Kerbala Journal for Engineering Sciences

https://kjes.uokerbala.edu.iq

Deep Learning Cross-Modal Learning for Audio-Visual Speech Recognition

Hayder Jasim Habil [*], Ibtehal Shakir Mahmoud [**], Dalal Abdulmohsin Hammood [***], Effariza Hanafi[****]

[*] College of Applied Arts, Middle Technical University (MTU), Baghdad, Iraq.
E-mail: hayderhjh@gmail.com
[**] Aliraqia University, Baghdad, Iraq.
E-mail: ibtehal.shaker@aliraqia.edu.iq
[***] Electrical Engineering Technical College, Middle Technical University (MTU), Baghdad, Iraq.
E-mail: dalal.hammood@mtu.edu.iq
[****] University Malaya; Kuala Lumpur.
E-mail: effarizahanafi@um.edu.my

Received: 17 June 2023; Revised: 25 June 2023; Accepted: 08 February 2024

Abstract

The ability to relate information about languages heard through visual and audio data is a crucial aspect of audio-visual speech recognition (AVSR), which has uses in data manipulation for audio-visual correspondence, including AVE-Net and SyncNet. The technique described in this research uses feature disentanglement to simultaneously handle the tasks listed above. By developing cross-modal standard learning methods, this model can transform visual or aural linguistic characteristics into modality-independent representations. AVE-Net and SyncNet can all be performed with the help of such derived linguistic expressions. Furthermore, audio and visual data output can be modified based on the required subject identity and linguistic content information. We do comprehensive trials on various recognition and synthesis tasks on both tasks separately, and that solution can successfully take on both audio-visual learning problems. The system gives great results in the enhanced video with 91.5% with 5 frames, while this will increase with the increase of frames with 99.03% with 15 frames, which is more efficient than the previous methods.

Keywords: CNNs, deep learning, AVE-Net, SyncNet, AVSR

1. Introduction

As seen in Fig.1, audio-visual speech recognition (AVSR) recognizes speech using visual clues such as lip movement. Speech synthesis has a subsection called audio-visual speech synthesis, or AVSR for short. This domain aims to make convincing talking-head movies and audio recordings. To this end, audio-visual speech synthesis can be utilized for learning tasks such as face-to-face, voice-to-voice, and voice-to-face conversion [1].[2][3][4][5][6][7][8]. To do audio-visual speech recognition or synthesis, extracting representative features from cross-modality input data (i.e., audio and visual information) is necessary. Information that maintains its modal form: Even though the extraction of linguistic representation is required for AVSR to accomplish its mission, certain information, such as subject identity, must still be preserved to facilitate data recovery and synthesis. Voice conversion [9], audio-visual speech separation [10], and synchronization of numerous speakers, when they are speaking [11] are all things that are made feasible as a result of the two sorts of representations discussed above. However, most available work only covers one or a limited number of jobs. Designing multi-task learning systems that take advantage of different modalities' inputs to handle the learning challenges discussed earlier would be beneficial. These problems are interconnected.

