International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 1, February, 2024, pp. 73-80
Journal homepage: http://iieta.org/journals/ria

Audio-Visual Source Separation Based Fusion Techniques

Noorulhuda Mudhafar*, Ahmed Al Tmeme

Department of Information and Communications Engineering, University of Baghdad, Baghdad 10070, Iraq

Corresponding Author Email: noor.abd2103m@kecbu.uobaghdad.edu.iq

https://doi.org/10.18280/ria.380108

Received: 12 September 2023
Revised: 10 November 2023
Accepted: 28 November 2023
Available online: 29 February 2024

Keywords:
audio-visual separation, deep learning, CNN, LSTM, complex ratio mask

ABSTRACT

A novel hybrid deep learning model for audio-visual source separation is introduced in this paper, with a specific focus on the precise isolation of a particular speaker's voice from video content. By leveraging both audio and visual characteristics, the achievement of accurate separation of the targeted speech signal is facilitated by our model. Notably, the incorporation of the speaker's facial expressions as an auxiliary cue for enhancing the extraction of their unique vocal qualities is emphasized. Proficiency in audio-visual speech separation and latent representations of distinctive speaker attributes, known as speaker embeddings, is simultaneously acquired by our model through unsupervised learning on unannotated video data. The model employed in this study is speaker-independent, wherein an initial stage of feature extraction is conducted for both audio and visual inputs prior to the subsequent deep modal analysis. The utilization of facial attribute features as an identifying code enables the identification of the speaker's frequency space or other audio properties. The model's efficacy was assessed through evaluation on the widely recognized AVspeech dataset yielding an improvement of 7.7 in terms of source-to-distortion ratio (SDR).

1. INTRODUCTION

Speech is the primary mode of communication for humans, involving the production and perception of sounds to convey information. However, speech signals can often be compromised by various factors, such as background noise or multiple speakers, which can make it challenging to understand a specific speaker in a crowded environment. Humans are often subjected to a combination of multiple conflicting sounds, necessitating their concentration on the pertinent source in order to differentiate it from the competing sources. This phenomenon, commonly referred to as the cocktail party effect or cocktail party problem [1], it describes the scenario in which an individual is attending to a single speaker while other speakers are also present. Audio-visual processing improves cocktail party problems by integrating auditory and visual information, including facial expressions and lip movements. Multimodal methods improve source localization, speech recognition, and noise robustness. Visual data, which has been demonstrated to be effective, provides a complete approach to isolating speakers in complex and congested environments, overcoming the limits of audio-only methods.

The practical uses of audio-visual source separation span across numerous fields. It facilitates the use of assistive devices for those with hearing impairments, boosts the clarity of speech during teleconferencing, and improves the post-production editing process in content development [2]. This technique also offers advantages to interactive entertainment, language learning platforms, and security systems through enhanced audio experiences, isolated pronunciations, and improved speaker identification in loud conditions, respectively. Solid evidence from observational research suggests that recognizing a speaker's facial cues enhances an individual's ability to cope with perceptual uncertainty in noisy circumstances. Visual information acquired from the speaker's face aids the listener in distinguishing and understanding speech in the presence of background noise and potential speaker confusion [2].

Early works for speech separation solely utilized the audio signal [3-7]. The audio only speech separation problem is inherently challenging due to its ambiguity, making it difficult to achieve satisfactory outcomes without additional information, for example, prior knowledge or certain microphone configuration. Another barrier encountered in audio-only speech separation is the label permutation problem, which pertains to the challenge of accurately associating each separated audio source with its respective speaker [8].

The limitation of audio only systems led to the development of Audio-Visual source separation systems, which leverage both auditory and visual information to perform the separation process. In recent studies, researchers have employed video recordings to address the challenging cocktail party problem. These approaches leverage the information contained in video data to separate the speech of a specific speaker. By utilizing the speaker's facial expressions or lip movements captured in the corresponding video, these audio-visual models demonstrate remarkable outcomes. However, it is important to note that the effectiveness of the systems that solely depends on lip movement relies heavily on the availability of high frame rate video. Therefore, their practical applicability is currently constrained to scenarios where such video resources are accessible.

