CINEMATIC ACOUSTIC-TO-GEOMETRIC MAPPING

J. Schoentgen * *and S. Ciocea* †
Institute of Modern Languages and Phonetics , CP 110, Université Libre de Bruxelles, 50 Avenue F.-D. Roosevelt, 1050 Bruxelles, Belgium.
* *National Fund for Scientific Research, Belgium,* † *Grant, U.L.B.*

ABSTRACT

The article describes a method of cinematic acoustic-to-geometric mapping. It directly calculates the cross-section areas of a vocal tract model from observed formant frequencies. The map is cinematic since it relates the time-derivatives of area function parameters and formant frequencies.

INTRODUCTION

The acoustic-to-geometric map relates eigenfrequencies of a vocal tract model to its area function. The area function is the link between the areas of the tract cross-sections and their distances from the glottis. Generally speaking, three approaches to acoustic-to-area mapping exist: inversion by table-look-up, inversion by means of optimization and inversion by means of linear prediction coefficients.

We here propose an alternative method. It uses analytical expressions of the time derivatives of the area function parameters to compute iteratively the parameter trajectories. The input data are experimentally obtained formant frequency trajectories. The output are lengths and cross-sections of the tubelets of an n-tubelet vocal tract model. Both tubelet lengths and cross-sections can vary with time. Inversion is mathematically separate from the choice, by means of additional constraints, of a unique area function. This, together with the iterative calculation of the parameter trajectories by means of time derivatives, guaranties that the trajectories are maximally smooth and the agreement between desired and generated formant frequencies is better than $0.01Hz$.

ANALYTIC ACOUSTIC-TO-AREA MAPPING

Conventionally, when the vocal tract shape is approximated by means of a concatenation of uniform tubelets, the link between resonance frequencies ω and tubelet cross sections S_i and lengths l_i is mathematically expressed by means of an algebraic equation $F(\omega, l_i, S_i) = 0$ [1].

It is assumed that formant frequencies ω_j have been experimentally obtained at time coordinates t_k, $t_a \leq t_k \leq t_b$. The acoustic-to-area mapping problem then is to determine the evolution with time of tubelet cross-sections and lengths so that the n-tubelet area function model generates eigenfrequencies $\omega_j(t_k)$ for $t_a \leq t_k \leq t_b$.

Hereafter, the set of area function parameters $\{S_i, l_i\}$ is designated by X_i. Provided that time interval $(t_{k+1} - t_k)$ is small, the link between parameters $(X_i)_{t_{k+1}}$ and $(X_i)_{t_k}$ can be formulated by means of their Taylor expansion.

$$(X_i)_{t_{k+1}} \simeq (X_i)_{t_k} + \left(\frac{dX_i}{dt}\right)_{t_k} (t_{k+1} - t_k). \quad (1)$$

Given X_i and derivatives $\frac{dX_i}{dt}$ at time coordinate t_a, area function parameter trajectories $(X_i)_{t_k}$ can be calculated by iteratively applying relation (1). Time derivatives $\frac{dX_i}{dt}$ can be obtained by means of equation $F(X_i, \omega_j) = 0$. Indeed, the so-called chain rule establishes a link (2) between time derivatives $\frac{dX_i}{dt}$ and $\frac{d\omega_j}{dt}$.

$$\sum_{i=1}^{n} \left(\frac{\partial F}{\partial X_i}\right)\left(\frac{dX_i}{dt}\right) + \left(\frac{\partial F}{\partial \omega}\right)\left(\frac{d\omega_j}{dt}\right) = 0. \quad (2)$$

Index n is the number of area function parameters, ω is the formant frequency variable and ω_j is the frequency of the first, second, third ... formant. The values of $\frac{d\omega_j}{dt}$ are arrived at by numerically derivating observed formant frequency trajectories. Expressions $\left(\frac{\partial F}{\partial X_i}\right)$ and $\left(\frac{\partial F}{\partial \omega}\right)$ are analytically obtained by means of equation $F(X_i, \omega) = 0$. The number of expressions (2) is equal to the number, m, of observed formants. These expressions are turned into a system of $m \times n$ linear algebraic equations by inserting the values of formant frequencies ω_j and area function parameters X_i of time coordinate t_k into analytic expressions $\frac{\partial F}{\partial X_i}$ and $\frac{\partial F}{\partial \omega}$.

$$\sum_{i=1}^{n} a_{ji} x_i = y_j, \quad j = 1, m. \quad (3)$$

Here $x_i = \left(\frac{dX_i}{dt}\right)_{t_k}$, $y_j = -\left(\frac{\partial F}{\partial \omega}\right)_{t_k,\omega_j}\left(\frac{d\omega_j}{dt}\right)_{t_k}$ and $a_{ji} = \left(\frac{\partial F}{\partial X_i}\right)_{(t_k,\omega_j)}$. Singular value decomposition delivers the general solution vector $\overline{x}$, even when $n \geq m$ [2].

