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], Gerasimos Potamianos [1]

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

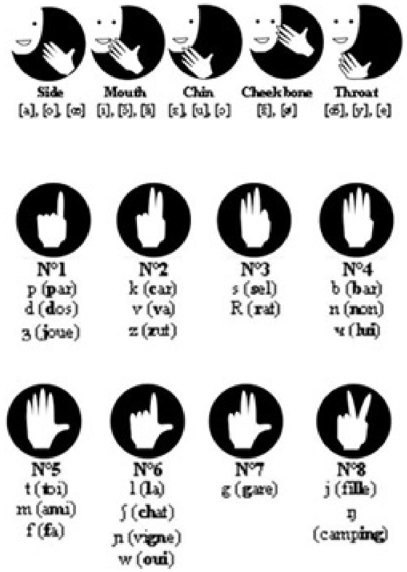
HCI INTERNATIONAL 2021 - International Conference on Human-Computer Interaction

HCI International 2021
24-29 July • Washington DC, USA

H.F.R.I.
Hellenic Foundation for Research & Innovation

Overview

- **Goal:**
  - ✓ Address **automatic cued speech recognition** (**CSR**) from **videos** with **no artificial markings**.
- **Challenges:**
  - ✓ **Phonetic information** from **simultaneous articulation** of **mouthing** patterns**, hand** positioning, and **gestures**.
  - ✓ **Asynchrony** between hand and lip **articulation**.
- **Our earlier approach [1]:**
  - ✓ **Tracking: hand** and **mouth** via a **hybrid method**.
  - ✓ **Features**: 3D-CNN **appearance** based and **positional** embeddings**.**
  - ✓ **Recognizer**: **time-depth separable** (**TDS**) **convolutional encoder** and **attentional convolutional decoder** [3].
- **Here:**
  - ✓ **Tracking:** via **OpenPose** framework [4]**.**
  - ✓ **Features:** investigate additional **benefit** of **2D** and **3D** (**regressed**) **skeletal keypoints.**
  - ✓ **Recognizer:** use **connectionist temporal classification** (**CTC**) [5] for **decoding.**



French CS articulation (figure modified from [2]).

5 possible hand positions

+

8 possible hand shapes

+

8 possible lip shapes *(not shown)*

34 distinct phonemes

[1] **Papadimitriou & Potamianos,** "A fully convolutional sequence learning approach for cued speech recognition from videos," ***EUSIPCO*** '**20**.

[2] **Attina *et al.***, "A pilot study of temporal organization in cued speech production of French syllables: rules for a cued speech synthesizer," ***Speech Comm.*** '**04**.

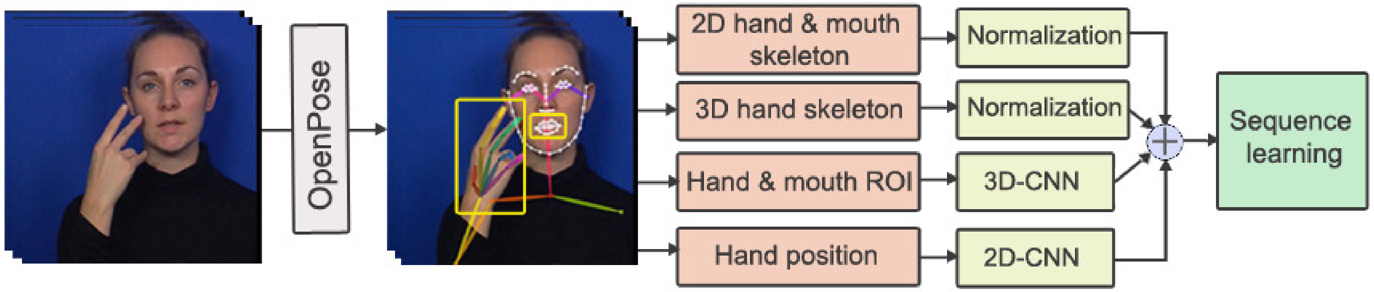
[3] **Hannun *et al.***, "Sequence-to-sequence speech recognition with time-depth separable convolutions", ***Interspeech*** '**19**.

