The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.4, pp.166~173

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.4.166

Lip and Voice Synchronization Using Visual Attention

Dongryun Yoon† · Hyeonjoong Cho††

ABSTRACT

This study explores lip-sync detection, focusing on the synchronization between lip movements and voices in videos. Typically, lip-sync detection techniques involve cropping the facial area of a given video, utilizing the lower half of the cropped box as input for the visual encoder to extract visual features. To enhance the emphasis on the articulatory region of lips for more accurate lip-sync detection, we propose utilizing a pre-trained visual attention-based encoder. The Visual Transformer Pooling (VTP) module is employed as the visual encoder, originally designed for the lip-reading task, predicting the script based solely on visual information without audio. Our experimental results demonstrate that, despite having fewer learning parameters, our proposed method outperforms the latest model, VocaList, on the LRS2 dataset, achieving a lip-sync detection accuracy of 94.5% based on five context frames. Moreover, our approach exhibits an approximately 8% superiority over VocaList in lip-sync detection accuracy, even on an untrained dataset, Acappella.

Keywords : Lip-Voice Synchronization, Visual Attention, Multi-Modal Transformer

시각적 어텐션을 활용한 입술과 목소리의 동기화 연구

윤 동 련† · 조 현 중††

요 약

본 연구에서는 얼굴 동영상에서 입술의 움직임과 음성 간의 동기화 탐지 방법을 제안한다. 기존의 연구에서는 얼굴 탐지 기술로 얼굴 영역의 바운딩 박스를 도출하고, 박스의 하단 절반 영역을 시각 인코더의 입력으로 사용하여 입술-음성 동기화 탐지에 필요한 시각적인 특징을 추출하였다. 본 연구에서는 입술-음성 동기화 탐지 모델이 음성 정보의 발화 영역인 입술에 더 집중할 수 있도록 사전 학습된 시각적 Attention 기반의 인코더 도입을 제안한다. 이를 위해 음성 정보 없이 시각적 정보만으로 발화하는 말을 예측하는 독순술(Lip-Reading)에서 사용된 Visual Transformer Pooling(VTP) 모듈을 인코더로 채택했다. 그리고, 제안 방법이 학습 파라미터 수가 적음에도 불구하고 LRS2 데이터 세트에서 다섯 프레임 기준으로 94.5% 정확도를 보임으로써 최근 모델인 VocaList를 능가하는 것을 실험적으로 증명하였다. 또, 제안 방법은 학습에 사용되지 않은 Acappella 데이터셋에서도 VocaList 모델보다 8% 가량의 성능 향상이 있음을 확인하였다.

키워드 : 입술-음성 동기화, 시각적 어텐션, 트랜스포머

1. Introduction

The task of lip-voice synchronization detection, referred to as *lip-sync detection*, involves detecting whether voice and lip movements in videos are synchronized. Lip-sync detection can be used to determine the consistency between visual and auditory signals for various types of video content, such as in press conferences, dubbed movies, singing, and synthetic face videos. When viewing videos online, hamans are highly sensitive to subtle lip-sync descrepancies [1]. However, detecting subtle yet crucial lip-sync nuances is an exceptional challenge. In particular, lip-sync detection has been used as a component in the training of talking face synthesis models that synthesize facial expressions to synchronize with given voices [2].

Previous studies typically receive the lower halves of facial images as input to the visual encoder, potentially including redundant areas for lip-sync detection, such as clothing, background, and the nose. To enhance lip-sync

※ 본 연구는 2023년 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1049202).
† 비 회 원 : 한국전자기술연구원 에너지IT융합연구센터 연구원
†† 종신회원 : 고려대학교 컴퓨터융합소프트웨어학과 교수
Manuscript Received : November 22, 2023
First Revision : February 5, 2024
Accepted : March 12, 2024
* Corresponding Author : Hyeonjoong Cho(raycho@korea.ac.kr)

detection accuracy by focusing on the essential information and minimizing redundancy, we propose a visual encoder designed to spatially prioritize lip areas. In essence, our proposal involves the utilization of a visual encoder that gives more weight to meaningful local features within the input image. As a baseline model, we have chosen the current State-of-The-Arts (SoTA) VocaList.

