READ MY LIPS... AND MY JAW!
HOW INTELLIGIBLE ARE THE COMPONENTS OF A SPEAKER'S FACE?

in Proceedings of the 4th EUROSPEECH Conference, Madrid, Spain

Bertrand LE GOFF, Thierry GUIARD-MARIGNY & Christian BENOIT
e-mail: benoit@icp.grenet.fr
Institut de la Communication Parlée
Université Stendhal
BP 25X, 38040 Grenoble Cedex 9, FRANCE

ABSTRACT

In adverse acoustic condition, speech intelligibility is dramatically improved when the speaker's face can be seen. Which components of the face best convey the visual information associated with speech gestures? We answer this question through a series of experiments where normal hearers had to identify three vowels [a,i,y] and six consonants [b, v, z, ʒ, R, l] in French nonsense words of the form VCVCV, presented audiovisually under five conditions of background noise and five conditions of visual displays (the real face of the reference speaker, his lips alone, and three synthetic displays of facial components: lips, jaw and lips, whole face.)

1 AUDIO-VISUAL INTELLIGIBILITY OF TALKING FACES

Everyday situations show that speech comprehension is greatly increased by vision of the speaker's face. Several studies have quantified the increase in speech intelligibility due to visual information (Sumby & Pollack, 1954 [8]; Erber, 1969 [3] Summerfield, 1979 [10]; Benoît et al. 1994 [1]; among others). Summerfield et al. (1989)[9], and McGrath (1985)[6], compared the visual intelligibility carried on by different components of the face, either natural or synthetic (lips, teeth and full-face). In the same vein, we compared the visual intelligibility carried on by natural and synthetic displays of the lips, jaw, and full-face, synchronized with French acoustic non-sense words uttered by a male speaker on whose face control parameters were measured in order to animate the synthetic models.

2 EXPERIMENTAL DESIGN

In order to quantify the visual intelligibility carried on by the different components of the face, we followed the same paradigm as in [1], [3], [8] and [10]. 18 non-sense words were presented to normal-hearers under six conditions of visual display (no video included) and five conditions of acoustic degradation. The same acoustic stimuli served in all five audio-visual conditions, with S/N ranging from -18 dB to +6 dB.

2.1 Stimuli

Five visual displays were tested : the front view of the speaker in black and white (Fig.1), a binary image of the lips only (i.e., the vermillon area: see Fig.2), a 3D model of the face (Fig.3), a 3D model of the skull and of the lips (Fig.4), and a 3D model of the lips (Fig.5). The three parametric models were animated through control parameters measured on the actual speaker's face.

The model of the face (Fig.2) was that first designed by Parke (1974)[7] and adapted to speech gestures by Cohen and Massaro (1993)[2], on which the original lips had been replaced with the above mentioned lip model. The whole model was animated through six parameters.

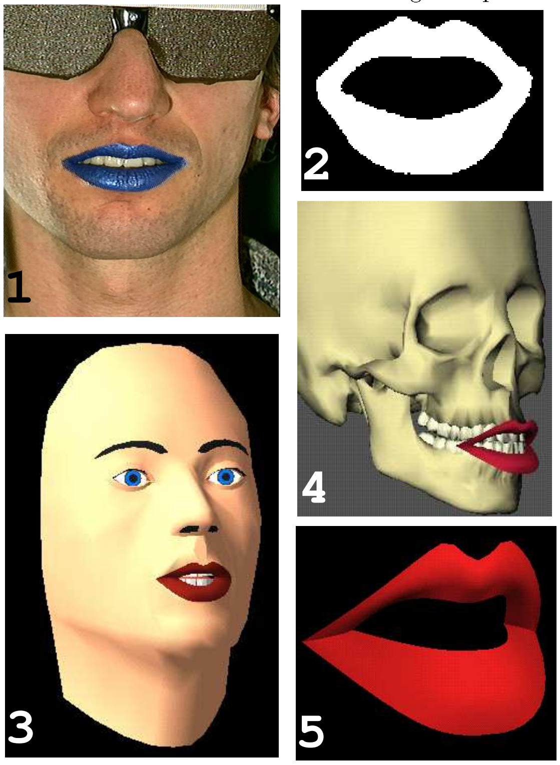

Fig.1 to 5: Visual displays used in experiments : 1. The natural face of the speaker, 2. binary images of the speaker's lips, 3. model of the face, 4. model of the jaw and model of the lips, 5. model of the lips.

