Retargeting cued speech hand gestures for different talking heads and speakers

Gérard Bailly, *Yu Fang*, *Frédéric Elisei*, *Denis Beautemps*

Dept. of Speech & Cognition, GIPSA-Lab, CNRS - Grenoble Universities, France
frederic.elisei@gipsa-lab.inpg.fr

Abstract

Cued Speech is a communication system that complements lip-reading with a small set of possible handshapes placed in different positions near the face. Developing a Cued Speech capable system is a time-consuming and difficult challenge. This paper focuses on how an existing bank of reference Cued Speech gestures, exhibiting natural dynamics for hand articulation and movements, could be reused for another speaker (augmenting some video or 3D talking heads). Any Cued Speech hand gesture should be recorded or considered with the concomitant facial locations that Cued Speech specifies to leverage the lip reading ambiguities (such as lip corner, chin, cheek and throat for French). These facial target points are moving along with head movements and because of speech articulation. The post-treatment algorithm proposed here will retarget synthesized hand gestures to another face, by slightly modifying the sequence of translations and rotations of the 3D hand. This algorithm preserves the co-articulation of the reference signal (including undershooting of the trajectories, as observed in fast Cued Speech) while adapting the gestures to the geometry, articulation and movements of the target face. We will illustrate how our Cued Speech capable audiovisual synthesizer – built using simultaneously recorded hand trajectories and facial articulation of a single French Cued Speech user – can be used as a reference signal for this retargeting algorithm. For the ongoing evaluation of our algorithm, an intelligibility paradigm has been retained, using natural videos for the face. The intelligibility of some video VCV sequences with composited hand gestures for Cued Speech is being measured using a panel of Cued Speech users.

Index Terms: Cued Speech, hand motion retargeting, augmented speech

1. Introduction

Cued Speech (CS) is a communication system that uses a manual component to complement the natural lip-reading ([1]). Firstly, CS is improving speech perception to a large extent for deaf people ([2] for the identification of the syllables, [3] 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 ([4]).

Producing Cued Speech content might be useful in many contexts (e-learning, telecommunication, CS perception experimentations…). A few synthesizers have been proposed commercially or in the literature for this purpose (see [5] for a review). Capturing hand motion or developing control module for such synthesizers is a time consuming process, especially when taking into account the peculiar timing of hand gestures relative to the face articulation. We believe that being able to adapt (retarget) existing hand gestures – captured or synthesized – to another face might be useful to study Cued Speech and address the applicative needs. Many existing 3D talking heads would become Cued Speech capable if they could be enhanced with such a retargeting algorithm. With suitable audiovisual algorithms from speech recognition and computer vision, even video footage of an oralizing speaker could be augmented in a Cued Speech version.

Part 2 will present details of French Cued Speech system as well as the ambitious TELMA scenario where a deaf Cued Speech user and a normal hearing people can have a mediated communication through a special audiovisual terminal. Part 3 will present the existing text-to-audiovisual-speech synthesizer that was used as a basis for the work presented here. Part 4 will present the specific retargeting algorithm proposed here, that allows us to adapt existing Cued Speech hand gestures to another 3D talking head. Part 5 will present the stimuli and the perception test that we developed as an evaluation framework for this algorithm. The preliminary results obtained with a few subjects will also be discussed here. Finally, part 6 will present the perspectives of our work.

2. Cued speech

Cued Speech 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 (for a complete study on the coordination between speech and the CS components, see [5]). Since Cornett ([1]), CS has been adapted to more than 50 languages worldwide. Whatever the language, the common principle is to complement each uttered consonant-vowel (CV) with a manual cue. A manual cue in this system contains two components: the shape of the hand and the hand placement relative to the face. Handshapes are designed to distinguish among consonants whereas hand placements 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. Figure 1 describes the system for French language.

