Journal of Phonetics (1995) **23**, 221–229

Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model

Pierre Badin, Denis Beautemps, Rafael Laboissière and Jean-Luc Schwartz

Institut de la Communication Parlée, URA CNRS N°368–INPG–Université Stendhal, 46 Avenue Félix Viallet, F-38031 Grenoble Cedex 01, France

(Received 9th February 1994, and in revised form 7th October 1994)

This study deals with the ill-posed problem of inversion of the articulatory-to-acoustic relationship, i.e., the recovery of vocal tract geometry from formant frequencies. A small database of articulatory-acoustic data has been established for one subject. A midsagittal-to-area function conversion model, which works both for vowels and fricative consonants, has been developed from these data. This model has finally been used as a major constraint for an optimization algorithm based on a gradient descent technique, in order to regularize the *ill-posed* inversion problem. Other spatial and temporal smoothing constraints have also been used. Single vocal tract configurations, as well as entire [VC] sequences, could be recovered using adequate initial conditions.

1. Introduction

In the domain of speech production research, more insight into speech production phenomena is still needed; particularly extensive knowledge of the articulatory-to-acoustic relation is a key factor in domains such as speech synthesis (Fant, 1991) or speech coding (Flanagan, Ishizaka and Shipley, 1980). It is a well-established fact that the articulatory-to-acoustic relation is a many-to-one function (cf. Atal, Chang, Mathews and Tukey, 1978), and that the inversion of such a function is an *ill-posed problem,* i.e., it is not certain that the solution exists, is unique, and continuously dependent on the initial conditions. Inversion of speech is thus a very challenging problem (Abry, Badin and Scully, 1994), with important perspectives such as the learning of articulatory gestures by a speech robot (Laboissière, Schwartz and Bailly, 1990), speech perception issues (Schwartz, Beautemps, Arrouas and Escudier, 1992), or speech low bit-rate coding (Rahim, Kleijn, Schroeter and Goodyear, 1991). It is known that a solution to an ill-posed problem can be found if the problem is regularized using constraints (Hadamard, quoted by Lavrentiev, 1967). Three types of constraints can be used to help find realistic solutions to the problem: spatial constraints (Boë, Perrier and Bailly, 1992), temporal constraints (Schroeter

0095-4470/95/010221 + 09 $08.00/0

and Sondhi, 1989) and good initial conditions for the optimization algorithms (Schroeter and Sondhi, 1994).

The aims of the present study were thus twofold: (1) developing a model of midsagittal-to-area function conversion for vowels and consonants based on a coherent set of geometric and acoustic data obtained for one subject; (2) performing the inversion of the articulatory-to-acoustic relationship for some [VC] sequences, using the midsagittal-to-area model as a major constraint.

2. A new model of midsagittal-to-area function conversion

The difficulty of measuring area functions directly on subjects has often been evoked in the literature (cf. Baer, Goore, Gracco and Nye's (1991) recent study). On the other hand, indirect methods are not completely satisfactory either, in the sense that it can not be proved that they yield the true area functions. For instance, Sondhi and Resnick's method (1983) to retrieve tract shapes by acoustic means may yield true area functions, but with the important restrictions that the talker has to be quiet, i.e. cannot actually phonate speech (depriving him/herself of acoustic feedback), and that the lip movement might be constrained due to the tight contact with the mouthpiece. Therefore a combined approach has been adopted: a set of midsagittal profiles and corresponding formants have been measured on a subject, and a model of midsagittal-to-area function conversion based on these data has been developed. The fact that the model is based on the strong constraint that it should work for sounds as different as vowels and steady-state consonants, and be coherent at both midsagittal and acoustic levels, should ensure the reliability of the area functions determined in such a way.

2.1. *A coherent set of geometric and acoustic data*

Midsagittal profiles were determined by conventional teleradiography for the sustained production of vowels [a, i, u] and voiceless fricatives [f, θ, s, ʃ, ç, x] by one French male subject. Front photographs of the lips and sound recordings were simultaneously obtained, and the formant frequencies of the different configurations were then extracted from the speech signal. Midsagittal functions, i.e. vectors of 50 midsagittal distances and corresponding section lengths, were then obtained from the midsagittal profiles, using a conventional semipolar coordinate grid (cf., e.g., Beautemps, Badin and Laboissière, 1995). The vocal tract midline was determined as the line joining the centers of gravity of each section limited by two contiguous lines of the grid. The midsagittal distance is defined as the distance between the upper and lower contours of the midsagittal profile measured on the lines perpendicular to the vocal tract midline.

