Can you "read tongue movements"? Evaluation of the contribution of tongue display to speech understanding

Yuliya Tarabalka * Pierre Badin Frédéric Elisei Gérard Bailly
GIPSA-lab, UMR 5216 CNRS – INPG – UJF – Université Stendhal
Département Parole & Cognition / ICP
46 avenue Félix Viallet, 38031 Grenoble Cedex 01, France
pierre.badin@gipsa-lab.inpg.fr

RESUME

De nombreux travaux ont établi que la vision des articulateurs typiquement visibles (lèvres, mâchoire, visage, partie antérieure de la langue, dents) facilite la compréhension de la parole par les humains, et augmente significativement le taux de détection de l'activité vocale et d'identification des mots. Pour autant, tout ne peut être "lu" sans ambiguïté avec la seule vue du visage. En particulier, la langue porte une part importante de l'information articulatoire mais n'est généralement pas visible. La Langue française Parlée Complétée (LPC) utilisée par une partie de la communauté des malentendants vise justement à compléter l'information labiale à l'aide d'un code qui normalise un ensemble de formes de la main et de positions par rapport au visage. Ainsi, les utilisateurs du LPC s'échangent des informations indirectes en particulier sur l'articulation de la langue. Ce codage, même s'il est efficace selon la théorie de l'information, est arbitraire et sans lien naturel avec la forme de la langue. Nous avons donc voulu tester l'hypothèse que la vision directe et complète de la langue – information a priori plus intuitive – peut être utilisée. Pour cela, nous avons mis en oeuvre la tête parlante virtuelle audiovisuelle du laboratoire, qui permet d'afficher tous les articulateurs de la parole, y compris la langue. Le mode de réalité augmentée choisi est un écorché de profil.

Nous avons construit un ensemble de stimuli audiovisuels VCV en déterminant les paramètres de contrôle de la tête parlante par inversion à partir des positions des bobines d'un articulographe électromagnétique fixées sur la langue, la mâchoire et les lèvres du sujet à partir duquel la tête parlante a été modélisée (clonage virtuel). Ces stimuli ont été présentés dans un test de perception audiovisuelle suivant quatre conditions: audio seul (AU), audiovisuel avec intérieur du visage *sans* langue (AVJ), avec intérieur du visage *avec* langue (AVT) et avec visage complet vu de l'extérieur (AVF). Chaque condition était présentée avec quatre niveaux RSB de bruit blanc ajouté au son: $-\infty$ (vidéo seule), –9 dB, +3 dB, $+\infty$ (signal sans bruit). Pour chaque stimulus, une réponse à choix forcé entre les huit consonnes était demandée. Pour tester les effets d'apprentissage, 12 sujets (groupe I) ont transcrit les stimuli avec des niveaux de bruits décroissants dans chaque condition, tandis que 11 autres sujets (groupe II) ont identifié les stimuli avec un bruit croissant (ce qui permettait un éventuel apprentissage de la relation entre audio et vidéo lorsque le signal audio était clair en début de test pour une condition donnée). Enfin, un autre ensemble de stimuli VCV (mode AVT, RSB = –9 dB) a été utilisé en fin de test pour évaluer les capacités de généralisation des sujets des deux groupes.