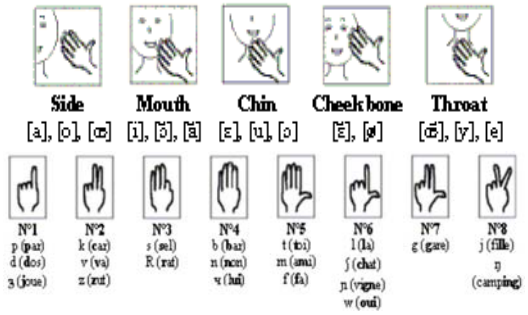

Figure 1: *CS Hand position (top) for vowels and CS handshapes (bottom) for consonants for French language (derived from [5]).*

2.1. Face and hand coordination in Cued Speech

The movement of the CS manual component must be closely related with speech and especially its face component to be effective for speech perception from visual output. Attina and colleagues ([5]) demonstrated that the CS manual component is not synchronized with the non-manual mouth movement of CS but anticipates it, and that CS readers take advantage of this anticipatory phenomenon. The authors showed that for CV syllables, the hand anticipates in order to allow the hand placement being attained in the consonant and thus largely before the corresponding vowel lip target (Figure 2). The CS technologies of automatic CS generation have to take into account this complex coordination of the CS manual component and the non-manual mouth component of speech to be effective for speech perception.

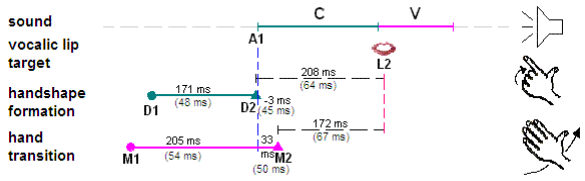

Figure 2: *Temporal pattern of coordination between sound, lips, handshape formation and hand transition for Cued Speech production in French language. Average value and standard deviation (between brackets) are shown (Adapted from Attina et al., [5]).*

2.2. The TELMA project

The demand of handicapped people to access to communication technologies is a major concern in modern society. The on-going TELMA research project ([6]) aims at easing telecommunication in French between a deaf person using CS and a normal-hearing speaker. More precisely, the project proposes to develop an automatic translation system of acoustic speech towards visual speech completed with CS using 3D CS synthesis 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 not familiar with Cued Speech with the help of the autonomous terminal TELMA. Compared with previous approaches of CS synthesis (see for example [5] and [7]), we do use a 3D model for the cueing hand, which is more easily adapted to various viewing conditions. The handshapes transitions, the temporal trajectories of their positions, as well as their coordination with the face articulation were accurately motion-captured in our system.

3. A Cued Speech synthesizer for French

A text-to-speech synthesizer (TTS) for French Cued Speech has already been developed in one of our previous research project [8].

3.1. Capturing 3D Cued Speech for synthesis

The shape and appearance of the face and the hand of a real French Cued Speech user have been captured by photogrammetry and plaster molding, while the dynamics of her articulation and gesture was recorded using a 120Hz motion capture system. The resulting 3D clone can be seen on Figure 3.

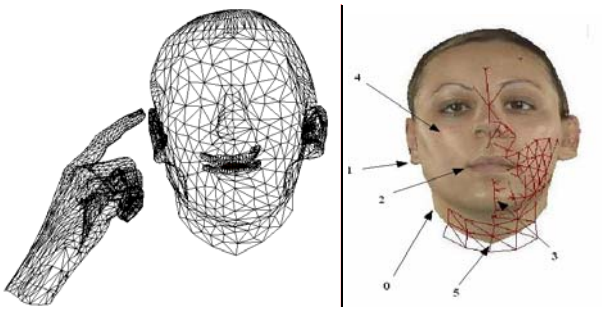

Figure 3: *The Cued Speech capable talking head. Left: wire frame rendering of hand, face and teeth. Right: textured rendering of the face with positions 1 to 5 as targeted by French Cued Speech gestures.*

3.1.1. Degrees of freedom of the face and the hand

The 3D models of the face and the hand of the Cued Speech user emerge from the recorded data by statistical analysis. The 3D face is controlled by 7 articulatory parameters and 6 movement parameters. The articulatory parameters are related to the speech degrees of freedom (jaw, lips…) and do influence the full face. The movement parameters correspond to the rotation and translation of the head skull relatively to the torso and also affect the neck area. The 3D hand has 9 articulatory parameters (to animate the fingers) and 6 movement parameters (moving and orienting the wrist).

It has to be noted that realizing a particular cue requires values for the hand articulations (key shape) but also adequate values of **both movement parameters sets** (face and hand) to align the correct finger with the correct fleshpoints target on the face. Not only are shapes and degrees of freedom collected with this data, but so do the temporal aspects of the trajectories and the Cued Speech specific prosody. A large corpus of sentences was recorded and processed to form a rich set of temporal trajectories that serves as gesture units in our dictionary-based synthesizer.

