

SEEING TONGUE MOVEMENTS FROM OUTSIDE

G. Bailly & P. Badin

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
{bailly,badin}@icp.inpg.fr

ABSTRACT

Previous studies [10, 11, 20] have shown that a part of the tongue posture variance may be recovered from visible facial movement. Confronting articulatory models of the vocal tract (VT) and of the facial movements to cineradiographic data for the same subject, we examine the visible consequences of speech articulation, and conversely we determined the characteristics of the VT that can directly and robustly be captured from facial movements.

1 INTRODUCTION

Multiple perceptual and developmental studies have shown that speech is clearly multimodal. If hearing-impaired can naturally compensate for their handicap with visual information, we all use visual information in a noisy environment [9, 19]. Even in case of perfect listening conditions, audiovisual integration helps the comprehension of a foreign language or of a semantically difficult passage [15]. We cannot simply escape from this integration [12] and data comparing the development of normally hearing children with deaf [18] or blind [13] children show that perception and production abilities are largely affected by a lack of exposure to multimodal stimuli. These results tend to show that modalities are both *complementary* and *redundant*. While each modality offers a more robust encoding of specific cues – e.g. the acoustic channel carries articulation mode better than place [12, 17], labial information is better carried by vision whereas lingual articulation is better carried by acoustics – phonetic information is also redundantly specified. Robert-Ribes *et al.* [17] demonstrate for example that all articulatory features are better transmitted in the audiovisual mode than in any unimodal communication. Quantitative estimation of redundant information delivered by each modality is thus of most importance to provide a reliable basis (1) for models of multimodal fusion that helps us control our own speech movements and identify those from others, (2) for estimating the deficits generated by the loss of a given modality.

In the following, we aim at estimating the quantity of information that may be recovered from vision only, and at estimating the expected precision of the recovery of VT movements from facial ones. A particular emphasis is placed on place and degree of lingual constriction.

2 PREVIOUS STUDIES

The initial experiments conducted at ATR [11, 20] characterized facial and tongue movements by the movements of respectively 12 to 18 OPTOTRAK infrared markers glued on the subject's face and 4 EMMA coils. Spatial trajectories of these fleshpoints were recorded during two different sessions where the speaker uttered the same target sentences. A subsequent alignment procedure using the acoustic signal and lip markers common to both experiments resulted in a correlation between common measures

superior to 0.93. The experiment conducted by Jiang *et al.* [10] consisted in a *simultaneous* recording of 18 infrared markers and 3 coils.

Despite different experimental protocols, these studies converge towards a correlation between the measured characteristics of the tongue shapes and multilinear predictors from facial movement ranging from 0.6 up to 0.8, the tongue tip movements being less predictable than the rear part of the tongue. Though these correlation values are surprisingly high, the authors do not propose any solid interpretation of what is really recovered and of what the acoustic consequences of these estimations are.

3 DATA AND METHODOLOGY

If the movements of a few fleshpoints provide interesting information on biomechanical deformation of the organ, they provide only partial and imprecise (especially for EMMA) information on the actual shape of the entire organ. Then the geometric-to-acoustic transform is highly nonlinear and errors in geometric estimation may – or may not – have drastic consequences depending on specific features of the VT configuration, e.g. position and cross-sectional area of the tongue constriction as well as the positioning of other articulators such as jaw, larynx or velum.

Cineradiographic data [6, 7] are used here to have access to the precise and detailed geometry of all relevant speech organs. Visible movements are accessed by means of the deformations of the speaker's profile whereas VT tract shapes are characterized by mid-sagittal contours of the different speech organs.

The dense information provided by this rich dataset of the basic allophonic variations of French vowels and consonants uttered by a male speaker was further characterized by articulatory modeling.

3.1 Facial articulatory model

A facial articulatory model has been developed for that speaker using an accurate and detailed cloning methodology developed at ICP [3, 5]: we capture the 3D trajectories of 168 colored beads glued on the subject's face by a photogrammetry procedure using two cameras and two mirrors. The movement of 30 additional points was determined by fitting manually a generic model of lips [16] on the images. An iterative linear analysis was performed on 33 target configurations. 6 linear predictors explained more than 97% of the variance of the 198 3D coordinates. Thanks to a jaw splint, the actual jaw height and protrusion are measured and imposed as two of 6 linear predictors (j_1 and j_2). The other predictors are lips rounding/spreading (11), lower lip raising/lowering (12), upper lip raising/lowering (13) and throat expanding/retracting (s_1).

