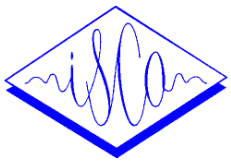

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

THE ACOUSTIC-ARTICULATORY MAPPING AND THE VARIATIONAL METHOD*

P. Jospa, and A. Soquet

Institut des Langues Vivantes et de Phonétique, Université Libre de Bruxelles,
50 av. F. D. Roosevelt, 1050 Bruxelles, Belgium.

Abstract

If work carried out in the framework of acoustic-articulatory inversion is taken into consideration, the usefulness of having a fast and accurate method to compute both the formant frequencies and the Jacobian matrix of the articulatory-acoustic transformation immediately becomes evident. For this purpose, a variational formulation of the resonance modes in the vocal tract is presented. This formulation is able to handle a labial condition that fits the opened as well as the closed tract at the lips, and at the same time gives us analytical expressions for the sensitivity functions, from which the Jacobian matrix can be deduce.

1. INTRODUCTION

A long-standing problem in speech processing is the estimation of vocal tract shape from the speech signal. Previous work on acoustic-articulatory inversion has shown that for a lossless tube with planar propagation, the acoustic measurements are not sufficient to determine a unique area function. Even in the presence of reasonable losses, Atal et al. [1] show that large modifications to the vocal tract shape can be made without changing the formant frequencies and bandwiths. The introduction of complementary constraints such as the use of an articulatory model, the specification of the lip opening or of the vocal tract length are often necessary to obtain satisfactory articulatory configurations. Several approaches to the realization of the acoustic-articulatory inversion have been reported so far. Two main approaches can be distinguished, based either on examples or on optimization.

- **Inversion based on examples**: Couples of vocal tract shapes and corresponding acoustic representations are taken as samples of the acoustic-articulatory mapping. These data can be used to build a look-up table (e.g. [1]) or to train one or several neural networks (e.g. [11], [10]). The principle of the table look-up approach is simple. First, a table is constructed with the examples. When an acoustic vector is presented, it is then matched with the entries of the table and the close matches enable a set of candidates to be found in the articulatory space. When using a neural network, the examples are used as training data so as to allow the network to approximate the non-linear relationship between the acoustic and the articulatory space. After training, when an acoustic vector is provided for the net, the output gives the corresponding articulatory configuration. Some further refinements can be added to example-based inversion in order to deal with ambiguities and insufficient acoustic matching. The ambiguities arise when several candidates have been selected. A selection criterium can be applied in order to satisfy constraints, for instance the continuity in the articulatory domain. A further optimization of the remaining candidate can then be applied in order to improve the acoustical matching (see next pragraph).
- **Inversion based on optimization**: The principle is to start from a given vocal tract shape (chosen at random or with a prior heuristic approach) and to tune the articulatory model so as to satisfy the acoustic constraints (e.g. [2], [11]). Neural controllers have also been used successfully in the framework of acoustic-articulatory inversion (e.g. [7], [9]). In this case, the optimization consists of the adaptation of the weights of the neural network. A major advantage of using neural networks lies in their learning capacity. After the net has been properly trained, it can be use to provide a first approximation of the vocal tract shape and as an optimization engine.

An important difficulty in the example-based approaches is precisely the collection of the examples. Indeed, the measurement of real vocal tract shapes is very expensive and difficult. Moreover, with current techniques, it is often not possible to measure the vocal tract shape and the corresponding speech signal simultaneously. This results in the fact that the number of measured data is not sufficient to provide an accurate description of the mapping. Therefore, people usually generate samples with a vocal tract model and base their inversion on these data, without any guarantee that they are using realistic shapes.

As stated above, the non-uniqueness problem has to be taken into account through the use of constraints on the articulatory configurations. For example-based approaches, the constraints have to be considered at a second stage by selection among the candidates. By constrast, for optimization based-approaches, these constraints can often be quite naturally introduced into the optimization procedure itself.

The main drawback of most of the optimization based approaches consists of the evaluation of the Jacobian matrix of the vocal tract model. This matrix — or an approximation of it — is used by optimization procedures based on gradient descent. Given the fact that the relations between the articulatory and acoustic space are often not monotonic over the whole working range, a constant approximation of the Jacobian matrix cannot guarantee a satisfying convergence of the optimization algorithm.

* This work was partially supported by a grant of European Community Contracts: SCIENCE PROGRAM, No. SC1*-CT92-0786, and by the ARC 93/98-168 project of the Communauté Française de Belgique.