To create a visual encoder with a targeted focus on lip areas, we drew inspiration from the lip-reading task, where the goal is to infer scripts based solely on visual information from mouth movements. Prajwal *et al.* introduced a lip-reading model that significantly improved accuracy compared to previous works [3,4,16]. They incorporated a Visual Transformer Pooling (VTP) module into their visual encoder, allowing spatial attention to lip areas. Leveraging the attention weights of the VTP encoder, which exclusively consider visual lip movements in critical regions for predicting scripted speech, we anticipate improved lip-sync performance. Notably, the VTP module employs a linear transformer designed to reduce computational complexity while preserving the original performance [5]. Consequently, we expect to overcome the problem of the increasing number of parameters observed in previous work, VocaList [6], with our proposed method.

In this study, the proposed *VTPVocaList* leverages the pre-trained VTP as part of the visual encoder to extract the feature vector from the lip areas directly related to lip-voice synchronization. We experimentally demonstrate that VTPVocaList surpasses previous models in terms of lip-sync detection accuracy on several test datasets, including LRS2, LRS3, and Acappella [7-9]. Significantly, VTPVocaList underwent training exclusively on the LRS2 dataset; nevertheless, it exhibited superior performance compared to earlier studies on three distinct test sets: LRS2, LRS3, and Acappella. Despite achieving SoTA performance, VTPVocaList manages to maintain this high standard with only 67% of the parameters present in the previous VocaList, thanks to the utilization of the pre-trained VTP module.

The remaining sections of this paper are structured as follows: Section 2 provides a summary of previous related studies, while Section 3 delves into the detailed architecture and components of VTPVocaList. Section 4 presents the experimental results. Following this, Section 5 outlines the conclusions, and finally, Section 6 discusses the limitations and suggests areas for future work.

2. Related Works

Early lip-sync detection models utilized multi-layer perceptrons to determine whether viseme-phoneme mapping is correct [10]. As training this model requires a dataset containing pairs of Visemes and Phonemes with ground-truth labels, the labeling process is expensive and cumbersome. To address these challenges, a pioneering self-supervised learning method called SyncNet was introduced [11]. Recent strides in lip-sync research have progressed, inspired by the structure and methodology originated from SyncNet. It utilized a Siamese network architecture comprising CNN-based visual and audio encoders trained in a self-supervised manner. A visual encoder and audio encoder extract feature vectors from facial videos and voices. During this process, positive and negative pairs are generated autonomously, consisting of aligned pairs and intentionally misaligned pairs in the context of lip-sync. Subsequently, the contrastive learning method is employed [9].

Among the lip-sync detection models built upon SyncNet, PM modifies the loss function employed by SyncNet. Additionally, Kim et al. enhanced SyncNet by introducing a classification model, thereby transforming the lip-sync detection problem into a classification problem [12]. AVST stacked a transformer-based synchronization block in addition to SyncNet. Following the extraction of visual and audio embeddings from the CNN-based encoders, SyncNet, the vanilla transformer architecture, syncrhonization block, is then utilized to capture their correlation. Instead of a vanilla transformer, VocaList replaced the transformer-based synchronization block with multimodal transformer architecture to learn the correspondence between video and audio embeddings based on cross-modal attention modules[6]. The attention mechanism based on multi-modal transformers computes attention weights between different modalities, such as audio and images. In this study, the VTP encoder, akin to Vision Transformer[13], subdivides images into smaller patch units and computes regional attention weights within the image.

Over time, [21] introduced an algorithm for lip-sync detection based on viseme-phoneme correspondence, without utilizing the SyncNet-based evaluation method.

SyncNet-based models have gradually improved lip-sync detection accuracy from 75.8% to 94.5% on the LRS2 dataset. Table 1 and Fig. 1 display the structural changes in lip synchronization models.

