Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 343-351
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5501

Human recognition by utilizing voice recognition and visual recognition

Sukaina Sh Altyar[a,*], Samera Shams Hussein[a], Mahir Jasem Mohammed[b]

[a]Department of Computer Science, College of Education for Pure Science University of Baghdad, Baghdad, Iraq
[b] Department of Religious Education, Iraqi Sunni Affairs, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

Audio-visual detection and recognition system is thought to become the most promising methods for many applications includes surveillance, speech recognition, eavesdropping devices, intelligence operations, etc. In the recent field of human recognition, the majority of the research be- coming performed presently is focused on the reidentification of various body images taken by several cameras or its focuses on recognized audio-only. However, in some cases these traditional methods can not be useful when used alone such as in indoor surveillance systems, that are installed close to the ceiling and capture images right from above in a downwards direction and in some cases people don't look straight the cameras or it cannot be added in some area such as W.C. or sleeping room. Thus, its commonly difficult to identify any movement or breakthrough process, on the other hand when need to pursue suspect when enter a building or party to identify his location and/or listen to his speech only and isolate it from other voices or noises, the other. Hence, the use of the hybrid combination technique is very effective. In this work, we proposed a multimodal human recognition approach that utilizes both the face and audio and is based upon a deep convolutional neural network (CNN). Mainly, to solve the challenge of not capturing part of the body, final results of recognizing via separate CNNs of VGG Face16 and ResNet50 are joined together depending on the score-level combination by Weighted Sum rule to enhance recognition performance. The results show that the proposed system success to recognise each person from his voice and/or his face captured. In addition, the system can separate the person voice and isolate it from noisy environment and determine the existence of desired person.

*Corresponding author
Email addresses: Sukaina.s.m@ihcoedu.uobaghdad.edu.iq (Sukaina Sh Altyar), Samera.s.h@ihcoedu.uobaghdad.edu.iq (Samera Shams Hussein), Mahir.jasem@yahoo.com (Mahir jasem Mohammed)

Received: August 2021 *Accepted:* September 2021

Keywords: Deep learning, Convolutional Neural Networks, Human Recognition, voice recognition, visual recognition.

1. Introduction

The automated human detection systems have grown to be significance in the last years, it is usefulness in various area such as surveillance, biometrics, etc. This trend has increased the interest of the research in the area of video analysis and, especially, in the field pattern recognition area [3]. Generally, the majority of human recognition systems operate typically at the visual level only, however, there are another information modalities that can be used (such as audio) to complementary information to identify and describe useful patterns" in a scene area. Hence, in the last few years, the Computer Vision studies have focused their efforts to utilize an audio-visual (AV) data combination for human detection area such as video surveillance and biometrics subfield only [4].
Artificial intelligence (AI) for video surveillance makes use of computer software programs that make an analyzation of the audio and images coming from video surveillance cameras to identify objects such as humans, vehicles, etc. The AI transmits an alert if it finds a trespasser breaking the "rule" such as when set that no person is permitted to present in that desired area for the duration of that time of day. [2] The AI program works by utilizing machine vision which is a set of mathematical methods or algorithms, that work the same as a series of questions or flowchart to eval- uate the object viewed with hundreds of thousands of saved reference images of persons in various angles, positions, postures and movements. The AI check if the recognized object is like the reference images, either it is close to in size (height in accordance with width), if it features has two arms and two legs, whether it moves with related speed, and if its vertical rather than horizontal. By taking inconsideration all the values from the several points, an overall rating is taken that gives the AI the probability that the detected object is or is not a human and in same if it the desired person. In case the value is greater than a limit that is set, in that case the alert is sent [5].

The proposed approach is employs audio and visual data to detect desired person and his voice. An audio-visual data of two different person was used. Overlapping time space block features has been calculated for time stamped video and audio data. The input video is able to be recomposed in a way that the audio related to specific persons is enhanced whilst all other sound is suppressed. A CNN has been designed and train that will take the recorded audio mixture, together with tight crops of recognized faces in every frame of the video to become as an input and divides the mixture in to separate audio streams for every recognized speaker. The proposed model utilizes visual information as well as audio to improve the detection quality (compared to visual or audio-only results).

