



# A Multi-Stream Framework Utilizing 3D Human Reconstruction for Cued Speech Recognition

Katerina Papadimitriou<sup>1,2</sup>, Gerasimos Potamianos<sup>1,2</sup>

<sup>1</sup>Department of Electrical & Computer Engineering, University of Thessaly, 38334 Volos, Greece

<sup>2</sup>Robotics Institute, Athena Research Center, 15125 Maroussi, Greece

k.papadimitriou@athenarc.gr, gpotam@athenarc.gr

## Abstract

In this paper, we propose a novel multi-stream framework for automatic cued speech recognition (ACSR) that directly processes the upper-body region, addressing hand-lip asynchrony without requiring explicit segmentation or synchronization. Our model integrates two distinct modalities: (i) an appearance-based stream leveraging the ResNet18 for feature extraction and (ii) a skeletal stream based on a modulated graph convolutional network (GCN). For graph construction, we incorporate, for the first time in ACSR, 3D pose parameters inferred from the PIXIE model. Both modalities are coupled with temporal convolution for short-range dynamics learning and a BiGRU encoder for long-term sequence modeling. In addition, we introduce an alignment module that combines CTC with two auxiliary losses, improving each modality performance and enabling effective late fusion during inference. Our model achieves state-of-the-art performance across three benchmark datasets, demonstrating its effectiveness.

**Index Terms:** cued speech recognition, 3D human reconstruction, PIXIE, GCN

## 1. Introduction

Cued speech (CS) is a visual communication system designed by Cornett [1] to facilitate speech perception for the deaf and hard-of-hearing. Unlike traditional speechreading, which suffers from viseme-to-phoneme ambiguity, CS offers a complete visual representation of spoken language, by combining lip contour patterns with hand positional and gestural cues. Specifically, hand configurations represent consonants, while hand positions relative to the mouth, along with mouthing patterns, encode vowels [2]. Given its structured nature, CS has been adopted in multiple languages, such as French, British English, and Mandarin Chinese [3].

Despite the growing research interest in ACSR, the task remains challenging, primarily due to the intricate interaction between hand and mouth cues, necessitating their simultaneous processing. Complexity increases with the natural asynchrony occurring between modalities, where the manual cue often precedes mouth articulation by approximately one syllable. Beyond multimodal complexity, data scarcity further exacerbates the problem. While various single-cue and multi-cue CS corpora have been collected [4–8], the availability of large-scale annotated datasets remains limited. The absence of extensive training corpora hinders the development of robust, data-driven

models, often resulting in reduced generalization across different cueders and linguistic contexts. In addition, the natural inter-cue articulation variability, hand positioning inconsistencies, and occlusions further hinder recognition. This paper addresses these challenges, developing a robust system capable of accurately predicting a sequence of phonemes from RGB video without prior knowledge of phoneme-level segmentation.

Accurate detection and tracking of both manual and non-manual cues is a critical aspect of ACSR. Early approaches employ artificial markers to facilitate articulation tracking [9, 10], while most contemporary ACSR schemes rely on computer vision techniques. For instance, some works [4, 11, 12] explore Kanade-Lucas-Tomasi lip tracking, as well as Gaussian mixture models (GMMs) and adaptive background mixture models for hand region segmentation. Other works, including our previous work [13], exploit the dynamic nature of gestures by incorporating motion-based algorithms for tracking. With advances in deep learning, recent approaches have shifted towards 2D and 3D human skeletal-based detection [14–17], leveraging pose regression models [18–21]. Given that hand articulation typically precedes lip articulation, many studies introduce explicit synchronization mechanisms, such as shifting the hand stream using a predefined temporal offset, as our previous work [13], or employing cross-modal attention techniques [22, 23]. Our first innovation here is that our proposed approach abstains from hand and mouth cue segmentation and synchronization, directly processing the entire upper-body region. This enables the model to learn synchronization implicitly, without ad-hoc alignment techniques. By processing the entire upper-body region, our framework automatically captures multi-cue dependencies, implicitly.

