FORMANTS AND VOWEL SOUNDS BY THE FINITE ELEMENT METHOD

Antti Hannukainen, Teemu Lukkari, Jarmo Malinen and Pertti Palo[1]
Institute of Mathematics
P.O.Box 1100, FI-02015
Helsinki University of Technology, Finland.
[1] Corresponding author: Pertti.Palo@tkk.fi

Abstract

We study computationally the dynamics of sound production in the vocal tract (VT). Our mathematical formulation is based on the three-dimensional wave equation, together with physically relevant boundary conditions. We focus on formant and pressure information in the VT. For this purpose, we make use of anatomical data obtained by MRI by other researchers. More precisely, we carry out modal analysis on a geometric form of [øː] produced by a native Swedish speaker. Our results show encouraging evidence for the validity of the presented numerical model of the VT.

Keywords: Formant, speech acoustics, vowel production, wave equation, finite element method, modal analysis, articulatory speech synthesis.

1. Introduction

We study speech production by a physically faithful model. Our approach is based on a Partial Differential Equation from mathematical physics that describes the wave propagation in three-dimensional domains. This equation is known as *the wave equation*, and it is given (together with appropriate *boundary conditions*) in equation (2) below. In mathematical systems theory, this model can be studied in terms of *conservative* and *energy-dissipative linear systems*; for such systems, see Malinen et al. (2006); Malinen and Staffans (2006: 2007) and the references therein.

Articulatory models are a good way to learn about the speech production process. It can be said that a decent articulatory speech synthesiser will make it possible to understand fine details of speech production process. While realising such a synthesiser is still far in the future, this work represents a step on the path leading there.

In the past, the VT acoustics has been modelled in a number of different ways. The celebrated *Kelly–Lochbaum model* makes use of *reflection coefficients* obtained from a variable diameter tube (Kelly and Lochbaum 1962). Such reflection coefficients appear in, e.g., models from geophysics and in interpolation theory (see Foias and Frazho 1990). We remark that the Kelly–Lochbaum model is closely related to the *horn model* described by the Webster equation (see Fant 1970). More advanced two- and three-dimensional descendants of the Kelly–Lochbaum model are the *transmission line networks* that have been developed by El Masri et al. (1996: 1998); Mullen et al. (2006). For a recent review of these models and related topics, see Palo (2006).

The wave equation model (2) is mathematically more refined and physically more realistic than any of the models described in the previous paragraph. Unfortunately, the analytic solution of (2) in complicated domains (such as the human VT) is not possible. Instead, some numerical method must be used for the approximate solution of this model.

For this purpose, we use the *Finite Element Method* (FEM), which is a popular and well established method in computational science. This is the approach used by e.g. Lu et al. (1993) and Dedouch et al. (2002), too.

For the present computational approach, a fairly detailed geometric model of the VT is necessary. Nowadays, accurate anatomical data can obtained by using Magnetic Resonance Imaging (MRI). We are indebted to Dr. Olov Engwall (KTH) for kindly providing us with the required data and the associated experimental formant information.

The purpose of this paper is to present the modal analysis in an anatomical configuration of [øː] as produced by a native Swedish speaker. We obtain computationally resonance frequencies, which correspond to formants. Moreover, these formants identify the vowel [øː] correctly in a larger set of measured data.

2. Acoustic Model

As mentioned above, the wave equation is a fundamental Partial Differential Equation in acoustics and other areas of physics. It describes wave motion in a homogeneous medium. Deriving the wave equation for sound pressure starts by assuming that the *total pressure* $P = P(\mathbf{r}, t)$ can be expressed as

$$P(\mathbf{r}, t) = P_0 + p(\mathbf{r}, t) \tag{1}$$

where P_0 is the *static pressure*, and $p = p(\mathbf{r}, t)$ is its perturbation at point $\mathbf{r} = (x, y, z)$ at time t. The static pressure P_0 is independent of t and $\mathbf{r}$, and p is assumed to be small compared to P_0. With this notation, our acoustic model[1] is given by

$$\begin{cases} p_{tt} = c^2 \Delta p & \text{inside the VT,} \\ \frac{\partial p}{\partial \nu} = 0 & \text{at the walls of the VT,} \\ p = 0 & \text{at the mouth,} \\ p_t + c\frac{\partial p}{\partial \nu} = u & \text{at the glottis.} \end{cases} \tag{2}$$

Here $u = u(\mathbf{r}, t)$ is the glottis input, and c is the sound velocity in air in the VT. Now the computational problem is to find the pressure function $p(\mathbf{r}, t)$ for a given glottal input function $u(\mathbf{r}, t)$.

