# Multimodal Speech: from articulatory speech to audiovisual speech

Slim Ouni

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Département de formation doctorale en informatique

# Multimodal Speech
## from articulatory speech to audiovisual speech

## *Parole Multimodale*
## *De la parole articulatoire à la parole audiovisuelle*

déposé le 20 Juin 2013

pour l'obtention d'une

**Habilitation à Diriger des Recherches de l'Université de Lorraine**

**(Spécialité informatique)**

par

Slim Ouni

**Composition du jury**

| | |
|---|---|
| *Rapporteurs :* | Olov Engwall, Professeur, Centre for Speech Technology, KTH, Suède<br>Catherine Pelachaud, DR CNRS, TELECOM ParisTech, France<br>Bernd Kröger, Professeur, RWTH Aachen University, Allemagne |
| *Examinateurs :* | Nick Campbell, Professeur, Trinity College Dublin, Irlande<br>Bernard Girau, Professeur, Université de Lorraine, France<br>Yves Laprie, DR CNRS, LORIA, France<br>Jean-Marie Pierrel, Professeur, Université de Lorraine, France |

**Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503**

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thesul.

# Table des matières

# Introduction

Spoken communication is inherently multimodal. The acoustic signal carries the auditory modality and the image carries the visual and gestural modalities (facial deformation). The speech signal is in fact the consequence of the deformation of the vocal tract under the effect of the movement of the jaw, lips, tongue, soft palate and larynx to modulate the excitation signal produced by the vocal cords or air turbulence. These deformations are visible on the face (lips, cheeks, jaw) through the coordination of different orofacial muscles and skin deformation induced by the latter. The visual modality can provide additional information to the acoustic signal, and it becomes essential if the acoustic signal is degraded, as is the case with hard of hearing or in a noisy environment. Other modalities may be related to speech, such as eyebrow movements and gestures that express different emotions. This latter modality that is suprasegmental can complete the acoustic or acoustic-visual message.

In this document, I present my main research during the last 10 years. Figure 1 gives an overview of my research activities. I consider speech as a multimodal object that can be studied from an articulatory, or acoustic, or visual standpoint. These different modalities can be combined and investigated together or two-by-two. The common point of these different aspects is that it is based on the data, and thus acquiring and processing data is a central point in my research. I am interested also in investigating the production of this multimodal object and its synthesis, but also on its perception by human perceiver in a face-to-face communication.

In a nutshell, I am mainly interested in the multimodal nature of speech communication that can be treated in two different ways:

1. consider both articulatory and acoustic components of speech. Indeed, I am interested in the articulatory characterization of speech sound and the study of the relationship between articulatory space and acoustic space. In particular, I am interested in the recovery of the temporal evolution of the vocal tract from the acoustic signal, also called acoustic-to-articulatory inversion (**Chapter 1**), and the study of articulatory characterization of speech by analyzing corpus articulatory data (**Chapter 2**).
2. study both acoustic and visual components. In this context, I am interested in the effect of the deformation of the vocal tract on the facial appearance that conveys the visual message. The acoustic-visual synthesis is a framework that allows us to study this aspect (**Chapter 3**). In addition, I study audiovisual intelligibility to better understand the mechanisms of audiovisual communication, but also to evaluate the acoustic-visual synthesis system (**Chapter 4**).

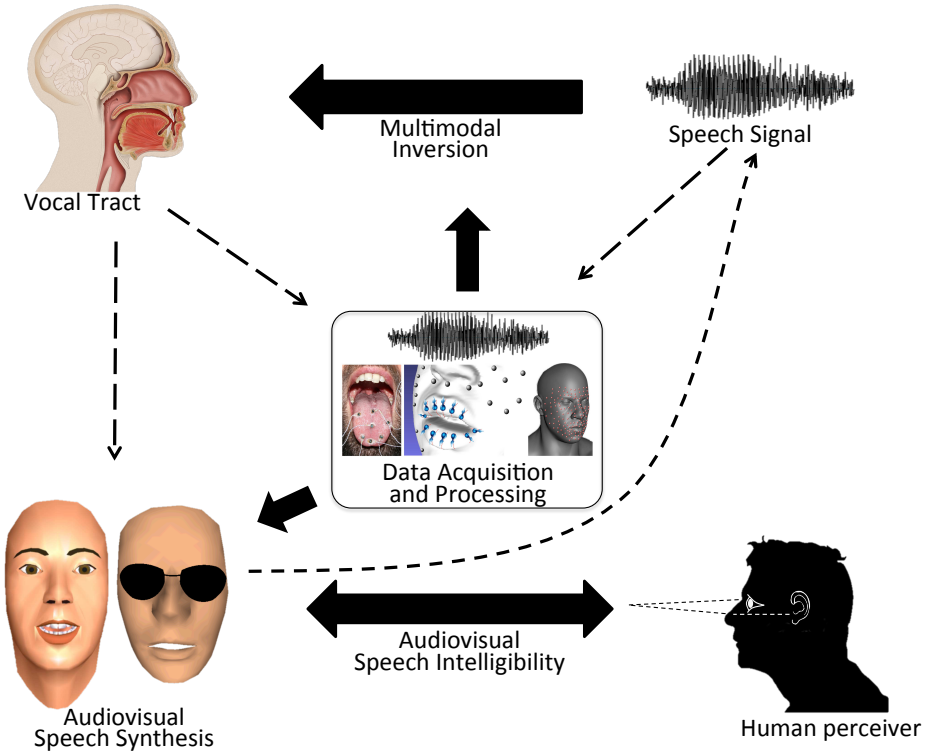



FIGURE 1 – Overview of my research activities

Acoustic-to-Articulatory inversion would enable knowing how a speech signal has been articulated. This potential knowledge could give rise to a number of breakthroughs in automatic speech processing, as in recognition and synthesis. Inversion can also have impact on applications as language learning and speech therapy. Inversion can be investigated using an analysis-by-synthesis approach or using articulatory data acquired using an acquisition system as an articulograph. In **Chapter 1**, I present three different approaches used to deal with the inversion problem. During my earlier work, articulatory-to-acoustic mapping has been addressed, where *Yves Laprie* and I proposed a representation of the articulatory space by a hypercube codebook. The inversion is performed using this codebook to provide a set of solutions combined with dynamic programming. Later on, I have shifted the focus on data by analyzing and considering real data for inversion, instead of data obtained by synthesis. This allows dealing with more realistic cases. Thus, I have investigated data-driven inversion techniques using articulatory corpora. During the postdoctoral stay of *Sebastien Demange*, we have introduced an inversion method that is based on episodic memory. The main advantage of this representation is that the memory models the real articulatory dynamics as observed. We have proposed a memory which is able to produce new articulatory trajectories that do not belong to the set of episodes the memory is based on. During the postdoctoral stay of *Asterios Toutios*, we addressed the problem of multimodal inversion, or how to retrieve the vocal tract dynamics from acoustic speech signal and the face.This inversion method is based on a statistical data-driven approach using support vector regression for the mapping from acoustic parameters to electromagnetic articulography (EMA) trajectories. The

data is based on synchronized acoustic and articulatory data streams using EMA. We extended the acoustics with visual information. We were targeting the invisible articulators, mainly the tongue.

To deal with data-driven methods, it is essential to acquire and process data. In fact, the nature of my research makes me highly concerned by acquisition and processing data as it is a central point in my work. Acquiring data and processing it, can be time-consuming and costly in terms of the effort required to carry out the acquisition and processing the data. This effort is unavoidable to make more progress in modeling the processes related to human communication. In **Chapter 2**, I presented my contribution in articulatory data acquisition techniques. I proposed a powerful tool, called VisArtico, which is an articulatory visualization software that is intended to visualize articulatory data acquired using an articulograph. I built a strong experience in EMA acquisition techniques, and several corpora have been developed and used to study, for instance, coarticulatory effect of pharyngealization in Arabic. As the data can also be visual, I have also discussed the motion capture aspects to be considered.

Articulatory and acoustic components are highly correlated, as the latter is the consequence of the former. In audiovisual speech, this should be also the case, where audiovisual speech synthesis should consider acoustic and visible components simultaneously. Therefore, audiovisual speech should be considered as a bimodal signal with two channels: acoustic and visual. This is actually the introduced vision in the ViSAC project. Within the framework of this project, I led this project that is a collaboration between two groups: Magrit group (*Brigitte Wrobel-Dautcourt* and *Marie-Odile Berger*, specialists in geometry capture and deformation) and Speech group (*Vincent Colotte*, our expert in acoustic speech synthesis). In addition, *Utpala Musti* has worked on the project during her Ph.D.; *Asterios Toutios* has developed the basic system during his postdoc stay ; *Blaise Potard* has worked as engineer to improve the stereovision acquisition system and *Caroline Lavecchia*, engineer, has developed the experimental evaluation platform. During the ViSAC project, we introduced a bimodal unit-selection synthesis technique that performs text-to-speech synthesis with acoustic and visual components simultaneously. It was based on the concatenation of bimodal diphones, units that consist of both acoustic and visual components. The results of the conducted evaluation showed that audiovisual speech provided by this synthesis technique is intelligible and acceptable as an effective tool of communication. It is worth noticing that the generation of facial animation together with the corresponding acoustic speech, is still considered in several works as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and facial animation. This was the case during my earlier work on parametric talking head where a multilingual talking head has been developed using extensive phonetics knowledge and successive perceptual experiments and modifications. All these aspects of audiovisual speech synthesis are presented in **Chapter 3**.

As speech communication is mainly face-to-face, we are naturally concerned by how effective is the audiovisual speech synthesis, which is critically dependent on the quality of the visual speech. I consider that it is important to study audiovisual intelligibility and the ability of the synthesis to send an intelligible message to the human receiver. In fact, the intelligibility of the audiovisual synthesis can be critical when considering applications addressed to hard-of-hearing humans or to learners of new languages. Thus, focusing on audiovisual intelligibility is extremely important. To assess the quality of audiovisual speech synthesis, objective perceptual experiments should be designed where human participants are usually asked to recognize the presented

linguistic items. My main goal is to investigate the mechanism of audiovisual intelligibility, what makes an audiovisual message intelligible, How the visual component contributes to the audiovisual perception and, more importantly, how to assess the audiovisual intelligibility. In **Chapter 4**, I present my work related to this topic. I have proposed a metric that allows the comparison of an animated agent relatively to a standard or a reference. This metric allows direct comparisons across different experiments and give measures of the benefit of a synthetic animated face relative to a natural face. I have also studied the importance of the audiovisual speech in language learning and how visual speech can contribute to a better perception. In particular, I am interested in investigating whether speech production and perception of new language would be more easily learned when using audiovisual instead of audio-only speech. In particular, I investigated whether viewing internal articulators (e.g. the tongue, not completely visible) is more beneficial for pronunciation training than seeing only the face from outside.

The first part of this document presents my main research works during last years, where I am studying speech as a multimodal object. The majority of my work presented here is published in international peer-reviewed journals. Thus, I present in this document an overview augmented with the necessary details to make the explanation clear and gives an idea of the conducted work. In the second part, I present my research program for the future. There will be a shift from neutral audiovisual speech to expressive audiovisual speech that will be studied at different levels: articulatory, visual and acoustic, taking into account all the information related to speech and those related to expression. I consider audiovisual speech as a unique multimodal signal carrying acoustic and visual components with the embedded expressivity. The main goal is to consolidate our knowledge and experience in the three fields (articulatory, acoustic and visual) to deal with expressive audiovisual speech synthesis from articulatory, acoustic and visual standpoints simultaneously using a global approach. The goal is to get much closer to natural behavior during face-to-face communication. I will also focus more on the realism of the dynamics, essential feature for intelligibility. All the details regarding my research directions are detailed in **Chapter 5**.

# My Research work

# 1

# Multimodal Inversion

## 1.1 Introduction

Estimating the vocal tract shape from a speech signal has received considerable attention because it offers new perspectives for speech processing. Recovering the vocal tract shape would enable knowing how a speech signal has been articulated. This potential knowledge could give rise to a number of breakthroughs in automatic speech processing. For instance, the location of critical articulators [Papcun et al., 1992, Ananthakrishnan and Engwall, 2011a] could be exploited to well characterize a given phoneme or to discard some acoustic hypotheses. This may improve speech synthesis [Sondhi, 2002], but also automatic speech recognition [Erler and Deng, 1993, Deng and Sun, 1994, Rudzicz, 2010]. In fact, articulatory features vary much more slowly and present less variability than speech acoustic features, and thus they should be more robust than acoustic parameterization especially in noisy environments. This might be one of the strongest motivations of many researchers showing interest in speech inversion. Actually, almost all related studies showed that adding articulatory information improves speech recognition results [Frankel and King, 2001, Richardson et al., 2003, Al Bawab et al., 2008]. As other possible applications to inversion, one can cite, for instance, language learning and speech therapy. In fact, using speech inversion could provide articulatory feedback. In this case, it is important to retrieve exactly the vocal tract shapes as produced by the speaker to utter a given acoustic speech signal. And finally, in the domain of phonetics, inversion would enable knowing how sounds were articulated without requiring medical imaging or other measurement techniques. I should note that to my knowledge, these applications are still not fully implemented and the work is either experimental or in a research stage.

## 1.2 Inversion: state of the art

The acoustic-to-articulatory inversion is a difficult problem mainly because of the non-uniqueness of the relationship between the articulatory and acoustic spaces and the non-linearity of this relationship [Atal et al., 1978, Charpentier, 1984b, Stevens, 1989, Ouni and Laprie, 2005a, Neiberg et al., 2008]. This challenging problem did not discourage researchers to work on developing methods to resolve the problem, since the work of [Atal et al., 1978] to our days.

A family of acoustic-to-articulatory inversion methods rely on an analysis-by-synthesis approach. That means the articulatory-to-acoustic mapping is represented either explicitly, by pairs stored in a codebook for a number of points which sample the articulatory space, or implicitly, by neural networks, for instance. These methods usually rely on an articulatory synthesizer built on an articulatory model. Several methods represent the mapping by a codebook. The quality of the representation influences strongly the recovered solutions as these trajectories use points of the codebook. The articulatory domain is usually quantized to represent all possible geometric configurations of the vocal tract. For instance, the sampling can be a random sampling [Larar et al., 1988, Schroeter et al., 1990, Boë et al., 1992] or using root-shape interpolation [Larar et al., 1988] (This method consists of sampling the articulatory space in a nonuniform manner by sampling the most probable regions, i.e., those corresponding to the most often observed vocal tract shapes).

Other methods implicitly use a codebook obtained, for instance, by training a neural network [Soquet et al., 1991, Laboissièrre and Galvàn, 1995]. The majority of these *codebook-like* methods are based on an articulatory synthesizer and thus it is a numerical series of approximations to the real mapping between the vocal tract shape and the acoustic. There is no guarantee that the recovered trajectories correspond to how the speaker really uttered speech, even when the corresponding acoustics perfectly match the input signal. In fact, the weak point of these methods is that codebooks do not take into account the dynamics: the collected pairs are independent and unrelated. There is no temporal information. The optimization techniques used in these methods can help to obtain better dynamics related to smoothness or to avoid extreme trajectories, but it is practically impossible to ensure that the obtained trajectories are what the speaker uttered.

Recently, databases of synchronized acoustic and articulatory data streams, using electromagnetic articulography (EMA) for instance, have become available. These corpora enable machine learning algorithms to perform acoustic-to-articulatory regression. The main techniques use support vector machines [Toutios and Margaritis, 2008], Gaussian mixture models [Toda et al., 2008], hidden Markov models (HMM) [Hiroya and Honda, 2004, Zhang and Renals, 2008], or artificial neural networks [Richmond, 2002, Richmond, 2006]. A mixture density network has been used in [Richmond, 2006] to obtain conditional probability densities of the acoustic features. Dynamic articulatory features were used in addition to static features to obtain a statistical trajectory model. [Toutios and Margaritis, 2008] used support vector regression to estimate EMA trajectories. The reported results were comparable to other statistical learning methods. This method however deals only with static aspects of EMA. [Toda et al., 2008] proposed a Gaussian mixture model to represent the mapping from static EMA features to mel-cepstral coefficients using a maximum likelihood estimation (MLE) to integrate the dynamic features. [Hiroya and Honda, 2004] presented an HMM based speech production model where HMMs represent articulatory features for each phoneme, as well as a mapping from articulatory features into acoustic features for each HMM state.

I believe that the main difficulty of inversion is the lack of a good representation of dynamics. The non-uniqueness problem [Ananthakrishnan, 2011] will very likely vanish if the dynamics are fully integrated within the inversion methods. In fact, [Qin and Carreira-Perpiñán, 2007], and to some extent [Neiberg et al., 2008], argued that natural human speech is produced with a unique vocal tract configuration and there are few cases of non-uniqueness. Phonetic context naturally imposes constraints on the vocal tract related to coarticulation. Ef-

fective modeling of articulatory dynamics seems essential to solving the inversion problem. These dynamics can be modeled with HMMs [Hiroya and Honda, 2004, Zhang and Renals, 2008] and neural networks [Richmond, 2002]; they can also be inferred from time derivative features [Toda et al., 2008]. For the codebook-based methods the dynamics are not modeled at all. Instead, continuity constraints are used during inversion. However, despite these constraints, the recovered articulatory trajectories show many discontinuities and need to be smoothed with signal processing techniques.

## 1.3 Multimodal Inversion: Contributions

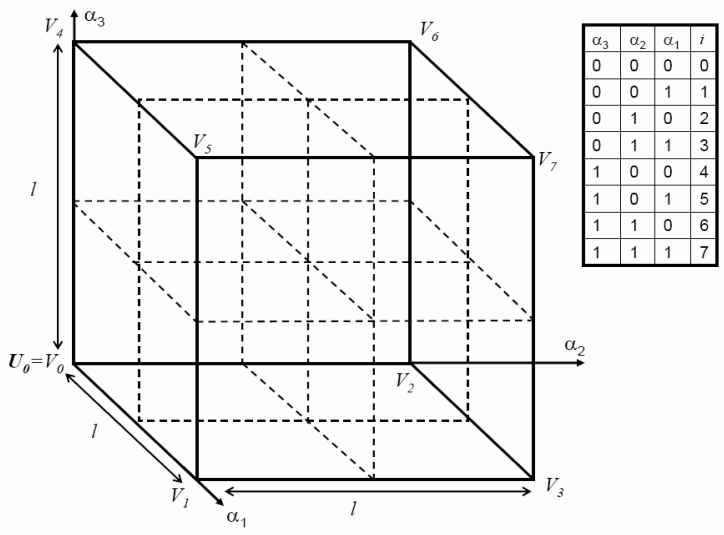
I have started working on acoustic-to-articulatory inversion for more than thirteen years. *Yves Laprie* and I, first, addressed the problem of representing articulatory-to-acoustic mapping. We proposed a representation of the articulatory space by hypercube codebook. We used a hierarchy of hypercubes. Each hypercube represents an articulatory region where the articulatory-to-acoustic mapping can be approximated by means of a linear transform [Ouni and Laprie, 2005a]. Then, the inversion is performed using a codebook to provide a set of solutions combined with dynamic programming and eventually additional constraints [Schroeter and Sondhi, 1994]. For instance, in [Ouni and Laprie, 2005a], a nonlinear smoothing algorithm together with a regularization technique were used to recover the best articulatory trajectory. This work was based on analysis-by-synthesis approach and relied on an articulatory synthesizer built on an articulatory model, that of [Maeda, 1979].

Recently, databases of synchronized acoustic and articulatory data streams using electromagnetic articulography (EMA) have become available. In addition, we acquired a 3D articulograph which motivated me to focus on analyzing and considering real data for inversion, instead of data obtained by synthesis. This allows dealing with more realistic cases. Within the framework of our research in inversion, I proposed two inversion techniques using observed articulatory data synchronized with acoustics. The first inversion method is based on a statistical data-driven method using support vector regression for the mapping from acoustic parameters to electromagnetic articulograph trajectories [Toutios and Margaritis, 2008] (the topic of **Asterios Toutios** during his postdoc). We extended the acoustics with visual information. We were targeting the invisible articulators, mainly the tongue [Toutios et al., 2011b]. The second inversion method is based on episodic memory (the topic of **Sebastien Demange** during his postdoc). The main advantage of this representation is that the memory models the real articulatory dynamics as observed. We have proposed a memory which is able to produce articulatory trajectories that do not belong to the set of episodes the memory is based on [Demange and Ouni, 2013].

In the following sections, I present our work dealing with model-based inversion, facial-acoustic-to-articulatory inversion and episodic memory-based inversion.

## 1.4 Model-based inversion

In our earlier work, we focused on modeling the articulatory-to-acoustic mapping using a codebook. A codebook is a collection of a vast number of vocal tract shapes given by articulatory or area function parameters indexed by their acoustic parameters. The acoustic parameters are obtained using an articulatory synthesizer. The articulatory space should be spanned so that the codebook represents all of the possible geometric of the vocal tract. However, this is difficult due to the fact that the relations between articulator positions and acoustics are non-linear [Fant, 1960, Stevens, 1972, Charpentier, 1984a]. In fact, there are articulatory regions




| $\alpha_3$ | $\alpha_2$ | $\alpha_1$ | $i$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 2 |
| 0 | 1 | 1 | 3 |
| 1 | 0 | 0 | 4 |
| 1 | 0 | 1 | 5 |
| 1 | 1 | 0 | 6 |
| 1 | 1 | 1 | 7 |

FIGURE 1.1 – For sake of clarity we represent a 3D hypercube. Note that the edge length is $\ell$ and $U_0$ is the origin of the hypercube. $V_i$ ($i = 0..7$) are the vertices of the hypercubes. The linearity test is performed on the segments $[V_i, V_j]$ where $i \neq j$. If the test fails the hypercube is split into 8 sub-hypercubes (8 is the number of the vertices in 3D). These sub-hypercubes are represented with dashed lines. The upper table gives the values of the parameter $\varphi_{ij}$ for the 8 vertices indexed from 0 to 7. [Ouni and Laprie, 2005a]

where a small variation in articulatory parameters produces a large variation of acoustic parameters. And conversely, there are some regions where a large variation in articulatory parameters does not produce any significant acoustic changes.

### Hypercube codebook inversion

We proposed to densely discretize the articulatory space only in the regions where the mapping is highly non-linear, using a hypercube structure to organize the codebook. Each hypercube represents an articulatory region where the articulatory-to-acoustic mapping can be approximated by means of a linear transform [Ouni and Laprie, 2005a]. In a given hypercube, the articulatory-to-acoustic mapping is considered to be linear. The linearity is evaluated by measuring the deviation between the acoustic values obtained by synthesis and those obtained by interpolation from codebook points. When the mapping is not linear, this hypercube is split into several equal sub-hypercubes and the linearity test is repeated for every new hypercube. This procedure is repeated recursively until the hypercube edge becomes smaller than a predefined value or no non-linearity higher than the predefined threshold exists anymore (See Figure 1.1).

The hypercube codebook is used to retrieve the articulatory parameters corresponding to the acoustic entry. All the hypercubes whose acoustic image contains the acoustic entry are examined. Recovering articulatory trajectories consists of choosing at each time an articulatory vector among those obtained by the inversion. This amounts to finding an articulatory path expressing the time evolution of the vocal tract shapes during the utterance to be inverted. The resulting articulatory trajectory should vary "slowly" (variations of articulatory parameters are small during an average pitch period, i.e. approximately 10 ms) and generates spectra as close as possible to those of the original speech. This corresponds to the satisfaction of two criteria: proximity to acoustic data and smoothness of articulatory trajectories. The overall inversion

algorithm that combines these two criteria, works as follows:

1. The first step of the inversion consists of recovering all of the inverse articulatory solutions at each instant of the utterance to be processed by exploring the codebook.
2. In the second step a non-linear smoothing algorithm, derived from a non-linear smoothing algorithm initially proposed by [Ney, 1983], finds smooth articulatory trajectories from the knowledge of the sets of inverse points recovered at each time frame.
3. The third step consists of regularizing the trajectories built by using the non-linear smoothing algorithm. This regularization is achieved through a variational method.

The details of this method are presented in [Ouni and Laprie, 2005a]. The main advantage of this method is that it ensures that all the possible inversion solutions can be explored, given an articulatory model and the frequency precision set for the acoustics being recovered, and does not implicitly favor any particular articulatory solution. The neutrality of the inversion enables evaluating several strategies for guiding the inversion process. We have implemented one strategy, where phonetic constraints have been incorporated in the inversion. These constraints have been arisen from phonetic knowledge [Potard et al., 2008].

Our work on inversion using hypercube codebook was a baseline for other works within the research group. For instance, *Blaise Potard*, during his PhD, has improved the hypercube codebook generation algorithm and used phonetic and visual constraints for inversion [Potard, 2008]; and *Julie Busset*, during her PhD, has used cepstral coefficient instead of formants as acoustic features [Busset, 2013].

We consider that the main drawback of the model-based Inversion is that there is no guarantee that the recovered trajectories correspond to how the speaker really uttered speech, even when the corresponding acoustics perfectly match the input signal. The optimization techniques used in these methods can help to obtain better trajectories to avoid excessive unrealistic trajectories, but it is practically impossible to ensure that the obtained trajectories are what the speaker uttered.

### Model-based *vs.* data-driven

For the above cited reasons, we are tempted to explore data-driven inversion methods. The advantage of such methods is that all the data is based on real data. It is possible, however, to make a link between the model-based inversion method and the articulatory data. To do this, a mapping between the parameters of the articulatory model and the acquired articulatory data should be found. We have developed a method to estimate the control parameters of the Maeda's model from EMA data [Toutios et al., 2011c]. This allows the derivation of full sagittal vocal tract slices from EMA sensors position. The proposed method first adapts the articulatory model to the speaker for whom EMA data was collected and an initial solution for the control parameters is determined by a least-squares method. Then, a dynamic smoothness of the articulatory trajectories is then applied using a variational regularization method. The evaluation of this method showed that formants synthesized on the basis of the estimated model-parameters were close to those tracked in real speech recorded synchronously with EMA (see [Toutios et al., 2011c] for details). This method can be considered as the missing link between model-based and data-driven articulatory techniques. In fact, all the conclusions that can be made or the applications that can be developed based on articulatory real data is highly dependent on the recorded speakers. The advantage of model-based inversion is that it is possible to adapt the data from one speaker to another. However, it cannot be efficient if the articulatory model is not very accurate and does not express precisely the real behaver of the vocal tract. This is far from being the case.

