

DETERMINING TONGUE ARTICULATION: FROM DISCRETE FLESHPOINTS TO CONTINUOUS SHADOW

Pierre Badin, Enrico Baricchi, & Anne Vilain

Institut de la Communication Parlée

46, Av. Félix Viallet, F-38031 Grenoble cedex 01, France

Tel.: +33 (0)4- 76.57.48.26 -- Fax: +33 (0)4- 76.57.47.10, E-mail: badin@icp.grenet.fr

ABSTRACT

The present study demonstrates the possibility to reconstruct complete midsagittal tongue shapes from the coordinates of three fleshpoints on the tongue and from the position of the jaw. The method is based on the inversion of an articulatory model made on the subject from cineradiographic images, and lead to an average reconstruction error of 1.26 mm.

1. CONTEXT AND OBJECTIVES

The development of coherent models of speech production calls for extended amounts of articulatory and acoustic data, with both high temporal and spatial resolutions. Cineradiography seems a good compromise between spatial and temporal resolutions: it provides continuous outlines of the vocal tract in the midsagittal plane at a reasonable – though a bit low – sampling frequency of 50 images per second. Other methods offer either better spatial information, such as the complete 3D images provided by MRI, at the cost of sampling frequencies of about 0.01 Hz (2 min. sustained articulations), or better temporal resolution, as the 400 Hz provided by electromagnetic articulometers on a reduced number of fleshpoints characterised by their X-Y coordinates. However, its use is extremely limited, due to its highly hazardous nature.

The aim of the present study was to show that it is possible to associate the benefit of both the good spatial resolution of cineradiography and the good temporal resolution of electromagnetic articulometers. Articulatory models rely on the fact that descriptions of tongue shape in terms of continuous contours are the redundant and can thus be reduced to a limited number of degrees of freedom, called articulatory parameters, that can be determined for instance by principal component analysis (cf. [2]). The present work exploits this basic concept, and postulates that the recovery of articulatory parameters can be performed without resorting to the measurement of the complete tongue shape, i.e. in other words, to reconstruct a continuous shadow of the tongue from a reduced number of tongue fleshpoints.

2. THE ARTICULATORY MODELS

The articulatory models used in this study are physiologically oriented linear statistical models of the vocal tract midsagittal contours, developed according to a methodology established by Maeda [4]. We use two such models: *Bergame*, developed by [2] based on a radiofilm of subject PB [1], and *Gentiane*, developed for this study, based a radiofilm of subject JLS.

The internal and external midsagittal vocal tract contours are represented by their intersections with a semipolar grid (cf. Fig. 1). In linear articulatory models, the abscissa of these intersection points, measured along the lines of the grid, are determined as linear combinations of

the articulatory command parameters. In particular, in the present models, tongue shape is entirely defined by six parameters: *jaw height* JH, *tongue body* TB, *tongue dorsum* TD, *tongue tip* TT, *tongue advance* TA, and *larynx height* LY. Some of the articulatory command parameters are merely centred and normalised versions of specific articulatory measurements, such as JH and LY, for *jawhei* and *larhei* respectively. Note that two parts of the grid are dynamically adjustable in order to follow the movements of the larynx, measured by parameter *larhei*, and of the *tongue tip* (in fact, the *tongue blade*, defined as the linguistic class *coronal articulation*), measured by parameter *tngadv*. Starting from the articulatory command parameters, the model thus establishes the grid limits, and predicts the articulatory measurements using the following equations:

$$\text{JAWHEI} = \text{jawhei_mean} + \text{JH} \times \text{jawhei_sd} \quad (\text{Eq. 1})$$

$$\text{LARHEI} = \text{larhei_mean} + \text{LY} \times \text{larhei_sd} \quad (\text{Eq. 2})$$

$$\text{TNGADV} = \text{tngadv_mean} + \text{Pred_TA} \times [\text{JH TB TD TA}] \quad (\text{Eq. 3})$$

where JAWHEI, LARHEI and TNGADV are the model reconstructed values of the measurements *jawhei*, *larhei* and *tngadv*, respectively, *_mean and *_sd the means and standard deviations of the corresponding parameters, and Pred_TA the vector of prediction coefficients for TNGADV.

