Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [a]

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alias, Oriol Guasch

GTM – Grup de Recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, 08022 Barcelona, Spain

{marc.freixes,marc.arnela,joanclaudi.socoro,francesc.alias,oriol.guasch}@salle.url.edu

Abstract

One-dimensional articulatory speech models have long been used to generate synthetic voice. These models assume plane wave propagation within the vocal tract, which holds for frequencies up to ~5 kHz. However, higher order modes also propagate beyond this limit, which may be relevant to produce a more natural voice. Such modes could be especially important for phonation types with significant high frequency energy (HFE) content. In this work, we study the influence of tense, modal and lax phonation on the synthesis of vowel [a] through 3D finite element modelling (FEM). The three phonation types are reproduced with an LF (Liljencrants-Fant) model controlled by the $R_d$ glottal shape parameter. The onset of the higher order modes essentially depends on the vocal tract geometry. Two of them are considered, a realistic vocal tract obtained from MRI and a simplified straight duct with varying circular cross-sections. Long-term average spectra are computed from the FEM synthesised [a] vowels, extracting the overall sound pressure level and the HFE level in the 8 kHz octave band. Results indicate that higher order modes may be perceptually relevant for the tense and modal voice qualities, but not for the lax phonation.

Index Terms: voice production, higher order modes, high frequency energy, glottal source modelling, LF model, numerical simulation, finite element method

1. Introduction

For many years, works on articulatory speech synthesis have considered a simplified one-dimensional (1D) representation of the vocal tract. This is built from the so-called vocal tract area functions, which describe the cross-sectional area variations along the vocal tract center midline (see e.g., [1]). Voice is then synthesised by simulating the propagation of acoustic waves within this 1D representation of the vocal tract (see e.g., [2, 3, 4]). However, 1D approaches assume plane wave propagation, so they can only correctly approximate the acoustics of the vocal tract in the frequency range below 4-5 kHz. Beyond this limit, not only planar modes get excited but also higher order propagation modes appear, which strongly change the high frequency energy (HFE) content of the spectrum [5, 6] compared to that from a 1D model. Although the high frequency range has not received much attention in the literature, some recent studies point out that the HFE may be relevant for voice quality, speech localisation, speaker recognition and intelligibility (see [7] and references therein).

On the other hand, three-dimensional (3D) models do not need to assume plane wave propagation, since they can directly deal with 3D vocal tract geometries to emulate the complex acoustic field generated during voice production [8, 9, 10]. However, higher order modes do not always appear even if a 3D acoustic model is used. As shown in [6], a straightened vocal tract based on circular cross-sections prevents the onset of such modes due to radial symmetry, in contrast to what occurs for realistic vocal tract geometries based on MRI data. Other vocal tract geometries simplifications were studied in that work, all of them showing large variations in the HFE while keeping a similar behaviour for low frequencies. One can then assert that the vocal tract shape is determinant for the HFE content of the generated sound. However, the vocal tract shape is not the only factor affecting the HFE. The type of phonation can also modify it, as shown for instance in [11] for sustained vowels with loud and soft phonation.

In this work we study the effect of tense, modal and lax phonation on the synthesis of vowel [a], paying special attention to the HFE content. These three phonation types are reproduced using an LF (Liljencrants-Fant) model [12]. Although this model cannot consider the interaction between the vocal tract and the vocal folds [13, 14], it has proved to be useful to explore the phonatory tense-lax continuum [15] by controlling the $R_d$ glottal shape parameter [16]. Regarding the vocal tract, we consider an MRI-based realistic geometry, and its simplified counterpart considering circular cross-sections in a straightened midline [6]. This allows us to somewhat “activate” and “deactivate” the higher order modes. Different versions of vowel [a] are generated by convolving the LF glottal source signals with the vocal tract impulses responses obtained using a 3D acoustic model based on the Finite Element Method (FEM) [17]. In order to analyse the relevance of the higher order propagation modes for the lax, modal and tense phonation, the long-term average spectra (LTAS) and the HFE levels of the synthesised vowels are computed and compared.

The paper is structured as follows. The methodology used to study the production of vowel [a] with the three phonation types and the two vocal tract geometries is explained in Section 2. Next, the obtained results are discussed in Section 3. Finally, conclusions and future work are presented in Section 4.

