Visual articulatory feedback for phonetic correction in second language learning

Pierre Badin, Atef Ben Youssef, Gérard Bailly, Frédéric Elisei, Thomas Hueber

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Grenoble University
961 rue de la Houille Blanche, BP 46, F-38402 Saint Martin d'Hères cedex, France
{Pierre.Badin, Atef.BenYoussef, Gerard.Bailly}@gipsa-lab.grenoble-inp.fr

Abstract

Orofacial clones can display speech articulation in an *augmented* mode, *i.e.* display all major speech articulators, including those usually hidden such as the tongue or the velum. Besides, a number of studies tend to show that the visual articulatory feedback provided by ElectroPalatoGraphy or ultrasound echography is useful for speech therapy. This paper describes the latest developments in acoustic-to-articulatory inversion, based on statistical models, to drive orofacial clones from speech sound. It suggests that this technology could provide a more elaborate feedback than previously available, and that it would be useful in the domain of Computer Aided Pronunciation Training.

Index Terms: Visual articulatory feedback, orofacial clone, acoustic-to-articulatory inversion, speaker normalisation, CAPT, ElectroMagnetic Articulography, language learning, speech therapy, Hidden Markov Models.

1. Introduction

It has become common sense to say that speech is not merely an acoustic signal but a signal endowed with complementary coherent traces such as visual, tactile or physiological signals [1]. Besides, it has been demonstrated that humans possess – to some degree – articulatory awareness skills, as measured *e.g.* by Montgomery [2]. Thomas & Sénéchal [3] results support the hypothesis that accuracy of articulation is related to quality of phoneme awareness in young children, while Kröger *et al.* [4] found that children older than five years are capable of interpreting vocal tract articulatory speech sound movements without any preparatory training in a speech adequate way. Finally, we have recently demonstrated that human subjects are able – to some extent – to make use of tongue shape vision for phonemic recognition, as they do with lips in lip reading [5]. All these findings suggest that visual articulatory feedback could help subjects to acquire the articulatory strategies needed to produce sounds that are new to them. The present paper thus describes the tools and methods currently being developed in our laboratory to offer such potentialities. We present the state-of-the-art in the domain of visual feedback for phonetic correction, we describe our orofacial clone and the inversion system that can control it to provide visual articulatory feedback, and conclude with some perspectives.

2. Visual feedback for phonetic correction

Interestingly, phonetic correction is involved in two domains, though with different specificities, *i.e.* second language learning and speech rehabilitation. In both domains, it has been attempted to provide learners / patients with various forms of signals that bear information on their spoken productions.

2.1. Speech Therapy

According to Bernhardt *et al.* [6] [7], "research has shown that visual feedback technologies can be effective tools for speech (re)habilitation, whether the feedback is acoustic or articulatory**"**. Acoustic information can be captured by a microphone and displayed as waveforms, intensity or fundamental frequency time trajectories, or still spectrograms [8] [9]. More elaborate devices can provide real time articulatory information: ElectroPalatoGraphy (EPG) [10] indicates the presence / absence of tongue-palate contacts in about 60-90 locations on the speaker's hard palate, while ultrasound echography [7] generates images of the tongue – in most cases in the midsagittal plane.

During clinic-based sessions conducted by Wrench *et al.* [10] the patient could use the visual feedback of tongue-palate contact patterns provided by EPG to establish velar and alveolar placement for different phonetic targets. Besides, these targets could be demonstrated by the speech therapist when also wearing an EPG-palate. They concluded that EPG is a potentially useful tool for treating articulation disorders as well as for recording and assessing progress during the treatment.

In the tradition of awareness and self-monitoring training approaches to phonological intervention, Bernhardt *et al.* [6] use the ultrasound machine to freeze specific images on the screen in order to allow patients to discuss and compare their own productions with target productions proposed by the speech therapists. They note that "the ultrasound images provide the patient with more information about tongue shapes and movements than can be gained with other types of feedback (the mirror, acoustic analysis, touch, EPG)." They remark also that, as auditory self-monitoring can be challenging for patients with hearing impairment, visual displays help them make judgments on their own productions.

Globally, most studies seem **"**to support the perspective that articulatory visual feedback facilitates speech habilitation for hearing impaired speakers across a variety of sound classes by providing information about tongue contact, movement, and shape" [11].