Therefore, the Jacobian elements have to be reevaluated for each configuration of the vocal tract. This reevaluation can be obtained thanks to a perturbation approach: every control parameter is slightly modified and the corresponding variation of the formant values are used to approximate the Jacobian elements (e.g. [7]). This approach is general and simple to implement, but is time-consuming because it requires as many acoustic computations as the number of articulatory control parameters.

In this paper, we propose to use a variational method that allows us to keep the advantages of the perturbation approach without its high computational load. Indeed, the physical laws that govern the first resonance modes in the vocal tract can be given a variational formulation [5] which has important advantages compared to the classical differential formulation. It enables the use of direct methods of variational calculus such as the Rayleigh-Ritz method, (cf. [3], for example), which are efficient, time saving, and simple to implement. In this way, a direct (non-linear) relation is explicitly obtained that links the first formant frequencies to the parameters of a given area function; seen the other way round, this link expresses the acoustic constraints which the parameters of the area function model must satisfy in order to produce the desired formant frequencies [6]. Moreover, the variational method provides explicit expressions for the sensitivity functions (mode sensitivity to local variations in the area function of the tract). This property can be successfully used for the inverse acoustic-articulatory mapping of different vocal tract models with optimization based approaches [8]. Indeed the Jacobian elements of different vocal tract models can easily be deduced from the sensitivity functions. Moreover, a new, somewhat more realistic lip condition is introduced. In the framework of the variational method, this lip condition works for a tract closed as well as open at the lips.

2. THE ACOUSTIC MODEL

2.1 Sondhi's model.

We adopt the acoustic model proposed by Sondhi [12]. Formally, the equivalence between the variational formulation and the differential one (which is expressed by the continuity and movement equations), is strictly established only for acoustic models which assume that tract wall admittance is distributed in proportion to area function. This is precisely the assumption that characterises the model put forward by Sondhi. Except for the lip condition which is described bellow, the other assumptions such as a quasi-plane wave propagation, are the classical ones.

2.2 Normal mode vibrating amplitude.

For the sake of convenience, we express the sound field by its velocity potential $\phi(x,t)$ defined by[1]: $-\partial_x\phi = v(x,t)$, where x is the spatial co-ordinate measured from the glottis along the longitudinal tract axis, t the time, and v the sound field particle velocity.

For a normal resonance mode at frequency $f = \omega/2\pi$, the velocity potential assumes the following form: $\phi(x,t) = \psi(x)e^{-\sigma t}\cos(\omega t)$, where ψ is the spatially distributed mode amplitude and σ the damping induced by losses in the vocal tract. Amplitude $\psi(x)$ obeys the following *self-adjoint* equation:

$$\frac{\partial}{\partial x}\left(A\frac{\partial \psi}{\partial x}\right) + \frac{\omega^2 - \omega_p^2}{c^2} A\psi = 0 \qquad (1)$$

[1] For the derivatives, notations $\frac{\partial f}{\partial x}$ or $\partial_x f$ were adopted without distinction.

which is a Webster equation in the frequency domain. In this equation $A(x)$ represents the area function at distance x from the glottis, c is the sound velocity (c = 34000 cm/s), and ω_p a frequency correction which expresses the wall admittance effect on the resonance mode frequencies ($\omega_p \approx 175 \cdot 2\pi\ s^{-1}$). Glottis impedance is assumed to be infinite, i.e.:

$$\partial_x\psi = 0 \quad \text{at } x = 0 \quad \text{(zero particle velocity)}. \qquad (2)$$

2.3 Lip condition.

Let L denote the vocal tract length and R_0 the lip aperture radius, i.e.: $R_0 \approx \sqrt{A(L)/\pi}$. The lip radiation impedance can be approached by the impedance of a vibrating piston of radius R_0 set on a spherical baffle of a radius of approximately 9 cm. In this case, it can be shown [8] that, in a first approximation, the lip condition takes the following form:

$$A(L)\partial_x\psi(L) + q\sqrt{A(L)}\psi(L) = 0, \qquad (3)$$

where q is a function of the ratio of the lip opening radius to the average of the head radius. As an initial estimate of function q we have adopted the following expression:

$$q = a\sqrt{A(L)} + b, \qquad (4)$$

with $a = -3.5\ \text{cm}^{-1}$, and $b = 35.0$.

Unlike the usual treatment, which consists of imposing condition $\psi(L') = 0$ with $L' = L + \Delta L$ and $\Delta L \approx (3/8)\sqrt{\pi A(L)}$, condition (3) makes it possible to deal with the tract regardless of whether it is open or closed at the lips. In fact, as we shall see below, condition (3) does not contradict the condition appropriate to the tract when closed at the lips, i.e. $\partial_x\psi(L) = 0$, and makes it possible to move continuously from open vowels to bilabial stops in the framework of the variational method presented in the following section.

