Journal of Scientific & Industrial Research
Vol. 79, March 2020, pp. 197-200

Development and use of a new Speech Quality Evaluation Parameter ESNR using ANN and Grey Wolf Optimizer

Tusar Kanti Dash[1,2,]* and Sandeep Singh Solanki[1]

[1]Electronics and Communication Engineering, Birla Institute of Technology, Mesra, India
[2]Electronics and Telecom Engineering, C V Raman College of Engineering, Bhubaneswar, India

Received 20 November 2019; revised 19 November 2019; accepted 23 January 2020

The performance of Speech Enhancement (SE) Algorithms is evaluated using various objective and subjective evaluation parameters. Recently, few objective evaluation parameters are developed for the measurement of speech quality and intelligibility. But still, there are ample scopes determining statistical parameters to predict the SNR of a noisy speech signal without using any reference of clean signal and noise. In this paper, this problem has been addressed and three types of Artificial Neural Networks (ANN) are developed for efficient prediction of the estimated SNR (E-SNR) of a given noisy speech signal. To further improve the accuracy of prediction of the SNR of the ANN, the coefficients of ANN are tuned using the bio-inspired optimization technique. In this paper, a popular and efficient Grey wolf Optimization is chosen for the purpose. Several audio features are studied and appropriate features are chosen as the inputs to the ANN. Finally, a comparative performance analysis is carried out using two standard speech databases and the best performing ANN and audio features are identified to provide the best ESNR.

Keywords: Speech Enhancement, Objective Evaluation Parameter, Speech Quality and Intelligibility, Artificial Neural Network, E-SNR

Introduction

The speech enhancement (SE) algorithms are used in the field of speech recognition, and speech communications and in hearing aids[1,2]. In these SE algorithms, the main goal is to reduce the noise level. But the noise level is not constant and varies for different conditions. Therefore, the performance of SE algorithms depends upon the accurate estimation of a noise level so that noise should be subtracted from the noisy signal without distorting the original clean speech signal[3]. It is often required to measure the performance of these algorithms using subjective or objective evaluation measures. Subjective listening tests are more reliable but can be time-consuming and require trained listeners. To overcome these difficulties several objective evaluation measures are designed for measuring speech quality and intelligibility[3]. The objective evaluation parameters can be broadly classified into three categories such as speech quality, intelligibility and statistical measures. Out of these statistical measures signal to noise ratio is a crucial parameter. Some statistical evaluation parameters based on SNR like Segmental SNR[4], SNR loss[5] have been proposed. But, scope exists for developing an algorithm to Estimated the SNR, which would predict the actual SNR of the noisy speech signal without using any reference of clean speech and noise. This problem has been addressed in the paper.

The rest of the paper is organized as follows. The materials and methods relating to the ANN, Grey wolf optimizer, speech data sets, audio features are presented in Section 2. In Section 3, the methodology of implementation and simulation results are obtained and discussed. Finally, the conclusion and scope for further research work are provided in Section 4.

Materials and methods

In this section, the details of the speech data sets, the ANN and GWO methods are discussed in detail.

Artificial Neural Network

In this section, the popular and low complexity ANN-based model Trigonometric Functional Link Artificial Neural Network (TFLANN) is dealt. It is used to develop a model to predict the relationship between the audio features and the SNR level[6]. The FLANN is a single neuron single layer-based architecture without any hidden layer. It involves lesser computational complexity and faster convergence