In fact, the 2D articulatory model is an approximation of the mid-sagittal section of the vocal tract, and the reconstruction of the 3D vocal tract is made by an approximate transformation. In addition, the articulatory model does not fit exactly the geometry of the particular speaker to be analyzed. To sum up, research in articulatory modeling needs to be conducted to hope for a better articulatory synthesizer and thus a better model-based inversion.

## 1.5 Multimodal Inversion: from Acoustics-Face to Articulatory

### 1.5.1 Introduction

Acoustic-to-articulatory inversion can be seen as estimating unseen features from observable ones. Classically, we estimate invisible articulatory features from "audible" acoustic ones. This is the case when we developed our data-driven inversion method [Demange and Ouni, 2013] (see section 1.6) that performs the mapping from acoustics to EMA with results in agreement with published state-of-the-art solutions related to acoustic-to-articulatory inversion problem [Ananthakrishnan and Engwall, 2011b, Ghosh and Narayanan, 2010, Richmond et al., 2003, Toda et al., 2008]. It is possible to extend this definition to define *multimodal inversion*. When adding another modality to the acoustics as input to inversion, we consider the method as multimodal. The additional modality should be observable, i.e., measurable or visible. Typically, in face-to-face communication, the message is conveyed by an acoustic signal and facial deformation. We can consider the case of retrieving the tongue deformation as main vocal tract shape from acoustic and facial features. The facial features can be the lips and the jaw, for instance.

Our hypothesis is that adding visual information regarding the lips and jaw to acoustics can improve the accuracy of predicting the tongue sensors, since such an addition can make the mapping less ambiguous. This information will regard EMA sensors on the lips and jaw in the acquisition of training data. We conducted a study to predict tongue shape from acoustics, lips and jaw using EMA data from MOCHA dataset [Wrench and Richmond, 2000], both for training and testing.

Adding visual information to an acoustic-to-EMA mapping setup using MOCHA was done in [Katsamanis et al., 2009] with no significant improvement to the prediction of the tongue sensor positions. Nevertheless, in that case the visual information was extracted from video images in a complex way, that might be a source of inaccuracies. We believe that using as visual information simply the positions of the EMA sensors on the jaw and lips, may be more informative. [Ben Youssef et al., 2010] studied the case where the tongue sensors are predicted using *only* information on the lip and jaw sensors, i.e. no acoustic information at all. As probably expected, the results were not good, suggesting that visual information alone is not enough to predict tongue position. Just for the sake of completeness, we replicate their experiments using our mapping method on MOCHA, as they used another dataset.

In the following sections, I present an overview of the method that retrieve the EMA positions of the tongue from acoustics and the EMA sensors of the lips and jaw. This method is based on support vector regression. I present also the experimental results. This work was done with Asterios Toutios and the complete study is detailed in [Toutios et al., 2011b].

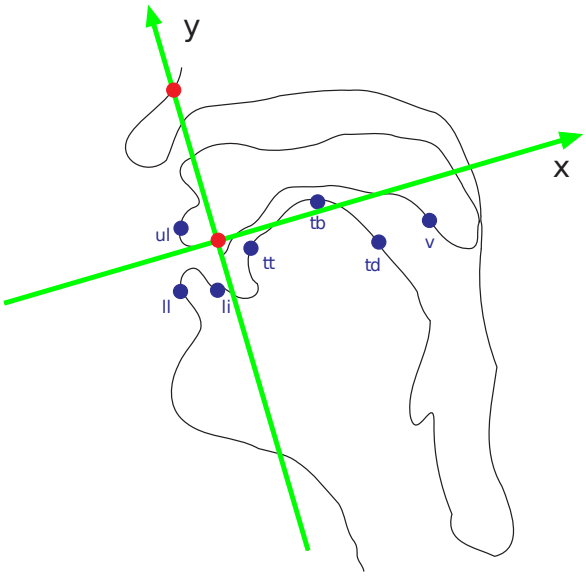



FIGURE 1.2 – Approximate positioning of sensors in the MOCHA database, and axes involved. The two coils at the bridge of the nose and the upper incisors are used for the normalization of the data from the rest. Seven coils, located at the lower incisors (*li*), upper lip (*ul*), lower lip (*ll*), tongue tip (*tt*), tongue blade (*tb*), tongue dorsum (*td*) and velum (*v*), offer useful location information, namely trajectories of the projections of their position on two axes on the midsagittal plane: one with direction from the front to the back of the head (x-axis) and one with direction from the bottom to the top of the head (y-axis).

### 1.5.2 Data and Method

MOCHA includes electromagnetic articulography (EMA) information for the coils shown in Figure 1.2. The information flows from individual coils on individual axes are referred to as *EMA channels*. We considered three different kinds of input information. For the *acoustic experiment*, the input information for each frame was its MFCC parameters. For the *facial experiment*, the input information for each frame was the values of the six EMA channels corresponding to sensors *li*, *ul*, and *ll*. For the *acoustic + facial experiment* the input information for each frame was the union of the previous two sets. In all cases, we constructed context input vectors spanning over 11 consecutive frames, centered around the output frame. The output information was the value of one of the EMA channels corresponding to sensors *tt*, *tb*, and *td*.

We trained the $\epsilon$-SVR algorithm with the gaussian kernel, using the LibSVM software [Chang and Lin, 2001]. The algorithm solves the following optimization problem:

$$\begin{gathered}\text{maximize} \\ -\varepsilon \sum_{i=1}^{n}(a_i^* + a_i) + \sum_{i=1}^{n}(a_i^* - a_i)y_i \\ -\frac{1}{2}\sum_{i,j=1}^{n}(a_i^* - a_i)(a_j^* - a_j)\exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2) \\ \text{subject to} \\ 0 \leq a_i, a_i^* \leq C,\ i = 1, \ldots, n \text{ and } \sum_{i=1}^{n}(a_i^* - a_i) = 0,\end{gathered} \tag{1.1}$$

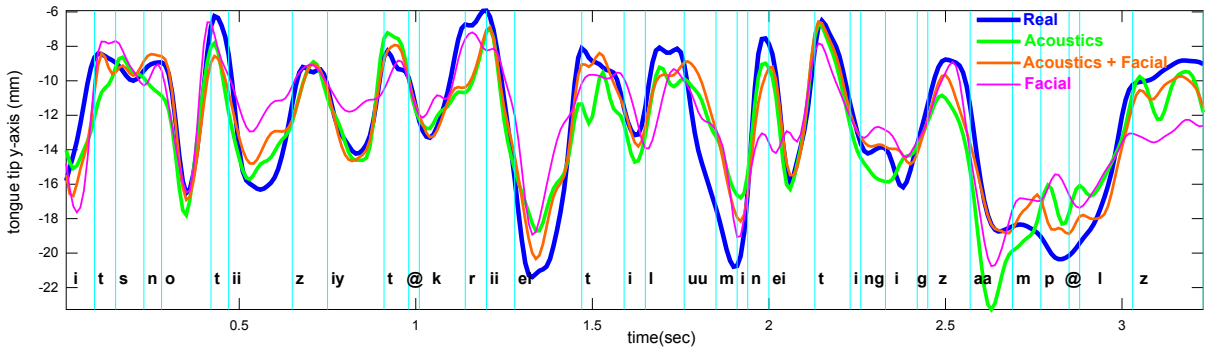



FIGURE 1.3 – Real and estimated trajectories describing the projection of tongue tip sensor position on axis $y$, as shown in Figure 1.2, for the utterance "It's not easy to create illuminating examples", spoken by speaker fsew0 of MOCHA. Vertical lines mark the boundaries between phones. Vertical axis denotes millimeters; horizontal axis denotes time in seconds. RMS errors and Pearson correlations between the estimated trajectories shown and the real trajectory are given in parentheses in Table 1.1. The MOCHA labeling convention is used for the names of the phonemes.

over the $n$ input vectors $\mathbf{x_i}$ and corresponding output scalars $y_i$, to provide with the mapping function

$$f(\mathbf{x}) = \sum_{i=1}^{n} (a_i^* - a_i) \exp(-\gamma \|\mathbf{x} - \mathbf{x_i}\|^2) + b \tag{1.2}$$

where $b$ is calculated from the Karush-Kuhn-Tucker conditions for the problem [Schölkopf and Smola, 2001]. The parameters $C$, $\varepsilon$, and $\gamma$ are to be selected by the experimenter. Based on [Cherkassky and Ma, 2004], we used

$$C = \max\left(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|\right) \tag{1.3}$$

where $\bar{y}$ and $\sigma_y$ are the mean and the standard deviation of the output values of training data, and

$$\varepsilon = 3\sigma_n \sqrt{\frac{\ln n}{n}} \tag{1.4}$$

where $n$ is the number of training examples, and $\sigma_n$ is the median value of $\sqrt{(y - \bar{y})^2}$ across the training output data. Finally, based on [Tsang et al., 2006], we used

$$\gamma = \frac{m^2}{\sum_{i,j=1}^{n} \|\mathbf{x_i} - \mathbf{x_j}\|^2}. \tag{1.5}$$

The results of the mapping function of Eq. 1.2 were z-scored values of the EMA channels. After testing, we inverted z-scoring and added back the filtered version of the channel mean corresponding to the utterance in question. As a final post-processing step, we smoothed the resulting trajectories using a low pass-filter at 20 Hz, i.e. the same filter we used at pre-processing.

### 1.5.3 Results

We experimented using the data of speaker fsew0 from MOCHA, i.e. a female with a Southern English accent. We performed cross-validation experiments over the 460 numbered utterances

**root mean squared error (mm)**

| channel | acoustic | acoustic + facial | facial |
|---|---|---|---|
| $tt_x$ | 2.34 (1.83) | 2.15 (1.74) | 3.52 (2.72) |
| $tt_y$ | 2.34 (1.98) | 2.11 (1.54) | 3.30 (2.39) |
| $tb_x$ | 2.19 (2.13) | 1.98 (2.18) | 3.16 (2.67) |
| $tb_y$ | 2.06 (2.23) | 1.95 (2.23) | 3.50 (4.23) |
| $td_x$ | 2.03 (2.08) | 1.81 (1.97) | 2.81 (2.51) |
| $td_y$ | 2.12 (1.91) | 2.06 (1.91) | 3.40 (3.48) |
| average | 2.18 (2.03) | 2.01 (1.93) | 3.28 (3.00) |

**Pearson correlation**

| channel | acoustic | acoustic + facial | facial |
|---|---|---|---|
| $tt_x$ | 0.813 (0.794) | 0.846 (0.811) | 0.543 (0.573) |
| $tt_y$ | 0.861 (0.879) | 0.888 (0.933) | 0.710 (0.811) |
| $tb_x$ | 0.806 (0.795) | 0.845 (0.791) | 0.573 (0.622) |
| $tb_y$ | 0.857 (0.895) | 0.873 (0.914) | 0.546 (0.402) |
| $td_x$ | 0.791 (0.757) | 0.839 (0.770) | 0.587 (0.562) |
| $td_y$ | 0.796 (0.837) | 0.809 (0.846) | 0.420 (0.323) |
| average | 0.821 (0.826) | 0.850 (0.844) | 0.562 (0.549) |

TABLE 1.1 – Outside the parentheses are cumulative results for the 460 utterances spoken by fsew0, after cross-validation experiments (see text for the explanation of the inputs used). Inside the parentheses are results for an example of the single utterance "It's not easy to create illuminating examples" which is also illustrated in Figures 1.3 and 1.4. Subscripts at sensor names denote the projection on the axes shown in Figure 1.2.

available, using five partitions. For evaluation we used the two metrics typical in works on acoustic-to-EMA mappings: root mean squared error

$$E_{RMS} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y'_i - y_i)^2} \tag{1.6}$$

where $m$ is the number of examples in the test set, and $y, y'$ are real and estimated values; and Pearson correlation

$$r = \frac{\sum_{i=1}^{m}\left(y'_i - \overline{y'}\right)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}\left(y'_i - \overline{y'}\right)^2 \sum_{i=1}^{m}(y_i - \overline{y})^2}} \tag{1.7}$$

where overlines denote mean values over the test set.

These results are summarized in Table 1.1. The numbers outside parentheses refer to the whole dataset of 460 utterances. The numbers between parentheses refer to the utterance "It's not easy to create illuminating examples" which was chosen randomly from the dataset, for illustration purposes. For this utterance, the real and estimated trajectories for the projection of the tongue tip sensor on the y-axis of Figure 1.2 are shown in Figure 1.3. For the same utterance, Figure 1.4 shows snapshots of an animation of the results.

Regarding overall results, we can see that the addition of facial cues to acoustics improves the performance on all six tongue channels, in terms of both metrics used. On average, the improvement in root mean squared error is 0.17 mm (7.8%), and the improvement in Pearson correlation is 0.29 units (3.5%). The performance of the system that uses only facial features as input is very poor, which verifies the findings of [Ben Youssef et al., 2010].

Regarding the single utterance, the presented results for both systems using acoustics (with or without facial features) are in general better than the cumulative results over the whole dataset, indicating that the specific utterance is a relatively good case among the 460. Nonetheless, the observed relative improvement after adding facial features is of similar importance compared to the whole (4.9% for root mean squared error, 2.8% for Pearson correlation, on average). Indeed, the trajectory estimated from the combination of acoustics and facial features shown in Figure 1.3 is, in broad terms, a slightly better match to the real trajectory, than the trajectory estimated only from acoustics. But if we focus on the animation presented in Figure 1.4 instead of just numbers (or a single trajectory), the improvement does not seem so important. The tongue contours based on the estimations from the combination of acoustics and facial features are closer to the real contours than their counterparts estimated just from acoustics, but only marginally so. Both sets of estimated contours present more or less the same problems in comparison to the real contours, for example the lack of making contact with the palate for velar consonants /k/ (2nd row, 6th column) or /g/ (5th row, 4th column), or the inverted overall tongue curvature for some instances of alveolar consonants like /t/ (in 1st row, 2nd column) and /n/ (in 1st row, 4th column and in 4th row, 4th column).

### 1.5.4 Application

Our initial motivation when first presented this work [Toutios et al., 2011b] was to use such a method in the context of adding a tongue and controlling it within a talking head system. In both facial and EMA acquisitions we concurrently record (and synchronize) the acoustic signal. We could obtain concurrent facial and *estimated* EMA data by building a system that maps acoustics to EMA information (using concurrently recorded acoustics and EMA) and then mapping the acoustic information recorded concurrently with facial data

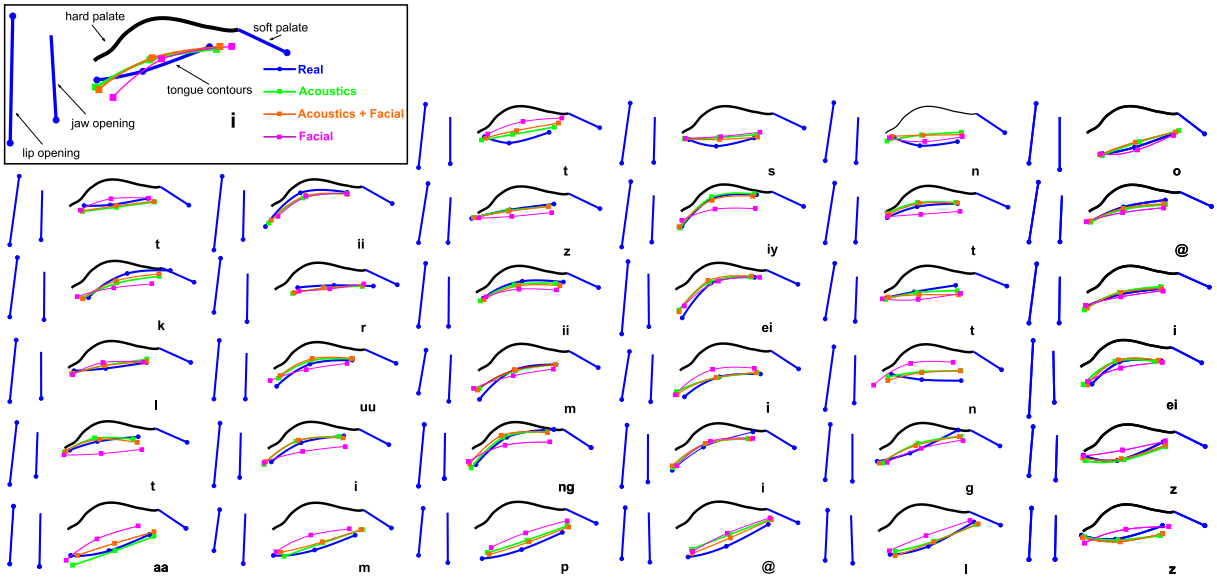



FIGURE 1.4 – Visualization of results for the phrase "It's not easy to create illuminating examples". Each subplot corresponds to approximately the middle of the phone duration, as indicated in Figure 1.3. For the tongue sensors, we show the real positions of the sensors and the three estimation. In each case the three sensors are shown connected by a simple cubic spline. For the sensors on lip, jaw, and velum, we show only real positions. The leftmost straight line approximates the lip opening (connecting upper and lower lip sensor) and the second straight line jaw opening (connecting the constant position of upper incisor sensor to that of lower incisor sensor). The straight line connecting the rightmost point of the hard palate with the velum sensor position approximates very roughly the soft palate. The shape of the hard palate was approximated from the full fsew0 dataset. RMS errors and Pearson correlations between real and estimated positions for the full utterance are given in parentheses in Table 1.1. The MOCHA labeling convention is used for the names of the phonemes.

to EMA information. This approach may augment the audiovisual corpus with data of the tongue.

After our experiments, and, most notably, after observing animations such as the one shown in Figure 1.4 we believe that, even *without* adding facial information, our acoustic-to-EMA mapping system is able to provide synthetic tongue trajectories useful for our purpose, i.e. a tongue animation that is intelligible to the interlocutor of the talking head. In fact, in the example of the Figure 1.4, many phoneme targets can be considered as acceptable, which might be considered the minimum that an inversion system should provide when used in language learning for instance. However, based on the shape of the tongue, it seems that depending on the method used, it is possible to have slightly different trajectories, that we cannot assume that they are equivalent. We still consider that the accuracy of the retrieved trajectories is highly important in sensitive applications, like providing articulatory feedback for speech training [Levitt and Katz, 2010] or driving an articulatory model for the purposes of articulatory synthesis [Toutios et al., 2011c].

## 1.6 Episodic memory-based inversion

We proposed an inversion method that aims to overcome the lack of a good representation of dynamics. This approach is based on the concept of episodic memory [Tulving, 1972]. An

episodic memory can be considered to be a codebook that includes a temporal dimension. While a codebook models the relationship between two static observations, an episodic memory can model the relationship between two sequences of observations. The main advantage of episodic memory is that it keeps track of the order of the observations and thus preserves the acoustic and articulatory dynamics of each episode. An episodic memory can deal with the non-linearity of the mapping function by using the one-to-one correspondence between the synchronized acoustic and articulatory observations. It can deal with the non-uniqueness of the mapping by exploiting the articulatory dynamics encoded through the time ordering of both the episodes and the observations. In addition, the need to smooth inferred articulator trajectories is greatly reduced compared to current codebook-based inversion methods.

In the following, I present an overview and the main idea behind the memory-based method. This work was done with Sebastien Demange and the complete description and the details of the results can be found in [Demange and Ouni, 2013].

### 1.6.1 Episodic modeling

We apply an episodic memory model to the acoustic-to-articulatory mapping problem, whereby episodes comprise the synchronized acoustic and articulatory feature sequences for each linguistic unit. Then, for inversion, the memory could exploit precisely the acoustic-articulatory relationship, as well as real acoustic and articulatory dynamics.

In speech inversion, the mapping is between two continuous spaces. This is an important issue for an episodic memory model because the memories need to contain many episodes of each linguistic unit in order to achieve adequate coverage of the variability present in speech.

For this reason, we propose to provide the memory with a mechanism to simulate many more episodes than the ones it contains. We define an episode as synchronized acoustic and articulatory realizations of a linguistic unit ($LU$). Let us consider two episodes $X$ and $Y$ of a given linguistic unit. $X$ and $Y$ are almost identical, differing only at the beginning and end, due to coarticulation effects. A C-Mem which does not contain any episode of $LU$ whose left and right contexts are those of $X$ and $Y$, respectively, will invariably fail to invert acoustic realizations of $LU$ in this context. However, the memory could perform better if it were allowed to go through the first part of $X$, then to switch to $Y$ at any time during the central, nearly identical part, and finally to go through the final part of $Y$.

Even though all episodes of a given linguistic unit are not identical, they can exhibit local articulatory similarities. Therefore, we propose to allow the memory to switch between any two episodes $X$ and $Y$ of the same linguistic unit during the inversion at times when observations of $X$ and $Y$ are similar. Care will be taken to produce realistic articulatory dynamics by considering similarities with regard to temporal alignments and with regard to the positions of the articulators. As the proposed memory will be able to generate episodes other than the ones in the database, we will refer to this memory as a generative memory (G-Mem).

### 1.6.2 Generative memory

We define an episode as synchronized acoustic and articulatory realizations of a particular linguistic unit, in this case, phonemes. The phoneme identity will be referred to as the class of the episode. In the following, we consider two given episodes $X$ and $Y$. $X$ is a particular realization of a given phoneme expressed as a sequence of $K$ articulatory-acoustic observations $X = (x_1, \ldots, x_K)$. Each observation $x_i = (x_i^{art}, x_i^{ac})$, where $i \in [1..N]$, is a synchronized pair composed of an articulatory observation $x_i^{art}$ and its corresponding acoustic observation $x_i^{ac}$.

The scalar articulatory observation $x_i^{art}$ can be a given articulator description or even the x- or y-coordinate of a sensor glued onto an articulator as used in our work. Similarly, $Y$ is another realization of the same phoneme expressed as a sequence of $N$ observations: $Y = (y_1, \ldots, y_N)$, where $y_i = (y_i^{art}, y_i^{ac})$.

**Local articulatory similarity**

Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978] is a general algorithm to find the shortest distance $D(X, Y)$ between two episodes $X$ and $Y$, which may vary in length. The episodes are warped non-linearly in order to minimize the effects of their temporal variability. Any given mapping leads to a particular alignment path $\Phi = (\Phi_1, \ldots, \Phi_M)$, where $M$ is the number of alignments. $\Phi_i = (\Phi_{x,i}, \Phi_{y,i})$ is the $i^{th}$ observation pairing along $\Phi$ with $\Phi_{x,i}$ and $\Phi_{y,i}$ as the indices in $X$ and $Y$ of the aligned observations. The distance between $X$ and $Y$ given $\Phi$ is the sum of the local distances $d(.,.)$ between the aligned observations along $\Phi$. The choice of this local distance $d(.,.)$ depends on the nature of the observations used to perform the mapping. Here, the articulatory observations are used because we are focusing on producing realistic articulatory trajectories. In our work, the articulatory observations are the x- and y-coordinates of sensors glued onto articulators in the midsagittal plane. Thus, Euclidean distance is used.

$D(X, Y)$ is the shortest distance over all $\Phi$:

$$D(X, Y) = \arg\min_{\Phi} \sum_{i=1}^{M} d(x_{\Phi_{x,i}}, y_{\Phi_{y,i}}) \tag{1.8}$$

Many variations of the algorithm have been proposed [Sakoe and Chiba, 1978] in order to prevent degenerate paths from occurring. In this work, we applied Itakura's constraints [Itakura, 1975]. These constraints make the DTW asymmetric, such that each observation in sequence $X$ is aligned with exactly one observation in $Y$. The mappings of many episodes onto $X$ result in many alignment paths of the same length (equal to the length of $X$). Therefore, the distances from these episodes to $X$ can be fairly compared and ranked. The other Itakura constraints impose bounds on the temporal deformation, in order to preserve a certain temporal consistency of the aligned observations.

Let $X^{art}$ be an articulatory trajectory of an episode $X$ expressed as a sequence of $K$ articulatory positions $(x_1^{art}, x_2^{art}, \ldots, x_K^{art})$. We define each articulatory observation $x_{i+1}^{art}$ as the natural articulatory target (local target) of $x_i^{art}$ since it has been observed to follow $x_i^{art}$. In fact, $x_{i+1}^{art}$ is a specific articulatory position, but we can suppose it could have been slightly different. Indeed, starting from $x_i^{art}$ at time $i$, the articulators could have reached a different position at time $i+1$ close to $x_{i+1}^{art}$ with no significant consequences to the acoustics. Then, for each $x_i$ we define an articulatory target interval $ATI_{x_i}$ as:

$$ATI_{x_i} = [x_{i+1}^{art} - \delta, x_{i+1}^{art} + \delta] \tag{1.9}$$

where $\delta$ is a given positive value.

We consider any articulatory position $y_j^{art}$ to be similar to any articulatory position $x_i^{art}$, if $y_j$ is aligned with $x_i$ when mapping $Y$ onto $X$, and if $y_j^{art}$ belongs to the articulatory target interval of $x_{i-1}^{art}$.

### Building the generative memory

We model a G-Mem as an oriented graph $\mathcal{G}_{G-Mem}$. The nodes are the pairs of synchronized acoustic and articulatory observations comprising the episodes. The oriented edges indicate the allowed transitions between the articulatory positions the memory can follow during the inversion process. They are created according to Algorithm 1.

For any pair of episodes $X$ and $Y$ of the same class we create an oriented edge from a given $x_i$ to a given $y_j$ if $y_j$ is similar to $x_{i+1}$ from an articulatory point of view (i.e, $y_j^{art}$ is similar to $x_{i+1}^{art}$) as defined previously:

$$\Phi_{y,i+1} = j \tag{1.10}$$

$$y_j^{art} \in ATI_{x_i} = [x_{i+1}^{art} - \delta, x_{i+1}^{art} + \delta] \tag{1.11}$$

In addition we impose that the asymmetric distance $D(X,Y)$ falls below a given threshold (proportional to the length of $X$):

$$D(X,Y) \leq \Delta * length(X) \tag{1.12}$$

The goal is to prohibit the memory from switching between episodes, which are globally very different (from an articulatory point of view) because it could lead the memory to produce unrealistic articulatory trajectories. As an example, consider the movements of the tongue tip, which might rise or fall during the production of a given phoneme. Although these trajectories are very different, it is likely that the tongue tip can reach similar positions midway through the fall and rise. Combining the fall and rise could possibly lead to a degenerate trajectory.