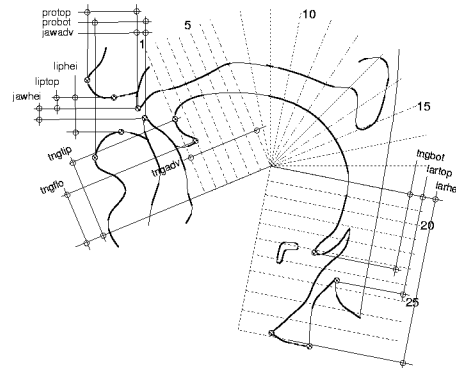


Fig. 1 – Example of midsagittal contours, semi-polar grid and articulatory measurements

It should be emphasised that TA was actually established as the predictor of the residues from *tngadv*, once the contributions of JH, TB, TD, and TT were removed. The relation between TA and *tngadv* depends thus on the other articulatory parameters. Finally, the tongue contour is constructed using the following linear combination:

$$\text{TNGCNT} = \text{tngcnt_mean} + \text{C_cnt} \times [\text{JH TB TD TT LY}] \quad (\text{Eq. 4})$$

where TNGCNT is the vector of tongue contour abscissa, *tngcnt_mean* the mean contour abscissa computed over the reference corpus, C_cnt a 4xN matrix of prediction coefficients that constitutes the core of the model (N is

the number of tongue contour points). The grid lines are numbered from 1 to N, from tongue tip to larynx, with N=27 for subject JLS and N=28 for subject PB.

3. THE EXPERIMENTAL DATA

The set of 1273 cineradiographic pictures acquired from subject JLS for this study¹ were processed in the same way as those of subject PB (cf. [1]), i.e. vocal tract midsagittal profile were traced by hand, digitised, and an number of articulatory measurements obtained (see Fig. 1). The crucial advantage of these new articulatory data is the presence of three small lead pellets glued on the subject's tongue along the midsagittal plane. The white spots generated by the pellets on the cineradiographic pictures were also manually traced, and the positions of their centres automatically measured. These three pellets will be referred to in the following as *tip pellet*, *mid pellet* and *back pellet*. The tip pellet was located at a distance of about 3.7 mm from the extremity of the tongue tip, while the distances were approximately 17.0 mm between tip pellet and mid pellet and 12.9 mm between mid pellet and back pellet.

Since the extremity of the tongue tip must be known to determine the grid position, we evaluated the possibility to predict it from the coordinates of the tip pellet. As expected, we found a strong correlation (-0.96) between *tngadv* and *Xplt_tip*, the horizontal coordinate of the tip pellet, and a negligible correlation with the vertical coordinate. We could finally establish, using linear regression analysis, the best prediction of *tngadv* by:

$$tngadv_pred = tngadv_mean - 0.90 \times Xplt_tip \quad (\text{Eq. 5})$$

The average reconstruction absolute error over the whole corpus was 1.07 mm (standard deviation 0.83 mm).

After this evaluation study, it is planned to acquire data for subjects JLS and PB with an articulometer equipped with five coils. One coil will serve as reference, and be attached to the anterior side of the upper incisor. A second coil will be attached to the lower incisors, and used to estimate jaw height. The three other coils will be glued on the tongue at approximately optimal positions.

4. THE INVERSION METHOD

4.1 Inversion principles

The aim of the inversion is the recovery of the articulatory parameters and thus the reconstruction of the tongue midsagittal shape from jaw height and three fleshpoints on the tongue. The parameter *jawhei* can be directly taken as the measure of the vertical position of the lower incisor edge on the jaw drawn from the radiofilm pictures. The three tongue fleshpoints are defined by the three lead pellets glued on the tongue.