[4] **Cao *et al.***, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," ***IEEE TPAMI*** '**21**.

[5] **Graves *et al.*,** "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," ***ICML*** '**06**.

Our Approach

- **Proposed deep learning-based approach architecture:**



- **Our main contributions:**
  - ✓ **2D skeletal data acquisition** of the CS interpreter via **OpenPose** framework.
  - ✓ **Hand** and **mouth** region **segmentation** through the 2D skeletal coordinates.
  - ✓ **3D hand skeletal** coordinates extraction by a **2D-to-3D hand-pose regression** architecture.
  - ✓ **Fusion** of various **feature** streams / representations of **manual** and **non-manual** articulators.
  - ✓ **Time-depth separable** (**TDS**) **convolution** block structure based encoder
  - ✓ **Connectionist temporal classification** (**CTC**) decoder.

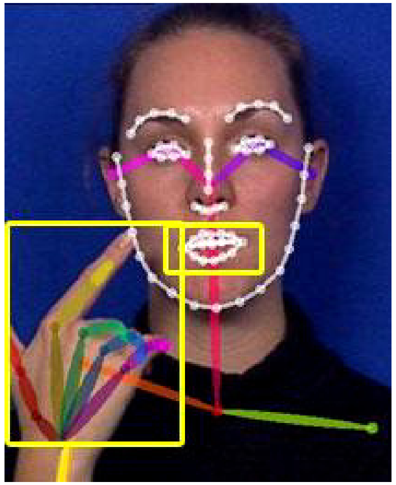
- **Results:**
  - ✓ Experiments on **2** publicly available **CS datasets.**
  - ✓ **Inclusion** of **skeletal data** to the feature fusion module **benefits** system **performance**.
  - ✓ **Better** than current **state-of-the-art CSR** methods.

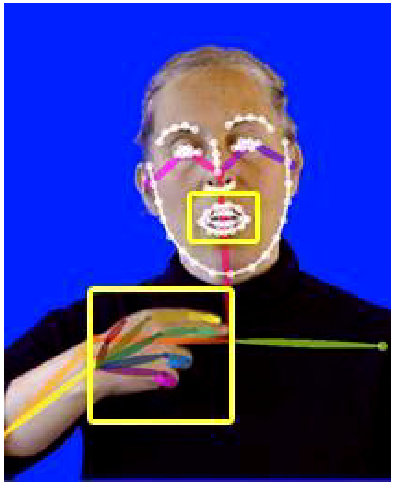
CSR System – Visual Front End (I)

- **Hand and mouth detection:**
  - ✓ Based on **OpenPose** detector of human joints.
  - ✓ Returns **25 body-pose** keypoints, **21** joints for **each hand**, and **70 face** keypoints.

- **2D hand and mouth keypoint features:**
  - ✓ Retain **21 joints** of the signing **hand** and **20 mouth keypoints.**
  - ✓ Apply **normalization** to their coordinates.
  - ✓ Obtain **82-dim features** (42-dim for the hand and 20-dim for the mouth).

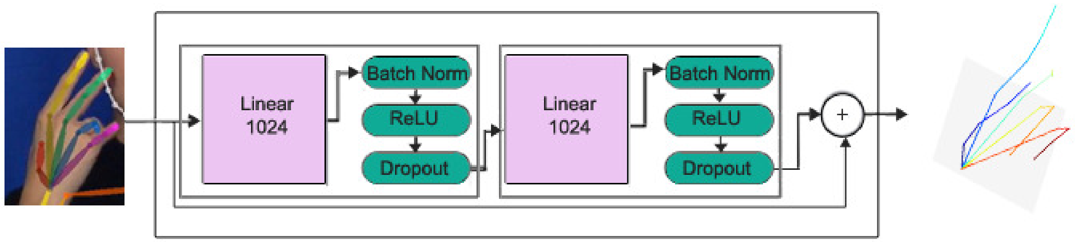
- **Hand and mouth appearance features:**
  - ✓ Extract region-of-interests (**ROIs**) based on the OpenPose skeleton.
  - ✓ Feed ROIs to **3D ResNet-34** network [6].
  - ✓ Obtain **512-dim** spatio-temporal **appearance features** for each **ROI** (512-dim for **hand** ROI and 512-dim for **mouth** ROI).



