

Towards the use of a *Virtual Talking Head* and of *Speech Mapping* tools for pronunciation training

Pierre Badin, Gérard Bailly, and Louis-Jean Boë

Institut de la Communication Parlée – UPRESA CNRS 5009 – INPG / Univ. Stendhal, Grenoble, France
badin@icp.inpg.fr, bailly@icp.inpg.fr, boe@icp.inpg.fr

Abstract

The *Speech Mapping* concept posits that speech sequences can be represented by trajectories in a multi-parametric space, whose elements are related to each other by relationships representing speech production mechanisms. This concept is viewed as a useful framework for pronunciation training, in a scheme where the teacher uses a *Virtual Talking Head*, to manipulate audio-visual speech stimuli in order to fulfil a double task: (1) evaluating and improving the learner's perception of the target language sounds, and (2) helping the learner produce the corresponding articulations by acquiring the internalisation of the relations between articulatory gestures and resulting sounds. We describe a set of data and models, including a virtual talking head, that can be useful for pronunciation training, present a few experiments supporting this approach, and suggest some directions for the future.

1. Introduction

A second language learner can be considered *phonologically deaf*, i.e. he/she may not be able to *hear*, or discriminate, speech sounds that do not belong to his/her own phonological inventory, or are too far away from it. Consequently, one can well imagine that if he/she can not discriminate these sounds, he/she can not produce them either. This situation is well known for deaf children having language development delays related to their impossibility to perceive speech sounds auditorily. These considerations support the necessity to take into account links between acoustics and speech perception in the process of language teaching.

On the other hand, supposing that the learner has acquired a good perception of the new sounds, he/she must shape his/her vocal tract and dynamically coordinate articulators to produce these specific acoustic targets by means of manoeuvres that may be new to him/her. One may postulate that the speech production act involves an explicit knowledge of the learner's own speech apparatus functioning (cf. e.g. [14]). This idea constitutes the basis of the *articulatory method* – or *cognitive approach* – to pronunciation training. It appears thus important for the second language teacher to be able to demonstrate – and for the learner to be able to explore and acquire – the *internalisation of the relations between articulatory gestures and resulting sounds*. This implies the necessity for the teacher of having a good working knowledge of phonetics and of the articulatory-acoustic relations ([14], pp. 11-12).

These considerations constitute one of the starting points of LeBel's [10] "Traité de correction phonétique

punctuelle [Treatise of punctual phonetic correction]", where he refers to "grands moyens [big means]" in the domain of phonetic correction. Three of them are directly related to perception and production: (1) *auditory discrimination* (one can pronounce well only what one can perceive well), (2) *articulatory and acoustic composition* (the learning process will be more efficient if the learner knows which articulator he/she should pay attention to in order to correct a specific problem), and (3) *combinatory phonetics* (various coarticulation effects can be used to induce the right articulatory gestures for a given phoneme).

2. *Speech Mapping*: a conceptual framework for language pronunciation training

The *Speech Mapping* concept posits that speech sequences can be represented by trajectories in a multi-parametric space, called *sensory-motor maps*, constituted of different spaces related to each other by relationships representing speech production mechanisms ([1], [5]). At present, we consider three spaces: the space of articulators positions, the geometric space (i.e. the space where constrictions / occlusions are specified), and the acoustic/auditory space (in practice the F1/F2 space). The relations between these spaces, which are not necessarily bijective, are implemented in a *virtual talking head*, which is an *anthropomorphic model of speech production*.

Considering the different aspects mentioned in the introduction, it seems thus that *Speech Mapping* would be a useful conceptual framework for pronunciation training. Figure 1 illustrates our conception of the situation. The sensory-motor maps play a central role: they constitute a kind of *common blackboard* where the learner, the teacher and the talking head can code, display or watch the multi-dimensional trajectories associated with speech sequences, and thus communicate. In order to simplify the relations between the three protagonists, we assume that normalisation problems are solved, i.e. that the talking head can be scaled so as to be able to reproduce any sound emitted by the learner, problem far from being solved at present.

At the segmental level, the learner's mother language background can be simplified and conceptualised as phonological maps, following the UPSID phonological descriptions of the world languages [15], and defined in terms of authorised regions of realisation (or dispersion ellipses around prototypes) in the sensory-motor maps [5]. Vowels are described in the F1/F2 space, while consonants are defined in the geometric space. The target language, i.e. the language to be learned, can be described in similar terms. The mission of the teacher is

then to help the learner acquire or develop the sensory-motor maps corresponding to speech sequences in the target language. The teacher's main tasks are thus: (1) evaluating and improving the learner's ability to perceive the target language vowels and consonants (through stable vowels, and then simple VC[VCV...] sequences), and (2) elaborating teaching strategies to help the learner exploring and finding the right articulatory gestures to produce what he/she has learned to perceive.

