CUED SPEECH HAND GESTURES RECOGNITION TOOL

Thomas Burger, Alice Caplier** and Stéphane Mancini***

* France Telecom R&D, 28, Ch. Vieux Chêne, Meylan, France
email: thomas.burger@francetelecom.com
** Laboratoire des Images et des Signaux, 46 avenue Félix Viallet, Grenoble, France
email: name.surname@.lis.inpg.fr, web: www.lis.inpg.fr

ABSTRACT

Automatic real-time translation of gestured languages for hearing-impaired would be a major advance on disabled integration path. In this paper, we present a demonstrator on Cued Speech hand gesture recognition.

Cued Speech is a specific visual coding that complements oral languages lip-reading. Its nature provides a simple gestures set which is likely to be automatically and reliably recognized in rather little constraint conditions.

A first PC-demonstrator illustrates the recognition process.

1. INTRODUCTION

As someone speaks, a hearing-impaired can try to guess the oral message by lip-reading. This is a difficult task, for different phonemes correspond to identical mouth shapes.

In order to improve the lip-reading efficiency Dr. Cornett developed the Cued Speech [1]. He proposed to add manual gestures to lip shapes so that each sound has an original visual aspect. Such a "hand & lip-reading" becomes as meaningful as the oral message.

Cued Speech is based on a syllabic decomposition: the message is formatted into a list of "Consonant-Vowel syllable" (a CV list). Each CV is coded with a specific gesture, which is combined to its lip shape, so that the whole looks unique and understandable. A gesture contains two pieces of information: a handshape (for the consonant coding – fig 1a) and a location around the face (for the vowel – fig 1b).

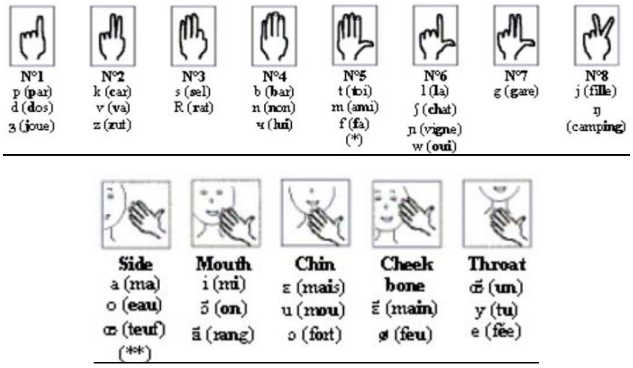


Figure 1: (a) The 8 handshapes for the 20 French consonants. **(b)** The 5 locations with respect to the face for the 15 French vowels.

Hand coding brings the same quantity of information than the lips movement: it is as difficult to lip-read without gestures as to understand the hand coding without lips movements. This symmetry explains why a single gesture codes several phonemes, which correspond to different lip shapes. Thus, there are only 8 handshapes and 5 locations for a combination of 40 CV-gestures.

We aim at automatically recognize in real-time a succession of Cued Speech gestures. By coupling such a device with an automatic lip-reading module and others various automates, a complete hearing-impaired translator could be feasible.

Our project undergoes few general restrictions:

- Cued Speech gestures occur in a plane space, which corresponds to ID photographs frame. So, we guess 2D acquisitions might provide enough information for the recognition process [2].
- The coding hand will always wear a one-coloured glove in order to make the hand segmentation easier. This is not a strong constraint for Cued Speech coders.
- We decided to add a 0th handshape corresponding to a closed wrist (absence of coding) for a total of 9 handshapes.

Then follow the restrictions dedicated to the proposed demonstrator:

- We are not going to focus on real time aspects yet.
- We only deal with the handshapes recognition issue. The location with respect to the face is beyond the scope of this paper.
- The dynamic aspect of Cued Speech can only be processed on fluent coder recorded data, for realistic coding is not trivial. On the contrary, anyone can present one of the 9 static possible handshapes in front of a webcam.

