MODEL-BASED FORMANT-TO-AREA MAPPING WITH LABIAL CROSS-SECTION CONSTRAINTS

J. Schoentgen(*), A. Kacha, and F. Grenez

L.I.S.T, CP 165/51, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium
(*) Fund for Scientific Research - FNRS, Belgium
email: jschoent@ulb.ac.be

ABSTRACT

The object of the presentation is model-based formant-to-area mapping. The objective is the examination of the morphological similarity between observed and mapped area functions that may involve constraints on the labial cross-section. The vocal tract model is a concatenation of truncated cones the cross-sections of which are inferred by means of an inverse mapper that is based on a locally linear relation between observed formant frequency increments and calculated cross-section increments. The corpus comprises eight speakers who have sustained 77 vowels. Results show that constraining the labial cross-section improves the similarity between recorded and calculated area functions. The increase is less than 10 %, however. Also, fixing the labial cross-sections for French rounded back vowels, the area functions of which are difficult to infer from acoustic data, does not qualitatively improve the similarity between observed and inferred area functions.

1. INTRODUCTION

The presentation concerns formant-to-area mapping, which aims at recovering the area function of the vocal tract from the first few formant frequencies. The area function is defined as the cross-section of the vocal tract at a given distance from the glottis. The state of the art in acoustic-to-articulatory inversion is one of research and development. That is, only a small minority of published articles have reported on acoustic-to-tract mapping as an instrument the performance of which is taken for granted. The motivation today as well as in the past is the non-invasive acquisition of articulatory or morphological features. Possible applications, discussed in the literature, are aids to the handicapped and speech therapy, second-language learning, as well as automatic speech recognition, compression, analysis or synthesis based on pseudo-articulatory features.

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. Here, acoustic accuracy is guaranteed implicitly because the convergence of the iterative inverse mapping warrants that the difference 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. The anatomical accuracy has been numerically expressed by means of a measure of similarity between the calculated and observed area functions. The results focus on the ability of labial constraints to improve the similarity between mapped and observed area functions.

Studies that have evaluated the anatomical plausibility of the recovered tract shapes are scarce. Indeed, statistical maps that relate observed spectral cues to a small number of articulatory positions report on the oral cavity only and their acoustic accuracy is unknown.

2. MODEL

2.1 Area function and tract transfer function models

For a truncated cone, the relation between complex input and output acoustic pressure p and volume velocity v is given by transfer matrix (1) [1]. It describes spectrally the wave propagation in truncated cones that diverge to the *right*, when the cone axis is horizontal. When the truncated cone diverges to the left, elements a, b, c and d are positioned differently in the transfer matrix. The wave propagation is assumed to be planar.

$$\begin{pmatrix} p_{in} \\ v_{in} \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} p_{out} \\ v_{out} \end{pmatrix} \tag{1}$$

The transfer matrix elements involve truncated cone height l and base and frustum radii r_2 and r_1. Symbol ω is the angular frequency variable and ρ and c the density of air and speed of sound at body temperature. Symbol j is the imaginary unit and symbol k the wavenumber $\frac{\omega}{c}$. Acoustic impedance z_0 is equal to $\frac{\rho c}{\pi r_1^2}$.

$$a = \frac{r_2}{r_1}\cos(kl) - \frac{1}{kl}\frac{r_2 - r_1}{r_1}\sin(kl)$$
$$b = j\frac{r_1}{r_2}z_0\sin(kl)$$
$$c = j\frac{1}{z_0}[\frac{r_2}{r_1} + (\frac{1}{kl})^2(\frac{r_2 - r_1}{r_1})^2]\sin(kl) - (\frac{r_2 - r_1}{r_1})^2\frac{1}{kl}\cos(kl)$$
$$d = \frac{r_1}{r_2}\cos(kl) + \frac{1}{kl}\frac{r_2 - r_1}{r_1}\sin(kl)$$

Several truncated cones may be concatenated to simulate a vocal tract area function. The total transfer matrix (2) of a concatenation of elementary tubelets is obtained by multiplying individual transfer matrices (1). The result is a global 2×2 matrix that relates the acoustic pressures and volume velocities at the lips and glottis.

