Audiovisual Perceptual Evaluation of Resynthesised Speech Movements

Matthias Odisio, Gérard Bailly

Institut de la Communication Parlée
INPG, 46, av. Félix Viallet, 38031 Grenoble Cedex 1, France
{matthias.odisio, gerard.bailly}@icp.inpg.fr

Abstract

We have already presented a system that can track the 3D speech movements of a speaker's face in a monocular video sequence. For that purpose, speaker-specific models of the face have been built, including a 3D shape model and several appearance models. In this paper, speech movements estimated using this system are perceptually evaluated. These movements are re-synthesised using a Point-Light (PL) rendering. They are paired with original audio signals degraded with white noise at several SNR. We study how much such PL movements enhance the identification of logatoms, and also to what extent they influence the perception of incongruent audio-visual logatoms. In a first experiment, the PL rendering is evaluated *per se*. Results seem to confirm other previous studies: though less efficient than actual video, PL speech enhances intelligibility and can reproduce the McGurk effect. In the second experiment, the movements have been estimated with our tracking framework with various appearance models. No salient differences are revealed between the performances of the appearance models.

1. Introduction

Virtual talking heads are expected to be useful in a number of applicate scenario involving communication with humans. An important issue is to evaluate the animation performance of these talking heads. Despite being very useful by itself for diagnosis, objective evaluation should be coupled with subjective evaluation; such a perceptual evaluation could tell if the generated movements can be fused with audio to enhance speech perception.

Several evaluation procedures have been proposed, including Turing tests [1], general communicational judgements such as quality, appeal ratings [2] or naturalness [3], or more specific properties of audiovisual speech such as better performance in identification tasks (of, *e.g.*, sentences, restrained vocabulary and mono-syllabic words [1], sentences [4], logatoms [5], or syllables [6]).

Most of these evaluation experiments are performed using the whole animation process including the modules responsible for the generation of the articulatory parameters (either provided by tracking natural sequences or by synthesis from text), the shape model that renders the face geometry on the screen and the appearance model that is responsible for the final rendering of each pixel of the face. Part of the controversial results obtained in [2][1], where poor intelligibility scores seem to contradict the excellent acceptability of the animation could be explained by the fact that evaluation scores are a complex by-product of the deficient behaviours of these three essential components. Viable movements may be judged unacceptable if rendered by an inadequate appearance model (*e.g.* a unique and flat texture) and a fine-grained appearance model may compensate for inappropriate control movements in a global quality judgment such as rating naturalness or adequacy.

We have already proposed elsewhere [7] using the point-light (PL) technique [8] in order to concentrate on the quality of driving signals while avoiding the problem of choice of a specific appearance model. PL have already proved to be effective for audiovisual integration [9] and intelligibility [10][11][6]. Moreover PL speech stimuli seem to activate common brain areas as actual speech [12].

In this paper, we are interested in evaluating perceptually the speech movements estimated with our video-based tracking system [13]. This system is outlined in the next section. Then, we describe two intelligibility tests that sketch out a framework for benchmarking motion capture systems against ground truth data. The first experiment sets up the benchmark by providing identification scores of original vowel–consonant–vowel (*VCV*) stimuli with three types of presentation: audio alone, audio–video and audio–PL. The second experiment provides identification scores obtained by various tracking systems operating on these data and rendered by PL.

2. Model-based tracking of facial movements

It is beyond the scope of this paper to describe in detail how the speech movements used in the perception experiments have been obtained. The tracking system is based on speaker-specific models of the face: a 3D shape model and several appearance models [13].

2.1. The 3D shape model

A speaker-specific articulatory model of a female French speaker was constructed. This model emerges from statistical analyses of hundreds facial fleshpoints 3D positions captured on a set of a few dozen typical articulatory configurations. This model is 3D, linear and controlled by seven articulatory parameters. It explains more than 96% of the learning data variance. We rely on this shape model to take into account the actual biomechanical constraints ruling and linking the skin tissue deformations all over the face.

