An analysis of produced versus predicted French Cued Speech keys

Brigitte Bigi

An analysis of produced versus predicted French Cued Speech keys

Brigitte Bigi

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

Abstract

Cued Speech is a communication system developed for deaf people to complement speechreading at the phonetic level with hands. This visual communication mode uses handshapes in different placements near the face in combination with the mouth movements of speech to make the phonemes of spoken language look different from each other. This paper presents an analysis on produced cues in 5 topics of CLeLfPC, a large corpus of read speech in French with Cued Speech. A phonemes-to-cues automatic system is proposed in order to predict the cue to be produced while speaking. This system is part of SPPAS - the automatic annotation an analysis of speech, an open source software tool. The predicted keys of the automatic system are compared to the produced keys of cuers. The number of inserted, deleted and substituted keys are analyzed. We observed that most of the differences between predicted and produced keys comes from 3 common position's substitutions by some of the cuers.

Keywords: cued speech, corpus, annotation

1. Introduction

The production of speech naturally involves lip movements; the acoustic information as well as the lipreading are part of the phonological representation of hearing people. For a better comprehension every sound of the language should look different but many sounds look alike on the lips when speaking. The term 'viseme' was introduced to refer to mutually confused phonemes that are deemed to form a single perceptual unit (Fisher, 1968; Massaro and Palmer Jr, 1998). In 1966, R. Orin Cornett invented the Cued Speech (Cornett, 1967), a visual system of communication; it adds information about the pronounced sounds that are not visible on the lips. Cued Speech (CS) is a communication system developed for deaf people to complement speech reading at the phonetic level with hands. It uses hand shapes in different placements near the face in combination with the mouth movements of speech to make the phonemes of spoken language look different from each other. Several studies have been conducted on CS to show how it can help speech perception for deaf or hard of hearing persons. It improves speech perception for hearing-impaired people and it offers a complete representation of the phonological system for hearing-impaired people ; among others, see (Nicholls and Mcgill, 1982; Leybaert and Alegria, 2003; Bayard et al., 2019). Cued Speech is then increasingly popular and has been adapted for more than 65 languages[1]. From both the hand position on the face to represent a vowel 'V' and handshapes to represent a consonant 'C', 'CV' syllables are coded. There are named either *keys* or *cues*. A single CV syllable will be generated or decoded through both the lips position and the key of the hand. Each time a speaker pronounces a 'CV' or '-V' syllable, a cue is produced. Other syllabic structures are produced with several cues - for example, a 'CCV' syllable is coded with the two consecutive keys 'C-' then 'CV'. As a consequence, when sounds look alike on the lips, they are cued differently. Thanks to this code, speech reading is encouraged since the Cued Speech keys match all of the spoken phonemes but phonemes with the same viseme have different keys. Once sounds are made visible and look different, it results in a better understanding of speech.

This paper investigates the automation of the production of keys. A rules-based system is proposed and is performed on time-aligned phonemes of CLeLfPC - Corpus de Lecture en Langue française Parlée Complétée (Bigi et al., 2022), a large open source corpus of French Cued Speech. This automatic annotation was manually checked and the differences between the predicted keys and the produced keys are analyzed.

2. French Cued Speech

The modality of cueing provides a level of visual access to deaf and hard-of-hearing people for spoken languages. Because CS fits the phonological level of a given spoken language, each language is cued differently because its CS chart is created from its phonemic representation and it follows the principles of cueing design defined by its inventor (Cornett, 1994).

The French Cued Speech is named "Langue française Parlée Complétée" - LfPC that literally means "Supplemented Spoken French Language". It makes use of the same 8 handshapes (consonants) and 5 hand positions on or around the face (vowels). Table 1 indicates the naming convention of the handshapes and Table 2 the ones of the hand positions. We used the same naming convention as the one of the British CS (BCS), except we propose to name the cheek bone vowel position (b) which does not exist in BCS. In addition, a 9th handshape is identified with (0) and a 6th hand position is identified with (n). They are respectively representing the neutral shape and neutral position. This is

[1] https://www.academieinternationale.org/list-of-cued-languages visited 2022-09

used along with long silences. Figure 1 illustrates both the positions of vowels and the handshapes for all phonemes.

id.	consonants	id.	consonants
(1)	/p/, /d/, /Z/	**(5)**	/m/, /t/, /f/, no consonant
(2)	/k/, /v/, /z/	**(6)**	/l/, /S/, /J/, /w/
(3)	/s/, /R/	**(7)**	/g/
(4)	/b/, /n/, /H/	**(8)**	/j/, /N/

