Al-Khwarizmi Engineering Journal, Vol. 19, No. 4, December, (2023)
P. P. 42- 55

Al-Khwarizmi Engineering Journal

An Overview of Audio-Visual Source Separation Using Deep Learning

Noorulhuda Mudhafar Sulaiman* **Ahmed Al Tmeme ****
Mohammed Najah Mahdi ***
*,** *Department of Information and Communications Engineering/Al-Khwarizmi College of Engineering/ University of Baghdad/ Baghdad/ Iraq*
***ADAPT Centre/ School of Computing/ Dublin City University/ Dublin D09 DXA0/ Ireland*
Corresponding **A**uthor *Email: noor.abd2103m@kecbu.uobaghdad.edu.iq
**Email: asattar@kecbu.uobaghdad.edu.iq
***Email: mohammed.mahdi@adaptcentre.ie

(Received 11 April 2023; accepted 18 June 2023)
https://doi.org/10.22153/kej.2023.06.003

Abstract

In this article, the research presents a general overview of deep learning-based AVSS (audio-visual source separation) systems. AVSS has achieved exceptional results in a number of areas, including decreasing noise levels, boosting speech recognition, and improving audio quality. The advantages and disadvantages of each deep learning model are discussed throughout the research as it reviews various current experiments on AVSS. The TCD TIMIT dataset (which contains top-notch audio and video recordings created especially for speech recognition tasks) and the Voxceleb dataset (a sizable collection of brief audio-visual clips with human speech) are just a couple of the useful datasets summarized in the paper that can be used to test AVSS systems. In its basic form, this review aims to highlight the growing importance of AVSS in improving the quality of audio signals.

Keywords: *Audio-visual, source separation, deep learning, CNN, datasets.*

1. Introduction

The process of distinguishing particular audio sources from a mixture of audio signals using visual indicators as further information is known as audio-visual source separation. This method differs from conventional ones that exclusively rely on the audio stream for source separation [1-5]. In other words, it is a technique that utilizes both auditory and visual information to separate individual sound sources from a mixed audio signal. AVSS algorithms generally take an input signal containing multiple sound sources and utilize visual information, such as lip movement or facial expressions, to separate the individual sources. The visual information can be captured from various sources, such as video recordings or real-time cameras. The main advantage of audio-visual source separation over traditional audio-only approaches is using visual signals to improve the separation quality. For example, visual cues can help separate speech from background noise [46] or separate multiple speakers in a video recording [16-17].

One advanced method for audio-visual source separation involves the use of deep learning techniques [10-12,14,22,34,37-38].

Here are some specific reasons why audio-visual source separation is important:

1-Enhancing speech intelligibility: In scenarios where speech is degraded by noise or overlapping sources, audio-visual source separation can be used to isolate the target speaker's voice and improve speech intelligibility [10-12, 17, 20].

2-Improving speech recognition: Audio-visual source separation can also improve speech recognition systems' accuracy, by separating out individual speakers and reducing interference

from other sources. In [25] the authors utilized a separation system to separate voices from a mixed audio signal, then fed the separated voices into an automatic speech recognition (ASR) system for further processing. In [40] the authors combined the spatial information from a microphone array with visual information from a camera to improve the separation and recognition of multiple speakers in a noisy and reverberant environment.
3-Enabling better audio and video processing: Audio-visual source separation can be used to separate out different sources of sound in a video recording, which can enable better post-processing of both the audio and video [23, 28, 47]. Overall, audio-visual source separation can help address some of the challenges in speech recognition by improving the quality and intelligibility of speech signals. Additionally, the development of audio-visual source separation methods has led to the creation of large-scale datasets that can be used for further research in related fields [17, 50-60].

2. Deep-learning for Audio-Visual Source Separation

Deep learning techniques have emerged as powerful tools in a wide range of applications [6-9]. Deep learning models consist of multiple processing layers that learn different representations of data to solve various problems. The field of audio-visual source separation witnessed the development of various deep learning approaches to help solve the audio separation problem in the presence of visual cues. Each approach has its advantages and limitations, depending on the application and data availability. In Figure 1, deep learning methods that are commonly used for AVSS are presented and explained as follows:

- Deep neural network (DNN): this deep learning architecture consists of multiple hidden layers for data processing. It is widely used for image classification, speech recognition, or natural language processing applications.
- Convolutional neural network (CNN): this deep learning architecture has the ability to detect patterns in images. It is widely used for image processing and object detection.
- Long short term memory (LSTM): this deep learning architecture can learn long-term dependencies between time-sequenced data. It is widely used for speech recognition and data analysis.
- Bidirectional long short term memory (BLSTM): this deep learning architecture can learn long-term dependencies between time sequential data in both backward and forward directions, making it more useful in sequential pattern analysis.

Deep learning approaches are used to extract audio and visual features and map them to different output signals. One approach is joint audio-visual processing, which processes audio and visual models jointly and uses the output to separate the target audio source, such as in [16, 19, 40]. Examples of joint audio-visual processing approaches include CNNs and recurrent neural networks (RNNs) that input audio and visual data [16]. Another approach is the multi-task learning approach, where audio and visual modalities are treated as separate tasks, and a deep neural network is trained to perform both tasks simultaneously.