*Author for Correspondence:
E-mail: tusarkantidash@gmail.com

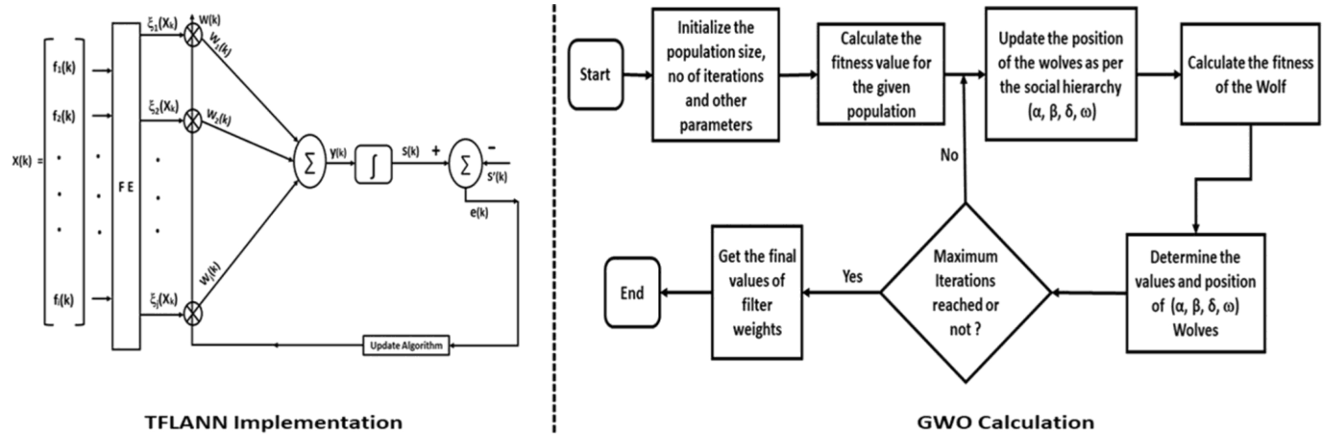
X(k) =
f₁(k)
f₂(k)
fᵢ(k)
F E
ξ₁(Xₖ)
ξ₂(Xₖ)
ξⱼ(Xₖ)
W(k)
W₁(k)
W₂(k)
Wⱼ(k)
Σ
y(k)
∫
S(k)
+
−
S'(k)
e(k)
Update Algorithm
TFLANN Implementation
Start
Initialize the population size, no of iterations and other parameters
Calculate the fitness value for the given population
Update the position of the wolves as per the social hierarchy (α, β, δ, ω)
Calculate the fitness of the Wolf
Determine the values and position of (α, β, δ, ω) Wolves
Maximum Iterations reached or not ?
No
Yes
Get the final values of filter weights
End
GWO Calculation

Fig. 1 — Block Diagram of TFLANN Implementation and GWO Algorithm Implementation

rate[7]. The block diagram of TFLANN model is presented in Figure 1. The weights of this model are updated according to Equation (1)

$$\xi\{x_j(k)\} = \{x_j(k), cos\ \pi x_j(k), sin\ \pi x_j(k) \ .... \\ cos\ N\pi x_j(k), sin\ N\pi x_j(k) \qquad \ldots (1)$$

$$y(k) = \xi(k)^T W,$$

Where $W = [w_1(l), w_2(l), \ldots w_m(l)]^T$ … (2)

During the training phase, when a particular input ***x(k)*** is applied, the model generates the output *s(k)*. It is then compared with the desired output *s'(k)*. By comparing *s(k)* and *s'(k),* the error term *e(k)* is found out as $e(k) = s(k) - s'(k)$. The weight matrix (***W***) is optimized according to the popular least mean square algorithm. So the accuracy of the prediction depends on the adaptive algorithm used for updating the connecting weights. In this proposed algorithm, the GWO is used for weights update. This process is dealt in the next section.