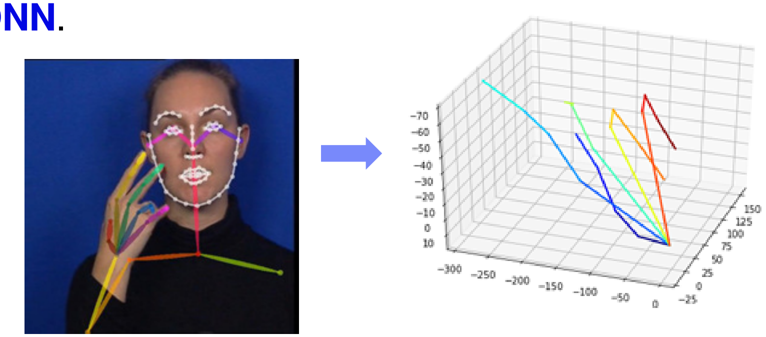


[6] **Hara *et al.***, "Learning spatio-temporal features with 3D residual networks for action recognition," ***CVPR*** **'17**.

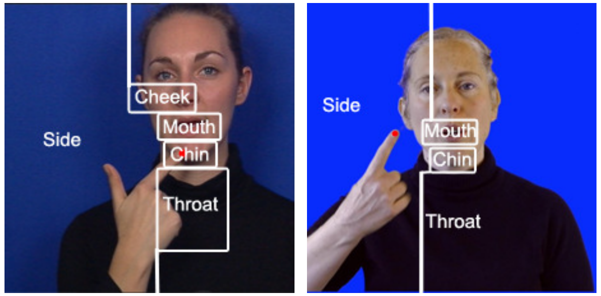
CSR System – Visual Front End (II)



- **3D hand keypoint features:**
  - ✓ **Regress 2D** hand joints to the **3D** space, using a two-layer **DNN**.
  - ✓ DNN **input**: 21 hand joint coordinates (**2D**) from OpenPose.
  - ✓ DNN **output**: 21 hand joint coordinates in **3D** space.
  - ✓ Apply **normalization** to the 3D joints.
  - ✓ Obtain **63-dim** features (21 x 3).



- **Hand positioning features:**
  - ✓ Employ 2D coordinates of the **upper-most hand skeletal joint**.
  - ✓ **2D-CNN** based **classification** of hand positioning relative to mouth.
  - ✓ Obtain 64-dim **hand positional embeddings**.
  - ✓ **5 positions** for **French** CS and **4 positions** for **British English** CS.



[7] **Parelli *et al.***, "Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos," ***ECCV-W*** **'20**.

CSR System – Feature Fusion and Sequence Model

- **Feature fusion (vector concatenation) yields 1233-dim feature vector:**

✓ **42**-dim for **2D hand keypoints**.

✓ **40**-dim for **2D mouth keypoints**.

✓ **63**-dim for **3D hand keypoints**.

✓ **512**-dim for **hand ROI** appearance (3D CNN).

✓ **512**-dim for **mouth ROI** appearance (3D CNN).

✓ **64**-dim for **hand positional embeddings**.

- **Sequence learning:**

✓ **Time-depth separable convolutional encoder** (**TDS**).

✓ **CTC loss** based **decoding**.

