Feature adaptation of hearing-impaired lip shapes: the vowel case in the Cued Speech context

Noureddine Aboutabit, Denis Beautemps, Olivier Mathieu, Laurent Besacier

Feature adaptation of hearing-impaired lip shapes: the vowel case in the Cued Speech context.

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

[1]Grenoble Images Parole Signal Automatique, département Parole & Cognition
46 Av. 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, France

Abstract

The phonetic translation of Cued Speech (CS) gestures needs to mix the manual CS information together with the lips, taking into account the desynchronization delay (Attina et al. [2], Aboutabit et al. [4]) between these two flows of information. This contribution focuses on the lip flow modeling in the case of French vowels. Previously, classification models have been developed for a professional normal-hearing CS speaker (Aboutabit et al., [7]). These models are used as a reference. In this study, we process the case of a deaf CS speaker and discuss the possibilities of classification. The best performance (92.8%) is obtained with the adaptation of the deaf data to the reference models.

Keywords: Lipreading, Lip Modeling, Vowel Classification, Cued Speech.

1. Introduction

To date, the benefit of visual information for speech perception (so called "lip-reading") is well known. However, even with high lip-reading performances, speech without knowledge of the semantic context can not be completely perceived. The best lip readers generally do not reach perfection. On average, only 40- 60 % of the phonemes are recognized by lip-reading for a given language (Montgomery & Jackson, [9]), and only 10-30 % of the words (Nicholls & Ling, [10]; Bernstein et al., [12]). The main reason for this is that the visual pattern is ambiguous. Anyway lip-reading remains for the orally educated deaf people the main modality to perceive speech. That is the reason why, Cornett ([1]) developed the Cued Speech system (CS) to complement the lip information.

CS is a visual communication system that uses handshapes placed in different positions near the face in combination with natural speech lip-reading to enhance speech perception from visual input. In this system, the speaker facing the perceiver moves his hand in close relation with speech (see Attina et al., [2] for a detailed study on CS temporal organization). The hand (with the back facing the perceiver) is a cue that corresponds to a unique phoneme when associated with a particular lip shape. A manual cue in this system contains two components: the shape of the hand and the hand position relative to the face. Handshapes are designed to distinguish among consonants whereas hand positions are used to distinguish 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). Firstly, CS is improving speech perception to a large extent for deaf people (Nicholls, [11]; Nicholls & Ling, [10] for the identification of the syllables, Uchanski et al., [13] for the identification of sentences, scores between 78 and 97 %). Secondly, CS offers a complete representation of the phonological system for deaf people exposed to this method since their youth, and therefore has a positive impact on the language development (Leybaert, [3]).

The demand of handicapped people to access to communication technologies is a major concern in modern society. The TELMA project (Beautemps et al., [14]) proposes to develop an automatic translation system of acoustic from speech towards visual speech completed with CS and inversely, i.e. from CS manual and lip components towards auditory speech. Thus, with this project, it will make possible to deaf users to communicate between them and with normal-hearing people with the help of the autonomous terminal TELMA. In this context, 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. The present contribution addresses the automatic lip flow modelling in the case of the French vowels produced by a hearing-impaired CS speaker. In a previous work, it has been shown that for a normal-hearing CS speaker, the automatic classification of the vowels with respect to the corresponding CS hand position was possible (accuracy of 89 %) with only three parameters derived from the inner contour of the lips at the instant of vowel lip target (Aboutabit et al., [7]). The normal-hearing CS speaker was certified in CS and was video recorded using a set of constraints, such as blue make up for the lips, head fixed in a helmet to avoid movement and safety goggles to protect the eyes of the strong highlight (Figure 2). However, what becomes the classification of the vowels produced by a CS speaker who is hearing-impaired and in the case of less constrained experimental set-up? And how, while being based on the modelling of the lip data of the normal-hearing cuer, considered as a reference, to process the between cuers variability?

