MULTISTREAM NEURAL ARCHITECTURES FOR CUED SPEECH RECOGNITION USING A PRE-TRAINED VISUAL FEATURE EXTRACTOR AND CONSTRAINED CTC DECODING

Sanjana Sankar, Denis Beautemps, and Thomas Hueber

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

ABSTRACT

This paper proposes a simple and effective approach for automatic recognition of Cued Speech (CS), a visual communication tool that helps people with hearing impairment to understand spoken language with the help of hand gestures that can uniquely identify the uttered phonemes in complement to lip-reading. The proposed approach is based on a pre-trained hand and lips tracker used for visual feature extraction and a phonetic decoder based on a multistream recurrent neural network trained with connectionist temporal classification loss and combined with a pronunciation lexicon. The proposed system is evaluated on an updated version of the French CS dataset CSF18 for which the phonetic transcription has been manually checked and corrected. With a decoding accuracy at the phonetic level of 70.88%, the proposed system outperforms our previous CNN-HMM decoder and competes with more complex baselines.

***Index Terms*—** Visual speech, cued speech, hearing impairment, multi-modality, neural network

1. INTRODUCTION

Cued Speech (CS) is a visual communication tool developed by Cornett [1] in 1967 to help people with hearing impairment to better understand the spoken language. It encodes speech as a combination of visible hand shapes (for consonants) and hand positions (for vowels) to highlight the uttered phoneme and complement lip-reading [2]. French CS or *Langue française Parlée Complétée (LPC)* [3] uses five hand positions to encode vowels and eight hand shapes for consonants as shown in Fig 1.

Automatic Cued Speech Recognition (ACSR) is based on transcripting visual cues of speech to text. The first ACSR systems were focused on isolated vowel and/or consonant recognition [4]. The cuers were artificially marked with hand landmarks and lip makeup before the video recording to simplify the image processing [4]. Currently, the research is focused on continuous ACSR, i.e., the decoding of connected

This work, as part of the Comm4CHILD project, has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No 860755. Authors would like to thank Sylla Camara for fruitful discussions.

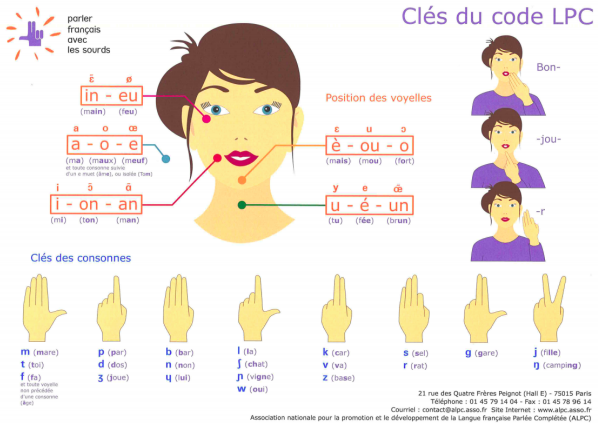


Fig. 1. Overview of the *Langue française Parlée Complétée* (Cued-speech system for French)

words [5]. In [6], we proposed the first system for continuous phoneme decoding from CS visual data. A feature extractor based on a multi-stream convolutional neural network (CNN) processing the raw regions of interest of hand and lips was combined with an HMM-GMM phonetic decoder. One of the major challenges in continuous ACSR is to deal with the asynchrony between hand and lip [7], i.e., the configuration of the hand pertaining to a certain phoneme can precede (or follow) that of the lips for the same phoneme by a variable delay ranging from a few milliseconds to several hundred milliseconds [8]. In [6], this issue has been addressed with a simple heuristic by considering the hand configuration observed at the beginning of the previous phoneme. However, this requires a temporal segmentation of the visual streams at the phonetic level. One of the goal of the present study is to get rid of this time-consuming task. To that purpose, we investigated the use of recurrent neural networks (Bi-directional Gated Recurrent Units (Bi-GRUs) [9]) trained with a Connectionist Temporal Classification (CTC) loss [10]. In particular, we compare different architectures for combining lips and hand information within the network.

