Vowel classification from lips: the Cued Speech production case.

Noureddine Aboutabit[1], Denis Beautemps[1], Laurent Besacier[2]

[1]Institut de la communication Parlée, CNRS UMR5009 /INPG/Université Stendhal
46 Av. Félix Viallet, 38031 Grenoble, cedex 1, France

[2]Communication Langagière et Interaction Personne Système, CNRS UMR5524
/UJF/INPG
385, rue de la bibliothèque-BP 53 – 38041 Grenoble Cedex 9

Abstract. *Cued Speech (CS) (Cornett, 1967) 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). Handshapes are designed to distinguish among consonants and hand positions among vowels. Due to the CS system, both hand and lip flows produced by the CS speaker carry a part of the phonetic information. This contribution focuses on the lip flow (lip parameters) and discusses lip-shape classification of vowels for each cued speech hand position. The lip parameters are extracted from the inner lip contour and still characterize the vowels in the context of CS production (see Aboutabit et al., 2006b). This paper will show how the distribution of lip parameters inside each group of CS hand positions allows vowel discrimination. A classification method based on Gaussian modeling is presented and results demonstrate a good performance of this classification (95,03% as learning score and 89% as test score).*

1. Introduction

Cued Speech (CS) (Cornett, 1967) 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 for a precise 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, 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., 2006a for an example of a complete analysis of the hand

flow). This paper focuses on the lip flow for vowels, with lip parameters extracted from the inner lip contour. Indeed these parameters still characterize the vowels (see Aboutabit et al., 2006b). An automatic method of lip target segmentation applied to the vowels allows to obtain the corresponding lip characteristics (parameters). For each CS hand position a Gaussian classification based on these parameters discusses the ability to identify vowels from a single measure instant. 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).

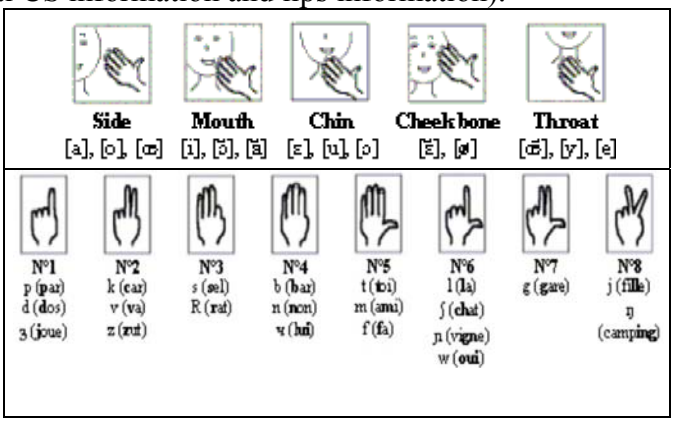

Figure 1. CS Hand position (top) for vowels and CS handshapes (bottom) for consonants (adapted from [1]).

1. Experimental set-up and data

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

Figure 2. Image of the speaker.

The French CS speaker is a female native speaker of French, certified in French CS. She regularly translates into French CS code in a school setting. 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. A second camera in zoom mode dedicated to the lips was synchronized with the first one but connected to a second betacam recorder. The lips were painted in blue, and blue marks were placed on the speaker's glasses as reference points (Figure 2).

At the beginning of the recording session, a set of LEDs was placed in the field of the camera and activated for further correspondence between the time codes of the two video recordings. In addition, 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 contour and to derive the corresponding characteristic parameters (Lallouache, 1991): 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.