Table 1: Handshapes identifiers and their corresponding consonants in X-SAMPA

id.	vowels	id.	vowels
(s)	/a/, /o/, /9/, /@/, no vowel	**(m)**	/i/, /O∼/, /a∼/
(c)	/E/, /u/, /O/	**(t)**	/y/, /e/, /9∼/
(b)	/e∼/, /2/		

Table 2: Hand position identifiers and their corresponding vowels in X-SAMPA

3. An automatic prediction system for cues

Despite the significant number of studies demonstrating the benefits of Cued Speech, studies on the automatic CS prediction are rather rare. The Massachusetts Institute of Technology has sought to address this problem in its realization of an Automatic Cue Generator (Bratakos, 1995; Sexton, 1997; Bratakos et al., 1998; Duchnowski et al., 1998). In a room, a speaker is filmed speaking without coding and an Automatic Speech Recognizer (ASR) uses the acoustic speech signal to determine which phoneme is being produced. Once the recognition is completed, in another room, the image of the filmed speaker with the synthesis keys according to the rules of the Cued Speech is displayed on a screen to the deaf individual. Several versions of this system were evaluated and it resulted in at least a small benefit to the cue receiver relative to speech-reading alone. However, the way they get the keys is neither fully described nor evaluated separately. Two French projects were also implementing a Text-to-Cued speech synthesizer between 2002 and 2006 but none of them neither described nor distributed the key generator.

In the scope of creating a Text-to-Cued system, the first required *new* step copes with time-aligned phonemes as input and produces an output with the cue names and their corresponding segmentation. Therefore, the problem we are dealing with is close to the syllabification of phoneme sequences we have previously investigated (Bigi et al., 2010). The phoneme sequences need to be automatically converted into key sequences and time-aligned from the corresponding phoneme time-alignments.

At a first stage, we have to create time groups from the time-aligned phonemes. 'Time Group' (TG) refers to an event sequence with a well-defined boundary condition (Gibbon, 2013). In the present context, a TG is an inter-break group where a break is a pause or any sound except a phoneme (laugh, noise, breath, etc).

The structure of CS assumes that a cue represents each CV combination as a handshape (C) and a specified position (V). Each phoneme of TG are then turned into its class: either labelled with C or V.

Given the sequence of class labels of a TG, the algorithm specifies a sequence of handshape-position pairs according to the rules of CS. Special rules are implemented for atypical class combinations such as VC, C, CC and CVC, instead of the regular 'CV' that makes a key. We developed a grammar corresponding to these rules and implemented this grammar in software a deterministic finite automata (DFA). For clarity, we show in Figure 2 the DFA of a single cue. The DFA accepts or rejects an input string of symbols, based on a deterministic algorithm. All states in consideration exist in a finite list and the abstract machine can only take on one of those states at a time.

When the sequence of class labels of a TG is segmented, we turn back the sequence of classes into phonemes. Each phoneme label is then mapped to its key code according either table 1 for a consonant or table 2 for a vowel. It results in a new time-aligned annotation at the CS key level. Figure 3 illustrates an example of such input and output. This automatic process is implemented in a Python package of SPPAS (Bigi, 2015) and distributed under the terms of the GNU GPL v3 license.

4. Dataset: cues annotation

CLeLfPC - Corpus de Lecture en LfPC, is a large open source multi-speaker dataset of Cued Speech (Bigi et al., 2022). It is under the terms of the CC-BY-NC-4.0, the Creative Commons Attribution-Non-Commercial 4.0 International License, and can be used for any research or teaching purpose about CS. The corpus is made of 4 hours of audio/video recordings: it is the largest available corpus of CS data. Among others, this corpus brings the following tangible benefits:

- an HD video quality of the whole speaker;
- 23 different participants, some are CS certified and some are not;
- 10 different topics, each one read by 2 or 3 participants;
- 4 different sessions in each topic: 32 isolated syllables, 32 isolated words or phrases, 7 up to 10 isolated sentences, a text.