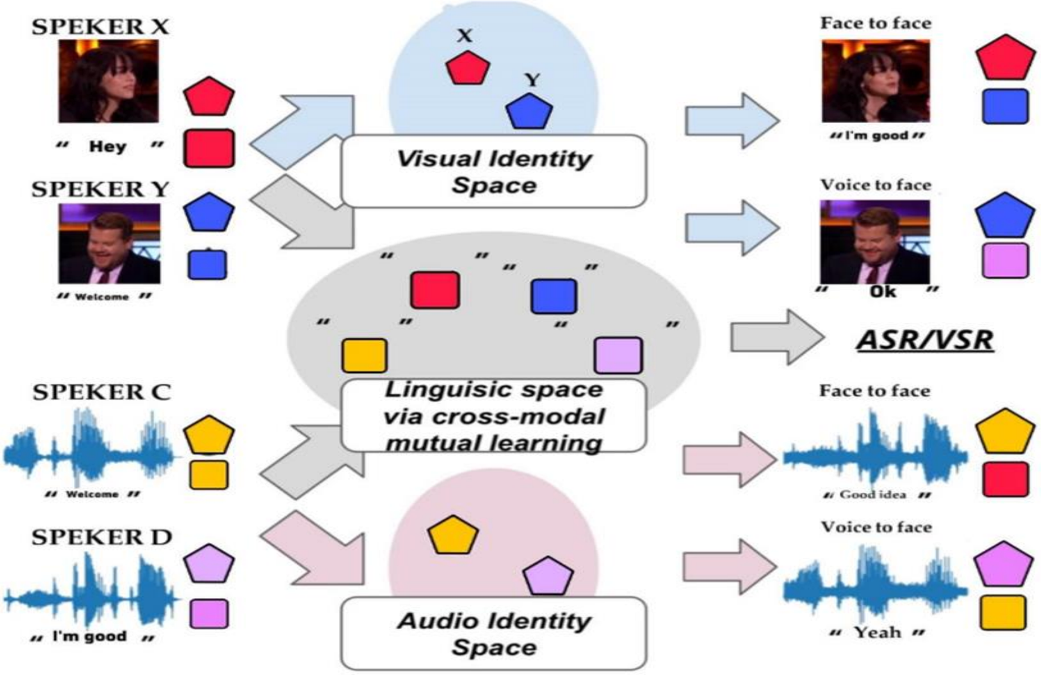

Figure 1. A demonstration of the simultaneous recognition and manipulation of audio and video speech.

Feature fusion is achieved by aligning modalities with long-term dependencies, which is the goal of transformer-based cross-modal mutual learning architecture (see Fig.2). First, framewise representations are extracted with the help of a front-end that has already been trained. The next step is to improve cross-modal interactions with a multi-modal encoder that pools query and critical weights. Our approach requires a linguistic codebook to be traversed by visual representations of distinct modalities, requiring the model to compile a compact linguistic picture for tokens of each modality, thereby capturing modality-invariant information. The cross-attention mechanism then aligns the various modalities through linguistic knowledge.

We introduced a novel training strategy for cross-modal matching and retrieval, allowing networks to undergo matching training without needing explicit class labels. This approach leverages the advantageous learning characteristics inherent in the cross-entropy loss. Our experimental results demonstrate superior performance in the audio-visual synchronization task compared to current state-of-the-art methods. The suggested embedding strategy substantially improves the visual speech recognition task and achieves performance on par with a fully supervised method employing the same architecture. Moreover, we anticipate the potential applicability of this method to other cross-modal tasks.

2. Related Work

The generation of talking-face movies based on predetermined facial or aural inputs has been the subject of multiple proposed methodologies; however, these systems still need to develop to a point where they are satisfactory. Word and identity labels are used as a guide [11], which uses a detangle technique to encode both the speaker's lip movements and the speaker's physical appearance to guarantee that each feature is clean; adversarial training uses a severe information bottleneck to block out data from one feature space and only use data from the other feature space. However, the depiction of head and facial expression movements as training goals for lip movement extraction is typically confined to word labels because the head and facial expression movements are intertwined. Since the speech function does not limit its attention solely to lip movement, it communicates this erroneous impression. As we can show, despite the advancements in end-to-end neural text-to-speech (TTS) methods, exemplified by Tacotron2, two persistent issues remain: 1) inefficiency in both training and inference and 2) challenges in modeling long dependencies with existing recurrent neural networks (RNNs). Inspired by the success of the

Transformer network in neural machine translation (NMT), our paper introduces a solution by adapting the multi-head attention mechanism to replace RNN structures and the original Tacotron2 attention mechanism.

Utilizing multi-head self-attention allows for the parallel construction of hidden states in the encoder and decoder, enhancing training efficiency. Moreover, a self-attention mechanism connects inputs at different times, effectively addressing long-range dependency challenges. According to phoneme sequences, our Transformer TTS network generates Mel spectrograms, employing a WaveNet vocoder for the final audio output. Experimentally, our Transformer TTS network accelerates training approximately 4.25 times faster than Tacotron2 while achieving state-of-the-art performance, surpassing Tacotron2 by 0.048. Human tests indicate a close approximation to human quality, scoring 4.39 compared to Tacotron2's 4.44 in MOS, which was the first approach. The second approach in this paper, we introduce a novel framework called the Disentangled Audio-Visual System (DAVS), designed to generate high-quality talking face videos using a disentangled audio-visual representation. Our approach begins by learning a joint audio-visual embedding space, wide, enriched with discriminative speech information obtained from word-ID labels. Subsequently, we employ adversarial learning to disentangle the wide space from the person-ID pid space. DAVS possesses several notable advantages compared to previous works: (1) The joint audio-visual representation is acquired through audio-visual speech discrimination, leveraging multiple supervisions. This disentangled representation notably enhances lip reading performance; (2) Audio-visual speech recognition and synchronization are seamlessly integrated into an end-to-end framework; (3) Most importantly, our framework enables the generation of talking face videos for arbitrary subjects with high quality and temporal accuracy. Both audio and video speech information can serve as input guidance for this purpose. This paper introduces a fresh training strategy for cross-modal learning, focusing on developing robust cross-modal embeddings through a multi-way matching task. By combining similarity-based methods like L2 distance loss with a multi-class cross-entropy loss, our approach naturally facilitates cross-modal retrieval, identifying the most relevant sample across different modalities.