3.1.2. Synthesizing new sentences and video dubbing

This is not the scope of this paper to recall how new cued speech sentences are synthesized by our system (see [8] for details). In the context of the present work, we just want to recall that the synthesizer can operate with text input as well as with a phonetic chain (with durations for every phoneme). The latter operating mode allows us to post-synchronize the cued-speech capable talking head to an existing audio recording, using speech recognition software and/or HMM networks to perform forced alignment. The synthesizer will first compute the sequence of configurations and positions for the hand, and how to schedule them relative to the acoustic boundaries: indeed, onsets are different for hand shaping, hand moving, acoustic signal and facial articulations (see [5] for a detailed accurate study of these coordinations). Then, the synthesizer will select gesture units and generate the flow of values for all the articulatory and movement parameters (face and hand). Figure 4 shows key instants of the synthesized output for a VCV (vowel-consonant-vowel) example.

3.2. Conclusions

Whereas this data-based process leads to the very life-like dynamic of our cued speech capable talking head, capturing, correcting and labeling all these units indeed is a time consuming procedure. This has so far limited the

development of new Cued Speech capable talking heads, as well as their applications in the society or for scientific researches. The algorithm proposed in the next section proposes to reuse previous work on Cued Speech modeling and synthesizing to develop new high quality Cued Speech capable systems.

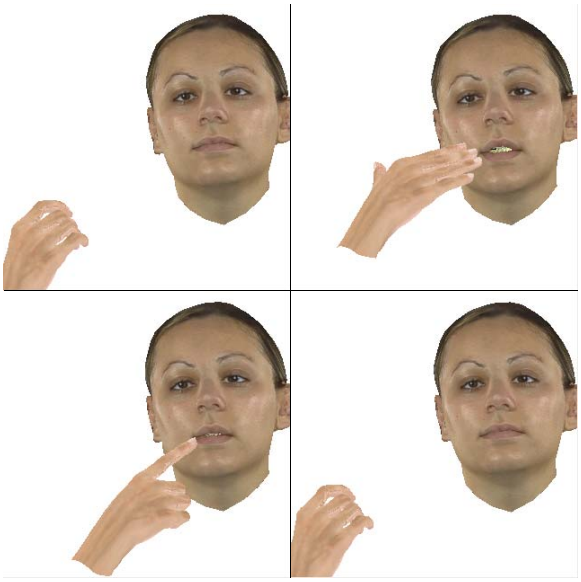

Figure 4: *Four frames of a synthesized CS [idi]. From left to right, top to bottom: t_0 initial rest position, t_1 instant of realization of the manual cue for [i], t_2 realization of the manual cue for [di], t_3 final frame with rest position and shape.*

4. A hand gesture retargeting algorithm for Cued Speech

We suppose here the existence of other 3D heads for which Cued Speech capability is missing but needed. Any such 3D head will be considered a **target face** for our algorithm. We will also consider **target sentences**: these are speech sentences for the target face where acoustic, phonetic chain and 3D positions of face and facial points are available. Such sentences are obtained by for example tracking audiovisual recording of a real speaker [9], or can be produced by audiovisual synthesis [8]. For our algorithm, the available Cued Speech hand and face models are the **reference models** and the synthesized Cued Speech outputs will be considered the **reference signals** that will be retargeted. The algorithm that we propose in this section will augment the target sentences in a Cued Speech audiovisual signal, by adding the missing hand trajectories. These hand trajectories will be adapted from the reference signals produced by the Cued Speech synthesizer (see Figure 5).

4.1. Retargeting the synthesized CS signals

The reference signals have been produced by the synthesizer, using the phonetic chain of the target sentence. The reference signals are therefore in synchrony with the target acoustic signal ("synchrony" of course includes expected anticipation effects with hand movement and shaping, lip rounding…)

This implies that the articulation of the fingers will be the same for the target than in the reference sequence. They should not be modified, as they correctly encode the shape of the hand key. On the other hand, the target face is likely to have a different geometry than the reference face. It is likely to have different facial articulations and another strategy for his or her head movements. Failing to adapt the position of the hand to suit the new face geometry and to compensate for the potential differences in head movements will fail to preserve the accurate and meaningful relative positions of hand and face that Cued Speech interpretation needs.