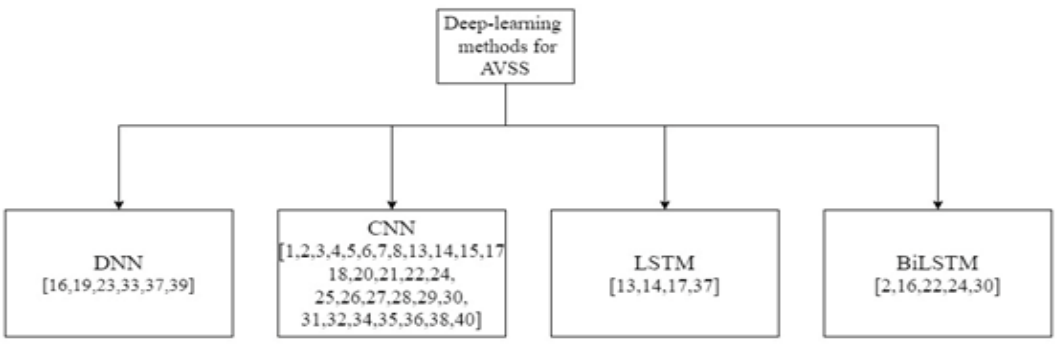

Fig. 1. Deep-learning methods for Audio visual source separation.

The pros and cons of the deep learning approach are listed in Table 1. It can be observed from Table 1 that the pros and cons reflect common observations and considerations in the field. However, it's important to note that variations and specific trade-offs can vary depending on the deep learning architectures, datasets, and problem settings used in different studies.

Table 1.
pros and cons of deep learning approaches

Pros	Cons
Can handle complex and high-dimensional data	Requires large amounts of data for training
Can learn complex relationships between audio and visual modalities	It can be computationally expensive and require powerful hardware
Can generalize well to new data	Can suffer from overfitting to training data
Can capture both low-level and high-level features	It may require significant effort for hyperparameter tuning and model selection

3. Recent AVSS Works

Different systems for distinguishing audio-visual inputs have been developed as a result of advancements in deep learning. These systems employ a blend of deep learning techniques, including DNNs, CNNs, LSTMs, and BLSTMs. The next part of this section is devoted to categorizing each of these works based on the specific deep learning technique that they use. An organized method for understanding and comparing different approaches to audio-visual source separation is provided by this classification. It also helps evaluate how certain topologies affect separation approaches, which helps to clarify the advantages and disadvantages of each.

The DNN-based works will be covered in the subsection that follows.

Takahashi et al [10] suggested a unique method that included end-to-end speech recognition in order to enhance voice separation performance. A DNN trained on mixed audio signals and their transcriptions is used in the suggested voice separation method, which incorporates end-to-end speech recognition to direct the separation process. The temporal dependencies are modeled and the input signal is converted into a feature representation using deep learning layers like fully connected and LSTM. Although the technique does not necessitate the pre-separation of audio and visual elements, its effectiveness may be impacted by failures in voice recognition.

Chung et al. [11] proposed a FaceFilter approach that enhances speech in a noisy audio recording using a single facial image of the speaker, employing a DNN with three subnetworks that extract and combine visual and audio features to produce a mask for the target speaker's voice. The method is advantageous in scenarios with limited visual data and faster processing times, but may be limited by assumptions of visible and stationary faces, leading to negative impacts on model performance in real-world scenarios, and may not be effective in partially obscured or poor lighting conditions.

Gu et al. [12] proposed a deep learning-based approach for target speech separation from multi-modal and multi-channel mixtures of speech and noise signals using audio, visual, and spatial information to improve accuracy. DNN combines audio spectrogram, visual feature map, and spatial inter-channel phase differences to predict the ideal ratio mask for speech separation, trained end-to-end for efficient separation without complex preprocessing.

Zhang et al. [13] proposed an approach to improve the MVDR beamformer for multi-channel target speech separation by replacing the conventional covariance matrix estimation with a DNN trained on a large-scale speech corpus. The approach is more robust for non-stationary noise and acoustic environments. However, the computational cost of training the DNN and computing the MVDR beamformer weights can be high.

Li et al. [14] Proposed an AV deep learning approach for multi-channel speech separation by jointly modeling audio-visual cues. It includes a neural network that estimates separation filters for target speech from multiple microphones and video frames, and a multi-task framework for dereverberation and speech recognition. Advantage: effective in removing noise and reverberation. Disadvantages: requires multi-channel audio signals, large amounts of training

data, and computational resources for deep neural networks.

Ong et al. [15] Proposed a real-time online multi-source separation method that uses both audio and visual information to estimate the position of speakers and separate target speech from background noise. They developed a deep neural network combining audio and visual processing features. This approach shows improved performance, especially in noisy and reverberant environments. However, it requires multiple microphones and cameras to capture the audio and visual data, and the quality of the separated speech may vary depending on the quality of the captured audio and video.

Furthermore, in this subsection, the CNN-based works will be discussed.

Afouras et al. [16] proposed a deep learning model for audio-visual speech enhancement using a two-stream CNN architecture to predict a cleaned audio signal by taking audio and visual inputs. Advantages include using both modalities, improving accuracy and outperforming existing models. Disadvantages include the requirement for both inputs and the need for a large amount of training data and computational resources.

Ephrat et al. [17] proposed an audio-visual deep neural network approach for separating multiple speakers in noisy environments. The approach consists of two streams: an audio stream that predicts ideal ratio masks (IRM) and a video stream that predicts attention masks to capture the spatial distribution of speakers. These streams are then combined using a multi-modal fusion network to generate the final separated speech signals. The network utilizes CNN and BLSTM layers to extract meaningful features and model temporal dynamics. An advantage of this approach is its ability to separate speech even when the target speaker is not visible in the camera's field of view. However, it relies on synchronized audio and video data, and its performance may be limited in extremely noisy environments or when multiple speakers are speaking simultaneously.