We will perform two versions of the demonstrator: in the first one, we will show how to process the dynamic aspect of coding, such as linking the continuous movements of the hand to the discrete chain of corresponding phonemes. In the second version, we will face our demonstrator to various static coders for simple recognition purpose.

In the second, third and fourth parts of this article, we will describe our general main processing steps: hand segmentation, handshape modelling and gesture classification. Finally, the fifth part gives a description of the proposed demonstrators.

2. HANDSHAPE SEGMENTATION

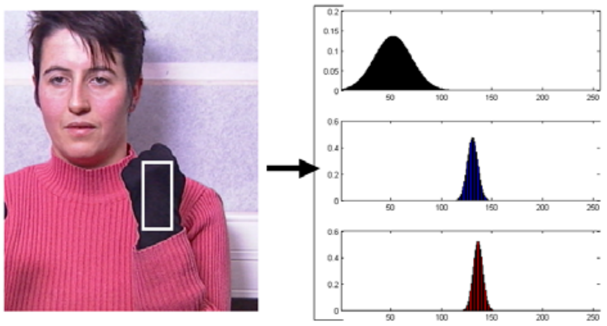


Figure 2: (a) Colour glove analysis, **(b)** Y, Cb & Cr projections of the 3D-Gaussian.

We suppose that the coding hand is wearing a one-coloured glove (of undetermined model) because, the segmentation of a bare hand moving in front of the face would be a too difficult issue regarding to the required accuracy and robustness. Moreover, such a glove copes with rings or nail varnish reflections by hiding them.

Before each use, one should learn the colour glove: on the first image, the statistical repartition in the YCbCr space of pixels from an inner glove rectangle (fig. 2a) allows computing the tri-dimensional Gaussian parameters, which define its colour (fig. 2b).

The next coming user's images (fig. 3a) are converted into a similarity colour map by computing pixels values under the Gaussian model: each pixel value of the colour map equals the Gaussian evaluation in the corresponding pixel with respect to its decomposition on the YCbCr coordinates (fig. 3b). At last, a threshold on the similarity map and a connectivity component labelling give the segmented handshape (fig. 3c).

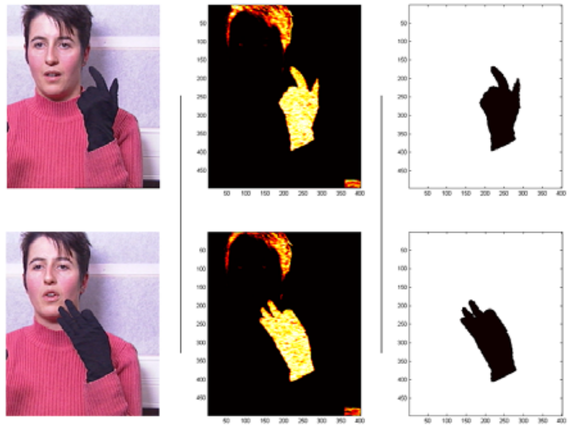


Figure 3: (a) original images, **(b)** similarity colour maps, **(c)** segmented handshapes.

3. HANDSHAPE MODELLING

After the segmentation, the handshape has to be turned into a set of parameters for the classification step. We will provide two sets of such parameters, which are combined into a structural model.

3.1. Mass Measurements Parameters

It comes from basic processing on the binary handshape: inertia moments and gravity centre, principal axes bounding box, mass histograms along principal axes…

Then, these low level measurements provide some other higher level information as the pointing finger (for the next coming location recognition), or some clues concerning stretched fingers (such as presence of the thumb …), which are combined with the next set of parameters.

3.2. Fingers Extraction Parameters

It is made of a selection of pixels, which correspond to plausible fingers. To extract them, we will process three successive transforms.

At first, let us compute the distance transform DIST of the handshape image, which worth is (fig. 4a) [3]:

$$DIST(pixel) = d(pixel, Handshape) \qquad \textbf{(1)}$$

Where the function *d(.,.)* computes the distance between the current pixel and the nearest pixel out of the handshape.

For morphologic reasons, the pixel with the biggest value in the distance transform image is the centre of a circle, which corresponds to the palm.

