ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 10, No. 1, January 2022, pp.71-13

Design of automatic speech recognition in noisy environments enhancement and modification

Ali Najdet Nasret[1], Abbas B. Noori[1], Arsen Ahmed Mohammed[2], Zuhair Shakor Mahmood[1]
[1] Electronic Department Kirkuk Technical Institute, Northern Technical University, Iraq
[2] Electrical department College of engineering. University of Kirkuk, Iraq

ABSTRACT

Recurrent neural networks (RNN) and feed-forward multi-layer perceptron's have been proposed for determining the absence and presence of speech in continuous voice signals when there is a variety of background noise levels present. The Aurora2 and Aurora3 were used to conduct detailed performance evaluations on vocal activity detection. When a Recurrent neural network feeds on automatic speech recognition particular features and acoustic features, the best outcomes can be achieved, according to this study. Aurora2 and the French, Romanian and Norway portions of the Aurora3 corpus is also proposed for detailed studies of ASR. When noise presence probability is utilized to change for encoding speech, phone subsequent probabilities are employed; the WER is reduced by 10.3 percent.

Keywords: RNN, Speech recognition, Acoustic features

Corresponding Author:

Ali Najdet Nasret
Electronic Department Kirkuk Technical Institute
Northern Technical University
Kirkuk, Iraq
alinajdet@ntu.edu.iq

1. Introduction

Large corpora with a wide range of speakers are used to train automatic speech recognition systems for multiple languages. Typically, they are collected in a controlled environment where there is little or no noise. When automatic speech recognition systems are operating in the presence of real-world noise, this often results in a decrease in performance. Erroneous word additions to the recognized sentence are possible when noise is present in non-speech signal segments, whereas Signal-to-Noise Ratio (SNR) regions Some phonemes may be masked., resulting in incorrect word removals. Recognition accuracy is often enhanced with the application of noise reduction techniques. These mistakes continue to be made despite the fact that they have been addressed. When training corpora in noisy environments, theoretically it would be possible to further minimize mistakes, but this is actually impractical. However, in fact, an accurate (VAD) Voice Activity Detector that is robust in noisy settings can be introduced to get some benefits. In order to determine the likelihood that speech would be heard, a variety of approaches have been proposed, including linear estimation and logistic regression [1]. In addition to HOS, log likelihood ratio, a priori speech absence probability is calculated using the smoothed power spectrum's minimum values., and Laplacian Gaussian model, other ways have been developed. [2] discusses features that can effectively identify speech in a variety of auditory settings. Each speech frame can be associated with an artificial neural network to determine the likelihood of speaking or silence in that frame. An artificial neural network was recently presented in [3] to do this. In acoustic Hidden Markov Models (HMM), scaled and utilized as a new observation, this likelihood. Noise Presence Probability in the absence of speech is also proposed in this study, but it has two important distinctions. The acoustic model's non-speech state's predicted strength informs a non-linear gain function that incorporates this likelihood. The non-voice subsequent probability calculated by the artificial neural network of the hybrid system can be altered using this function in

an ANN/HMM ASR hybrid system. Thus, Adjusting the prior probability of other phones does not change their relative values calculated by a model learned through discriminative learning. There is less uncertainty between speech and non-speech as a result of this. The neural VAD is the ANN that calculates NPP. Noise and signal-to-noise ratio are all taken into account while training it to distinguish voice from non-voice in various noise settings (SNR). ASR system and suggested VAD are described in Section 2. There are three sections: Section 3 details the ANN that was used to compute the likelihood of noise existence, and Section 4 reports findings achieved using Aurora2 as well as its French, Romanian and Norway components.

