# Evaluation of phonetic constraints used in acoustic-to-articulatory inversion

**Blaise Potard**[1] **, Yves Laprie**[1] **, Anne Bonneau**[1] *

[1]LORIA-CNRS 615 rue du jardin botanique
54600 Villers-lès-Nancy FRANCE

potard,laprie,bonneau@loria.fr

***Abstract.*** *One of the main challenges in acoustic-to-articulatory inversion is the incorporation of constraints in order to reduce the under-determination of the problem. This paper is dedicated to the evaluation of phonetic constraints we proposed in a previous work. The vocal tract shapes recovered for three vowels uttered by a female speaker, the speech signal together with X-ray images are available for, were analyzed. It turns out that the phonetic constraints derived from standard phonetic knowledge are quite efficient to keep relevant vocal tract shapes. In addition, acoustic-to-articulatory inversion appears to be an efficient evaluation tool to explore the acoustical properties of an articulatory model.*

## 1. Introduction

Articulatory models and acoustical simulations of the vocal tract are now widely used to investigate speech production and to a lesser extent to deal with articulatory synthesis and acoustic-to-articulatory inversion. Articulatory models play a central role because they enable a drastic reduction of the number of parameters necessary to describe the vocal tract. The model of Mermelstein (1973), for instance, requires the knowledge of only nine parameters and that of Maeda (1979) only seven. However, this complexity reduction is not sufficient and several kinds of additional constraint have been designed. Many of them consist of imposing a principle of economy on the dynamics of articulatory parameters. This often amounts to search trajectories of articulatory parameters among slow time varying functions by minimizing, the speed, the acceleration or the jerk (third derivative) of articulatory parameters. These constraints are easily implemented by means of minimal path algorithms, like dynamic programming.

Even if other constraints have been proposed (Sorokin et al. (2000) for instance) they are often not easily usable from a practical point of view, either because they involve a number of constants to be adjusted by hand, or because they require data that do not exist or are speaker dependent. We thus proposed (Potard and Laprie, 2005) to derive constraints from standard phonetic knowledge. The main advantages are that this knowledge is speaker independent, and has been evaluated and adjusted on a large number of phonetic investigations.

*This work is part of the ASPI project funded by the IST Programme of the Commission of the European Communities as project number IST-2005-021324.

Even if these constrains seem reasonable we were interested in evaluating them by comparing results they provide with articulatory data. Unfortunately there are very few data available that associate articulatory data together with the speech signal. We used those described in the book of Bothorel et al. (1986) because the speech signal is available for one of the female speakers (PB) whose X-ray images are presented. The quality of the speech signal is not very good, due to the noise of the X-ray machine, but sufficient to evaluate formant frequencies.

The evaluation relies on the acoustic-to-articulatory inversion framework we developed in a previous work (Ouni and Laprie, 2005). It consists of recovering all the vocal tract shapes corresponding to the 3-tuple of formants measured in the vowels uttered by the subject of Bothorel. The inversion process provides all the inverse solutions that give formants very close to the original ones. The evaluation of constraints consists of examining which are the shapes they favour, and conversely those which are strongly penalized, with respect with phonetic knowledge. In addition to the evaluation of constraints this acoustic-to-articulatory inversion framework turns out to be a very efficient tool to investigate the acoustical behaviour of an articulatory model.

In the second section we will present the phonetic knowledge used and how it has been represented. In the third section the inversion procedure and the inverse solutions for three vowels will be presented. The vocal tract shapes recovered are compared to X-ray data and the results are discussed in the forth section.

## 2. Design of phonetic constraints

Phonetic constraints are derived from standard phonetic knowledge (Ladefoged, 2001; Marchal, 1980) 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 (Maeda et al., 2002). Each constraint is put 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 acoustic regions chosen are sufficiently wide to be robust to speaker variability. These constraints return numerical values, decreasing from one, when the constraint is satisfied, to zero.

Tab. 1 summarizes the phonetic description for the 10 non-nasal French vowels. $D$ stands for “tongue dorsum position”, $O$ for “mouth opening”, and $P$ for “lip protrusion”. Lip stretching is not involved because protrusion and lip stretching vary in opposite directions and thus are fairly redundant. 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 between 6 and 8 for vowels.

