

A HMM recognition of consonant-vowel syllables from lip contours: the Cued Speech case.

Noureddine Aboutabit¹, Denis Beautemps¹, Jeanne Clarke¹, Laurent Besacier²

¹ ICP, Département Parole et Cognition de GIPSA-lab, CNRS UMR 5216 / Université Stendhal / INPG, 46 Avenue Félix Viallet, 38031 Grenoble Cedex 1, France.

² Laboratoire d'Informatique de Grenoble, UMR 5217 - 681 rue de la passerelle - BP 72 - 38402 Saint Martin d'Hères

Noureddine.Aboutabit@gipsa-lab.inpg.fr

Abstract

Cued Speech (CS) is a manual code that complements lip-reading to enhance speech perception from visual input. The phonetic translation of CS gestures needs to combine the manual CS information with information from the lips, taking into account the desynchronization delay (Attina et al., 2004 [1], Aboutabit et al., 2006 [2]) between these two flows of information. This paper focuses on HMM recognition of the lip flow for Consonant Vowel (CV) syllables in the French Cued Speech production context. The CV syllables are considered in term of viseme groups that are compatible with the CS system. The HMM modeling is based on parameters derived from both the inner and outer lip contours. The global recognition score of CV syllable reaches 80.3%. This study shows that the errors are mainly observed on consonant groups in the context of high and mid-high rounded vowels ([ɔ̃, y, o, ø, u]). In contrast, CV syllables for anterior non rounded vowels ([a, ɛ̃, i, œ, e, ɛ]) and for low and mid-low rounded vowels ([ã, ɔ, œ]) are well recognized (in average 87%).

Index Terms: Cued Speech, HMM CV syllables recognition, lip modeling

1. Introduction

Cued Speech (CS) (Cornett, 1967 [7]) is a visual communication system that uses handshapes placed in different positions near the face in combination with the natural speech lip-reading to enhance speech perception from visual input. In this system, the speaker moves his or her hand in close relation with speech (see Attina et al., 2004 [1] for a detailed study on CS temporal organization). The hand (with the back facing the perceiver) is a cue that uniquely determines a phoneme when associated with the corresponding lip shape. A manual cue in this system is made up of two components: the shape of the hand and the hand position relative to the face. Handshapes are designed to distinguish among consonants and hand positions among vowels. A single manual cue corresponds to phonemes that can be discriminated with lip shapes, while phonemes with identical lip shapes are coded with different manual cues (see figure 1 which describes the complete system for French). In the framework of communication between hearing and hearing impaired people, the automatic translation of CS components into a phonetic chain is a key issue. Due to the CS system, both hand and lip flows produced by the CS speaker carry a part of the phonetic information. Thus the recovering of the complete phonetic information needs to

constrain the process of each flow by the other one (see Aboutabit et al., 2006 [2] for an example of a complete analysis of the hand flow). This paper focuses on the lip flow for French Consonant-Vowel (CV) syllables, defined as lip parameters extracted from the inner and outer lip contours. More precisely, the CV syllables are considered in term of visemes for a HMM modeling. A grouping into visemes of the vowels has been established in a previous study (see Aboutabit et al., 2006 [3]). For the consonants, many sets of viseme are tested in order to find out the grouping that gives the best recognition rate with the constraint of being compatible with the CS grouping. In addition, the contribution of two new lip parameters relative to the pinching of the upper and lower lips is evaluated.

This exploratory work is part of the TELMA project that aims to translate the CS gestures into phonetic chain (from the merging of manual CS information and lips information).