Jaw height command JH can be determined directly from *jawhei* by inverting Eq. 1. Larynx height command LY can not be determined directly. Moreover, its contribution to the tongue contours (Eq. 4) is limited to the six grid lines closest to the glottis, and in particular is zero on the tongue points in the vicinity of the pellets. Therefore, it can not be recovered, and was set to zero. The consequences of this limitation are evaluated below.

Finally, a multi-objective goal attainment optimisation algorithm based on sequential quadratic programming

was used (AttGoal function from the MatlabTM package) to determine simultaneously the remaining four parameters TB, TD, TT and TA. Three goals consisted in minimising the Euclidean distances between the measured positions of the three pellets and the tongue contour produced by the articulatory model. The fourth goal was the minimisation of the difference between the TNGADV value produced by the model (Eq. 3) and the corresponding value *tngadv_pred* predicted from the coordinates of the tip pellet (Eq. 5).

4.2 Optimal conditions for the inversion

The location of the three pellets along the tongue midsagittal line is obviously determining for the precision of tongue reconstruction. In order to determine the optimal locations of the pellets, expressed in terms of grid line indices, simulation experiments were carried out. Six configurations of tongue typical of the speech material, [a i u ø z J], were used as a set of *synthetic test tongues*, i.e. tongues produced by articulatory models from known command parameters JH, TB, TD, TA, TT, LY. For each combination of three grid line indices, and for each test tongue, three *synthetic test pellets* were taken as the intersections of the test tongue with these lines. The optimisation algorithm was given as input: (1) the coordinates of the three test pellets, (2) the value of TNGADV computed by the model from the test command parameters. The search was initialised with these command parameters augmented with values randomly chosen in a range [-1 -0.5] or [+0.5 +1], ensuring a perturbation large enough to realise the test in fair and realistic conditions.

The accuracy of the inversion can be estimated by the difference between the test and recovered tongue contours. Two distances were thus defined: (1) a *point to point* vector of the distances between the N corresponding points of test and recovered tongues; (2) a *contour to contour* distance vector defined as the minimum between (i) the vectors of distances from the points of the test tongue to the smoothed contour of the recovered tongue and (ii) the vectors of distances from the points of the recovered tongue to the smoothed contour of the test tongue. The major difference between these two measures is that the latter does not take into account fleshpoints differences, but only tongue shadow differences, whereas the first one is also sensitive to the displacement of fleshpoints, even though tongue shadows are identical.

In order to test specifically the ability of the algorithm to recover the tongue shape, JH, and LY were given constant values corresponding to the test configurations.

In a preliminary evaluation, TA was also kept constant to its value in the test set, and only TB, TD and TT were recovered by the inversion algorithm. All possible combinations of pellet positions between 1 and N were explored. As expected, the lowest reconstruction errors were obtained when the tip pellet was close to the tip, the back pellet just above the epiglottis, and the mid pellet in the tongue dorsum region. These combinations were however unrealistic, and thus another exploration was carried out for a restricted number of combinations.

For this new evaluation, the tip pellet was limited to grid lines 1 to 4, and the back pellet to grid line 16 (about 60.0 mm from the extremity of the tongue for JLS, and 85.0 mm for PB). TA was also determined by the inversion algorithm simultaneously with TB, TD and TT. A number of good combinations were found for each of the four positions of tip pellet. However, we excluded

¹ Sequences of vowels [a i u y] and consonants [b d g J v].

position in grid line 1 because it would hamper too much the subject's articulation, and position 4 because it would reduce too much the accuracy of the prediction of *mgadv* needed in the inversion. Finally, the best choice for JLS was (2, 7, 11), leading to maximal point to point distances less than 0.22 mm for the limited test set, and was (2, 7, 12) for PB, corresponding to a maximal error of 0.39 mm. Note that the positions actually used for the radiofilm were about (2, 6, 9).

