Egyptian Informatics Journal 23 (2022) 121–131

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

A speech separation system in video sequence using dilated inception network and U-Net

Ghada Dahy *, Mohammed A.A. Refaey, Reda Alkhoribi, M. Shoman

Faculty of Computers and AI, Cairo University, Egypt

ARTICLE INFO

Article history:
Received 7 January 2022
Revised 15 May 2022
Accepted 4 September 2022
Available online 29 September 2022

Keywords:
Speech separation
Surpassing Speech Noise
Speech Analysis using STFT
Cepstral Coefficients
Log Filterbank Energies and Spectral
Subband Centroid
Audio-visual speech

ABSTRACT

In this paper, an audio-visual model for separating a speech of the target speaker from a combination of other speakers' speeches is proposed. It can be used in speech separation, automatic speech recognition systems (ASR) and also in creating single speaker speech databases. Speech separation is complicated problem using audio information only so visual and auditory signals are combined to complete the separation process. The proposed model consists of four modules, two for audio signal, one for visual feature and the last one used to concatenate the features resulted from the previous three modules to generate the separated signals. Our proposed model improved Short-time objective intelligibility (STOI) with 11%, Perceptual Evaluation of Speech Quality (PESQ) with 24%, and Frequency-weighted Segmental SNR (fwSNRseg) with 16% compared with previous works. It also improved Csig' which is the predicted rating of speech distortion with 13% and 'Covl' which is the predicted rating of overall quality with 18% compared with previous audio-visual models.

1. Introduction

Voice data has an important role in our current days, and speech communication becomes more frequently method daily, as using software to send voice messages and controlling mobile phone setting and applications.

In real-world environments, speech occurs simultaneously with acoustic interference, as background noise. The interference usually affects in speech perception, and results in performance degradation. Speech separation is being one of the essential problems in the field of audio processing, it is considered as the task of focusing on the sound of the target speaker in a noisy environment and mutating "silencing" speeches of other speakers and background noise. It is also considered the process of extracting all interfered speech sources in a given mixed speech signal.

With increasing speech technologies, as Apple Siri and Amazon Alexa, the speech separation problem has become more of an important and interest. Actually, speech separation is considered as a pre-processing phase that is used in many speech applications such as noise separation from speech signals, ASR and creation of speech databases. Actually, machines cannot recognize speech relatively well in noisy environments, there is a significant deterioration in performance of recognition in crowding environment with sounds. Progress in the speech separation area relies heavily on the development of appropriate speech databases.

Human auditory system has the ability that enables them to distinguish sounds from complex mixture of signals through their experience but machines can't distinguish between them that known as the "cocktail party effect" as human could understand a conversation partner in crowded environment with other speakers in background. Although a significant progress has been made on this problem, it is a widely regarded challenge. The main goal of this paper is to recover a better quality speech of a target speaker from a mixed background signals. We hope in the future having hearing aids and earbuds that able to remove audio sources that we don't want to hear to them.

It is difficult to model the process of speech separation effectively because the time frequency characteristics of human's sound

* Corresponding author at: Teaching Assistant, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt.

E-mail addresses: g.dahy@fci-cu.edu.eg (G. Dahy), m.refaey@fci-cu.edu.eg (M.A.A. Refaey), r.abdelwahab@fci-cu.edu.eg (R. Alkhoribi), m.essmael@fci-cu.edu.eg (M. Shoman).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2022.09.001

vary widely. In an acoustic environment like a cocktail party, it is difficult to follow one speaker in the presence of other speakers and background noises. Researchers have shown that the advantage of using visual streams beside audio signal as an assistant factor to analyze mixed sounds in a noisy environment. As we have a limited resource, it is difficult to train the proposed model with datasets having millions of recorded videos so a subset samples were taken from TheOxford-BBC LRS2 Dataset.

Most Speech separation systems usually operate on the short time Fourier transform (STFT), researchers have tried to analyze speech in frequency domain so we used STFT components to analyze input audio signals. There is a great attention in the past few years to enhance accuracy of speech separation systems by suggesting new network structure or using new loss function depending resulted data from STFT.

In this paper, we introduced a new solution for speech separation which achieved an interesting performance comparing against other old methods. The proposed solution combines visual and auditory signals to complete the process of separation. The selected visual features are used to guide in Isolating audio of the desired speaker and to improve the quality of separation process. Supporting model with other audio features beside visual feature and STFT components as the Mel Frequency Cepstral Coefficients, Log Filterbank Energies and Spectral Subband Centroid can enhance the quality of the separated speech. It combines four modules, two to analyze audio signal, one to learn visual feature and last module is used to concatenate the features resulted from the three previous modules to generate the separated signals. To complete the process of separation, we built a dataset consists of 15,625 samples splitted into 14,063 for training and 1562 for testing. U-net and inception network deep learning structure are used to support our proposed model to learn sound of target speaker.

The main contributions of this paper are building audio separation model of two parallel speakers, using visual features as a guide in the separation process and analyzing audio signals in more level to enhance the performance of separated sounds.

The rest of this paper is organized as follows. We first review several related speech separation approaches in Section 2. Then, we elaborate our proposed model for sound separation in Section 3. In Section 4, we describe the experimental setup, report the experimental results, and discuss our findings. Finally, we conclude our work in Section 5.

2. Related works

There are many researchers who try to solve this problem using audio information only. Yannan et al. [1] demonstrated unsupervised co-channel speech separation. Their proposed DNN framework could well segregate speech from mixtures of two unseen speakers with different genders because the distance between speakers of the same gender is smaller than those between speakers of different genders. They used log power spectral an input feature to DNN. They implemented the system under the MMSE criteria which minimize the mean squared error between the DNN outputs and the reference clean features of both female and male speakers. Their system achieved PESQ better than GMM (Gaussian mixture model).

There are some researchers have tried to separate sounds using reference audio of the target speaker. Quan Wang et al. [2] presented a novel system that separates the voice of a target speaker from multi-speaker signals, by making use of a reference signal from the target speaker. They achieved this by training two separate neural networks. Their system consists of two separately components, the speaker encoder and the VoiceFilter. The speaker encoder produced a speaker embedding from an audio sample of the target speaker. The second system is named VoiceFilter is developed for speech enhancement. Their network can generate a magnitude spectrogram of the target speaker; they directly add a noisy phase to the estimated magnitude and finally apply ISTFT to get the final result. Their network was trained to minimize the difference between the masked magnitude spectrogram and the target spectrogram. In the VoiceFilter, they need a clean reference utterance of each target to compute the speaker embedding so they generated their dataset from VCTK dataset. They randomly selected 99 speakers for training and 10 for testing. To evaluate the performance of their system, they used two metrics: the speech recognition word error and source to distortion ratio. The system reduced word error rate to 23.4 %. Using visual information from target speaker as expression from his/her face's muscles is considered one of the best methods to improve the accuracy of speech separation.

