Investigating the dynamics of hand and lips in French Cued Speech using attention mechanisms and CTC-based decoding

Sanjana Sankar, Denis Beautemps, Frédéric Elisei, Olivier Perrotin, Thomas Hueber

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

`firstname.lastname@gipsa-lab.grenoble-inp.fr`

Abstract

Hard of hearing or profoundly deaf people make use of cued speech (CS) as a communication tool to understand spoken language. By delivering cues that are relevant to the phonetic information, CS offers a way to enhance lipreading. In literature, there have been several studies on the dynamics between the hand and the lips in the context of human production. This article proposes a way to investigate how a neural network learns this relation for a single speaker while performing a recognition task using attention mechanisms. Further, an analysis of the learnt dynamics is utilized to establish the relationship between the two modalities and extract automatic segments. For the purpose of this study, a new dataset has been recorded for French CS. Along with the release of this dataset, a benchmark will be reported for word-level recognition, a novelty in the automatic recognition of French CS.

Index Terms: cued speech, hearing impaired, assistive technology, explainability, corpus, multimodality

1. Introduction

Cued Speech (CS) is a visual communication technique developed by Cornett [1] in 1967 to facilitate the interpretation of spoken language for people with hearing impairment. As shown in Figure 1, it uses a combination of hand shape and hand position with respect to the lips to represent different phonemes being uttered while speaking a language. It helps people who are hard of hearing to differentiate between phonemes which look similar at the lips and are difficult to distinguish while lip-reading, like [b], [p] and [m]. In this work, the focus is on French CS or Langue française Parlée Complétée (LfPC) [2], which uses five hand positions to disambiguate vowels and eight hand shapes for consonants. There have been several attempts to develop systems for automatic CS recognition (ACSR) in recent years to provide speech technologies that aid people with hearing disabilities. In early studies [3, 4], the focus was on isolated vowel and/or consonant recognition and artificial landmarks are used to facilitate the segmentation and tracking of hand and lips. Later studies have moved on to the more challenging task of *continuous* cued-speech recognition[5, 6, 7, 8, 9]. However, all these studies provide decoding strategies of LfPC at the phonetic level and not at the word level which is a prerequisite for the implementation of a practical solution (e.g. for a dictation application or when combined with a Natural language understanding system). Investigating new decoding strategies of CS at the word level is one motivation of the present study. However, the only available corpus of LfPC is the CSF18 corpus [5] which was designed for the purpose of generation (see [10] for its use in the context of Cued-speech synthesis) rather than for recognition. Its lin-

guistic material was designed to cover all possible diphones of the French language in a minimum number of sentences. As a consequence, most of the sentences of this corpus present very rare word sequences and unusual syntactic structures, (e.g."A giant staple may have hit his beautiful speedboat"). These sentences are not suitable for recognition as they are very unlikely for a statistical language model (LM). Also, a new dataset with more realistic linguistic content is needed. To this end, such a dataset has been designed and recorded to be released with this paper. On this dataset, a new model is proposed and evaluated for decoding LfPC at the word level based on (i) automatic hand and lips pose extraction, (ii) RNN-based phonetic decoding (iii) beam search decoding procedure with lexical constraint and n-gram language modeling, and (iv) top-K neural rescoring using a pre-trained GPT2 model.

In addition to this technological objective, this study also aims to analyze such a decoder to better understand the production of CS by humans. In particular, the focus is on the complex dynamic relationships between the hand and lips. In fact, there is an inherent asynchrony between the onset of the hand movement w.r.t to the lip movement. In [11], it was established that the hand can precede the lips by a delay of a few milliseconds to several hundred milliseconds. To take into account this asynchrony in ACSRs, [5] assumed an average anticipation of the hand and used a heuristic to synchronize the 2 modalities. More recently, [8, 6, 9] have used bi-directional RNNs to capture useful contextual information for both hand and lips streams before their joint modeling. In this study, it is proposed to use an ACSR system to elucidate the temporal relationships between the hand and lips. To this end, temporal self-attention mechanisms for hand and lip feature streams are integrated within an RNN-based phonetic decoder and the resulting attention maps are analyzed a posteriori. The working hypothesis for this article is the following: in the case of an anticipatory gesture of the hand w.r.t lips, the model should be able to learn to pay attention to the onset of this gesture. Therefore, by detecting any deviation from the diagonal of the attention map, these onsets could be extracted automatically. It should provide a way to visualize hand-lip asynchrony in CS but also to segment hand and lip movements automatically, a time-consuming task and difficult when done manually. Such an automatic segmentation process could thus be useful for the fundamental research on the production and perception of CS [12, 13, 14, 15].