2.2. Language learning

Oppositely to speech therapy, most of the literature in Computer Aided Pronunciation Training (CAPT) seems to deal visual feedback that does not involve explicit articulatory information (for a recent survey on spoken language technology for education, cf. [12]). Menzel *et al.* [13] mention that "usually, a simple playback facility along with a global scoring mechanism and a visual presentation of the signal form or some derived parameters like pitch are provided." But they pinpoint that a crucial task is left to the student, *i.e.* identifying the place and the nature of the pronunciation problem. According to them, automatic speech recognition (ASR) is often used to localise the errors, and even to perform

an analysis in terms of phone substitutions, insertions or omissions, as well as in terms of misplaced word stress patterns. But they note that "if the feedback is provided to the student through a multimedia-based interface, all the interaction is carried out using only the orthographic representations". Though more and more precise and flexible ASR systems have allowed progress in CAPT [14], [15], it may be interesting to explore the potentialities of visual articulatory feedback.

A limited but interesting series of studies has used Virtual Talking Heads (VTH) controlled by text-to-speech synthesis to display speech articulators – including usually hidden ones such as the tongue. These displays are meant to demonstrate targets for helping learners acquiring new or correct articulations, though they actually do not provide a real feedback of the learner's articulators as in speech therapy.

Massaro & Light [16] found that using a VTH as a language tutor for children with *hearing loss* led to some quantitative improvement in their performances. Later, using the same talking head, Massaro *et al.* [17] showed that visible speech could contribute positively to acquiring new speech distinctions and promoting active learning, though they could not conclude about the effectiveness of showing *internal* articulatory movements for pronunciation training.

Engwall [18] implemented an indirect visual articulatory feedback by means of a wizard-of-Oz set-up, in which an expert phonetician chose the adequate pre-generated feedback with a VTH meant to guide the learner to produce the right articulation. He found that this helped French subjects improve their pronunciation of Swedish words, though he did not perform any specific evaluation of the benefit of the vision of the tongue.

Other studies investigated the visual information conveyed by the vision of internal articulators. Grauwinkel *et al.* [19] reported 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.* [4] asked children older than 5 years to mimic the mute speech movements displayed by a VTH for different speech sounds, and found that were capable of interpreting vocal tract articulatory speech sound movements without any preparatory training in a speech adequate way.

Badin *et al.* [5] have recently shown that naive untrained subjects can make use of the direct and full vision of the tongue provided by a VTH to improve their consonant identification in audiovisual VCVs played with a low Signal-to-Noise Ratio or no speech sound at all. They noticed that *tongue reading* was implicitly learned during the audiovisual perception tests, suggesting that, as *lip reading*, it could be trained and useful in various speech training domains.

Note finally the only experiment that we are aware of in speech therapy, where Fagel *et al.* [20] attempted to correct lisping for a few children, and found that using a VTH to demonstrate the correct (prototypic) pronunciation of the /s z/ sounds did significantly enhance their speech production.

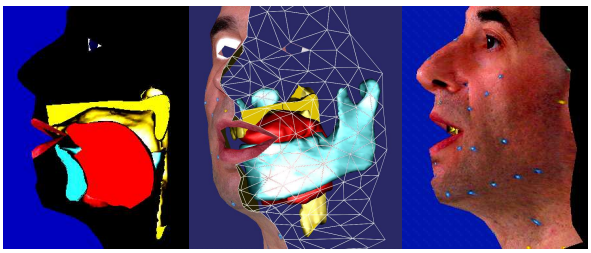

Figure 1: *Examples of OFC display. The face, the jaw, the tongue and the vocal tract walls including the hard and soft palates can be distinguished when the skin is not represented.*

We can conclude from this short survey that: (1) the direct vision of tongue by means of a VTH can be used, even by naive subjects, and can be trained, (2) visual articulatory feedback is effective in speech (re)habilitation, and (3) visual articulatory feedback has never been experimented with in the domain of CAPT. The next sections of this paper describe the tools developed in our laboratory that may allow to implement and test a visual articulatory feedback for CAPT, that is considerably more elaborate than EPG or echography, as it provides the complete set of articulators.

3. Orofacial clones and augmented speech

Except for echography, which is however restricted to a limited part of the tongue, there are at present no real time medical imaging systems capable of displaying the whole set of articulators in animation with a reasonable time and frequency resolution. The modelling approach offers an interesting alternative: 3D fine grained articulators model can be built from static volume data such as MRI or CT, and can be controlled through motion capture devices such as ElectroMagneticArticulography (EMA) that provides only a few articulators points, but at a good sampling frequency.

