Conformer-Based Multi-Modal Learning for Cued Speech Recognition from Videos

Katerina Papadimitriou*† and Gerasimos Potamianos*†

*Department of Electrical & Computer Engineering, University of Thessaly, 38334 Volos, Greece
†Robotics Institute, Athena Research Center, 15125 Maroussi, Greece
k.papadimitriou@athenarc.gr, gpotam@athenarc.gr

***Abstract*—This paper presents a novel multimodal framework for automatic cued speech recognition (ACSR) based on a Conformer architecture that directly processes whole upper-body video frames, eliminating the need for explicit segmentation or synchronization of hand and lip regions. Unlike prior approaches that separately process hand and mouth cues, our model jointly learns appearance and skeletal representations, effectively handling hand-lip asynchrony. Our framework consists of two branches (streams): (i) an appearance-based stream, utilizing the ResNet18 model for RGB feature extraction, and (ii) a skeletal-based stream, employing a modulated graph convolutional network (GCN) to process 3D joint coordinates extracted via MediaPipe. To learn temporal dependencies, we integrate a temporal convolutional network (TCN) for short-range temporal modeling and a Conformer encoder for long-range sequence learning. In addition, to enhance feature alignment and improve phoneme sequence learning, we incorporate an auxiliary loss, ensuring robust multimodal fusion. Extensive evaluations on three benchmark cued speech datasets, including French, British English, and Mandarin Chinese, demonstrate that our model achieves state-of-the-art performance, outperforming existing approaches.**

***Index Terms*—cued speech recognition, 3D skeleton, graph convolutional network, Conformer, alignment module**

I. Introduction

Cued speech (CS) constitutes a visual-based communication tool introduced by Cornett [1] to enhance speech perception for individuals who are deaf or hard-of-hearing. Unlike conventional lipreading, which is often ambiguous due to the many-to-one mapping between phonemes and visemes, CS provides an unambiguous visual representation of spoken language by coupling lip articulation with distinct hand configurations and placements. In particular, consonants are encoded through distinct hand gestures, while vowels are represented by specific hand positions relative to the mouth, together with mouthing patterns [2]. Figure 1 illustrates the encoding process for French CS, where eight different handshapes at five unique positions are employed to cover its phonemic inventory [3].

Performing ACSR from videos constitutes a challenging task due to: (i) the multimodal complexity of jointly processing both manual and non-manual cues simultaneously, while handling the inherent asynchrony between them (hand cues often lead the corresponding lip movements); (ii) the articulation variability across cuers, differences in video recording conditions and occasional articulation occlusions; and (iii) the limited availability of large-scale annotated CS datasets [4]–[8].

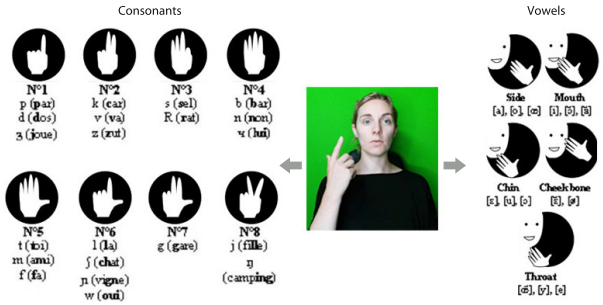

Fig. 1. French Cued Speech: Hand shapes and positioning for consonants and vowels (modified from [3]).

The first attempts to address ACSR in the literature relied on artificial markers and colored gloves in order to facilitate hand and lip tracking, simplifying segmentation in controlled environments [9]. Traditional recognition systems then mapped hand-crafted visual features to phonemes using GMMs or HMMs. Such approaches, however, struggled to generalize beyond controlled settings [9], [10]. As computer vision techniques evolved, researchers moved towards markerless tracking methods, for example Kanade-Lucas-Tomasi lip tracking along with statistical segmentation techniques for hand tracking, such as Adaptive Background Mixture Models [4], [11], [12]. In terms of feature extraction and classification techniques, recent ACSR works combine CNN-based visual feature learners [13]–[15] with hybrid classifiers, such as GMM-HMMs [4], [12], or recurrent neural networks in conjunction with CTC decoders [11], [13], [16]–[18]. Various works investigate skeletal-based body representations [16], [17], employing pose estimation models [19]–[22] to deduce joint coordinates from the hands and lips, while others combine hand positioning relative to the mouth with both appearance and skeletal features [15]. Still, aligning the hand and mouth modalities remains non-trivial, with some methods introducing algorithmic fusion schemes to dynamically synchronize hand and lip features [13], [14].

