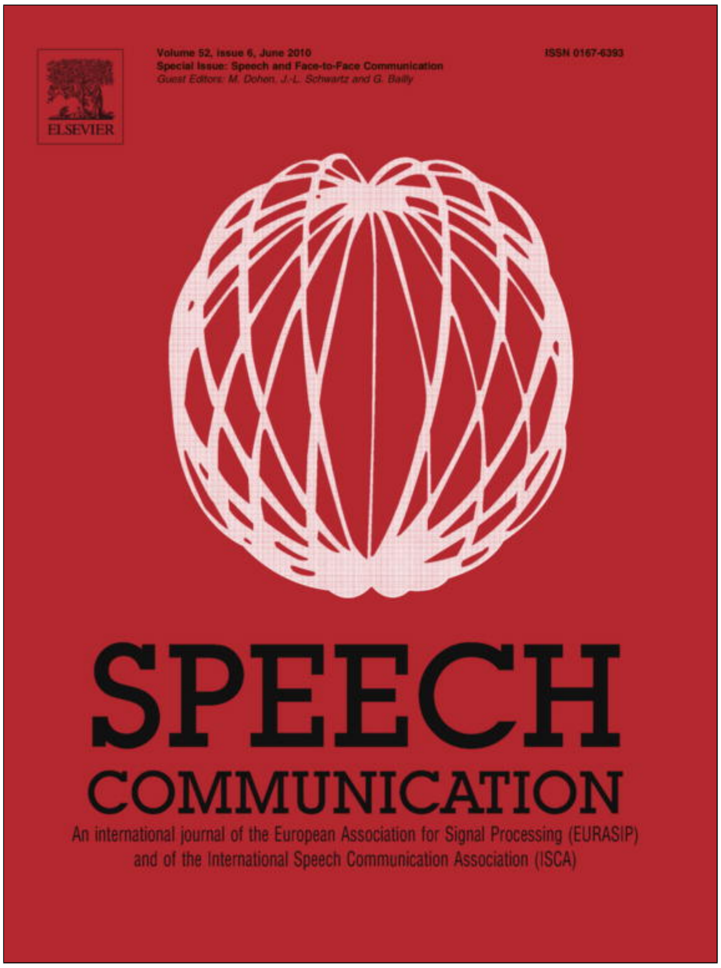
ELSEVIER
Volume 52, Issue 6, June 2010
Special Issue: Speech and Face-to-Face Communication
Guest Editors: M. Dohen, J.-L. Schwartz and G. Bailly
ISSN 0167-6393
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 52 (2010) 493–503

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

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

Pierre Badin [*], Yuliya Tarabalka, Frédéric Elisei, Gérard Bailly

Département Parole et Cognition/ICP, GIPSA-lab, UMR 5216 CNRS, Grenoble University, France

Received 22 April 2009; received in revised form 1 March 2010; accepted 3 March 2010

Abstract

Lip reading relies on visible articulators to ease 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 is thus whether the direct and full vision of the tongue allows *tongue reading*. We have therefore generated a set of audiovisual VCV stimuli with an audiovisual talking head that can display all speech articulators, including tongue, in an *augmented speech* mode. The talking head is a virtual clone of a human speaker and the articulatory movements have also been captured on this speaker using ElectroMagnetic Articulography (EMA). These stimuli have been played to subjects in 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 indicate: (1) the possibility of implicit learning of tongue reading, (2) better consonant identification with the cutaway presentation with the tongue than without the tongue, (3) no significant difference between the cutaway presentation with the tongue and the more ecological rendering of the complete face, (4) a predominance of lip reading over tongue reading, but (5) a certain natural human capability for tongue reading when the audio signal is strongly degraded or absent. We conclude that these tongue reading capabilities could be used for applications in the domains of speech therapy for speech retarded children, of perception and production rehabilitation of hearing impaired children, and of pronunciation training for second language learners.

Keywords: Lip reading; Tongue reading; Audiovisual speech perception; Virtual audiovisual talking head; Augmented speech; ElectroMagnetic Articulography (EMA)

1. Introduction

Vision is obviously involved in speech perception as soon as the beginning of life. The importance of the visual speech input in language acquisition has been well documented by Mills (1987) in her review on language acquisition by blind children, with data showing that these children have difficulty in learning easy-to-see but hard-to-hear contrasts such as [m] *vs.* [n]. Other relevant data show the predominance of bilabials at the first stage of language acquisition (Vihman et al., 1985), reinforced in hearing impaired children (Stoel-Gammon, 1988), but less dominant in blind ones (Mulford, 1988).

A number of studies has established more quantitatively how the vision of visible articulators (lips, jaw, face, tongue tip, teeth) eases speech understanding, and how significantly it increases the performance of detection and identification of words in noise (Erber, 1975). Sumby and Pollack (1954) and Benoît and Le Goff (1998), among others, have quantified the gain in speech intelligibility provided by lip reading in comparison with the sole acoustic signal.

Though very useful, the mere vision of the lips and face provides incomplete phonetic information: for instance, laryngeal activity or movements of tongue and velum are

[*] Correspondence to: Pierre Badin, GIPSA-lab/DPC, ENSE3, 961 rue de la Houille Blanche, BP 46, F-38402 Saint Martin d'Hères cedex, France. Tel.: +33 (0)476 57 48 26; fax: +33 (0)476 57 47 10.

E-mail addresses: Pierre.Badin@gipsa-lab.grenoble-inp.fr (P. Badin), Yuliya.Tarabalka@gipsa-lab.grenoble-inp.fr (Y. Tarabalka), Frederic.Elisei@gipsa-lab.grenoble-inp.fr (F. Elisei), Gerard.Bailly@gipsa-lab.grenoble-inp.fr (G. Bailly).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.03.002

not visible. Indeed, the tongue, that carries an important part of the articulatory information, can – at best – be only very partially seen in usual conditions. "Cued Speech", elaborated by Cornett (1967), used by some hearing impaired speakers, aims precisely at complementing lip information by a set of hand gestures and shapes that provides most of the missing phonetic information, in particular related to mode of articulation and tongue articulation. This coding system, although efficient in terms of information theory, is arbitrary, and not directly related to tongue movements. Considering the fact that humans possess to some degree articulatory awareness skills, as measured *e.g.* by Montgomery (1981), it can be hypothesised that human subjects[1] might be able to make use of tongue vision for phonemic recognition, as they do with the lips in lip reading. The central question of our study is thus to determine whether direct and full vision of the tongue – information presumably more intuitive than hand cues – can be natively used and processed by human subjects, in a way similar to that in lip reading. The present article, which is an extended version of Badin et al. (2008), describes our experiments and our findings.

