Journal of Theoretical and Applied Information Technology
29th February 2024. Vol.102. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NOVEL FEDERATED-LEARNING BASED ADVERSARIAL FRAMEWORK FOR AUDIO-VISUAL SPEECH ENHANCEMENT

MOHAMMED AMIN ALMAIAH[1, 2, 3], AITIZAZ ALI[4], RIMA SHISHAKLY[5], TAYSEER ALKHDOUR[6], ABDALWALI LUTFI[7], MAHMAOD ALRAWAD[7]

[1]Department of Computer Science, Aqaba University of Technology, Aqaba 11947, Jordan

[2]Applied Science Research Center, Applied Science Private University, Amman 11931, Jordan

[3]King Abdullah the II IT School, the University of Jordan, Amman 11942, Jordan

[4]School of IT, UNITAR International University, Malaysia

[5]College of Business Administration, Ajman University, Ajman 346, United Arab Emirates

[6]College of Computer Science and Information Technology, King Faisal University, Al-Ahsa 31982, Saudi Arabia

[7]College of Business, King Faisal University, Al-Ahsa, 31982, Saudi Arabia.

Corresponding author: Dr. Tayseer Alkhdour, talkhdour@kfu.edu.sa

ABSTRACT

Current speech enhancement (SE) techniques operate in the spectral domain, utilizing either edge computing or the cloud. Most existing frameworks offer solutions for a limited number of noise conditions and rely on first-order statistics. To address these limitations, researchers have explored machine learning approaches to learn complex functions and train large datasets. However, these models typically rely on centralized servers like the cloud, which raises security concerns. Furthermore, running such training models on edge devices is challenging due to their limited battery power and privacy issues. In this study, we propose a federated learning-based SE framework for multiple clients, using two speakers, to overcome these challenges. Our proposed framework offers a decentralized model that allows for both local and global training of data. Moreover, it is well-suited for adversarial networks and private clinics as it preserves privacy on edge devices and in the cloud, facilitating SE in a distributed fashion. The proposed model enables multiple clients to train their data independently and send the aggregated training model to the cloud. In contrast to existing approaches, our method operates at the waveform level, training the model end-to-end and incorporating two speakers with different noise conditions into a single model. This allows for sharing model parameters with multiple clients using federated learning. Our approach provides improved security, speed, and reduced battery usage for various clients using hearing aids, resulting in enhanced robustness and other speech-centric design choices to improve speech quality securely.

Keywords: *Speech Enhancement; Federated Learning; Cloud computing; Deep learning AV dataset; SDG.*

1. INTRODUCTION

Speech enhancement (SE) methods are gaining more importance due to the ap- plication of hearing aids and the issue of hearing loss in human beings. In this research, we propose a novel federated learning-based speech separation model that supports single and multiple speech channels. The proposed model utilizes two approaches: first, it employs a pre-trained model on the central server, and then it distributes a locally trained model to the client side using the federated learning approach. This proposed approach provides privacy preservation, secu-

rity for private data, and cost-effectiveness in terms of hearing aid or edge device battery life.

The Minimum Variance Distortion Less Response (MVDR) beamformer can be utilized for voice augmentation when there is only one intended speech source. An essential quantity required for computing the MVDR beamformer is a vector of acoustic transfer functions from the desired voice source to all of the micro- phones. Instead of using the acoustic transfer function vector directly, Relative Transfer Functions (RTF) are employed, which are normalized with respect to a reference microphone. By making assumptions about the properties of the microphones, their positions, the placement of the speaker, and the acoustics of the room, the RTF vector can be calculated for a Large Microphone Array (LMA). In assistive hearing devices, it is often assumed that the position of the desired speech source is known, and this knowledge can be utilized to construct an RTF vector [1].

Neural network designs that can accurately distinguish between distinct sound sources in sound mixtures have recently been developed thanks to recent developments in deep learning. Music separation [2], speech [3, 4], and speech augmentation [5, 6] can all be achieved using discriminative separation models with supervised training. A clean source waveform for supervised training in several areas can be cumbersome or perhaps impossible to acquire. The use of contrastive learning [7] or weak labels including sound-class information [2, 8] have been other less supervised approaches. There have also been good findings with unsupervised techniques using spatial information from numerous micro- phones [9, 6] and visual cues [7]. Synthetic mixtures of mixtures have showed tremendous promise for single-channel sound separation and voice improvement through the use of MixIT [10]. Despite the fact that these works reduce the re- liance on supervised data, they nevertheless necessitate the availability of large consolidated audio data collections for IID training on a single device.