In this paper we present a hybrid deep learning model that is speaker-independent which follows a two-step process, beginning with feature extraction for both audio and visual inputs, followed by deep modal analysis. The utilization of facial attribute features as a means of identification allows for the discernment of the speaker's frequency spectrum or other auditory characteristics like tone, pitch, and voice timbre. The proposed model architecture depends upon the utilization of one-dimensional Convolutional Neural Networks (1D CNNs), a Dense Neural Network, and Long Short-Term Memory (LSTMs). The objective is to extract a single speech signal from a mixture of multiple auditory components, such as other speakers and noise from the surroundings. The training of our model took place on the extensive and demanding AVspeech dataset. Remarkably, the proposed method yielded top-tier results in audio-visual source separation on the AVspeech dataset, surpassing previous state-of-the-art approaches.

The remainder of this paper is structured as follows: in Section 2 we have presented the related work, while Section 3 delves into greater detail regarding the proposed audio-visual model. Section 4 outlines the experiments conducted and presents the corresponding results. Finally, Section 5 offers a conclusion to the paper.

2. RELATED WORK

The emergence of deep learning sparked advancement in a variety of fields [9-14], and the audio visual source separation is one of them. Speech signal separation is the main goal of the majority of audio-visual separation techniques. The permutation problem, which is a significant obstacle in speech separation, involves correctly assigning separated components to different speakers.

Lu et al. [15] addressed the permutation problem by including visual information after the initial separation to fine-tune how the components are separated. The same group of researchers [16] then proposed a method that integrates visual data with an audio-based deep clustering framework, resulting in the development of an audio-visual deep clustering model that is specifically made for distinguishing speech signals.

Ephrat et al. [17] introduced a novel approach that integrates spectrograms with facial embeddings of all speakers detected in the audio sample. The purpose of this method is to design a model that can generate a complex mask that can be applied to the baseline spectrogram. This makes it possible to rebuild the complex spectrogram for each individual speaker and separate them from the other speakers in the audio mix.

Chung et al. [18] used an alternative approach in cases where solely a profile image of the speaker is accessible, instead of a video representation, a different strategy is used. In this case, learned identification embeddings produced by leveraging a pretrained model specifically designed for facial images can aid in isolating the speaker's voice from the audio mixture. The model's prior learning on facial features facilitates the successful separation process.

Gao and Grauman [19] proposed an approach that consists of audio and visual networks that are concurrently trained using a cross-modal consistency loss to ensure that the audio and visual features are properly synchronized. In essence, the method employs a sequence of facial images and a combined audio input containing lip movements. Subsequently, it conducts a prediction process to generate a complex mask, which contributes to the overall processing procedure.

Zhang et al. [20] proposed a novel Audio-Visual speech Separation (AVSS) model that focuses on isolating a target speaker's voice from a mixture of speakers. The method consists of two major steps: first, extracting speech-related visual features using joint audio-visual representation learning with supervision loss, and then improving these features via adversarial training. To capture temporal dependencies in audio-visual data, the AVSS model, which is implemented in the time domain, incorporates Temporal Convolutional Neural Networks (TCNN) blocks. The goal is to separate speech signals successfully in complex audio-visual settings.

Makishima et al. [21] proposed a model with two sub-networks for visual and auditory inputs. Embeddings are produced through different sub-networks, which are subsequently integrated and processed by a decoder network. The alignment of audio and visual data uses a loss function called cross-modal correspondence to produce better results. The blending of audio and visual components is enhanced by this alignment, producing superior outcomes.

Li et al. [22] created a novel audio-visual deep learning technique that combines auditory and visual data to detect speech from many channels. The separation filters that extract the desired speech from a mixed input of microphones and video frames are constructed by a neural network. Additionally, they utilized a multi-task architecture to simultaneously perform voice detection and dereverberation tasks.

Ong et al. [23] designed a cutting-edge method for real-time online multi-source separation that combines audio and visual data to determine the speakers' locations and separate the intended speech from background noise. They developed a deep neural network that integrates both visual and auditory features to carry out this processing.

Oya et al. [24] proposed a method of audio-visual source separation that uses bounding boxes as a form of supervision. This method requires two phases: initially, it uses object identification to generate the necessary bounding boxes, and then it uses a neural network to carry out the actual separation process.

Gu et al. [25] proposed a method for estimating beamforming filters in both the temporal and frequency domains that makes use of deep learning. This method employs a frequency-domain beamforming network and a time-domain beamforming network that work jointly to improve target speech while suppressing interfering sources. Table 1 illustrates key features, benefits, and limitations of previous related work and ours.

The work [17] is the closest to our work. The key difference is that they use two streams, one for audio and one for video, each processed by separate deep models. These streams are then fused and passed to a joint audio-visual model. In contrast, our model processes audio and visual features using a single joint deep learning model.