Ariel Ephrat et al. [3] designed and trained a dilated convolution neural network model which takes an audio of the mixed sounds, with all detected faces of each frame in the video as an input, and splits the mixed sounds into sets of separated audios for all detected speakers. The model used visual Embedding features to improve the source separation quality and also to support their model in keeping track with all speakers in the video. Their model takes from the user marked face of the target speaker which he/she wants to separate as a guide in separation process. They introduced a new dataset which consists of thousands of hours vides recorded from web. To generate the training data, they mixed clean audios of two different speakers. Their proposed network was implemented by using TensorFlow. It has high SDR compared with other previous visual methods when trained in their data sets. They showed that their audio system is robust to separate voices of the same gender. They presented the importance of visual information in increasing the system accuracy even if it is small.

There are end-to-end learning methods for audio source separation operates directly in the time domain and also may be used in frequency domain as Time Audio Separation Network (TAS-Net) [4,5], U-Net [6,7], Multi-Resolution Network [8] and WaveNet [9]. TAS-Net depends on taking the raw data from user and separates it to unmixed signal. There is a great attention in the current days to TAS-Net in sounds separation depends on audio or audio-visual information. U-Net contains two paths. First path is the contraction path (also called the encoder) which is used to capture the context in the audio. The encoder is just a traditional stack of convolutional and max pooling layers. The second path is the symmetric expanding path (also called the decoder). U-net focuses only on audio information to separate signals. It operates in time and frequency domain. Most researches related to U-net are in time domains and depend on separating songs from background noise as drums and guitar. There are researches in frequency domain which use U-net for speech enhancement and separating background noise. Multi-Resolution network consists of two main parts encoder and decoder. Encoder consists of set of convolution layers and a set of filters applied to each layer after that the resulted feature maps from each layer are concatenated to generate the input of the next layer after applying activation function on them. Decoder uses the reversed operation of encoder to generate the final output. Most of researches uses multi- resolution network are in frequency domain but we can benefit from the filter sets idea in increasing the accuracy of the time domain signal by analyze the sound using different resolution. Inception Modules [10] are used in Convolutional Neural Networks to allow more efficient computation and deeper Networks.

3. Audio-visual speech separation model

The proposed model introduced architecture for speech separation when a list of speaker is available where the number of speakers in this paper is two. It is considered as audio visual model that separates speech relative to face embedding features of each speaker. It could achieve better performance even using limited size dataset.

The proposed model consists of four modules. It takes visual embedding features of the detected faces and mixed audios as inputs, and generates compressed complex spectrogram $R_{(t,f)}$, related to each speaker as shown in Fig. 1. The Complex spectrogram of each speaker $M_{(t,f)}$ is calculated relative to STFT components of mixed speech $Y_{(t,f)}$ and clean speech of the target speaker $S_{(t,f)}$ as in Eq. (1).

$$M_{(t,f)} = \frac{S_{(t,f)}}{Y_{(t,f)}} \tag{1}$$

$$R_{(t,f)} = 10\frac{1 - e^{-0.1M}}{1 + e^{-0.1M}} \tag{2}$$

where t and f are indexes of time and frequency respectively.

The proposed model completes the process of separation by implementing two modules for audio, one for visual streams and the fourth one for feature concatenation. The visual streams are taken from speaker's detected face in video's frames, and the audio data is taken from mixed speech in the input video.

For the visual stream module, a pre-trained face recognition model is used to generate face embeddings per frame in the input video for each detected faces that called A Unified Embedding for Face Recognition and Clustering 'FaceNet' which is considered as a facial recognition system. It achieved high face recognition accuracy in many benchmark face dataset as Youtube Face Labeled Faces in the Wild datasets. It is directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity [15], then the proposed module learns a visual feature using the same idea of filters decomposition in inception network that enable us to analysis input by using multiple types of filters that are concatenated and forwarded to the next block or layer instead of using only one filter. The Inception block is designed in a way which needs less computational complexity and fewer parameters compared with using only one filter. This module consists of four inception block where each inception block consists of concatenation of three convolution neural networks (CNN) and MaxPooling2D.

Another Inception network with different number of filters is used to train audio stream analysis module which takes Mel Frequency Cepstral Coefficients, Log Filterbank Energies and Spectral Subband Centroids of mixed sounds as an input.

STFT of the mixed signal is calculated to train short time speech frequency analysis module by using a similar U-net convolutional