Hou et al. [18] proposed deep learning-based audio-visual speech enhancement using CNNs for audio and visual processing. The multimodal fusion network combines the two branches to enhance speech intelligibility with lip movements. Advantages include handling noisy speech signals and avoiding handcrafted features. However, using lip movements as a visual modality may not be suitable for all scenarios.

Lu et al. [19] proposed an approach that uses deep neural networks for audio-visual speech separation. The network includes CNN and RNN layers for audio and CNN layers for visual processing. An attention-based multimodal fusion network is estimated by combining the two branches, and estimate the ideal ratio mask for speech separation. The main disadvantage is the need for aligned audio-visual data.

Zhao et al. [20] proposed an audio-visual source separation method using a CNN. It uses a two-stage approach: first, the CNN predicts the magnitude spectrogram of the target speech signal from visual features; second, the predicted spectrogram is used to estimate the speech signal. Visual features are extracted from a pre-trained network and fine-tuned on mixture frames.

It can separate speech even in noisy or reverberant environments. However, it requires a large amount of training data.

Gabbay et al. [21] proposed a method for enhancing speech signals in noisy environments using visual information from a video recording of speakers. They used deep learning to extract facial features from a CNN and integrated them into a deep RNN for speech separation. Achieved state-of-the-art performance, but limitations include quality of video, accuracy of facial feature extraction, and computational resources required.

Gogate et al. [22] proposed a method that uses a CNN to extract visual features and an RNN to process audio signals to estimate a mask separating the speaker's voice from background noise. Deep learning improves mask estimation and speech separation. The method is speaker-independent but computationally expensive for real-time use.

Morrone et al. [23] proposed a method for enhancing speech using audio and visual information in multi-talker environments. The deep neural network architecture includes audio and visual branches, with CNN and RNN in the audio branch and face detection and landmark localization algorithms in the visual branch. The attention mechanism combines the two branches for final results. The disadvantage is that the accuracy of facial landmark detection can affect performance.

Wu et al. [24] proposed a method to separate speech signals of multiple speakers in an audio-visual recording using deep neural networks. The method jointly learns audio and visual features of the speech signals and performs separation in the time domain. Visual features are extracted using CNN and temporal dynamics are modeled using LSTM. A deep clustering network is used to fuse audio and visual features. The method preserves the temporal characteristics of speech signals, but

may not work well in cases with similar speakers or occlusions in the visual data and may have long processing times.

Gogate et al. [25] proposed that the DNN model combines binaural cues and visual input to improve the speech quality of unknown speakers in a noisy environment. Multi-stream CNN considers the temporal and spatial dynamics of binaural audio and visual lip images to estimate spectral masks for each channel. Binaural audio signals preserve spatial information, improving speech separation. Disadvantages include the complexity of training due to supervised and unsupervised learning methods.

Nguyen et al. [26] proposed a method to improve audio-visual sound source separation using object-level priors. They used a deep neural network with convolutional and deconvolutional layers to learn audio and visual features and estimate masks for each sound source. The method involves two stages: object-level localization and object-aware separation. In the first stage, an object detection model is used to localize the sound sources in the video frames. The audio signals are separated in the second stage based on the estimated object locations and their corresponding masks. The advantage is that it can better exploit object-level information and improve separation performance. Still, the method relies on an object detection model, which can be challenging in complex scenes with multiple objects or occlusions.

Li et al. [27] presented a model for extracting speech signals from mixed sound sources in a single-channel recording using both audio and visual data. It makes use of an attention mechanism and is composed of fully connected layers, bidirectional LSTM, and CNN. For best results in terms of performance, it may require a large amount of training data and be computationally costly.

Gu et al. [28] presented a deep learning method that combines audio and visual data to distinguish desirable speech from distracting sources. In order to improve separation performance, the model integrates an optical encoder network for extracting visual information, a CNN for processing audio features, and a multi-modal fusion network. Utilizing both aural and visual modalities, the method gains a more comprehensive representation of the input signals. Multiple microphones can also be used by the model to acquire spatial information and improve separation accuracy. It is important to keep in mind, nevertheless, that this method might need a significant quantity of training data and that using several modalities and channels could result in higher computational complexity.

Gan et al. [29] proposed an audio-visual source separation approach using a deep neural network that leverages a musician's body movements to separate instruments. Two-part model: video analysis network and visual-audio separation network. Video analysis extracts keypoint coordinates and global context using context-aware Graph CNN to generate latent representation. Visual-audio separation uses visual features to separate audio. Does not require prior knowledge or training data, but assumes each instrument is associated with specific body movements. The method requires a motion capture system, but it is not practical in real-world scenarios.

Zhu et al. [30] proposed an audio-visual source separation method consisting of two stages: visual feature extraction and sound source separation. A pre-trained CNN extracts visual features from the input video frames, which are used to compute attention maps for each sound source. A cascaded opponent filter network (COFN) is used for sound source separation, with attention maps guiding the separation process. The approach improves separation performance, but requires a pre-trained CNN and significant computational resources for training.