At the episode boundaries the memory is only subject to the articulatory continuity requirement expressed by equation (1.11). Let $Z = (z_1, z_2, \ldots, z_P)$ be the episode, which was observed after $X$. Then, an edge from $x_K$ to the first observation $w_1$ of any episode $W$ of any class is created if $w_1^{art} \in ATI_{x_K} = [z_1^{art} - \delta, z_1^{art} + \delta]$. If the episode $X$ is the last of a record, its natural articulatory target is unknown and equation (1.11) cannot be satisfied; thus no edge to any other episode is possible. Note that a C-Mem only accounts for these transitions between episode boundaries. So, a C-Mem can be seen as a particular case of G-Mem for which $\Delta$ is set to zero.

### Recovering articulatory trajectories

As the nodes of the oriented graph $\mathcal{G}_{G-Mem}$ are bimodal (composed of an acoustic and an articulatory observation), each path within the graph corresponds to a particular articulatory trajectory and also to the acoustics that would have been produced by the articulatory trajectory. Thus, acoustic-to-articulatory inversion is performed by searching the path within $\mathcal{G}_{G-Mem}$ that best matches the acoustic speech signal to be inverted. The estimated articulatory trajectory is extracted from the articulatory observations of the visited nodes along this path.

All search paths can start only at nodes that represent the first observation of an episode. During inversion a breadth-first search is performed, applying the Viterbi algorithm. That is, at each step, the $K$-best paths obtained at the previous step are propagated through the $\mathcal{G}_{G-Mem}$ along the oriented edges defining allowed articulatory movements. The $K$-best propagated paths are kept while the others are discarded and the process is repeated up to the end of the speech signal.

The winning path is the one with the best acoustic score selected from all paths ending at a node that corresponds to the final observation of an episode. The score of each path is expressed as the sum of acoustic distances between the speech frames and the acoustic observations of the visited nodes along the path, computed on a predefined acoustic window $\mathcal{W}$.

TABLE 1.2 – Overview of the corpora.

| Corpora | Sets | Durations | Sentences | Phones |
|---|---|---|---|---|
| *fsew* | train | 16 min 35 sec | 368 | 11179 |
| | dev | 1 min 57 sec | 46 | 1324 |
| | test | 2 min 5 sec | 46 | 1457 |
| *msak* | train | 13 min 59 sec | 368 | 11179 |
| | dev | 1 min 41 sec | 46 | 1324 |
| | test | 1 min 45 sec | 46 | 1457 |
| *mdem* | train | 8 min 24 sec | 319 | 6355 |
| | dev | 1 min 2 sec | 40 | 817 |
| | test | 1 min 3 sec | 40 | 814 |

### 1.6.3 Inversion experiments

#### Corpora

All the experiments presented in this work were carried out on the following two corpora of synchronized acoustic speech signal and articulatory trajectories. We used the EMA corpus of MOCHA [Wrench and Hardcastle, 2000] (see Figure 1.2) and we recorded a French EMA corpus *mdem* using an articulograph (AG500, Carstens Medizinelektronik). A male French speaker (*mdem*) was recorded uttering 400 phonetically balanced sentences. The audio is provided as waveforms sampled at 16 kHz, and the EMA data consist of 2D coordinates in the mid-sagittal plane. We used 6 sensors fixed in the mid-sagittal plane on the lower lip (ll), upper lip (ul), tongue tip (tt), tongue body (tb), tongue dorsum (td), and tongue back dorsum (tbd). The phonetic segmentation was obtained from a word-level transcription of the sentences, a dictionary containing several pronunciation variants for each word and a set of French monophone HMMs trained on several hours of speech and adapted to *mdem*'s voice. The segmentation was obtained by force aligning the phone HMMs onto the acoustics given the sentence word transcription and the pronunciation dictionary.

Each corpus was split into training, development and test sets. For MOCHA, care was taken that the selected utterances for each set corresponded exactly to the ones used by [Richmond, 2002], as this split was also used in [Toutios and Margaritis, 2008] and [Zhang and Renals, 2008]. Information about the different sets are given in table 1.2. Note that the durations only account for usable speech (without the start and end silences). Figure 1.5 shows the distributions of the articulatory samples for each speaker and coil.

#### Experiment design

We have implemented three different inversion methods: a codebook-based approach, as described in [Suzuki et al., 1998], and two memory-based approaches (C-Mem and G-Mem) as described above. We have chosen to compare the memory-based approaches to this codebook-based approach, as they differ only in the manner of inferring the dynamics of the recovered articulatory trajectories. While the memories can model and use observed trajectories, the codebook only relies on continuity constraints.

In fact, this codebook method consists in looking up a set $\Gamma_i$ of the $N$ best entries in the codebook for each given acoustic signal sample $Y_i$. The best entries are the ones, which minimize the acoustic distance to $Y_i$. In fact, this distance is the average acoustic distance computed over

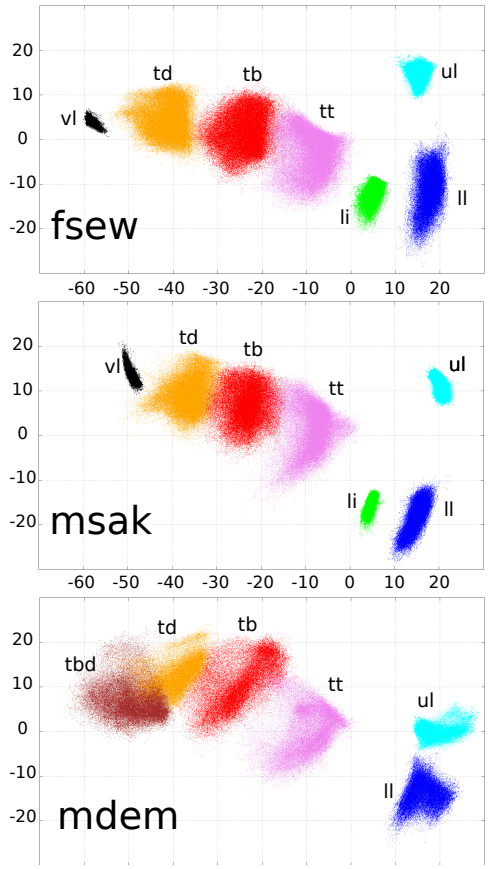



FIGURE 1.5 – EMA data distribution. The three graphs represent the articulatory data distributions for the three speakers: *fsew*, *msak* and *mdem*. The coordinates are expressed in millimeters.

a window of a predefined length $\mathcal{W}$. The articulatory trajectory is obtained by looking for the path through $\Gamma_i$ that minimizes the weighted sum of the acoustic distances and squared distances between subsequent articulatory parameters.

We consider the recovery of the articulatory trajectories for each coil, and along both the $x$ and $y$ axes, as independent inversion problems. Thus, the experiments presented here consist of fourteen (for *fsew* and *msak*) and twelve (for *mdem*) distinct inversion problems. For each inversion problem, a dedicated codebook, C-Mem and G-Mem are built and optimized.

The parameters of the codebook are the length of the spectral window, and the weight of the articulatory constraints with regard to the acoustic distances. During the inversion, the 1000 best codebook entries are considered at any given time. For G-Mem, the parameters to be set are $\delta$, the half ATI length, and $\Delta$, the maximum allowed articulatory distance between two episodes $X$ and $Y$ for allowing the memory to switch from one to the other. Note that the parameter $\Delta$ is equal to zero for a C-Mem, so that each episode can only be combined with itself. An acoustic window was used to compute the acoustic distances. The length of this window is a parameter to be optimized. Euclidean distance is used for both acoustic and articulatory distance calculation.

Two types of experiment have been carried out: with and without phonetic constraints. Note that the G-Mem used phonetic segmentation only for building the memory, and the segmentation is not needed for inversion. The purpose of adding experiments with phonetic constraints is to ascertain what extra information is added for use of the phonetic knowledge to the different

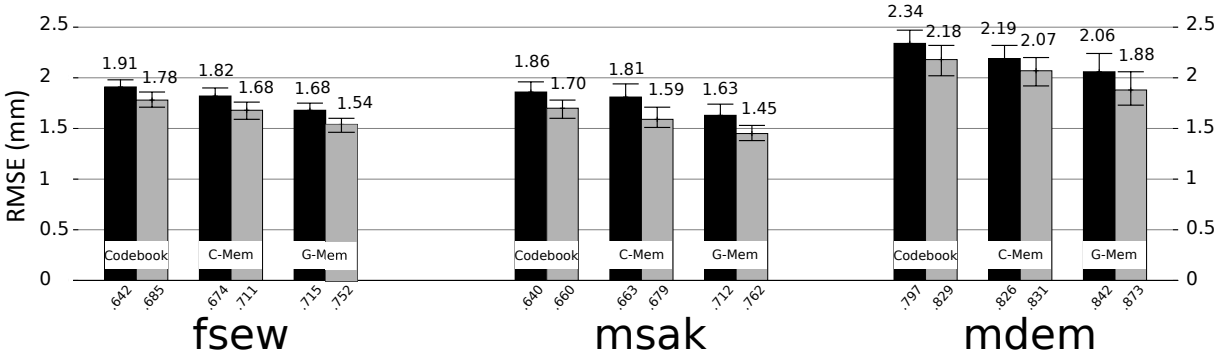



FIGURE 1.6 – Overall RMSE (in millimeters) for each corpus using the Codebook, C-Mem, and G-Mem based methods with (grey) and without (black) the reference phonetic segmentations. The error bars represent the 99% confidence intervals. The RMSE values are indicated above the bars. For each experiment, Pearson's correlations are provided below the bars.

inversion methods. The reference phonetic segmentation is obtained by force aligning HMM acoustic models onto the speech signal according to the transcription of the uttered words. In the phonetically constrained mode, the estimated articulatory position at time $t$ has to come from a codebook entry or an episode of the same phoneme as the one indicated by the segmentation at time $t$.

For the three systems, the parameters have been jointly optimized through a grid search optimizing RMSE on the development set. Since the articulators move in different ranges, different sets of parameters were obtained for all inversion problems. Globally, the length of the acoustic window $\mathcal{W}$ is approximately 150 ms for the codebook and 90 ms for the memories. $\delta$ ranges from a few hundredths of a millimeter (for the velum) to at most one millimeter, and $\Delta$ ranges from a tenth of a millimeter to two millimeters for different articulators. Finally, the memory search beam width is set to 10 000 for the memories and 1000 for the codebook.

### 1.6.4 Inversion results

Figure 1.6 shows the results of the three inversion methods, with (grey) and without (black) phonetic constraints. The bars show the overall RMSE (means over the coils and $x$- and $y$-coordinates) in millimeters for each corpus; the RMSE values are indicated above the bars. The respective Pearson's correlations are given below the bars. The figure shows that the memory-based approaches always outperform the codebook-based method. This suggests the articulatory dynamics are modeled more effectively in the memory-based systems than by continuity constraints used by the codebook for speech inversion. It also illustrates how much an episodic memory can benefit from the generalization capability of the G-Mem, as it always outperforms the C-Mem. The best Pearson's correlation scores were obtained using a G-Mem. Without any phonetic knowledge, an overall RMSE of 1.65 mm and a correlation of .714 were obtained on MOCHA with the proposed G-Mem, while an RMSE of 1.81 mm and 1.88 mm, and a correlation of .668 and .641 were obtained with the C-Mem and codebook, respectively. Using the phonetic segmentation of the test recordings, the RMSE decreased to 1.50 mm and the correlation increased to .757 for the G-Mem. The relative improvements due to the phonetic segmentation of the acoustic signal was roughly the same across the three inversion methods and across the three speakers.

Figure 1.7 shows the tongue tip movements (thick curves) along the vertical (up/down) axis

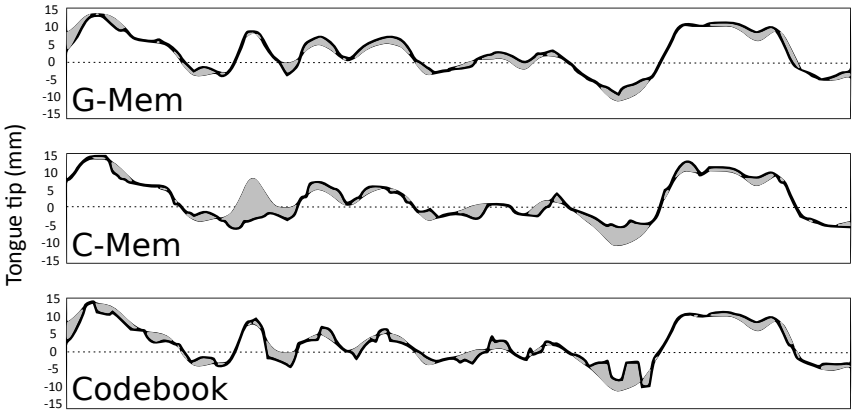



FIGURE 1.7 – Tongue tip movements (thick curves) along the vertical (up/down) axis recovered by each of the three approaches from the two second long French utterance "juste quelques extrémités de branches gelées" (*only a few frozen branch tips*) uttered by the speaker *mdem*. For each of the three graphs the reference trajectory is provided as the thin curve and the estimation errors are emphasized with the filled areas between the estimated and reference trajectories.

recovered by each of the three approaches from a two second long speech signal corresponding to the French sentence "juste quelques extrémités de branches gelées" (*only a few frozen branch tips*). For each of the three graphs, the reference trajectory is provided as the thin curve and the estimation errors are emphasized as filled areas between the estimated and reference trajectories. Though optimized, the dynamic articulatory constraints of the codebook-based approach do not prevent discontinuities, as the recovered articulatory trajectory is very jerky compared to the smooth and continuously varying reference trajectory. The results obtained by both the C-Mem and G-Mem are visibly better. One might have expected smoother trajectories using the C-Mem, as the result is expressed as a concatenation of natural articulatory episodes. Through deeper analysis of the decoding paths within the C-Mem, we have noticed that most of the time the C-Mem does not contain episodes, which acoustically match the test signal well enough. In order to counteract this lack of good episodes, the C-Mem tends to select many short episodes. Indeed, the sum of the acoustic distances of short episodes, which locally match the test signal well, is usually smaller than the acoustic distance of a longer episode with partial acoustic mismatches. Finally, the G-Mem succeeds in estimating the articulatory movements accurately. The combination of episodes significantly reduces the estimation error. Furthermore, the articulatory dynamics embodied in the G-Mem transition graph contribute to the naturalness of the resulting articulatory movements.

**Smoothness and naturalness of articulatory trajectories**

Many proposed solutions to the inversion problem include low-pass filtering the recovered articulatory trajectories. This filtering aims to remove rapid changes (usually of low amplitude) that come from errors during the acoustic-to-articulatory mapping, and that can be considered noise. Indeed, all articulators have their proper velocity and can move more or less quickly, but they all move continuously. Comparing the recovered trajectories with their smoothed version is therefore one possible way of assessing how well the articulatory dynamics have been approximated. For each coil, in both directions, we have determined the best cut-off frequency minimizing the RMSE on the test set. Figure 1.8 summarizes the best case scenario smoothing results. As in Figure 1.6, the bars represent the RMSE, and Pearson's correlations are provided below the bars.

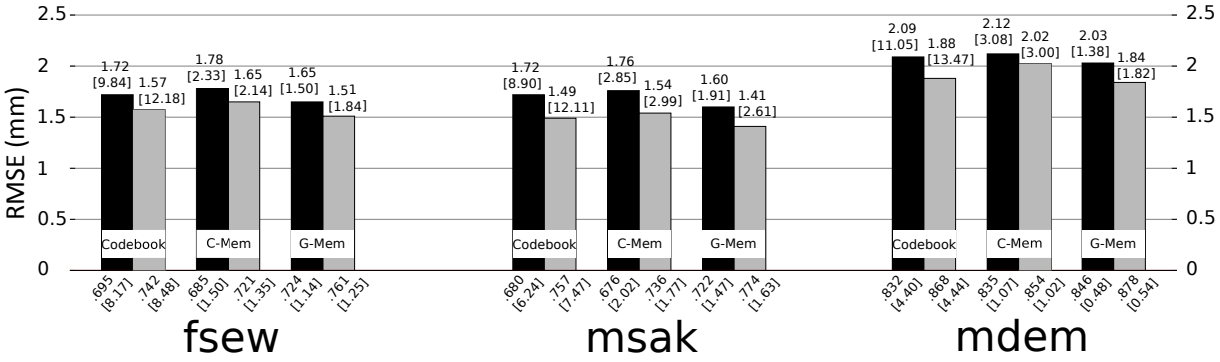



FIGURE 1.8 – Overall RMSE (in millimeters) for each corpus using the Codebook, C-Mem, and G-Mem based methods with (grey) and without (black) the reference phonetic segmentations (as in Figure 1.6) after optimal smoothing of the recovered articulatory trajectories. The RMSE values are indicated above the bars. For each experiment, Pearson's correlations are provided below the bars. The numbers in square brackets are the relative percentage improvements over the non-smoothed trajectories.

The numbers in square brackets are the relative percentage improvements over the non-smoothed trajectories.

The most obvious effect is that the codebook-based method significantly benefits from this filtering. Its RMSE improves by approximately 10%, while the improvements for the memory-based approaches do not exceed 3%. The same trend can be observed for the Pearson's correlations. This indicates that the memory-based approaches really do take advantage of the observed dynamics of the episodes. Applying articulatory continuity constraints during the inversion does not yield the same benefits. Note that similar observations have been reported in [Toda et al., 2008]. Indeed, the authors proposed an MLE-based mapping that accounts for correlation between frames. They reported significant improvements over their baseline GMM-based mapping, but also showed that the low-pass filtering effect became negligible. A relative improvement of 9.36% for RMSE was obtained over the baseline GMM-based mapping with a low-pass filter, but only .72% over the MLE-based mapping.

## 1.7 Inversion, evaluation, application

The estimation errors using a G-Mem are very encouraging compared with the state of the art. [Hiroya and Honda, 2004] reported RMSEs of 1.50 and 1.73 mm with and without phonetic segmentations, respectively, using a HMM-based production model. However, we cannot directly compare our results with theirs, as they used a Japanese database. On MOCHA, [Toda et al., 2008] used Gaussian mixture models to map the acoustic space onto the articulatory space. They reduced the RMSE from 1.58 to 1.40 mm by applying a maximum likelihood estimation (MLE) of the dynamic features. [Zhang and Renals, 2008] obtained an RMSE of 1.71 mm using a trajectory HMM. They included velocity features in their acoustic front end and performed speech recognition prior to inversion to provide their system with a phonetic segmentation. Even without phonetic segmentation, the G-Mem performs slightly better. [Al Moubayed and Ananthakrishnan, 2010] proposed a memory-based method. They used a linear regression on the local neighborhood of the codebook entries to map the acoustic input frames onto the articulatory space. They also used MLE of the dynamic features to improve the results. An RMSE of 1.52 mm was reported using this method. Finally, [Richmond, 2006] reported an RMSE of 1.40 mm using trajectory mixture density neural networks. We share with

[Zhang and Renals, 2008], [Richmond, 2006] and [Al Moubayed and Ananthakrishnan, 2010] the same train, dev and test sets and all reported RMS errors range from 1.40 to 1.73 mm on this corpus.

Recently Richmond applied the exact same methodology to MOCHA and to a newly acquired dataset and found a decrease of root mean squared error from 1.54 mm to 0.99 mm [Richmond, 2009]. This means that the quality of the data can influence drastically the results and probably that RMSE is not a very good metric.

I believe that using RMSE as a unique performance measure is not sufficient. For this reason, I advocate for proposing another measure which provides an accurate evaluation, probably at the phoneme levels, instead of a global score. In fact, some errors for some articulators during phoneme uttering may be less critical than others. For instance, in the case of bilabial phonemes, the complete closure is critical and should be highly penalized in a given evaluation metric. It is worth noting that [Ananthakrishnan and Engwall, 2011a] proposed an automatic method for segmenting the articulatory movements into articulatory gestures. This method accounts for the notion of critical articulators. That is, the production of a particular phoneme depends mainly on the movements of few articulators (the critical articulators), which have to reach precise articulatory target positions. The other articulators can move more freely and thus can anticipate the production of the next phoneme to be uttered. Then, the movements of the articulators overlap with each other and contribute to coarticulate all the uttered sounds. I believe that this is a good idea that can be considered to define probably an evaluation metric based on the notion of critical articulators.

Using acoustic-to-articulatory inversion techniques as a visual feedback is pronunciation training or in speech therapy is very promising practical application. However, the accuracy of the inversion is highly important to provide the good feedback.

## 1.8 Summary and Contribution

### 1.8.1 Summary

In this chapter, I presented my contribution in the field of articulatory-to-acoustic inversion, i.e., how to estimate the vocal tract shape from a speech signal. I presented three different approachs to address this problem.

- **Model-based Inversion:** The articulatory-to-acoustic mapping has been addressed. We have proposed a representation of the articulatory space by a hypercube codebook. The inversion is performed using this codebook to provide a set of solutions combined with dynamic programming. We used a nonlinear smoothing algorithm together with a regularization technique to recover the best articulatory trajectory.
- **Multimodal Inversion - from acoustic-Face to Articulatory:** This inversion method is based on a statistical data-driven approach using support vector regression for the mapping from acoustic parameters to electromagnetic articulography (EMA) trajectories. The data is based on synchronized acoustic and articulatory data streams using EMA. We extended the acoustics with visual information. We were targeting the invisible articulators, mainly the tongue.
- **Episodic memory-based Inversion:** This inversion method is based on episodic memory. The main advantage of this representation is that the memory models the real articulatory dynamics as observed. We have proposed a memory which is able to produce articulatory trajectories that do not belong to the set of episodes the memory is based on.

### 1.8.2 Perspectives

As shown above, several inversion methods provide similar performance when using RMSE. I argued that this measure is not sufficient. As future work, I propose to develop a metric that provides an accurate evaluation at the phoneme level for instance, to replace the RMSE global score. This metric will be adapted to the articulatory context. When the evaluation metric is improved and the inversion methods are very accurate, it is possible to envision using acoustic-to-articulatory inversion techniques as a visual feedback in pronunciation training or in speech therapy, where such inversion method can pilot a talking head from a speech signal. This actually one of the motivations behind research works on inversion.

### 1.8.3 Related projects and contributions

- Supervised **Asterios Toutios**, during his postdoc stay. He worked on articulatory model adaptation using EMA (co-supervised with Yves Laprie) , and on a statistical data-driven inversion method using as input acoustic and facial information.
- Supervised **Sebastien Demange**, during his postdoc stay (ATER). He worked on inversion using episodic memories with EMA data.
- Participated in supervising **Imen Jemaa**, during her Ph.D studies (joint supervision "cotutelle"). She worked on formant tracking, highly important when dealing with model-based inversion. Acoustic features are formants, that is difficult to track in some cases, and thus, they may impact the quality of the inversion.
- Supervised **Jonathan Demange**, during his master project. He worked on fitting ultrasound data to an articulatory model to evaluate inversion.
- Supervised **Mathew Wilson**, during his master project. He worked on fitting EMA data to an articulatory model.

- **Participation in the ANR Project (2009-2013): ARTIS** - Articulatory inversion from audiovisual speech for augmented speech (Collaboration LORIA, GIPSA, ParisTech)
- **Participation in Regional Project (2007-2009): MODAP** - Articulatory modeling of Speech (collaboration LORIA - Supelec)

### 1.8.4 Selection of related publications


- Demange, S. and Ouni, S. (2013). An episodic memory-based solution for the acoustic-to-articulatory inversion problem. *The Journal of the Acoustical Society of America*, 133(5):2921–2930
- Toutios, A., Ouni, S., et al. (2011b). Predicting tongue positions from acoustics and facial features. In *Interspeech 2011*, pages 2661–2664, Florence, Italy
- Toutios, A., Ouni, S., and Laprie, Y. (2011c). Estimating the Control parameters of an Articulatory Model from Electromagnetic Articulograph Data. *The Journal of the Acoustical Society of America*, 129(5):3245–3257
- Demange, S., Ouni, S., et al. (2011). Continuous episodic memory based speech recognition using articulatory dynamics. *Proceedings of Interspeech, Florence, Italy*, pages 2305–2308
- Toutios, A., Ouni, S., and Laprie, Y. (2008). Protocol for a model-based evaluation of a dynamic acoustic-to-articulatory inversion method using electromagnetic articulography. In *The eighth International Seminar on Speech Production (ISSP'08)*, pages 317–320, Strasbourg, France
- Potard, B., Laprie, Y., and Ouni, S. (2008). Incorporation of phonetic constraints in acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 123(4):2310–2323
- Ouni, S. and Laprie, Y. (2005b). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 118(1):444–460
- Ouni, S. (2005). Can we retrieve vocal tract dynamics that produced speech? toward a speaker articulatory strategy model. *Interspeech 2005-Eurospeech*, pages 1037–1040

# Entr'acte

# 2

# Data acquisition and processing

## 2.1 Introduction

Over the past years, techniques for the acquisition of articulatory and visual data have been maturing steadily. Recent developments have enabled significant improvements in data acquisition, practically revolutionizing the foundations of research in speech, with increasing impact on related fields, such as speech technology, clinical assessment of speech, language learning, etc. Electromagnetic Articulography (EMA) captures articulatory movements in three dimensions (3D) with a high temporal resolution by tracking tiny sensors attached to speech articulators such as the tongue, teeth, and lips. Magnetic Resonance Imaging (MRI) is a non-invasive, hazard-free medical imaging technique allowing high-resolution scans of the vocal tract. Several other modalities, such as ultrasound tongue imaging, electropalatography, electroglottography, video recording (conventional, stereo-vision), optical motion capture, air flow and pressure measurements, are also in active use.

Due to their sheer diversity and significant technical overhead, using these techniques can be a time-consuming and costly endeavor, both in terms of the facilities needed and the effort required to carry out the acquisition and subsequent processing of the data. These factors have so far limited the widespread exploitation of articulatory data where they would offer added value, such as in related clinical research fields or technological applications.

