THE RELATION BETWEEN VOCAL TRACT SHAPE AND FORMANT FREQUENCIES CAN BE DESCRIBED BY MEANS OF A SYSTEM OF COUPLED DIFFERENTIAL EQUATIONS

J. Schoentgen, A. Soquet, V. Lecuit, S. Ciocea

Laboratory of Experimental Phonetics, Institute of Modern Languages and Phonetics
Université Libre de Bruxelles, CP 110, 50, Av. F.-D. Roosevelt
B-1050 Brussels, jschoent@ulb.ac.be

ABSTRACT

The objective is to present coupled differential equations that relate the vocal tract shape to its eigenfrequencies. The shape of the vocal tract is described either directly by means of an area function model, or indirectly by means of an articulatory model. Some consequences of the formalism are discussed in relation to phonetic gestures or targets and the quantal principle of speech production.

1. INTRODUCTION

The objective of the presentation is to describe a mathematical framework for discussing relations between formant frequencies, area function models, articulatory models and notions such as targets, gestures, and the quantal principle of speech production. This mathematical formalism is a generalization of an earlier development by means of which we obtained an analytical solution to the formant-to-area mapping problem [1]. This mapping method has been tested on speech produced at different speaking rates by healthy and dysarthric speakers, and has been used to post-synchronize sustained speech signals and tract shapes obtained via magnetic resonance imaging [2,3,4,5]. The plausibility of the tract shapes inferred under a variety of experimental conditions argues in favor of the adequacy of the formalism proposed for the quantitative study of the relation between articulatory models, area functions and formant frequencies.

The formalism is founded on a system of coupled differential equations that relates the shape of the vocal tract and its eigenfrequencies. The tract shape is described directly via its area function, or indirectly by means of an articulatory model. The area function is the tract cross-section as a function of the distance from the glottis.

The text is organized as follows. Section 2 develops the differential link between the tract cross-sections and eigenfrequencies. Section 3 extends these relations to parametric models of the area function, articulators or articulatory postures. Section 4 focusses on the relations between acoustic and articulatory targets. Within the same framework, section 5 contrasts phonetic targets and gestures. Finally, section 6 discusses possible synergistic relations between targets or gestures on the one hand and the quantal principle of speech production on the other. Targets or gestures and the quantal principle, in fact, determine different terms in the shape/frequency relations. These terms could therefore control observed formant patterns synergistically.

2. DIFFERENTIAL FORMULATION OF THE RELATION BETWEEN THE AREA FUNCTION AND ITS EIGENFREQUENCIES

The differential formulation is derived from Webster's equation that describes the loss-free propagation of a plane acoustic wave within a conduit of arbitrary shape. Analytical solutions of the Webster equation exist for a variety of tube shapes, including the cylinder and the exponential horn [6,7]. In the case of a cylinder, the eigenmodes are described via matrix relation (1). The eigenmodes are solutions of Webster's equation and describe an acoustic field that oscillates with the same frequency and phase everywhere inside the cylinder. The knowledge of the eigenmodes is foundational since any acoustic field described via the same equation is a weighted sum of the eigenmodes. This mathematical property explains the correspondence of the tract's eigenfrequencies to observed formant frequencies.

The acoustic pressures and volume velocities at the input and output of a cylinder are noted as p_i, p_o, u_i and u_o. Parameters S_j and l_j refer to the cylinder cross-sections and lengths, i designates the square root of –1, c the speed of sound and ρ the density of air. Symbol ω is the eigenfrequency variable.

$$\begin{pmatrix} p_i \\ u_i \end{pmatrix} = \begin{pmatrix} \cos\frac{\omega l_j}{c} & \frac{i\rho c}{S_j}\sin\frac{\omega l_j}{c} \\ i\frac{S_j}{\rho c}\sin\frac{\omega l_j}{c} & \cos\frac{\omega l_j}{c} \end{pmatrix} \begin{pmatrix} p_o \\ u_o \end{pmatrix} \quad (1)$$

The shape of the vocal tract can be approximated by means of a concatenation of N cylinders. Relation (2) between acoustic pressures and volume velocities at the glottis and lips (p_{glot}, p_{lips}, u_{glot}, u_{lips}) is therefore obtained via the multiplication of N transfer matrixes (1).

