Characterizing and Classifying Cued Speech Vowels from Labial Parameters

D. Beautemps, T. Burger, & L. Girin

Institut de la Communication Parlée
CNRS UMR5009 /INPG/Université Stendhal, Grenoble, France
denis.beautemps@icp.inpg.fr

Abstract

As part of the *THIMP* project (Telephony for Hearing-IMpaired People), we aim at automatically analyzing Cued Speech [1] and translating it into oral spoken language. This work focuses on vowel classification and will be part of this transcoding process as a preprocessing step of the input data analysis. Its objective is to identify vowels produced by a speaker pronouncing and coding in Cued Speech a set of French sentences, knowing:

- The Cued Speech Hand Placement,
- The analysis of defined Labial Parameters.

Here, we will show that the crossing of these two sources of information allows to automatically identify vowels. These results have to be compared to performances of hearing-impaired people in perception of Cued Speech.

1. Introduction

Manual Cued Speech (CS) [1] is an effective method used to enhance speech perception for hearing-impaired people orally educated. CS is designed to complement speech lipreading and is based on the association of lip shapes with cues formed by the hand placed at specific location. While uttering, the speaker uses one of his hand to point out specific positions around the mouth, palm toward him, so that the speech reader can see the back of the hand simultaneously with lips. The manual cues are formed along two parameters: Hand Placement (HP) and Handshape. Hand Placements code groups of vowels (*Fig. 1*) whereas Handshapes allow to distinguish among groups of consonants.

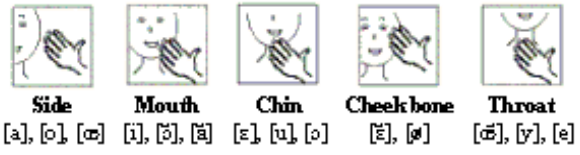

Figure 1: Hand Placement for French vowels (from [2]).

Attina and colleagues [2] observed in their studies on the production of French CS that the information on CV syllable transmitted by the hand is completely available at the beginning of the syllable, thus quasi synchronically with the lips in the case of consonant but largely in advance for the vowels. Moreover, Cathiard and colleagues [3] showed that the advance of the CS manual information is perceptively detected in a "gating" experiment in which manual and lips information on CV syllables are progressively presented to speech readers using CS.

In a decoding task in which the phonetic structure has to be recovered, the fusion of manual and lip information is a main issue. Fusion models for speech perception are classically used for audiovisual speech identification, that is joint perception of audio speech (eventually in background noise) and lipreading (see e.g. [4], [5] for a complete review). In the CS case, the two sources of information (hand gesture and lip movement) are both perceived by vision. Moreover, none of them carry out the complete information on the phonetic code contrarily to the auditory channel in noise-free environment. It is rather the intersection of the two channels (hand and lips) that allows phonetic identification. Apart the study of the optimal integration structure (see the fusion problem pre- or post-classification in audiovisual speech), one of the difficulties is to take into account the temporal component, i.e. the specific synchrony between information carried out by the hand and lips.

To give a first contribution to this problem of CS sensors fusion, this work focuses on the French vowel classification from lip outer contours [6], given that the vowels to be identified are grouped in a CS Hand Placement category. The information transmitted by the hand was thus separated from lip information and was considered as known (in the *THIMP* project, other works are currently being investigated in automatic Hand Placement analysis and identification). Labial Parameters were derived from the 3D labial contours (*Fig. 2*).

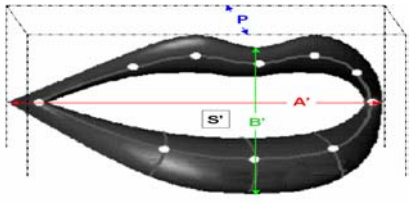

Figure 2: Lip Outer Width (**A'**), Lip Outer Aperture (**B'**), Lip Outer Area (**S'**) & Lip Upper Protrusion (**P**) parameters.

2. The Classifying method

2.1. The Gaussian Classifier

A Gaussian model was used to classify vowels using the Labial Parameters information: For each vowel, these Labial Parameters are supposed to follow normal laws. This kind of classifier was chosen mainly for its simplicity. Moreover, it may be generalized in future studies with Hidden Markov Models, as we may want to complete it with a lexical or contextual analysis.