The next step is to find out the pixels of the distance transform image, which correspond to the watershed lines on a three-dimensional view (fig. 4b). Such a transform allows a good separation between the fingertips to appear (fig 4c), which is useful to model the digits relative positions. Here is described an efficient way to compute it: we consider the pixels, which are near enough from the local maximum value to belong to the watershed. Let us consider a N-by-N neighbouring for each pixel *(x, y)*, and *locMax(N, x, y)* its local maximum. Let *tol* a tolerance rate (around 15%) and *WST* the image transform of the original image *IM*.

$$WST(x,y) = \begin{cases} 1 \;\; if \;\; IM(x,y) \geq locMax(N,x,y) \times (1 - tol) \\ 0 \;\; else \end{cases} \qquad \textbf{(2)}$$

The last transform keeps the pixel which neighbouring is composed of a low rate (around 1/3) of pixels belonging to the previous transform image. This operation selects the thinnest elements of the image, which correspond to plausible fingers (fig. 4d).

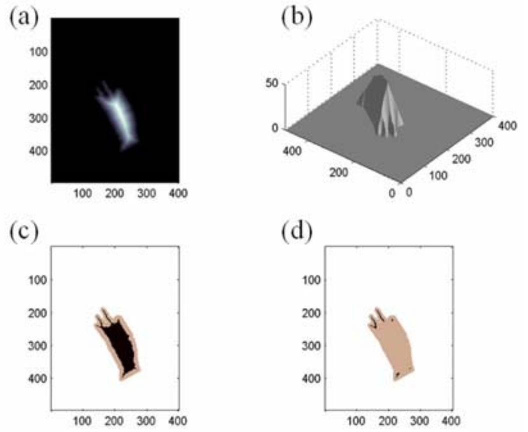


Figure 4: (a) Distance transform, **(b)** 3D representation, **(c)** approximated watershed image, **(d)** plausible digits image.

3.3. Structural Model & Plausible Fingers Analysis

The structural model is produced on the two previous sets of parameters (mass measurement and fingers extraction): each group of connected pixels corresponding to a plausible finger is evaluated regarding to these parameters, and some other criteria such as:

- Its self-mass measurements: gravity centre, orientation, inertia moments, surface…
- Its position toward the palm and other fingers,
- Its possibility to be the thumb.

These evaluations lead to classify each plausible finger in one of these clusters:

- **LongFingers**, which are known to correspond to single complete, stretched and isolated fingers (fig. 5a): each one counts as **1** finger.
- **SmallFingers**, which might be one finger tip (count as **1**) or fake alarms (count as **0** - fig. 5b).
- **FatFingers**, which are quiet thick and could correspond to **1**, **2** or **3** fingers (fig. 5c).
- **WrongFiner**, which are too small, bad oriented, or echoing an obvious other finger. They are erased (fig 5d).

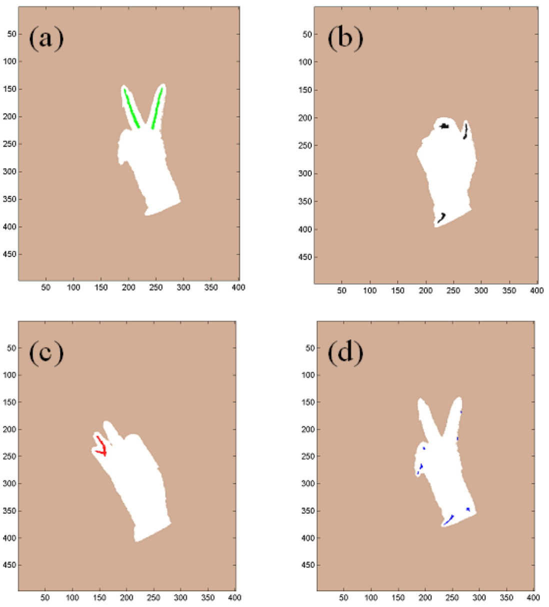


Figure 5: (a) LongFingers**, (b)** SmallFingers, **(c)** FatFingers, **(d)** WrongFiner