5. EVALUATION OF THE METHOD

5.1 Simulation tests on synthetic data

Once optimal pellet positions were found for each subject, a series of simulation test was carried out to evaluate the algorithm on the whole speech material.

The robustness of the inversion method in relation with the initialisation of the articulatory parameters TB, TD, TT and TA, was very high. For the inversion of speech sequences, the articulatory parameters found in the preceding frame were used to initialise the current search, so as to use temporal smoothing properties of the speech articulators to speed up algorithm convergence.

5.1.1 Basic inversion evaluation

The first simulation test aimed at extending the previous test to the whole corpus. Sequences of synthetic test tongues were thus extracted by inversion from the radiofilms (cf. [2] for details), and the corresponding TNGADV computed². Inversion was carried out, and reconstruction errors were computed. Table I gives the maximum *mx* and average *M* over the speech material for: (1) the average of the residual tongue-pellet distances³ *dtp*; (2) the RMS values *dpp* and *drm* of the vectors of point to point, respectively contour to contour, distances between test and reconstructed tongues; (3) the absolute differences between test and recovered TNGADV *ta*, and between articulatory parameters.

		ta	dtp	dpp	drm	TB	TD	TT	TA
JLS	mx	1.01	0.72	1.44	1.00	0.32	0.25	1.04	0.66
JLS	M	0.08	0.07	0.12	0.09	0.02	0.02	0.04	0.04
PB	mx	0.52	0.50	0.80	0.69	0.27	0.18	0.75	0.48
PB	M	0.07	0.07	0.12	0.11	0.03	0.03	0.07	0.05

Table I – Reconstruction errors for the global evaluation test (distances in mm)

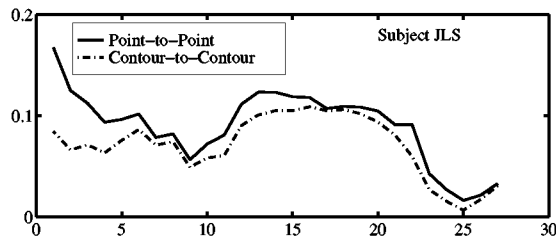


Fig. 2 – RMS distances between test and recovered tongues averaged over all configurations (in mm)

Reconstruction errors were rather small. Note in particular the low average RMS point to point distance of 0.12 mm for both subjects (the maxima for point to point distances were 4.16 mm for JLS and 1.43 mm for PB, while the contour to contour maximum distances

² Since parameter TA is defined as the residue of the *mgadv* measure, this TNGADV is identical to *mgadv*.

³ Note that the tongue-pellet distances were always found nearly identical for the three pellets.

were respectively 2.26 and 1.34 mm). The distribution of these errors along the tongue are displayed in Fig. 2.

5.1.2 Influence of *mgadv* prediction

The previous tests were carried out using a TNGADV parameter identical to the measured *mgadv*. In the real experiment, this measurement is not available, and the *mgadv*_pred predicted from the horizontal position of the tip pellet will have to be used. It was thus important to estimate the influence of the prediction errors of *mgadv*_pred on the accuracy of the inversion. The previous test has thus been replicated, using this prediction. The results, given in Table II, show that TB and TD are robust against perturbations on *mgadv*, but less so TT, and not TA. Globally, the reconstruction errors on the tongue are approximately doubled (compare Table I and II, and see Fig. 3).

	ta	d	dpp	drm	TB	TD	TT	TA
mx	7.37	1.83	3.24	1.62	0.37	0.93	1.49	2.59
M	0.90	0.10	0.44	0.22	0.04	0.04	0.20	0.30

Table II – Reconstruction errors for a realistic *mgadv* prediction (distances in mm)