3.2 VT articulatory model

The cineradiographic data have been initially used to build a complete midsagittal articulatory model including the ability to

compute sound from articulatory movements [6, 7]. The deformation of a mobile grid intersecting VT walls and speech organs in 70 points determines the length and areas of 35 elementary acoustic tubes. An iterative linear analysis was performed on 1222 configurations: 9 linear predictors – including jaw height (JH semantically identical to the facial j1), larynx height (LY), 4 intrinsic degrees-of-freedom for the tongue (TB, TD, TT and TA) and 3 for the lip tube (LH, LP, LV) - explained more than 96% of the variance of the 70 sagittal coordinates. Note that a model trained using only 22 targets has almost the same explanatory power and precision than the one trained on the entire data set [7]. Note also that a further analysis of 3D tongue geometry using MRI data [2, 3] revealed that no additional predictors were necessary to predict lateral tongue movements for speech.

3.3 Coordinate reference frame

All data are referenced to a unique coordinate system bound to the skull. The origin of the system is set midsagittally at the lower edge of the upper incisors. The xz plane is taken as the bite (or occlusal) plane. For facial data, this plane is materialized by a Plexiglass plate in which the upper dental cast has been imprinted. This plate is used by the subject in all movement capture experiments. For X-ray data, a reference tracing of the upper incisor and hard palate is used to compensate for head movements. Again the same dental cast enable us to relate this landmark to the reference bite plane.

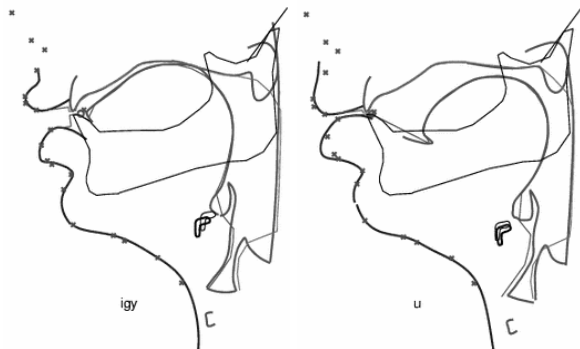


Figure 1: Adjusting the profile of the 3D facial model and the mid-sagittal VT model to X-ray data (here at the maximum occlusion in [igy] and at the center of realization of [u]). The original X-ray contours (dark) and the optimal mid-sagittal contour (light) are superposed with 3D facial fleshpoints of the 3D facial model closest to the profile (crosses). The original and reconstructed tongue contours are nearly indistinguishable except in the laryngeal region.

3.4 Analysis-by-synthesis procedure

Parameters of both facial and VT models are adjusted to the X-ray tracings of 45 target configurations chosen at the center realizations of all vocalic and consonantal allophones uttered during the cineradiography.

The average RMS reconstruction error of the tongue profile is less than 1mm except in the pharyngeal region where it raises to 1.5mm. 14 flesh points of the 3D facial model (including 6 points in the vermilion) closest to the face midline are selected (see crosses in Figure 1) and the 6 parameters of facial model optimized so as to minimize the distance between the X-ray

tracing of the speaker’s profile and the projection of these 14 points to the facial midline. The resulting mean RMS distance is 0.86 mm.

4 VT shape from facial movements

For each X-ray tracing we thus collected the values of 6 facial parameters and 9 VT parameters. These parameters describe, with a precision close to the millimeter, the geometry of the entire VT, the face as well the position of important articulators such as the jaw and the larynx. The individual correlation between each facial and VT parameters is given in Table 1. A multilinear prediction of each VT parameter from all visible elementary movements is also given.

4.1 Global results

These results evidence the quality of our coordinate reference frame, data acquisition and model construction: parameters common to both models are predicted with a correlation higher than 0.98. Although directly accessible here, lower teeth position is in fact not included in the 14 selected facial points: the almost perfect recovery of the true jaw height (JH) from the estimated optimal value (j1) evidences that jaw aperture can effectively be recovered from outside while a single point (e.g. chin) is obviously not sufficient.