2. literature-survey

- Cees et al (2010) [6], proposed an objective intelligibility strategy that presents a higher correlation of (rho=0.95) along with the intelligibility of both noisy speech and TF-weighted noisy speech. Their method achieved a significantly enhanced performance, much more sophisticated, objective measures. In addition, it is depending on an intermediate intelligibility strategy for short-time (around 400ms) TF-regions and utilizes a simple DFT- based TF-decomposition.

- Torfi et al (2017) [7], proposed a new method that compounded 3D convolutional architecture for sound and video stream networks together with convolutional fusion in spectral dimension (by using of 3D convolutional together with pooling operations) and merging between the networks. Their results show that the proposed model using various data sets is better than

the various present approaches for audio and image matching, furthermore, this model can minimize the number of parameters significantly in comparison to the earlier methods. The results also shows that when utilized CNN it can improve the performance of the learning, in addition, the combined of local speech representative qualities are proven for being more probable for audio-visual Speech Recognition by using CNN.

- Grais and Plumbley (2018) [1], presented an audio separating model merge fully convolutional neural networks together with long short-term memory. They proved that when utilized both networks it can got two advantages, is when use the fully convolutional neural networks it can achieve efficient result for extracting beneficial features from the audio data and the use of long short-term memory are appropriate at modeling the temporal construction of the audio signals. From results, it appears that the combined of long short-term memory and fully convolutional neural networks gained significantly better separation and higher performance as compared to making use of every one individually.

3. The Proposed Method

The proposed model combines each motion (optical flow) in addition to appearance (gray image) info of the face area of human to proper spectrogram masks. From figure 1, the proposed audio-visual model takes each person voices and their face images. The two voice signals are mixed then use short-time Fourier transform (STFT) algorithm to create spectrogram. Additionally, every face image of desired person together with spectrum image will be trained by using Dilated CNN (DCNN) and the output result is "audio-visual combination image. After that we utilize Long Short-Term Memory (LSTM) and obtained Fully Convolutional Network (FC) layer and after that create complex mask and then get back the spectrogram of every person voice. Finally, the model recognizes every person with his voice.

The proposed system structure can be described in follows:

- Input features: The proposed model has involved (audio & video) features as input. Taking into consideration that audio file has a number of speakers, then there is a need to compute the STFT of corresponding to the seconds of an audio segments. In such a case, every time-frequency involves the real and imaginary parts of a complex number, everyone is utilized as input. Even so, it can make use of power-law compression to be able to prevent loud audio from complex soft audio. Identical processing is also applied for both the noisy and clean signals and reference signals. In the inference time, this method can also be used on randomly long segments of audio file. In such cases in which many speakers are recognized in a frame, the proposed model needs to accepting several voice streams as input.

- ANN model: This model is utilized deep learning network that is dependent on [2], that combining LSTM and FCN in order to separate sounds from mixed signal. The FCN uses a CNN to transform image pixels to a pixel classis. FCN is differing from CNN, that it is converts the heigh and width of the feature map of medium layer and getting back to the input image size via the transposed convolution layer, to be sure that the estimations involve a one-to-one linked for input image.

  - The audio stream from the beginning computes the STFT of the input audio signal to obtain a spectrogram. Next used DNN to learns an audio representation.

- The visual streams represent an input thumbnail of recognized faces in each frame in the video, and as well the audio stream is representing an input of video sound-track, that will include a combination of human audio and noises from background. Commonly, the visual streams natural herb face embeddings for each thumbnail making use of a model of pretrain faces recognition, next we learn a visual characteristic by using of a similar DNN.
- A combining of audio-visual representation is after that produced by cobined the visual features with learned audio and is hence further processed utilizing a bidirectional LSTM and 3 FC layers.
- The output of network are a complicated spectrogram mask for each individual that is multiplied together with the noisy input and converted back to waveforms to obtain a a speech signal for most person's voice in mixed speech.