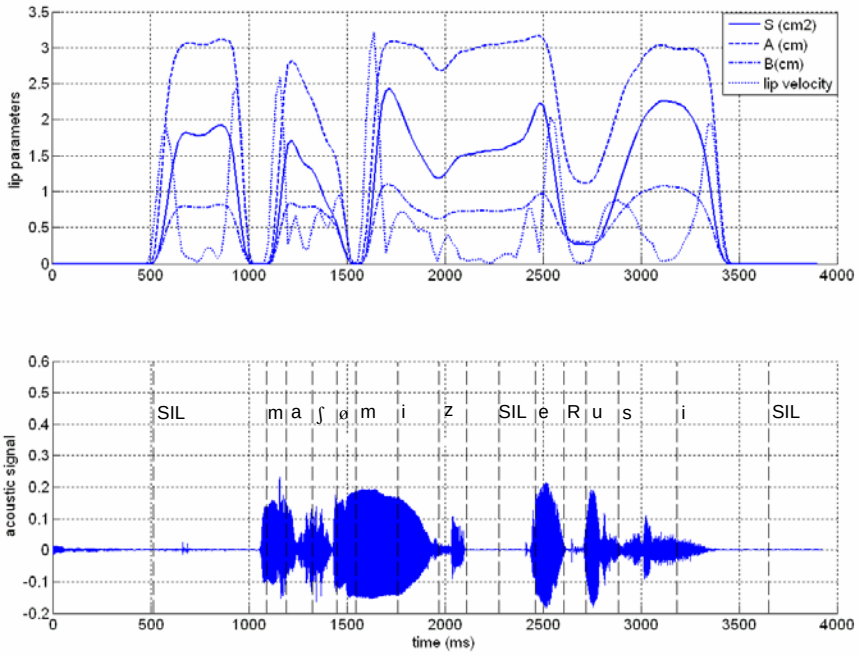

Figure 3. Top: A (dashed line), B (dash-dot line) and S (solid line) parameters extracted from the inner contour and the corresponding lip velocity (dotted line). Bottom the corresponding acoustic realization. SIL: acoustic silence.

The acoustic signal was automatically labeled at the phonetic level using forced alignment (see Lamy, 2004 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).

This process resulted in a set of temporally coherent signals: the 2D hand position (see Aboutabit, 2006a) the lip width (A), the lip aperture (B) and the lip area (S) values every 20 ms, and the corresponding acoustic signal with the associated phonetic chain temporally marked. Figure 3 shows an example of the different data flows for a sentence.

3. Lip target segmentation

The lips were characterized at the instant the lip target was attained. The automatic definition of this instant is based on the temporally marked phonetic chain. Recall that the phonetic chain marks the acoustic realization. Note that the beginning and the end of each phoneme are obtained automatically with a forced alignment; this labeling may therefore include errors or fuzzy phone frontiers. Moreover, it is well known that the lip can anticipate the acoustic realization. Thus, in the automatic process of lip target calculation, the middle of the phoneme interval is considered as a first estimation of the instant of vocalic target. The target instant is finally obtained at the nearest instant of minimum lip velocity. In the case of important anticipation the research process is limited by the end frontier of the phone acoustic realization. Lip velocity (see Figure 3, dotted line) is estimated from the lip area S parameter as the difference between two successive values normalized by the sample periodicity (20 ms). Note that S is highly correlated to the crossing of A by B ($r = 0.99$).

The algorithm for vocalic lip target instant detection is thus as follows: (1) calculation of the lip velocity from S parameter, (2) detection of all the local minima, (3) determination of the mid-point of the vowel from the phonetic chain (4) choice of the nearest instant of lip velocity local minimum.

4. lip-reading ambiguity and Cued Speech complementarity : the vocalic case

The previous algorithm was applied to obtain the (A, B, S) parameters at the instant of vowel lip target. A Mahalanobis distance was computed on the basis of the A, B and S lip parameters and was used to trace the hierarchical cluster tree (dendrogram) from the vowel distribution. The dendrogram consists of many U-shaped lines connecting objects (vowels or group of vowels) in a hierarchical tree. The height of each U represents the distance (using the Mahalanobis distance) between the two objects being connected. Figure 4 shows how the vowels are grouped into three categories (visemes, Benoit, 1992) in conformity with the phonetic description of the vowels (anterior non rounded vowels [a, ɛ̃, i, œ̃, e, ɛ], high and mid-high rounded [ɔ̃, y, o, ø, u] low and mid-low rounded vowels [ɑ̃ , ɔ, œ]). From this, we can conclude that the extracted lip parameters from the inner contour still accurately characterize the phonetic content in the CS context of speech production. Moreover, vowels can't be differentiated with only lip information. This illustrates the lip-reading ambiguity that deaf people are confronted with.

