EVALUATION OF MOVEMENT GENERATION SYSTEMS USING THE POINT-LIGHT TECHNIQUE

G. Bailly, G. Gibert & M. Odisio

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
{bailly,gibert,odisio}@icp.inpg.fr

ABSTRACT

We describe a comparative evaluation of different movement generation systems capable of computing articulatory trajectories from phonetic input. The articulatory trajectories here pilot the facial deformation of a 3D clone of a human female speaker. In this paper we test the adequacy of the predicted trajectories in accompanying the production of natural utterances. The performance of these predictions are compared to the ones of natural articulatory trajectories produced by the speaker and estimated by an original video-based motion capture technique. The test uses the point-light technique [26, 27].

1 INTRODUCTION

A system able to produce audiovisual speech from phonetic input generally consists in three modules: (a) a movement generation system that plans articulatory movements according to the phonological task, (b) a shape model that specifies how the geometry of the face is affected by these movements and (c) an appearance model that specifies how the skin texture - or more generally the face appearance – renders this shape deformation. Not all facial animation systems separate out these steps nor identify these intermediary representation spaces (articulatory, shape and appearance) for building up a synthetic animation. For example, image-based techniques consisting in overlaying facial regions [8] or morphing between target images [16] extracted from real videos do not impose a priori to distinguish between shape and appearance. Similarly, systems using visemes [16] as elementary units do not always distinguish between a "high-level" parametric control and a finer "low-level" shape deformation model. For a more extensive presentation of models and modules currently used, please refer to our recent reviews [2, 3].
In the following we present a test that aims at evaluating the quality of the movement generation system together with the shape model. The appearance model is here reduced to point-lights placed at facial fleshpoints defining the facial geometry.

2 EVALUATING ANIMATIONS

2.1 Intelligibility and cognitive load

The most common benchmark for evaluating virtual animations consists in measuring the gain of intelligibility that the video signal offer in a noisy environment [5, 19]. The expected results should reproduce strong properties of natural speech: at all noise levels and even with clear speech, audiovisual performance is always superior to monomodal (audio or video-only) perception [15, 31]. This multimodal integration can also help comprehension especially when listening to a foreign language or a passage with difficult semantic content [23].
The McGurk illusion [20] shows that we simply cannot avoid this innate audiovisual integration and that we are very sensitive to audiovisual discrepancies [13, 28] and incoherent or impoverished information provided by the video versus audio channels. Despite their long-standing experience of audiovisual perception and successful implementation of Baldi, Massaro and colleagues recognize that they "failed to replicate the prototypical McGurk fusion effect" with their talking head [18, p.22], although prior evaluation of Baldi exhibited a quite satisfactory gain of intelligibility and proved his efficiency for language learning and perceptual rehabilitation.
A more systematic evaluation was performed at ATT [21] on 190 subjects to show the benefit of audiovisual communication. The third experiment of this study aimed at comparing the *appeal* ratings for three different synthetic faces driven by the sample synthetic audiovisual control parameters: (a) a standard flat 3D talking head, (b) a texture mapped 3D talking head and (c) a sample-based talking face. Subjects were not particularly seduced by synthetic faces: the best score was obtained by (a) while (c) obtained the worst rating. Surprisingly attempting to increase naturalness resulted in inverse satisfaction. These results seem to contradict the results of the first experiment evaluating the intelligibility of digits in noise where (a) and (c) performed equally well. However actual and estimated times to complete the task were both significantly higher for (c). Although offering quite acceptable intelligibility gains, synthetic faces seem thus to require more cognitive effort and more mental resources than natural speech and some synthetic faces more than others.
If incoherent or impoverished audiovisual stimuli require more processing time and result in increased cognitive load, it seems interesting to separate out the contributions of the different generation modules to the overall quality. Evaluation procedures - including the ones previously mentioned – cannot distinguish between the adequacy of the movement generation system, the shape and appearance models in replicating the underlying motor control and biophysics of natural faces.

