How far are vowel formants from computed vocal tract resonances?

Aalto, Daniel

2012-08-29


http://hdl.handle.net/10138/39259

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
How far are vowel formants from computed vocal tract resonances?

D. Aalto$^{1,3}$, A. Huhtala$^2$, A. Kivelä$^2$, J. Malinen$^2$, P. Palo$^3$, J. Saunavaara$^4$, M. Vainio$^3$

1 Dept. of Signal Processing and Acoustics, Aalto University, Finland
2 Dept. of Mathematics and Systems Analysis, Aalto University, Finland
3 Inst. of Behavioural Sciences (SigMe group), University of Helsinki, Finland
4 Medical Imaging Centre of Southwest Finland

jarmo.malinen@aalto.fi

August 30, 2012

Abstract

We compare numerically computed resonances of the human vocal tract with formants that have been extracted from speech during vowel pronunciation. The geometry of the vocal tract has been obtained by MRI from a male subject, and the corresponding speech has been recorded simultaneously. The resonances are computed by solving the Helmholtz partial differential equation with the Finite Element Method (FEM).

Despite a rudimentary exterior space acoustics model, i.e., the Dirichlet boundary condition at the mouth opening, the computed resonance structure differs from the measured formant structure by $\approx 0.7$ semitones for [i] and [u] having small mouth opening area, and by $\approx 3$ semitones for vowels [a] and [ae:] that have a larger mouth opening. The contribution of the possibly open velar port has not been taken into consideration at all which adds the discrepancy for [a] in the present data set. We conclude that by improving the exterior space model and properly treating the velar port opening, it is possible to computationally attain four lowest vowel formants with an error less than a semitone. The corresponding wave equation model on MRI-produced vocal tract geometries is expected to have a comparable accuracy.

Keywords. Formant analysis, acoustic resonance computation, FEM, MRI.
1 Introduction

The purpose of this paper is to evaluate the accuracy of vowel simulations based on the wave equation model (1). We use 3D vocal tract (VT) geometries that have been obtained by Magnetic Resonance Imaging (MRI) from a native male speaker of Finnish while he pronounces prolonged vowels [a], [i], [u], and [œ]. The evaluation is carried out by comparing the computed resonances of (1) with the measured formants, extracted from sound samples, instead of comparing simulated and actual speech signals. In this work, the sound samples have been recorded simultaneously with the MRI, using the equipment and the arrangements detailed in [5, 6, 21, 27]. This is in contrast to, e.g., our earlier work [15] where only a single anatomic configuration (corresponding to Swedish [œ]) was taken from the data set of [13].

We use the same wave equation model for vowels as in [15], namely

\[
\begin{aligned}
\Phi_{tt} &= c^2 \Delta \Phi \text{ on } \Omega, \quad \Phi = 0 \text{ on } \Gamma_1, \\
\frac{\partial \Phi}{\partial \nu} &= 0 \text{ on } \Gamma_2, \quad \Phi_t + c \frac{\partial \Phi}{\partial \nu} = 2 \sqrt{\frac{c}{\rho_0}} u \text{ on } \Gamma_3,
\end{aligned}
\]

(1)

where \( \Omega \subset \mathbb{R}^3 \) is the interior of the VT whose boundary \( \partial \Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \) consists of the mouth opening \( \Gamma_1 \), the VT tissue walls \( \Gamma_2 \), and an (imaginary) control surface \( \Gamma_3 \) placed right above the glottis. The parameters \( c = 350 \text{ m/s} \) and \( \rho_0 = 1.225 \text{ kg/m}^3 \) are the speed of sound in and the density of dry air at \( T = 305 \text{ K} \), respectively. The functions in (1) are as follows: \( u = u(r, t) \) is the incoming acoustic power (per unit area) at glottis input, \( \frac{\partial \Phi}{\partial \nu} = \nu \cdot \nabla \Phi \), and \( \nu \) is the exterior unit normal on \( \partial \Omega \). In time domain simulations, we compute the velocity potential \( \Phi(r, t) \) for a given glottal input function \( u(r, t) \) produced by a source such as described in [2, 3]. From the velocity potential, the sound pressure and (perturbation) velocity can be extracted as the partial derivatives \( p' = \rho_0 \Phi_t \) and \( \mathbf{v} = \nabla \Phi \). The physics of model (1) is further explained in [15] and the references therein.