The previous grouping into three vocalic visemes is compatible with the grouping of the five CS hand positions excepted for one case. For example, the vowels [a, œ, o] of the CS side position are included in the three different visemes, and thus are well discriminated, as seen in figure 5.

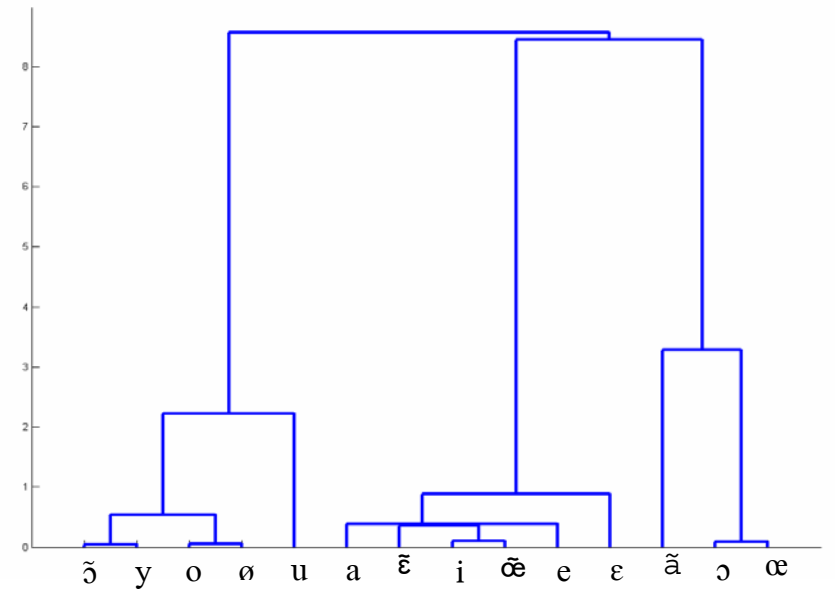

Figure 4. Hierarchical cluster tree of the vowels.

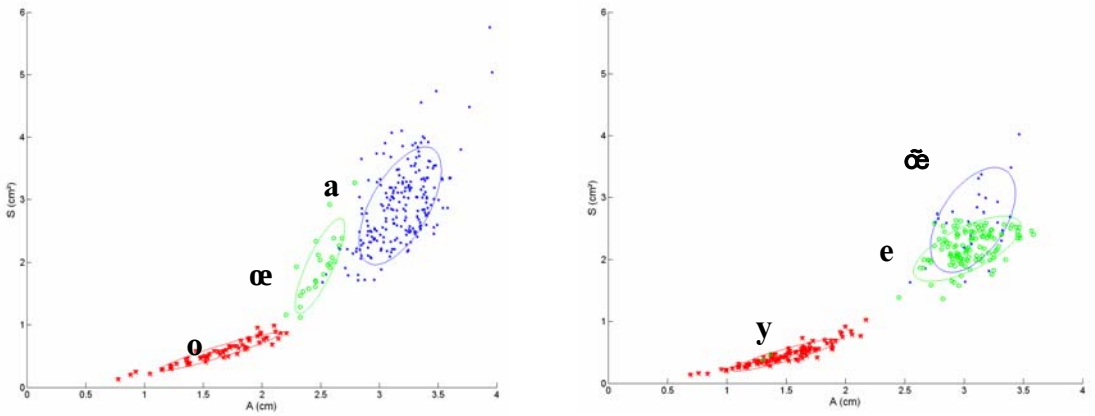

Figure 5. French vowels with the CS side (left) and throat (right) hand positions in the [A(cm), S(cm²)] plan, 1.5 standard deviation ellipsis.