A distributed and privacy-preserving architecture called Federated Learning (FL) proposed in this research allows each client to train on their own data and then send updates to a central server. Each communication round, the central server gathers and distributes a new model based on those updates. By contin- uing this method, one can train a global model without violating the privacy of clients even if the data are not supplied IID to the clients. Audio models for keyword spotting, automatic speech recognition, and sound event detection have also been trained using FL. When several clients are present or the IID assumption is violated, supervised data on the client side is required, and these techniques, as well as most FL configurations, are less successful. FL algorithms that require less supervision have been developed as a result of this issue [11–13].

Single-channel speech augmentation has been studied for decades, but it is still a challenging problem to solve in many systems, such as automatic speech recognition (ASR), hearing aids, and hands-free mobile communications. Algorithms that have been around for a long time use statistical methods such as noise spectrum estimation and speech signal estimation. These procedures de- pend on an accurate estimation of the noise level. The statistical assumption that speech and noise are one and the same leads to considerable amounts of musical noise artifacts in most conventional techniques, which are unable to detect non-stationary noise. Recently, various data-driven systems have circumvented the notion that speech and noise processes have unique distortions. An ASR system successfully uses DNNs for acoustic modeling, which is now being used for speech improvement, is the basis for this technique. CASA provides an ideal binary mask (IBM) or ideal ratio mask (IRM) that may be calculated using DNN in the frequency domain for monoaural speech separation. In spite of the fact that these strategies improve the amplitude range of speech, they leave the phase spectrum unaltered. Additive noise degrades the clarity and intelligibility of speech, and SE seeks to remedy this [2]. Its primary use is to enhance the quality of mobile phone calls in loud surroundings. There are key applications connected to hearing aids and cochlear implants, where improving the signal before amplifying it considerably reduces discomfort and improves intelligibility [3]. In

voice recognition and speaker identification systems, speech augmentation has been effectively

The existing approaches consist of spectral subtraction [7, 2], Wiener filtering [2], statistical model-based techniques [8], and subspace algorithms [6]. In the literature mostly NN model is used to improve the quality of the speech. Moreover, another example of SE which is called denoising auto-encoder are consider prominent techniques for SE. However, RNNs are also utilised. Incorporating temporal context into the recurrent denoising auto encoder has increased its performance. In some of the existing model LSTM are also used with SVM for training the model and keeping the previous record in order to update the model with more information. The contributions of this research are as follows:

1. A novel federated learning-based SE model is proposed for multiple clients in an adversarial network.

2. The proposed model achieves low end-to-end latency by fast training on audio-visual data using the federated learning approach.

3. The proposed approach allows multiple clients to train their private data using a local training model and aggregates the models in the cloud after receiving them from all the clients for SE.

4. The proposed approach enables the model to learn from multiple speakers and noise types and aggregate them in the cloud, resulting in a more generalized and smarter system compared to the benchmark model."

2. BACKGROUND

In this section, we discuss the related studies on SE and privacy preservation approaches based on federated learning. It is feasible to establish a secure region on the main processor that ensures high confidentiality and integrity for any stored or processed data or code. Trusted Execution Environments (TEEs) achieve robust isolation and attestation of secure compartments by enforcing a dual-world view, where even compromised or malicious system software (also known as the Rich Operating System Execution Environment or REE) cannot gain access to the safe world. An alternative to training the entire Deep Neural Network (DNN) model from start to finish is layer-by-layer training or greedy layer-wise training. Furthermore, another model proposed in reference [10] is based on federated learning for privacy. The primary objective of this model is to reduce the disclosure of private information from the input data. To achieve this, the model uses the updated gradients of the model trained with additional data to train a binary classifier on both the server and the client side of the attack, which is also referred to as a side-channel attack. It is easier to carry property information, which refers to the feature information of the input data, in a stronger aggregate. Even when clients in federated learning only receive several snapshots of the broadcast global models that have been linearly aggregated from participating clients' updates, it is still possible to preserve property information effectively.

2.1 Federated Learning (FL)

Federated Learning (FL) is a machine learning approach that allows a model to be trained across multiple decentralized devices or servers holding local data samples without exchanging them. The model is trained locally on each device, and only the model updates (not the raw data) are sent to a central server, which aggregates these updates to improve the global model. Adversarial techniques, on the other hand, typically involve introducing adversarial examples or perturbations to the input data to deceive or mislead a machine learning model. This is often used in the context of improving the robustness of models against attacks. It refers to the idea of defending federated learning models against adversarial attacks. Adversarial training within a federated learning framework might involve training models on data that has been deliberately manipulated to deceive the model, helping it become more robust to such attacks.

Federated Learning-based Adversarial techniques can offer several advantages in the context of machine learning:

(1) Improved Robustness: Training models using adversarial examples within a federated learning framework can enhance the robustness

of the model against adversarial attacks. This means the model becomes more resistant to intentional manipulations of input data designed to mislead the model.

(2) Privacy Preservation:

Federated Learning itself is known for its privacy-preserving characteristics. By training models locally on devices without exchanging raw data, user privacy is maintained. Combining this with adversarial techniques can further ensure that even in the presence of adversarial attacks, sensitive information is protected.

