ANATOMICAL PLAUSIBILITY OF AREA FUNCTIONS INFERRED BY ANALYTIC FORMANT-TO-AREA MAPPING

A. Kacha[1], F. Grenez[1], J. Schoentgen[2,1]

[1]Department Waves and Signals, Université Libre de Bruxelles, Brussels, Belgium
[2]National Fund for Scientific Research, Belgium

{akacha, fgrenez, jschoent}@ulb.ac.be

ABSTRACT

The presentation concerns the evaluation of the anatomical plausibility of vocal tract shapes calculated by means of an analytic formant-to-area map. A constraint that requests that the jerk of the evolving model parameters is minimal has been used to select a single shape among the infinitely many area functions that are compatible with the observed formant frequencies. A similarity measure between observed and inferred cross-sections has been computed to express the plausibility of the recovered shapes quantitatively. Results show that vowel qualities involving double articulations have been the most likely to give rise to dissimilarities between acoustically inferred and directly measured vocal tract cross-sections.

Keywords: Formant-to-area inversion, area function models.

1. INTRODUCTION

The presentation concerns formant-to-area mapping, which aims at recovering the area function of the vocal tract from the formant frequencies. The area function is defined as the cross-section at a given distance from the glottis.

The method that has been evaluated in this presentation explicitly inverts the mathematical relations between parameters of the vocal tract model and its natural frequencies, so that the natural frequencies are turned into "causes" and the model parameters into "effects".

The method involves linear relations between increments of the parameters of a model and increments of the corresponding natural frequencies. This relation is pseudo-inverted and additional constraints are used to select a unique solution automatically.

The evaluation of a method of acoustic-to-articulatory inversion may involve tests of its acoustic accuracy as well as anatomical plausibility. Anatomical plausibility of acoustically inferred area functions is required when the user expects to be informed about the vocal tract shape per se.

Within the framework of the method that is presented here, acoustic accuracy is guaranteed implicitly because the convergence of the iterative inverse mapping warrants that the distance between observed and calculated formant frequencies has been kept below a threshold.

The presentation therefore concerns the evaluation of the anatomical plausibility. Experiments have been carried out by means of published acoustic and articulatory data and the anatomical accuracy has been numerically expressed by means of a measure of similarity between the calculated and observed area functions.

2. METHODS

2.1. Vocal tract model

The vocal tract has been modelled by means of a concatenation of loss-less truncated right cones of length l and radii r_1 and r_2. Matrix (1) gives the relations between the complex input and output acoustic pressures p and volume velocities v of a single conical pipe that is expanding towards the right of its virtual apex, when its axis is assumed to be horizontal [5].

$$\begin{pmatrix} p_{in} \\ v_{in} \end{pmatrix} = \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix} \begin{pmatrix} p_{out} \\ v_{out} \end{pmatrix} \tag{1}$$

$$A_i = \frac{r_2}{r_1}\cos kl - \frac{\delta}{r_1 kl}\sin kl, \quad B_i = j\frac{r_1}{r_2} z_0 \sin kl$$

$$C_i = j\frac{1}{z_0}\left[\frac{r_2}{r_1} + \left(\frac{\delta}{r_1 kl}\right)^2\right]\sin kl - \frac{1}{kl}\left(\frac{\delta}{r_1}\right)^2 \cos kl$$

$$D_i = \frac{r_1}{r_2}\left(\cos kl + \frac{\delta}{r_1}\frac{1}{kl}\sin kl\right)$$

Symbol k denotes the wave number, ρ the density of air, c the speed of sound, j the imaginary unit and s the area of the frustum; $\delta = r_2 - r_1$ and $z_0 = \rho c/s$. For a converging cone, which expands to the left, the elements of matrix (1) are re-ordered.

The global transfer matrix, which relates the acoustic pressures and volume velocities at the lips and glottis, has been obtained by multiplying transfer matrices (1) of the individual pipes. To obtain the natural frequencies, the acoustic pressure is requested to be zero at the lips, leading to $D = 0$, where D is the right lower element in the global transfer matrix. The natural frequencies are found by numerically searching for the first three frequencies for which D is zero. They correspond to the frequencies of the oscillatory modes that cause the first three formants.

The influence of the vocal tract losses due to yielding walls and acoustic radiation at the lips have been taken into account via empirical corrections of the lowest natural frequency and of the total tract length. Other types of losses have a negligible influence on the formant frequencies [3,6].

2.2. Formant-to-area inversion

Initialization

The algorithm is iterative and must be initialized. The initial tract shape has comprised a fixed cross-section of *1 cm²* adjacent to the glottis and cross-sections of *5 cm²* elsewhere. The formant trajectories between initial and observed formant frequencies have been determined by interpolation.

The following steps, from the computation of the direct map to the final iteration, have been carried out recursively in the order of presentation.