Datasets and Experimental Setup

- **French CS dataset [8]:**
  - ✓ 2 repetitions of 238 **French sentences** performed by a **professional CS interpreter**.
  - ✓ **11,770 phonemes** in total belonging to 34 classes.
  - ✓ **Upper-body RGB** video data available at 50 fps and 720x576-pixel resolution.
  - ✓ **8 lip** patterns, **8 handshapes**, and **5** different hand **positions** (**34 phonetic classes**).

- **British English CS dataset [9]:**
  - ✓ 97 **British English sentences** recorded by a **professional CS interpreter**.
  - ✓ **Upper-body color** video images available at 25 fps and 720x1280-pixel resolution.
  - ✓ 4 **hand positions** for the **12 monophthongs**, 4 **hand slips** for the **8 diphthongs,** and 8 **hand shapes** for the **24 consonants** (**44 phonetic classes**).

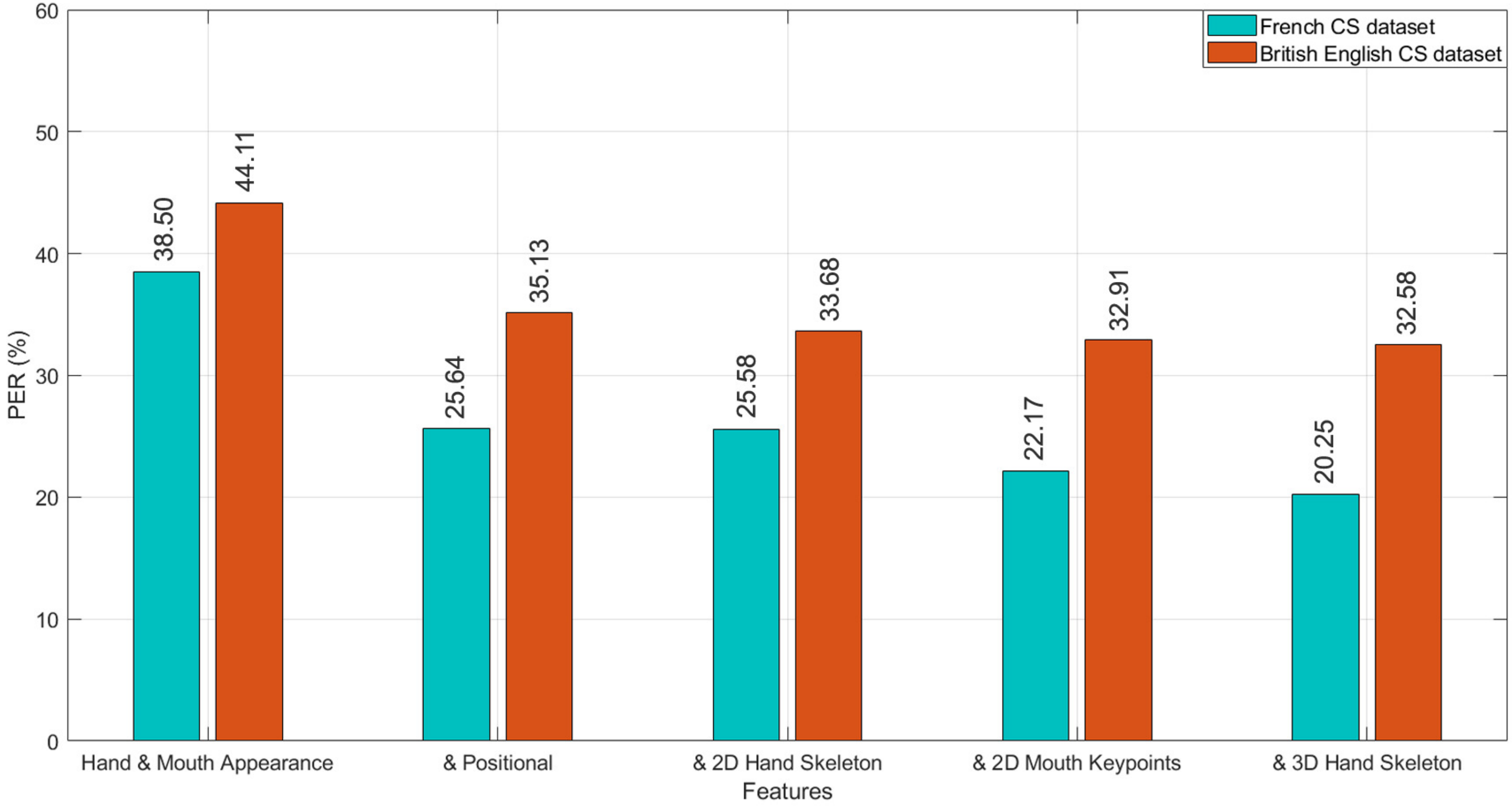
- **Experimental framework:**
  - ✓ **Ten-fold cross-validation**.
  - ✓ **80%** of each fold used for **training**, **10%** for **validation**, and **10%** for **testing**.
  - ✓ **Phonetic error rate** (**PER**, **%**) reported.