Annotations are under construction but some are already available under the terms of the same license. Five different topics read by participants of level 5 (highly experimented) or 6 (CS certified) were annotated. The 4 sessions of all the 5 topics were time-aligned at the phonetic level, following a semi-automatic procedure. Using SPPAS (Bigi, 2015; Bigi and Priego-Valverde, 2019), Inter-Pausal Units - e.g. sounding segments separated by silences, were identified. The orthographic transcription was then performed manually with Praat (Boersma and Weenink, 2018) by the first author of this paper, and the boundaries of the IPUs were manually verified at the same time. The text transcription was automatically normalized and converted to phonemes. The automatic graphemes-to-phonemes conversion results were manually verified then automatically time-aligned with the recording. The resulting time-aligned

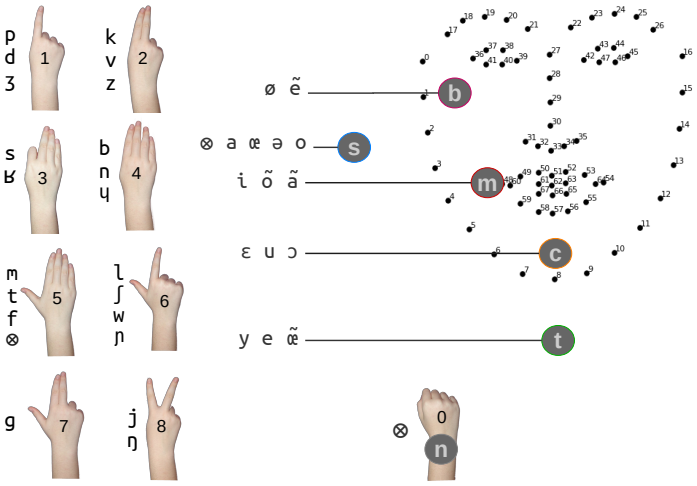

Figure 1: French Cued Speech coding scheme with phonemes in IPA, and a special character to represent "no speech".

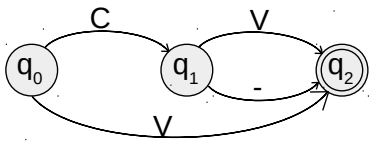

Figure 2: Grammar of a CS key.

phonemes were manually verified with Praat by the first author.

The automatic prediction system for cues was then used in order to get the time-aligned predicted CS keys annotation like illustrated in the first 3 tiers of Figure 3. The videos were viewed in slow motion in order to identify differences between the keys that were predicted by the system and the keys that were coded. It resulted in a new annotation with the time-aligned produced CS keys, represented in the 4th tier of Figure 3. Table 3 indicates the distribution of the 4143 produced keys according to the key structure and session. In addition, 476 neutral handshape and hand position were observed. Table 4 indicates the mean duration

	N	C	V	CV
syllables	165	0	0	159
words	168	187	66	621
sentences	89	309	145	1013
text	54	361	145	1137
total	476	**857**	**356**	**2930**
percent		*20.69%*	*8.59%*	*70.72%*

Table 3: Produced cues in 5 topics of CLeLfPC.

and standard deviation of the produced keys. The 'i1' and 'i2' flags refer to the following reading instructions given to CLeLfPC cuers:

i1 the syllables and the words/phases have to be read clearly, like to teach CS to someone else;

i2 the sentences and the text should be read as naturally as possible, like to tell or read someone a story.

		CV		C	V
		μ	stdev	μ	μ
i1	syllables	0.354	0.105		
i1	words	0.317	0.104	0.182	0.194
i2	sentences	0.268	0.085	0.153	0.170
i2	text	0.250	0.085	0.137	0.154

Table 4: Duration in 5 topics of CLeLfPC.

Perhaps somewhat unsurprisingly, the average duration highlights differences between 'i1' and 'i2'. Duration of 'i2' are about 25% lower than those of 'i1'. It has to be noticed that these are the duration of the phonemes clusterized into cues like illustrated in Figure 3, not the duration of the cues themselves.

5. Predicted versus produced keys

The aims of a comparison between the predicted keys and the produced ones by cuers are twofold. On the one hand, this analysis could reveal implicit rules, i.e. rules of common use that constitute exceptions to the rules of the general definition in order to implement a prediction system closer to the real coding habits. On the other hand, it allows to describe the CS coding as it is practiced, quantifying errors and qualifying them.

We firstly compared the annotations quantitatively. The differences are stated below according 3 categories:

insertion The cuer added 8 keys compared to the predicted ones;

deletion The cuer did not code 47 keys compared to the predicted ones;

substitution The cuer and the prediction system coded 183 keys differently.

The number of inserted and deleted keys is very small relatively to the number of substitutions, and almost anecdotal relatively to the corpus size.