Une série d'analyses a permis de dégager les résultats suivants. Les scores de reconnaissance du groupe II sont significativement plus élevés que ceux du groupe I, ce qui conforte l'idée que le groupe II a bénéficié d'un apprentissage implicite plus fort. Toutes les conditions de présentation vidéo améliorent la compréhension de parole par rapport à l'audio seul. Les scores pour l'ensemble des niveaux RSB se classent, pour chaque groupe, avec des différences statistiquement significatives, dans l'ordre décroissant : AVF, AVT, AVJ, AU. Pour chaque RSB, AVF est significativement mieux décodé que AVJ : les sujets préfèrent un rendu écologique des mouvements à un écorché. La condition AVT n'est pas significativement mieux perçue que la condition AVJ sauf lorsque le signal audio est absent, pour le groupe II, qui a bénéficié d'un apprentissage implicite plus fort : dans ce cas le score AVT est supérieur de 18% au score AVJ. Ce résultat suggère que la lecture "linguale" peut prendre le relais de l'information audio lorsque cette dernière n'est plus suffisante pour compléter la lecture labiale. Le taux de reconnaissance relativement élevé du test de généralisation, ainsi que la différence globale de performance entre les deux groupes semble montrer par ailleurs qu'un apprentissage rapide peut être réalisé. Ces résultats très préliminaires sont à compléter par des tests plus systématiques impliquant notamment des mesures d'attention visuelle, pour confirmer que nos capacités naturelles de lecture linguale sont faibles, ou qu'elles sont simplement dominées par celles en lecture labiale. Nous envisageons cependant d'élaborer des protocoles pour montrer que l'apprentissage de la lecture linguale est rapide et facile. Notre objectif futur est donc d'utiliser les capacités de parole augmentée de notre tête parlante virtuelle pour des applications dans les domaines de l'orthophonie pour les enfants atteints de troubles de parole, de la réhabilitation en perception et production pour les enfants handicapés auditifs, et de la correction phonétique pour les apprenants de langue seconde.

Keywords

Lip reading, audiovisual speech perception, virtual audiovisual talking head, hearing losses, augmented speech.

* At present: Norwegian Defence Research Establishment (FFI)
P.O. Box 25 N-2027 Kjeller, Norway – yuliya.tarabalka@ffi.no

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 ([10]). Sumby and Pollack [20] as well as Benoît *et al.* [5], among others, have quantified the gain in speech intelligibility provided by “lip reading” in comparison with the sole acoustic signal.

However, the mere vision of the lips and face just provides incomplete phonetic information: laryngeal activity as well as movements of tongue and velum are not visible. Indeed, the tongue, that carries an important part of the articulatory information, can generally not be seen completely. “Cued Speech”, elaborated by Cornett [8], used by an increasing number of hearing impaired speakers aims precisely to complement the lip information by means of a set of hand shapes and positions in relation to the face that provides most of the missing phonetic information, in particular related to tongue articulation. This coding system, although efficient in terms of information theory, is arbitrary and is not directly related to tongue movements. Therefore, we have attempted to determine if direct and full vision of the tongue – information presumably more intuitive – can be intuitively used and processed by human subjects, in a similar way to “lip reading”.

The literature on this subject is rather scarce. The first attempt that we are aware of that tests the ability of subjects to use tongue vision was made by Tye-Murray *et al*. [21]. They conducted experiments to determine whether increasing the amount of visible articulatory information could affect speech comprehension, and whether such artefacts are effectively beneficial. The experiments involved videofluoroscopy, which allows movements of the tongue body, lips, teeth, mandible, and often velum to be observed in real-time during speech. Subjects were asked to speechread videofluoroscopic and video images. The results suggest that seeing supralaryngeal articulators that are typically invisible does not enhance speechreading performance. Subjects always performed better with the video than videofluoroscopy medium. Subjects performed as well when the tongue was not visible in the videofluoroscopic records as when it was visible. Training did not improve subjects' ability to recognize speech presented with videofluoroscopy. These conclusions should however be considered with caution, as the quality and the interpretability of videofluoroscopic images is not very high.

Massaro *et al*. [13] used Baldi, their computer-animated talking head, as a language tutor for speech perception and production for individuals with hearing loss. Baldi could, among other features, display articulation by making the skin transparent to reveal the tongue, teeth, and palate [7]. Seven children with hearing loss were trained on both perception and production for a total of 6 hours spread over 21 weeks. The proportion of correct identifications measured through a forced choice reference test (on 8 series of minimal pairs or triplets of words) increased from 0.64 at pre-test to 0.86 at post-test on average, which evidences some clear learning effect. Note that this study did not explicitly test the spontaneous / innate ability to interpret tongue movements produced by real speakers.

Bälter *et al*. [3] propose strategies for using ARTUR, their CAPT (Computer Aided Pronunciation Training) system for phonetic correction. ARTUR is based on a virtual talking head that can display both visible and non visible articulators. Their test using a Wizard of Oz type of control of the talking head was well received by the three children involved. What the children liked most was the correction (the instruction on how to improve the pronunciation) e.g. on how to move the tongue more forward or backward to produce the instructed sound. However, no quantitative evaluation was conducted.

