French Cued Speech rhythm: first findings on the relationship between hand position and segments' duration

Mélanie Lancien, Brigitte Bigi

French Cued Speech rhythm: first findings on the relationship between hand position and segments' duration

Mélanie Lancien*, Brigitte Bigi†

* ATILF, Université de Lorraine, CNRS, UMR7118, 44 Av. de la Libération, 54000 Nancy, France.
†LPL, CNRS, Aix Marseille Université, UMR7309, 5 Av. Pasteur, 13100 Aix-en-Provence, France.
*melanie.lancien@univ-lorraine.fr, †brigitte.bigi@cnrs.fr

Abstract

Cued Speech (CS) is a visual system that clarifies spoken language by combining lipreading with hand gestures encoding phonological segments. Each CS key, roughly corresponding to a CV syllable, combines a hand shape (consonant) and a spatial position on the face (vowel); these associations are universal across languages. In French CS, 21 consonants use eight hand shapes and 14 vowels use five positions. Despite its potential, CS remains understudied by linguists, leaving questions about its use, its structure, and the speech-gesture synchronization. This paper begin to explore these relationships by focusing on CS rhythm, primarily through the study of spoken vowels and syllables durations and their link to the hand's target position on the face, hand's shape, type of syllable, and type of reading activity. The experiment uses annotated video data from the CLeLfPC corpus, comprising over 3,700 keys produced by five experienced users. A statistical analysis via generalized mixed-effects models revealed that Hand Target Position significantly affected vowel and syllable duration ($p < 0.005$, $R^2 = 9\%$ and 6%). These findings contribute to a better understanding of the spatiotemporal organization of LfPC and may inform future models of gesture planning and cue synthesis.

Keywords: Cued Speech, hand position, syllable, speech-gesture synchronization

1. Introduction

Cued Speech (CS hereafter) is a visual communication system developed by R. Orin Cornett in 1966 to improve speech perception/comprehension in deaf and hard-of-hearing individuals by disambiguating speech sounds for which the visible part of the articulation looks similar (Cornett, 1967). Although lipreading provides visual access to some speech articulation information, it is very limited and carry phonetic ambiguities (for instance [y] and [u] are not distinguishable by lipreading only). Cued Speech addresses this limitation by using hand shapes and placements near and/or on the face to encode consonants and vowels, respectively, thereby making visually similar speech sounds more distinguishable. This *bimodal encoding* allows each phoneme to be visually distinct, thus significantly improving access to the segmental structure of spoken language. Since CS operates at the phonemic level, each language requires a dedicated set of hand shapes and positions, derived from its phoneme inventory and following the original guidelines set by the system's creator.

Cued Speech has been adapted to over 65 languages to match their specific phonological systems. In French, it is known as *Langue française Parlée Complétée* (LfPC), or "Completed Spoken French Language", and encode each CV (consonant-vowel) syllable as a single visual unit, called a key, formed by the combination of lip movements and a specific hand shape+position compound. More complex syllables, such as CCV or VC, are represented by sequences of successive keys (i.e. using a "zero vowel" hand position to signify a single consonant or the first consonant of a cluster). This system allows all syllables of spoken French to be perceived visually in a clear and systematic way.

In practice, producing CS requires the hand to simultaneously adopt the expected shape and reach the correct position near/on the face. Each of these configurations corresponds to a phoneme, and together with the lip movement, they form a complete key. The hand must arrive in time and with the correct form and placement for the key to be clearly understood. This creates a constraint of temporal coordination between speech articulation and hand gesture.

Among the few studies on cued speech, most have shown that lips and hands are not precisely synchronized. The hand generally begins its motion before the corresponding acoustic event, a phenomenon known as anticipatory movement. This illustrates the complexity of Cued Speech as a multimodal communication code involving tightly timed audiovisual cues. Previous studies have explored this asynchrony from several perspectives. For instance, work by (Attina, 2005) and (Aboutabit, 2007) proposed synchronization models based on isolated CV syllables, describing the typical timing relationships between the acoustic onset of a syllable and the moments when the hand begins to move and reaches its target position.

