Multimodal Fusion and Sequence Learning for Cued Speech Recognition from Videos

Katerina Papadimitriou[1], Maria Parelli[2], Galini Sapountzaki[3], Georgios Pavlakos[4], Petros Maragos[2], and Gerasimos Potamianos[1]

[1] Department of Electrical & Computer Eng., University of Thessaly, Volos, Greece
[2] School of Electrical & Comp. Eng., National Technical University of Athens, Greece
[3] Department of Special Education, University of Thessaly, Volos, Greece
[4] Electrical Eng. & Computer Sciences, University of California, Berkeley, CA, U.S.A.

aipapadimitriou@uth.gr, maryparelli@gmail.com, gsapountz@sed.uth.gr, pavlakos@berkeley.edu, maragos@cs.ntua.gr, gpotam@ieee.org

Abstract. Cued Speech (CS) constitutes a non-vocal mode of communication that relies on lip movements in conjunction with hand positional and gestural cues, in order to disambiguate phonetic information and make it accessible to the speech and hearing impaired. In this study, we address the automatic recognition of CS from videos, employing deep learning techniques and extending our earlier work on this topic as follows: First, for visual feature extraction, in addition to hand positioning embeddings and convolutional neural network-based appearance features of the mouth region and signing hand, we consider structural information of the hand and mouth articulators. Specifically, we utilize the OpenPose framework to extract 2D lip keypoints and hand skeletal coordinates of the signer, and we also infer 3D hand skeletal coordinates from the latter exploiting own earlier work on 2D-to-3D hand-pose regression. Second, we modify the sequence learning model, by considering a time-depth separable (TDS) convolution block structure that encodes the fused visual features, in conjunction with a decoder that is based on connectionist temporal classification for phonetic sequence prediction. We investigate the contribution of the above to CS recognition, evaluating our model on a French and a British English CS video dataset, and we report significant gains over the state-of-the-art on both sets.

Keywords: Cued speech recognition · convolutional neural networks · time-depth separable convolutional encoder · connectionist temporal classification · OpenPose · skeleton · 2D-to-3D hand-pose regression.

1 Introduction

Speechreading is essential to speech perception for the hearing impaired, albeit inaccurate due to the confusability of visual speech patterns, as multiple phonemes share identical mouthing (visemes). To address this problem, Cornett [6] introduced the cued speech (CS) communication system, complementing

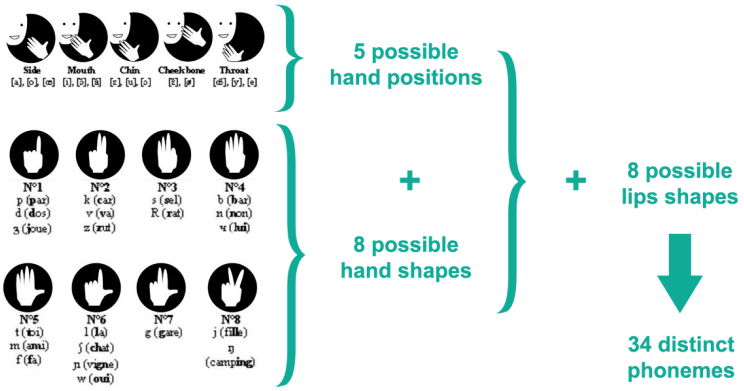

Fig. 1. French CS phonetic encoding system (figure adapted from [1])

mouthing patterns with hand positional and gestural cues. In CS, the simultaneous articulation of mouthing patterns, hand-shapes, and hand positioning relative to the mouth provides a complete visual representation of the spoken language phonological system that is valuable to the speech and hearing impaired. Not surprisingly, CS has been adopted in many languages and dialects. For instance, as also shown in Fig. 1, French CS comprises 5 hand positions that encode vowels, as well as 8 hand-shapes that encode consonants in conjunction with 8 lip contour patterns, yielding 34 phonemes [15]. Similarly, CS for British English encapsulates 4 hand positions for monophthongs (12 monophthongs) and 4 hand slips for diphthongs (8 diphthongs) encoding, as well as 8 hand-shapes for the encoding of 24 consonants in conjunction with lip patterns (44 phonemes in total). Example video frames of CS articulation in French and British English are shown in Fig. 2, obtained from corresponding corpora [21, 22].