$$\begin{pmatrix} p_{glot} \\ u_{glot} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} p_{lips} \\ u_{lips} \end{pmatrix} \quad (2)$$

The glottis and the lips are the boundaries of the vocal system where conditions are imposed from the outside. The simplest conditions are $u_{glot} = 0$ (infinitely rigid wall) and $p_{lips} = 0$ (infinitely pliable medium). Matrix (2) then reduces to a nonlinear algebraic equation (3) [6].

$$D(S_j, l_j, \omega) = 0 \quad (3)$$

Equation (3) implicitly defines function $\omega(S_1 \ldots S_N, l_1 \ldots l_N)$ that relates tract shape and formant frequencies. Equation (3) can therefore be used to calculate eigenfrequencies ω_k of the tract as a function of a given shape.

The form of equation (3) remains the same for other models of the tract shape or acoustic propagation [13,14].

The tract shape evolves while connected speech is being produced. Assuming that Webster's equation remains valid, it is possible to explicate the temporal evolution of the vocal tract and its eigenfrequencies by means of a temporal derivation of equation (3). The result is a system of coupled equations (4).

$$\frac{dD}{dt} = \sum_{j=1}^{N} \frac{\partial D}{\partial S_j}\frac{dS_j}{dt} + \frac{\partial D}{\partial \omega_k}\frac{d\omega_k}{dt} = 0,\ k = 1....M \qquad (4)$$

M is the number of eigenfrequencies of interest (typically 3) and N the number of concatenated cylinders. To simplify the notation, the lengths of the cylinders are assumed to be fixed. We have also found that the formant-to-area mapping based on equations (4) is numerically more stable when the cylinder lengths are kept constant. A possible explanation is that, whereas the lengths are part of the arguments of the sines and cosines in matrix (1), the cross-sections are multiplicative factors only.

In practice, the partial derivatives in equations (4) can be calculated either analytically or numerically via equation (3). The system of differential equations (4) can then be solved iteratively by conventional linear methods. The initial conditions must be chosen so as to satisfy equation (3), i.e. D=0. Normally, the evolving tract shape is known and the eigenfrequencies are calculated. In the framework of the inverse problem, i.e. formant-to-area mapping, the formant frequencies are observed and the tract shapes inferred via equations (4) combined with additional constraints since, generally speaking, the number of equations (4) is below the number of unknown cross-sections.

3. DIFFERENTIAL FORMULATION OF THE RELATION BETWEEN ARTICULATORY MODELS AND THEIR EIGENFREQUENCIES

Often, the vocal tract shape is not directly described by means of cross-sections, but via an articulatory model or via a parametric model of the area function. The purpose of these models is the control of the vocal tract shape by means of a small number of parameters that are thought to be phonetically or articulatorily relevant.

A parametric model of the area function is the stylized 3D-contour of the vocal tract described by means of a feeble number of independent variables that fix the tract length, the position of the main point of constriction, the lip tubelet length and shape, etc.

On the contrary, the vast majority of the articulatory models are geometric and bi-dimensional representations in the mid-sagittal plane of the forms and postures of the articulators. In fact, most articulatory models do not refer to the masses or elasticities of three-dimensional articulators. We will henceforth designate by θ_i the parameters of the model and by δ_i the inter-sagittal widths, which measure the distances between the upper and lower boundaries of the bi-dimensional vocal tract contour.

The link between the inter-sagittal distances and the tract cross-sections must then be formulated via heuristically-derived algebraic expression (5) [8].

$$S_j = \alpha\delta_j^{\beta} \quad (5)$$

The relation between T articulatory parameters θ, and N inter-sagittal widths δ, can be written formally as follows.

$$\frac{d\delta_j}{dt} = \sum_{i=1}^{T} \frac{\partial\delta_j}{\partial\theta_i}\frac{d\theta_i}{dt},\ j = 1...N \qquad (6)$$

Mathematically speaking, expression (6) is the result of the multiplication of matrix $\partial\delta_j/\partial\theta_i$ with a vector that contains the temporal derivatives of the articulatory parameters. Generally speaking, the matrix is not square since the number of parameters, θ, is not equal to the number of distances, δ. If the matrix were square, i. e. the numbers of cross-sections and model parameters were the same, the transform of the articulatory parameters into cross-sections would be a reversible change of coordinates, and matrix $\left(\partial\delta_j/\partial\theta_i\right)$ would be Jacobian.