Grey Wolf Optimizer

Grey wolf optimizer (GWO) is a meta-heuristic algorithm based on the concept of social leadership and hunting skills of Grey wolves [8,9]. This algorithm is designed on the basis of the following three steps. (1) Following the prey (2) encircling prey and (3) attacking the prey. The block diagram of the implementation of the GWO is shown in Figure 1. Four types of grey wolves such as α, β, δ, ω are used for determining the leadership hierarchy. After the implementation of the fitness function, α wolf is taken as the best solution, β and δ are considered as the next two best solutions. The remaining wolves are taken as ω. The encircling behavior is represented as $\vec{D} = \left|C.\overrightarrow{X_P}(k) - \vec{X}(k)\right|$ and $\vec{X}(k+1) = \overrightarrow{X_P}(k) - A.\vec{D}$, where $\overrightarrow{X_P}(k)$ and $\vec{X}(k)$ represents the position vectors of the prey and the wolf for the *k*th iteration. The coefficients are denoted as A and C. The optimum solution is taken the prey that is to be hunted. The wolves update their positions according to the positions of the α, β, δ using the hunting model [$\overrightarrow{D_\alpha} = \left|C_1\,\overrightarrow{X_\alpha}(k) - A_1\overrightarrow{D_\alpha}\right|$ and $\overrightarrow{X_1} = \overrightarrow{X_\alpha} - A_1\overrightarrow{D_\alpha}$], where A_1 and C_1 are the coefficients with a different set of random numbers. This process is repeated for the β, δ using the same hunting model. The position of any wolf is updated as $\vec{X}(k+1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3}$. The value of A plays a crucial role in exploitation and exploration. If $|A| < 1,$ then the grey wolves attack the prey (exploitation) and when $|A| > 1$, then grey wolves get separated from each other (exploration). The values of the coefficients A and C are calculated as $= 2a\,.r_1 - a$, where a is linearly decreased from 2 to 0 during the iterations and $r_1$is a random number in the range of 0 to1. This process is continued until the maximum number of iterations is reached.

Speech Dataset

For the simulation estimation study, 150 noisy files from NOIZEUS[10] database (suburban train noise, babble, car, exhibition hall, restaurant, street, airport, and train-station noise at 0,5,10,15 dB SNR) and 50 noisy files from the SPEAR[11] database (factory, pink, volvo 340, bursting and white noise at 0 to 19 dB SNR level) are combined.

Audio Features selection using Neighborhood Component Feature Selection

For the SNR prediction of noisy speech, several audio features are used and the details are discussed in this section. To find out the appropriate features

Neighborhood Component Feature Selection method is used.

The NCFS[12] is the nearest neighbor-based feature weighting algorithm, which is used to calculate the feature weights for the specific classification task. It is based on the maximization of the expected leave-one-out classification accuracy. The audio features selected for classification[13,14] are used in this proposed estimation of SNR. The details of the six best audio features[13] and their corresponding feature weight are listed in Table 1.

Simulation and Results Discussion

The detailed steps of implementation of the proposed evaluation parameter (E-SNR) and the evaluation of the accuracy in prediction are discussed in this section.

Proposed Scheme of Implementation

Figure 2 depicts the block diagram of the proposed scheme of ESNR implementation. To accurately predict the SNR three stages are used: in the first step the audio features are identified which are related to the change in the change in noise level. In the second stage, the relationship between these features and the SNR levels are predicted using the ANN. In the third stage to increase the accuracy of prediction the adaptive algorithm of the ANN, the Grey Wolf Optimizer is employed for weights of ANN calculation.

At first, the audio features mentioned in Section 2.4 are calculated for the speech data corresponding to SNR values of 0 to 19 dB SNR levels. In this way, a set of input-output data is generated. Out of these input-output data sets, 80% of the data are used for training of the TFLANN model and the remaining 20% is used for validation purposes. The weights of the TFLANN (***W***) are updated using the GWO algorithm. Each of the training set is applied to the model and the estimated output is obtained. It is then compared with the corresponding target output obtained from the theoretical response at that particular instant to produce the error. Using the mean square error (MSE) produced during the training phase is computed and is plotted as a function of the number of iterations in Figure 3. After the MSE attains the minimum possible value, the training process is terminated. The final values of W(k) is obtained in the model are stored.