Since CS information is primarily delivered by mouthing and gestural patterns, its automatic recognition from video data necessitates the integration of lipreading [28] and sign language recognition techniques [2, 29]. The topic has attracted recent interest in the literature, facilitated by the availability of CS data resources [19, 21, 22]. For example, on the visual front-end side of automatic CS recognition systems, early approaches rely on artificial markings for detecting the articulators of interest [15, 16], while more recent works utilize deep learning for lip tracking and hand region segmentation [21, 25], possibly assisted by a traditional image pre-processing pipeline [25]. This process is typically followed by appearance-based visual feature extraction, most often by means of convolutional neural networks (CNNs) [21, 23, 25]. On the back-end side, most phonetic sequence modeling approaches employ hidden Markov models [1, 15, 16, 23] or more recently a deep learning-based attentional encoder-decoder [25]. In addition to the above, an important CS aspect is the inherent asynchrony between hand-shape and mouthing articulation. Indeed, as shown in [1], the former precedes the latter by roughly one syllable. The issue is also considered in [20, 22, 23].

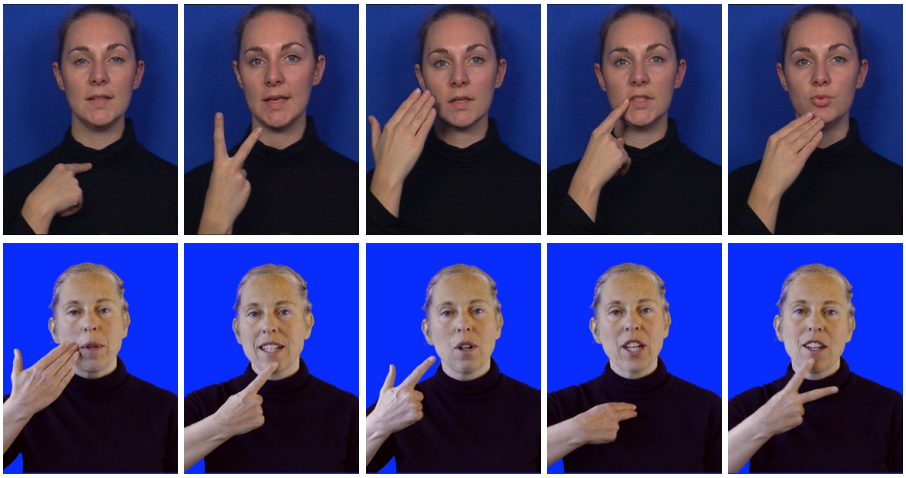

Fig. 2. Example video frames from the French CS dataset [21] (upper row) and the British English CS database [22] (lower row) that are used in this paper, showing various combinations of hand shapes, mouthing patterns, and hand positions.

In this paper, we address the problem of automatic CS recognition from upper-body videos with no artificial markings, by significantly extending our earlier work on this topic [25]. That CS recognition system commenced with a hybrid approach for mouth and hand region tracking (based on a traditional image pre-processing pipeline and 2D-CNNs), it then extracted appearance features of these regions by employing 3D-CNNs, as well as hand positional embeddings relative to the mouth based on 2D-CNNs, and finally concatenated these three visual feature streams and fed them to a deep attentional encoder-decoder [25].

Here, we modify the aforementioned system in multiple ways: We utilize the OpenPose framework [30] for skeletal data acquisition of the CS interpreter, and, by extension, for hand and mouth region segmentation. We then consider additional feature streams that capture structural information of the articulators of interest in order to investigate their benefit to CS recognition. Specifically, we first consider the 2D lip points and 2D hand skeletal coordinates of the signer, provided by OpenPose. Further, we infer 3D hand skeletal coordinates from the 2D ones, by exploiting a powerful architecture [24] that we recently used for 2D-to-3D hand-pose regression in sign language recognition [26], thus enriching knowledge about the trajectory of hand movement by enabling its observation in 3D. Finally, we modify the sequence learning model, by considering its time-depth separable (TDS) convolution block structure [11, 25] used to encode the fused visual features, in conjunction with a decoder that is based on connectionist temporal classification (CTC) [10] for phonetic sequence prediction. Note that our approach does not rely on explicit synchronization of the hand and mouth feature streams prior to their fusion, instead expecting our model to learn such implicitly.