During the learning phase, a 4-dimensional (in the {**A'**,**B'**,**S'**,**P**}-space) Gaussian model was built for each of the vowels using the occurrences of the vowels in the learning corpus (*Fig. 3a*).

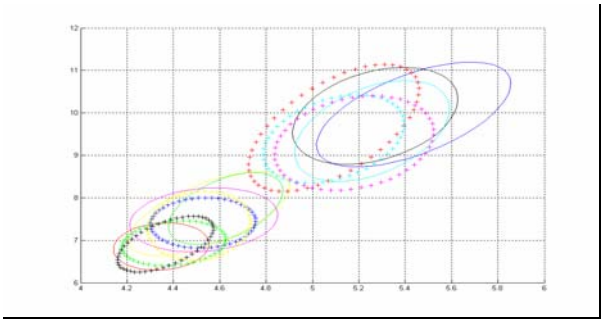

Figure 3a: Result of the modeling in the [**A'**(cm),**S'**(cm²)] plan for all the vowels - 1.5 standard deviation ellipsis.

In the data corpus, all the {**A'**,**B'**,**S'**,**P**} set that correspond to vowels, should be classified correctly among all the vowels in their HP group (*Fig. 3b*). Thus, in the decision phase, the Hand Placement being given, the classifier calculates the probability of the {**A'**,**B'**,**S'**,**P**} quadruplet, for each of the three vowels of the HP group (two vowels in case of Cheek bone HP). The quadruplet is affected to the vowel with the highest probability value.

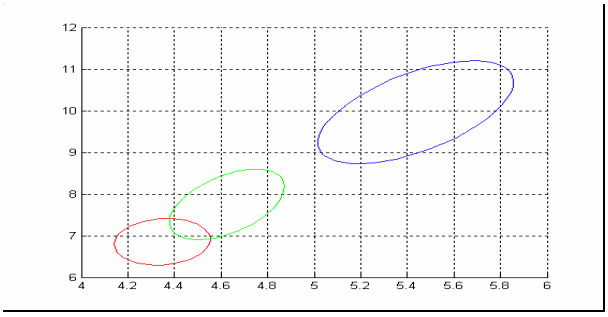

Figure 3b: French vowels of the Chin HP in the [**A'**(cm),**S'**(cm²)] plan. From left to right: [**u**], [**ɔ**] and [**ɛ**] - 1.5 standard deviation ellipsis.

The same corpus was used for the two phases, but the learning phase only generated the probabilistic models and none of the result was saved for the decision phase. Thus, the predictive power of the classifier is slightly lowered, but it is non significant considering to our goal: We aimed at showing that *the Labial Parameters, known as perceptively pertinent ([7], [8]), were efficient enough for an automatic classification task based on a Hand Placement preprocessing*.

2.2. The vowels corpus

The Corpus was made of 182 sentences containing 1974 vowels (*Fig. 4*). Concerning the vowel repartition through HP Groups, some groups had unbalanced composition that needed to be balanced by normalizing the corresponding Gaussian laws. These HP Group are written into brackets ( ):

- The Side Group was so unbalanced, that the classification would not have been efficient without normalizing all the Gaussian laws of the group.
- The Throat Group contained the [œ̃] vowel, which had a wide dispersion. Therefore, we should normalize the classes of this group too, in order to prevent that class to become a "dustbin".

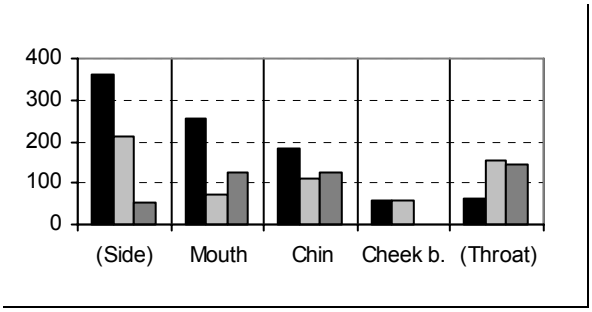

Figure 4: Vowels repartition in the data corpus.

