Can you "read tongue movements"?

Pierre Badin, Yuliya Tarabalka, Frédéric Elisei, Gérard Bailly

Can you “read tongue movements”?

Pierre Badin, Yuliya Tarabalka, Frédéric Elisei, Gérard Bailly

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Universités de Grenoble,
961 rue de la Houille Blanche, D.U. - BP 46, F-38402 Saint Martin d'Hères cedex, France
Pierre.Badin@gipsa-lab.inpg.fr

Abstract

Lip reading relies on visible articulators to ease audiovisual speech understanding. However, lips and face alone provide very incomplete phonetic information: the tongue, that is generally not entirely seen, carries an important part of the articulatory information not accessible through *lip reading*. The question was thus whether the direct and full vision of the tongue allows *tongue reading*. We have therefore generated a set of audiovisual VCV stimuli by controlling an audiovisual talking head that can display all speech articulators, including tongue, in an *augmented speech* mode, from articulators movements tracked on a speaker. These stimuli have been played to subjects in a series of audiovisual perception tests in various presentation conditions (audio signal alone, audiovisual signal with profile cutaway display with or without tongue, complete face), at various Signal-to-Noise Ratios. The results show a given implicit effect of tongue reading learning, a preference for the more ecological rendering of the complete face in comparison with the cutaway presentation, a predominance of lip reading over tongue reading, but the capability of tongue reading to take over when the audio signal is strongly degraded or absent. We conclude that these tongue reading capabilities could be used for applications in the domain of speech therapy for speech retarded children, perception and production rehabilitation of hearing impaired children, and pronunciation training for second language learners.

Index Terms: Lip reading, tongue reading, audiovisual speech perception, audiovisual talking head, hearing losses, augmented speech.

1. Introduction

A large number of studies has established that the vision of visible articulators (lips, jaw, face, tongue tip, teeth) eases speech understanding, and significantly increases the detection and identification performance of words in noise ([1]). Sumby and Pollack [2] as well as Benoît *et al.* [3], among others, have quantified the gain in speech intelligibility provided by lip reading in comparison with the sole acoustic signal. However, lips and face alone provide very incomplete phonetic information: the tongue, that generally can not be completely seen, carries an important part of the articulatory information that can not be accessed through traditional lip reading. Given the general articulatory awareness human skill, i.e. the ability to know the shape and position of one’s own articulators, it sounded interesting to investigate whether direct and full vision of the tongue can be used and processed by human subjects, in a similar way to lip reading.

The literature in this domain is rather scarce. Massaro *et al.* [4] used their computer-animated talking head, that can display articulation by making the skin transparent, to train children with hearing loss on both perception and production, and found evidence of some clear learning effect. Note, however, that their study did not explicitly test the spontaneous / innate ability to interpret tongue movements produced by real speakers. Bälter *et al.* [5] proposed strategies for phonetic correction based on their virtual talking head, that can display both visible and non visible articulators. The informal tests conducted in this preliminary study were well received by the three children involved. Recently, Fagel *et al.* [6] assessed the visual information conveyed by the dynamics of internal articulators. They found that displaying motion of internal articulators did not lead to significant improvement of identification scores at first, but that a short training session in which vocal tract movements were explained did significantly increase visual and audiovisual speech intelligibility.

The purpose of the present study was therefore twofold: (1) assessing the degree of spontaneous or innate ability of subjects for *tongue reading*, i.e. their ability to recover information from tongue vision without prior learning, and (2) testing their ability to rapidly learn this skill. The talking head developed at the department was thus used in a audiovisual perception test based on the noise degradation paradigm used by [2] or [7].

2. The talking head and its control

In order to ensure the ecological quality of the stimuli, we have built the audiovisual stimuli using original natural speech sounds and articulatory movements recorded synchronously by an ElectroMagnetic Articulography (EMA) device on one subject. The recorded movements are used to drive a talking head based on extensive measurements on the same subject.