In the past, the VT acoustics has been modelled in many different ways. Electrical transmission lines have been used already in [10], and the classical Kelly–Lochbaum model in [17] makes use of reflection/transmission coefficients of a variable diameter tube. The 3D wave equation for linear wave propagation as well as the related Helmholtz equation for acoustic resonances have been known for a long time; see, e.g., [16]. The Kelly–Lochbaum model is closely related to Webster’s equation in, e.g., [2, 14] but the latter can be easily deduced from variable-impedance electrical transmission lines as well as from the wave equation in 3D tubular domains as shown in great generality in [19]. More advanced models are the transmission line networks that
have been applied for speech in, e.g., [11, 12, 22]; see also [4] for a purely mathematical treatment.

At their best, all of these modelling paradigms are known to produce very good simulated speech even though they are based on radically simplified representations of the underlying anatomic geometry $\Omega$ with the sole exception of the wave equation. In most applications, simplifications are even desirable as it may improve conceptual clarity and reduce computational burden. There are, however, situations where the faithful representation of the VT geometry is required, e.g., when modelling the effects of anatomical abnormalities and maxillofacial surgery on speech [8, 23, 25, 32, 33, 34].

2 Background and motivation

In our earlier work [15], the same numerical computations were carried out using a minimal data set: a single MRI-based anatomic geometry $\Omega$ corresponding to the Swedish vowel $[\phi]$. The computed resonances were compared to the first four formants of all Swedish vowels (including $[\phi]$) that were extracted from speech samples of the same test subject. The speech samples were not recorded simultaneously with the MRI because of technological restrictions but the subject was in a similar supine position during both MRI and speech recording; see [13].

We made the following observations in [15]:

1. The computed resonances $R_1...R_4$ corresponding the formants $F_1...F_4$ of $[\phi]$ are systematically $3\frac{1}{2}$ semitones too high compared to the measured values;

2. the formant ratios of the computed and measured data (i.e., $R_i/R_1$ and $F_i/F_1$ for $i = 2, 3, 4$) correspond to each other quite well; and

3. if the systematic error in $R_1...R_4$ of $[\phi]$ is compensated by linear scaling, then the scaled, computed data gets identified correctly as $[\phi]$ in the measured formant table from the same subject.

Two main potential sources were identified for the discrepancy: (i) the Dirichlet boundary condition on $\Gamma_1$ in (11) results in too short acoustic length of the computational VT; and (ii) the minimal data set used in [15] is insufficient to draw any conclusions on the error sources. The purpose of this work is to exclude the latter possibility (ii) by extending and improving the data set in an essential manner. We also aim at a deeper understanding of the sources of descrepancy to guide future model improvements and to understand the quality of simulation that one can reasonably expect to attain.
We remark that the formant computation of [15] was later validated by completely independent FEM resonance computations that were based on the generalized Webster’s model instead of (1); the computed resonances corresponding to $F_1...F_4$ were practically the same as reported in [1, Table 3.1 on p. 31]. As shown in [19], the generalized Webster’s model is a low-frequency approximation of the wave equation in a tubular domain. In the case of human VT, the approximation remains accurate for formants $F_1...F_3$ and even for $F_4$ at least in some vowel configurations where cross-mode resonances do not dominate; see [15, Fig. 1].

We comment on the interesting parallel work [32] at the end of the paper.

3 Model and methods

As explained in [15, p. 3], the resonances of Eq. (1) can be solved by finding the complex frequencies $\lambda$ such that the Helmholtz problem

\begin{equation}
\begin{cases}
\lambda^2 \Phi_\lambda = c^2 \Delta \Phi_\lambda & \text{on } \Omega, \\
\Phi_\lambda = 0 & \text{on } \Gamma_1, \\
\frac{\partial \Phi_\lambda}{\partial \nu} = 0 & \text{on } \Gamma_2, \\
\lambda \Phi_\lambda + c \frac{\partial \Phi_\lambda}{\partial \nu} = 0 & \text{on } \Gamma_3
\end{cases}
\end{equation}

is solvable for some nonzero eigenfunctions $\Phi_\lambda(r)$. It is known that all such eigenvalues $\lambda$ form an infinite sequence \( \{\lambda_j\}_{j \in \mathbb{Z} \setminus \{0\}} \) with $|\lambda_j| \to \infty$ as $|j| \to \infty$, $\text{Re } \lambda_j \leq 0$, and $\text{Im } \lambda_{-j} = -\text{Im } \lambda_j$. The lowest formants $F_1$, $F_2$, ..., correspond to the numbers $R_j = \text{Im } \lambda_j$ for $j = 1, 2, \ldots$ in the order of increasing imaginary parts.