$$\begin{pmatrix} p_{glot} \\ v_{glot} \end{pmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{pmatrix} p_{lips} \\ v_{lips} \end{pmatrix} \tag{2}$$

The eigenmode conditions are that the volume velocity is zero at the glottis and the acoustic pressure zero at the lips. Inserting these conditions into matrix (2) leads to matrix element $D = 0$. Formally, equation $D = 0$ implicitly maps the vocal tract shape onto the eigenmode frequencies. In practice, these are found by numerically searching for all frequency values that zero element D.

2.2 Direct locally linear map

To ease the discovery of an inverse map, implicit direct map $D = 0$ that is nonlinear is replaced by an explicit direct map that is linear. A formulation of the direct local link (3) between model parameter and eigenfrequency increments ΔP_i and ΔF_i is possible in terms of Jacobian matrix J, which involves first-order partial derivatives of eigenfrequencies F_i with respect to model parameters P_i.

$$\begin{bmatrix} \frac{\partial F_1}{\partial P_1} & \frac{\partial F_1}{\partial P_2} & \cdots & \frac{\partial F_1}{\partial P_N} \\ \frac{\partial F_2}{\partial P_1} & \frac{\partial F_2}{\partial P_2} & \cdots & \frac{\partial F_2}{\partial P_N} \\ \frac{\partial F_3}{\partial P_1} & \frac{\partial F_3}{\partial P_2} & \cdots & \frac{\partial F_3}{\partial P_N} \end{bmatrix} \begin{pmatrix} \Delta P_1 \\ \Delta P_2 \\ \cdots \\ \Delta P_N \end{pmatrix} = \begin{pmatrix} \Delta F_1 \\ \Delta F_2 \\ \Delta F_3 \end{pmatrix} \tag{3}$$

Assuming that they exist, the partial derivatives can be estimated numerically by increasing each model parameter by a small amount with reference to the present shape and recording the corresponding change in eigenmode frequencies.

2.3 Inverse locally linear map

In the context of inverse mapping, the formant frequency increments ΔF_i are observed and the model parameter increments ΔP_i calculated. In practice, Jacobian matrix J in (3) is expected to be non-square, because the number of model parameters must be at least twice the number of observed formants in vocal tract models that enable independently controlling as many eigenmode frequencies as there are observed formants.

This suggests computing the generalized inverse of J, which is obtained via singular value decomposition. It consists in a break-up of matrix (3) into a product of three matrices $J = UWV^T$. Matrices U and V are square and orthogonal, and matrix W is diagonal. Matrix V^T is the transpose of V.

Inverse (4) of direct relation (3) involves matrices U and V, as well as the generalized inverse W^{-1}, which is obtained by zeroing diagonal elements $\frac{1}{w_{ii}}$ if the absolute values of elements w_{ii} are smaller than a threshold.

Among the infinitely many possible ones, solution (4) involves the smallest increments $\Delta P_{s,i}$ for a triplet of observed increments ΔF_i. That is, solution (4) minimizes the speed of deformation of the evolving vocal tract, which is a popular constraint in the framework of acoustic-to-articulatory mapping. An alternative constraint involves minimizing the deformation of the vocal tract with regard to a quasi-neutral reference shape.

$$\begin{pmatrix} \Delta P_{s,1} \\ \Delta P_{s,2} \\ \cdots \\ \Delta P_{s,N} \end{pmatrix} = VW^{-1}U^T \begin{pmatrix} \Delta F_1 \\ \Delta F_2 \\ \Delta F_3 \end{pmatrix} \tag{4}$$

Inverse maps may involve 2-dimensional articulatory models, which are less flexible than general-purpose area function models because they comprise lip, jaw and tongue positions and contours, to constrain inverse solutions. These models do not guarantee either obtaining anatomically plausible solutions under all circumstances. Also, they may be unable to restitute observed acoustic data exactly.

2.4 Static constraints

Most parameters must stay within physiologically or physically plausible intervals. Also, one may wish to insert into the map articulatory data that have been recorded directly.