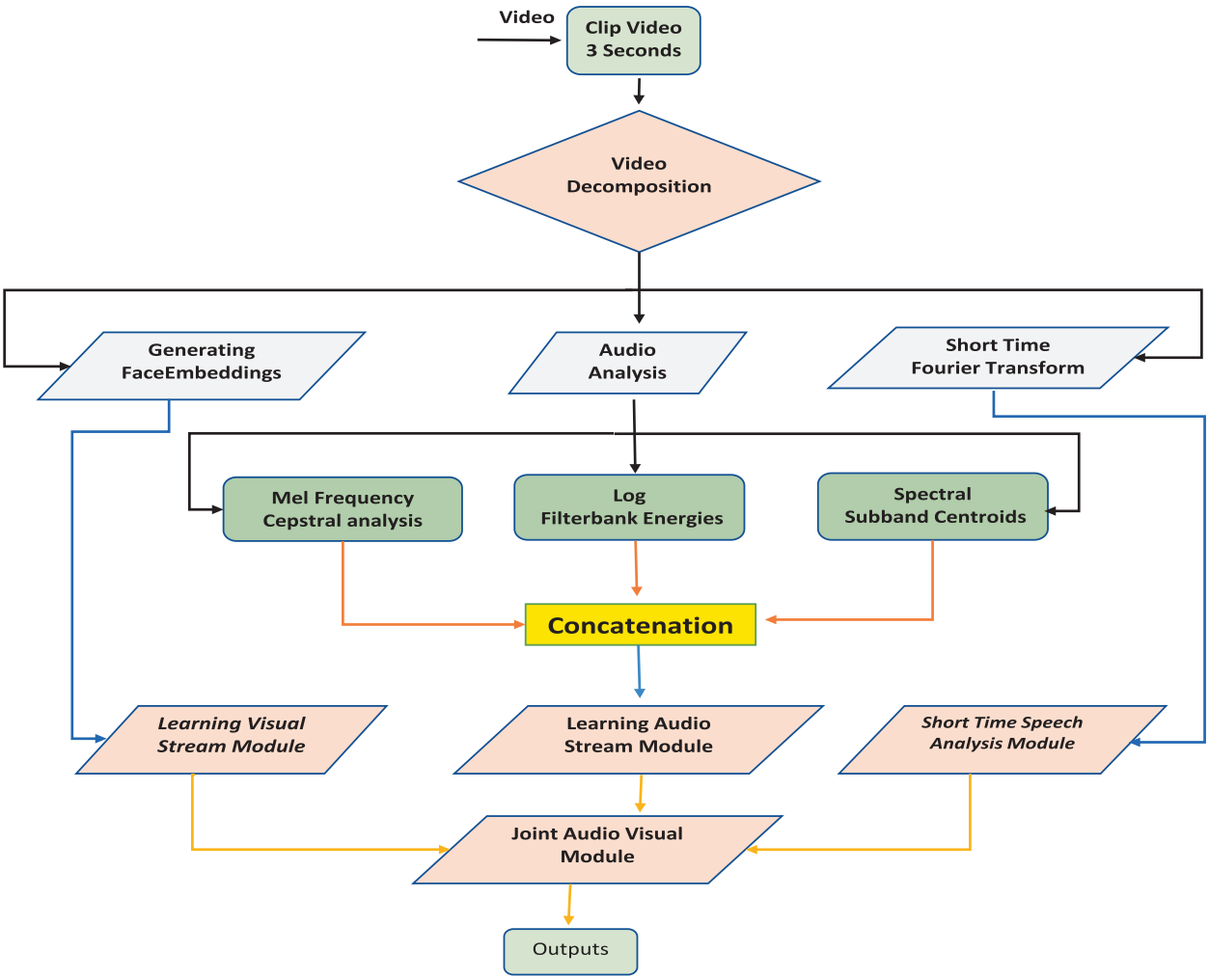

Fig. 1. Structure of the Proposed Model Which Consists of Four Modules to Isolate Speech of the Target Speakers.

NN. U-net was used for medical image segmentation and its structure is very shallow and simple as it repeat blocks of CNNs,Relu and maxPooling for decoding and encoding path so it is very easy to implement. It achieved high segmentation performance on various benchmarks and also it can work well with little amount of data. It has better performance in case of using it in speech separation and enhancement.

The audio-visual architecture is created by combining the learned visual embeddings features and audio features, then this combination is processed by using a bidirectional long short-term memory 'BLSTM' and three dense layers. BLSTM able to analysis input sequence from back to front and from front to back that enable the network to generate a context of the input where the network able to connect to past and future. The outputs of the network are complex spectrogram masks for all detected speaker in the video, then multiplied by the signal of the input audio and finally, they are converted back to waveforms to get an isolated speech of each detected speaker.

3.1. Learning visual stream module

This module is used to guide the overview model in the process of separation as the model would retrieve the voice of the target speaker based on the his/her face features.Firstly a pre-processing step is applied on the scanned images of the recoded videos to prepare the input of this module. In this step, a pre-trained FaceNet network takes detected faces from input image as input and output is a vector of 1792 values which represent the most important features crossponding to each face. In machine learning, this vector is called embedding feature vector then the proposed module takes as an input the concatenation of face embeddings features which are 75x1792Xnumber of speakers, as the number of frames in the input video is 75 frames, and FACENET network obtains 1792 face embeddings features for each frame.

The proposed module is trained by using inception block as it is built by different scales of convolution kernels, then contact in the end to fuse features of different scales. It depends on dilated Convolutional Neural Networks to allow more efficient computation with different dilated factor in each block.

The dilated convolution is called convolution with a dilated filter. It can expand the input by setting holes between its consecutive elements. It works as convolutions but using pixel skipping to cover a large area from the input and discover relations between elements of the input easily. The classical CNN consumes too much computation resources but the dilated CNN could reduce the time of training by 12.09 % [11] and also it has better performance in image classification and segmentation.

Mainly in this module, we benefit from advantages of Inceptions blocks and dilated convolution by constructing architecture that consists of four inception blocks where each inception block has different dilation factor. This module uses only four blocks because the updated version from FACENET is used for extracting input features of face and the process of increasing the dilation factor would be harmful to small embedding features retrieved from FACENET [12].

Each block consists of concatenation of three dilated convolution neural network and MaxPooling2D layers with size N1XM1, the first dilated CNN uses N filters with size N2XM2, second dilated CNN uses M filters with size N3XM3, third dilated CNN uses L filters with size N4XM4, then the previous four layers are concatenated, Batchnormaliztion and relu activation function are applied to the concatenated features as in Fig. 2. Max pooling is done as part of inception block to reduce over-fitting by providing an abstracted form of the representation. Each inception block has different dilation rates to support exponential expansion in the context of the receptive field and this led to no loss of resolution.

Finally, a CNN layer with K filters is applied on features map resulted from Batchnormlization layer of the last inception block.

3.2. Learning audio stream module

To improve the process of speech separation, a concatenation of the Mel Frequency Cepstral Coefficients, Log Filterbank Energies and Spectral Subband Centroids of the input audio which are 302X65 as the number of Mel Frequency Cepstral Coefficients are 302X13, Log Filterbank Energies are 302X26 and Spectral Subband Centroids are 302X26. To analysis these features and argues the important of each one in separation process.

The proposed module takes the concatenation of these features as an input to six inception blocks with different dilation rate as the width of the input feature is 65 that means, if the number of inception blocks are increased than six, the last one will scan borders all the time. Using dilated CNN in the process of analysis audio feature with different dilation factor in each inception block is con-