When expressions (5) and (6) are inserted into coupled differential equations (4), new equations are obtained that relate articulatory parameters and eigenfrequencies

$$\sum_{j=1}^{N} \frac{\partial D}{\partial S_j}\alpha\beta\delta_j^{\beta-1}\left(\sum_{i=1}^{T}\frac{\partial\delta_j}{\partial\theta_i}\frac{d\theta_i}{dt}\right) + \frac{\partial D}{\partial\omega_k}\frac{d\omega_k}{dt} = 0,\ k = 1...M.$$

(7)

Equations (8), which refer to the sensitivities of the eigenfrequencies to small changes in the tract cross-sections, are arrived at by dividing equations (7) through derivatives $\partial D/\partial\omega_k$. A theorem in fact states that the quotients of derivatives so formed are derivatives $\partial\omega_k/\partial S_j$ [12].

$$\sum_{j=1}^{N} \frac{\partial\omega_k}{\partial S_j}\alpha\beta\delta_j^{\beta-1}\left(\sum_{i=1}^{T}\frac{\partial\delta_j}{\partial\theta_i}\frac{d\theta_i}{dt}\right) + \frac{d\omega_k}{dt} = 0,\ k = 1...M \qquad (8)$$

It is obvious that the values of temporal derivatives $d\theta_i/dt$ should be determined either directly or indirectly at the planning stage of the speech production model. The other terms in the equations are obtained either numerically or analytically by means of the articulatory model or equation D=0. The numerical approximations of the partial derivatives are arrived at by forming the ratio of small changes in the dependent and independent variables. Finally, to solve equations (8) numerically, initial conditions must be chosen that are solutions of equation D=0.

In sections 4 and 5 we discuss by means of model (8) the relations between acoustic and articulatory targets, gestures and the quantal principle of speech production.

4. ARTICULATORY VERSUS ACOUSTIC TARGETS

An articulatory target is a reference state of the vocal organs or vocal tract shape that a speaker attempts to attain while producing a phonetic segment [16]. Similarly, an acoustic target is a reference formant pattern [9]. Often, targets are thought of as quasi-stationary postures in which the articulators remain for some time once the target has been reached. This, however, leaves out targets that are dynamic.

Acoustically speaking, reaching a static target involves $d\omega_k/dt = 0,\ k = 1...M$. This means that in equations (8), the expression to the right of the plus sign vanishes. Possible solutions of the new equations are:

1) $d\theta_i/dt = 0,\ i = 1...T$. This means that an articulatory target has been reached and that the articulatory postures are momentarily frozen.

2) $\sum_{i=1}^{T} \frac{\partial \delta_j}{\partial \theta_i} \frac{d\theta_i}{dt} = 0,\ i = 1...T, j = 1...N$. This condition holds when articulators move without changing the tract shape. For instance, the lips are spreading while the lower jaw is moving upwards or the tongue is moving upwards while the jaw is moving downwards. These manoeuvres are obviously inadequate for the implementation of phonetic contrasts.

3) $\partial\omega_k / \partial S_j = 0,\ k = 1...M, j = 1...N$. This condition describes a hyposensitive link between cross-sections and frequencies. When this condition is fulfilled, small changes in the tract shape do not entail any changes in the eigenfrequencies. The theory that assigns a special phonetic status to these shapes is known as the quantal principle of speech production [15].

Solutions 1) to 3) may obviously combine so as to give rise to a desired reference formant pattern. The discussion of possible synergetic relations is postponed to the discussion section.

5. PHONETIC TARGETS VERSUS GESTURES

Solutions 1) to 3) discussed in the previous section point to shortcomings in a description of phonetic segments by means of static phonetic targets (or feature bundles). Indeed, once (context-dependent) phonetic targets have been determined at the planning stage of the production model, other co-articulatory phenomena and inter-articulatory timings that cannot be taken into account at the planning stage must be fixed in an ad hoc manner at an intermediary level, i. e an ill-defined step between the planning of phonetic targets and model (8). This step is nonetheless crucial since it assigns values to derivatives $d\theta_i/dt$ so that equations (8) can be solved.