Formally, external static constraints are inserted by a change of variables $P = \frac{A+B}{2} + \frac{A-B}{2}\tanh(P_v)$. Symbol P_v designates virtual parameters, which are free to vary and P control parameters, which are constrained to stay within an interval. Constants A and B fix the limits of the interval of variation. They may differ for each parameter. Direct linear link (3) is then replaced by an expression that involves the product of two Jacobian matrices. The only difference is that virtual parameters P_v replace physiological parameters P. All other operations are identical to those that have been outlined above.

The default constraints are the following. The downstream truncated cone cross-section adjacent to the glottis is constrained to remain in the interval $0cm - 1.5cm^2$; the remaining cross-sections are in the interval $0cm - 12cm^2$ and the default tract length is between $16 - 18cm$.

2.5 Error correction

Inverse map (4) is used to compute *temporal* parameter increments from observed *temporal* frequency increments. One expects small errors to occur that would accumulate while iterating the mapping. Frequency errors are therefore corrected as follows. Formant *errors* δF (i.e. the difference between mapped and observed formant frequencies) together with the corresponding parameter *errors* δP also obey relations (3). Parameter errors can therefore be predicted from observed formant errors and used to correct the model parameters so that the formant errors decrease. This error correction step can be repeated several times. In practice, this means that if the temporal iteration proceeds without warning, the natural frequencies of the computed tract shapes are guaranteed to agree with the observed formant frequencies to within a fixed tolerance, which may be as small as $0.01Hz$.

2.6 Iteration and initialisation

Once temporal increments ΔP_i of the model parameters have been obtained at time t_m as a function of temporal increments ΔF_i of the observed formant frequencies, model parameters at time t_{m+1} can be calculated by adding the increments to the parameters obtained at time t_m. That is, the calculated motion of the model parameters can be obtained from the observed motion of the formant frequencies by iterating direct and inverse maps (3) and (4). This iteration must be initialized. In practice, initialization is carried out with a modeled vocal tract shape the natural frequencies of which are known. A typical choice is the quasi-uniform tract. Formant frequency motion is then obtained by interpolating between the known initial frequencies and the first observed frequencies, from whereon the iteration proceeds via the observed formant data. The number of time steps included in the interpolation is chosen to keep the formant increments small so that the approximations involved in map (3) apply.

3. METHODS

3.1 Similarity

Mapped and observed area function cross-sections have been compared quantitatively via intercorrelation. The intercorrelation has been retained as a measure of similarity because it is unaffected by affine transforms of the cross-sections. Before computing similarities, the computed cross-sections have been linearly interpolated to equalize the number of computed and observed cross-sections. The interpolated conical cross-sections have then been transformed into right cylinders of the same volume, because observed area functions have been reported as concatenations of cylinders. A preliminary experiment has shown that log-transforming the cross-sections causes the similarities between observed and computed tract shapes to be somewhat lower.

3.2 Experiments and statistical processing

For each vowel quality and speaker, formant-to-area inversion has been carried out with an area function modeled by 8, 12 and 16 truncated cones. Each inversion has been completed with the slow deformation and minimal deformation constraints that are discussed above, as well as two additional kinetic constraints. Also, each has been performed with the default length fixed to 17*cm* or to the observed tract length, as well as the lip area cross-section set to the observed cross-section or left free to vary in the limit of the static interval 0 to 12 cm^2. The purpose has been to test whether speaker-specific anatomical constraints enable increasing the similarity between inferred and observed cross-sections. The lip cross-section has been chosen because it may be estimated non-invasively via video imaging, for instance. The total number of inversions carried out per speaker and vowel category has therefore been equal to 48. The results that are reported here exclusively focus on the effect of the labial constraint.

The raw similarities have been summarized by means of their quartiles. The effects of kinetic and anatomical constraints and model complexity have been explored via multiple linear regression analyses of the morphological similarities obtained for each speaker and vowel quality. Finally, inter-constraint differences have been tested statistically by means of non-parametric methods because the number of data has been small and assumptions of equal variance, independence and Gaussian distribution have not been met under all circumstances.