2. Use probability of noise presence

The design presented in Figure 1 is used to integrate the neural VAD with the hybrid systems. Hybrid HMM-NN ASR defined in [3] is the phonetic expert. With a huge multi-condition training set and a speech enhancement algorithm that has already reduced stationary noise, the Neural VAD is the Voice/Noise expert. It is specifically built and taught to distinguish between voice and noise. Consideration has been given to the use of a Recurrent Neural Network (ANN) and a Multi-Layer Perceptron (MLP). After removing the stationary noise, the J-RASTA, signal energy and Perceptual Linear Prediction Coefficients (PLPC) filtering are extracted from the speech signal. For Voice/Noise discrimination, (J-Rasta PLP) [4] some more characteristics difference between the voice signal energy, Spectral entropy and (MSE) are examples of these properties. P (C|Y) is calculated using the spectral parameters as well as their first derivative and second derivative, which the phonetic specialist employs. The Voice/Noise expert, which calculates NPP, receives the same parameters as before, along with the supplementary features. When the phonetic expert calculates P(C|Y), he uses this probability, along with an speculation of the inherent intensity of the background noise , to calculate gain of a non-linear that modulates P(C|Y).

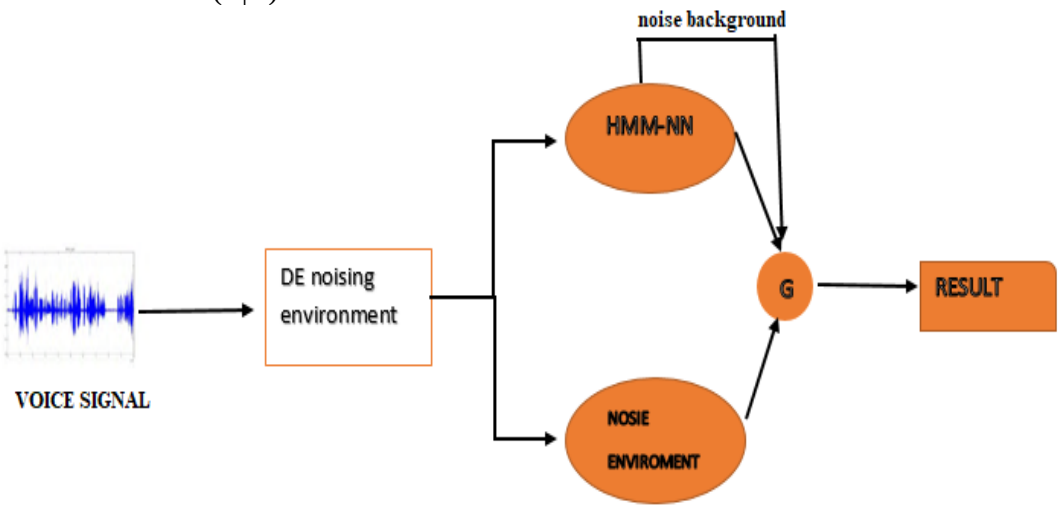

Figure 1. Proposed system design frame work

An output node in the ASR system's ANN offers an estimate of the background noise's posterior probability P (BGN|Y). Many insertion and deletion errors are caused by a miscalculation of $P(BGN|Y)$. Given the input signal Y, the neural Voice Activity Detector computes NPP = P (noise-only|Y). For the absence of speech, noise presence probability is used to instead of $P(BGN|Y)$ and generate a new subsequent estimation $\acute{P}(BGN|Y)$.the expression of $\acute{P}(BGN|Y)$ is written in the below

$$\acute{P}(BGN|Y) = G(NPP, NS) \cdot P(BGN|Y) \tag{1}$$

As an example, NS estimates the "intrinsic strength" of the artificial neural network O/P node that determine the subsequent probability $P(BGN|Y)$ in this case. The $G(NS \cdot Npp)$ function is defined as:

$$G(NS \cdot Npp) = \begin{cases} 1 + (\alpha_{max} - 1) \cdot [(N_s - 0.6) \cdot 2] \cdot (1 - Npp) & if \quad N_s \geq 0.6 \\ 1 - (1 - \alpha_{min})[(0.6 - N_s) \cdot 2] \cdot Npp & if \quad N_s < 0.6 \end{cases} \tag{2}$$

Where the gain function's α_{max} and α_{min} values are min and Max, respectively. For each ANN model, and even within a single language, the "intrinsic strength" of the BGN node varies based on the training material. It is possible to assess "intrinsic strength" in two method.