The second major challenge in continuous ACSR is the lack of datasets. To the best of our knowledge, the largest available corpus of CS data is the CSF18 corpus released with our previous study [6]. Since this corpus remains relatively

small (i.e. 476 sentences), we hypothesize that training the feature extractor from scratch (i.e. a CNN in our previous study) may not be an efficient strategy. Therefore, we investigate in the present study the use of a pre-trained feature extractor. We use the Mediapipe [11] toolkit to infer automatically a set of landmarks of the hand and lips from a raw image sequence of a CS cuer.

Since our previous study [6], other approaches have been proposed for decoding continuous CS data at the phonetic level. In [12], Papatimitriou et al. proposed a fully convolutional model with a time-depth separable block and attention based decoder. In [13], Wang et al. make use of pre-trained teacher model by exploring different methods for cross-modal knowledge distillation. Both approaches are evaluated on the CSF18 corpus and outperformed the CNN-HMM baseline by a significant margin. However, we identified a potential bias in some of those studies due to the particular structure of the linguistic material of the CSF18 corpus. One of the goal of this paper is thus to quantify and report the impact of this potential bias on the overall performance. Moreover, we also show that a performance gain could be obtained by correcting the phonetic transcription of the CSF18 corpus so that it reflects what the CS cuer has actually coded.

Finally, all previous studies in ACSR focus on phonetic decoding. However, a practical system should be able to deliver to the user the most likely sequence of words. As a first step toward such a system, we also investigated the use of the Token Passing algorithm combined with a pronunciation lexicon. The key contributions of this work are (i) a light architecture for continuous decoding of CS based on a pre-trained feature extractor and a multistream RNN which does not need a temporal segmentation of the visual stream at the phonetic level and can compete with the more complex approaches recently reported in the literature, (ii) the release of a corrected version of the CSF18 which now reflects more accurately the CS encoded keys and can lead to a significant performance gain, and (iii) a first attempt toward automatic decoding of CS at the word level.

2. METHODOLOGY

Visual features are extracted automatically with the use of Mediapipe toolkit. The key regions of interest for feature extraction in an ACSR system are the hands and the lips. MediaPipe Hands and MediaPipe Face Mesh provide pre-trained solutions that can estimate the 3D geometry of the visuals. These solutions were adapted for the purpose of this experiment. MediaPipe Hands is a high-fidelity hand and finger tracking solution that provides 21 3D landmarks of a hand. MediaPipe Face Mesh is a face geometry solution that estimates 468 3D face landmarks. For feature extraction, the raw input image sequences for 476 French sentences of the CSF18 corpus were fed to the system. For the purpose of CS, only 42 2D landmarks that attribute to the lips and 21 2D landmarks on the hands were considered for this study. The 2D landmarks of hands and lips pertain to the x and y coordinates of their position in the image. A Principal Component Analysis (PCA) [14] of the x-y coordinates was done and normalized for each of the two streams dedicated to hand and lips. The first 20 principal components that can summarize up to 99% of the information contained in the features are selected for each stream in order to reduce the data taxation on the model. The finger tracking for the index or the middle finger was also done separately to retain the information for hand position which encodes the vowel. Since in French CS, either the index finger or the middle finger is always seen for all the target gestures, the extremity of these fingers are tracked and fed to the 3rd input stream in Fig 2 (b). This is done in order to accentuate the information for vowel decoding.