Another key aspect of ACSR is the visual representation of manual and non-manual articulation, as well as the modalities used. Most studies extract RGB appearance features from lip and hand regions using 2D- or 3D-CNNs [12–14, 17, 22, 23]. Recent works have explored skeletal-based representations [15, 16], employing pose estimation models to derive 3D joint coordinates, sometimes combining hand position relative to the mouth with appearance or skeletal features [14, 17]. However, most skeleton-based methods focus mainly on coordinate-based embeddings, neglecting rotational dynamics. Our approach extends our previous works [13, 14], which incorporate 3D joint embeddings, by introducing, for the first time in ACSR, 3D pose rotation parameterization, yielding a richer skeletal representation that combines rotation parameters, positional keypoints, and facial expression features. For that purpose, we leverage the PIXIE human pose and shape reconstruction model [24] to infer 3D joint-rotation parameters and facial expression features, as well as the MediaPipe holistic model [20] to generate 3D joint coordinates. To model skeletal representations, we

This research work was supported by the project “Applied Research for Autonomous Robotic Systems” (MIS5200632) which is implemented within the framework of the National Recovery and Resilience Plan (NNRP) “Greece 2.0” (Measure: 16618- Basic and Applied Research) and is funded by the European Union-NextGenerationEU.

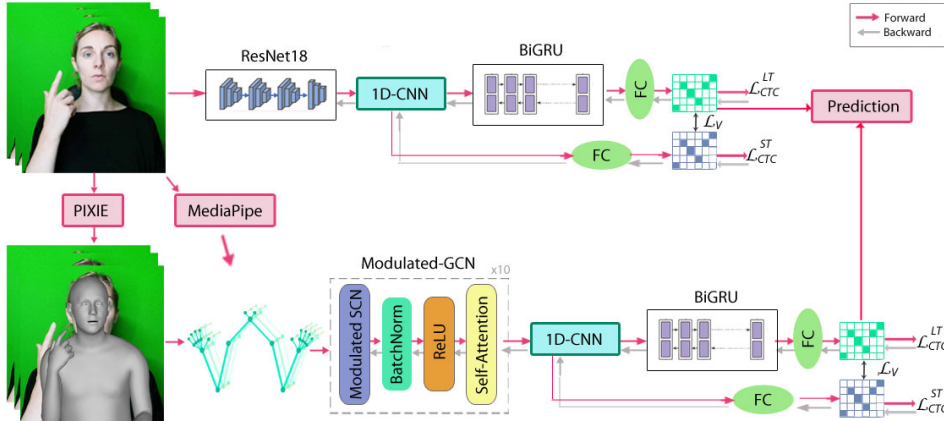


Figure 1: Architecture of the proposed ACSR system. RGB frames and skeletal features are processed in parallel streams, using a ResNet18 and modulated-GCN, respectively, each followed by 1D-CNN and BiGRU modules. Both streams are supervised under a primary CTC loss, an auxiliary short-term CTC loss, and a KL-divergence loss. During inference, their representations are fused, yielding the final phoneme predictions.

adopt a modulated GCN [25], which enhances feature adaptability through dynamic weight modulation. Moreover, we couple the modulated GCN with an attention mechanism that focuses on critical skeletal joint movements, further enhancing feature representations. Assuming that integrating multiple modalities could complement each other, we introduce a multimodal framework with two parallel streams: one processing RGB appearance features and the other handling 3D pose data. Note that RGB appearance features are extracted via the ResNet18 [26].