1- The instantaneous value of P (BGN|Y).
2- Average P(BGN|Y) when no voice activity is detected by the VAD

The following is an expression for the noisy speech model in the frequency domain:

$$Y(t,k)=D(t,k)+X(t,k) \quad (3)$$

Where D(t,k) represent noise signal ,X(t,k) represent speech signal and Y(t,k) represent noisy signal On the other side non-linear system such as the fourth root or square root is used to further compress this number in order to provide a more convenient rescaling of the dynamics. Differences between the two are inconsequential. In both notions of intrinsic strength. Gain function states that if the intensity of the background noise is intrinsically strong, Noise zones are less amplified compared to voice regions.; the opposite is true if the BGN is weak. Afterwards, all of the NN acoustic model's other outputs P (Ci | Y) are normalized to sum to one by modulating gain with P (BGN|Y) to obtain $\acute{P}(BGN|Y)$. Figure 2 shows the gain function's form as a function of NPP and an estimate of noise strength (NS). Because noise and poor intrinsic BGN acoustic model strength are both indicated by the VAD, this model's maximum gain (α_{max}) is reached with noise presence probability =1.0 and NS=0.0. To get the smallest gain (α_{min}), NS=1.0 and NPP=0.0 (i.e., when the Voice Activity Detector tells that there is inherent strength of the BGN model and voice is high) must be achieved. That is why when NPP=1.0 (i.e. Voice Activity Detector tell that there is noise but BGN acoustic model intrinsic strength are both at 1.0), P (BGN|Y) does not change at all. NPP and NS are both 0.0 when the Voice Activity Detector indicates that there is a voice, but there is already a low inherent strength to the BGN acoustic model.