Our innovative training strategy involves training the network for the multi-way matching task without explicit class labels while still harnessing the beneficial learning characteristics of the cross-entropy loss. We showcase the effectiveness of this strategy in audio-visual synchronization, specifically in locating the most relevant audio segment given a short video clip. The models trained for multi-way matching demonstrate the ability to generate powerful representations of both auditory and visual information, with applications extending to other tasks. Additionally, we highlight superior performance in a visual speech recognition task compared to embeddings obtained through pairwise objectives.

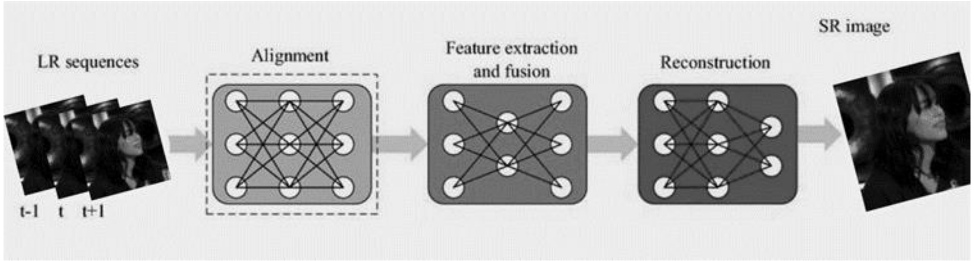

Figure 2. Architecture for learning that is cross-modal and uses transformers.

Information about verbal communication. On the other hand,[4] [6] builds movies by directly integrating audio and visual identification features. In addition, they apply an additional discriminator to ensure that the audio and video remain in sync with one another. For its part, [2] uses the input audio as background information and direction to generate facial landmarks voice technology fraud for synthesizing talking-face movies. The currently available methodologies have produced promising results in two applications: text-to-voice production and the transfer of style from voice to speech. Vector quantization (VQ) and instance normalization (IN), which separate linguistic information from the audio signal while embedding the speaker style information, have been employed in previous works as the speaker style information bottleneck for voice-to-voice style transfer. VQ and IN can divorce linguistic information from the audio signal while preserving speaker-style information.[12] [9][3]. To increase accuracy, however, these designs would need additional components because they do not consider that a speaker's tone of voice might cause a signal to become corrupted. Using sequence-to-sequence learning in conjunction with attention processes, auto-regressive Mel-spectrogram synthesis was utilized to meet the requirements of the text-to-voice production challenge [13]. It has been demonstrated that voice-to-text translation may produce excellent results; however, research into face-to-voice translation is still in its early stages because of the difficulty presented by homophones. Recently, [5] developed a model dependent on the face sequence but with an overall design conceptually comparable to the text-to-voice paradigm. This model was generated for each speaker. The models, however, cannot be expanded to do voice synthesis for many speakers since the framework architecture requires a significant quantity of training data to be provided for each speaker. The Recognition of Speech Through the Use of Sounds and Images Visual speech recognition (VSR) is a task that is more difficult than audio speech recognition (ASR) because of the variety and ambiguity of lip movements among different speakers. However, past research has shown excellent performance in ASR. Researchers in [14] and [15] argue for word-level voice recognition using recurrent input sequence modules. However, in most cases, one will be required to deal with the convergence problem of the connected recurrent networks. Recent studies have begun using temporal convolution networks to improve learning effectiveness [16]. Previous studies [17] and [18] on voice recognition at the sentence level have concentrated on using sequence-to-sequence learning to identify the character or set of words within a whole sentence. It is difficult to extract language characteristics that can be used for discrimination from longer films. As a result, a pretraining feature extractor that uses techniques at the word level is utilized. Earlier efforts, such as [19][20], eliminated knowledge from models learned from audio or audio-visual data to direct the VSR ones due to the superior performance of ASR. However, the methodologies currently used need to consider the possibility of transferring VSR model knowledge to ASR models.