To derive (2) from "first principles", one needs to assume that some thermodynamic equation of state (such as $pV = nRT$ for ideal gas) holds, and that the entropy is kept constant. The topmost equation $p_{tt} = c^2 \Delta p$ in (2) — the wave equation itself — can be derived by a long linearisation argument involving the continuity equation, Euler equation and thermodynamic state equations; see, e.g., Fetter and Walecka (1980: Chapter 9).

The wave equation model (2) is sophisticated enough to capture most of the relevant properties of wave propagation in three-dimensional geometry (e.g., to detect cross modes). However, it does not model turbulence, shock formation, or losses due to viscosity and heat conduction. It can also be used as the theoretical starting point in deriving the Webster equation mentioned above.

[1] In standard mathematical notation, a variable as a subscript indicates differentiation with respect to that variable, Δp is the Laplacian of p, i.e. $\Delta p = p_{xx} + p_{yy} + p_{zz}$, and $\frac{\partial p}{\partial \nu} = \nu \cdot \nabla p$ stands for the derivative in the direction of the outer normal vector ν of the surface.

We also need to take into account the walls and both ends of the VT in the model (2). For this purpose, boundary conditions on these surfaces must be prescribed, and these are the three remaining equations in (2). We regard the mouth as an open end of an acoustic tube, and this is modelled by the *Dirichlet boundary condition* $p(\mathbf{r}, t) = 0$ for all $\mathbf{r}$ in the mouth opening for all times t. On the walls of the VT, we use the same *Neumann boundary condition* $\frac{\partial p}{\partial \nu} = 0$ that one would use at the closed end of a resonating tube. The validity of these two boundary conditions is discussed by Fetter and Walecka (1980: pp. 306-307). At the glottis, we use a special *scattering boundary condition* that specifies the ingoing sound pressure wave. Some motivation for this boundary condition can be found in the examples given by Malinen (2004: pp. 25–34).

Once the equation and boundary conditions are given, we proceed to solve the problem numerically. In our case, this means that given the ingoing wave $u(\mathbf{r}, t)$ at the glottis, we would like to compute the pressure distribution $p(\mathbf{r}, t)$ inside the VT. In this paper we solve an easier (yet relevant) problem related to the model (2); namely, we determine the resonance frequencies corresponding to a particular configuration of the VT. By Malinen and Staffans (2006: Theorem 2.3), the resonances of model (2) can be solved as follows: find the complex frequencies λ and their nonzero *eigenfunctions* $p_\lambda(\mathbf{r})$ such that the equations

$$\begin{cases} \lambda^2 p_\lambda = c^2 \Delta p_\lambda \text{ in VT}, \\ \frac{\partial p_\lambda}{\partial \nu} = 0 \text{ on walls}, p_\lambda = 0 \text{ on mouth, and } \lambda p_\lambda + c \frac{\partial p_\lambda}{\partial \nu} = 0 \text{ on glottis} \end{cases} \tag{3}$$

are satisfied. We remark that the equations (3) have nontrivial solutions $p_\lambda \neq 0$ only for some discrete values of λ. The imaginary parts of such particular λ's correspond to the angular frequencies of the formants.

3. Finite Element Method

The Finite Element Method (FEM) is an energy minimising interpolation method; see, e.g., Johnson (1987) for an elementary treatment. It can be used to approximately solve the variational forms of both the full time dependent problem (2) and the resonance problem (3).