As solving Eqs. (1) and (2) analytically is possible only in a radically simplified geometry [31], we solved the problem numerically by the Finite Element Method (FEM). This is the approach used in [18], [24], [9], [29], [35], and by many others. Eqs. (2) were solved in variational form as given in [15, Eq. (5)] using a custom implementation of FEM programmed in MATLAB. We used piecewise linear shape functions on tetrahedral meshes. The tetrahedral meshes were generated using TetGen [30] from a triangular surface mesh. As a result, we obtained the matrices $A$ and $B$ for a high-order eigenvalue problem $Ay(\lambda) = \lambda By(\lambda)$ as explained in [15, Eq. (6)]. The lowest eigenvalues $\lambda_j$, $j = 1, 2, 3, 4$ were then computed using the \textit{eigs} routine of MATLAB. It takes around 3 seconds on a workstation with an Intel Xeon X3450 processor to build the matrices and to solve the eigenvalue problem.

The imaginary parts of the computed $\lambda_j$ are given in Table 1 together with the number of elements used in each VT geometry. The computed values are good approximations of the eigenvalues defined in Eqs. (2) when the number of elements is high enough. It was observed with the anatomic
geometry of [u] that using four times as many elements does not change the numerical result, and thus the resonances given in Table 1 can be regarded as accurate in this respect.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
<th>$R_4$</th>
<th># of elem.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>720</td>
<td>1547</td>
<td>2721</td>
<td>4138</td>
<td>47514</td>
</tr>
<tr>
<td>[i]</td>
<td>246</td>
<td>2135</td>
<td>3592</td>
<td>4667</td>
<td>37335</td>
</tr>
<tr>
<td>[u]</td>
<td>324</td>
<td>659</td>
<td>2262</td>
<td>3091</td>
<td>50579</td>
</tr>
<tr>
<td>[œ]</td>
<td>562</td>
<td>1612</td>
<td>2519</td>
<td>3602</td>
<td>53087</td>
</tr>
</tbody>
</table>

Table 1: Resonances (in Hz) of the Helmholtz problem by FEM, and the number of the elements in each geometry.

4 Geometric data from MRI

The raw MRI data was collected from a native male Finnish speaker while he pronounced prolonged vowels [a], [i], [u], and [œ] in a supine position.

As a result of these pilot experiments in June 2010, a combined data set of 53 simultaneously recorded MRI data and sound samples was produced. Out of this data set, four samples were chosen based on a visual quality assessment of spectrogram data as well as the requirement that the $F_1-F_2$ vowel space should be covered in a satisfactory manner. The spectrograms of these samples can be found in [27, p.64, p.66, p.68, and p.79]. The first three of these samples were pronounced at the fundamental frequency $f_0 = 110$ Hz and the last one at $f_0 = 137.5$ Hz.

A Siemens Magnetom Avanto 1.5T scanner was used in these experiments. A 12-element Head Matrix Coil was combined with a 4-element Neck Matrix Coil in order to cover the anatomy of interest. 3D VIBE (Volumetric Interpolated Breath-hold Examination) [28] was found out to be the most suitable MRI sequence for rapid 3D acquisition. As the naming of the sequence suggests, it was originally developed for fast 3D imaging of the abdominal region where breath-hold during the scan is essential. Sequence parameters were optimized in order to minimize the acquisition time, and we were able to carry out MRI with 1.8 mm isotropic voxels in just 7.6 s.