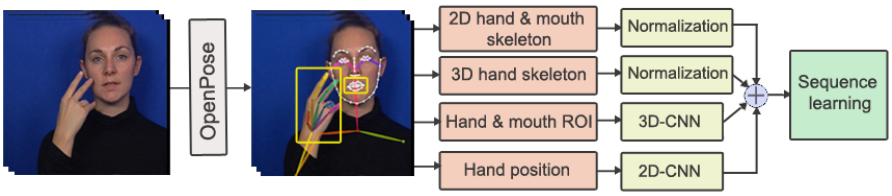

Fig. 3. Architecture of the introduced CS recognition system that generates phonemes from CS videos, following the detection of hand and mouth articulators (left), the extraction and fusion of various feature streams (middle), and sequence learning for phoneme prediction (right).

We evaluate our proposed system on two publicly available CS datasets in French [21] and British English [22], each containing a single subject (see also Fig. 2). We compare our approach against alternative sequence learning models and investigate the combination of various of the aforementioned visual feature streams for CS recognition. Our proposed system turns out superior, significantly exceeding the state-of-the-art on the two datasets that was reported in our earlier work [25]. In particular, we observe a significant absolute phoneme error rate (PER) reduction of 8.87% (from 29.12% to 20.25%) in the French CS corpus and 3.67% (from 36.25% to 32.58%) in the British English CS set.

2 The CS Recognition System

We next present our proposed system for CS recognition from videos. The system is schematically depicted in Fig. 3 and contains multiple components: Visual detection of the articulators, visual feature extraction of multiple streams relating to hand and mouth articulation, their fusion, and, finally, sequence learning for phoneme prediction. All system modules are detailed next.

2.1 Hand and mouth detection

Since CS relies on manual articulation together with mouthing patterns, it is clear that a successful CS recognition system should be able to accurately track both articulators in space and time. For this purpose, we utilize the OpenPose framework [30], which relies on deep convolutional pose models to provide a detailed representation of the human body in the form of multiple 2D keypoints. In particular, OpenPose can estimate up to 137 "human skeleton joints" in the 2D image pixel coordinate system, yielding 70 facial, 25 body-pose, and 42 hand-pose (21 for each hand) keypoints, as also shown in Fig. 4(a),(d).

We further employ the hand and mouth keypoints to generate respective hand and mouth regions-of-interest (ROIs), as also depicted in Fig. 4(c),(f). We then feed these ROIs (after appropriate rescaling) to CNNs for appearance feature generation, as discussed in Section 2.4. Note that occasionally OpenPose