To employ FEM, we first require a digitised geometric description for the boundary of the VT. Then we need to partition, using common software tools, the volume of the VT to an *element mesh* consisting of sufficiently many tetrahedrons. In this paper, we use a mesh of $n = 64254$ elements and piecewise linear shape functions. This number of elements is large enough to give accurate results in our setting.

Once the element mesh is completed, the implementation of FEM (in order to solve either (2) or (3)) is an exercise in computer programming using, e.g., MATLAB environment. Thus we obtain three $n \times n$ matrices, namely the *stiffness matrix* $\boldsymbol{M}$, the *mass matrix* $\boldsymbol{N}$, and an additional matrix $\boldsymbol{P}$ representing the glottis boundary condition in (3).

In the final step, we manipulate large systems of linear equations described by $\boldsymbol{M}$, $\boldsymbol{N}$, and $\boldsymbol{P}$. When treating problem (3), we solve the following linear algebra problem: find all complex numbers λ and corresponding nonzero vectors $\mathbf{x}(\lambda)$ such that

$$\lambda^2 \boldsymbol{M}\mathbf{x}(\lambda) + \lambda c \boldsymbol{P}\mathbf{x}(\lambda) + c^2 \boldsymbol{N}\mathbf{x}(\lambda) = 0 \tag{4}$$

is satisfied where c is the sound velocity. With some manipulations (Saad 1992), equation (4) can be written in the form

$$\boldsymbol{A}\mathbf{y}(\lambda) = \lambda \boldsymbol{B}\mathbf{y}(\lambda), \tag{5}$$

where $\boldsymbol{A} = \begin{bmatrix} -c\boldsymbol{P} & -c^2\boldsymbol{N} \\ \boldsymbol{I} & 0 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{M} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}$, and $\mathbf{y}(\lambda) = \begin{bmatrix} \lambda\mathbf{x}(\lambda) \\ \mathbf{x}(\lambda) \end{bmatrix}$. This eigenvalue problem can be immediately solved using, e.g., MATLAB.

The numbers λ computed from (5) are good approximations of the λ's appearing in (3), provided that the number n of elements is high enough. We also remark that for this numerical formulation, there are as many such numbers λ as there are elements in the mesh. However, only those that have smallest imaginary parts are interesting as they correspond to the lowest formants F1, F2, etc..

4. Data

Figure 1 shows a sliced representation of the VT geometry that we have used as the basis of our analysis. There are 29 slices, each consisting of 51 points, and they define the VT from glottis to mouth. For faster computation, the slices were down-sampled by taking into account only every fourth point.

The raw MRI data was collected from a native male speaker of Swedish while he pronounced a prolonged vowel in supine position. Engwall and Badin (1999) describe the MR imaging procedure and image post-processing. The vowel articulation was close to [øː]. Corresponding formant measurement data is also available on the same subject, and it is reported in the same article. The formants were estimated from speech recorded on a different occasion but with the same subject in a similar supine condition.

5. Results

The formants we obtained by solving (4) are shown in table 1. The computed formants F1 to F4 are roughly $3\frac{1}{2}$ semitones too high compared to the measured values. This offset between measured and computed formants has been estimated based on the first four formants. The bottom row in table 1 shows the computed formants multiplied by 0.817, which corresponds to a difference of $3\frac{1}{2}$ semitones. We will discuss the physical background of this discrepancy in section 6 below.

Table 1: Computed, measured, and scaled formants for [øː] in kHz

	F1	**F2**	**F3**	**F4**
Computed	0.68	1.35	2.71	3.79
Measured	0.50	1.06	2.48	3.24
Scaled	0.56	1.11	2.22	3.10

We also obtained from (4) the resonance modes p_λ (see (3)) corresponding to the formants F1-F4. These perturbation pressures are not given here in any physically relevant scale. Rather, they have been normalised so that the maximum deviation from the static pressure P_0 is either 1 or -1. Figure 2 shows isobars for the modes. Figures 3 and 4 show the pressure distributions of the modes. Figures 2 and 3 are plotted along a cross-sagittal mid-line cut shown in figure 1.