In summary, the key contributions of this paper are (i) releasing a new dataset for LfPC adapted to word-based decoding, (ii) providing a first baseline of an automatic decoder of LfPC at the word level and (iii) investigating the use of self-attention mechanisms to provide a fine-grained analysis of hand-lips temporal relationships and a method for segmenting CS data.
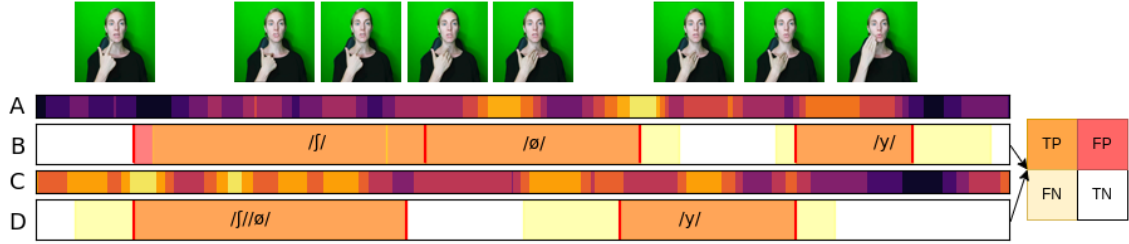
Figure 1: *Automatic segmentation of hand and lip movements in LfPC from the attention maps of a CTC-based decoder. (A) & (C) show the attention scores along the DTW-based attention path for lips and hand respectively while (B) & (D) show the automatic vs. manual segmentation for lips and hand respectively. True positives (TP, in orange) correspond to automatically detected segments that overlap with manually annotated ones. The color scale for (A) & (C) is as shown in Fig.3.*

2. Methodology

2.1. Model architecture

The model architecture for the ACSR task is presented in Fig.2. Similarly to [8], it has 3 streams corresponding to lips, hand shape and hand position respectively. For each stream individually, a first Bidirectional Gated Recurrent Unit (Bi-GRU) captures contextual information from the sequence of visual features. However, to interpret how the model exploits this contextual information, a temporal self-attention mechanism is added after each (Bi-GRU). This mechanism is based on the scaled-dot product attention defined in [16] as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V = W_{att}V \quad (1)$$

where K, Q, and V (key, query, and value) are embedded representations learned from the input sequence (in our case, the output of the Bi-GRU). Here, W_{att} is the attention score at each time step for all the frames of a given sequence. The outputs of each self-attention mechanism are then concatenated together, passed to a Bi-GRU (processing the 3 streams of information jointly) and finally to the output *softmax* layer. The decoder is trained only with phonetic sequences as a target using the CTC loss [17] and importantly, without any temporal segmentation at the phonetic level (the dimension of the final softmax layer is $k + 1$ where k is the number of phonetic classes, one output encoding the CTC *blank* token).

2.1.1. Word-level decoding

Word-level decoding is achieved using a beam search decoder. During beam search, phonetic hypotheses are iteratively expanded with the next possible phonemes. At each time step, only the hypotheses with the highest scores are retained. A pronunciation dictionary associating words with phonetic transcription is used to constrain the decoding (only words from the dictionary can be generated). An N-gram language model provides a likelihood score for each sequence of hypotheses during the beam-search decoding. An auto-regressive neural model based on Transformer (GPT) is finally used a posteriori to re-score the top-K sequences produced by the beam search decoder. This re-scoring is done by computing its perplexity w.r.t each candidate sequence $X^{(k)} = (x_0, \dots, x_N)$ (classically tokenized), defined as:

$$PPL(X^{(k)}) = \exp{-\frac{1}{t}\sum_i^t \log p_\theta(x_i^{(k)} \mid x_{<i}^{(k)})} \quad (2)$$

where θ is the parameter set of the LM. The candidate sequence for which the perplexity is minimal is finally selected.