This work has received funding from the EU's Horizon Europe research and innovation programme under grant no. 101070381 (project: PILLAR-Robots).

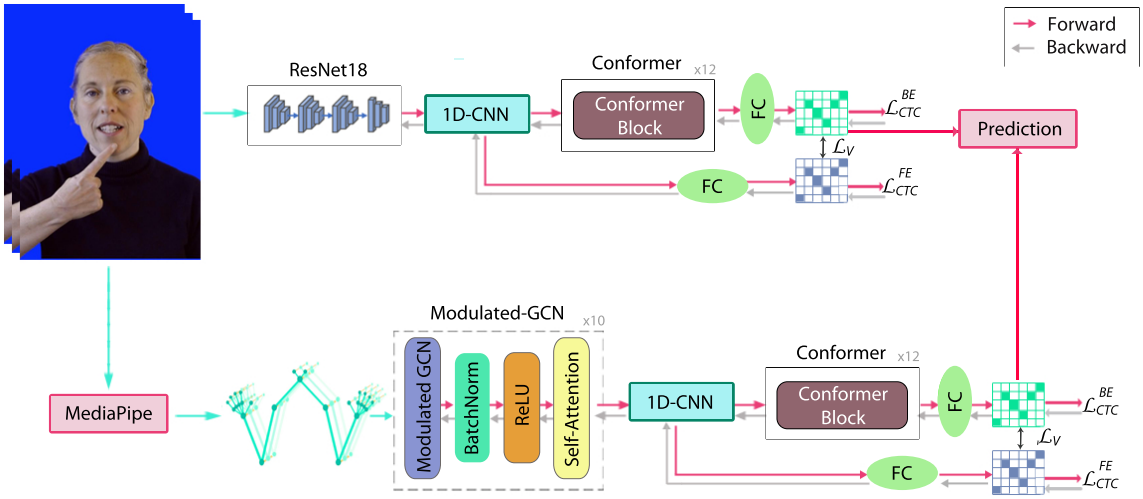

Fig. 2. Architecture of the proposed ACSR system that integrates both RGB and skeletal streams to predict phoneme sequences. The system utilizes a ResNet18 for appearance feature extraction, a modulated GCN for skeletal representation, and a 1D-CNN/Conformer network for sequence learning. The final prediction is generated using CTC loss functions along with an auxiliary loss to enhance both long-range and short-range phoneme alignment.

To improve upon these approaches, our early work in [3] introduced a multi-stream 3D-CNN front-end, coupled with a time-depth separable convolutional encoder with an attention-based decoder. This model performed direct fusion of the modalities without explicit synchronization, allowing the network to implicitly learn the hand-lip timing interaction. Building on this, our more recent work has turned to human pose estimation models [23], which provide structured representations of hand and lip movements by extracting joint coordinates from these regions. Such skeletal-based representations enhanced the modeling of spatial relationships between lip and hand cues, offering a more robust way to align their features.

Here, in this paper, we extend our previous work by introducing several novel components in our approach. First, we employ the ResNet18 model [24] as the CNN backbone for RGB data feature extraction, allowing for efficient representation of spatial CS information. Second, we incorporate a modulated GCN [25], [26] to model skeletal features, leveraging the MediaPipe model to extract 3D joint coordinates. Modulated GCNs can learn meaningful geometric relationships that are not easily captured by raw RGB data features, also enabling a more effective representation of hand placement relative to the mouth, compared to previous methods that only relied on coordinate-based embeddings. Third, for sequence learning, we integrate a TCN for short-range modeling and a Conformer encoder [27] for long-range sequence learning in CS videos. In particular, the Conformer encodes global temporal relationships through attention, while modeling local context via convolution, thus providing a structured and effective approach to capturing both local and global temporal dependencies. In addition, Conformers can better generalize with limited data, thereby mitigating the impact of CS data scarcity. Fourth, to ensure robust sequence learning, we employ well-established loss functions, further enhancing the performance of the temporal modeling components. In particular, we introduce an auxiliary loss (KL-divergence) that serves as a regularization mechanism, ensuring feature consistency across modalities and enabling robust multimodal fusion.