2.1. The talking head

Our virtual talking head is made of the assemblage of individual three-dimensional models of diverse speech organs (tongue, jaw, lips, velum, face, etc) built from MRI, CT and video data acquired from a single subject and aligned on a common reference coordinate system related to the skull. The jaw, lips and face model described in [8] is controlled by two jaw parameters (*jaw height*, *jaw advance*), three lip parameters (*lip protrusion* common to both lips, *upper lip height*, *lower lip height*). The three-dimensional jaw and tongue model developed by Badin *et al.* [9] is driven mostly by five parameters: *jaw height* (common with the lips / face model), *tongue body*, *tongue dorsum*, *tongue tip vertical* and *tongue tip horizontal*. Note that the geometry of the models is defined by three-dimensional surface meshes whose vertices are associated with *flesh points*, i.e. points that can be identified on the organs.

2.2. Control of the talking head from EMA recordings

The control parameters of the various articulatory models of the talking head can be recovered by inversion from the midsagittal coordinates of a sufficient number of vertices of these 3D model meshes (see [10] for more details). An ElectroMagnetic Articulograph (EMA) device was thus used to record time trajectories of fleshpoints associated with these specific vertices by means of small electromagnetic receiving coils attached to the articulators in the midsagittal plane: one for the jaw, three along the tongue; one for each lip. The resulting control parameters were then used to build animations of the talking head.

3. Elaboration of the perception test

3.1. Corpus

The implementation of any perception test faces the dilemma between the highest number of stimuli and the necessarily limited duration that is practical for subjects to endure.
With the aim to assess the contribution of tongue vision, we collected the identification scores of all non-nasal French voiced consonants /b d g v z ʒ ʁ l/. The consonants were embedded in symmetrical VCV vocalic contexts with the set of vowels /a i u y/. This set contains the three cardinal vowels and the /y/ which has a labial shape almost identical to that of /u/ but differs from it by tongue placement. [u] and [y] have the strongest front-back lingual contrast and tongue movements are indeed expected to be highly informative and discriminant. The main corpus is finally composed of 32 VCV stimuli. Two additional corpora were recorded: a corpus with vowels /ɛ e ɔ/ was used for a generalisation test and a corpus with /œ/ for the familiarisation step (see further).

3.2. The presentation conditions and SNRs

In order to assess the contribution of the tongue vision, we designed a test with conditions where the tongue was visible contrasting with a condition where the tongue was not displayed. As we were also interested in comparing the contribution of the lips and face with the contribution of the tongue, we tested four presentation conditions:

- Audio signal alone (AU)
- Audio signal + cutaway view of the virtual head along the sagittal plane *without* tongue (AV*J*) (the *J*aw and vocal tract walls – palate and pharynx – are however visible) (see PB_ada.avi)
- Audio signal + cutaway view of the virtual head along the sagittal plane *with T*ongue (AV*T*) (see Figure 1, or PB_phrm6.avi)
- Audio signal + complete *F*ace with skin texture (AV*F*) (see Figure 1, or PB_ibi.avi)

The cutaway presentation on Figure 1 shows, in addition to the lips, the jaw, the tongue, the hard palate, the velum, and the back of the vocal tract wall from nasopharynx down to larynx. A profile view was chosen as it provides maximal information on the tongue, given that the angle of view does not change very much lip reading scores.
The contribution of the various visual elements was assessed according to the noise degradation paradigm: the identification score of the consonants in context is measured for different levels of white noise added to the audio signal. For each presentation condition, four Signal to Noise Ratios (SNRs) were generated: –∞ (i.e. no audio), –9 dB, +3 dB, +∞ (i.e. no noise).

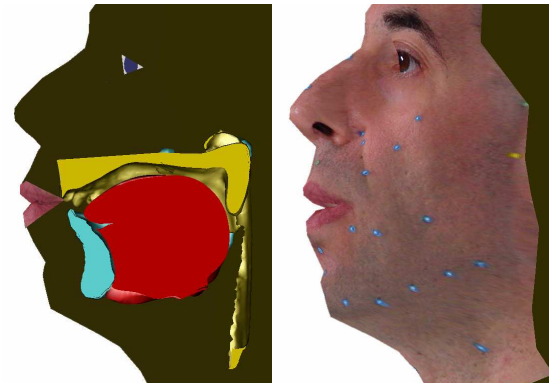

Figure 1 *: Examples of presentation conditions for the audiovisual test: cutaway view of the head including tongue (left) vs. complete face with skin texture (right).*