2.2. *The new model*

The proposed model is an extension of the Heinz and Stevens (1965) "$\alpha\beta$ model". It has been motivated by the fact that we did not succeed in extending Perrier, Boë and Sock's (1992) $\alpha\beta$ model to consonants, designed for vowels. The major improvement of the proposed model over Perrier *et al.* is that α continuously depends on both midsagittal distance $d(x)$ and vocal tract position x along the midline. For each section, the area $A(x)$ is defined as $A(x) = \alpha(d, x) \cdot d(x)^{\beta}$ (except for the larynx region which, because its pyramid-shaped structure can not be correctly mapped by the $\alpha\beta$ model, is represented by a fixed uniform tube of 1.8 cm^2 area and 2 cm length, following Fant's [1960] suggestion). The coefficient β is fixed at 1.5 in reference to the bi-parabolic geometric model of the shape of the vocal tract (Perrier *et al.*, 1992). The curve $\alpha(d, x)$ is obtained by linear interpolation between two bounds $\alpha_{inf}(x)$ and $\alpha_{sup}(x)$ not dependent on d: $\alpha(d, x) = \alpha_{inf}(x)$ if $d < d_{inf}$ and $\alpha(d, x) = \alpha_{sup}(x)$ if $d > d_{sup}$. Values of respectively 1 and 2 cm for d_{inf} and d_{sup} are used, in agreement with Perrier *et al.* (1992). The curves $\alpha_{inf}(x)$ and $\alpha_{sup}(x)$ (see Fig. 1) are defined by Fourier series of sines and cosines of argument $n\omega x$, where $\omega = \pi / l_{tot}$, with a limited number of coefficients, l_{tot} being the vocal tract total length. Both curves are described by their fundamental and their first three harmonics, leading to a total of 14 coefficients. In order to avoid null or negative values, the α coefficients are limited to a minimal threshold set to 0.01.

2.3. *Network implementation*

The model described above had to be fitted to the subject: a unique set of two curves $\alpha_{inf}(x)$ and $\alpha_{sup}(x)$ must be determined for the whole corpus in order that the four or five first formants computed from the midsagittal functions through the new midsagittal-to-area conversion model and the acoustic model fit the formant frequencies measured on the subject optimally.

The problem reduces thus to the determination of the two sets of seven Fourier coefficients defining the curves α_{inf} and α_{sup}, and minimizing the quadratic error between actual and desired formants for all configurations. The solution is obtained by an optimization algorithm based on a gradient descent technique, which requires the implementation of the models to be networks of analytically differentiable basis functions. The total chain of transformations of the models is shown in Fig. 2 (left), where the various functional components can be distinguished. The propagation of acoustic plane waves in the vocal tract is classically modeled as a series of uniform