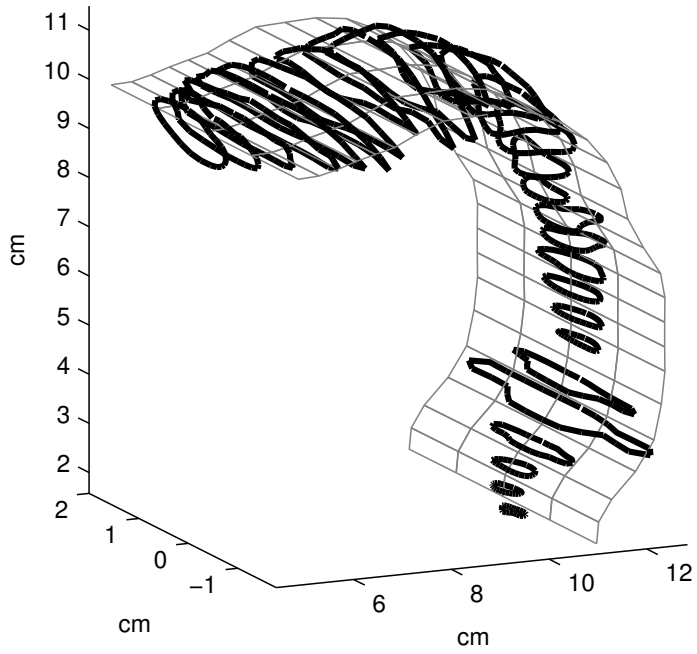

Figure 1: The vocal tract represented as 29 cross-sectional slices (bold lines) and the surface defining the cross-sagittal mid-line cut (grey lines) used in figures 2 and 3. Mouth is the last slice on the left and glottis the bottommost slice on the right.

6. Conclusions

Let us compare the computed and measured data in more detail. For this purpose, we present table 2 by Engwall and Badin (1999) that contains the formants of long vowels in the Swedish language.

Table 2: Formants (in kHz) of a Swedish speaking subject in supine position (Engwall and Badin 1999)

IPA	F1	F2	F3	F4	IPA	F1	F2	F3	F4
uː	0.34	0.80	2.32	3.20	oː	0.40	0.74	2.40	3.14
ɑː	0.56	0.94	2.74	3.24	æː	0.76	1.34	2.44	3.60
ɛː	0.60	1.62	2.46	3.54	eː	0.34	2.10	2.60	3.52
iː	0.32	1.58	3.04	3.72	yː	0.30	1.54	2.84	3.50
ɯː	0.36	1.72	2.54	3.28	øː	0.50	1.06	2.48	3.24

The vowels from table 2, together with the scaled and computed $[\text{øː}]_{s,c}$ from table 1, are plotted in the (F2, F1)-plane in figure 5. Clearly, $[\text{øː}]_{s,c}$ is closer to measured [øː] than to any other measured vowel, *except* possibly [ɑː]. To further clarify the situation, let us consider the formants F1 to F4 for $[\text{øː}]_{s,c}$, [øː], and [ɑː] as vectors: $[\text{øː}]_{s,c} = (0.56, 1.11, 2.22, 3.10)$, $[\text{øː}] = (0.5, 1.06, 2.48, 3.24)$, and $[\text{ɑː}] = (0.56, 0.94, 2.74, 3.24)$. Then the Euclidean distance between $[\text{øː}]_{s,c}$ and [øː] is 0.31, but the distance between $[\text{øː}]_{s,c}$ and [ɑː] is significantly larger, equalling 0.57. This difference is explained by F3,