In our case, retargeting the hand gestures for another face is a matter of correcting the position and orientation of the 3D hand and faces.

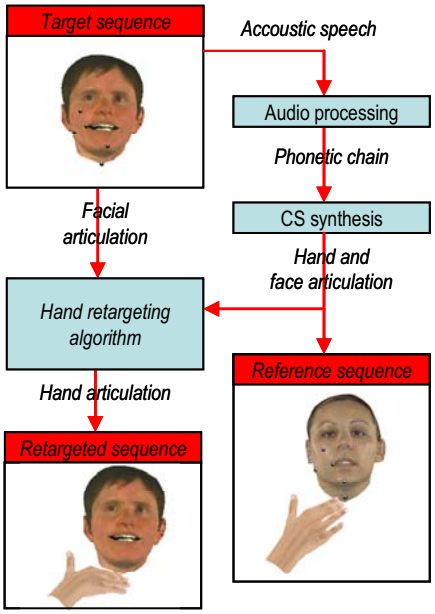

Figure 5: *The retargeting methodology: The phonetic chain derived from the target sequence is used to synthesize a reference Cued Speech articulation. This will be used to adapt the 3D hand to the target facial articulation.*

4.2. Correcting for the rotations

In our head modeling framework, both reference and target talking heads orientations are expressed in the same frame set: with a null movement, any of our 3D head model will face the camera straightaway. The Cued Speech synthesizer does generate head movements for the reference talking head that probably convey prosodic information and might help segmentation of the Cued Speech units. To get a chance of being perceived, these movements should be transferred to the target model: either as extra target head movements or as extra hand movement (reverse movement) to preserve the spatial relationship with the head.

4.3. Correcting for the translation

To correct the hand in translation, we defined 6 facial locations on the CS-capable reference head (see right part of Figure 3). The last five locations correspond to the facial positions specified for French Cued Speech (side, lip corner, chin, cheek and throat, as in Figure 1). R_0 plays an analog role for a rest position (hand sideway down from the face for our reference speaker).

Analogs of these six points must be also defined for a target face. Along time, during a speech sequence, these target points are moving as $T_i(t)$ whereas the six reference points for the same speech sequence are moving as $R_i(t)$.

For every sequence, we define six possible translations, as functions of time that would displace one of the target points to the position of its reference counterpart:

$$\overrightarrow{D_i}(t) = \overrightarrow{R_i(t)T_i(t)}, \forall i \in [0,5] \quad (1)$$

Translating the reference hand with $\overrightarrow{D_i}(t)$ for a complete sequence will ensure that translated keys at position *i* (for example i=3 for chin) will effectively address the chin in the retargeted sequence. As we need to correct for other target locations also, we define a composite transformation as:

$$\overrightarrow{D}(t) = \sum_{i=0}^{i<6} w_i(t).\overrightarrow{D_i}(t) \quad (2)$$

where every $w_i(t)$ acts as a weight for every individual translation. The set of $w_i(t)$ should follow the following properties:

$$w_i(t) \in [0,1], \forall i \in [0,5] \quad (3)$$

$$\sum_{i=0}^{i<6} w_i(t) = 1 \quad (4)$$

For every segmentation intervals $[t_j, t_{j+1}[$ delivered by the Cued Speech synthesizer for the successive hand positions, we just define values for the $w_i(t)$ functions according to:

$$w_i(t \in [t_j, t_{j+1}[) = \frac{t_{j+1} - t}{t_{j+1} - t_j} P_{i,j} + \frac{t - t_j}{t_{j+1} - t_j} P_{i,j+1} \quad (5)$$

where $P_{i,j} = 1$ if hand cue position is *i* at time t_j, and $P_{i,j} = 0$ else. This formulation will construct linear ramps that peak at value 1 only when a translation is to be applied with the corresponding key.