3. THE VARIATIONAL METHOD

3.1. Variational formulation.

Let us posit:

$$\lambda = \frac{\omega^2 - \omega_p^2}{c^2}, \qquad (5)$$

and consider the following expression:

$$I[\varphi;\omega] = \int_0^L A\left((\partial_x\varphi)^2 - \lambda\varphi^2\right)dx + q\sqrt{A(L)}\varphi^2(L) \qquad (6)$$

where $\varphi(x)$ is an arbitrary regulary function. This integral is a functional of φ. The calculus of variations shows that this functional is stationary (here, minimum) vis-à-vis the small arbitrary variations $\delta\varphi$ of φ if, and only if, φ conforms to both Webster equation (1) and conditions (2)-(3) at the tract extremities. There is no a priori need for functions φ to fulfil boundary conditions (2) and (3) at either the glottal or the labial extremities; these conditions are automatically verified as soon as φ minimises functional I defined by (6). For this reason conditions (2) and (3) are described as the "natural boundary conditions" of the variational problem: $\delta I[\varphi;\omega]/\delta\varphi = 0$. The solutions to this variational problem thus identify with the spatial amplitude distributions of the tract resonance modes, i.e. $\varphi = \psi(x)$.

For $\varphi \equiv \psi(x)$, the first term of the right-hand side of (6) represents the "integral of action" of the sound field associated with a resonance mode; the second term constitutes a load term at the labial end, introduced to satisfy condition (3) at the lips. When the lips are closed (bilabial stops), functional (6) remains valid. In fact, when $A(L) = 0$ (lips closed), functional (6) reduces to:

$$I_0[\psi;\omega]=\int_0^L A\left((\partial_x\psi)^2-\lambda\psi^2\right)dx.$$

For this functional I_0, the natural boundary conditions are: $\partial_x\psi(0)=0$ and $\partial_x\psi(L)=0$; these conditions belong to a tract closed at either end. The variational formulation, and the computation method derived from it, are consequently applicable to both bilabial stops and oral vowels.

Since amplitude $\psi(x)$ verifies (2), (3) and (4), it can easily be shown that:

$$I[\psi;\omega]=0. \qquad (7)$$

3.2. Resonance modes computation

The variational computation of the modes can be expressed in the following terms:

(i) The determination of the first N eigenfrequencies $\omega=\omega_n$ which permit the existence of the first N functions $\psi=\psi_n$, $(n=1,\ldots,N)$ that cancel the first variation of functional $I[\psi;\omega]$ defined by (6).

(ii) The identification of these eigenfunctions in a complete set of functions $\varphi(x)$ (*test functions*).

In practice, the choice of the test functions is limited to an incomplete sub-set of permissible functions so that the results obtained are generally approximations.

The problem so formulated can be solved by the Rayleigh-Ritz method [3]. The Rayleigh-Ritz method involves the choice of a set of M linearly independent functions (co-ordinate functions), i.e. $\eta_m(x)$, $(m=1,\ldots,M)$. Test functions φ are defined as linear combinations of these co-ordinate functions, i.e.:

$$\varphi(x;c_1\ldots c_M)=\sum_{m=1}^{M}c_m\eta_m(x).$$

Functional (6) is then approximated via an ordinary quadratic function in parameters c_m. The variational problem thus boils down to one of seeking the minimum of a quadratic function. As co-ordinate functions we selected the M first Chebyshev polynomials. Other well adapted co-ordinate functions are the trigonometric functions.

The Rayleigh-Ritz method establishes a direct algebraic link between the approximate values of the first tract resonance frequencies (formant frequencies) and the tract geometry, via the quantities:

$$V_{l,m}=\int_0^L A\eta_l\eta_m dx \qquad (8a)$$

and:

$$K_{l,m}=\int_0^L A(\partial_x\eta_l)(\partial_x\eta_m)dx+q\sqrt{A(L)}\eta_l(L)\eta_m(L). \qquad (8b)$$

For a given choice of co-ordinate functions these quantities depend on the area function alone. This link assumes the form of an M degree polynomial equation in λ (or in ω^2):

$$\det\left[K_{l,m}-\lambda V_{l,m}\right]=0 \qquad \text{(characteristic determinant)}, \qquad (9)$$

the first roots of which supply the approximate values of the first tract resonance frequencies.

As far as the computation of the first three modes is concerned, a satisfactory degree of accuracy is attained with a relatively small number M of co-ordinate functions, i.e. $M\approx 10$. It will also be noted that the method does not call for the computation of the transfer function. It consequently allows a significant saving in computing time.

