector machine (SVM), have been used to classify psychoacoustic metrics such as loudness, roughness, and annoyance of vehicle noise (Chen *et al.*, 2011; Liu *et al.*, 2015). A review of the related literature shows that the ANN is more effective for predicting the SQ in intelligent SQE systems, thanks to its good performance and adaptability to complex nonlinear problems alongside its self-learning and self-organization characteristics (Fausett, 1994; Żwan, 2008).

In (Xing *et al.*, 2016), a SQ model was designed to evaluate non-stationary vehicle interior noise using a back propagation neural network (BPNN) model. Results showed good accuracy and efficiency of the model in mimicking the human hearing system, however, just two psychoacoustic indices including loudness and sharpness were predicted. As to the best knowledge of the authors, other previous models (except (Xing *et al.*, 2016)) estimated the pleasantness as a neural network output, by feeding the psychoacoustic metrics into the network. The previous approaches will not decrease computational cost because calculating the psychoacoustic metrics is the main computational load of predicting the pleasantness, thus, they are not suitable for real-time applications. Since SQ can be considered to estimate the overall pleasantness or annoyance of noise, one of the targets of this paper is to develop a SQE model and predict the objective psychoacoustic metrics to incorporate with active sound quality control (ASQC) system in the ongoing research. Consequently, there is a necessary need for presenting a new SQ prediction model that has the lowest computational load with the highest accuracy for real-time implementation of active noise control (ANC) system (Kuo, Morgan, 1996). Accordingly, traditional algorithms and commercial software cannot be used due to time delay execution.

The model presented in (Xing *et al.*, 2016) estimates only the loudness and sharpness, and there is a weak correlation between energy index and roughness and tonality, as the neural network input and outputs, respectively. Thus, this model cannot be used to estimate the roughness and tonality indices. However, the roughness and tonality are required for predicting the pleasantness. Therefore, in this manuscript, two other indices including the mean and standard deviation of the scalogram of sound signals are also added to the other input of the ANN, which is the energy index, for estimating the roughness and tonality as well as loudness and sharpness.

The BPNN can be trapped in local optima or engaged with a slow convergence rate because of not selecting proper primary weights and biases (Gori, Tesi, 1992; Jaddi, Abdullah, 2018; Zhang *et al.*, 2007). In (Zhang *et al.*, 2015a), the genetic algorithm (GA) and particle swarm optimization (PSO) were compared, in terms of efficiency of optimizing the primary coefficients used in a BPNN for predicting the pleasantness metric of vehicle interior noise. The case study investigated in that work was stationary. It is necessary to develop a technique to decompose components based on time-frequency features of non-stationary signals. In order to address these limitations, in this paper, evolutionary optimization methods optimize the ANN and then it is combined with the WPT in order to predict the psychoacoustic metrics of non-stationary noises with a low computational cost. All of the samples used herein are non-stationary which are decomposed to analyze over time and frequency domains. The results are validated by being compared against those of other investigations and models, demonstrating fast convergence and high accuracy of the proposed model in predicting the SQ at a low computational cost. The proposed algorithm can be implemented using embedded field-programmable gate array (FPGA) boards in microphones to directly measure the sound quality metrics in different appliances and devices.

The rest of this paper is organized as follows: Subsec. 2.1 introduces the psychoacoustic metrics. Theoretical foundations of the WPT and ANN are explained in Subsecs 2.2 and 2.3. The applied sound database is described in Sec. 3. Section 4 provides a derivation of the proposed SQE model based on the WPT and the optimized ANN. In this section, the simulation results

are presented to validate the performance of the proposed model. Conclusions are drawn in Sec. 5.

2. Background theory

2.1. Psychoacoustic metrics

Psychoacoustics deals with how a human perceives the received sound. In the middle ear, the sound waves are transformed into mechanical vibrations, which are then transferred into electrical signals once subjected to nonlinear filtering in the internal ear. Aural comprehension is formed in the human brain through the neural system. The biological structure of the basilar membrane in the internal ear is the basis of psychoacoustic effects. The most popular psychoacoustic metrics include loudness, sharpness, roughness, and tonality. These parameters and their attributes are summarized in Table 1. A detailed description is available in the literature (Fastl, Zwicker, 2007).

Table 1. Psychoacoustic metrics and their attributes.

Psychoacoustic metrics	Attributes
Loudness	This auditory characteristic reflects the effect of energy content on the human ear (Fastl, Zwicker, 2007).
Sharpness	This auditory characteristic is calculated as a weighted loudness focused on high-frequency contents for quantitative modelling (Fastl, Zwicker, 2007).
Roughness	This auditory characteristic is the subjective perception of amplitude modulation of a sound pressure signal, which is obtained by measuring the time variation of the loudness spectrum with modulating frequencies ranging from 20 to 300 Hz (Fastl, Zwicker, 2007).
Tonality	This auditory characteristic shows the presence of the tonal component in the content of broadband noise (Hasting, Davies, 2002; Kim *et al.*, 2012).

2.2. Wavelet Packet Transform

Wavelet transform refers to decomposing a signal using wavelets. Wavelets are a family of orthogonal functions that are obtained by scaling and translating a mother wavelet. In order to avoid redundancy in the wavelet function of CWT, its discrete version DWT is usually used in engineering applications. Based on the DWT, WPT can be extracted to provide a computationally efficient alternative to CWT with sufficient frequency resolution. Thus, the wavelet packet spectrum can be used to perform time-frequency analysis on non-stationary signals (Wang *et al.*, 2007; Zeng *et al.*, 2008). By selecting nodes at different levels of the WPT tree, a suitable model is obtained for the purpose of the present research.

2.3. Artificial Neural Network

Artificial Neural Networks (ANNs) are based on biological neural systems. An ANN consists of neurons organized in the input, output, and hidden layers. Neurons are connected to one another via sets of weights. During the learning process, ANN varies the weights and biases continuously. The main advantage of ANN in prediction processes lies in its ability to estimate strong nonlinear correlations. In order to overcome the associated challenges with ANN, new heuristic optimization methods or evolutionary algorithms have been used to optimize the ANN structure (Beheshti *et al.*, 2014; Razmjooy *et al.*, 2013; Zhang *et al.*, 2015b).

In this section, four ANN-based prediction models including BPNN, genetic algorithm-back-propagation neural network (GA-BPNN), particle swarm optimization-back-propagation neural network (PSO-BPNN) and imperialist competitive algorithm-back-propagation neural network (ICA-BPNN) which are optimized by back propagation (BP), GA, PSO and imperialist competitive algorithm (ICA), respectively, are considered. Taking into account the extracted features of a signal as the network input and psychoacoustic metrics as the network output, the SQ prediction performance could be compared among the developed neural network models. Figure 1 demonstrates the three-layer feed-forward neural network with one input layer, one hidden layer, and one output layer, that has been established in this paper. If the output of ANN fails to reach the predefined desired target, the network error criterion is computed and propagated backward to adjust the weights using the BP algorithm.