Figure 6: *Frames 32 and 56 of the retargeted [idi] sequence. Left image shows the hand realization of the initial [i] cue (isolated vowel, with handshape 5) Right image corresponds to the instant of hand realization of the final [di]*

4.4. Behavior of the algorithm

The proposed algorithm will not modify the translation (respectively the rotation) of the originally synthesized signal if the targeted talking head is the reference one. This good property also means that the algorithm does not enforce that the CS targets get reached if they were not recorded so.

By construction, adding a translation or rotation to the target head sequence gets the exact same transformations added to the 3D retargeted sequence.

Figure 6 shows a few retargeted frames constructed with our algorithm. The handshapes are ones from the reference sequence that was already illustrated in Figure 4, but their orientation and position have been slightly modified by the algorithm to address the target 3D face. From a Cued Speech interpretation point of view, they exhibit the right semantic.

4.5. Adapting the algorithm to target images

Instead of targeting a 3D model that can be rendered with any head movement, we could also use videos of a talking face as the signal to be Cued Speech augmented, as was already done but in the 2D case [5,7]. Every image of the video could serve as the background where we incrust our 3D hand. This is feasible if we have a projection model (even an approximate camera model) of the scene. This also supposes that we can recover the 3D position of the 6 target points in the video image and the 3D rotations of the head. This can be done using computer vision algorithms, a model of the speaker (our case), or by tracking dedicated markers at the right places. Of course, the target head movements are imposed by the images and we do not plan to modify them. They might be quantitative if the speaker did record free expressive speech without any constraint (or if the camera is moving). In these cases, it is important that the orientation of the 3D hand gets corrected before it gets incrusted in the image.

First, the hand should undergo an extra rotation to preserve the relative rotation that is applied to the reference head and will not be applied to the target head: we apply the inverse of this transformation to the hand. Secondly, the hand should undergo the exact same 3D rotation that the target head presents in the data. With these two transformations, the target hand and face frames will display the same relative rotational differences than the references. The translation correction step, applied in the skull-related frame, remains exactly the same. Figure 7 illustrates some results.

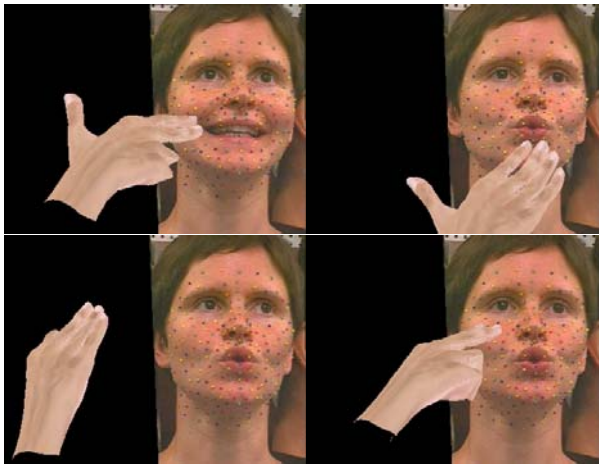

Figure 7: *Examples of hand retargeting, using images from VCV sequences. From left to right, top to bottom key instants where the synthesized hand is in shape and in position for the CV of [igi] [utu] [oso] [øzø]*

5. An evaluation framework for French Cued Speech

Thanks to the way the translation algorithm works, retargeted hand keys should be at the correct position at the realization instants. The influence on the gesture dynamic is less obvious though. We decided to conduct a subjective evaluation of the intelligibility of the retargeted hand movements. We could have used a 3D talking head as the target of the previous algorithm, but the test might suffer from inaccuracies with tongue and teeth positions from the clone. Instead, we used video stimuli for the face that were augmented with 3D hand gestures by our algorithm.

5.1. The stimuli

We used an audiovisual corpus recorded for a previous experiment dedicated to face modeling. In this corpus, the face of a French female was covered with many glued markers and recorded while uttering VCV non sense words (where V=[a,o,i,e,u,ø] and C=[b,d,f,g,k,l,m,n,p,r,s,ʃ,t,v,z,ʒ]). Secondly, the video recordings of the natural face were augmented automatically with the CS manual component by our algorithm of 3D hand gestures. We thus obtained 96 avi movies with the natural face and 96 avi movies containing the face augmented with the CS manual component.