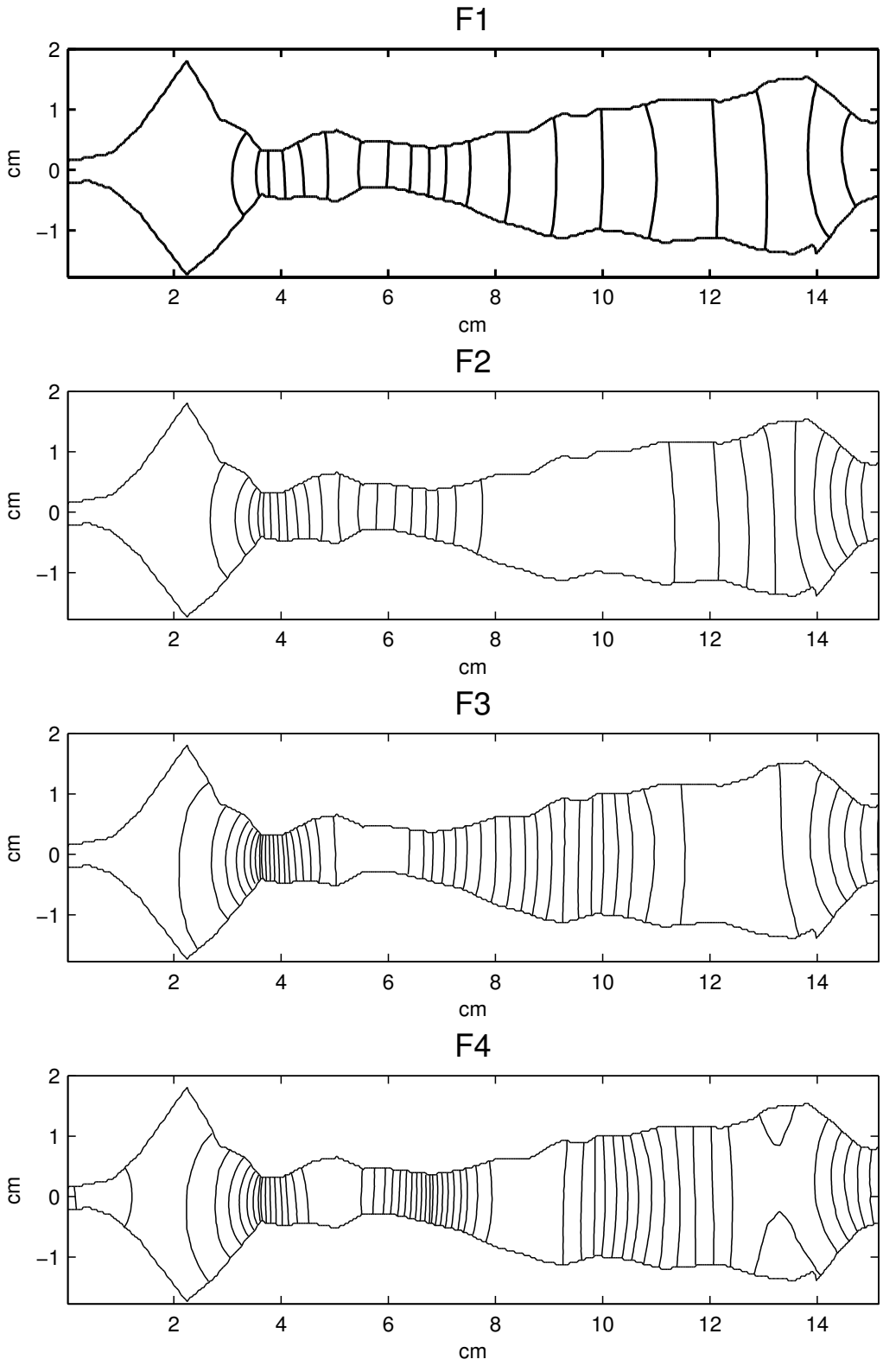

Figure 2: Isobars corresponding to F1-F4 along a mid-line cut

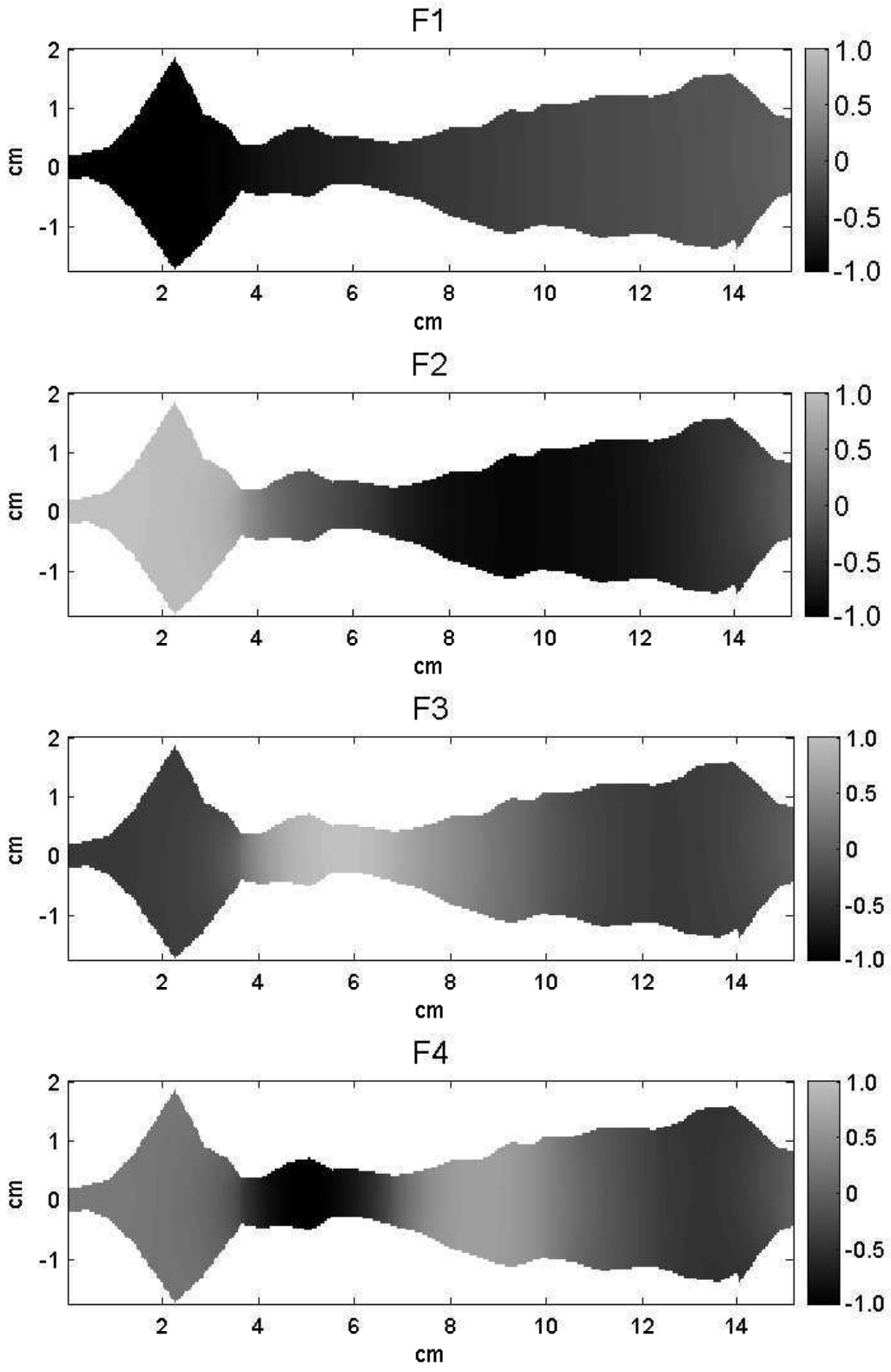

Figure 3: Pressure distributions for F1-F4 along a mid-line cut

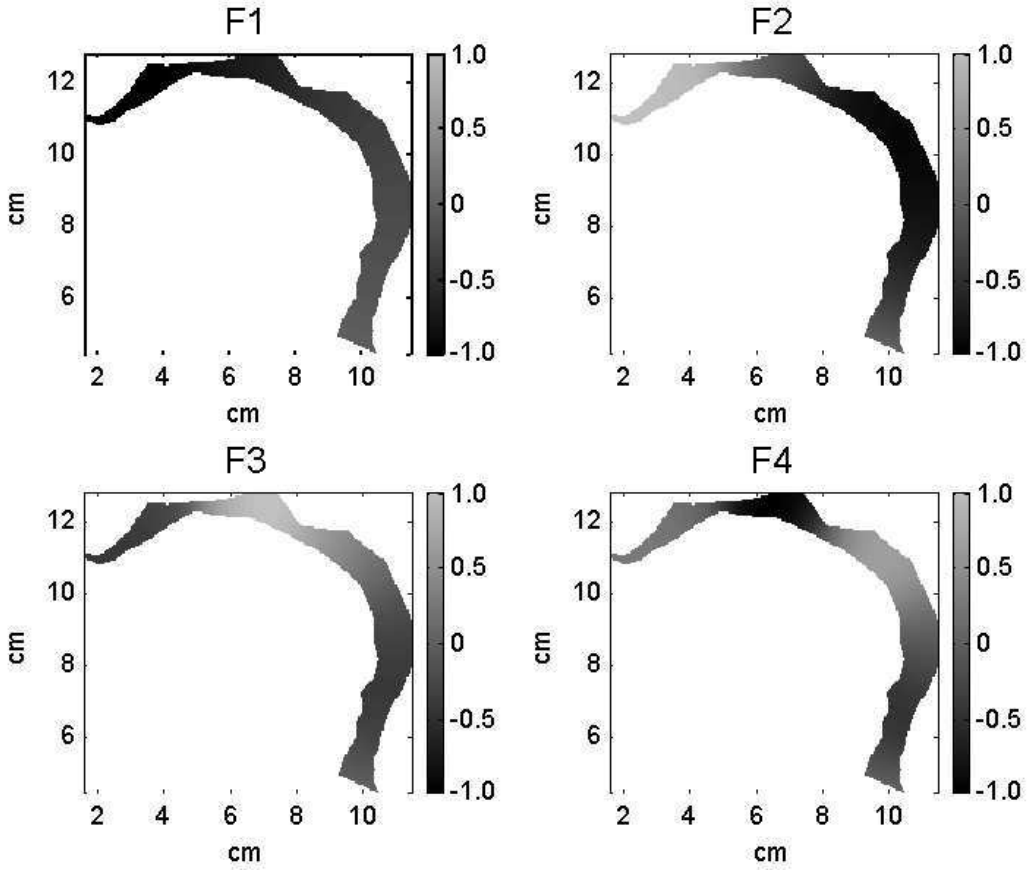