Tan et al. [31] proposed a two-stage strategy for speech enhancement that included a separation module for noise reduction and a dereverberation module for room reverberation suppression. A CNN is used in the architecture for audio-visual feature extraction, while a BLSTM is used for dereverberation. The approach has the benefit of being able to handle weakly labeled data and uses a multimodal fusion layer for feature fusion. However, accuracy depends on the strength of audio-visual connection, which in complicated scenarios may not always be dependable.

Qu et al. [32] proposed a method for improving speech separation that combines auditory and visual modalities. A CNN is used to retrieve features from both modalities, which are subsequently integrated into a shared embedding space with the help of a multi-modal fusion network. Then, using the joint embedding space as a reference, a target speech separation network is trained to extract the target speaker's speech from the mixture. The approach is adaptable and

may be applied with a variety of input modalities; however, it may need high-quality video input, which is not always feasible in real-world scenarios.

Rahman et al. [33] proposed a weakly supervised audio-visual sound source detection and separation method using video-level labels for joint learning of visual and auditory segments. It consists of a video frame semantic segmentation path and a spectrogram mask prediction path implemented using an attention U-Net architecture. The final mask is constructed using multi-modal audio-visual features. It can handle multiple sound sources and does not require precise temporal annotations. It has a lower separation performance than fully-supervised methods, and the quality of visual features can affect the separation performance.

Gao et al. [34] proposed a deep learning approach for audio-visual speech separation. It consists of two networks: an audio network and a visual network, both are CNNs. The networks are jointly trained using a cross-modal consistency loss. The approach demonstrates good performance even in challenging scenarios but requires paired audio and video data during training. However, its performance may be limited in cases of significant overlap between speakers or when the speakers are not visible in the video.

Majumder et al. [35] proposed an audio-visual source separation approach using a robot equipped with a camera and microphones. A deep learning model consisting of CNN predicts each sound source's arrival direction, and the robot moves its head to focus on the desired sources. Audio signals are then separated using a deep neural network. The approach allows for real-time audio-visual source separation in dynamic environments but may be disruptive in quiet or crowded spaces.

Liu et al [36] proposed a two-stage feature fusion approach for audio-visual speech separation. Stage one uses a CNN to extract features from both audio and visual modalities, fused using a gated fusion mechanism. Stage two uses an RNN to separate mixed speech signals into individual sources. Advantages include handling complex scenes and improved performance. Disadvantages include dependence on synchronization and increased computational complexity.

Tian et al. [37] proposed an audio-visual sound source separation approach that uses visual information to separate target sound sources. Object candidates are obtained using faster R-CNN, and a cyclic co-learning framework jointly optimizes sounding object visual grounding (SOVG) and sound separation (SS) tasks. The SOVG module uses a pre-trained object detection network to locate the target sounding object and project visual features onto a shared embedding space with audio features. The SS module separates sound sources based on the shared embedding space, and the estimated source is used to update the SOVG module for better visual grounding. The approach can handle multiple sound sources but may be limited by the object detection network's accuracy and visual information quality.

Makishima et al. [38] proposed a deep learning method that utilizes both auditory and visual information to separate speech signals from an audio-visual mixture. It consists of two sub-networks, a visual network and an auditory network, which generate embeddings that are concatenated and passed through a decoder network. The model introduces a cross-modal correspondence loss function to align auditory and visual information and improve performance. One disadvantage of the approach is its high dependency on visual data, making it unreliable when it is limited.

Nguyen et al. [39] proposed an unsupervised technique based on audio-visual generative modeling for the purpose of speech separation. They used a variational auto-encoder (VAE) to learn a variable generative model from clean speech. The model uses visual information (lip movements) associated with each speaker to separate the audio streams through a visual network. The system does not require the clear identification of acoustic signals, but its performance is highly dependent on visual information and has a high computational cost.

Lee et al. [40] proposed a deep learning modality that uses CNNs to extract audio and visual information. The outputs of CNNs are then integrated into a joint representation by using a cross-modal affinity function. The target speech signal is reconstructed using the combined representation. The approach relies purely on the connection between the audio and visual signals and does not require prior information of the speakers' speech patterns or microphone placements. One disadvantage of the approach is its high dependency on the strong coupling between audio and visual data input.

Zhu et al. [41] proposed an object category-based technique for visual sound source separation. At first, each pixel in the video recording is categorized into a specific item

category by using a pre-trained object identification network. The audio stream and item category map are then sent into a neural network, which is trained to perform the separation. The approach can increase separation performance by utilizing object category information but it depends on its availability and can only conduct single-frame separation without taking temporal information into account.

Gu et al. [42] proposed a deep learning method to estimate beamforming filters for time and frequency domains. The method consists of a frequency-domain beamforming network and a time-domain beamforming network, which enhance the target speech and suppress interfering sources. The approach is more flexible and adaptable to different situations than traditional beamforming methods but requires a two-stage approach, which may increase computational complexity and training time.

Oya et al. [43] proposed a method for audio-visual source separation that uses bounding boxes as supervision. The method has two stages: object detection to obtain bounding boxes, and a neural network for separation. It does not require manual annotation but relies on the accuracy of the detection model.

Zhu at al. [44] proposed a method for audio-visual sound source separation and localization using self-supervised motion representations. Learns motion features from video data without explicit annotations and incorporates them into a deep neural network for separation and localization. The visual motion pattern of sound sources is used to estimate location.

Learns motion features with CNN to predict future frames from past frames.

It is scalable, efficient, and eliminates the need for manual feature engineering. Performance depends on video quality and resolution, which may be limited in real-world applications with low-resolution or noisy videos.