The final key component in ACSR concerns sequence learning, as well as the alignment of feature representations to phoneme sequences. Early methods rely on GMM-HMMs [4, 12], while recent approaches employ BiLSTMs [22] and BiGRUs [15, 16] in conjunction with CTC decoding. Time-depth separable convolutional encoders [13, 14] and Transformer-based models [17, 23] have also been investigated for cross-modal fusion. However, Transformers require large-scale training data, which are not readily available in CS. We have also explored attention-based encoder-decoder architectures [13], however, these often struggle with alignment issues. To address this, we adopt a temporal convolution layer, which captures short-term motion dynamics between adjacent frames, in conjunction with a BiGRU encoder [27] that effectively learns long-range dependencies. For alignment, instead of using a student-teacher model like in [22], our method combines CTC with two auxiliary loss functions, enhancing the representation capacity of both modalities. This constitutes another innovation of our work. Notably, the two modalities are trained separately, and their scores are properly fused during inference.

To summarize, the main contributions of this work are: (i) the design of a novel multi-stream framework for ACSR that directly processes upper-body videos, without requiring separate hand-lip processing and explicit synchronization; (ii) the first use of 3D pose rotation parameterization in ACSR, coupled with a modulated GCN and an attention mechanism, enhancing skeletal modeling and improving motion dynamics learning; and (iii) the integration of an innovative alignment module that combines CTC with two auxiliary losses, enhancing feature representation capacity and improving multimodal fusion.

We evaluate our approach on three CS datasets: a French [7], a British English [5], and a multi-cuer Mandarin Chinese [6] corpus. Our system outperforms state-of-the-art

methods on all datasets, with a 12.35% absolute error reduction on French, 12.37% on British English, and 11.3% on Mandarin, compared to the next best result, under a multi-cuer (MC) experimental paradigm. Further, we also report recognition results on the Mandarin CS dataset under a cuer-independent (CI) setup for the first time.

## 2. Methodology

To address ACSR, we introduce a multi-stream framework, illustrated in Fig. 1. Our system involves two main components: (i) an appearance modality utilizing a 2D-CNN-based model coupled with a 1D-CNN and a BiGRU encoder for spatio-temporal feature extraction; and (ii) a skeletal modality relying on a GCN module followed by a 1D-CNN and a BiGRU encoder for modeling temporal motion dynamics. Both modalities are trained separately, each using an alignment module that combines CTC with two auxiliary losses, ensuring robust multimodal fusion during inference. Further details follow.

### 2.1. Appearance Representation Modeling

Unlike prior works [12, 13, 16, 22] that segment hand and mouth regions for separate processing, our approach preserves holistic spatial information of visual cues. Specifically, our model focuses on the entire upper-body region, allowing the simultaneous capture of hand and mouth articulation while reducing dependence on precise segmentation of these regions. To dynamically define the upper-body region, we utilize MediaPipe’s holistic human pose detector [20], which estimates 543 whole-body keypoints, including 33 for the torso, 468 for the face, and 21 for each hand. If the MediaPipe detector fails, missing values are filled using the last detected ones. In particular, we crop the upper-body region using the minimum and maximum  $x$  and  $y$  coordinate values of the detected keypoints.

To learn spatial embeddings from the extracted upper-body region, we adopt the ResNet18 [26] model. Pretrained on ImageNet [28], ResNet18 captures high-level feature representations and effectively recognizes intricate patterns in visual data, which are crucial for our task. The upper-body images are rescaled to  $256 \times 256$  pixels before being fed into the network. After global average pooling, each frame is represented by a 512-dimensional (dim) feature vector.

## 2.2. Rotation-Aware 3D Skeletal Modeling

To improve the efficacy of our system, we integrate an additional stream processing the cuer’s skeletal data. In particular, we employ a GCN-based module, comprising a modulated GCN [25] coupled with a self-attention mechanism [29]. As depicted in Fig. 1, the model commences with graph construction, which is based on 3D joint-position and 3D joint-rotation features. In particular, for generating the 3D keypoints of the cuer pose, we employ the MediaPipe model [20] (see also Section 2.1). Regarding 3D joint-rotation parameter regression, we adopt the PIXIE 3D human pose and shape reconstruction model [24]. Specifically, the PIXIE model, employing a moderator to estimate the 3D hand, body, and shape parameters, outputs 55 joints corresponding to the SMPL-X kinematic tree, namely 15 for each hand, 23 for the body, 1 for the head, and 1 for the jaw. Each joint’s rotation is parameterized as a 6D vector encoding 3D orientation, except for the jaw, which uses Euler angles. Additionally, PIXIE extracts 50 facial expression features. Since the hands and mouth are the primary articulation in CS, we perform skeleton graph reduction, selecting 10 nodes per hand and 7 for the upper body, including jaw.