The literature on this subject is rather scarce. The first study that we are aware of and that tests the ability of subjects to perceive and process tongue movements was performed by Tye-Murray et al. (1993). They conducted experiments to determine whether increasing the amount of visible articulatory information could influence speech comprehension, and whether such artefacts are effectively beneficial. The experiments involved profile view videofluoroscopy, which allows movements of the tongue body, lips, teeth, mandible, and often velum, to be observed in real-time during speech, as well as profile view videoscopy of the same speaker. Subjects were asked to speech read 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 with videofluoroscopy. They performed equally well whenever the tongue was visible in the videofluoroscopic records or not. Training did not improve their ability to recognise speech presented with videofluoroscopy. These conclusions should however be considered with caution, as the quality and the interpretability of videofluoroscopic images was not very high.

More recently, this idea has been revisited by several authors, thanks to the development of audiovisual talking heads that can display a subject's natural looking face as well as usually hidden internal articulators.

Grauwinkel et al. (2007) reported the results of a study investigating the visual information conveyed by the dynamics of internal articulators. Intelligibility of synthetic audiovisual speech produced by a talking head seen from profile, with and without active movement of the internal articulators (tongue and velum), was compared through perception tests at an audio SNR of 0 dB on a corpus of synthesised German Vowel–Consonant–Vowel (VCV) sequences where V = [a i u] and C = [b d g v z ʃ l m n ŋ]. Results showed that displaying the movements of internal articulators did not lead to significant improvement of identification scores at first, but that training did significantly increase visual and audiovisual speech intelligibility.

Kröger et al. (2008) conducted a visual perception experiment where a group of children was instructed to identify (by mimicry) mute animations of vocalic and consonantal speech movements displayed by a 2D midsagittal visual articulatory model. Four synthetic [neutral-V] transitions with V = [a i y u] and eleven [aCa] sequences where C = [p t k f s ʃ ç x m n ŋ] were generated by means of a dominance model of coarticulation. Identification scores of about 20% (the chance level being 7% for 15 stimuli) supported the hypothesis that some of the information provided by the vision of vocal tract internal articulators might be intuitively perceived by children without any prior learning.

Wik and Engwall (2008) conducted a similar study aiming to investigate the benefit of the vision of internal articulators for speech perception. They asked subjects to identify words in acoustically degraded Swedish sentences for three different presentation conditions: audio only, audiovisual with a front view of a synthetic face, and audiovisual with both front face view and a side view where tongue movements were visible. They reported that the *augmented reality* side view did not help subjects perform better overall than with the front view only, but that it seemed to have been beneficial for the perception of palatal plosives, liquids and rhotics, especially in clusters. Their results indicate that subjects have difficulty in exploiting intra-oral animations in general, but that information on some articulatory features can be extracted and have impacts on speech perception.

The use of computer-animated talking heads as language tutors for speech rehabilitation or pronunciation training has started to spread, but in most cases the specific contribution of tongue or internal articulators to learning or rehabilitation is not extracted from the overall contribution of the complete talking head.

Massaro and Light (2003) investigated the effectiveness of a virtual talking head for teaching non-native phonetic contrasts, by comparing instruction illustrating the internal articulatory processes of the oral cavity versus instruction providing an external view of the talking head's face. The view of the internal articulators, which were again controlled by text-to-speech synthesis and not by motion capture, did not show any additional benefit. Massaro and Light (2004) used the same computer-animated talking head as a language tutor for speech perception and production for individuals with hearing loss; they measured some quantitative improvement of the children's performances, but did not assess explicitly the specific benefit of displaying internal articulators.

[1] The traditional term for *subject* when referring to aural perception is *listeners*. In the case of audiovisual perception, a more appropriate term would be *listener–viewer*. However, for simplicity's sake, the general term *subject* will be used throughout the article.

A Swedish computer-animated talking head that displays internal articulators was used by Bälter et al. (2005) for the phonetic corrections of children, and by Engwall (2008) for second language pronunciation training of Swedish difficult phonemes by French speaking subjects. They found that training with the talking head improved the speech production performance of the students, but they did not perform any explicit evaluation of the benefit of the vision of the tongue.

In a limited study involving a few children in pronunciation correction, Fagel and Madany (2008) showed that a 3D talking head was an applicable tool for speech therapy, though they did not attempt to assess explicitly the usefulness of the vision of the internal articulators.

The purpose of the present study is 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 specific learning, and (2) if the first question is inconclusive, to test their ability to rapidly learn this skill. The talking head developed at the Speech and Cognition Department of GIPSA-lab was used in an audiovisual perception test based on the paradigm of noise degradation of the audio signal used by Sumby and Pollack (1954) or Benoît et al. (1996).

2. The talking head and its control from EMA recordings

The approach used by Tye-Murray et al. (1993), *i.e.* displaying the actual shape of the tongue recorded from a real human speaker pronouncing the desired corpus of words, is in principle the most appropriate method to determine the tongue reading ability of subjects. However, this approach has two major drawbacks: (1) the videofluoroscopic images represent the sagittal projection of the whole head of the speaker, and thus the individual articulators are not very easy to identify and to track; (2) due to health hazards induced by X-rays, the amount of speech material that can be safely recorded is very limited. An alternative approach, made possible by recent progress in real-time MRI imaging, could produce midsagittal images of the vocal tract, but with a poor rate and a low resolution (*cf. e.g.* Narayanan et al. (2004), who obtained a raw image rate of 9 Hz, interpolated to 24 Hz, and an image resolution of 2.7 mm per pixel). In order to overcome these problems, while maintaining the ecological quality of the stimuli, we build the audiovisual stimuli for our perception experiments using original natural speech sounds and articulatory movements recorded synchronously by an ElectroMagnetic Articulography (EMA) device on one speaker. The recorded movements were then used to drive a virtual talking head based on extensive measurements on the same speaker.