This temporal variability in hand movement raises a fundamental question: is there a systematic correlation between the spatial characteristics of the movement, specifically the target position the hand must reach, and the duration of the key? In other words, does the final destination of the hand influence the time the movement lasts? Such a correlation would suggest that motor planning in LfPC is not only driven by phonological constraints but also by biomechanical or spatiotemporal constraints. Investigating this relationship is crucial for better understanding the gestural dynamics of CS and could inform future efforts in automatic cue generation and improved training protocols for learners and instructors.

This study addresses that question by analyzing the duration of keys in relation to their target position, using data

from the open CLeLfPC corpus (Bigi et al., 2022). The analysis focuses on productions by experienced cuers under controlled conditions, and aims to quantify this relationship using the theoretical spatial target and temporal measurements. Our methodology builds on existing annotation protocols and leverages multimodal data. The findings contribute to a better understanding of the articulatory organization of Cued Speech and open new perspectives for modeling its motor implementation.

2. French Cued Speech

As with all CS systems, LfPC encodes spoken syllables using a combination of lip movements and manual cues, where each key is defined by a specific hand shape and a spatial position near/on the face. The hand shape inventory consists of eight distinct forms, each associated with a subset of French consonants (see Figure 1). A ninth hand shape, referred to as the neutral shape, is also used in contexts where no consonant is present or during silences. As for the vowels, they are encoded by five distinct hand placements around the face (see Figure 2). An additional neutral position is used when the hand is at rest, outside of active cueing.

To facilitate corpus annotation and automatic processing, each hand shape and position is assigned a symbolic label, following conventions from prior work on automatic key generation for LfPC (Bigi, 2023). Hand shapes are numbered from (1) to (8), as shown in Figure 1. Additionally, (0) represents the neutral shape used in the absence of a consonant or during silences. Positions are labeled using lowercase letters: (s) for side, (m) for mouth, (c) for chin, (t) for throat, and (b) for cheekbone, as shown in Figure 2. The neutral position, labeled (n), corresponds to the hand at rest, outside of active cueing.

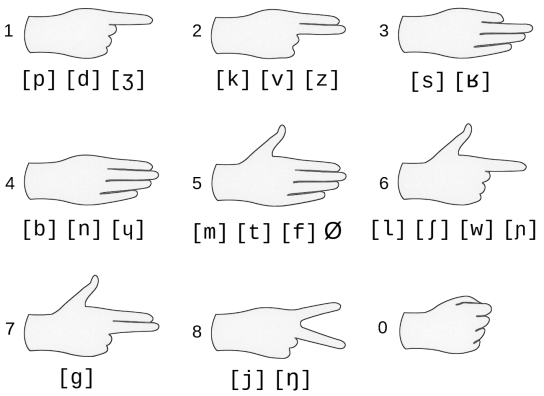

Figure 1: Hand shapes representing consonants

3. Corpus and Annotations

3.1. The CLeLfPC corpus

The present study is based on CLeLfPC (Corpus de Lecture en Langue française Parlée Complétée), a large open-access multimodal dataset of French Cued Speech. This corpus contains high-quality audio and video recordings of 23 participants, recorded under controlled conditions and designed to cover a wide range of keys and key

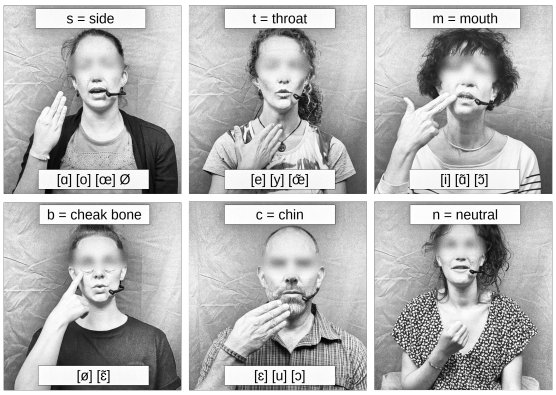

Figure 2: Positions representing vowels

sequences. The speakers vary in experience and training, with some certified in LfPC and others less experienced.