Table 1. Related Works

Year	Structure	Model	Training Method	Backbone Model	Accuracy on LRS2
2014	-	[12]	Supervised	MLP	-
2016	A	SyncNet	Self-supervised	CNN	75.8%
2018		PM			88.1%
2020		SyncNet*			91.0%
2021	B	AVST		Transformer	92.0%
2022	C	VocaList		Multi-Transformer	92.8%
2023	D	VTPVocaList(ours)		Multi-Transformer + Visual Attention	***94.5%***

Related Works in Lip synchronization. The simplified model structures (A,B,C and D) are shown in Figure 1. SyncNet* is improved model used in Wav2Lip. Details are in Related workst.

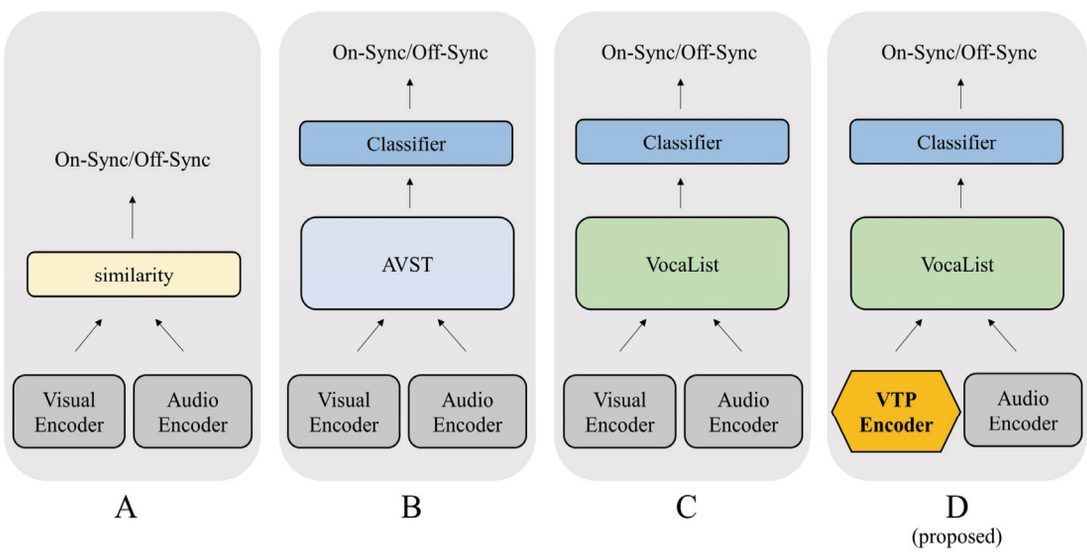

Fig. 1. Four different structures for lip-sync detection. In the second column of Table 1, symbols A, B, C, and D are linked. At first, A is composed of visual encoder and audio encoder for feature extraction, respectively. Then B is added the synchronization block that learns the correlation between features. C replaced the synchronization block with a more advanced multi-modal transformer. In this study, D is proposed and replaced with a VTP encoder that focuses on regional attention to lip movements

3. Method

3.1 Rationale

Most existing lip-sync detection models rely on a rough face-cropping method that detects a face and crops the lower half of the image for input. However, the lower half of facial images may include redundant areas, such as necks and clothes, in addition to the region of interest (ROI) around the mouth [15]. Moreover, the performance of the face-cropping methods depends on the angles of the target faces. Prajwal *et. al.* addressed this issue in the context of lip-reading tasks [16]. Instead of the rough face-cropping method, they utilize the transformer-based model VTP, which can learn where to direct visual attention on a given face image. After training, VTP mostly accords visual attention to the mouth area, and if necessary, also some level of visual attention to other facial areas for lip-reading. We adopted a pre-trained VTP module as the visual encoder for lip-synchronization detection.

3.2 VTPVocaList Architecture

The proposed method is designed based on VocaList. This section describes the entire vocalist pipeline of VTPVocaList. And Fig. 2 is shows VTPVocaList Architecture.