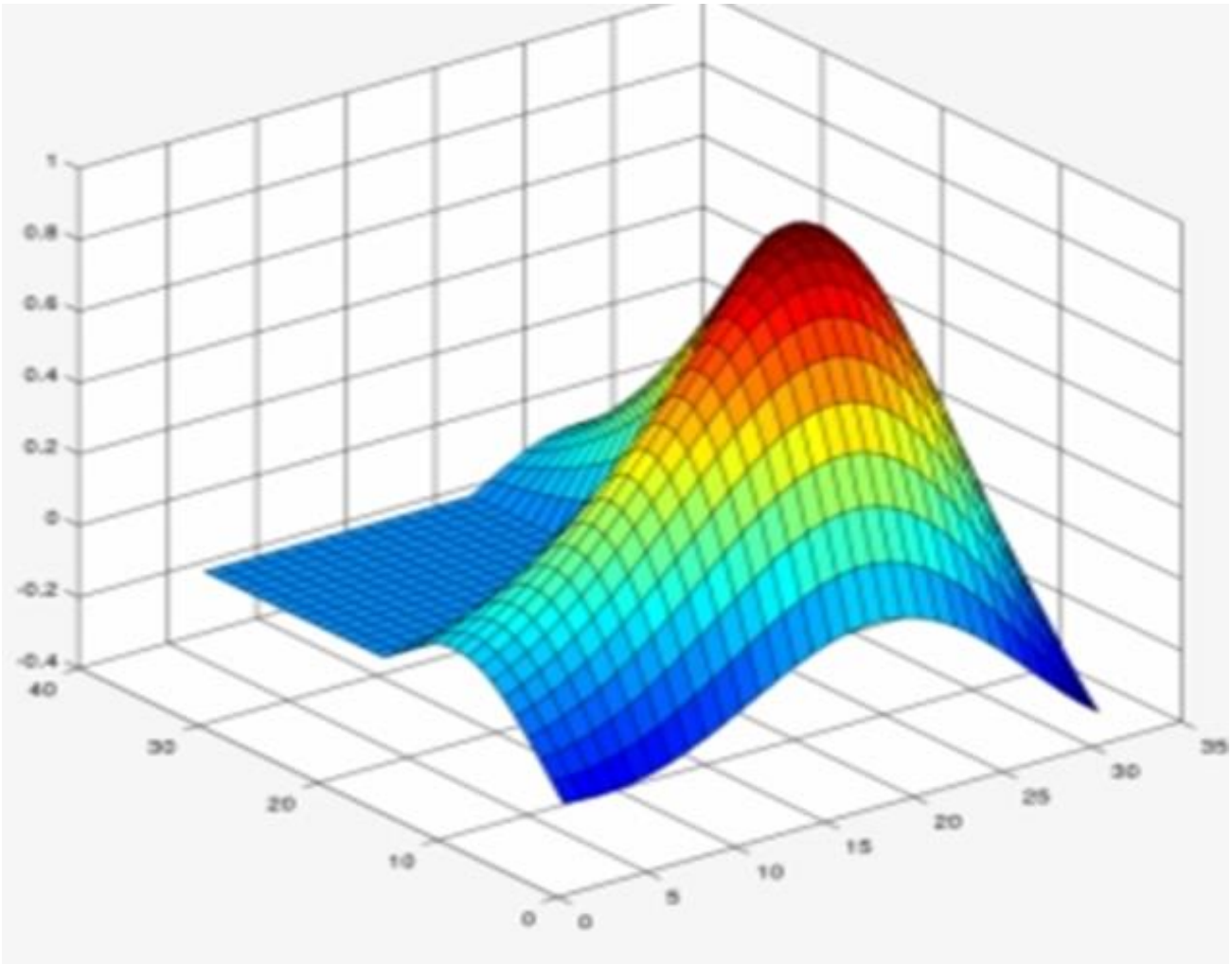

Figure 2. Represent the NS and NPP gain

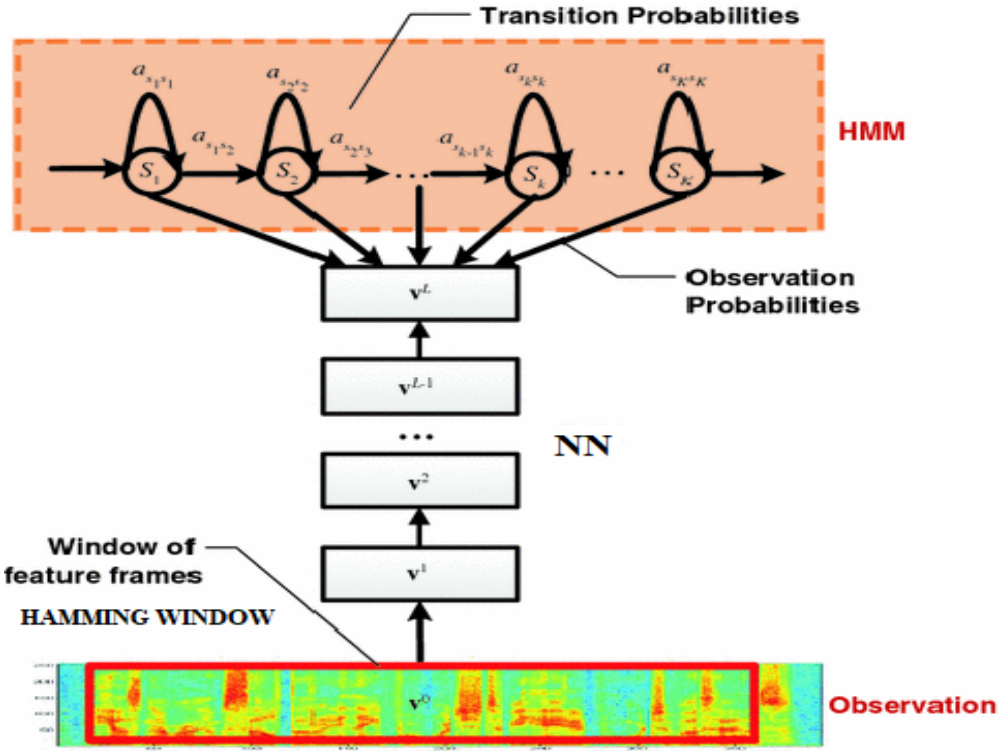

Figure 3. Hidden markov model based on neural network

3. Structures for the neural VAD

Presence of only background noise, a two-output MLP has been developed. Seven nearby frames feed the artificial neural network input; each frame comprises the 13 cepstral coefficients and total energy, as well as first derivative and third time derivative; and seven adjacent frames feed each frame. Other ancillary factors have been taken into consideration. Pitch trackers can extract periodicity and pitch, and other metrics such as signal noise ratio, the difference between (MSE) and current energy are available.

In the following, the term "AuxF" refers to a set of complementary features that are included. The first hidden layer is composed of 315 units, each of which is locally related to the properties of the concentration frame as well as the right and left contexts of the four frames under consideration. With 52 units, the second concealed layer is completely connected to the first. A multi-layer perceptron's and a Recurrent neural network with feed-back on the second layer nodes have both been implemented as neural Voice Activity Detector, with the MLP being the simpler of the two. Voice and Background Noise posterior probabilities are computed using two units in a soft ax layer, which is the result of the algorithm's output. It is necessary to train the neural VAD with a multi-style corpus, which contains many languages as well as different There are many distinct kinds of noise, as well as varying decibel levels. For the neural Voice Activity