

Incorporation of phonetic constraints in acoustic-to-articulatory inversion

Blaise Potard, Yves Laprie, Slim Ouni

▶ **To cite this version:**

Blaise Potard, Yves Laprie, Slim Ouni. Incorporation of phonetic constraints in acoustic-to-articulatory inversion. Journal of the Acoustical Society of America, 2008, 123 (4), pp.2310-2323. 10.1121/1.2885747 . inria-00112226

HAL Id: inria-00112226
https://inria.hal.science/inria-00112226

Submitted on 10 Nov 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Incorporation of phonetic constraints in acoustic-to-articulatory inversion

Yves Laprie* and Blaise Potard†

Speech Team,

LORIA,

Vandœuvre-Lès-Nancy.

(Dated: November 7, 2006)

Abstract

This study investigates the use of constraints upon articulatory parameters derived from standard phonetic knowledge in the context of acoustic-to-articulatory inversion. These speaker independant "phonetic" constraints are introduced and investigated in an existing inversion framework. The validity of these constraints is assessed by comparing synthetic vocal tract shapes and real vocal tract shapes obtained from X-ray images. Beyond the scope of phonetic constraints, this study also provides an extensive exploration of the acoustical properties of Maeda's articulatory model.

I. INTRODUCTION

Acoustic-to-articulatory inversion remains an open challenge in speech analysis. Although there is a wide range of potential applications, there is as of yet no clear answer to whether or not inversion is possible for all the sounds of speech[19]. However, there do exist numerical simulations that cover both articulatory and acoustical phenomena involved in speech production and which enable the synthesis of acoustical artificial signals close to natural speech. These tools, especially those generating a speech spectrum, are often used to perform inversion. Indeed, most of the existing approaches to acoustic-to-articulatory inversion are analysis-by-synthesis methods.

The key difficulty is that an infinity of vocal tract shapes can produce any given spectrum. In order to reduce the number of inverse solutions, methods of acoustic-to-articulatory inversion incorporate explicit or implicit constraints.

Sorokin, for instance[20], presents seven possible kinds of constraints: limitations in the contractive force of muscles involved in speech production, anatomy of the vocal tract or equivalently, ranges of articulatory parameters, interdependencies between muscles, i.e. interdependent variations of the articulatory parameters, interdependency between transversal and mid-sagittal dimensions of the vocal tract, aerodynamic constraints with respect to the kinds of sound produced, level of the acoustical deviation tolerated between analyzed and resynthesized sounds according to style and rate of speech, and lastly, a constraint concerning the complexity of planning and programming of the articulatory control.

Some of these constraints, those upon articulatory parameters and, to a certain extent, those upon the transversal dimension estimated from the mid-sagittal profile of the vocal tract, can be incorporated directly in the analyzing model, in the form of an articulatory model. They can rely only on pure geometrical primitives, like those

*Electronic address: Yves.Laprie@loria.fr
†Electronic address: Blaise.Potard@loria.fr

of Coker[5] and Mermelstein[16]. Despite their flexibility and their ability to copy the natural vocal tract, geometrical models cannot render deformation modes of human vocal tracts. Therefore, models derived from X-ray images of a human vocal tract through a factor analysis procedure (Maeda[12] or Gabioud[8]), or more recently from MRI images,[6] are now preferred. Building articulatory models from a single speaker could be a strong limitation due to the difficulty of acquiring data for several speakers. However, deformation modes seem to be sufficiently speaker independent to ensure a satisfactory generality to these articulatory models. On the other hand, a prior size adjustment (the overall vocal tract length or both mouth and pharynx lengths) should be performed to adapt these articulatory models to a new speaker[9]. The adapted articulatory model imposes geometrical constraints which enable the dimension of the solution space to be reduced, down to seven for instance with Maeda's model. Even if the benefit of using an articulatory model is obvious it should be kept in mind that the corresponding constraints may be biased by the geometrical mismatch between the speaker retained to build the articulatory model, and the speaker who uttered speech to be inverted.

Other constraints proposed by Sorokin dealing with contractive forces, interdependencies between articulatory parameters or the complexity of the articulatory parameters are far too complex to be exploited because there is almost no data available. Indeed, they would require medical investigation technologies that do not exist yet, or that cannot be used easily, e.g. electromyography.

For this reason we investigate the incorporation of constraints upon articulatory parameters derived from standard phonetic knowledge. Their main advantage is to capture speaker independent constraints and a general human expertise about the use of a vocal tract absent from any articulatory/acoustical simulation of speech production. In order to obtain a precise assessment of the role of constraints, the acoustical properties of Maeda's articulatory model will be studied through the place and degree of constriction of vowels. Beyond the scope of this study about phonetic constraints, this

study provides an extensive exploration of the acoustical properties of an articulatory model.

Section II briefly presents the acoustic-to-articulatory method we previously developed. Then section III presents the phonetic constraints and their implementation. As it is important to get a good knowledge of the acoustical behavior of the articulatory model, section IV studies places of articulation and constriction areas of the vocal tract shapes recovered by inversion for French vowels. Then section V describes the derivation of compensatory effects from inversion results for five vowels.

II. PRESENTATION OF THE ACOUSTIC-TO-ARTICULATORY FRAMEWORK

Our inversion method[17] relies on Maeda's articulatory model which uses seven parameters to describe the shape of the vocal tract (see Fig.1). Each articulatory parameter varies between -3σ and 3σ where σ is the standard deviation. This model was obtained by analyzing X-ray images of a female speaker uttering small sentences. Only images corresponding to vowels were analyzed. The articulatory model and the acoustical simulation form the analyzing model used in our approach of inversion.

The articulatory-to-acoustic mapping is represented in the form of an articulatory table which associates vectors of articulatory parameters, i.e. 7-tuples in the case of Maeda's model, with their corresponding 3-tuples of the first three formant frequencies. This table thus represents the synthesis facet of the inversion. It is used to recover all the possible 7-tuples of articulatory parameters corresponding to the formant frequencies extracted from a vowel signal at each time frame. One crucial issue is the acoustical and articulatory resolution of such articulatory tables. The strength of our table lies in its quasi-uniform acoustic resolution. This property originates in the construction method of the table, which evaluates the linearity of the articulatory-to-acoustic mapping at each step. Unlike other methods used to represent the articulatory-to-acoustic mapping, this construction method ensures that no articulatory region compatible with

a given 3-tuple of formants will be forgotten (unless a very strong non-linearity in the mapping was missed during the exploration used during table construction).

Together with specific search algorithms this table enables the recovery of all 7-tuples (with respect to a prior acoustical/articulatory resolution) of articulatory parameters which can generate a given 3-tuple of formants.

If a speech segment has to be inverted so as to recover articulatory trajectories, a second stage is applied to sets of inverse solutions recovered at each time frame of the speech segment. This second stage consists of reconstructing articulatory trajectories that are sufficiently regular along time. This is achieved by a dynamic programming algorithm that minimizes a cost function that represents the overall "distance" covered by articulatory parameters. A final stage improves the articulatory regularity and the acoustical proximity of formants derived from the determined model vocal tract from the original formants measured on the speech segment. More details of the inversion method can be found in[17].

III. INCORPORATION OF PHONETIC CONSTRAINTS FOR VOWELS

Phonetic constraints are derived from standard phonetic knowledge[15] about the articulation of French vowels. This knowledge, and thus the expression of phonetic constraints, is about tongue dorsum position, mouth opening, lip stretching and protrusion. Each constraint is on one vowel, and consequently its relevancy depends on the vowel considered, or in a more general way, on an acoustic region in the formant space. Since the aim of our study is to derive constraints with very little speaker-specific data, the regions chosen are quite large. In order to account for the inter-speaker variability these constraints return numerical values, decreasing from one, when the constraint is perfectly satisfied, to zero.