Simulation Results

The simulations are carried out using the MATLAB platform in Intel core i5 processor. The comparison of the actual and estimated SNRs using

Table 1 — Details of the audio features used in the proposed algorithm

Sl No	Audio Feature	Details	Feature Weight
1	*Spectral Centroid (f_1)*	It is expressed as the ratio between the frequency weighted sum of the power spectrum and its unweighted sum[14].	4.28
2	*Spectral Skewness (f_2)*	It gives the information on the symmetry of the distribution of the spectral magnitude values over the arithmetic mean[14].	5.67
3	*Spectral Spread (f_3)*	It provides information regarding the concentration of the power spectrum around the spectral centroid[14].	4.45
4	*Spectral Tonal Power Ratio (f_4)*	It is calculated as the ratio between the tonal power and the overall power[14].	2.97
5	*Entropy (f_5)*	The absolute value of entropy is expressed as $f_4(n) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{p(n)} \log 10\ (\mathrm{p(n)})$, where *p(n)* is the probability of occurrence of the *nth* sample value.	4.15
6	*Standard Deviation (f_6)*	Standard deviation is the calculation of the spreading of the input signal from the mean value.	5.99

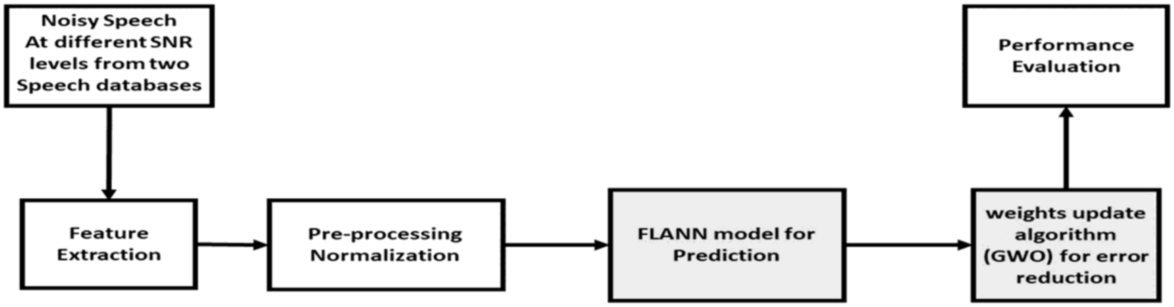
Noisy Speech At different SNR levels from two Speech databases
Feature Extraction
Pre-processing Normalization
FLANN model for Prediction
weights update algorithm (GWO) for error reduction
Performance Evaluation

Fig. 2 — Block Diagram of the proposed Implementation of ESNR Calculation

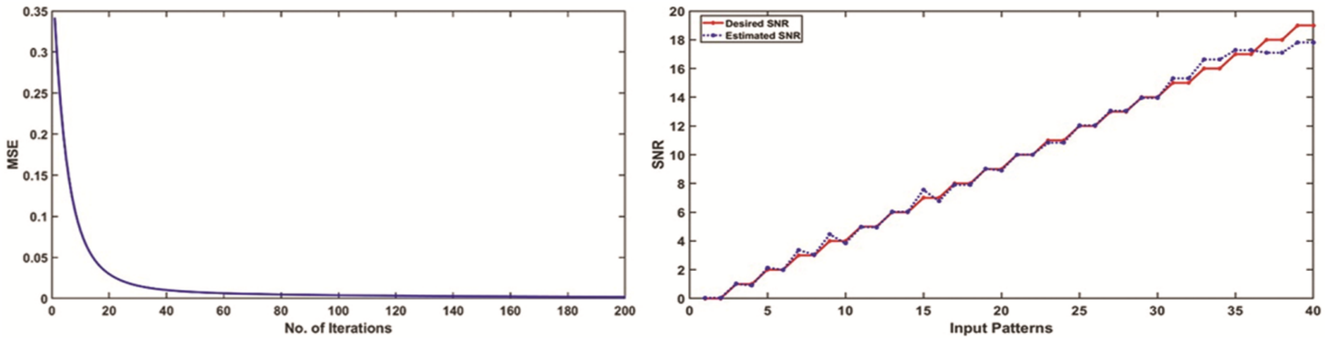

Fig. 3 — The Convergence characteristics obtained during the training of TFLANN model and comparison of actual and estimated SNR values for different noise levels using the TG method