3. PREPARATION WORK

This section prepares the audience for the audio and video accompanying it. When comparing the results of the new training strategy to those of the old one, we used the same inputs and layer configurations that SyncNet[21] employed. The network receives media in increments of 0.2 seconds, typically for audio streaming. At regular intervals of 10 milliseconds and with a frame duration of 25 milliseconds, inputs based on extracted Mel-frequency cepstral coefficients (MFCCs) are delivered into the audio stream. Converting video to audio will always produce undesirable side effects, such as background noise and speech distortions. 13 frames are moving in the temporal direction, and 13 cepstral coefficients are moving in the spatial order, resulting in an input image of 13 by 20 pixels. The network was built with the VGG-M [21] CNN architecture, and the filters were modified to handle the increased auditory input size of the streaming visuals. For example, the video of a face cropped and played into the visual stream has dimensions 224 by 224 and a frame rate 25. Over 0.2 seconds, the network takes in the visual information, which consists of five RGB frames that have been stacked. According to the VGG-M [12], the first layer of the visual stream employs a filter size of 5x7x7 rather than the more typical 7x7 to better capture motion information across all five frames. This process is done by using a more significant number of samples per filter.

A comprehensive study of the visual evolution may be found in Fig. 3. Instructional Strategies and Procedures 2.2 Unsupervised exploration of the audio and visual data is intended to lead to the discovery of cross-modal embeddings of the data. The suggested approach is taught through the use of a multi-way matching scheme, while the two baselines are trained through the use of a paired correspondence task. The beginning point provided by SyncNet is utilized. During the training process for the first iteration of SyncNet [21], a contrastive loss is applied. It causes the distance between features to grow when mismatched pairs of inputs are being processed but causes it to shrink when processing matched pairs. When a team does not fit, the video and audio were recorded various times from the same face track.

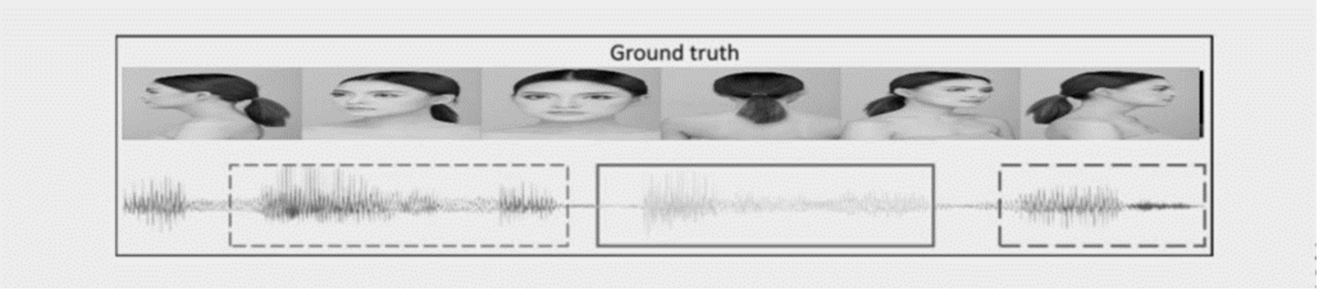

Figure 3. A method of approaching self-supervised learning through sampling.

The method calls for the precise adjustment of a hyper-parameter that is referred to as "margin." The AVE-Net serves as the initial point of contact. The Audio-visual Embedding Network, also known as AVE-Net [21][22], was developed specifically for cross-modal retrieval. It uses the outputs of both the audio and visual networks as its inputs. After the input vectors have been L2-normalized, they are sent to a fully connected layer and a SoftMax layer, both are responsible for computing the Euclidean distance between the two normalized embeddings. A Multiple-category system is proposed, and as illustrated in Fig. 4, the fully-connected layer is accountable for learning the threshold beyond which the features are no longer considered to correspond to one another.