Table 1. Summary of related work

Method	Key Features	Benefits	Limitations
Lu et al. [15]	Utilizing optical flow features to fine-tune component separation after initial separation.	Solving the permutation problem.	Need for audio-visual synchronization.
Lu et al. [16]	AVDC network that clusters audio and visual data using convolutional and recurrent layers then utilize clustering approach to separates speech signals.	Separate speech signals under adverse situations.	Computationally expensive.
Ephrat et al. [17]	Dual-stream design for audio and video processing. Streams use different deep models for their tasks. These streams' results are then integrated and sent to a joint audio-visual model.	Ability to separate speech while the intended speaker is out of view.	Restricted in loud environment or when numerous speakers are speaking.
Chung et al. [18]	Identification embeddings were obtained using the speaker's profile image. This was done using a face-specific pretrained model to help isolate the speaker's voice from the audio mix.	Fast processing time.	Assumptions of visible and stationary faces limit the model's real-world performance.
Gao and Grauman [19]	Two distinct networks, specifically an audio network and a visual network, both of which are CNNs. The networks are jointly trained by employing a cross-modal consistency loss.	Performs well in challenging scenarios.	Limited performance When speakers overlap or are not visible in the video.
Zhang et al. [20]	Time-domain AVSS model that uses Temporal Convolutional Neural Networks (TCNN) blocks to capture audio-visual data temporal dependencies.	Enhancing speech-related visual features through adversarial training.	The degradation of performance in complex scenarios.
Makishima et al. [21]	Two sub-networks generate audio-visual embeddings, which are concatenated and delivered via a decoder network.	Cross-modal correspondence loss function that synchronizes the visual and audio data	Visual data shortage may degrade system reliability.
Li et al. [22]	Estimates target speech separation filters from multiple microphones and video frames. A multi-task framework addresses dereverberation and voice recognition tasks.	The efficacy of noise and reverberation removal	The utilization of multi-channel audio signals.
Ong et al. [23]	Real-time online multi-source separation that estimates speaker position and separates target speech from background noise using audio and visual data.	Excellent performance in noisy, reverberant situations	The utilization of many microphones and cameras.
Oya et al. [24]	Two-step approach begins with object detection to acquire bounding boxes for supervision then a neural network assists separation.	Manual annotation is not necessary.	Depends on detection model accuracy.
Gu et al. [25]	Beamforming filters estimation in time and frequency by obtaining frequency-and time-domain beamforming networks. These networks boost target speech and minimize interference.	Flexible and adaptable to diverse scenarios.	Could increase computational complexity and training time.
Our Proposed modal	Speaker-independent, utilizing a pre-features extraction stage for both audio and visual features before proceeding on to the deep modal.	Dimensionality reduction with stable training.	System complexity increases slightly.

3. PROPOSED AUDIO-VISUAL MODEL

In this work we introduced a hybrid deep learning method for audio-visual speech separation. Our approach is speaker-independent and leverages a multi-stream architecture to analyze visual streams containing detected faces alongside with auditory inputs. By integrating both visual and audio information, our model aims to effectively separate speech signals from complex audio-visual mixtures involving two speakers. A combination of Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCCs) is used for audio feature extraction. The STFT is a technique for analyzing a signal's frequency content throughout short, overlapping time intervals it provides a time-localized frequency information when the frequency components of a signal fluctuate over time [26], while MFCCs possess the capability to effectively capture significant components of the spectral composition of audio signals, namely within the frequency ranges that hold the utmost relevance to human auditory perception [27]. Principal Component Analysis (PCA) is a dimension reduction technique with the primary objective of effectively reducing the dimensionality of a given dataset and identifying its most important elements by transforming it into a new, uncorrelated space defined by principle components [28]. It reduces computing complexity, training time, and overfitting risks for high-dimensional data. STFT and MFCC help extract robust characteristics, especially in speech and speaker recognition. STFT and MFCC are vital to identify speaker features and adjust to different acoustic settings. High dimensionality in short-time Fourier transform (STFT) analysis is addressed by MFCC. A logarithmic compression and mel-scale filterbank are applied to the STFT magnitude spectrum to compress and efficiently represent the audio signal. This combination also allows multimodal fusion, which integrates visual signals to improve performance. Lossy compression is used to convert the MFCC's two-dimensional output into a one-dimensional vector representation using vector quantization. As we are dealing with large dataset Principal Component Analysis (PCA) is used for visual feature extraction before the deep learning model. This method reduces computing complexity, training time, and overfitting risks for high-dimensional data. PCA reduces noise and speeds training convergence by prioritizing signal over noise. By reducing the correlation between highly correlated features, it successfully addresses multicollinearity. The audio and visual feature vectors that are first extracted are subsequently combined and passed to the deep model. For audio-visual source separation, the proposed deep modal is made up of several layers and is mainly based on combining one-dimensional convolutional neural networks (1D CNNs) and long short-term memory (LSTMs). It is more reliable to use 1D CNNs for feature extraction because they can perform translation invariance and multimodal integration. This is because they are better at capturing localized patterns and temporal hierarchies in sequential input. The integration of LSTM enhances the model's ability to capture long-term dependencies and understand sequential context, which is of utmost importance in source separation tasks. LSTM memory cells store and utilize information effectively over long sequences, which make it easier for 1D CNNs to learn local patterns. The output of the deep model is a set of complex spectrogram masks, each of which corresponds to a detected face in the video (the proposed model is illustrated in Figure 1) We can isolate the speech signals of each speaker while efficiently suppressing any interfering signals from other sources by multiplying the noisy input spectrograms with these masks. This method allows for a distinct separation of each speaker's voices.