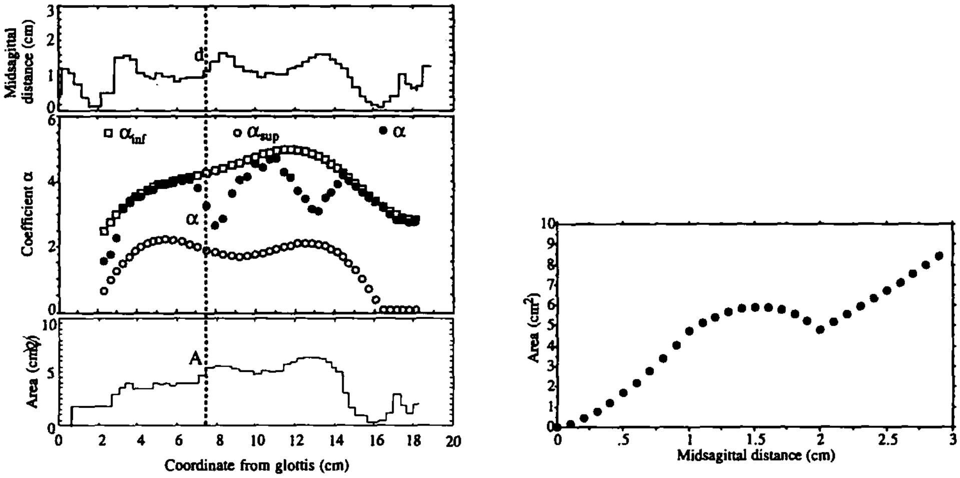

Figure 1. Obtaining the area function from the midsagittal function. The different steps of the new model are shown on the left side of the picture: $\alpha(x, d)$ is interpolated between $\alpha_{inf}(x, d)$ and $\alpha_{sup}(x, d)$, and $A(x) = \alpha(x, d)d^{\beta}$. An example of $A(x)$ curve for $x = 10$ cm is shown on the right side.

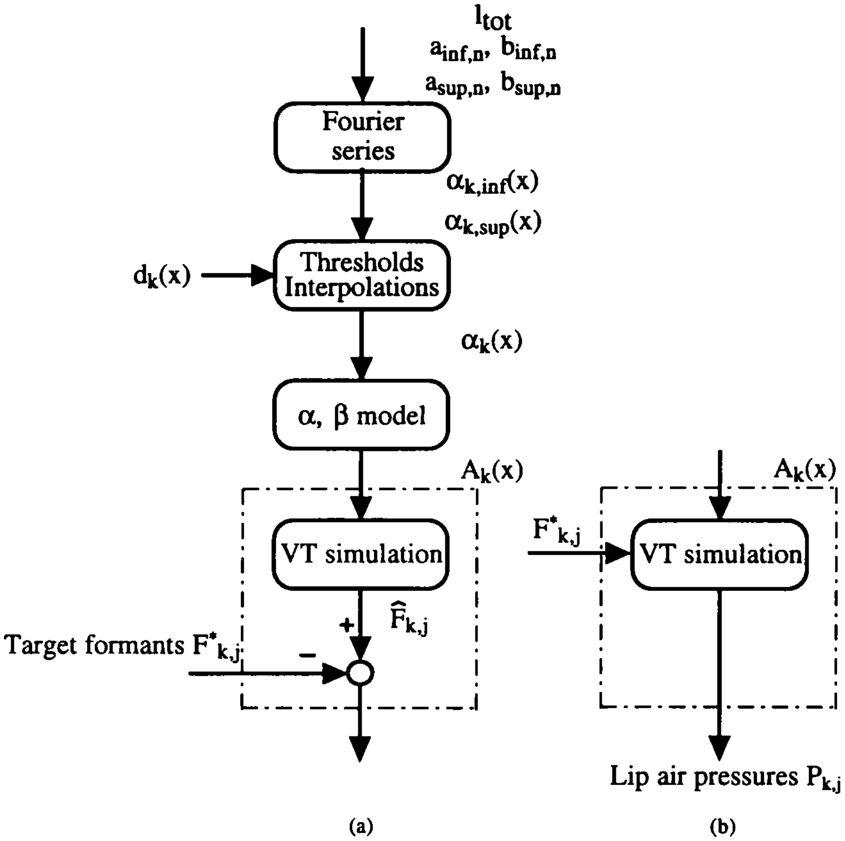

Figure 2. Graphs implementing the midsagittal-to-area and acoustic wave propagation models: (a) complete model; (b) modified VT simulation).

tube sections. The gradient of the error E with respect to the Fourier coefficients cannot be easily derived because the algorithm implementing the area-to-formants model cannot be described in an analytical way. In order to overcome this difficulty, the amplitude of the air pressure along the vocal tract, and particularly at the lips, is computed recursively, from a sinusoidal excitation signal at the glottis, of unit amplitude and given frequency; this computation, performed in the frequency domain, is based on the fact that the ratio of the pressure derivatives on each side of the junction between two sections is inversely proportional to the ratio of the areas. A lossless case is assumed; the inductive part of lip radiation and wall vibration impedances are taken into account. This wave propagation model is represented in Fig. 2 (right). Using the fact that the air pressure is zero at the lips for all resonance frequencies of the vocal tract, the error between the desired frequencies $F^*_{k,j}$ and the attained frequencies $F_{k,j}$, where k $(1 \leq k \leq 9)$ is the configuration index and j $(1 \leq j \leq 5)$ the formant index, is finally evaluated as:

$$E = \sum_{k=1}^{9} e_k^2 \quad \text{with} \quad e_k^2 = \sum_{j=1}^{5} P_{k,j}^2$$

where $P_{k,j}$ is the air pressure at the lips.