The implementation of these constraints is described in Potard and Laprie (2005).

| Vowel | D | O | P |
|---|---|---|---|
| i | D6 | 01 | P1 |
| e | D6 | 02 | P1 |
| ɛ | D6 | 03 | P1 |
| a | D7 | 04 | P1 |
| y | D6 | 01 | P4 |
| ø | D6 | 02 | P3 |
| œ | D6 | 03 | P2 |
| u | D8 | 01 | P4 |
| o | D8 | 02 | P3 |
| ɔ | D8 | 03 | P2 |

**Table 1.** *Phonetic description of French vowels.*

## 3. Inversion of vowels

Tab.2 gives the three vowels inverted with their first three formant frequencies.

| 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 |
| 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 |

**Table 2.** *Vowel and phonetic context, first three formants (Hz), average error (Hz) and number of inverse solutions for the three French vowels of PB (female speaker) inverted.*

Formants were extracted from a spectrum computed by the "true envelope" algorithm (Halle, 1983) which is an iterative cepstral smoothing that takes into account only spectral peaks, i.e. mainly harmonics. All the vowels have been listened to in order 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 that 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.

In this work we used the articulatory model designed by Maeda (1979) from X-ray images of vowels uttered in small sentences. 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 (the darker the gray level, the higher the satisfactory level) 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.

## 4. Discussion

Examining Figs. 1, 3 and 5 we observe some key properties of the constriction location. 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. 1, 3 and 5). However, despite this local spreading, points representing the location of the main constriction are organized in a small number of compact regions, always less than three.

Second, the results are in good agreement with the data of Wood (1979) for both the constriction locations and area. Wood's data also confirm that the constriction location of /a/ can be spread over a large part of the pharynx.

Third, phonetically relevant and irrelevant vocal tract shapes share common places of articulation. This comes from the fact that the acoustical properties of vowels, i.e. the way of organizing the two main cavities to obtain the expected resonances, 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. 2, 4 and 6, i.e. examples of vocal tract shapes recovered with inversion, and comparing them to the original X-ray mid-sagittal slices (right image of Figs. 1, 3 and 5) obtained by Bothorel et al. (1986) enables a finer analysis of the impact of phonetic constraints.

The places of articulation correspond with phonetic knowledge and the results provided by two tube vocal tract models of vowels proposed by Fant (1960). 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. 2, 4 and 6. 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 above. 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). The least realistic slices generally correspond to extreme positions of the articulators. The upper left slice of Fig. 2, 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.

Similarly, the worst vocal tract shapes of /i/ and /u/ both correspond to very unlikely configurations. For /y/ (not shown in this paper), 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 constriction locations exist for /u/. The second one (represented by the third mid-sagittal slice of Fig. 6) is located in the lower part of the pharynx, and thus not in agreement with other knowledge about the articulation of /u/. However, further examination of the area function, shows that the entire pharynx actually corresponds to a narrow tube, what is more consistent with the investigation of (Savariaux and Orliaguet, 1995).

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. 4 and 3.b the second vocal tract shape recovered presents 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. 3.b) compared to other speakers of the study of Bothorel et al. Furthermore, there is no obvious articulatory phenomenon that could explain 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.