Detector's benefit, a noise-decreasing method [6] is applied during front-end processing to decrease stationary noise. Two classes of phonetic labels based on acoustic models and the conventional acoustic models, the forced segmentation of the training and test corpora was performed on the corpora (voice, noise). Several tests have been carried out in order to evaluate the performance of the neural Voice Activity Detector. Table I contains the results of the study.

Table 1. Aurora 3 French

VAD Type	Del	Correct	Ins
NNVAD	0.31	0.9	0.1
NVAD+AuxF	0.61	0.31	0.054
RNN -VAD+AuxF	0.34	0.78	0.0567
VAD NE Etsi	0.81	0.569	0.012

Table 2. Aurora 3 Romanian

VAD Type	Del	Correct	Ins
NNVAD	0.12	0.89	0.021
NVAD+AuxF	0.68	0.51	0.041
RNN -VAD+AuxF	0.24	0.71	0.04
VAD NE Etsi	0.91	0.8	0.026

Table 3. Aurora 3 Norway

VAD Type	Del	Correct	Ins
NNVAD	0.22	0.8	0.031
NVAD+AuxF	0.78	0.32	0.0222
RNN -VAD+AuxF	0.25	0.65	0.0178
VAD NE Etsi	0.81	0.71	0.0658

- Etsi NE VAD: this is the Voice Activity Detector that is used in the ETSI -AFE noise estimation module.
- NVAD: this is the neural Voice Activity Detector Spectral Attenuation noise reduction and based on J-Rasta PLP features;
- NVAD + AuxF: Auxiliary features have been introduced to this version;
- RNNVAD + AuxF: The neural Voice Activity Detector beats the energy-based ETSI Noise-Est Voice Activity Detector, according to the results of the experiment. The results reveal that, among the neural VAD variations, the auxiliary features (fundamental frequency, periodicity, signal noise ratio, entropy, and noise energy difference) often enhance the performance, as predicted by the researchers. Using an RNN, you can achieve even greater results.