2.2 Movement generation and pure motion stimuli

Despite successful applications made by the animation industry, the laws governing the complex biological movements of living species still escape to our understanding of motor control principles, physical modeling and their interactions with the environment. Even the most recent complex computer animations rely on the capture of real biological movements from real environments, living animals or humans, that are further morphed onto avatars or more realistic clones. Automatic motion capture devices deliver typically the trajectories of a few dozen florescent dots glued on the moving object or organ. A further analysis of these data constitutes the basis of most movement generation modules.

Johansson [17] has shown that the observation of as less as 100ms of a movement – where the only visible elements are these point-lights – suffice to identify the underlying human activity. This sensitivity to biological motion seems quite innate or at least extremely precocious [6, 7]. This rapid and global sensitivity has also been shown for the perception of the person's gender [9], of complex actions – from instrumental ones (grasping, throwing…) to more social actions such as dancing [10-12]. Moreover any deviation from expected patterns is interpreted as intended: even a professional mime cannot cheat observers on the actual weight of a carried object [29, 30]. This decision is all the easier as the movement correspond to familiar behavior [4]. It is therefore not surprising that untrained observers can lipread vowels, syllables, and some simple words from point-light faces [27] with the same benefit in terms of signal-to-noise ratio as full video stimuli [25]. Rosenblum et al [26] also show that the Mc Gurk effect could also be reproduced by point-light stimuli.
Although point-light images contain no obvious facial features such as skin, teeth, or the shadows produced in an open mouth, this sort of "*pure motion stimulus*" provides visual speech information that can be integrated well with auditory speech.

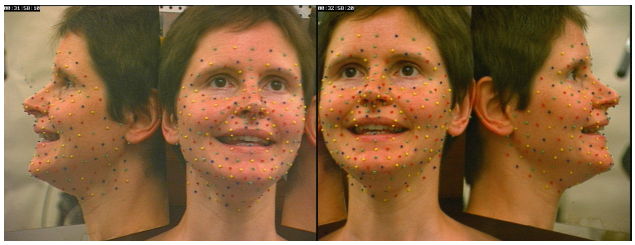

Figure 1: Gathering flashpoint positions using a photogrammetric method. Here 245 colored beads have been glued on the subject's face.

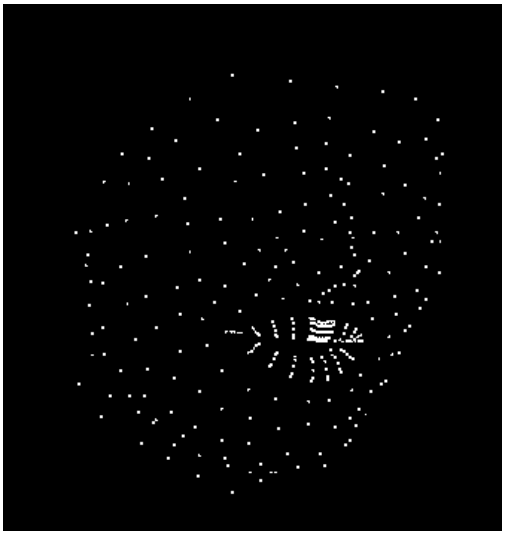

Figure 2: Distribution of facial point-lights from Figure 1.

3 SPEAKER-SPECIFIC SHAPE MODEL

We conducted a point-light experiment where natural articulatory trajectories are compared with synthetic trajectories computed by different movement generation systems from phonetic input. Both natural and synthetic articulatory trajectories pilot the same data-driven speaker-specific linear shape model [24].
Using a very simple photogrammetric method – previously used by Parke to build his initial model [22] - and up-to-date calibration procedures, we recorded 120 prototypical configurations of a French female speaker whose face was marked with 245 glued colored beads (on the cheek, mouth, nose, chin and front neck areas), as depicted in Figure 1. In a coordinate system linked with the bite plane, every viseme is thus characterized by a set of 245 3D points including positions of the lower teeth and of 30 points characterizing the lip shape (for further details see [14, 24]). We show that 6 linear predictors explain 97% of the variance of the data. Of course jaw opening, lip protrusion and lip opening are part of these parameters, that constitute our parametric control.