The resulting graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is processed by a GCN module, where  $\mathcal{V}$  represents the set of nodes corresponding to the skeletal joints  $J$ , and  $\mathcal{E}$  corresponds to the structural edges within the skeleton. For each node  $i$ , we concatenate two types of information into a  $D$ -dim feature vector  $\mathbf{q}_i \in \mathbb{R}^D$ , with  $D = 9$ : the 3D joint position, extracted from MediaPipe and the 3D rotation parameters derived from PIXIE. To capture the intricate relational dynamics of the cuer’s skeletal joints, we apply a modulated GCN model that integrates both weight and affinity modulation techniques. For weight modulation, a learnable weight matrix  $\mathbf{L} \in \mathbb{R}^{D' \times J}$  is introduced for each node  $i$ , and the GCN forward propagation is defined as follows:

$$\mathbf{Q}_{out} = \sigma((\mathbf{L} \odot (\mathbf{W}\mathbf{Q}_{in}))\hat{\mathbf{A}}),$$

where  $\odot$  denotes element-wise multiplication,  $\mathbf{Q}_{in} \in \mathbb{R}^{D \times J}$  represents the input features,  $\mathbf{Q}_{out} \in \mathbb{R}^{D' \times J}$  is the updated feature vector,  $\sigma(\cdot)$  is the activation function,  $\mathbf{W} \in \mathbb{R}^{D' \times D}$  indicates the learnable weight matrix, and  $\hat{\mathbf{A}}$  is the normalized adjacency matrix. The adjacency matrix  $\mathbf{A} \in \{0, 1\}^{J \times J}$  specifies the graph’s connectivity, where a value of 1 indicates direct connection, and 0 indicates no link. Affinity modulation is achieved by adding a learnable mask  $\mathbf{B} \in \mathbb{R}^{J \times J}$  to  $\mathbf{A}$ .

To enhance GCN capacity to model dynamic dependencies, we complement it with a self-attention module, which involves spatial, temporal, and channel attention [30]. To prevent overfitting, we also incorporate a DropGraph technique [31]. Our system consists of ten such units, followed by a global average pooling layer in the spatial domain, yielding a 256-dim feature vector. To also incorporate information concerning mouthing patterns, we concatenate the resulting feature vectors with the 50 facial expression parameters. The resulting features are then fed into a linear projection layer, yielding 512-dim features.

## 2.3. Temporal Modeling

To effectively model both short-term and long-range dependencies, the resulting feature maps of both modalities are propagated into a temporal processing pipeline, consisting of a temporal convolution followed by a BiGRU model. In particular, a 1D-CNN is applied to the extracted appearance and skeletal embeddings in order to capture local temporal context. To learn long-term motion dynamics, the extracted representations are

then fed into a 2-layer BiGRU model with hidden state dimensionality of 512. The resulting features are then passed through a dense fully-connected layer with softmax activation, generating the probability distributions over phoneme classes.