4. Recognition results

Test sets from the Aurora2 and Aurora3 corpora's were used in ASR tests that were carried out in accordance with the architecture described in Section 2. The results, expressed as a percentage of WER, are presented in Tables II. The Phonetic Expert is a standard Loquendo hybrid Hidden Markov model –neural network that has been released for the French, Romanian, Norway, and untied state English languages, as well as for other languages. NNVAD is the Voice/Noise Expert in Table and it is the one with no auxiliary characteristics, as indicated by the name. Because it represents the best compromise between accuracy and computing complexity, it has been selected for the recognition studies. As a matter of fact, the fundamental frequency features increase the outcomes at the expense of approximately double the amount of time spent on the front-end computations. Table II contains row headers that relate to the following conditions:

- EM -SA: standard J-Rasta-PLP based on signal noise ratio dependent modified Malah -Ephraim- spectral attenuation [11];
- EM -SA-NPP-1: but with a gain function-modified phone posterior probability and an NS value calculated as the mean of background noise state probability NPP>0.6 (α_{min} = 1.6 and m=0.0);
- EM -SA-NPP-2: same as before, but with the phone's following probability being adjusted by the gain function, the SA-EM-NPP-2 model includes NS, which is computed as the square root of the instantaneous background noise state probability (α_{min} = 1.6 and m=0.0).

The confidence intervals for the word error rate are shown in parenthesis.

Table 4. Aurora 2 French

VAD Type	Del	Correct	Ins
NNVAD	0.87	0.89	0.036
NVAD+AuxF	0.7	0.63	0.0542

VAD Type	Del	Correct	Ins
RNN -VAD+AuxF	0.41	0.21	0.0457
VAD NE Etsi	0.81	0.421	0.0632

Table 5. Aurora 2 Romanian

VAD Type	Del	Correct	Ins
NNVAD	0.32	0.8	0.011
NVAD+AuxF	0.45	0.21	0.0222
RNN -VAD+AuxF	0.41	0.65	0.0145
VAD NE Etsi	0.452	0.545	0.0254

Table 6. Aurora 2 Norway

VAD Type	Del	Correct	Ins
NNVAD	0.22	0.8	0.011
NVAD+AuxF	0.58	0.21	0.0222
RNN -VAD+AuxF	0.25	0.32	0.0234
VAD NE Etsi	0.732	0.569	0.0654

Using NPP to re-modulate the posterior probability of background noise and phonemes has been found to be highly successful, especially when noise induces a significant insertion and deletion rate. The pre-plosive delay in geminate plosives is filled by noise in the Aurora3 French digits, which causes the word to be deleted or split into two wrong words. The best overall setup, EM -SA-NPP-2, yields the most improvement for Aurora3 Romanian (10.3% E.R.) However, when using the optimum setup, the improvement is consistent across Aurora2 and Aurora3 (7.8 percent). When the residual error is mostly caused by substitutions, the technique has limited effect because it is designed to focus on deletions and insertions only. This is the case in Norway with Aurora3 as an example. In order to make a direct comparison with the ETSI Noise Est Voice Activity Detector values.

5. Conclusions

Using a Phonetic expert and a Noise /voice expert to improve automatic speech recognition accuracy in loud contexts has been proposed in this study. When the outputs of the hidden markov model –neural network Hybrid system are modulated by results from a noise-resistant Neural VAD, the integration is achieved. Multiple test sets, both noisy (Aurora3 and Aurora2) and clean (Aurora3), have been used to evaluate various neural VAD architectures and features (TIMIT). With the standard ETSI NoiseEst VAD, the results are always inferior. Once noise has been reduced to an acceptable level, the neural VAD was utilized to re-modulate the background noise and phoneme likelihoods. Test instances in which this issue is more prevalent have yielded consistent decreases in errors.

6. References

[1] S. Jafarlou, S. Khorram, V. Kothapally and J. H. L. Hansen, "Analyzing Large Receptive Field Convolutional Networks for Distant Speech Recognition," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 252-259, doi: 10.1109/ASRU46091.2019.9003805.

[2] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding and K. Yu, "Adaptive Very Deep Convolutional Residual Network for Noise Robust Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 8, pp. 1393-1405, Aug. 2018, doi: 10.1109/TASLP.2018.2825432.