Once the eigenvalues (the ω_n^2) have been calculated, the computation of the eigenfunctions involves no more than the resolution of a system of $M-1$ linear equations for each of the modes.

4. MODE FREQUENCIES SENSITIVITY

4.1. Sensitivity Functions

Let us denote by $[1_A]$ and $[1_x]$ the unity of area and the unity of length respectively. The mode frequency sensitivity function, i.e. $S_n(x)\equiv[1_A]^{-1}\delta f_n/f_n$, expresses the rate of frequency shift caused by a weak (first order) area function perturbation situated in x and represented by a Dirac delta distribution: $\delta A(x)=[1_A]\delta(x'-x)$. The analytical expression of these sensitivity functions is easily obtained from result (7):

$$\int_0^L A\left((\partial_x\psi_n)^2-(2\pi/c)^2\left(f_n^2-f_p^2\right)\psi_n^2\right)dx+q\sqrt{A(L)}\left(\psi_n(L)\right)^2=0, \qquad (7bis)$$

where f_n designates the nth. mode frequency and f_p the wall vibration eigenfrequency. A simple first order perturbation calculation applied to (7bis) provides the expression sought. For $x\neq L$, we have:

$$S_n(x)=\frac{\left((c/2\pi)^2\left(\partial_x\psi_n(x)\right)^2-\left(f_n^2-f_p^2\right)\psi_n^2(x)\right)[1_x]}{2f_n^2\int_0^L A\psi_n^2dx}. \qquad (10a)$$

For $x=L$, we have:

$$S_n(L)=\frac{\left((c/2\pi)^2\left(\partial_x\psi_n(L)\right)^2-\left(f_n^2-f_p^2\right)\psi_n^2(L)\right)[1_x]}{4f_n^2\int_0^L A\psi_n^2dx}+$$

$$\frac{(c/2\pi)^2\left(\partial_{A(L)}q+\left(q/2A(L)\right)\right)\sqrt{A(L)}\psi_n^2(L)}{2f_n^2\int_0^L A\psi_n^2dx}. \qquad (10b)$$

The sensitivity functions are expressed entirely in terms of unperturbed quantities f_n, $\psi_n(x)$ and $A(x)$; they are valid for the open as well as for the closed tract at the lips. The numerical computation of the sensitivity functions follows on immediately once f_n and $\psi_n(x)$ have been computed by the method summarised in the previous paragraph. Examples of sensitivity functions obtained by this method are illustrated in Figure 1.

4.2. Jacobian of the acoustico-articulatory transform

Let $A(x;\alpha_1,\ldots,\alpha_K)$ denotes the area function generated by an arbitrary articulatory (or area function) model controlled by K "articulatory" parameters $\alpha_1,\ldots,\alpha_K$. The Jacobian of the acoustico-articulatory transform is defined by the following NxK matrix J, with:

$$J_{n,k}=\frac{\partial f_n}{\partial\alpha_k} \qquad (11)$$

and N being the number of resonance frequencies (formants) considered $(N=3)$. From the definition of the sensitivity functions (see §4.1) it is easy to express the Jacobian matrix elements as:

$$J_{n,k}=f_n\int_0^L S_n(x)\frac{\partial A(x)}{\partial\alpha_k}dx, \quad n=1,..,N \text{ and } k=1,..,K. \qquad (12)$$

The Jacobian is thus easy to calculate once the sensitivity functions are obtained, and this calculation involves unperturbed acoustic quantities f_n and $\psi_n(x)$ only.