A three-layer back-propagation neural network can estimate any continuous nonlinear function (Hecht-Nielsen, 1992). Consider the three-layer BPNN shown in Fig. 1, where n, h, and m are the numbers of neurons in the input, hidden, and the output layers, respectively. In this paper, the transfer functions f_1 and f_2 were examined to minimize the training error, i.e. mean square error (MSE) between the predicted output and the desired value, as defined by Eq. (1)

$$\mathrm{MSE} = \frac{1}{2}\sum_{k=1}^{q}\frac{e_k}{(q\cdot m)}, \tag{1}$$

where $e_k = \sum_{i=1}^{m}\left(y_i^k - d_i^k\right)^2$, q is the number of total training samples, m is the number of outputs, y_i^k and d_i^k are, respectively, actual and desired outputs at the i-th output node for the k-th training sample (Fausett, 1994).

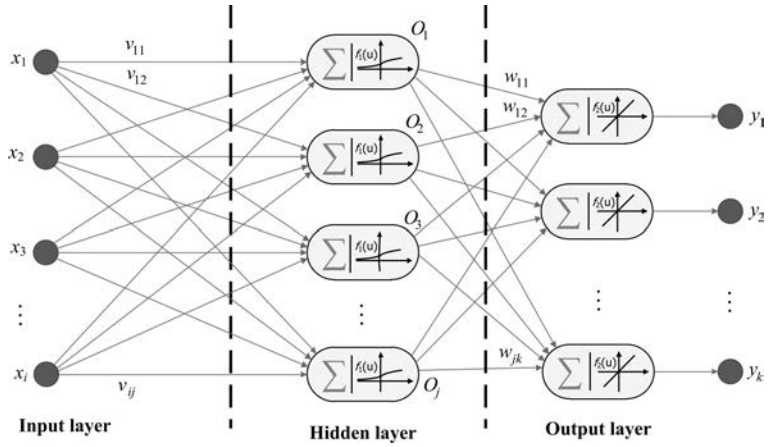

Fig. 1. A three-layer feed-forward neural network.

2.4. Optimization of primary weights of BPNN

Primary weights and thresholds are the most effective parameters on the performance of a neural network. Since gradient descent algorithm is usually used for minimizing the error function in the BP approach, random selection of the primary weight and threshold coefficients causes two main shortcomings: trapping in local minima and low convergence rate (Gori, Tesi, 1992; Jaddi, Abdullah, 2018; Zhang *et al.*, 2007). Therefore, global search methods have been applied to overcome these shortcomings. In this paper, GA, PSO, and ICA methods are used to optimize the primary values of the weights and thresholds, with their performances compared to one another.

The number of individuals is set as follows:

$$N = n_l n_h + n_h n_o + n_{h_{\text{bias}}} + n_{o_{\text{bias}}}, \tag{2}$$

where n_i, n_h, and n_o refer to the numbers of neurons in the input, hidden and output layers of the neural network, respectively, and $n_{h_{\text{bias}}}$ and $n_{o_{\text{bias}}}$ are the numbers of biases in the hidden and output layers, respectively.

3. Establishment of sound database

In order to evaluate the proposed model, a sound database is selected based on incorporating the model with ASQC system for noise control of the Neonatal intensive care unit (NICU) in our ongoing research. Moreover, one of the criteria for ranking the NICU is its environmental sound quality (Dunn *et al.*, 2013). NICU noise consists of equipment noise and human activity noise. Medical staff activities and treatment operations can generate noise. Caregiving routines involving talking, laughing and neutral emotional states can add to the neonate's environment noises. Moreover, the crying also increases NICU noise level (Olbrych, 2010). Therefore, in this paper, the Oxford Vocal Sounds (OxVoc) database is used for extracting the sound indices; the sound samples include natural affective vocal sounds from infants and adults (Parsons *et al.*, 2014). This database consists of a total of 173 non-verbal sounds including happy, sad and neutral emotional states. The main feature of this database is that it includes high-quality noise-free vocalizations from different naturalistic situations. In this paper, the psychoacoustic metrics of nonverbal affective sounds are investigated for the first time. The use of nonverbal vocalizations is very important in psychological feature recognition since the vocalizations involve no accent, individual features, and issues of authenticity. Thus, the vocalizations are being increasingly utilized today. The sampling frequency is set to 44 100 Hz for using the mentioned samples in the developed sound feature extraction model.

4. Sound feature extraction process

Since features of non-stationary audio signals should be specified in the time and frequency domains, selecting an appropriate time-frequency analysis approach is very important in the SQE. As the human auditory threshold is in the range of 20 Hz to 20 kHz, a three-order high-pass Butterworth filter is designed for eliminating the infrasound, which is below 20 Hz of the used sound signals. The resolution of WPT is linear in frequency domain and the frequency ranges are not overlapped. Therefore, it can be used for signal analysis at different frequency bands according to the human hearing system. In order to reach this goal and remove the redundant data resulted from the WPT, suitable nodes at multi levels of wavelet packet decomposition tree should be selected. Based on the sampling rate and the frequency partition of the critical bands, a nine levels wavelet packet decomposition tree is designed, as shown in Fig. 2. Then, a set of WPT nodes is selected manually to create 24 critical bands as shown in Table 2. Daubechies wavelet function (db35) is used in this paper.

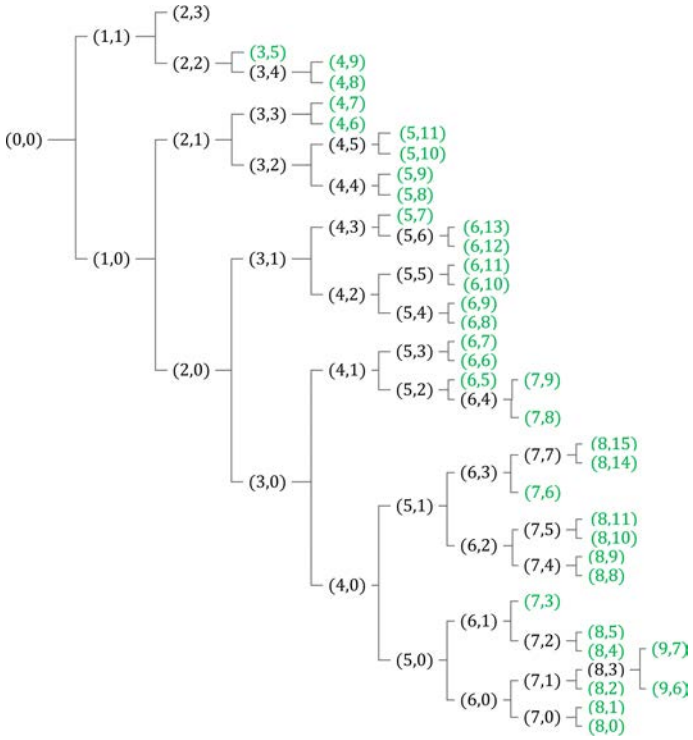

Fig. 2. The selected nodes from the wavelet packet decomposition tree and the corresponding 24 critical bands.