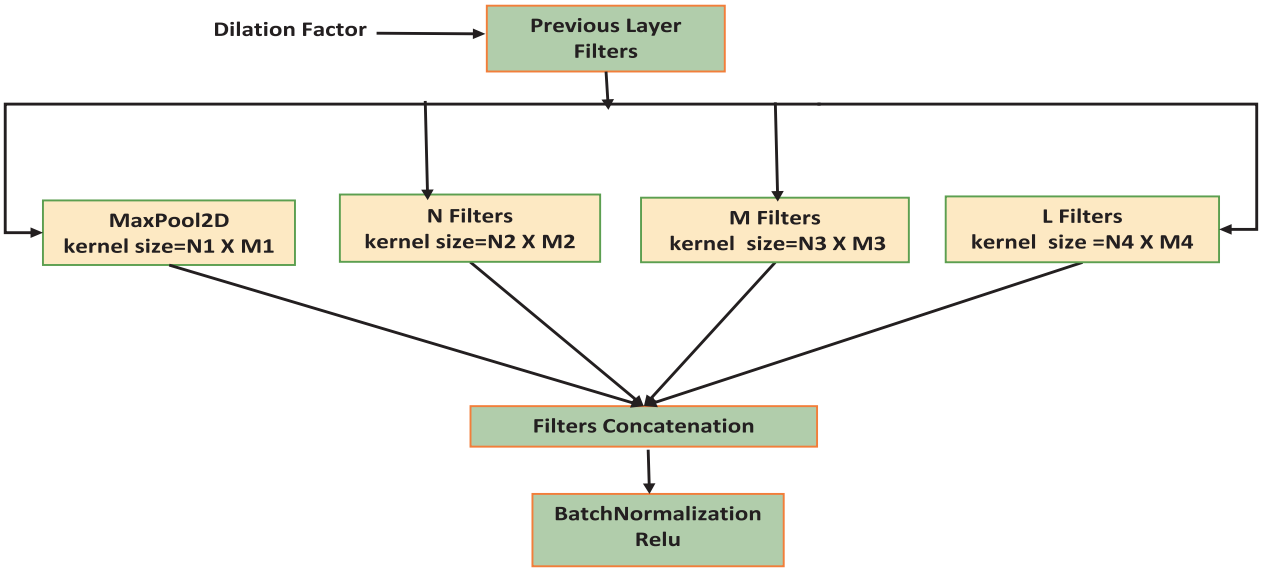

Fig. 2. Inception Block layers.

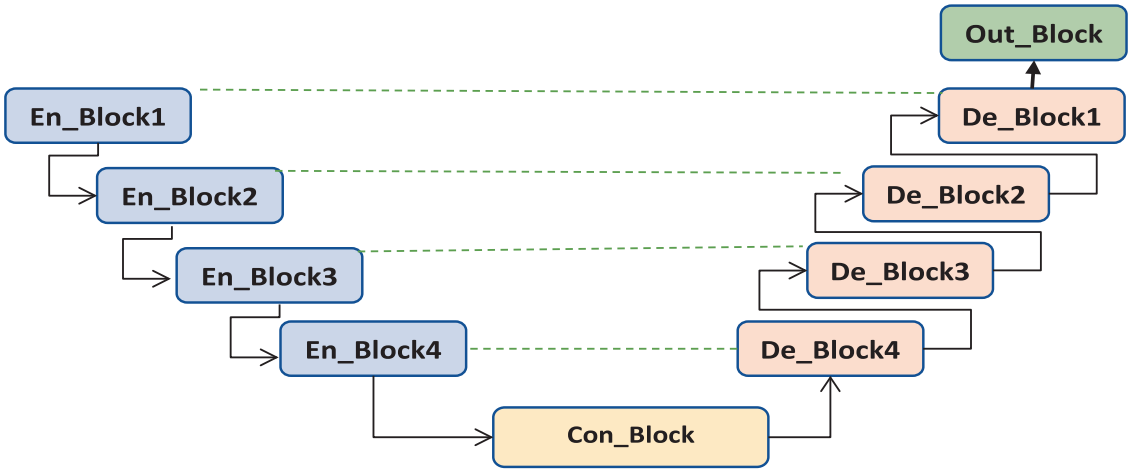

Fig. 3. Short Time Speech Analysis Module by using U-Net Network.

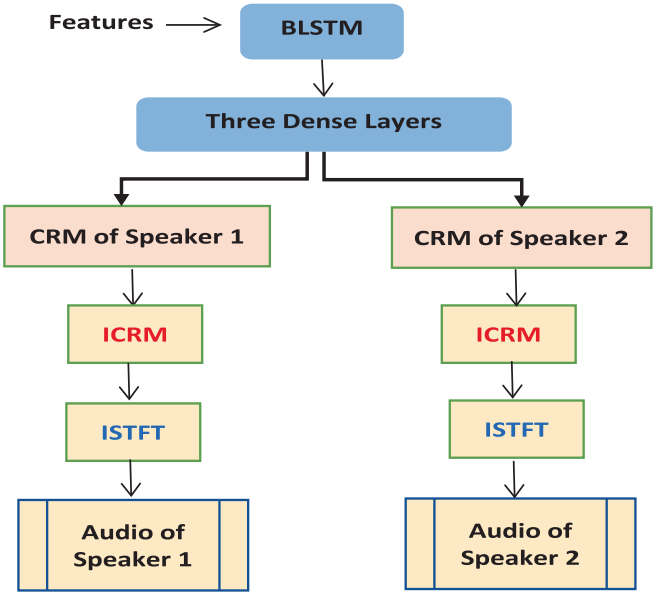

Fig. 4. A joint Audio-Visual Module.

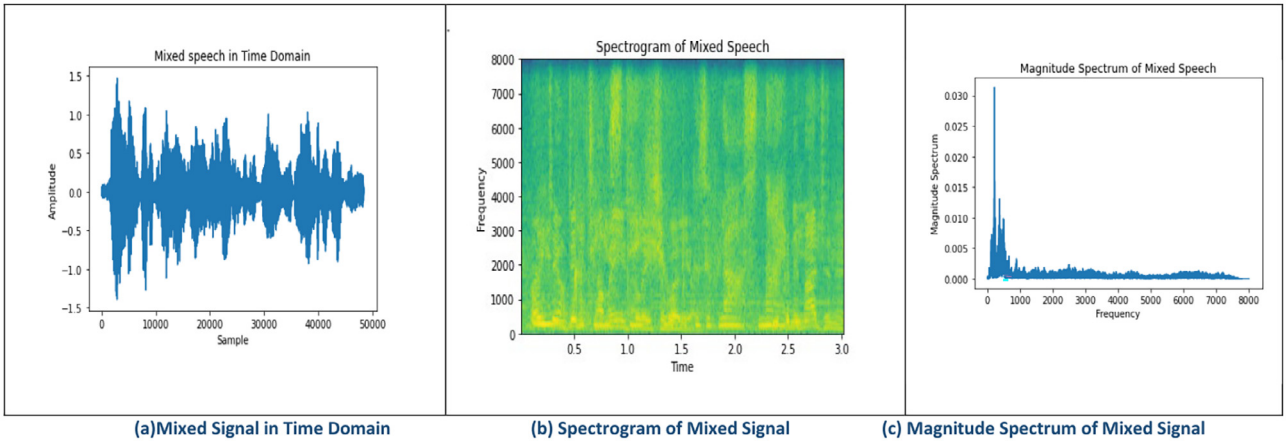

Fig. 5. Mixture speech of two speakers.

sidered an important process to resolution and context of the input.

Each block consists of concatenation of three dilated CNN and MaxPooling2D layers. The first dilated CNN uses N filters with size N2XM2, second dilated CNN uses M filters with size N3XM3, third dilated CNN uses L filters with size N3XN4 and MaxPooling2D with size N1XM1, then the previous four layers are concatenated, Batch-normaliztion and Relu activation function are applied on the concatenated features. The proposed module uses different kernel size of filters to analyze and track audio features in different scales.