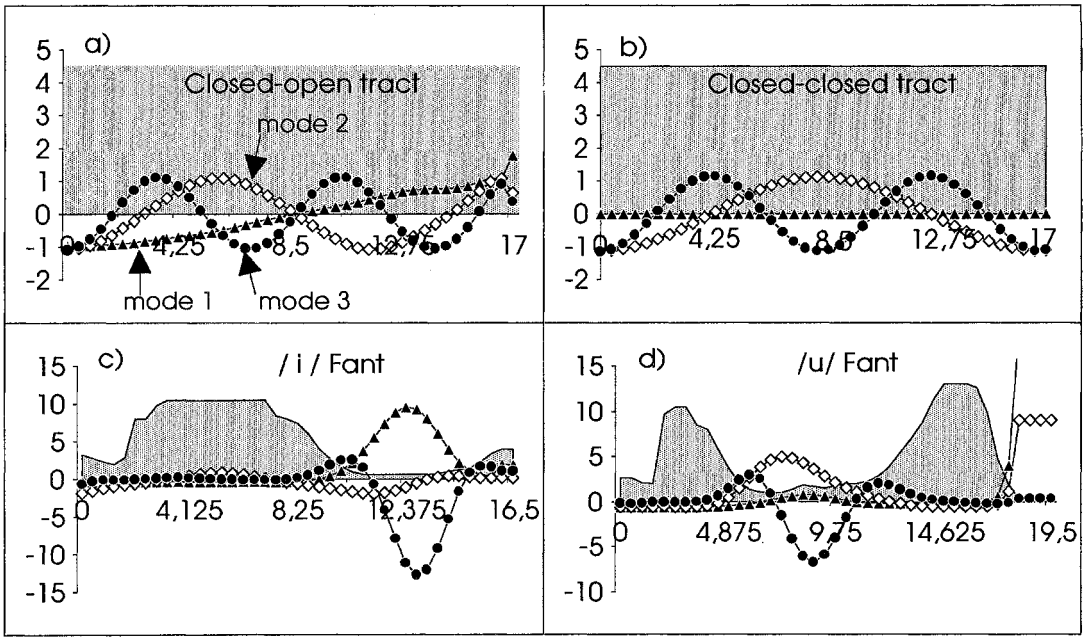

Figure 1a-d:
Sensitivity function for the first three resonance mode frequencies (formants). X axis: distance from the glottis (cm). Y axis: area function (cm2), and sensitivity function (multiplied by 2 for better visibility; in (1/cm2)(1/hz)).

5. CONCLUSIONS

A variational formulation of the resonance modes in the vocal tract is presented. This formulation establishes a direct and explicit (non linear) link between the first formant frequencies and the parameters of a given area function model. The condition adopted at the labial end fits the opened as well as the closed tract at the lips. The variational method gives us analytical expressions for the sensitivity functions, from which the Jacobian matrix is easily deduced.

REFERENCES

[1] B. S. ATAL, J. J. CHANG, M. V. MATHEWS, AND J. W. TUKEY, (1978) "Inversion of articulatory-to-acoustic transformation in the vocal-tract by a computer-sorting technique". *J. Acoust. Soc. Am.*, 63:1535-1555.

[2] J. L. FLANAGAN, K. ISHIZAKA, AND K. L. SHIPLEY, (1980) "Signal modelsfor low bit-rate coding of speech". *J. Acoust. Soc. Am.*, 68:780-791.

[3] S.H. GOULD (1966). "Variational Methods for Eigenvalue Problems", London: Oxford University Press.

[4] P. JOSPA. (1974). "On the use of critical formant frequencies in designing an articulatory model". in: Speech Communication., Ed.: G, Fant, *Proc. Speech Comm. Seminar*, Stockholm, Vol.2, pp 285-292.

[5] P. JOSPA (1982). "Theory of evolving normal modes and the vocal tract. Part I - Basic relations and adiabatic invariance". *Acustica*, 52(2):86-94.

[6] P. JOSPA (1991). "Des paramètres formantiques au profil articulatoire". In *Proc. of the XIIth International Congress of Phonetic Sciences*, pp 210-214, Aix-en-Provence.

[7] P. JOSPA, A. SOQUET, AND M. SAERENS, (1992) "Acoustical sensitivity functions and the control of the vocal tract model". Signal Processing VI: Theories and Applications, J. Vanderwalle, R. Boite, M. Moonen, A. Oosterlinck (editors), Morgan Kaufmann Publishers, pp 319-327.

[8] P. JOSPA, A. SOQUET, AND M. SAERENS, (1994) "Variational formulation of the acoustico-articulatory link and the inverse mapping by means of a neural network". to appear in *Levels in Speech Communication Relations and Interactions*. Amsterdam: Elsevier.

[9] R. LABOISSIERE, J.-L. SCHWARTZ, AND G. BAILLY, (1991) "Motor control for speech skills: a connectionist approach". In Touretzky, Elman, Sejnowski, and Hinton (Eds.), Proc. of the 1990 Connectionist Models Summer School, pp 319-327.

[10] M. G. RAHIM, C. C. GOODYEAR, W. B. KLEIJN, J. SCHROETER, AND M. M. SONDHI, (1993) "On the use of neural networks in articulatory speech synthesis". *J. Acoust. Soc. Am.*, 93(2):1109-1121.

[11] K. SHIRAI, (1993) "Estimation and generation of articulatory motion using neural networks". *Speech Comm.*, 13:45-51.

[12] M. SONDHI (1974). "Model for wave propagation in a lossy vocal tract". *J. Acoust. Soc. Am.*, 57(5):1070-1075.