2.3. Lip parameters

The data were made of the (x, y, z) coordinates of 63 points placed on the left side of a CS speaker's face and recorded while the speaker was pronouncing and coding in CS the set of 182 French sentences. When needed, the points have been corrected by using a virtual 3D model in order to be more coherent and more robust: The coordinates of this set of 63 points were linearly predicted by a set of 7 Face-Deformation Parameters derived from a factor analysis with a maximum error less than 2 mm in the reconstruction of the data [9].

Among all the face points, the A01..A06 and ST02 were the points considered to compute the Labial Parameters: A01 and A02 are the extreme points of the Cupidon's Arch, the next coming A03, A04, A05 and A06 define the left half external border of the lips. ST02 is a reference point (*Fig. 5*).

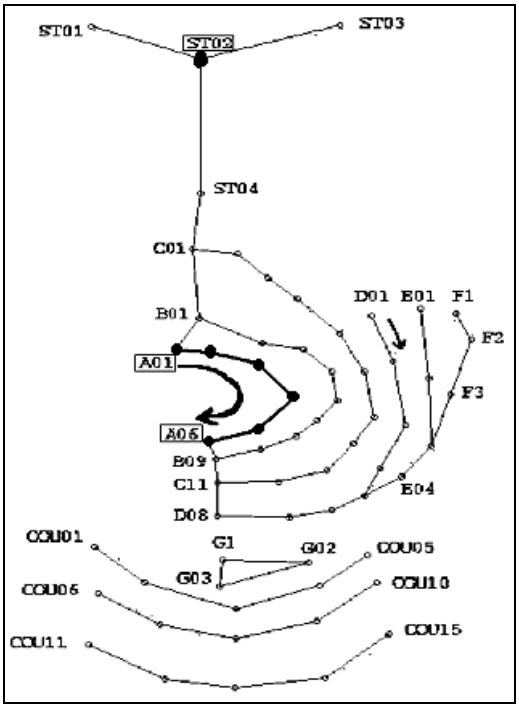

Figure 5: The 63 marked points on the left side CS speaker's face, with the 7 specific points [9].

The inner parameters are known to be the more efficient for vowel perception [7], but it was impossible to compute them, since no reflectors could be placed on the vermillion of the lips. Thus the Labial Parameters were derived from the outer contour of the lips, since the correlation between the inner and outer parameters is known to be very strong [10].

The Labial Parameters **A'**, **B'** and **S'** were computed from a two-dimensional projection of the 3D lip contour.

Let A12, the middle of the Cupidon's Arch [A01, A02]. Thus, the {A12, A02, A03…A06} chain exactly defines the left side of the lips. Let PM, the symmetrical plan of the face, containing {ST02, A12, A06}. We computed the 4 Labial Parameters as follows:

- **B':** B' is just the Euclidean norm of the [A12,A06] segment.
- **A':** Let A04' be the A04 orthogonal projection on PM. A' is twice the Euclidean norm of the [A04, A04'] segment.

Let PP be the plan containing the {A04, A04'} points and the (A12, A06) axis. Basically, this plan is parallel to a mirror in front of which the speaker is performing. This is on PP that the lips border is supposed to be projected.

- **P:** Let $\overrightarrow{Npp}$ be the PP normal vector. Let $\overrightarrow{A12}$ the vector whose coordinates are those of the A12 point. P is simply defined as the norm of the following scalar product:

$$P = \left| \overrightarrow{Npp} \cdot \overrightarrow{A12} \right| \quad (1)$$

- **S':** Let's project {A12, A02, A03, A04, A05, A06, A04'} on PP, and compute the flat area of the convex hull of these 7 points by the triangles method. The result is multiplied by 2 to give S'.

$(A' \times B')$ and S' are correlated at 99.58%. Moreover, for **M'** parameters, defined as, $S/(A \times B)$, the mean is,

$$\overline{M'} = \overline{\left(\frac{S'}{A' \times B'} \right)} = 0.6812 \quad (2)$$

whereas the literature [5] gives for $M = S/(A \times B)$,

$$\overline{M} = \overline{\left(\frac{S}{A \times B} \right)} \approx 0.75 \quad (3)$$

which seems logical as the vermilion is fatter beside the axis where **B'** is measured.