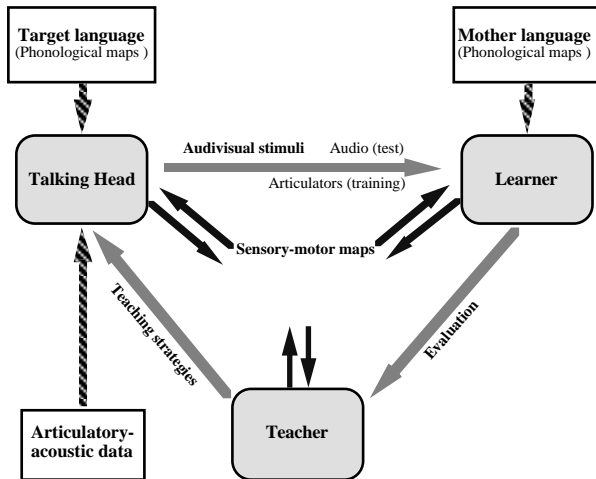


Figure 1. Interactions between second language learner, teacher and the talking head.

3. The Speech Mapping tools

The *Speech Mapping* tools consist of a coherent set of data and models that formalise or represent relationships between the different spaces of representation of speech. We briefly describe them in the present section.

3.1 Articulatory-acoustic data

One of our approaches to speech production studies consists in accumulating into a single coherent framework complementary data obtained from various experimental setups for a few reference subjects uttering the same speech material in controlled conditions. This aims at providing as complete as possible a picture of the different mechanisms involved at different levels in the speech production chain and at constructing a comprehensive model. The most important methods used are: *cineradiography* that produces limited but extremely valuable sets of midsagittal vocal tract contours [2], *pneumotachometry* that provides air flow at the lips and intraoral pressure, *video labiometry* that furnishes a geometric description of lips from front and profile views [2], *electromagnetic articulometry* that delivers the X/Y coordinates in the midsagittal plane of a few points attached the tongue or to the jaw [3], and *Magnetic Resonance Imaging* that results in full 3D geometric descriptions of sustained articulations [4].

3.2 The Virtual Talking Head

The *talking head* under development at ICP is a *virtual anthropomorphic robot* based on physical modelling of the articulatory, aerodynamic and acoustic phenomena involved in the audio-visual production of speech. This talking head has been developed from the data described

in section 3.1. The core of the speech production model is a linear midsagittal physiologically oriented articulatory model, built by “guided principal component analysis” [6]. The jaw is assumed to have one degree of freedom only, i.e. its height. Tongue shape is defined by five parameters, in addition to jaw height, while lips are represented by a single tube of controlled length and height. The midsagittal function produced by the model is converted into area function by means of a conversion method, and then sounds produced by time-domain reflection-type line analogue or formants computed by a frequency-domain model [6].

An extension of this midsagittal model to a 3D description of vocal tract shape has been recently elaborated [4], following the same methodology used for the 2D model. The 3D geometry of the vocal tract is described as closed planar contours determined as the intersections of the vocal tract 3D contour with planes orthogonal to the midsagittal plane (cf. Figure 2). In a first approximation, this model can be controlled by the same parameters as the 2D model. In addition, a 3D model of the lips has been worked out from geometrical analysis of the natural lips of a reference subject [9], and can also be driven by two parameters, i.e. lip protrusion and lip height. The lip display appears in Figure 3.

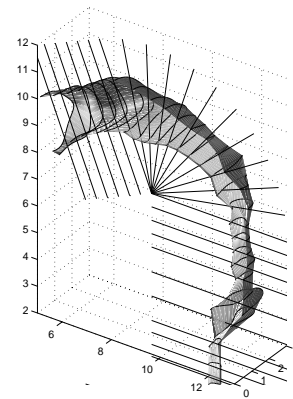


Figure 2: Example of 3D vocal tract shape with the semipolar grid.

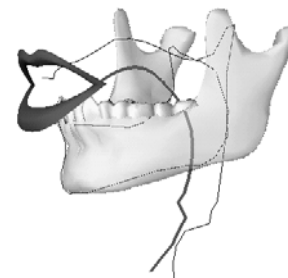


Figure 3. Display of 2D tongue, and 3D lips and jaw.

Finally, the talking head and its articulators can be made visible by means of various displays. Tongue, lips and jaw can be visualised in 2D and 3D. The face itself can be completely absent, displayed as a semitransparent skin, presented as a bony skull or as a complete face. Figure 3 presents an example of 3D representation of the jaw and lips, along with a 2D representation of midsagittal vocal tract contour. More realistic and complete displays are planned for the future.

3.3 Controlling the Talking Head

The control parameters of the talking head can be derived in various ways.