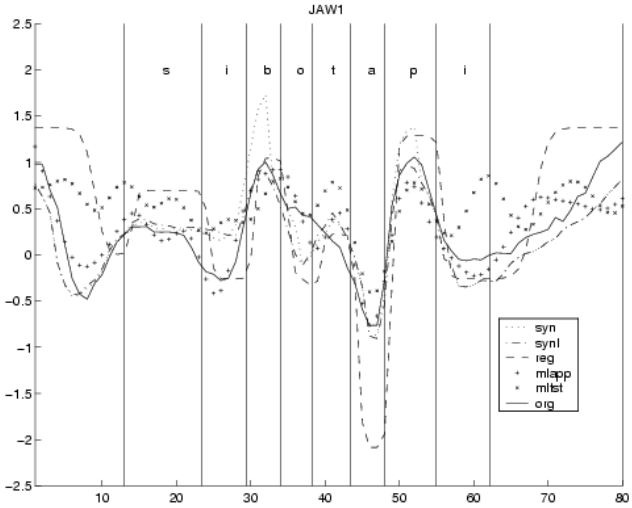

Figure 3: JAW1 (jaw rotation) trajectories predicted by different movement generation systems for the sentence "Six beaux tapis". Org is the trajectory produced by our female speaker.

4 GENERATING MOVEMENTS

4.1 Training and test corpus

All movement generation systems have at their disposal a training material and will be tested on a disjoint set of ten utterances. Training and test material result from the recording of audiovisual speech sequences using a multi-camera video capture system delivering uncompressed image sequences as in Figure 1. The trajectories of the six articulatory parameters are estimated at 50 frames/s using an analysis-by-synthesis procedure described in [14] where the RMS distance between each image and a simple appearance model using blending/morphing of three textures from the training corpus is used in an closed-loop estimation procedure. A simplex-based optimization technique estimates the set of articulatory parameters that minimizes this RMS distance at each frame.
The training corpus consists in 66 utterances, 96 VCV stimuli - where C is one of 16 French consonants and V is one of the 6 vowels /a,i,u,e,œ,o/ - and the 120 prototypical visemes used for building up the shape model. The test corpus consists in 10 utterances. The 76 phonetically-balanced sentences have been designed so as the diphones of the test utterances are at least present once in the training utterances in order to enable diphone-based audiovisual concatenative synthesis (see below).

4.2 Movement generation systems

All movement generation systems receive as input the same phonetic string augmented by phoneme durations. All training and test stimuli have been hand-labeled and segmentation results are available to all systems with the articulatory trajectories of the training material. Five movement generation systems have been tested.

1. *Syn* consists in a diphone-based audiovisual concatenation system. Diphones are multi-represented: candidate diphones are selected using a standard dynamic pro-

gramming technique. The local distance is the RMS distance between the Line Spectrum Pairs (available in the characterization of the audio signals: we use a LPPSOLA technique for the audio naipulations) across each boundary and does not take into account any articulatory distance. Intra-diphone articulatory trajectories are warped synchronously with acoustic frames.

2. *Synl* is similar to *Syn* except that a subsequent articulatory smoothing compensates for the jumps observed at the inter-diphone boundaries. The *anticipatory* smoothing procedure computes a linear interpolation of the observed jump during the previous diphone.
3. *Reg* computes trajectories using the Ohman's coarticulation model (for more details refer to [14]): rapid consonantal closures are superposed with slower vocalic articulations. Closure targets for each articulatory parameter are computed by a linear model using the *underlying vocalic* values for the jaw and the considered parameter as predictors. The coarticulation model is parameterized using all available consonantal targets. A Movement Expansion Model [1] describes the timing of the inter-vocalic transitions.
4. *Mltst* computes articulatory trajectories from the acoustics. As in [32], a linear regression links low-pass filtered (10Hz) LSP trajectories with articulatory movements using the 66 utterances (4051 frames) as training material.
5. *Mlapp* is similar *to Mltst* except that correlation is increased by using the test utterance as training material.