Our data are quite in line with the studies already mentioned: the correlation coefficients for the intrinsic tongue shape parameters range from 0.37 to 0.74. These values would be even higher if jaw height had not been subtracted from the tongue shape variance. Note also that the whole tongue shape is considered here. The tongue dorsum (TD) and advance (TA) are the worst predictable parameters. TD, that bunches/flattens the tongue, is recruited for palatal/velar constriction. TA, that advances/retracts the apex, is also an important parameter for the fine control of coronal constrictions. Note finally that larynx height (LY) is also recovered-

	j1	l1	l2	l3	j2	s1	LR
LH	.50	-.02	.84	.83	-.08	.11	.99
LP	.13	.96	.34	.02	.09	.33	.98
JH	.99	.19	.44	.40	.04	.15	.99
TB	.24	.07	-.24	.01	.35	-.24	.71
TD	-.11	.20	.22	.18	-.50	-.12	.64
TT	.33	.34	.39	.37	-.01	-.24	.74
TA	.04	-.10	-.01	-.18	-.17	.03	.37
LY	-.00	.57	-.26	-.46	-.13	.25	.84
LV	-.00	.02	-.47	.55	-.26	-.50	.99

Table 1: From visible DoFs to underlying articulatory DoFs: correlation coefficients between individual parameters and between each articulatory DoF and a multilinear prediction (LR) using all visible DoFs. Tongue Advance (TA), Tongue Dorsum (TD) cannot be recovered from facial deformation. Correlations higher than 0.8 are highlighted in gray.

4.2 A closer view

To have a better inside view on the critical consequences of a loss of 0.3 on a correlation coefficient, we synthesized the predicted tongue shapes as displayed on Figure 2. Results are very heterogeneous: configurations associating a jaw/tongue/lips synergy along the axis closed/front (e.g. [i]) vs. open/back (e.g. [a]) are accurately recovered, while most configurations requir-

ing constrictions deviating from this synergy are poorly predicted. This later case includes most consonants in open contexts and velars in closed context as well as labialised vowels.

A more global landscape of the strong deterioration of maximal control and acoustic spaces of speech sounds is illustrated by Figure 3 and Figure 4. The characteristics of the predicted lingual constriction converge clearly along an axis closed/front vs. open/back while closed and protruded configurations centralize. Figure 4 shows that the acoustic consequence of these biases are drastic, especially for the formants of rounded vowels and consonant loci.

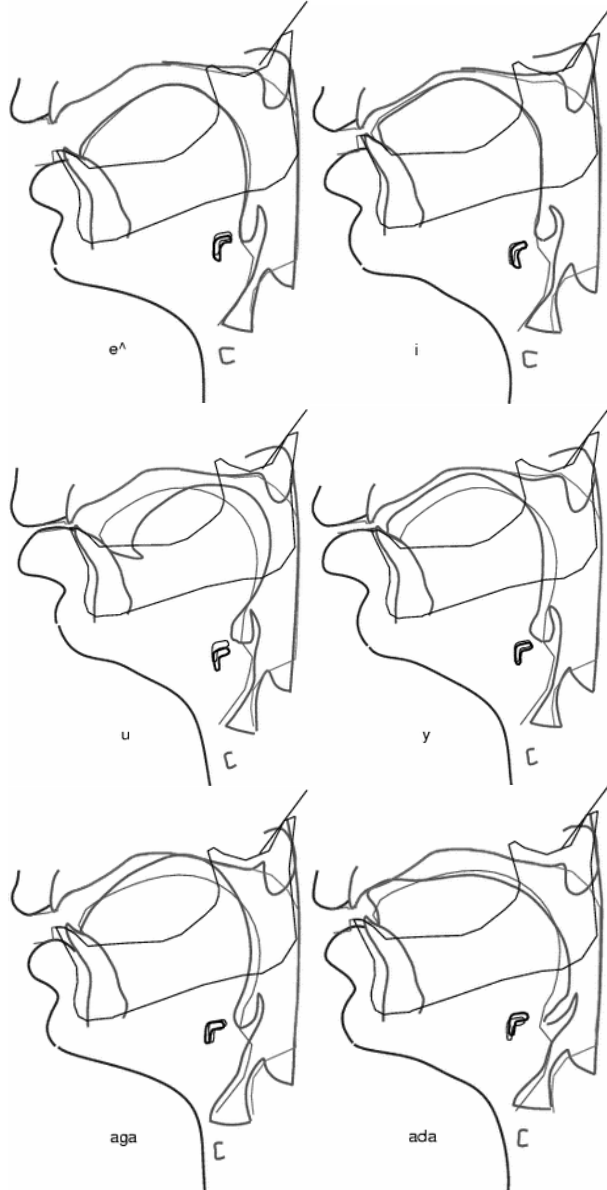


Figure 2: Predicting vocal tract configurations from the face. Top: two successful examples for vowels [e] and [i], that follow the general synergy open/back and closed/front. Middle: for the two labial doubles [u] and [y], the inverse model predicts quasi identical tongue shapes. Bottom: failing to predict vocal tract constriction for /g/ [aga] and /d/ in [ada]. Same conventions as for Figure 1.