(3) Decentralized Security:

Since federated learning involves training models on decentralized devices, the security of the overall system becomes more resilient. Adversarial techniques can be used to defend against attacks at the local level, and the federated learning aggregation process helps in learning from a diverse set of local models.

(4) Adaptability to Dynamic Environments:

Federated Learning allows models to be trained on data from various sources, making it adaptable to dynamic and evolving environments. Adversarial training can help the model adapt to changes and uncertainties by making it more robust to unexpected variations in the input data.

(5) Collaborative Learning without Centralized Data:

Federated Learning enables collaborative model training without the need for centralizing sensitive data. Adversarial techniques add an extra layer of defense against malicious attempts to compromise the learning process, ensuring the collaboration remains secure and privacy is maintained.

(6) Transferability of Knowledge:

Adversarial training in federated learning can improve the generalization of models across different local datasets. This transferability of knowledge helps create models that perform well on various data distributions, making them more versatile.

2.2 Audio-Visual Speech

Audio-visual speech refers to the integration of both auditory (sound) and visual (image or video) information in the context of speech processing. This multidimensional approach takes advantage of both the audio and visual cues present in speech signals to enhance various applications, including speech recognition, speaker identification, and human-computer interaction.

Audio-visual speech processing faces several challenges, spanning both technical and practical aspects. Some of the main challenges include:

(1) Synchronization:

Aligning audio and visual streams accurately is crucial for effective audio-visual speech processing. Variations in the delays between audio and video signals can impact the quality of synchronization and affect the overall performance of the system.

(2) Variability in Speech Production:

Individuals exhibit significant variability in their speech production, including differences in lip movements, facial expressions, and pronunciation. Developing models that can generalize across diverse speakers and speaking styles is a challenge.

(3) Lip Synchronization and Speech Dynamics:

Capturing accurate lip synchronization with speech dynamics is challenging due to the rapid and complex nature of speech movements. The variability in lip shapes during different phonemes and the influence of coarticulation make this a challenging aspect of audio-visual speech processing.

(4) Noisy Environments:

In real-world scenarios, audio signals can be affected by various types of noise. Integrating visual information becomes crucial for improving robustness in noisy environments. However, the effectiveness of visual cues may be limited in extremely challenging acoustic conditions.

(5) Limited Availability of Datasets:

Annotated datasets for audio-visual speech processing are often limited in size and diversity. Training models that can generalize well across various speakers, languages, and environmental conditions requires access to comprehensive datasets, which may be scarce.

(6) Cross-Modal Variability:

Variability between the audio and visual modalities, such as differences in pronunciation and facial expressions, poses challenges for aligning and integrating information from both sources effectively. Developing models that can handle cross-modal variability is a key research area.

(7) Real-Time Processing:

Achieving real-time audio-visual speech processing is challenging, particularly when dealing with complex models or resource-intensive algorithms. Applications like live video conferencing and human-computer interaction demand low-latency systems.

(8) Ethical Considerations:

There are ethical considerations related to privacy when dealing with visual data, particularly in applications that involve capturing and processing facial expressions. Ensuring responsible and privacy-preserving approaches is crucial in the development of audio-visual speech systems.

(9) Multilingual and Cross-Cultural Challenges:

Variations in speech across languages and cultures add complexity to the development of audio-visual speech processing systems. Models need to be robust enough to handle diverse linguistic and cultural contexts.

By addressing these challenges requires a multidisciplinary approach that combines expertise in signal processing, computer vision, machine learning, and cognitive science. Thus, this research aims to improve the accuracy, robustness, and efficiency of audio-visual speech processing systems in various applications.

3. METHODOLOGY

The proposed methodology consists of server-side training and client-side training using a transfer function. Furthermore, we utilized the LSTM approach to locally train the data and SVM to classify both the speaker and the background noise. The proposed model takes an image or speech as input, which is accompanied by background noise mixed with the source. A denoising technique, as described in Figure 1, is applied.

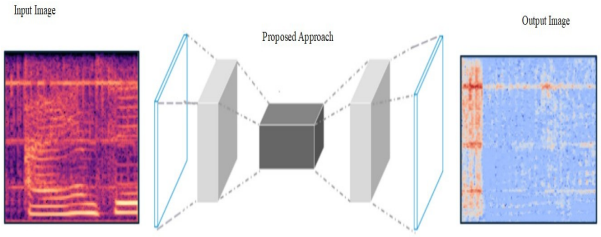


Figure 1. Proposed Framework For Speech Enhancement

Fig.2 illustrates the process of noise cancellation during the experimental analysis. The input spectrogram is fed into the model, and the predicted noise model is subtracted from it, resulting in a denoised spectrogram. The output signal is then obtained, free from noise. The proposed approach utilizes the spectrogram subtraction method by employing the predicted noise model.