In the following paragraph, we will therefore discuss the consequences of a few basic assumptions concerning derivatives $d\theta_i/dt$. The consequences of these assumptions attest the relevance of notions introduced elsewhere in the framework of gestural models [10]. We will so confirm, for instance, the distinctions between inter-articulatory and inter-gestural timing, and phonetic gestures and observed movement patterns.

A simple model that describes a to and fro movement between two static positions of an articulator is logistic equation (9) [11].

$$\frac{d\theta_i}{dt} = a_i(\theta_i - \theta_{i1})(\theta_i - \theta_{i2}) \qquad (9)$$

Static solutions of equation (9) are $\theta_i^* = \theta_{i1}$ or $\theta_i^* = \theta_{i2}$. The values of transition parameters a_i determine the rapidity of the evolution towards static positions θ_{i1} or θ_{i2} and the signs of the parameters the stability of these positions. Indeed, postures or motions only qualify as targets or referential states when they are stable, i.e. robust vis-à-vis small perturbations.

Inserting motion (9) into equation (8) leads to kinetic model (10).

$$\sum_{j=1}^{N} \frac{\partial \omega_k}{\partial S_j} \alpha\beta\delta_j^{\beta-1} \left(\sum_{i=1}^{T} \frac{\partial \delta_j}{\partial \theta_i} a_i(\theta_i - \theta_{i1})(\theta_i - \theta_{i2}) \right) + \frac{d\omega_k}{dt} = 0,$$

$$k = 1...M \qquad (10)$$

Model (10) suggests that when more than one articulator is involved, a distinction must be made between articulatory targets (i.e. stable positions, or motions, towards which individual articulators evolve) and gestural targets (i.e. stable shapes, or motions, towards which the vocal tract evolves as a whole). Other distinctions that arise naturally are as follows.

Firstly, active articulators do not move in unison when transition parameters a_i are chosen more or less at random, and, since the timing has been left undetermined, the stable postures that are eventually reached have much in common with the conventional phonetic targets discussed in section 4. Under these assumptions, model (10) is at best able to simulate sustained phonetic segments.

Secondly, successive articulatory targets must be activated serially when, on the contrary, the simulation of connected phonetic segments is involved. In the present model (10), transition parameters a_i must then be chosen so that different articulators move in a coordinated manner towards their assigned targets. Anatomical or acoustic reference patterns are otherwise produced either incompletely, or with a high level of variability. The need for the temporal coordination of acoustic or geometric movement patterns on the one hand and the temporal and spatial coordination of articulators on the other leads to a distinction between inter-gestural and inter-articulatory timing. Most gestural models assign the former to the planning stage and the latter to coordinative structures, the purpose of which is to tie several articulators into a whole in the pursuit of a gestural goal [17].

Thirdly, when gestures are predicted as overlapping, several of them may compete for the control of the same articulator. This means, for instance, that in the framework of model (10), derivatives $d\theta_i/dt$ are determined by more than one target. As a consequence, model (9) must be rewritten to take multiple targets into account. The static solutions of equation (11) are obtained by putting the derivative equal to zero and solving the second-order algebraic equation so obtained.

$$\frac{d\theta_i}{dt} = a_{i1}(\theta_i - \theta_{i1})(\theta_i - \theta_{i2}) + a_{i2}(\theta_i - \theta_{i1})(\theta_i - \theta_{i3}) \qquad (11)$$

A consequence is that gestures and gestural targets must be defined at a level distinct from model (10). The reason is that solutions of equations (10) that determine the observable geometric and acoustic patterns are co-governed by the solutions of equations (11) which combine the influences of multiple targets and transition parameters. In other words, the context-dependence of observable movement patterns arises via the blending of multiple overlapping gestures while the context-independent gestures are assigned to a planning stage distinct from model (10).

To sum up, the gestural model discussed here has been chosen for illustrative purposes only. The goal has not been to substitute it for other, better established, models [17]. Instead, the purpose was to show that a few basic assumptions concerning the control of a simple differential model naturally lead to categories introduced elsewhere via other models. The present discussion appears to confirm that when attempts are made to fill in the gaps left by the conventional phonetic target model, especially with respect to articulatory timing and co-articulation, a natural distinction arises between inter-articulatory coordination and the coordination of geometric or acoustic movement patterns (called gestures). Similarly, overlapping gestures are abstract patterns that are defined at a level distinct from the one at which articulatory coordination and gestural blending are implemented.

