Bridging the Gap: Design and Evaluation of an Automated System for French Cued Speech

Brigitte Bigi

Bridging the Gap: Design and Evaluation of an Automated System for French Cued Speech

Brigitte Bigi
Laboratoire Parole et Langage, CNRS, Aix Marseille Univ, 13100 Aix-en-Provence - France
brigitte.bigi@cnrs.fr

Abstract

Access to spoken language remains a challenge for deaf and hard-of-hearing individuals due to the limitations of lipreading. Cued Speech (CS) addresses this by combining lip movements with hand cues—specific shapes and placements near the face—making each syllable visually distinct. This system complements cochlear implants and supports oral language, phonological awareness, and literacy. This paper introduces the first open-source system for automatically generating CS in video format. It takes as input a video recording, the corresponding audio signal, and an orthographic transcript. These elements are processed through a modular pipeline, which includes phonetic mapping, temporal alignment, spatial placement, and real-time rendering of a virtual coding hand. The system is multilingual by design, with current resources focused on French. An evaluation under varied conditions showed decoding rates up to 92% for manually coded stimuli, and averages exceeding 80% for automatically generated ones. Visual clarity of hand shapes proved more critical than timing or angle. Stylized designs and frontal views enhanced decoding performance, while attempts at naturalistic rendering or visual effects often degraded it. These findings indicate that visual abstraction improves readability. This work provides a reproducible and scientifically grounded framework for visual phonetic encoding, and delivers a practical tool for education, accessibility, and research.

1 Introduction

1.1 Visual Access to Spoken Language through Cued Speech

Speech production involves both acoustic and visual cues. While lip movements convey useful information, many phonemes appear identical on the lips and form so-called "visemes"—groups of phonemes that are visually indistinguishable (Fisher, 1968; Massaro and Palmer Jr, 1998). As a result, lipreading remains highly ambiguous: correct word identification rarely exceeds 30% (Rönnberg, 1995; Rönnberg et al., 1998).

To address this limitation, R. Orin Cornett introduced Cued Speech (CS) (Cornett, 1967), a visual communication system designed to make each phoneme visually distinct. CS combines lip movements with hand cues—specific handshapes and positions placed around the face—that encode consonants and vowels. It provides full visual access to spoken language and supports phonological awareness, literacy development, and spoken language acquisition in deaf or hard-of-hearing individuals (Clarke and Ling, 1976; Neef and Iwata, 1985). CS has since been adapted to over 65 languages[1].

Cued Speech is widely used by speech-language pathologists to support early language acquisition in deaf children. Among others, in France, it is promoted by the Association pour la Langue française Parlée Complétée (ALPC)[2], and in the US by the National Cued Speech Association[3]. Numerous studies have shown that CS enhances access to phonological structure, supports literacy development, and fosters inclusive education (Leybaert and Charlier, 1996; Colin et al., 2017; LaSasso et al., 2010). Together, these findings highlight its importance in supporting language acquisition pathways for deaf learners.

Building on its demonstrated benefits for access to spoken language, Cued Speech and Sign Languages serve distinct linguistic and cultural functions. They are not mutually exclusive: while some deaf children follow a sign language pathway, access to reading and writing typically requires exposure to spoken language. By offering a precise visual representation of sounds, Cued Speech supports this process. It is therefore relevant to all deaf learners aiming to acquire spoken language,

[1] https://www.academieinternationale.org/
[2] https://alpc.asso.fr/
[3] https://cuedspeech.org/

whether or not they use a sign language. This distinction is essential to avoid misinterpretations: Cued Speech is not a language and is not intended to replace natural sign languages such as LSF, but to complement them when access to spoken language is required or preferred.

Following the general principles of CS, the French adaptation was developed in the 1970s. It uses eight handshapes to encode consonants and five facial positions to encode vowels. Each syllable is represented by a combination of lip movement and a hand cue, also called a key, formed by a handshape–position pair. A simple syllable like CV or V is coded by a single key, while more complex structures, such as CCV, require multiple successive keys: for example, a 'C-' followed by a 'CV' structure. To illustrate this system, Figure 1 shows the handshapes used for consonants, and Figure 2 shows the vowel positions around the face. Both figures include the neutral position used when no speech is pronounced.