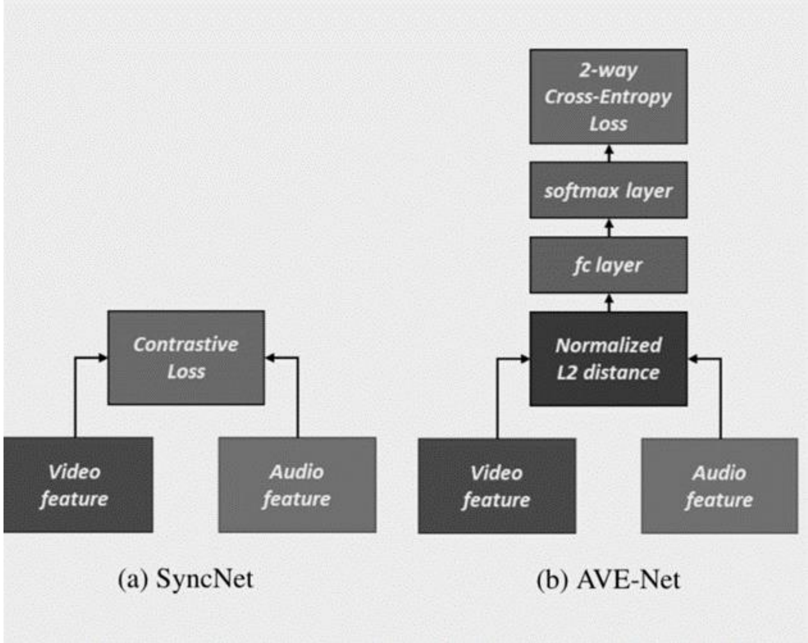

Figure 4. A comparison of the already utilized training methods.

Unlike earlier approaches that relied on pairwise losses, the proposed embeddings are learned using a multi-way matching challenge. The only situation in which pairwise defeats are helpful is in binary matching. As a result, they do not take any additional information seriously. On the other hand, the multi-way matching method not only regulates the distance between pairs but also uses critical information buried in the data sequence to train the model. A single visual feature and a significant number of audio characteristics make up the inputs for the learning criterion, which can be accomplished by utilizing a feature-matching task with N dimensions. When comparing audio and visual features, calculating the Euclidean distance returns N different distances. The network is trained using a cross-entropy loss on the inverse of this distance, and a SoftMax layer is used to enhance the similarity between pairings that match while at the same time decreasing the similarity between pairings that do not match. The many different methods of training are outlined in this paragraph. Even though all N video and audio frames belong to the same race, the temporal alignment of only one is correct. This exercise aims to teach the network to focus more on words' meaning than on the recognition of individual words. Fig. 4 visually represents the sampling process.

4. EXPERIMENTAL RESULTS

We evaluate the suggested system's effectiveness compared to a conventional lip-syncing approach and an audio-visual application known as synchronization of sound and picture. The synchronization of audio and video is a cross-modal retrieval problem. We were given a video clip and instructed to choose an audio clip from a collection that accurately represented the video's time signature. To accomplish this, we first analyze the video to determine its constituent parts, and then we evaluate those parts concerning the audio (using a 5-frame window). When there is a minimum distance between features in both streams, we consider them to be synchronized with one another. However, the article [21] contends that because only some samples give discriminative information, relying on more than one visual component may be necessary to determine the appropriate offset. For instance, there could be mute video portions consisting of five frames. In addition, we do tests using context windows that are larger than five frames, averaging distances over a large number of video samples (while maintaining a temporal stride of one frame). To train the network, we use the LRS2 lip-reading pre-train set. In total, 1,243 test clips and 96,318 training clips are included in the LRS2 dataset [23]. We conduct training and assessment of the models using the Lip Reading in the Wild (LRW) dataset. This dataset comprises word-level speech and video segments extracted from British television. It encompasses a vocabulary of 500 and includes more than 500,000 utterances. Among these, 25,000 are specifically set aside for testing purposes. Because longer video clips are required to train networks with bigger N (the candidate audio clips are sampled without overlap), there is a trade-off between the number of classes (or candidate audio features) N and the number of accessible video clips for training, where N is the number of courses or candidate audio features.

We conduct tests using various data, reporting accuracy, and available video clip counts from Fig. 5 to determine the N value that gives the best results. The accuracy of the synchronization has been validated to be within a range of 15 video frames, provided that the projected offset is within 1 video frame of the ground truth. The odds of hitting the target with a shot in the dark are only 9.7 percent likely to do so under the current circumstances. We also compute the sync offset across all optical input frames by averaging the feature distances to account for noise when dealing with input lengths K that are more significant than 5. Results. The results of experiments are documented and shared on the internet. Computer programs that can recognize spoken words.