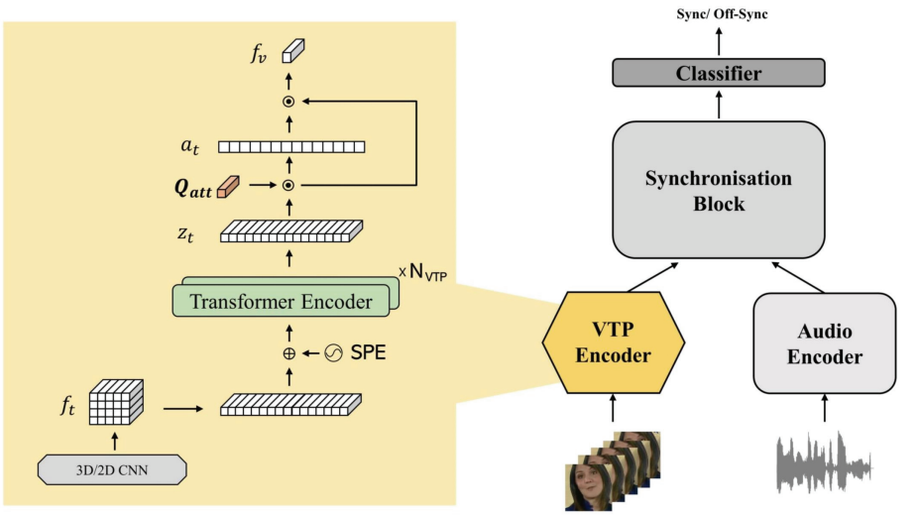

Fig. 2. VTPVocaList Architecture

1) VTP Encoder

As mentioned earlier, the VTP encoder is a model specifically trained to concentrate on lip movements without incorporating sound information in a lip-reading model [16]. The lip-reading model underwent a two-step training approach. Initially, the entire model underwent end-to-end training, after which the VTP encoder was frozen. Subsequently, the remaining script prediction Encoder-Decoder model was fine-tuned in the second stage of training. Hence, we adopted the well-pretrained VTP encoder, deemed effective, for the lip sync detection model. Both lip-reading model and VTP encoder were trained on three extensive datasets: LRS2, LRS3, and TEDx [7,8,16].

Fig. 3 depicts the visualizations of attention weights extracted from the VTP encoder on the LRS2 dataset. By utilizing visual inputs with emphasized lip regions, the VTP encoder helped enhance lip sync detection accuracy.

As a preprocessing procedure, VTPVocaList detects and crops facial areas by using a face tracker[14] after receiving an array of video frames. The initial part of the VTP encoder consisting of 3D/2D CNN layers extracts a local visual feature for each preprocessed frame. 3D/2D CNN is applied to 5 frames with a 1 frame stride.

Extracted feature is flattened and projected onto $f_t \in R^{hw \times c}$, where $t \in \{1,2,\ldots, T-1, T\}$, and T is the number of frames. The hw is (24,24) and c is 128. The VTP encoder then adds spatial positional encodings (SPE) to the visual feature vectors and passes them to the transformer encoders to produce a self-attended feature map Z_t.

Fig. 3. Visualization of the Attention Weights Extracted from the VTP Encoder on the LRS2 Dataset

$$Z_t = encoder_v(f_t + SPE_{1:hw}) \in R^{hw \times d}$$

After obtaining Z_t, VTP encoder computes the inner product between Z_t and the query vector Q_{att} to achieve the visual attention weight a_t. Then, Z_t and a_t are weighted and averaged to obtain the final feature vector $f_t^v \in R^{512}$, as follows:

$$a_t = softmax(Q_{att}^T Z_t) \in R^{hw \times 1}$$

$$f_t^v = \frac{1}{hw}\sum_{i=1}^{hw} a_t^i Z_t^i \in R^d$$

The VTP encoder uses a linear transformer that reduces computational complexity compared to the vanilla transformer [17]. The number of transformer encoder layers is 8 and a detailed description of VTP can be found [16].