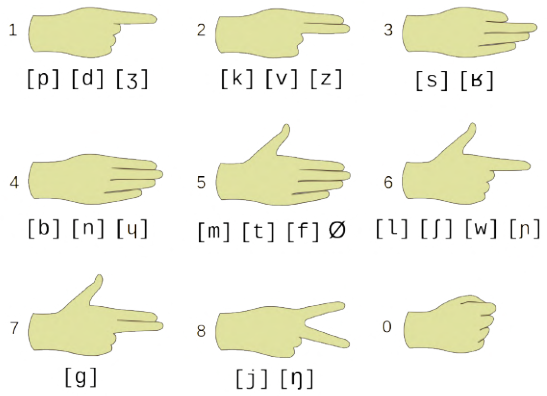

Figure 1: Handshapes representing consonants

Figure 2: Positions representing vowels

Below is a concrete example showing how a sentence is encoded into cues:

```
text:              Tu es gris.
phones:            t y e g ʁ i
CV sequence:       C V V C C V
cues-structure:    C-V.-V.C-.C-V
cues code:         5-t.5-t.7-s.3-m
```

The internal consistency of Cued Speech makes it well-suited for automation. Generating cues from speech or text opens the door to a wide range of applications: cued videos for learning and access, training tools for families and educators, and greater availability of CS in contexts where trained coders are not present. More broadly, automatic cueing can support language acquisition in deaf children, improve communication in mixed hearing environments, and reinforce lipreading skills.

This paper presents the first complete and shareable system for automatic CS generation. It takes as input a video recording, its audio signal, and a transcript, and produces a new version of the video in which a synchronized virtual hand encodes the CS transcription. The architecture was built entirely from scratch, formalizing each stage of the process from segmentation to cue rendering. It is designed for multilingual use and has been implemented and tested for French. The full system is open-source, and all components have been evaluated with end-user testing.

1.2 Related Works

The first attempt to automate cue generation, AutoCuer, was developed by R. Orin Cornett himself (Cornett et al., 1977). Between 1995 and 2000, a series of studies at the Massachusetts Institute of Technology (MIT) explored real-time automatic cueing for American English (Bratakos, 1995; Sexton, 1997; Bratakos et al., 1998; Duchnowski et al., 2000). These remain the most advanced documented efforts in the field. Their system relied on speaker-dependent automatic speech recognition to extract phonemes from live recordings, which were then converted into hand cues and displayed as a virtual hand overlaid on the video. Evaluations showed significant gains in decoding accuracy, with some conditions yielding scores twice as high as lipreading alone. However, many components required manual adjustments (Sexton, 1997): cue positions were initialized by hand, transitions were interpolated without formal modeling, and the mapping rules were not described in reusable form. The lack of published code or reproducible design has prevented further development or reuse.

To date, no operational or open-source tool exists for automatic CS generation in any language, despite increasing scientific interest and documented benefits.

1.3 Foundations and Scope

Developing a complete system was a necessary step, independently of data availability. It provided the opportunity to define a structured architecture, implement a fully functional version, and formalize the modeling of each component. The resulting system is transparent, reproducible, and already usable in real conditions. It operates with minimal computational cost, can be refined through expert feedback, and offers a solid basis for future improvements, including data-driven modules once annotated resources become available.

A French Cued Speech corpus has recently been collected and partially annotated (Bigi et al., 2022), but the annotation process is still ongoing due to the precision required.

This work then marks the beginning of a long-term effort to build a reliable and extensible framework for automatic CS generation. It defines a shared foundation for future developments in augmented video production and evaluation.

2 System Description

While many studies describe individual aspects of CS production—such as articulation, speech coordination, timing, or spatial organization—formal descriptions remain rare. Few are presented in a way that supports computational modeling or system implementation. The literature describes many aspects of CS production. However, formal accounts of its speech coordination, timing, and spatial organization remain rare. Few works address these questions, and the descriptions are rarely framed in terms of computational modeling.

In this work, the cueing process was analyzed by combining published linguistic descriptions (Attina, 2005; Aboutabit, 2007) with structured discussions conducted with experienced coders. This led to the identification of four core processing components, which structure the system: determining what to code (i.e., the sequence of keys from the phoneme transcription), when to display the cues (synchronization with the speech signal), where to place the hand (spatial positioning, angle, and size), and how to render it visually (hand design).

The four components are interdependent: timing depends on phoneme alignment, spatial positioning requires both timing and content, and visual rendering builds on all previous stages. This structure is the result of the analysis described above. It defines an architecture for cue generation and supports the implementation of a consistent and extensible system. The same framework has guided the present system and can serve as a reference for future developments.

For example, the system is *multilingual by design* in the sense that language-specific knowledge is externalized into modular, open-format resource files. The core components—covering normalization, phonetic transcription, alignment, and cue generation—are implemented in a language-independent way. Language-specific resources, such as dictionaries, acoustic models, and cueing rules, are handled through separate, editable files. This modular architecture follows the same strategy as adopted in SPPAS for text normalization (Bigi, 2014), phonetic transcription (Bigi, 2016), and alignment (Bigi and Meunier, 2018). Its applicability to multiple languages has already been validated in these components (Lancien et al., 2020; Bigi et al., 2021; Pakrashi et al., 2023), and is here extended to the novel task of Cued Speech generation.