## 2.4. Training Strategy

The appearance and skeletal modalities are trained separately, using a CTC loss  $\mathcal{L}_{CTC}^{LT}$  (LT: long-term) that is applied to the generated probability distributions of Section 2.3. This loss maps the sequence representations with the phoneme sequence without requiring frame-level labels. Moreover, we introduce an additional CTC loss  $\mathcal{L}_{CTC}^{ST}$  (ST: short-term) applied to the posterior probability distributions extracted from the 1D-CNN layer. The latter posteriors are generated by passing the short-term temporal features through a fully-connected layer and a softmax activation. Inspired by [32], we also incorporate a KL-divergence loss formulated as:  $\mathcal{L}_V = \text{KL}(\text{softmax}(\mathbf{D}_{ST}), \text{softmax}(\mathbf{D}_{LT}))$ , where  $\mathbf{D}_{ST}$  and  $\mathbf{D}_{LT}$  denote the short-term and long-term posteriors, respectively. This additional loss function mitigates inconsistencies in the learned representations and enhances compatibility between short-term and long-term feature learning. The final training objective is computed as:  $\mathcal{L}_M = \mathcal{L}_{CTC}^{LT} + \mathcal{L}_{CTC}^{ST} + 0.5\mathcal{L}_V$ .

## 2.5. Ensemble Module

During inference, the two modalities are combined through an ensemble module. Specifically, the posterior probabilities generated by the final fully-connected layer of each modality are fused. To achieve effective fusion, each modality is assigned a distinct weight, depending on its performance during validation. The weighted posteriors are then summed to produce the final scores:  $p_{\text{fused}} = 1.0p_{\text{app}} + 0.8p_{\text{skel}}$ .

## 3. Experimental Framework

The proposed model is evaluated on three datasets: French CS [7], British English CS [5], and Mandarin Chinese CS [6]. For the French CS dataset, we use the official split, with 979 training videos, further using 108 videos for validation, and 108 test videos. The dataset provides 720×576-pixel RGB frames at 50 fps and includes 34 phonetic classes, represented via 8 lip patterns, 8 handshapes, and 5 hand positions. The British English CS dataset comprises 98 sentences, and we apply 5-fold cross-validation, splitting the data into 60% for training, 20% for validation, and 20% for testing. The dataset provides 720×1280-pixel RGB frames at 25 fps and involves 44 phonemes, encoded through various lip patterns, hand positions and shapes. The Mandarin Chinese CS dataset consists of 4,000 videos (1,000 sentences) from 4 cuers, recorded at 1280×720-pixel frames and 30 fps, covering 40 phonemes. We evaluate the model under two settings: (i) The MC split, using 4-fold cross-validation with 60%, 20% and 20% of videos being used for training, validation, and testing, and (ii) The CI split, where each fold includes data from three cuers for training and validation (80%-20%), while testing is performed on the fourth cuer videos.

System training involves 50 epochs and a batch size of 2. Optimization is performed using the Adam optimizer [33] with an initial learning rate of 0.0001 reduced by a factor of 0.5 per iteration, and a weight decay of 0.0001. To enhance generalization, data augmentation techniques, such as random cropping and horizontal flipping, are applied. In addition, pose position and rotation parameters are normalized within the range [0, 1]. Experiments are conducted on an Nvidia RTX 3090 GPU.

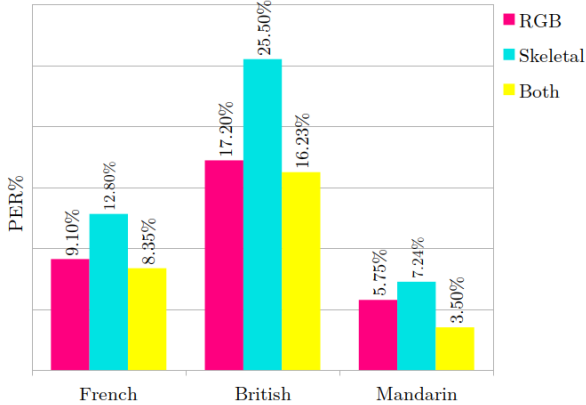


Figure 2: Performance comparison of different modality configurations in terms of PER (%) on the French, British English, and Mandarin Chinese CS datasets (MC setting).

## 4. Experimental Results

In this section, we present the performance evaluation of our proposed approach. Performance is quantitatively assessed on the datasets of Section 3, using phoneme error rate (PER, %) as the evaluation metric.