[8] **Liu *et al*.**, "Visual recognition of continuous cued speech using a tandem CNN-HMM approach," ***Interspeech* '18**.

[9] **Liu *et al*.**, "Automatic detection of the temporal segmentation of hand movements in British English cued speech," ***Interspeech* '19**.

Experimental Results (I)

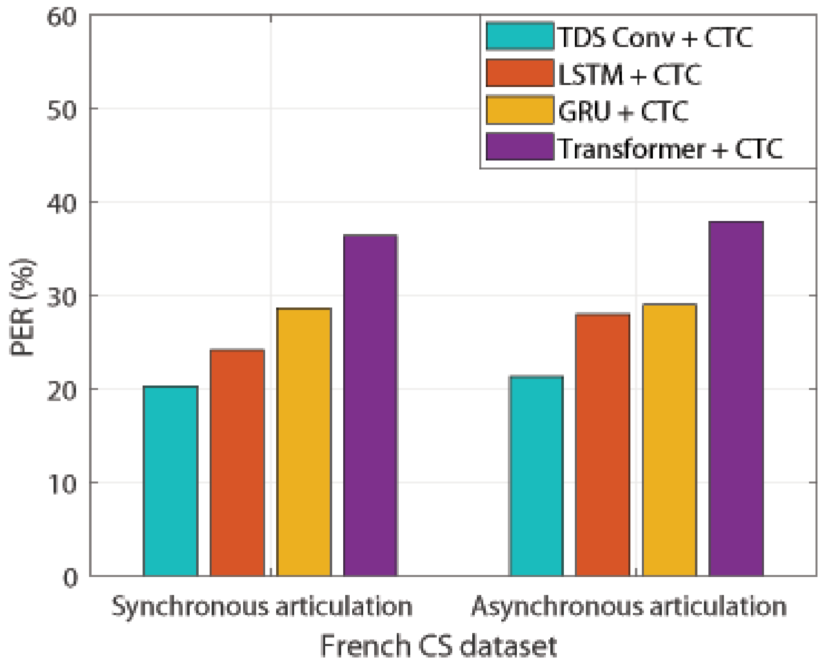
- Evaluation of various **feature stream combinations**.
- **Fusion** of **all feature streams** yields the best results on both CS corpora.