During the optimization phase, nine identical networks such as the one described in Fig. 2 were operated in parallel, in order to optimize the Fourier coefficients over the whole corpus of the nine configurations. The error backpropagated through the networks was the sum of the errors computed by each network for each configuration, including extra cost functions related to geometric parameters, i.e., constraining the lip area to be close to the measured values, and constraining constriction areas to be close to their corresponding aerodynamically equivalent areas available for [f, θ, s, ʃ, ç] (cf. Beautemps *et al.* (1995)).

2.4. *Results of the model optimization*

Fig. 1 shows an example of the α functions obtained for [s]. It is important to note that, in all cases, the α_{sup} curve lies entirely under the α_{inf} one, which means that, for a given x coordinate, $\alpha(d, x)$ decreases when d increases and vice-versa, in the range between d_{inf} and d_{sup}. The function $A(d)$ is thus expected to attain a maximum value for a certain d_{max}, as exemplified for $x = 10$ cm on the right side of Fig. 1.

Fairly good formant frequencies were obtained for [a, i, f, s, θ, ç], but mismatches affected [u, ʃ, ç]. Therefore minor corrections on midsagittal distances have been realized. These adjustments ensured a much better fit (3.2% error on the average), even though some errors were still too high (up to 16% for F_1 of [i]). The highest errors are related to F_1, which could be explained partly by a higher relative measurement error (the absolute error is constant). Note that the formant errors are not normalized in the classical way, but that their weighting depends on the pressure distribution in the vocal tract near the lips, which depends on each configuration; this could also explain the relatively large error of F_1 for [i]. The mean errors are 8.0% for F_1, 3.1% for F_2, and 1.6% for F_3.

3. Vocal tract geometry recovery

In this section, different attempts to recover vocal tract geometry from measured formants for the same subject used for determining α are described. The articulatory space of the midsagittal functions (vectors of 50 sections) has been chosen as input space because it can be directly compared with experimental data; an articulatory model would have had less parameters, but would have raised the need for a geometric normalization between the subject and the articulatory model. The output space consists of vectors of five formants. As the number of degrees of freedom of the input space is much higher than that of the output space, constraints are needed to regularize the problem and to reduce the size of the set of possible solutions known as articulatory fibers since Atal *et al.* (1978). The inversion has been performed with the help of the optimization algorithm already used for the model parameters determination (see above), using spatial and temporal constraints.

The midsagittal-to-area conversion model is a major source of constraint. Fig. 1 (right) shows that the function $A(d)$ is not monotonous and reaches a certain local maximum for a given value d_{max} of d. This property, in conjunction with the constraint that prevents midsagittal distances from being too large (implemented as a cost function J_1), reduces the range of possible variations of the corresponding area function: indeed, if a need appears to increase the area at some location in the vocal tract in the optimization procedure, the cost function J_1 will limit the increase of midsagittal distance, and the optimal solution will likely be around the local maximum of the $A(d)$ function. One can thus see that the midsagittal-to-area conversion model plays indirectly the role of constraint in the optimization procedure. In addition to this constraint, another constraint of *spatial smoothing* is used, implemented as the cost function J_2. The definitions of J_1 and J_2 are:

$$J_1 = \frac{1}{2}\sum_{i=2}^{N} d_i^2(t) \quad \text{and} \quad J_2 = \frac{1}{2}\sum_{i=2}^{N} [d_i(t) - d_{i-1}(t)]^2,$$

where $d_i(t)$ is the midsagittal distance of section i at instant t, and N the number of sections of the configuration considered.

3.1. *Recovery of single configurations*

The influence of context on the ability of the method to recover the midsagittal function from formants has been evaluated for fricatives. Moreover, in order to supplement the corpus, inversion has been used to determine the midsagittal and area function of [y] for the subject.