3.3 Corpora

Four published corpora have been used for testing ([2] - [5]). They report observed area functions and formant frequencies of 8 speakers sustaining a total of 77 vowels. Acoustic and morphological data have been used as published. An exception is [2] who reports the tract cross-sections of a female speaker without reporting her formant frequencies. Therefore, we have obtained these a posteriori by means of a tract model into which the observed cross-sections have been inserted.

4. RESULTS

4.1 Acoustic accuracy

The acoustic accuracy, that is, the maximum difference between the observed and mapped formant frequencies has been equal to $0.1 Hz$.

4.2 Summary of morphological similarity data

Table 1 reports the median similarity between observed and mapped tract cross-sections. The number N of concatenated conical pipes that simulate the vocal tract area function has been 8, 12 or 16. The default tract length has been 17*cm* and the slow deformation constraint has been applied. Each table entry describes 77 vowels sustained by a total of 8 speakers.

Table 1: Median similarity between observed and mapped tract cross-sections per experimental condition (all vowels)

	model complexity		
labial constraint	$N = 8$	$N = 12$	$N = 16$
no	0.83	0.81	0.81
yes	0.86	0.86	0.85

Table 2 and Table 3 respectively report the median similarity between observed and mapped tract cross-sections for 25 rounded back vowels and 52 front and unrounded back vowels sustained by a total of 8 speakers.

Table 2: Median similarity between observed and mapped tract cross-sections per experimental condition (rounded back vowels)

	model complexity		
labial constraint	$N = 8$	$N = 12$	$N = 16$
no	0.72	0.72	0.68
yes	0.74	0.74	0.67

Table 3: Median similarity between observed and mapped tract cross-sections per experimental condition (front and unrounded back vowels)

	model complexity		
labial constraint	$N = 8$	$N = 12$	$N = 16$
no	0.88	0.87	0.86
yes	0.89	0.89	0.89

4.3 Regression analyses

The purpose of the linear regression analyses has been exploratory. The objective has been to survey the effect of the labial cross-section constraint on the similarities between recorded and computed morphological data. The constraint has been handled as a dummy variable. Other variables that have been involved (e.g speaker, vowel category, default tract length and kinetic constraints) have been considered influencing the outcome uncontrollably.

Table 4 reports for how many vowels the insertion of the labial constraint has statistically significant increased (+) or decreased (−) morphological similarity, or has no statistically significant effect (.). The count is broken down according to whether all or only rounded back vowels have been involved. Table 4 shows that for a large majority (≥ 90%) of vowels a significant increase of the similarity, or no significant effect has been observed when the free variation of the lip section has been replaced by the fixed observed lip cross-section.

Results reported in Table 4 can only be indirectly compared to results reported in Tables 1 - 3 and 5. The reason is that regression analyses have involved similarity data obtained under a wider range of conditions than those that have been involved in Tables 1 - 3 and 5. Some of these conditions have had deleterious effects on morphological similarity. These negative effects have been successfully compensated for by the labial constraint, which therefore has a more favorable impact on the data reported in Table 4.

Table 4: Number of vowels for which the insertion of the labial constraint has statistically significantly increased (+) or decreased (−) morphological similarity, or has had no statistically significant effect (.)

all			rounded back		
+	−	.	+	−	.
33	8	36	16	2	7

4.4 Statistical comparison

Formant-to-area mappings have been performed with the lip cross-section fixed to the observed value or left free to vary. Linear regression analyses have suggested that fixing the lip section to their measured value increases similarity between observed and mapped area functions. Three (one for each number N of conical ducts) paired (fixed versus free lip cross-section) Wilcoxon Signed Ranks tests or Sign Tests have been carried out on similarities involving 77 vowels (i.e. data summarized in Table 1). Table 5 summarises the Wilcoxon Signed Rank Test results. One sees that for $N = 12$ and 16 the similarity differences between maps with and without labial constraints have been statistically significant.