For the single exception, the CS throat position, the [œ̃] and [e] vowels are also in the same viseme group but in contrast to the previous case, their distribution is still very closed in the (A, S) and (A, B) plans (see Figure 5 for the (A, S) plan). For this case, the discrimination between [œ̃] and [e] from the lips might be slightly tricky, even with the CS hand position information. The [œ̃] realizations are not sufficiently opened. This observation can be explained by the fact that the CS speaker does not seem to differentiate [œ̃] from [ɛ̃] with the lips, even though these two phonemes are cued with two different CS hand positions. Indeed the CS speaker produced similar realizations of lip shapes for [œ̃] and [ɛ̃], as shown by the small distance between the two corresponding distributions (see Figure 5). The ambiguity is maintained due to the

coding choice of the CS speaker. In this case, the complete discrimination needs a higher level of processing.

5. Classification

According to the previous section, knowing both lip information and CS hand position, the vowels can be discriminated a priori excepted the lip confusion between [œ̃] and [e]. To evaluate this discrimination, a Gaussian classification based on the lip inner contour parameters (A, B, S) was carried out for each vowel group corresponding to a CS hand position. For this classification, the data were obtained with the lip target segmentation method presented previously and applied to 124 sentences repeated twice, that resulted in 1167 vowels for the learning phase and 1105 vowels for the test phase. Data size of both learning and test phases are presented in table1.

	a	o	œ	ɛ̃	ø	i	ã	ɔ̃	ɛ	u	ɔ	œ̃	y	e
Learning	216	63	24	26	110	176	67	41	96	69	32	26	97	124
test	199	57	23	24	97	168	66	42	83	68	31	25	96	126

Table 1. Selected vowels and sample size by vowel for both the learning and the test.

	A (cm)	B (cm)	S (cm²)
ø	1,6 (0,31)	0,48 (0,11)	0,55 (0,2)
œ	2,5 (0,14)	1,1 (0,22)	1,94 (0,51)
ɛ	3,4 (0,24)	1,32 (0,13)	3,13 (0,42)
ɔ	2,45 (0,15)	1,06 (0,23)	1,82 (0,47)
a	3,18 (0,23)	1,29 (0,23)	2,9 (0,63)
ã	2,28 (0,21)	0,79 (0,18)	1,25 (0,37)
e	3,03 (0,3)	1,02 (0,13)	2,18 (0,35)
i	3,13 (0,25)	1,22 (0,17)	2,67 (0,52)
ɛ̃	3,05 (0,22)	1,11 (0,26)	2,41(0,66)
o	1,64 (0,34)	0,49 (0,11)	0,57 (0,2)
ɔ̃	1,45 (0,35)	0,42 (0,11)	0,45 (0,21)
u	1,17 (0,24)	0,35 (0,1)	0,3 (0,12)
œ̃	3,08 (0,24)	1,2 (0,21)	2,64 (0,56)
y	1,49 (0,30)	0,42 (0,1)	0,46 (0,17)

Table 2. Mean and standard deviation (between brackets) of lip parameters (A, B, S) for the vowels of the learning data.

The learning phase consists in building a 3 dimensional Gaussian model of each vowel. Table 2 presents, for each vowel, both mean and standard deviation vectors necessary to define the corresponding model.

A first level of the classifier performance was estimated with the modeling scores of the learning data (Figure 6). The average score was 95.03%. The 4.97 % of error were partially caused by the undiscriminated [œ̃] and [e] groups of the throat position (for 2.71%). The precision of automatic phone labeling and the lip target segmentation methods explain the residual error. The test data were used for a second level performance evaluation (Figure 7). The average score was 89%. In addition to the errors explained just previously, supplementar errors (for 4.5%) were caused by both the phonetic transcription and CS speaker coding errors. More precisely these errors were observed in the distinction of [o] vs. [ɔ], [e] vs. [ɛ] and [ø] vs. [œ].