$$\overline{x} = \overline{d^*} + \sum_{j=1}^{n-m} \overline{d_j}\lambda_j. \quad (4)$$

Parameters λ_j may take any real value. Vector $\overline{d^*}$ is a particular solution of system (3) and vectors $\overline{d_j}$ are columns of a matrix which singular value decomposition splits off from matrix $\{a_{ji}\}$. The selection of a unique solution is carried out by means of additional constraints. A possible constraint is the requirement that the generalized potential energy of a given area function is a minimum. The generalized potential energy is the greater the farther away an area function is from the "neutral" area function. The definition is the following.

$$E_p = \frac{1}{2}\sum_{i=1}^{n} k_i (X_i - X_{i0})^2. \quad (5)$$

Coefficients k_i are pseudo spring constants which are, for sake of convenience, put equal to 1. X_{i0} are the "neutral" area function parameters.

Inserting solution (4) into generalized potential energy (5) and combining with Taylor expansion (1), the potential energy at time coordinate t_{k+1} becomes the following.

$$E_p = \frac{1}{2}\sum_{i=1}^{n}[(X_i)_{t_k} + (d_i^* + \sum_{j=1}^{n-m} d_{ij}\lambda_j)(t_{k+1} - t_k) - X_{i0}]^2. \quad (6)$$

Computing the extremum condition $\frac{\partial E_p}{\partial \lambda_j} = 0$ leads to a system of $n - m$ linear algebraic equations with $n - m$ unknowns λ_j, system which can be solved by conventional means.

Finally, once the optimal λ-values have been determined, solution (4) yields the values of $\frac{dX_i}{dt}$ at time coordinate t_k, which, inserted into Taylor expansion (1), is used to compute area function parameters X_i at time coordinate t_{k+1}. The procedure then starts all over again with the estimation of $\frac{dX_i}{dt}$ at time coordinate t_{k+1}.

ERROR CORRECTION

Iteratively applied estimation steps accumulate small errors over time. In order to keep error accumulation to a minimum, we applied several additional stratagems, namely *a*) initialization by means of reference area functions, *b*) linearization and *c*) iterative correction of parameters X_i. *a*) Initial area function parameters $(X_i)_{t_0}$ were those that generated the first observed m-tuplet of formant frequencies. Reference formant frequencies were frequencies for which the matching area functions were known. Possible reference frequency m-tuplets were the formant frequencies of the French vowels [a], [e], [i], [o], [u]. Smooth trajectories were constructed by means of interpolation between reference and initially observed formant frequencies. Thus, the inversion procedure started with known reference formant frequencies and area function parameters and iteratively calculated parameters X_i along

interpolated formant trajectories till the initial area function parameter values $(X_i)_{t_0}$ were obtained. From then on, the method followed observed formant trajectories. *b*) The purpose of linearization was to improve the quality of Taylor approximation (1). Several authors have drawn attention to the fact that the relation between the logarithm of the area function and the logarithm of the eigenfrequencies is nearly linear around the uniform area function [3]. A change of variables $X_i \rightarrow \ln X_i$ and $\omega_j \rightarrow \ln \omega_j$ was therefore performed in equation $F = 0$ and in Taylor expansion (1). *c*) The purpose of the iterative error correction presented here was to adjust area function parameters X_i so as to suppress any remaining small disagreements between observed and generated formant frequencies. When parameters X_i at time coordinate t_k were known, relation (1) gave an estimate $(X_i)^{(1)}$ of parameters X_i at time coordinate t_{k+1}. But, formant frequencies $(\omega_j)^{(1)}_{t_{k+1}}$ generated by vocal tract model $(X_i)^{(1)}_{t_{k+1}}$ were generally slightly different from observed formant frequencies $(\omega_j)_{t_{k+1}}$. Therefore, advantage was taken of the fact that quantities $\Delta(\omega_j)^{(1)}_{t_{k+1}} = (\omega_j)^{(1)}_{t_{k+1}} - (\omega_j)_{t_{k+1}}$ and $\Delta(X_i)^{(1)}_{t_{k+1}} = (X_i)^{(1)}_{t_{k+1}} - (X_i)_{t_{k+1}}$ were related by a formula similar to formula (2). As a consequence, given $\Delta\omega_j$, solving an algebraic system analogue to (2) yielded corrections $\Delta(X_i)^{(1)}_{t_{k+1}}$ which were added to the previously calculated $(X_i)^{(1)}_{t_{k+1}}$. The procedure was repeated till the difference $\Delta(\omega_j)^{(p)}_{t_{k+1}}$ was as small as desired (i.e. 0.01 Hz). Typically, p was equal to 2 or 3.