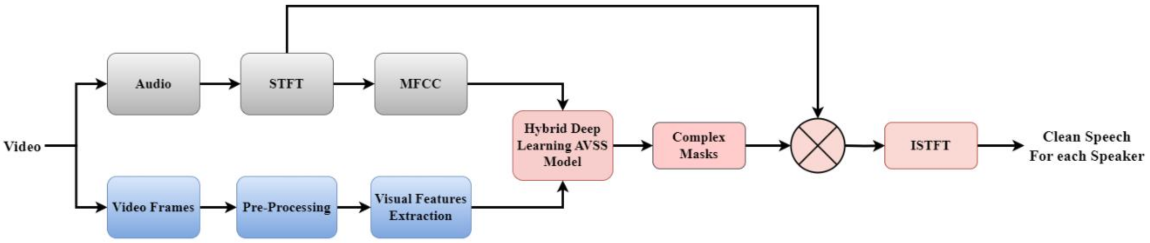
Video
Audio
STFT
MFCC
Video Frames
Pre-Processing
Visual Features Extraction
Hybrid Deep Learning AVSS Model
Complex Masks
ISTFT
Clean Speech For each Speaker

Figure 1. Proposed framework

The mixture audio signal in the time domain is given by:

$$x(t) = s_1(t) + s_2(t) \tag{1}$$

The spectrogram of the audio signal in the frequency domain at frequency bin f and time frame τ is given by:

$$X(f,\tau) = S_1(f,\tau) + S_2(f,\tau) \tag{2}$$

We predict complex ratio masks (CRM) to accomplish separation, separating the corresponding speakers' clean speech from the mixture x.

The estimated sound sources in the time domain at time t are given by:

$$\begin{aligned} \hat{s}_1(t) &= x(t) * m_1(t) \\ \hat{s}_2(t) &= x(t) * m_2(t) \end{aligned} \tag{3}$$

The estimated sound sources in the frequency domain are given by:

$$\begin{aligned} \hat{S}_1(f,\tau) &= X(f,\tau) * M_1(f,\tau) \\ \hat{S}_2(f,\tau) &= X(f,\tau) * M_2(f,\tau) \end{aligned} \tag{4}$$

3.1 Audio and video data representation

Our model takes both audio and visual data as input. We calculate the short-time Fourier transform (STFT) for 3-second audio segments to derive audio features.

After the STFT we calculate Mel-frequency cepstral coefficients (MFCC) for further audio spectral feature extraction, it captures sound signal spectral features by converting the audio spectrum into a compact, perceptually useful form. Each time-frequency (TF) bin comprises both the real and imaginary components of a complex number, both of which are utilized as input. Additionally, we apply power-law compression, which involves applying a non-linear transformation to the audio signal in order to reduce amplitude variations between loud and quiet parts, hence preventing distortion and clipping while retaining perceived quality to prevent loud noise from overpowering faint audio signals [29]. Furthermore, to ensure consistency, the same processing is done to both the noisy signal and the clean reference signal.

Given a video recording that may contain multiple speakers, first we convert the video into 75 frames per speaker, and then convert to grayscale images with a size of 128 × 720. The conversion to gray scale is highly sufficient since we are dealing with large number of image frames thus less processing cost [30].