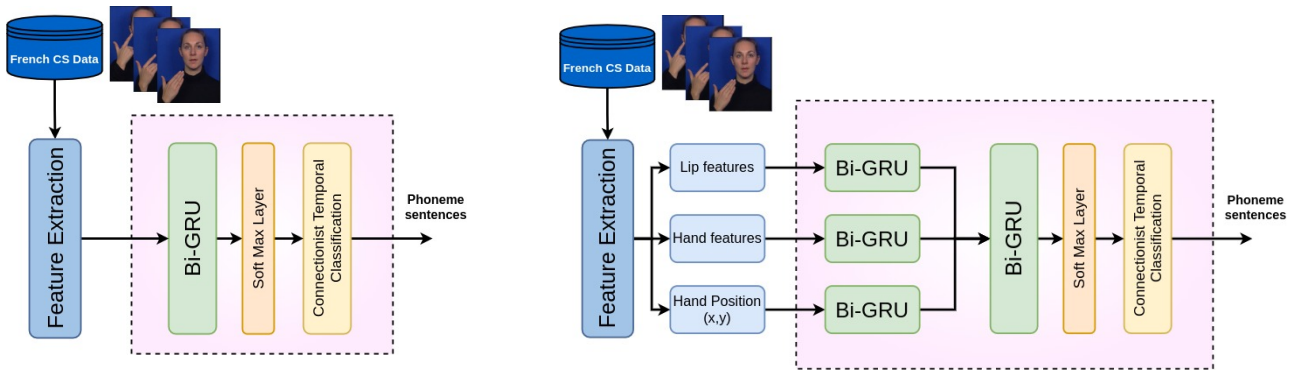
2.1. Model Architecture

In this study, we use Bi-GRUs to capture the intrinsic dynamics of hand and lips in CS. Three architectures are investigated: **(A)** Early Fusion ACSR, Fig.2(a), **(B)** 2-Streams (only the Lips and Hand streams in Fig.2(b)), and **(C)** 3-Streams ACSR, Fig.2(b). In **(B)** and **(C)**, we employ middle fusion. The extracted features for each stream are fed to a Bi-GRU that caters specifically to each stream. The output of the stream-wise GRUs are then concatenated and fed to a second layer of Bi-GRU as shown in Fig. 2(b). The purpose of including the second Bi-GRU layer after concatenation in **(B)** and **(C)** is to intrinsically learn the dynamics between each stream. Bi-GRUs can capture both past and future information. Therefore, we employ Bi-GRU to learn long-term dependencies between various time steps of the different streams. This step is crucial to automatically learn the time lag between the hand gesture and lip movement in CS.

A fully connected softmax layer then provides the posterior distribution for each phonetic class. This network is trained with a CTC loss which decodes directly the most possible sequence of phonemes without the need of temporal segmentation of the visual input data.

2.2. Decoding Strategies

Two decoding strategies were investigated. In the first one, referred to as "unconstrained", a greedy CTC decoder is used to recover the most likely phonetic sequence. In short, the decoder first concatenates the most probable phonemes at each timestep and then removes the duplicate phonemes and the CTC blank tokens. The second strategy, hereafter the "constrained" one, exploits a pronunciation lexicon and is based on the Token Passing algorithm [15], adapted for CTC [16]. The use of a lexicon allows us (i) to introduce prior linguistic knowledge in the decoding and thus potentially resolve ambiguities in the CS data, (ii) to recover a sequence of words which is a crucial step toward a practical system of AVCSR.



(a) Early Fusion ACSR

(b) 3-stream ACSR system

Fig. 2. Proposed architectures for automatic Cues Speech Recognition

3. EXPERIMENT

3.1. Dataset

The CSF18 [6, 17] is a dataset consisting of 238 French sentences that has been uttered twice. The French CS cuer was recorded at 50 fps and the image is of 720 X 576 pixels resolution. We conducted experiments on the publicly available CSF18 dataset. This corpus constitutes of 34 phonetic classes - 14 vowels and 20 consonants that can be encoded with 8 hand shapes and 5 hand positions by the cuer. We then updated this corpus by manually correcting the phonetic transcription so that it matches the CS keys encoded by the cuer, wherever it differed from the phonemes transcribed based on speech. In this process, 3 new phonetic classes were introduced in the corpus- the clusters *"gn"*, *"ng"* and the semi-vowel *"ui"* as shown in Fig.1. This helps in the correct labeling of the gestures and hand positions used by the cuer. The updated corpus is henceforth referred to as CSF18V2 and is made available with the DOI: 10.5281/zenodo.1206000).

3.2. Model Training

The first ACSR model shown in Fig.2(a) consists of a single layer of Bi-GRU with 128 fully connected hidden units. The 2-Streams and 3-Streams ACSR models consist of 1-layer Bi-GRU for each stream and the second layer of Bi-GRU after concatenation as shown in Fig.2(b) constitutes of 256 hidden units. The model is trained with Adam Optimizer [18] with an initial learning rate of 0.001 and decreased by a factor of 2 when necessary. The training patience is set to 10 and the CTC Loss is used to calculate the loss between all the possible phonetic transcriptions at train time. The mini-batch size is fixed to 16. We have setup 3 different experiments and all the models are trained using the same optimization criteria and loss function. At test time, and for the constrained decoding strategy only, a pronunciation lexicon was built by keeping the 1105 words of the CSF18 corpus.