3.3. The protocol

The principle of the test was the following. An audiovisual stimulus is played to the subject once, without repetition, by means of a computer with a 17' TFT screen and high quality headphones at a comfortable listening level. The task of the subject is to identify the consonant, in a forced choice test, among the eight possible consonants /b d g v z ʒ ʁ l/. No repetition was allowed and subjects were instructed to answer as soon as possible. In order to familiarise the subjects with the test procedure, the session starts with a demonstration of the four presentation conditions and a series of five dummy tests (with vowel /œ/, which is not used in the real test).
The complete test is made of a 16 successive series, each defined by its condition, its SNR, and its stimuli, as described in Table I. For each series, the stimuli are presented in a randomised order different for each subject, preceded by two dummy stimuli with vowel /œ/ to help the subject getting accustomed with the new conditions.
As the aim of the test was to determine the spontaneous ability of the subjects to get information from different visual conditions, care was taken to avoid learning as much as possible. As the 32 VCV sequences were the same in the first 15 series, the tests were administrated in the order of increasing visual information: AVJ providing more information that AU, AVT more information than AVJ. No specific hypothesis was made about the AVF condition in relation to AVJ and AVT. Within each visual condition, it can be assumed that the association between sound and image would be more efficiently learned for high SNRs than for low ones. The subjects were thus divided in two groups to assess this hypothesis: group I subjects received the tests with increasing SNRs within each visual condition, while group II subjects received the tests with decreasing SNRs within each visual condition. The possible difference in the results will be used to test the implicit learning that occurs when no noise is added.
The last series, made of stimuli never played previously in the test, was used to assess the generalisation abilities of our subjects, i.e. if they did learn to *tongue read*, and verify if they did not learn the stimuli per se.

Table I. *Characteristics of the series of tests.*

Stimuli	Condition	SNR Gr I	SNR Gr II
/b d g v z ʒ ʁ l/ × /a i u y/	AU AU AU	+∞ +3 dB –9 dB	–9 dB +3 dB +∞
/b d g v z ʒ ʁ l/ × /a i u y/	AVJ AVJ AVJ AVJ	+∞ +3 dB –9 dB –∞	–∞ –9 dB +3 dB +∞
/b d g v z ʒ ʁ l/ × /a i u y/	AVT AVT AVT AVT	+∞ +3 dB –9 dB –∞	–∞ –9 dB +3 dB +∞
/b d g v z ʒ ʁ l/ × /a i u y/	AVF AVF AVF AVF	+∞ +3 dB –9 dB –∞	–∞ –9 dB +3 dB +∞
/b d g v z ʒ ʁ l/ × /ɛ e o/	AVT	–9 dB	–9 dB

3.4. The subjects

We have selected French subjects, with no known hearing nor non corrected sight losses, without prior experience in speech organs study nor analysis. The subjects from group I (7 females and 5 males, mean age 27.2 years) performed the tests in the decreasing SNR order, while the subjects from group II (4 females and 7 males, mean age 26.9 years) performed the tests in the increasing SNR order. A complete test session lasted between 30 and 50 minutes.

4. Results

4.1. Informal comments

Before presenting the results in details, it is worth summarizing the informal comments made by the subjects. Some reported that watching simultaneously the movements of the lips and of the tongue was not easy; a possible compromise was to focus the gaze on the incisors region in order to maintain the tongue in one side of the visual field of view and the lips in the other side. Subjects reported also that, whenever the sound was present, even with a high level of noise, they felt that the vision of the tongue was not very useful, but that in the video only condition (SNR = –∞), the tongue was very helpful for recognizing the consonant. The last session (i.e. the generalisation test) was deemed easier that the other series of test for the same conditions.

4.2. Main test

Figure 2 represents the mean identification scores, i.e. the percentage of consonants correctly identified for the 16 test series, separately for the two groups of subjects. Note first that the results for the AU and AVF conditions are coherent with those obtained by [3] for comparable lips / face presentation conditions. An important remark is that the standard deviations of the scores may be rather large (up to 13.4). Therefore, careful ANOVA analysis is required to draw valid conclusions. We found that the scores of group II are higher than those of group I (significant difference $F(1,10)=35.59$, $p<0.0001$). All audiovisual conditions have significantly higher identification scores than the AU condition.