Direct map

To obtain equations that are linear and pseudo-invertible, explicit solutions of equation $D = 0$ have been approximated by their first-order Taylor expansion, which have been put in matrix form.

$$\begin{bmatrix} \frac{\partial F_1}{\partial P_1} & \frac{\partial F_1}{\partial P_2} & \cdots & \frac{\partial F_1}{\partial P_N} \\ \frac{\partial F_2}{\partial P_1} & \frac{\partial F_2}{\partial P_2} & \cdots & \frac{\partial F_2}{\partial P_N} \\ \frac{\partial F_3}{\partial P_1} & \frac{\partial F_3}{\partial P_2} & \cdots & \frac{\partial F_3}{\partial P_N} \end{bmatrix} \begin{bmatrix} \Delta P_1 \\ \Delta P_2 \\ \cdots \\ \Delta P_N \end{bmatrix} = \begin{bmatrix} \Delta F_1 \\ \Delta F_2 \\ \Delta F_3 \end{bmatrix} \quad (2)$$

Assuming that the partial derivatives exist, the matrix elements can be numerically computed via the ratio $\delta F/\delta P$ obtained by incrementing each model parameter P by a small amount δP and recording the corresponding change δF of the natural frequencies F. Map (2) is only valid for small parameter and frequency increments, for which the matrix elements can be assumed to be constant.

Inverse map

The purpose of the inverse map is to estimate model parameter increments ΔP from observed formant frequency increments ΔF. The inverse of the matrix in (2) does not exist, however, because the number of model parameters must necessarily be greater than the number of observed formants to be able to control the corresponding number of natural frequencies independently.

This suggests computing a generalized inverse via singular value decomposition, which decomposes any matrix into a product of three matrices: UWV^T. Matrices U and V are square and orthogonal, W is a diagonal matrix and V^T denotes the transpose of matrix V. Matrices U and V are invertible by conventional methods. The generalized inverse W^{-1} is obtained by zeroing the elements $1/w_{ii}$ of W^{-1} when the absolute value of w_{ii} is below a small threshold.

Pseudo-inverse solutions

A special solution ΔP_s of the inverse problem is obtained via the following expression.

$$\Delta P_S = U^T W^{-1} V \Delta F \quad (3)$$

$$\Delta P_S = \left[\Delta P_{S,1}, \Delta P_{S,2}, ..., \Delta P_{S,N}\right]^T,\ \Delta F = \left[\Delta F_1, \Delta F_2, \Delta F_3\right]^T$$

The general solution is obtained from the special one as follows. Symbol λ_j denotes real parameters.

$$\Delta P = \Delta P_s + \sum_{j=1}^{N-3} \lambda_j v_j,\ v_j = \left[v_{1,j}, v_{2,j}, \cdots, v_{N,j}\right]^T \quad (4)$$

The *N-3* column vectors v_j in (4) are the columns of matrix V that correspond to zeroed diagonal elements in matrix W.

Kinematical constraints

Because parameters λ_j may assume any real values, the number of solutions (4) is infinite. A unique solution must therefore be obtained by imposing

additional constraints, which are formulated so as to produce linear equations the solutions of which are the values of free parameters λ_j. The results that are presented here have been obtained by minimizing the jerk, which is the third-order rate of change of the model parameters. Other constraints are possible that give similar results. Generally speaking, human limb movement, including articulator motion, is subject to a constraint of minimal jerk.

Tolerance

Small errors are expected to accumulate during the recursive computation of the evolving model parameters. Differences between computed natural frequencies and observed or interpolated formant frequencies have therefore been corrected by locally reinterpreting the frequency and parameter increments in (2) as small errors and the corresponding corrections. The correction step has been repeated several times till the differences between computed and observed frequencies are below a threshold of *0.1 Hz*.

Iteration

Once the increments ΔP of the model parameters have been obtained as a function of the observed increments of the formant frequencies, the model parameters at time *n+1* are calculated from the model parameters at time *n* as follows.

$$P_{i,n+1} = P_{i,n} + \Delta P_{i,n} \qquad (5)$$

Evaluation of the anatomical plausibility of the inferred shapes

Observed and computed cross-sections have been inserted into two arrays and a numerical similarity has been calculated to express the likeness between observed and inferred area functions [2]. A similarity equal to *1* corresponds to a perfect agreement. When the numbers of computed and observed cross-sections have not been equal, the arrays have been aligned by zero-order linear interpolation.

Corpora

The corpus has comprised observed (via MR imaging) area functions, approximated by means of eight short cylinders of unequal lengths, and the first three formant frequencies corresponding to ten French vowels sustained by two male speakers (MS1 and MS2) and two female speakers (FS1 and FS2) [4].