3.3. Train-Test split

As mentioned in 1, one of the goals of this paper is to quantify a potential bias in the performance due to the linguistic content of the CSF18 corpus. Since this corpus consists of repeated sentences, a conventional shuffling of the data is very likely to create a large linguistic overlap between training and test set. In another words, a significant number of test and train sentences will share the same text. Neglecting this potential risk can bring a significant difference in the accuracy rate, thus biasing the evaluation of the method. In this study, we carefully control this aspect. First, we kept the order of the sentences unchanged and use 90% of the dataset for training and the remaining 10% for testing. Then, k-fold cross-validation was performed to obtain the average performance over all possible splits without any overlap. Second, we voluntarily shuffle the dataset before splitting into training and test set. This allows us to quantify the potential bias related to the linguistic overlap between training and test sets.

3.4. Results and Discussion

The evaluated results are expressed in terms of Accuracy $Acc.(\%) = (N - D - S - I)/N$, where N, D, S, I are the number of phonemes in the test set, deletions, substitutions and insertions. The statistical significance of this measurement was assessed by calculating the Binomial proportion confidence interval $\Delta_{95\%}$ using the Wilson formula.

Results of all conducted experiments are presented in Table 1. First, correcting the phonetic transcription of the CSF18 corpus led to a performance gain of 6% (64.9% vs. 70.9% when considering 3-streams and unconstrained decoding). In addition, we report, here, the minimum accuracy (68.52%), the average accuracy (70.9%) and the best accuracy (72.33%) obtained across 10 splits with the 3-streams architecture on the CSF18-V2 corpus. This corpus should now be a reference corpus for future studies in the field.

We now compare the different architectures used to com-

Study	Features	Corpus	Model	Shuffling	Acc.(%)
Liu et al. [6]	CNN	CSF18	HMM-GMM	No	61.5
Papadimitriou et al. [12]	CNN		Fully Conv.	Yes	70.9
Wang et al. [13]	CNN		Cross-distillation	No	74.2
Ours	**Mediapipe**	CSF18 V2	Early Fusion	No	58.8
		CSF18 V2	2-streams Bi-GRUs (unconstrained)	No	63.5
		CSF18	3-streams Bi-GRUs (unconstrained)	No	64.9
		CSF18	3-streams Bi-GRUs (unconstrained)	Yes	77.1
		CSF18 V2	**3-Streams Bi-GRUs (unconstrained)**	**No**	**70.9**
			3-Streams Bi-GRUs (unconstrained)	Yes	79.2
			3-Streams Bi-GRUs (constrained)	**No**	**70.2**

Table 1. Results comparison for existing models and ours with and without shuffling train-test split which can bias the evaluated results. Early Fusion uses Fig.2(a) architecture, 2-Streams is Fig.2(b) with only Lips and Hand streams, and 3-Streams (also constrained) is the same as Fig.2(b). For all experiments, $\Delta_{95\%}$ confidence interval is around 4%. The results for the other approaches are reported as is from the papers [6],[12],[13].

bine hand and lip information within our model. At first, and similar to our previous study, the early fusion strategy provided the the worst results (58.8%). The 3-streams approach outperforms significantly the 2-streams one (e.g. 70.9% vs. 63.5%). Although the hand features include information about the position of all the points in hand, re-enforcing the hand position features by providing the third stream with index/middle finger tracking helps the model further. This highlights the importance of hand position in CS, an expected outcome as the hand position encodes the vowel information. The success of the middle fusion models provide reason to believe that the advancement of the hand with respect to the lips and its variability are learnt by the system.

Our results also highlight the bias due to the dataset shuffling while partitioning it into training and test subsets. This apparently harmless procedure can lead to significant gain of performance of more than 8% (70.9% vs. 79.2% and 64.9% vs. 77.1%). This bias should be considered carefully when comparing the different studies in the field.

We now discuss the performance of the proposed approach w.r.t other studies. First, the proposed method outperforms our previous CNN-HMM approach [6] by a margin of 4.8% (60.1% vs. 64.9%). Let us recall that the proposed method does not rely on the temporal segmentation of the visual data at the phonetic level. Then, to compare our approach with the method reported in the [12], we have to consider the biased condition due to dataset shuffling (as it seems to be the case in this study, see section IV-A in [12]) and the experiments conducted on the original CSF18 corpus. For these particular experimental conditions, our proposed method outperforms the one reported in [12] (77.1% vs. 70.9% [12]). However, when considering the original CSF corpus, our system performs significantly worse than the one proposed in [13] (64.9% vs. 74.2%) which is based on a more complex architecture (pre-trained teacher model with knowledge distillation toward a student model). When considering the corrected CSF18V2 corpus, our method reaches a comparable level of performance (70.9%). Thus, it would be interesting for future work to quantify the extra benefit given by knowledge distillation approach of [13] on this updated dataset.