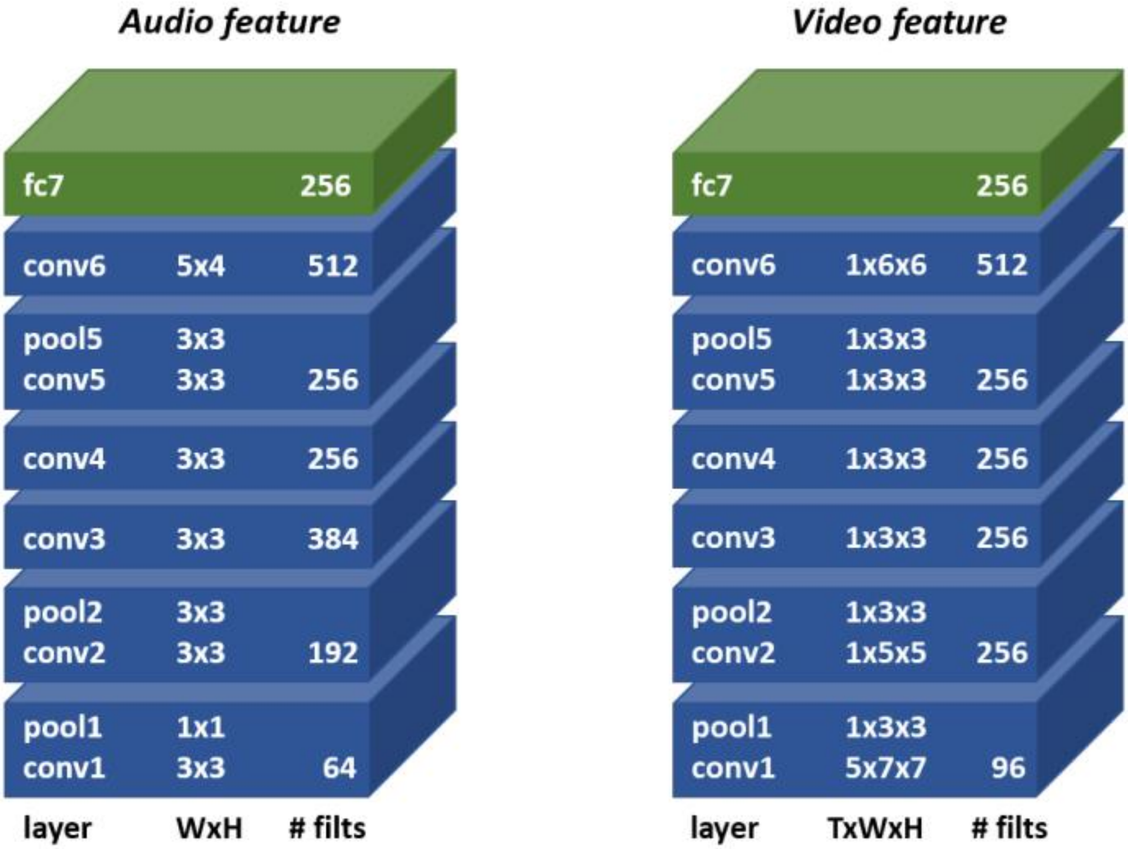

Figure 5. Architecture of the trunk for both audio and visual streams

Table 1. The number of visual frames used to calculate an average distance for all of the frames.

# Frames	SyncNet	AVE-Net	Proposed method
5	75.8%	74.3%	91.5%
7	82.4%	80.4%	94.12%
9	87.8%	86.6%	94.9%
11	91.7%	90.1%	97.26%
13	94.4%	93.8%	97.85%
15	96.7%	95.6%	99.03%