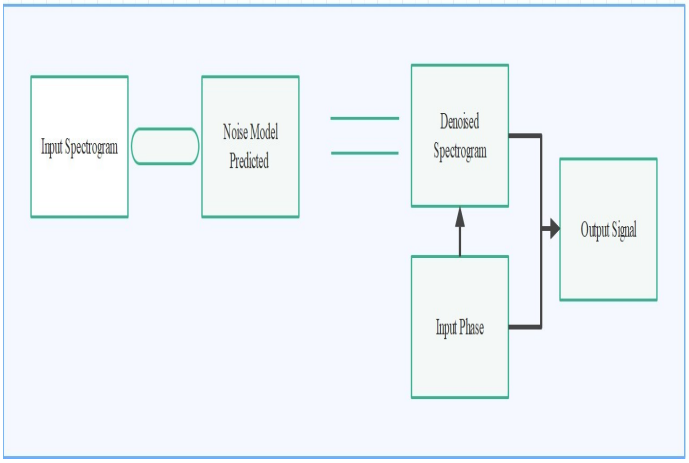


Figure 2. Proposed Framework For Speech Enhancement

In Figure 3 we have discussed the working of the proposed FL approach. It can be seen from the framework consist of edge layer and cloud layer where the global model is pre-trained in the cloud and the local model is deployed over multiple client on edge devises or hearing aid. The local client data contain noisy data and hence the local model is used to train the client data using SE technique based on deep learning. The central server in this case consider as an organization who provide speech services or hearing aid to the clients. The central server using the cloud layer receive updated model from multiple clients and aggregate it. Using the aggregated model the server provide better SE using the updated model.

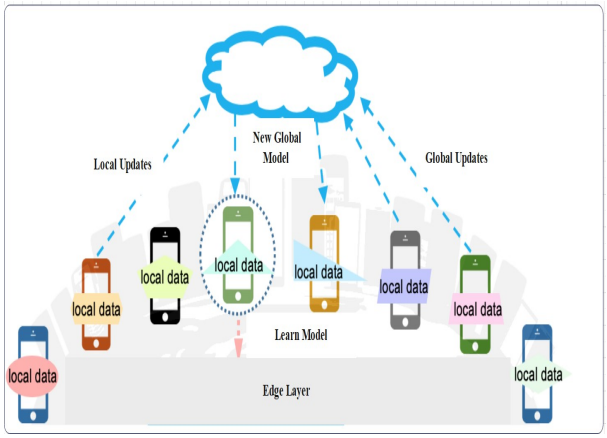


Figure 3. Overview Of Federated Learning For Speech Enhancement.

3.1 Server Side Communication

The server in the proposed approach holds the model weights 0(rg). The model weight is distributed to each and every client in each round r of the training. The server shares the pre-trained model with each client using Federated Learning (FL). The client utilizes an edge node to train the local data, specifically speech data with noise. A network separation parameter is distributed among all the network clients using the function f(;0(r)g). Additionally, each client is assigned a time slot that distinguishes it from other clients in the time domain, using T samples for r rounds. Therefore, during the update phase, a client c sends the updated individual model f(;0(r)g) to the server. Finally, the global model is updated as 0(r+1)g. Moreover, this model is redeployed to the client using FL for up to r rounds.

3.2 Client side Communication

In this study, we aim to elucidate how individual clients utilize their own pri- vate data for training voice improvement independently. For the client, we assume that it only has access to its private data Dc, which comprises two com- ponents, namely Dc = (Dmc, Dnc). Dnc exclusively consists of pure noise, while Dmc contains a mixture of speech and noise. As explained in [10], each client gen- erates a MoM (MoM) x = s+1+1+1+n2 by employing a noisy speech example m=s+n1+Dmc and a clean noise recording n2Dnc. The separation model consistently estimates M=3 sources. Considering that each client has access to clean noise recordings n2Dnc, there are two distinct scenarios for Dmc, corresponding to each client's private data.

4. EXPERIMENTAL SETUP

To evaluate the performance of our approach, we utilized a publicly available dataset including WHAM [14] and LibriFSD50K [15] as the following:

(1) WHAM: (Waveform-based Speech Activity Detection and Enhancement) is a system designed for speech activity detection and enhancement in noisy and overlapping speech scenarios. It aims to improve the performance of automatic speech recognition (ASR) systems by addressing the challenges posed by back- ground noise and multiple speakers. In WHAM, the waveform of the audio signal is utilized for speech activity detection. It employs a deep learning architecture, typically based on convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to classify whether a given time frame contains speech or non-speech. By accurately identifying speech segments, WHAM can effectively separate and enhance the desired speech signal.