Each participant read on a predefined topic containing four sessions: 32 CV syllables, 32 words or phrases, 7 to 10 isolated sentences, and one short text. For each participants, recordings were made under two distinct cueing instructions:

- a pedagogical condition (syllables and words/phrases reading), where participants were asked to produce cues as clearly as possible,
- a natural condition (sentences and texts reading), where they were instructed to cue as they would in spontaneous communication.

The data used in this study come from a subset of this corpus made of the material from five speakers who achieved high coding proficiency.

3.2. Corpus processing

In order to perform the analysis, the audio was segmented into speech segments (Bigi and Priego-Valverde, 2019), orthographic transcriptions were done, normalized and converted into phonemes, then aligned with the audio signal using SPPAS (Bigi, 2015). All alignments were manually verified and corrected as necessary. Based on this first layer, a phoneme segmentation was first generated semi-automatically. Syllables were then automatically generated from the time-aligned phonemes (Bigi et al., 2010). Cued Speech keys were also derived from the phonetic sequences using the automatic key generation module of SPPAS (Bigi, 2025). These automatically predicted keys were manually reviewed and corrected to match the cues actually produced, based on video observation. This correction step ensures that the annotation represents the performed key rather than the theoretical one.

Hand position transitions were annotated entirely manually by reviewing the front-facing video frame by frame. Each movement toward a vowel target position was annotated with two temporal landmarks: the onset of the transition and the moment when the hand reached and stabilized at the target position. This annotation thus provides access both to the timing of transitions and to the exposure intervals, defined as the duration during which the

hand remained stable in its final position. For each annotated key (i.e., handshape–position pair), the duration was computed from the temporal boundaries of the associated phoneme(s). For CV structures, this corresponds to the interval between the onset of the consonant and the end of the vowel.

3.3. Statistical analysis method

The data analysis relies on the two acoustics measures: 1/ the duration of the acoustic vowel, and 2/ the duration of the acoustic (C)(C)(C)V(C)(C) syllable. Due to the distribution type of these dependent variables, a generalized linear mixed model with a log distribution family was needed (R Core Team, 2021; Venables and Ripley, 2002).

The model incorporated the following four independent variables:

- *HandShape*: a 8-level variable like illustrated in Figure 1
- *HandTargetPosition*: the theoretical placement of the hand with 6-levels like illustrated in Figure 2[1]
- *SyllableStructure*: a 8-level variable describing the phonological structure of the spoken syllable, i.e., CCCV, CCVC, CCV, CVCC, CVC, CV, VC, V
- *ActivityType*: a 2-level variable indicating the type of reading, i.e., sentence or text reading

SpeakerID, the anonymized code given to each recorded speaker, was added as a random effect to account for inter-speaker variability.

Hence, the two generalized linear mixed models were built as follow:

$$Duration \sim HandShape + HandTargetPosition + SyllableStructure + ActivityType + (1|SpeakerID)$$

Due to the natural repartition of phones in French, some hand shapes were very marginal, namely shapes 7 and 8, corresponding to [g] and [j,ŋ] respectively. To ensure reliable computation of means and analysis of variation, these data were excluded from the dataset analyzed in the following section. This subset of the corpus being quite small, interactions between those variables could not be included in the model. Table 1 shows an overview of the dataset repartition among speakers and type of syllable structure. The first line of each cell is the raw number of occurrence, the second is the line percentage and the third is the column percentage.