2) Audio Encoder

The audio encoder consists of 2D CNN layers with residual skip connections. For 25 fps video clips, the audio encoder receives 16 mel-spectrograms, x_a, which match five frames. The mel-spectrogram was obtained using 80 mel-filters with hop size of 200 and window size of 800 from audio signals at 16 kHz sampling rate. Similar to the visual encoder, the audio encoder receives $x_t^a \in R^{80\times 16}$ mel-spectrograms and returns a 512-dim feature vector f_t^a.

$$x_t^a = encoder_a(x_t^a) \in R^d$$

3) Synchronization block

The synchronization block is a multi-modal transformer-based architecture introduced using VocaList. This approach was inspired by the cross-modal transformer architecture proposed in [18]. The objective of the design was to learn the correlated features of different modalities such as audio, video, and text. The synchronization block was divided into three transformer encoders: AV, VA, and hybrid fusion transformer. The keys, queries, and values of both the AV and VA transformer encoders come from the visual and audio encoders, and the outputs of both are passed to the hybrid fusion transformer encoder as input. The mathematical formulation is as follows:

$$g_t^{av} = AV-Transformer(f_t^a, f_t^v, f_t^v)$$
$$g_t^{va} = VA-Transformer(f_t^v, f_t^a, f_t^a)$$
$$out_t = F-Transformer(g_t^{av}, g_t^{va}, g_t^{va})$$

4) Classifier

The 512-dimensional output of the hybrid fusion transformer passes through a max-pooling layer and an activation layer, and then through an additional classifier and linear layer, to determine whether the video clip and audio segment are synchronized. A detailed description of both the audio encoder and synchronization block can be found in [6].

$$SyncScore_t = Classifier(out_t)$$

4. Experimnets

4.1 Training

As proposed for SyncNet, training was conducted in a self-supervised manner using contrastive learning [9]. For contrastive learning, two types of pairs, positive and negative, are randomly selected, where the positive pairs include well-synchronized video clips and audio segments, whereas the negative pairs include mis-aligned video clips and audio segments. Fig. 4 shows the process that makes the positive pairs and negative pairs. Using binary cross-entropy loss, VTPVocaList was trained as a binary classification model. And we utilized the pretrained VTP encoder on LRS2 and LRS datasets, consequently the VTP visual encoder remains frozen throughout the entire training process. We utilized the pretrained VTP encoder, assuming it was well-trained to focus solely on lip movements without auditory information.

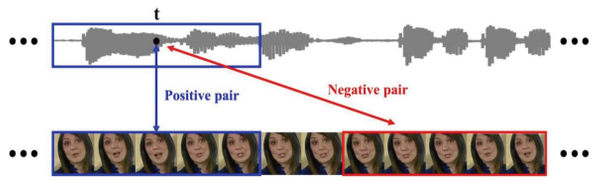

Fig. 4. Self-supervised Learning Method with Positive and Negative Pairs

The model training process took approximately three hours with an RTX 3090 GPU. All experiments ran on an Anaconda3 virtual environment with Python 3.9, CUDA 11.2 versions, and Jupyter Notebook.

4.2 Dataset

1) LRS2

The Oxford BBC lip-reading sentences 2 (LRS2) dataset contains hundreds of thousands of spoken sentences from the UK BBC broadcasting network [7]. The total dataset is divided into pretrain, training, validation, and test sets. In particular, the pretrain set contains 96K utterances and 2,064K words, and the test set contains 1,243 utterances and 6,663 words. We used the pretrain set for training and the test set for testing VTPVocaList.

2) LRS3

LRS3 is a large-scale audio-visual dataset collected from TED and TEDx videos [8]. The LRS3 dataset has not been used in the previous works. Although we did not use LRS3 for training, we used it as a test dataset to measure the lip-sync detection accuracy. LRS3 Test set contains 1,452 utterances and 11K words.

3) Acappella

Acappella is a dataset containing solo-singing videos

gathered from YouTube, covering a wide distribution of singers and languages [19]. It is divided into training, validation, and test set, and the test set is further divided into 'seen-heard' and 'unseen-unheard' subsets. The singers in seen-heard subset included those in the training set, and the unseen-unheard subset contained the rest. The 'unseen-unheard' subset was used for evaluation. Note that we did not use the Acappella dataset for training.

4.3 Evaluation Protocol

For a fair comparison, we followed the evaluation protocol suggested in previous studies [20]. As the video clip was shifted by a temporal stride of 1 from -15 to +15, the lip-sync detection model generated the corresponding sync scores. The prediction offset was determined based on the highest synchronization scores. If the offset was within a threshold ±1 frame for LRS2 and LRS3, lip-sync detection was considered correct. Accuracy is calculated as the number of windows where lip sync is considered correct divided by the total number of windows in the dataset.