The skull/jaw model (Fig.4) was that elaborated by Guiard-Marigny and Ostry (1995)[4]. The above mentioned lip model was overlayed onto the 3D skull. It is animated through six control parameters.

The lip model (Fig.5) was that developped at the ICP (Guiard-Marigny et al., 1994[5]). It is animated through five control parameters.

The speech material consisted of the natural acoustic utterances of a French speaker. All three synthetic models were animated through the same (five or six) control parameter directly measured on the speaker's face (from the front and profile views), at a 25 ips rate. No delay affected the original synchrony between audio and video.

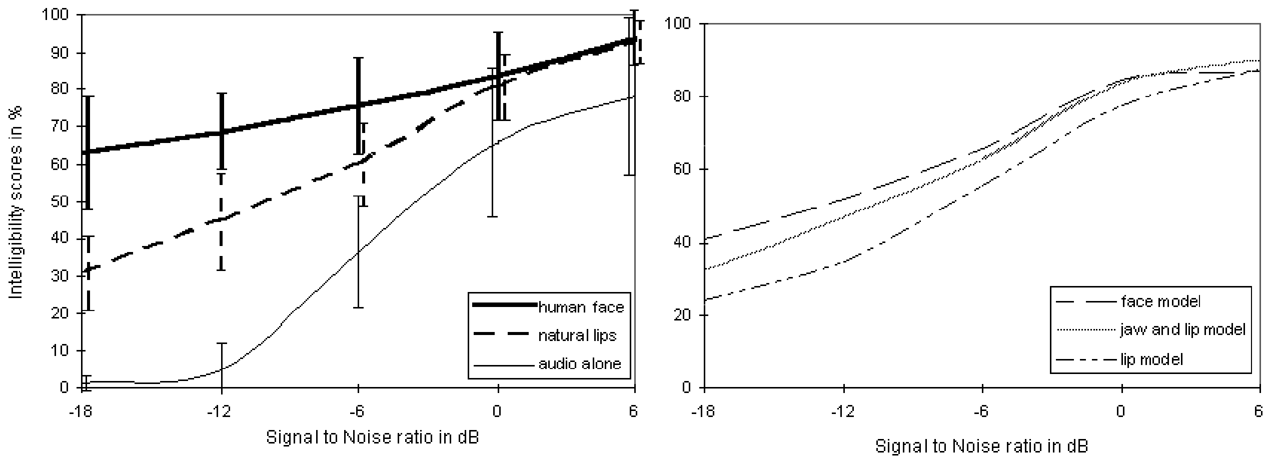

Fig.6: Intelligibility scores obtained in the identification of 18 stimuli, as a function of acoustic degradation, depending on the mode of presentation.

The corpus was made of VCVCV nonsense French words. V was /a/, /i/ or /y/. C was /b/, /v/, /z/, /ʒ/, /R/ or /l/. The whole test consisted of six sub-tests.
S0 no video : The eighteen different sentences were first digitized at Fs=16 kHz. They were then acoustically degraded by addition of white noise, at five S/N levels, by 6 dB steps. There were overall 90 audio stimuli. A pseudo-random order was used for presentation. Ten extra stimuli were appended before the actual test so that subjects could adapt to the test conditions. These acoustic stimuli served as a reference to the next five experimental conditions where the natural face or the synthetic models were synchronized to the audio part.
S1 natural face : The front view of the lower part of the face was displayed on a 15 inch monitor. The actual width of the whole face on the screen was roughly 10 cm. The video rate was 25 ips (PAL format) with a VHS-like quality. A visual alone condition was added to the five audio-visual conditions. This sub-test was made of 108 stimuli.
S2 natural lips : binary images of the speaker's lips only were displayed at 25 ips in white on a dark screen on a 19 inch monitor. The actual width of the lips on the screen was roughly 10 cm.
S3 synthetic face : the face model was Gouraud shaded and animated at 25 ips on a 19 inch monitor. It was displayed at a 20 degrees angle view from the sagittal plane. The actual width of the whole face on the screen was roughly 10 cm.
S4 synthetic skull : the skull model was Gouraud shaded and displayed at 25 ips on a 19 inch monitor at a 20 degrees angle view from the sagittal plane. The actual width of the whole skull on the screen was roughly 10 cm.
S5 synthetic lips : the lip model was Gouraud shaded and displayed at 25 ips on a 19 inch monitor at a 20 degrees angle view from the sagittal plane. The actual width of the lip model on the screen was roughly 10 cm.