Tab. I summarizes the phonetic description for the 10 non-nasal French vowels designed within the context of modeling labial coarticulation[18]. D stands for "tongue

dorsum position", O for "mouth opening", S for "lip stretching", and P for "lip protrusion". The coding is straightforward: the higher the number, the higher the value associated with the given constraint. For example, a constraint O_1 means that the mouth has a small opening and a value of O_4 means a very big opening. These data are average values of the way native French speakers articulate vowels, and thus may be different from the way a particular speaker articulates French sounds. Note that for the main place of articulation of vowels, corresponding to D in the case of vowels, the range of possible values is a sub-domain of the values acceptable for consonants (from 0 for /p,b,m/ to 9 for /ʁ, ɹ/). This explains why D only ranges betwee n 6 and 8 for vowels.

A. Translation of phonetic constraints in the articulatory model

In most articulatory models, translating simple phonetic features into parameters of the model can be quite complex. In our case we use Maeda's model[12], in which the parameters are easily interpretable from a phonetic point of view. Consequently, expressing phonetic constraints in terms of articulatory parameters is straightforward: lip protrusion and tongue dorsum position are already parameters of the model, and the mouth opening is a linear combination of two parameters (jaw position, and intrinsic lip opening).

Actually, the mouth opening constraint also uses the tongue position in order to take into account compensatory effects described in[13]: Maeda observed that for non-rounded vowels (/i/, /a/, /e/), the tongue position and the jaw opening had parallel effects on the acoustic image, and therefore were mutually compensating. He also observed that this compensatory effect was indeed used by his test subjects. Furthermore, it appeared that the direction of compensation did not depend on the vowel pronounced: there was a linear correlation

$$Tp + \alpha Jw = \text{ Constant}$$

where Tp is the tongue position, Jw the jaw position, and the α the linearity coefficient that is the same for both /a/ and /i/. The other vowels were not studied because there were not enough occurrences of them in the X-ray database. Maeda observed this compensation in both his subjects (but the coefficients of correlation were of course different). The coefficient we used for PB was the one Maeda found experimentally on X-ray data, which was approximately equal to 0.66. This compensatory effect allowed Maed a to explain most of the articulatory variability for /a/ and /i/.

The mouth opening is thus given by:

$$min(Tp + \alpha Jw, Lh)$$

where Lh is the lip aperture.

B. Acoustic space partitioning

An acoustic domain has to be defined for each phoneme, where the phonetic constraints are considered to be valid (i.e. a domain where articulatory configurations which respect the given constraints are likely to be observed). Since we are currently using the first three formant frequencies as data for acoustic-to-articulatory inversion, these domains are regions in a 3-D space. We tested different models for the partitioning of the acoustic space: Voronoi diagram around the vowels (cf. Fig. 2) and Voronoi diagram weighted by the standard deviation of each formant frequencies (cf. Fig. 3). The second gives slightly better results.

C. Phonetic scoring

Now that we have partioned the acoustic space, we still have to explain how a phonetic score can be associated to each inverse solution. Basically, a given acoustic vector is attached to an "ideal articulatory domain", as defined by the constraints in Tab. I, corresponding to the region of the acoustic space it belongs to. Then each inverse solution V corresponding to this 3-tuple can be given a "phonetic score" according to

the distance of the articulatory vector to the "ideal domain". A simple way to do that would be to compute the norm of the vector defined by the point and its orthogonal projection onto the domain. Actually, we compute a score relative to each type of constraint: tongue dorsum, mouth opening, lip stretching and protrusion.

The computation of the score depends on two values: the target value of the constraint considered $\theta(v,t)$, where v is the vowel and t is the type of constraint, and a margin $\sigma(v,t)$, which defines a validity interval $I(v,t) = [\theta(v,t) - \sigma(v,t); \theta(v,t) + \sigma(v,t)]$. If the value of the constraint for V is within $I(v,t)$, then it gets a perfect score (1) for that type of constraint. Otherwise, it gets a positive score less than 1 which exponentially decreases from 1 according to the distance to $I(v,t)$. The overall phonetic score is simply a linear combination of the 4 types of constraints, to get scores within the interval $[0;1]$ (1 being the best score). All constraints get equal weights, except the lip stretching which gets a null weight, because Maeda's model cannot account for lip stretching, since it was designed using X-ray images of sagittal profiles of the vocal tract.

IV. RECOVERY OF THE PLACE OF ARTICULATION OF FRENCH VOWELS

A. Choice of parameters to represent articulation of vowels

Boë et al.[3] used Maeda's articulatory model to study the place of articulation for French vowels. However, their work still did not provide a complete set of solutions, since instead of using an inversion method, they used a limited number of random articulatory configurations. Indeed, it should be noted that the 30 000 articulatory configurations used by Boë et al. correspond approximately to the choice of only 4 values for each of the articulatory parameters.

Besides the classical way of describing the vocal tract by specifying the position of articulators, it is possible to calculate representative geometrical measures. In fact,

Stevens and House (3) and Fant (4) point out that the most important characteristics of vowels from an acoustic phonetic point of view are the position of the main constriction between the tongue and the vocal tract wall and the degree of constriction (cross-sectional area) at that position.

B. Improving the acoustic faithfulness of Maeda's model

Since the objective is to get a precise evaluation of the articulatory and acoustical behavior of Maeda's model, we chose to invert vowels of the same female speaker (PB) whose X-ray images were used to build the articulatory model. Indeed, as the acoustic signal was recorded during shooting of the ten short sentences used to build the articulatory model, these data are very interesting to assess the relevance of the inversion method. However, even if the overall resynthesized formant trajectories are quite similar to those extracted from speech, we observed a non-negligible deviation between original and resynthesized formants. Actually, the geometrical precision of the model depended on two scale factors that have been set arbitrarily because the calibration of the X-ray machine was not known precisely. The adjustment of these scale factors was not possible in 1979 when the model was constructed because it would have required too much computation time. Therefore, we sampl ed ranges of reasonable values for these two factors to find better values. It turned out that the ad-hoc area increase set to 40% in the original model can be removed provided that the scaling factor is set to 196 (instead of 187). The overall frequency error thus decreases from 114Hz to 54Hz for formants F1 and F2. The remaining acoustical errors probably originate in the calculation of the area function from the mid-sagittal slice, which exploits α β coefficients proposed by Heinz and Stevens[11].

C. Experiments

Tab.II gives the five vowels inverted with their first three formant frequencies.

Formants were extracted from a spectrum computed by the "true envelope" algorithm[10] which is an iterative cepstral smoothing that takes into account only spectral peaks, i.e. mainly harmonics. All the vowels have been listened to to ensure that the vowels are perceptively correct. Formants F2 and F3 of /u/ were particularly difficult to find because the energy of this vowel is weak which means this vowel was dominated by the noise of the X-ray machine. The occurrence retained corresponds to a stronger /u/ and a slightly less intense noise.

For each vowel, the possible vocal tract shapes are recovered by applying the inversion procedure to its formants. We imposed an acoustical precision of 30 Hz to F1, 50 Hz to F2 and 75 Hz to F3 . To check the accuracy of the inversion results, we resynthesized spectra, evaluated formants and compared them against formants measured in original vowels.

In this study, we present the results according to two parameters: cross-sectional area of the main constriction (A_c, cm^2), also called degree of constriction, and the position of the main constriction in the vocal tract (X_c, cm), also called place of articulation. These parameters are obtained by retrieving the vocal tract section where the cross-sectional area is minimal. We do not consider the constriction formed at the lower part of the pharynx (close to the larynx at 2 cm from the glottis). Neither do we consider the constriction formed at the lips. As we mentioned above, the constriction considered is the lingual constriction: formed by the tongue and external vocal tract wall.