The nature of my research makes me highly concerned by acquisition and processing issues. In fact, in my work, I use articulatory data, visual data and acoustic data. For articulatory data, I'm mainly working with EMA data using an articulograph, and to lesser extent, I am interested in MRI and ultrasound data. These data are used in multimodal audiovisual speech synthesis, speech production and speech inversion. Thanks to a lengthy work of several months, I built a strong experience with the acquisition of articulatory data, and I am continuously working on:

– determining optimal protocols and guidelines for acquiring articulatory data (e.g., placement of sensors for EMA), which at the same time allow diverse research questions to be addressed;

– dealing with the question of synchronizing and merging articulatory data from different acquisition sessions, different modalities and eventually different speakers: defining how to reduce the heterogeneity of the data;

– addressing recurrent issues with acoustic recordings during articulatory data acquisition (e.g., electromagnetic interference, perturbation of articulation due to the intrusion of sensors, or unnatural acquisition postures);

During the last five years, I tried to be more active in the community to improve techniques and to exchange experiences in this field. For instance, I proposed a special session on "articulatory data acquisition and processing" during Interspeech 2013, that was accepted and that attracted several participants (exceptionally, the session was organized on two-session long). In addition, I'm continuously encouraging the use and distribution of software processing tools to ease the exploitation of articulatory data and resources. In particular, I developed a powerful tool, called *V*isArtico, an articulatory visualization software that is intended to visualize articulatory data acquired using an articulograph. It allows visualizing the tongue shape, the lips, the jaw and the palate. It allows animating the vocal tract synchronously with audio and in real-time. It is proposed for researchers that need to visualize and analyze the data acquired from the articulograph with no excessive processing. The software is a cross-platform software. I conducted the design and supervised the development. This software was published and freely distributed to academic researchers.

In this chapter, I present briefly *V*isArtico, the visualization software, followed by some comments on multimodal data acquisition using EMA and motion capture.

## 2.2 VisArtico: Articulatory visualization

### 2.2.1 Introduction

A small collection of programs are already available to inspect and visualize EMA data, including EMATOOLS [Nguyen, 2000], MVIEW [Tiede, 2010], Carstens *JustView* [Carstens Medizinelektronik, 2006], and a few others. Unfortunately, some of the existing software tools are no longer maintained or even available for download. Others can be used only on computers running certain versions of Windows, or require a (fairly expensive) license of the commercial MATLAB computing and simulation platform. The latter prerequisite involves significant initial effort with a steep learning curve, and can intimidate or frustrate the non-technically minded user who is interested only in analyzing articulatory speech data.

For these reasons, I have conducted the design and supervised the development of *V*isArtico, a lightweight, easy-to-use software tool which allows visualizing EMA data, and which can be run on any computer that supports Java. The software has been designed so that it can directly use the data provided by the articulograph [Carstens Medizinelektronik, 2004] to display the articulatory coil trajectories, synchronized with the corresponding acoustic recordings. Moreover, *V*isArtico not only allows viewing the coils but also enriches the visual information by indicating clearly and graphically the data for the tongue, lips and jaw, and offers some advanced functionality. In the following sections, we describe the software and its main features.

### 2.2.2 Presentation

*VisArtico* is a user-friendly software which allows visualizing EMA data. We recall that EMA acquisition technique using an articulograph tracks the positions of small electromagnetic coils attached to the speech articulators. The positions and orientations of these coils are calculated by

measuring the electrical currents produced within multiple low-intensity electromagnetic fields. This technique is known to present no risk to the health of the speaker.

This visualization software has been designed so that it can directly use the data provided by the articulograph to display the articulatory coil trajectories, synchronized with the corresponding acoustic recordings. Moreover, *VisArtico* not only allows viewing the coils but also enriches the visual information by indicating clearly and graphically the data for the tongue, lips and jaw, and offers some advanced functionality.

The first announcement of the software was at Interspeech 2012 [Ouni et al., 2012] during the session "Speech Tools". We presented a first version that supports only Carstens AG500 articulograph. Participants showed interest in VisArtico and asked for additional features. For this reason, we improved the software in addition to the support of the NDI Wave articulograph. During Interspeech 2013, I present the latest improvements of *VisArtico* and taking this opportunity to exchange with attendees and getting their feedback about the software.

### 2.2.3 *VisArtico* main features

The current version presents several new features and improvements. The main user-interface is presented in figure 2.1. The main features of *VisArtico* are the following:

- Supporting Carstens and NDI articulographs (AG500, AG501 and NDI Wave). These are currently the two existing family of articulographs widely used.
- Displaying the raw data. The basic fearure of VisArtico is to display the data as provided by the articulograph, which is the Cartesian coordinates $(x, y, z)$ and spherical coordinates $(\phi, \theta)$, as well as the reported RMSE for each EMA coil.
- Visualizing the tongue shape, the lips, the jaw and the palate. The tongue contour is shown as a spline interpolation through the coils on the tongue. a front view of the lips is also available to observe the degree of lip opening. The accuracy of the displayed shape of the lips depends on the number of corresponding coils (between two and four).The angular shape of the jaw is provided as an approximation.The palate contour can also be displayed if the palate trace data is available.
- Determining automatically the midsagittal plane of the speaker, allowing a better interpretation of the midsagittal view.
- Determining automatically the contour of the palate. When that information is not recorded, VisArtico can then calculate an approximation of the missing palate contour, using a simple but effective algorithm, which predicts the contour from the convex hull of the tongue coil positions.
- Providing several possible views: (1) temporal view, (2) 3D spatial view and (3) 2D midsagittal view
- Animating and playing the three views synchronously with audio and in realtime.
- Displaying different articulatory trajectories in addition to the acoustic signal (spectrogram) and eventually phonetic labels.
- Providing coil velocity and acceleration (displaying first and second derivative of a giving trajectory).
- Providing meaningful articulatory information.
- Displaying segmentation information at several levels (see figure 2.2. It is possible to create, read and modify a segmentation file in emphVisArtico. It is possible to display several levels of labeling (for instance, one for phonemes, one for words, etc.). Currently, *VisArtico* can read the *ali* format (Winsnorri), *textGrid* format (Praat) and *lab* format (x-lab). An additional simple format has been added *seg* (each line has: $start_timeend_timelabel$).

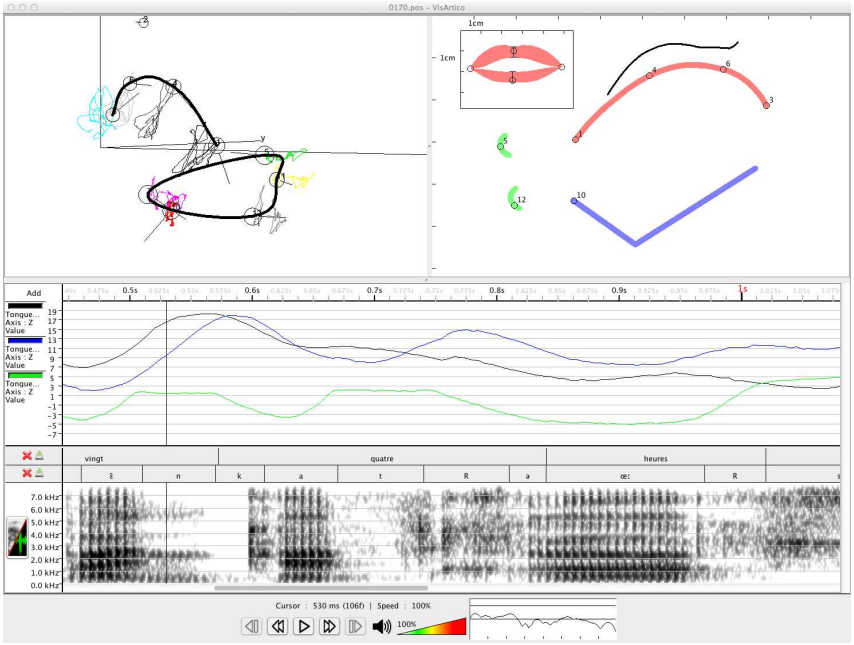



FIGURE 2.1 – Main user interface of *VisArtico*. Three different views: (1) the 3D-view (top left panel), (2) the midsagittal view (top right panel) with the tongue contour (red), jaw (blue), and lips (green, and in the front-view insert, red), (3) the temporal view is shown in the lower half of the window, displaying several channel trajectories, a phonetic segmentation, and the acoustic spectogram

- Improving the quality of the data by providing tools to filter out some outliers. It is possible to remove noise using low-pass FFT filter, or based on RMSE threshold.
- Using a playlist to make it easier to handle several files at the same time.
- Exporting graphic and text files of the data. The exported graphic is a vector-graphic file of the 2D view or the 3D view that can be customized. The text file is a selection of all or part of the data to export in readable format.

### 2.2.4 Present and future

Until today, 76 researchers from 16 countries have downloaded *VisArtico*. This interest shows that the community was waiting for such application. In fact, VisArtico is very useful for the speech science community and makes the use of articulatory data more accessible. It will expand the user base for electromagnetic articulography to researchers who are not necessarily computer experts. A dedicated website, `http://visartico.loria.fr/`, is online, where a presentation of *VisArtico* and detailed user guide are available. The software is written in java and, thus, is a cross-platform application (i.e., running under Windows, Linux and Mac OS). It is freely available to researchers and students. An Inria ADT (technological development action) project, has been accepted to make further improvements to make *VisArtico* very useful and widely used. For instance, we are planning the support of other EMA data formats and the addition of advanced data analysis and filtering.

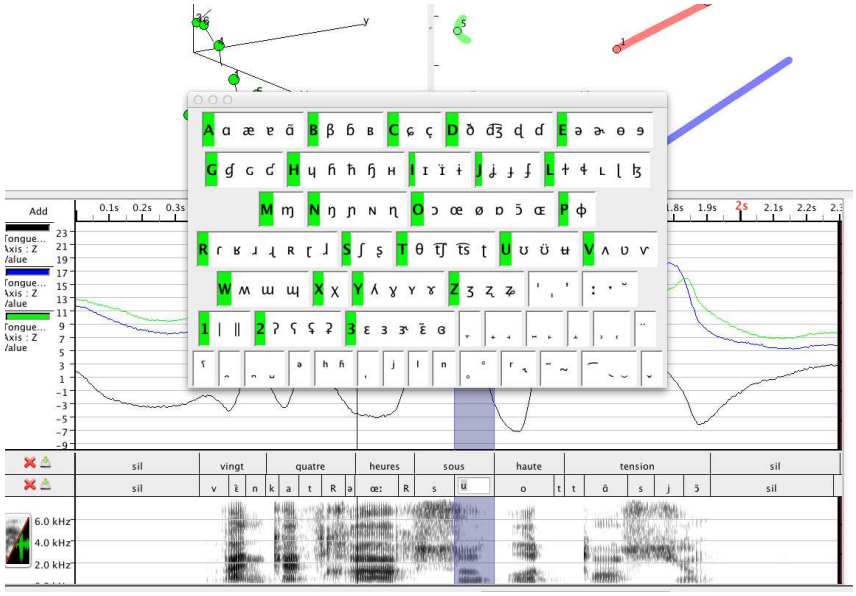



FIGURE 2.2 – Visualizing and editing phonetic labels. It is possible to display, edit or create segmentation files.

## 2.3 Multimodal acquisition techniques

I am led to use multimodal data in current and future work. The acquisition and processing the data is essential in my work related to articulatory-based speech production and audiovisual speech synthesis. In the following, I present how to consider the use of EMA data and motion capture in the context of multimodal data acquisition and processing. This strategy will be applied during my future work in this field.

### 2.3.1 EMA data

Currently, I'm using a 12 channel 3D-articulograph AG500 and in I will use in the near future the 24-channel AG501. Usually, the articulograph is used to track the movement of tongue, jaw, lips and head at a sampling rate of 200Hz (up to 400Hz) which gives very good temporal resolution. Typically, three sensors are used for head positions, four for lips, one for the jaw and four sensors were glued on the tongue. It should be noticed that in classical use of EMA, sensors are usually glued on the mid-sagittal plane of the tongue, where 2 to 4 sensors are glued on the middle of the tongue, from the tip to the back (see Figure 2.3). For instance, we conducted a study on pharyngealization where the sensors were located at distances of approximately 1.6, 3.6, 5.2 and 7cm from the tongue tip respectively in the midsagittal plane [Ouni and Laprie, 2009, Embarki et al., 2011b]. In practice, it is very difficult to glue more than 4 sensors on the mid-sagittal plane, because uttering speech becomes difficult and the acoustics will be adversely affected.

It is worth considering the 3D EMA to investigate and model 3D tongue and 3D lips. To have accurate behavior of the tongue and the lips, one needs to consider gluing sensors, as much as reasonably possible, on the tongue, and on the lips together. It is likewise considering the possibility of gluing up to 10 sensors on the lips, and 10 sensors on the tongue. This is

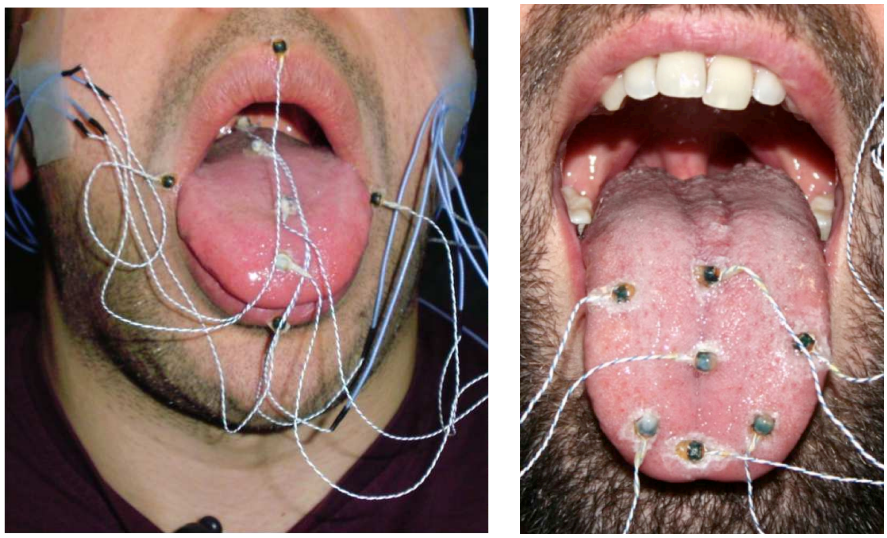


FIGURE 2.3 – EMA sensors topology. (left) Sensors are glued on the midsagittal plane of the tongue. In addition, 4 sensors are glued on the lips. (right) Sensors covering a part of the surface of the tongue, for the 3D-tongue modeling. Here, 7 sensors are glued on the tongue.

a sensitive task, as more sensors one has, more risk of detaching it, increases. Recording the tongue and lips simultaneously allows keeping a link between the internal articulatory gestures and the external ones, which may lead to extensive studies of the correlation between these two kind of gestures. A careful attention should be paid to the topology of the sensors on the tongue and lips, to capture speech gestures accurately. An advantage of 3D EMA, which counters some of the drawbacks of data sparsity, lies in the fact that along with the spatial positions of the sensors, their orientation is measured as well (thus it has 5 dimensional coordinates: 3D position + 2 angles). This additional information is well suited for transformational projection onto geometric structures larger than the sensors themselves. In addition, EMA is more efficient for lip tracking than using motion capture as the sensors glued on the lips can be tracked even when totally occluded. This is not the case with markers of motion capture. The EMA data will be post-processed to correct or remove any error related to the acquisition technique. In fact, EMA data may be noisy and thus should be filtered and smoothed. In some cases, when EMA sensors come into contact, which may happen for lips, the position may be wrong or absent and needs to be corrected or interpolated. In addition, the use of the angle orientation of the EMA sensor information, known as sensitive to noise, needs to be verified and corrected. The acoustics recorded during EMA acquisition is slightly degraded due to the presence of the sensors glued on the tongue. Post-processing methods to overcome this problem should be advised.

### 2.3.2 Motion Capture

Regarding the facial data, one can consider using a motion capture system (as Vicon or Qualisys, for instance) composed of 4 cameras controlled by motion capture software (as Nexus for Vicon and Track Manager for Qualisys). I already started testing the Vicon motion capture system to see how effective this system for such a task. The system allows tracking (1.5 to 3mm) reflective markers on the face. The software presents a user friendly interface to track and visualize the motion in real time. It is possible to synchronize audio or other sources with

the acquisition system. It allows processing and reconstructing the 3D data fast. It also offers the possibility to review, to correct and to tune the reconstructed and labeled 3D facial data. The motion capture data will be post-processed to correct or remove any error related to the acquisition technique. In fact, some markers may be occluded and may need tools to correct the tracking result. This is can be a lengthy work. When considering recording large corpora, the recording should be over several sessions. In fact, the arrangement of the markers on the face should be identical from session to session. It is worth considering this problem and see how to overcome it.

## 2.4 Conclusion

It is possible to consider other techniques using the more recent technological advances. For instance, some researchers started investigating the use of the kinect (or kinect-like system) in audiovisual speech recognition [Galatas et al., 2012]. The kinect is a motion sensing input device that can capture both video, as well as depth data at the same resolution. An IR camera, a laser and a structured light technique are used to acquire depth data. The captured depth is in the range 35 cm to 600 cm, where the precision decreases with the distance. It is a low-cost and very promising technique for the near future.

Articulatory data acquisition and processing is an important issue that should be carefully considered in the fields of speech production and audiovisual speech synthesis. The quality of the data can influence drastically the results. For instance, recently [Richmond, 2009] applied an inversion method to MOCHA and to a newly acquired dataset and found a decrease of root mean squared error from 1.54 mm to 0.99 mm [Richmond, 2009].

When dealing with multimodal data, two important problems have to be addressed: synchronization and integration. It is very often the case that each modality has its own sampling rate, that is not very accurate in some cases, and some modalities may have some delay relative to another modality, which need to be considered and incorporated in the calculation of the exact time frame for the different modalities. It is also the case that different modalities can have different spatial scale and reference. To integrate the data obtained from different modalities, it is necessary to use registration techniques.

The purpose of this chapter is to emphasis the importance of data acquisition and processing in the field of multimodal speech production and synthesis. Acquiring data and processing it can be time-consuming and costly in terms of the effort required to carry out the acquisition and processing the data. This effort is unavoidable to make more progress in modeling the processes related to human communication. Nevertheless, I am continuously working on improving the acquisition techniques and investigating methods to make the process easier.

## 2.5 Summary and Contribution

### 2.5.1 Summary

The purpose of this chapter is to emphasis the importance of data acquisition and processing in the field of multimodal speech production and synthesis. Acquiring data and processing it, can be time-consuming and costly in terms of the effort required to carry out the acquisition and processing the data. This effort is unavoidable to make more progress in modeling the processes related to human communication. In this chapter, I have presented my contribution in articulatory data acquisition techniques. I have proposed a powerful tool, called VisArtico, which is an articulatory visualization software that is intended to visualize articulatory data acquired using an articulograph. I built a strong experience in EMA acquisition techniques, and several corpora have been developed and used to study, for instance, coarticulatory effect of pharyngealization in Arabic. As the data can also be visual, I have also discussed the motion capture aspects to be considered.

### 2.5.2 Perspectives

As presented in my research program (Chapter 5), when dealing with multimodal speech, we need to consider several acquisition techniques, to capture vocal tract movement, facial deformation, and acoustics. Two main difficulties that need to be addressed are modality synchronization and integration. In the case of studying expressive audiovisual speech, some regions of the face need to be well acquired. We should also consider studying the correlation of one modality with the other. Finally, I consider that should be Improving the acquisition techniques a continuous work.

### 2.5.3 Related projects and contributions

- **Development and distribution of *VisArtico*:** it has been developed and distributed freely. It is used by 76 researchers from 16 countries. *VisArtico* not only allows viewing the coils but also enriches the visual information by indicating clearly and graphically the data for the tongue, lips and jaw, and offers some advanced functionality. `http://visartico.loria.fr`
- **Principal Investigator of an Inria ADT (technological development action) project (09/2013 - 08/2015)** . This project will make further improvements to make *VisArtico* very useful and widely used. For instance, we are planning the support of other EMA data formats and the addition of advanced data analysis and filtering. An engineer will participate in this project.
- **Participation in the project Equipex ORTOLANG (2012-2015)** - Tools and resources for a better optimized language processing - The project aims to offer a platform for articulatory data. In particular, we will acquire the latest articulograph AG501 during the last quarter of 2013
- **Participation in the project CPER MODAP (2007 - 2009)** - Articulatory speech modeling.

- Supervised **Ingmar Steiner**, during his postdoc stay. He worked on modeling 3D tongue using EMA.
- Supervised **Aymen El Amraoui**, during his master project. He worked on articulatory modeling of Arabic sounds.

– Supervised **Hatem Azzouni**, during his master project. He studied the articulatory and acoustic characteristics of arabic language.
– Supervised **Arnaud Paris and Aminata Leye** during their first year master project. They worked on the reconstruction of mid-sagittal tongue using EMA.

### 2.5.4 Selection of related publications


– Steiner, I., Richmond, K., and Ouni, S. (2013). Speech animation using electromagnetic articulography as motion capture data. In *AVSP 2013 - International Conference on Auditory-Visual Speech Processing*
– Ouni, S., Mangeonjean, L., Steiner, I., et al. (2012). Visartico: a visualization tool for articulatory data. In *Interspeech*
– Savariaux, C., Badin, P., Ouni, S., and Wrobel-Dautcourt, B. (2012). Étude comparée de la précision de mesure des systèmes d'articulographie électromagnétique 3d wave et ag500. *Actes des 29èmes Journées d'Etude de la Parole*, pages 513–520
– Steiner, I., Ouni, S., et al. (2011a). Investigating articulatory differences between upright and supine posture using 3d ema. In *9th International Seminar on Speech Production (ISSP'11)*, Montreal, Canada
– Steiner, I., Ouni, S., et al. (2011b). Towards an articulatory tongue model using 3d ema. In *9th International Seminar on Speech Production-ISSP'11*, pages 147–154
– Toutios, A., Ouni, S., and Laprie, Y. (2011c). Estimating the Control parameters of an Articulatory Model from Electromagnetic Articulograph Data. *The Journal of the Acoustical Society of America*, 129(5):3245–3257
– Embarki, M., Ouni, S., Yeou, M., Guilleminot, C., and Al Maqtari, S. (2011b). *Instrumental Studies in Arabic Phonetics*, volume 319 of *Current Issues in Linguistic Theory*, chapter Acoustic and electromagnetic articulographic study of pharyngealisation: Coarticulatory effects as an index of stylistic and regional variation in Arabic, pages 193–216. Zeki Majeed Hassan and Barry Heselwood, Amsterdam, j. benjamins publishing company edition
– Embarki, M., Ouni, S., and Salam, F. (2011a). Speech clarity and coarticulation in modern standard arabic and dialectal arabic. In *International Congress of Phonetic Sciences*, pages 635–638

# 3

# Audiovisual Speech Synthesis

## 3.1 Introduction

Speech communication is usually understood as the process of sending and receiving oral messages between people. We consider human-produced speech as a bimodal signal with two channels: acoustic and visual. The acoustic signal is the acoustic consequence of the deformation of the vocal tract under the effects of jaw, lips and tongue movements; and the visual signal is the consequence of this same deformation, which affects the shape of the face. Since some of the articulators directly correspond to facial features, it is quite reasonable to find out that acoustics and facial movements are correlated [Barker and Berthommier, 1999, Yehia et al., 1998].

Research in audiovisual speech intelligibility has shown the importance of the information provided by the face especially when audio is degraded [Sumby and Pollack, 1954, Le Goff et al., 1994, Ouni et al., 2007]. Moreover, [Le Goff et al., 1994] have shown that when audio is degraded or missing, the natural face provides two thirds of the missing auditory intelligibility, their synthetic face without the inner mouth (without the tongue) provides half of the missing intelligibility and the lips restores a third of it. For audiovisual synthesis, this suggests that one should pay careful attention to model the part of the face that participates actively during speech, i.e., mainly the lips and lower part of the face.

Audiovisual speech synthesis is a sub-field of the general areas of speech synthesis and computer facial animation. I started working on this field since my postdoctoral stay at the Perceptual Science Laboratory, University of California at Santa Cruz. I worked on improving the talking head Baldi. We extended the system to be multilingual. This work is presented in the following section. The audiovisual system Baldi is based on a parametric facial animation technique. The goal of this technique is to develop a polygon (wireframe) model with realistic motions without duplicating the musculature of the face to control this mask. This technique is also called, terminal analogue synthesis because its goal is to simply use the existing speech synthesis system (final product) to control the facial articulation of speech rather than illustrating the physiological mechanisms that produce it.

In the vast majority of recent works, parametric audiovisual synthesis and data-driven audiovisual speech synthesis, i.e., the generation of facial animation together with the corresponding acoustic speech, is still considered as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the facial

animation [Bailly et al., 2003, Theobald, 2007b, Liu and Ostermann, 2009, Edge et al., 2009]. However, achieving perfect synchronization between these two streams is not straightforward and presents several challenges related to audio-visual intelligibility. In fact, humans are acutely sensitive to any incoherence between audio and visual animation. This may occur as an asynchrony between audio and visual speech [Dixon and Spitz, 1980], or a small phonetic distortion compared to the natural relationship between the acoustic and the visual channels [Green and Kuhl, 1989, Green and Kuhl, 1991, Jiang et al., 2002, Jiang et al., 2005]. In the field of audiovisual synthesis, it has been shown that the degree of coherence between the auditory and visual modalities has an influence on the perceived quality of the synthetic visual speech [Mattheyses et al., 2009]. All these studies suggest the importance of keeping the link between the two highly correlated acoustic and visual channels. To reduce the possible existence of incoherency during audiovisual facial animation, I conducted research within the framework of the project ViSAC, to achieve synthesis with its acoustic and visual components simultaneously. Therefore, we consider audiovisual speech as a bimodal signal with two channels: acoustic and visual. This bimodality is kept during the whole synthesis process.

In the following sections, I first present an overview of my earlier work with developing multilingual talking head, then I introduce the conducted work on bimodal audiovisual speech synthesis.