Table 5 shows the details of such differences for each one of the 5 speakers. We can observe that for two of them (AM,

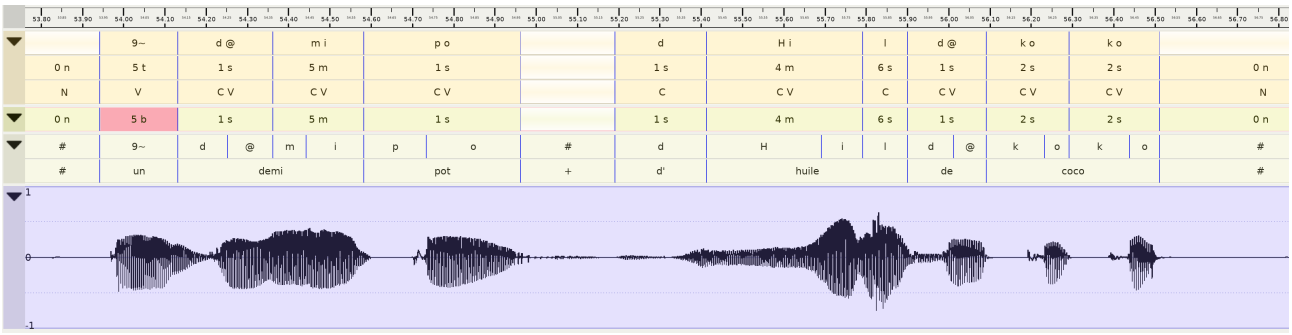

Figure 3: Annotations and waveform of two TG extracted from CLeLfPC. From bottom to top: tokens, phonemes, manually checked CS code, automatic CS keys structure, automatic CS code, automatic CS keys.

ML) there's only a few number of differences, which means that the predicted system and these speakers are consistent in their key production. A detailed analyses of the difference will give some clues to understand in which specific situations the other three speakers are coding differently.

speaker:	CH	VT	AM	ML	LM
insertion	1	4	2	0	1
deletion	16	2	7	4	18
substitution	35	74	12	6	56

Table 5: Produced keys that don't match the predicted ones, depending on the cuer

5.1. Insertion

Among the 8 inserted keys, three are errors of the cuer but five are related the liaison phenomenon. For example, the tokens "pour un" (*for a*) is pronounced /puR9∼/ then the automatic system predicts a sequence of two keys: /pu/ and /R9∼/. However, the cuer is coding a sequence of three keys corresponding to: /pu/ then /R/ then /9∼/. In this case, both coding solutions are acceptable, but this situation is very rare, so it does not need to be taken into account into the prediction system.

5.2. Deletion

Among the 47 keys the cuer did not code compared to the predicted ones, 41 are 'C' and 6 are 'CV'. So, isolated vowels are always coded which is not surprising given that they are the nucleus of syllables. Only 3 of the un-coded sounds are related to the instruction 'i1', so the high majority were from sentences and text. The removed 'C' keys are during 0.102 seconds in average which represents 67% of the average duration of the coded ones. However, 32% of the coded 'C' are during less than 0.102 seconds. As a consequence, we can observe that the un-coded isolated consonants are frequently short but it does not make it a rule for a prediction system because the majority of the short isolated 'C' are coded. We can formulate the hypothesis that, sometimes, the cuer has not had enough time to move the hand at the side position with the expected handshape. As shown in Table 5, among the 5 cuers, two are significantly un-coding the consonants: 18 deletion for LM and 16 for CH. The most frequently un-coded consonants are /t/ (9 times), /R/ (8 times), /p/ (5 times) and /l/ (4 times).

5.3. Substitution

Key substitutions are representing 4.4% of the produced keys, so their analysis is important, particularly because it has never been done in previous studies on CS. As shown in Table 5, three cuers (CH, VT, LM) are producing 90% of the substitutions. Like before, we observe an effect of the instruction. None of the substitutions are occurring during the syllable sessions and only 24 are occurring during the word ones. The high majority of substitutions is from sentences (65) and text (94).
Among the 183 substitutions, 16 are 'C' (8.7 %), 22 (12 %) are 'V' and 145 (79.2 %) are 'CV'. Proportionally to their frequency, it seems that substitutions mostly concern the position (the vowel) than the handshape (the consonant). This tendency is confirmed by the following detailed analysis of the predicted 'CV' keys compared to the produced ones. Among the 145 'CV' cued keys that don't match with the predicted ones:

- 1 substitutes both the shape and the position;
- 6 substitute the shape only;
- 138 substitute the position only.