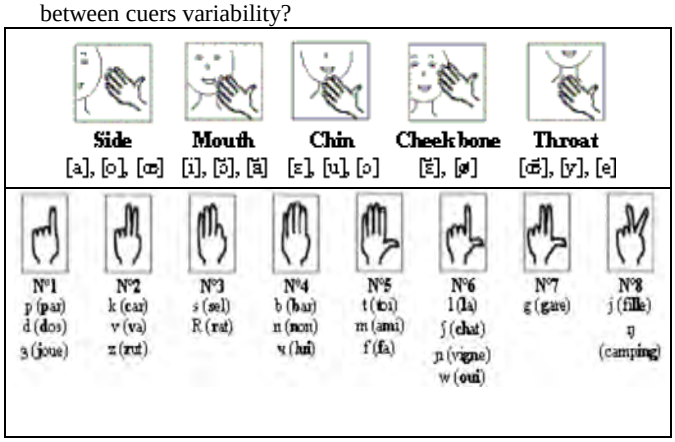

Figure 1: *CS Hand position for the vowels and handshapes for the consonants (adapted from Attina et al. [2]).*

2. Speech material

The speech material in this experiment is derived from a video recording of a profoundly deaf young woman coding in French Cued Speech and participant in an experimentation of a telephone conversation with a normal-hearing speaker. The conversation was realized according to a Wizard of Oz paradigm. This participant, whom one will call the hearing-impaired cuer hereafter, practices CS daily in particular to communicate with other deaf people. The Wizard of Oz experimentation consisted here in simulating a distant interaction between the hearing-impaired and the normal-hearing participant by creating an illusion of a real telephone. The hearing-impaired cuer believed that her CS gestures were automatically recognized and transformed into speech, which is transformed remotely on the telephone line towards the normal-hearing participant. The normal-hearing participant stayed in a room next to the place of the hearing-impaired cuer and he received directly the video of the hearing-impaired cuer without using telephone. The normal-hearing participant was a complicit. Under these experimental conditions, the hearing-impaired cuer was seated and was free of his movements and of the choice of his lexicon within a framework of communication theme (a trip reservation at an agency and a fixing RDV at a medical secretary). The conditions of lighting allowed the hearing-impaired cuer not to have need for eyes protection. On the other hand, in a similar way as with the professional normal-hearing cuer, the information of the lips and the hand was marked by artifices (make-up of the lips in blue, landmarks on the back of the hand and at the extremity of the fingers, reference landmark on the face).

Figure 2 illustrates the experimental set-up. Using the Image/Speech tool of the Speech & Cognition Department of GIPSA-lab, the images of the videotapes of the recording were digitized like Bitmap images every 20 ms, in synchrony with the corresponding audio part, digitized at 44100 Hz.

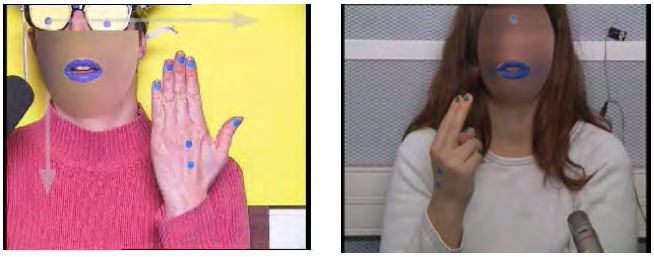

Figure 2: *CS speakers. On left: The normal-hearing cuer; On the right: the hearing-impaired cuer.*

The lip information was extracted directly from the images using a process which locates first the inner and the outer contours of the lips, and then derives the temporal evolutions of the parameters A, B and S (respectively lip width, lip aperture, lip area of the contour). 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 (Lamy et al., [5], Vaufreydaz et al., [6]). This whole process resulted in a set of temporally coherent signals: the 2D hand position (see Aboutabit et al., [4]) the lip width (A), the lip aperture (B) and the lip area (S) values every 20 ms for the inner contour, and the corresponding acoustic signal with the associated phonetic chain temporally marked.