In summery the proposes model need to have following parts:

- Audio streams: this part is composed of dilated convolutional layers.
- Neural Network: The audio streams are combined by concatenating the feature maps of every stream, that are consequently fed into a Long Short-Term Memory (LSTM) followed by 3 FC layers.
- Isolation: the proposed model should be able to isolate multiple speakers, each represented by an audio stream. For this case, a separate and dedicated model need to be trained for every number of speakers.

The algorithm steps can be described as in follows:

⋆ Dataset: The "CSR-I (WSJ0)" dataset [3] has been used in this study, where we select 5 speakers voices. The dataset consists of 880 voices to be as trained sets, 219 to be as development sets and 132 to be as evaluation. The most datasets voices possess 16kHz sampling rate.

⋆ Preprocessing Stage: In this part the model has preprocess the input dataset files and converted the WSJ0 audio files which is in ".sph" format to wav e (.wav) format.

⋆ Mixed Human Voices and Noise: In order to simulate the real speech conservation, we mix the two person voices and some noise and generate audio with a sample rate of 8000.

⋆ Extract Features from Histogram: In this par the STFT has been • utilized in order to extract useful features, after that it transformed into the records format that require to implement as input for TensorFlow ops. This stage involves the following steps:

  – Input the speakers audio file (wav format), then transforms it to 32 bit float values, and then re-forms it dependent upon the number of channels.
  – Determine the STFT of a multiple channel and multiple speaker time signal. It may possibly to place more zeros for fade-in and fade-out and require to create an STFT signal that will make it probable for most desirable reconstruction.
  – Calculate the inverse STFT to accurately reconstruct the time signal.
  – Evaluate the STFT frames obtained from sound samples throughout the time domain.

⋆ Training Network: Training Stage: In this stage, a BLSM algorithm with RNN algorithm to separate multiple speaker speech from single sound record. In this study we have used speeches of different persons, were we record speech of two persons an save it as wav files format)to be an input. Then an Ideal Binary Mask has been used to separate them. After that an ANN is trained by making use of Keras. Every single pickle file contains a dictionary having keys X, S, N, in which each key holds a v list, where the columns are the features.

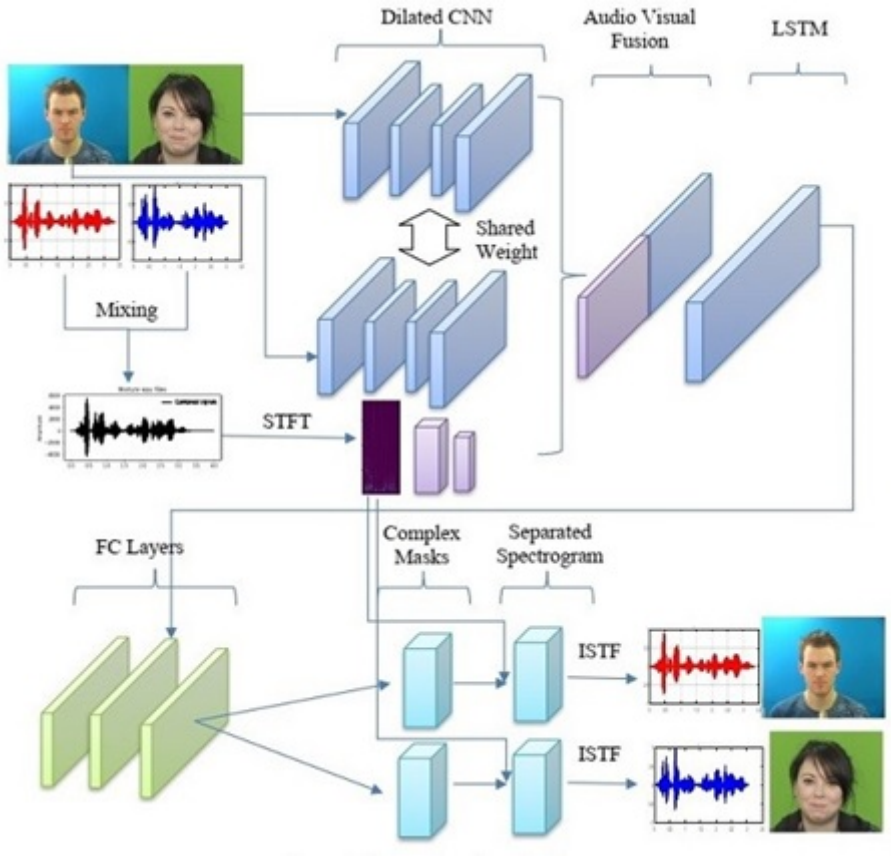

Figure 1: Proposed Model Structure.