2.2. Attention-based temporal segmentation

In accordance with the working hypothesis, the model should pay attention to the onset of hand and/or lips movements, providing a way to segment those movements. Therefore, for a given test sentence, the attention maps for hand and lip streams are extracted. From these attention maps, an optimal path - hereon referred to as the attention path - along which the cumulative attention is maximum is extracted using the Dynamic Time Wrapping (DTW) algorithm (the attention score between each pair of frames of the sequence is used as the local distance). A Sakoe-Chiba band of 30 frames is used to constrain the attention path to be not too far from the diagonal of the attention map. Frames at which the attention map deviates from the diagonal are expected to match the onset of hand/lip movements. This simple heuristic also avoids setting an arbitrary threshold on the value of the attention scores. When there are several consecutive frames detected as candidates for the movement onset, the middle frame is considered. We then constrain the number of segments to be not more than the number of predicted phonemes and assign the boundaries to the predicted phonemes. Finally, the respective position of the onsets of hand and lip movements for each phoneme reveals the asynchrony between these two modalities of CS.

For evaluating the accuracy of the proposed segmentation technique, "temporal intersection over union" $tIoU$ is calculated as described in [18] and defined as:

$$tIoU = \frac{1}{|S|}\Sigma_{s \in S}\frac{l_{ts} \cap l_{ps}}{l_{ts} \cup l_{ps}}, \quad (3)$$

where S is the set of all the segments from manual annotation (ground truth), l_{ts} is the true annotation segment s is assigned to, and the output of the intersection (or union) operation is the number of frames of intersection (or union) between the segment spanned by l_{ps} which is the predicted segment.

3. Experiments

3.1. CSF2022 Dataset

The new dataset consists of 1 087 sentences recorded by a professional French cuer with typical hearing. These sentences were selected from the SIWIS database [19] and are originally from various sources, such as French parliament debates,
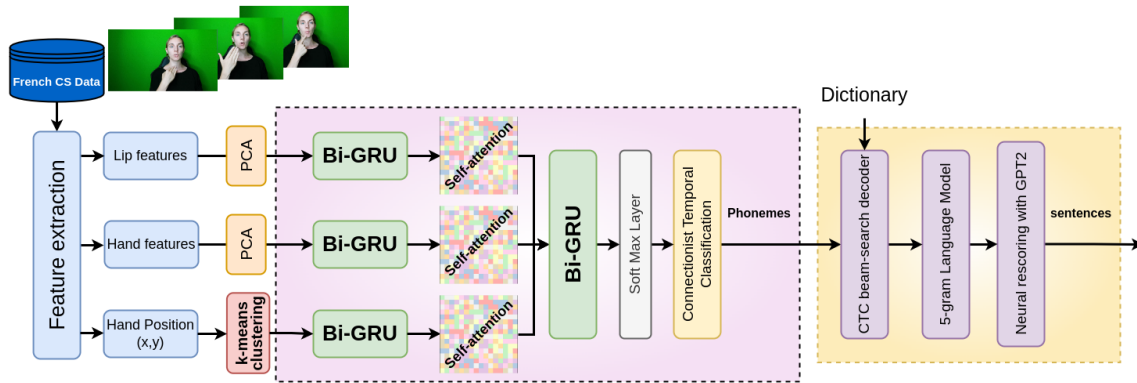
Figure 2: *Model architecture*