Table 1
Filters of Inception Network.

Parameters	Learning Visual Stream Module	Learning Audio Stream Module
N	50	25
M	50	50
L	100	100
K	256	–
N1 × M1	3 × 1	3 × 3
N2 × M2	1 × 1	1 × 1
N3 × M3	3 × 1	3 × 3
N4 × M4	5 × 1	5 × 5
Dilation Rates	[1,2,4,7]	[1,2,4,8,16,32]

Table 2
Filters of Unet Network.

Block	*Layers[Filter Size = 5x5]*	*Input Size*	*Output Size*
En_Block1	CNN1	304 × 256 × 2	304 × 256 × 64
	CNN2	304 × 256 × 64	304 × 256 × 64
	MaxPooling	304 × 256 × 64	152 × 128 × 64
En_Block2	CNN1	152 × 128 × 64	152 × 128 × 128
	CNN2	152 × 128 × 128	152 × 128 × 128
	MaxPooling	152 × 128 × 128	76 × 64 × 128
En_Block3	CNN1	76 × 64 × 128	76 × 64 × 256
	CNN2	76 × 64 × 256	76 × 64 × 256
	MaxPooling	76 × 64 × 256	38 × 32 × 256
En_Block4	CNN1	38 × 32 × 256	38 × 32 × 512
	CNN2	38 × 32 × 512	38 × 32 × 512
	MaxPooling	38 × 32 × 512	19 × 16 × 512
Con_Block	CNN	19 × 16 × 512	19 × 16 × 1024
De_Block4	UpSampling	19 × 16 × 1024	38 × 32 × 512
	CNN1	38 × 32 × 512	38 × 32 × 512
	CNN2	38 × 32 × 512	38 × 32 × 512
De_Block3	UpSampling	38 × 32 × 512	76 × 64 × 256
	CNN1	76 × 64 × 256	76 × 64 × 256
	CNN2	76 × 64 × 256	76 × 64 × 256
De_Block2	UpSampling	76 × 64 × 256	152 × 128 × 128
	CNN1	152 × 128 × 128	152 × 128 × 128
	CNN2	152 × 128 × 128	152 × 128 × 128
De_Block1	UpSampling	152 × 128 × 128	304 × 256 × 64
	CNN1	304 × 256 × 64	304 × 256 × 64
	CNN2	304 × 256 × 64	304 × 256 × 64
Out_Block	CNN	304 × 256 × 64	304 × 256 × 2

3.3. Short time speech analysis module

In this module, audio data which is recoded in input video is analyzed and understood by using real and imaginary components of audio files after converting them to frequency domain.

The input of this module is a concatenation of real and imaginary components resulted from STFT which are 304X256x2. Frequency components are fed into U-Net to learn audio features. UNet is a convolutional neural network architecture that built by expanding with few changes in the CNN architecture, first used for biomedical image segmentation as has a great effect in o analyzing and tracking features in different sizes and also it has a good performance in the speech separation. The U-net architecture is symmetric and consists of two major parts contracting path, Expansive Path, a convolution process between contracting and expansion paths named Con_Block and finally a convolution layer named Out_Block is applied after expansion path as shown in Fig. 3. The left part which is called contracting path consists of four processes named En_Block1, En_Block2, En_Block3 and En_Block4 respectively. Each process consists of two 2d convolutional and max pooling process which halves down size of input features. The right part which is called expansive path consists of four process named De_Block1, De_Block2, De_Block3, and De_Block1 respectively. It is going to upsized features of the previous process to its original size. The size of the input features is upsized, then the upsized features are concatenated with the corresponding features from the contracting path to generate a concatenated features which are taken as input to next layers in the same process which has two 2d convolutional layers, and finally the resulted image is prepared to be taken as input to start the next process.

3.4. A joint audio-visual module

Features concatenation is the method that has ability to transform the features in the separated columns into a single column of feature vector. It is necessary and essential process to generate new feature set that control the separation process. In this module, Learning Visual Stream, Learning Audio Stream and Short Time

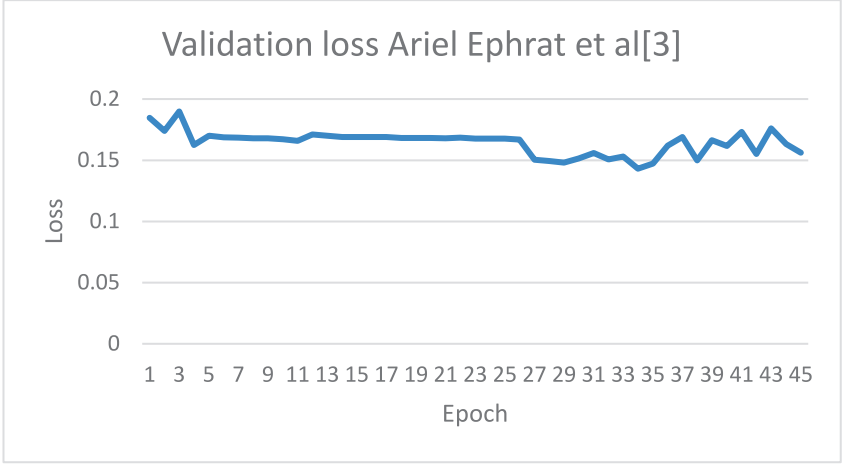

Fig. 6. Validation Loss Ariel of Ephrat et al. [3].

Speech Analysis modules are concatenated by combining the feature maps of the previous three modules, then BLSTM are applied on the concatenated features followed by three Dense layers.

Outputs of this module are compressed complex masks of speaker's. All results of our experiments depend on CRM to separate audio segment of the target speaker. To retrieve the speech of the target speakers, firstly the predicted CRM resulted from model is decompressed by using equation (3), and then we calculate ICRM of decompressed CRM. The final speech of each speaker is computed by using ISTFT of ICRM as in Fig. 4.

$$D(R) = \frac{1}{0.1}\log\left(\frac{10-R}{10+R}\right) \tag{3}$$

where D is decompressed CRM resulted from model, *ICRM* is the inverse of decompressed, (R) is compressed CRM.

4. Experiments and results

In order to compare our results with the Audio-Visual model of the previous work, we implemented Ephrat et al. [3] Speech Separation model, we train and test it using the same dataset in our proposed model.

Table 3
We compare our speech separation results in case of training our proposed network using Experiment 1 and Experiment 2 with Cocktail party model [3], using the Short-time objective intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ) and Frequency-weighted Segmental SNR (fwSNRseg) to compare the quality of models.