Figure 3 presents the full processing pipeline, from user inputs to the final coded video. It illustrates the modular organization of the system and the sequence of required operations. The first stages involve automatic processing of the input transcript, audio, and video using the open-source SPPAS toolkit (Bigi, 2015), including normalization, phonetization, forced-alignment, and face analysis. These annotations are used without manual correction and provide the foundation for reproducible experiments. The subsequent steps implement the proposed framework, computing the sequence of keys, their temporal and spatial properties, and rendering the virtual hand accordingly.

2.1 What to Code

The first component of the system determines the sequence of keys to be produced from the phoneme transcription. Each key encodes a consonant–vowel association as a pair of handshape and position. Based on the aligned phoneme sequence, the system infers the structure and associates each segmental unit with a key of type C-, -V, or CV. A deterministic finite automaton (DFA) formalizes

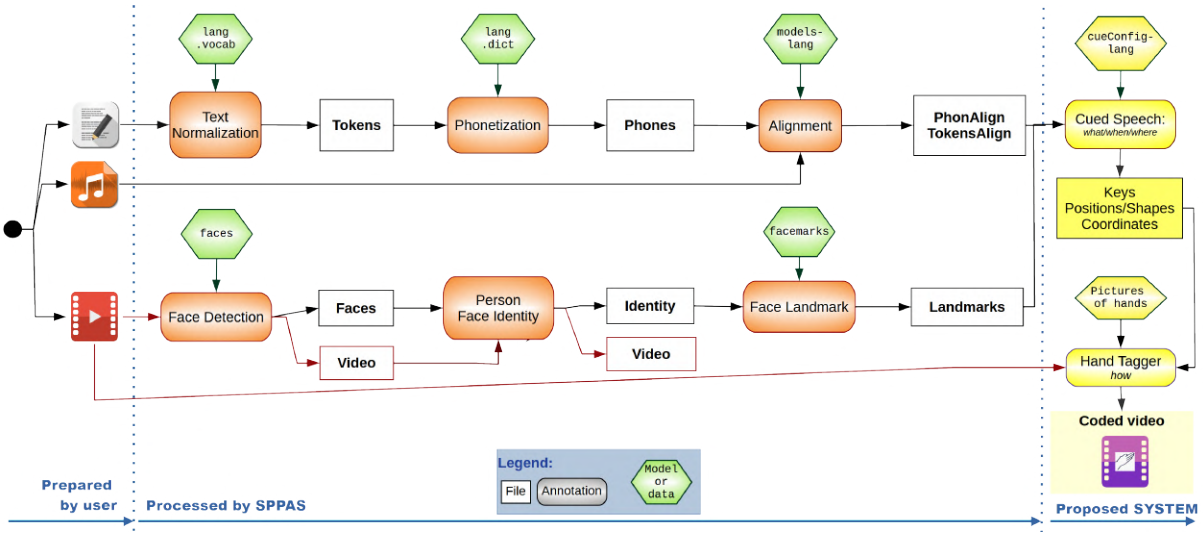

Figure 3: Workflow of the full process: from the user's data to the coded video

all valid transitions and decomposes complex syllables into successive keys.

This component was previously described and evaluated in a dedicated study (Bigi, 2023). On a manually annotated corpus, the predicted sequences aligned closely with those produced by expert coders, with most deviations reflecting individual preferences rather than systemic errors. The DFA-based system was found to be both reliable and sufficient. A web-based text-to-cue converter[4], developed in collaboration with the deaf community, provides public access to this module for educational and training use.

2.2 When to Display the Cues

Once the sequence of keys is defined, the next step is to determine their temporal coordination with speech. It is already known that the hand must anticipate the associated phonemes to allow visual decoding. This principle has been consistently supported in the literature (Cornett, 1967; Bratakos et al., 1998; Duchnowski et al., 1998, 2000) and confirmed by French studies (Cathiard et al., 2003; Attina, 2005; Aboutabit, 2007), which highlight the role of anticipation in perception.

Four timing models were implemented: three drawn from previous work, and a fourth developed specifically for this system. The notation introduced in (Attina, 2005) is used throughout. A1 marks the acoustic onset of the key—consonant onset in 'C-' or 'CV' keys, vowel onset in '-V' keys. A3 marks the acoustic offset—vowel end in 'CV' or '-V' keys, consonant end in 'C-' keys. M1 and M2 represent the start and end of the manual transition. The interval A1–A3 corresponds to the acoustic duration of the key, while M1 and M2 are the time points to be predicted by the models.

Model 1 reproduces the configuration described in (Duchnowski et al., 1998), in which the hand appears 100 ms before the phoneme, with no transition phase. This model was implemented for reference purposes but was not included in the experimental protocol, as later studies (Duchnowski et al., 2000) have shown that Model 2 yields better results. **Model 2** introduces a fixed transition of 150 ms, so that the hand reaches its target 100 ms prior to the phoneme onset.