Lips and jaw can be animated through analysis / synthesis [11]. The nine articulatory parameters can be derived by inversion from formants and possibly lip area [12], for the reference subject. Parameters controlling jaw and tongue shape can also be determined from the X/Y coordinates of three coils of an electromagnetic articulometer attached to the tongue, in addition to the jaw height measured from a coil attached to the lower incisors [3]; this enables the acquisition of midsagittal vocal tract contours in much larger amounts than it is possible from cineradiography, or even direct monitoring from the subject.

Finally, a French text-to-audio-visual-speech synthesiser which coarticulation rules were automatically extracted from an extensive analysis of the same reference subject's data is also available [11]. This system produces a complete face display based on articulators positions (jaw, lips, tongue tip / blade).

3.4 The SMIP

A version of the 2D model, the *SMIP (Speech Maps Interactive Plant*, [7]), allows to interactively vary the articulatory control parameters, visualise the effects on the midsagittal contours and on the formant frequencies in the F1/F2 and F2/F3 planes (cf. Figure 4), and to listen to sustained voiced sounds generated from these formants. A older version of this tool has been used for years for teaching phonetics, and is expected to be useful in pronunciation training.

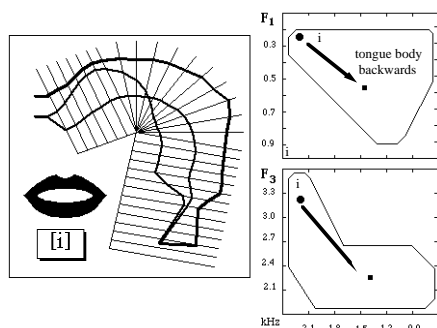


Figure 4. Example of SMIP displays showing the effect of tongue body backing on formants.

3.5 Auditory sensory-motor maps

Another aspect of *Speech Mapping* lies in the listener's perception of the acoustic space, in relation with the underlying articulatory space. In an experiment conducted to study this aspect [8], an expert phonetician, capable to recognise and produce any of the 28 vocalic prototypes of the UPSID data base [15], was randomly presented vowel stimuli produced with the help of the SMIP and covering the F1/F2 acoustic space, and was instructed to indicate the prototype closest to each stimulus. Figure 5 shows the resulting *auditory sensory-motor map* for some of the 28 prototypes. The striking point is the contrast between the low dispersion of the results for the peripheral vowels and the high dispersion of the central ones. Similar tests have been

also conducted with Japanese and Arabic learners of French [15], for the set of French target vowels. These tests, repeated at five weeks interval, have shown dispersions in the learner's maps lower for the second test than for the first one, indicating a gradual decrease of their phonological deafness. Note that this experiment has in particular confirmed that labiality is much less correctly perceived than vowel height.

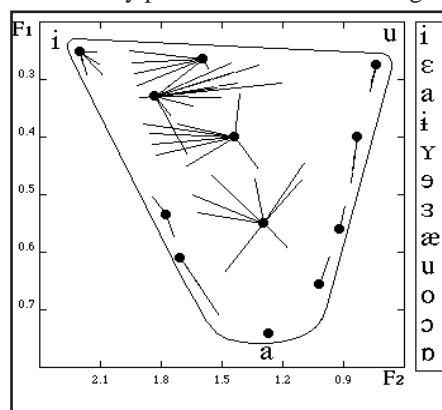


Figure 5. Example of the auditory sensory-motor map of an expert phonetician.

4. Hints towards teaching strategies

In reference to Figure 1, we give hints towards strategies that a teacher could develop within the framework of *Speech Mapping*. Basically, the teacher will use the talking head to manipulate in a controlled way speech articulations and associated audio-visual speech signals.

The teacher will use the SMIP to generate appropriate stimuli in order to evaluate the learner's ability to discriminate vowels in the target language, and to reduce progressively his/her phonological deafness by helping him built the auditory map of the target language starting from his own language phonological inventory. This work will be particularly based on the UCLA database (UPSID) of the phonological systems of the world languages maintained at ICP [15]. The evaluation of the progress made by Japanese and Arabic learners of French reported in [8] constitutes a first example of this strategy.

The teacher will also use his/her knowledge of the articulatory-acoustic relations to guide the learner during the acquisition of the appropriate articulatory gestures, starting also from elements in his/her mother language background. The teacher can experiment with the talking head to find out facilitating strategies. In particular, the properties of *virtual reality* of the talking head, i.e. its ability to display usually hidden articulators such as the tongue will be useful, to demonstrate for instance the tongue backing needed to go from [y] to [u] in French.