## 3.2 Parametric Facial Animation

One advantage of parametric facial animation synthesis is that it is possible to carry out much faster calculations of the changing surface shapes in the polygon models than those for muscle and tissue simulations.

Baldi is a descendant of Parke's software and his particular 3-D talking head [Parke, 1975]. The model resolution has been increased, additional control parameters have been added, asymmetric facial movements have been allowed, a complex tongue has been trained, a coarticulation algorithm implemented, and finally (see Figure 3.1). Baldi can be driven by text-to-speech synthesis system or natural speech, to generate auditory/visual speech. The animation is mainly vertex-based or using geometric functions such as rotation (e.g. jaw rotation) or translation of the vertices in one or more dimensions. Many part of the face such as forehead shape, neck or cheek, and also some expressive parameters such as smiling are based on interpolation. The synthetic tongue is based polygon surfaces defined by coronal or sagittal curves. The teeth, tongue, and palate interactions during speech use a collision detection algorithm to prevent the tongue from going through the teeth and palate.

Baldi uses phonemes (in contrast to our in ViSAC, where we use diphone). In this scheme, an utterance is successive phonemes, and each phoneme is represented as a set of target values for the control parameters such as rounding, jaw rotation, etc. I recall that coarticulation is the description the interaction of phonemes which are influenced by the surrounding context. A coarticulation model is a speech production model using rules that describe the relative dominance of the characteristics of the speech segments [Cohen and Massaro, 1993]. Each phoneme is specified by a target value for each facial control parameter which has a temporal dominance functions dictating the influence of that phoneme over the control parameter. These dominance functions determine independently for each control parameter its importance realtive to the target and neighboring phonemes, which will in turn determine the final control values.

### 3.2.1 Multilingual Talking Head

My main work during the postdoctoral stay was on designing and developing a multilingual talking head. Adding languages to a talking head system is beneficial to extending research in bimodal speech and to develop applications in other languages (as visual speech can facilitate the learning on new languages.). For instance, it is possible to investigate the variability of articulation of a phoneme across different languages, to investigate the influence of visual speech across different languages, and to study the importance of visual information for perception of non-native speech.

We have introduced a platform to extend the capabilities of Baldi to speak other languages than English [Ouni et al., 2005]. We have imposed an independent implementation and application of the acoustic speech module and the visual speech synthesis module by using a client/server architecture. This scheme enables an efficient extension of text-to-speech synthesis and facial animation to many different languages with minimum development effort. For example, the client doesn't need any modification when a new text-to-speech engine is added to the server.

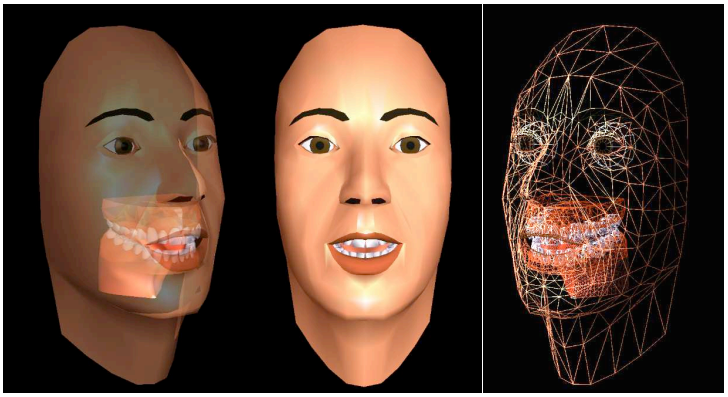


FIGURE 3.1 – Talking head Baldi. Three different views. In the middle, the standard Baldi; to the left, Semi-transparent Baldi (which allow to see the inner articulation: tongue, palate and teeth); to the right, the wireframe.

This strategy simplifies the extension to other languages.The development of the phoneme set and the corresponding target and coarticulation values allow synthesizing several other languages. We used empirical evaluation of the visual speech synthesis based on recognition by human observers to assess and to improve the quality of the synthesis. These experiments allow evaluating the intelligibility of the speech synthesis relative to natural speech. The goal of the evaluation was to learn how the synthetic visual talker performs in comparison to natural talkers and to improve the synthesis based on a given comparative analysis to reduce disparity with natural visual speech. In the following sections, I present the general scheme for this multilingual talking head and I present how we conducted the work to evaluate the system for Arabic.

#### General scheme for a multilingual talking head

Several requirements have to be achieved at the acoustic level and visual level, to have a multilingual talking head.

– **Auditory:** For any new language to be added to the talking head system, a text-to-speech (TTS) engine capable of producing that language is required (or a database of natural

speech that has been segmented and phonemically labeled. Often there are several TTS engines available for a given language.

– **Visual:** New phonemes will be required as well as revised definition for the existing phonemes, to have Baldi speak a new language accurately. Target values must be defined for each of the animation control parameters and the dominance functions (coarticulation parameters) must be specified, for each phoneme in a specific language.

In this client/server architecture, the server handles the acoustic speech component and the client deals with the visual animation component. The architecture allows the system to be flexible, distributable and usable via a network, in addition to the common benefits of modular software. The client software can be any application built around the talking head. The server module provides a standard and unique interface to various TTS systems or to a natural speech corpus (Figure 3.2). It is fairly easy to add a new language when the corresponding TTS engine is available. To include a TTS engine in the system, the minimum requirements are that the TTS engine provide the phonemes, their durations, and the synthesized acoustic speech. We applied this architecture to extend the capabilities of Baldi to speak new languages. In addition to English, we extended Baldi to speak Arabic, French, German, Italian, Chinese Mandarin, Castilian Spanish, Mexican Spanish, Korean and Swedish, Brazilian Portuguese. In terms of improving the visual speech for languages other than English, the most complete work has been done for Arabic. In the following sections, I present our method to define the articulation and coarticulation for each phoneme for this language.

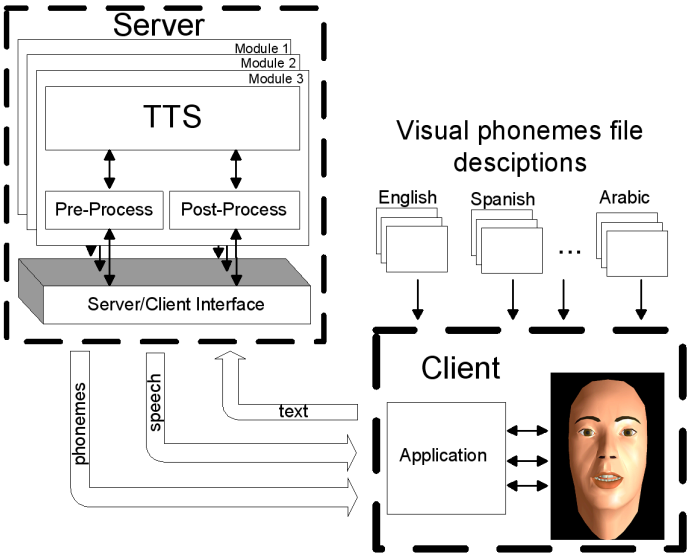



FIGURE 3.2 – Client/Server architecture system. The server is a unique interface between different TTSs and the Client (talking head). Each TTS has a corresponding module in the server that pre-processes the input and post-processes the output. The client is an application built around the talking head. The visual phonemes used for the talking head are defined in a description file. The client sends the text to the server, and in return, the server sends back the speech corresponding to the text and the phonemes with their duration.

### 3.2.2 An Arabic talking head

#### Articulation and coarticulation in Arabic

An important articulatory feature of Arabic is the presence of pharyngeal and pharyngealized phonemes. Two pharyngeal fricatives (/ʕ/ and /ħ/) are characterized by the constriction formed between the tongue and the lower pharynx in addition to the rising of the larynx. Three uvulars (/x/, /ʁ/ and /q/) are characterized by a constriction formed between the tongue and the upper pharynx for /x/ and /ʁ/ and a complete closure for /q/ at the same level. We consider these five consonants as pharyngeals. In addition, there are four pharyngealized phonemes: /$s^ʕ$/, /$t^ʕ$/, /$d^ʕ$/ and /$ð^ʕ$/. These phonemes have a secondary articulatory gesture that is a pharyngealization to a primary articulatory gesture that is oral dental; the oral dental consonants are /s/, /t/, /d/ and /θ/.

The main characteristic of the pharyngealization is the rearward movement of the back of the tongue. Thus, the vocal tract shape presents an increased oral cavity and a reduced pharyngeal cavity because of the retraction of the body and the root of the tongue toward the back wall of the pharynx [Al-Ani, 1970, Ghazeli, 1977, Jakobson, 1962]. The pharyngealized consonants also induce a considerable backing gesture in neighboring segments. It is mostly noticed for the adjacent vowels (the pharyngealized consonants affect the neighboring vowels in such a way that they will be transformed in their pharyngealized version.

Pharyngealization is both an intrasyllabic and intersyllabic phenomenon ([Ali and Daniloff, 1972, Ghazeli, 1977]. The coarticulatory effect of pharyngeal and pharyngealized consonants can affect just a single syllable or several. It is not easy to determine the extent of the coarticulation effect of the pharyngealized and pharyngeal phonemes on their neighboring consonants and vowels. The pharyngealization is observed in the pharyngealized consonants /$s^ʕ$/, /$t^ʕ$/, /$d^ʕ$/ and /$ð^ʕ$/ in all of the Arabic dialects. However, for pharyngeal phonemes, this varies from one dialect to another. The pharyngealization may also affect in certain cases /l/, /r/, /j/. As expected, researchers are not unanimous about the properties of these pharyngeal and pharyngealized phonemes in Arabic and its various dialects, their effects on other segments, and the mechanism used for pharyngeal consonant production [Al-Ani, 1970, Elgendy, 2001, Ghazeli, 1977].

#### Modeling Arabic Visual phonemes

We have used a variety of resources to define the visual phonemes as detailed measurements of the Arabic articulation and coarticulation are sparse and rare. I present two successive developments of the talking head: version 1 (a preliminary version) and version 2 (an improved version). For the first version, we roughly mapped the Arabic phonemes into their English closest ones, as the latter have already been defined and evaluated for English [Cohen et al., 2002], and we used the English phoneme that was closest visually for some other phonemes.

The goal of the first version was to establish a performance baseline of the intelligibility of a synthetic talking head of Arabic in noise. This method reduced the amount of time required to implement a working system, although we recognize that it is not ideal. The visual speech synthesis can be then improved based on perceptual experiment results, which is a strategy that has been used successfully in English (Massaro, 1998, Chapter 13) and can be productive especially when objective measures of natural articulation are absent. The results of the recognition experiment indicated which phonemes need improvement. Furthermore, successive perceptual

experiments and modifications ideally allow the eventual evolution of completely accurate synthetic visual speech. Based on the first perceptual evaluation study, the preliminary version was used to identify which phonemes need more modifications than others. We then improved the visual speech animation of the Arabic phonemes to create version 2, based on the results of the evaluation made on version 1 along with other the results found in the literature on Arabic phonetics and articulation.

**Evaluation**

To evaluate the Arabic talking head, we designed recognition experiments which help to determine how easily perceivers can speechread the face and how much the face adds to intelligibility of auditory speech presented in noise. The results of the different experiments were detailed in [Ouni et al., 2005]. The main finding was that the empirical evaluation enabled the improvement of the visual speech synthesis from one version to another. The talking head presented higher intelligibility than the first version. In addition, finer analysis at the phoneme level showed that the recognition of almost all the phonemes had progressed from version 1 to version 2. It should be mentioned that the Arabic talking head performed as good as the English talking head that was trained on real data [Ouni et al., 2005].

## 3.3 Acoustic-Visual Bimodal Synthesis

My earlier work described above is based on parametric talking head. This system belongs to the classic approach of audiovisual synthesis. As highlighted, in the vast majority of recent works, the generation of facial animation together with the corresponding acoustic speech, is still considered as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the facial animation. I have presented the drawbacks for a such approaches in section 3.1.

Within the framework of the project ViSAC (*Acoustic-Visual Speech Synthesis by Bimodal Unit Concatenation*) we perform synthesis with its acoustic and visual components simultaneously. Therefore, we consider audiovisual speech as a bimodal signal with two channels: acoustic and visual. This bimodality is kept during the whole synthesis process. The setup is similar to a typical concatenative acoustic-only speech synthesis, with the difference that here the units to be concatenated consist of visual information alongside acoustic information. The smallest segmental unit adopted in our work is the diphone. The advantage of choosing diphones is that the major part of coarticulation phenomena is captured in the middle of the unit, and the concatenation is made at the boundaries, which are acoustically and visually steadier. This choice is in accordance with current practices in concatenative acoustic-speech synthesis [Hunt and Black, 1996, Taylor, 2009].

I believe that an ideal audiovisual speech synthesis system should target the human receiver as its final and principal goal. Therefore, we focus on those aspects of audiovisual speech that make it more intelligible. These involve the dynamics of the lips and the lower part of the face: given that the lips are accurately animated, articulation and coarticulation will reproduce similar behavior as that of the real speaker. To achieve this goal, we are using a straightforward but efficient acquisition technique to acquire and process a large amount of parallel audiovisual data to cover the whole face by 3D markers. As can be seen in Figure 3.3, a large number of these markers mainly covers the lower face to allow accurate reconstruction of the lips and all the area around them.

At the current stage of our long term research goal, we do not provide a full talking head with a high rendering resolution. We do provide a bimodal synthesis method that can serve as the core of a larger system which will animate a high resolution mesh of the face with the inner vocal tract, using our simultaneous bimodal synthesis technique. Hence, our attempts are directed towards synthesizing realistic acoustic-visual dynamics that is coherent and consistent in both domains simultaneously: audio and visual. In the following, I present the ViSAC synthesis method and its evaluation. This work was a collaboration between two groups: Magrit group (*Brigitte Wrobel-Dautcourt* and *Marie-Odile Berger*) and Speech group (*Vincent Colotte*). In addition, *Utpala Musti* has worked on the project during her Ph.D. ; *Asterios Toutios* has developed the basic system during his postdoc stay ; *Blaise Potard* has worked as engineer to improve the stereovision acquisition system and *Caroline Lavecchia*, engineer, has developed the experimental evaluation platform.

### 3.3.1 Data acquisition and modeling

Visual data acquisition was performed simultaneously with acoustic data recording, using a classical stereovision system developed by the research group Magrit [Wrobel-Dautcourt et al., 2005]. The recorded corpus consisted of the 3D positions of 252 markers painted on the face of the speaker and the concurrently recorded speech signal, for 319 medium-sized French sentences, covering about 25 minutes of speech, uttered by a native male speaker. The positions of the markers were captured using a low-cost 3D facial data acquisition infrastructure [Wrobel-Dautcourt et al., 2005], with a sampling rate of 188.27 Hz. Acoustics were recorded at 16 kHz. The visual acquisition and 3D processing were performed by the research group Magrit, specialized in vision. The acoustic and visual components of the corpus are post-processed, and the corpus is analyzed linguistically. The final result is stored in a database as diphone entries. The details of the acquisition can be found in [Toutios et al., 2010a].

We applied PCA on a subset of markers: in the lower part of the face (jaw, lips, and cheeks – see Figure 3.3). The movements of markers on the lower part of the face are tightly connected to speech gestures. We retained the 12 first principal components, shown in Figure 3.4, which explain about 94% of the variance of the lower part of the face. The first two components account both for combined jaw opening and lip protrusion gestures. For the first component, as the jaw closes lips protrude. The effect is reversed for the second component: as the jaw *opens*, lips *protrude*. The third component accounts for lip opening, after removal of the jaw contribution, which is in good agreement with the lip opening parameter used in Maeda's articulatory model [Maeda, 1990]. Components 4 and 5 capture lip spreading. Component 6 is a smiling gesture. Components 7 to 12 seem to account for extremely subtle lip deformations, which we believe are idiosyncratic characteristics of our speaker. For the less significant components (4 to 12), it is not entirely clear whether they correspond to secondary speech gestures, or to facial expression features. Preliminary experiments indicated that retaining as few as three components could lead to an animation which would be acceptable, in the sense that it would capture the basic speech gestures and would filter out almost all the speaker specific gestures. However, such an animation would lack some naturalness, which is mostly captured by secondary components. Besides, I am in favor of keeping the specificity of the speaker gestures. Retaining 12 components leads to animations that are natural enough for all these purposes.

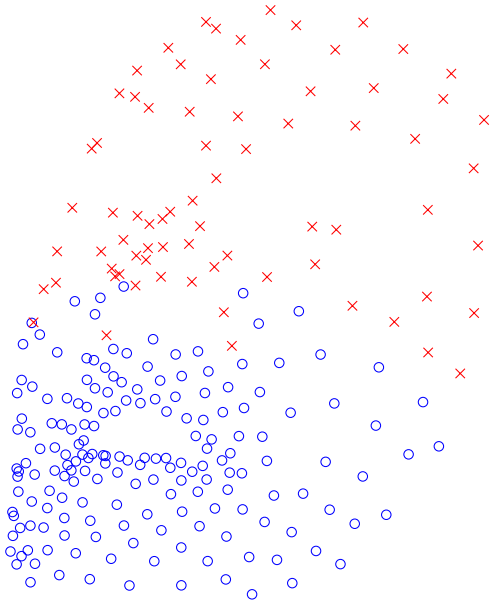


FIGURE 3.3 – 3D positions of the 252 markers. 178 of these markers (plotted in blue circles) are covering the lower face. The remaining markers (plotted in red crosses) do not reflect explicit speech gestures, in our case.

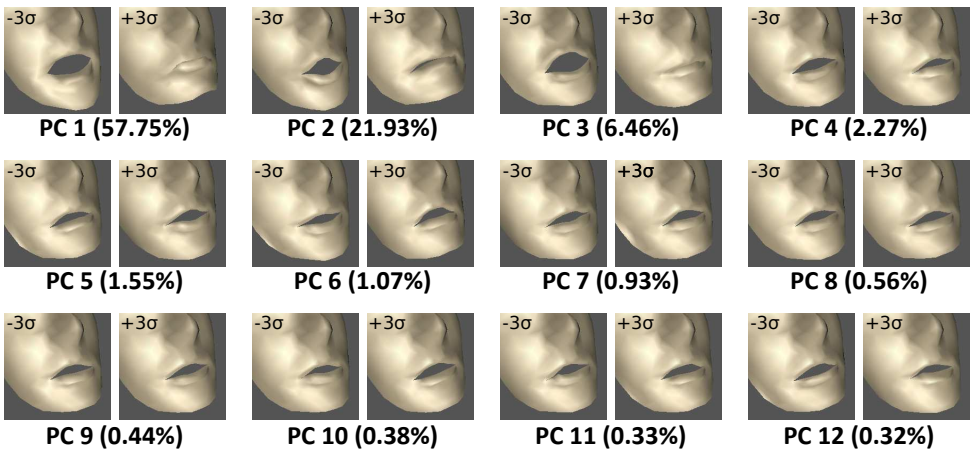



FIGURE 3.4 – The 12 first principal components of the facial data and the percentage of variance each of them explains. Each pair of images shows the deformation of the face when the corresponding component assumes a value of -3 (left) or +3 (right) standard deviations.

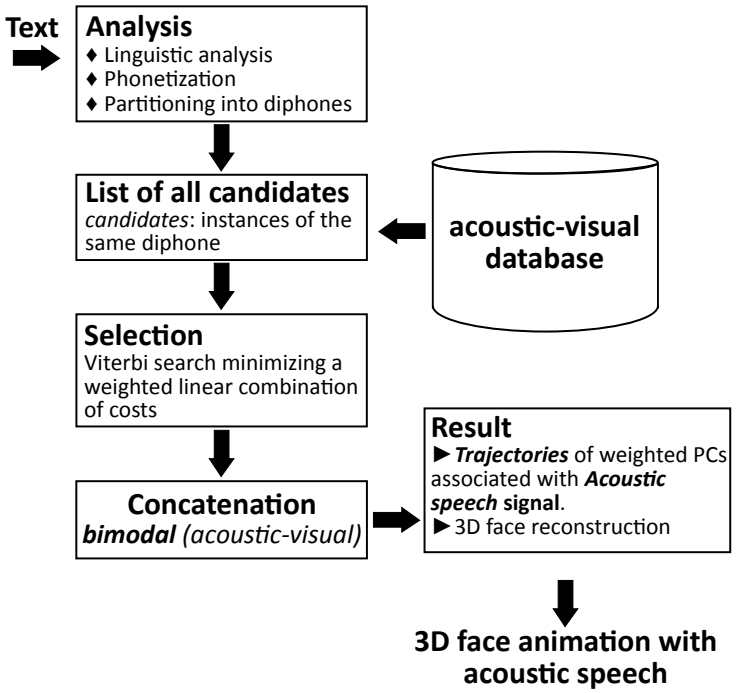



FIGURE 3.5 – Bimodal Text to speech synthesis processing.

### 3.3.2 Bimodal Text-to-Speech synthesis

The overall bimodal synthesis process is presented in Figure 3.5. The different steps of the text-to-speech synthesis (TTS) are similar to those in typical concatenative acoustic TTS [Clark et al., 2007]. The engine of our bimodal acoustic-visual synthesis relies on the acoustic-TTS system [Colotte and Lafosse, 2009], especially, for the necessary text analysis step.

At execution time, a text to be synthesized is first automatically phonetized and partitioned into diphones. For each diphone, all possible candidates from the database must have the same phonemic label. A special algorithm is available to handle cases when there are no instances of the same diphone in the database. The selection among these candidates is operated by resolution of the lattice of possibilities using dynamic programming. The result of the selection is the path in the lattice of candidates which minimizes a weighted linear combination of four costs, i.e.

$$C = w_{tc}TC + w_{jc}JC + w_{vc}VC + w_{dvc}DVC \tag{3.1}$$

where $TC$ is the target cost, $JC$ is the *acoustic join cost*, defined as the acoustic distance between the units to be concatenated, and is calculated using acoustic features at the boundaries of the units to be concatenated: fundamental frequency; spectrum; energy; and duration specification. $VC$ is the *visual join cost* calculated using the values of the PC trajectories at the boundaries of the units to be concatenated, i.e.

$$VC = \sum_{i=1}^{12} w_i \left(P_{i,1} - P_{i,2}\right)^2 \tag{3.2}$$

where $P_{i,1}$ and $P_{i,2}$ are the values of the projection on principal component $i$ at the boundary between the two diphones (see Figure 3.6). The weights $w_i$ should reflect the relative importance of the components, and we choose them to be proportional to the eigenvalues of PCA analysis,

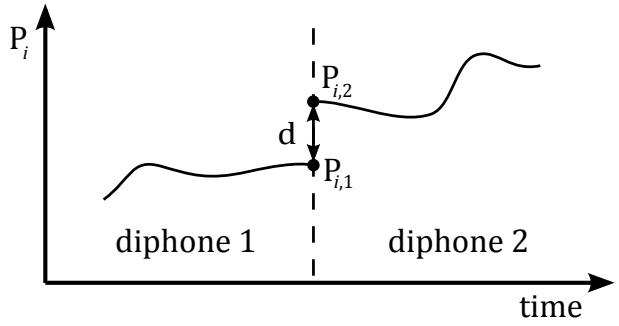



FIGURE 3.6 – Illustration of the visual cost calculation. The purpose is to minimize the distance $d$ between the points $P_{i,1}$ and $Pi, 2$ at the boundary of the two concatenated diphones.

in accordance with [Liu and Ostermann, 2009]. Therefore, as shown in Figure 3.4, the weights put a lot of emphasis on the first few components. Finally, the *derivative join cost DVC* is calculated in the same manner as $VC$, using the derivatives of the PC trajectories. Derivatives were calculated using a five-point stencil approximation.

The weights $w_{tc}$, $w_{jc}$, $w_{vc}$, $w_{dvc}$ are fine-tuned using an optimization method, which involves a series of simple metrics that compare a synthesized utterance to a set of test utterances. See [Toutios et al., 2011a] for the details of this optimization method and the description of the metrics.

In the acoustic domain, the selected diphone sequence is concatenated using a well-studied technique, where pitch values are used to improve the join of diphones. In the visual domain, we applied an adaptive local smoothing around joins which present discontinuities. If the first ($\Delta$) or second ($\Delta\Delta$) derivatives at a given sample of a synthesized visual trajectory lie out of the range defined by $\pm 3$ standard deviations (measured across the whole corpus) then this sample is judged as problematic. We traverse a visual trajectory $x_i$, and check $\Delta$ and $\Delta\Delta$ at each sample $i$. If one of them is out of the desired range, we replace samples $x_{i-k}$ to $x_{i+k}$ by their 3-point averaged counterparts, using incremental values for $k$, until $\Delta$ and $\Delta\Delta$ at sample $i$ are within the desired range. This technique reduces the irregularities at the boundaries based on the observed articulatory behavior of our speaker.