(2) LibriFSD50K: LibriFSD50K (LibriSpeech-Farfield Speaker Detection 50K) is a dataset

specifically created for far-field speaker detection tasks. Speaker detection aims to identify and distinguish different speakers in an audio recording. Far-field speaker detection refers to scenarios where the microphone is placed at a significant distance from the speakers, leading to challenges such as back-ground noise, reverberation, and reduced signal quality. LibriFSD50K is based on the larger LibriSpeech dataset, which consists of approximately 1,000 hours of read English speech data from audiobooks. LibriFSD50K focuses on creating a subset of LibriSpeech with simulated far-field conditions, including artificially added noise and reverberation, to emulate real-world far-field scenarios. It contains around 50,000 utterances from over 3,700 speakers.

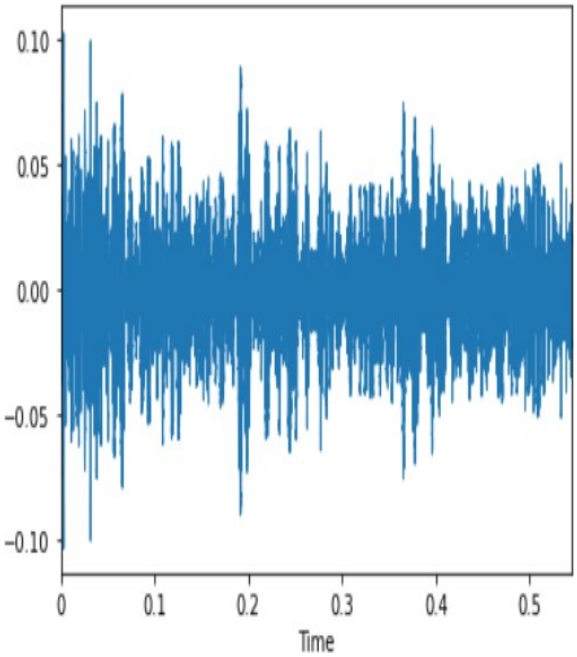


Figure 4. The Original Speech Signal.

Additionally, we incorporated various types of noises, such as cafe noise, in- sect noise, and bus noise, recorded from different speakers. The dataset was divided into training, testing, and validation sets, with a ratio of 70:20:10, respectively. For the creation of the noisy training set, we considered a total of 5 different environmental conditions [16]. Each condition included 4 signal-to-noise ratios (SNR) - 15 dB, 10 dB, 5 dB, and 0 dB, which were used to evaluate the performance of the proposed model. The model was trained using different sentences spoken by each training speaker under each condition. To construct the test set, we utilized 20 different conditions. These conditions involved mixing 5 types of noise with the original speech and applying 4 SNR levels (18 dB, 12 dB, 8 dB, and 2 dB) to each mixture. The model underwent training for 10 epochs, with the learning rate observed up to 0.0001. An effective batch size of 20 was used during the training process.

5. Analysis and Results

In this section, we present the simulation results obtained from the experiment and methodology implementations. The experiment was conducted using PyTorch and speech libraries. Figure 4 represent the original speech signal received from the speaker.

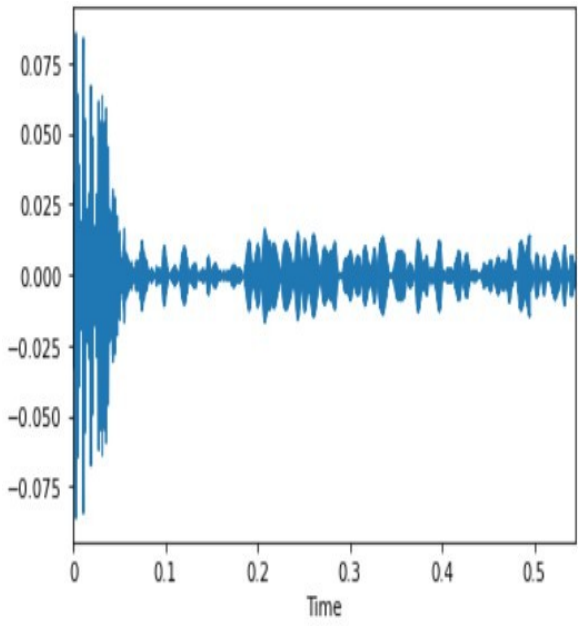


Figure 5. The Noise Is Predicted Using Time Series Fourier Transform.

The length of the signal is o.5 sec Fig .4 represent the background noise, in this case we consider the noise generated from the cafe, bus and insect.

In Figure 5 the noise model is predicted using time series Fourier transform. Figure. 6 represent

Journal of Theoretical and Applied Information Technology
29th February 2024. Vol.102. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the speech signal obtained with the noise detection using federated learning approach.