2.2 Procedure

All subjects were normal-hearers. 14 subjects ran sub-tests S0, S1, S3 and S5. 20 subjects ran sub-tests S0, S2, S4 and S5.

The order of presentation of the sub-tests was balanced accross subjects in each group. Subjects were recommended to respond to both the vowel and the consonant, as much as they could guess it. A "?" response was tolerated, however.

3 RESULTS AND DISCUSSION

3.1 Global intelligibility

A test word was considered correct only if both the vowel and the consonant were correctly identified. As for auditory and visual intelligibility of natural stimuli, the results obtained in this experiment are in agreement with those by Benoît et al. (1994)[1]. Adding the video image of the natural face shows a dramatic gain in intelligibility over presenting the audio alone, as seen in Figure 6. The synthetic models also contribute strongly to improve the intelligibility of auditory speech. Scores in lipreading conditions are not presented in Figure 6 simply because they are strictly identical to those obtained under the most degraded acoustic condition (S/N = -18 dB). In fact, no acoustic cues could even be detected at this noise level. Figure 6 shows that the synthetic lips account for one-third of the intelligiblity carried on by the whole natural face, whatever the acoustic degradation. The synthetic face accounts for the two thirds of it. The jaw/lip model carries on roughly the same intelligibility as the natural lips alone, i.e. half the intelligibility carried on by the natural face.

The contribution of the synthetic lips and of the synthetic face to visual speech intelligibility is certainly impressive when considering the large contribution of the whole natural face. Five parameters are sufficient to animate the lip model alone. Even without the teeth, the tongue, the chin and the skin, the intelligibility carried on by the lip model is striking, and this is obtained with a very small quantity of information. As for the face model, a sixth parameter by itself almost doubles the visual intelligibility provided by the synthetic lips to the perceiver.

Here again, there is no control of the tongue, but the teeth and the chin are animated, and the structure of the face is coherently displayed. The visual information provided by the face around the lips allows subjects to disambiguate confusions among spread vowels (/i/ vs. /a/)

which are largely mixed up through the lips alone. The jaw-lip model is controlled by the same six parameters as the whole face model. However, the former is less intelligible than the latter. This is certainly due to the unnaturalness of the simultaneous display of lips superimposed to a jaw bone.

Overall, the whole natural face restores two thirds of the missing information when acoustics is degraded or missing; the facial model (tongue movements excluded), the jaw model and the natural lips restores about half of it; and the lip model restores a third of it. This is strong evidence that a very low bit rate of information (five or six parameters 25 times per second) is sufficient to transmit a great deal of the visual information carried on by the speaker's natural face, even though tongue gestures are not yet controlled.

3.2 Consonant confusions

Consonant confusions are presented in Table 1 at S/N = -12 dB where differences are at their maximum.

Whatever the consonant, there is a very strong disambiguation due to visual information. The disambiguation power follows the same hierarchy as that of global intelligibility, from the lips alone to the natural face, through the synthetic face.

/b/ is the consonant best identified audio-visually. Except for the natural face, it is given as a response to many /v/ stimuli, especially with the lip model (33%).

/ʒ/ identification is not improved when vision of the lip model is added to audio. However, /ʒ/ is rather well identified with the synthetic face, the natural face or the jaw model. With the lip model and binary lips, there are many confusions between /ʒ/ and /R/ in spread-vocalic context. Moreover, not only /ʒ/ is never identified in a /i/ context, but it leads subjects to identify the vowel /i/ as an /a/ (/iʒiʒi/ is perceived as /aRaRa/). Adding the chin (or the jaw) and the teeth disambiguate both the carrier vowel and the carrier consonant.

The two liquids /l/ and /R/ are mixed up in all conditions. A main reason is obviously that there is no tongue associated with the lip model, binary lips and jaw model, and that the tongue is not controlled in the face model. Even with the human face, confusions occur in /i/ and /y/ contexts, that is when the lip opening is too small for subjects to see the vertical tongue movement characteristic of /l/.