Formant trajectories were tracked for nonsense words [aza] and [aʒa] uttered by the subject. The recovery of the midsagittal functions for the fricatives [z] and [ʒ] from their formants measured during the most stable part of the trajectories was then attempted. The midsagittal distances obtained for the sustained [a] were used as initial conditions for the optimization method. For each fricative, the section lengths were given the values of the corresponding lengths of the sustained voiceless fricatives in the corpus (indeed, no major articulatory differences exist between voiced and voiceless fricative cognates). Fig. 3 shows that the midsagittal and area functions measured for the sustained [s] are fairly close to those recovered for the [z] in vocalic context.

3.2. *Recovery of trajectories*

The experiments described in this section extend the previous experiments to whole trajectories of formants and midsagittal functions. Formants of the [az] and [aʒ] transitions extracted from [aza] and [aʒa] sequences were measured at regular points in time. The principle previously applied for single configurations was used, a major difference being that the midsagittal function recovered for a given time slot was used as the initial condition for the next time slot; the solution for the first time slot was given as the midsagittal function of [a] slightly tuned so that its formants were fitting those of the first time slot. This choice of initial conditions for each local inversion clearly acts as a *temporal smoothing constraint.* The lengths of each section were interpolated—and frozen—between those of the initial vowel and those of the target fricative. Thus, only the midsagittal distances for each slot were recovered. The only spatial constraint used, apart from the indirect constraining effects of the midsagittal-to-area function model itself, was the smoothing constraint J_2. The midsagittal functions recovered for the target consonants are close to the reference ones, particularly in the region of the constriction. Fig. 4 shows an example of the trajectories of midsagittal and area functions for [az]. It has been verified that the

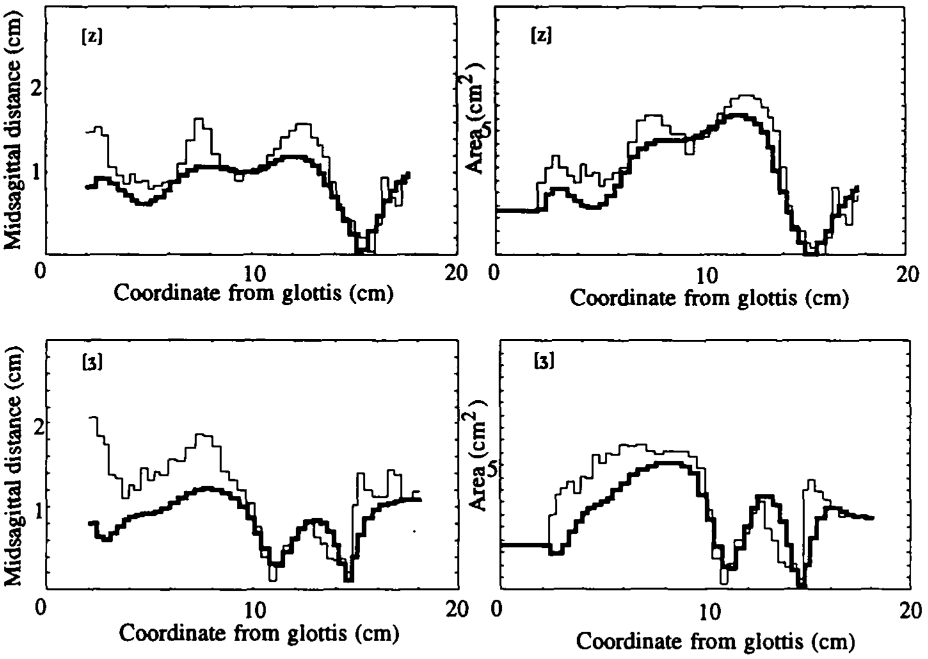

Figure 3. Measured (—) and recovered (—) midsagittal (left) and area (right) functions for [z] (top) and [ʒ] (bottom).

trajectories of the constriction area for both consonants are close to the aerodynamically equivalent areas measured on the same subject uttering the same sequences. These areas were obtained with the help of a *Rothenberg mask*, using the *orifice equation* (cf., e.g., Scully, 1971). These preliminary examples show that it is possible, for some [VC] sequences, to recover geometric trajectories from formant trajectories using simple local regularization criteria (initial conditions and spatial smoothing).