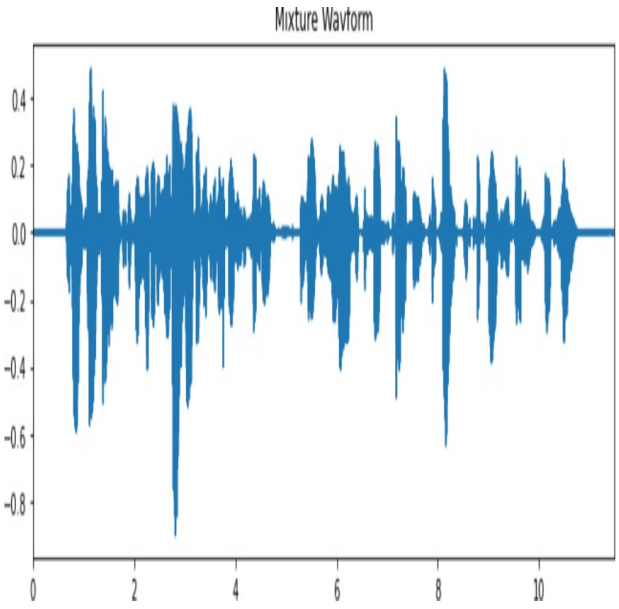
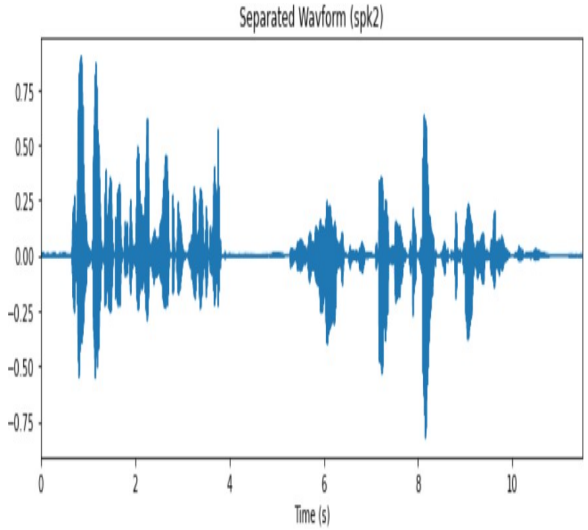


Figure 6. The Noise Detection Using Federated Learning Approach

In Figure 7 we the proposed model obtained the noise signal spectrogram model which later is used in the denoising process. Figure 7 shows obtained the noise signal spectrogram model



which later is used in the denoising process.

Figure 7. The Noise Signal Spectrogram Model Which Later Is Used In The Denoising Process.

Figure 8 represent the histogram of mixture speech based on the original signal with background noise in DB. Moreover, Figure 9 represent the enhanced speech signal spectrogram obtained through IMR.

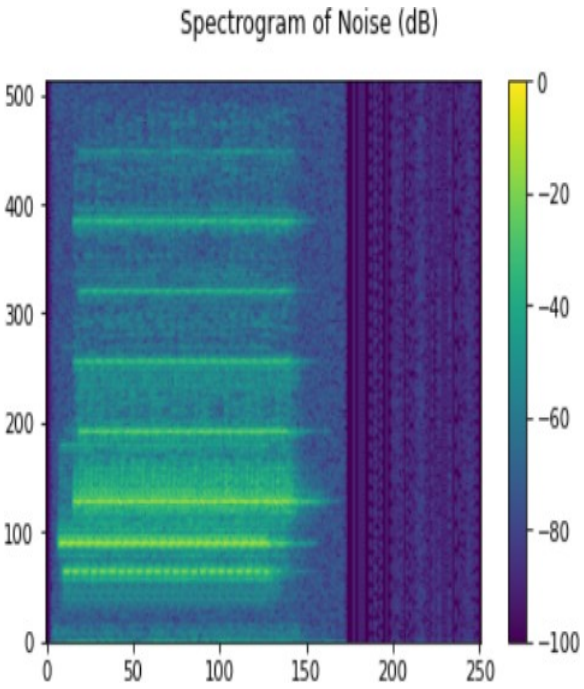


Figure 8. Spectrogram Representing Noise.

