Comparative study of linear and non-linear models for viseme inversion: Modeling of a cortical associative function

Frédéric Berthommier
Institut de la Communication Parlée, INPG
46 Av. Félix Viallet, Grenoble, France
e-mail : bertho@icp.inpg.fr

Abstract
The strong association existing between the audio speech features and the state of mouth opening is exploited for inversion in a comparative framework, using linear and non-linear models. At first, an associative map between an array of visemes and the audio features is constructed following a statistical learning process. The visemic mapping is self-organized and after convergence, the conditional mean of audio features is associated to each of them. Since the viseme states form a 2-dimensional continuum, the principle of the non linear inversion models is to drive a continuous trajectory across the output space, using less continuous audio inputs. Two strategies are proposed in order to smooth the output sequence. The first one consists in filtering (reshaping) the input trajectory and the second one is the driving of a traveling wave. A comparative study including linear and non linear models shows that the second strategy is plausible for modeling an associative cortical function.

1. Introduction
The inversion problem was tackled ten years ago by ICP researchers in the context of the EC SpeechMars project. The aim of this project was to build a plausible model, the articulation, able to learn articulatory movements from speech sounds, and to integrate knowledge about the format of sensori-motor representations as well as observations about perceptual processes. Among others, a computational model was proposed by Bailly [1]. This was based on 3 different phonological representations situated at the cortical level. After the finding of the role of low-level coherence between audio and video signals for speech detection [6,8] and improvement of the audio-visual intelligibility [10], a model of low-level audio-visual interaction has been proposed [3], retaining the principle of linear association described by Yehia et al. [12]. This model mainly exploited the inference of audio data from a real video recording and there was a lack of an alternative technique for quantifying the performance of the reciprocal inference making the inversion from speech sounds to lips movements. A new retro-marking method [4], presented again in this paper, authorizes a detailed quantification of the inversion performance and guides the development of non-linear inversion techniques. In this domain, non linear modeling is a priori a way of improvement, as we experimented in [2], but the point of view defended here is that the linear statistical model suffers of a lack of representational support, and that the non linear characteristic is contingent. The characteristics of the models which are derived from this representational point of view are highly suggestive about the functioning of associative cortical areas, and a parallel with the most recent advances about the support of consciousness [5,9] is also reasonable.

2. Material and methods
2.1 Database
The database was recorded for developing an audio-visual speech recognition system [7] based on natural images. This is a repetition of a subset of Numbers95 (OGI) by a single female speaker, in a soundproof room condition. The mouth region is well centered and fixed. In the database content, 64*78RGB images are available after spatial sub-sampling, at 50fps. The audio is down-sampled from 22KHz to 11KHz. The dataset is composed of about 40000 BMP images taken in the same continuous sequence. The first half is used for training and a part of the second half (about 6 min) is what we call the test database. The speech sequences are composed of groups (sentences) of English digits and numbers, among 30 different words, separated by silence periods, during which the speaker’s mouth stays close.

The three audio representations are (1) the spectrally coarse Sb4, using the output amplitude expressed in dB RMS of a filterbank composed of four quasi-rectangular filters (nb=4) as in [3,4], (2) the DCT of a 16-filterbank (nb=16), and (3) the 12 Line Spectral Pairs-RMS amplitude (nb=25). These audio features X are extracted in 40ms half overlapping hamming windows synchronized with the video. The video parameters are extracted using the full DCT of the initial 64*78 images stored in the database, converted in gray-levels. The full block of the first 2*12 (288) DCT values has been selected for video parameters, in order to maximize the content of available information.

2.2 The retro-marking technique
Following the so called geometrical approach of video parameterization, the ABS parameters (6 parameters, Figure 1. bottom) describe the state of mouth opening and these were classically used at ICP thanks to their excellent properties for carrying explicitly the lip-reading information. Consistently, our first research about the acoustic-visual associations involved this type of parameterization [2]. Practically, the extraction of ABS parameters from natural images requires a quite complex processing, susceptible to also introduce artifacts. In order to recover ABS parameters from the unmarked images of our video database, a new ‘retro-marking’ technique is proposed (Figure 1). A transformation function is established relating the DCT and ABS parameter spaces, via a Self Organizing Map (SOM). The SOM algorithm was applied on the training section of the video database, with 288 DCT parameters and 10*10 output vectors. Despite the high number of input parameters, the convergence was easy to reach, showing that the intrinsic dimensionality is very low (~ 2). Then, the 10*10 components of the bi-dimensional map (SOM DCT) were transformed by inverse DCT and smoothing for visualization. This map is composed, as the input database, of about half of close mouth states corresponding to silent periods. A first axe noted Oi (Opening index, horizontal in Figure 1, represents the degree of mouth opening and the second one the mouth rounding, noted Rj (Rounding index, vertical). The visualization of the 100 components allows drawing by hand on each of them the 8 points needed for defining the 6 ABS parameters and building a second map (SOM ABS, Figure 1). Then the transformation function can be applied over any section of the database because a common SOM label V(Oi,Rj) is attributed to each frame using the SOM DCT. The related ABS parameters are selected from the SOM ABS using these labels. Particularly, the training section is marked retroactively (suggesting the name of the method: ‘retro-marking’).
3. Linear and non linear models