Pham et al. [45] proposed a novel approach for training a cross-modal retrieval framework using video data. The framework is refined through three loss functions: separation loss, object-consistency loss, and cross-modal loss. By incorporating visual guidance, this method enhances source separation performance, especially in situations where the audio signal is degraded or the sources are closely positioned. However, it is crucial to remember that this method depends on auditory and visual data, which might not always be available or practical in real-world situations.

Other authors created a new network architecture for audio-visual source separation that is also based on deep learning:

Xu et al. [46] proposed a method using a neural network called Minus-Plus Net (MPN) to separate sound sources from mixture signals using audio and visual input. The MPN recursively separates the sources in a coarse-to-fine manner and can handle complex mixtures of sources. The approach has advantages in improved performance and the separation of multiple sources, but relies on pre-segmentation and has longer processing times.

Zhao et al. [47] proposed a deep learning-based method for audio-visual sound separation using visual motion features. The approach involves two stages: training a Motion-to-Sound Network (MTSN) to map motion features to sound sources, then refining the estimated sources using a Sound Refinement Network (SRN) with spectrograms and original audio. The method is flexible and adaptable but may not work well in low-light or noisy environments and can struggle with visually similar or closely located objects.

Lu et al. [48] proposed an audio-visual method to separate speech signals from a mixture of multiple speakers using a deep neural network architecture called AVDC. The AVDC network clusters audio and visual features using convolutional and recurrent layers, then applies a clustering algorithm to separate speech signals. The method works with one modality and is computationally expensive, but can separate speech signals in challenging scenarios.

Gao et al. [49] proposed a method for audio-visual source separation using a co-separation network, which learns to separate sounds from different visual objects by jointly modeling audio and visual features. The approach is real-time and generalizable to different scenarios without prior knowledge of sound sources. Still, it requires synchronized audio and video inputs and may degrade in the presence of occlusions or incomplete visual information.

Table 2 illustrates deep-learning features for AVSS in terms of complexity, data availability, and other features.

Table 2.
Deep learning features for AVSS

Feature	Deep Learning-based AVSS
Type of Model	Neural networks (CNNs, RNNs, GANs)
Training Data Availability	Require large amounts of labeled data
Model Complexity	Complex models with high computational requirements
Performance	Can achieve state-of-the-art performance on many tasks
Robustness to Noise	Can handle complex, noisy audio-visual scenes
Generalization	Can generalize well to unseen data
Training Time	Longer training times due to the complexity of the models

4. Datasets

An audio-visual dataset is a collection of data that includes both audio and visual information. This dataset type is often used in machine learning and computer vision applications where audio and visual features are needed to train a model. Audio-visual datasets can include video recordings with accompanying audio tracks or separate audio and visual files that are synchronized to each other. It is important to follow certain techniques and protocols to create a dataset for audio-visual source separation. This includes gathering a diverse range of data, accurately labeling the audio and visual components, preprocessing the signals, dividing the dataset into appropriate subsets, maintaining a balanced distribution of samples, and applying data augmentation techniques. Implementing these steps makes the dataset more representative, reliable, and suitable for training and evaluating audio-visual source separation models. Table 3 shown below illustrates some of the commonly known datasets.

Table 3.
Audio-visual datasets

Dataset	Description
GRID	Audio-visual speech recognition dataset consisting of 33 speakers uttering 1000 phrases. Contains video, audio, and aligned phonemic transcriptions [50].
Lombard Grid	Audiovisual Lombard speech corpus, freely available for download. It contains 5400 utterances (2700 Lombard and 2700 plain reference utterances), produced by 54 talkers [51].
TCD TIMIT	Designed for continuous audio-visual speech recognition research. Consists of high-quality audio and video footage of 62 speakers reading a total of 6913 phonetically rich sentences [52].
OuluVS	The OuluVS database includes video and audio data of 20 subjects consisting of 10 phrases [53].
OuluVS2	A multi-view audiovisual database for non-rigid mouth motion analysis. It includes more than 50 speakers uttering three types of utterances [54].
Voxceleb	An audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube [55].
Voxceleb2	Contains over 1 million utterances for 6,112 celebrities, extracted from videos uploaded to YouTube [56].
LRW	Consists of up to 1000 utterances of 500 different words, spoken by hundreds of different speakers [57].
LRS	Consists of thousands of spoken sentences from BBC television. Each sentence is up to 100 characters in length [58].
LRS3	Consists of thousands of spoken sentences from TED and TEDx videos [59].
AVA-ActiveSpeaker	Contains about 38.5 hours of face tracks, and the corresponding audio [60].
Avspeech	Large-scale audio-visual dataset comprising speech video clips with no interfering background noises contains roughly 4700 hours of video segments [17].