The example of a lip-tube compensation experiment [13] provides a good support of the idea that demonstrating articulatory gestures can help the learner. At the occasion of work on the nature of the internal representations of speech for speakers, Savariaux and colleagues [13] conducted an experiment where subjects equipped with a large lip-tube were instructed to produce an isolated French rounded vowel [u]. The study shown

that most subjects realised the strong backing of the tongue needed to achieve the goal only when given clear hints towards the solution, i.e. to produce an [o] and to shift gradually to the good compensated [u]. Our interpretation of these results is that learners can be greatly helped by hints about the right articulatory manoeuvres. Therefore, we are convinced that the possibility to visualise these manoeuvres should be useful for the learner, while the teacher could be helped in discovering the right gestures by experimenting with a system such as the SMIP.

5. Perspectives

We have proposed possibilities to use the *Speech Mapping* tools developed in the context of speech production studies for improving phonetic correction and pronunciation training.

The possibility, offered by the talking head, of simulating the articulatory-acoustic mapping could allow the teacher to experiment various pedagogical solutions. The teacher could test, through virtual reality simulations, the efficiency of the solutions that are traditionally proposed in the literature to transform the sound system of the learner's mother language into the target sound system, and to assess the problems related to erroneous articulatory manoeuvres. The efficient solutions would then be stored in a library or *good ideas* to be used by the teacher or even the learner. In the future, the teacher will be able to edit the trajectories of the talking head articulatory control parameters that can be extracted either from the database, obtained by inversion or synthesised by an audio-visual text-to-speech synthesis system, in order to elaborate strategies to facilitate the learner's acquisition of the target language segmental pronunciation.

Acknowledgements

This work has been partly funded by the ARASSH (Rhônes-Alpes Regional Agency for Social and Human Sciences), in the framework of the project: "A Virtual Talking Head: Data and Models in Speech Production".

References

- [1] Abry C and Badin P (1996). *Speech Mapping* as a framework for an integrated approach to the sensory-motor foundation of language, *Proc. 1st ETRW on Speech Production Modeling*, May 1996, Autrans, 175-184.
- [2] Badin P, Gabioud B, Beautemps D, Lallouache TM, Bailly G, Maeda S, Zerling JP, and Brock G (1995). Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model. *Proc 15th ICA*, Trondheim, Jun 1995, IV:349-352.
- [3] Badin P, Baricchi E, and Vilain A (1997). Determining tongue articulation: from discrete flechpoints to continuous shadow. *Proc Eurospeech '97*, Rhodos, Sep 97, 1:47-50.
- [4] Badin P, Pouchoy, L, Bailly G, Raybaudi M, Segebarth C, Lebas JF, Tiede M, Vatikiotis-Bateson E, and Tohkura Y (1998). Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM, *Proc XXIIèmes JEPS*, Martigny (accepted).
- [5] Bailly G (1998). Learning to speak. Sensori-motor control of speech movements, *Speech Communication*, 22(2-3):251--267.
- [6] Beautemps D, Badin P, Bailly G, Galván A, and Laboissière R (1996). Evaluation of an articulatory-acoustic model based on a reference subject, *Proc. 1st ETRW on Speech Production Modeling*, May 1996, Autrans, 45-48.
- [7] Boë LJ, Gabioud B, and Perrier P (1995). The SMIP: An interactive articulatory-acoustic software for speech production studies. *Bulletin de la Communication Parlée*, 3:137-154.
- [8] Chauny C (1996). Catégorisation et espace perceptif des sons de la parole: vers des cartes sensori-motrices, Mémoire de DEA, Université Stendhal, Grenoble.
- [9] Guiard-Marigny T, Adjoudani A, and Benoît C (1996). 3D Models of the lips and jaw for visual speech synthesis. *Progress in speech synthesis*, Van Santen et al., Eds. Springer-Verlag.
- [10] LeBel JG (1990). Traité de correction phonétique ponctuelle: essai systémique d'application, C.I.R.A.L., Université Laval, Québec, Canada.
- [11] LeGoff B. and Benoît C (1997). A French-speaking synthetic head. *Proc Audio Visual Speech Processing Workshop*, Rhodes, Sep 1997, 145-148.
- [12] Mawass K, Badin P, and Bailly G (1997). Synthesis of fricative consonants by audiovisual-to-articulatory inversion. *Proc Eurospeech '97*, Rhodos, Sep 97, 3:1359-1362.
- [13] Savariaux C, Boë LJ, and Perrier P (1997). How can the control of the vocal tract limit the speaker's ability to produce the ultimate perceptive objectives of speech? *Proc Eurospeech '97*, Sep 1997, Rhodos, 2:1063-1066.
- [14] Stellio JF (1996) Eléments théoriques et suggestions pratiques pour améliorer la prononciation du français d'apprenants étrangers, Mémoire de DEA, Université Stendhal, Grenoble.
- [15] Vallée N, Boë LJ, & Payan Y (1995). Vowel prototypes for UPSID's 33 phonemes. XIIIth ICPHS, 1:424-427.