Table 2. Frequency range of the selected nodes in the WPT model corresponding to the 24 critical bands.

Critical band rate, z (Bark)	Critical bands [Hz]	Frequency range of the WPT [Hz]	Number of the selected nodes in the wavelet tree
1	(0–100)	(0–86)	(8, 0)
2	(100–200)	(86–172)	(8, 1)
3	(200–300)	(172–258),(258–301)	(8, 2), (9, 6)
4	(300–400)	(301–344),(344–430)	(8, 4), (9, 7)
5	(400–510)	(430–516)	(8, 5)
6	(510–630)	(516–688)	(7, 3)
7	(630–770)	(688–775)	(8, 8)
8	(770–920)	(775–861),(861–947)	(8, 9), (8, 10)
9	(920–1080)	(947–1033)	(8, 11)
10	(1080–1270)	(1033–1205),(1205–1291))	(7, 6), (8, 14)
11	(1270–1480)	(1291–1378),(1378–1550)	(7, 8), (8, 15)
12	(1480–1720)	(1550–1722)	(7, 9)
13	(1720–2000)	(1722–2067)	(6, 5)
14	(2000–2320)	(2067–2411)	(6, 6)
15	(2320–2700)	(2411–2756)	(6, 7)
16	(2700–3150)	(2756–3100)	(6, 8)
17	(3150–3700)	(3100–3445), (3445–3789)	(6, 9), (6, 10)
18	(3700–4400)	(3789–4134), (4134–4478)	(6, 11), (6, 12)
19	(4400–5300)	(4478–4823), (4823–5512)	(5, 7), (6, 13)
20	(5300–6400)	(5512–6201)	(5, 8)
21	(6400–7700)	(6201–6890), (6890–7579)	(5, 9), (5, 10)
22	(7700–9500)	(7579–8268), (8268–9746)	(4, 6), (5, 11)
23	(9500–12000)	(9746–11025), (11025–12403)	(4, 7), (4, 8)
24	(12000–15500)	(12403–13789), (13789–16537)	(3, 5), (4, 9)

Considering the Zwicker's model for calculating the loudness and sharpness of the sound, the sound energy distribution can be measured in the time-frequency domain to extract the sound features. The model presented in (Xing *et al.*, 2016) estimates only the loudness and sharpness, but the correlation analysis between the energy matrix as inputs and the roughness and tonality as outputs, shows that there is a weak correlation between the neural network inputs and outputs in this model. Table 3 shows the RMS error and correlation coefficient of the ANN model presented in (Xing *et al.*, 2016). As seen in Table 3, this model cannot be used to estimate roughness and tonality. However, the roughness and tonality are required for predicting the pleasantness. Thus, in this paper, in addition to energy matrix, two statistical features, namely mean and standard deviation of selected nodes in WPT model output, are used as neural network inputs to estimate the tonality and roughness metrics. The mean scalogram is a representation of the sound energy deformations and the standard deviation of scalogram reveals the temporal attributes and time fluctuations of the signal.

Table 3. The RMS error and correlation coefficient based on energy criteria input and objective psychoacoustic metric outputs by ANN model.

	Loudness	Sharpness	Roughness	Tonality
R^2	0.97317	0.93410	0.23934	0.45203
RMS	0.08190	0.10135	0.61151	0.39609

It is essential to analyze non-stationary sounds over time and frequency domains. The temporal masking effects in the human auditory system are considered by setting the resolution of the sound in the time domain to 50 ms, and the frequency masking is observed by setting up the frequency interval into 24 critical bands. Thus, each sound signal is partitioned into 24 by $T/50$ ms blocks, where T is the signal length and 50 ms is the frame length commonly used in psychoacoustics analysis.

A schematic presentation of the presented model for extracting the sound features is shown in Fig. 3. First, the sound signal is passed through the high-pass filter, and then it is divided into M frames of 50 ms in time length. Each frame is divided into 24 sub-signals by the multi-level node selection of the WPT. The energy value E_i is obtained for each sub-signal in a discrete form with the following relationship:

$$E_i = \sum_t [a_i(t)]^2 \Delta t, \tag{3}$$

where $a_i(t)$ and Δt are the amplitude of the i-th sub-signal, and the time interval of $a_i(t)$, respectively.

The total extracted feature matrix would be the input of the BPNN model, which is made by juxtaposing the feature blocks of the n signals with a size of $n \times (24 \times 3 \times T/50)$. The output of the neural network, which is the sound quality matrices (SQM), can be expressed as follows:

$$\text{SQM} = [\text{loudness sharpness roughness tonality}]^{\text{T}}. \tag{4}$$

4.1. Development of the sound quality prediction model

In order to evaluate the sound quality, it is necessary to map the extracted sound features to the related psychoacoustic metrics. Accordingly, the BPNN model can be used for this purpose. As previously mentioned, in order to overcome the challenges of the BPNN model, some evolutionary optimization algorithms (GA, PSO, and ICA) are used to obtain initial weights and thresholds of the BPNN model. The presented model used to predict the objective psychoacoustic parameters goes through the following steps (Fig. 4):

Step 1: Identification of the input and output nodes. Considering the energy, mean and standard deviation scalogram of the output sub-signals of the multi-level node selection WPT method, the number of neurons in the input layer is found to be $3 \times 24 = 72$. Four outputs there exist in the output of the ANN to represent the four psychoacoustic metrics: loudness, sharpness, roughness, and tonality.

Step 2: Selection of neurons in the hidden layer empirically via a trial and error approach.

Step 3: The linear function is selected as the output transfer function f_2, and two sigmoid transfer functions including logarithmic sigmoid and hyperbolic tangent functions can be used as the transfer function f_1 of the hidden layer.

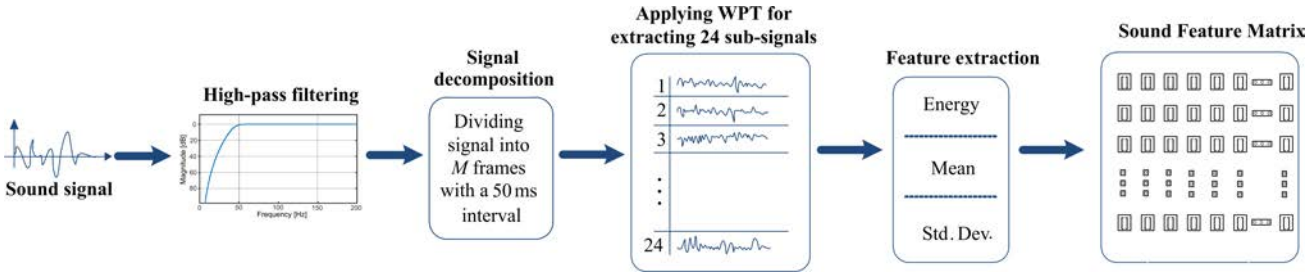

Fig. 3. Schematic presentation of the WPT model for sound feature extraction.