PM initially introduced the context frame evaluation method, which involves averaging the surrounding embedding vectors in cases where meaningful information for determining lip sync within a 5-frame window cannot be found. PM and VocaList extended the context frame length from 5 to 15 or 25 depending on the dataset, and the average offsets were computed for comparison with the threshold. In particular, the threshold was set to ±5 for the Acappella dataset, following the experiments in a previous study on VocaList. They argued that determining lip sync in singing voices is much more challenging compared to typical speech datasets due to the greater prevalence of vowels in singing.

4.4 Experiments Results

After training VTPVocaList on the LRS2 training set, we evaluated its performance on three test datasets: LRS2, LRS3, and Acappella. Using the experimental measurements from previous literature restricted our selection of counterparts. For LRS2, SyncNet, PM, AVST, and VocaList were selected as counterparts. For LRS3, VocaList was selected as the counterpart. For Acappella, SyncNet and VocaList were selected as counterparts. Lip-sync detection accuracy was measured by varying the context frame length from 5 to 15 (or 25).

The accuracy evaluated on the test set of the LRS2 dataset is presented in Table 2. The results indicate an improvement in accuracy as the model advances, and a corresponding increase in accuracy with longer context frame lengths. Although VocaList and VTPVocaList demonstrated similar performance with 15 frames, VTP VocaList, leveraging a pretrained VTP encoder, has approximately half the number of trainable parameters compared to VocaList. The numbers within parentheses and * represent the parameter count, including the pretrained VTP encoder. Even when combined, this count is over 30% less than that of VocaList.

Table 3 presents the accuracy evaluated on the test set of the LRS3 dataset. LRS3, which has not been used in previous studies and for which the code is not publicly available, was exclusively compared with the evaluable

Table 2. Comparison of Lip-sync Detection Accuracy on LRS2

Model	Trainable Params	Context frame length					
		5(0.2s)	7(0.28s)	9(0.36s)	11(0.44s)	13(0.52s)	15(0.6s)
SyncNet	13.6M	75.8	82.3	87.6	91.8	94.5	96.1
PM	13.6M	88.1	93.8	96.4	97.9	98.7	99.1
AVST	42.4M	92.0	95.5	97.7	98.8	99.3	99.6
VocaList	80.1M	92.8	96.7	98.4	99.3	99.6	***99.8***
VTP VocaList (ours)	41.4M (54.3M)*	***94.5***	***97.6***	***99.0***	***99.5***	***99.7***	***99.8***

Table 3. Comparison of Lip-sync Detection Accuracy on LRS3

Model	Context frame length					
	5(0.2s)	7(0.28s)	9(0.36s)	11(0.44s)	13(0.52s)	15(0.6s)
VocaList	75.25	82.37	86.69	89.55	91.29	92.45
VTP VocaList (ours)	***79.38***	***84.88***	***88.14***	***90.33***	***91.63***	***92.48***

Table 4. Comparison of Lip-sync Detection Accuracy on Acappella

Model	Trainded dataset	Context frame length				
		5(0.2s)	10(0.4s)	15(0.6s)	20(0.8s)	25(1s)
SyncNet	Acappella	57.7	63.9	69.9	75.1	78.7
VocaList	LRS2	56.7	65.1	72.2	77.2	81.2
VocaList	Acappella	58.8	65.4	71.6	76.5	80.5
VTP VocaList (ours)	LRS2	***66.8***	***74.7***	***80.6***	***84.8***	***87.8***

VocaList. Even when evaluated using LRS3, which was not used in training, VTPVocaList demonstrates superior performance.

Table 4 shows the results of the comparison on Acappella 'unseen-unheard' test set. Notably, VTPVocaList trained on LRS2 outperforms VocaList trained on the Acappella training set.