Table 2 — Details of the MSE values calculated using different models for ESNR calculation

Sl No	Models for Prediction	MSE
1	TG	0.009
2	TL	0.03
3	PL	0.08
4	CL	0.055

the TFLANN with the GWO (TG) method is shown in Figure 3. The proposed model consists of mainly two parts the FLANN architecture and the update algorithm to find the appropriate weights of FLANN after convergence is achieved. The performance of the proposed model, TG is compared with those obtained by the other three models such as TFLANN with LMS (TL), Polynomial FLANN with LMS (PL) and Chebyshev FLANN with LMS (CL). The MSE obtained for all the four models are listed in Table 2. The results from Table-2 exhibit minimum MSE of the proposed method which thus demonstrates that the proposed estimation of SNR is a potential method.

Conclusion

In the area of speech processing field, noise level estimation and classification play important roles but still more investigations are required to yield improved performance. In this direction, this paper has presented a new speech quality evaluation parameter ESNR which can be used to better predict the SNR value of any unknown noisy speech signal without any reference of a clean speech signal. The accuracy of the prediction performance of ESNR is compared with the other three FLANN based prediction models using two standard speech databases. It is observed that the proposed TFLANN and GWO based ESNR prediction provides the best results with minimum error. In the future, the outcome of the proposed feature analysis can be combined with the speech enhancement algorithms and the performance can be assessed and compared with other standard methods.

References

1 Deepa D & Shanmugam A, Enhancement of noisy speech signal based on variance and modified gain function with PDE preprocessing technique for digital hearing aid, *J Sci Ind Res*, **70(5)** (2011) 332–337.

2 Dash T K & Solanki S S, Comparative study of speech enhancement algorithms and their effect on speech intelligibility, *Int Conf on IEEE* (2017) 270-276.

3 Loizou P C, *Speech Enhancement: Theory and Practice* (CRC press, Taylor & Francis Group, Boca Raton, FL) 2007.

4 Mermelstein P, Evaluation of a segmental SNR measure as an indicatorof the quality of ADPCM coded speech, *J Acoustical Soc Am*, **66** (1979) 1664–1667.

5 Ma J & Loizou P C, SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech, *Speech Comm*, **53** (2011) 340–354.

6 Majhi R, Panda G & Sahoo G, Development and performance evaluation of FLANN based model for forecasting of stock markets, *Expt Syst App*, **36** (2009) 6800–6808.

7 Panda S, Development of ANN Based Improved Model of Amplitude Response in Suppression State of Axonal Memory, *J Sci Ind Res*, **78(10)** (2019) 659-663.

8 Mirjalili S, Mirjalili S M & Lewis, Grey wolf optimizer, *Adv Engg Soft*, **69** (2014) 46-61.

9 Das D, Sadiq A S, Mirjalili S, & Noraziah A, Hybrid Clustering-GWO-NARX neural network technique in predicting stock price. *J Phy*, **892** (2017) p. 012018.

10 Loizou P, NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms. *Speech Comm*, **49** (2017) 588–601.

11 Wan E, Nelson A, & Peterson R, Speech enhancement assessment resource (SpEAR) database. *CSLU, Oregon Grad Inst Sci Tech* (2002).

12 Yang W, Wang K, & Zuo W, Neighborhood Component Feature Selection for High-Dimensional Data. *J of Comp*, **7(1)** (2012) 161–168.

13 Dash T K & Solanki S S, Investigation on the Effect of the Input Features in the Noise Level Classification of Noisy Speech, *J Sci Ind Res*, **78(12)** (2019) 868-872.

14 Lerch A, *An introduction to audio content analysis: Applications in signal processing and music informatics* (John Wiley & Sons, Inc., Hoboken, New Jersey) 2012.