METHODS

Our cinematic acoustic-to-area mapping method was applied to a Kelly-Lochbaum model which consisted of a concatenation of 6 uniform equal-length tubelets. The section of tubelet S_1 at the glottis was fixed at 2.5 cm^2. The cross-sections of the other five tubelets were variable and total length L depended on the area of lip tubelet S_6 ($L\ (cm) = 22 - S_6\ (cm^2)$).

The method was tested on 1170 transitions in an [iVi] context of the first three formants of vowels [a] and [e]. Vowels [a] and [e] were stressed or unstressed and the French carrier sentence was produced by a male speaker at 10 speaking rates controlled by a metronome. Each combination of vowel, stress pattern and rate was repeated 30 times.

RESULTS

The objectives of the test were the following. Firstly, to check that acoustic-to-area mapping did not, in addition to noise stemming from the formant frequency measurements, introduce noise into cross-section trajectories. Secondly, to determine the kind and quantity of gross errors. A gross error occurred when measured and generated formant frequencies differed by more than 0.01 Hz after 10 error correction iterations. Thirdly, to study the dependence of area function parameters on speaking rate and accent pattern. Indeed, a qualitative agreement between dependencies obtained by inversion and reported elsewhere would lend further support to the inversion method presented here.

The results were the following. (1) Shimmer values were computed for the first three formant frequency trajectories and for the trajectories of areas S_2 and S_5. Average shimmer was not greater in the area function trajectories than in the formant trajectories. (2) The number of experimentally obtained formant frequency triplets was equal to 44 879. Gross errors occured 273, 233 and 189 times for the first, second and third formant respectively. (3) Figure 1, for example, represents the trajectories of sections S_i for stressed vowels [a] and [e] in an [iVi] context produced at the fastest (120 beats/sec) and at the slowest (48 beats/sec) rate. It is seen that for vowel [a], "fast" and "slow" trajectories differ considerably more than for vowel [e]. A possible explanation is that, since vowels [a] and [i] differ both on frontness and closeness whereas vowels [e] and [i] only differ on closeness, contrast was easier to implement for the vowel [e] in the case of fast rates. The asymmetry of the trajectory of lip-tubelet cross-section S_6 is possibly related to the fact that, in the carrier sentence, $[i]_2$ was followed by [m].

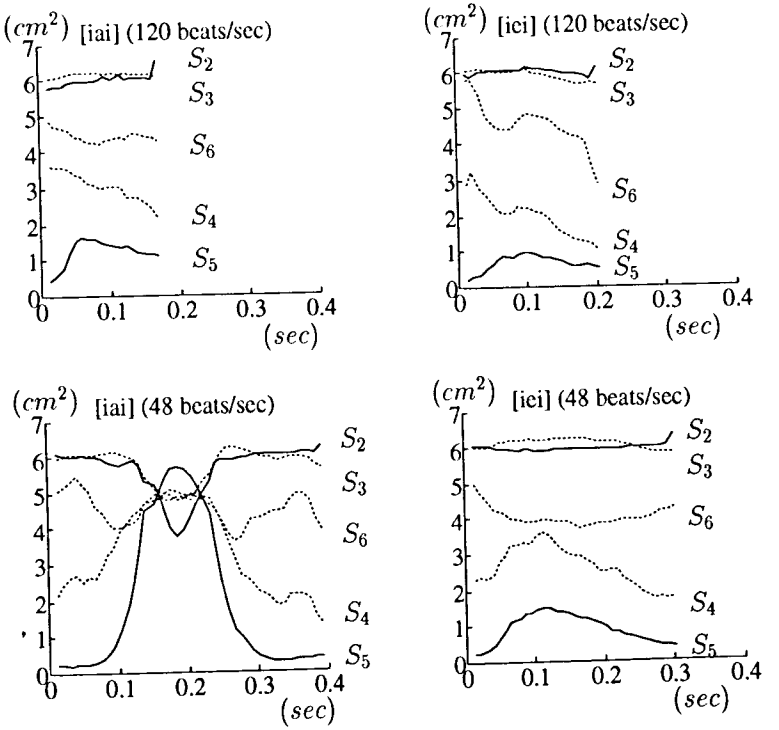

Figure 1.

References

[1] L.J. Bonder. The n-tube formula and some of its consequences. *Acustica*, 52:216–226, 1983.

[2] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes - The Art of Scientific Computing*. Cambridge University Press, New York, 1987.

[3] M.R. Schroeder. Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of Acoustic Society of America*, 41:1002–1010, 1967.