4. Results and Discussion

In this section, we will introduce our experiment details and results

- INPUT DATA: Initially, we record a video file from each person in low noise audio environment. The person's faces and their voices signal is used as an input for NN. Figure 2 illustrated the face and voice single of each speaker.
- Mixing the Speakers Signal: In this part we mixed the voices of speakers (2 speech signals) to made a mixed voice signal. Figure 3.
- Spectrogram: The following step is to generate spectrogram signal for speech of each speaker and then mixed their voice signal, as shown in figure 4.
- Voice Masks: The next process is to get voice masks of two person and the mixing signal to used then as input to network training. Figure 5 show the voices masks generation results.

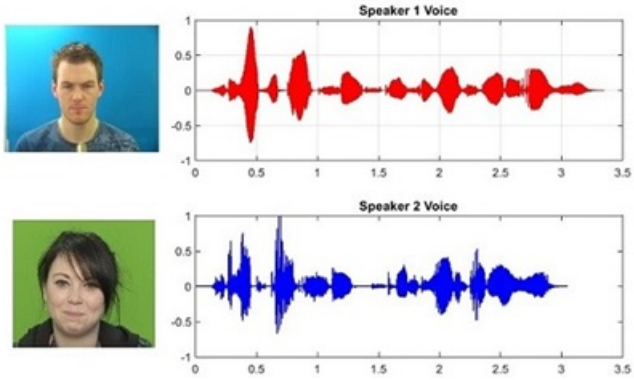

Figure 2: Speakers faces and their voice signals

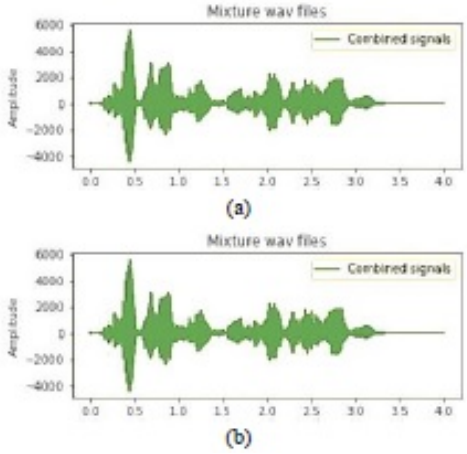

Figure 3: Mixed voices signal results of two speakers.

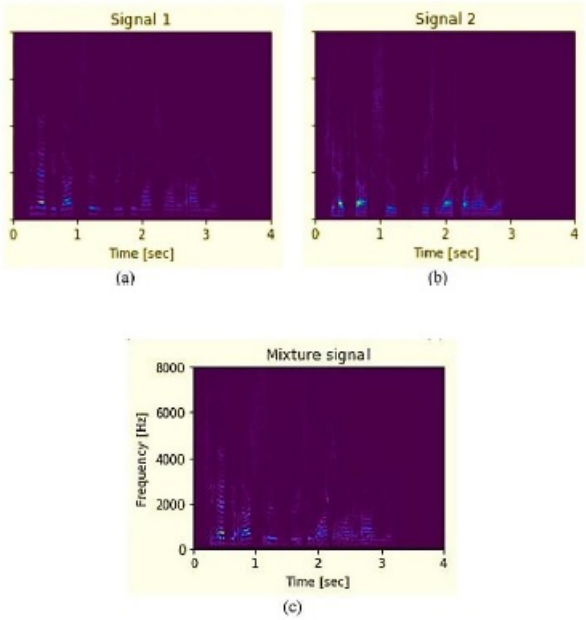

Figure 4: Spectrum signal of (a) speaker 1, (b) speaker 2, (c) mixed speakers voices.