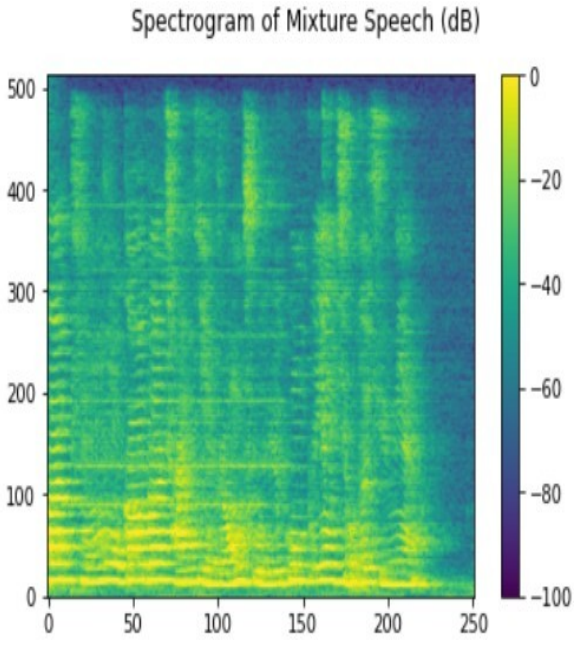


Figure 9. The Enhanced Speech Signal Spectrogram Obtained Through IMR

Journal of Theoretical and Applied Information Technology
29th February 2024. Vol.102. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 10 represents the spectrogram of the clean signal obtained after denoising the spectrogram model from predicted model.

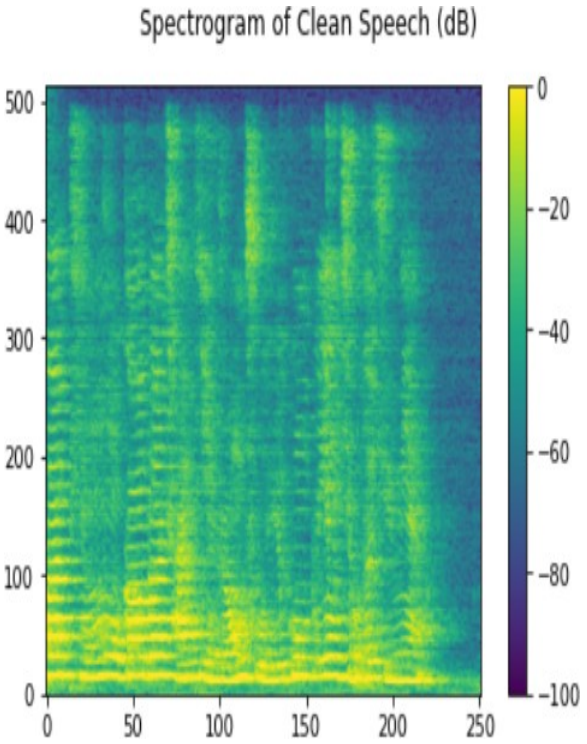


Figure 10. The spectrogram of the clean signal obtained after denoising the spectrogram model from predicted model.

Furthermore, Figure11 represent the IRM spectrogram for individual client speech which is the target speech of the framework.

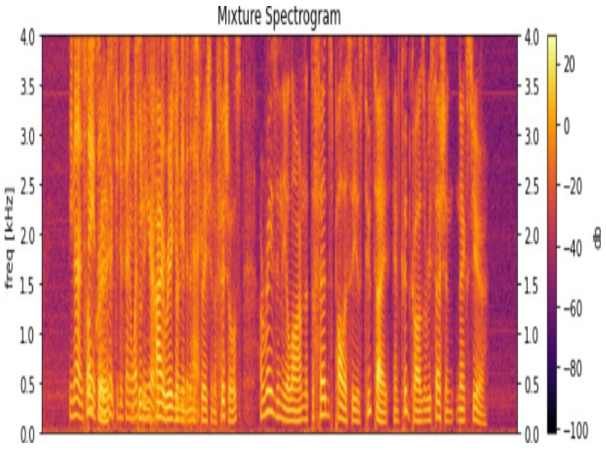


Figure 11. The IRM spectrogram for individual client speech.

In Figure 12 we have carried out comparative analysis of the proposed framework versus the centralized and traditional decentralized framework. It can be observed from the simulation results that for the same number of epochs the proposed model achieve more accuracy as compared to the centralized and traditional decentralized model.

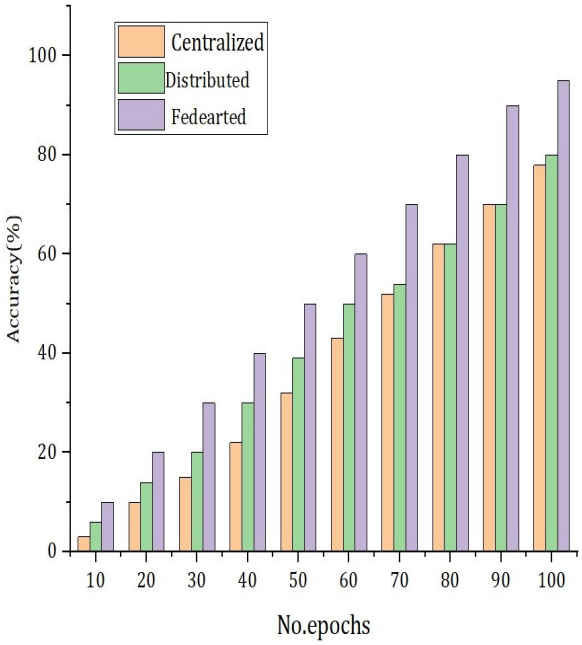


Figure 12. Comparative Analysis of the proposed model versus the centralized and Distributed Approach.