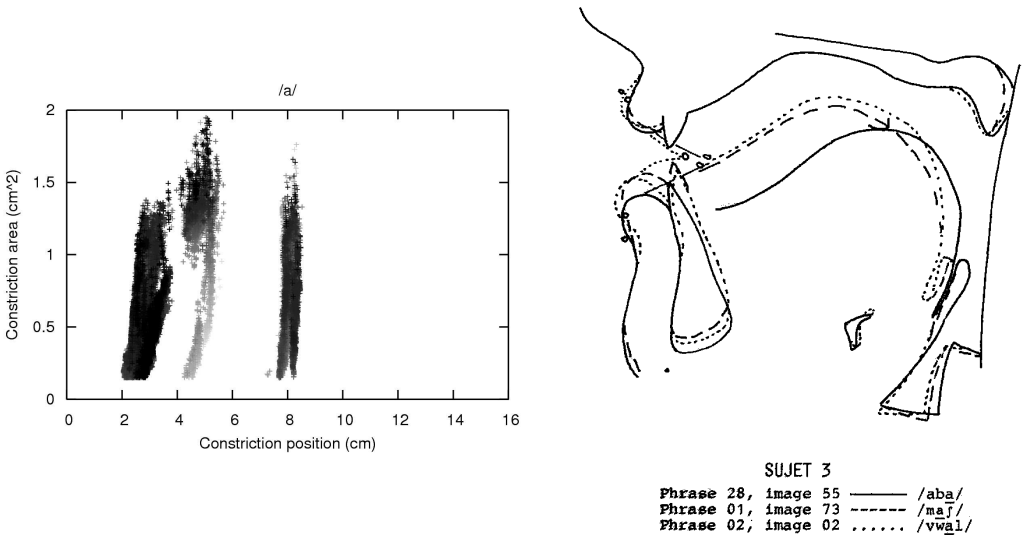


**Figure 1.** *Vowel /a/* **a:** inverse solutions represented by their constriction position (cm) and area ($cm^2$). **b:** X-ray mid-sagittal slice.

## 5. Concluding remarks

This evaluation shows that the phonetic constraints proposed are relevant and penalize not realistic vocal tract shapes. The key point is that standard phonetic knowledge enables

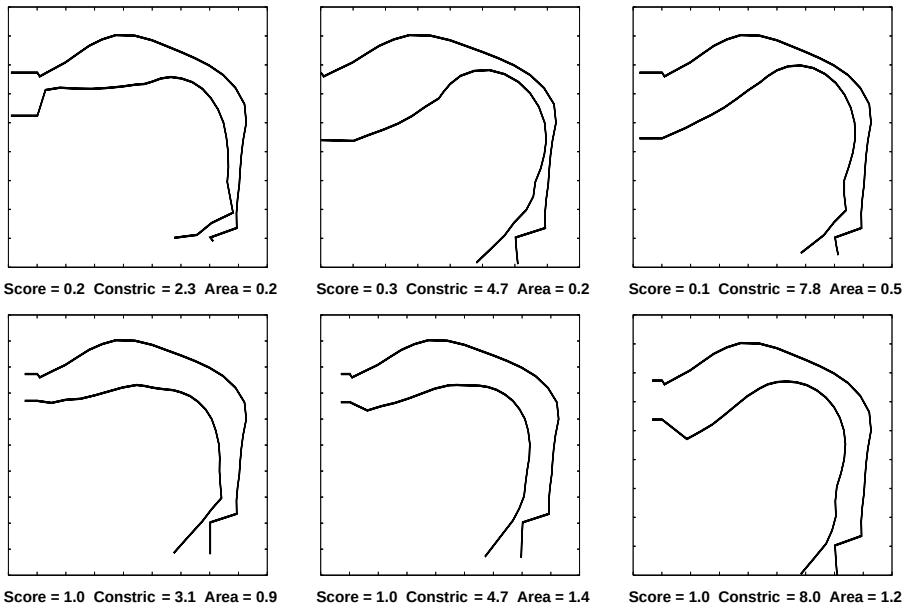


**Figure 2.** *Mid-sagittal slices of the vocal tract for /a/. For each slice the phonetic score, the maximum constriction place (Constric) w.r.t. the glottis and the area in* $cm^2$ *is given.*

interdependencies between articulators to be captured efficiently. In addition, this work enables the evaluation of the articulatory model itself. Indeed, the number of vocal tract shapes recovered strongly depends on the vowel. The inversion procedure, and especially the exploration of the null space of the articulatory to acoustic mapping (see (Ouni and Laprie, 2005)), 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. Figures of Tab.2 clearly show that the articulation of vowel /u/ requires more articulatory precision than /i/, and /i/ more than /a/ which is consistent with phonetic knowledge. In our case, despite the favourable 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 which probably exaggerates the imbalance between /a/ and /u/. The very small number of inverse solutions for /u/ compared to /a/ probably means that the articulatory model and/or some physical parameters should be slightly adjusted.

## References

Bothorel, A., Simon, P., Wioland, F., and Zerling, J.-P. *Cinéradiographies des voyelles et consonnes du Français (Cineradiographies of vowels and consonants in French).* Travaux de l'institut de Phonétique de Strasbourg, 1986.