5 COMMENTS

As a common frame for shaping vocal tract (labial/lingual) constrictions, the jaw is the most evident main supplier of redundancy between audible and visible movements: JH explains 16.7% of the variance of the 3D geometry of the tongue shape while the same parameter j1 explains 16.4% of the variance of the 3D facial geometry [3]. The rough *placing* of the tongue in the mouth is thus predictable using this clear redundant information together with more global lips/tongue synergies. What a predictor of tongue shape clearly misses is the final *shaping* of the tongue. As both placing and shaping [1] are needed to lead to the final relevant acoustic result, it can be concluded that the present prediction of tongue shape from face geometry may not be sufficient.

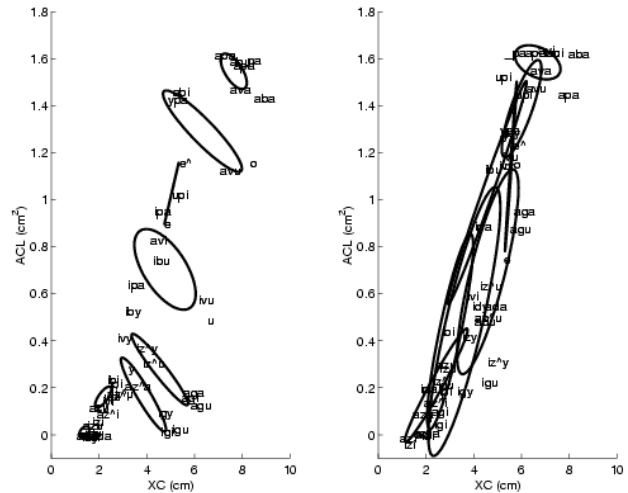


Figure 3: Characterizing the original (left) vs. recovered (right) lingual constriction. XC is the distance between the upper incisors and the main constriction, while ACL is the corresponding cross-sectional area. The mapping does not recover the constrictions occurring (mainly for [g]) in the velopalatal region and broaden the place of articulation of original front and mid-front articulations.

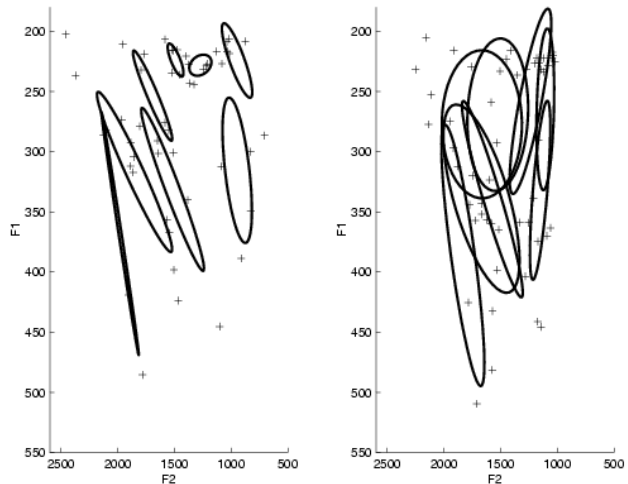


Figure 4: Characterizing the original (left) vs. recovered (right) acoustic targets. Except for vowels with a “neutral position” of the tongue, large deviations of predicted formant patterns can be observed.

An incomplete capture of essential facial correlates of tongue shaping could be put forward to invalidate our simulations. Imperceptible movements of the cheeks have been advocated for explaining the surprisingly “good” recovery of tongue shape [10, 20]. However, the analysis of the present data shows that correlation levels compatible with previously published results may be obtained without such subtle movements. We also point out that, for all the 3D facial models we have been built so far, jaw, lips and throat predictors explained more than 95% of the global variance, and that the residual variance of the cheeks never exceeded twice the measurements errors due to the experimental settings (camera calibration and bead tracking procedure). And this despite the fact that most analysed targets were hyperarticulated sustained articulations that should have exaggerated the resulting facial deformation.

Another criticism may concern the disparity between the nature of the characteristic measurements used to describe the face and VT: intersecting points between X-ray contours and a carefully designed deformable grid does not provide a direct access to fleshpoints. Badin *et al.* [4] have however shown that our articulatory model can accurately predict EMMA trajectories of tongue movements produced by this speaker.

6 CONCLUSIONS

The present data here suggest that visible characteristics of speech production do provide both redundant and complementary information on the sounds actually produced. Given the phonological structure of the language studied, visible movements have been shown to provide some information on the place of articulation of underlying speech organs. Jaw movements as well as larynx height could be reliably estimated from visual information. Visual information appears however insufficient to recover the proper lingual constriction. This confirms a posteriori the importance of the information provided by the manual cued speech [8] or gathered by the hand placement on the face in TADOMA [14] to supply lip reading with additional cues on place and mode of articulation.