5. Conclusion

In this study, several lip-sync detection models following SyncNet were summarized. The architectures of these models comprise four parts: a visual encoder, an audio encoder, a module to learn the correlation between features, and a classifier. For lip-sync detection, this study proposes *VTPVocaList*. which leverages the pre-trained VTP as part of the visual encoder. Using the attention mechanism of the transformer model, VTP enables the visual encoder to learn to focus on lip areas that are directly related to lip-voice synchronization. The experiments on three well-known datasets showed that VTPVocaList surpasses existing models in terms of lip-sync detection accuracy, even with fewer trainable parameters than its counterparts.

The VTPVocaList is designed against joint training for the VTP module, deeming the pre-trained VTP module effective in focusing on lip regions without relying on sound information. Despite surpassing the accuracy of prior studies, the rate of improvement becomes marginal as the length of the context frame used in the evaluation method increases. This diminishing improvement rate is attributed to the encoder's ability to extract abstract and informative features while filtering out redundancies. In harnessing local and visual attention information within image frames via the VTP module, we acknowledge the potential for future research to explore attention utilization across temporal sequences, offering a promising avenue to further diminish feature redundancies.

References

[1] Y. Shalev and L. Wolf, "End to end lip synchronization with a temporal autoencoder," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[2] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," *Proceedings of the 28th ACM International Conference on Multimedia*, pp.484-492, 2020.

[3] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.7613-7617, 2021.

[4] T. Makino et al., "Recurrent neural network transducer for audio-visual speech recognition," In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, pp.905-912, 2019.

[5] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fas autoregressive transformers with linear attention," In: *International Conference on Machine Learning. PMLR*, pp.5156-5165, 2020.

[6] V. S. Kadandale, J. F. Montesinos, and G. Haro, "VocaLiST: An audio-visual synchronisation model for lips and voices," In: *Interspeech*, pp.3128-3132, 2022.

[7] 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.

[8] T. Afouras, J. S. Chung, and A. Zisserman. "LRS3-TED: a large-scale dataset for visual speech recognition," In: *arXiv preprint arXiv:1809.00496*, 2018.

[9] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE. Vol.1, pp.539-546, 2005.

[10] B. V. Mahavidyalaya. "Phoneme and viseme based approach for lip synchronization," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol.7, No.3, pp.385-394, 2014.

[11] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," In: *Workshop on Multi-view Lip-reading*, ACCV. 2016.

[12] Y. J. Kim, H. S. Heo, S. W. Chung, and B. J. Lee, "End-to-end lip synchronisation based onpattern classification," In: *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp.598-605, 2021.

[13] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] A. Bulat and G. Tzimiropoulos. "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)," In: *International Conference on Computer Vision*, 2017.

[15] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, pp.356-363, 2020

[16] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5162-5172, 2022.

[17] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," In: *International Conference on Machine Learning*, PMLR, pp.5156-5165, 2020.

[18] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L-P Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," In: *Proceedings of the Conference Association for Computational Linguistics*, Meeting, NIH Public Access, pp.6558, 2019.

[19] J. F. Montesinos, V. S. Kadandale, and G. Haro, "Acappella: audio-visual singing voice separation," In: *32nd British Machine Vision Conference*, BMVC 2021, 2021.

[20] S.-W. Chung, J. S. Chung, and H.-G. Kang. "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.3965-3969, 2019.

[21] H. Gupta, "Perceptual synchronization scoring of dubbed content using phoneme-viseme agreement," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Dongryun Yoon

https://orcid.org/0009-0006-7554-3722
e-mail : yddr1209@korea.ac.kr

Bachelor's degree in Computer Convergence Software Engineering from Korea University in 2021. Master's degree in Computer Science from Korea University in 2023. Researcher at the Korea Electronics Technology Institute since 2023. Her research interests includes Computer Vision and Material Informatics.

Hyeonjoong Cho

https://orcid.org/0000-0003-1487-895X
e-mail : raycho@korea.ac.kr

Bachelor's degree in Electronic Engineering from Kyungpook National University in 1996. Master's degree in Electrical and Electronic Engineering from Pohang University of Science and Technology in 1998. Ph.D. in Computer Engineering from Virginia Tech in 2006. Professor at the Department of Computer Convergence Software at Korea University since 2016. His research interests include Machine Learning, Computer Vision, and Time Series Data Analysis.