4. HANDSHAPE CLASSIFICATION

4.1. The current basic classifier

Let us attribute a value of 1 for each of the 4 fingers and a value of 0.5 for the thumb. Then, we have got this correspondence table between the total value **Vref** and the handshapes **HS** (handshapes numbering is given on fig.1):

V_{ref}	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
HS	0		1	6	2 8	7	3		4	5

Table 1: Configurations encoding for recognition

Each handshape has a specific corresponding value **V** (except handshapes 2 & 8 which for a thresholding on the angle between the two stretched fingers is discriminating).

Let us compute an interval including the real number of digits, based on the structural modelling hypothetic number of each kind of finger.

For example, let us consider a handshape with a SmallFinger (**V** = {0,1}), a FatFinger (**V** = {1,2,3}) and 2 LongFingers, including a thumb, (**V** = {1}+{0.5}). It may not be the structural model of the handshape # 0, 1, 2, 3, 4 and 8, for there is a thumb. Nor may it be the #6 as we have a minimum of 3 stretched fingers including the thumb. It may be the # 5 (the FatFinger corresponding to 3 real fingers) or the # 7 (the FatFinger corresponding to 1 real finger). Let us find this result, as the classifier would do it. We will compute the values corresponding to that structural model:

$$\{0;1\}+\{1;2;3\}+\{1.5\}=\{2.5;\ 3.5;\ 4.5;\ 5.5\} \qquad \textbf{(3)}$$

Plausible handshapes correspond to the V_{ref} set associated to admissible configurations (cf. table 1):

$$\{V_{ref}\}=\{0;\ 1;\ 1.5;\ 2;\ 2.5;\ 3;\ 4;\ 4.5\} \qquad \textbf{(4)}$$

There intersection, which is …

$$\{2.5;\ \ 3.5;\ 4.5;\ 5.5\} \ \cap\ \{V_{ref}\} = \{2.5;\ 4.5\} \qquad \textbf{(5)}$$

… gives the values corresponding to the plausible handshapes. Here, # 7 and # 5.

This stage leads us to a subset of plausible configurations which cardinal usually belonged to [1; 3]. Then, the final clustering is simply performed via naïve ad-hoc rules, but we are looking forward to improving it.

4.2. Some Improvements On The Classification Layer

The first improvement would take place into the Evidence Theory. This theory is a generalisation of probabilities ([4], [5], [6]). It allows to design classifiers [7], which cope with contradictory sources of information, or to deal with incomplete knowledge, such as the possibility not to choose between several clusters by grouping them in a super-cluster. Moreover, it allows the combination of information of different kind to make a decision. We are looking forward to designing a classifier based on this theory to cluster the structural model of the handshape.

The second one would deal with hand tracking. The recognition could be set more robust by taking into account the dynamic aspect and rhythm of the hand movement. This can both be done thanks to the Bayesian Theory (Markovian Filters) or through The Evidence Theory (Transfer Belief Model [8]).

5. DEMONSTRATOR SPECIFICATIONS

The first part of the demonstrator deals with the interactive aspect of our topic. Video sequences come from a webcam with a non-trained coder performing static gestures corresponding to one of the 9 handshapes. Each image of a video sequence is recognized thanks to the process previously explained.

The Graphical User Interface (GUI - fig.6) is made of several elements. The first one is a simple camera output. The second one is a graphical scheme of the structural modelling, which is made of:

- Handshape axis,
- Palm and wrist representations,
- LongFingers (normal lines), FatFingers (thick lines) and SmallFingers (light lines).
- Thumb tip (green) and pointing finger tip (red).

The last element of the GUI displays a set of 9 icons, beyond which the ones corresponding to the plausible handshapes clusters are enlighten.