The tissue/air interface from the MR data was extracted by combining sagittal DICOM sequence -images to form a 3D voxel image. A triangular surface mesh of the interface was then extracted using custom MATLAB code. The three boundary components $\Gamma_1$, $\Gamma_2$, and $\Gamma_3$ were identified manually in the triangle mesh so that the different boundary conditions could
be applied in right places. Since teeth are not visible in MRI (and hence, they are not part of the computational geometry of this work), some resulting artefacts had to be corrected manually. The velar port was open in the geometries of \([\alpha]\) and \([i]\), and the resulting hole in the surface model was manually closed. A shaded representation of a typical surface mesh is presented in Fig. 1.

The geometric error in the tissue/air interface is a fraction of the voxel-based resolution of the original MRI data: interpolating in 2D sections by the gray-scale values of pixels results in about 1.8 bits of additional information compared to the MRI pixel size, corresponding to the geometric error of \(\approx 0.5\) mm with the current voxel size of 1.8 mm; see [6]. We conclude that the resonances in Table 1 do not contain essential errors due to inaccuracies of surface geometries.

5 Sound recording and formant extraction

The interior of a MRI machine is a challenging environment for speech recording. We used the recording arrangement discussed in [21, 27] and the experimental arrangements described in [5, Section 2]; see also [6].

Let us briefly describe the recording arrangement. A two-channel sound collector samples the speech and noise signals in a dipole configuration. The sound collector is an acoustically passive, non-microphonic device which does not cause artifacts in the MR images. The sound signals are coupled to a RF-shielded microphone array by acoustic waveguides of length 3 m. There
are acoustic impedance terminations at the both ends of the waveguides to sufficiently reduce longitudinal resonances. The microphone signals are coupled to an amplifier that is situated outside the MRI room. This analogue electronics is used to optimally subtract the noise signal from the contaminated speech signal in real time, and the cleaned-up signal is fed back to test subject’s earphones. The same audio signal is digitized by a 24bit ADC, and the residual longitudinal resonances of the waveguides are compensated numerically in the post-processing stage.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
<th>( F_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>651 ± 7</td>
<td>1024 ± 35</td>
<td>2647 ± 117</td>
<td>3679 ± 36</td>
</tr>
<tr>
<td>( i )</td>
<td>247 ± 9</td>
<td>2183</td>
<td>3304 ± 46</td>
<td>4407 ± 251</td>
</tr>
<tr>
<td>( u )</td>
<td>306 ± 37</td>
<td>675 ± 39</td>
<td>2173 ± 13</td>
<td>3242 ± 139</td>
</tr>
<tr>
<td>( \phi )</td>
<td>483 ± 35</td>
<td>1249 ± 74</td>
<td>1994 ± 50</td>
<td>3188 ± 17</td>
</tr>
</tbody>
</table>

Table 2: Formants (in Hz) computed as means of those extracted from the beginnings and ends of the samples. The upper (lower) sign refers to the beginning (resp., the end) of the sample.

For this work, the formants \( F_1 \)...\( F_4 \) were first extracted separately from the beginnings and the ends of the sound samples where the acoustic MRI noise is absent. The extraction was done by LPC using MATLAB similarly to the approach explained in [27, Chapter 6]. However, the present values were obtained by applying LPC analysis to FFT power spectra of the signals with the algorithm detailed in [20]. The residual waveguide resonances were removed from the spectra before LPC analysis. The results were compared visually to the peaks of the smoothed spectra to detect crude errors. The final results in Table 2 are the averages of these two values, augmented with their half distances. The formant data for \( i \) is subject to following remarks: (i) the LPC found a peak at 855 ± 133 Hz but this was removed from the data set as an outlier; (ii) \( F_2 \) could not be extracted from the beginning sample by the LPC even though it can be found easily in spectral curves by visual inspection; (iii) the LPC finds very strong double peaks about 500 Hz apart at \( F_3 \), and the \( F_3 \)-values given in Table 2 are defined as their means.