3.1 Linear model

Initially proposed by Yehia et al. [12], the linear model allows a prediction of audio to video parameters and conversely, after a simple training step. In [3], this model was symmetrically applied for predicting temporal envelopes and speech enhancement, as well as for the direct estimation of audio video parameters. These are not useful for a performance quantification based on the comparison between estimated and effective video frames. Using the new retro-marking technique, we will test the prediction of the AB parameters (S, S' are excluded) which carry the main information about the mouth opening state. For each of the 3 types of audio parameters previously defined, the linear transformation matrix $T_0$, relating audio data $X$ to video data $Y$, is estimated from the two aligned data sets of the training section (the set of four AB parameters was derived using the retro-marking technique):

$$T_0 = (Y - \mu_Y)(X - \mu_X)^T(X - \mu_X)^{-1}$$

The size of $T_0$ is $nbp \times 4$, nbp the number of audio parameters. The means are calculated over the training section of the database. Then, the estimation of the four video parameters per frame, at 50 fps, using the audio channel of the test section ($X$), follows a linear rule (in which the video mean is derived from the training section, and the audio mean from the current test signal):

$$\hat{Y} = T_0(X - \mu_X) + \mu_Y$$

A pre-filtering is applied on all audio parameters, with a cutoff frequency at 5 Hz, for smoothing in the vocal tract motion range. The linear model exploits the significant correlations existing between audio and video parameters in this low frequency domain.

3.2 Model 1: reshaping trajectories

For non linear modeling, the aim is to incorporate the SOM representation in the prediction process itself, and to use the map as an associator between audio inputs and discrete mouth opening states. The introduction of an intermediate representation disrupts the direct relationship existing between audio and video parameter spaces, but has the double interest (1) to allow a better control of the association process, and (2) to substitute to the output representation a map of pointers, and then to manipulate these pointers. The challenge is to compensate the disruption, and the approach of Model 1 is to regularize the output trajectories, knowing that the SOM representation is composed of neighbor states.

For each of the hundred nodes of the SOM, a SOM audio state $X(Oi,Ri)$ is evaluated during the learning stage. The conditional mean of audio data is calculated over the training database at 50 fps, knowing $V(Oi,Ri)$:

$$X(Oi,Ri) = \text{mean}(X|V(Oi,Ri)) - \mu$$

As for the linear model, the audio data is centered using the mean derived from the training section.

For testing this mapping between audio and video states (Figure 2), the audio frames $X$ are filtered and centered (now using the audio mean of the test data). The non linear ‘winner take all’ rule is applied to estimate an index $A(Oi,Ri)$, with an Euclidean distance function:

$$A(Oi,Ri) = \text{argmin}(\text{dist}(X - \mu, X(Oi,Ri)))$$

The whole distance matrix of a silence frame is shown Figure 3 (the z axis is negatively oriented for having a high level for a small distance). About half of the cells (on the right part of the map) have a high response and this is consistent with the mouth closure observed in the SOM ABS map (Figure 1).

The 40000 cells of the SOM ABS map (Figure 1) show that for each cell of the SOM (50x50), each of the 4 parameters is calculated (50, 45, 40, 35, 30, 25, 20, 15, 10, 5, 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50) and the silence frame is shown Figure 3 (the z axis is negatively oriented for having a high level for a small distance). About half of the cells (on the right part of the map) have a high response and this is consistent with the mouth closure observed in the SOM ABS map (Figure 1).

Figure 2: Block diagram of the model

Figure 3: Distance matrix of SOM audio (Sb4) for a silence frame, smoothed for showing a simulation of the model (see on website).

The Figure 4 (top) shows that the trajectory of $A(Oi,Ri)$ is apparently chaotic. Consistently, with a statistic of Euclidean distance between successive locations, Figure 8, we find that the mean of Audio natural jump is about 4 cells. In order to regularize this trajectory, a post-processing is applied on $A(Oi,Ri)$, having two supplementary stages (Figure 2): Anchorage and Index filtering. The anchorage redirects the close mouth states (about half of the video states) to a single one, which is fixed in the center of the array. The goal of the anchorage step is (1) to reduce the great variability due to the one-to-many relationship between the (non invertible) audio silence state and the many video close mouth states and, (2) at the same time to initialize trajectories in the center location. Then, to reshape speech trajectories, the two estimates of Oi and Ri are filtered independently by moving average, and rounded to discrete integer values.
3.3 Model 2: dynamical generation