5. Conclusions

The rapid advancement of deep learning approaches has revolutionized the field of audio-visual source separation, enabling significant progress in separating speech and other audio sources from complex audio-visual mixtures. DNNs, CNNs, LSTM, and BLSTM have emerged as powerful tools to address the challenges posed by audio-visual sources' multi-modality and dynamic nature. Researchers should consider the task requirements, available data, model

complexity, computational resources, interpretability needs, previous research, and time/resource constraints when choosing a method. Through careful evaluation of these factors, researchers can make informed decisions and choose a suitable method. It is important to note that BLSTM alone may not be sufficient, but it plays a crucial role as a key component in many state-of-the-art architectures. The combination of BLSTM with other modules like CNNs and DNNs has demonstrated promising outcomes in enhancing separation performance. Thus, the synergistic utilization of these techniques within an integrated system yields the most promising results, rather than relying on any single technique in isolation. Moreover, incorporating these models with additional deep learning techniques, such as attention mechanisms and generative adversarial networks, has remarkably improved the separation process. Large-scale datasets and the availability of high-performance computing resources have facilitated the training of deep models and the development of more complex architectures that can handle the variability of acoustical conditions. Despite these advancements, significant challenges remain, including handling variable speech overlap and room acoustics. There is a need for robust models that can generalize across different acoustic environments and effectively handle various types of audio-visual mixtures. In conclusion, the use of deep learning approaches in audio-visual source separation has enabled significant progress in the field and holds great promise for further advancements in the future. Continued research and development of these techniques will be essential to address the remaining challenges and facilitate the adoption of audio-visual source separation in practical applications.

Acknowledgments

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

References

[1] A. Al-Tmeme, W. L. Woo, S. S. Dlay and B. Gao, "Underdetermined Convolutive Source Separation Using GEM-MU with Variational Approximated Optimum Model Order NMF2D," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 35-49, Jan. 2017. http://dx.doi.org/10.1109/TASLP.2016.2620600.

[2] Woo, W.L.; Dlay, S.S.; Al-Tmeme, A.; Gao, B. "Reverberant signal separation using optimized complex sparse nonnegative tensor deconvolution on spectral covariance matrix". Digit. Signal Process. 2018, 83, 9–23. http://dx.doi.org/10.1016/j.dsp.2018.07.018

[3] Al-Tmeme, A.; Woo, W.L.; Dlay, S.; Gao, B. "Single channel informed signal separation using artificial-stereophonic mixtures and exemplar-guided matrix factor deconvolution". Int. J. Adapt. Control. Signal Process. 2018, 32, 1259–1281. http://dx.doi.org/10.1002/acs.2912.

[4] Ahmed Al-Tmeme, W.L. Woo, S.S. Dlay, and B. Gao, "Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models," J. Acoust. Soc. Am, 138, 3411, 2015. http://dx.doi.org/10.1121/1.4923156.

[5] Amer, R., and Al Tmeme, A. "Hybrid deep learning model for singing voice separation". Mendel 27, 2 (2021), 44–50. http://dx.doi.org/10.13164/mendel.2021.2.044.

[6] Mahmood, Israa N. and Hasanen S. Abdullah, "Telecom Churn Prediction Based on Deep Learning Approach" (2022) 63(6) Iraqi Journal of Science. http://dx.doi.org/10.24996/ijs.2022.63.6.32.

[7] Jameel, Humam Khaled and Ban Nadeem Dhannoon, "Gait Recognition Based on Deep Learning" (2022) 63(1) Iraqi Journal of Science. http://dx.doi.org/10.24996/ijs.2022.63.1.36.

[8] Al-Akkam, Reem Mohammed Jasim and Mohammed Sahib Mahdi Altaei, "Plants Leaf Diseases Detection Using Deep Learning" (2022) 63(2) Iraqi Journal of Science. http://dx.doi.org/10.24996/ijs.2022.63.2.34.

[9] Hussein, Noor Alhuda Khalid and Basad Al-Sarray, "Deep Learning and Machine Learning via a Genetic Algorithm to Classify Breast Cancer DNA Data" (2022) 63(7) Iraqi Journal of Science. http://dx.doi.org/10.24996/ijs.2022.63.7.36.

[10] N. Takahashi, M. K. Singh, S. Basak, P. Sudarsanam, S. Ganapathy and Y. Mitsufuji, "Improving Voice Separation by Incorporating End-To-End Speech

Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 41-45, http://dx.doi.org/10.1109/ICASSP40776.2020.9053845.

[11] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-visual speech separation using still images," Proc. of Interspeech, 2020. http://dx.doi.org/10.21437/Interspeech.2020-1065.

[12] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," IEEE Journal of Selected Topics in Signal Processing,2020. http://dx.doi.org/10.1109/JSTSP.2020.2980956.

[13] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADLMVDR: All deep learning MVDR beamformer for target speech separation," ICASSP, pp. 6089–6093, 2021. http://dx.doi.org/10.1109/ICASSP39728.2021.9413594.

[14] G. Li, J. Yu, J. Deng, X. Liu and H. Meng, "Audio-Visual Multi-Channel Speech Separation, Dereverberation and Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022, pp. 6042-6046, http://dx.doi.org/10.1109/ICASSP43922.2022.9747237.

[15] J. Ong, B. T. Vo, S. Nordholm, B. -N. Vo, D. Moratuwage and C. Shim, "Audio-Visual Based Online Multi-Source Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1219-1234, 2022. http://dx.doi.org/10.1109/TASLP.2022.3156758.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement,"2018. http://dx.doi.org/10.21437/Interspeech.2018-1400.

[17] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audiovisual model for speech separation," ACM Trans. Graph., pp. 112:1–112:11, 2018. http://dx.doi.org/10.1145/3197517.3201357.

[18] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 2, pp. 117–128, 2018. http://dx.doi.org/10.1109/TETCI.2017.2784878.

[19] R. Lu, Z. Duan, and C. Zhang, "Listen and look: audio–visual matching assisted speech source separation," IEEE Signal Processing Letters, vol. 25, no. 9, pp. 1315–1319,2018. http://dx.doi.org/10.1109/LSP.2018.2853566.