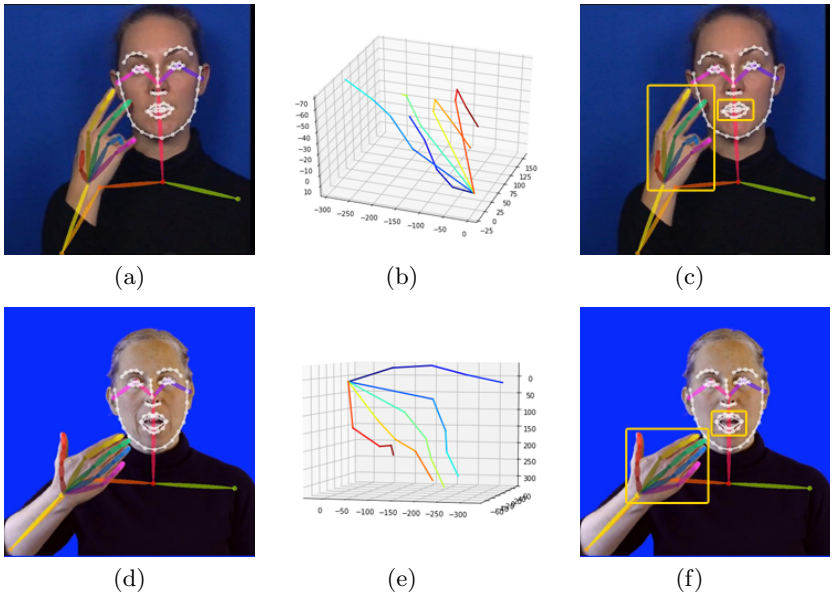

Fig. 4. Examples of articulator detection and keypoint feature extraction on the French CS dataset (upper row) and the British English CS corpus (lower row). Shown, column-wise, left-to-right: (a,d) 2D "skeletal" joints returned by OpenPose; (b,e): inferred 3D keypoints of the signing hand; (c,f): bounding boxes of the signing hand and mouth regions-of-interest derived based on the corresponding OpenPose 2D keypoints.

fails, most likely due to the fact that only part of the signer's body is visible in the datasets considered here (see also Fig. 2). In such cases, we revert to the detection, tracking, and ROI extraction scheme of our earlier work [25].

2.2 2D hand and mouth keypoint features

Our CS system exploits 41 keypoints returned by OpenPose, namely 21 skeleton joints of the signing hand (this happens to be the right hand in the two datasets), and 20 facial keypoints associated with the lip region, all provided as 2D coordinates. This yields 42-dimensional (dim) features for the hand and 40-dim features for the mouth (82-dim in total). These features are normalized before being fed to the fusion module, to counter possible variations in the subject and camera relative positions. Specifically, the 2D points of interest are converted to a local coordinate system with the wrist keypoint and the upper-middle lip keypoint being the respective origins. In addition, all keypoints are further normalized based on the distance between the left and right shoulder joints. Note that in case OpenPose fails to return the desired keypoints, the missing feature streams are filled by the previous existing ones.

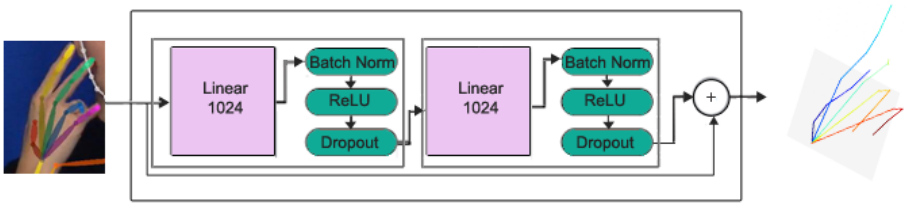

Fig. 5. Model architecture for 3D hand skeleton generation from corresponding 2D information (figure adapted from [26]).

2.3 3D hand keypoint features

In addition to 2D hand keypoint features, we investigate the benefit of exploiting more detailed hand skeletal information by inferring the 3D hand skeleton, thus enriching the system with hand trajectory information in 3D. Our approach extracts the desired 3D hand joints by regressing the 2D hand joint locations to the 3D space [26]. Specifically, after extracting the 2D human skeleton of the hand via OpenPose, we feed its 2D hand coordinates to the hand-pose regression model, producing a series of hand keypoints in the 3D space. We zero-center both 2D and 3D poses around the wrist joint, so as to ensure translation invariance. The regression model, depicted in Fig. 5, is a deep neural network with two layers, each containing two basic blocks that share a residual connection. The network basic building block is a linear layer, followed by batch normalization, a rectified linear unit (ReLU) activation, and dropout. Incorporating batch normalization and dropout increases model robustness to noisy detections, whereas residual connections improve model generalization.

The model yields 21 3D joints for the signing hand, thus producing 63-dim feature vectors (see also Fig. 4(b),(e)). Note that prior to their fusion with other feature streams, these are normalized based on the distance between the hand shoulder and elbow joints, and regarding the wrist as the system origin.

2.4 Hand and mouth appearance features

In addition to the aforementioned features, we extract spatio-temporal appearance features from the ROIs of the signing hand and mouth, as in our earlier work [25]. Specifically, we resize each ROI to 96×96 pixels and apply a 3D-CNN feature learner on three temporally adjacent ROIs of the signing hand or mouth. For this purpose, we utilize the 3D ResNet-34 network [12], which contains 3D convolutions ($3 \times 3 \times 3$) and 3D pooling and is pre-trained on the Kinetics dataset [3], obtaining feature maps from the output of its global average pooling layer. This process yields 512-dim feature vectors for each of the hand and mouth ROIs.