3.2. Visual Features

French novels and semantically unpredictable sentences (the latter not being used at test time for word-based decoding). The experimental setup also includes LED lights (placed next to the cuer's face) and a green background. The cuer was also requested to utter the sentences while cueing. The sentences were prompted on the screen using a dedicated web-based application. In the mid-term perspective of recording data 'in the wild', i.e. professional CS interpreters recording data in various environments and devices (e.g. at home with their smartphone), the videos were recorded using a simple webcam (1920x1080 pixels). The targeted framerate was set to 60 fps. Audio was recorded at 44.1 kHz (PCM, 16bits) but is not used in this study. In total this new dataset contains 97mn of CS data (i.e. 3 times more than the CSF18 corpus). The phonetic transcription of the recorded sentences was obtained automatically using the LIA-Phon phonetizer [20] and verified/corrected manually (36 phonetic categories are used to describe French). To evaluate the accuracy of the attention-based segmentation, a set of 20 sentences was manually annotated with the ELAN software [21] to provide the true segmentation boundaries for lips, hand shape and hand position based on the video. The dataset (including raw data, hand/lips pose estimation, manual annotations, and dedicated train and test sets) will be made available publicly in association with this paper.

3.2. Visual Features

As mentioned above, the videos are recorded with a target framerate set to 60fps via a web application that facilitates the recording of data in the wild. However, due to the use of a webcam (and probably also due to the encoding codec), the frame rate is found to be variable when extracting the images (i.e. 60Hz with a few missing frames). Each extracted image is then processed by the Mediapipe software [22] which provides 21 x-/y-coordinates for hands and 42 x-/y-coordinates for the lips. The resulting sequence of visual feature is re-interpolated to match the rate of 60 fps. The dimension of the visual feature vectors is then reduced to 20 for each stream using PCA (explaining more than 99% of the variance). For the hand position features, the data is clustered using the k-means algorithm and the cluster labels (one-hot encoded) are used as the feature for the third stream. 8 distinct clusters are chosen to account for the 5 hand positions and a few transition states.

3.3. Implementation details

Each Bi-GRU of the ACSR model has 256 units. The embedding size of the weight matrix used to obtain the Q, K and V matrices in the (single-head) self-attention layer is set to 256. The Adam optimizer is initialized with a learning rate of 0.001 and reduced on a plateau. The model is trained for 120 epochs with a batch size of 16.

For the word-level decoding, a pronunciation dictionary is created with all the words in the new CSF2022 dataset to which a list of the 1000 most frequent words used in French is added. This list was extracted from *Spacy.io* online tool, resulting in a 3.2k words lexicon. For the beam search decoding, the pretrained 5-gram LM provided in the Voxpopuli corpus [23] is used. Evaluations are conducted on a test set of 108 sentences selected randomly (avoiding semantically unpredictable sentences). To the best of our knowledge, these sentences are not in the training set of this LM (which is rather based on European Parliament debates). The CTC beam search decoder available in PyTorch (version 1.13.1) is used in this study. The hyperparameters *beam_width*, *word_score* and *lm_weight* are set respectively to 1000, 0 and 0.2, after being optimized on a subset of the training set of 30 sentences. For the rescoring procedure, the top-30 sequences of words from the beam search decoder are considered and the pre-trained GPT2 model for French proposed by [24] and available on *HuggingFace* is used.

4. Results and Discussion

4.1. Recognition performance

Since the proposed system is partially based on the phonetic decoder described in [8], the latter is used as a baseline and its performance on the new corpus is evaluated. The results are presented in Table 1. First, thanks to the temporal attention mechanisms, the phonetic decoder is more easily interpretable while also maintaining performance as the baseline(Acc=77.5% vs. 76.6%). Several ablations studies are conducted to assess the impact of the different modules involved in the decoding at the word level. It can also be observed that constraining the vocabulary slightly degrades the performance at the phonetic level (Acc=74.1% vs. 76.6%). This is mainly due to a higher number of deletion errors and can be explained by the lack of information about the "liaison", which occurs quite often in the French language. This issue is alleviated when using the 5-gram LM (Acc=76.8%).