Initially, we investigate the impact of each modality by comparing the performance of the RGB- and skeletal-based streams alone, as well as their fusion. As illustrated in Fig. 2, the appearance stream consistently outperforms the skeletal stream alone across all datasets. Nevertheless, the combination of both modalities further enhances recognition accuracy, leading to absolute PER reductions of 0.75% on the French dataset, 0.97% on British English, and 2.25% on Mandarin Chinese, compared to the RGB-only model. This suggests that while the RGB modality effectively captures discriminative features, the skeletal modality enriches the representations, improving recognition accuracy.

In addition, in Table 1, we report an ablation study evaluating different aspects of our model, highlighting the benefits of using the BiGRU, as well as incorporating 3D joint-rotation parameters into the skeletal graph and integrating facial expression features. Furthermore, we examine the impact of auxiliary loss functions on system performance. Results show that the BiGRU outperforms both BiLSTM and Transformer-based models, demonstrating its effectiveness in capturing long-term temporal dependencies. Further, integrating joint-rotation features into the skeletal graph together with facial expressions yields better performance than position-only modeling, while

Table 1: Ablation study evaluating the impact of auxiliary loss functions, sequence learning models, and different skeletal representations on ACSR performance. The evaluation is conducted in terms of PER (%) on the French CS dataset. The following notation is used: Temporal convolution (TCN), joint-position (JP), joint-rotation (JR), and facial expressions (F).

Backbone Models		$\mathcal{L}_{CTC}^{LT}$	$\mathcal{L}_{CTC}^{ST}$	$\mathcal{L}_V$	French
RGB	ResNet18+TCN/BiLSTM	✓	✓	✓	10.00
	ResNet18+TCN/Transformer	✓	✓	✓	12.60
	ResNet18+TCN/BiGRU	✓	✓	✓	<b>9.10</b>
Skel.	JP+MGCN+TCN/BiGRU	✓	✓	✓	18.60
	JR/F+MGCN+TCN/BiGRU	✓	✓	✓	16.42
	JP/JR/F+MGCN+TCN/BiGRU	✓	✓	✓	<b>12.80</b>
Both		✓			19.25
	Ours	✓	✓		11.58
		✓	✓	✓	<b>8.35</b>

combining all leads to further reductions in PER. Finally, the incorporation of the two auxiliary loss functions significantly enhances system performance, leading to absolute PER reductions of 10.90%, which demonstrates the effectiveness of our alignment-driven optimization approach.

In Table 2, we compare our model against state-of-the-art approaches on the French and British English CS datasets. Our method significantly outperforms all considered alternatives, achieving the lowest PER of 8.35% on the French and 16.23% on the British CS datasets. Notably, while previous methods rely on hand-mouth segmentation and synchronization, our model learns more robust feature representations directly from the upper-body region, showing the advantage of a holistic approach over region-specific processing. Finally, in Table 3, we evaluate our model on the Chinese CS dataset under both MC and CI setups. Our approach outperforms all systems, achieving 3.50% PER for the MC data split. Moreover, the proposed model yields 28.63% PER under the CI experimental framework, representing the first such result in the literature. Note that the large variation in system performance across the three languages mainly stems from the considerable disparities in dataset sizes, which directly impact the model’s learning and generalization capabilities.

## 5. Conclusions

In this work, we introduce a novel multi-stream framework for ACSR. Our approach integrates appearance spatial features, extracted via ResNet18, with skeletal features modeled through a modulated GCN. The skeleton graph is constructed using 3D joint positions inferred from MediaPipe, as well as joint-rotation and facial expression parameters derived from PIXIE. Short-term temporal dependencies are captured through a 1D-CNN, while a BiGRU encoder is used for long-term sequence learning. In addition, we incorporate a CTC-based visual alignment module with two auxiliary losses. Our findings highlight the effectiveness of processing the entire upper body, allowing the model to learn hand-lip synchronization implicitly, while the integration of 3D pose rotation parameterization enhances skeletal feature expressiveness. Additionally, our alignment mechanism significantly improves feature representation. Our system outperforms state-of-the-art methods across three benchmark datasets.