3. RESULTS AND DISCUSSION

Fig. 1 shows from left to right, as an example, the calculated cross-sections of vowels [a], [i] and [u] for male speaker MS1. The number of cross-sections has been equal to eight. The best agreement, between computed and reference shapes, is observed for vowel [a]. The mismatch that is observed for vowel [u] is explained as follows. Vowel [u] is produced with equally narrow constrictions at the velum and lips. Because the acoustic properties of the vocal tract are mainly determined by narrow constrictions, trade-offs are possible between the constrictions at the velum and lips so that qualitatively different shapes may be obtained from a given set of formant frequencies [1].

When comparing concatenations of 8, 12 or 16 conical tubelets, the best results in terms of similarity have been obtained for 12 tubelets. The corresponding similarity matrices for two male and female speakers sustaining ten French vowels are given in Tables 1 to 4. The leftmost columns in Tables 1 to 4 denote the calculated (i.e. targeted) vowel qualities and the uppermost lines denote the observed (i.e. reference) qualities. For instance, the number *0.5* reported at the intersection of column 3 and line 2 in Table 2 is the similarity between the observed cross-sections of vowel [e] and the cross-sections that are computed on the base of the formant frequencies of vowel [a].

Inspection of Tables 1 to 4 shows that the similarities between targeted and reference cross-sections have been maximal for 29 vowels out of 40 (4 x 10). Targets [u] and [o] have been missed each thrice for [ɔ], i.e. for the corresponding open-mid vowel. The same targets have been missed once for [œ], that is, for the corresponding open-mid front vowel. Rounded target [y] has been missed twice for un-rounded [i]. Finally, [ø] has been obtained once in place of [œ].

All missed targets have been rounded and the qualities that have failed to attain their target for more than one speaker have been close or close-mid. Averaging Tables 1 – 4 and ranking according to similarities between target and reference qualities indeed confirms that targets [y] (two misses), [o] and [u] (four misses) have been the least similar to their reference qualities. A

possible explanation is given in the first paragraph of the section.

Figure 1: Computed (thin line) and measured (thick line) cross-sections corresponding, from left to right, to vowels [a], [i] and [u] for male speaker MS1. The horizontal axis is the distance from the glottis in cm; the vertical axis is the cross-sections in cm-square. The reproduced shapes are cylindrical because the Figure only displays the cross-sections that have been compared.

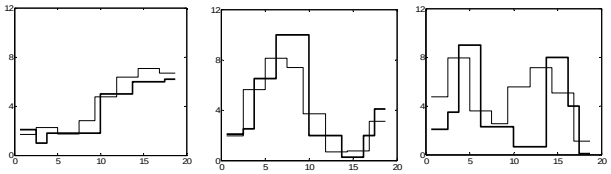

Table 1: Similarities between computed and measured vocal tract cross-sections corresponding to ten French vowels produced by male speaker MS1.

	[a]	[e]	[i]	[u]	[o]	[œ]	[ø]	[y]	[ɛ]	[ɔ]
[a]	**0.86**	0.53	0.51	0.56	0.62	0.74	0.59	0.48	0.69	0.80
[e]	0.52	**0.82**	0.75	0.52	0.56	0.61	0.79	0.69	0.69	0.56
[i]	0.48	0.77	**0.78**	0.51	0.53	0.58	0.74	0.71	0.63	0.53
[u]	0.61	0.45	0.45	0.62	0.61	0.64	0.52	0.49	0.55	**0.66**
[o]	0.68	0.50	0.49	0.58	0.68	0.75	0.59	0.52	0.60	**0.76**
[œ]	0.71	0.62	0.59	0.57	0.63	**0.80**	0.72	0.58	0.75	0.74
[ø]	0.57	0.76	0.72	0.52	0.59	0.69	**0.86**	0.70	0.77	0.61
[y]	0.45	0.69	0.68	0.50	0.53	0.57	0.70	**0.70**	0.59	0.51
[ɛ]	0.67	0.68	0.64	0.52	0.59	0.76	0.77	0.60	**0.87**	0.67
[ɔ]	0.79	0.53	0.52	0.57	0.63	0.78	0.61	0.50	0.68	**0.82**

Table 2: Similarities between computed and measured vocal tract cross-sections corresponding to ten French vowels produced by male speaker MS2.