Recently, Fagel *et al*. [11] reported results of a study investigating the visual information conveyed by the dynamics of internal articulators. Intelligibility of synthetic audiovisual speech with and without visualization of the internal articulator movements was compared. Additionally identification scores were contrasted before and after a short training session in which vocal tract movements were explained, once with and once without motion of internal articulators. Results show that displaying motion of internal articulators did not lead to significant improvement of identification scores at first, and that training did significantly increase visual and audiovisual speech intelligibility. After the learning lesson with all internal articulatory movements, the visual recognition could be enhanced to a higher degree than the audiovisual recognition. The absolute increase of visual recognition could not be integrated completely into audiovisual recognition. The authors show that this could be due to redundant information conveyed by auditory and visual sources of information.

Vision is obviously involved in speech perception in everyday life as soon as the beginning of speech. The importance of the visual speech input in language acquisition has been well documented by Mills [14] in her review on language acquisition in blind children, with data showing that blind children have difficulty in learning an easy-to-see and hard-to-hear contrast such as [m] vs. [n]. Other relevant data are those showing the predominance of bilabials at the first stage of language acquisition [22], reinforced in hearing impaired children [19], but less clear in blind ones [16].

We may assume that the general articulatory awareness skill (as measured e.g. by Montgomery [15]) would allow subjects to make use of tongue vision for phonemic recognition, as it does for lip reading. The purpose of the present study was therefore twofold: (1) to question the spontaneous or innate ability of subjects to “tongue read”, i.e. their ability to recover information from tongue vision without prior learning, and (2) if the first question is inconclusive, to test 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 [20] or [4]. But if the rendering of movements is artificial, the movements themselves were captured synchronously with the acoustic signal on a real subject.

2. THE TALKING HEAD AND ITS CONTROL

The approach used by Tye-Murray *et al*. [21] was in principle the most appropriate method to determine the tongue reading ability of subjects displaying the actual shape of the tongue recorded from a real human speaker pronouncing the desired corpus of words. However, this approach presents two major drawbacks: (1) the videofluoroscopic images represent the sagittal projection of the whole head of the subject, and thus the individual articulators are not very easy to identify and to track; (2) due to health hazards due to X-rays, the amount of speech material that can be recorded is very limited. In order to overcome these problems, while maintaining 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 [17] 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 model of the velum geometry presented in [18] is controlled essentially by a single parameter that drives the opening / closing movements of the nasopharyngeal port. Finally, the three-dimensional jaw and tongue model developed by Badin *et al*. [2] is mostly driven by five parameters: the main effect of the *jaw height* parameter (common with the jaw height parameter of the lips / face model) is a rotation of the tongue around a point located in its back; the next two parameters, *tongue body*, and *tongue dorsum*, control respectively the *front-back* and *flattening-arching* movements of the tongue; the last two other parameters, *tongue tip vertical* and *tongue tip horizontal* control precisely the shape of the tongue tip. Figure 1 illustrates one possible display of this virtual talking head. Note that the geometry of all the models (except for the inner part of the lips) is defined by three-dimensional surface meshes whose vertices are associated with *flesh points*, i.e. points that can be identified on the organs.

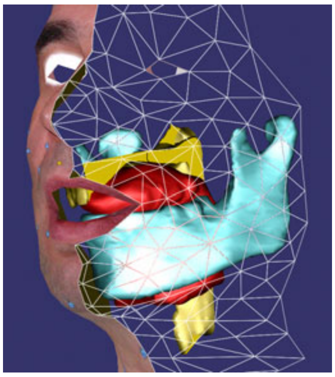

Figure 1 – Example of virtual talking head display. The face, the jaw, the tongue and the vocal tract walls including the hard and soft palates can be distinguished.

2.2 Control of the talking head from EMA recordings

The parameters used to control the various articulatory models can be considered as the *degrees of freedom* of the articulators [1], i.e. the specification, for each articulator, of the limited set of movements that it can execute independently of the other articulators. As our articulatory models are linear, the 3D coordinates of each vertex are linear combinations of the control parameters. These coordinates can thus be simply obtained by multiplying the vector of control parameters by the matrix of the models coefficients. Recovering, by inversion, the control parameters of the tongue and lips / face models can therefore be done from a sufficient number of independent geometric measurements using the (pseudo-) inverse of the model coefficient matrix.