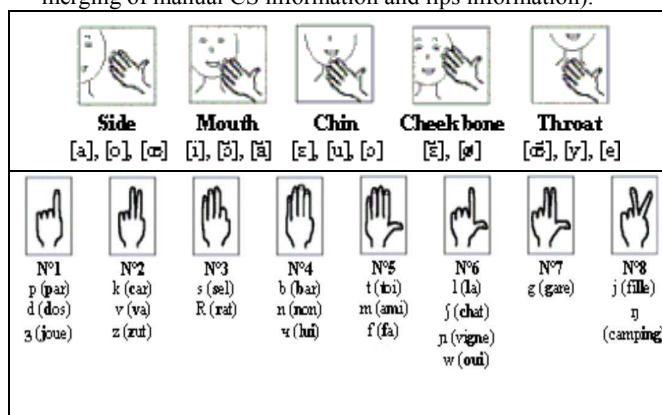


Figure 1: CS Hand position (top) for vowels and CS handshapes (bottom) for consonants (adapted from [1]) in French.

2. Data

The data was obtained from a video recording of a speaker pronouncing and coding in French CS a set of 267 sentences, repeated at least twice.

The French CS speaker is a native female speaker of French, certified in French CS. She regularly translates into French CS code in a school. The recording was made in a sound-proof booth at Institut de la Communication Parlée (ICP), at 50 frames/second for the image video part. The speaker was seated and wore a helmet that served to keep her head in a fixed position and thus in the field of the camera. She wore opaque glasses to protect her eyes against a halogen floodlight. The camera in large focus was used for the hand and the face and was connected to a betacam recorder. The

lips were painted in blue, and blue marks were placed on the speaker's glasses as reference points (Figure 2).



Figure 2: Image of the speaker.

A square paper was recorded for further pixel-to-centimeter conversion. Using ICP's Face-Speech processing system, the audio part of the video recording was digitized at 22,050 Hz in synchrony with the image part, the latter being stored as Bitmap frames every 20 ms. A specific image processing was applied to the Bitmap frames in the lip region to extract the inner and outer contours and to derive the corresponding characteristic parameters (Lallouache, 1991 [4]): lip width (A), lip aperture (B) and lip area (S). These parameters were converted using a pixel-to-centimeter conversion formula. Finally the parameters were low-pass filtered.

The acoustic signal was automatically labeled at the phonetic level using forced alignment (see Lamy, 2004 [5] for a description of the speech recognition tools used for this). Since the orthographic transcription of each sentence was known, a dictionary containing the phonetic transcriptions of all words was used to produce the sequence of phonemes associated with each acoustic signal. This sequence was then aligned with the acoustic signal using French ASR acoustic models trained on the BRAF100 database (Vaufreydaz, 2000 [6]).

This process resulted in a set of temporally coherent signals: the 2D hand position (see Aboutabit, 2006 [2]) the lip width (A), the lip aperture (B) and the lip area (S) values every 20 ms for both inner and outer contours, and the corresponding acoustic signal with the associated phonetic chain temporally marked. In addition, two supplementary parameters relative to lip morphology were extracted: the pinching of the upper (Bsup) and lower (Binf) lips.

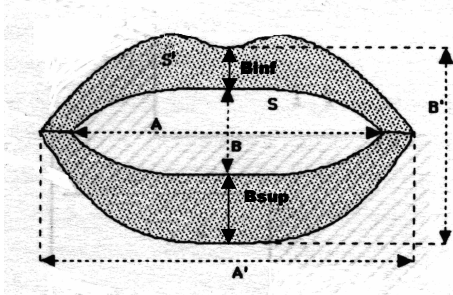


Figure 3: Lip parameters: inner Width (A), inner Aperture (B), inner Area (S), Outer Width (A'), Outer Aperture (B'), Outer Area (S'), upper lip pinching (Bsup) and lower lip pinching (Binf).