/z/ is auditorily identified below chance, and /ʒ/ is then the most frequent response. This is obviously due to the background noise used. Surprisingly enough, our jaw model helps subjects to disambiguate /z/ and /ʒ/ better than the face model. Nevertheless, a significant amount of /ʒ/ responses to /z/ stimuli remains when a natural face is presented. In fact, these confusions only occur in the /y/ context. /y/ has such an important coarticulatory effect that all consonants (but /b/ and /v/) look very similar when surrounded by two /y/'s. Therefore, audio-visual confusion of consonants presented in a /y/ context is mostly based on auditory similarities.

3.3 Vowel confusions

The vowel confusions are presented in Table 2 at S/N = -12 dB where differences are at their maximum. Complementarity between audition and vision is clearly seen in Table 2. In the auditory mode, /a/ is seldom confused with /i/ or /y/, whereas /i/ and /y/ are largely mixed up. On the opposite, /y/ is seldom confused with /a/ or /i/ in the (audio-)visual mode, whereas there are many confusions between /a/ and /i/ in the audio-visual mode.

As stated above, there is almost no auditory confusion between /a/ and /i/. However, an /a/ response is given to an /i/ stimulus in almost 25% of the cases, whereas /a/ is almost always identified when the lip model, natural lips or jaw model is simultaneously displayed. This effect remains to a smaller extent with the natural face, where an /i/ stimulus still leads to /a/ responses in 12% of all cases. 70% of these latter confusions are observed with /izizi/ (perceived as /azaza/ by half of the subjects.)

Table 1. Confusion matrices of consonants, irrespective of the response on the vowel (S/N = -12 dB). Stimuli are presented in rows. Percepts are presented in columns. Scores are in percent.

Audio only

	b	ʒ	l	R	v	z	?
b	26	12	3	7	11	3	**38**
ʒ	7	24		17	3	11	**38**
l	14	7	3	17	9		**50**
R	9	3	5	31	7	7	**38**
v	14	3	5	17	11	3	**47**
z	9	17	5	3	7	5	**54**

Natural face

	b	ʒ	l	R	v	z	?
b	**90**				10		
ʒ		**83**	3	3	6	5	
l		5	**60**	24	3	5	3
R		14	7	**59**	3	17	
v	7	7		3	**80**		3
z		20	3	3	5	**66**	3

Binary images of lips

	b	ʒ	l	R	v	z	?
b	**87**				3	2	8
ʒ	2	27	20	**30**	3	5	13
l	2		30	**52**	2	3	11
R		5	17	**53**	3	8	13
v	20	3	7	2	**63**		5
z	2	2	17	7	15	**50**	8

Lip model

	b	ʒ	l	R	v	z	?
b	**77**		2		14		7
ʒ	3	12	15	**32**	7	7	25
l	2	5	13	**41**	10	3	25
R	2	12	13	**43**	10	5	15
v	33	2	5	7	**38**	3	12
z	7	13	5	13	17	**35**	10

Face model

	b	ʒ	l	R	v	z	?
b	**86**		3		7		4
ʒ		**72**	7	7		7	7
l		5	32	**50**	3	3	7
R	5		33	**45**	3	7	7
v	17	5	9	9	**50**	3	7
z		14	14	12	17	**36**	7

Jaw and lip model

	b	ʒ	l	R	v	z	?
b	**86**		2	2	8		2
ʒ		**48**	10	19	5	8	10
l	2	17	22	**47**	2	3	7
R	3	5	18	**58**	2	7	7
v	23	13	10	2	**38**	7	7
z	8	12	17	3	3	**50**	7

Table 2. Confusion matrices of vowels, irrespective of the response on the consonant (S/N = -12 dB). Stimuli are presented in rows. Percepts are presented in columns. Scores are in percent.