[20] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in Proc. of ECCV, 2018. http://dx.doi.org/10.1007/978-3-030-01246-5_35.

[21] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in Proc. of ICASSP, 2018. http://dx.doi.org/10.1109/ICASSP.2018.8462527.

[22] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "DNN driven speaker independent audio-visual mask estimation for speech separation," in Proc. of Interspeech, 2018. http://dx.doi.org/10.21437/Interspeech.2018-2516.

[23] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6900–6904. http://dx.doi.org/10.1109/ICASSP.2019.8682061.

[24] J. Wu, Y. Xu, S. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in Proc. IEEE Autom. Speech Recognit. Understanding Workshop, 2019, pp. 667–673. http://dx.doi.org/10.1109/ASRU46091.2019.9003983.

[25] Mandar Gogate et al. "Deep Neural Network Driven Binaural Audio Visual

Speech Separation". In: International Joint Conference on Neural Networks (IJCNN). IEEE. 2020, pp. 1–7. http://dx.doi.org/10.1109/IJCNN48605.2020.9207517.

[26] Q. Nguyen, J. Richter, M. Lauri, T. Gerkmann and S. Frintrop, "Improving mix-and-separate training in audio-visual sound source separation with an object prior," 2020 (ICPR). http://dx.doi.org/10.1109/ICPR48806.2021.9412174.

[27] C. Li and Y. Qian, "Deep Audio-Visual Speech Separation with Attention Mechanism," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7314-7318, http://dx.doi.org/10.1109/ICASSP40776.2020.9054180.

[28] R. Gu et al., "Multi-modal multi-channel target speech separation," IEEE J-STSP, 2020. http://dx.doi.org/10.1109/JSTSP.2020.2980956.

[29] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. "Music gesture for visual sound separation". In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10475–10484, 2020. http://dx.doi.org/10.1109/CVPR42600.2020.01049.

[30] Lingyu Zhu and Esa Rahtu. , "Visually guided sound source separation using cascaded opponent filter network" Proc. of ACCV, 2020. http://dx.doi.org/10.1007/978-3-030-69544-6_25.

[31] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 3, pp. 542–553, 2020. http://dx.doi.org/10.1109/JSTSP.2020.2987209.

[32] L. Qu, C. Weber, and S. Wermter, "Multimodal target speech separation with voice and face references," Proc. of Interspeech, 2020. http://dx.doi.org/10.21437/Interspeech.2020-1697.

[33] T. Rahman and L. Sigal, "Weakly-Supervised Audio-Visual Sound Source Detection and Separation,"IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 2021, pp. 1-6. http://dx.doi.org/10.1109/ICME51207.2021.9428196.

[34] R. Gao and K. Grauman. "VisualVoice: Audio-visual speech separation with cross-modal consistency". In CVPR, 2021. 8, 45. http://dx.doi.org/10.1109/CVPR46437.2021.01524.

[35] Majumder, S., Al-Halah, Z., Grauman, K.: "Move2Hear: Active audio-visual source separation". In: ICCV (2021). http://dx.doi.org/10.1109/ICCV48922.2021.00034.

[36] Y. Liu and Y. Wei, "Multi-Modal Speech Separation Based on Two-Stage Feature Fusion," IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 800-805, http://dx.doi.org/10.1109/ICSIP52628.2021.9688674.

[37] Y. Tian, D. Hu and C. Xu, "Cyclic Co-Learning of Sounding Object Visual Grounding and Sound Separation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 2744-2753. http://dx.doi.org/10.1109/CVPR46437.2021.00277.

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

[39] V. -N. Nguyen, M. Sadeghi, E. Ricci and X. Alameda-Pineda, "Deep Variational Generative Models for Audio-Visual Speech Separation," IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 2021, pp.1-6 http://dx.doi.org/10.1109/MLSP52302.2021.9596406.

[40] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn, "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1336–1345.

http://dx.doi.org/10.1109/CVPR46437.2021.00139.

[41] Lingyu Zhu and Esa Rahtu. "Leveraging category information for single-frame visual sound source separation," In IEEE 2021 9th European Workshop on Visual Information Processing (EUVIP), pages 1–6. http://dx.doi.org/10.1109/EUVIP50544.2021.9484036.

[42] R. Gu, S. -X. Zhang, Y. Zou and D. Yu, "Towards Unified All-Neural Beamforming for Time and Frequency Domain Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 849-862, 2023. http://dx.doi.org/10.1109/TASLP.2022.3229261.

[43] T. Oya, S. Iwase and S. Morishima, "The Sound of Bounding-Boxes," 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 9-15. http://dx.doi.org/10.1109/ICPR56361.2022.9956384.

[44] Lingyu Zhu and Esa Rahtu. "Visually guided sound source separation and localization using self-supervised motion representations" In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1289–1299, 2022. http://dx.doi.org/10.1109/WACV51458.2022.00223.

[45] D. -H. Pham, Q. -A. Do, T. T. -H. Duong, T. -L. Le and P. -L. Nguyen, "End-to-end Visual-guided Audio Source Separation with Enhanced Losses," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 2022, pp. 2022-2028. http://dx.doi.org/10.23919/APSIPAASC55919.2022.9980162.

[46] Xudong Xu, Bo Dai, and Dahua Lin, "Recursive visual sound separation using minus-plus net," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 882–891. http://dx.doi.org/10.1109/ICCV.2019.00097.