Best performances at the word level are obtained by

Table 1: *Performances of the proposed system for automatic decoding of French Cued-Speech, both at the phonetic and word levels, in terms of correctness (Corr) and accuracy (Acc, which takes into account the insertion errors). Here, $Acc. = 100(1 - WER)\%$. The Wilson confidence interval $\Delta_{95\%}$ is also reported, assuming a binomial distribution of the errors.*

Configuration	Level	Corr $\pm\Delta_{95\%}$	Acc $\pm\Delta_{95\%}$
Baseline phonetic decoder [8]	Phone	81.3 ± 1.6	77.5 ± 1.8
Phonetic decoder (baseline + self-attention)	Phone	81.2 ± 1.6	**76.6** ± 1.8
Phonetic decoder + Lexicon	Phone	77.0 ± 1.6	74.1 ± 1.8
	Word	41.8 ± 3.9	27.7 ± 3.9
Phonetic decoder + Lexicon + 5-gram LM	Phone	79.4 ± 1.6	76.8 ± 1.8
	Word	54.5 ± 3.9	52.9 ± 3.9
Phonetic decoder + Lexicon + 5-gram LM + GPT-based rescoring	Phone	79.2 ± 1.6	76.5 ± 1.8
	Word	**61.9** ± 3.9	**58.5** ± 3.9
Phonetic decoder + Lexicon + Dict + 5-gram LM + 30-best	Phone	82.8 ± 1.5	80.4 ± 1.6
	Word	71.4 ± 3.5	70.7 ± 3.5

combining the 5-gram model and the GPT-based rescoring, (Acc=58.5%). It can be noted that the 5-gram model improves the accuracy by 25.2% compared to the lexicon-only configuration (Acc=52.9% vs. 27.7%) and that the GPT-rescoring resulted in an additional 5.6% improvement (Acc=58.5% vs. 52.9%). The analysis of the decoded sequences reveals that an important number of errors are due to deletion or insertion of short words (e.g."vous **n'**avez donc rien inventé" instead of "vous avez donc rien inventé **et**", or syntax errors (e.g. "à leur famille" instead of "à leur**s** famille**s**") which are unfortunately not always corrected by GPT-2. To confirm these interpretations, the performance obtained when considering the 30-best hypotheses from the beam search decoding is reported while keeping only the one for which the accuracy at the word level is maximum. The performance is much higher (Acc=70.7% at the word level) but obviously, this configuration cannot be used in practice since the ground truth is not known at decoding time. However, it provides an estimate of the performance that could be achieved if ambiguities at the semantic level could be eliminated (e.g. by exploiting a larger linguistic context).
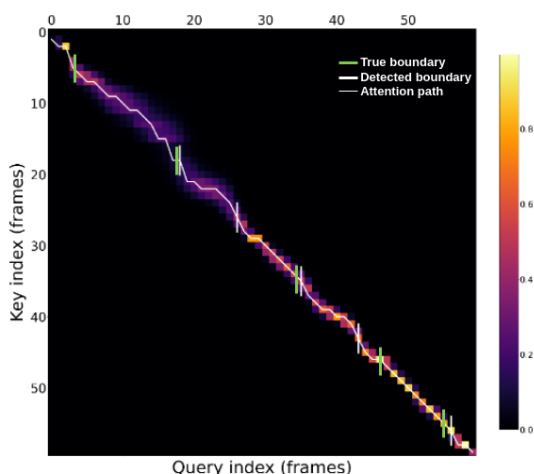
4.2. Attention-based segmentation

Figure 3: *Attention map for hand shape with corresponding DTW attention path and detected movement onsets (for one sentence from the test set)*

Fig.3 shows the onset of hand shapes as detected from the attention map along the attention path and from manual annotation. It can be seen from the graphs that the detected segmentation onsets are close to the manually annotated onset. On average, the automatically detected segmentation onset falls within 5 frames of the ground truth (i.e $\pm$ 40ms). Similar results are observed for the lips, but surprisingly, the attention map for the hand position is much less interpretable. This needs further study and the focus here is on lips and hand shape.