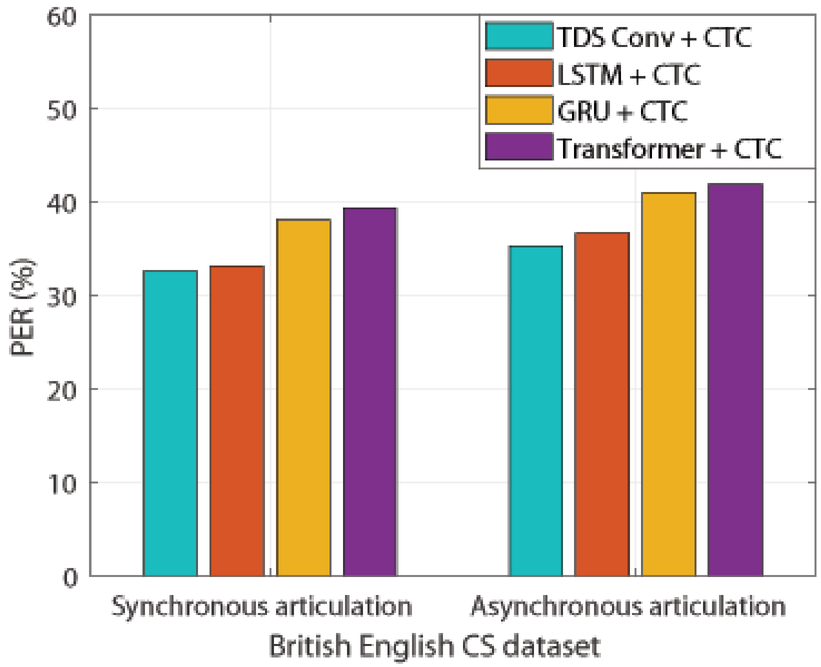


- **Significant improvements** on both datasets compared to our **earlier model** (state-of-the-art):
  - ✓ **8.87%** absolute **PER reduction** (from 29.12% to 20.25%) for **French** CS.
  - ✓ **3.67%** absolute **PER reduction** (from 36.25% to 32.58%) on **British English** CS.

Experimental Results (II)

- **Proposed** model **comparison against** various **sequence learning models** on **both CS sets:**
  - ✓ A one-layer **long short-term memory (LSTM) encoder** coupled with **CTC decoding**.
  - ✓ A one-layer **gated recurrent unit (GRU) encoder** and **CTC decoding**.
  - ✓ **A Transformer encoder** complemented with a **CTC decoder**.
- **Two** feature **fusion schemes** employing all feature streams:
  - ✓ **Synchronous articulation**: All features concatenation discarding asynchrony.
  - ✓ **Asynchronous articulation**: Hand-related feature streams artificially delayed by a fixed amount in time.
- The **proposed model** yields the **best results** on both sets when there is **no enforced time shift**.





Experimental Results (III)

- **Performance evaluation** of the proposed model **under** a number of **variations**:

✓ Replace the 3D-CNN with a **2D-CNN** (ResNet-18 [10]) for **appearance feature extraction**:

- **Degraded PER** by over **2% absolute** for the **French CS** dataset.
- **Degraded PER** by about **3.5% absolute** for the **British English CS** dataset.

✓ The **number of TDS blocks** in the TDS convolutional encoder:

- **Increase the number of channels** keeping the **same receptive field**.
- **Worse PERs** on both corpora.

[10] **He *et al*.**, "Deep residual learning for image recognition," ***CVPR*** **'16.**

Conclusions

- **Proposed** a **deep learning** model for effective **CS recognition from upper-body videos:**
  - ✓ **Spatio-temporal feature extraction** and **fusion.**
  - ✓ **State-of-the-art deep-learning based** sequence **learning model**.
- **Highlighted** how the **incorporation** of **multiple** representation **streams, TDS convolutional encoder** and **CTC decoding improves feature learning** performance.
- **Inclusion** of **skeletal data** to the feature fusion module **benefits system performance**.
- **Inferred 3D hand skeletal** data **boosted CS recognition** when added on top of all other spatio-temporal streams.
- **Demonstrated** that the proposed model **outperforms** other **sequence learning architectures**.

THANK YOU!

Questions? Pls. contact:

aipapadimitriou@uth.gr gpotam@ieee.org

Acknowledgments

The **research work** was supported by the **Hellenic Foundation for Research and Innovation (H.F.R.I.)** under the **"First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant"** (Project Number: 2456).