For each vowel, the results are presented in two different forms: constriction area according to the position of the main constriction, and mid-sagittal slices of characteristic vocal tract shapes recovered. The position of the main constriction varies between 0 cm (glottis) and 16 cm (lips). In order to keep constriction areas consistent with the production of vowels, we eliminated shapes which present a constriction area of less than 0.2 cm^2. We did not eliminate any other solutions from these diagrams. However, in order to save space, these diagrams are presented with the values of the phonetic

constraints (rendered by gray levels) presented above. In addition, three or six characteristic vocal tract slices are given for each of these places of articulation in order to get an idea of the vocal tract shape.

The first observation is about the number of vocal tract shapes recovered. The inversion procedure, and especially the exploration of the null space (see[17]), roughly samples the articulatory space in a uniform fashion. This means that the number of solutions is tightly connected to the extent of the articulatory region corresponding to vowels. If there were no mismatch between the analyzing model and the human vocal tract, these figures would directly represent the degree of precision required to articulate a vowel. In our case, despite the favorable situation, i.e. the analyzing model was derived from images of the speaker being inverted, and the attention we paid to the adaptation of the analyzing model, there is some model mismatch. However, figures of Tab.II clearly show that the articulation of vowels /i, y, u/ and especially /u/ requires more articulatory precision than /e/ and /a/. The small number of solutions for /u/ is probably due to t he difficulty of the model to copy the natural articulation. This problem seems more general than the model of Maeda since /u/ is always more difficult to copy than other vowels with the articulatory synthesizers available (that of Birkholz for instance[2]). Another explanation could be the construction of the articulatory model itself. Indeed, the number of solutions for /a/ is approximately half that of /e/ although one would rather expect a larger number of solutions for /a/ because of its well known articulatory variability. Therefore, it is possible that the factor analysis[14] applied to vocal tract shape contours penalizes vowels located at the extremities of the articulatory domains /a, i, u/ and to a lesser extent /y/ because they are farther from the average vocal tract shape.

Examining Figs. 4, 7, 13 and 16 we observe some key properties of the places of articulation. First, the discretization of the vocal tract, and consequently of the area function, gives rise to discrete points of articulation (which correspond to the vertical lines in the Figs. 4, 7, 13 and 16). However, despite this local spreading, places of

articulation are organized in a small number of compact regions, always less than three. In some cases, these regions merge together as the area of the constriction increases. This is particularly visible in the case of /e/.

Second, the computation of the articulation place given by the point where the area function is minimal depends on both the motionless vocal tract wall and the tongue. Some places of articulation which seems somewhat distant, especially for /a/, actually correspond to very similar vocal tract shapes. The three places of articulation of /a/ all correspond to the pharyngeal part of the vocal tract as exhibited by the mid-sagittal slices.

Third, the results are in good agreement with the data of Wood[21] for both the places of articulation and the constriction area. The constriction area of /e/ is on average greater than that of /i/ as shown by the vocal tract configurations presented by Wood. Wood's data also confirm that the place of articulation of /a/ can be spread over a large part of the pharynx.

Fourth, phonetically relevant and irrelevant vocal tract shapes share common places of articulation. This comes from the fact that the acoustical properties of vowels put very strong constraints onto the places of articulation. Consequently, irrelevant vocal tract shapes cannot be eliminated from the knowledge of their places of articulation.

Examining Figs. 5, 8, 11, 14 and 17, i.e. examples of vocal tract shapes recovered with inversion, and comparing them to the original X-ray mid-sagittal slices (Figs. 6, 9, 12, 15 and 18) obtained by Bothorel et al.[4] provides a finer analysis of the vocal tract shapes.

First, the places of articulation correspond with phonetic knowledge and the results provided by two tube vocal tract models of vowels proposed by Fant[7]. Despite this good agreement with the two tube approximation, it turns out that there exists a large articulatory variability allowed by the articulatory model, as shown by the mid-sagittal slices of Figs. 5, 8, 11, 14 and 17. Some of this variability only corresponds to realistic vocal tract shapes. For each of the vowels studied, the first mid-sagittal slices shown are

the least realistic according to the phonetic constraints presented below. One example of good and bad slices is given for each place of articulation of /a/ (i.e. roughly 3 cm, 4.7cm and 8 cm from the glottis). It has been shown that the least realistic slices correspond to extreme positions of the articulators. The upper left slice of Fig. 5, for instance, presents a wide lip opening together with a small jaw opening, and a very low position of the tongue which gives a strong constriction close to the glottis. Clearly, this vocal tract shape cannot be realized by a human speaker, or at least it is very unlikely.

In the case of /e/, the three least realistic vocal tract shapes, Fig. 8 (upper row), present a fairly high protrusion. Besides, the first two examples present a high lip opening together with the tongue in a high position, which seems difficult to realize for a human speaker.

Similarly, the worst vocal tract shapes of /i/ and /u/ both correspond to very unlikely configurations. For /y/, the worst configurations correspond to a very small protrusion and lip opening together with a low position of the apex and a compact shape of the tongue.

Two places of articulation exist for /u/. The second one (represented by the third mid-sagittal slice of Fig. 14) is located in the lower part of the pharynx. However, Fig. 19, which gives the area function, shows that the entire pharynx actually corresponds to a narrow tube. As it can be noted the number of vocal tract shapes giving this place of articulation is substantially smaller than for the first place of articulation. This means that this kind of vocal tract shape cannot be reached as easily as the first one from an articulatory point of view.

Second, vocal tract shapes recovered correspond very well with original X-ray mid-sagittal slices. This is all the more important since the acoustical simulation is unable to copy original formant data with a high precision as mentioned before. Despite this acoustical mismatch the inversion procedure is able to capture speaker specificities as shown by the inversion of the vowel /i/. As shown in Figs. 11 and 12 the second vocal

tract shape recovered represents a lip opening value substantially bigger than expected for /i/. However, it turns out that this female speaker realizes /i/ with a fairly large lip opening (as shown by the dotted contour in Fig. 12) compared to other speakers of the study of Bothorel et al. Furthermore, there is no obvious articulatory phenomenon that could explains this large lip opening. Therefore, even if the second mid-sagittal slice presents a slightly bigger value of lip opening than that observed in the X-ray contour, it is consistent with the articulation of the human speaker.

V. DERIVATION OF COMPENSATORY ARTICULATORY EFFECTS THROUGH INVERSION

The previous data show the place of articulation of five vowels for a female speaker and prove that there exist only a limited number of articulation places. These data can also be used to derive compensatory effects available to a human speaker by describing the spreading of all the vocal tract shapes giving a vowel. For that purpose, principal component analysis (PCA) was applied to all the shapes recovered for a given vowel and presenting a sufficient phonetic score. Compared to the notion of "articulatory fiber" introduced by Atal et al.[1] this approach gives a global overview of all the shapes corresponding to a vowel. The eigenvectors provided by PCA describe the spreading of vocal tract shapes, and consequently vocal tract deformations that keep formants unchanged. We tested two levels of constraint satisfaction: 0.2 and 0.9. The first score (0.2) allows very unrealistic shapes to be eliminated and the second (0.9) only allows shapes in full agreement with phonetic constraints to be kept.

It turns out that eigenvectors are quite close for both scores (although eigenvalues differ). This is probably due to the fact that good and poor vocal tract shapes according to the phonetic measure share the same places of articulation.

The results are given in Fig. 20 for the first level of satisfaction (phonetic score greater than 0.2). For clarity's sake, only the first two components are presented. Tab. III gives the amount of variance explained by the first two eigenvectors. As can

be seen in this table, a greater phonetic constraint generally increases the contribution of the first two coefficients, sometimes dramatically (for example, the first coefficient of /a/).