2.2. Appearance models and Tracking framework

An analysis-by-synthesis loop is used to recover the articulatory parameters that best fit an analysed image. The dissimilarity function being minimised is measured as the difference between two sets of appearance descriptors: a set computed in the analysed image and a set synthesised according to articulatory parameters. Three facial appearance models were built: two texture mapping models *tex_lin* (driven by the articulatory parameters) and *tex_cst* (a constant texture), and an articulatory-driven model of local appearance *la_lin*.

Three sets of articulatory movements were obtained by tracking monocular video speech material using each of these three appearance models. These movements are evaluated in Experiment II.

2.3. Acquisition of the ground truth reference movements

Also, a direct inversion of the shape model was computed from semi-automatically labelled fleshpoints image positions. This reference movements set is assumed to be the ground truth data.

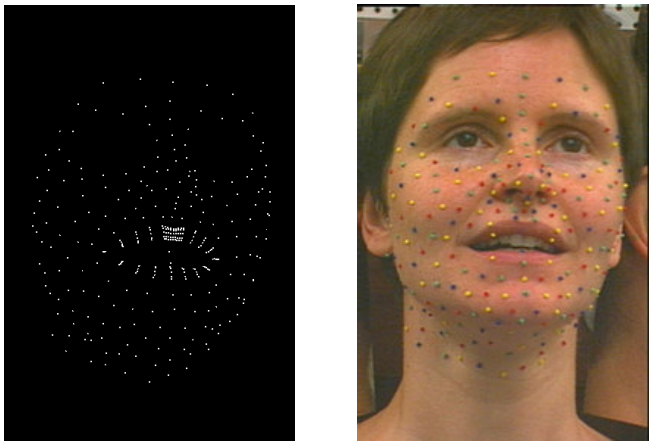

Figure 1: *plf* (left) and *natural* (right) visual systems

3. Experiment I: Evaluation of the Point-Light rendering

3.1. Method

3.1.1. VCV *stimuli*

The stimuli set consists of over-articulated *VCV* logatoms without a carrier sentence: *V* is one of { a, i, u } and *C* is one of 10 voiced consonants { b, d, g, v, z, ʒ, ʁ, l, m, n }. The speaker is recorded in the conditions of the face models construction, illustrated on Figure 1. For the acoustics, the natural audio signals are combined with an adaptive white noise level, for each SNR in { -24, -18, -12, -6, 0, 6 } dB. Three visual systems are tested: audio alone, original video record and PL face (corresponding to ground-truth data, see Section 2.3), hereafter called *nil*, *natural* and *plf*. Moreover, as a prospective study, the stimuli set comprises McGurk stimuli [14], *i.e.* a [aga] video dubbed by the carefully warped audio stimuli [aba].

3.1.2. Procedure

The experiment is divided in three successive sessions, corresponding to the visual systems *nil*, *natural* and then *plf*. In each session, all 180 stimuli are played in a random order. Five additional training stimuli are played at the beginning of the three sessions. Moreover, before the *plf* session, participants are told that they will see a talking face represented as a moving set of PL; a video showing a morphing between the *natural* and *plf* visual systems during a sentence is also played. The stimuli are played on a 15" laptop computer screen with headphones, and the animation window size is 384×576 pixels. The displayed face is slightly oriented (-10° y-axis rotation) and it is approximately 12 cm high. The *natural* videos are produced using Adobe Premiere, with a 25 fps frame rate. The *plf* animations are rendered in white on a black background. They are generated on the fly at 50 fps using graphic facilities offered by the standard 3D acceleration card. A first rendering pass of the whole face polygonal mesh computes the Z-buffer; the displayed point-lights correspond actually to the illumination of fleshpoints facing the camera and that are not masked by the head; two point-lights are set for the upper and the lower teeth (see Figure 1). A graphic user interface is designed. The identified consonant is indicated by a click on the corresponding labelled button (a forced-choice task). A progression index indicates the part of the session currently completed. Immediately after a choice is given, the next stimulus is played. Participants are instructed to choose relatively quickly. On one hand, this ensures that the whole experiment lasts about 35 minutes. On the other hand, this enhances the chance of occurrence of an after-effect (*i.e.* when the perception of a stimulus is perturbed by the perception of previous stimuli), especially when a *very noisy* stimulus follows a *clear* stimulus; however, we hope such a bias is compensated on average by the different random orders for each participant.