Model 3 adjusts anticipation values based on the consonant–vowel structure of the key. It is derived from French-language studies (Attina, 2005), which reported consistent variation in cue timing across key types. Transitions are defined as proportions of the A1–A3 interval, assuming an average duration of 400 ms. For 'CV' and 'C-' keys, M1 starts 62% before A1 and M2 occurs 10% after A1. For '-V' keys, M1 starts 46% before A1 and M2 occurs 21% after A1.

Model 4 was developed specifically for this system. It extends previous models by incorporating finer adjustments derived from coder expertise and by explicitly modeling transitions involving the neutral position, which are absent from earlier systems. The model adapts timing to speech rate and defines transition points as proportions of the A1–A3 interval.

For the first key, corresponding to a transition from the neutral zone to a facial position, M1 oc-

[4] https://auto-cuedspeech.org/textcue.html

curs 140% before A1 and M2 20% before A1. For the second key, these values are 125% and 15% before A1. For the third, 100% and 10%. For subsequent keys, M1 is set to 90% and M2 to 5% before A1. For the return to neutral, M1 is delayed to 20% after A1, and M2 to 80% after M1.

2.3 Where to Place the Hand in the Video?

This component determines the position, angle, and size of the hand relative to the speaker's face for each frame of the video.

The vowel positions were first defined by expert coders on a theoretical face, then formalized using the 68-point facial landmark model given by SPPAS. Each position is computed as a function of facial landmarks. The formulas used for the positions of French Cued Speech were derived in collaboration with expert coders and adapted to ensure consistency across speakers and morphologies. They are summarized in Table 1 and illustrated in Figure 4. No variability was introduced at this stage: for each frame, the fingertip is placed at the target coordinates.

	x =	y =
n	x_8	$y_8 + 4 \cdot (y_8 - y_{57})$
b	$x_4 + \frac{1}{2} \cdot \lvert x_{36} - x_0 \rvert$	$y_1 - \frac{1}{3} \cdot \lvert y_1 - y_0 \rvert$
c	x_8	$y_8 - \frac{1}{5} \cdot \lvert y_8 - y_{57} \rvert$
m	$x_{48} - \frac{1}{4} \cdot \lvert x_{48} - x_4 \rvert$	y_{60}
s	$x_0 - \frac{2}{3} \cdot \lvert x_8 - x_0 \rvert$	$y_4 - \frac{1}{2} \cdot \lvert y_4 - y_3 \rvert$
t	x_8	$y_8 + 1.2 \cdot \lvert y_8 - y_{57} \rvert$

Table 1: Estimated positions from facial landmarks

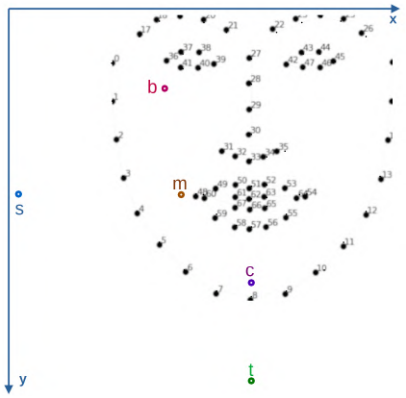

Figure 4: Estimated positions relatively to the landmarks

Hand orientation is also controlled to improve visual realism. Three models were implemented. **Model 0** uses a fixed angle of 60°, serving as a baseline (Figure 5). **Model 1** introduces expert-defined

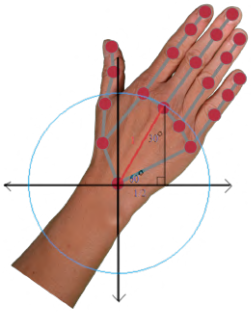

Figure 5: Hand angle of Model 0 is 60°.

variations by position. Excluding the neutral zone, the average angle is 71.2°, with a standard deviation of 9.3°. **Model 2** uses a data-driven approach: five annotated frames per position were manually selected to estimate average orientations. It yields an average angle of 61.8° and a standard deviation of 12.5°. Detailed values are given in Table 2.

Position	**Model 1**	**Model 2**
n (chest)	50°	50°
b (cheek bone)	75°	62°
c (chin)	67°	59°
m (mouth)	73°	56°
s (side)	83°	83°
t (throat)	58°	49°

Table 2: Hand angles (in degrees) for Models 1 and 2.

The hand size is scaled proportionally to face height and remains fixed throughout. Transitions between positions follow a straight-line trajectory at constant speed. Handshape transitions occur at the midpoint of this trajectory, using a three-frame fade between the two handshapes. These simplifications reflect a design choice: only one spatial parameter is introduced at a time for evaluation.

This component of the system then produces a complete 2D hand trajectory of the hand, it's angle and it's size, for each frame of the given video.