Table 2: PER (%) comparison of state-of-the-art on the French and the British English CS datasets. The following notation is used: hand (H), mouth (M), and hand position (P).

Model	Feature streams	French	British
Fully Conv [13]	H+M+P	-	36.25
TDS-CTC [14]	H+M+P+Skel.	-	32.58
Student CTC [22]	H+M+P	-	28.6
CB + VLA [23]	H+M	-	33.6
3S-BiGRUs [15]	H+M+Skel.	20.7	-
3S-BiGRUs+LM [16]	H+M+Skel.	25.8	-
<b>Ours</b>	Full Frame	<b>8.35</b>	<b>16.23</b>

Table 3: PER (%) comparison of state-of-the-art on the Chinese CS dataset under MC and CI settings. The following notation is used: hand (H), mouth (M), and hand position (P).

Model	Feature streams	MC	CI
Student CTC [22]	H+M+P	68.2	-
CB + VLA [23]	H+M	24.5	-
FedCSR [34]	H+M+Skel.	14.8	-
<b>Ours</b>	Full Frame	<b>3.50</b>	<b>28.63</b>

## 6. Acknowledgements



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

## 7. 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. Beaudemps, 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. 1144–53, 2005.
- [3] S. J. Gardiner-Walsh, K. Giese, and T. P. Walsh, “Cued speech: Evolving evidence 1968–2018,” *Deafness & Education International*, vol. 23, no. 4, pp. 313–334, 2021.
- [4] L. Liu, T. Hueber, G. Feng, and D. Beaudemps, “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. Beaudemps, and T. Hueber, “CSF22,” 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-cue corpus,” in *Proc. Interspeech*, 2023, pp. 3407–3411.
- [9] P. Heracleous, D. Beaudemps, and N. Aboutabit, “Cued speech automatic recognition in normal-hearing and deaf subjects,” *Speech Communication*, vol. 52, no. 6, pp. 504–512, 2010.
- [10] P. Heracleous, D. Beaudemps, and N. Hagita, “Continuous phoneme recognition in cued speech for French,” in *Proc. Eusipco*, 2012, pp. 2090–2093.
- [11] L. Liu, G. Feng, and D. Beaudemps, “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. Beaudemps, 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] K. Papadimitriou and G. Potamianos, “A fully convolutional sequence learning approach for cued speech recognition from videos,” in *Proc. Eusipco*, 2021, pp. 326–330.
- [14] 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. UAHCI*, 2021, pp. 277–290.
- [15] S. Sankar, D. Beaudemps, 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.
- [16] S. Sankar, D. Beaudemps, 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.
- [17] L. Liu, L. Liu, and H. Li, “Computation and parameter efficient multi-modal fusion transformer for cued speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1559–1572, 2024.
- [18] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] L. Liu and L. Liu, “Cross-modal mutual learning for cued speech recognition,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [24] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, “Collaborative regression of expressive bodies using moderation,” in *Proc. 3DV*, 2021, pp. 792–804.
- [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. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.
- [27] C. Yu, L. Tianrui, J. Zhen, and Y. Chengfeng, “BGRU: A new method of Chinese text sentiment analysis,” *Journal of Physics: Conference Series*, vol. 13, no. 06, pp. 973–981, 2019.
- [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] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, “Skeleton aware multi-modal sign language recognition,” in *Proc. CVPRW*, 2021, pp. 3408–3418.
- [31] 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.
- [32] 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.
- [33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.
- [34] Y. Zhang, L. Liu, and L. Liu, “Cuing without sharing: A federated cued speech recognition framework via mutual knowledge distillation,” in *Proc. ACM MM*, 2023, pp. 8781–8789.