First, it clearly appears that the first deformations keeping formants unchanged for the vowels studied explain the spreading of places of articulation as shown in Figs. 4, 7, 13, 10 and 16.

The first deformation of /a/ corresponds to narrowing the pharynx region and thus to changing the place of articulation among one of the three constriction regions, or from one of these regions to another. It should be noted that this deformation is somewhat surprising because one would rather expect the tongue position jaw compensatory effect (lowering the jaw while moving the tongue back). Actually, this first deformation occurs in a very small acoustical region (less than 25 Hz on each formant) compared to the acoustical region allowable for /a/, and it corresponds more to a deformation that keeps the overall geometrical shape of the vocal tract unchanged rather than to a true articulatory compensatory effect.

The second deformation plays only a marginal role considering the variance it explains.

Unlike deformations of /a/ those of /e/ have got comparable weights. They correspond to a slight change of the constriction location. The first one also exhibits deformations larger at lips than at other regions of the vocal tract. Both of them involve deformations in the larynx region which compensate for deformations at lips. It should be noted that the involvement of the larynx also occurs for /i/, /y/ and /u/.

Deformation modes of /i/ show that the degree of freedom on lip aperture and protrusion is larger (related to the allowable domain for each articulatory parameter) than that of other articulators.

The first deformation of /u/ corresponds to the change of the constriction location from the palatal region to the pharynx region. The small number of inverse solutions in the pharynx region explains that this first deformation focuses more on the palato-velar

region than really between the palato-velar and pharynx regions. The second mode of deformation clearly shows a wide variability on the degree of constriction.

Finally, the two modes of deformation of /y/ show that the location of the constriction remains quasi constant; the first allowable deformation is mainly on the aperture while the second is on the volume of the front and back cavities without changing neither the overall shape nor ratio between the two cavities.

VI. CONCLUDING REMARKS

This work clearly shows the importance of phonetic constraints which enable relevant vocal tract shapes to be recovered. These constraints are all the more important since they very efficiently supplement constraints provided by the analyzing articulatory model itself.

A key point is that these constraints are derived from speaker independent knowledge. They thus capture interdependencies between articulators linked with physiological properties which can be neither investigated nor modelled directly because it would require invading medical acquisition means together with a complex muscular tridimensional modeling of the vocal tract.

It turns out that results obtained for a speaker, where X-ray images and speech signal are simultaneously available, are in very good agreement with X-ray contours even if this kind of evaluation can be carried out in very few cases, i.e. when a sufficient articulatory and acoustic data are available. However, the mismatch between the analyzing model and the vocal tract of the speaker involved in the experiment prevents a very precise acoustical inversion to be carried out. This suggests that incorporation of additional constraints, for instance upon the position of visible articulators, should be accompanied by the reduction of the acoustical precision imposed for inversion.

Even if it is not the central object of this paper, it clearly appears that acoustic-to-articulatory inversion offers a very efficient means to investigate the acoustical and articulatory properties of an articulatory model. It is especially interesting to determine

which compensatory effects are available. It should be noted that the compensatory effects studied here only correspond to a precise three-tuple of formant frequencies. This means that the compensatory effects measured in this way are under-evaluated compared to what they are when considering the entire acoustic domain allowed for a vowel. The corollary is that there is a large amount of variability in the planning of articulatory gestures by speakers.

References

[1] 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", JASA **63**, 1535–1555.

[2] Birkholz, P. and Jackèl, D. (**2003**). "A three-dimensional model of the vocal tract for speech synthesis", in *15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain*, 2597–2600.

[3] 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.

[4] Bothorel, A., Simon, P., Wioland, F., and Zerling, J.-P. (**1986**). *Cinéradiographies des voyelles et consonnes du Français* (Travaux de l'institut de Phonétique de Strasbourg).

[5] Coker, C. H. (**1976**). "A model of articulary dynamics and control", Proceedings of the IEEE **64**, 452–460.

[6] Engwall, O. (**1999**). "Modelling of the vocal tract in three dimensions", in *Proc. EUROSPEECH*, 113–116 (Budapest).

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

[8] Gabioud, B. (**1994**). "Articulatory models in speech synthesis", in *Fundamentals of Speech Synthesis and Speech Recognition*, edited by E. Keller, chapter 10 (John

Wiley & Sons, West Sussex, England).

[9] Galván-Rodrigez, A. (**1997**). *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives* (Thèse de l'Institut National Polytechnique de Grenoble).

[10] Halle, P. (**1983**). "Techniques cepstrales améliorées pour l'extraction d'enveloppe spectrale et la détection du pitch", in *Actes du séminaire "Traitement du signal de parole"*, 83–93 (Paris).

[11] Heinz, J. M. and Stevens, K. N. (**1965**). "On the relations between lateral cineradiographs, area functions and acoustic spectra of speech", in *Proceedings of the 5th International Congress on Acoustics*, A44.

[12] Maeda, S. (**1979**). "Un modèle articulatoire de la langue avec des composantes linéaires", in *Actes 10èmes Journées d'Etude sur la Parole*, 152–162 (Grenoble).

[13] Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marschal (Kluwer Academic Publishers).

[14] Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", in *Speech production and speech modelling*, edited by W. Hardcastle and A. Marchal, 131–149 (Kluwer Academic Publisher, Amsterdam).

[15] Marchal, A. (**1980**). *Les sons et la parole* (Guérin, Montréal).

[16] Mermelstein, P. (**1973**). "Articulatory model for the study of speech production", JASA **53**, 1070–1082.

[17] Ouni, S. and Laprie, Y. (**2005**). "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion", JASA **118**, 444–460.

[18] Robert, V., Wrobel-Dautcourt, B., Laprie, Y., and Bonneau, A. (**2005**). "Strategies of labial coarticulation", in *Interspeech, Lisboa*.

[19] 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**, 133–150.

[20] Sorokin, V., Leonov, A., and Trushkin, A. (**2000**). "Estimation of stability and accuracy of inverse problem solution for the vocal tract", Speech Communication **30**, 55–74.

[21] Wood, S. (**1979**). "A radiographic analysis of constriction locations for vowels", Journal of Phonetics **7**, 25–43.

Vowel	D	O	S	P
i	D6	01	S4	P1
e	D6	02	S3	P1
ɛ	D6	03	S2	P1
a	D7	04	S1	P1
y	D6	01	S1	P4
ø	D6	02	S1	P3
œ	D6	03	S1	P2
u	D8	01	S1	P4
o	D8	02	S1	P3
ɔ	D8	03	S1	P2

TABLE I. *Phonetic description of French vowels.*

Vowels	context	$F1$	$F2$	$F3$	$\Delta F1$	$\Delta F2$	$\Delta F3$	#
a	t**a**bac	749	1701	2785	19.1	25.0	24.4	103578
e	t**e**s habits	458	2341	3070	10.7	27.8	35.8	208502
i	rouss**i**es	349	2305	3345	15.8	19.4	54.1	52799
u	b**ou**gies	367	1050	2495	22.6	49.8	10.7	5147
y	d**u** gué	341	1956	2523	11.1	60.3	27.4	21748

TABLE II. *Vowel and phonetic context, first three formants (Hz), average error (Hz) and number of inverse solutions for the five French vowels of PB inverted.*

Score	> 0.2		> 0.9	
Vowel	Coef. 1	Coef. 2	Coef. 1	Coef. 2
a	0.33	0.25	0.84	0.06
e	0.30	0.23	0.42	0.30
i	0.39	0.24	0.53	0.19
u	0.67	0.17	0.59	0.25
y	0.41	0.27	0.57	0.23

TABLE III. *Variances of the first two components obtained through PCA, applied to the vowel samples with phonetic scores greater than, respectively, 0.2 and 0.9.*