The analysis has shown that the condition factor is significant for the three audiovisual conditions, and that all three conditions are significantly different from each other, with the ranking AVF > AVT > AVJ. Note however, that the individual differences for each SNR between the AVT condition and the other two audiovisual conditions are not significant, with two exceptions.

We have also found that the scores for the AVF condition are significantly higher than those for the AVJ condition for each SNR ($p<0.05$ for each pair, with one exception). This result was initially not expected, since the articulatory information is the same in both conditions. It might be ascribed to the fact that the skin texture of the face provides a supplementary source of information related to the redundant nature of the movements of the jaw, lips and cheeks. Another interpretation would be that subjects prefer an ecological (naturally looking) rendering to a cutaway presentation. It may also be a consequence of learning of the limited set of stimuli, as the tests with the AVF condition are administrated after those with the AVJ and AVT conditions.

An important conclusion is that the scores for the AVT condition are not significantly higher than those for the AVJ condition. An interesting exception occurs for group II, when no sound is present in the stimuli: in this case, the score for the AVT condition is significantly higher than that for the AVJ condition ($F(1, 10)=9.28$; $p<0.05$), with a score difference of 18%. This is in agreement with the informal comments reported above.

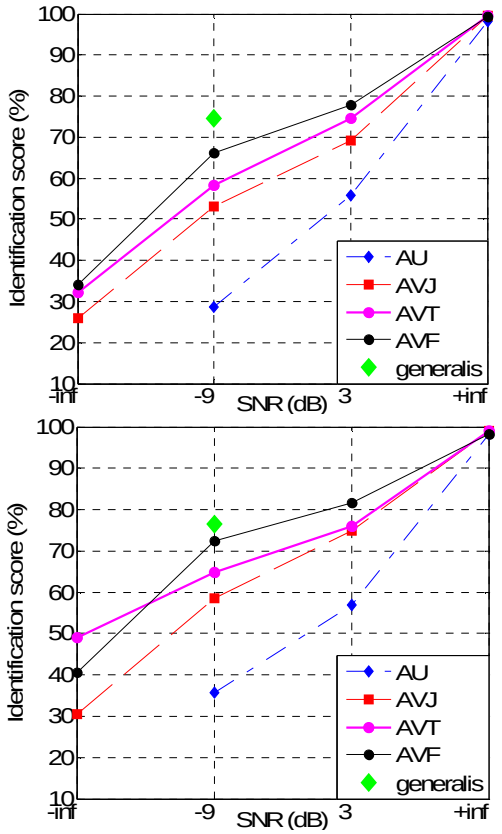

Figure 2 – Mean identification scores as a function of SNR (top: group I; bottom, group II) for the different conditions (from bottom to top): AU, AVJ, AVT, AVF. The isolated diamond indicates the score for the generalisation test.

4.3. Generalization test

As the same set of 32 VCV sequences was used in the first 15 series of the session, we had to verify that part of the implicit learning that may occur throughout the session was not due to the learning of the stimuli themselves rather than to tongue reading learning of. The generalisation test aimed thus at verifying that the good scores obtained with the main test would hold with new stimuli never presented before.

The scores for the generalisation series using a different set of vowels are significantly higher than the corresponding ones of the main test (AVT, SNR = –9 dB) for both groups (group I : $F(1, 10)=23.68$; $p<0.001$; group II : $F(1, 10)=8.92$; $p>0.01$). This finding seems to confirm the hypothesis that subjects acquire implicitly tongue reading skills during the test session. This interpretation should however be considered with caution. Indeed, Benoit *et al*. [11] have shown that vocalic context influences the intelligibility of the adjacent consonants: the improvement of the score may thus also be ascribed to the fact the vocalic contexts used in the generalisation tests would have facilitated the identification.

The conclusion that some implicit learning occurred is also supported by the fact that subjects in group II, who could benefit more from implicit learning as the were played the audiovisuals stimuli with low SNRs first, performed better than subjects in group I. Another argument is the fact that the score difference between the two groups for the generalisation test is not significant ($F(1, 10)=0.61$; $p>0.44$) since all the subjects have had the same tests when starting the generalisation test.

5. Conclusions et perspectives

5.1. Conclusions

Using ecological audiovisual stimuli obtained by controlling a virtual talking head from articulatory movements measured on a speaker, we performed an audiovisual test in order to assess the comprehension benefit that human subjects can get from seeing the tongue in an augmented speech condition. The study has yielded the following results.