Training/Testing	STOI	PESQ	fwSNRseg
Ariel Ephrat et al. [3]	0.72	1.75	10.90
Experiment 1	0.77	1.85	11.95
Experiment 2	0.80	2.17	12.59

Table 4
Using 'Csig' which is the predicted rating of speech distortion and 'Covl' which is the predicted rating of overall quality to compare the proposed with Cocktail part model [3].

Training/Testing	Csig)	Covl
Ariel Ephrat et al. [3]	3.17	2.40
Experiment 1	3.28	2.51
Experiment 2	3.58	2.83

To generate our dataset of two-speaker, 250 videos are taken from the BBC (LRS2) Dataset [13]. The BBC dataset consists of thousands of spoken sentences from BBC television. Each sentence is up

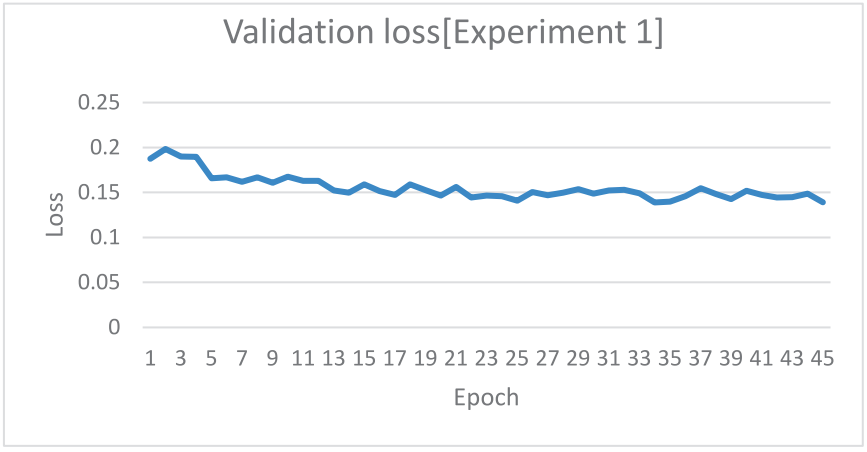

Fig. 7. Experiment 1 validation loss.

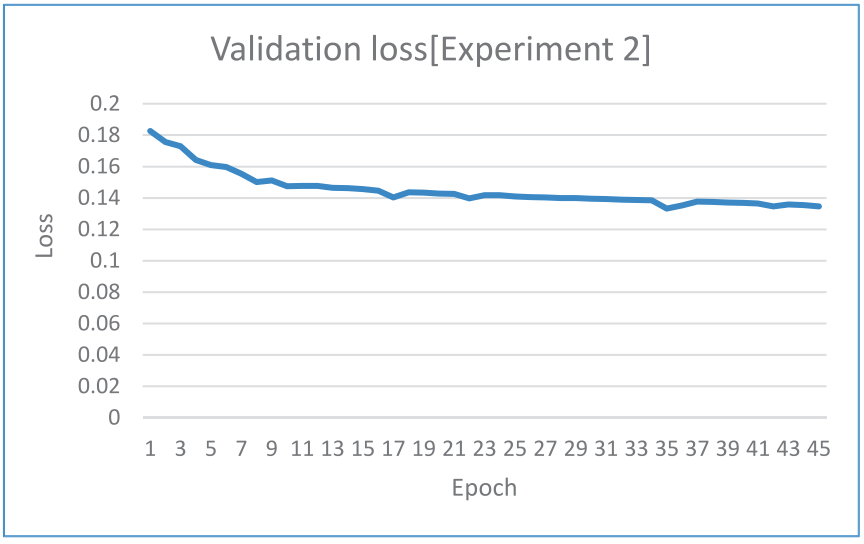

Fig. 8. Experiment 2 validation loss.

to 100 characters in length. The video segments have a length between 3 and 10 s. The dataset was built by combining clean speech of two speakers from different videos: combined speech = *Speaker*1 + *Speaker*2where *Speaker*1 and *Speaker*2 are clean speeches as in Fig. 5. The total generated samples in our datasets are 15,625 samples splitted into 14,063 for training and 1562 for testing.

The proposed model takes visual embedding and audio features as inputs. OpenCV library is used to detect speaker's face in the input video. OpenCV is the most popular library for computer vision. Originally written in C/C++, it now provides bindings for Python [14].

Audio files are resampled to 16 kHz, then STFT is computed of 3-second audio segments of the input video. Every time–frequency (TF) bin consists of real and imaginary parts of a complex number