In summary, the key contributions of this work include: (i) the development of an innovative multimodal ACSR framework that integrates both appearance and skeletal information from whole upper-body frames, unlike prior approaches that process hand and lip regions separately; (ii) the introduction of a modulated GCN for more effective skeletal feature representation; (iii) the employment of a Conformer for sequence modeling, providing robust handling of long-range dependencies; and (iv) the incorporation of an auxiliary loss to optimize multimodal fusion and improve phoneme sequence learning. It is worth noting that our work represents the first ever use of GCNs, Conformers, and the KL-divergence loss function to the problem of ACSR in the literature.

Experimental results demonstrate that our model outperforms previous approaches, achieving state-of-the-art recognition on three benchmark CS datasets: French [7], British English [5], and multi-cuer Mandarin Chinese [6]. Specifically, our model yields absolute error reductions of 9.45%, 10.19%, and 7.91% on the three datasets, respectively, compared to the next-best results in the literature. Note that we also report a cuer-independent ACSR result in our experiments.

II. The Proposed Multi-Modal Framework

To tackle ACSR, we introduce a novel multi-stream architecture, as depicted in Figure 2. The system consists of two branches (streams): (i) an appearance-based stream, which employs a 2D-CNN for visual feature extraction; and (ii) a skeletal stream, which processes 3D skeletal joints using a modulated GCN. In both branches, the modules are followed by a 1D-CNN and a Conformer encoder to model both spatial and temporal dependencies. The two streams are trained independently using an alignment module that relies on CTC loss functions and an auxiliary KL-divergence loss, ensuring effective stream fusion during inference.

A. Appearance Modeling

Unlike our previous works [3], [23], where the hand and mouth regions were processed independently, our method preserves the holistic spatial information of visual cues. Specifically, our model processes the entire upper-body region, enabling the simultaneous capture of both hand and mouth articulations without requiring explicit synchronization between the two regions. To extract the upper-body area, we utilize the MediaPipe holistic human pose detector [21], which estimates 543 keypoints across the body, including 33 for the torso, 468 for the face, and 21 for each hand. In particular, a bounding box is computed based on the minimum and maximum x and y coordinates of the relevant keypoints. If the MediaPipe detector fails, missing positions are replaced with the last detected ones. To extract spatial feature embeddings from the upper-body region, we use a ResNet18 model [24] pretrained on ImageNet [28]. Upper-body images are resized to 256×256 pixels before being fed into the network. The output feature maps undergo global average pooling, yielding a 512-dimensional (dim) vector per frame.

B. Graph-Based 3D Skeletal Modeling

To enhance the effectiveness of our framework, we incorporate a supplementary stream that processes the skeletal data of the cuer. This stream operates on 3D joints deduced from MediaPipe [21], utilizing a GCN-based module to capture the spatial dependencies between articulatory joints. In particular, we employ a modulated GCN [25], [26] coupled with a self-attention mechanism [29]. Unlike traditional pose learning methods that might rely on a large set of joints, our model simplifies the skeletal graph by retaining only the most relevant 3D joints. Specifically, we focus on 10 joints for each hand and 7 for the upper-body, including 3 facial joints (see also Figure 2), which are critical for accurate phoneme recognition. Once these 27 3D joints are extracted, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is generated, where $\mathcal{V}$ denotes the set of 27 joints serving as nodes, and $\mathcal{E}$ defines the edges connecting these nodes. Each node i is associated with a feature vector $\mathbf{q}_i \in \mathbb{R}^D$, where $D = 3$, representing its position in the 3D space. Subsequently, the graph is passed through a modulated GCN, where the network learns the complex dependencies between the joints. The GCN uses node-specific learnable modulation vectors to adapt a shared weight matrix, and employs affinity modulation by learning a mask over the adjacency matrix to better capture task-relevant joint relationships.