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 an OroFacial Clone (OFC) based on extensive measurements on the same speaker.

3.1. The OroFacial Clone

The OFC currently developed in our department is the assemblage of individual three-dimensional models of various speech organs of the same speaker (*cf.* [21] [5] for a detailed description). These models are built from Magnetic Resonance Imaging (MRI), Computer Tomography (CT) and video data acquired from this speaker. The jaw, lips and face model is controlled by two jaw parameters (*jaw height, jaw advance*), and three lip parameters (*lip protrusion, upper* and *lower lip heights*). The velum model is essentially controlled by one parameter that drives the opening / closing movements of the nasopharyngeal port. The jaw and tongue model is primarily driven by five parameters: the main effect of the *jaw height* parameter 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 other two parameters, *tongue tip vertical* and *tongue tip horizontal* control precisely the shape of the tongue tip. Interestingly, it was found that these components correspond roughly to muscle synergies used in speech production.

Figure 1, which shows possible displays of this OFC, illustrates the *augmented speech* capabilities offered by the vision of the internal articulators.

Figure 2 exemplifies in more detail the behaviour on the tongue model by demonstrating the *tongue dorsum* component effects, in particular tongue grooving and tongue bunching.

3.2. Animation of the OFC from EMA

The first method that we have implemented to produce animations of our OFC is based on the concept of *motion*

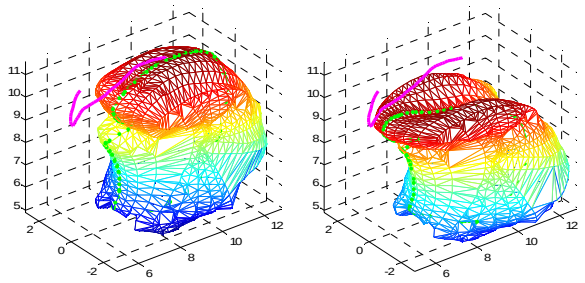

Figure 2: *Illustration of the* tongue body *component of the 3D tongue model. Note the bunching (left) and the grooving (right).*

capture used in the film animation domain [22], where the movement of a small number of markers attached to specific locations of articulators are monitored and acquired. In our case, due to the difficulty of accessing internal articulators such as the tongue or the velum, we use ElectroMagnetic Articulography (EMA). EMA is an experimental method that infers the coordinates of small electromagnetic coils from the magnetic fields that they receive from electromagnetic transmitters [23]. We record the midsagittal coordinates of a set of six coils as indicated in Figure 3 (left): a *jaw coil* is attached to the lower incisors, while a *tip coil*, a *mid coil* and a *back coil* are attached to the tongue, as illustrated in Figure 3 (right); 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 incisors and to the nose serve as references. After appropriate scaling and alignment, the coordinates of the coils are obtained in the same coordinate system as the models. As demonstrated in [5], this information is sufficient to *inverse* the articulatory models of the OFC, *i.e.* to *recover* the control parameters that give the best fit in the midsagittal plane between the modelled 3D surfaces and the measured coils coordinates.

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, as illustrated in www.gipsa-lab.fr/~pierre.badin/SpeechCommTongueReading/pb_phrm6.avi and www.gipsa-.../pb_tongue_phrm6_side.avi

4. HMM-based acoustic-to-articulatory inversion

Though motion capture with EMA provides very valuable articulatory data and realistic animations, it is rather invasive, as illustrated in Figure 3 (right). This limits its use to laboratory experiments, and prevents any practical visual articulatory feedback applications. This is why we have developed an alternative approach based on acoustic-to-articulatory inversion that finally allows driving our OFC. We have improved a method based on Hidden Markov Models (HMMs) developed in a previous study [24].

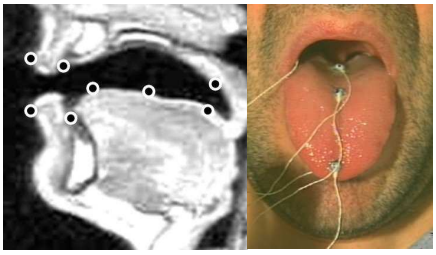

Figure 3: *Illustration of the possible EMA coils locations (left) and of the speakers' tongue.*

4.1. Acoustic and articulatory data

This statistical method is based on joint acoustic-articulatory data. Using the EMA setup described above, we have recorded a French speaker producing a variety of Vowel-Consonant-Vowel sequences, short CVC words, and longer sentences, amounting altogether to about 17 minutes of speech corresponding to about 5000 allophones. As the HMM-based method implies explicit phone models, we have first labelled the phones based on the audio signal and a forced alignment procedure. The 36 phonemes include the 14 oral and nasal French vowels, 16 consonants, 3 semivowels, the schwa and short and long pauses.