3. Vowels Classification

3.1. An upper bound for the classification score

In a first step, the classification was done from the 7 Face-Deformation Parameters. As they explain the entire variation of the modeled data, a classifier that would work on this 7-dimensions space would be the most accurate one, regarding the corpus. This gave us an upper bound for the Labial Parameters' result.

3.2. Results of Labial Parameters-based classification

Compared to that bound, the results with the Labial Parameters were only, in the worst case, 5 points smaller on each HP Group, (knowing that we only use 4 parameters instead of 7).

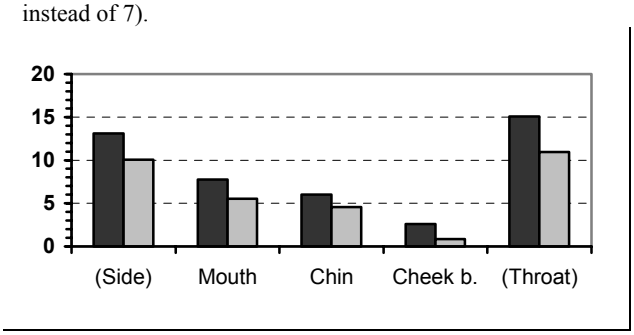

Figure 6: Error rate - Comparison between the Labial Parameters (Black) and the upper bound from the Face-deformation Parameters (Gray).

3.3. Efficiency of the Labial Parameters

We could compare the 2 classifiers regarding the validity of the parameters. On one hand we have the 4 parameters and on the other hand, we have 7 of them.

After this, we have compared the 4 best Face-Deformation Parameters between the 7 (considering the explanation of the variance), and the 4 Labial Parameters: Even if a tiny advantage is given to the Labial Parameters, the difference is so small that it can be considered as non-significant. It has been shown that the Labial Parameters are used in human perception ([7] and [8]), but nothing proves that they are the most efficient: One could imagine that human just use these parameters because they make the bimodal integration easier by facilitating the projection in the motor space [5]. It might be the case, *but this experiment nonetheless proves on an algorithmic point of view the efficiency of the linguistic models through automatic speech analysis.*

3.4. Interest of the Labial Parameters

Here is the main difference between the two sets of parameters we used: The Face Deformation Parameters set is made of quasi-non correlated parameters as the result of a factorial analysis on the global corpus, whereas the set of Labial Parameters has *a priori* nothing to deal with orthogonality.

Thus, if the number of Face-Deformation Parameters is positively related to the classification rate, it is not *a priori* the case for Labial Parameters: Maybe, in some sub-groups, one of these parameters is redundant, or, worse, maybe it can bring confusion.

So, let us analyze each group of 3 parameters chosen between **A'**, **B'**, **S'** & **P**. There are 4 subsets to be analyzed. For each HP Group, the result is the same: One subset is always clearly better than the others (around the same classification rate as the complete 4 parameters set, and often better, which confirms the hypothesis of the fourth parameters bringing confusion), whereas the others remain far below. The 4 Labial Parameters have therefore to be considered more as a redundant set of vectors than a base of our 4-dimensional

space. One should therefore optimize the choice of the parameters set according to the vowel sub-group.

Hereafter is the comparison between the 4 best Face-Deformation Parameters and the 3 best Labial Parameters for each group (*Fig. 7*): The Labial Parameters allow a lower error rate. It now reaches less than 15% in the worst case: *Our classifier is efficient*. Moreover, 6 Face-Deformation Parameters are needed to reach the same classification score as with the 3 chosen Labial Parameters.

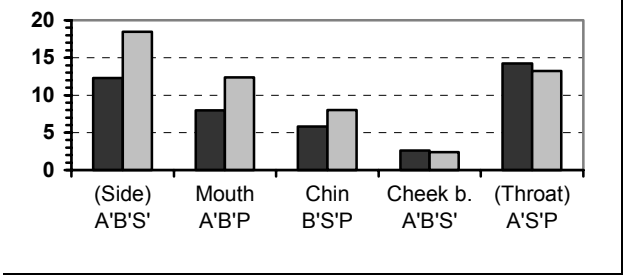

Figure 7: Error rate - Comparison between the 3 best Labial Parameters (black) and the 4 best Face-Deformation Parameters (gray).