[47] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The sound of motions," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1735–1744. http://dx.doi.org/10.1109/ICCV.2019.00182.

[48] R. Lu, Z. Duan, and C. Zhang, "Audio-visual deep clustering for speech separation," IEEE ACM Trans. Audio Speech Lang. Process., vol. 27, no. 11, pp. 1697–1712, 2019. http://dx.doi.org/10.1109/TASLP.2019.2928140.

[49] Ruohan Gao and Kristen Grauman, "Co-separating sounds of visual objects," In Proc. ICCV, 2019. http://dx.doi.org/10.1109/ICCV.2019.00398.

[50] M. Cooke, J. Barker, S. Cunningham, X. Shao. "An audio-visual corpus for speech perception and automatic speech recognition" The Journal of the Acoustical Society of America, vol.120, no.5, pp.2421–2424, 2006. http://dx.doi.org/10.1121/1.2229005.

[51] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, G. J. Brown. "A corpus of audio-visual Lombard speech with frontal and profile views" The Journal of the Acoustical Society of America, vol.143, no.6, pp.EL523–EL529, 2018. http://dx.doi.org/10.1121/1.5042758.

[52] N. Harte, E. Gillen. "TCD-TIMIT: An audio-visual corpus of continuous speech". IEEE Transactions on Multimedia, vol.17, no.5, pp.603–615, 2015. http://dx.doi.org/10.1109/TMM.2015.2407694.

[53] G. Y. Zhao, M. Barnard, M. Pietikainen. "Lipreading with local spatiotemporal descriptors". IEEE Transactions on Multimedia, vol.11, no.7, pp.1254–1265, 2009. http://dx.doi.org/10.1109/TMM.2009.2030637.

[54] I. Anina, Z. H. Zhou, G. Y. Zhao, M. Pietikäinen. "OuluVs2: A multi-view audiovisual database for non-rigid mouth motion analysis". In Proceedings of the 11th IEEE International Conference and Workshops on Auto- matic Face and Gesture Recognition, IEEE, Ljubljana, Slovenia, pp.1−5, 2015. http://dx.doi.org/10.1109/FG.2015.7163155.

[55] A. Nagrani, J. S. Chung, A. Zisserman. "VoxCeleb: A large-scale speaker identification dataset". In Proceedings of the 18th Annual Conference of the International Speech Communication

Association, Stockholm, Sweden, pp.2616−2620, 2017. http://dx.doi.org/10.21437/Interspeech.2017-950.

[56] J. S. Chung, A. Nagrani, A. Zisserman. "VoxCeleb2: Deep speaker recognition". In Proceedings of the 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, pp.1086−1090, 2018. http://dx.doi.org/10.21437/Interspeech.2018-1929.

[57] J. S. Chung, A. Zisserman. "Lip reading in the wild". In Proceedings of the 13th Asian Conference on Computer Vision, Springer, Taipei, China, pp.87−103, 2017. http://dx.doi.org/10.1007/978-3-319-54184-6_6.

[58] J. S. Chung, A. Senior, O. Vinyals, A. Zisserman. "Lip reading sentences in the wild". In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Honolulu, USA, pp.3444−3453, 2017. http://dx.doi.org/10.1109/CVPR.2017.367.

[59] J. S. Chung, A. Zisserman. "Lip reading in profile". In Proceedings of British Machine Vision Conference 2017, BMVA Press, London, UK, 2017. http://dx.doi.org/10.5244/C.31.155.

[60] J. Roth et al., "Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 4492-4496 http://dx.doi.org/10.1109/ICASSP40776.2020.9053900.

نظرة عامة على فصل المصادر السمعية والبصرية باستخدام التعلم العميق

نورالهدى مظفر سليمان* احمد التميمي محمد نجاح مهدي *****

*,** قسم هندسة المعلومات والاتصالات/ كلية الهندسة الخوارزمي/ جامعة بغداد

***مركز ADAPT / كلية الحاسبات/ جامعة دبلن

*البريد الالكتروني:noor.abd2103m@kecbu.uobaghdad.edu.iq

**البريد الالكتروني:asattar@kecbu.uobaghdad.edu.iq

***البريد الالكتروني: mohammed.mahdi@adaptcentre.ie

الخلاصة

يعطي البحث المقدم هنا لمحة عامة على أنظمة فصل المصادر السمعية والبصرية (AVSS) القائمة على تقنيات التعلم العميق. حقق AVSS نتائج استثنائية في عدد من المجالات ، بما في ذلك تقليل مستويات الضوضاء ، وتعزيز التعرف على الكلام ، وتحسين جودة الصوت. تمت مناقشة مزايا وعيوب كل نموذج من نماذج التعلم العميق خلال البحث حيث يقوم بمراجعة مختلف التجارب الحالية على AVSS. مجموعة بيانات TCD TIMIT (التي تحتوي على تسجيلات صوتية ومرئية من الدرجة الأولى تم إنشاؤها خصيصًا لمهام التعرف على الكلام) ومجموعة بيانات Voxceleb (مجموعة كبيرة من المقاطع السمعية والبصرية المختصرة للكلام البشري) هي احد مجموعات البيانات المفيدة التي تم تلخيصها في هذا البحث و التي يمكن استخدامها لاختبار أنظمة AVSS. الغرض من هذه المراجعة في شكلها الأساسي هو تسليط الضوء على الأهمية المتزايدة لـ AVSS في تحسين جودة الإشارات الصوتية.