3. HMM consonant vowel recognition

From the previous data, a large corpus of CV sequences was constructed with lip parameters. In each sentence, all CV sequences are extracted by using the acoustic labeling. These sequences are then grouped as CV visemes. Indeed, vowels can be sorted into three visemes where each vowel has a similar lip shape (Aboutabit et al., 2006 [3]): high and mid-high rounded (V1= [ɔ̄, y, o, ø, u]), anterior non rounded vowels (V2= [a, ɛ̃, i, œ, e, ɛ]), low and mid-low rounded vowels (V3= [ã, ɔ, œ]). For the consonant case, the grouping based on front lip shapes is not so evident since the production of a large set of consonants is influenced, as well known, by the vocalic context. However, consonants can be grouped into five articulatory classes: bilabial consonants (C1= [p, b, m]), labiodental consonants (C2= [f, v]), dental consonants (C3= [t, d, s, z, n, l]), palatal consonants (C4= [ʃ, ʒ, ɲ]) and velar consonants (C5= [k, g, R]).

3.1. Experiment 1: HMM recognition with modified articulatory consonant grouping

The previous articulatory grouping of consonants is compatible with the CS handshapes grouping except the [ʃ, ʒ, ɲ] group for which [ʃ] and [ɲ] are coded with the same CS handshape. Thus, it seems that adding the [ɲ] consonant to the [k, g, R] group is appropriate. This modified consonant grouping is used with the previous vowel viseme grouping to construct a CV corpus for this first experiment. For each CV group, a three state HMM is trained using the first repetition of the sentences while the second one is devoted to the test data. It results in 2357 CV syllables for the learning phase and 1766 CV syllables for the test phase.

The HMM modeling is based on six lip parameters: inner width (A), inner aperture (B), inner area (S), outer width (A'), outer aperture (B'), outer area (S').

The global recognition rate is 65.3%. This result may be improved by another modification of the consonant grouping. The main errors come from the confusion between C3 and C5 consonant groups.

3.2. Experiment 2: HMM recognition with improved consonant grouping

The experiment 2 aims to evaluate the influence of a new modification of the consonant grouping. Indeed, after preliminary tests it appeared that the [l] and the [R] consonants improved the global recognition score when they are in the same consonant group. Thus, the [l] consonant is shifted to the [k, g, R] group. To deal with the problem of the CS compatibility (the [ɲ] and the [l] consonants are coded with the same CS handshape), the [ɲ] consonant must be shifted to [t, d, s, z, n] group. Finally, the final consonant groups for this experiment are: C1= [p, b, m], C2= [f, v], C3= [t, d, s, z, n, ɲ], C4= [ʃ, ʒ] and C5= [k, g, R, l].

Note that the choice of the switch between C3 and C5 is motivated by the fact that their associated consonants are not mainly articulated at the lips.

With those groups, the recognition rate increases to reach 75.9% (which means an increase of 10.6% in comparison with the first experiment). However, significant errors remain in the discrimination of CV classes containing C3 and C4 consonants in the context of V1 vowels. Indeed, the recognition rate of CV classes containing C4 and C3

consonants are lower than the global rate with respectively 64.4% and 70.5%.

It should be recalled that these rates are obtained by using the six previous lip parameters. They seem to be not sufficient to take into account the specific control of the lip shapes especially for C4 consonants.

3.3. Experiment 3: lip pinching effect

Experiment 3 aims to evaluate the contribution of both the upper and lower lip pinching to recognize CV syllables in a HMM recognition test and more precisely to better modelize the C4 consonant group. So, The HMM recognition test is based this time on eight lip parameters. In addition to the six parameters used in the previous experiment, the pinching of upper and lower lips (respectively Bsup and Binf) is measured at one point (more precisely in the mid-lips). The CV corpus is still based on the same grouping of consonants as the experiment 2.

In this experiment 3, the performance reaches 80.3% as global recognition rate. This result shows that the pinching parameters improve the recognition accuracy by more than 4% in average. This enhancement differs from a class to another. Indeed, Table 1 shows the recognition rates for each of the consonant group. Interestingly, the recognition rate for CV sequences containing C4 consonants benefits more than 10% (see table 1). Although C3 consonants are not articulated at lips, the recognition rate is improved. This proves that the lip pinching brings supplementary information to discriminate CV sequences.