2.4 How to Represent the Hand in the Video?

The final module of the system handles the visual rendering of the cueing hand, based on the timing and spatial information computed in the previous stages. This component determines how the hand appears in the video and offers several options in terms of style and visual clarity.

Four handsets were integrated into the system. Two are based on photographs: a male hand set ('l'), and a female hand set ('b') shown in Figure 5.

The other two use 2D illustrations: 'd' displays a uniform yellow shape, while 'c' assigns colors to specific keys to reduce visual confusions—key 3 is pink, key 8 is blue, and the neutral hand is white; all others remain yellow. These assignments build on prior work (Duchnowski et al., 1998) indicating that color can help distinguish keys that are visually similar but phonologically distinct.

Figure 6 shows examples of these handsets, along with enhancement filters described below.

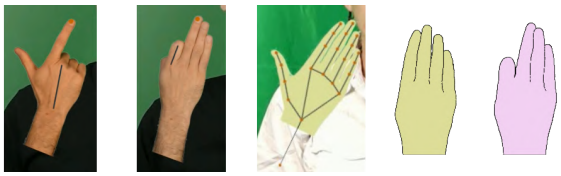

Figure 6: Some hands configurations: "l+1", "l+2", "d+3", "d", "c"

Three visual enhancements were implemented to explore whether additional graphic information could improve the visual distinction between similar handshapes. Each one is exclusive and applies to a single rendering at a time. The first one adds a dot at the fingertip target and a line along the index for keys 3 and 8, to improve distinction from keys 4 and 2, similarly to the 'c' handset. The second one draws a line along the back of the hand and a dot at the target point, highlighting orientation. The 3rd one overlays the full 21-point hand sights with connecting lines, as illustrated in Figure 5.

This rendering module supports both realistic and stylized outputs and can be adapted to user needs or preferences.

2.5 System Summary

The system covers the full pipeline of automatic CS generation. Starting from a video, an audio signal, and an orthographic transcript, it performs phoneme alignment, transformation into keys, synchronization of each key with the speech signal, analysis of facial landmarks, determination of hand angle, hand size, handshape transitions, spatial transitions between positions, and visual rendering.

The process results in a synchronized and augmented video, where a virtual hand encodes the Cued Speech transcription with precise timing and positioning. All elements—phonetic inference, timing models, spatial computation, and graphical output—are integrated into a reproducible framework. This combination of coverage and modularity is, to our knowledge, the first of its kind.

This framework is implemented in Python and released under an open-source license. Its graphical user interface and user-friendly installation process allow non-specialists to use it.

3 System Evaluation

The system was evaluated through a decoding task with eight deaf participants, all fluent in French Cued Speech and familiar with video-based cueing. The goal was to assess the readability of automatically generated cues and to compare different configuration options. The task consisted in watching short cued videos and writing down what was decoded. Their responses were scored using SCLite, designed for evaluating ASR output. It aligns each decoded transcription with a reference using utterance IDs and computes word-level scores: correct (Corr), substituted (Sub), deleted (Del), and inserted (Ins). In this setting, the reference is the recorded sentence, and the hypothesis is the participant's transcription.

Decoding accuracy was then used as a proxy for system performance. This metric was deliberately chosen to reflect the perceptual clarity of the generated cues, independently of participant-specific inference or language comprehension skills. Although comprehension-based tasks might better reflect communicative effectiveness, they would confound system output quality with individual-level interpretation strategies. By focusing on transcription alignment, the evaluation isolates the contribution of the system itself, ensuring a more rigorous and interpretable measure of cue readability.

3.1 Experimental Conditions

The evaluation was conducted during the 2024 annual internship organized by the ALPC. Eight deaf adults participated on a voluntary basis and gave informed consent. All participants watched a standardized instructional video before the session. The protocol was anonymous, non-intrusive, and approved by the organizing institution.

Each participant decoded 20 silent videos: five manually coded by a professional (used as a reference set), and fifteen automatically generated using the system with different configurations. To control for inter-participant variability, each participant was assigned to a single experimental dimension: timing, angle, hand appearance, or visual enhance-

ment. This allowed for within-subject comparisons across three variants per parameter. Each system configuration was identified by a four-character code: the first digit refers to the timing model (2, 3, or 4), the second to the angle model (0, 1, or 2), the third to hand appearance ('b', 'c', or 'd'), and the fourth to optional enhancements (1, 2 or 3). Participants were divided into four groups:

- **Group 1 – Timing:** P1 and P5 decoded sets 2.1.l.0, 3.1.l.0, and 4.1.l.0.
- **Group 2 – Angle:** P2 and P6 decoded sets 4.0.l.0, 4.1.l.0, and 4.2.l.0.
- **Group 3 – Appearance:** P3 and P7 decoded sets 4.1.b.0, 4.1.c.0, and 4.1.d.0.
- **Group 4 – Enhancement:** P4 and P8 decoded sets 4.1.l.1, 4.1.l.2, and 4.1.l.3.