To further refine the model's ability to capture dynamic joint interactions, we integrate a self-attention mechanism into the GCN. This enables the network to dynamically focus on important features across spatial, temporal, and channel dimensions. To improve generalization and prevent overfitting, we incorporate a DropGraph technique [30], which randomly drops edges (20%) in the graph during training. Our system employs 10 GCN layers, each augmented with a self-attention mechanism. The resulting skeletal representations undergo global average pooling, resulting in a 512-dim feature vector per frame.

C. Temporal Sequence Modeling

To model both short-range articulation patterns and long-range phonetic dependencies, the extracted feature maps from both branches are each processed separately by a temporal modeling pipeline, consisting of a temporal convolutional layer (1D-CNN) followed by a Conformer encoder [27]. The Conformer block is composed of four key modules stacked together: multi-head self-attention, convolution, and two feed-forward modules. This hybrid structure allows the model to efficiently capture both long-range phoneme dependencies and local articulatory patterns. The self-attention mechanism processes the input sequence $\mathbf{s} \in \mathbb{R}^{T\times 512}$ by projecting it into queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) via linear transformations. These are split across n_h attention heads, yielding $\mathbf{Q}^h, \mathbf{K}^h, \mathbf{V}^h \in \mathbb{R}^{T\times(512/n_h)}$ for each head h. Each head computes scaled dot-product attention independently as:

$$\textbf{attn}^h = \frac{\mathbf{Q}^h(\mathbf{K}^h)^\top}{\sqrt{512/n_h}}, \quad \text{where} \quad \textbf{attn}^h \in \mathbb{R}^{T\times T}.$$

The attention scores $\mathbf{attn}^h$ are normalized via softmax to produce weights, which are then applied to the corresponding values $\mathbf{V}^h$. The outputs from all heads are concatenated and projected through a final linear layer. Local feature extraction is further refined by the Conformer's convolutional module, while its dual feed-forward layers maintain a balance between local and global context. Our sequence model comprises 12 stacked Conformer blocks. The final representations are projected via a fully connected layer followed by softmax, yielding phoneme probability distributions.

D. Multimodal Alignment and Training Strategy

The appearance and skeletal modalities are trained independently using a CTC loss $\mathcal{L}_{CTC}^{BE}$, applied to the back-end predictions for phoneme alignment without requiring frame-level annotations. To improve temporal modeling, an additional CTC loss $\mathcal{L}_{CTC}^{FE}$ is applied to front-end posteriors derived from short-range features processed by the 1D-CNN and a fully connected layer with softmax. Inspired by [31], we also add a KL-divergence loss $\mathcal{L}_V = \text{KL}(\text{softmax}(\mathbf{D}_{FE}), \text{softmax}(\mathbf{D}_{BE}))$ to align front-end and back-end distributions. The final objective is: $\mathcal{L}_M = \mathcal{L}_{CTC}^{BE} + \mathcal{L}_{CTC}^{FE} + 0.5\mathcal{L}_V$.

E. Fusion and Inference Strategy

During inference, the appearance and skeletal modalities are combined at the probability level through a late fusion strategy. In particular, each branch independently processes input sequences through its respective 1D-CNN and Conformer encoder, generating modality-specific predictions. The posterior probability distributions produced by the final fully connected layer of both streams are then merged using a weighted fusion scheme. Specifically, each modality is assigned a distinct weight, heuristically chosen based on its validation performance, ensuring that the more reliable modality contributes more significantly to the final prediction. In

TABLE I
PER (%) COMPARISON FOR DIFFERENT MODALITIES ACROSS DATASETS.