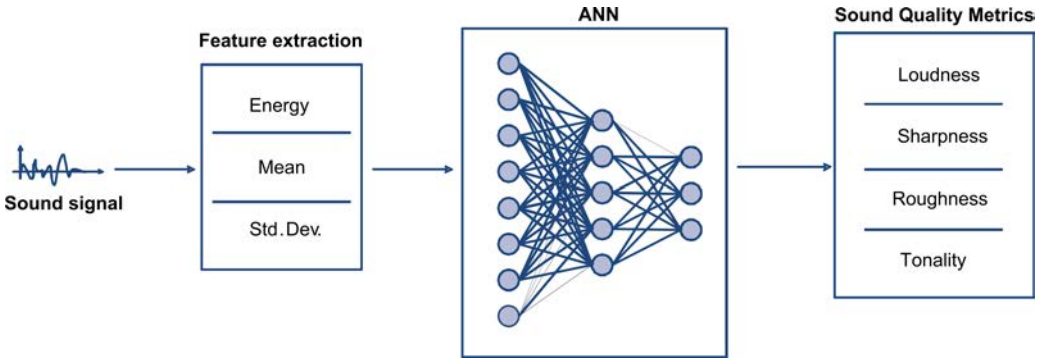

Fig. 4. Schematic of the combined BPNN and WPT model used for predicting the objective psychoacoustic metrics.

Step 4: Optimization of the initial weights and thresholds of the BPNN model using the evolutionary optimization algorithms.

Step 5: Training the BPNN.

Step 6: Computing hidden layer values and output layer values.

Step 7: Calculating MSE of the network (Eq. (1)).

Step 8: Updating the weights and thresholds of the neural network.

Step 9: If the MSE is less than a predetermined value, the BP algorithm stops, otherwise, the algorithm returns to *Step 2*.

4.2. Architectural design of optimized BPNN

In order to train the neural network, objective psychoacoustic metrics and sound feature matrices are computed by programming in MATLAB software. Seventy percent of the sound samples are randomly separated and are used to train the network; fifteen percent of the data are used for validation purposes, and the remaining formed a basis for testing the neural network. In order to achieve a precise prediction model, it is necessary to determine the network structure and the training parameters.

The optimal number of neurons in the hidden layer and the transfer functions are determined by trial and error. Accordingly, the number of neurons is experimentally changed from 8 to 50 according to the number of inputs and outputs of the neural network. In order to identify appropriate activation functions f_1 and f_2 in the hidden and output layers, different combinations of sigmoid functions (tansig, logsig) and linear function (purelin) are considered. The predicated RMS errors for different feasible combinations of the transfer functions are listed in Table 4. As seen, the combination of (logsig, purelin) gives the lowest mean RMS error. The effect of the number of neurons on the accuracy of the final BPNN model is shown for various output parameters based on the RMS errors in Table 5. The lowest RMS error value for the predicted psychoacoustic metrics is achieved at $h = 12$. Consequently, 12 neurons are used in the hidden layer, with the transfer functions f_1 and f_2 set to logarithmic sigmoid (logsig) and pure line (purelin), respectively.

Table 4. The RMS errors of the network outputs for different transfer functions.

Neuron number	All data	A	B	C	D
(tansig, purelin)					
8	0.1515	0.0980	0.1243	0.1617	0.2015
12	0.1764	0.1462	0.1630	0.1714	0.2173
16	0.1604	0.1121	0.1422	0.1650	0.2072
20	0.1754	0.1469	0.1628	0.1702	0.2146
25	0.1668	0.1264	0.1514	0.1666	0.2113
Mean	0.1661	0.1259	0.1487	0.1670	0.2104
(logsig, purelin)					
8	0.1506	0.0976	0.1162	0.1668	0.1998
12	0.1419	0.0824	0.1098	0.1568	0.1928
16	0.1685	0.1298	0.1528	0.1690	0.2117
20	0.1694	0.1290	0.1541	0.1699	0.2133
25	0.1529	0.0991	0.1309	0.1609	0.2019
Mean	0.1567	0.1076	0.1328	0.1647	0.2039
(tansig, tansig)					
8	0.1530	0.1209	0.1227	0.1593	0.1964
12	0.1517	0.1181	0.1292	0.1589	0.1900
16	0.1666	0.1401	0.1521	0.1672	0.2008
20	0.1680	0.1481	0.1421	0.1664	0.2074
25	0.1871	0.1470	0.1578	0.1742	0.2512
Mean	0.1653	0.1349	0.1408	0.1652	0.2092
(logsig, tansig)					
8	0.1442	0.0909	0.1163	0.1583	0.1905
12	0.1448	0.0922	0.1127	0.1555	0.1962
16	0.4048	0.7551	0.1376	0.1595	0.2019
20	0.6131	0.7551	0.9326	0.1573	0.1978
25	0.4042	0.7551	0.1348	0.1593	0.1995
Mean	0.3422	0.4897	0.2868	0.1580	0.1972

A – loudness, B – sharpness, C – roughness, D – tonality.

Table 5. Sensitivity analysis of the RMS errors of the BPNN model versus number of neurons in the hidden layer.

Neuron number	All data	A	B	C	D
7	0.150397	0.091415	0.120058	0.164543	0.201575
8	0.15063	0.097577	0.11625	0.166771	0.199774
10	0.157641	0.10973	0.132403	0.1668	0.20496
11	0.167282	0.125133	0.147472	0.170043	0.213572
12	0.141947	0.082369	0.109833	0.156811	0.192764
14	0.166157	0.125029	0.150333	0.167564	0.210053
15	0.16141	0.111813	0.145899	0.165064	0.207792
16	0.168519	0.129807	0.152776	0.169028	0.21174
18	0.168553	0.132291	0.151586	0.169371	0.21089
20	0.169388	0.129047	0.154095	0.169933	0.213292
25	0.152948	0.099071	0.130878	0.160866	0.201869
30	0.165381	0.122711	0.148726	0.167142	0.210452
35	0.164204	0.121196	0.148776	0.166221	0.208325
40	0.167162	0.124484	0.152715	0.168412	0.211167
45	0.167234	0.130515	0.152295	0.165501	0.210356
50	0.168237	0.132168	0.151485	0.166759	0.21211

A – loudness, B – sharpness, C – roughness, D – tonality.

Table 6. Optimal parameters for the four SQ prediction models designed in this research.