4.2. Inversion based on HMMs

The acoustic-to-articulatory inversion based on HMMs involves two stages: a stage of phonemic recognition from the acoustics, and a stage of synthesis of the articulatory trajectories.

In order to take coarticulation into account, HMMs are used to model allophones, *i.e.* phones in context. Therefore, we used the Mahalanobis distance in the articulatory space to generate dendrograms and determine six coherent vocalic and eleven consonantal allophonic context classes. Multistream three-states HMMs of allophones are jointly trained from the synchronous streams of articulatory data acquired by EMA and speech spectral parameters; the acoustic part of the feature vectors consist of 12 Mel-Frequency Cepstral Coefficients and of the logarithm of the energy, along with their first time derivatives, while the articulatory part is made of the coordinates of the six active coils and their first time derivatives. The phoneme recognition system is fed with the speech signal and uses the acoustic part of the HMMs to deliver the chain of the recognised phoneme and the associated state durations (see more details in [24]). This information is then used by a trajectory formation procedure based on the articulatory part of the HMMs to resynthesise the articulatory data using the HTS software [25].

4.3. Evaluation

To test the validity of our inversion method, we computed the Root Mean Square Error (RMSE) between the measured and recovered coil coordinates over the whole corpus, using a jackknife method with five partitions to avoid the bias of training and testing on the same speech material. We obtained an RMSE of 1.7 mm, which is identical to the reference score of [26].

5. Speaker normalisation

The inversion algorithm described in the previous sections can determine with a fair approximation, from a speech input, the midsagittal coordinates of six markers attached to the main articulators. We have seen in Section 3.2 that it is subsequently possible to recover the control parameters of the OFC and thus to animate it. This proves the feasibility of a visual articulatory feedback system that provides a nearly complete view of the articulators, thus extending considerably the EPG- and ultrasound echography-based systems. We recall that all the data and models involved in the present system are based on the same French speaker. In order to be useful in real speech therapy or CAPT, such a system should be able to cope with any new speaker, being it a teacher or a learner. We are thus faced to a speaker normalisation or adaptation problem: we need to be able to adapt the present OFC to the morphology and articulatory strategies of other speakers on

the one hand, and to convert the coil coordinates of one speaker into the coil coordinates of another speaker on the other hand. The first problem has been recently studied by Ananthakrishnan *et al.* [27], who have assessed various three-mode factor analysis techniques to model the variations of midsagittal vocal tract contours obtained from MRI images for three French speakers articulating 73 vowels and consonants. We will explore the possibility of adapting speaker adaptation techniques used in Automatic Speech Recognition or in Voice Conversion to solve the second problem, either at the signal level or at the articulatory space level. We are also aware that deviant articulation of students may affect the possibility to rely on the method based on one fairly consistent speaker, and will therefore explore the possibility to include a number of well documented erroneous articulations in the training corpus of this reference speaker.

6. Conclusions and perspectives

We have described the components of a system based on speech sound to articulation inversion that can provide visual articulatory feedback for one specific speaker. The next challenge is to extend this study to a larger collection of other speakers, and also to develop (near)-real time algorithms.

The possibilities of the system will be demonstrated at the conference, with the hope that this will receive a positive feedback from the CAPT community, along with the expression of specific needs.

7. Acknowledgements

We thank Christophe Savariaux for the EMA recordings. This work has been partially supported by the French ANR-08-EMER-001-02 grant in the framework of the *ARTIS* project "Articulatory inversion from audio-visual speech for augmented speech presentation".

8. References

[1] Bailly, G., Badin, P., Beautemps, D., and Elisei, F., "Speech technologies for augmented communication," in Computer Synthesized Speech Technologies: Tools for Aiding Impairment, Mullennix, J. and Stern, S., Eds.: IGI Global, Medical Information Science Reference, 2010, pp. 116-128.