Table 5: Differences between similarities for fixed and free labial cross-sections; summary of Wilcoxon Signed Rank Tests, N: number of concatenated conical ducts, N.S./S.: not statistically / statistically significant

all vowels		
$N = 8$	$N = 12$	$N = 16$
N.S.	S.	S.

The Sign test has confirmed the results obtained by means of the Wilcoxon Signed Rank test. A Bonferroni correction that accounts for multiple comparisons does not suggest modifying the general conclusion.

5. DISCUSSION

5.1 Inter-vowel category differences

The morphological validity of the computational recovery of area functions depends on vowel category. Back rounded vowels have been the most difficult to retrieve (Table 2). Tests which are not reported here show that inter-vowel category differences have been statistically significant for all speakers. The special status of French back rounded vowels with regard to acoustic-to-area mapping is a known problem that has been discussed in the literature [6]. At present, the consensus appears to be that two qualitatively different posterior rounded area functions exist the first three formants of which must be identical. These categories are phonetically relevant. One observes, for instance, that the similarity between observed and mapped area functions is higher for Russian than for French [u]. This is expected knowing that Russian [u] is more posterior and French [u] more anterior, and that the acoustic-to-area mapper that is used here appears favoring area function models that are posteriorly constricted (Figures 3 and 4). Also, results reported in Table 2 suggest that fixing the labial cross-section to its observed value does not enable disambiguating sustained ”anterior” and ”posterior” [u]. Anterior rounded vowels appear to be less of a problem.

5.2 Model complexity

Formant-to-area mappings have been carried out with the number of elementary ducts equal to 8, 12 or 16. Exploratory data analyses (e.g. upper halves of Tables 1 to 3) and statistical tests (which are not reported here) suggest that increasing the number of elementary ducts, the succession of which simulate the area function, decreases the similarity between observed and mapped area functions, or has no statistically significant effect. The explanation is that increasing the number of concatenated ducts increases the number of unknown quantities that must be calculated by means of a fixed number of formant frequencies. One therefore expects the similarities to decrease.

Exceptions may be observed, however, for which a model complexity increase does not decrease the similarity between observed and mapped cross-sections. The lower halves of Tables 1 and 3 indeed suggest that when the labial constraint is active the similarities may not decrease when the number of ducts increases. The trend towards decreasing similarities with increasing complexity may indeed be neutralized by a mechanism that is explained in the section hereafter.

5.3 Observed versus free labial cross-sections

Formant-to-area mappings have been performed with the lip cross-section fixed to the observed value or left free to vary. The summary of raw data, exploratory data analyses and statistical tests suggest that inserting the labial cross-section improves the similarity between observed and mapped area functions, or has no statistically significant effect. An explanation is that inserting the observed labial cross-section increases its morphological veracity of the area function model and therefore the similarities between observed and mapped cross-sections.

A possible interaction between the labial constraint and model complexity can be observed in Tables 1 to 3 and 5. First, the lower halves of Tables 1 to 3 show that the expected decrease of the similarities with an increasing number of cross-sections is slowed down when the labial constraint is active. Second, Table 5 shows that the labial constraint does not improve statistically significantly the similarity between original and mapped vocal tracts when the number of elementary cones equals 8. Both observations may be explained as follows. The elementary ducts have been of equal length and fixing the labial cross-section makes itself felt not only at the labial end, but also tract-internally. For instance, if the elementary ducts were cylindrical and their number equal to N then the labial constriction would fix the area function over a distance equal to L/N, which would be the longer the smaller the number N of elementary cylinders. Labial cross-section constraints and model complexity therefore have contradictory requests with regard to the number of cross-sections that are free to vary. With regard to model complexity, the number of elementary ducts the cross-sections of which are unknown should be as small as possible. With regard to the insertion of labial constraints, the same number should be as large as possible to decrease the constraint's tract-internal effects. These contradictory requests may explain the data patterns that are observed in Tables 1-3 and 5.