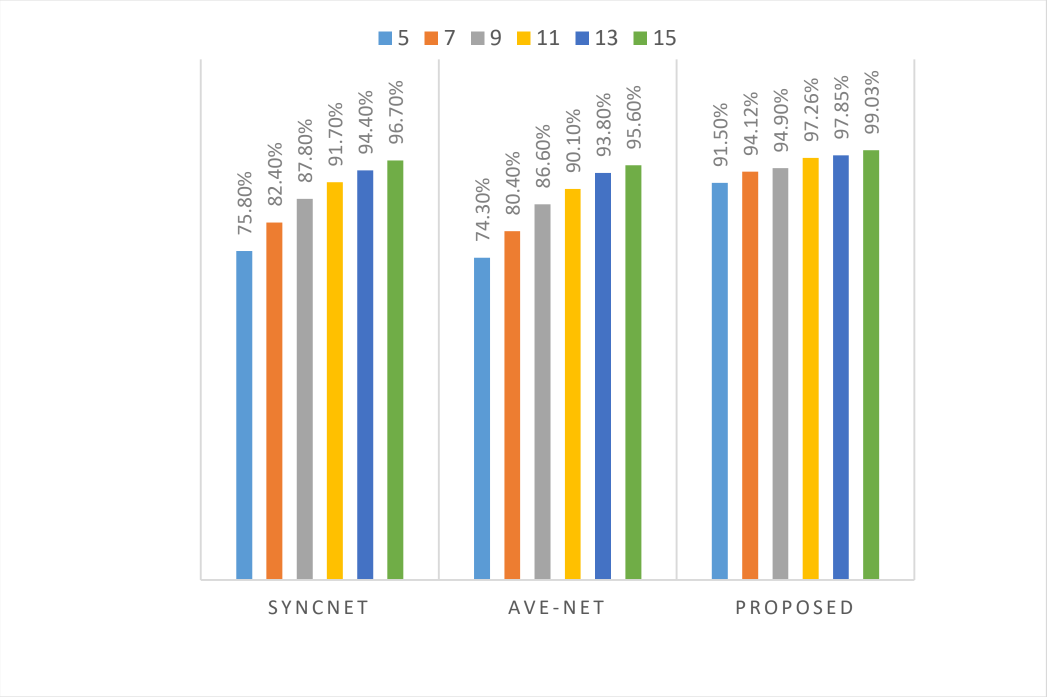

Figure 6. A chart that displays the visual frames for which the distances are an average over time.

The network acquires an accurate embedding of the graphical data included in the input video through its learning process. In this experiment, the transfer learning process will demonstrate that the learned embeddings of the matching network may be successfully transferred to other tasks, such as visual voice recognition. We show this on a word-level recognition problem and compare the results of networks trained using the suggested self-supervised strategy to learn the embeddings with those prepared using the traditional fully-supervised method. Both approaches are used to understand the embeddings. Dataset. For training and testing purposes, we use the Lip Reading in the Wild (LRW) dataset, comprising word-level audio and video samples from British television. The vocabulary is 500 words, and the dataset has approximately 500,000 phrases, with 25,000 of those phrases set aside for examination. The network's visual stream (which was covered in Section 2) serves as a guide for the front-end development. First, we suggest a full-stack structure with a 500-way SoftMax classification layer and two layers that perform temporal convolution.

The information on this network topology, labelled TC-5 in Fig. 6, can be found in Table 1. Following the naming conventions of the previous networks, the number '5' represents the temporal receptive field of the feature extractor. When trained end-to-end (E2E), the TC-5 model outperforms the networks in terms of its performance. In the 'pre-trained' (PT) trials, where visual features are extracted in advance, only the last layers of the 500-way classification job are supervised and certified to ensure that the results are accurate, as shown in Table (1) and Fig. 6. It can be shown that the proposed method is better than the other two algorithms as it shows when using 5 frames, the proposed algorithm provides 91.5% compared to 74.3% for AVE-Net SyncNet 75.8%, and this value increases with the video has more frames to be 99.03% for the proposed algorithm, and this shows that our proposed algorithm is faster with great enhanced to the images of the video compared to previous algorithms. So it can work in frames video to give big changes to low-quality video.

5. Conclusions

We suggested a unified framework for the recognition and synthesis of audio-visual speech within the scope of this research project, which aimed to investigate the topic. We progress in cross-modal mutual learning to harmonize linguistic information across visual and auditory input, ultimately resulting in AVE-Net and SyncNet having a representation independent of modality. Our technology can modify intra- or cross-modality data outputs with appropriate audio or visual information provided that modality-specific identifying qualities from either modality are kept where the case is regardless of whether the data outputs are manipulated intra- or cross-modality. A comprehensive set of tests was carried out on the problematic AVE-Net and SyncNet benchmark datasets. The findings of this research demonstrated, both qualitatively and numerically, that our model is more effective than the audio-visual learning techniques considered to be state-of-the-art in the context of recognition and synthesis tasks.

References

[1] Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip Movements Generation at a Glance. In Proceedings of the European Conference on Computer Vision (ECCV).

[2] Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7832–7841.

[3] Van den Oord, A.; Vinyals, O.; and kavukcuoglu, k. 2017. Neural Discrete Representation Learning. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

[4] KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; and Jawahar, C. 2019. Towards automatic face-to-face translation. In Proceedings of the 27th ACM International Conference on Multimedia, 1428–1436.