Dataset	RGB	Skeletal	Both
French	12.40	18.60	**11.25**
British	20.32	31.52	**18.41**
Mandarin	9.27	12.47	**6.89**

TABLE II
ABLATION STUDY EVALUATING THE IMPACT OF FEATURE LEARNERS, SEQUENCE LEARNING AND AUXILIARY LOSS FUNCTIONS ON ACSR PERFORMANCE REGARDING THE RGB STREAM ALONE. THE EVALUATION IS CONDUCTED IN TERMS OF PER (%) ON THE FRENCH CS DATASET.

Appearance Features	Sequence Learning	$\mathcal{L}_{CTC}^{BE}$	$\mathcal{L}_{CTC}^{FE}$	$\mathcal{L}_{V}$	French
VGG	TCN/LSTM	✓	✓	✓	13.23
	TCN/Transformer	✓	✓	✓	14.41
	LSTM	✓	✓	✓	16.11
	Transformer	✓	✓	✓	17.58
ResNet18	TCN/LSTM	✓	✓	✓	12.97
	TCN/Transformer	✓	✓	✓	13.56
	LSTM	✓	✓	✓	15.27
	Transformer	✓	✓	✓	16.78
	TCN/Conformer	✓			22.94
	TCN/Conformer	✓	✓		14.89
	Ours	✓	✓	✓	**12.40**

particular, the weighted sum of these posteriors is computed as: $\mathbf{p}_{\text{fused}} = 1.0\,\mathbf{p}_{\text{app}} + 0.8\,\mathbf{p}_{\text{skel}}$.

III. EXPERIMENTAL FRAMEWORK

We evaluate our proposed approach on three CS datasets: French [7], British English [5], and Mandarin Chinese [6]. For the French CS dataset, we follow the official data split, using 979 videos for training, 108 for validation, and 108 for testing. This dataset contains high-resolution (1920×1080) videos at 60 fps, representing 34 phonetic classes. The British English CS corpus consists of 98 sentence-based videos, and we employ a 5-fold cross-validation strategy, dividing the data into 60% training, 20% validation, and 20% testing. The videos are recorded at 1280×720 resolution with a frame rate of 25 fps, encoding 44 phonemes. Finally, the Mandarin Chinese CS dataset includes 4,000 videos (1,000 sentences) from 4 different cuers, captured at 1280×720 resolution and 30 fps. For the latter dataset, we evaluate our model under two experimental setups: (i) A multi-cuer (MC) split, where the dataset is divided into 4 folds with 60% allocated for training, 20% for validation, and 20% for testing, and (ii) a cuer-independent (CI) split, where training and validation data come from three cuers, while the fourth cuer's data are exclusively used for testing. This process is repeated four times, reporting the average performance across all folds.

Our model is trained for 50 epochs with a batch size of 2, employing the Adam optimizer [32] with an initial learning rate of 0.0001, which is reduced by a factor of 0.5 after each iteration. To improve generalization, we apply data augmentation techniques, including random cropping and horizontal flipping. Additionally, skeletal joint coordinates are normalized to the image plane based on the image width and height. All experiments are conducted on an NVIDIA RTX 3090 GPU.

TABLE III
COMPARISON OF PER (%) FOR STATE-OF-THE-ART METHODS ON THE FRENCH AND BRITISH ENGLISH CS DATASETS, WITH THE FOLLOWING NOTATION: HAND (H), MOUTH (M), AND HAND POSITION (P).

Model	Feature streams	French	British
Fully Conv [3]	H+M+P	-	36.25
TDS-CTC [23]	H+M+P+Skel.	-	32.58
Student CTC [13]	H+M+P	-	28.6
CB + VLA [14]	H+M	-	33.6
3S-BiGRUs [16]	H+M+Skel.	20.7	-
3S-BiGRUs+LM [17]	H+M+Skel.	25.8	-
Ours	Full Frame	**11.25**	**18.41**

TABLE IV
COMPARISON OF PER (%) FOR STATE-OF-THE-ART METHODS ON THE CHINESE CS DATASET UNDER MC AND CI SETTINGS, WITH THE FOLLOWING NOTATION: HAND (H), MOUTH (M), HAND POSITION (P).