5.2. The perception test

The perception test was divided into two sessions. The first one was dedicated to the evaluation of the reception of the lip-reading with the stimuli of the natural face. This session has been called *the lip-reading condition*. This condition was used at first and was considered as a base line. In the second session, the addition of the CS manual component was evaluated with the use of the stimuli augmented by the CS manual component. This session was called *the CS condition*. The stimuli of each condition were presented once, following a random order. The participants of the experiment were asked to identify both the consonant and the vowel of the VCVs stimuli presented as avi movies through a home made Matlab interface. The participants were two adults profoundly deaf. They use to practice CS daily, since their youth.

Table 1. *Confusion Matrix of the vowels,* in the lip-reading condition. *In line, responses to the stimuli of the first column. The table cumulates the responses of the two participants (16 presentations x 2 participants, for each kind of vowel).*

		Responses					
		[a]	[o]	[e]	[i]	[u]	[ø]
Stimuli	[a]	30				2	
	[o]		22			4	6
	[e]			4	26	2	
	[i]			4	28		
	[u]				3	29	
	[ø]				1	31	0

Table 2. *Confusion Matrix of the vowels,* in the CS condition. *In line, responses to the stimuli of the first column. The table cumulates the results of the two participants (16 presentations x 2 participants, for each kind of vowel).*

		Responses					
		[a]	[o]	[e]	[i]	[u]	[ø]
Stimuli	[a]	32					
	[o]		31				1
	[e]			27	5		
	[i]				32		
	[u]					32	
	[ø]						32

5.3. The results

The results for vowel and consonant identification have been split into different confusion matrixes.

5.3.1. Identification of vowels

Table 1 shows the results on vowel identification for the *lip-reading condition*. It cumulates the responses of the two participants. If the [a] vowel is well identified, table 1 shows that the [e] and the [ø] vowels are largely perceived as [i] and [u] respectively. These results are not surprising since these groups of vowels belong to the same visemes.

The CS condition (see Table 2) shows the benefit of the addition of the synthesized CS manual component that allows quasi complete disambiguation.

5.3.2. Identification of consonants

For each condition, the participant had to identify 6 times each kind of consonant C inside six various vowel contexts. As previously done for the vowels, the responses of the two participants have been computed in confusion matrixes.

As for the vowel case, the CS condition (Table 4) shows the benefit of the addition of the synthesized CS manual component that allows quasi complete disambiguation. Only an exception has to be noticed in the case of the [k] consonant, mainly perceived as [z], another consonant but belonging to the same CS handshape. The CS manual component in that case is not called into question. The lip-reading is the source of the error, probably because of the view angle and/or the condition of lighting not allowing well seeing the front view of the vocal tract. Indeed, the confusion could occur with the [v] consonant that belongs also to the same CS handshape. But this consonant is well articulated at the lips (with a well visible contact of the lower lip with the upper incisives) in contrary to [k] and [z] for which the place of articulation is not at the lips but inside the vocal tract (a specific contact of the tongue with the palate in case of the [k] consonant and a precise alveolar constriction between the tongue and the upper tooth) and thus not easy visible.

Table 3. *Confusion Matrix of the consonants,* in the lip-reading condition. *In line, responses to the stimuli of the first column. The table cumulates the results of the two participants.*

		Responses														
		b	d	f	g	k	l	m	n	p	r	s	ʃ	t	v	z	ʒ
Stimuli	b	1						1		10							
	d			1					2			1		7			1
	f			9							1	1			1		
	g				3	3			1		2			1		1	1
	k			1	3	1	1		1		1	1		1		1	1
	l				2		3	1			5			1			
	m	1						5		6							
	n		1						2				2	4	1	1	1
	p	2						1		9							
	r				1		1	1		1	6			1			1
	s			1								4	1	2	1	3	
	ʃ				1							1	7	1		1	1
	t									1		2	1	6		2	
	v			10											2		
	z											5	1	3		3	
	ʒ											4	3	1		3	1

Table 4. *Confusion Matrix of the consonants,* in the CS condition. *In line, responses to the stimuli of the first column. The table cumulates the results of the two participants.*

Stimuli \ Responses	b	d	f	g	k	l	m	n	p	r	s	ʃ	t	v	z	ʒ
b	12															
d		11														1
f			12													
g				12												
k					2						1			1	8	
l						12										
m	1						11									
n								12								
p									12							
r								1		10	1					
s								1			11					
ʃ			2		1							9				
t													12			
v														12		
z											1			1	10	
ʒ		4										1				7