It is effortless to attribute the major part of the errors to the fact that C3 and C5 consonants are not principally articulated at lips. Thus, if these two groups are gathered in the same consonant group, practically the global recognition rate does not change (80.58%). But, if they are not considered in the test, the recognition rate increases clearly (90.41%). This later on CV recognition is similar to the classification rate obtained by a Gaussian classification for vowels only (Aboutabit et al. 2006 [3]). Then, the residual error can be partly explained by the confusion between vowels.

Table 1. Recognition rates by consonant grouping of CV syllables obtained in experiment 3 compared to those resulted in experiment 2.

| | CV with C1 | CV with C2 | CV with C3 | CV with C4 | CV with C5 |
|-----------------------------------|------------|------------|------------|------------|------------|
| Recognition Rate without pinching | 85.5% | 81.7% | 70.5% | 64.5% | 80.1% |
| Recognition rate with pinching | 89.9% | 79.4% | 75.8% | 74.8% | 81.7% |

4. Discussion

The three experiments have shown good recognition rates for CV syllables based on consonants which are articulated or not at lips. Even though the consonants are not articulated at lips, these rates demonstrate that the choice of the lip parameters is pertinent.

The best recognition rate is obtained in experiment 3. However, some errors remain. Table 2 presents the confusion matrix for experiment 3 results. In this table, each column illustrates the distribution of CV identification (in line). The diagonal corresponds to the number of correct recognized CV.

Firstly the confusions between consonant groups and vowel groups explain the error with different proportion for each CV syllable. Thus, for CV syllables with C1 consonant group, the 10.1% error is caused principally by the confusion between groups of vowels (9.4%). This proves the relevance of the choice of lip parameters to account for the bilabial occlusion. Inversely, for the other CV syllables, the confusion between consonant groups results in a large part of the errors. For example, for expected CV syllables with C3 consonants, the errors caused by the confusion between consonants reach 20% while those caused by the confusion between vowels is only 4.2%.

On the other hand, the opening gesture is well detected. In fact, the transition from the consonant towards anterior non rounded and low and mid-low rounded vowels seems to bring relevant information to discriminate CV syllables. In contrast, in the context of V1, i.e. high rounded vowels, CV syllables are largely less recognized except the case of C1V1. Note that the rounded gesture is well identified (very few errors on the identification of V1). Then, the transition from the consonant towards V1 group of vowels seems not sufficient to discriminate the consonant using the chosen lip parameters. This result is coherent with the well known co-articulation effect with high rounded vowels (V1 group) that modifies the lip realization of the preceding consonant (Abry & Boë, 1986 [8]) in the manner that the consonant can be masked.

The results on C2V3 can not be analyzed since the size is too low.

5. Conclusions

In conclusion, the HMM modeling of CV syllables based on lip parameters gives interesting performances. More precisely, the consonant viseme grouping finally obtained and the addition of the lip pinching parameters benefit to the CV recognition.

Moreover, this study shows that the errors are mainly observed on consonant groups in the context of non rounded vowels. In contrast, CV syllables for high and mid-high rounded vowels ([ɔ̃, y, o, ø, u]) and for low and mid-low rounded vowels ([ã, ɔ, œ]) are well recognized (in average 87%).

In this work, the modeled CV syllables are in context of sentences. The effect of the context on the recognition score should be analyzed in order to optimize the selected observations of CV syllables.

In perspective, a supplementary parameter related to the movement inside the mouth (such as the tongue gesture) could improve the recognition score.

6. Acknowledgements

Many thanks to Sabine Chevalier, our CS speaker, for having accepted the recording constraints. This work is supported by the French TELMA project (RNTS / ANR).