The data of Table 2 can be found in [26] except \( \phi \) (which is from a different series where \( f_0 = 137.5 \) Hz) and \( F_3, F_4 \) of \( i \) (which required a higher order LPC run). These results without numerical compensation of the residual longitudinal resonances of the waveguides can be found in [27, Tables 6.2 and 6.3 on p.49–50]. We remark that a typical mean deviation in vowel formant frequencies (when measured in ideal conditions in an anechoic chamber rather than inside a MRI machine) is in the class of 0.5 semitones;
Table 3: Discrepancy (in semitones) between computed resonances and mean formant frequencies from Table 2. Positive number implies that the computed resonance is higher than the measured formant.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
<th>$D_4$</th>
<th>mean discr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>1.7</td>
<td>7.1</td>
<td>0.5</td>
<td>2.0</td>
<td>2.8</td>
</tr>
<tr>
<td>[i]</td>
<td>-0.1</td>
<td>-0.4</td>
<td>1.4</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>[u]</td>
<td>1.0</td>
<td>-0.4</td>
<td>0.7</td>
<td>-0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>[œ]</td>
<td>2.6</td>
<td>4.4</td>
<td>4.0</td>
<td>2.1</td>
<td>3.3</td>
</tr>
</tbody>
</table>

6 Results and conclusions

Just as in [15], our new data indicates that the computed resonances $R_i$ from (2) tend to be higher than the measured formants $F_i$. The discrepancy given in Table 3 is in semitone scale to make the comparison easy with the “$3\frac{1}{2}$ semitone rule” that was discovered in [15]. Recall that difference of frequencies $F$ and $R$ is $D = 12 \ln(R/F)/\ln(2)$ semitones.

The main sources of discrepancy in these computations and experiments are as follows: (i) less than perfect performance of the test subject in the MRI machine, (ii) sporadic problems in formant extraction by the LPC, and (iii) physically unrealistic boundary conditions in Eqs. (1)–(2) especially at the mouth opening.

The mean discrepancy in Table 3 is at its largest for vowels [a] and [œ] where the computed resonances are consistently higher than the measured formants. Also, the mouth opening is largest for these vowels in our data set, and the Dirichlet boundary condition on $\Gamma_1$ in (2) is expected to be the most significant error source. All this is in good agreement with the results and the conclusions of [15].

The particularly significant error in $F_2$ of [a] can be explained by the fact that the velar port of the subject was unexpectedly open in the MR image, and the anti-node of the standing pressure wave (corresponding to $F_2$) would be at the velar port if it was closed. The nasality of the pronunciation is clearly heard from the speech sample. In the computational geometry, however, the hole was closed manually which leads to the perfectly reflecting Neumann boundary condition for the velar opening. It is physically more realistic to use a similar boundary condition for the velar opening as on $\Gamma_3$ in (2).

It is worth noting that the computational model performs strikingly well
for [u]: The discrepancy is of the same order as the fluctuations in formant values in sustained vowel production [6]. The velar port is closed in this MRI geometry. Also, the mouth opening is very small which results in relatively smaller error due to the Dirichlet boundary condition.

We conclude that the error profile in Table 3 supports earlier observations in [15], and it can be qualitatively explained by considering the underlying physics. A more sophisticated exterior space model (compared to the Dirichlet boundary condition) is likely to remove most of the formant discrepancy in vowels where the mouth opening is large. Complications related to the open velar port should be treated by taking into account the nasal tract resonating structures. This can be done by including them in the computational geometry or by setting an improved boundary condition at the velar port opening.

The results of [32] support the view that computed and measured resonances of a plastic model VT do not differ significantly from each other. Rather than carrying out speech recordings in MRI, the authors produce 3D physical printouts from MRI geometries by fast prototyping techniques for Japanese vowels [a], [i], [u], [e], [o]. Separately imaged teeth geometries were manually aligned with the soft tissue geometries. The formant structure is measured from the plastic models in ideal conditions, and the same configuration is used for transfer function simulations by the Finite-Difference Time-Domain (FDTD) method.

It is observed that the formant frequencies of the computed and measured transfer functions differ from each other by less than 3.2% (i.e., 0.55 semitones) which is comparable to our results on vowels [i] and [u]. Such an indirect experimental arrangement excludes most of the sources of discrepancy considered in our work, and the results of [32] can therefore be regarded as a limit what is reasonable to expect in a computational modelling effort of true vowel utterance. The effect of anatomic details such as piriform fossae, epiglottic valleculae and inter-dental spaces were computed, and the contribution of the latter two was found to be almost negligible in [32].

7 Acknowledgements

The authors were supported by the Finnish Academy grant Lastu 135005, the Finnish Academy projects 128204 and 125940, European Union grant Simple4All, Aalto Starting Grant, and Åbo Akademi Institute of Mathematics.
References