For the face detection part, we use the multitask cascaded convolutional neural network (MTCNN) [31], which is designed to accurately detect faces and localize facial landmarks, such as the eyes, nose, and mouth in the images, we perform face detection in each frame (75 frames for each speaker in a 3-second video at 25 FPS) and extract face embedding for each face detected per frame. After face detection the frames are resized to a size of 160 × 160. The reasoning is that these embeddings preserve crucial information for recognizing countless faces while eliminating unimportant variations, like lighting differences between photos. For visual features extraction we used Principal component analysis (PCA) to characterize the pattern with the fewest number of features and to lower the dimensionality of the feature space without losing the most critical information for discriminating.

Our method generates a multiplicative spectrogram mask that represents the correlations between clean speech and background noise in both time and frequency. To achieve this, we employ a complex ratio mask (CRM), which quantifies the relationship between the complex clean and noisy spectrograms. The CRM consists of real and imaginary components, which are assessed independently in the real domain. This allows us to accurately separate the desired speech from unwanted interference, leading to improved audio quality [32]. The complex ratio mask is given by:

$$\begin{aligned} M_{real} &= \frac{X_{real} \,.\, S_{real} + X_{imag} \,.\, S_{imag}}{|X^2| + \epsilon} \\ M_{imag} &= \frac{X_{real} \,.\, S_{imag} - X_{imag} \,.\, S_{real}}{|X^2| + \epsilon} \\ M &= \left[M_{real}, M_{imag}\right] \end{aligned} \tag{5}$$

3.2 Network architecture

In our proposed model, both audio and visual features are processed jointly. After preprocessing the visual and auditory data, we fuse the extracted features using concatenation fusion, combining these features before passing them to the deep model (as depicted in Figure 2). Our deep learning model consists of a total 36 layers, 11 Convolutional layers with filter size of (16, 32, 64, 128, 256, 256, 512, 1024, 512, 128, 128) respectively for feature extraction, 11 MaxPooling layers with pool size 1 and stride of 1 to reduce the spatial size of the extracted features, 8 LeakyRelu layers with alpha of value 0.3, 2 LSTM layers to learn the long term dependencies between the sequenced data, 2 dense layers with linear activation functions which we used as a collector for strong and close to strong features, and one dense layer that works as fully connected layer with softmax as activation function and finally one flatten layer that converts the multidimensional data into a one dimensional so it can be passed to the fully connected layer. The model deep learning layers specifications and parameters are listed in Table 2.

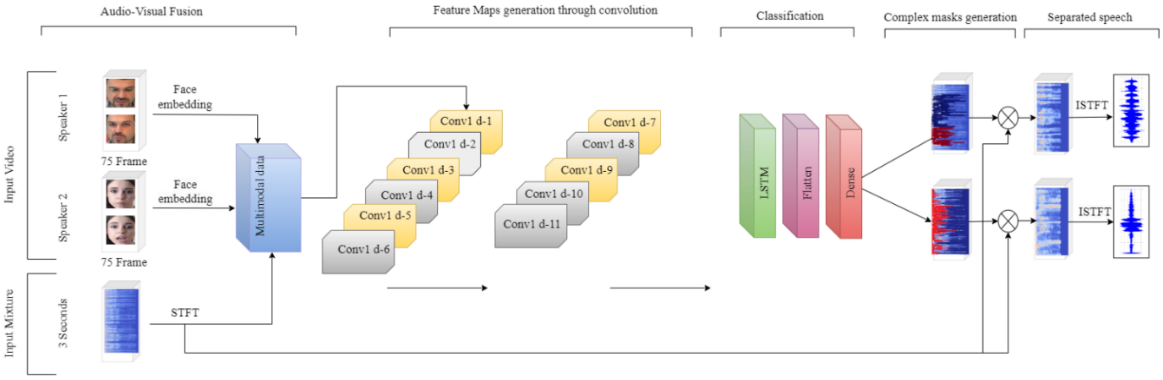

Figure 2. Deep neural network architecture

Table 2. The model layers details

Layer	Filter Size	Kernal Size	Stride	Num. of Para.
Conv1d-1	16	5	1	96
MaxPool1D-1				0
Max_Pooling1d_2				0
LeakyReLu_1				
Conv1d_2	32	3	1	1568
Max_Pooling1d_3				0
LeakyReLu_2				
Conv1d_3	64	3	1	6208
Max_Pooling1d_4				0
LeakyReLu_3				
Conv1d_4	128	3	1	24704
Max_Pooling1d_5				0
LeakyReLu_4				
Dense_1				16512
Conv1d_5	256	3	1	98560
Max_Pooling1d_6				0
LeakyReLu_5				
Conv1d_6	256	3	1	196864
Max_Pooling1d_7				0
LeakyReLu_6				
Dense_2				131584
Conv1d_7	512	3	1	786944
LeakyReLu_7				
Conv1d_8	1024	3	1	1573888
LeakyReLu_8				
Conv1d_9	512	3	1	1573376
Conv1d_10	128	5	1	327808
Max_Pooling1d_8				0
LSTM_1				1312768
Max_Pooling1d_9				0
Conv1d_11	128	5	1	327808
Max_Pooling1d_10				0
LSTM_2				22960
Max_Pooling1d_11				0
Flatten_1				0
Dense_3				18692