Type of algorithm	Parameters	Value
BP	Number of the hidden layer	1
	Number of neurons in the input layer	72
	Number of neurons in the hidden layer	12
	Number of neurons in the output layer	4
	Transfer function of the input-hidden layer	Logsig
	Transfer function of the hidden-output layer	Purelin
	Training function	Levenberg-Marquardt
	Momentum factor	0.9
	Learning rate	0.5
	Training target of MSE	0.001
	Testing performance	MSE
GA	Population size of GA	150
	Max generation	100
	Crossover factor	0.5
	Mutation factor	0.02
PSO	Population size of PSO	200
	Max generation	100
	Acceleration factors	2
	Inertial factor	0.7 to 0.4
	Particle dimension	928
ICA	Number of Countries	200
	Number of Initial Imperialists	50
	Number of Decades	50
	Revolution Rate	0.3
	ξ	0.02
	γ	0.5
	β	2

Prior to the network training, all data should be normalized to the interval [−1, 1] to remove the effect of data magnitudes and prevent large prediction errors. The normalization is using Eq. (5):

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{5}$$

where x_i' is the normalized value.

Final optimal parameters of the proposed method are reported in Table 6. Performance curve of the four well-trained models is shown in Fig. 5, where the horizontal axis denotes the number of iterations and the vertical axis shows the prediction MSE of the network. This figure reveals that the ICA-BPNN model converges to the target MSE at 640 iterations. The convergence rate of this model is approximately 1.5 times, 1.8 times, and 2.2 times faster than the convergence rate of the PSO-BPNN model, GA-BPNN model, and standard BPNN model, respectively.

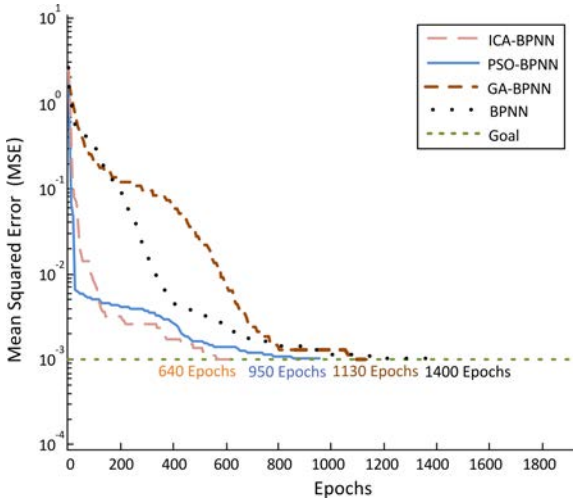

Fig. 5. Performance curve of the four SQ prediction models.

4.3. Discussion and model verification

In Table 7, RMS errors and R^2 values are tabulated for BPNN, GA-BPNN, PSO-BPNN, and ICA-BPNN models in the testing phase. By comparing the results, we concluded that the PSO-BPNN model has higher accuracy and lower RMS error than the other three models. So that the corresponding mean RMS value of PSO-BPNN is 79%, 70%, and 67% of those of ICA-BPNN, GA-BPNN, and standard BPNN model, respectively. The mean correlation coefficient indicates that the developed model is highly reliable for SQE applications.

Plots of comparison between real and prediction outputs of the BPNN model are shown in Fig. 6 for loudness, sharpness, roughness and tonality metrics. According to the figures, the prediction outputs of the BPNN model show good agreement with the real data. Better conformity can be found for the loudness and sharpness results. Since the roughness and tonality values of the applied sound data are relatively small, the corresponding graphs exhibit a lot of compaction.

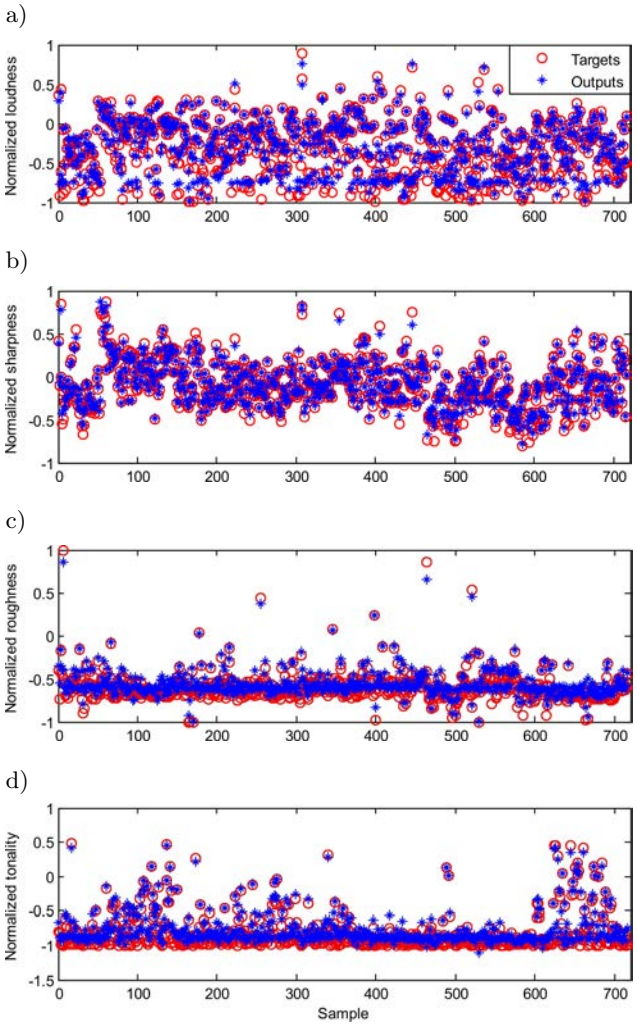

Fig. 6. Comparison between real and predicted SQE model for: a) normalized loudness, b) normalized sharpness, c) normalized roughness, and d) normalized tonality.

Table 7. Performance of BPNN, GA-BPNN, PSO-BPNN, and ICA-BPNN models.

Model	Loudness		Sharpness		Roughness		Tonality		Mean	
	RMS	R^2	RMS	R^2	RMS	R^2	RMS	R^2	RMS	R^2
BPNN	0.08237	0.98237	0.10983	0.98883	0.15681	0.88528	0.19276	0.93396	0.13544	0.94761
GA-BPNN	0.09138	0.98084	0.09859	0.98925	0.14259	0.88564	0.18443	0.93323	0.12925	0.94724
PSO-BPNN	0.09210	0.98526	0.10473	0.98936	0.09885	0.88234	0.06543	0.93104	0.09028	0.947
ICA-BPNN	0.09151	0.98203	0.08504	0.98885	0.09717	0.8919	0.18247	0.93519	0.11405	0.94949

Three typical sound samples (Signals 1 to 3) including cry, laughter, and neutral vocalizations of adults are selected. Based on the presented algorithm in Fig. 3, the feature matrix is computed for each sound signal and then fed to the presented model. Table 8 presents the corresponding mean percent error and RMS error to each signal for each optimized model in the simulation phase.