We have thus used the vertical and horizontal coordinates in the midsagittal plane of a set of six coils of an EMA system: a *jaw coil* was attached to the lower incisors, while a *tip coil*, a *mid coil* and a *back coil* were respectively attached at approximately 1.2 cm, 4.2 cm and 7.3 cm from the tongue extremity; an *upper lip coil* a *lower lip coil* were attached to the boundaries between the vermilion and the skin in the midsagittal plane. Extra coils attached to the upper incisor and to the nose served as references. After appropriate scaling and alignment, the coordinates of the coils were obtained in the same coordinate system as the models. No inter subject normalisation was necessary, since the same subject is used both for the models and the EMA measurements.

A specific vertex of the 3D tongue mesh was then chosen and associated to each tongue coil in such a way as to minimise the maximum of the distance between the vertex and the coil for a set of 22 articulations representative of the articulatory space of the subject (cf. [18] for more details). The vertices of the lips / face model surface mesh located at the boundary between the vermilion and the skin for each lip in the midsagittal plane were naturally associated with the lips coils. As a first approximation, the left-right coordinates of all the vertices in the midsagittal plane were assumed to be zero, which turned out to be valid assumption. The three tongue vertices and the two lip vertices associated with the EMA coils have respectively six and four independent coordinates controlled by five parameters each. These control parameters can finally be obtained from the (pseudo-) inverse matrices of the sub-models that predict the coordinates of these specific vertices.

As the *jaw height* and *jaw advance* parameters are directly proportional to the vertical and horizontal coordinates of the jaw coil, they were computed first in practice, and their linear contributions were removed from the measured EMA coordinates. The four remaining tongue parameters were next obtained by the pseudo-inverse of the [4 parameters x 6 coordinates] matrix for the three tongue coils, and the three remaining lips / face parameters were obtained by the pseudo-inverse of the [3 parameters x 4 coordinates] matrix for the two lips coils. The pseudo-inverse matrices give only sub-optimal solution to the inversion problem, since the number of control parameters is lower than the number of measured coordinates. The mean estimation error, defined as the RMS of the distance between the EMA coils and their associated tongue vertices over a set of 44 articulations was 0.17 cm.

Although in most cases this inversion procedure yields satisfactory results, the resulting tongue contours sometimes cross the hard palate contours (cf. Figure 2), as a consequence of the not modelled nonlinear effect of tongue tip compression when in contact with the palate. In such case, the four tongue parameters are slightly adjusted, using a constrained optimisation procedure (Matlab *fminimax*) that minimises the distance between the coils and the three specific tongue model vertices, with the constraint of preventing the tongue contour from crossing the palate contour. Figure 2 illustrates the results.

Note finally, that for practical reasons, no coil was attached to the velum in this experiment, and thus no nasal sounds were involved in the study.

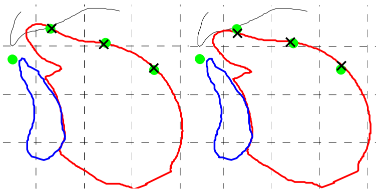

Figure 2 – Midsagittal contours of tongue and jaw models for /da/, coils location (grey dots) and model vertices attained by inversion. Left: the result obtained by the simple pseudo-matrix inversion procedure; right, the result after the constrained adjustment.

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.

As the aim was to assess the contribution of tongue vision, and assuming that the voiced and voiceless cognates present only very minor visible differences, we have chosen to collect the identification scores of all French voiced consonants except nasal ones, i.e. /b d g v z ʒ ʁ l/. In order to compare different rendering conditions, 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 o/ was used for the 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. We were also interested in comparing the contribution of the lips and face with the contribution of the tongue. Therefore, we have tested four presentation conditions:

1. Audio signal alone (AU)
2. 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)
3. Audio signal + cutaway view of the virtual head along the sagittal plane *with T*ongue (AV*T*) (see Figure 3)
4. Audio signal + complete *F*ace with skin texture (AV*F*) (see Figure 3)

The cutaway presentation on Figure 3 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. Since no EMA coil was attached to the velum, it was not animated in the present study. A profile view was chosen as it provides the most information on the tongue, given that the angle of view does not change very much lip reading scores (cf. e.g. [12]).