3.1.3. Participants

Seventeen participants took part to the experiment (13 males and 4 females; aged from 20 to 26, mean age 22.8 ± 1.7). They are native speakers of French and naive to the purpose of the experiments. Their (corrected) visual accuracy is estimated using the optometric Parinaud test. All participants read fluently at level 3, indicating a good visual accuracy for near vision. They are not screened for acoustic accuracy but report normal hearing; their audio ability for the identification task is evaluated with the *nil* visual system.

3.2. Results

3.2.1. Identification

The percentages of correct consonant identifications are represented on Figure 2. Globally, scores with *plf* are worse than with *natural*. Interestingly, some participants still perform nearly as well with both visual systems. A two-way repeated-measures analysis of variance (ANOVA) shows significant main effects for the visual systems ($F(2, 32) = 289.72$, $p < .001$) and for the audio level ($F(5, 80) = 395.97$, $p < .001$), and a significant interaction ($F(10, 160) = 13.224$, $p < .001$). Post-hoc comparisons using Holm adjusted p values show at each audio level significant differences between all the visual systems. Intelligibility, as the proportion of correct responses, is about the error rate; it tells us nothing about how the incorrect responses are distributed. Figure 2 also shows a measure of the error dispersion [15]. The error dispersion for the stimuli represents "the effective number of error categories per stimulus"; an interesting property is that it is rather insensitive to the error rate. Dispersion of the incorrect responses is greater with *plf* than with *natural*. However, the structure remains the same; a sketch of this degradation is given in Table 1. Salient properties of these results are: the poor salience of [z], the loss of the salience of [ʒ] in rounded context and the salience of nasal consonants ([m] and [n] are never confused with [b] or [d]). The asymmetry in the [l]–[ʁ] confusions is certainly due to the trill [ʁ] produced by the French speaker (the fricative [ʁ] being more frequently used in French).

3.2.2. Relative visual informational contribution

The relative visual informational contributions [16] of systems *natural* and *plf* are approximately constant across the SNR, being 58% for *natural* and 36% for *plf*. This leads to the following empirical relation: the information transmitted by our PL is approximately equal to the information transmitted by the actual video decreased by 22% of the information not transmitted by the audio channel.

3.2.3. McGurk effect

Table 2 shows the percentage of responses identical to the audio consonant for the McGurk stimuli. This table also contains the responses for [ava]: unexpectedly, as the SNR increases, the labiodental [v] in [a] vocalic context is more and more heard as the bilabial [b], whereas the video and the PL movements are judged by the authors to be very good. Upon further consideration, we believe that this particular stimulus could be considered as a McGurk combination (hereafter noted audio [ab*a] – video [ava]).

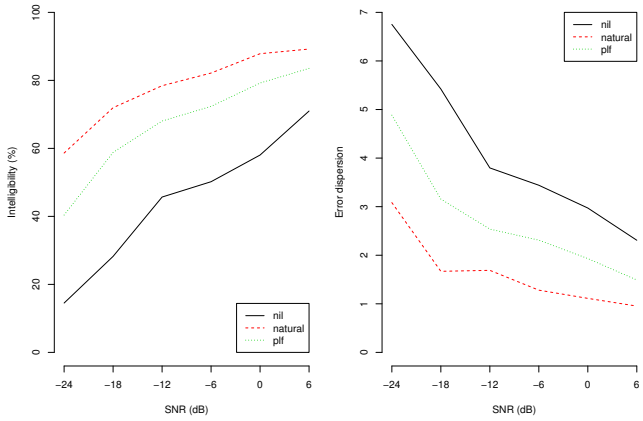

Figure 2: Consonant identification score (left) and error dispersion (right), function of audio SNR, for each visual system