To evaluate the security resistance feature of the proposed model, we conducted a threat model analysis. The proposed federated learning (FL) model involves multiple client devices training a locally initialized deep neural network (DNN) and transmitting their local model parameters to a centralized server. The server then aggregates these parameters to create a global model. The primary objective of adversaries is to exploit vulnerabilities and capture sensitive information during the transaction process. We consider two types of attacks: active and passive. The active attack refers to actions initiated by the owner of the central server who has access to the updated model. On the other hand, the passive attack, also known as a side channel attack, originates from the clients' side. Assuming the server owner is trustworthy and possesses access to the updated global model, we encounter certain

challenges. If a layer does not fit within the available Trusted Execution Environments (TEEs), the network design must be altered either by reducing the number of layers or adjusting the training batch size accordingly. To establish a secure connection between the server TEE and each client device TEE, we employ a slightly modified version of the key ex- change protocol or an authenticated variant of Transport Layer Security (TLS). Another critical assumption is that data will be transmitted to and received from the TEE by the central server. Failure to adhere to this assumption by a malevolent server would compromise the system's functionality but not the security and privacy properties of our solution. It is worth noting that the conventional TEE threat model does not classify this type of Denial-of-Service (DoS) attack as a significant threat.

6. CONCLUSION

In this study, we have implemented an end-to-end federated learning-based method for SE within an adversarial framework. The model utilizes a hybrid deep learning approach, combining LSTM with SVM, and is deployed in both the local and global domains. The global model refers to the cloud, while the local model trains the speech data on the edge node. The global domain operates as an encoder-decoder fully-convolutional structure, enabling fast denoising of wave- form chunks. The results demonstrate the robustness of the proposed approach, which is also applicable to multiple clients. Potential future work includes exploring better convolutional structures and integrating adversarial techniques with lightweight learning to reduce end-to-end latency. Additionally, the proposed model can be extended to handle more complex noisy environments, including dynamic noise. Further experiments are required to compare the performance of the proposed approach with other benchmark models. In the future, we plan to integrate the proposed approach with different SE methods such as MIMO and other masking techniques, aiming to improve the model's quality and latency even further.

ACKNOWLEDGMENT

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Project No. Grant No. 5590).

REFERENCES

[1]. I. Dokmani´c, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," Proceedings of the National Academy of Sciences, vol. 110, no. 30, pp. 12 186–12 191, 2013.

[2]. F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: Learn- ing to separate sounds with weak supervision," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2386–2399, 2020.

[3]. Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, "Speech enhancement with weakly labelled data from audioset," arXiv preprint arXiv:2102.09971, 2021.

[4]. L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," Computers & Industrial Engineering, vol. 149, p. 106854, 2020.

[5]. Z. Hong, J. Wang, X. Qu, J. Liu, C. Zhao, and J. Xiao, "Federated learning with dynamic transformer for text to speech," arXiv preprint arXiv:2107.08795, 2021.

[6]. F. Granqvist, M. Seigel, R. Van Dalen, A. Cahill, S. Shum, and M. Paulik, "Im- proving on-device speaker verification using federated learning with privacy," arXiv preprint arXiv:2008.02651, 2020.

[7]. D. Guliani, F. Beaufays, and G. Motta, "Training speech recognition models with federated learning: A quality/cost framework," in ICASSP 2021-2021 IEEE Inter- national Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 3080–3084.

[8]. J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise compo- nents in the wavelet domain," in 1997 IEEE international conference on acoustics, speech, and signal processing, vol. 2. IEEE, 1997, pp. 1323–1326.

[9]. F. Lai, Y. Dai, S. Singapuram, J. Liu, X. Zhu, H. Madhyastha, and M. Chowd- hury, "Fedscale: Benchmarking model and system performance of federated learn- ing at scale," in International Conference on

Machine Learning. PMLR, 2022, pp. 11 814–11 827.

[10]. L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in 2019 IEEE symposium on security and privacy (SP). IEEE, 2019, pp. 691–706.

[11]. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE signal processing magazine, vol. 37, no. 3, pp. 50–60, 2020.

[12]. J. Koneˇcny`, H. B. McMahan, F. X. Yu, P. Richt´arik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[13]. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–19, 2019.

[14]. G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," arXiv preprint arXiv:1907.01160, 2019.

[15]. P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794, 2021.

[16]. 16. M. Shaheen, M. S. Farooq, T. Umer, and B.-S. Kim, "Applications of federated learning; taxonomy, challenges, and research trends," Electronics, vol. 11, no. 4, p. 670, 2022.