List of Figures

FIG. 1 *Parameters of Maeda's articulatory model: P1 (jaw position, vertical movement) P2 (tongue dorsum position that can move roughly horizontally from the front to the back of the mouth cavity) P3 (tongue dorsum shape, i.e. rounded or unrounded) P4 (apex position ; this parameter only deforms the apex part of the tongue by moving it up or down) P5 (lip height) P6 (lip protrusion) P7 (larynx height)* 26

FIG. 2 *Voronoi diagram model.* 27

FIG. 3 *Weighted Voronoi diagram model.* 28

FIG. 4 *Vowel /a/* 29

FIG. 5 *Mid-sagittal slices of the vocal tract for /a/. For each slice the phonetic score, the maximum constriction place (Constric) w.r.t. the glottis and the area in cm^2 is given.* 30

FIG. 6 *X-ray mid-sagittal slices for /a/* 31

FIG. 7 *Vowel /e/* 32

FIG. 8 *Mid-sagittal slices of the vocal tract for /e/.* 33

FIG. 9 *X-ray mid-sagittal slices for /e/.* 34

FIG. 10 *Vowel /i/* 35

FIG. 11 *Mid-sagittal slices of the vocal tract for /i/.* 36

FIG. 12 *X-ray mid-sagittal slices for /i/.* 37

FIG. 13 *Vowel /u/* 38

FIG. 14 *Mid-sagittal slices of the vocal tract for /u/.* 39

FIG. 15 *X-ray mid-sagittal slices for /u/.* 40

FIG. 16 *Vowel /y/* 41

FIG. 17 *Mid-sagittal slices of the vocal tract for /y/. For each slice the phonetic score, the maximum constriction place (Constric) w.r.t. the glottis and the area in cm^2 is given.* 42

FIG. 18 *X-ray mid-sagittal slices for /y/* 43

FIG. 19 *Area function for /u/ with a pharyngeal constriction* 44

FIG. 20 *First two deformation modes keeping formant frequencies unchanged for /a,e,i,u,y/. The bold solid contour corresponds to the average vocal tract shape for the vowel studied. The thin and dotted contours correspond to* $\pm 3\sigma$ *(standard deviation).* 45

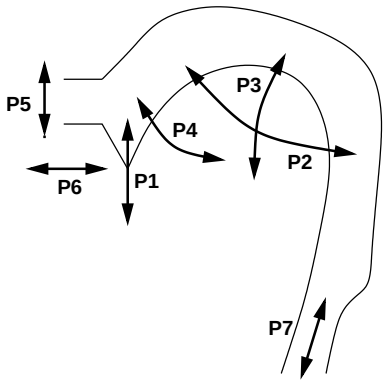


FIG. 1.

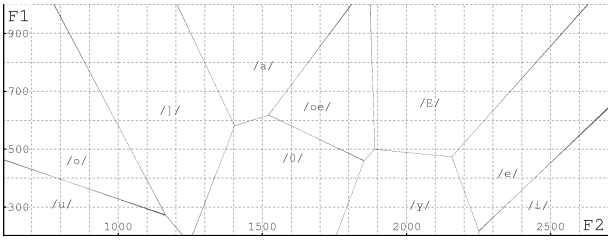


FIG. 2.

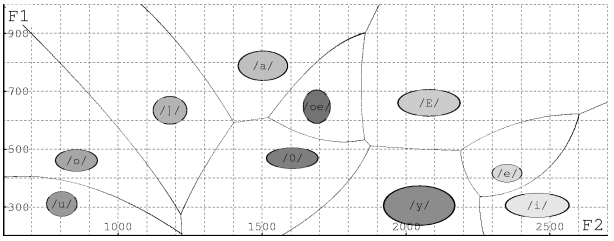


FIG. 3.

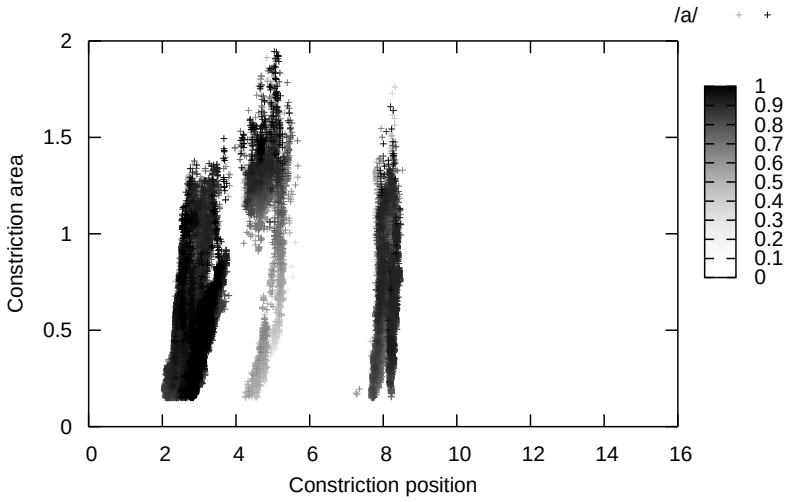

/a/
Constriction area
Constriction position
0
0.5
1
1.5
2
0
2
4
6
8
10
12
14
16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

FIG. 4.

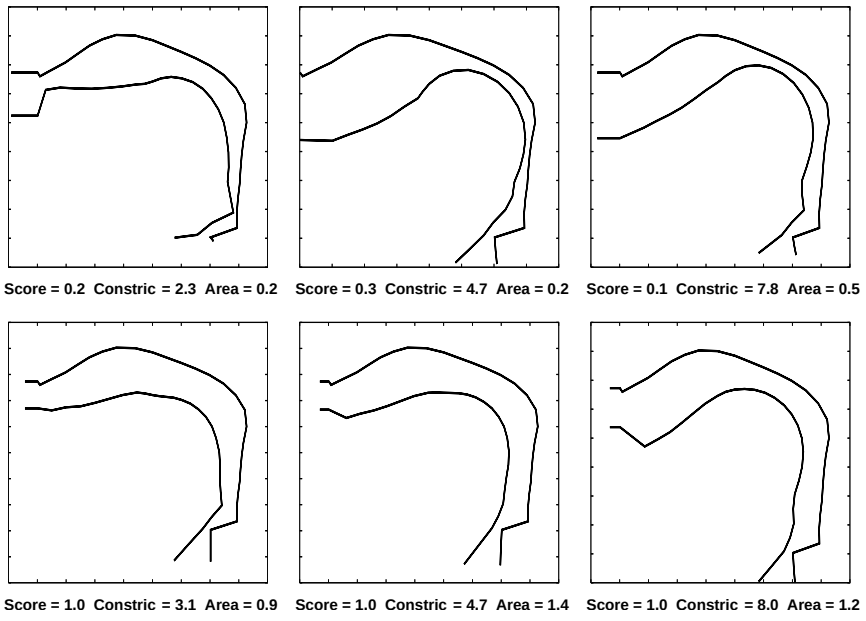

Score = 0.2 Constric = 2.3 Area = 0.2
Score = 0.3 Constric = 4.7 Area = 0.2
Score = 0.1 Constric = 7.8 Area = 0.5
Score = 1.0 Constric = 3.1 Area = 0.9
Score = 1.0 Constric = 4.7 Area = 1.4
Score = 1.0 Constric = 8.0 Area = 1.2

FIG. 5.