Poorer fusion responses with *plf* are observed than with the *natural* face. For the [aba–aga] combination, McNemar's χ^2 tests show significant differences at all SNR between *natural* and *nil* and between *natural* and *plf*. Significant differences are found between *plf* and *nil* at -24 dB ($\chi^2(1) = 5.00$, $p = .025$) and at -18 dB ($\chi^2(1) = 6.40$, $p = .011$). For the [ab*a–ava] combination, differences between *natural* and *nil* are significant for all SNR but -24 dB. Differences between *natural* and *plf* are significant for SNR greater than -18 dB. Significant differences are found between *plf* and *nil* at -6, 0 and 6 dB (*e.g.*, at -6 dB: $\chi^2(1) = 5.44$, $p = .020$).

Table 1: Pooled confusion matrix for experiment I with *natural* (top) and *plf* (bottom). Stimuli are given in the columns.

nat	b	m	v	d	n	l	z	ʒ	g	ʁ
b	*297*	9	2	2						
m	3	*288*			5	3			1	1
v	4		*293*	1	2	1		1	7	3
d	1	1	1	*217*	3		*38*	3	*32*	1
n		2			*190*	11	2		3	1
l		5		1	*83*	*220*	1	4	6	
z			7	26	5	3	*163*	*53*	25	2
ʒ			2	11	4	5	*70*	*227*	23	5
g	1			*41*	1	2	22	5	*202*	2
ʁ		1	1	7	13	*61*	10	13	7	*291*
plf	b	m	v	d	n	l	z	ʒ	g	ʁ
b	*280*	11	*41*	8			2	1	8	
m	3	*252*	2		18	6	1		2	2
v	14	5	*221*	14	10	4	16	26	7	6
d			7	*185*	1	4	*55*	6	*32*	3
n	2	20			*152*	16	2	1	2	4
l		14	2	1	*82*	*208*	2	10	2	17
z	2	1	12	27	4	6	*116*	*51*	10	5
ʒ			11	16	16	8	*46*	*172*	20	7
g	4	2	7	*48*	4	5	*57*	15	*213*	9
ʁ	1	1	3	7	19	*49*	9	24	10	*253*

3.3. Discussion

As could be expected, the perceptual enhancement provided by the PL is less effective than the actual video. PL seem to be globally-degraded visible speech: compared to video, PL are less intelligible ([m] and [n] are more confused for example) and have poorer audiovisual integration performance. Interestingly, in similar acoustic conditions, we observe the same integration results with our synthetic talking face as those observed with a human speaker in [9]. Further studies including distinct PL presentations of articulators should help to investigate how this degradation could be due to the PL modality *per se*, to the absence of the tongue, or to other factors.

Table 2: Percentage of correct responses (based on audio) for the *McGurk-related* stimuli of experiment I.

Visual system	Stimuli	SNR (dB)					
		-24	-18	-12	-6	0	6
A–V incongruent							
natural	aba–aga	6	6	12	24	29	41
plf	aba–aga	12	47	82	88	94	94
natural	ab*a–ava	0	0	0	6	0	0
plf	ab*a–ava	12	18	29	47	47	47
Audio alone							
nil	aba	29	94	100	100	100	100
nil	aga	71	71	88	100	100	100
nil	ab*a	6	35	59	88	88	82
A–V congruent							
natural	aba–aba	100	100	100	100	100	100
plf	aba–aba	94	100	100	100	100	100
natural	aga–aga	47	82	94	100	100	100
plf	aga–aga	71	82	94	100	100	100

It should be noted that our PL rendering technique (each dot is a 2×2 pixels square, equivalent to $1mm$ in diameter in the 3D world) differs from classical studies where the subject is videotaped under low illumination with his/her face blackened with make-up and retro-reflective large dots (typically $3mm$ in diameter in [9]) glued on his/her lower face. Some of these setups are likely to induce additional 3D kinematic information during speech, because: *(i)* an imperfect chromakey of natural videos could leave traces of head and skin motion [11]; *(ii)* the face surface and normals to this surface change, and so apparent geometries of the dots change as well (*e.g.* a dot located near lip corner will appear as a circle in spread articulations and as an ellipsis in rounded articulations). Our PL are true 2D points moving in the screen. Nevertheless, our PL are also not a canonical point-light display [1]because of the large number of points: even statically, such a display can easily be identified as a face and thus it also *contains* some pictorial information.