Struct.	Speaker					Total
	AM	CH	LM	ML	VT	
CCCV	1 8.3 % 0.4 %	4 33.3 % 1.7 %	1 8.3 % 0.3 %	5 41.7 % 1.5 %	1 8.3 % 0.3 %	12 100 % 0.8 %
CCVC	9 20.5 % 3.7 %	11 25 % 4.6 %	14 31.8 % 4.8 %	2 4.5 % 0.6 %	8 18.2 % 2.7 %	44 100 % 3.1 %
CCV	33 14.9 % 13.4 %	41 18.6 % 17 %	42 19 % 14.3 %	56 25.3 % 16.8 %	49 22.2 % 16.3 %	221 100 % 15.6 %
CVCC	6 40 % 2.4 %	0 0 % 0 %	4 26.7 % 1.4 %	4 26.7 % 1.2 %	1 6.7 % 0.3 %	15 100 % 1.1 %
CVC	47 18.4 % 19.1 %	47 18.4 % 19.5 %	59 23 % 20.1 %	55 21.5 % 16.5 %	48 18.8 % 16 %	256 100 % 18.1 %
CV	128 18.2 % 52 %	122 17.4 % 50.6 %	139 19.8 % 47.3 %	158 22.5 % 47.4 %	156 22.2 % 52 %	703 100 % 49.7 %
VC	3 12 % 1.2 %	0 0 % 0 %	5 20 % 1.7 %	6 24 % 1.8 %	11 44 % 3.7 %	25 100 % 1.8 %
V	19 13.8 % 7.7 %	16 11.6 % 6.6 %	30 21.7 % 10.2 %	47 34.1 % 14.1 %	26 18.8 % 8.7 %	138 100 % 9.8 %
Total	246 17.4 % 100 %	241 17 % 100 %	294 20.8 % 100 %	333 23.6 % 100 %	300 21.2 % 100 %	1414 100 % 100 %

Table 1: Data repartition among speakers and type of syllable structure, in sentence and text readings

4. Analysis

Both models yielded low AIC values with -957.9171 for syllable duration and -2069.324 for vowel duration, which indicates a good model fit. The conditional R^2 values (Stoffel et al., 2021) were approximately 0.9, suggesting that the five independent variables' effects accounted for around 90% of the variance in duration. Only the marginal R^2 is reported in the following sections, as it reflects the proportion of variance explained by the fixed effects alone, i.e., the four non-random independent variables, unlike the conditional R^2 which includes both fixed and random effects.

The significance of each independent variable was assessed via type 3 Anovas computed on the models (Fox and Weisberg, 2019). Then, post-hoc tests (Tukey HSD) (Lenth, 2025) and partial R^2 values were computed to get a finer picture of the dynamics at play. Note that all results plots illustrating durations variance were generated using raw data and not models' predictions to ease the readers' interpretation.

4.1. Hand Position effect on Syllable Duration

Only syllables of the types V and CV were included in this analysis, as they involve the realization of a single CS key. In contrast, complex syllable structures (e.g., CVC, CCV, etc) require multiple successive keys, making direct duration comparisons across types methodologically inconsistent. The restriction to single-key syllables ensures that measured durations correspond to a single cue — so that a single shape and a single reached position, allowing for more reliable interpretation of timing effects. A total of 417 such syllables were retained for this analysis.

The model on syllable duration demonstrates that *HandTargetPosition*, *SyllableStructure* and *ActivityType* all had a significant effect on the syllable duration ($p < 0.004$). Marginal R^2 measures indicate that those variables explain about 75% of the variance in duration. The syllable structure explained 59% of the 75%, which is expected and follows the same pattern as most results on speech syllables. *HandTargetPosition* explained 6% of the observed variance and *ActivityType* 2%.

Post hoc tests showed that the effect of *HandTargetPosition* only lies in the difference between the side and the throat position — "s", and "t" on Figure 3, with an average 20 milliseconds shorter for side ($p < 0.05$). As for the

[1]Further exploration is already being done with $[x, y]$ coordinates of the actual hand positions.