6. DISCUSSION AND CONCLUSION

Developments in sections 4 and 5 suggest that the quantal principle of speech production and models that concentrate on phonetic targets or phonetic gestures are not mutually exclusive. Indeed, the quantal principle refers to the sensitivity of formant frequencies to articulatory changes. These sensitivities appear in expressions (8) or (10), for instance, through partial derivatives that weight articulatory targets, or articulatory motions under the control of phonetic gestures. The quantal principle and gestures or targets may consequently entertain synergistic relations since articulatory zones designated by the quantal principle would favor the production of stable acoustic patterns even when targets or gestures were planned or implemented with feeble precision or doubtful stability. The mathematical framework that has been used here suggests this as a possibility only. The reality of this scheme depends, obviously, on the choice of articulatory model.

7. REFERENCES

[1] Schoentgen J., S. Ciocea (1997), « Kinematic formant-to-area mapping », Speech Communication, Vol. 21, 227-244.

[2] Ciocea S., J. Schoentgen, L. Crevier-Buchman (1997), « Analysis of dysarthric speech by means of formant-to-area mapping », 5th European Conference on Speech Communication and Technology, Rhodes, Grèce, 2547-2550.

[3] Pitermann M., S. Ciocea, J. Schoentgen (1996), « Influence de la vitesse d'élocution et de l'accent sur des cibles vocaliques estimées aux niveaux acoustique et quasi articulatoire », Comtes-rendus des XX^{èmes} Journées d'Etudes sur la Parole, Société Française d'Acoustique & European Speech Communication Association, Avignon, France, 95-98.

[4] Schoentgen J., S. Ciocea (1997), « Post-synchronization via formant-to-area mapping of synchronously recorded speech signals and area functions », 5th European Conference on Speech Communication and Technology, Rhodes, Grèce, 1799-1802.

[5] Ciocea S., J. Schoentgen (1998), " Formant-to-area mapping as a method of acoustic and articulatory data fusion in the context of automatic speaker recognition", Proceedings Workshop on Speaker Recognition and its Commercial and Forensic Applications, Avignon, France, 33-36

[6] Bonder L. (1983), « The n-tube formula and some of its consequences », Acoustica Vol. 52, 216-226.

[7] Ciocea S, J. Schoentgen, F. Bucella (1996), « A comparative study in the framework of formant-to-area mapping of continuous and discontinuous area function models of the vocal tract », ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing, Mierlo, Pays-Bas, 83-87.

[8] Heinz J., Stevens K. (1965), « On the relation between lateral cineradiographs area functions, and acoustic spectra of speech », 5th International Congress of Acoustics, Liège, Belgium, A44.

[9] Lindblom B.E. (1963), « Spectrographic study of vowel reduction », The Journal of the Acoustical Society of America, Vol. 35, 1773-1781.

[10] Browman C., L. Goldstein (1986), "Towards an articulatory phonology", Phonology Yearbook, vol. 3, 219-252

[11] Tomassone R., S. Audrain, E. Lesquoy, C. Miller (1992), « La régression - nouveaux regards sur une ancienne méthode statistique », Inra et Masson, Paris, 108-109.

[12] Margenau H., G. Murphy (1964), The Mathematics of Physics and Chemistry, D. Van Nostrand Co., Princeton, U.S.A., 6-7

[13] Ciocea S. (1997), « Détermination de la fonction d'aire du conduit vocal par inversion acoustico-géométrique semi-analytique », Thèse de doctorat, Université Libre de Bruxelles.

[14] Jospa P., Soquet A., Saerens M. (1995), « Variational formulation of the acoustico-articulatory link and the inverse mapping by means of a neural network », Levels in speech communication, Elsevier, Amsterdam, 103-113.

[15] Stevens K. (1989), « On the quantal nature of speech », Journal of Phonetics, Vol. 27, 3-45.

[16] Farnetani E. (1997), "Co-articulation and connected speech processes", The Handbook of Phonetic Sciences, Hardcastle and Laver (Eds), Blackwell, Oxford, UK, 371-404

[17] Lofqvist A. (1997), "Theories and models of speech production, The Handbook of Phonetic Sciences, Hardcastle and Laver (Eds), Blackwell, Oxford, UK, 405-426