[5] Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020a. Learning Individual Speaking Styles for Accurate Lip-to-Speech Synthesis. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020b. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In Proceedings of the 28th ACM International Conference on Multimedia, MM ’20, 484–492. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

[7] Song, Y.; Zhu, J.; Li, D.; Wang, A.; and Qi, H. 2019. Talking Face Generation by Conditional Recurrent Adversarial Network. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 919–925. International Joint Conferences on Artificial Intelligence Organization.

[8] Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; andWang, X. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. In AAAI Conference on Artificial Intelligence (AAAI).

[9] Ding, S.; and Gutierrez-Osuna, R. 2019. Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion. In Proc. Interspeech 2019, 724–728.

[10] Gao, R.; and Grauman, K. 2021. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. arXiv preprint arXiv:2101.03149.

[11] Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; andWang, X. 2019. TalkingFace Generation by Adversarially Disentangled Audio-Visual Representation. In AAAI Conference on Artificial Intelligence(AAAI).

[12] Chou, J.-c.; Yeh, C.-c.; and Lee, H.-y. 2019. One-shotVoice Conversion by Separating Speaker and Content Representations with Instance Normalization. arXiv preprint arXiv:1904.05742.

[13] Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. NeuralSpeech Synthesis with Transformer Network. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):6706–6713.

[14] Feng, D.; Yang, S.; Shan, S.; and Chen, X. 2020. Learn an Effective Lip Reading Model without Pain. arXiv preprint arXiv:2011.07557.

[15] Stafylakis, T.; Khan, M. H.; and Tzimiropoulos, G. 2018. Pushing the boundaries of audio-visual word recognition using Residual Networks and LSTMs. Computer Vision and Image Understanding, 176-177: 22–32.

[16] Ma, P.; Wang, Y.; Shen, J.; Petridis, S.; and Pantic, M. 2021. Lip-Reading With Densely Connected Temporal Convolutional Networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2857–2866.

[17] Ma, P.; Wang, Y.; Shen, J.; Petridis, S.; and Pantic, M. 2021. Lip-Reading With Densely Connected Temporal Convolutional Networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2857–2866.

[18] Garcia, B.; Shillingford, B.; Liao, H.; Siohan, O.; de Pinho Forin Braga, O.; Makino, T.; and Assael, Y. 2019. Recurrent Neural Network Transducer for Audio-Visual Speech Recognition. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop.

[19] Ren, S.; Du, Y.; Lv, J.; Han, G.; and He, S. 2021. Learning From the Master: Distilling Cross-Modal Advanced Knowledge for Lip Reading. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13325–13333.

[20] Zhao, Y.; Xu, R.; Wang, X.; Hou, P.; Tang, H.; and Song, M. 2020. Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04): 6917–6924.

[21] Soo, W; Joon, S and Hong, May 2019. Perfect Match: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1520-6149.

[22] Chih, Y; Wan, F; Cheng, Y; and Yu, W; 2022 “Cross-modal mutual learning for audio-visual speech recognition and manipulation,” in Proceedings of the 36th AAAI Conference on Artificial Intelligence, vol. 22

[23] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

التعلم العميق عبر الوسائط للتعرف على الكلام السمعي البصري

الخلاصة: *تعد القدرة على ربط المعلومات حول اللغات المسموعة من خلال البيانات المرئية والمسموعة جانبًا حاسمًا في التعرف على الكلام السمعي البصري (AVSR) ، والذي يستخدم في معالجة البيانات للمراسلات السمعية والبصرية ، بما في ذلك AVE-Net و SyncNet. تستخدم التقنية الموصوفة في هذا البحث فك التشابك للتعامل مع المهام المذكورة أعلاه في وقت واحد. يمكن لهذا النموذج تحويل الخصائص اللغوية المرئية أو السمعية إلى تمثيلات مستقلة عن الطريقة من خلال تطوير أساليب التعلم القياسية عبر الوسائط. يمكن إجراء كل من AVE-Net و SyncNet بمساعدة مثل هذه التعبيرات اللغوية المشتقة. علاوة على ذلك ، يمكن تعديل إخراج البيانات الصوتية والمرئية بناءً على هوية الموضوع المطلوبة ومعلومات المحتوى اللغوي. نجري تجارب شاملة على مهام التعرف والتوليف المختلفة في كلتا المهمتين بشكل منفصل ويمكن لهذا الحل أن يعالج مشاكل التعلم السمعي البصري بنجاح.*