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

[3] Thomas, E.M. and Sénéchal, M., "Articulation and phoneme awareness of 3-year-old children," Applied Psycholinguistics, vol. 19, pp. 363-391, 1998.

[4] Kröger, B.J., Graf-Borttscheller, V., and Lowit, A., "Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders," presented at Interspeech 2008, Brisbane, Australia, 2008.

[5] Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G., "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," Speech Communication, vol. 52, pp. 493-503, 2010.

[6] Bernhardt, B.M., Gick, B., Bacsfalvi, P., and Adler-Bock, M., "Ultrasound in speech therapy with adolescents and adults," Clinical Linguistics & Phonetics, vol. 19, pp. 605-617, 2005.

[7] Bernhardt, B.M., Bacsfalvi, P., Adler-Bock, M., Shimizu, R., Cheney, A., Giesbrecht, N., O'connell, M., Sirianni, J., and Radanov, B., "Ultrasound as visual feedback in speech habilitation: Exploring consultative use in rural British Columbia, Canada," Clinical Linguistics & Phonetics, vol. 22, pp. 149-162, 2008.

[8] Neri, A., Cucchiarini, C., and Strik, H., "Feedback in computer assisted pronunciation training: technology push or demand pull?," presented at ICSLP-2002, Denver, Co, USA, 2002.

[9] Menin-Sicard, A. and Sicard, E., "Evaluation et rééducation de la voix et de la parole avec Vocalab," Glossa, vol. 88, pp. 62-76, 2006.

[10] Wrench, A., Gibbon, F., McNeill, A.M., and Wood, S., "An EPG therapy protocol for remediation and assessment of articulation disorders," presented at ICSLP-2002, 2002.

[11] Bernhardt, B.M., Gick, B., Bacsfalvi, P., and Ashdown, J., "Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners," Clinical Linguistics & Phonetics, vol. 17, pp. 199-216, 2003.

[12] Eskenazi, M., "An overview of spoken language technology for education," Speech Communication, vol. 51, pp. 832-844, 2009.

[13] Menzel, W., Herron, D., Morton, R., Pezzotta, D., Bonaventura, P., and Howarth, P., "Interactive pronunciation training," ReCALL, vol. 13, pp. 67-78, 2001.

[14] Chun, D.M., "Come ride the wave: But where is it taking us?," Calico Journal, vol. 24, pp. 239-252, 2007.

[15] Cucchiarini, C., Neri, A., and Strik, H., "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," Speech Communication, vol. 51, pp. 853-863, 2009.

[16] 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.

[17] Massaro, D.W., Bigler, S., Chen, T., Perlman, M., and Ouni, S., "Pronunciation training: the role of eye and ear," presented at Interspeech 2008, Brisbane, Australia, 2008.

[18] Engwall, O., "Can audio-visual instructions help learners improve their articulation? — An ultrasound study of short term changes," presented at Interspeech 2008, Brisbane, Australia, 2008.

[19] 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, Antwerp, Belgium, 2007.

[20] Fagel, S. and Madany, K., "A 3-D virtual head as a tool for speech therapy for children," presented at Interspeech 2008, Brisbane, Australia, 2008.

[21] Badin, P., Elisei, F., Bailly, G., and Tarabalka, Y., "An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data," in Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098), Perales, F.J. and Fisher, R.B., Eds. Berlin, Heidelberg, Germany: Springer Verlag, 2008, pp. 132–143.

[22] Joon, J.S., "A preliminary study of human motion based on actor physiques using motion capture," presented at Sixth International Conference on Computer Graphics, Imaging and Visualization, 2009. CGIV '09, Tianjin, 2009.

[23] Perkell, J.S., Cohen, M.M., Svirsky, M.A., Matthies, M.L., Garabieta, I., and Jackson, M.T.T., "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," Journal of the Acoustical Society of America, vol. 92, pp. 3078-3096, 1992.

[24] Ben Youssef, A., Badin, P., Bailly, G., and Heracleous, P., "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models," presented at Interspeech 2009, Brighton, UK, 2009.

[25] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[26] Hiroya, S. and Honda, M., "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE Trans. Speech and Audio Processing, vol. 12, pp. 175-185, 2004.

[27] Ananthakrishnan, G., Badin, P., Valdés Vargas, J.A., and Engwall, O., "Predicting unseen articulations from multi-speaker articulatory models," presented at Interspeech 2010, Makuhari, Japan, 2010.