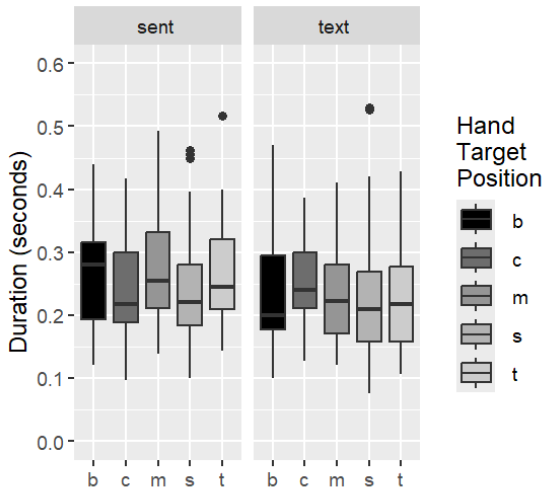

Figure 3: Dispersion of the raw syllable duration (seconds) for the sentence (”sent” on the left) and text reading (”text” on the right), for each hand position (shade of gray), only for V and CV structures.

ActivityType difference, overall the sentence reading task showed syllable durations about 20ms longer ($p < 0.004$) — ”sent” and ”text” on Figure 3. Eventually, for *SyllableStrucutre* V is 80ms shorter than CV ($p < 0.0001$)

4.2. Hand Position effect on Vowel Duration

The model on vowel duration demonstrates that *HandShape*, *SyllableStructure* and *HandTargetPosition* all have a significant effect on the vowel duration ($p < 0.006$). Marginal R^2 measures indicate that those variables explain about 35% of the variance, 9% of the variance is explained by the *HandTargetPosition*, 10% by the *SyllableStructure* and 10% by the *HandShape*.

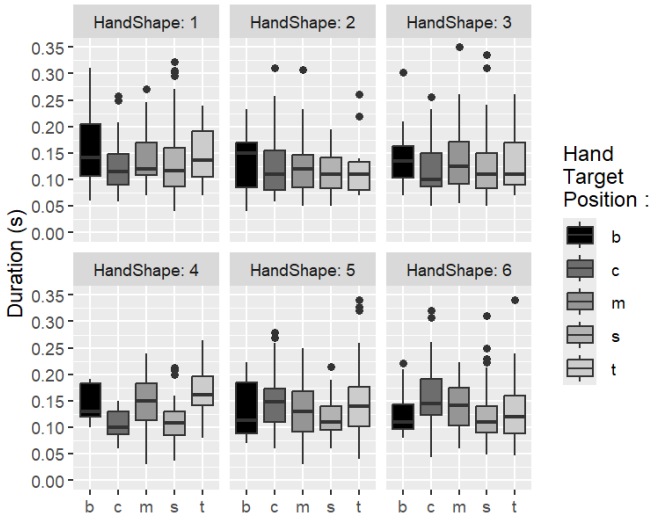

Figure 4: Dispersion of the raw vowel duration (seconds) for the each hand shape (grid), for each hand position (shade of gray).

Post hoc tests showed that the effect of *HandShape* lies in the difference between, on one side hand shape 6 — [l,w,ʃ,ɲ], and on the other side hand shapes 3 and 5 — [s, ʁ] and [m,t,f,∅], ($p < 0.038$). Marginally, positions 3 and 4 ([b,n,ɥ]) also seam to differ (p=0.071). As for *HandTargetPosition*, only ”s” (side) contrasts with ”t” and ”b” (throat and cheek bone), side showing shorter durations ($p < 0.018$).

Eventually, the difference regarding syllables structures mostly concerns CCV vs. CV and CVC, CCV syllables being shorter by an average of 130ms and 150ms respectively ($p < 0.01$).