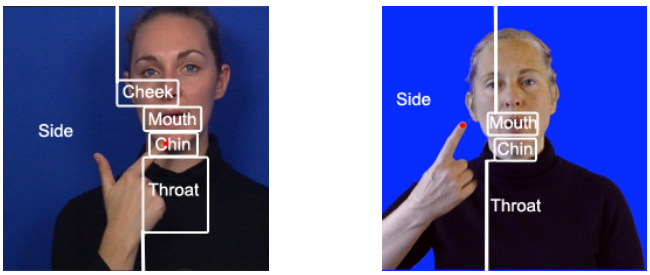

Fig. 6. Example frames of the French CS (left) and British English CS (right) datasets, showing the location-based area division of possible signing hand positions relative to the mouth, which we use to obtain hand positional embeddings in Section 2.5.

2.5 Hand position detection and representation

As discussed in Section 1, hand positioning relative to the mouth plays a crucial role in CS. For this purpose, we use the 2D coordinates of the upper skeletal joint of the signing hand to detect the hand relative position, and then we pass this information through a five-layer 2D-CNN with three fully-connected layers to extract 64-dim hand positional embeddings. Specifically, the CNN is a multi-class model, with each class corresponding to the several possible location areas of the signing hand relative to the mouth. There are five such classes (location areas) for French CS and four for British English CS, as also depicted in Fig. 6.

2.6 Feature fusion

The aforementioned feature streams are fused by simple vector concatenation, producing a 1233-dimensional feature vector for each video frame: 42 for 2D hand skeletal features, 40 for the mouth keypoints, 63 for the 3D hand skeletal stream, 512 for hand appearance, 512 for mouth appearance, and 64 for hand positional embeddings. These fused vectors are then passed to the sequence learning module for predicting the phonetic sequence of the CS video.

2.7 Sequence learning

Viewing phoneme recognition in continuous CS videos as a sequence-to-sequence prediction task, we address it by employing a TDS convolutional encoder [11, 25], followed by CTC decoding [10]. Specifically, the resulting latent-representation vectors generated in Section 2.6 are modeled by a TDS convolutional encoder, which comprises two blocks: a 2D convolution over time, followed by a fully-connected block. In particular, the first sub-block involves a 2D convolutional layer complemented with a ReLU non-linearity and a normalization layer, while the fully-connected layer block consists of two convolutions with ReLU non-linearity in between and a normalization layer. The TDS convolutional encoder output is later subjected to linear projection followed by a log-softmax, yielding a probability distribution over all the possible phoneme labels prior to computing the CTC loss.

3 Experimental Evaluation

3.1 Datasets and experimental framework

As already mentioned, our experiments are conducted on two single-subject, continuous CS corpora, namely the French CS dataset [21] and the British English CS database [22]. All experiments are carried out using ten-fold cross-validation, with 80% of each fold used for training, 10% for validation, and 10% for testing.

In more detail, the French CS dataset contains 238 French sentences, each repeated twice, yielding a 476-sentence set with 11,770 phonemes in total belonging to 34 classes, and it is performed by a professional CS interpreter with no hearing disorders. The collected RGB video data include the upper body of the subject and are available at 50 frames per second (fps) and a 720×576-pixel resolution. On the other hand, the British English CS dataset is significantly smaller, containing only 97 sentences (with 44 phonetic classes) and is recorded by a professional CS speaker with no hearing impairment. The collected RGB video data include the upper body of the subject and are available at 25 fps and a 720×1280-pixel resolution.

3.2 Implementation details

We implement our system in the PyTorch framework [27] and carry out its training using GPU acceleration.

For 3D hand skeleton network (Section 2.3) training, we use the Rendered HandPose Dataset [33], a large-scale 3D hand pose dataset based on synthetic hand models [33]. This dataset utilizes 3D human models with corresponding animations from Mixamo 2 [9], while the software Blender 3 [5] is used for image rendering. It features 20 characters performing 39 actions, and different camera locations are selected randomly for each frame. The dataset provides 41,258 images for training and 2,728 images for evaluation with a resolution of 320×320 pixels. We train the network for 150 epochs using Adam optimizer [18], a batch size of 64, a starting learning rate of 0.001, and exponential decay. The weights of the linear layers are set by Kaiming He initialization [13].