3.3 Implementation details

All audio data is converted to a 16kHz sampling rate. The Short-Time Fourier Transform (STFT) is performed using a Hann window with duration of 25ms, a hop length of 10ms with approximately 60% overlapping between consecutive windows, and an FFT size of 512. This calculation yields an input audio feature represented by a matrix of dimensions 257×298×2, containing scalar values.

Before training and inference, we adjust the face embeddings from each video to match a frame rate of 25 frames per second (FPS) by either removing or duplicating embeddings, this result in a visual stream with 75 face embeddings as the input. During training, we have employed 100 epoch with a batch size of 64 samples and conduct training for 5 million steps (batches) using the Adam optimizer [33], which adjusts learning rates dynamically and incorporates momentum to efficiently optimize model parameters during training, with a learning rate of 0.001.

4. EXPERIMENT AND RESULTS

4.1 Dataset

To evaluate and train our system, we have utilized the AVspeech dataset [17], which is a comprehensive audio-visual dataset featuring an abundance of speech recordings without any interfering background noises. Each segment within this dataset varies in length from 3 to 10 seconds and prominently features a single speaking individual whose face is clearly visible in the video, accompanied by audible speech in the soundtrack.

This dataset comprises a substantial collection of more than 4700 hours of video clips, showcasing approximately 150,000 unique speakers from diverse language and demographic backgrounds. The richness and diversity of this dataset provide valuable resources for assessing and enhancing the performance of our system.

Our model is trained on segments of 3 seconds in length; however, during predictions, our model is capable of handling video segments of varying lengths. In the experiment, a partitioning scheme was employed where 70% of the available dataset was designated as the training set, while the remaining 30% was allocated as the test set, no validation set was utilized, as neither parameter adjustment nor early stopping were conducted.

4.2 Evaluation

The evaluation of the proposed method requires precise measurements of signal-to-distortion (SDR) and source-to-interference ratio (SIR), which are critical metrics for measuring the quality of separated signals. We compared our findings to those obtained in a previous studies conducted by [17, 20].

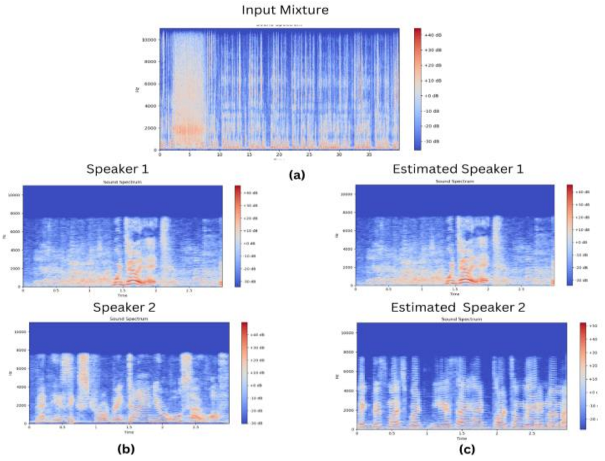

Figure 3. Spectrograms (a) The audio mixture. (b) Clean speaker 1, 2. (c) The estimated sources

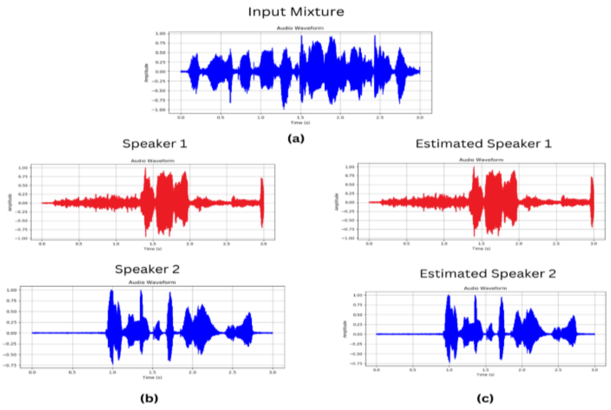

Figure 4. Waveforms (a) The audio mixture. (b) Clean speaker 1, 2. (c) The estimated sources

The results of our evaluation, as shown in Table 3, demonstrated that our suggested method achieved better SDR and SIR values on the challenging AVspeech dataset. This performance indicates an improvement in the quality and accuracy of audio-visual source separation. The spectrograms and waveforms of the original, unprocessed audio as well as the estimated outputs of our model are clearly demonstrated in Figures 3 and 4.