REFERENCES

[1] Abry, C., Laboissière, R., Loevenbruck, H., Cathiard, M.A., and Schwartz, J.-L. (2000) *Glide production and control in the TWO-component vowel model*. in *5th Seminar on Speech Production: Models*. Kloster Seeon, Germany. p. 37-40.

[2] Badin, P., Bailly, G., Raybaudi, M., and Segebarth, C. (1998) *A three-dimensional linear articulatory model based on MRI data*. in *Proceedings of the International Conference on Speech and Language Processing*. Sydney, Australia. p. 417-420.

[3] Badin, P., Bailly, G., Revéret, L., Baciú, M., Segebarth, C., and Savariaux, C. (in print) *Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images*. *Journal of Phonetics*.

[4] Badin, P., Baricchi, E., and Vilain, A. (1997) *Determining tongue articulation: from discrete fleshpoints to continuous shadow*. in *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes - Greece. p. 47-50.

[5] Badin, P., Borel, P., Bailly, G., Revéret, L., Baciú, M., and Segebarth, C. (2000) *Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images*. in *Proceedings of the 5th*

Speech Production Seminar. Kloster Seeon - Germany. p. 261-264.

[6] Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.-P., and Brock, G. (1995) *Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model*. in *International Congress on Acoustics*. Trondheim - Norway. p. 349-352.

[7] Beautemps, D., Badin, P., and Bailly, G. (2001) *Degrees of freedom in speech production: analysis of cineradio- and labio-films data for a reference subject, and articulatory-acoustic modeling*. *Journal of the Acoustical Society of America*, **109**(5): p. 2165-2180.

[8] Cornett, R.O. and Daisey, M.E. (1992) *The cued speech resource book for parents of deaf children*. Raleigh, NC: The National Cued Speech Association, Inc. pages.

[9] Erber, N.P. (1975) *Auditory-visual perception of speech*. *Journal of Speech and Hearing Disorders*, **40**: p. 481-482.

[10] Jiang, J., Alwan, A., Bernstein, L., Keating, P., and Auer, E. (2000) *On the Correlation between facial movements, tongue movements and speech acoustics*. in *International Conference on Speech and Language Processing*. Beijing, China. p. 42-45.

[11] Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., and Yehia, H. (1999) *Audio-visual synthesis of talking faces from speech production correlates*. in *EuroSpeech*. p. 1279-1282.

[12] McGurk, H. and MacDonald, J. (1976) *Hearing lips and seeing voices*. *Nature*, **26**: p. 746-748.

[13] Mulford, R. (1988) *First words of the blind child*, in *The emergent lexicon: The child's development of a linguistic vocabulary*, M.D. Smith and J.L. Locke, Editors. Academic Press: New-York. p. 293-338.

[14] Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Delhorne, L.A., Braida, L.D., Pemberton, J.C., Mulcahey, B.D., and Washington, D.L. (1992) *Analytic study of the Tadoma method: improving performance through the use of supplementary tactual displays*. *Journal of Speech and Hearing Research*, **35**: p. 450-465.

[15] Reisberg, D., McLean, J., and Goldfield, A. (1987) *Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli*, in *Hearing by Eye: The Psychology of LipReading*, B. Dodd and R. Campbell, Editors. Lawrence Erlbaum Associates: Hillsdale, New Jersey. p. 97-113.

[16] Revéret, L. and Benoît, C. (1998) *A new 3D lip model for analysis and synthesis of lip motion in speech production*. in *Auditory-visual Speech Processing Workshop*. Terrigal, Australia. p. 207-212.

[17] Robert-Ribes, J., Schwartz, J.-L., Lallouache, T., and Escudier, P. (1998) *Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise*. *Journal of the Acoustical Society of America*, **103**(6): p. 3677-3689.

[18] Stoel-Gammon, C. (1988) *Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: A comparison of consonantal inventories*. *Journal of Speech Hearing Disorders*, **53**: p. 302-315.

[19] Sumbly, W.H. and Pollack, I. (1954) *Visual contribution to speech intelligibility in noise*. *Journal of the Acoustical Society of America*, **26**: p. 212-215.

[20] Yehia, H.C., Rubin, P.E., and Vatikiotis-Bateson, E. (1998) *Quantitative association of vocal-tract and facial behavior*. *Speech Communication*, **26**: p. 23-43.