An example of the predicted trajectories is given Figure 3. *Reg* tends clearly to hyper-articulate. A sixth movement generation systems *Inv* is simulated by just inverting (multiplying by –1) the natural articulatory trajectories. Mean correlation coefficients between natural and synthetic articulatory trajectories for all 10 test sentences are given in Tableau 1 below.

Tableau 1: Mean correlation coefficients between original motion capture parameters and those predicted by different generation systems. Note that Reg *and* Mltst *have the lowest correlations*

Parameters Systems	Jaw1	Jaw2	Lips1	Lips2	Lisp3	Skin1
Syn	0.85	0.62	0.66	0.70	0.65	0.65
Synl	0.84	0.70	0.73	0.65	0.64	0.63
Reg	0.32	0.66	0.27	0.32	0.37	0.46
Mlapp	0.85	0.88	0.90	0.89	0.87	0.89
Mltst	0.44	0.44	0.55	0.41	0.30	0.53
Inv	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00

5 POINT-LIGHT EXPERIMENT

5.1 Procedure

All point-light animations are paired with the natural acoustic stimuli. All stimuli are generated in real-time using our MOTHER OPENGL® animation software [24] using graphic facilities offered by most standard 3D acceleration cards. A first pass drawing an invisible polygonal mesh is used to initialize the z-buffer: point-lights correspond effectively to the illumination of fleshpoints facing the camera and not obscured by the head (see Figure 2). Point-lights are drawn white on a black background. The animation window size is 576x768 pixels and is displayed on a 17" monitor. The user interface is designed using MATLAB® GUI. The latency between the onset of the audiovisual stimuli and key stroke is measured.

Subjects are asked to rate on a five points scale (incoherent, unsatisfactory, average, satisfactory, excellent) the degree of coherence between acoustics and the proposed facial motion. No head motion is added and the face is presented from front.

Subjects are made familiar with the point-light presentation with three natural stimuli from the training utterances. A sinusoidal (1 Hz) axial rotation (45 degrees apart from front view) of the head - pronouncing the longest utterance of the training corpus with natural movements - is added here. All subjects reported seeing a natural talking face.

5.2 Results

Results are compiled in Figure 4. As expected, the original audiovisual stimuli (*Org*) and their inverted version (*Inv*) lie at the extremes of the MOS scale. The difference between *Org* and *Synl* scores is not significant () as also between *Reg* and *Syn* () and between *Mlapp*, *Mlreg* and *Inv* (). The most surprising result is that the acoustic-to-articulatory system *Mlapp* is rated unacceptable – as the fair *Mtst* synthesis case - despite its high degree of correlation with the original trajectories delivered by *Org*. On the contrary, *Reg* generates quite acceptable trajectories despite poor correlation coefficients. Mean test durations tend to be shorter for the extremes of the rating scale.

Subjects reported that a series of stimuli was adequate in terms of phonetic features but not "natural" (i.e. too hyperarticulated). We verified that *Reg* had indeed longer decision lags.

Note also that all trajectories were rated with reference to the natural signal. If the evaluation would have used the synthetic signal delivered synchronously with the articulatory trajectories by the audiovisual generation in *Syn* and *Synl,* one could suspect that these systems would have reached the highest scores!

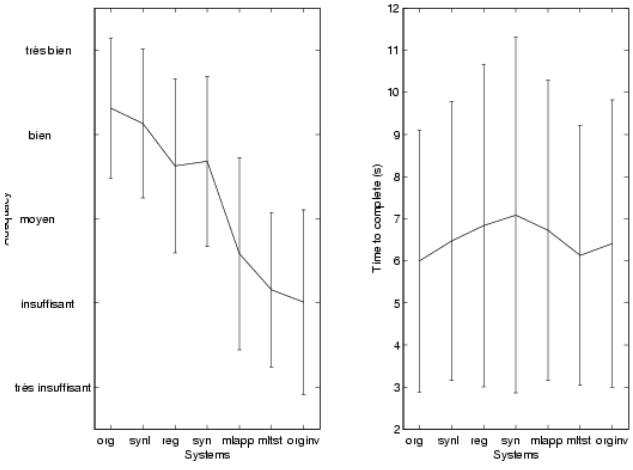

Figure 4: Results of the point-light experiments. Left: mean MOS according to generation systems. Right: mean stimuli-response latency.