The maximum mean percent error for loudness is 6.2729%, with a maximum RMS of 0.1443, while the highest mean error for sharpness is 6.7971% with a maximum RMS of 0.1524. The corresponding values to roughness are 6.3741% and 0.1480, respectively, while those of tonality are 6.5799% and 0.1574, respectively. The highest mean percent error is that of PSO-BPNN (4.4726%), which is 0.9182% lower than that of ICA-BPNN model (5.3908%), 1.0152% lower than that of GA-BPNN model (5.4878%), and 1.8122% lower than that of the BPNN model (6.2848%). In addition, the highest mean RMS of PSO-BPNN is calculated to be 0.1006, i.e. 22%, 28%, and 45% lower than those of ICA-BPNN (0.1228), GA-BPNN (0.1288), and BPNN model (0.1464), respectively. An overview of the results demonstrates that maximum mean percent error of the developed models is 6.3%, indicating very good accuracy for calculating the loudness, sharpness, roughness and tonality indices.

Based on the above reviews, it can be concluded that, among others, the multi-level node selection of the WPT combined with the PSO-BPNN is the most accurate model for predicting the SQ of the signal, although the convergence rate of ICA-BPNN is faster, as seen in Fig. 5. The developed prediction models are considerably superior over traditional models for computing psychoacoustic indices and have good generalizability for direct SQE. By applying the presented models, especially for non-stationary signals, the SQE can be carried out conveniently rather than performing the complex subjective evaluations and time-consuming and costly tests. To compare the computational load in computing the psychoacoustic indices, the calculating times of four metrics including loudness, sharpness, roughness, and tonality between the traditional algorithms, commercial software and the proposed model are shown in Table 9. A computer with processor Intel(R) Core(TM) 2 Duo CPU 2.66 GHz and RAM 4 GB performs all psychoacoustic calculations. The results clearly demonstrate the superiority of the proposed method in reducing the computational burden for predicting the quality of non-stationary sounds.

In addition, the computed psychoacoustic indices for Signal 1 are compared with the results from the corresponding mathematical models in Fig. 7. The figure shows that there is a good agreement in the variation tendencies between BPNN model outputs and the results from mathematical models.

Table 8. Loudness, sharpness, roughness, and tonality errors between real data and predicted values using the presented models for three selective signals.

Model	Loudness		Sharpness		Roughness		Tonality		Mean	
	Mean [%]	RMS	Mean [%]	RMS	Mean [%]	RMS	Mean [%]	RMS	Mea [%]	RMS
BPNN	6.2729	0.1400	6.7971	0.1524	5.7872	0.1480	6.2820	0.1453	6.2848	0.1464
GA-BPNN	5.7456	0.1336	5.1816	0.1231	6.2719	0.1392	4.7522	0.1194	5.4878	0.1288
ICA-BPNN	4.8549	0.1138	5.6464	0.1244	5.6732	0.1250	4.4641	0.1114	5.1597	0.1187
PSO-BPNN	3.6269	0.0819	3.2486	0.0778	4.2600	0.0980	3.2578	0.0806	3.5983	0.0846
BPNN	5.9475	0.1390	5.3186	0.1280	6.2028	0.1465	6.5799	0.1449	6.0122	0.1396
GA-BPNN	5.6212	0.1286	5.8412	0.1317	6.1609	0.1408	4.1354	0.1006	5.4396	0.1254
ICA-BPNN	5.5848	0.1236	5.4748	0.1255	5.8174	0.1334	4.6864	0.1085	5.3908	0.1228
PSO-BPNN	4.4812	0.1031	4.5459	0.1015	3.6822	0.0869	3.5839	0.0812	4.0733	0.0932
BPNN	6.1599	0.1443	6.6779	0.1348	6.3741	0.1423	5.6989	0.1574	6.2277	0.1447
GA-BPNN	5.3350	0.1251	5.3076	0.1226	4.7834	0.1154	4.7828	0.1205	5.0522	0.1209
ICA-BPNN	5.4491	0.1226	4.5428	0.1092	4.7462	0.1105	4.9357	0.1198	4.9184	0.1155
PSO-BPNN	5.4239	0.1157	4.2089	0.0971	4.1514	0.0972	4.1063	0.0924	4.4726	0.1006

Table 9. The computational time of calculating loudness, sharpness, roughness, and tonality in various models for another test cry vocalization of adult "Adultfemale_cry02".

Psychoacoustic Metrics Code	MATLAB Code (mathematical models)	LabVIEW Sound and Vibration Toolkit	Proposed model
Computational time [s]	1.6659	1.5609	0.7141