These results can be compared with linguistic issues: Cathiard [11] pointed the Labial Parameter that was the most important on lip-reading according to each phoneme: for each HP Group, this crucial parameter is always within our 3 remaining parameters.

4. Discussion and conclusion

As a simple stochastic learning system, our classifying method is the first step toward a Bayesian Filter. The next coming improvements toward a Bayesian Filter are:

- Definition of an orthogonal (or supposed so by hypothesis) set of parameters in order to apply the Bayesian rules in an easy way. The first solution is to demonstrate that the sub-sets of 3 parameters follow independent stochastic for each HP Group, and then generating different Bayesian Filter for each. Another solution is to deal with new parameters. Even if they are not completely uncorrelated, they might be able to be considered as independent enough in a first approximation.
- Use of a stochastic law for the French vowels occurrences.
- Contextual analysis thanks to HMM.
- As this basic "vowel classifier" is bound to evolve toward a significant automatic CS learning processor, it would be interesting to reread the results with an "Artificial Intelligence" point of view:

The 4 Labial Parameters are closer from a redundant set of vectors than a base of our 4-dimensional space. Moreover, their interest remains in their use by human for lipreading. Let us relate these two remarks:

Live beings' vision is known to work on redundant bases of decomposition, because this redundancy allows a certain standard of robustness: It allows, for instance, to interpolate hidden borders objects or to rebuild missing information. The classifier illustrates basically this phenomenon: one can choose the best decomposition form between those that are possible, and then palliate to some noisy component, or simply choose the most powerful one according to the kind of data. The main interest of this remark lay in the confidence we can have in our method (which consists in choosing the best parameters between a redundant set, instead of using a PCA to extract efficient parameters), as this natural behavior works on a stochastic automat that was not programmed for it.

5. Acknowledgements

This work benefitted from interesting discussions with G. Bailly (mainly for the data-mining), G. Gibert, M.A. Cathiard and J.-L. Schwartz. This work is supported by the French THIMP project (so-called "Projet TELMA").

6. References

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

[2] Attina, V., Beautemps, D. and Cathiard, M.-A., Odisio, M., "Toward an audiovisual synthesizer for Cued Speech: Rules for CV French syllables", *Int. Conf. Audio-Visual Speech processing*, 227-232, 2003.

[3] Cathiard, M.-A., Bouaouni, F., Attina, V., Beautemps, D., "Etude perceptive du décours de l'information manuo-faciale en Langue Française Parlée Complétée ", in Proceedings of XXV Journées d'Etude sur la Parole, 2004.

[4] J. Robert-Ribes, J., Schwartz, J.L., and Escudier, P., *A comparison of models for fusion of the auditory and visual sensors in speech perception*, Artificial Intelligence Review, Kluwer Academic Publishers, Norwell, 323-346, 1995.

[5] Robert-Ribes, J., *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à l a reconnaissance automatique des voyelles*, PhD thesis, Institut National Polytechnique de Grenoble, 1991.

[6] Abry, C. Boë, L.J., "Laws for lips", *Speech Communication*, 5, 97-104, 1986.

[7] Jackson, P.L., Montgomery, A.A. and Binnie, C.A., "Perceptual dimensions underlying vowel lipreading performance", *J. Speech Hearing Research.*, 19, 796-812, 1976.

[8] Montgomery, A.A., and Jackson, P.L., "Physical characteristics of the lips underlying vowel lipreading performance", *J. Acoust. Soc. Amer*., 73(6), 2134-2144, 1983.

[9] Gibert, G., Bailly, G., Beautemps, D., Eliséi, F., and Brun, R., " Analysis and synthesis of the 3D movements of the head, face and hands of a speech cuer", *J. Acoust. Soc. Amer*., submitted.

[10] Lallouache, M.T., Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres, PhD thesis, Institut National Polytechnique de Grenoble, 1991.

[11] Cathiard, M.A., "Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français". Mémoire de Maîtrise, Université Grenoble II, 1988.