6 COMMENTS AND CONCLUSION

Results of this point-light experiment shows that subjects are quite sensitive to the coherence between the movement of facial fleshpoints (evidenced here by point-lights) and an acoustic signal. We show that audiovisual concatenative synthesis using a simple anticipatory smoothing procedure has the potential of generating high quality movements. Too simple acoustic-to-articulatory mapping models generate quite unacceptable articulatory movements despite rather high correlation coefficients. This confirms that audiovisual perception is quite sensitive to the phasing between crucial events that concatenative synthesis preserves, that coarticulation models oversimplify and that acoustic-to-articulatory inversion has poor chance to recover.

The experiment described here involves uses natural driving audio stimuli: subjects expect thus a higher quality of motion generation than should be requested in case of synthetic acoustic signals. We also foresee to obtain still much higher degree of satisfaction in case of joint audiovisual concatenative synthesis! More alternative systems and presentation procedures should also be tested including prediction of head motion and other facial movements, presentation angles or density of points. We do think that point-light experiments should be considered as a standard benchmarking procedure for further proposals.

ACKNOWLEDGMENTS

This work was supported by the RNRT Artus Project. The authors thank their colleagues C. Abry and J.-L. Schwartz and L. D. Rosenblum for their fruitful comments on the audiovisual point-light experiment.

REFERENCES

[1] Abry, C. and Lallouache, T. (1995) *Modeling lip constriction anticipatory behaviour for rounding in French with the MEM (Movement Expansion Model)*. in *Proceedings of the International Congress of Phonetic Sciences*. Stockholm - Sweden. p. 152-155.

[2] Bailly, G. (2002) *Audiovisual speech synthesis. From ground truth to models*. in *International Conference on Speech and Language Processing*. Boulder - Colorado

[3] Bailly, G. (submitted) *Audiovisual speech synthesis.* International Journal of Speech Technology.

[4] Beardworth, T. and Bukner, T. (1981) *The ability to recognize oneself from a video recording of one's movement without one's body.* Bulletin of the Psychonomic Society, **18**: p. 19-22.

[5] Benoît, C. and Le Goff, B. (1998) *Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP.* Speech Communication, **26**: p. 117-129.

[6] Bertenthal, B.I., Proffitt, D.R., and Cutting, J.E. (1984) *Infant sensitivity to figural coherence in biomechanical motions.* Journal of Experimental Child Psychology, **37**: p. 213-230.

[7] Bertenthal, B.I., Proffitt, D.R., and Kramer, S.J. (1987) *Perception of biomechanical motions by infants: Implementation of various processing constraints. Special Issue: The ontogenesis of perception.* Journal of Experimental Psychology: Human Perception and Performance, **13**: p. 577-585.

[8] Bregler, C., Cowell, M., and Slaney, M. (1997) *VideoRewrite: driving visual speech with audio*. in *SIGGRAPH'97*. Los Angeles, CA. p. 353-360.

[9] Cutting, J.E., Proffitt, D.R., and Kozlowski, L.T. (1978) *A biomechanical invariant for gait perception.* Journal of Experimental Psychology: Human Perception and Performance, **4**: p. 357-372.

[10] Dittrich, W.H. (1993) *Action categories and recognition of biological motion.* Perception, **22**: p. 15-23.

[11] Dittrich, W.H. (1999) *Seeing biological motion - Is there a role for cognitive strategies?*, in *Lecture Notes in Artificial Intelligence: Gesture-Based Communication in Human-Computer Interaction*, A.e.a. Braffort, Editor. Springer Verlag: Berlin. p. 3-22.