6. Perspectives

The first perceptual tests have shown that our target audiovisual stimuli were not perfectly adequate for the evaluation task, because of an inadequate viewing angle and some visible hyper-articulations. For a future perception test with more subjects, we plan to record a new audiovisual corpus dedicated to this evaluation with a more frontal view, normal articulation and about 10 markers only (which would look more ecological). While still using VCVs to first identify some potential weaknesses of our Cued Speech augmenting system, it might be interesting to also use semantically unpredictable sentences and compare the rate of keywords identification with that of previous studies by Duchnowski and colleagues [7]. In the context of the TELMA project, where the normal-hearing speaker might use fast speech and many head movements, we should study how these impact on the intelligibility of the augmented CS signal and how to limit their negative impact if any. For example, we should test whether slowing down the speech rate, as an audiovisual pre-treatment on some selected frames, might help the CS readers to better catch some otherwise too fast transitions. It is likely that difficult ones, maybe some consonantal or vowel clusters, might need this improvement when augmenting acoustic free speech to intelligible CS.

Some modifications of the retargeting algorithm might also be tested. We plan to measure the effect of using non-linear weighting functions of the translations, such as a sigmoid one. The translation vector associated with the CS "side" position is actually computed like the other translation vectors (as the observed displacement between the reference and target points). But these are placed on the face also, which is not as pertinent as for the other vectors. Replacing the computed "side" translation by a constant one, specific to every reference/target face pair, might increase the quality of some hand transitions involving the "side" position.

7. Conclusions

In this paper, we presented an algorithm that can adapt existing Cued Speech hand gestures – previously recorded or carefully synthesized – to a new face with different geometry and unconstrained head movements. Whereas our algorithm is a simple one, it does preserve the eventual hypo-articulation of the original hand gestures rather than enforcing artificial targets or creating new gestures with artificial dynamics.

The algorithm only requires that the positions of the 6 reference points get available at every instants, as well as the head movement. We have shown how this algorithm is straightforwardly used with other 3D talking heads, and how to use with real videos of a speaker.

We also presented an evaluation framework to measure the intelligibility of the retargeted hand gestures. Stimuli were produced for French Cued Speech, mixing 3D gestures for the hand and real video for the face, as would be done in an augmented speech application. The test needs to be complemented with more subjects, but the preliminary results indicate that the retargeting of hand position as computed by our algorithm preserves the very good intelligibility of the reference signals.

8. Acknowledgements

Many thanks to the two participants of the test experiment for having accepted to participate to the different phases of the evaluation. We are also very grateful to the various subjects that accepted to be cloned. Christophe Savariaux and Alain Arnal were also invaluable for their technical expertise in recording the cloned subjects. Motion capture was performed at Attitude Studio, France. Parts of this work were funded by the French RNRT ARTUS project or by the French ANR/RNTS TELMA project.

9. References

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

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

[3] Uchanski, R.M., Delhorne, L.A., Dix, A.K., Braida, L.D., Reed, C.M., Durlach, N.I. Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech. Journal of Rehabilitation Research and Development 31(1), 20-41, 1994.

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

[5] Attina, V., Beautemps, D., Cathiard, M.A., Odisio, M. A pilot study of temporal organization in Cued Speech production of French syllables: rules for Cued Speech synthesizer. Speech Communication 44, 197-214, 2004.

[6] 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., Jutten, C., Le, V.B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.F., Tribout, M., Vidal, S.. TELMA : Telephony for the Hearing-Impaired People. From Models to Users Tests. In: proc. ASSISTH'2007, Toulouse, France, 2007.

[7] Duchnowski, P., Lum, D., Krause, J., Sexton, M., Bratakos, M. & Braida, L. D. Development of speechreading supplements based on automatic speech recognition. IEEE Transactions on Biomedical Engineering, 47 (4), 487-496, 2000.

[8] Gibert, G., Bailly, G., Beautemps, D., Elisei, F. & Brun, R. Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech. Journal of Acoustical Society of America,118 (2), 1144-1153, 2005.

[9] Odisio, M. & Bailly, G.. Tracking talking faces with shape and appearance models. Speech Communication, 44(1-4), 63-82, 2004.