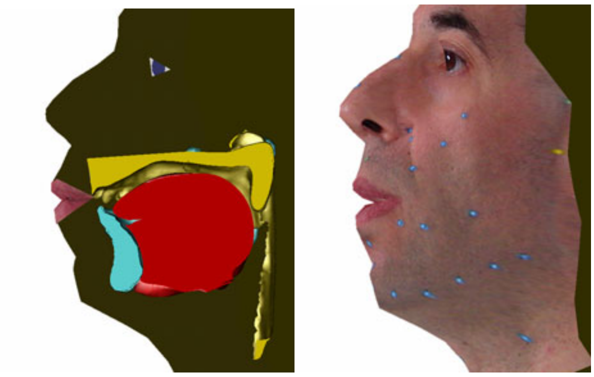

Figure 3 – Examples of presentation conditions for the audiovisual test: cutaway view of the head including tongue (left) *vs.* complete face with skin texture (right).

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 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). The special combination of no audio with the AU condition was discarded. The intermediate SNRs were set to –9 dB and +3 dB in order to yield identification scores in the AU condition of respectively 33 % and 66 % gathered in a preliminary test. Note that the noise spectrum was a white, that the level of the noise added was estimated with reference to average energy of the speech signal over vocalic parts of each stimulus, and that the level of each stimulus was normalised in order to avoid unjustified changes between signals played to the subjects.

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 PC 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: after the presentation, the subject has to choose among the eight possible consonants /b d g v z ʒ ʁ l/ by pointing to the appropriate box with the computer mouse. No repetition was allowed and subjects were instructed to answer as soon as possible. The next stimulus is played about one second after the response. The subjects were instructed to answer randomly in case of hesitation between different answers.

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) in the AVT condition (cutaway view *with* tongue) with a –9 dB SNR. This is not a training period: the right answers are not given.

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 familiarised with the new conditions (these two dummy stimuli are not taken into account in the results).

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 23 French subjects, with no known hearing nor non corrected sight losses, without prior experience in speech organs study nor analysis. The 12 subjects from group I (7 females and 5 males, mean age 27.2 years) performed the tests in the decreasing SNR order, while the 11 subjects from group II (4 females and 7 males, mean age 26.9 years) performed the tests in the increasing SNR order.

The duration of a complete test session ranged from 30 to 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 after their sessions. Some subjects 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 4 represents the mean identification scores, i.e. the percentage of consonants correctly identified for the 16 different test series, separately for the two groups of subjects.

Note first that the results displayed in Figure 4 for the AU and AVF conditions are coherent with those obtained by [5] for different lips / face presentation conditions: for instance [5] found, at an SNR = −9 dB, an increase of the identification score of 34 % from the AU condition to a condition of full synthetic head vision, while we get an increase of about 37 % from the AU to the AVT condition.

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.

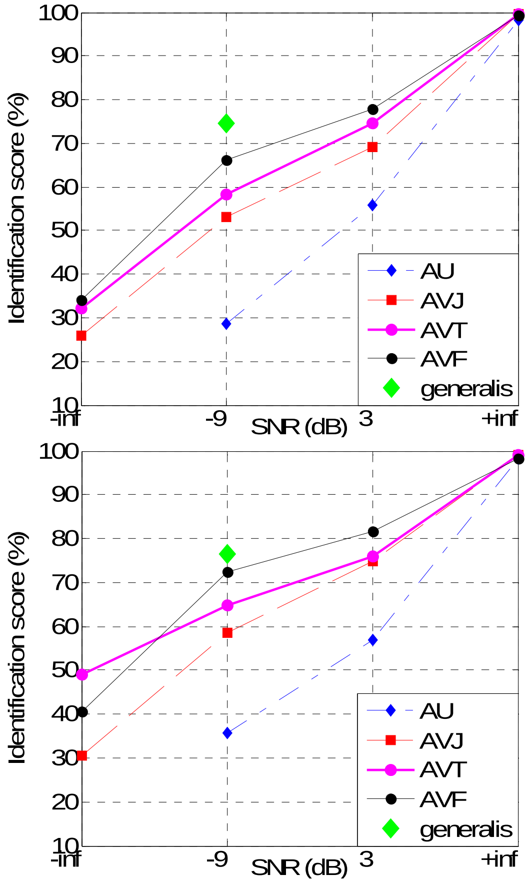

Figure 4 – Mean identification scores as a function of SNR (top: group I; bottom, group II) for the different conditions (from bottom to top): AU, AVJ, AVT, AVF. The isolated diamond indicates the score for the generalisation test.