Finally, we discuss the performance w.r.t to the CTC decoding strategy. First, the use of a $\sim$1000 words pronunciation lexicon (constrained decoding) is not as helpful as expected (no significant difference observed between the two decoding strategies, i.e. 70.9% vs. 70.2%). Moreover, the performance at the word level remains very low ($\sim$25% correctness). Nevertheless, some sentences were perfectly re-transcribed at the word level and others were partially correctly transcribed, e.g. "*le bedeau euphorique secoue l'anneau un jour par an*" decoded as "*le bedeau euphorique se qu' l' nos un jours part en*". These results may suggest that accurate decoding of CS at the word level would be possible after integrating more prior linguistic knowledge via the use of a statistical language model.

4. CONCLUSION AND PERSPECTIVES

To conclude, this study has explored and proposed a light model for ACSR which outperforms our previous approach and compete with other recent (and more complex) approaches. The use of an efficient pre-trained feature extractor allowed to reduce the complexity of the (visual) phonetic decoder. Also, cleaning the CSF18 dataset to account for differences in cueing and pronunciation of certain phonemes improved significantly the performance. However, the decoding performance at the word level, as well as the introduction of prior linguistic knowledge via the use of a pronunciation lexicon were not as helpful as expected. Recent neural language models, and the use of larger and multi-speaker datasets could lead to significant improvements in decoding performance.

5. REFERENCES

[1] R. O. Cornett, "Cued speech," vol. 112, no. 1, pp. 3–13, 1967.

[2] Lynne E. Bernstein, Paula E. Tucker, and Marilyn E. Demorest, "Speech perception without hearing," *Perception & Psychophysics*, vol. 62, no. 2, pp. 233–252, Jan. 2000.

[3] R. Orin Cornett, "Adapting cued speech to additional languages," *Cued Speech Journal*, 1994.

[4] Panikos Heracleous, Denis Beautemps, and Noureddine Aboutabit, "Cued speech automatic recognition in normal-hearing and deaf subjects," *Speech Communication*, vol. 52, no. 6, pp. 504–512, June 2010.

[5] Panikos Heracleous, Denis Beautemps, and Norihiro Hagita, "Continuous phoneme recognition in cued speech for french," in *Proc. of EUSIPCO*, 2012, pp. 2090–2093.

[6] Li Liu, Thomas Hueber, Gang Feng, and Denis Beautemps, "Visual Recognition of Continuous Cued Speech Using a Tandem CNN-HMM Approach," in *Proc. of Interspeech*, 2018, pp. 2643–2647.

[7] Li Liu, Gang Feng, Denis Beautemps, and Xiao-Ping Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 292–305, 2021.

[8] Virginie Attina, Denis Beautemps, Marie-Agnès Cathiard, and Matthias Odisio, "A pilot study of temporal organization in cued speech production of french syllables: rules for a cued speech synthesizer," *Speech Communication*, vol. 44, no. 1-4, pp. 197–214, Oct. 2004.

[9] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. of EMNLP*. 2014, Association for Computational Linguistics.

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification," in *Proc. of ICML*. 2006, ACM Press.

[11] MediaPipe Team and chuoling, "Mediapipe," https://github.com/google/mediapipe, 2019.

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

[13] Jianrong Wang, Ziyue Tang, Xuewei Li, Mei Yu, Qiang Fang, and Li Liu, "Cross-Modal Knowledge Distillation Method for Automatic Cued Speech Recognition," in *Proc. of Interspeech 2021*, 2021, pp. 2986–2990.

[14] Karl Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[15] Stephen John Young, NH Russell, and JHS Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," in *Technical report*.

[16] Alex Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer Berlin Heidelberg, 2012.

[17] Li Liu, Gang Feng, and Denis Beautemps, "Automatic temporal segmentation of hand movements for hand positions recognition in french cued speech," in *Proc. of ICASSP*, Apr. 2018, pp. 3061–3065.

[18] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015.