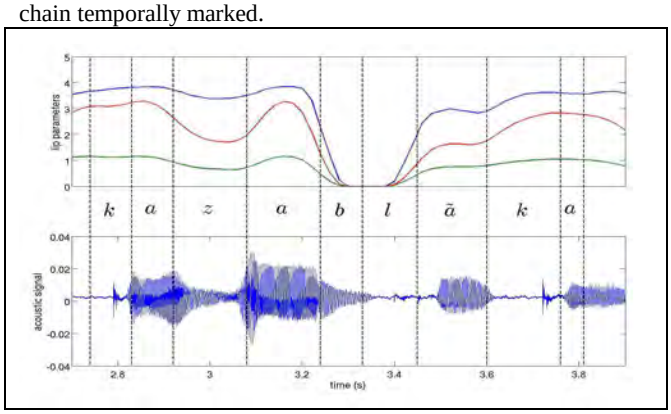

Figure 3: *The A, B and S lip parameters of the inner contour with the corresponding acoustic realization.*

3. The modeling

3.1 Lip target detection of the vowels

The lips are considered at the instant of attained lip target, marked at the instant of minimum of lip movement. 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. It is imperative to realize that this alignment requires knowing the phonetic transcription of phonemes. However, the final objective of this study is to recognize these phonemes from visual features merged with information from the hand. We used this alignment in order to control the experimental process. If necessary, there were other approaches to segment the speech signal without recourse to the phonetic transcription (Golipour & O'shaughnessy, [15]). Therefore, this labeling may 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 vowel target. The target instant is finally obtained at the nearest instant of minimum lip velocity. The lip velocity is calculated from the difference on S(t) for two successive instants separated by 20 ms (S is the area lip parameter calculated from the inner contour, and low-pass filtered for the lip velocity calculation). This algorithm was applied to all of the sequences to obtain the instants of vowel lip-target, so-called L2 instant with respect to Attina et al. ([2]) nomenclature. Table 1 presents the number of vowels thus obtained for each vocalic category, for the hearing-impaired cuer.

Table 1 : *selected vowels and sample size by vowel.*

Vowel	a	o	œ	ɛ̃	ø	i	ɑ̃	ɔ̃	ɛ	u	ɔ	y	e
size	200	63	19	40	46	162	59	57	118	58	29	45	130

3.2 Vowel grouping : visemes

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 et al., [16]) in conformity with the phonetic description of the vowels (anterior non rounded vowels [a, ɛ̃, i, e, ɛ], high and mid-high rounded [ɑ̃ , ɔ̃, y, o, ø, u] low and mid-low rounded vowels [ɔ, œ]).

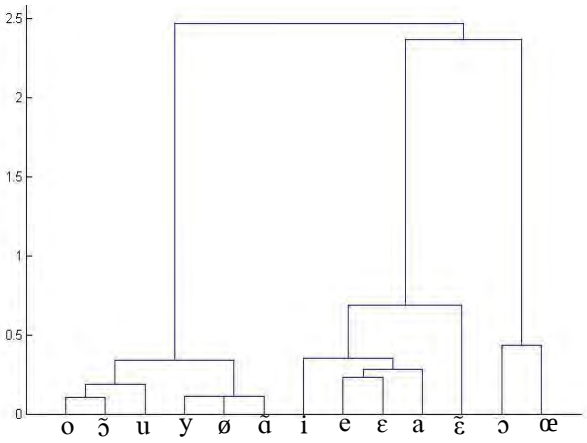

Figure 4: *Hierarchical cluster tree of the vowels (Adapted from Sacher et al., [8]).*

We obtain here the same vowel grouping as for the reference cuer except for the vowel [ɑ̃] which switched to the high and mid-high rounded vowels group.