**Synthesis Example**

To get a better sense of the effect of the bimodal synthesis techniques on the outcome of the audiovisual synthesis, we present an example of synthesis in Figure 3.7. The trajectories of the first principal component for a synthesized utterance are presented in several synthesis scenarios. The first example (a) shows the case where only the acoustic cost is minimized. Several discontinuities are visible that result in visible jerks during the animation of the face. On the contrary, in the visual-only (b) and bimodal (c) cases, the resulting visual trajectories are sufficiently continuous. The synthesized acoustic speech of the visual-only result, while still intelligible, has several problems related to duration of diphones, intonation and some audible discontinuities at boundaries between diphones. The three cases (a), (b) and (c) of Figure 3.7 are using non-optimized weights. The result using optimized weights [Toutios et al., 2011a] is presented in (d). When using a different set of weights, several selected diphones are different, which is reflected in both acoustic and visual channels. The adaptive visual smoothing produced smoother animation (e). Figure 3.7 (f) shows a comparison of the synthesized trajectory with recorded trajectory. All the half-phones (the two half-phones of a diphone) of the synthesized sentence and the recorded

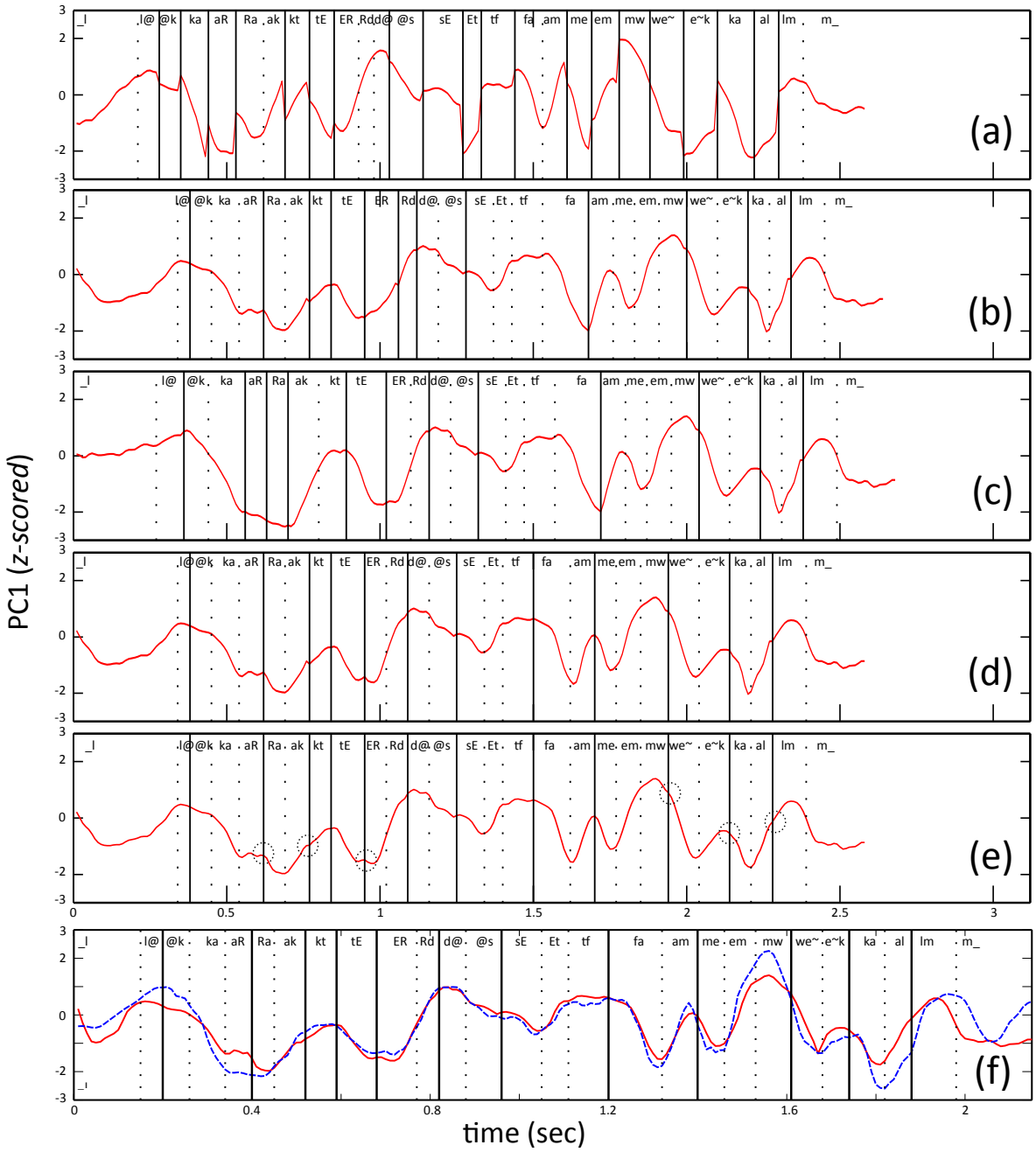



FIGURE 3.7 – Visual trajectories. First visual principal component (in z-scored units) for the sentence "Le caractère de cette femme est moins calme". When: (a) only acoustic join costs is minimized; (b) only visual cost minimized; (c) both acoustic and visual costs minimized using non-optimized weights; (d) then using optimized weights without processing at the visual joins. (e) Synthesized using the optimized weights, after processing visual joins. Note the corrected details marked with circles. (f) Original recorded trajectory (dashed) compared to the synthesized trajectory (solid) in (e). In (f), the duration of the diphones were adjusted to be able to make such comparison. Horizontal axes denote time in seconds. The boundaries between diphones are marked. Dashed lines indicate that the combination of the two diphones exists consecutively in the corpus and is extracted "as is" from it, solid lines otherwise. SAMPA labels for diphones are shown.

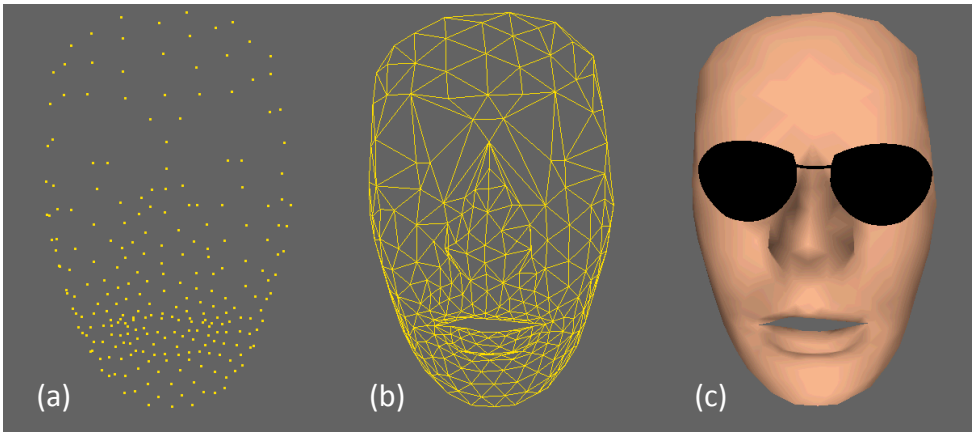



FIGURE 3.8 – Rendering Examples of the Face. (a) the 3D vertexes; (b) the triangulated mesh ; (c) smoothed mesh: the final result. The visual output of the synthesis process is the 3D vertexes that are then rendered as a smoothed mesh with skin-like color.

sentence were resampled individually to make the number of visual samples equal. It is worth noticing that the synthesized trajectories is following the same trends as the recorded trajectory.

### 3.3.3 Perceptual and Subjective Evaluations

Evaluating an audiovisual synthesis technique is always subtle and needs careful attention to draw the correct conclusion. As in our work we are manipulating both channels, acoustic and visual, the problem is twofold. Both audiovisual speech (animation) and acoustic speech need to be evaluated. It is probably possible to provide some conclusion on the quality of the visual synthesis based on the obtained visual trajectories shown in Figure 3.7, for instance. The trajectories are smooth and are similar to some test utterances. We used a cross-validation technique to evaluate the synthesis by comparing the synthesized sentences with the original ones [Musti et al., 2011a]. We used root mean square error (RMSE) and correlation coefficients for the evaluation. The results showed high correlation coefficients and the RMSE was very low.

However, I consider that the main evaluation criterion should be the intelligibility and the ability of the synthesis to send an intelligible message to the human receiver. I am convinced and advocate for the kind of evaluation. More details on my work on audiovisual intelligibility is presented in the next chapter. The audiovisual speech intelligibility focuses mainly on how well both audio and visual channels are integrated, and any mismatch or asynchrony influence human perception. If the acoustic or visual channel does not have a good quality, both acoustic and visual channels together might provide an overall result with higher intelligibility compared to taking each channel separately. When dealing with acoustic speech intelligibility, the focus is not just how comprehensible speech is (the degree to which speech can be understood), but also how natural and how pleasant the acoustic speech sounds.

It is not easy to conceive a method to evaluate both channels simultaneously. For this reason, we designed a perceptual experiment to evaluate the intelligibility of synthesized visual speech, and a subjective experiment to evaluate the intelligibility of synthesized acoustic speech. Even though, both experiments seem to be independent, they are implicitly highly correlated. The synthesis quality of one channel is related to the synthesis quality of the other channel, due to the synthesis design. Therefore, the perceptual experiment also provides hints on how good the acoustic speech is, and the subjective experiment will also provide insights on how good the

| | Audio | | Audiovisual | |
|---|---|---|---|---|
| | Hi N. | Lo N. | Hi N. | Lo N. |
| *Out-of-Corpus* | .34 | .40 | .40 | .45 |
| *In-Corpus* | .59 | .69 | .65 | .72 |

TABLE 3.1 – Mean scores under each condition (Hi N.: high noise, Lo N.: low noise), split into two set of stimuli: out-of-corpus words (39 words) and in-corpus words (11 words).

visual speech is.

We carried out a human audiovisual speech recognition experiment and a subjective MOS (mean opinion score) experiment. For the first experiment, the two presentation conditions were unimodal auditory speech and bimodal audiovisual speech. In the unimodal auditory presentation, participants can hear only the audio of the synthesized words. In the bimodal audiovisual presentation, participants can see and hear the synthesized face pronouncing the different words. For all the stimuli, the acoustics was paired with 2 different white noise signals where the average values of the speech-to-noise ratio (SNR) were either 6 dB or 10 dB. The noise was added to the stimuli to make it difficult, to some extent, to recognize the words based on audio only. For the second experiment, we used the MOS test to *subjectively* measure the quality of the synthesis as perceived by participants. Twenty synthesized acoustic-visual sentences were presented (without any added noise), and participants were asked to rate each sentence by answering five questions related to the quality, naturalness, prosody and asynchrony.

**Perceptual Evaluation Results**

Across the two noise conditions, the performance of the audiovisual presentation improved compared with unimodal audio presentation, and the difference was significant (Low noise level: (Audio: M=.47, SD=.08 ; Audiovisual: M=.51, SD=.09) t(76)=-2.25, p=.03; High noise level: (Audio: M=.4, SD=.09 ; Audiovisual: M=.46, SD=.1) t(76)=-2.79, p=.007). Although this was the minimum that one can expect from such a technique, but this suggests that visual synthesis present good coherence with audio regardless of the size of the corpus.

To refine the analysis, we also provide the results of *in-Corpus* (data as recorded from the original speaker) and *out-of-Corpus* (the result of the synthesis) sets, presented in Table 3.1. The results should be seen just as an indication on the intelligibility performance and not as a deep analysis since the number of items in *in-Corpus* set is smaller than that of *out-of-Corpus*. The purpose of introducing these two sets is to be able to compare the performance of the acoustic-visual synthesis compared with the face of the speaker used to record the corpus. It should be noted that, in this evaluation, we are not using

the video of the real face of our speaker, but a 3D reconstruction of the 252-vertex-face based on the recorded data. Thus, in our case, we replace the real face by the dynamics or the articulation of the speaker. For this reason, we are interested in comparing the synthetic face to the speaker's articulation. We continue to denote the reconstructed face from the original data as *the natural face*.

To estimate the quality of the synthetic face, we used the metric proposed by [Sumby and Pollack, 1954], to quantify the visual contribution to intelligibility. The metric is based on the difference between the scores of the bimodal and unimodal auditory conditions,

and measures the visual contribution $C_v$ in given noise condition, which is

$$C_v = \frac{C_{AV} - C_A}{1 - C_A} \tag{3.3}$$

where, $C_{AV}$ and $C_A$ are the unimodal auditory and the audiovisual intelligibility scores. We have used this metric , as several researchers, for evaluation purpose [Le Goff et al., 1994, Ouni et al., 2005]. We propose to use this metric not to compare synthetic face against natural face, but, for each kind of face, we compute its visual contribution to intelligibility. For the natural face, $C_v = .146$ in high noise level, and $C_v = .097$ in low noise level. For the synthetic face, $C_v = .091$ in high noise level, and $C_v = .083$ in low noise level. This suggests that the visual contribution to intelligibility of the synthetic face is very close to that of the natural face in the same condition. This is actually influenced by the quality of the audio.

Table 3.1 shows the improvement made by the synthetic face compared to that of using only the natural audio. The difference in performance between synthetic and natural audios shows that the acoustic synthesis has a scope for improvement to reach natural audio performance. In all cases, the perceptual experiment clearly shows that visual animation is not conflicting with audio, and there is no doubt of its intelligibility.

**Subjective Evaluation Results**

The main results of the subjective experiment are the followings. The proposed bimodal synthesis technique does not introduce any mismatch or asynchrony between the audio and visual channels. The acoustic prosody is acceptable. However, the rating related to the naturalness of the voice, is low. This can be explained by the size of the corpus where some diphones have a small number of candidates to propose during the selection step. We were expecting low rating for the naturalness of the face, as the vertexes of the face are not those of a high resolution face, and the face has no teeth or tongue. However, it seems that having good dynamics can overcome the sparseness of the vertexes. This can also be explained by the fact that humans are tolerant when we are not very close to the uncanny valley [Mori, 1970]. The detailed results of the evaluations can be found in ([Ouni et al., 2013]).

### 3.3.4 Conclusion

In this chapter, I have presented the conducted work on audiovisual speech synthesis. First, I started working on parametric talking head where a multilingual talking head has been advised. Extensive phonetics knowledge and successive perceptual experiments and modifications allowed the evolution of completely accurate synthetic audiovisual speech. Then, I have led the ViSAC project, where a bimodal unit-selection synthesis technique that performs text-to-speech synthesis with acoustic and visual components simultaneously was developed. It was based on the concatenation of bimodal diphones, units that consist of both acoustic and visual components. During all the steps, both components were used together. The good coverage of the lower face by an important number of markers allowed good dynamics of the lips. A perceptual and subjective evaluations of the bimodal acoustic-visual synthesis has been presented. The results showed that audiovisual speech provided by this synthesis technique is intelligible and acceptable as an effective tool of communication. The use of bimodal units to synthesize audiovisual speech seems to be a very promising technique that should probably be generalized in future projects as an effective audiovisual speech synthesis technique.

Regarding the acoustic quality, the bimodal speech synthesis quality is still not as good as that of the state-of-the-art acoustic synthesis systems. In fact, the latter is usually trained on three hours or more of acoustic speech, much larger than the 25-minute corpus used in the presented work. To reach equivalent quality, bimodal corpus should obviously be at equivalent size compared to that of the corpora typically used in acoustic speech synthesis. This means that an effort should be made in improving the acquisition technique to be able to acquire larger bimodal corpus. Regarding the visual synthesis, it is worth noticing that we are not yet presenting a complete talking head as we are for now just synthesizing the face and the lips concurrently with the acoustic speech. We are currently focusing on synthesizing the dynamics of the face, to assess that it is possible in practice to provide a synthesis technique where both acoustic and visual channels are considered as one unique bimodal signal. The development of a complete talking head is a part of my research program.

## 3.4 Summary and Contribution

### 3.4.1 Summary

In this chapter, I presented my contribution in the field of audiovisual speech synthesis. First, I started working on parametric talking head where a multilingual talking head has been advised. Extensive phonetic knowledge and successive perceptual experiments and modifications allowed the evolution of completely accurate synthetic audiovisual speech. Then, I have led the ViSAC project, where a bimodal unit-selection synthesis technique that performs text-to-speech synthesis with acoustic and visual components simultaneously was developed. It was based on the concatenation of bimodal diphones, units that consist of both acoustic and visual components. The results of the conducted evaluation showed that audiovisual speech provided by this synthesis technique is intelligible and acceptable as an effective tool of communication.

### 3.4.2 Perspectives

The main perspectives of this work will be largely detailed in my research program (Chapter 5). I will focus on expressive audiovisual speech synthesis, which represent more realistic face-to-face communication. The realism of this communication needs to improve the quality of the lip and tongue modeling where the dynamics needs to be stressed. In the proposed project, I will use my experience in the field of articulatory speech modeling and audiovisual speech synthesis to address all the related problems to this project.

### 3.4.3 Related projects and contributions

- **Principal Investigator of National Project ANR (Jeunes Chercheurs) ViSAC (2009-06/2013)** The program ANR Jeunes Chercheurs is highly competitive and the number of the accepted projects is very limited. This program requires that the P.I. should be involved in the project to 80% of his time of research. I have under my supervision a Ph.D student, a postdoc and an engineer. This project involved the members of two research groups: Vincent Colotte, Brigitte Wrobel-Dautcourt and Marie-Odile Berger.

- Supervised **Utpala Musti** (in collaboration with Vincent Colotte), during her Ph.D. She worked on acoustic-visual speech synthesis using bimodal units.
- Supervised **Asterios Toutios**, during his postdoc stay. He worked on developing the bimodal synthesis system, the main synthesis system of the project ViSAC.
- Supervised **Ingmar Steiner**, during his postdoc stay. He worked on developing 3D tongue model using EMA data, to be used in the context of audiovisual speech. This was a preparatory work for my research program in the future which can be considered as a feasibility study.
- Supervised **Caroline Lavecchia**, as an engineer. She worked on developing an evaluation platform on the web and she conducted the evaluation experiments.

### 3.4.4 Selection of related publications

- Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M.-O., and Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):16

– Ouni, S., Cohen, M. M., and Massaro, D. W. (2005). Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication*, 45:115–137
– Steiner, I., Richmond, K., and Ouni, S. (2013). Speech animation using electromagnetic articulography as motion capture data. In *AVSP 2013 - International Conference on Auditory-Visual Speech Processing*
– Musti, U., Colotte, V., Toutios, A., and Ouni, S. (2011b). Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer. In *Proc. 10th International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 49–55
– Toutios, A., Musti, U., Ouni, S., and Colotte, V. (2011a). Weight Optimization for Bimodal Unit-Selection Talking Head Synthesis. In ISCA, editor, *12thAnnual Conference of the International Speech Communication Association - Interspeech 2011*, Florence, Italie
– Musti, U., Toutios, A., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. (2010). Hmm-based automatic visual speech segmentation using facial data. In *Interspeech 2010*, pages 1401–1404
– Toutios, A., Musti, U., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. (2010a). Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units. In *Interspeech*, Makuhari, Japan
– Toutios, A., Musti, U., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. (2010b). Towards a true acoustic-visual speech synthesis. In *9th International Conference on Auditory-Visual Speech Processing-AVSP2010*

# 4

# Audiovisual Speech Intelligibility

## 4.1 Introduction

Several studies have shown that face-to-face communication is more efficient than situations where audio speech is uniquely used. One explanation is that the face improves intelligibility, in particular when the acoustic signal is degraded by noise or distracting prose (see [Sumby and Pollack, 1954, Benoit et al., 1994, Jesse et al., 2000]). Animated agents, that we also refer to as virtual 3D animated talking heads, have the potential to facilitate communication. They are beneficial in particular for hard-of-hearing individuals. Furthermore, an animated agent could be an intermediate between two persons communicating remotely when their video information is not available. [Beskow et al., 2004] has developed a system where an animated agent is driven by voice in telephone conversations. Language learning applications as vocabulary tutoring or pronunciation training can also use animated agents [Bosseler and Massaro, 2003, Wang et al., 2012, Wik and Hjalmarsson, 2009].

When dealing with audiovisual synthesis and seeing the potential of animated agents, their effectiveness is critically dependent on the quality of their visual speech. For this reason, I consider it is important to study audiovisual intelligibility and the ability of the synthesis to send an intelligible message to the human receiver. In fact, the intelligibility of the audiovisual synthesis can be critical when considering applications addressed to hard-of-hearing humans or to learners of new languages. In fact, this population is acutely sensitive to the quality of the articulation and to any incoherence between audio and visual animation. For hard-of-hearing, the visual component of speech is their main means of communication. In language learning, learners cannot master the pronunciation of a given language if they learn the wrong articulation.

For all these reasons, focusing on audiovisual intelligibility is extremely important. Naturally, my interest in audiovisual speech intelligibility started since I worked on audiovisual speech synthesis. In fact, to assess the quality of audiovisual speech synthesis, objective perceptual experiments should be designed where human participants are usually asked to recognize the presented linguistic items. My main goal is to investigate the mechanism of audiovisual intelligibility: what makes an audiovisual message intelligible? How does the visual component contribute to the audiovisual perception? and more importantly, how to assess the audiovisual intelligibility? How to compare the individual differences in speech intelligibility of different talkers?

In the field of audiovisual intelligibility, I have proposed a metric that allows comparing the intelligibility of an animated agent relatively to a standard or reference. This metric allows direct

comparisons across different sessions of experiments and measures the benefit of a synthetic face in comparison with the natural face, or any two conditions. It allows also quantifying how this benefit can vary depending of the type of synthetic face, different individuals, the test items, and applications. I have also studied the importance of the audiovisual speech in language learning and how visual speech can contribute to a better perception. In particular, I am interested in investigating whether speech production and perception of new language would be more easily learned when using audiovisual instead of audio-only speech. In particular, investigating whether viewing internal articulators (e.g. the tongue, palate, etc., not completely visible) is more beneficial for learning than a seeing the face from outside.

In the following sections, I present my contribution in the fields of audiovisual speech intelligibility and the visual contribution in pronunciation training.

## 4.2 Measuring Audiovisual intelligibility

### 4.2.1 Introduction

Providing a metric to measure audiovisual speech synthesis facilitates the evaluation of the visual intelligibility effectiveness of an agent. In earlier studies, [Sumby and Pollack, 1954] showed that speech intelligibility improved dramatically when seeing the speaker's facial and lip movements relative to not viewing the speaker. [Sumby and Pollack, 1954] proposed a metric to describe the contribution of the face relative to the auditory speech presented alone. In collaboration with Massaro and his colleagues [Ouni et al., 2007], I have conducted a study where an invariant metric has been proposed that provides a constant measure of the visual speech contribution across all levels of performance, and therefore would be independent of noise level. Additionally, this metric evaluates the effectiveness that describes intelligibility relative to a reference.

The goal is to extend the metric of [Sumby and Pollack, 1954] to quantify the advantage of a synthetic face relative to the advantage provided by a natural face. In the following sections, I first introduce the relative visual contribution metric, then I present how we used and evaluated the metric in three objective perceptual experiments, where a synthetic talker was compared to a natural talker, synthetic lips only versus a synthetic full face, and a natural talker's lips only versus a natural full face.

#### Visual contribution metric

To tackle this problem, [Sumby and Pollack, 1954] introduced a visual contribution metric that was supposed to provide a measure independent of the noise level. This metric has been used by [Le Goff et al., 1994]. In addition, we already used this metric in [Ouni et al., 2005] to compare results across different experiments. The metric corresponds to the difference between the scores from the bimodal and unimodal auditory conditions, and measures the visual contribution $C_v$ to performance in a given the speech-to-noise ratio condition, which is

$$C_v = \frac{C_{AV} - C_A}{1 - C_A} \tag{4.1}$$

where $C_{AV}$ and $C_A$ are the scores of bimodal audiovisual and unimodal auditory intelligibility. In this formula, $C_{AV}$ is expected to be greater than or equal to $C_A$. Qs can be seen in above, $C_v$ can vary between 0 and 1. Sumby and Pollack considered that $C_v$ is approximately constant over a range of speech-to-noise ratios. We have shown in [Ouni et al., 2007] that they did not

use inferential statistics to be able to conclude that their metric is not. Thus, we don't consider [Sumby and Pollack, 1954] metric as constant.

### 4.2.2 Relative Visual Contribution Metric

Sumby and Pollack's metric provides the contribution of a single talker. In the context of animated agents, the evaluation is made relatively to a natural talker. Measuring the quality of an animated agent should be assessed relative to a natural talking head reference. A completely inefficient agent is expected to have an intelligibility equal to or worse than the unimodal auditory condition. We consider an agent efficient when its intelligibility would be equal to the reference. In our work, we modified [Sumby and Pollack, 1954] equation, and proposed a direct measure of the intelligibility of an animated agent relative to that of a natural talker. Equation (4.1) is based on the perfect performance of the reference. In the context of animated agent evaluations, it is possible to consider several natural talker as references. In our work the reference was one talker, as the main goal is the implementation and testing an invariant metric. Thus, a metric that considers the natural talking head performance as the reference is introduced. First, I start by defining $C_v^r$ , the relative visual deficit to measure the missing information (the gap between the visual contribution of the natural face and the visual contribution of the synthetic face). $C_v^r$ is defined as follows:

$$\overline{C_v^r} = \frac{C_N - C_S}{1 - C_A} \tag{4.2}$$

where $C_N$ , $C_S$ , and $C_A$ are bimodal natural face, bimodal synthetic face and unimodal auditory intelligibility scores. Then we deduce the relative visual contribution $C_v^r$:

$$C_v^r = 1 - \frac{C_N - C_S}{1 - C_A} \tag{4.3}$$

and:

$$C_v^r + \overline{C_v^r} = 1 \tag{4.4}$$

To use this metric meaningfully, the unimodal auditory recognition scores should not be perfect

$$0 < (1 - C_A) < 1 \tag{4.5}$$

If this inequality does not hold, we cannot measure the benefit of visual speech as it implies that the unimodal auditory condition is not degraded. In these experiments, the acoustic signal channel should be degraded or some noise should be added. Thus, this metric allows evaluating the performance of a synthetic talker compared to a natural talker when the acoustic channel is degraded.

#### Interpretation of the relative visual contribution metric

– $\mathbf{C_v^r > 1}$: this is the case where the synthetic face outperforms the natural face. This result could simply mean that the natural talker reference has lower intelligibility than normal, or that the synthesized visual speech provides additional information. A hyperrealism can be the case when performance of the synthetic face is better than the natural face. In this case, the synthesis face might deliver additional cues not found in natural speech as making the nose red to signal nasality, an air stream coming from the mouth to signal

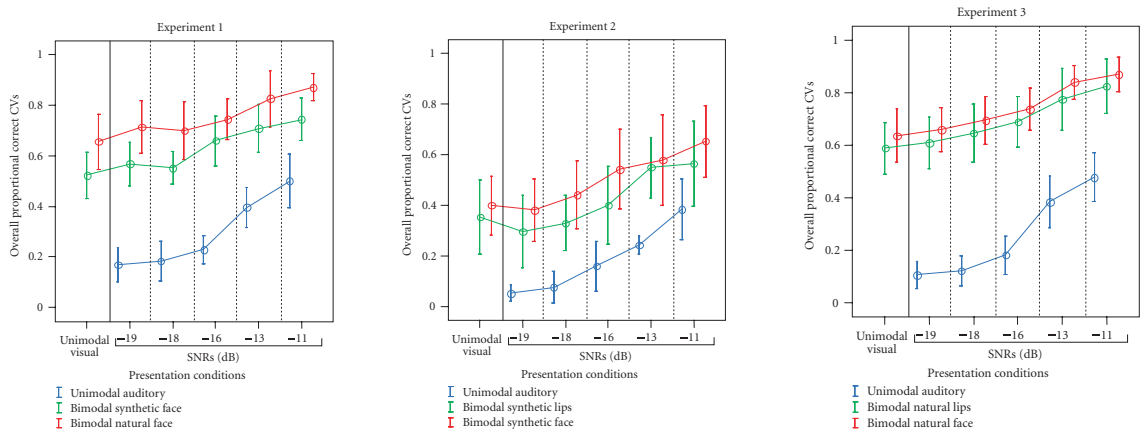



FIGURE 4.1 – Overall proportional correct CVs across five noise levels (SNR in dB) in three conditions per experiment: (Experiment 1) unimodal auditory, bimodal AV- synthetic face, and bimodal AV-natural face; (Experiment 2) unimodal auditory, bimodal AV- synthetic lips, and bimodal AV-synthetic face; (Experiment 3) unimodal auditory, bimodal AV- natural lips, and bimodal AV-natural face. Error bars represent the mean +/- 1 standard deviation. The figure includes also visual-only results.

frication and neck vibration to signal voicing ; that can obviously be visible in a natural face.