For hand and mouth appearance feature extraction (Section 2.4), we apply a 3D ResNet-34 [12], trained by stochastic gradient descent with momentum at 0.9 with an initial learning rate of 0.1 decreased by a factor of 0.001. We perform 500 complete passes over the data with a mini-batch size of 256 images.

For the sequence learning model of Section 2.7, we employ a TDS convolutional encoder with two 3-channel, three 5-channel, and six 7-channel TDS blocks with kernel sizes 3×1. Additionally, we compare our approach to a number of alternative sequence models, differing in encoder type. Specifically, we evaluate our system using a one-layer long short-term memory (LSTM) [17] encoder and a one-layer gated recurrent unit (GRU) [4] encoder, both with 256 hidden units, as well as a Transformer [32] encoder with hidden dimensionality equal to 512.

We conduct the training of all sequence learning models by the Adam optimizer [18] with a learning rate of 0.003 decayed by a factor of 0.85 and use a

Table 1. Phoneme error rate (%) on the French and British English CS datasets, employing various feature stream combinations in conjunction with the sequence learning model of Section 2.7.

Feature Streams						CS Datasets	
Hand 3D-CNN (512-dim)	Mouth 3D-CNN (512-dim)	Positional 2D-CNN (64-dim)	2D Hand Skeleton (42-dim)	2D Mouth Keypoints (40-dim)	3D Hand Skeleton (63-dim)	French	English
✓	✓					38.50	44.11
✓	✓	✓				25.64	35.13
		✓	✓	✓		38.87	47.10
		✓		✓	✓	37.54	45.89
✓	✓	✓	✓			25.58	33.68
✓	✓	✓		✓		24.93	33.29
✓	✓	✓			✓	25.06	33.50
✓	✓	✓	✓	✓		22.17	32.91
✓	✓	✓	✓	✓	✓	**20.25**	**32.58**

batch size of 128. We employ 0.1 dropout and 0.1 label smoothing [31]. During decoding, we apply the beam search strategy of [8] with beam width equal to 3.

3.3 Results

The performance of our proposed approach for continuous CS recognition from videos is reported in Table 1. There, the phoneme error rate (PER) (%) obtained by the introduced sequence learning model relying on the TDS convolutional encoder and CTC decoding and operating on various feature stream combinations is shown on both CS corpora. It is apparent that the best results are achieved when all feature streams are concatenated, showcasing the benefit of incorporating multiple feature representations into the CS recognition system.

Comparing the best results of the table to our earlier work [25] that represents the state-of-the-art in the field, we obtain significant improvements on both datasets: An 8.87% absolute PER reduction (from 29.12% to 20.25%) for French CS and a 3.67% one (from 36.25% to 32.58%) on British English CS. Such improvements can be attributed to the redesign of both visual feature extraction and sequence learning modules of our system. Indeed, from Table 1 it is obvious that the earlier used appearance and hand positional features alone lag behind the much richer feature representation proposed here. Further, the introduced sequence learning model improves over our earlier model that was based on the TDS convolutional encoder and attentional convolutional decoder, achieving a PER reduction of 3.48% (from 29.12% to 25.64%) for French CS and 1.12% (from 36.25% to 35.13%) for British English CS, revealing the power of CTC decoding in CS recognition (note that these results refer to the combination of appearance and hand positional features, since only these were considered in [25]).