The identification scores of group II are significantly higher that those of group I. This supports the idea that group II has benefited from a stronger implicit learning due the presentation of the audiovisual stimuli with a clear sound before those degraded by noise. All audiovisual conditions yield speech comprehension rates higher than the simple audio condition. The scores for all SNR levels rank, for each group, with statistically significant differences, in the following decreasing order : AVF, AVT, AVJ, AU. For each SNR, AVF is significantly better decoded than AVJ, which would mean that subjects prefer an ecological rendering to a cutaway view of the talking head.

The AVT condition is not significantly better perceived than the AVF condition, except when the audio signal is absent, for the group II, who benefited from a stronger implicit learning: in this case, the AVT score is higher by 18% than the AVJ score. This finding suggests that *tongue reading* can take over the audio information when this latter is not sufficient to supplement lip reading. Moreover, the relatively high identification score for the generalisation test, as well as the global performance difference between the groups seems to indicate that fast learning is possible.

Note that the similar study conducted very recently by Fagel *et al*. [6], using a less elaborate tongue model and synthetic movements, arrived to similar conclusions.

These preliminary tests need to be complemented by more systematic ones, involving in particular measures of visual attention, in order to confirm that our natural abilities for *tongue reading* are weak, or simply dominated by those for lip reading that are permanently practised right from birth.

5.2. Perspectives in speech rehabilitation and pronunciation training

As a follow up of this study, we envisage to elaborate learning protocols to show that the acquisition of *tongue reading* skills can be fast and easy.

Our aims in the future are thus to use the augmented speech capabilities of our virtual talking head for applications in the domains of (1) speech therapy for speech retarded children, as more and more asked by speech therapists, (2) perception and production rehabilitation of hearing impaired children as started by [4], and (3) pronunciation training for second language learners, as discussed by [12].

6. ACKNOWLEDGMENTS

We sincerely thank Ch. Savariaux and C. Vilain for helping us with the EMA recordings, and F. Berthommier, A. Rochet-Capellan and J.-L. Schwartz for helpful discussions on visual perception and statistics.

7. REFERENCES

[1] N. P. Erber, "Auditory-visual perception of speech," *J. Speech and Hearing Disorders*, vol. XL, pp. 481-492, 1975.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, pp. 212-215, 1954.

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

[4] D. W. Massaro and J. Light, "Using visible speech to train perception and production of speech for individuals with hearing loss," *JLSHR*, vol. 47, pp. 304-320, 2004.

[5] O. Bälter, O. Engwall, A.-M. Öster, and H. Kjellström, "Wizard-of-Oz Test of ARTUR - a computer-based speech training system with articulation correction," presented at the 7th Int. ACM SIGACCESS, Baltimore, 2005.

[6] K. Grauwinkel, B. Dewitt, and S. Fagel, "Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech," presented at Interspeech'2007, 2007.

[7] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, "Which components of the face to humans and machines best speechread ?," in *Speechreading by Humans and Machines*, vol. 150, *NATO ASI:* D. G. Stork and M. E. Hennecke, Eds. Berlin: Springer Verlag, 1996, pp. 315-328.

[8] M. Odisio, G. Bailly, and F. Elisei, "Tracking talking faces with shape and appearance models," *Speech Communication*, vol. 44 (1-4), pp. 63-82, 2004.

[9] P. Badin and A. Serrurier, "Three-dimensional linear modeling of tongue: Articulatory data and models," presented at ISSP7, Ubatuba, SP, Brazil, 2006.

[10] Y. Tarabalka, P. Badin, F. Elisei, and G. Bailly, "Can you “read tongue movements”? Evaluation of the contribution of tongue display to speech understanding," presented at ASSISTH'2007, Toulouse, France, 2007.

[11] C. Benoît, T. Mohamadi, and S. Kandel, "Effects of phonetic context on audio-visual intelligibility of French," *JSHR*, vol. 37, pp. 1195-1203, 1994.

[12] O. Engwall, "Feedback strategies of human and virtual tutors in pronunciation training," *TMH - Quaterly Progress Status Report - Stockholm*, vol. 48, pp. 11-34, 2006.