Model	Feature streams	MC	CI
Student CTC [13]	H+M+P	68.2	-
CB + VLA [14]	H+M	24.5	-
FedCSR [33]	H+M+Skel.	14.80	-
Ours	Full Frame	**6.89**	**33.14**

IV. EXPERIMENTAL RESULTS

We next report our experiments. We measure performance using the Phoneme Error Rate (PER, %) on the datasets of Section III. We begin by analyzing the contribution of individual modalities, followed by an ablation study investigating the impact of sequence modeling choices, skeletal feature representations, and auxiliary loss functions. Finally, we compare our approach against state-of-the-art methods.

To understand the role of each modality, we compare the RGB-only and skeletal-only models with the fully fused system. As shown in Table I, the RGB modality outperforms the skeletal modality across all datasets. However, fusing both modalities further enhances recognition accuracy, yielding absolute PER reductions of 1.15% on French, 1.91% on British English, and 2.38% on Mandarin Chinese, compared to the RGB-only model. This indicates that while the RGB modality captures more discriminative features, the skeletal modality provides complementary information that improves phoneme recognition.

To evaluate the impact of different components, we conduct an ablation study (Table II), focusing on sequence modeling choices and auxiliary loss functions. In particular, we compare the performance of the combination of TCN with Conformer, LSTM, and Transformer encoders, as well as their conjunction with VGG and ResNet18 feature learners to assess their effectiveness in sequence modeling. Results indicate that the Conformer consistently achieves the lowest PER, demonstrating its ability to model both short-range articulation patterns and long-range phoneme dependencies. The LSTM encoder, while effective in capturing temporal relationships, shows limitations in handling long sequences, whereas the Transformer struggles due to its lack of local feature modeling. We also observe that when the TCN is removed from the pipeline, the PER increases, confirming that the TCN is essential for capturing short-range articulation patterns. Additionally, we assess the

effect of the KL-divergence loss ($\mathcal{L}_V$) on phoneme alignment. Results show that removing this auxiliary loss leads to a noticeable increase in PER, validating the effectiveness of our alignment-driven optimization strategy.

Table III compares our model against state-of-the-art ACSR approaches on the French and British English CS datasets. Our model achieves the lowest PER of 11.25% on French and 18.41% on British English, significantly outperforming existing methods. Unlike prior approaches that rely on explicit hand-mouth segmentation and synchronization mechanisms, our framework directly processes the full upper-body region, leading to more robust feature representations. We further evaluate our approach on the Mandarin Chinese CS dataset under both MC and CI settings (Table IV). Our model achieves a PER of 6.89% in MC and 33.14% in CI.

V. Conclusions

In this work, we introduced a Conformer-based multimodal framework for ACSR that effectively captures both appearance and skeletal articulation, addressing hand-lip asynchrony without requiring explicit segmentation. By processing whole upper-body frames, our model leverages a ResNet18 for appearance modeling and a modulated GCN for skeletal representation, both refined through a TCN and Conformer encoder for robust sequence learning. With the integration of a KL-divergence auxiliary loss, our method significantly enhances multimodal feature alignment, improving phoneme sequence prediction. Extensive evaluations on three benchmark datasets confirm that our system outperforms all existing approaches, achieving state-of-the-art performance in ACSR.

References

[1] R. O. Cornett, "Cued speech," *American Annals of the Deaf*, vol. 112, no. 1, pp. 3–13, 1967.

[2] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun, "Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech." *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1144–1153, 2005.

[3] K. Papadimitriou and G. Potamianos, "A fully convolutional sequence learning approach for cued speech recognition from videos," in *Proc. EUSIPCO*, 2021, pp. 326–330.

[4] L. Liu, T. Hueber, G. Feng, and D. Beautemps, "Visual recognition of continuous cued speech using a tandem CNN-HMM approach," in *Proc. Interspeech*, 2018, pp. 2643–2647.

[5] L. Liu, J. Li, G. Feng, and X. Zhang, "Automatic detection of the temporal segmentation of hand movements in British English cued speech," in *Proc. Interspeech*, 2019, pp. 2285–2289.

[6] L. Liu and G. Feng, "A pilot study on Mandarin Chinese cued speech," *American Annals of the Deaf*, vol. 164, no. 4, pp. 496–518, 2019.