The five manually coded reference videos were presented first. The fifteen system-generated clips followed, in a fixed interleaved order balancing topic and condition. Playback issues affected two participants (three clips for P1, two for P2) due to local hardware errors. Since all videos had been generated beforehand, only playback was affected and the evaluation protocol remained valid. This is reported here in accordance with FAIR principles.

3.2 Global Decoding Performance

Table 3 presents the decoding scores for the control set (professionally coded) and for the system-generated output (all configurations combined). Manual coding achieved 92.3% accuracy. The system, with no participant training or adaptation, reached 80.7%.

SPK	Corr	Sub	Del	Ins	Err
Control	92.3	5.2	2.5	2.3	10.0
All sets	80.7	9.7	9.6	1.3	20.6

Table 3: Participant decoding scores

These results were obtained using strict word-level alignment. Minor spelling differences were counted as substitutions, and no correction was applied to participant input. The control score reflects the best achievable performance under these conditions and serves as an oracle reference.

That the system reaches over 80% under the same constraints is a key finding. Participants had no prior exposure to the system and received no training. Despite this, several decoded videos scored near the reference level. The output is therefore not only intelligible but already close to expert quality for a majority of sentences.

The most frequent errors were deletions, increasing from 2.5% in the control set to 9.6% with system output. Substitutions also rose, though to a lesser extent. Informal debriefings suggest that fast speech segments were harder to decode, especially when hand transitions compressed timing contrasts.

To our knowledge, this is the first publicly documented benchmark comparing professional and system-generated Cued Speech. These results show that automatic cue generation is not only feasible, but already yields intelligible output close to expert performance. This first benchmark sets a high baseline for future systems and provides a reproducible framework for comparison.

The 80.7% score reported above reflects an average across multiple system variants. It includes different timing strategies, spatial models, hand appearances, and visual enhancements. The result therefore combines heterogeneous outputs, some of which led to higher decoding scores than others.

3.3 Detailed evaluation and discussion

The three sentence sets used in the experiment yielded average decoding scores of 83.6%, 84.4%, and 74.9%, respectively, indicating noticeable differences in difficulty. Without normalization, such variation would interfere with the analysis of model-specific effects. To control for these biases, all scores were normalized by participant and by sentence set. This adjustment accounts for individual decoding ability and for intrinsic difficulty of the material. Final results are reported as z-scores: a positive value indicates that the participant decoded better than their own average, and a negative value indicates below-average decoding accuracy.

3.3.1 Group 1 – Timing Models

Participant P1 showed slightly negative performance on the baseline (model 2), and slightly positive scores on models 3 and 4 ($z = -0.07$, $+0.04$, $+0.06$). P5 had the best result on model 2 ($z = +0.08$), followed closely by model 4 ($+0.02$), with model 3 performing lower (-0.04). Overall, model 4 seems less sensitive to speaker or material, while model 3 is more sensitive to speaker or sentence variation.

3.3.2 Group 2 – Angle Models

For P2, model 1 yielded the best performance ($+0.03$), followed by model 2 (-0.04), while model 0 performed neutrally (-0.001). P6 achieved highest scores on models 0 and 1 ($+0.07$ and $+0.06$), with lower performance on model 2 (-0.03). The results suggest that moderate expert-defined angle variation (model 1) provides a good compromise between visual consistency and realism, while corpus-derived angles (model 2) may introduce instability.

3.3.3 Group 3 – Hand Appearance

P3 had a slight preference for the 'd' design ($+0.01$), with lower results on the 'b' and 'c' designs (-0.09, -0.03). P7 also favored 'd' ($+0.12$), followed by 'b' ($+0.05$), and had a neutral response to 'c' (-0.01). Unlike earlier findings reported in (Duchnowski et al., 1998), our results do not replicate a consistent benefit from color coding: one participant improved with the 'c' design, while another performed better without it. These observed trends confirm that the simplified, high-contrast 'd' illustrations enhance decoding performance, likely due to their visual clarity and reduced ambiguity.

3.3.4 Group 4 – Visual Enhancements

P4 showed balanced performance across the three enhancement types (z-scores ranging from 0.0 to +0.04), while P8 experienced a sharp decline, particularly on Skeleton (-0.19). These results suggest that while visual enhancements may assist some users, they may also introduce distracting or overly complex visual elements, especially for less experienced decoders.

3.3.5 Discussion

The experimental results converge on a configuration that *favors clarity over realism*. The most effective combination includes a fixed anticipation model refined by phonetic context (Model 4), expert-defined orientation values (Model 1), and a stylized 2D design with strong visual contrast ('d'). This setup does not aim to reproduce natural hand movement but rather to enhance cue discriminability. It consistently produced the best decoding scores across participants and conditions. Visual enhancements overlays did not improve performance and occasionally introduced confusion, suggesting that additional graphic elements may interfere with the perception of essential features. These findings support the adoption of a simplified, controlled rendering strategy as the system's default configuration for future use.