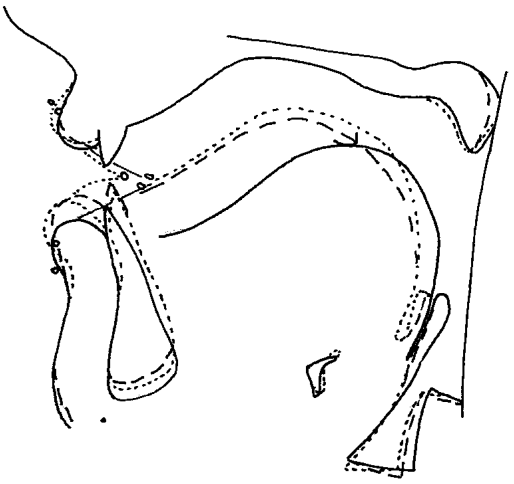

SUJET 3
Phrase 28, image 55 ——— /aba/
Phrase 01, image 73 ------ /maʃ/
Phrase 02, image 02 ..... /vwal/

FIG. 6.

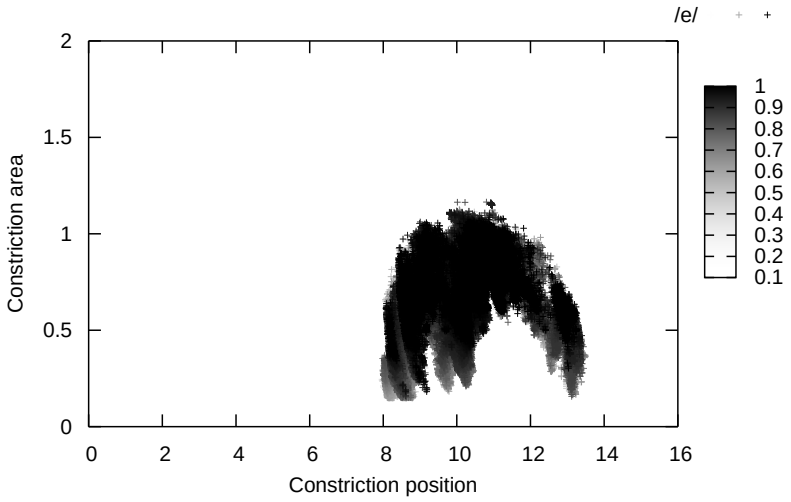

/e/
Constriction area
Constriction position
0
0.5
1
1.5
2
0
2
4
6
8
10
12
14
16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

FIG. 7.

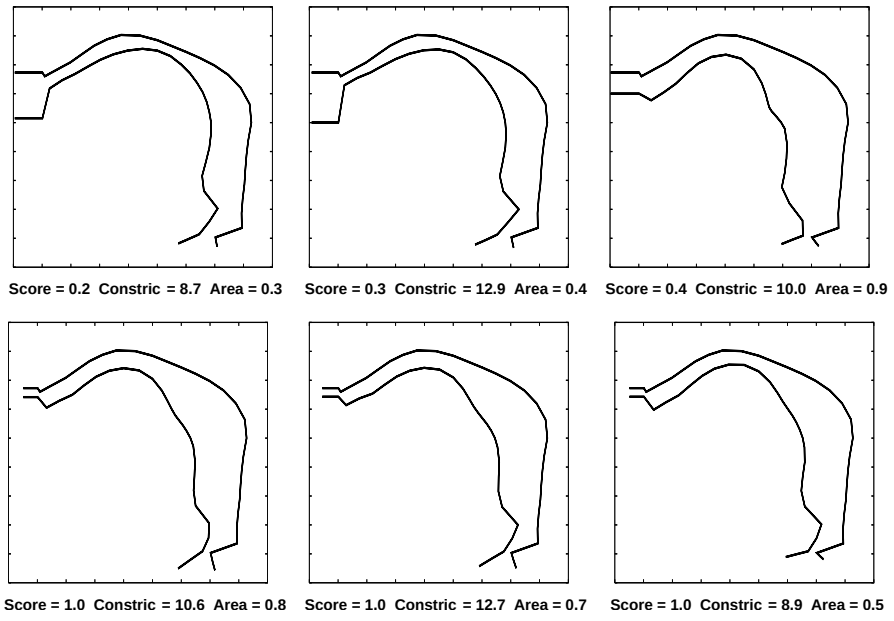

Score = 0.2 Constric = 8.7 Area = 0.3
Score = 0.3 Constric = 12.9 Area = 0.4
Score = 0.4 Constric = 10.0 Area = 0.9
Score = 1.0 Constric = 10.6 Area = 0.8
Score = 1.0 Constric = 12.7 Area = 0.7
Score = 1.0 Constric = 8.9 Area = 0.5

FIG. 8.

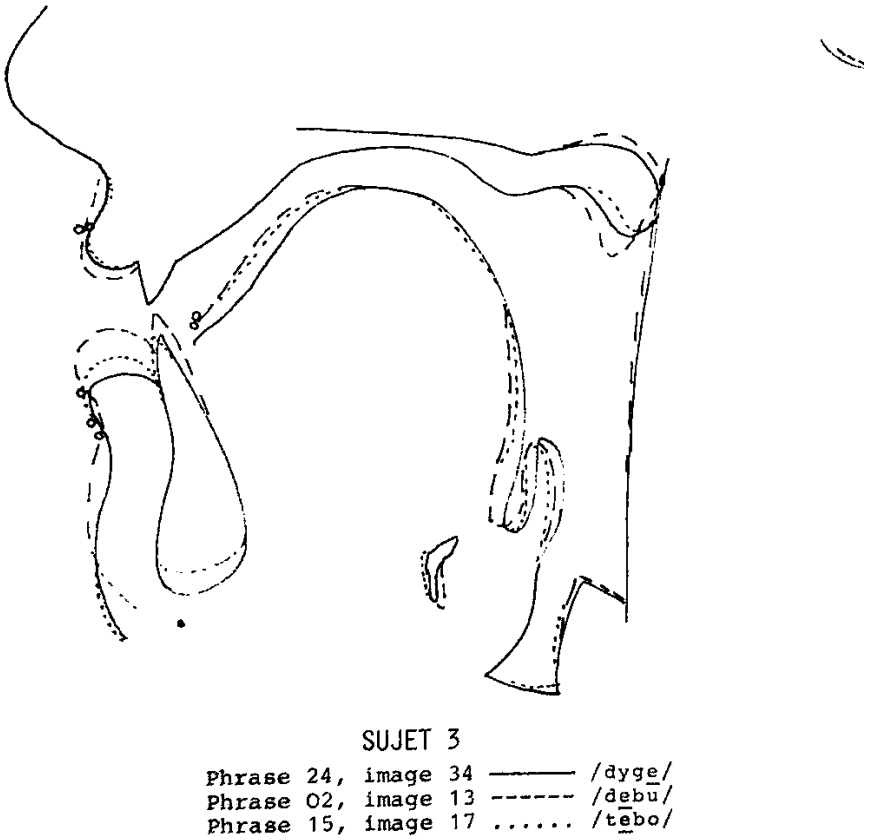

SUJET 3
Phrase 24, image 34 ——— /dyge/
Phrase 02, image 13 ------ /debu/
Phrase 15, image 17 ..... /tebo/

FIG. 9.

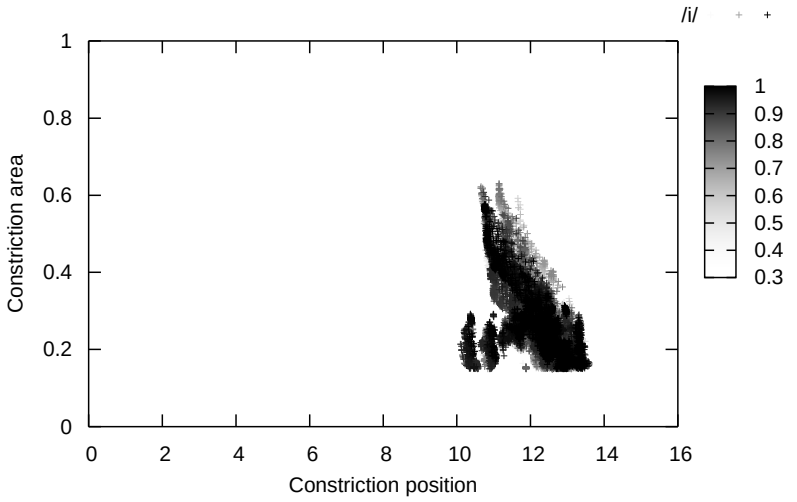

/i/
Constriction area
Constriction position
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
14
16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3

FIG. 10.