2.1. The talking head

Our virtual talking head is the assemblage of individual 3D models of various speech organs (jaw, tongue, lips, velum, and face) of the same speaker. These models are built from Magnetic Resonance Imaging (MRI), Computer Tomography (CT) and video data acquired from this speaker and aligned on a common reference coordinate system related to the skull. The jaw, lips and face model, described in Odisio et al. (2004), is controlled by two jaw parameters (*jaw height*, *jaw advance*), and three lip parameters (*lip protrusion LP* common to both lips, *upper lip height UL*, *lower lip height LL*). The velum model presented in Serrurier and Badin (2008) is essentially controlled by one parameter that drives the opening/closing movements of the nasopharyngeal port. Finally, the 3D jaw and tongue model developed by Badin and Serrurier (2006) is primarily driven by five parameters: the main effect of the *jaw height* parameter *JH* (common with the jaw height parameter of the lip/face model) is a rotation of the tongue around a point located in its back; the next two parameters, *tongue body* (*TB*), and *tongue dorsum* (*TD*), control respectively the *front–back* and *flattening–arching* movements of the tongue; the last other two parameters, *tongue tip vertical* (*TTV*) and *tongue tip horizontal* (*TTH*) control precisely the shape of the tongue tip. Note that the *jaw height* parameter is common to all soft organs models and is also used, as well as the *jaw advance* parameter, to control the movements of the jaw itself. Fig. 1 illustrates one possible display of this virtual talking head that can produce *augmented speech*, as it can display more than a real face. Note that the geometry of all the articulated models (except for the inner part of the lips) is defined by 3D surface meshes whose vertices are associated with *flesh points*, *i.e.* points that can be identified on the organs.

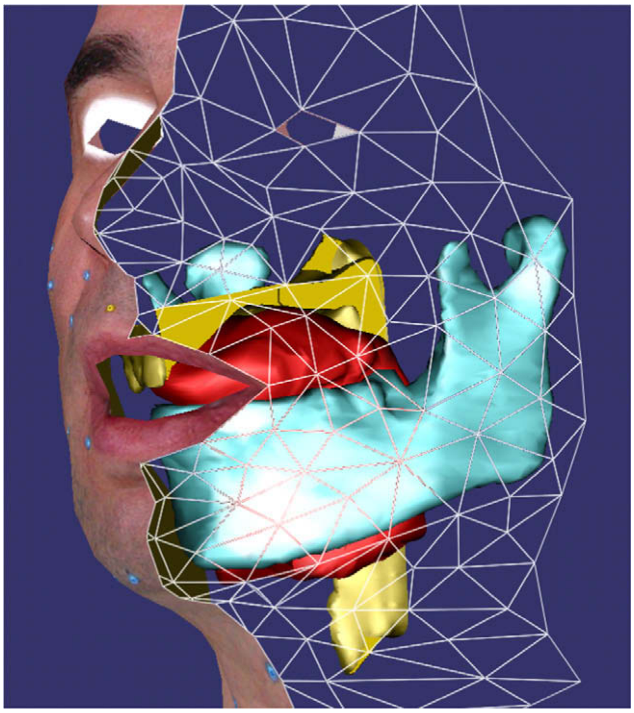

Fig. 1. Example of virtual talking head display. The face, the jaw, the tongue and the vocal tract walls including the hard and soft palates can be distinguished.

2.2. Control of the talking head from ElectroMagnetic Articulograph (EMA) recordings

Badin and Serrurier (2006) have quantitatively assessed the possibility to infer the 3D shape of the tongue from its midsagittal shape. Using an optimisation procedure, they determined the set of control parameters of a 3D tongue surface model that gives the best fit *in the midsagittal plane* to measured 3D tongue surfaces: they found – over the 46 articulations of the corpus used to build the 3D tongue model – an average RMS 3D reconstruction error of 0.25 cm, which compares well the 0.22 cm corresponding to the modelling error itself. This testifies to the very good predictability of the 3D tongue surface mesh from its 2D midsagittal contour. This is why, in the present work, we have determined the control parameters of the 3D models from the midsagittal coordinates of a small number of EMA sensors placed on the sagittal plane of the jaw, tongue and lips.

As the articulatory models involved in our talking head 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 model coefficients, as exemplified for the tongue in Eq. (1):

$$\mathrm{Tng3D}(1:3\times \mathrm{np},1) = \mathrm{Mod}(1:3\times \mathrm{np},1:5)\times \mathrm{art}(1:5,1) + \overline{\mathrm{Tng3D\_m}}(1:3\times \mathrm{np},1) \qquad (1)$$

where Tng3D(1:3 × np, 1) is the vector of the tongue vertex 3D coordinates (np is the number of vertices), Mod(1:3 × np, 1:5) the matrix of coefficients of the linear model, art(1:5, 1) the vector of articulatory control parameters (JH, TB, TD, TTV, TTH), and $\overline{\mathrm{Tng3D\_m}}$ (1:3 × np, 1) the mean of Tng3D(1:3 × np, 1) on the corpus. Recovering, by inversion, the control parameters of the tongue and lip/face models can therefore be done from a sufficient number of independent geometric measurements using the (pseudo-) inverse of the models coefficient matrices, as will be described further.

ElectroMagnetic Articulography (EMA) is an experimental method that infers the coordinates of small electromagnetic coils from the magnetic fields that they receive from electromagnetic transmitters (*cf.* Perkell et al., 1992 or Hoole and Nguyen, 1997). We have used the vertical and horizontal coordinates in the midsagittal plane of a set of six coils of a 2D EMA system (Carstens AG200): a jaw coil was attached to the lower incisors, whereas three coils were attached to the tongue tip, the tongue middle, and the tongue back at approximately 1.2, 4.2, and 7.3 cm, respectively, from the extremity of the tongue; an upper lip coil and 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 incisors and to the nose served as references. The coordinates of these receiving coils were recorded at a sampling frequency of 200 Hz and subsequently low pass filtered at 20 Hz. The speech sound was recorded on a digital audio recorder at 22,005 Hz, and later synchronised with the EMA tracks by means of the square pulse signal produced by the EMA system. After appropriate scaling and alignment, the coordinates of the coils were obtained in the same coordinate system as the models. No inter speaker normalisation was necessary, since the same speaker was used for both the models and the EMA measurements.

Two video cameras recorded front views of the speaker's face covered with small markers, but the records were not used in the present study. Note that the markers visible on the reconstructed skin texture in Figs. 1 and 3 are actually those used to build the skin texture, and not those used in the present experiment.

Each tongue coil is associated with a specific vertex of the 3D tongue model surface mesh. Each vertex was thus determined in such a way as to minimise the maximum of its distance to the corresponding tongue coil for a set of 22 articulations representative of the articulatory space of the speaker (as done by Serrurier and Badin (2008) for the velum). The lips coils were naturally associated with the vertices of the lip/face model surface mesh located at the boundary between the vermilion and the skin for each lip in the midsagittal plane.