This regularization process of the non linear estimates has the consequence to blur the fine relationships existing between audio states and mouth opening states. It does not take into account well the overall spectro-temporal structure of the audiovisual association which is the main point emphasized after the application of linear models [3]. The aim of Model 2 is to preserve this structure by generating the trajectories dynamically, using the same mapping and the same audio inputs as for Model 1. The principle of model 2 (Figure 5 and 6) is to detect a priori the audio silence state, and to initialize and reset an iterative search of the local minimum within a neighborhood Nd:

\[ E_{\text{r}}(O,R) = \arg\min \text{dist}(X_{0:1}-\mu; N_d(X(O,R),E(O,R))) \]

Let remark that pre- and post-filtering are avoided in this scheme, and that the smoothness constraint is implicitly applied via the choice of an appropriate neighborhood Nd, which limits the possible jumps. The knowledge of the overall structure is also introduced here with a simple criterion. An analysis of the 10*10 SOM audio map (Sb4) indicates that the opening dimension is related to the overall amplitude level, and that the rounding dimension is related to the spectral shape (mainly the slope). Moreover, we assume that the pattern of variation of speech amplitude is quite regular (fast attack, slow decay), whereas the spectral variations depend on more complex factors (degree of co-articulation, stress, language etc.). Then, at 50 fps, the jumps are limited to \( E(O,R) = \{1, 2\} \) along the opening (temporal) dimension. A free parameter \( d \) is introduced, which limits the neighborhood at \( E(O,R) = \{1-d, d\} \) along the rounding (spectral) dimension. Simulations are performed with \( d=1 \) and \( d=5 \), and we observe Figure 8 that the mean/std jump of the ‘Video Natural’ is well restored with Model 2.

4. Evaluation

4.1 Quantitative comparison for test database inversion

The quantitative performance comparison of these models involves three types of audio parameters because differences are suspected. For the speech enhancement task [3, 4], a representation favouring the temporal dimension (Sb4) has been found to outperform a representation favouring the spectral dimension (e.g., LSP). In Figure 7, we see that the LSP is the best representation for feeding the linear model (Corr. Coeff. top, and % Rel. Diff., bottom, vary in inverse direction), particularly when the (audio) silence is included. Remarkably, the advantage of the LSP is persistent with the Model 1, but it disappears with the Model 2. Globally, the inversion performance of Model 1 is better to this of the linear model, but the Model 2 is clearly worse in the ‘silence excluded’ condition. The Model 2 is less sensitive to the input format, and less adapted for the inversion task, because the local search causes a loss of information, but it could preserve the main informative (phonetic) content.
4.2 Simulations using synthetic speech

The previous results are difficult to confirm out of our test database, because this needs a complete processing of a new audio-visual database. But this is possible to apply these models with any audio speech source, tentatively across different languages. To test the performance in stable conditions and with a large vocabulary, a new audio database, composed of 400 different di-syllabic French words, has been synthesized with the TTS Speechstim®. The measure of the average/std jump of \( A(O|R_i) \) is similar to this of the natural speech database (Audio synthetic vs. natural, Figure 8), but the response histograms (Figure 9) show that the input distribution \( A(O|R_i) \), noted Audio) is preserved only with the model 2 and \( d=5 \). Otherwise, this is blurred, and the distribution of Model 2 with \( d=1 \) has two clear branches as explained Figure 6. Finally, the full video synthesis process is based on the pointing of stored images (as proposed by Sacks [9] for explaining the mental representation of movement), and we observe strong co-articulation effects between successive movements: the resultant video speech is coherent temporally, but sometimes gets a wrong place of articulation.

Acknowledgements: This work is a part of the CTI-STIC project "Etude psychoacoustique et modélisation computationnelle des mécanismes de décodage acoustico-phonétiques à partir de la parole dégradée spectralement et temporellement". I thank L. Rebult, M. Heckmann and C. Savariaux for the elaboration of the audio-visual database. For having access to the multimedia content of this paper, see http://www.escp.mpj.fr/~bertho/02/p/icsip04/icsip04.html.

Figure 7: Comparative study of the inversion of the four AB parameters from LSP, DCT and Sh4. Top: correlation coefficient between estimate and test AB components; Bottom: % absolute differences between estimate and test database AB vectors.

Figure 8: Comparative measure of the average/std jump across the SOM; 'no jump' excluded; with Sh4 only.

Audio-Visible response distributions for French syllable words.

Figure 9: Comparative response histograms (SOM axes, smoothed) obtained with synthetic French Fournier's lists (400 words); silence state excluded; audio sh4. In the bottom left graph, the schwa [i] has been indicated as a possible audio-visual state, which is preserved by Model 2 (\( d=5 \)) among others.

5. References