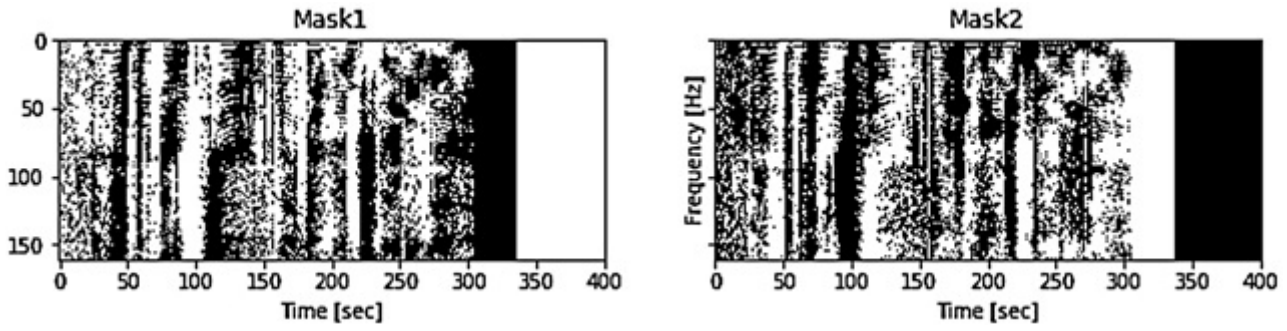

Figure 5: Speaker's voice masks.

- Training Process: the output result from previous stage is feed to proposed network (specified in section 3) to trained and generate weights. The training results is shown in figure 6.