[3] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang and W. Liu, "Boosting Noise Robustness of Acoustic Model via Deep Adversarial Training," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5034-5038, doi: 10.1109/ICASSP.2018.8462093.

[4] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita and S. Araki, "Mask-based MVDR Beamformer for Noisy Multisource Environments: Introduction of Time-varying Spatial Covariance Model," ICASSP

2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6855-6859, doi: 10.1109/ICASSP.2019.8683092.

[5] T. Kang, S. Lee, M. Haghigat, D. Abramson and M. Flynn, "A 50μW 4-channel 83dBA-SNDR Speech Recognition Front-End with Adaptive Beamforming and Feature Extraction," 2021 IEEE Custom Integrated Circuits Conference (CICC), 2021, pp. 1-2, doi: 10.1109/CICC51472.2021.9431579.

[6] N. Ito, R. Ikeshita, H. Sawada and T. Nakatani, "A Joint Diagonalization Based Efficient Approach to Underdetermined Blind Audio Source Separation Using the Multichannel Wiener Filter," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1950-1965, 2021, doi: 10.1109/TASLP.2021.3079815.

[7] A. Ghazi, S. Aljunid, S. Z. S. Idrus, A. Fareed, A. Al-dawoodi, Z. Hasan, R. Endut, N. Ali, A. H. Mohsin, and S. S. Abdullah, "Hybrid Dy-NFIS & RLS equalization for ZCC code in optical-CDMA over multi-mode optical fiber," Periodicals of Engineering Natural Sciences, vol. 9, no. 1, pp. 253-276, 2021.

[8] P. A. Tapkir, M. R. Kamble, H. A. Patil and M. Madhavi, "Replay Spoof Detection using Power Function Based Features," 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018, pp. 1019-1023, doi: 10.23919/APSIPA.2018.8659582.

[9] Z. S. Mahmood, A. N. N. Coran, and A. Y. Aewayd, "The Impact of Relay Node Deployment In Vehicle Ad Hoc Network: Reachability Enhancement Approach," in 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-3: IEEE.

[10] B. Mohammed, R. Chisab, and H. Alrikabi, "Efficient RTS and CTS Mechanism Which Save Time and System Resources," international Journal of Interactive Mobile Technologies, vol. 14, no. 4, pp. 204-211, 2020.

[11] H. Salim, and N. A. Jasim, "Design and Implementation of Smart City Applications Based on the Internet of Things," International Journal of Interactive Mobile Technologies (iJIM), vol. 15, no. 13, pp. 4-15, 2021

[12] Z. S. Mahmood, A. N. N. Coran, A. esam Kamal, and A. B. Noori, "Dynamic Spectrum Sharing is the Best Way to Modify Spectrum Resources." 2021 Asian Conference on Innovation in Technology (ASIANCON), IEEE, pp. 1-5, 2021

[13] Z.S.Mahmood, A.N.Nasret, and O.T. Mahmood, “Separately excited DC motor speed using ANN neural network,” In AIP Conference Proceedings , vol. 2404, no. 1, p. 080012,2021

[14] A. Noori, A. Kamal, S .Mohammed,A .Humada, “Design and Implementation of Biquad Filters Using CMOS Circuit Based Active Elements,” International Review of Electrical Engineering (IREE), 14 (2), pp. 141-147, (2019) .doi:https://doi.org/10.15866/iree.v14i2.16373

[15] A. Kamal, A. Nasret, Z. Mahmood, “Design of Multiband Slot Patch Antennas for Modern Wireless Applications,” *International Journal on Communications Antenna and Propagation (IRECAP)*,10(5), pp.353-359, (2020). doi:https://doi.org/10.15866/irecap.v10i5.19071

[16] M. Hassan, A. Nasret, M. Baker and Z. Mahmood, " Enhancement automatic speech recognition by deep neural networks," Periodicals of Engineering Natural Sciences, vol. 9, no. 4, pp. 921-927, 2021.