4.3 Generalisation test

Since the same set of 32 VCV sequences has been used in the first 15 series of the session, we have 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 the learning of tongue reading. 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*. [6] 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.* [11], but using a less elaborate tongue model and synthetic movements arrived to similar conclusions.

These very preliminary tests need to be complemented by more systematic tests, 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

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 [13], and (3) pronunciation training for second language learners, as discussed by [9].

6. ACKNOWLEDGMENTS

We sincerely thank Christophe Savariaux and Coriandre Vilain for helping us with the EMA recordings, and Frédéric Berthommier, Amélie Rochet-Capellan and Jean-Luc Schwartz for helpful discussions about visual perception and statistics.

7. REFERENCES

[1] Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C., "Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533-553, 2002.

[2] Badin, P. and Serrurier, A., "Three-dimensional linear modeling of tongue: Articulatory data and models," presented at Proceedings of the 7th International Seminar on Speech Production, ISSP7, Ubatuba, SP, Brazil, 2006.

[3] Bälter, O., Engwall, O., Öster, A.-M., and Kjellström, H., "Wizard-of-Oz Test of ARTUR - a Computer-Based Speech Training System with Articulation Correction," presented at Proceedings of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility, Baltimore, 2005.

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

[5] Benoît, C. and Le Goff, B., "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.

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

[7] Cohen, M.M., Beskow, J., and Massaro, D.W., "Recent developments in facial animation: an inside view," presented at Proceedings of the International Conference on Auditory-Visual Speech Processing / Second ESCA ETRW on Auditory-Visual Speech, Terrigal-Sydney, Australia, 1998.

[8] Cornett, O., "Cued Speech," *American Annals of the Deaf*, vol. 112, pp. 3-13, 1967.

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

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

[11] Grauwinkel, K., Dewitt, B., and Fagel, S., "Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech," presented at Interspeech'2007 - Eurospeech - 9th European Conference on Speech Communication and Technology, 2007.

[12] IJsseldijk, F.J., "Speechreading performance under different conditions of video image, repetition, and speech rate," *Journal of Speech and Hearing Research*, vol. 35, pp. 466-471, 1992.

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

[14] Mills, A.E., "The development of phonology in the blind child," in *Hearing by eye: the psychology of lipreading*, Dodd, B. and Campbell, R., Eds. London: Lawrence Erlbaum Associates, 1987, pp. 145-161.

[15] Montgomery, D., "Do dyslexics have difficulty accessing articulatory information?," *Psychological Research*, vol. 43, 1981.

[16] Mulford, R., "First words of the blind child," in *The emergent lexicon: The child's development of a linguistic vocabulary*, Smith, M.D. and Locke, J.L., Eds. New-York: Academic Press, 1988, pp. 293-338.

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

[18] Serrurier, A. and Badin, P., "A three-dimensional articulatory model of nasals based on MRI and CT data," *Journal of the Acoustical Society of America*, in revision.

[19] Stoel-Gammon, C., "Prelinguistic vocalizations of Hearing-Impaired and Normally Hearing subjects. A comparison of consonantal inventories," *Journal of Speech and Hearing Disorders*, vol. 53, pp. 302-315, 1988.

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

[21] Tye-Murray, N., Kirk, K.I., and Schum, L., "Making typically obscured articulatory activity available to speechreaders by means of videofluoroscopy," *NCVS Status and Progress Report*, vol. 4, pp. 41-63, 1993.

[22] Vihman, M.M., Macken, M.A., Miller, R., Simmons, H., and Miller, J., "From babbling to speech: A re-assessment of the continuity issue," *Language*, vol. 61, pp. 397-445, 1985.