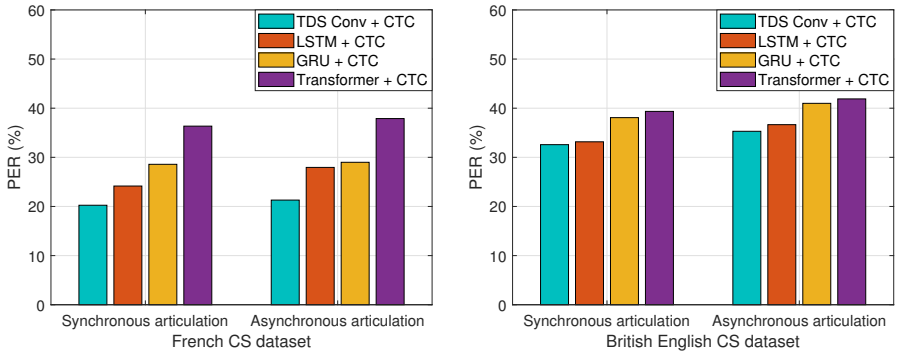

Fig. 7. Comparative evaluation of various sequence learning models on both datasets in terms of PER (%) using all feature streams fused synchronously or with a fixed delay (asynchronously).

Concerning the various feature combinations considered in Table 1, we observe that discarding hand positional embeddings results in the worst PERs on both datasets, demonstrating their importance to CS recognition. Notably, substituting hand and mouth appearance features with the respective skeletal data yields significantly higher PERs on both datasets than their combination. This demonstrates that skeletal data constitute descriptive representations conveying valuable information that can complement the corresponding appearance features, and thus their combined use is essential. Moreover, the 3D hand skeleton seems to be a robust representation, since it performs better than the corresponding 2D hand skeleton when added to the fusion module, and its incorporation on top of all other streams boosts system performance. It can also be seen that the 2D mouth keypoints representation performs well as additional mouth articulatory information, reducing PER on both datasets. This is primarily due to the fact that facial keypoints are more robustly detected by OpenPose than the hand joints that are often occluded. Finally, it can also be observed that there is a significant difference in PERs between the two datasets, most likely due to the limited size of the British English set.

Next, in Fig. 7, we investigate the performance of the various sequence learning models described in Section 3.2 on both CS datasets, when employing all feature streams (1233-dim vectors). Due to the asynchrony between hand and mouth articulation with the former preceding by approximately one syllable [1], we consider two feature fusion schemes: one that concatenates all features disregarding this asynchrony (as we do in our proposed system), referred to as "synchronous articulation" in Fig. 7, and another one, where the hand-related feature streams are artificially delayed by a fixed amount in time in the hope of better matching the mouth-related streams (referred to in the figure as "asynchronous articulation"). Specifically, we use a delay of 12 frames for French CS (as proposed in [1]) and 15 frames for the British English set. As it can be observed from Fig. 7, the proposed sequence learning model (TDS convolutional

encoder and CTC decoding) yields the best results on both sets when features are directly concatenated with no enforced time shift. It can also be seen that the worst results for both datasets are obtained by the Transformer encoder-based model, while the LSTM encoder gives significantly better results compared to the respective GRU model, but still lagging our model. Further, enforcing a time delay of the hand features yields consistently worse results across all models compared to synchronous fusion.

Lastly, we investigated the performance of our model under a number of variations in the appearance feature learner and the number of TDS blocks in the TDS convolutional encoder. Specifically, we replaced the 3D-CNN with a 2D-CNN (ResNet-18 [14]) for appearance feature extraction of the hand and mouth ROIs. That model uses 3×3 convolutional kernels, downsampling with stride 2, and is pretrained on the ImageNet corpus [7]. This modification degraded PER significantly, by over 2% absolute (from 20.25% to 22.74% PER) for French CS and by about 3.5% (from 32.58% to 36.12% PER) for British English CS. Regarding the TDS convolutional encoder, we increased the number of channels keeping the same receptive field from (3, 5, 7) to (10, 12, 14), (10, 14, 18), and (10, 10, 14), but in all cases we ended up with worse PERs on both corpora.

4 Conclusions

In this paper, we investigated the incorporation of multiple representation streams into a state-of-the-art deep-learning based sequence learning model for CS recognition from upper-body videos. In particular, our CS recognition system relied on spatio-temporal feature extraction and fusion learned via a TDS convolutional encoder, followed by CTC decoding without the use of any explicit stream synchronization. We highlighted how the inclusion of skeletal data to the feature fusion module benefits system performance. Notably, inferred 3D hand skeletal data boosted CS recognition when added on top of all other spatio-temporal streams. The conducted evaluation on two CS datasets demonstrated that the proposed model outperformed other sequence learning architectures, surpassing the state-of-the-art in the field.