Figure 4: Pressure distributions for F1-F4 in the mid-sagittal plane

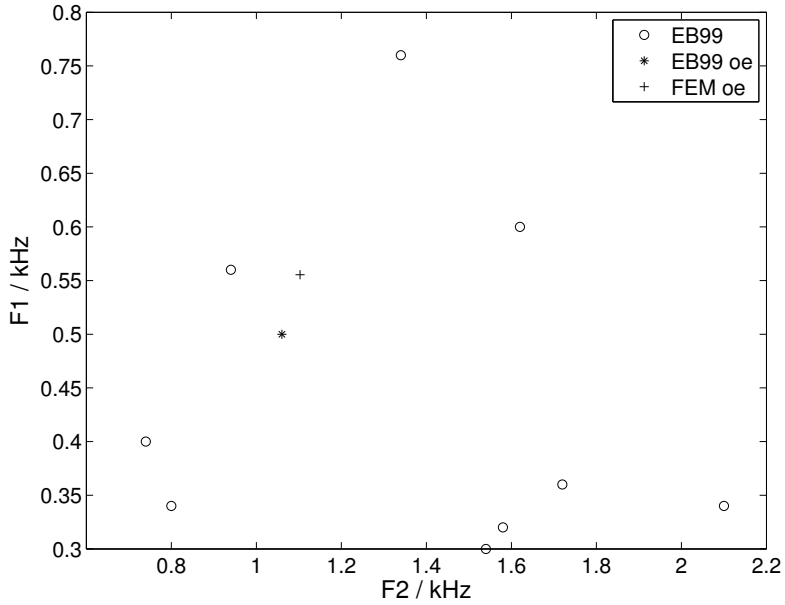

Figure 5: Vowels in the (F2,F1)-plane. *FEM oe* (+) is the scaled, computed [øː], *EB99 oe* (*) is the measured [øː] and *EB99* (o) are other measured vowels. (EB99 denotes Engwall and Badin (1999).)

since the fourth formants are almost the same. We conclude that the *first two* formants classify the scaled, computed vowel $[\text{ø:}]_{s,c}$ almost correctly. Moreover, if we look at *all four* available formants, even the remaining ambiguity disappears.