As expected, the resulting vertices were found close to the midsagittal plane, and their left–right coordinates were therefore assumed to be zero. The three tongue vertices and the two lip vertices associated with the EMA coils have respectively six and four coordinates controlled by five parameters each. The equations of the "sub-models" restricted to these specific vertices for the tongue can then be expressed as follows:

$$\mathrm{Tng3D}(1:2\times \mathrm{nc},1) = \mathrm{Mod}(1:2\times \mathrm{nc},1:5)\times \mathrm{art}(1:5,1) + \overline{\mathrm{Tng3D\_m}}(1:2\times \mathrm{nc},1) \qquad (2)$$

where Tng3D(1:2 × nc, 1) is the vector of the coordinates in the midsagittal plane of the tongue vertices associated with the three tongue coils (nc = 3), and Mod(1:2 × nc, 1:5) the matrix of coefficients of the linear model restricted to the coordinates of corresponding vertices. As the *jaw*

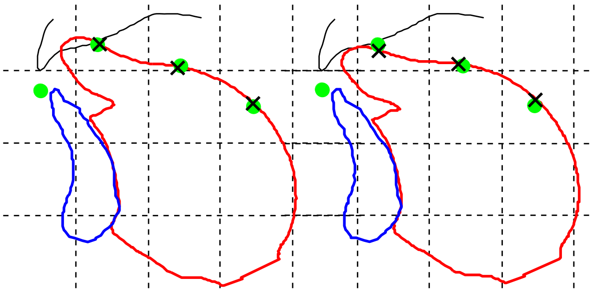

Fig. 2. Midsagittal contours of tongue and jaw models for /d^a/, coils location (green dots) and model vertices (black crosses) attained by inversion. Result obtained by the simple pseudo inversion matrix procedure (left); result after the constrained adjustment (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

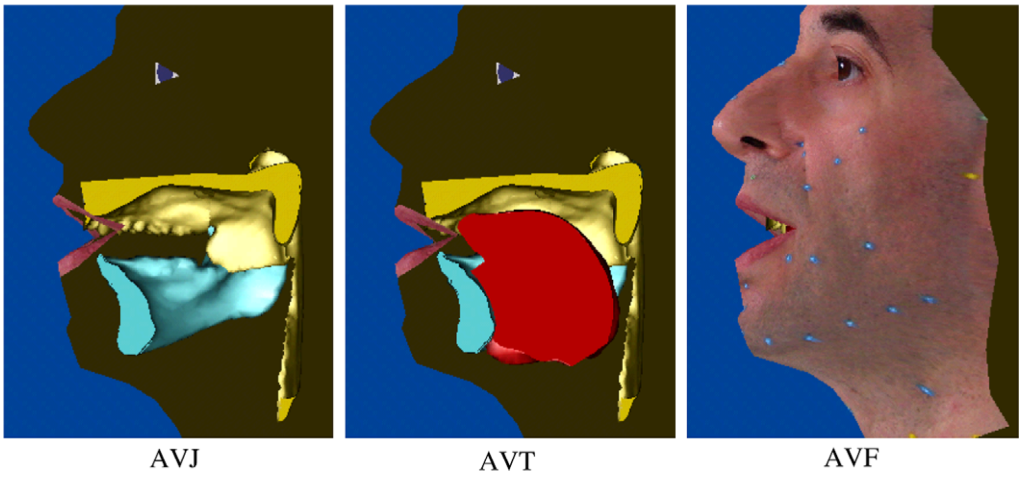

Fig. 3. Examples of presentation conditions for the audiovisual test: cutaway view of the head without tongue (AVJ, left), cutaway view of the head including tongue (AVT, centre), complete face with skin texture (AVF, right).

height and *jaw advance* parameters are directly proportional to the vertical and horizontal coordinates of the jaw coil, they were computed first; their linear contributions were then subtracted from the measured EMA coils coordinates before further analysis, which ensured a better precision. The control parameters could finally be obtained from the (pseudo-) inverse matrices of the sub-models, as expressed in Eq. (3):

$$\begin{aligned}\mathrm{art\_inv}(2:5,1) = {} & \mathrm{Mod\_Pinv}(2{:}5, 1{:}2 \times \mathrm{nc}) \\ & \times (\mathrm{Tng3D\_res}(1{:}2 \times \mathrm{nc},1) \\ & - \overline{\mathrm{Tng3D\_m}}(1{:}2 \times \mathrm{nc},1)),\end{aligned} \qquad (3)$$

where Mod_Pinv(2:5, 1:2 × nc) is the pseudo-inverse of matrix Mod(1:2 × nc, 2:5), Tng3D_res(1:2 × nc, 1) is the residue of Tng3D(1:2 × nc, 1) after subtraction of the *jaw height* contribution, and art_inv(2:5, 1) the remaining inversed articulatory control parameters (TB, TD, TTV, TTH). The pseudo-inverse matrices generate only sub-optimal solutions 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 46 articulations was 0.17 cm.

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

In order to verify that the three coils midsagittal coordinates contain enough information for the reconstruction of the 3D tongue shape, we compared the RMS reconstruction error of the 3D shapes from the first six components of the PCA of the entire midsagittal contour with that obtained from the three coils midsagittal coordinates only. The three coils did not lead to a significantly higher error, which justified their use for inversion, and thus our approach.

Although in most cases this inversion procedure yielded satisfactory results, the resulting tongue contours sometimes crossed the hard palate contours (*cf.* Fig. 2), as the non-linear effect of tongue tip compression when in contact with the palate is not modelled at present. In such cases, the four tongue parameters were slightly adjusted, using a constrained optimisation procedure that minimises the distance between the coils and the three specific tongue model vertices, the constraint being to prevent the tongue contour from crossing the palate contour. Fig. 2 exemplifies the results.

The lip data were processed in a similar way: the three lip/face parameters ([LP, UL, LL]) were obtained by the pseudo-inverse of the [3 parameters × 4 coordinates] matrix for the two lips coils, after subtraction of the *jaw height* and *jaw advance* contributions.

An important advantage of this approach for animation control based on motion capture is that the articulatory dynamics is entirely preserved. This results in very naturally moving animations, in contrast to approaches based on a resynthesis of these movements from acoustic or phonetic specifications, which may produce less realistic movements.