5.1.3 Influence of LY

As seen above, LY can not be predicted in realistic conditions, and is thus set to zero. To evaluate the consequence of this, the evaluation described in section 5.1.2 was replicated, using LY = 0. Fig. 3 shows a considerable increase of the error in the larynx region, however below 1 mm. This relatively limited error can be explained by the fact that, even in the pharynx region, the major predictor of the tongue contour is JH, the factor LY explaining less than 15% of the total variance, except for the lowest two grid lines where it reaches about 50%. Note that the rest of the vocal tract is absolutely not perturbed. This localised error results in an increase of the mean RMS contour to contour error from 0.22 to 0.40 mm. Note finally that the major error on tongue contour due to a fixed LY is the larynx height error itself (the span of *larhei* is about 20 mm).

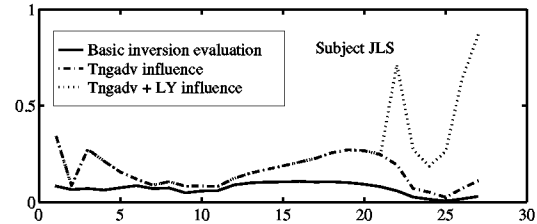


Fig. 3 – RMS contour to contour distances between test and recovered tongues averaged over all configurations (mm)

5.2 Tests on real data

The data acquired from the JLS corpus associate tongue shapes and pellets attached to fleshpoints, and thus offer the possibility to recover tongue contours from pellets coordinates by our inversion method, and to estimate the distance between the recovered contours and the contours directly determined from the original radiofilm tracings.

In this experiment on real data, the inversion algorithm was given the coordinates of the three pellets as input. It used *mgadv*_pred estimated from the tip pellet, and LY was set to zero. Results are summarised in Table III and Fig. 4. Three contour to contour distances were computed: *drx* between the recovered and the original contours, *drm* between the recovered and the test tongue

contours derived from the original contours (cf. above), and dxm between the original contours and the test contours. Fig. 4 shows large differences between recovered and original tongues on the lowest two sections of larynx. In this region the data are noisy, due to the difficulty to draw accurately the outlines, and thus the articulatory model does not fit the original contours so well. On the other hand, the differences between recovered and test tongue contours are larger than those between the recovered and test contours: this can be ascribed to problems in the inversion procedure used to establish the reference test contours.

		ta	d	drx	drm	dxm	TB	TD	TT	TA
org	mx	9.19	1.10	4.87	5.42	3.74	1.52	3.16	3.53	3.79
org	M	0.99	0.19	1.57	2.21	1.55	0.38	0.59	0.82	0.83
tst	mx	9.33	0.97	4.14	4.00	3.74	1.42	1.25	3.40	2.84
tst	M	0.98	0.12	1.74	0.99	1.55	0.24	0.21	0.68	0.43

Table III – Reconstruction errors on real data
(distances: mm; org=original pellets; tst=test back pellet)

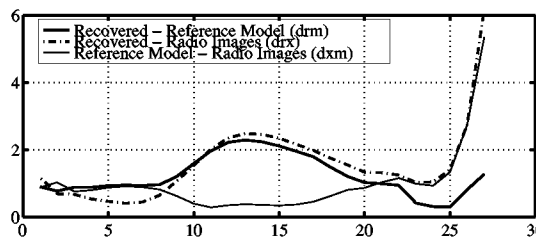


Fig. 4 – RMS contour to contour distances between original, test reference and recovered tongue contours averaged over all configurations (in mm)

Table III shows still acceptable errors on TB, but too large ones for TD, not mentioning TT and TA. Further analysis has shown that, in some cases, the recovered tongue is too much bunched backwards (cf. Fig. 6), likely because the back pellet was not positioned optimally, in particular not far enough towards the back of the tongue. To check this hypothesis, another test was carried out, where the real back pellet was replaced by a synthetic test pellet located on grid line 11. Table III and Fig. 5 show that this position, which is perfectly realistic, reduces considerably the reconstruction errors.