These results highlight that controlled visual simplicity can effectively outperform realism by enhancing usability and reducing cognitive load in accessibility-focused systems.

4 Conclusion

Despite the documented benefits of Cued Speech for phonological awareness and literacy, no operational system has yet addressed its automatic generation in a reproducible and open manner. The only prior effort explicitly targeting cue generation in video, developed at MIT in the late 1990s, remains undocumented, non-reproducible, and is no longer maintained.

This paper presents the first functional and publicly available system for automatic Cued Speech generation. It targets French and implements a modular pipeline structured into four components: determining what to code, when to display, where to place, and how to render. Each step is formally defined and operational, from phoneme alignment to video rendering with an integrated virtual hand. The system provides explicit control over linguistic content, synchronization, spatial placement, and visual output.

Evaluation with deaf participants confirmed that the output is readable and effective: decoding accuracy averaged 80.7%, compared to 92.3% for professionally coded videos. This result was obtained without participant training or adaptation. Among the tested parameters, hand appearance had the strongest impact. The highest scores were obtained with a stylized 2D design, limited angle variation, and no visual enhancement. These findings indicate that intelligibility benefits from simplification rather than natural imitation.

This work defines a complete and reproducible framework for Cued Speech generation in video. Moreover, it provides a usable tool with a graphical interface, ready for practical use and offering a reference baseline for future systems. The system is already integrated into the actively maintained software platform SPPAS, and has been successfully used by non-technical users in applied settings.

The next step will involve inserting transitional frames when needed, to reduce deletion errors and improve comfort. The goal is to better match the rhythm of the speaker with the decoding strategies used by human coders.

Limitations

This study presents the first fully documented and reproducible system for automatic CS generation. However, several limitations must be acknowledged.

First, the system has been implemented and evaluated only for French. While the architecture is designed to support multiple languages, further work is needed to confirm its adaptability to different phonological inventories and cueing conventions. This is currently being addressed through the ongoing adaptation of the system to American English.

Second, although the evaluation protocol was carefully designed, the number of participants remains limited. This constraint, inherent to the difficulty of recruiting expert Cued Speech users, may affect the generalizability of some findings.

Third, while the current design provides transparency and control, it may miss fine-grained variations observed in natural cueing. To address this, a follow-up project has been launched to explore targeted data-driven modeling, restricted to cases where statistical learning is justified — in line with principles of ecological minimalism and methodological necessity.

Finally, two aspects of the system have been fixed *a priori* and remain to be systematically evaluated: the precise spatial placement of hand positions around the face, and the trajectory modeling, which currently assumes straight-line motion at constant speed. While hand cue positions are algorithmically inferred from facial landmarks, we acknowledge that systematic validation against manual annotations remains limited due to the complexity of recruiting trained evaluators. Preliminary cross-checks on held-out data indicate promising consistency, and ongoing work is extending this evaluation as resources permit.

Ethical Considerations

This study did not involve the collection of any sensitive or identifying information. Participation was voluntary, based on informed consent, and fully anonymous. Participants were not evaluated; rather, their responses served solely to assess the intelligibility of the system's outputs.

The experiment followed the principles of the ALPC association's internal ethics charter, which promotes respect, autonomy, and non-discrimination in all interactions with deaf participants and their families.

Acknowledgements

This research was supported by FIRAH (Fondation Internationale de la Recherche Appliquée sur le Handicap), project APa2022_022[5]. The author gratefully acknowledges FIRAH's trust and commitment to supporting inclusive, applied research.

Warm thanks also go to the members of the French ALPC association for their generous collaboration, their trust, and their invaluable insights throughout the development of this work.

References

N. Aboutabit. 2007. *Reconnaissance de la Langue Française Parlée Complété (LPC): décodage phonétique des gestes main-lèvres.* Ph.D. thesis, Institut National Polytechnique de Grenoble - INPG.

V. Attina. 2005. *La Langue Française Parlée Complétée: Production et Perception.* Ph.D. thesis, Institut National Polytechnique de Grenoble - INPG.

B. Bigi. 2014. A multilingual text normalization approach. *Human Language Technology Challenges for Computer Science and Linguistics, LNAI 8387*, pages 515–526.

B. Bigi. 2015. SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, 111–112:54–69.

B. Bigi. 2016. A phonetization approach for the forced-alignment task in SPPAS. *Human Language Technology. Challenges for Computer Science and Linguistics, LNAI 9561*, pages 515–526.

B. Bigi. 2023. An analysis of produced versus predicted french cued speech keys. In *10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, ISBN: 978-83-232-4176-8*, pages 24–28, Poznań, Poland.