4. Experiment II: Evaluation of tracking with the appearance models

4.1. Method

All participants of Experiment I were recruited for Experiment II a week later. They faced the same interface and the same identification task. Movements (corresponding to the same *VCV* stimuli as in Experiment I) were estimated by model-based tracking using different appearance models: *tex_lin*, *tex_cst*, and *la_lin*. As a control condition, the stimuli also include the *plf* movements of Experiment I, renamed here *ground_truth*. The resulting movements are rendered with point-lights and played with the natural audio signals degraded with three different SNR { -24, -18, 0 } dB [2]. All the stimuli are presented in random order, in only one session; participants are told that they can take a break whenever they feel the need.

4.2. Results

4.2.1. Control with Experiment I

An average identification improvement of 4% is observed between the two experiments, *i.e.* considering results for *plf* and *ground_truth* at SNR { -24, -18, 0 } dB, reflecting participant reports

[1] In a structure-from-motion paradigm, a point-light display should not provide any cue on the underlying 3D structure in absence of motion.

[2] Additional movements corresponding to two other appearance models not described in this paper were also part of Experiment II. With three SNR, the number of stimuli is the same in both experiments.

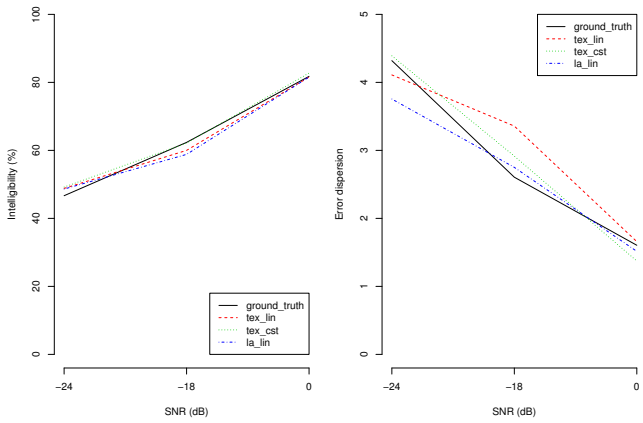

Figure 3: Consonant identification score (left) and error dispersion (right), function of SNR audio level for each tracking system

that this task was easier. A two-way repeated-measures ANOVA shows significant main effects for the visual systems ($F(1,16) = 8.2004, p = .011$) and for the audio level ($F(2,32) = 254.99, p < .001$), and no interaction ($F(2,32) = 0.8782, p$ value exceeded the .05 level, unless otherwise stated). Further one-way repeated-measures ANOVA at each audio level shows a significant difference at -24 dB ($F(1,16) = 7.427, p = .015$), and non-significant differences at -18 dB and at 0 dB. Moreover, the error dispersion of the incorrect responses for *ground_truth* is, in average, 0.48 categories less than for *plf*. Such differences could be explained by the fact that the *ground_truth* stimuli could occur throughout Experiment II whereas the *plf* stimuli only occurred during the third session of Experiment I (and thus after the tedious audio-only session).

4.2.2. *Identification and McGurk effect*

The percentage of correct consonant identifications is represented on Figure 3. Globally, the scores are very close to each other. A two-way repeated-measures ANOVA shows a significant effect for the audio level ($F(2,32) = 427.0,\ p < .001$) and no significant effect for the visual systems ($F(3,48) = 1.101$). But dispersions of incorrect responses are different, especially at the lowest SNR. The appearance model *la_lin* has the lowest error dispersion, and thus is more coherent. These differences in error dispersion do not seem to be due to particular stimuli but rather reflect global tendencies.
Concerning the McGurk stimuli, although *tex_cst* seems to perform better, McNemar's χ^2 tests show no significant differences.