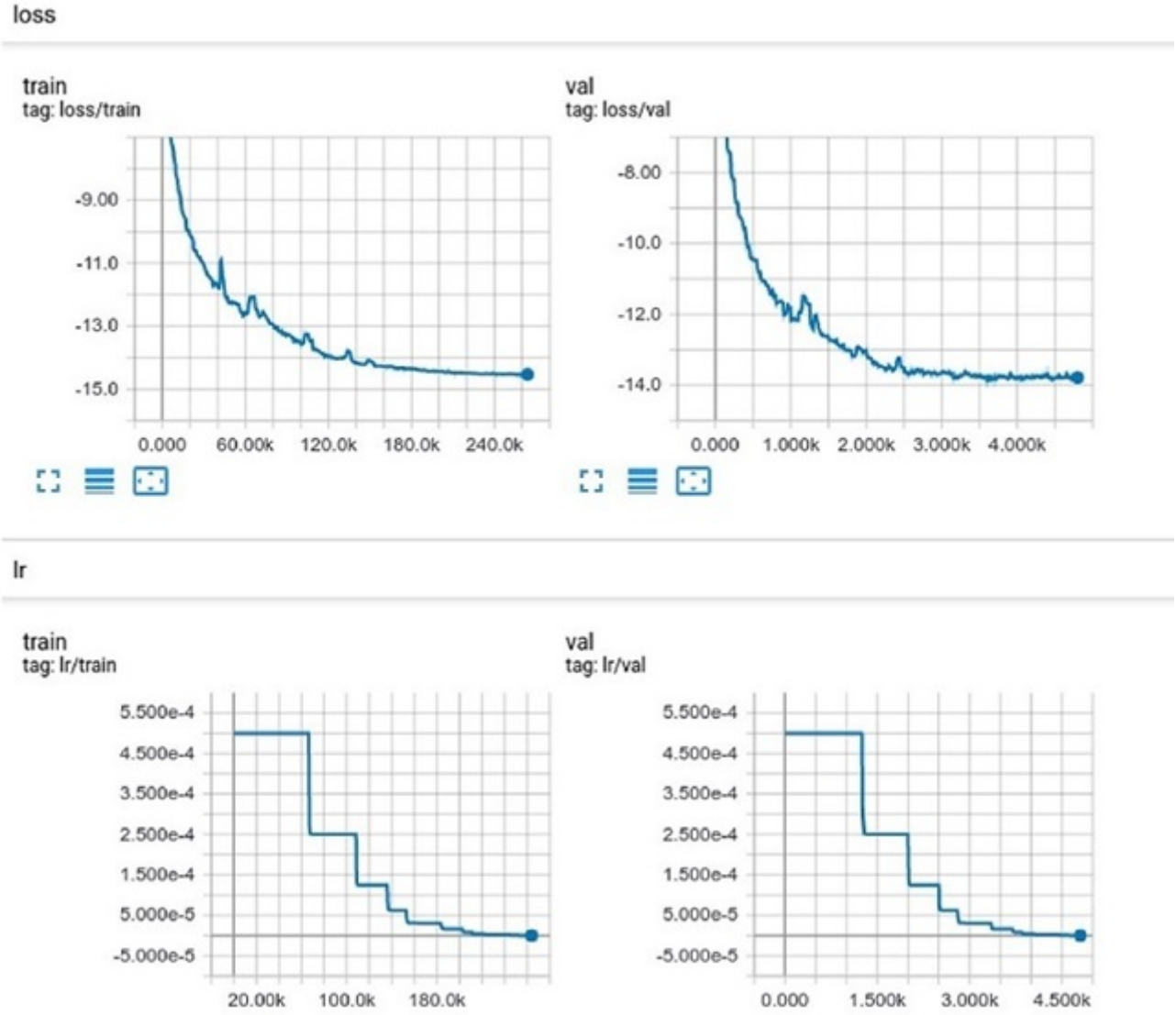

Figure 6: The recognition results for train/validation loss results.

- Recognized Signal Results: In this part, the speakers signal is detected and isolated from

environment (speakers voices and other noise). From result it can observe that the recovered voice signal of speaker one (figure 7) and recovered voice signal from speaker 2 (figure 8) is very close to their original voice signals.

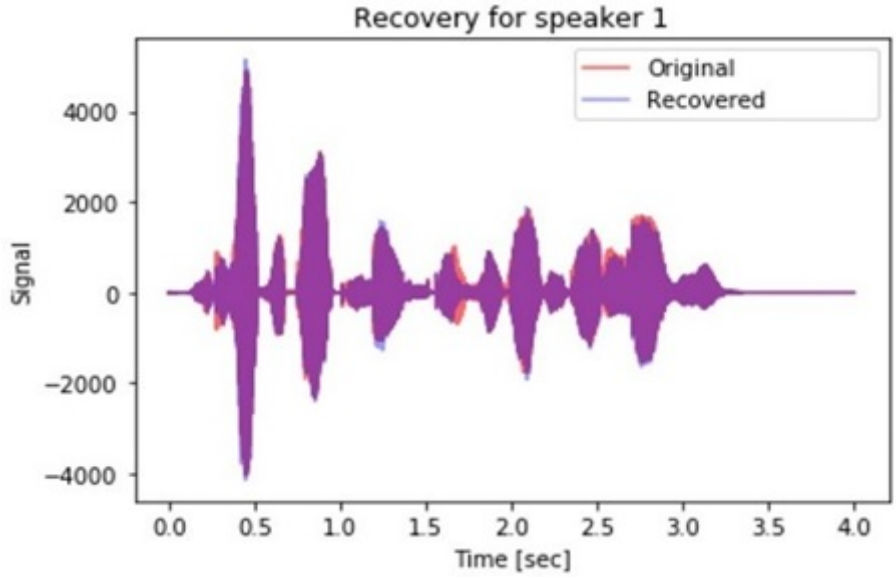

Figure 7: Recovered signal for speaker one.