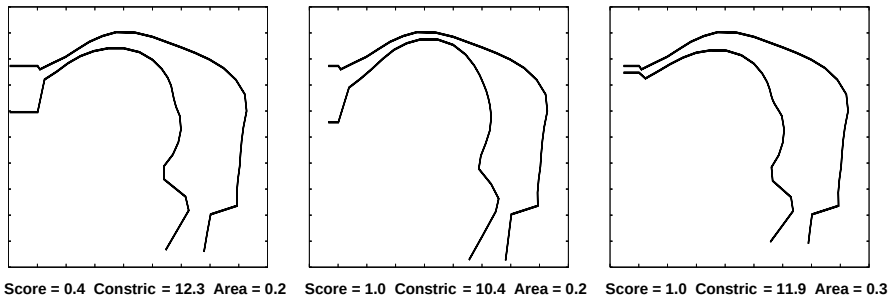

Score = 0.4 Constric = 12.3 Area = 0.2
Score = 1.0 Constric = 10.4 Area = 0.2
Score = 1.0 Constric = 11.9 Area = 0.3

FIG. 11.

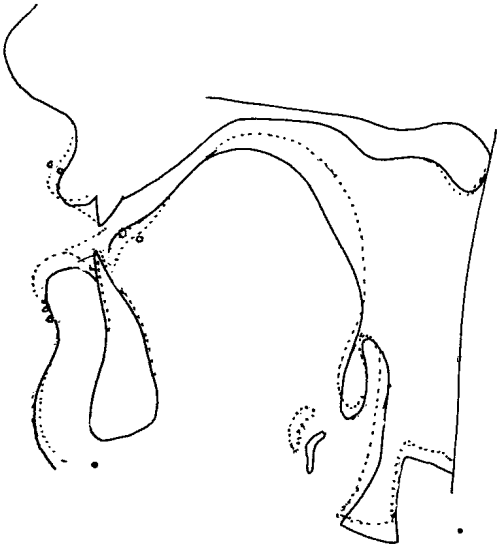


FIG. 12.

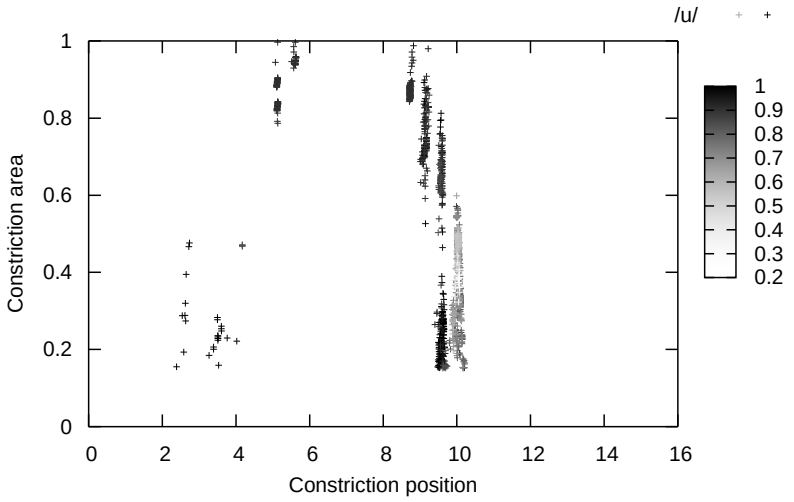

/u/
Constriction area
Constriction position
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
14
16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

FIG. 13.

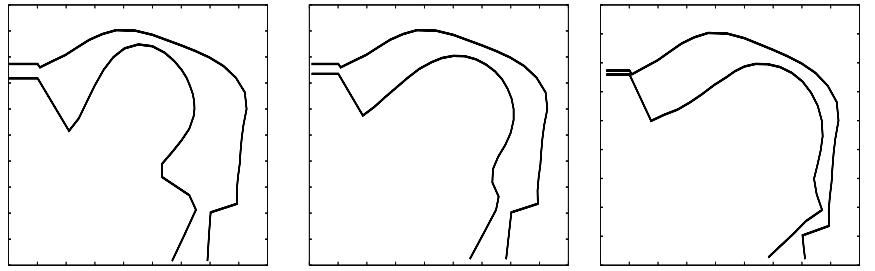

Score = 0.4 Constric = 10.0 Area = 0.5
Score = 0.9 Constric = 9.1 Area = 0.8
Score = 1.0 Constric = 2.7 Area = 0.5

FIG. 14.

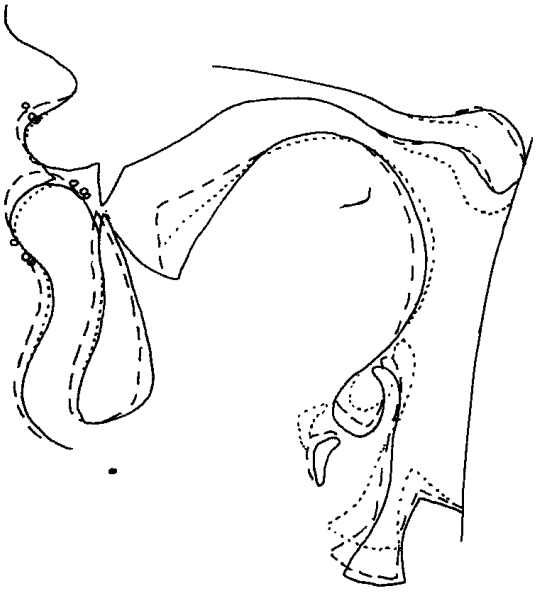

SUJET 3
Phrase 03, image 44 ——— /iku/
Phrase 17, image 56 ------ /ʃu/
Phrase 01, image 102 ..... /ʀusi/

FIG. 15.

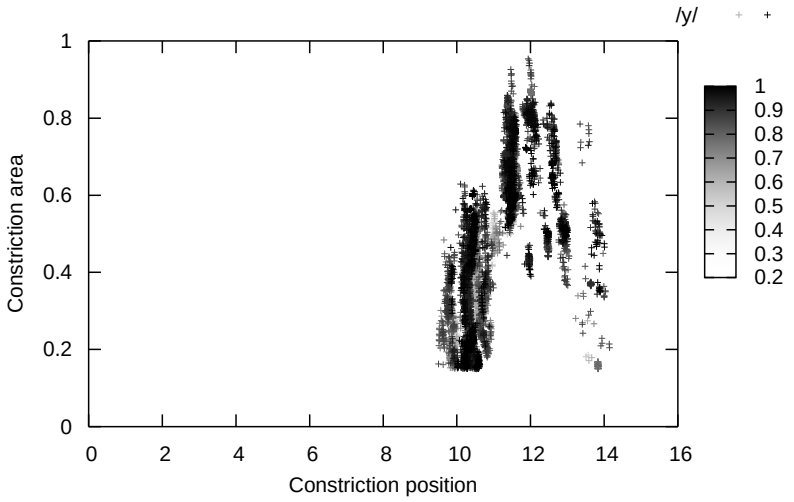

/y/
Constriction area
Constriction position
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
14
16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

FIG. 16.