3. The audiovisual perception tests

The present study of the contribution of the vision of various articulators to speech perception is based on the *paradigm of noise degradation of the audio signal* used by Sumby and Pollack (1954) or Benoît et al. (1996): the identification score of audiovisual speech stimuli by a panel of listener–viewer subjects is measured for different levels of noise added to the audio signal. This section describes the tests that were conducted.

3.1. Corpus

Carrying out perception tests always faces the dilemma of finding a balance between the largest number of stimuli and the necessarily limited duration that subjects can endure.

As the aim of the study was to assess the contribution of tongue vision, and assuming that the voiced and voiceless cognates present only very minor visible differences, we chose to collect the identification scores of all French voiced oral consonants, *i.e.* /b d g v z ʒ ʁ l/ (no nasal articulations are considered in the present study). The consonants were embedded in symmetrical VCV sequences with vowels V = /a i u y/. This set was constituted of the three cardinal vowels which are the most contrasted ones, complemented with /y/, which has a labial shape nearly identical to that of /u/, but differs from it by a strong front lingual position. The main corpus was finally composed of 32 VCV stimuli. Three additional sets of VCV sequences, with vowels /ɛ e o/, were used for a generalisation test, and one set with /œ/ are used for a familiarisation step (see further).

3.2. The presentation conditions and SNRs

In order to assess the contribution of the tongue display, the test contrasted a condition in which the tongue was visible with a condition in which 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, four presentation conditions were finally used (as illustrated in Fig. 3):

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*)
4. Audio signal + complete synthetic *F*ace model with synthetic skin texture (AV*F*)

The cutaway presentations in Fig. 3 show, the lips (pink), the jaw (cyan), the tongue (red), the hard palate (yellow), the velum (yellow), and the back of the vocal tract wall from nasopharynx down to larynx (yellow). Since no measurement was performed on the velum, it was not animated in the present study. A profile view was chosen as it provides most tongue information. This choice was also valid for the lips, as Cathiard et al. (1996) found that the angle of view of the speaker's face has nearly no influence on the identification scores in audiovisual perception tests, confirming results obtained by IJsseldijk (1992). We could have also included a front view condition to test the visibility of the tongue behind the teeth, but the principal aim of the work was to study the full vision of the tongue, and a fifth condition would have increased the load of the perception test for the subjects.

The identification score of the consonants in the VCV sequences were measured for different levels of noise added to the original audio signal recorded during the EMA data acquisition. 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 impossible combination of no audio with the AU condition was obviously 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% as measured in a preliminary test. Note that (1) the noise spectrum was white, (2) the level of the noise added was estimated with reference to the speech signal energy averaged over the vocalic parts of each stimulus, and (3) the sound level of each stimulus was normalised in order to avoid unjustified changes of overall level between signals played to the subjects (no specific auditory weighting was used for the normalisation).

> Video examples are available at http://www.gipsa-lab.fr/~pierre.badin/SpeechCommTongueReading/index.html
> pb_ada + 3db.avi: [ada], +3 dB SNR, cutaway view of the head without tongue
> pb_ibi-9 dB.avi: [ibi], −9 dB SNR, complete face with skin texture
> pb_ulu.avi: [ulu], clear audio signal, cutaway view of the head with tongue
> pb_phrm6.avi: "La langue peut prendre une position avancée comme dans le 'i', ou encore une position reculée comme dans le 'ou'. . ., ou une position basse comme dans le 'a'. [English: "The tongue can take a forward position as in 'i', or still a backward position as in 'oo', or a low position as in 'a']" [lalɑ̃gpøpʁɑ̃dʁynpozisjɔ̃avɑ̃sekɔmdɑ̃lø_i_uɑ̃kɔʁynpozisjɔ̃ʁøkylekɔmdɑ̃lø_u_uynpozisjɔ̃baskɔmdɑ̃lø_a], +∞ SNR (clear audio), cutaway view of the head with tongue.

3.3. The protocol

The test ran as follows. An audiovisual stimulus was played to the subject once, without repetition, by means of a personal computer with a 17′ thin-film transistor (TFT) screen and high quality headphones at a comfortable listening level (this level was adjusted by each subject, and was thus potentially different for each subject). The task of the subject was to identify the consonant in a forced choice test: after the presentation, the subject had to choose among the eight possible consonants /b d g v z ʒ ʁ l/ by clicking in the appropriate box with the computer mouse. No repetition was allowed. The subject was instructed to answer as fast as possible. The next stimulus was played about one second after the response. The subject was instructed to answer randomly when (s)he could not decide.

In order to allow subjects to get accustomed with the test procedure, the session started with the play of demonstration sentences for the four presentation conditions. It was followed by a dummy test with a set of five VCV sequences (with vowel /œ/, which was not used in the real tests) in the AVT condition (cutaway view *with* tongue) with a 9 dB SNR; this dummy test was not considered as a training session, as the correct answers were not given.

The main test was made of 15 successive sets of presentations of the same 32 VCV sequences (all combinations where V = /a i u y/, and C = /b d g v z ʒ ʁ l/). Each set was defined by its presentation condition and its SNR, as described in Table 1.

For each set, the stimuli were presented in a randomised order, different for each subject, preceded by two dummy stimuli with vowel /œ/ to help the subject to get familiarised with the new conditions (the answers associated with these stimuli were not taken into account in the results).

As the experiment aimed to determine the spontaneous ability of the subjects to get information from different visual conditions, we tried to avoid learning as much as possible. Since the 15 sets contained the same 32 VCV sequences, stimuli were presented in the order of increasing visual information, AVJ providing more information than AU, and AVT more information than AVJ. No specific hypothesis was made about the AVF condition in relation to AVJ and AVT. It was assumed that – within each visual condition – the association between sound and image would be more efficiently learned for high SNRs than for low ones. In order to assess this hypothesis, the subjects were divided in two groups: group I subjects received the tests with increasing SNRs (*i.e.* decreasing noise) within each presentation condition, while group II subjects received the tests with decreasing SNRs (*i.e.* increasing noise) within each presentation condition. The possible differences in the results were used to test the implicit learning that occurs when no noise is added.

Finally, in order to assess the generalisation abilities of the subjects, *i.e.* if they did learn to tongue "read", and to verify whether they did not learn the stimuli *per se*, a supplementary set of presentations, using stimuli never presented previously in the test (24 VCV sequences, where V = /ɛ e o/, and C = /b d g v z ʒ ʁ l/), was run (*cf.* last line of Table 1).