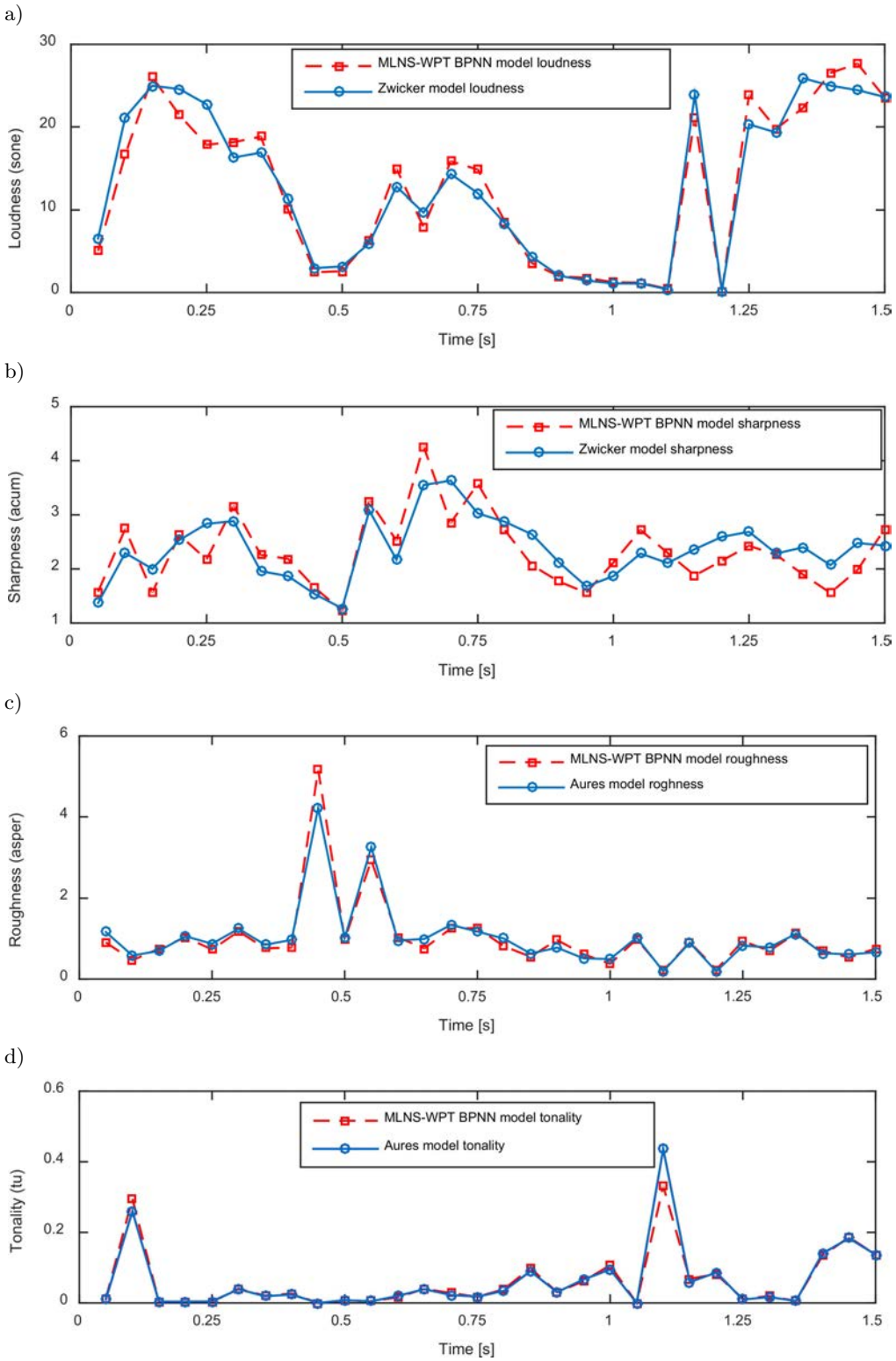

Fig. 7. Comparison of time-varying SQE results between corresponding mathematical model outputs and the presented model outputs for another test cry vocalization of adult "Adultfemale_cry02": a) loudness, b) sharpness, c) roughness, and d) tonality.

5. Conclusion

In this paper, a modified intelligent model was presented to determine objective sound metrics for using in SQE directly by combining the WPT method and an optimized ANN. This model can be combined with current microphones using embedded FPGA boards to measure sound quality in various applications. The

main advantages of the proposed model are in its low computational cost and high accuracy compared to previous models. These advantages lead to using this measuring technique in real-time applications. Multilevel nodes selection in the WP tree was used to extract feature matrices of sound signals corresponding to human auditory critical bands. The feature matrices include energy, mean and standard deviation values of the sub-signals. Then, these features were fed to the ANN in order to predict the sound quality metrics. Various optimization algorithms including BP, GA, PSO, and ICA were studied to find the best primary weights and thresholds in the ANN. Results demonstrate that the PSO is the most effective algorithm in finding the optimum. The overall comparison between the developed model outputs and the corresponding psychoacoustic models show that maximum mean error is as low as 6.3% and the correlation coefficient is more than 0.9, while the computational cost is lower than the previous models. The newly presented model provides an important and effective tool in analyzing SQ in real-time applications such as incorporating with ASQC of nonstationary noises and gives a reliable technique for studies related to human hearing.

References

1. Aures W. (1985a), *Calculation method for the sensory euphony of any sound signals* [in German: *Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale*], Acta Acoustica United with Acustica, **59**, 2, 130-–141.
2. Aures W. (1985b), *Method for calculating auditory roughness* [in German: *Ein Berechnungsverfahren der Rauhigkeit*], Acta Acoustica United with Acustica, **58**, 5, 268—281.
3. Beheshti Z., Shamsuddin S.M.H., Beheshti E., Yuhaniz S.S. (2014), *Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis*, Soft Computing, **18**, 11, 2253–2270. doi: 10.1007/s00500-013-1198-0.
4. Blauert J., Jekosch U. (1998), *Product-sound quality: A New aspect of machinery noise*, Archives of Acoustics, **23**, 1, 105–124.
5. Błazejewski A., Kozioł P., Łuczak M. (2014), *Acoustical analysis of enclosure as initial approach to vehicle induced noise analysis Comparatevely using STFT and wavelets*, Archives of Acoustics, **39**, 3, 385–394, doi: 10.2478/aoa-2014-0042.
6. Carletti E. (2013), *A perception-based method for the noise control of construction machine*s, Archives of Acoustics, **38**, 2, 253–258, doi: 10.2478/aoa-2013-0030.
7. Chen X., Hu H., Liu F., Gao X.X. (2011), *Image reconstruction for an electrical capacitance tomography system based on a least-squares support vector machine and a self-adaptive particle swarm optimization algorithm*, Measurement Science and Technology, **22**, doi: 10.1088/0957-0233/22/10/104008.
8. Dunn M.S., Erickson D., Avenue H., Gregory S. (2013), *Recommended standards for newborn ICU design, eighth edition*, Journal of Perinatology, **33**, S2–S16, doi: 10.1038/jp.2013.10.
9. Fastl H., Zwicker E. (2007), *Psychoacoustics: facts and models*, Springer, Berlin, Germany, 3rd ed., retrieved from http://dx.doi.org/10.1007/978-3-540-68888-4.
10. Fausett L. (1994), *Fundamentals of Neural Networks*, Prentice-Hall, Englewood Cliffs, NJ.
11. Gori M., Tesi A. (1992), *On the Problem of Local Minima in Recurrent Neural Networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **14**, 76–86, doi: 10.1109/34.107014.
12. Hasting A., Davies P. (2002), *An examination of Aures's model of tonality*, Proceeding on Sound Quality Symposium, **29**, 4–9.
13. Hecht-Nielsen R. (1992), *Theory of the backpropagation neural network*, [in:] H. Wechsler, V. Fairfax [Eds.], *Neural networks for perception: computation, learning,* architectures, vol. 2, pp. 65–93, Harcourt Brace & Co., Orlando, FL, http://dl.acm.org/citation.cfm?id=140639.140643.
14. Huang H.B., Li R.X., Huang X.R., Yang M.L., Ding W.P. (2015), *Sound quality evaluation of vehicle suspension shock absorber rattling noise based on the Wigner-Ville distribution*, Applied Acoustics, **100**, 18–25, doi: 10.1016/j.apacoust.2015.06.018.
15. Jaddi N.S., Abdullah S. (2018), *Optimization of neural network using kidney-inspired algorithm with control of filtration rate and chaotic map for real-world rainfall forecasting*, Engineering Applications of Artificial Intelligence, **67**, 246–259, doi: 10.1016/j.engappai.2017.09.012.
16. Kaczmarek T., Preis A. (2010), *Annoyance of time-varying road-traffic noise*, Archives of Acoustics, **35**, 3, 383–393, doi: 10.2478/v10168-010-0032-2.
17. Kim E.Y., Lee Y.J., Lee S.K. (2012), *Sound metric design for evaluation of tonal sound in laser printer*, International Journal of Precision Engineering and Manufacturing, **13**, 1349–1358, doi: 10.1007/s12541-012-0178-0.
18. Klonari D., Pastiadis K., Papadelis G., Papanikolao G. (2011), *Loudness assessment of musical tones equalized in a-weighted level*, Archives of Acoustics, **36**, 2, 239–250, doi: 10.2478/v10168-011-0019-7.
19. Kuo S., Morgan D. (1996), *Active noise control systems: algorithms and DSP implementations*, Wiley, New York, NY, USA.
20. Leite R.P., Paul S., Gerges S.N.Y. (2008), *A sound quality-based investigation of the HVAC system noise of an automobile model*, Applied Acoustics, **70**, 1–10, doi: 10.1016/j.apacoust.2008.06.010.