Acknowledgments

This research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members & Researchers and the procurement of high-cost research equipment grant" (Project "SL-ReDu", Project Number 2456).

References

1. Attina, V., Beautemps, D., Cathiard, M., Odisio, M.: A pilot study of temporal organization in cued speech production of French syllables: rules for a cued speech synthesizer. Speech Communication **44**(1), 197–214 (2004)

2. Bheda, V., Radpour, D.: Using deep convolutional networks for gesture recognition in American Sign Language. CoRR **abs/1710.06836** (2017)
3. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? A new model and the Kinetics dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (2017)
4. Cho, K., Merrienboer, B.V., Gülçehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (2014)
5. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org
6. Cornett, R.O.: Cued speech. American Annals of the Deaf **112**(1), 3–13 (1967)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
8. Freitag, M., Al-Onaizan, Y.: Beam search strategies for neural machine translation. CoRR **abs/1702.01806** (2017)
9. Fuse, M.: Mixamo: Quality 3D Character Animation In Minutes (2015), [online]: https://www.mixamo.com
10. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning (ICML) (2006)
11. Hannun, A., Lee, A., Xu, Q., Collobert, R.: Sequence-to-sequence speech recognition with time-depth separable convolutions. CoRR **abs/1904.02619** (2019)
12. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3D residual networks for action recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3154–3160 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1026–1034 (2015)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Heracleous, P., Beautemps, D., Aboutabit, N.: Cued speech automatic recognition in normal-hearing and deaf subjects. Speech Communication **52**(6), 504–512 (2010)
16. Heracleous, P., Beautemps, D., Hagita, N.: Continuous phoneme recognition in cued speech for French. In: Proceedings of the European Signal Processing Conference (EUSIPCO). pp. 2090–2093 (2012)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
19. Liu, L., Feng, G.: A pilot study on Mandarin Chinese cued speech. American Annals of the Deaf **164**(4), 496–518 (2019)
20. Liu, L., Feng, G., Beautemps, D.: Automatic temporal segmentation of hand movements for hand positions recognition in French cued speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3061–3065 (2018)

21. Liu, L., Hueber, T., Feng, G., Beautemps, D.: Visual recognition of continuous cued speech using a tandem CNN-HMM approach. In: Proceedings of Interspeech. pp. 2643–2647 (2018)
22. Liu, L., Li, J., Feng, G., Zhang, X.: Automatic detection of the temporal segmentation of hand movements in British English cued speech. In: Proceedings of Interspeech. pp. 2285–2289 (2019)
23. Liu, L., Feng, G., Beautemps, D., Zhang, X.P.: Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. IEEE Transactions on Multimedia **23**, 292–305 (2021)
24. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2659–2668 (2017)
25. Papadimitriou, K., Potamianos, G.: A fully convolutional sequence learning approach for cued speech recognition from videos. In: Proceedings of the European Signal Processing Conference (EUSIPCO). pp. 326–330 (2021)
26. Parelli, M., Papadimitriou, K., Potamianos, G., Pavlakos, G., Maragos, P.: Exploiting 3D hand pose estimation in deep learning-based sign language recognition from rgb videos. In: Proceedings of the ECCV 2020 Workshops. pp. 249–263 (2020)
27. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Proceedings of the NIPS-W (2017)
28. Potamianos, G., Marcheret, E., Mroueh, Y., Goel, V., Koumbaroulis, A., Vartholomaios, A., Thermos, S.: Audio and visual modality combination in speech processing applications. In: Oviatt, S., Schuller, B., Cohen, P., Sonntag, D., Potamianos, G., Krüger, A. (eds.) The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations, pp. 489–543. Morgan-Claypool (2017)
29. Rao, G.A., Syamala, K., Kishore, P.V.V., Sastry, A.S.C.S.: Deep convolutional neural networks for sign language recognition. In: Proceedings of the Signal Processing and Communication Engineering Systems (SPACES). pp. 194–197 (2018)
30. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4645–4653 (2017)
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017)
33. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4913–4921 (2017)