Fant, G. *Acoustic Theory of Speech Production*. The Hague: Mouton & Co., 1960.

Halle, P. Techniques cepstrales améliorées pour l'extraction d'enveloppe spectrale et la

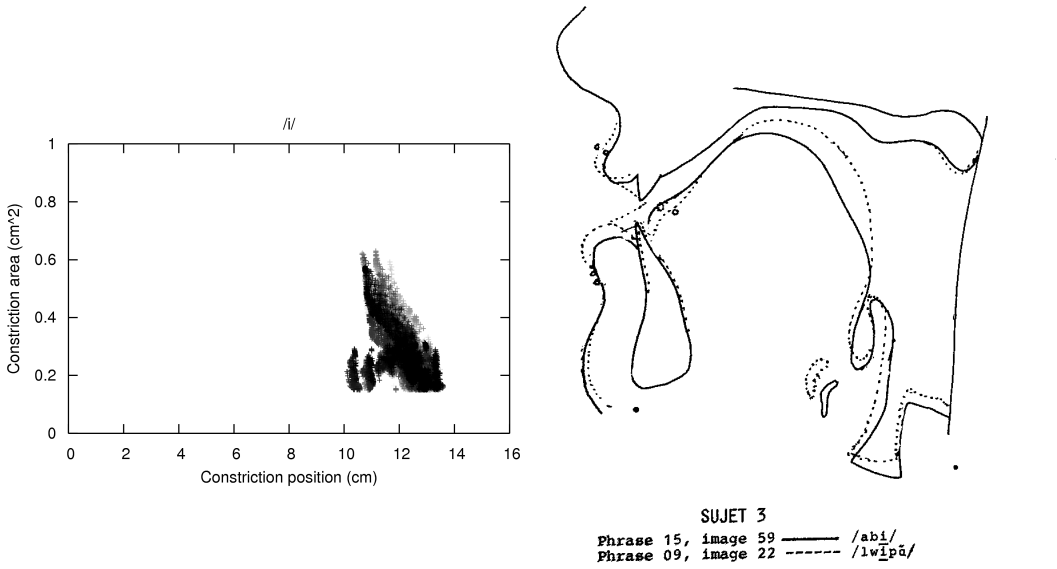


**Figure 3.** *Vowel /i/* **a:** inverse solutions represented by their constriction position (cm) and area ($cm^2$). **b:** X-ray mid-sagittal slice.

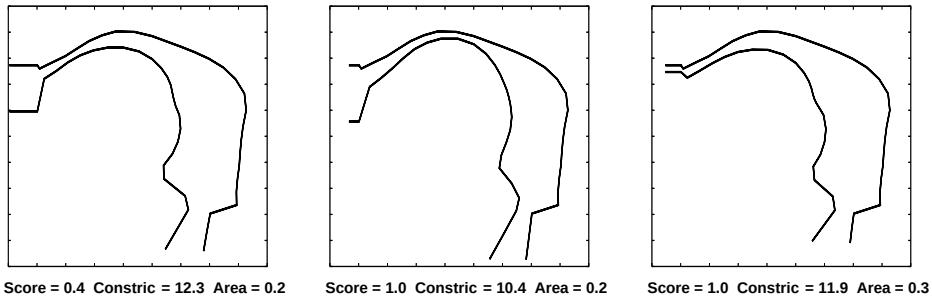


**Figure 4.** *Mid-sagittal slices of the vocal tract for /i/.*

détection du pitch (improved cepstral techniques for spectral enveloppe extraction and pitch detection). In *Actes du séminaire "Traitement du signal de parole"*, pages 83–93, Paris, 1983.

Ladefoged, P. *A Course in Phonetics, 4th edition*. Heinle, 2001.

Maeda, S. Un modèle articulatoire de la langue avec des composantes linéaires (an articulatory model of the tongue with linear components). In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

Maeda, S., Toda, M., Carlen, A. J., and Meftahi, L. Functional modeling of the face during speech production. In *Actes des Journées d'Étude sur la parole, Nancy*, pages 112–115, June 2002.