4.3. Results interpretation

The effect of *HandTargetPosition* on both vowel and syllable level, particularly for the “side” and “throat” positions. The observed differences follows intuitive expectations as one might assume that transitions toward the side would allow faster articulation — given the absence of physical hand-face contact and thus potentially quicker shape transitions — the data do systematically support this assumption.

An ambiguity arises when interpreting the effect of *HandShape* on vowel length. Hand shape 6 which is used for [l, w, ʃ, ɲ], tends to be associated with slightly longer vowel durations (mean = 134 ms). In comparison, shapes 3 and 5, used respectively for [s, ʁ] and [m, t, f, ∅], show mean durations of 129 ms and 134 ms. The scale of these differences remains small. Given this limited contrast, part of this effect may result of alignment imprecision introduced by the automatic speech segmentation used for data annotation.

The effect of *ActivityType* might be the result of a difference in the instruction given to the speakers. For the sentence reading, they had to go back to a neutral position, with their hand on the chest, between every sentence. Thus, syllables at the beginning and at the end of the sentences might have a longer production time, biasing the data towards longer segments in sentence reading.

Finally, the effect of the syllable structure can be accounted for by the number of phonemes and their position in the key. Syllables made of 1 phoneme, i.e., the vowel, or syllables beginning by consonant clusters, i.e., CCV, show shorter durations. The observed shortening of CCV syllables — by approximately 130 to 150 milliseconds compared to CV and CVC structures — suggests that the presence of an initial consonant cluster reduces the temporal space available for the vowel. One plausible interpretation is that the second consonant in the onset cluster requires articulatory realization within the same initial key, thereby constraining the time allocation for the vocalic nucleus. As a result, the vowel may be shortened to accommodate the full cluster within a limited temporal window.

5. Conclusion

Cued Speech (CS) enhances visual speech perception by combining lipreading with hand gestures that encode phonological information. Each key is defined by a specific hand shape and a target position near/on the face, corresponding, respectively, to consonants and vowels.

This study investigated the relationship between the spatial characteristics of cueing gestures and speech temporal properties, focusing specifically on whether the final target position influences vowel and/or syllable duration. A detailed analysis of durations distributions revealed that *HandTargetPosition* significantly affects durations, particularly between the "side" and "throat" positions. This paper supports the assumption that transitions toward the side of the face allow faster articulation thanks to the absence of physical hand-face contact leading to quicker hand shape transitions.

These findings lay the groundwork for assessing potential correlations between gesture, speech duration and spatial targets, and contribute to a better understanding of the time constraints and motor organization underlying French CS production.

6. Reproducibility

Both data and source code referenced in this paper comply with the principles of open science. The experimental script source code is released under the GNU Affero General Public License v3 (AGPLv3) 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`.

7. References

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

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

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

Bigi, B, 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*. Poznań, Poland.

Bigi, B, 2025. Bridging the gap: Design and evaluation of an automated system for french cued speech. In *International Conference on Natural Language and Speech Processing*. Odense, Danemark.

Bigi, B, C Meunier, I Nesterenko, and R Bertrand, 2010. Automatic detection of syllable boundaries in spontaneous speech. In *Language Resource and Evaluation Conference*. La Valetta, Malta.

Bigi, B and B Priego-Valverde, 2019. Search for interpausal units: application to cheese! corpus. In *9th Language & Technology Conference*. Poznań, Poland.

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

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

Fox, J and S Weisberg, 2019. *An R Companion to Applied Regression*. Thousand Oaks CA: Sage, 3rd edition.

Lenth, Russell V., 2025. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.11.2-8.

R Core Team, 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stoffel, Martin A., Shinichi Nakagawa, and Holger Schielzeth, 2021. partr2: Partitioning r2 in generalized linear mixed models. *PeerJ*.

Venables, W. N. and B. D. Ripley, 2002. *Modern Applied Statistics with S*. New York: Springer, 4th edition. ISBN 0-387-95457-0.