Table 1
Characteristics of the sets of tests.

Stimuli	Set #	Condition	SNR group I	SNR group II
/b d g v z ʒ ʁ l/	1	AU	−9 dB	+∞
× /a i u y/	2	AU	+3 dB	+3 dB
	3	AU	+∞	−9 dB
/b d g v z ʒ ʁ l/	4	AVJ	−∞	+∞
× /a i u y/	5	AVJ	−9	+3 dB
	6	AVJ	+3	−9 dB
	7	AVJ	+∞	−∞
/b d g v z ʒ ʁ l/	8	AVT	−∞	+∞
× /a i u y/	9	AVT	−9 dB	+3 dB
	10	AVT	+3 dB	−9 dB
	11	AVT	+∞	−∞
/b d g v z ʒ ʁ l/	12	AVF	−∞	+∞
× /a i u y/	13	AVF	−9 dB	+3 dB
	14	AVF	+3	−9 dB
	15	AVF	+∞	−∞
/b d g v z ʒ ʁ l/	16	AVT	−9 dB	−9 dB
× /ɛ e o/				

3.4. The subjects

We selected 23 native French subjects, with no known hearing problems or non corrected sight losses, with no prior experience in either speech organs study or analysis. The 12 subjects from group I (7 females and 5 males, mean age 27.2 years) performed the tests in the increasing SNR order, while the 11 subjects from group II (4 females and 7 males, mean age 26.9 years) performed the tests in the decreasing SNR order.

The duration of a complete test session ranged from 30 to 50 min, depending on the subjects.

4. Results

4.1. Informal comments

Before presenting the results in detail, it is worth summarizing the informal comments made by the subjects after the tests. Some subjects reported that watching simultaneously the movements of the lips and of the tongue was not an easy task; a possible compromise was to focus the gaze on the incisors region in order to maintain the tongue on one side of the visual field of view and the lips on the other side. Subjects also reported 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 recognising the consonant. The last set (*i.e.* the generalisation test) was deemed easier than the other sets of test for the same conditions.

4.2. Main test

Fig. 4 represents the mean identification scores, *i.e.* the percentage of consonants correctly identified for the 16 different test sets, separately for the two groups of subjects.

Note first that the results displayed in Fig. 4 for the AU and AVF conditions are coherent with those obtained by Benoît and Le Goff (1998) for different lip/face presentation conditions: for instance they 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%). ANOVA analysis was performed to draw valid conclusions, as follows. The experiment had a 4 × 4 × 8 design with the following within subject factors: presentation condition (four levels: AU, AVJ, AVT, AVF), SNR (four levels: −∞, −9 dB, +3 dB, +∞), and consonant (eight levels: /b d g v z ʒ ʁ l/). One between subject factor (two levels: group

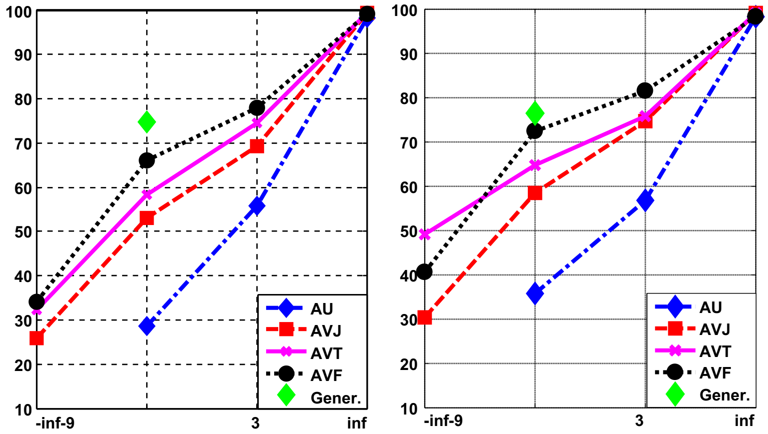

Fig. 4. Mean identification scores (%) as a function of SNR (left: group I; right, group II) for the different conditions: AU, AVJ, AVT, AVF. The isolated diamond indicates the score for the generalisation test.

I, group II) was also involved. Note that, as the combination of the AU condition with the $-\infty$ SNR does not exist (no audio, no video), the scores for this combination were set to 12.5%, in accordance with the hypothesis of a random choice among the eight consonants; this allowed to use statistical tools more extensively. A three-way repeated measures analysis of variance (ANOVA) was conducted on the identification scores with the above within subject and between subject factors.

Significant within subjects main effects were found for presentation condition ($F(3, 63) = 153.04$, $p < .001$), SNR ($F(3, 63) = 1872.37$, $p < .001$), and consonant ($F(7, 147) = 86.88$, $p < .001$).

A significant between subject effect was also found ($F(1, 21) = 10.992$, $p = .003$): the mean score of group II (65.5%) is higher than that of group I (61.5%), which could reflect the fact that group II subjects may have benefited from implicit learning.

Detailed ANOVA analysis of contrasts for the presentation condition showed that the mean identification score is significantly ($p < .001$) greater for all audiovisual conditions (68.0%) than for the AU condition (49.8%), as expected. A more important result is that the mean score was found significantly ($p < .001$) greater for AVT (69.1%) than for AVJ (63.7%): this demonstrates that the objective information brought by the vision of the tongue is indeed perceived by naive subjects. More surprisingly, the mean score are greater for AVF (71.2%) than for AVT, but not significantly ($p = .099$), which implies however that the scores for AVF are significantly higher than those for AVJ. This result was initially not expected, since the articulatory information is the same in both conditions (basically jaw position and lip shape). 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 perceive or decode more efficiently an ecological (naturally looking) rendering than a cutaway presentation. It may also be a consequence of learning of the limited set of stimuli, as the tests with the AVF condition were administrated after those with the AVJ and AVT conditions.

ANOVA analysis of the contrasts for the SNRs showed that the mean identification scores are significantly ($p < .001$) different for all SNRs (29.5%, 54.6%, 70.7% and 98.9%, respectively for $-\infty$, -9 dB, $+3$ dB, $+\infty$), which was expected, as the scores naturally increase with the SNR.

Several significant interactions between factors were found. All interactions with SNR are significant: the effect of the various factors were found stronger at low SNRs than at high ones, which could be expected, since at high SNRs all scores are close to 100%.

In particular, the presentation condition has a significant interaction with SNR ($F(9, 189) = 35.99$, $p < .001$): its influence appears more marked for low SNRs, which is in line with the result well established for lip reading that the contribution of vision increases when the SNR decreases (*cf. e.g.* Benoît et al. (1996)).