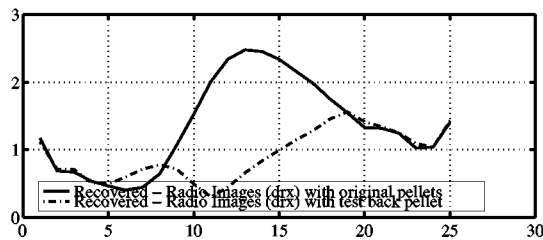


Fig. 5 – RMS contour to contour distances between recovered tongue contours (in mm) with different back pellet positions

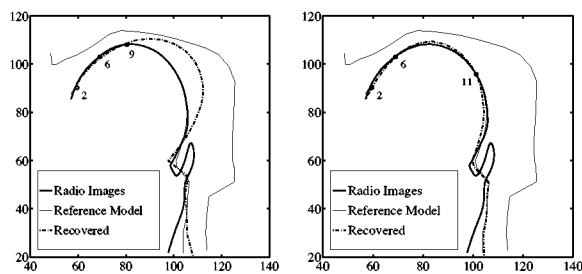


Fig. 6 – Effect of back pellet location

6. DISCUSSION AND CONCLUSION

We have shown the possibility and assessed the accuracy of the reconstruction of the whole tongue shape from three fleshpoints measured on the tongue and from the jaw position. The method is based on the inversion of an articulatory model established for each subject from a set of cineradiographic images. A similar study had been conducted by Kaburagi & Honda [3], who used an ultrasonic method to determine the tongue contours simultaneously with the coordinates of four fleshpoints tracked with an electromagnetic articulometer.

Our method involves making a radiofilm and establishing an articulatory model for the subjects that we study. This is obviously heavy to implement, compared to ultrasonic tongue imaging; however ultrasound techniques do not allow complete tongue images: the location of tongue tip in Kaburagi & Honda's figures seems very uncertain, and the contours in the pharynx region are limited to the equivalent of grid lines 15 or 16 in our models.

Another important difference is the use of an articulatory model to perform the inversion: Kaburagi & Honda use a simple multilinear regression to reconstruct tongue contours from pellet positions.

Kaburagi and Honda have found an *estimation error* of 1.24 mm, computed as the RMS value of all point to point distances of all tongue shapes in their speech material. With an equivalent measure, we have found, using the back pellet position in grid line 11 an estimation error of 1.80 mm if we consider the complete tongue contours, and of 1.26 mm when excluding the lowest two larynx points. The present method allows thus a more extensive reconstruction of the tongue shape with a comparable accuracy.

Our method provides the possibility to combine the advantages of both cineradiography and electromagnetic articulometry, and opens perspectives to acquire large amounts of articulatory data crucial for studies in speech motor control, articulatory to acoustic inversion and articulatory synthesis.

ACKNOWLEDGEMENTS

This work has been partially supported by a European LEONARDO grant to E. Baricchi. The radiofilms have been made in collaboration with the Strasbourg Phonet Inst. at the Schiltigheim Hosp. We sincerely thank people involved in this work at different degrees: C. Abry, D. Beutemps, G. Brock, B. Gabioud, Agn s Hennel, T.M. Lallouache, P. Simon, J.P. Zerling, and the subject J.L. Schwartz.

REFERENCES

- [1] Badin, P., Gabioud, B., Beutemps, et al., (1995) Cineradiography of VCV sequences: Articulatory-acoustic data for a speech production model. 15th ICA, IV, 349-352.
- [2] Beutemps, D., Badin, P., Bailly, et al., (1996) Evaluation of an articulatory-acoustic model based on a reference subject. 4th Speech Production Seminar, pp. 45-48. Autrans, France.
- [3] Kaburagi, T. and Honda, M. (1994). Determination of sagittal tongue shape from the positions of points on the tongue surface. JASA 96(3), 1356-1366.
- [4] 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. (Eds.), *Speech Production and Modelling*. (pp. 131-149). Kluwer: Acad. Publishers.