In the end, we observed 160 vowel substitutions among the 183 referenced ones, that is 87.4 % of the substitutions, 3.86 % of the produced cues of the corpus. A position substitution therefore represents the major difference between predicted and produced keys.
A large number of the vowel substitution (88, that is 48 %) concerns the phoneme /@/ which is coded at position (b) instead of (s). The (b) position is the one of the vowel /2/ but /2/ is never coded at (s) position like /@/. When phonetically realized, schwa (/@/) is a mid-central vowel with some rounding. Many authors consider it to be phonetically identical to /2/ (Anderson, 1982). In the internal position, the acoustic analysis carried out in the reading of a list of words demonstrated the quasi-acoustic identity (Racine et al., 2016). The differences with /2/ are that schwa duration is reduced or that it can be omitted. Such reduction of schwa in French highly depends on the accent: schwa is one of the phenomena that makes it possible to differentiate the northern and southern varieties of French. We observed

that two cuers are significantly coding /@/ at (b) position: VT 45 times and LM 40 times; however VT coded it at (s) position 55 times and LM 45 times like expected by the key rules production. We then sought to understand why they use both solutions (s) and (b), and we found the answer by looking at the words:

- LM: *"de"* is 14 times at (b) against 2 times at (s);
- VT: *"de"* is 11 times at (b) against 5 times at (s);
- VT: *"le"* is 10 times at (b) and never at (s);
- LM: *"le"* is 6 times at (b) against once at (s);
- VT: *"ne"* is 4 times at (b) and never at (s);
- LM: *"que"* is 4 times at (b) and never at (s).

Another significant substitution concerns the vowel /e/ which is coded 32 times at position (c) instead of (t). The (c) position is the one of the vowel /E/. Here again, two speakers are mostly coding this way: 18 times VT and 9 times CH. However, we did not observed any particular trend that could explain this difference in coding. We only found that it affects some words more that others but not systematically. These words are: *c'est* (6 times), *les* (5 times), *ses* (4 times) and *des* (4 times).
The last significant substitution concerns the vowel /9∼/ which is coded at position (b) 17 times instead of (t). The (b) position is the one of the vowel /e∼/. This difference is mainly observed in the word *un* of CH speaker (12 times) who is coding this word only 2 times in (t).

6. Discussion and Conclusion

This paper presented an automatic system to predict CS keys from phonemes. An automatic annotation of cues was performed on 5 topics of CLeLfPC, a large open source corpus of French Cued Speech. This annotation was manually verified to obtain the keys produced by the cuers. An analysis of the differences between the predicted keys and the produced ones allowed to validate the automatic system: this analysis did not reveal implicit rules. Moreover, there is few information on how CS is produced by human coders, so this paper has contributed in this area. This study highlighted some cuer habits. The most significant difference comes from position substitution of some specific phonemes in some specific words by some of the cuers.
Next work will focus on the analysis of duration and timing of the sequences of cues in the time-groups and on the temporal and spatial organization of the code in its speech co-production. It will require to manually re-check the time-alignment of phonemes by an expert phonetician and to time-align the keys with the video in order to annotate the moments they are produced.

References

Anderson, S-R, 1982. The analysis of french shwa: or, how to get something for nothing. *Language*:534–573.

Bayard, C, L Machart, A Strauß, S Gerber, V Aubanel, and J-L Schwartz, 2019. Cued speech enhances speech-in-noise perception. *The Journal of Deaf Studies and Deaf Education*, 24(3):223–233.

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

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 inter-pausal 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.

Boersma, P and D Weenink, 2018. Praat: doing phonetics by computer [computer program], version 6.0.37, retrieved 14 march 2018 from http://www.praat.org/.

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

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

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

Cornett, R-O, 1994. Adapting cued speech to additional languages. *Cued Speech Journal*, 5:19–29.

Duchnowski, P., L.-D. Braida, M.-S. Bratakos, D.-S. Lum, M.-G. Sexton, and J.-C. Krause, 1998. A Speechreading aid based on phonetic ASR. In *5th International Conference on Spoken Language Processing*. Sydney, Australia.

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

Gibbon, D, 2013. TGA: a web tool for Time Group Analysis. In *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*. Aix-en-Provence, France.

Leybaert, J and J Alegria, 1990. Cued speech and the acquisition of reading by deaf children. *Cued Speech Journal*, 4:24–38.

Leybaert, J and J Alegria, 2003. The role of cued speech in language development. *Oxford handbook of deaf studies, language, and education*, 1:261.

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

Nicholls, G-H and D-L Mcgill, 1982. Cued speech and the reception of spoken language. *Journal of Speech, Language, and Hearing Research*, 25(2):262–269.

Racine, I, J Durand, and H N Andreassen, 2016. PFC, codages et représentations: la question du schwa. *Corpus*, 15.

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