3.3 Classification

In accordance with the modeling suggested in Aboutabit et al. ([7]), a three dimensional mono-Gaussian classifier of the three lip parameters is considered for each of the five CS hand positions. This classifier applied directly to the data of the hearing-impaired cuer gives 62.3% as average accuracy. Thus, to improve the classification in the case of the hearing-impaired cuer, two approaches were studied: with or without retraining. In the first experiment, the corpus is divided into two sets of data, one for the estimation of the values of the classification parameters (mean and covariance matrix), the second one for the evaluation. In the second approach, the data of the hearing-impaired cuer are adapted to the referent models. For the adaptation, we considered two approaches. In the first one, a simple translation of the average values toward the reference ones has been applied, without changing the standard deviation. The second one completes the translation by the normalisation of the standard deviations (see equation b).

$$(a)\ \tilde{X}_S = X_S + (m_{NE} - m_S) \quad (b)\ \tilde{X}_S = (X_S - m_S)\frac{\sigma_{NE}}{\sigma_S} + m_{NE}$$

With $\tilde{X}_S$: the normalized lip parameter; X_S : the lip parameter; m_S, m_{NE}: mean values for respectively the hearing-impaired and reference cuer; σ_S, σ_{NE}: standard deviation values for respectively the hearing-impaired and reference cuer.

In addition, we applied these two approaches of adaptation to three different regroupings of the vowels. In the first one (R1), all the vowels are considered in a same group and a same adaptation processing is applied to all the vowels. In the second ones (R2), the vowels are gathered by groups of the three referent visemes: the non-rounded vowels [a, ɛ̃, i, œ̃, e, ɛ], the rounded vowels [ɔ̃, y, o, ø, u] and the mid-low rounded vowels [ɑ̃ , ɔ, œ] (Aboutabit et al., [7]). Finally, in the third one (R3), each of the vowel category is considered alone.

4. Results and Discussion

4.1 Classification with retraining

Let us recall that here, a Gaussian classifier of the labial parameters is considered for each CS hand position. The average accuracy is 82.5% (see on figure 4 the distribution of this score), a little lower than the 89 % obtained for the reference cuer, but is still comparable.

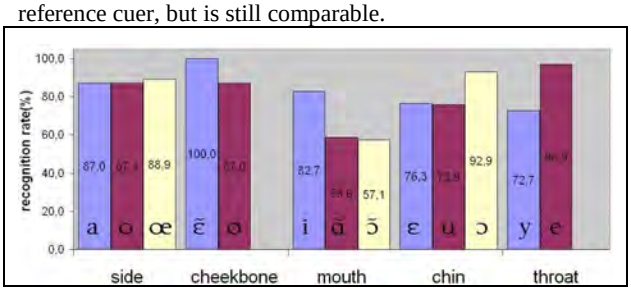

Figure 5: *Vowels classification per CS Hand position.*

The mouth hand position is the category for which the rate is drawn downwards, caused by the lip-reading ambiguity between the vowels [ɑ̃] and [ɔ̃] which are not well differentiated in the production of the hearing-impaired cuer (these two vowels are in the same viseme as demonstrated by the previous dendrogram).

4.2 Classification without retraining

Two main tendencies appear (Table 3). The use of the translation alone is less effective than with the addition of the reduction of the standard deviations. Indeed, for all the regroupings (R1, R2 or R3), the accuracy is lower than the score of 82.5% obtained previously. In addition, the adaptation with the R1 regrouping gives a quasi-identical score in both cases of adaptation, which remains clearly lower than the accuracy obtained for R2 or R3. Only the cases R2 and R3 give comparable rates even higher than the score of 82.5 %. Finally, it should be noted that the scores for R2 and R3 are very close what gives finally a premium to the R2 condition since in the best of the cases (Translation + reduction), only six coefficients (3 averages and 3 standard deviations) are to be applied for the adaptation, in comparison of 26 for the R3 condition.

Table 3: *Recognition rate (%) of vowels per adaptation kind and the regrouping level.*

		Shift			***Shift + reduction***	
Vowel	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*

a	83,0	82,0	83,0	83,5	89,0	93,0
o	63,5	85,7	84,1	60,3	95,2	93,7
œ	0,0	31,6	10,5	5,3	84,2	84,2
ɛ̃	90,0	90,0	92,5	90,0	90,0	90,0
ø	60,9	73,9	76,1	54,3	100	100
i	88,9	91,4	90,1	90,1	90,1	93,2
ɑ̃	61,0	27,1	40,7	52,5	83,1	83,1
ɔ̃	5,3	28,1	59,6	5,3	96,5	86,0
ɛ	66,1	85,6	89,8	64,4	94,9	94,9
u	69,0	87,9	89,7	67,2	96,6	98,3
ɔ	48,3	69,0	82,8	55,2	100	100
y	26,7	53,3	31,1	22,2	91,1	95,6
e	96,2	98,5	96,9	95,4	98,5	100
% global	**70,4**	**77,8**	**79,8**	**69,4**	**92,8**	**93,9**