	[a]	[e]	[i]	[u]	[o]	[œ]	[ø]	[y]	[ɛ]	[ɔ]
[a]	**0.88**	0.50	0.45	0.49	0.61	0.64	0.62	0.45	0.64	0.73
[e]	0.51	**0.86**	0.71	0.55	0.59	0.65	0.76	0.65	0.76	0.60
[i]	0.46	0.75	**0.79**	0.52	0.55	0.61	0.67	0.73	0.63	0.55
[u]	0.63	0.52	0.49	0.65	0.67	**0.73**	0.64	0.53	0.58	0.70
[o]	0.63	0.55	0.52	0.62	0.68	**0.80**	0.68	0.51	0.60	0.75
[œ]	0.62	0.67	0.62	0.64	0.62	**0.84**	0.84	0.56	0.71	0.72
[ø]	0.61	0.73	0.64	0.59	0.61	0.76	**0.88**	0.61	0.82	0.69
[y]	0.40	0.63	**0.69**	0.52	0.54	0.55	0.57	0.68	0.54	0.47
[ɛ]	0.61	0.70	0.61	0.54	0.58	0.69	0.81	0.60	**0.87**	0.65
[ɔ]	0.74	0.59	0.52	0.58	0.63	0.76	0.73	0.49	0.67	**0.79**

4. CONCLUSION

Results suggest that problematic vowel qualities have been those that involve double articulations, which cannot be resolved by acoustic data alone. The vowels that have been tested are static because anatomical data had not been recorded for connected speech. Whether co-articulation is an aid or obstacle to formant-to-area mapping therefore remains an open question.

Table 3: Similarities between computed and measured vocal tract cross-sections corresponding to ten French vowels produced by female speaker FS1.

	[a]	[e]	[i]	[u]	[o]	[œ]	[ø]	[y]	[ɛ]	[ɔ]
[a]	**0.86**	0.51	0.44	0.54	0.60	0.70	0.56	0.40	0.68	0.77
[e]	0.51	**0.78**	0.72	0.52	0.55	0.61	0.77	0.68	0.66	0.57
[i]	0.47	0.74	**0.78**	0.51	0.51	0.55	0.70	0.73	0.59	0.53
[u]	0.61	0.46	0.43	0.62	0.62	0.65	0.53	0.46	0.55	**0.67**
[o]	0.68	0.49	0.46	0.57	0.68	0.75	0.59	0.44	0.60	**0.77**
[œ]	0.71	0.59	0.54	0.56	0.62	**0.78**	0.70	0.50	0.73	0.75
[ø]	0.57	0.73	0.67	0.52	0.59	0.69	**0.86**	0.61	0.77	0.64
[y]	0.44	0.70	**0.72**	0.50	0.50	0.55	0.65	0.70	0.56	0.50
[ɛ]	0.68	0.67	0.58	0.52	0.60	0.77	0.77	0.53	**0.87**	0.69
[ɔ]	0.78	0.54	0.50	0.58	0.64	0.80	0.62	0.45	0.68	**0.82**

Table 4: Similarities between computed and measured vocal tract cross-sections corresponding to ten French vowels produced by female speaker FS2.

	[a]	[e]	[i]	[u]	[o]	[œ]	[ø]	[y]	[ɛ]	[ɔ]
[a]	**0.84**	0.50	0.43	0.53	0.60	0.59	0.55	0.44	0.65	0.74
[e]	0.49	**0.84**	0.70	0.52	0.56	0.63	0.76	0.67	0.71	0.55
[i]	0.44	0.72	**0.80**	0.53	0.51	0.59	0.68	0.69	0.61	0.51
[u]	0.60	0.46	0.43	0.62	0.62	0.63	0.56	0.50	0.57	**0.68**
[o]	0.67	0.51	0.47	0.60	0.68	0.74	0.64	0.52	0.63	**0.77**
[œ]	0.59	0.64	0.59	0.55	0.59	0.81	**0.83**	0.61	0.76	0.69
[ø]	0.57	0.74	0.65	0.54	0.59	0.76	**0.91**	0.67	0.81	0.64
[y]	0.42	0.67	0.64	0.50	0.54	0.63	0.66	**0.71**	0.58	0.49
[ɛ]	0.62	0.67	0.56	0.53	0.58	0.74	0.78	0.58	**0.89**	0.66
[ɔ]	0.77	0.53	0.48	0.58	0.63	0.68	0.64	0.49	0.69	**0.83**

5. REFERENCES

[1] Boë, L. J., Perrier, P., Bailly, G. 1992. The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *J. Phonetics*. 20, 27-38.

[2] Chandon J., Pinson S., 1981, Analyse typologique, Masson, 64-66.

[3] Ciocea, S. 1997. Semi-analytic formant-to-area mapping. *PhD thesis*, Université Libre de Bruxelles.

[4] George, M. 1997. Analyse du signal de parole par modélisation de la cinématique de la fonction d'aire du conduit vocal. *PhD thesis*, Université Libre de Bruxelles.

[5] Scavone, G. P. 1997. An acoustic analysis of singe-reed instruments with emphasis on design and performance issues and digital wave guide modeling techniques. *PhD thesis*, Stanford University, USA.

[6] Schoentgen, J., Ciocea, S. 1997. Kinematic formant-to-area mapping. *Speech Com*. 21, 227-244.