4. Conclusions and perspectives

A small coherent articulatory-acoustic database has been built for one subject. A midsagittal-to-area function conversion model based on these data has been developed: this model works both for vowels and fricative consonants, which is an original feature. However, some limitations should be noticed: the larynx has been simplified into a single uniform tube, because of its complex geometry; the lip horn could be better taken into account with the help of an *acoustic equivalent* (cf. Badin, Motoki, Miki, Ritterhaus and Lallouache, 1994). The obtained area functions should be compared with 3D data obtained on the same subject, using Magnetic Resonance Imaging, Ultrasonic Imaging, Electropalatography, etc. The methodology used to optimize the model should be extended to other subjects.

Using this model, it has been possible to recover vocal tract geometry, specially the constriction region, from measured formants for single configurations. As well, entire midsagittal function trajectories could be recovered from formant trajectories for some [VC] sequences. However, an important limitation should be

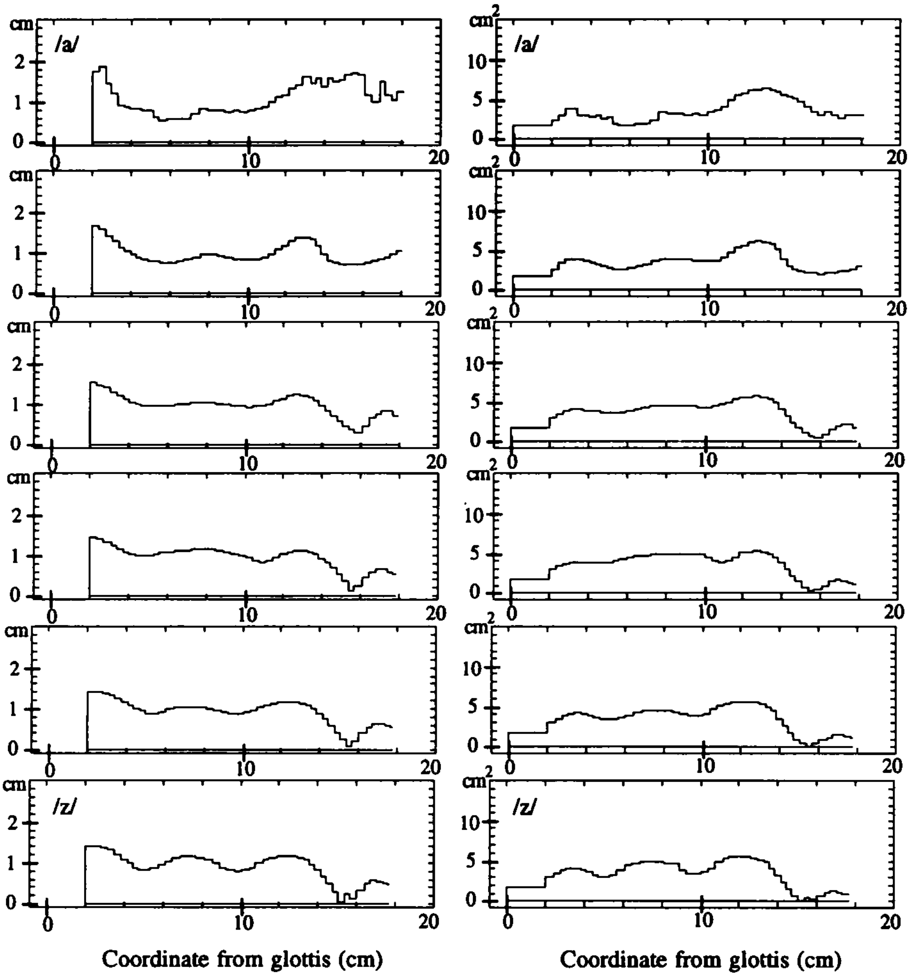

Figure 4. Midsagittal (left) and area (right) functions recovered for [az]; sampled every 40 ms from [a] (top) to [z] (bottom).

overcome: vocal tract length has not been recovered. A parametric model of the vocal tract length along its midline should be developed, in order to reduce the number of length parameters to be recovered.