Audio only

	a	i	y	?
a	**48**	2	6	44
i	9	19	**38**	34
y	6	14	**47**	33

Natural face

	a	i	y	?
a	**98**	1		1
i	12	**86**		2
y			**100**	

Binary images of lips

	a	i	y	?
a	**95**			5
i	37	**45**	10	8
y	1	1	**93**	6

Lip Model

	a	i	y	?
a	**91**	1	1	7
i	22	**53**	13	12
y	1		**93**	6

Face model

	a	i	y	?
a	**89**	11		
i	2	**98**		
y	1	1	**98**	

Jaw and lip model

	a	i	y	?
a	**92**	3	2	3
i	15	**73**	7	5
y		1	**98**	1

In fact, this is mostly due to the individual utterances /izizi/ and /azaza/ selected as stimuli for our experiment. As shown by Benoît et al. (1994)[1], /a/ shows smaller lip and jaw opening, and takes the shape of an /i/ when surrounded by /z/. They also noticed that /z/ has the same shape when surrounded by /i/ or /a/. Those statistical observations were obtained from a multi-dimensional analysis of ten utterances of /izizi/ and /azaza/. We looked back at the original data. It turns out that the /azaza/ used in our intelligibility test is amongst those with the largest lip and jaw opening. The /izizi/ here selected is also the one with the largest lip opening (and with an average jaw opening.) Differences in lip (resp. jaw) opening between our two stimuli /izizi/ and /azaza/ are in the range of 1 mm (resp. 0.5 mm.) It is thus not surprising that this hyper-articulated /azaza/ has been correctly identified, whereas half of the subjects perceived the hyper-articulated /izizi/ as another /azaza/, when they had the oportunity to see the natural face. This effect is emphasized with the lip model and the natural lips, where subjects cannot see the (even small) jaw movements.

More surprisingly, the face model allows subjects to correctly identify /i/ on the one hand, and to respond with an /i/ percept to an /a/ stimulus. Those errors occur when /a/ is coarticulated with labial consonants in the two thirds of the cases. The fact that /izizi/ is not here perceived as /azaza/ is probably due to an insufficient control of the chin displacements by the face model.

4 CONCLUSION

We quantified the visual intelligibility of several parametric models controlled for speech production. We also compared the relative contribution of single articulator (lips, jaw) to visible speech intelligibility. Sumby and Pollack (1954)[8] proposed an index of the visual contribution to the missing auditory information. This index is remarkably constant over the acoustic conditions of degradation. The natural face restores the two thirds of the missing information. The facial model restores almost half of it. The lip model restores a third. The jaw/lip model and the natural lips stand inbetween those last two models.

All the syntetic models we tested and evaluated are controlled through five or six parameters. If each parameter is coded with 1 byte, the transmission rate of visual information is in the order of 1 kbit/sec. It is remarkable that such a low bit-rate of information allows a six-parameter model of the face to compensate for the two thirds of the acoustic information when it is missing. Nevertheless, intelligibility scores compared across the jaw/lip model and the facial model tests highlight that visual information is at its best only when it is ultimately displayed through an ecological mode of presentation.

Acknowledgments:
This study was supported by the French CNRS and by the European ESPRIT-BRA project No 8579 "MIAMI".

References

[1] Benoît, C., Mohamadi,T. & Kandel, S. Audio-Visual Intelligibility of French speech in noise. *Journal of Speech & Hearing Research*, 37, 1195-1203, 1994.

[2] Cohen, M.M., & Massaro, D.W. Modelling coarticulation in synthetic visual speech. Proceedings of *Computer Animation 93*, Magnetat-Thalmann & Thalmann Eds, Genève, Suisse, 1993.

[3] Erber, N.P. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech & Hearing Research*, 12,423-425, 1969.

[4] Guiard-Marigny, T., & Ostry, D.J. *Three-Dimensional Visualization of Human Jaw Motion in Speech.* to appear, 1995.

[5] Guiard-Marigny, T., Adjoudani, A., & Benoit, C. *A 3-D model of the lips for visual speech synthesis Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis* New Paltz, New York, U.S.A., Sept. 1994

[6] McGrath, M. *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces*, Ph.D. Thesis, Univ. of Nottingham, UK, 1985.

[7] Parke, F. I. *A parametric model for human faces.* PhD Dissertation, University of Utah, Department of Computer Sciences, 1974.

[8] Sumby, W.H., & Pollack, I. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215, 1954.

[9] Summerfield, Q., MacLeod, A., McGrath, M., & Brooke, M. Lips, teeth, and the benefits of lipreading in *Handbook of Research on Face Processing*, A.W. Young & H.D. Ellis Editors, Elsevier Science Publishers, 223-233, 1989.

[10] Summerfield, Q. Visual perception of Phonetic gestures. in *Modularity and the Motor Theory of Speech Perception*, I.G. Mattingly & M. Studdert-Kennedy, Eds., Lawrence Erlbaum Addociates Pubs, Hillsdale, 117-137, 1991.