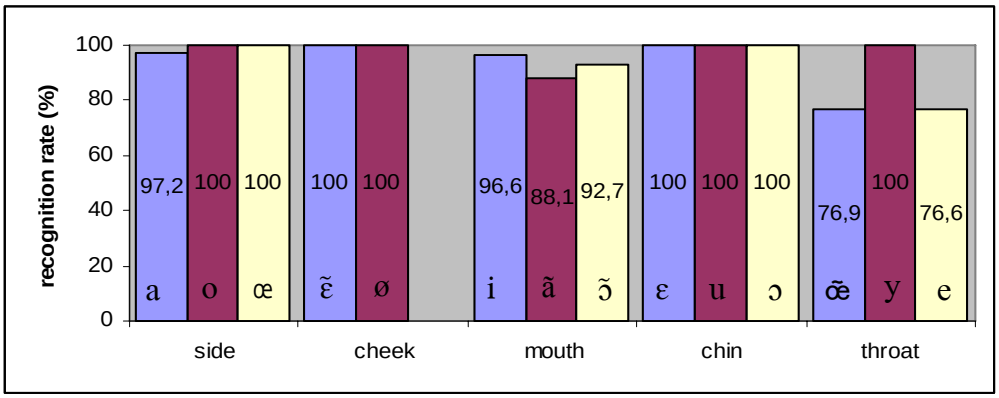

Figure 6. Scores of correct classification for the learning data.

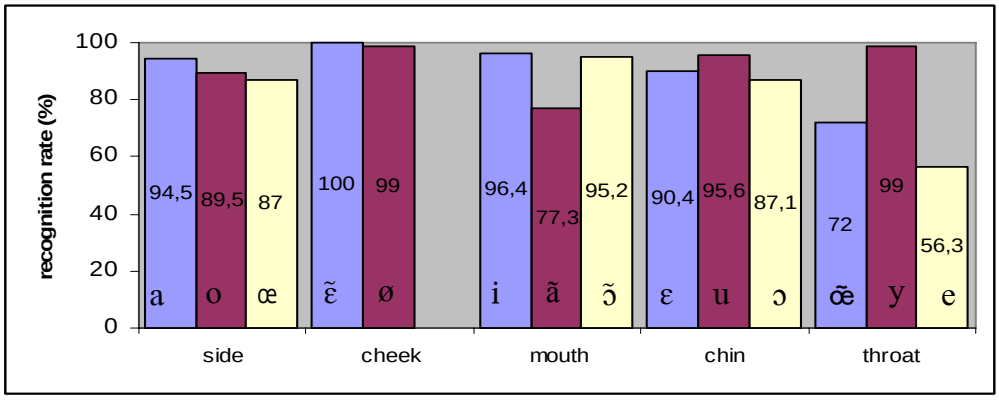

Figure 7. Scores of correct classification for the test data.

6. Conclusion and perspectives

This study demonstrates that when the CS hand position was given, high scores of vowel identification are obtained with only one measure instant, defined by the lip target segmentation method. The global score of 89% has to be compared to the CS perceptual effectiveness score of 83.5% as obtained by Nicholls and Ling (1982) in their study on the reception of CV and VC syllables with hearing-impaired children. However, identification errors, caused by the confusion between some vowels ([ɛ̃] vs. [œ̃], [o] vs. [ɔ], [e] vs. [ɛ] and [ø] vs. [œ]) and the imprecision of the lip target detection

system, subsist. It could be minimized by an improvement of the automatic acoustic labeling precision and the quality of the lip contour extraction.

This approach needs to be extended to the case of consonants. In this case, since consonant realization at the lips are generally influenced by the vocalic context, the lip information at the instant of vowel lip target should be a key issue.

7. Acknowledgments

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).

References

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, pp. 197-214, 2004.

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*, 2006a.

Aboutabit, N., Beautemps, D. and Besacier, L. Characterization of Cued Speech vowels from the inner lip contour. *In Proceedings of ICSLP'06*, 2006b.

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.

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.

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.

Benoit, C., Lallouache, T., Mohamadi, T., and Abry C. "A set of French visemes for visual French speech synthesis", in G. Bailly, C. Benoît and T.R. Sawallis (Editors). *Talking Machines: Theories, Models and Designs,* pages 485-504. Amsterdam: Elsevier SC. Publishers, 1992.

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

Nicholls, G., Ling, D. Cued Speech and the reception of spoken language. *Journal of Speech and Hearing Research*.25, 262-269, 1982.