– $\mathbf{C_v^r \leq 1}$: this should be the most frequent outcome, as it is difficult to animate a synthetic talking face performing as good as a natural face. The value of $C_v^r$, however, provides a readily interpretable metric indexing the quality of the animated talker. The value of $C_v^r$ is the visual contribution of the synthetic talker relative to that of a natural talker. For $C_v^r$, the value should be read as the visual contribution of the synthetic face compared to the natural face independently of the auditory conditions of degradation. For example, a value of 80% means that the synthetic face has reached 80% of the natural face visual performance and that the quality is approaching real visual speech as this measure increases from 0 to 1.

### 4.2.3 Relative Visual Contribution in noise experiments

We expect that this metric would be invariant. The critical assumption underlying the metric is that it remains constant with differences in unimodal auditory performance (of course, ceteris paribis, when all other experimental conditions are constant). This allows also to compare similar experiments with different auditory performance. This assumption has been tested by carrying out a first experiment that compare a natural talker against a synthetic talker at 5 different noise levels to modulate baseline performance. We chose a natural talker who is known to have highly intelligible visual speech (see [Bernstein and Eberhardt, 1986, Massaro, 1998]). Then we carried out two extra experiments comparing a full face to lips-only to provide additional results to test for an invariant metric. This is a showcase where the metric can be used to assess how informative a particular part of the face compared to another part or to the full face is, in addition to comparing a natural talker to a synthetic talker. This type of result would be helpful to improve the synthesis of a particular region of the synthetic face, for example. We have chosen these conditions to give an important performance differences between the reference and the test.

**Results** The results are shown in Figure 4.1. This figure presents the overall percentage correct identification as one of the 27 CV syllables in the three experiments across five noise levels. For the first experiment, the three conditions were: unimodal auditory, bimodal AV-synthetic face, and bimodal AV-natural face. It is worth noticing that the performance improved with decreases in noise level (see Figure 4.1). Natural and synthetic faces gave an important advantage relative to the auditory condition. Performance for the synthetic face fell somewhat short of that for the natural face. For the two other experiments, intelligibility improved with decreases in noise level, both the full face and the lips-only gave an important advantage relative to the auditory condition. The full face gave better performance than lips-only when comparing both natural and synthetic talkers, however, the difference was much smaller for the natural face.

The detailed results can be found in [Ouni et al., 2007]. The main finding in this study is that our relative visual contribution metric did not differ over noise levels, $F(4,140) = 0.89$. Nor did noise level interact with experiments, $F(8, 140) = 0.88$. This shows clearly that the proposed metric remained invariant across noise levels, which was not the case for the Sumby and Pollack metrics.

### 4.2.4 A potential metric to measure intelligibility

The presented work confirm that animated synthetic talkers have not yet achieved the accuracy of natural talkers which is consistent with several other research (see [Massaro, 1998, Beskow et al., 2004, Ouni et al., 2005]). Identifying the important components of the face for visual speech perception (see [Summerfield, 1979, Benoit et al., 1994, Preminger et al., 1998] will improve the synthetic visual speech We found that the lips only were almost as effective as the full face for the natural face but much less than for the synthetic face. In addition, information from the face other than the mouth area can be used for visual speech perception. Currently, I started conducting research in this direction. Improved metrics for quantifying the contribution of visual speech with a good method should advance our understanding of audiovisual speech perception.

## 4.3 Audiovisual contribution to pronunciation training

### 4.3.1 Introduction

Audiovisual speech can be important and beneficial in language learning and visual speech can contribute for a better perception and thus better learning. In second language learning literature, researchers usually make the link between perception and production in how to discriminate non-native phones from native phonemes. In these cases, articulatory configurations of these different phonemes are usually discussed [Best et al., 2001, Hazan et al., 2006, Flege et al., 1995]. Recently, interest has increased among researchers to apply speech-production-based techniques for pronunciation training in second language learning. The main idea behind these techniques is that training learners on how to articulate non-native phonemes, or showing them the articulatory differences between native and non-native phonemes, can help to improve their production, and perhaps, their perception of these sounds. The articulatory improvement can be assessed either directly by measuring the learner's articulation using articulatory visualization techniques, or indirectly through the evaluation of the learner's acoustic realization.

In the following sections, I present my point of view regarding the benefits of using talking heads (also denoted virtual embodied conversational agents (ECAs)) in pronunciation training to

improve the production of second language learners. In particular, I am arguing whether seeing ECAs in views that cannot be provided by seeing a natural speaker are helpful for improving pronunciation. First, I present an overview of some studies in addition to my own contribution in this field where talking heads were used in pronunciation training. As the tongue is an important organ in speech articulation, I present then the study where I investigated human awareness of controlling their tongue body gestures, an important issue when dealing with audiovisual contribution in the context of pronunciation training.

**ECA as a visual pronunciation training tutor**

An ECA specialized in pronunciation training can be capable of showing the articulation of each sound from different views. For example, a midsagittal view, where a 2D view of the tongue, palate and velum are displayed, or a semi-transparent 3D view, where it is possible to see through the skin, and therefore to observe the inner articulators such as the tongue and the velum.

During recent years, some studies examined the benefits of using ECAs in pronunciation training to improve the production of new language learners. Of special interest was to assess whether seeing ECAs in views that cannot be provided by seeing a natural speaker are helpful for improving pronunciation. For example, French participants were asked to pronounce some Swedish words by observing a talking head with a view of the tongue [Engwall, 2012]. During training, instructions were given to explain the articulation presented by the animation of the talking head. Pronunciation improvement was assessed through ultrasound by measuring articulation. Participants' articulation improved through the training. The contribution of seeing the tongue and receiving instructions can, however, not be assessed here.

[Badin et al., 2010] assessed the benefit from seeing the tongue to better perceive sounds. They performed audiovisual perception experiments in a noisy environment with different viewing conditions. The main presentation conditions were the following: Participants either saw a midsagittal view of the jaw, vocal tract walls, palate and pharynx with the tongue or received the same view but without seeing the tongue. This was compared to a condition where a profile view of the full face was provided during training. An auditory-only condition where no speaker was visual, was also added. One of the results of this study is that the full face appears to provide the best perception results.

Similarly, [Grauwinkel and Fagel, 2007] found that showing the tongue and other articulators in comparison to not showing them did not provide a significant benefit in a consonant identification task. However, a short training where articulatory gestures were explained improved recognition. In another experiment, [Grauwinkel et al., 2007] used the visualization of inner vocal tract in a learning lesson addressed to three children with Sigmatismus interdentalis to improve their sibilant production. In an experiment presented in [Kröger et al., 2008] the visual recognition of mute uttered phonemes by children (five to eight years old) when presenting a 2D- or 3D-model was very low (19% to 22%).

[Massaro and Light, 2003] assessed the additional contribution of seeing the internal articulatory processes compared to simply seeing the face of a talking head for teaching non-native phonetic contrasts to Japanese learners of American English. The task was to identify and produce /r/ and /l/ in American English. Training either involved showing the face or also included showing the internal articulatory processes of the oral cavity. A pre-test/post-test

design was used. The participants' realizations were scored; where each acoustically produced word was scored by a human judge as correct or incorrect, without any knowledge of the experimental conditions. Although both speech identification and production improved, showing the internal articulators did not show an additional benefit.

In collaboration with Massaro and his colleagues, we have conducted an experimental study in the same context using a contrastive teaching approach [Massaro et al., 2008]. An important question was raised in this study is whether learners engaged in the pronunciation training differently as a function of the practice conditions. When watching a frontal view of the the talking head, learners might be likely to materially practice pronunciation in coordination with the ECA. Showing the sagittal view might be helpful, as they might even practice the gesture of a particular articulator, of which they might not otherwise be aware or only passively aware. We have designed a lesson to teach native English speakers phoneme pairs of a second language, where one phoneme in a pair was highly similar to a native phoneme, while the second phoneme was not. The idea behind this approach was that learners may benefit in their pronunciation training of the novel phoneme when they can see how it is produced differently from a phoneme they know how to produce. Figure 4.2 shows the two phoneme pairs that were trained: one in Arabic (/k/, /q/) and one in Mandarin (/i/, /y/). For Arabic, a midsagittal view was used, as the phonemes were velar and uvular, in comparison with a front view of the face. For Mandarin a front view of the talking face was used, as the phonemes were bilabials, in comparison with audio only. Although learners showed some improvements in the different conditions, for the Arabic talking head the benefit was larger for the full face than when showing the tongue. (the complete study can be found in [Massaro et al., 2008])

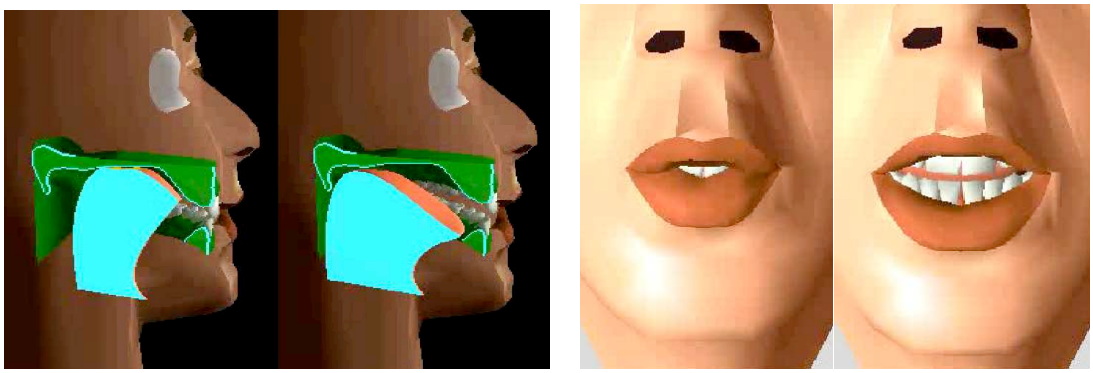


FIGURE 4.2 – (Left) Illustration of the internal articulatory processes of /k/ and /q/. (Right) Frontal view of /ly/ and /li/ at the point of maximum articulation of the vowel.

### Seeing the tongue and controlling it

The pronunciation training results show that although seeing inner articulators may help pronunciation, it does not seem to provide an additional benefit *a priori*. The general assumption behind showing articulations to learners is that they will imitate or implicitly improve their perception of the to-be-learned phonemes and thus their production. In fact, there is evidence for a strong link between perception and production in the motor regions of the brain. [Fadiga et al., 2002] showed that while listening to speech, listeners show an increase of motor activities of tongue muscles when the heard words involved tongue movements.

[Watkins et al., 2003] found similar results for lip muscles, when lip movements were observed in addition to listening to speech. However, in the case of second language learning, this perception-production link seems less useful when the relevant phoneme that is to be learned is absent in the learners' first language. The existence of a highly confusable native phoneme will provide additional difficulties [Best et al., 2001, Iverson et al., 2003]. One remedy here could be to provide instructions to learners on how to reach a target from a position of a phoneme in their native language by, for example, moving their tongue in some direction. However, as shown to some extent in the studies discussed above, there is little evidence that learners can correctly follow such instructions. In other words, we think that learners cannot easily reproduce some tongue movements just by illustrating or describing the gestures. In fact, pronunciation trainings based on illustrating tongue movement for some phonemes cannot be successful if learners are not able to reproduce those movements, even if they understand the animations. For instance, a French /r/, an English /r/, an Arabic /q/ or a German /ç/ are not easy to pronounce just by showing how to do so for learners for whom these phonemes do not exist in their native language.

Thus, important questions rise: can humans consciously control precisely the movement of the body of their tongue when asked to imitate or reproduce a tongue gesture? Are humans aware of their tongue gestures? Is it easy for humans to perform tongue movements mechanically? Does training with visual feedback of articulatory movements improve tongue control awareness?

To answer these questions, I have designed an experimental study based on two groups: a control group (10 participants) and an experimental group (14 participants). The general scheme of this study was a pre-test/post-test design. The control group did not receive any feedback, and the experimental group had a short training session (about 15 to 20 minutes) where participants observed their tongue movements in real time using an ultrasound machine. Each group had pre-test and post-test sessions. Their realizations were recorded using an ultrasound machine and evaluated offline by observing how well they succeeded in achieving the different gestures. This experimental design can be seen as two experiments. In the first one, I investigated how well the participants succeeded in achieving the different tongue gestures, without any *a priori* knowledge. In this experiment I examined only the pre-test sessions. In the second experiment, the goal was to investigate whether a short training session improves their awareness of their own tongue gestures. In this experiment, we examined the pre-test and post-test sessions and we compared the performance of the experimental group against the control group. The complete study can be found in [Ouni, 2013].

### 4.3.2 Tongue Control Study

#### Experimental Design

**Tongue Gestures** The 12 tongue movements are presented in Figure 4.3 and can be organized in two sets. The purpose of the first set of gestures was to observe how humans could control the motion of their tongue in various directions. The second set allowed assessing to what degree it is possible for speakers to move the body of their tongue from a known position of a phoneme of their native language to a new position. Therefore, this study addresses the question to what extent speakers can control the movement of the body of their tongue and the degree to which each set of gestures are easy or difficult to accomplish. Note that the starting position of the tongue is a neutral position, i.e. the tongue lying on the jaw and not touching the palate.

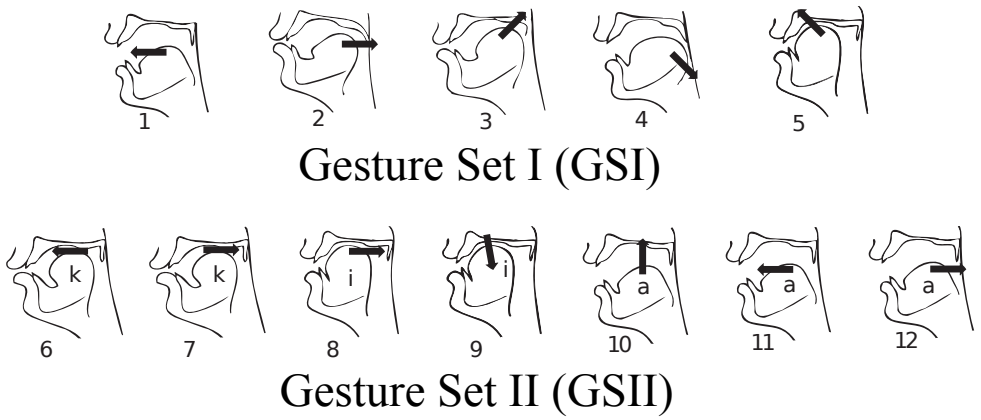



FIGURE 4.3 – The different directions used for the two sets of gestures. The vocal tract is just for illustration to show the tongue movement. The arrows show the direction of each gesture.

**Ultrasound tongue observation** In this study, we observed the motion of the tongue using ultrasound imaging. An ultrasound transducer (probe) placed against the chin produces a beam across the tissues of the tongue and is reflected at the surface of the tongue when it makes contact with air. The placement of the probe allows obtaining a midsagittal view of the vocal tract at a frame rate of 66 images per second.

**Pre-test and post-test sessions** The realizations of the participants were recorded using an ultrasound machine to be analyzed offline. Participants did not receive any visual feedback, i.e., they did not see the ultrasound images during the pre-test and post-test sessions. They handled the ultrasound probe themselves. They were instructed how to handle it correctly. Furthermore, the experimenter checked the orientation of the probe on the screen before asking the participant to perform a given gesture. The recording was analyzed offline. Instructions were given to participants by showing the direction of the tongue movement by hand. Before starting the experiment, participants were explained what the task was, and what was meant by the body of the tongue, making it clear that the body of the tongue gestures are different from those of the tip of the tongue. In addition, the production of some phonemes was recorded to serve as a reference in the data analyses. However, they were not given any information about the aim of the study before and during the experiment.

**Training session** A group of participants was involved in a very short observation and practice session (about 15 to 20 minutes). During this session, participants had the possibility to observe the movement of their tongue as displayed in real time by the ultrasound machine. They were able to practice the 12 predefined gestures. The experimenter provided them with a description of the different gestures and explained how to read an ultrasound image by showing the palate, the tongue and the overall shape of the vocal tract. In this study, the purpose was not to provide any training on how to control the tongue. The aim was to investigate whether a visual feedback session can be sufficient to improve the awareness of tongue gestures.

**Two experiments** In this study, I present two closely related experiments. In the first experiment, I investigate whether humans are aware of their tongue gestures, and whether there is a performance difference in reproducing set GSI compared to set GSII. In the second experiment, I investigate whether a short training session would improve participants' awareness of their own tongue gestures. The experimental design was as follow: (1) a pre-test session; (2) a training session for one group (test group) and nothing for the second group (control group); (3) a post-test session. For the first experiment, I consider only the pre-test sessions, and thus to increase the reliability of the result, the two groups were pooled. For the second experiment, I considered the three sessions, where one group of 14 participants (*Ultrasound group*) has a training session, and the other group of 10 participants (*Control group*) has no training. The control group has only a paper containing the list of gestures, but they were not given any instructions or asked to do any practice. During the two-test session, no feedback was provided. The only difference between pre-test and post-test is that the sequences of the presented gestures were randomized.

**Results**

In the following, I present the main results.

For the first experiment, the results showed that reproducing a specific gesture was not easy or obvious (M = 5.77, SD = 2.24). There was no participant who was able to reproduce all the gestures correctly. Although the selected gestures did not present a particular difficulty and are physically easy to produce by the different muscles, the participants were not able to control their tongue body movement and succeed in reproducing the different gestures.

For the second set GSII, I should note that the articulation of the starting phoneme was very accurate across all participants. However, the execution of the following gesture was in many cases not successful. This shows that starting from a very well-known position does not help that much, as participants did not seem to be aware of the place of articulation of this gesture, but just executed a "pre-recorded" movement.

The goal of the second experiment is to see the effect of a very short observation and practice session in improving participants' realization of the tongue gestures. This improvement was measured by comparing the performance of the Ultrasound group with the performance of the Control group.

Figure 4.4 shows the progress gained from pre-test to post-test, for the two groups. It is clear that almost all the 12 gestures (except gesture 9 and 10) gained improvement across Ultrasound group participants. However, the gestures realized by the Control group did not gain much improvement. Moreover, the scores of the set GSI actually deteriorated. One can speculate that the Control group does not have any hint on these gestures, and they are not even able to reproduce them. For the set GSII, as the Control group starts from a known position, this may reduce the possibility of making many wrong gestures, and thus the difference between pre- and post-test is reduced.

For the Ultrasound group, production was overall better after (M = 6.34, SD = 1.99) than before training (M = 5.72, SD = 2.22). The Control group gestures did not present any improvement during the pre-test (M = 5.86, SD = 2.23) and the post-test (M = 5.86, SD = 2.72). This result was significant (F (1, 42) = 3.82, p = 0.005) and there was an interaction between group and test (F (1, 42) = 5.69, p = 0.02).

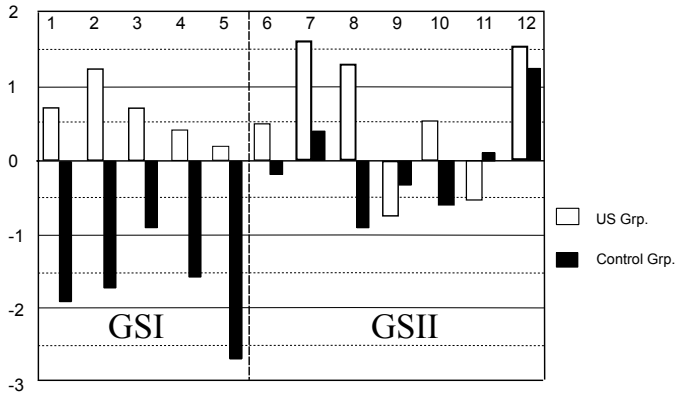



FIGURE 4.4 – Mean score differences of the 12 gestures across Pre-test Post-test. Results are presented for Ultrasound group and Control group.

### 4.3.3 Pronunciation training and visual feedback

The key finding of this study is that controlling the movement of the tongue body is to some extent difficult. Producing specific tongue gestures that were not learned during native language acquisition or second language learning processes is not an easy task. Humans are not really aware of the mechanism of articulating a given segment of speech as soon as they succeed in reaching the needed target. The coordination of the different tongue muscles to produce a given phoneme or word was acquired during early stages of language acquisition, after several repetitions and the retention became permanent. In this study, the gestures starting from places where the articulation is well known, as it is in their phonetic repertoire, did not help to reach a given target position. This implies that it is very likely that pronouncing phonemes during continuous speech follows an already learned articulatory path, and that it is not very easy to split up this path in elementary gestures. This suggests that pronunciation-training methods based on contrasts would not be very effective, if the purpose is to transfer the tongue movement illustrated by a talking head to the learner by imitation. Visual feedback seems to be helpful for these training methods. In fact, this study showed that learners benefit from having visual feedback available, even if only during an extremely short session of practice. I should highlight, however, that I could not confirm that the improvement is totally related to the visual feedback. In fact, the participants of the Control group were not asked to perform any particular task, as effectively practicing the two sets of gestures, and I cannot tell if some of them did do so or not.

During the training session, which is a trial and error process, participants were capable of increasing their awareness of their tongue gestures. Learners can visualize their own tongue movement and readjust a particular gesture based on the observation and the given instructions. During this practice session, participants were consciously trying to control the different gestures starting with awkward movements and they made many errors before starting to produce some correct gestures. I should notice that the effect of this session of practice lasts even when the feedback was removed, i.e., during the post-test session. Another finding is that visual feedback helped to increase the awareness of participants' articulations of phonemes of their first language used in this experiment. In fact, during the tongue observation sessions, participants stated that they did not know that the tested phonemes are produced in such way. This implies that

pronunciation training based on contrasts can be efficient if it is preceded or combined with visual feedback of the learner's articulation.

As final remark, pronunciation training based on illustrating speech articulation should take into account the awareness of learners of the used gestures. More generally, we highly recommend the use of some training based on visual feedback of the learner's articulation preceding the use of ECAs as tutor in language learning lessons. In fact, we believe that the use of ECA in language learning is very effective when used in the right conditions. We recommend that pronunciation training should be based on some explicit real-time visual feedback or preceded by an awareness task of the articulation gestures. As this is not an easy task, future studies should focus on how to integrate visual feedback techniques efficiently in the learning process. In addition, the persistence of the training using some visual feedback technique is not known and should be evaluated in dedicated studies.

## 4.4 Summary and Contribution

### 4.4.1 Summary

In the field of audiovisual intelligibility, I have proposed a metric that allows the comparison of an animated agent relatively to a standard or a reference. This metric allows direct comparisons across different experiments and give measures of the benefit of a synthetic animated face relative to a natural face, or any two conditions, and how this benefit varies as a function of the type of synthetic face, the test items, different individuals, and applications. I have also studied the importance of the audiovisual speech in language learning and how visual speech can contribute to a better perception. In particular, I am interested in investigating whether speech production and perception of new language would be more easily learned when using audiovisual instead of audio-only speech. In particular, investigating whether viewing internal articulators (e.g. the tongue, palate, etc., not completely visible) is more beneficial for pronunciation training than seeing the face from outside.

### 4.4.2 Perspectives

In the conducted experiments, we found that the lips-only were almost as effective as the full face for the natural face but much less than for the synthetic face. It is worth investigating how the information from the face other than the mouth area can be used for visual speech perception. Currently, I started conducting research in this direction. The aim is to quantify the information perceived form different regions of the face. It is also worth investigating how face-to-face communication can be very effective and how to provide the means to evaluate this effectiveness accurately.

### 4.4.3 Related projects and contributions

- Visit to Perceptual Science Laboratory at University of California, Santa Cruz for a month on June 2006. I have worked during this visit with Dominic Massaro and Michael Cohen on the development of the intelligibility metric. I have conducted perceptual experiments.

- Supervised **Jeremy Miranda**, during his master project. He studied the variability intra- and inter-speakers in audiovisual intelligibility.
- Supervised **Imen Jemai**, during her master project. She worked on the extraction of visual cues and the evaluation of audiovisual intelligibility
- Supervised **Fatma Ferchichi**, during her master project. She worked on the hierarchical visual classification of arabic phonemes.
- Supervised **Raja Fdhila**, during her master project. She worked on audiovisual speech intelligibility in Arabic.

- Invited speaker at the Workshop "Corpus et Outils en Linguistique, Langues et Parole" 1-2/07/2013, where I presented a talk on EMA articulatory acquisition and processing.
- Invited speaker at Institut de Phonétique de Strasbourg (2009), where I gave a talk on talking heads as a framework to study audiovisual speech.
- Invited speaker at the Natural Language Processing and Language Learning Workshop (2010), where I presented my work on Tongue control and its implication in pronunciation training.