Table 3. Comparing with prior AVspeech dataset results

Separation Results on the AVspeech Dataset		
Model	**SDR**	**SIR**
Ephrat et al. [17]	10.3	-
Zhang et al. [20]	12	-
Proposed model (audio only)	17.8	15.2
Proposed model (audio-visual)	19.7	16

5. CONCLUSIONS

This paper presents a speaker-independent hybrid deep learning model that employs a two-stage procedure. The first stage involves extracting features from both audio and visual inputs through the use of STFT with MFCC for audio features and PCA for visual features, while the second stage involves conducting deep learning modal analysis. The integration of facial attribute data improves the ability to distinguish between speakers by capturing not just vocal characteristics but also subtle facial expressions. This leads to a more thorough and relatable comprehension of the distinct audio sources. The model architecture described in this study relies on the incorporation of one-dimensional Convolutional Neural Networks (1D CNNs), a Dense Neural Network, and Long Short-Term Memory (LSTMs). The primary goal is to isolate a single speech signal from a mixture of other auditory elements, including additional speakers and ambient noise.

The effectiveness of the proposed model was assessed using the well-known AVspeech dataset, achieving a notable source-to-distortion ratio (SDR) of 19.7, yielding an improvement of 7.7 SDR from recent work on the AVspeech dataset.

ACKNOWLEDGMENT

This work is aided by the information and communications engineering department at Al-Khawarizmi Engineering College, University of Baghdad.

REFERENCES

[1] Bednar, A., Lalor, E.C. (2020). Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. NeuroImage, 205: 116283. https://doi.org/10.1016/j.neuroimage.2019.116283

[2] Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D., Jensen, J. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 1368-1396. https://doi.org/10.1109/TASLP.2021.3066303

[3] Al-Tmeme, A., Woo, W.L., Dlay, S.S., Gao, B. (2016). Underdetermined convolutive source separation using GEM-MU with variational approximated optimum model order NMF2D. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1): 35-49. https://doi.org/10.1109/TASLP.2016.2620600

[4] Woo, W.L., Dlay, S.S., Al-Tmeme, A., Gao, B. (2018). Reverberant signal separation using optimized complex

sparse nonnegative tensor deconvolution on spectral covariance matrix. Digital Signal Processing, 83: 9-23. https://doi.org/10.1016/j.dsp.2018.07.018

[5] Al-Tmeme, A., Woo, W.L., Dlay, S.S., Gao, B. (2018). Single channel informed signal separation using artificial-stereophonic mixtures and exemplar-guided matrix factor deconvolution. International Journal of Adaptive Control and Signal Processing, 32(9): 1259-1281. https://doi.org/10.1002/acs.2912

[6] Al Tmeme, A., Woo, W.L., Dlay, S.S., Gao, B. (2015). Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models. The Journal of the Acoustical Society of America, 138(6): 3411-3426. https://doi.org/10.1121/1.4923156

[7] Amer, R., Al Tmeme, A. (2021). Hybrid deep learning model for singing voice separation. Mendel, 27(2): 44-50. https://doi.org/10.13164/mendel.2021.2.044

[8] Isik, Y., Roux, J.L., Chen, Z., Watanabe, S., Hershey, J.R. (2016). Single-channel multi-speaker separation using deep clustering. arXiv Preprint arXiv: 1607.02173. https://doi.org/10.48550/arXiv.1607.02173

[9] Moraboena, S., Ketepalli, G., Ragam, P. (2020). A deep learning approach to network intrusion detection using deep autoencoder. Revue d'Intelligence Artificielle, 34(4): 457-463. https://doi.org/10.18280/ria.340410

[10] Thayumanasamy, I., Ramamurthy, K. (2022). Performance analysis of machine learning and deep learning models for classification of alzheimer's disease from brain MRI. Traitement du Signal, 39(6): 1961-1970. https://doi.org/10.18280/ts.390608

[11] Al-Akkam, R.M.J., Altaei, M.S.M. (2022). Plants leaf diseases detection using deep learning. Iraqi Journal of Science, 801-816. https://doi.org/10.24996/ijs.2022.63.2.34