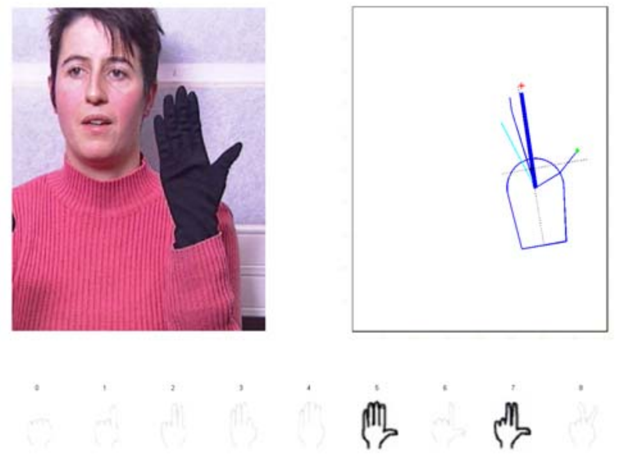
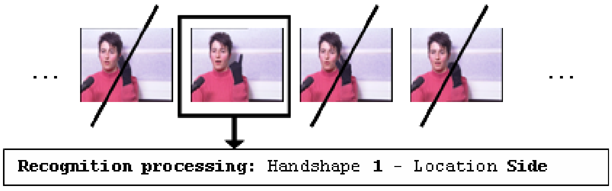


Figure 6: Handshape Recognition Demonstrator GUI.

The second version of the demonstrator runs on recorded data with a professional coder performing on it. It is an easy way to picture how the device would work in real conditions. Moreover, the realness of movement rhythms in such a video may allow the use of a Target Detector, which is able, with only a few frames delay, to extract the important pictures before analysing them. An important picture is a picture where the hand gesture reaches exactly the handshape to be realised in contrary to a transition between two handshapes.

The Target Detector is an algorithm, which heavily relies on the Event Detector basis described in [9]. The Event Detector is inspired on retinal and human cortex processing and quantifies the motion per image regarding the previous few images.

In our case, we are chasing the images where the hand slows down in order to reach a handshape realization. As the main moving objects of our video picture are the hand and the fingers, the local motion minima of the global image actually correspond to fully realized handshapes and locations. Then, the corresponding images (the so-called important one) are extracted from the whole film (fig. 7). They are the only images to be analysed, as they contain the whole series of gestures the coder realized.



Recognition processing: Handshape 1 - Location Side

Figure 7: Targets Detector extracting important pictures.

6. CONCLUSION

We presented a demonstrator on automatic Cued Speech manual gesture recognition. As a first prototype, few aspects of the processing scheme are fully efficient, whereas others need improvement. Segmentation and modelling layers are mature, as:

- Their visual rendering are easily understandable,
- They provide a good enough basis for the following layers.

On the contrary, two aspects of the classification layer have to be improved: its accuracy (thanks to the evidence theory), and its robustness (thanks to tracking techniques).

Finally, we will cope with the real time aspects, as they might need dedicated hardware architecture.

ACKNOWLEDGMENTS:

This work was supported by the Similar network of excellence**.**

REFERENCES

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

[2] A. Caplier, L. Bonnaud, S. Malassiotis, & M. G. Strintzis, "Comparison of 2D and 3D Analysis For Automated Cued Speech Gesture Recognition", *SPECOM 2004.*

[3] T. Morris, O. S. Elshehry, *Hand segmentation from live video*, in The 2002 Intl. Conference on Imaging Science, Systems, and Technology, UMIST, Manchester, UK, 2002.

[4] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey, 1976.

[5] P. Smets. "The combination of evidence in the transferable belief model". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5): 447–458, 1990.

[6] T. Denoeux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7): 1095–1107, 1997.

[7] T. Denoeux. "A k-nearest neighbour classification rulebased on Dempster-Shafer theory". *IEEE Transactionson Systems, Man and Cybernetics*, 25(5): 804–813, 1995.

[8] P. Smets and R. Kennes. "The transferable belief model". *Artificial Intelligence*, 66(2): 191–234, 1994.

[9] A. Benoit, A. Caplier, “Motion Estimator Inspired from Biological Model for Head Motion Interpretation”, *WIAMIS05*, Montreux, Suisse, April 2005