5.4 Conclusion

The acoustic precision of the inverse mapper has been $0.1 Hz$. The tract cross-sections of rounded back vowels have been the most difficult to infer computationally. Increasing model complexity in terms of the number of free model parameters has decreased the similarity between observed and mapped area functions statistically significantly, or has had no statistically significant effect. Labial constraints and model complexity interact. Inserting labial constraints into the inverse mapper has increased the similarity between observed and mapped area functions statistically significantly or has had no statistically significant effect. Numerically speaking, the increase in morphological similarity has been less than 10%.

Acknowledgements

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324.

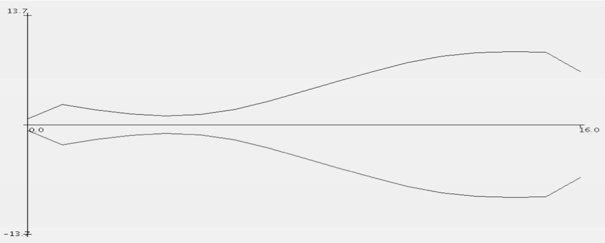

Figure 1: Mapped area function, French vowel [a], female speaker; the area function model comprises 16 truncated cones; the intercorrelation with the observed cross-sections equals 0.93; the vertical and horizontal axes are ±$13.7 cm^2$ and $0-16 cm$, respectively.

REFERENCES

[1] G. P. Scavone, "An acoustic analysis of singe-reed instruments with emphasis on design and performance issues and digital wave guide modeling techniques," *PhD Thesis*, Standford: Standford University, 1997.

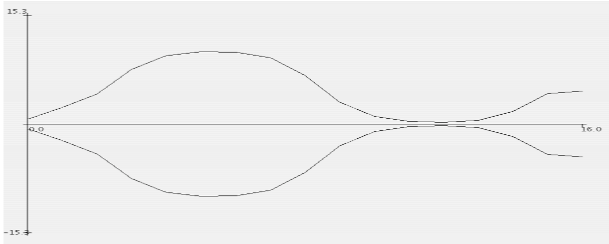

Figure 2: Mapped area function, French vowel [i], female speaker; the area function model comprises 16 truncated cones; the intercorrelation with the observed cross-sections equals 0.92; the vertical and horizontal axes are ±$15.3 cm^2$ and $0-16 cm$, respectively.

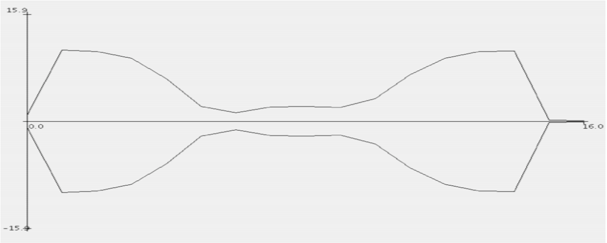

Figure 3: Mapped area function, French vowel [u], female speaker; the area function model comprises 16 truncated cones; the intercorrelation with the observed cross-sections (Figure 4) equals 0.66; the vertical and horizontal axes are ±$15.9 cm^2$ and $0-16 cm$, respectively.

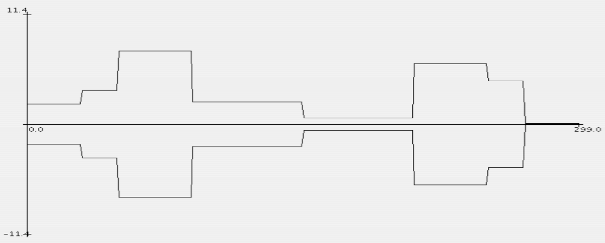

Figure 4: Observed area function based on MRI data, sustained French vowel [u], female speaker; the area function comprises 8 cylinders of unequal lengths; the vertical and horizontal axes are ±$11.4 cm^2$ and $0-299$ in arbitrary units, respectively.

[2] B. Story, I. Titze, E. Hoffman,"Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 100, pp. 537-554, 1996.

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

[4] M. Mrayati, "Contributions aux études sur la parole," *PhD Thesis*, Grenoble : Institut Polytechnique de Grenoble, 1976.

[5] G. Fant, "Acoustic Theory of Speech Production," The Hague: Mouton & Co, 1960.

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