A significant interaction was found between presentation condition and consonant ($F(21, 441) = 20.43$, $p < .001$). This interaction can be largely explained by the contrast between the consonants with clear labial cues /b v/ and the others /d g z ʒ ʁ l/.

A significant interaction was found between SNR and consonant ($F(21, 441) = 43.17$, $p < .001$): the score differences between low and high SNRs depend on the consonant, and are smaller for consonants with clear labial cues.

Finally, a significant triple interaction was found between condition, SNR and group ($F(9, 189) = 2.79$, $p = 0.004$). This interaction can be observed in Fig. 4 that shows that the effect of the SNR on the differences between conditions (and especially between AVJ and AVT) depends on the group. This can be considered as an argument in favour of the greater learning of group II.

Average identification scores displayed in Fig. 4 conceal large individual differences. Fig. 5 shows extreme examples for group II subjects. On the one hand, some

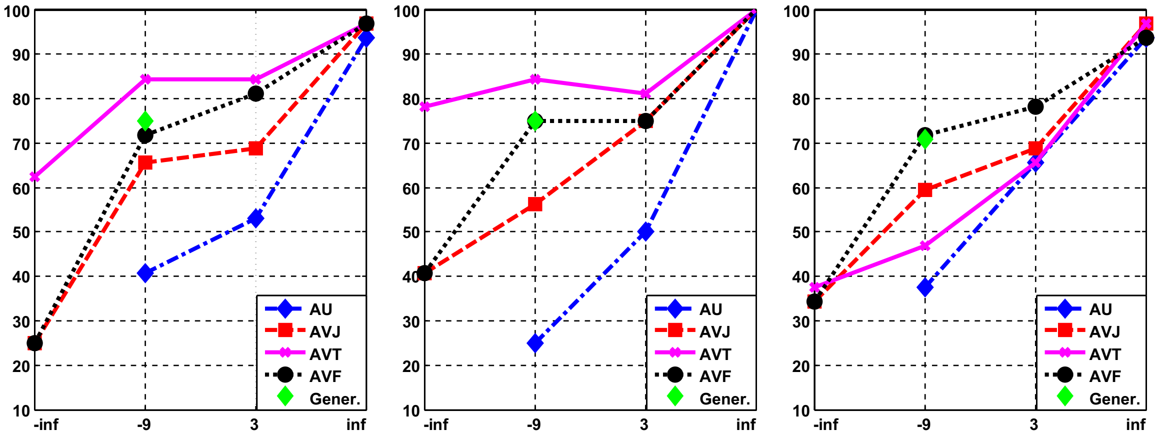

Fig. 5. Identification scores (%) as a function of SNR for individual subjects of group II: good tongue readers (left, centre), poor tongue reader (right) (same conventions as in Fig. 4).

subjects make a very efficient use of tongue vision and gain up to 40% identification score in absence of audio signal. On the other hand, some subjects seem perturbed by the tongue display and gain less than 5% in identification scores in the AVT condition compared to the other audiovisual conditions in absence of audio signal. Note that this difference between *good* tongue readers and *poor* tongue readers shows up essentially in the complete absence of audio.

4.3. Generalisation test

Since the same 32 VCV sequences were used in the first 15 sets of the session, it was needed to verify that part of the implicit learning that may have occured throughout the session was not due to the learning of the stimuli themselves but rather to the learning of tongue reading. The generalisation test thus aimed at verifying that the good scores obtained with the main test still hold with new stimuli never presented before. Though it would have been pertinent to run this test in all SNR and presentation conditions, we restricted it to a unique condition (AVT, SNR $= -9$ dB) in order to maintain the total duration of the test acceptable for the subjects.

As can be seen from Fig. 4, the scores for the generalisation set using a different set of vowels are indeed significantly higher than the corresponding ones for 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 implicitly acquire tongue reading skills during the test session. This interpretation should however be considered with caution. Indeed, Benoît et al. (1994) showed that the vocalic context influences the intelligibility of the adjacent consonants: the improvement of the score may thus also be ascribed to the fact that the more neutral vocalic contexts of the generalisation test would have facilitated the identification.

The conclusion that some implicit learning occurred is also supported by the fact that subjects in group II, who benefit more from implicit learning as they rated the audiovisual stimuli with high 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 had the same tests when starting the generalisation test.

4.4. Discussion

The identification scores of group II are significantly higher that those of group I, principally due to low SNRs (*cf.* the triple interaction mentioned above). This supports the idea that group II benefited from a stronger implicit learning due to the presentation of the audiovisual stimuli with a clear sound before those degraded by noise. All audiovisual conditions yielded speech comprehension rates higher than the simple audio condition. The scores for all SNR levels rank, for each group, with statistically significant differences (except between AVF and AVT), in the following decreasing order: AVF, AVT, AVJ, AU. AVF was significantly better decoded than AVJ, which would mean that subjects perceive better an ecological rendering than 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 group II, who benefited from a stronger implicit learning: in this case, the AVT score is higher by 10% than the AVF score and even 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 of uni-sensory tongue reading is possible, though the integration of tongue reading in multisensory perception is not easy.*

5. Conclusions and perspectives

The main objective of the present study was to assess whether the direct and full vision of the tongue allows tongue reading. Audiovisual VCV stimuli obtained by controlling a virtual talking head from articulatory movements tracked on one speaker, and thus having a natural articulatory dynamics, were used in audiovisual perception tests. The results indicate: (1) the possibility of implicit learning of tongue reading, (2) better consonant identification with the cutaway presentation with the tongue than without the tongue, (3) no significant difference between the cutaway presentation with the tongue and the more ecological rendering of the complete face, (4) a predominance of lip reading over tongue reading, but (5) a certain natural human capability for tongue reading when the audio signal is strongly degraded or absent.

These results are in line with those obtained in a study with very similar goals but a quite different technical approach performed by Wik and Engwall (2008): articulatory synchronisation from audio *vs.* control from EMA recordings for the articulatory control of the talking head, 60 short Swedish sentences *vs.* 32 French VCV sequences for the corpus, repetitions allowed *vs.* no repetitions allowed for the presentation, retrieving words in a sentence *vs.* identifying the consonant by a forced choice for the task, spectral details reduction by means of a channel vocoder *vs.* white noise addition for the audio degradation. Their conclusions are quite close to ours: “the augmented reality side view did not help subjects perform better overall than with the front view only”. They quote a high intersubject variability; some of their subjects did clearly benefit from the tongue view, with up to 30% better word recognition, while others were rather perturbed (some of our subjects got 40% better consonant identification). They also hypothese that the additional repetition may allow users to take information from both face views into account, which is coherent with the informal comments of our French subjects about the difficulty to follow simultaneously lips and tongue. Their subjects appear to be able to learn extracting some information about phonemes from the intra-oral articulation.