7. References

- [1] Attina, V., Beautemps, D., Cathiard, M. A. and Odisio, M. "A pilot study of temporal organization in Cued Speech production of French syllables: rules for Cued Speech synthesizer." *Speech Communication*, Vol. 44, 2004, pp. 197-214.
- [2] Aboutabit, N., Beautemps, D. and Besacier, L., "Hand and Lips desynchronization analysis in French Cued Speech: Automatic segmentation of Hand flow". In *Proceedings of ICASSP'06*, 2006.
- [3] Aboutabit, N., Beautemps, D. and Besacier, L., "Vowels classification from lips: the Cued Speech production case". In *Proceedings of ISSP'06*, 2006.
- [4] Lallouache, M.-T. "Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des

lèvres," Doctoral dissertation, Institut National Polytechnique de Grenoble, Grenoble 1991.

[5] Lamy, R., Moraru, D., Bigi, B., Besacier, L. "Premiers pas du CLIPS sur les données d'évaluation ESTER," In *Proc. of Journées d'Etude sur la Parole*, Fès, Maroc, 2004.

[6] Vaufreydaz, D., Bergamini, J., Serignat, J. F., Besacier, L. & Akbar, M. "A New Methodology for Speech Corpora Definition from Internet Documents". *LREC2000*, 2nd International Conference on Language Resources and Evaluation. Athens, Greece, pp. 423-426, 2000.

[7] R.O. Cornett, "Cued Speech," *American Annals of the Deaf*, 112, pp. 3-13, 1967.

[8] Abry, C. & Boč, L.-J., "Laws for lips," *Speech Communications*, 5, pp. 97-104, 1986.

Table 2. *Confusion matrix of experiment 3 results*

| | | Expected CV syllables | | | | | | | | | | | | | | |
|-------------------------|------|-----------------------|------------|-----------|-----------|-----------|----------|------------|------------|-----------|-----------|-----------|-----------|------------|------------|-----------|
| | | C1V1 | C1V2 | C1V3 | C2V1 | C2V2 | C2V3 | C3V1 | C3V2 | C3V3 | C4V1 | C4V2 | C4V3 | C5V1 | C5V2 | C5V3 |
| Recognized CV syllables | C1V1 | 85 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | C1V2 | 0 | 137 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | C1V3 | 6 | 18 | 45 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | C2V1 | 0 | 0 | 0 | 22 | 1 | 1 | 11 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 |
| | C2V2 | 0 | 0 | 0 | 0 | 77 | 5 | 2 | 3 | 1 | 0 | 2 | 0 | 0 | 1 | 0 |
| | C2V3 | 0 | 0 | 0 | 2 | 0 | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | C3V1 | 0 | 0 | 0 | 7 | 1 | 0 | 171 | 2 | 1 | 18 | 0 | 0 | 23 | 0 | 0 |
| | C3V2 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 235 | 0 | 0 | 2 | 1 | 0 | 13 | 0 |
| | C3V3 | 0 | 0 | 0 | 0 | 1 | 0 | 13 | 10 | 64 | 0 | 0 | 0 | 2 | 2 | 2 |
| | C4V1 | 0 | 0 | 0 | 1 | 0 | 0 | 36 | 0 | 3 | 23 | 0 | 0 | 26 | 0 | 0 |
| | C4V2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 67 | 1 | 0 | 7 | 0 |
| | C4V3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 6 | 0 | 2 | 11 | 0 | 1 | 0 |
| | C5V1 | 1 | 0 | 1 | 2 | 1 | 1 | 23 | 1 | 0 | 4 | 0 | 0 | 122 | 1 | 2 |
| | C5V2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 22 | 0 | 0 | 1 | 0 | 0 | 314 | 0 |
| | C5V3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 12 | 7 | 40 |
| | sum | | 92 | 155 | 50 | 34 | 85 | 12 | 258 | 284 | 77 | 48 | 74 | 13 | 191 | 348 |
| Recognition rate (%) | | 92% | 88% | 90% | 65% | 91% | 42% | 66% | 83% | 83% | 48% | 91% | 85% | 64% | 90% | 91% |