[7] S. Sankar, D. Beautemps, and T. Hueber, "The CSF22 database," Sep. 2023. [Online]. Available: http://dx.doi.org/10.5281/zenodo.8392607

[8] L. Gao, S. Huang, and L. Liu, "A novel interpretable and generalizable re-synchronization model for cued speech based on a multi-cuer corpus," in *Proc. Interspeech*, 2023, pp. 3407–3411.

[9] P. Heracleous, D. Beautemps, and N. Hagita, "Continuous phoneme recognition in cued speech for French," in *Proc. EUSIPCO*, 2012, pp. 2090–2093.

[10] P. Heracleous, D. Beautemps, and N. Aboutabit, "Cued speech automatic recognition in normal-hearing and deaf subjects," *Speech Communication*, vol. 52, no. 6, pp. 504–512, 2010.

[11] L. Liu, G. Feng, and D. Beautemps, "Automatic temporal segmentation of hand movements for hand positions recognition in French cued speech," in *Proc. ICASSP*, 2018, pp. 3061–3065.

[12] L. Liu, G. Feng, D. Beautemps, and X. Zhang, "A novel resynchronization procedure for hand-lips fusion applied to continuous French cued speech recognition," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[13] J. Wang, Z. Tang, X. Li, M. Yu, Q. Fang, and L. Liu, "Cross-modal knowledge distillation method for automatic cued speech recognition," in *Proc. Interspeech*, 2021, pp. 2986–2990.

[14] L. Liu and L. Liu, "Cross-modal mutual learning for cued speech recognition," in *Proc. ICASSP*, 2023, pp. 1–5.

[15] L. Liu, L. Liu, and H. Li, "Computation and parameter efficient multimodal fusion transformer for cued speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1559–1572, 2024.

[16] S. Sankar, D. Beautemps, and T. Hueber, "Multistream neural architectures for cued speech recognition using a pre-trained visual feature extractor and constrained CTC decoding," in *Proc. ICASSP*, 2022, pp. 8477–8481.

[17] S. Sankar, D. Beautemps, F. Elisei, O. Perrotin, and T. Hueber, "Investigating the dynamics of hand and lips in French cued speech using attention mechanisms and CTC-based decoding," in *Proc. Interspeech*, 2023, pp. 4978–4982.

[18] S. Sankar, "Automatic recognition and generation of French Cued Speech using deep learning," Ph.D. Thesis, Université Grenoble Alpes, 2024.

[19] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[20] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. CVPR*, 2017, pp. 4645–4653.

[21] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for perceiving and processing reality," in *Proc. CV4ARVR*, 2019.

[22] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos," in *Proc. ECCVW (SLRTP)*, 2020, pp. 249–263.

[23] K. Papadimitriou, M. Parelli, G. Sapountzaki, G. Pavlakos, P. Maragos, and G. Potamianos, "Multimodal fusion and sequence learning for cued speech recognition from videos," in *Proc. HCII*, 2021, pp. 277–290.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[25] Z. Zou and W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *Proc. ICCV*, 2021, pp. 11 457–11 467.

[26] K. Papadimitriou and G. Potamianos, "Sign language recognition via deformable 3D convolutions and modulated graph convolutional networks," in *Proc. ICASSP*, 2023, pp. 1–5.

[27] A. Gulati, C. Chiu, J. Qin, J. Yu, N. Parmar, R. Pang, S. Wang, W. Han, Y. Wu, Y. Zhang, and Z. Zhang, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[29] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[30] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling GCN with DropGraph module for skeleton-based action recognition," in *Proc. ECCV*, 2020, pp. 536–553.

[31] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proc. ICCV*, 2021, pp. 11 522–11 531.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[33] Y. Zhang, L. Liu, and L. Liu, "Cuing without sharing: A federated cued speech recognition framework via mutual knowledge distillation," in *Proc. ACMMM*, 2023, pp. 8781–8789.

This project is funded by the European Union under Horizon Europe (grant No. 101136568 - project HERON).

Funded by the European Union