[12] Mahmood, I.N., Abdullah, H.S. (2022). Telecom churn prediction based on deep learning approach. Iraqi Journal of Science, 2667-2675. https://doi.org/10.24996/ijs.2022.63.6.32

[13] Jameel, H.K., Dhannoon, B.N. (2022). Gait recognition based on deep learning. Iraqi Journal of Science, 63(1): 397-408. https://doi.org/10.24996/ijs.2022.63.1.36

[14] Hussein, N.A.K., Al-Sarray, B. (2022). Deep learning and machine learning via a genetic algorithm to classify breast cancer DNA data. Iraqi Journal of Science, 3153-3168. https://doi.org/10.24996/ijs.2022.63.7.36

[15] Lu, R., Duan, Z., Zhang, C. (2018). Listen and look: Audio-visual matching assisted speech source separation. IEEE Signal Processing Letters, 25(9): 1315-1319. https://doi.org/10.1109/LSP.2018.2853566

[16] Lu, R., Duan, Z., Zhang, C. (2019). Audio-visual deep clustering for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11): 1697-1712. https://doi.org/10.1109/TASLP.2019.2928140

[17] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv Preprint arXiv: 1804.03619. https://doi.org/10.48550/arXiv.1804.03619

[18] Chung, S.W., Choe, S., Chung, J.S., Kang, H.G. (2020). Facefilter: Audio-visual speech separation using still images. arXiv Preprint arXiv: 2005.07074. https://doi.org/10.21437/Interspeech.2020-1065

[19] Gao, R., Grauman, K. (2021). Visualvoice: Audio-visual speech separation with cross-modal consistency. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15490-15500. https://doi.org/10.1109/CVPR46437.2021.01524

[20] Zhang, P., Xu, J., Shi, J., Hao, Y., Qin, L., Xu, B. (2021). Audio-visual speech separation with visual features enhanced by adversarial training. In 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1-8. https://doi.org/10.1109/IJCNN52387.2021.9533660

[21] Makishima, N., Ihori, M., Takashima, A., Tanaka, T., Orihashi, S., Masumura, R. (2021). Audio-visual speech separation using cross-modal correspondence loss. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6673-6677. https://doi.org/10.1109/ICASSP39728.2021.9413491

[22] Li, G., Yu, J., Deng, J., Liu, X., Meng, H. (2022). Audio-visual multi-channel speech separation, dereverberation and recognition. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6042-6046. https://doi.org/10.1109/ICASSP43922.2022.9747237

[23] Ong, J., Vo, B.T., Nordholm, S., Vo, B.N., Moratuwage, D., Shim, C. (2022). Audio-visual based online multi-source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 1219-1234. https://doi.org/10.1109/TASLP.2022.3156758

[24] Oya, T., Iwase, S., Morishima, S. (2022). The sound of bounding-boxes. In 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, pp. 9-15. https://doi.org/10.1109/ICPR56361.2022.9956384

[25] Gu, R., Zhang, S.X., Zou, Y., Yu, D. (2022). Towards unified all-neural beamforming for time and frequency domain speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 849-862. https://doi.org/10.1109/TASLP.2022.3229261

[26] Vincent, E., Virtanen, T., Gannot, S. (2018). Audio Source Separation and Speech Enhancement. John Wiley & Sons. http://doi.org/10.1002/9781119279860

[27] Abdul, Z.K., Al-Talabani, A.K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10: 122136-122158. https://doi.org/10.1109/ACCESS.2022.3223444

[28] Ghojogh, B., Samad, M.N., Mashhadi, S.A., Kapoor, T., Ali, W., Karray, F., Crowley, M. (2019). Feature selection and feature extraction in pattern analysis: A literature review. arXiv Preprint arXiv: 1905.02845. https://doi.org/10.48550/arXiv.1905.02845

[29] Reiss, J.D., McPherson, A. (2014). Audio Effects: Theory, Implementation and Application. CRC Press. https://doi.org/10.1201/b17593

[30] Bala, R., Braun, K.M. (2003). Color-to-grayscale conversion to maintain discriminability. Color Imaging IX: Processing, Hardcopy, and Applications, SPIE, 5293: 196-202. https://doi.org/10.1117/12.532192

[31] Zhang, K., Zhang, Z., Li, Z., Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10): 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[32] Wang, Z., Wang, X., Li, X., Fu, Q., Yan, Y. (2016). Oracle performance investigation of the ideal masks. In

2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1-5. https://doi.org/10.1109/IWAENC.2016.7602888

[33] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv Preprint arXiv: 1412.6980. https://doi.org/10.48550/arXiv.1412.6980