### 4.4.4 Selection of related publications


- Ouni, S. (2013). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning*, 2013(16):1–15
- Miranda, J. and Ouni, S. (2013). Mixing faces and voices: a study of the influence of faces and voices on audiovisual intelligibility. In *AVSP 2013 - International Conference on Auditory-Visual Speech Processing*
- Ouni, S., Cohen, M. M., Ishak, H., and Massaro, D. W. (2007). Visual contribution to speech perception: measuring the intelligibility of animated talking heads. *EURASIP J. Audio Speech Music Process.*, 2007(1):3–3
- Ouni, S. (2011). Tongue gestures awareness and pronunciation training. In *12th Annual Conference of the International Speech Communication Association-Interspeech 2011*
- Massaro, D. W., Bigler, S., Chen, T., Perlman, M., Ouni, S., et al. (2008). Pronunciation training: the role of eye and ear. In *Proceedings of Interspeech*, pages 2623–2626

# Research Program

# Bibliographie


[Al-Ani, 1970] Al-Ani, S. H. (1970). *Arabic phonology*. Mouton The Hague, Paris.

[Al Bawab et al., 2008] Al Bawab, Z., Raj, B., and Stern, R. (2008). Analysis-by-synthesis features for speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4185–4188.

[Al Moubayed and Ananthakrishnan, 2010] Al Moubayed, S. and Ananthakrishnan, G. (2010). Acoustic-to-articulatory inversion based on local regression. In *Proc. Interspeech*, pages 937–940, Makuhari, Chiba, Japan.

[Albrecht et al., 2005] Albrecht, I., Schröder, M., Haber, J., and Seidel, H.-P. (2005). Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8(4):201–212.

[Ali and Daniloff, 1972] Ali, L. H. and Daniloff, R. G. (1972). A contrastive cinefluorographic investigation of the articulation of emphatic-non emphatic cognate consonants. *Studia Linguistica*, 26(2):81–105.

[Ananthakrishnan, 2011] Ananthakrishnan, G. (2011). *From Acoustics to Articulation: Study of the acoustic-articulatory relationship along with methods to normalize and adapt to variations in production across different speakers*. PhD thesis, KTH, School of Computer Science and Communication, Stockholm, Sweeden.

[Ananthakrishnan and Engwall, 2011a] Ananthakrishnan, G. and Engwall, O. (2011a). Mapping between acoustic and articulatory gestures. *Speech Communication*, 53(4):567–589.

[Ananthakrishnan and Engwall, 2011b] Ananthakrishnan, G. and Engwall, O. (2011b). Mapping between acoustic and articulatory gestures. *Speech Communication*, 53(4):567 – 589.

[Atal et al., 1978] Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63(5):1535–1555.

[Badin et al., 2010] Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G. (2010). Can you 'read'tongue movements? evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, 52(6):493–503.

[Bailly et al., 2008] Bailly, G., Bégault, A., Badin, P., et al. (2008). Speaking with smile or disgust: data and models. In *Interspeech*, pages 111–116.

[Bailly et al., 2003] Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003). Audiovisual speech synthesis. *International Journal of Speech Technology*, 6(4):331–346.

[Bailly et al., 2009] Bailly, G., Pelachaud, C., et al. (2009). Parole et expression des émotions sur le visage d'humanoïdes virtuels. *Traité de la réalité virtuelle: Volume 5: les humains virtuels*, pages 187–208.

[Barker and Berthommier, 1999] Barker, J. and Berthommier, F. (1999). Evidence of correlation between acoustic and visual features of speech. In *ICPhS*, San Francisco, USA.

[Ben Youssef et al., 2010] Ben Youssef, A., Badin, P., and Bailly, G. (2010). Can tongue be recovered from face? The answer of data-driven statistical models. In *Interspeech*, Makuhari, Japan.

[Benoit et al., 1994] Benoit, C., Mohamadi, T., and Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of french. *Journal of Speech and Hearing Research*, 37(5):1195–1203.

[Bernstein and Eberhardt, 1986] Bernstein, L. and Eberhardt, S. (1986). Johns hopkins lipreading corpus videodisk set.

[Beskow et al., 2004] Beskow, J., Karlsson, I., Kewley, J., and Salvi, G. (2004). Synface-a talking head telephone for the hearing-impaired. In *Proceedings of 9th International Conference on Computers Helping People with Special Needs (ICCHP '04)*, pages 1178–1186, Paris, France.

[Beskow and Nordenberg, 2005] Beskow, J. and Nordenberg, M. (2005). Data-driven synthesis of expressive visual speech using an mpeg-4 talking head. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 793–796.

[Best et al., 2001] Best, C. T., McRoberts, G. W., and Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109:775.

[Black et al., 2012] Black, A. W., Bunnell, H. T., Dou, Y., Kumar Muthukumar, P., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S., and Vaughn, C. (2012). Articulatory features for expressive speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4005–4008. IEEE.

[Boë et al., 1992] Boë, L.-J., Perrier, P., and Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20:27–38.

[Bosseler and Massaro, 2003] Bosseler, A. and Massaro, D. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of Autism and Developmental Disorders*, 33(6):653–672.

[Busset, 2013] Busset, J. (2013). *Acoustic-to-articulatory inversion using cepstral coefficients*. PhD thesis, University de Lorraine, Nancy, France.

[Carstens Medizinelektronik, 2004] Carstens Medizinelektronik (2004). *AG500 Data Format and Data Structure*. Lenglern, Germany.

[Carstens Medizinelektronik, 2006] Carstens Medizinelektronik (2006). *JustView: AG500 measuring environment display*. Lenglern, Germany.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Charpentier, 1984a] Charpentier, F. (1984a). Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic non-linearities. *Speech Communication*, 3:291– 308.

[Charpentier, 1984b] Charpentier, F. (1984b). Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic nonlinearities. *Speech Communication*, 3(4):291–308.

[Cherkassky and Ma, 2004] Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 17:113–126.

[Chuang and Bregler, 2005] Chuang, E. and Bregler, C. (2005). Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG)*, 24(2):331–347.

[Clark et al., 2007] Clark, R., Richmond, K., and King, S. (2007). Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, 49(4):317 – 330.

[Cohen and Massaro, 1990] Cohen, M. M. and Massaro, D. W. (1990). Synthesis of visible speech. *Behavior Research Methods, Instruments, & Computers*, 22(2):260–263.

[Cohen and Massaro, 1993] Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. *Models and techniques in computer animation*, 92.

[Cohen et al., 2002] Cohen, M. M., Massaro, D. W., and Clark, R. (2002). Training a talking head. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 499. IEEE Computer Society.

[Colotte and Lafosse, 2009] Colotte, V. and Lafosse, A. (2009). Soja: French text-to-speech synthesis system.

[consortium, 2009a] consortium, P. (2009a). Gv-lex (`http://www.gvlex.com/en`).

[consortium, 2009b] consortium, P. (2009b). Semaine project (`http://www.semaine-project.eu`).

[Demange and Ouni, 2013] Demange, S. and Ouni, S. (2013). An episodic memory-based solution for the acoustic-to-articulatory inversion problem. *The Journal of the Acoustical Society of America*, 133(5):2921–2930.

[Demange et al., 2011] Demange, S., Ouni, S., et al. (2011). Continuous episodic memory based speech recognition using articulatory dynamics. *Proceedings of Interspeech, Florence, Italy*, pages 2305–2308.

[Deng and Sun, 1994] Deng, L. and Sun, D. (1994). Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of english sounds. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume i, pages I/45 –I/48 vol.1.

[Dixon and Spitz, 1980] Dixon, N. F. and Spitz, L. (1980). The detection of audiovisual desynchrony. *Perception*, 9:719–721.

[Edge et al., 2009] Edge, J. D., Hilton, A., and Jackson, P. (2009). Model-based synthesis of visual speech movements from 3d video. *EURASIP Journal on Audio, Speech, and Music Processing*.

[Elgendy, 2001] Elgendy, A. (2001). *Aspects of pharyngeal coarticulation*. LOT.

[Embarki et al., 2011a] Embarki, M., Ouni, S., and Salam, F. (2011a). Speech clarity and coarticulation in modern standard arabic and dialectal arabic. In *International Congress of Phonetic Sciences*, pages 635–638.

[Embarki et al., 2011b] Embarki, M., Ouni, S., Yeou, M., Guilleminot, C., and Al Maqtari, S. (2011b). *Instrumental Studies in Arabic Phonetics*, volume 319 of *Current Issues in Linguistic Theory*, chapter Acoustic and electromagnetic articulographic study of pharyngealisation: Coarticulatory effects as an index of stylistic and regional variation in Arabic, pages 193–216. Zeki Majeed Hassan and Barry Heselwood, Amsterdam, j. benjamins publishing company edition.

[Engwall, 2002] Engwall, O. (2002). Evaluation of a system for concatenative articulatory visual speech synthesis. In *International Conference on Speech and Language Processing, Boulder-Colorado*.

[Engwall, 2012] Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, 25(1):37–64.

[Erler and Deng, 1993] Erler, K. and Deng, L. (1993). Hidden Markov model representation of quantized articulatory features for speech recognition. *Computer Speech and Language*, 7(3):265–282.

[Fadiga et al., 2002] Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a tms study. *European Journal of Neuroscience*, 15(2):399–402.

[Fant, 1960] Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton & Co., The Hague.

[Fernandez and Ramabhadran, 2007] Fernandez, R. and Ramabhadran, B. (2007). Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pages 34–39.

[Flege et al., 1995] Flege, J. E., Munro, M. J., and MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, 97:3125.

[Frankel and King, 2001] Frankel, J. and King, S. (2001). Asr - articulatory speech recognition. In *Proc. Eurospeech*, pages 599–602, Aalborg, Denmark.

[Galatas et al., 2012] Galatas, G., Potamianos, G., and Makedon, F. (2012). Audio-visual speech recognition incorporating facial depth information captured by the kinect. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2714–2717.

[Ghazeli, 1977] Ghazeli, S. (1977). *Back consonants and backing coarticulation in Arabic*. PhD thesis, University of Texas at Austin.

[Ghosh and Narayanan, 2010] Ghosh, P. and Narayanan, S. (2010). A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 128(4):2162–2172.

[Grauwinkel et al., 2007] Grauwinkel, K., Dewitt, B., and Fagel, S. (2007). Visualization of internal articulator dynamics and its intelligibility in synthetic audiovisual speech. *Proc. ICPhS, Saarbrücken*.

[Grauwinkel and Fagel, 2007] Grauwinkel, K. and Fagel, S. (2007). Visualization of internal articulator dynamics for use in speech therapy for children with sigmatismus interdentalis. In *Int. Conf. on Auditory-Visual Speech Processing*.

[Green and Kuhl, 1989] Green, K. P. and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics*, 45:34–42.

[Green and Kuhl, 1991] Green, K. P. and Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17:278–288.

[Guiard-Marigny et al., 1996] Guiard-Marigny, T., Tsingos, N., Adjoudani, A., Benoit, C., and Gascuel, M.-P. (1996). 3d models of the lips for realistic speech animation. In *Computer Animation'96. Proceedings*, pages 80–89. IEEE.

[Hazan et al., 2006] Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., and Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, 119:1740.

[Hiroya and Honda, 2004] Hiroya, S. and Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech and Signal Processing*, 12(2):175–185.

[Hunt and Black, 1996] Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP*, Atlanta, USA. IEEE.

[Itakura, 1975] Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):67–72.

[Iverson et al., 2003] Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87(1):B47–B57.

[Jakobson, 1962] Jakobson, R. (1962). *Selected writings*, volume 1, chapter "Mofaxxama", the emphatic phonemes in Arabic, pages 510–522. Mouton.

[Jesse et al., 2000] Jesse, A., Vrignaud, N., Cohen, M., and Massaro, D. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, 5(2):95–115.

[Jiang et al., 2002] Jiang, J., Alwan, A., Keating, P. A., Auer, E. T., and Bernstein, L. E. (2002). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Applied Signal Processing*, 11:1174–1188.

[Jiang et al., 2005] Jiang, J., Bernstein, L. E., and Edward T. Auer, J. (2005). Realistic face animation from sparse stereo meshes. In *AVSP*, British Columbia, Canada.

[Katsamanis et al., 2009] Katsamanis, A., Papandreou, G., and Maragos, P. (2009). Face active appearance modeling and speech acoustic information to recover articulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):411–422.

[King and Parent, 2005] King, S. A. and Parent, R. E. (2005). Creating speech-synchronized animation. *Visualization and Computer Graphics, IEEE Transactions on*, 11(3):341–352.

[Kröger et al., 2008] Kröger, B., Graf-Borttscheller, V., and Lowit, A. (2008). Two and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In *Interspeech, 9th Annual Conference of the International Speech Communication Association*.

[Laboissièrre and Galvàn, 1995] Laboissièrre, R. and Galvàn, A. (1995). Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences. In *Proceedings ICPhS*, volume 1, pages 358–361, Stockholm.

[Larar et al., 1988] Larar, J., Schroeter, J., and Sondhi, M. (1988). Vector quantization of the articulatory space. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(12):1812 –1818.

[Le Goff et al., 1994] Le Goff, Guiard-marigny, T., Cohen, M., and Benoit, C. (1994). Real-time analysis-synthesis and intelligibility of talking faces. In *In 2nd International conference on Speech Synthesis*, pages 53–56.

[Levitt and Katz, 2010] Levitt, J. and Katz, W. (2010). The Effects of EMA-Based Augmented Visual Feedback on the English Speakers' Acquisition of the Japanese Flap: A Perceptual Study. In *Interspeech*, Makuhari, Japan.

[Ling et al., 2010] Ling, Z.-H., Richmond, K., and Yamagishi, J. (2010). Hmm-based text-to-articulatory-movement prediction and analysis of critical articulators. In *INTERSPEECH*, pages 2194–2197, Chiba, Japan.

[Liu and Ostermann, 2009] Liu, K. and Ostermann, J. (2009). Optimization of an Image-Based Talking Head System. *EURASIP Journal on Audio, Speech, and Music Processing*.

[Maeda, 1979] Maeda, S. (1979). Un modele articulatoire de la langue avec des composantes linéaires. *Actes 10emes Journées d'Etude sur la Parole*, pages 152–162.

[Maeda, 1990] Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic.

[Massaro, 1998] Massaro, D. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press, Cambridge, Mass, USA.

[Massaro and Light, 2003] Massaro, D. and Light, J. (2003). Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pages 2249–2252, Geneva, Switzerland.

[Massaro et al., 2008] Massaro, D. W., Bigler, S., Chen, T., Perlman, M., Ouni, S., et al. (2008). Pronunciation training: the role of eye and ear. In *Proceedings of Interspeech*, pages 2623–2626.

[Mattheyses et al., 2009] Mattheyses, W., Latacz, L., and Verhelst, W. (2009). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Audio, Speech, and Music Processing*.

[Miranda and Ouni, 2013] Miranda, J. and Ouni, S. (2013). Mixing faces and voices: a study of the influence of faces and voices on audiovisual intelligibility. In *AVSP 2013 - International Conference on Auditory-Visual Speech Processing*.

[Mori, 1970] Mori, M. (1970). The Uncanny Valley. *Energy*, 7:33–35.

[Munhall et al., 2004] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137.

[Musti et al., 2011a] Musti, U., Colotte, V., Toutios, A., and Ouni, S. (2011a). Introducing Visual Target Cost within an Acoustic-Visual Unit-Selection Speech Synthesizer. In *International Conference on Auditory-Visual Speech Processing - AVSP2011*, Volterra, Italy.

[Musti et al., 2011b] Musti, U., Colotte, V., Toutios, A., and Ouni, S. (2011b). Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer. In *Proc. 10th International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 49–55.

[Musti et al., 2010] Musti, U., Toutios, A., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. (2010). Hmm-based automatic visual speech segmentation using facial data. In *Interspeech 2010*, pages 1401–1404.

[Neiberg et al., 2008] Neiberg, D., Ananthakrishnan, G., and Engwall, O. (2008). The acoustic to articulatory mapping: non-linear or non-unique? In *Proc. Interspeech*, pages 1485–1488, Brisbane, Australia.

[Ney, 1983] Ney, H. (1983). A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2):163–173.

[Nguyen, 2000] Nguyen, N. (2000). A MATLAB toolbox for the analysis of articulatory data in the production of speech. *Behavior Research Methods, Instruments, & Computers*, 32(3):464–467.

[Ouni, 2005] Ouni, S. (2005). Can we retrieve vocal tract dynamics that produced speech? toward a speaker articulatory strategy model. *Interspeech 2005-Eurospeech*, pages 1037–1040.

[Ouni, 2011] Ouni, S. (2011). Tongue gestures awareness and pronunciation training. In *12th Annual Conference of the International Speech Communication Association-Interspeech 2011*.

[Ouni, 2013] Ouni, S. (2013). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning*, 2013(16):1–15.

[Ouni et al., 2007] Ouni, S., Cohen, M. M., Ishak, H., and Massaro, D. W. (2007). Visual contribution to speech perception: measuring the intelligibility of animated talking heads. *EURASIP J. Audio Speech Music Process.*, 2007(1):3–3.

[Ouni et al., 2005] Ouni, S., Cohen, M. M., and Massaro, D. W. (2005). Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication*, 45:115–137.

[Ouni et al., 2013] Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M.-O., and Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):16.

[Ouni and Laprie, 2005a] Ouni, S. and Laprie, Y. (2005a). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of Acoustic Society of America*, 118(1):444–460.

[Ouni and Laprie, 2005b] Ouni, S. and Laprie, Y. (2005b). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 118(1):444–460.

[Ouni and Laprie, 2009] Ouni, S. and Laprie, Y. (2009). Studying pharyngealisation using an articulograph. In *International Workshop on Pharyngeals and Pharyngealisation*.

[Ouni et al., 2012] Ouni, S., Mangeonjean, L., Steiner, I., et al. (2012). Visartico: a visualization tool for articulatory data. In *Interspeech*.

[Papcun et al., 1992] Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., and Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the Acoustical Society of America*, 92(2):688–700.

[Parke, 1975] Parke, F. I. (1975). A model for human faces that allows speech synchronized animation. *Computers & Graphics*, 1(1):3–4.

[Pelachaud et al., 2001] Pelachaud, C., Magno-Caldognetto, E., Zmarich, C., and Cosi, P. (2001). Modelling an italian talking head. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*.

[Pelachaud and Poggi, 2002] Pelachaud, C. and Poggi, I. (2002). Subtleties of facial expressions in embodied agents. *The Journal of Visualization and Computer Animation*, 13(5):301–312.

[Potard, 2008] Potard, B. (2008). *Acoustic-to-articulatory inversion with constraints*. PhD thesis, University Henri Poincare, Nancy, France.

[Potard et al., 2008] Potard, B., Laprie, Y., and Ouni, S. (2008). Incorporation of phonetic constraints in acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 123(4):2310–2323.

[Preminger et al., 1998] Preminger, J., Lin, H.-B., Payen, M., and Levitt, H. (1998). Selective visual masking in speechreading. *Journal of Speech, Language, and Hearing Research*, 41(3):564–575.

[Qin and Carreira-Perpiñán, 2007] Qin, C. and Carreira-Perpiñán, M. (2007). An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. In *Proc. Interspeech*, pages 74–77, Antwerp, Belgium.

[Richardson et al., 2003] Richardson, M., Bilmes, J., and Diorio, C. (2003). Hidden-articulator Markov models for speech recognition. *Speech Communication*, 41(2-3):511–529.

[Richmond, 2002] Richmond, K. (2002). *Estimating Articulatory Parameters from the Speech Signal*. PhD thesis, Centre for Speech Technology Research, Edinburgh, UK.

[Richmond, 2006] Richmond, K. (2006). A trajectory mixture density neural network for the acoustic-articulatory inversion mapping. In *Proc. Interspeech*, pages 577–580, Pittsburgh, PA, USA.

[Richmond, 2009] Richmond, K. (2009). Preliminary inversion mapping results with a new EMA corpus. In *Interspeech*, Brighton, UK.

[Richmond et al., 2003] Richmond, K., King, S., and Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, 17:153–172.

[Rudzicz, 2010] Rudzicz, F. (2010). Correcting errors in speech recognition with articulatory dynamics. In *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49.

[Savariaux et al., 2012] Savariaux, C., Badin, P., Ouni, S., and Wrobel-Dautcourt, B. (2012). Étude comparée de la précision de mesure des systèmes d'articulographie électromagnétique 3d wave et ag500. *Actes des 29èmes Journées d'Etude de la Parole*, pages 513–520.

[Schölkopf and Smola, 2001] Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

[Schroeter et al., 1990] Schroeter, J., Meyer, P., and Parthasarathy, S. (1990). Evaluation of improved articulatory codebooks and codebook access distance measures. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 393 –396 vol.1.

[Schroeter and Sondhi, 1994] Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. on Speech and Audio Processing*, 2(1, Part. II):133–150.

[Sondhi, 2002] Sondhi, M. M. (2002). Articulatory modeling: a possible role in concatenative text-to-speech synthesis. In *Proc. IEEE Workshop on Speech Synthesis*, pages 73–78.

[Soquet et al., 1991] Soquet, A., Saerens, M., and Jospa, P. (1991). Acoustic-articulatory inversion based on a neural controller of a vocal tract model: further results. In T. Kohonen, K. Mokisara, O. S. and Kangas, J., editors, *Artificial Neural Networks*, pages 371–376. North Holland: Elsevier.

[Steiner et al., 2011a] Steiner, I., Ouni, S., et al. (2011a). Investigating articulatory differences between upright and supine posture using 3d ema. In *9th International Seminar on Speech Production (ISSP'11)*, Montreal, Canada.

[Steiner et al., 2011b] Steiner, I., Ouni, S., et al. (2011b). Towards an articulatory tongue model using 3d ema. In *9th International Seminar on Speech Production-ISSP'11*, pages 147–154.

[Steiner et al., 2013] Steiner, I., Richmond, K., and Ouni, S. (2013). Speech animation using electromagnetic articulography as motion capture data. In *AVSP 2013 - International Conference on Auditory-Visual Speech Processing*.

[Steiner et al., 2010] Steiner, I., Schröder, M., Charfuelan, M., and Klepp, A. (2010). Symbolic vs. acoustics-based style control for expressive unit selection. In *Proceedings of Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*.

[Stevens, 1972] Stevens, K. (1972). *Human communication: A unified view*, pages 51–66. McGraw Hill, New York.

[Stevens, 1989] Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17:3–45.

[Strom et al., 2006] Strom, V., Clark, R., and King, S. (2006). Expressive prosody for unit-selection speech synthesis. In *Proceedings of Interspeech*.

[Sumby and Pollack, 1954] Sumby, W. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212.

[Summerfield, 1979] Summerfield, A. (1979). Use of visual information for phonetic perception. *Phonetica*, 36(4-5):314–331.

[Suzuki et al., 1998] Suzuki, S., Okadome, T., and Honda, M. (1998). Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 2251–2254, Sydney, Australia.

[Taylor, 2009] Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge Univ. Press.

[Theobald, 2007a] Theobald, B. (2007a). Audiovisual speech synthesis. In *International Congress on Phonetic Sciences*, pages 285–290.

[Theobald, 2007b] Theobald, B.-J. (2007b). Audiovisual speech synthesis. In *ICPhS*, Saarbrucken, Germany.

[Tiede, 2010] Tiede, M. K. (2010). MVIEW: Multi-channel visualization application for displaying dynamic sensor movements.

[Toda et al., 2008] Toda, T., Black, A., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3):215–227.

[Toutios and Margaritis, 2008] Toutios, A. and Margaritis, K. (2008). Contribution to statistical acoustic-to-EMA mapping. In *16th European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland.

[Toutios et al., 2011a] Toutios, A., Musti, U., Ouni, S., and Colotte, V. (2011a). Weight Optimization for Bimodal Unit-Selection Talking Head Synthesis. In ISCA, editor, *12thAnnual Conference of the International Speech Communication Association - Interspeech 2011*, Florence, Italie.

[Toutios et al., 2010a] Toutios, A., Musti, U., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. (2010a). Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units. In *Interspeech*, Makuhari, Japan.

[Toutios et al., 2010b] Toutios, A., Musti, U., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., and Berger, M.-O. (2010b). Towards a true acoustic-visual speech synthesis. In *9th International Conference on Auditory-Visual Speech Processing-AVSP2010*.

[Toutios et al., 2011b] Toutios, A., Ouni, S., et al. (2011b). Predicting tongue positions from acoustics and facial features. In *Interspeech 2011*, pages 2661–2664, Florence, Italy.

[Toutios et al., 2008] Toutios, A., Ouni, S., and Laprie, Y. (2008). Protocol for a model-based evaluation of a dynamic acoustic-to-articulatory inversion method using electromagnetic articulography. In *The eighth International Seminar on Speech Production (ISSP'08)*, pages 317–320, Strasbourg, France.

[Toutios et al., 2011c] Toutios, A., Ouni, S., and Laprie, Y. (2011c). Estimating the Control parameters of an Articulatory Model from Electromagnetic Articulograph Data. *The Journal of the Acoustical Society of America*, 129(5):3245–3257.

[Tsang et al., 2006] Tsang, I., Kwok, J., and Zurada, J. (2006). Generalized core vector machines. *Neural Networks, IEEE Transactions on*, 17(5):1126–1140.

[Tulving, 1972] Tulving, E. (1972). Episodic and semantic memory. *Organization of Memory*, pages 381–402.

[Wang et al., 2012] Wang, L., Qian, Y., Scott, M. R., Chen, G., and Soong, F. K. (2012). Computer-assisted audiovisual language learning. *Computer*, 45(6):38–47.

[Watkins et al., 2003] Watkins, K. E., Strafella, A. P., Paus, T., et al. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994.

[Wik and Hjalmarsson, 2009] Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication*, 51(10):1024–1037.

[Wrench and Hardcastle, 2000] Wrench, A. and Hardcastle, W. (2000). A multichannel articulatory database and its application for automatic speech recognition. In *Proc. 5th International Seminar on Speech Production*, pages 205–308, Kloster Seeon, Bavaria.

[Wrench and Richmond, 2000] Wrench, A. and Richmond, K. (2000). Continuous speech recognition using articulatory data. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 145–148, Beijing, China.

[Wrobel-Dautcourt et al., 2005] Wrobel-Dautcourt, B., Berger, M., Potard, B., Laprie, Y., and Ouni, S. (2005). A low-cost stereovision based system for acquisition of visible articulatory data. In *AVSP*, British Columbia, Canada.

[Yehia et al., 1998] Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43.

[Yehia et al., 2002] Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30(3):555–568.

[Zhang and Renals, 2008] Zhang, L. and Renals, S. (2008). Acoustic-articulatory modeling with the trajectory HMM. *Signal Processing Letters*, 15:245–248.