5. Conclusion

The lip shapes produced by the CS hearing-impaired cuer in the case of the vowels can also be classified by a simple mono-Gaussian classifier tool. The best results are obtained in the case of the data adaptation towards the reference models of the normal-hearing cuer. The best of the cases (92.8 % of the R2 condition, for the adaptation "Translation + reduction") gives a performance identical (even higher) to that of the reference cuer. The significance of this result can be better appreciated when one realizes that the cuer analyzed here is profoundly deaf and than the experimental set-up was less constrained compared to the normal-hearing cuer of reference. Even if only one subject could be tested, this study shows that the idea to model finely a reference cuer and then to adapt any other cuer on this reference seems to be a profitable step. This contribution opens the way towards more complex logatomes as for example Consonant-Vowel syllables.

6. Acknowledgements

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

7. References

[1] Cornett, R.O., “Cued Speech,” American Annals of the Deaf, 112, pp. 3-13, 1967.

[2] 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, 44, pp. 197-214, 2004.

[3] Leybaert, J., Phonology acquired through the eyes and spelling in deaf children. Journal of Experimental Child Psychology, 75, 291-318, 2000.

[4] Aboutabit, N., Beautemps, D. and Besacier, L., “Hand and Lips desynchronization analysis in French Cued Speech: Automatic segmentation of Hand flow”. In Proc. of ICASSP, 2006.

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

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

[7] Aboutabit, N., Beautemps, D. and Besacier, L., “Vowels classification from lips: the Cued Speech production case”. In Proceedings of ISSP’06, 2006

[8] Sacher, P., Beautemps, D., Cathiard, M.-A. and Aboutabit, N., "Analyse de la production d’un codeur LPC sourd". In Proc. of Journées d’Etude de la Parole, Avignon, France, 2008.

[9] Montgomery, A. A. and P. L. Jackson. “Physical characteristics of the lips underlying vowel lipreading performance”. Journal of the Acoustical Society of America, 73(6):2134–2144, 1983.

[10] Nicholls, G. and Ling, D., “Cued speech and the reception of spoken language”. Journal of Speech and Hearing Research, 25:262–269, 1982.

[11] Nicholls, G., "Cued Speech and the Reception of Spoken Language." Master's thesis, McGill University, Montreal, 1979.

[12] Bernstein, L.E., Demorest, M.E. and Tucker, P.E., “Speech perception without hearing”. Perception & Psychophysics 62(2), 233-252, 2000.

[13] Uchanski, R.M., Delhorne, L.A., Dix, A.K., Reed, C.M., Braida, L.D. and Durlach, N.I. "Automatic Speech Recognition to Aid the Hearing Impaired: Current Prospects for the Automatic Generation of Cued Speech". Journal of Rehabilitation Research and Development, Vol. 31, pp. 20-41, 1994.

[14] Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Le, V.B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.F., Tribout, M. and Vidal, S., “TELMA : Telephony for the Hearing-Impaired People. From Models to User Tests”. In: proc. ASSISTH'2007, Toulouse, France, 2007.

[15] Golipour, L. and O'Shaughnessy, D.,”A new approach for phoneme segmentation of speech signals”. In: Proc. Interspeech’07, Antwerp, Belgium, 2007.

[16] Benoit, C., Lallouache, T., Mohamadi, T. and Abry, C., “A set of French visemes for visual speech synthesis”. In: Bailly, G., Benoit, C. (Eds.), Talking Machines: Theories, Models and Designs. Elsevier Science Publishers, Amsterdam, pp. 485-504, 1992.