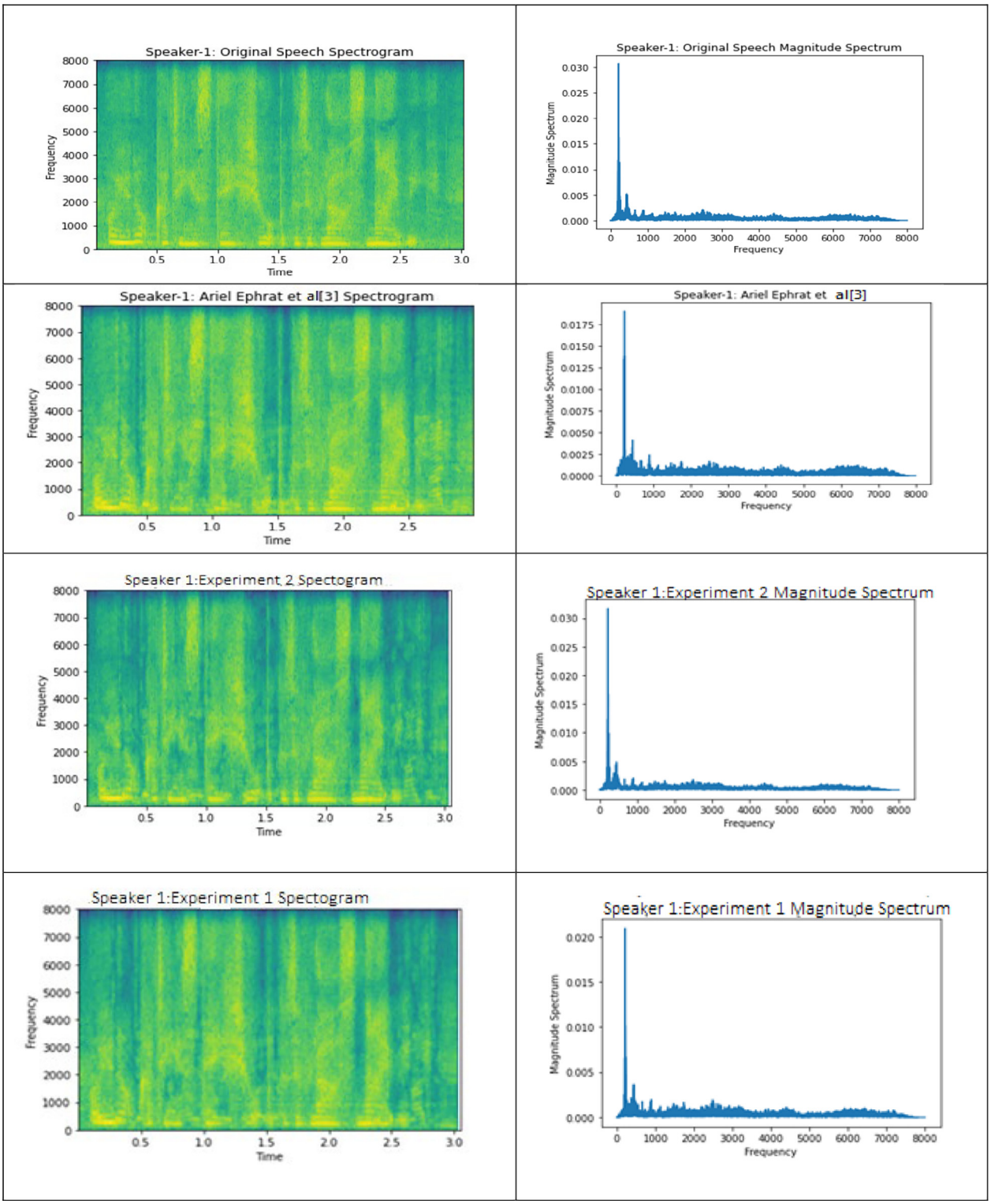

Fig. 9. Spectrogram and Magnitude Spectrum of Speaker-1.

as they are used as an input to train our model. STFT is computed on input audio files using a Hann window of length 25 ms, hop length of 10 ms, and FFT size of 512.

We use python_speech_features [16] library to compute Mel Frequency Cepstral Coefficients, Log Filterbank Energies and Spectral Subband Centroids of 3-seconds input segments which are

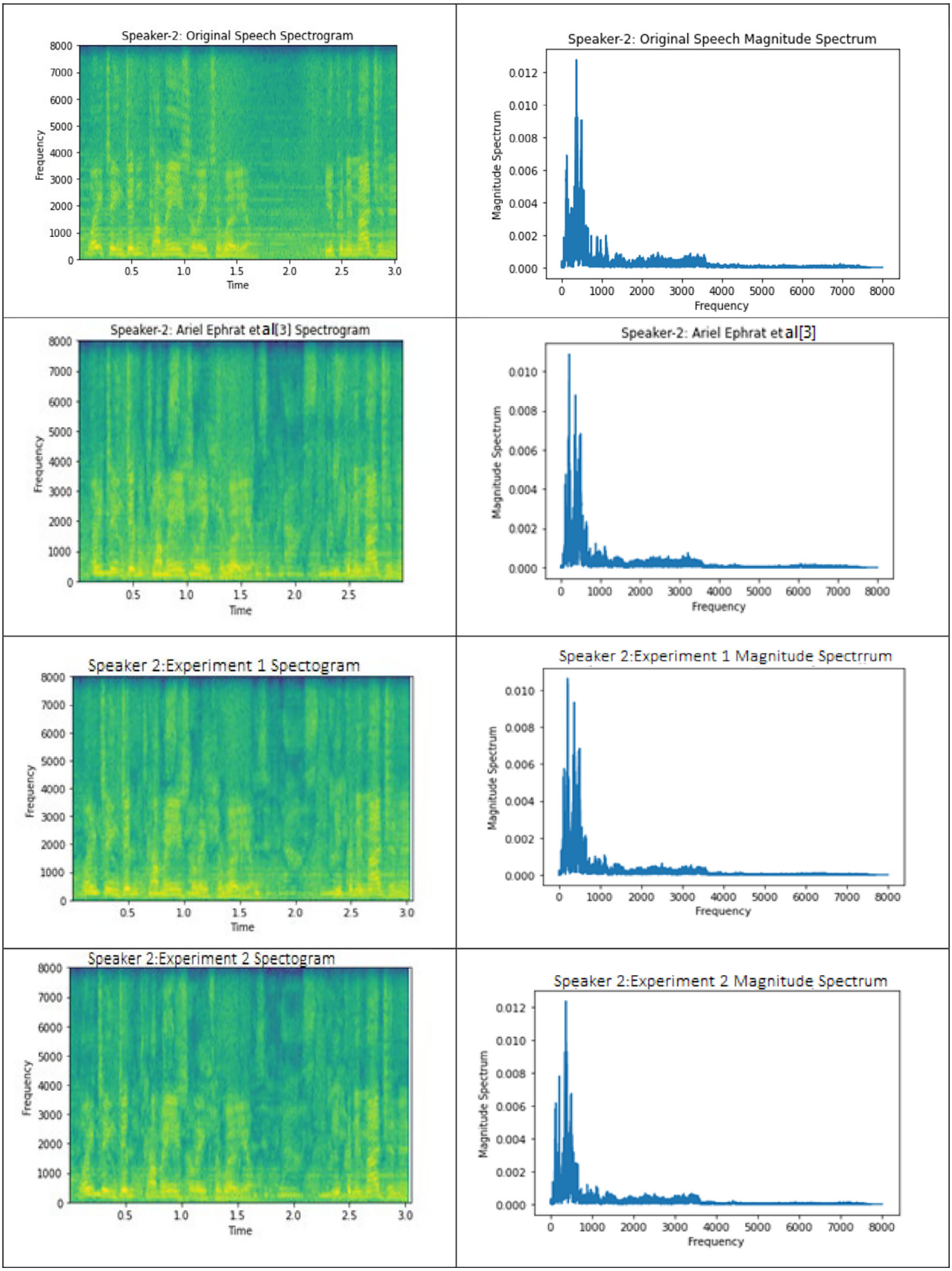

Fig. 10. Spectrogram and Magnitude Spectrum of Speaker-2.

combined, then used as an input to our proposed model. We use the mean squared error (L) between compressed clean spectrogram and the enhanced spectrogram to train our proposed model as in equation (4).

$$L(A,P) = \frac{\sum_i^N (A(i) - P(i))^2 - 0.1\sum_{j\neq i}^N (A(i) - P(j))}{S} \quad (4)$$

where A is compressed clean spectrogram, P is the enhanced spectrogram N is the number of speakers and S is the size of CRM.