The scope of this preliminary study was limited to the articulatory-to-acoustic relationship, and did not intend to apply to running speech, which would raise the problem of automatic formant tracking. No complete spectrum match has been attempted either, in the frame of this study, as this would imply recovering also both source location and source spectrum, still two challenging problems. However, despite its limitations, our approach has proven to be promising for the recovery of consonantal articulation, when adequate constraints are used.

This work has been partly supported by the French CNRS and the European ESPRIT project *SPEECH MAPS*. The authors are thankful to the editor Anders Löfqvist and to the reviewer Juergen Schroeter for their stimulating comments and advice.

References

Abry, C., Badin, P. & Scully, C. (1994) Sound-to-gesture inversion in speech: The Speech Maps approach. ESPRIT Research Report N°6975. In K. Varghese, S. Pfleger & J. P. Lefèvre (Eds), *Advanced speech applications* pp. 182–196. Berlin: Springer Verlag.

Atal, B. S., Chang, J. J., Mathews, M. V. & Tukey, J. W. (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer–sorting technique. *Journal of the Acoustical Society of America*, **63**, 1535–1555.

Badin, P., Motoki, K., Miki, N., Ritterhaus D. & Lallouache T. M. (1994) Some geometric and acoustic properties of the lip horn. *Journal of the Acoustical Society of Japan* (*E*), **15** (4), 243–253.

Baer, T., Goore, J. C., Gracco, L. C., & Nye, P. W. (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America*, **90**, 799–828.

Beautemps, D., Badin, P. & Laboissière, R. (1995) Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. *Speech Communication*, **16**, 27–47.

Boë, L. J., Perrier, P. & Bailly, G. (1992) The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, **20**, 27–38.

Fant, G. (1960) *Acoustic Theory of Speech Production*. 's-Gravenhage: Mouton & co.

Fant, G. (1991) What can basic research contribute to speech synthesis? *Journal of Phonetics*, **19**, 75–90.

Flanagan, J. L., Ishizaka, K. & Shipley, K. L. (1980) Signal models for low bit–rate coding of speech. *The Journal of the Acoustical Society of America*, **68**, 780–791.

Heinz, J. M. & Stevens, K. N. (1965) On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. In *Proceedings of the 5th International Congress of Acoustics*, Paper A44.

Laboissière, R., Schwartz, J. L. & Bailly, G. (1990) Motor control for speech skills: a connectionist approach. In D. S. Touretzki, J. L. Elman, T. L. Sejnowski & G. E. Hinton (Eds), *Connectionist models, Proceedings of the 1990 Summer School*, pp. 319–327. San Mateo, California. Morgan Kaufmann Publishers.

Lavrentiev, M. M. (1967) *Some improperly posed problems of mathematical physics*. Berlin, Heidelberg, New York: Springer.

Perrier, P., Boe, L. J. & Sock, R. (1992) Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modeling the transition with two sets of coefficients. *Journal of Speech and Hearing Research*, **35**, 53–67.

Rahim, M. G., Kleijn, W. B., Schroeter, J. & Goodyear, C. C. (1991) Acoustic to articulatory parameter mapping using an assembly of neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, **Paper S7.20**, pp. 485–488.

Schroeter, J. & Sondhi, M. M. (1989) Dynamic programming search of articulatory codebooks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, **Paper S11.8**, 588–591.

Schroeter, J. & Sondhi, M. M. (1994) Techniques for estimating vocal tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, **2(1)**, **Part II**, 133–150.

Schwartz, J. L., Beautemps, D., Arrouas, Y. & Escudier, P. (1992) Auditory analysis of speech gestures. In M. E. H. Schouten (Ed.), *The Auditory Processing of Speech*, pp. 239–252. Berlin, New York: Mouton de Gruyter.

Scully, C. (1971) A comparison of /s/ and /z/ for an English speaker. *Language and Speech*, **14**, 187–200.

Sondhi, M. M. & Resnick, J. R. (1983) The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis. *Journal of the Acoustical Society of America*, **73**, 985–1002.