Note finally that a few other studies (Grauwinkel et al. (2007) or Massaro et al. (2008)) obtain converging conclusions on the possible benefit of displaying internal articulators movements, but with the necessity to train the subjects to this type of view.

These studies need to be complemented by more systematic tests, involving in particular measures of visual attention by means of eye-tracking systems, in order to confirm that the human natural abilities for tongue reading are weak, or simply dominated by those for lip reading. 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.

Besides, our aims in the future are 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, and (3) pronunciation training for second language learners.

Acknowledgments

We sincerely thank two anonymous reviewers for their pertinent comments and suggestions. We are very much indebt to Marion Dohen for her careful editorial work and for her significant help with the statistical analyses. We also 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.

References

Badin, P., Serrurier, A., 2006. Three-dimensional linear modeling of tongue: Articulatory data and models. In: Yehia, H.C., Demolin, D., Laboissière, R. (Eds.), Seventh International Seminar on Speech Production, ISSP7. Ubatuba, SP, Brazil, UFMG, Belo Horizonte, Brazil, pp. 395–402.

Badin, P., Tarabalka, Y., Elisei, F., Bailly, G., 2008. Can you “read tongue movements”? In: Interspeech 2008. Brisbane, Australia, pp. 2635–2638.

Bälter, O., Engwall, O., Öster, A.-M., Kjellström, H., 2005. Wizard-of-Oz test of ARTUR – a computer-based speech training system with articulation correction. In: Seventh International ACM SIGACCESS Conference on Computers and Accessibility. Baltimore, pp. 36–43.

Benoît, C., Guiard-Marigny, T., Le Goff, B., Adjoudani, A., 1996. Which components of the face do humans and machines best speechread? In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines. Springer-Verlag, Berlin, pp. 315–328.

Benoît, C., Le Goff, B., 1998. Audio–visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP. Speech Communication 26, 117–129.

Benoît, C., Mohamadi, T., Kandel, S., 1994. Effects of phonetic context on audio–visual intelligibility of French. Journal of Speech and Hearing Research 37, 1195–1203.

Cathiard, M.-A., Lallouache, M.T., Abry, C., 1996. Does movement on the lips mean movement in the mind ? In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines. Springer Verlag, Berlin, pp. 211–219.

Cornett, O., 1967. Cued speech. American Annals of the Deaf 112, 3–13.

Engwall, O., 2008. Can audio–visual instructions help learners improve their articulation? An ultrasound study of short term changes. In: Interspeech 2008. Brisbane, Australia, pp. 2631–2634.

Erber, N.P., 1975. Auditory-visual perception of speech. Journal of Speech and Hearing Disorders XL, 481–492.

Fagel, S., Madany, K., 2008. A 3-D virtual head as a tool for speech therapy for children. In: Interspeech 2008. Brisbane, Australia, pp. 2643–2646.

Grauwinkel, K., Dewitt, B., Fagel, S., 2007. Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech. In: Interspeech'2007 – Eurospeech – 9th European Conference on Speech Communication and Technology. Antwerp, Belgium, pp. 706–709.

Hoole, P., Nguyen, N., 1997. Electromagnetic articulography in coarticulation research. Forschungsberichte des Instituts für Phonetik und Spachliche Kommunikation der Universität München, FIPKM, vol. 35, pp. 177–184.

IJsseldijk, F.J., 1992. Speechreading performance under different conditions of video image, repetition, and speech rate. Journal of Speech and Hearing Research 35, 466–471.

Kröger, B.J., Graf-Borttscheller, V., Lowit, A., 2008. Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In: Interspeech 2008. Brisbane, Australia, pp. 2639–2642.

Massaro, D.W., Bigler, S., Chen, T., Perlman, M., Ouni, S., 2008. Pronunciation training: the role of eye and ear. In: Interspeech 2008. Brisbane, Australia, pp. 2623–2626.

Massaro, D.W., Light, J., 2003. Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In: Eurospeech 2003. Geneva, Switzerland, pp. 2249–2252.

Massaro, D.W., Light, J., 2004. Using visible speech to train perception and production of speech for individuals with hearing loss. Journal of Speech, Language, and Hearing Research 47, 304–320.

Mills, A.E., 1987. The development of phonology in the blind child. In: Dodd, B., Campbell, R. (Eds.), Hearing by eye: the psychology of lipreading. Lawrence Erlbaum Associates, London, pp. 145–161.

Montgomery, D., 1981. Do dyslexics have difficulty accessing articulatory information? Psychological Research 43 (2).

Mulford, R., 1988. First words of the blind child. In: Smith, M.D., Locke, J.L. (Eds.), The Emergent Lexicon: The Child's Development of a Linguistic Vocabulary. Academic Press, New-York, pp. 293–338.

Narayanan, S.S., Nayak, K., Lee, S., Sethy, A., Byrd, D., 2004. An approach to real-time magnetic resonance imaging for speech production. Journal of the Acoustical Society of America 115 (4), 1771–1776.

Odisio, M., Bailly, G., Elisei, F., 2004. Tracking talking faces with shape and appearance models. Speech Communication (Special Issue on Audio Visual Speech Processing) 44 (1–4), 63–82.

Perkell, J.S., Cohen, M.M., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.T., 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. Journal of the Acoustical Society of America 92, 3078–3096.

Serrurier, A., Badin, P., 2008. A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. Journal of the Acoustical Society of America 123 (4), 2335–2355.

Stoel-Gammon, C., 1988. Prelinguistic vocalizations of hearing-impaired and normally hearing subjects. A comparison of consonantal inventories. Journal of Speech and Hearing Disorders 53, 302–315.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26 (2), 212–215.

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

Vihman, M.M., Macken, M.A., Miller, R., Simmons, H., Miller, J., 1985. From babbling to speech: a re-assessment of the continuity issue. Language 61 (2), 397–445.

Wik, P., Engwall, O., 2008. Can visualization of internal articulators support speech perception? In: Interspeech 2008. Brisbane, Australia, pp. 2627–2630.