We have two experiments which are implemented using TensorFlow. To evaluate our experiments, we use as a helper Python Speech Enhancement Performance Measures (Quality and Intelligibility) project [17], predicted rating of speech distortion and overall quality project [18].

4.1. Experiment 1

In this experiment, the proposed model takes visual streams of the target speaker and STFT components of mixed speech as input where the output is corresponding compressed CRM of target speaker. A combination between visual stream, speech frequency analysis and joint audio visual modules are used to build the overall architecture.

Visual stream module uses combination of inception blocks where each inception block uses the concatenation of learnable parameters in Table 1 that consist of 50 filters with size 1X1 **"where size means number of scanned neighbors information"** to track and focus on pure embedding features, 50 filters with size 3x1, 100 filters with size **5x1" number of filters is doubled as the scanned area relative to neighbors increased from 3 to 5"** to analysis the relation between neighbors features in different levels and finally maxpooling with size 3x1. Table 2 shows the UNet filters relative to each layer in speech frequency analysis module where each block in encoder and decoder path consists of two convolutions layers with the same number of filters and maxpooling2D with size 2x2 to halve input features.The joint audio-visual module consists of BLSTM with 400 neurons followed by three Dense layers with 600 neurons.

In this experiment, a batch size of six samples and Adam optimizer for 2343 batches with learning rate of $1\mathbf{X}10^{-5}$ are used to complete the separation process. It is observed that experiment 1 has better validation loss compared to Ariel Ephrat et al. [3]. It has better Speech Enhancement Performance, predicted rating of speech distortion and overall quality compared to Ariel Ephrat et al. [3]. The spectrogram and magnitude spectrum of each speaker are better than Ariel Ephrat et al. [3].

4.2. Experiment 2

In this experiment, the proposed model takes Cepstral Coefficients, Log Filterbank Energies and Spectral Subband Centroid as an input feature beside visual streams and STFT components. It is trained by using a concatenation of the audio stream analysis module beside three modules mentioned in experiment1.

In audio stream analysis module, each inception block consist of concatenation of 25 filters with size 1x1, then the number of filters is duplicated when the size of filter increased so the audio module uses 50 filters with size 3x3 and 100 filters with size 5x5. Every inception block in this module uses maxPooling with size 3x3 as in Table 1.

This experiment uses Adam optimizer with learning rate $1\mathbf{X}10^{-5}$ and batch size equal two. It has better validation loss compared to previous work and experiment 1 as in Figs. 6–8. Speech Enhancement Performance predicted rating of speech distortion are considered best values compared to previous work and Experiment 1 as shown in Tables 3 and 4. The spectrogram and magnitude spectrum of each speaker have best values compared to experiment 1 and Ariel Ephrat et al. [3] as in Figs. 9 and 10.

5. Conclusion

In this paper, we presented combined architecture between voice and audio data for speech separation which has better performance comparing to pervious work. We generate a new dataset by using 250 videos from (LRS2) Dataset which contains visible speaker and clean speech. The proposed model used a deep neural network structure consists of u-net and inception networks to learn the separation process. Our proposed model consists of two modules for audio, one for visual streams and the fourth one for feature concatenation where the visual streams data of each speaker is taken from speaker's detected faces, and the audio stream data is taken from mixture of speakers' speech in the input audio. To evaluate its performance, we used python Speech Enhancement Performance Measures as Short-time objective intelligibility, Perceptual Evaluation of Speech Quality and Frequency-weighted Segmental SNR and predicted rating of speech distortion and the predicted rating of overall quality. The experiments illustrated that our proposed model improved STOI with 11 %, PESQ with 24 %, and fwSNRseg with 16 % compared with previous works. It also improved Csig' with 13 % and 'Covl' with 18 % compared to other visual models. In our view, we will promise future directions include enhancing quality of target speaker's speech by taking the resulted signal from this system as an input to speech enhanced system and we will take in our consideration removing any background noises mixed with the speech of the target speaker.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Yannan Wang, Jun Du, Li-Rong Dai, Chin-Hui Lee, Unsupervised single-channel speech separation via deep neural network for different gender mixtures, in: *Proceeding in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA),* Jeju, Korea, 13-15 December 2016.

[2] Quan Wang, Hannah Muckenhirn, , Prashant Sridhar, Zelin Wu, John Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, Ignacio Lopez Moreno, VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking, *Proceeding in International Speech Communication Association INTERSPEECH,* Graz, Austria, 15-19 September 2019.

[3] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, Michael Rubinstein, Looking to Listen at the Cocktail Party:A Speaker-Independent Audio-Visual Model for Speech Separation, *ACM Trans. Graph. Journal*, Vol:37, No: 4, PP: 112:1-112:11,2018.

[4] Luo Yi, Mesgarani N. Conv-TasNet: surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 2019;27(8):1256–66.

[5] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, Dong Yu, Time Domain Audio Visual Speech Separation, *Proceedings in the 20th Annual Conference of the International Speech Communication Association INTERSPEECH 2019*, Graz, Austria, 15-19 September. 2019.

[6] Tomasz Grzywalski, Szymon Drgas, Application of recurrent U-net architecture to speech enhancement, *Proceeding in 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland,19-21 September 2018.

[7] Daniel Stoller, Sebastian Ewert, Simon Dixon Wave-U-Net: a multi-scale neural network for end-to-end audio source separation, *Proceeding in the 19th International Society for Music Information Retrieval Conference (ISMIR 2018*), Paris, France 23-27 September 2018.

[8] Emad M. Grais, Dominic Ward, Mark D. Plumbley, Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders, *Proceeding in 2018 26th European Signal Processing Conference (EUSIPCO),* Rome, Italy, 3-7 September 2018.

[9] Dario Rethage, Jordi Pons, Xavier Serra, A wavenet for speech denoising, *Proceedings in IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, April 15-20, 2018.

[10] Christian Szeged, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, *Proceedings in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA,7-12 June 2015.
[11] Lei X, Pan H, Huang X. A dilated CNN model for image classification. IEEE Access 2019;7:124087–95.
[12] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, Shuhei Hikosaka, Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery, *Proceedings in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV),* Lake Tahoe, NV, USA, 12-15 March 2018.
[13] https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html last accessed:20/2/2021.
[14] Mastering Computer Vision with TensorFlow 2.x: Build advanced computer vision applications using machine learning and deep learning techniques book.
[15] Florian Schroff, Dmitry Kalenichenko, James Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering, *Proceeding in E EE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA, 7-12 June 2015.
[16] https://github.com/jameslyons/python_speech_features library last accessed: 22/10/2021.
[17] https://github.com/schmiph2/pysepm last accessed: 10/11/2021.
[18] https://github.com/usimarit/semetrics/blob/master/README.md last accessed 10/11/2021.