Marchal, A. *Les sons et la parole (Sounds and Speech)*. Guérin, Montréal, 1980.

Mermelstein, P. Articulatory model for the study of speech production. *JASA*, 53:1070–1082, 1973.

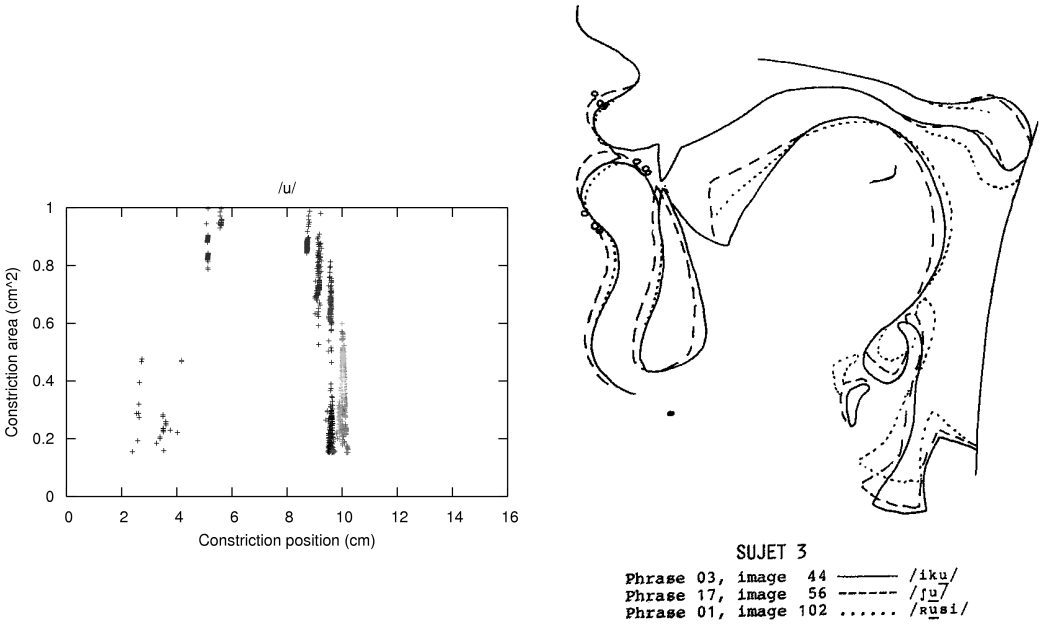


**Figure 5.** *Vowel /u/* **a:** inverse solutions represented by their contriction position (cm) and area ($cm^2$). **b:** X-ray mid-sagittal slice.

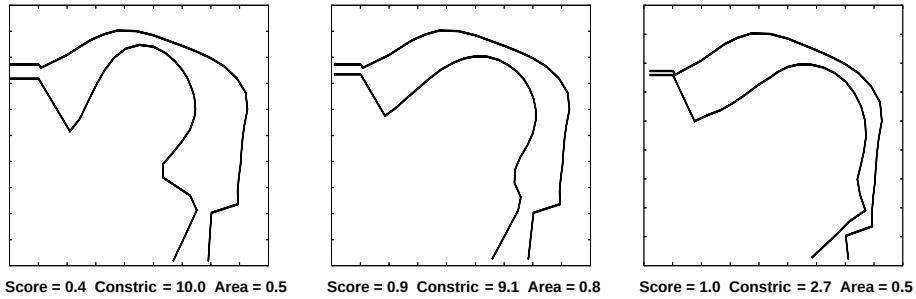


**Figure 6.** *Mid-sagittal slices of the vocal tract for /u/.*

Ouni, S. and Laprie, Y. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *JASA*, 118(1):444–460, 2005.

Potard, B. and Laprie, Y. Using phonetic constraints in acoustic-to-aticulatory inversion. In *Interspeech, Lisboa*, pages 3217–3220, September 2005.

Savariaux, C. Perrier, P. and Orliaguet, J.-P. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube : A study of the control space in speech production. *JASA*, 98:2428–2442, 1995.

Sorokin, V., Leonov, A., and Trushkin, A. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30:55–74, 2000.

Wood, S. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics*, 7:25–43, 1979.