We remark that figures 2 and 3 supports the hypothesis that a weak cross-mode resonance related to F4 should appear in the oral cavity.

As we pointed out earlier, the computed formants F1 to F4 differ from the corresponding measured formants by $3\frac{1}{2}$ semitones. Having said that, the *ratios* between the computed formants and the measured formants match each other very well. There is a simple physical explanation why such a discrepancy is to be expected. In model (2), we use the Dirichlet boundary condition on the lip opening. This results in a vibrational node at the opening. In reality, such a node would appear further away outside the mouth since we are surely able to hear the sound outside of a speakers VT. In that sense, the real life VT is effectively longer than the one described by model (2), resulting in lower formants. To get rid of this phenomenon, we should also model the surrounding acoustic space.

Acknowledgements

We would like to thank Olov Engwall from KTH, Stockholm, for providing the articulation geometry for this study.

Antti Hannukainen has been supported by the Academy of Finland.

References

Dedouch, K., Horacek, J., Vampola, T., Svec, J., Krsek, P., and Havlik, R. (2002). Acoustic modal analysis of male vocal tract for Czech vowels. In *Proceedings Interaction and Feedbacks '2002*, pages pp. 13 – 19.

El Masri, S., Pelorson, X., Saguet, P., and Badin, P. (1996). Vocal tract acoustics using the transmission line matrix (TLM) method. In *Proceedings of the 4th International Conference on Spoken Language Processing*, pages pp. 953 – 956.

El Masri, S., Pelorson, X., Saguet, P., and Badin, P. (1998). Development of the transmission line matrix method in acoustics. applications to higher modes in the vocal tract and other complex ducts. *Int. J. of Numerical Modelling*, 11:133 – 151.

Engwall, O. and Badin, P. (1999). Collecting and analysing two- and three-dimensional MRI data for swedish. *TMH-QPSR*, (3-4/1999):pp. 11–38.

Fant, G. (1970). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Fetter, A. and Walecka, J. (1980). *Theoretical mechanics of particles and continua*. McGraw–Hill.

Foias, C. and Frazho, A. E. (1990). *The commutant lifting approach to interpolation problems*, volume 44 of *Operator Theory: Advances and applications*. Birkhäuser Verlag.

Johnson, C. (1987). *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press.

Kelly, J. and Lochbaum, C. (1962). Speech synthesis. In *Proceedings of the 4th International Congress on Acoustics*, pages Paper G42: 1–4.

Lu, C., Nakai, T., and Suzuki, H. (1993). Finite element simulation of sound transmission in vocal tract. *J. Acoust. Soc. Jpn. (E)*, 92:pp. 2577 – 2585.

Malinen, J. (2004). Conservativity of time-flow invertible and boundary control systems. Helsinki University of Technology, Institute of Mathematics Research Reports, A479.

Malinen, J. and Staffans, O. J. (2006). Conservative boundary control systems. *J. Diff. Eq.*, 231(1):pp. 290 – 312.

Malinen, J. and Staffans, O. J. (2007). Impedance passive and conservative boundary control systems. *Complex Analysis and Operator Theory*. to appear.

Malinen, J., Staffans, O. J., and Weiss, G. (2006). When is a linear system conservative? *Quart. Appl. Math.*, 64:pp. 31 – 91.

Mullen, J., Howard, D., and Murphy, D. (2006). Waveguide physical modeling of vocal tract acoustics: Flexible formant bandwith control from increased model dimensionality. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3):pp. 964 – 971.

Palo, P. (2006). Review of articulatory speech synthesis. Master's thesis, TKK.

Saad, Y. (1992). *Numerical methods for large eigenvalue problems*. Manchester University Press, Manchester.