B. Bigi and C. Meunier. 2018. Automatic segmentation of spontaneous speech. *Revista de Estudos da Linguagem. International Thematic Issue: Speech Segmentation*, 26(4).

B. Bigi, A-S. Oyelere, and B. Caron. 2021. Resources for automated speech segmentation of the african language naija (nigerian pidgin). *Human Language Technology. Challenges for Computer Science and Linguistics, LNAI 12598*, pages 164–173.

B. Bigi, M. Zimmermann, and C. André. 2022. CLeLfPC: a Large Open Multi-Speaker Corpus of French Cued Speech. In *Proceedings of The 13th Language Resources and Evaluation Conference*, pages 987–994, Marseille, France.

[5] https://www.firah.org/fr/appel-a-projets-general-2022.html

M. S. Bratakos. 1995. *The effect of imperfect cues on the reception of cued speech*. Ph.D. thesis, Massachusetts Institute of Technology.

M. S. Bratakos, P. Duchnowski, and L. D. Braida. 1998. Toward the automatic generation of cued speech. *Cued Speech Journal*, 6:1–37.

M.-A. Cathiard, V. Attina, and D. Alloatti. 2003. Labial anticipation behavior during speech with and without cued speech. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1935–1938, Barcelona, Spain.

B. R. Clarke and D. Ling. 1976. The effects of using cued speech: A follow-up study. *Volta Review*, 78(1):23–34.

S. Colin, J. Ecalle, E. Truy, G. Lina-Granade, and A. Magnan. 2017. Effect of age at cochlear implantation and at exposure to cued speech on literacy skills in deaf children. *Research in developmental disabilities*, 71:61–69.

R. O. Cornett. 1967. Cued speech. *American annals of the deaf*, pages 3–13.

R. O. Cornett, R. Beadles, and B. Wilson. 1977. Automatic cued speech. In *Research Conference on Speech Processing Aids for the Deaf*, pages 224–239, Gallaudet College (USA).

P. Duchnowski, L. D. Braida, D. Lum, M. Sexton, J. Krause, and S. Banthia. 1998. Automatic generation of cued speech for the deaf: status and outlook. In *International Conference on Auditory-Visual Speech Processing*, Sydney, Australia.

P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida. 2000. Development of speechreading supplements based on automatic speech recognition. *IEEE transactions on biomedical engineering*, 47(4):487–496.

C. G Fisher. 1968. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804.

M. Lancien, M-H. Côté, and B. Bigi. 2020. Developing resources for automated speech processing of quebec french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5323–5328, Marseille, France. European Language Resources Association.

C. J. LaSasso, K. L. Crain, and J. Leybaert. 2010. *Cued speech and cued language development for deaf and hard of hearing children*. Plural Publishing.

J. Leybaert and B. Charlier. 1996. Visual speech in the head: The effect of cued-speech on rhyming, remembering, and spelling. *The Journal of Deaf Studies and Deaf Education*, 1(4):234–248.

D. W. Massaro and Stephen E. Palmer Jr. 1998. *Perceiving talking faces: From speech perception to a behavioral principle*. Mit Press.

N. A Neef and B. A. Iwata. 1985. The development of generative lipreading skills in deaf persons using cued speech training. *Analysis and intervention in developmental disabilities*, 5(4):289–305.

M. Pakrashi, B. Bigi, and S. Mahanta. 2023. Resources creation of bengali for sppas. In *10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, ISBN: 978-83-232-4176-8*, pages 218–222, Poznań, Poland.

J. Rönnberg. 1995. Perceptual compensation in the deaf and blind: Myth or reality? *Compensating for psychological deficits and declines: Managing losses and promoting gains*, pages 251–274.

J. Rönnberg, S. Samuelsson, B. Lyxell, R Campbell, B Dodd, and D Burnham. 1998. Conceptual constraints in sentence-based lipreading in the hearing-impaired. *The psychology of speechreading and auditory–visual speech*, pages 143–153.

M. G. Sexton. 1997. *A video display system for an automatic cue generator*. Ph.D. thesis, Massachusetts Institute of Technology.

A Reproducibility

All data and source code referenced in this paper comply with the principles of open science. The source code of the proposed system is released under the GNU Affero General Public License v3 (AGPLv3). It is part of SPPAS and can be downloaded at https://sourceforge.net/projects/sppas/.

The experimental scripts are also made available under the same license and can be obtained from the author upon request.

The datasets used in this work are distributed under both the Open Database License v1.0 (ODbL) and the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) licenses. They can be downloaded at https://hdl.handle.net/11403/clelfpc/v10.

Software and Evaluation Tools:

- The full speech segmentation pipeline, including text normalization, phonetic transcription, and alignment, was performed using SPPAS, version 4.11 (https://sppas.org/),
- Evaluation metrics were computed using SCTK 2.4.12 (https://github.com/usnistgov/SCTK).