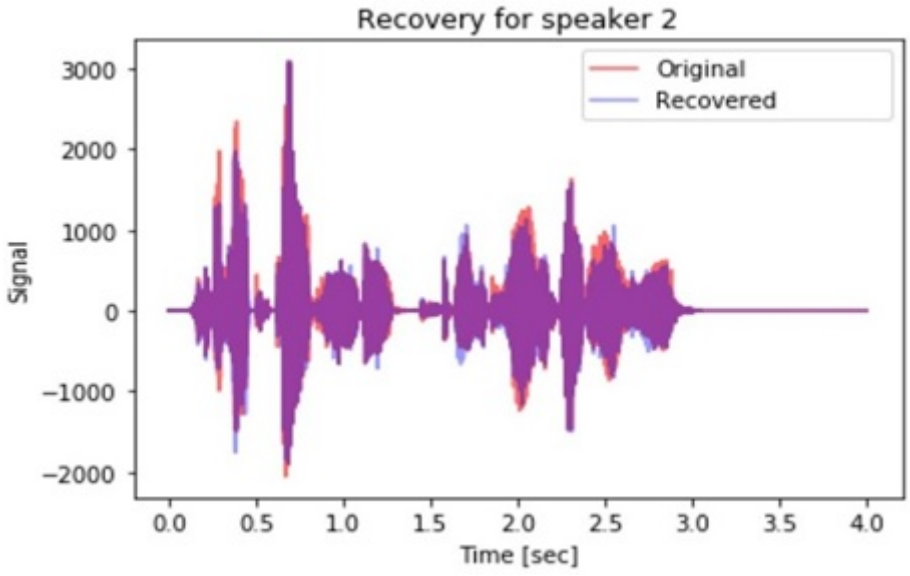

Figure 8: Recovered signal for speaker two.

From results it can see that the program starts by getting persons voices and their face images (in this work we input two persons) then we mixed their voices together. After that we generate histogram images for voice signal of each person in addition to mixed voices signal. Follow that we generated the voice mask for each person voice and for mixed voices to use as input to net- work. The model is then combined each motion along with its gray-scale images of the face area of human to generate the spectrogram masks. From figure 1, the proposed audio-visual model gets each person

voices and their face images. The two voice signals are mixed then use STFT algorithm to generate spectrogram. hence, each speaker face image and his spectrum image is trained by used DCNN in which the result is "audio-visual fusion image. The next step is to use LSTM and got FC layer then generate complex mask then retrieve the spectrogram of each person voice and the final result detect each person with his voice. The results shown in figure 6 and 7 approve that the system is successfully separated mixed voices of each person with signal close to the original voice of desired person and the system can identified each pension and track him with his voice and isolate him from background voices in addition to track the persons of person in either his voice or his face image.

5. Conclusion

In this paper, we have proposed an audio-visual human detection model. The model is based on utilized CNN in which it is combining FCN and a LSTM for voice separation. Two persons has been used to test our model in which each person has its own vide records. The results show that the system is successfully detect human from visual only, audio only or audio-video. In addition, the model can isolate desired person voice from environmental speech and noises in which the system can continuously track desired human even when visual tracking is not available.

References

[1] E. M. Grais and M. D. Plumbley, *Combining Fully Convolutional and Recurrent Neural Networks for Single Channel Audio Source Separation*, In Audio Engineering Society Convention 144. Audio Engineering Society, 2018.

[2] M. H. Kolekar, *Intelligent Video Surveillance Systems: An Algorithmic Approach*, CRC Press, 2018.

[3] Y. Kortli, M. Jridi, A. Al Falou and M. Atri, *Face recognition systems: A Survey*, Sensors, 20 (2020) 342.

[4] J. Kotus, K. Lopatka, A. Czyz˙ewski, G. Bogdanis and June, *Audio-visual surveillance system for application in bank operating room*, Int Conf Multimedia Commun Serv Secur., Springer, Berlin, Heidelberg, 2013, pp. 107-120.

[5] G. O'Regan, *Artificial Intelligence and Applications*, Springer, 2018.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens and J. A. Jensen, *short-time objective intelligibility measure for time-frequency weighted noisy speech*, In 2010 IEEE Int. Conf. Acoustics, Speech and Signal processing, IEEE, (2010) pp.4214-4217.

[7] A. Torfi, S. M. Iranmanesh, N. Nasrabadi and J. Dawson, *3d convolutional neural networks for cross audio-visual matching recognition*, IEEE Access, 5 (2017) 22081–22091.