21. Liu H., Zhang J., Guo P., Bi F., Yu H., Ni G. (2015), *Sound quality prediction for engine-radiated noise*, Mechanical Systems and Signal Processing, **56**, 277–287, doi: 10.1016/j.ymssp.2014.10.005.

22. Majeed S.A., Husain H., Samad S.A. (2015), *Phase autocorrelation bark wavelet transform (PACWT) features for robust speech recognition*, Archives of Acoustics, **40**, 1, 25–31. doi: 10.1515/aoa-2015-0004.

23. Mallat S. (2009), *A wavelet tour of signal processing*, Academic Press, 3^{rd} ed., Burlington, MA, doi: 10.1016/B978-0-12-374370-1.X0001-8.

24. Miskiewicz A., Rogala T., Szczepańska-Antosik J. (2007), *Perceived roughness of two simultaneous harmonic complex tones*, Archives of Acoustics, **32**, 4, 737–748.

25. Olbrych S. (2010), *Noise pollution in the NICU*, Case Western Reserve University, retrived from https://case.edu/med/epidbio/mphp439/NoisePollution_NICU.pdf.

26. de Oliveira L.P.R., Janssens K., Gajdatsy P., Van der Auweraer H., Varoto P.S., Sas P., Desmet W. (2009), *Active sound quality control of engine induced cavity noise*, Mechanical Systems and Signal Processing, **23**, 2, 476–488, doi: 10.1016/j.ymssp.2008.04.005.

27. Parfieniuk M., Baszun J., Petrovsky A.A. (2006), *Computing of masking thresholds for audio coders based on a quaternionic 4-band wavelet packet transform*, Archives of Acoustics, **31**, 1, 155–165.

28. Parmanen J. (2007), *A-weighted sound pressure level as a loudness/annoyance indicator for environmental sounds – Could it be improved?*, Applied Acoustics, **68**, 58–70, doi: 10.1016/j.apacoust.2006.02.004.

29. Parsons C.E., Young K.S., Craske M.G., Stein A.L., Kringelbach M.L. (2014), *Introducing the Oxford Vocal (OxVoc) Sounds database: A validated set of non-acted affective sounds from human infants, adults, and domestic animals*, Frontiers in Psychology, **5**, 562, doi: 10.3389/fpsyg.2014.00562.

30. Pleban D. (2010), *Method of acoustic assessment of machinery based on global acoustic quality index*, Archives of Acoustics, **35**, 2, 223–235.

31. Pleban D. (2014), *Definition and measure of the sound quality of the machine*, Archives of Acoustics, **39**, 1, 17–23, doi: 10.2478/aoa-2014-0003.

32. Qin J., Sun P. (2015), *Applications and comparison of continuous wavelet transforms on analysis of A-wave impulse noise*, Archives of Acoustics, **40**, 4, 503–512, doi: 10.1515/aoa-2015-0050.

33. Razmjooy N., Mousavi B.S., Soleymani F. (2013), *A hybrid neural network imperialist competitive algorithm for skin color segmentation*, Mathematical and Computer Modelling, **57**, 848–856. doi: 10.1016/j.mcm.2012.09.013.

34. Silva M.C.G. (2002), *Measurements of comfort in vehicles*, Measurement Science and Technology, **13**, 41–60.

35. Szczepańska-Antosik J. (2008), *Roughness of two simultaneous harmonic complex tones in various pitch registers*, Archives of Acoustics, **33**, 1, 73–78.

36. Vencovský V. (2016), *Roughness prediction based on a model of cochlear hydrodynamics*, Archives of Acoustics, **41**, 2, 189–201, doi: 10.1515/aoa-2016-0019.

37. Wang Y.S. (2009), *Sound quality estimation for non-stationary vehicle noises based on discrete wavelet transform*, Journal of Sound and Vibration, **324**, 3, 1124–1140, doi: 10.1016/j.jsv.2009.02.034.

38. Wang Y.S., Lee C.M., Kim D.G., Xu Y. (2007), *Sound-quality prediction for nonstationary vehicle interior noise based on wavelet pre-processing neural network model*, Journal of Sound and Vibration, **299**, 4, 933–947, doi: 10.1016/j.jsv.2006.07.034.

39. Wang Y.S., Shen G.Q., Xing Y.F. (2014), *A sound quality model for objective synthesis evaluation of vehicle interior noise based on artificial neural network*, Mechanical Systems and Signal Processing, **45**, 1, 255–266, doi: 10.1016/j.ymssp.2013.11.001.

40. Xing Y.F.F., Wang Y.S.S., Shi L., Guo H., Chen H. (2016), *Sound quality recognition using optimal wavelet-packet transform and artificial neural network methods*, Mechanical Systems and Signal Processing, **66–67**, 875–892, doi: 10.1016/j.ymssp.2015.05.003.

41. Zeng X., Zhao W., Sheng J. (2008), *Corresponding relationships between nodes of decomposition tree of wavelet packet and frequency bands of signal subspace*, Acta Seismologica Sinica, **21**, 1, 91–97, doi: 10.1007/s11589-008-0091-x.

42. Zhang E., Hou L., Shen C., Shi Y., Zhang Y. (2015), *Sound quality prediction of vehicle interior noise and mathematical modeling using a back propagation neural network (BPNN) based on article swarm optimization (PSO)*, Measurement Science and Technology, **27**, 1, 15801, doi: 10.1088/0957-0233/27/1/015801.

43. Zhang J.R., Zhang J., Lok T.M., Lyu M. R. (2007), *A hybrid particle swarm optimization-back-propagation algorithm for feedforward neural network training*, Applied Mathematics and Computation, **185**, 2, 1026–1037, doi: 10.1016/j.amc.2006.07.025.

44. Żwan P. (2008), *Automatic singing quality recognition employing artificial neural networks, Archives of Acoustics*, **33**, 1, 65–71, http://acoustics.ippt.gov.pl/index.php/aa/article/view/631.