4.3. Discussion

This experiment fails to reveal any salient differences in intelligibility and integration between the tracking behaviours of the appearance models *tex_lin*, *la_lin* and *tex_cst*. These performances are very satisfying because they are similar to those with the ground truth data. [3] In the same constrained video conditions, objective evaluations of these appearance models performed in [13] has drawn the same conclusion. In less controlled situations however, the objective performance of the several appearance models varied. Perceptual tests based on movements tracked on more challenging videos would likely bring to light discriminative features.

[3] These so-called ground truth movements were actually regularised by the shape model which cannot reproduce *accurately every* facial postures. This could weaken the *biological movements* assumption for these stimuli.

5. Conclusion

Intelligibility tests of movements rendered as point-light and estimated by tracking of (rather easy) video sequences have shown no significant differences between the three appearance models and with ground truth data. Moreover, these re-synthesised tracked movements can reproduce the McGurk effect to some extent.
We plan to extend benchmarking towards less controlled stimuli and unmarked faces. Further experiments would better characterise the audiovisual integration efficiency of our virtual talking head.

6. Acknowledgements

We thank H. Lœvenbruck as the subject of this study, F. Elisei, B. Holm, G. Gibert, C. Savariaux and A. Arnal for their valuable input, the students at ENSERG who participated in the experiments, and P. Welby who proofread an early version of this paper.

7. References

[1] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," Massachusetts Institute of Technology, Cambridge, USA, AI Memo 2003-003, Feb. 2003.

[2] I. S. Pandzic, J. Ostermann, and D. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330–340, 1999.

[3] B.-J. Theobald, J. A. Bangham, I. Matthews, and G. Cawley, "Evaluation of a talking head based on appearance models," in *Proc. of AVSP*, St Jorioz, France, Sept. 2003, pp. 187–192.

[4] C. Siciliano, A. Faulkner, and G. Williams, "Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners," in *Proc. of AVSP*, St Jorioz, France, Sept. 2003, pp. 205–208.

[5] T. Guiard-Marigny, D. Ostry, and C. Benoît, "Speech intelligibility of synthetic lips and jaw," in *Proc. of ICPHS*, vol. 3, Stockholm, Sweden, Aug. 1995, pp. 222–225.

[6] M. M. Cohen, R. L. Walker, and D. W. Massaro, "Perception of synthetic visual speech," in *Speechreading by Humans and Machines*, ser. Computer and Systems Sciences, D. G. Stork and M. E. Hennecke, Eds., vol. 150. Springer, 1996, pp. 153–168.

[7] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," in *IEEE Workshop on Speech Synthesis*, Santa Monica, USA, Sept. 2002, pp. 27–30.

[8] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[9] L. D. Rosenblum and H. M. Saldaña, "An audiovisual test of kinematic primitives for visual speech perception," *J. of Exp. Psy.: Human Perception and Performance*, vol. 22, no. 2, pp. 318–331, 1996.

[10] L. D. Rosenblum, J. A. Johnson, and H. M. Saldaña, "Point-light facial displays enhance comprehension of speech in noise," *JSHR*, vol. 39, no. 6, pp. 1159–1170, 1996.

[11] T. R. Bergeson, D. B. Pisoni, and J. T. Reynolds, "Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants," in *Proc. of AVSP*, St Jorioz, France, Sept. 2003, pp. 55–60.

[12] A. Santi, P. Servos, E. Vatikiotis-Bateson, T. Kuratate, and K. Munhall, "Perceiving biological motion: Dissociating talking from walking," *J. of Cognitive Neuroscience*, vol. 15, no. 6, pp. 800–809, 2003.

[13] M. Odisio and G. Bailly, "Shape and appearance models of talking faces for model-based tracking," in *Proc. of AVSP*, St Jorioz, France, Sept. 2003, pp. 105–110.

[14] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[15] R. J. J. H. van Son, "A method to quantify the error distribution in confusion matrices," in *Proc. of EUROSPEECH*, Madrid, Spain, 1995, pp. 2277–2280.

[16] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *JASA*, vol. 26, no. 2, pp. 212–215, Mar. 1954.