In Fig.1, an example of automatic segmentation of both lips and hand movements is presented for one of the manually annotated sentences. While not perfect, it is clear that the proposed method segments the visual data in a consistent manner w.r.t to the manual annotation. This tendency is confirmed by the relatively high $tIoU$ (averaged over all the manually annotated sentences, see Eq.3) which was found to be 60.3% for hand and 69.6% for the lips. From the analysis of the attention maps and from the automatic segmentation results, it can be observed that the temporal relationships between the hand and lips in CS are complex. The hand gesture may precede the lip gesture by several hundred milliseconds. However, the two articulators are sometimes required to synchronize. In such cases, the two attention paths overlap and are close to the diagonal.

5. Conclusions

Although the decoding performances at the word level are still quite far from those of an automatic acoustic speech recognition system, the proposed ACSR system is a first step toward a system that can be used in practice. In addition, it is shown that the attention maps of an ACSR can be used to automatically detect (with a minimum of post-processing) the onset of hand and lips movement in Cued-speech, providing potential useful information on its underlying organisation. In our future work, this technique will be applied to a multi-cuer database, in order to study potential speaker-specific strategies in hand-lips coordination in cued speech.

6. Acknowledgements

7. References

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

[2] ——, "Adapting cued speech to additional languages," *Cued Speech Journal*, 1994.

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

[4] N. Aboutabit, D. Beautemps, and L. Besacier, "Hand and lip desynchronization analysis in French Cued Speech: Automatic temporal segmentation of hand flow," in *Proc. of ICASSP*, vol. 1, 2006, pp. I–I.

[5] L. Liu, G. Feng, D. Beautemps, and X.-P. Zhang, "Resynchronization 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.

[6] 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. of Interspeech 2021*, 2021, pp. 2986–2990.

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

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

[9] L. Liu and L. Liu, "Cross-modal mutual learning for cued speech recognition," 2022. [Online]. Available: https://arxiv.org/abs/2212.01083

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

[11] V. Attina, D. Beautemps, M.-A. Cathiard, and M. 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.

[12] J. C. Krause, K. A. Pelley-Lopez, and M. P. Tessler, "A method for transcribing the manual components of Cued Speech," *Speech Communication*, vol. 53, no. 3, pp. 379–389, Mar. 2011.

[13] C. Hage and J. Leybaert, "The effect of Cued Speech on the development of spoken language." in *Advances in the spoken language development of deaf and hard-of-hearing children*. Oxford University Press, 2006, p. 193–211.

[14] L. Machart, A. Vilain, H. Loevenbruck, and L. Ménard, "Influence of french Cued Speech on consonant production in children with cochlear implants: an ultrasound study," in *in Proc. of International Seminar on Speech Production (ISSP)*, 2020.

[15] C. Bayard, L. Machart, A. Strauß, S. Gerber, V. Aubanel, and J.-L. Schwartz, "Cued Speech Enhances Speech-in-Noise Perception," *The Journal of Deaf Studies and Deaf Education*, vol. 24, no. 3, pp. 223–233, 02 2019.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NIPS*, 2017, p. 6000–6010.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification," in *Proc. of ICML*. ACM Press, 2006.

[18] P. Peng and D. Harwath, "Word discovery in visually grounded, self-supervised speech models," *Interspeech 2022*, pp. 2823–2827, 2022.

[19] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The siwis french speech synthesis database? design and recording of a high quality french database for speech synthesis," Idiap, Tech. Rep., 2017.

[20] F. Bechet, "Liaphon : Un système complet de phonétisation de textes," *Traitement Automatique des Langues*, no. 42, p. 47–67, 2001.

[21] M. P. I. for Psycholinguistics The Language Archive The Netherlands, "Elan (version 6.4) [computer software]," 2022. [Online]. Available: https://archive.mpi.nl/tla/elan

[22] 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. of Third Workshop on Computer Vision for AR/VR at CVPR 2019*, 2019.

[23] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. of the 59th ACL Annual Meeting (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[24] A. Simoulin and B. Crabbé, "Un modèle Transformer Génératif Pré-entraîné pour le français," in *Traitement Automatique des Langues Naturelles*, P. Denis, N. Grabar, A. Fraisse, R. Cardon, B. Jacquemin, E. Kergosien, and A. Balvet, Eds. Lille, France: ATALA, 2021, pp. 246–255.