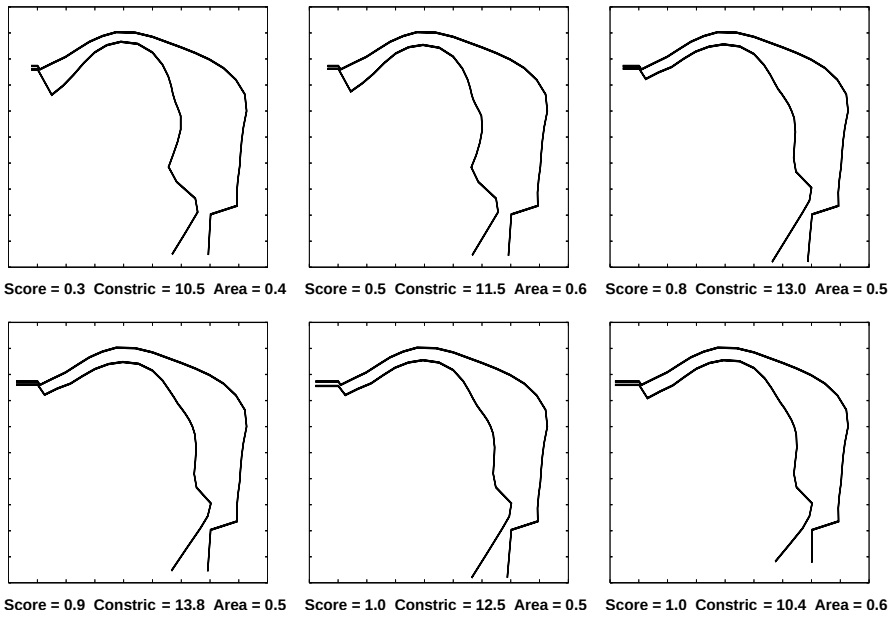

Score = 0.3 Constric = 10.5 Area = 0.4
Score = 0.5 Constric = 11.5 Area = 0.6
Score = 0.8 Constric = 13.0 Area = 0.5
Score = 0.9 Constric = 13.8 Area = 0.5
Score = 1.0 Constric = 12.5 Area = 0.5
Score = 1.0 Constric = 10.4 Area = 0.6

FIG. 17.

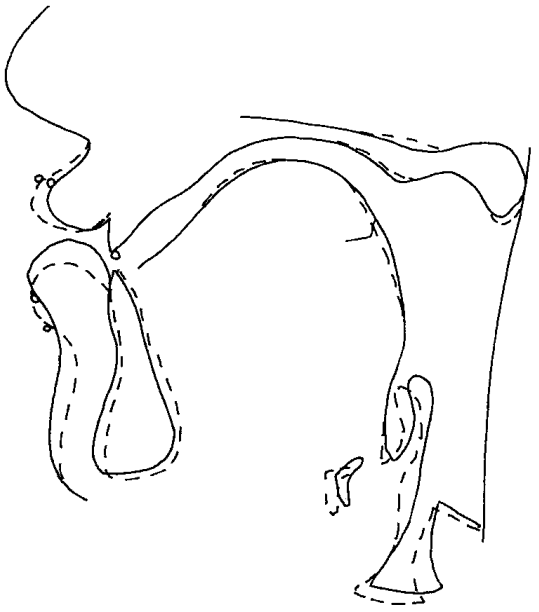


FIG. 18.

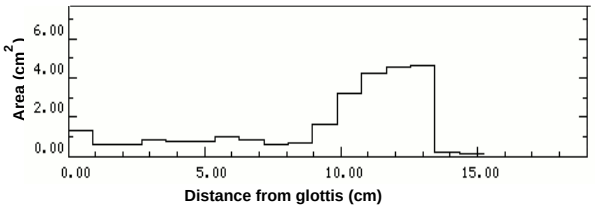

6.00
4.00
2.00
0.00
Area (cm^2)
0.00
5.00
10.00
15.00
Distance from glottis (cm)

FIG. 19.

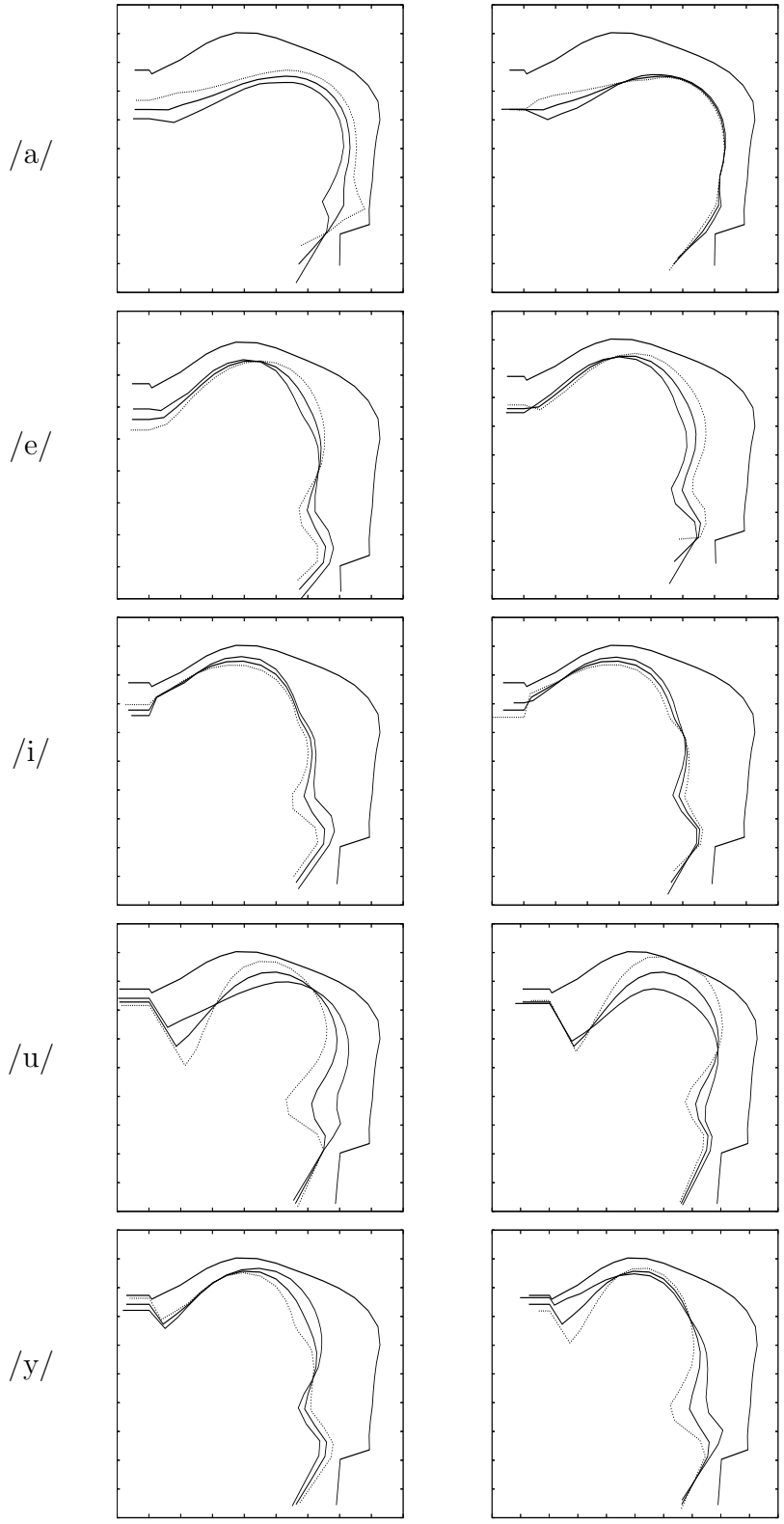


FIG. 20.