[12] Dittrich, W.H., Troscianko, T., Lea, S.E.G., and Morgan, D. (1996) *Perception of emotion from dynamic point-light displays represented in dance.* Perception, **25**: p. 727-738.

[13] Dodd, B. (1979) *Lipreading in infants: Attention to speech presented in and out of synchrony.* Cognitive Psychology, **11**: p. 478-484.

[14] Elisei, F., Odisio, M., Bailly, G., and Badin, P. (2001) *Creating and controlling video-realistic talking heads*. in *Auditory-Visual Speech Processing Workshop*. Scheelsminde, Denmark. p. 90-97.

[15] Erber, N.P. (1975) *Auditory-visual perception of speech.* Journal of Speech and Hearing Disorders, **40**: p. 481-482.

[16] Ezzat, T. and Poggio, T. (1998) *MikeTalk: a talking facial display based on morphing visemes*. in *Computer Animation*. Philadelphia, PA. p. 96-102.

[17] Johansson, G. (1973) *Visual perception of biological motion and a model for its analysis.* Perception and Psychophysics, **14**: p. 201-211.

[18] Massaro, D. (1998) *Illusions and issues in bimodal speech perception*. in *Auditory-Visual Speech Processing Conference*. Terrigal, Sydney, Australia. p. 21-26.

[19] Massaro, D.W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.*Cambridge, MA: MIT Press.

[20] McGurk, H. and MacDonald, J. (1976) *Hearing lips and seeing voices.* Nature, **26**: p. 746-748.

[21] Pandzig, I., Ostermann, J., and Millen, D. (1999) *Users evaluation: synthetic talking faces for interactive services.* The Visual Computer, **15**: p. 330-340.

[22] Parke, F.I. and Waters, K. (1996) *Computer Facial Animation*.Wellesley, MA, USA: A.K. Peters.

[23] Reisberg, D., McLean, J., and Goldfield, A. (1987) *Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli*, in *Hearing by Eye: The Psychology of LipReading*, B. Dodd and R. Campbell, Editors. Lawrence Erlbaum Associates: Hillsdale, New Jersey. p. 97-113.

[24] Revéret, L., Bailly, G., and Badin, P. (2000) *MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation*. in *International Conference on Speech and Language Processing*. Beijing - China. p. 755-758.

[25] Rosenblum, L.D., Johnson, J.A., and Saldaña, H.M. (1996) *Visual kinematic information for embellishing speech in noise.* Journal of Speech and Hearing Research, **39**(6): p. 1159-1170.

[26] Rosenblum, L.D. and Saldaña, H.M. (1996) *An audiovisual test of kinematic primitives for visual speech perception.* Journal of Experimental Psychology: Human Perception and Performance, **22**(2): p. 318-331.

[27] Rosenblum, L.D. and Saldaña, H.M. (1998) *Time-varying information for visual speech perception*, in *Hearing by Eye: Part 2, The Psychology of Speechreading and Audiovisual Speech*, R. Campbell, B. Dodd, and D. Burnham, Editors. Earlbaum: Hillsdale, NJ. p. 61-81.

[28] Rosenblum, L.D., Schmuckler, M.A., and Johnson, J.A. (1997) *The McGurk effect in infants.* Perception & Psychophysics, **59**(3): p. 347-357.

[29] Runeson, S. and Frykholm, G. (1981) *Visual perception of lifted weight.* Journal of Experimental Psychology: Human Perception and Performance, **7**: p. 733-740.

[30] Runeson, S. and Frykholm, G. (1983) *Kinematic specification of dynamics as an informational basis for person and action perception: Expectation, gender recognition, and deceptive intention.* Journal of Experimental Psychology: General, **112**: p. 585-615.

[31] Sumby, W.H. and Pollack, I. (1954) *Visual contribution to speech intelligibility in noise.* Journal of the Acoustical Society of America, **26**: p. 212-215.

[32] Yehia, H.C., Rubin, P.E., and Vatikiotis-Bateson, E. (1998) *Quantitative association of vocal-tract and facial behavior.* Speech Communication, **26**: p. 23-43.