2. Methodology

Figure 1 depicts the process followed to synthesise six versions of vowel [a]. These were obtained by convolving three glottal source signals with the FEM impulse responses of two vocal tract geometries that produce this vowel. In particular, and as mentioned before, we used the realistic vocal tract and the simplified straightened simplification with circular cross-sections from [6] (see Section 2.1), and computed their impulse responses $h(t)$ using the FEM (see Section 2.2). The glottal source signals $u_{gl}(t)$ were generated by means of an $R_d$ controlled LF model. The values $R_d = 0.3$, 1 and 2.7 were selected from the $R_d$ range [0.3, 2.7] (see [16]) to reproduce a tense, a modal, and a lax phonation, respectively (see Section 2.3).
Synthesis of vowel [a] with a realistic vocal tract geometry (above) and its simplified counterpart of circular cross-sections in a straightened midline (below). Three phonation types are considered to reproduce a tense (dashed red line), a modal (solid black line) and a lax (dotted green line) voice production. The output pressure signal $p(t)$ is computed as the convolution of the glottal source $u_g(t)$ with the vocal tract impulse response $h(t)$ obtained from 3D FEM simulations.

For each vowel, the LTAS was computed as the Welch’s power spectral density estimate, with a 15 ms hamming window, 50% overlap and a 2048-point FFT. The overall energy levels and the HFE levels in the 8 kHz octave band were also extracted as in [11]. The 16 kHz octave band was not considered, since HFE changes in this frequency range were found almost perceptually irrelevant in [11].

2.1. Vocal Tract Geometries

Two vocal tract geometry simplifications of vowel [a] have been employed in this work, namely, the realistic configuration and the simplified straight vocal tract with circular shape (see Fig. 1). These geometries were obtained in [6] by simplifying the MRI-based vocal tract geometries in [18]. In a nutshell, the procedure consisted in the following. First, the subglottal tube, the face and the lips were removed from the original geometry (see [10] for a detailed analysis of the lips influence on simulations). Moreover, side branches such as the piriform fossae and valleculae were occluded (see e.g. [9, 19] for their acoustic effects). Cross-sections were next extracted as typically done to generate 1D area functions, but preserving their shapes and locations in the vocal tract midline. The realistic configuration was generated by linearly interpolating the resulting cross-sections. As shown in [6], this simplification provides very similar results to the original MRI-based vocal tract geometry without branches.

In the simplified straight vocal tract configuration, the cross-sectional shapes were modified to be that of a circle, preserving the same area. These circular cross-sections were located in a straightened version of the vocal tract midline and then linearly interpolated. The two configurations are hereafter referred as the realistic and the simplified vocal tracts.

2.2. Vocal Tract impulse response

The impulse response of each vocal tract geometry was computed using a custom finite element code that numerically solves the acoustic wave equation,

$$\partial^2_t p - c_0^2 \nabla^2 p = 0,$$

combined with a Perfectly Matched Layer (PML) to account for free-field propagation [17]. In Eq. (1) $p(x, t)$ is the acoustic pressure, $\partial^2_t$ stands for the second order time derivative, and $c_0$ is the speed of sound which is set to the usual value of 350 m/s. A Gaussian pulse was introduced on the glottal cross-sectional area as an input volume velocity $u_g(t)$. This pulse is of the type

$$u_g(t) = e^{-[(t - T_{hyp})/0.2T_{hyp}]^2} [\text{m}^3/\text{s}],$$

with $T_{hyp} = 0.646/f_c$ and $f_c = 10$ kHz. Wall losses were considered by imposing a boundary admittance coefficient of $\mu = 0.005$ on the vocal tract walls. A 20 ms simulation was then performed capturing the acoustic pressure $p_g(t)$ at a node located outside of the vocal tract, 4 cm away from the mouth aperture center. The sampling frequency was set to $f_s = 8000$ kHz, which ensures a restrictive stability condition of the Courant-Friedrich-Levy type required by explicit numerical schemes (see [17] for details on the numerical scheme).

A vocal tract transfer function $H(f)$ was computed from each simulation to compensate for the slight energy decay in frequency of the Gaussian pulse. This is defined as

$$H(f) = \frac{P_o(f)}{U_o(f)},$$

with $P_o(f)$ and $U_o(f)$ being the Fourier Transform of $p_g(t)$ and $u_g(t)$, respectively. $H(f)$ was computed up to 12 kHz, to allow the calculation of HFE level in the 8 kHz octave band [11]. The vocal tract transfer functions $H(f)$ for the realistic and the simplified geometries of vowel [a] are shown in Fig. 2 (also reported in [6], but only up to 10 kHz). As can be observed, planar modes are mainly produced below 5 kHz giving place to the first vowel formants. Beyond this value, higher order modes can also propagate, resulting in the more complex spectrum of...
the realistic geometry. Note, however, that these modes do not appear in the spectrum of the simplified configuration. The radial symmetry of this geometry prevents their onset [5, 6].

Finally, the inverse Discrete Fourier Transform was applied to the vocal tract transfer functions $H(f)$ to obtain the vocal tract impulse responses $h(t)$ of the two geometries (see Fig. 1).

### 2.3. Voice Source Signal

An LF model [12] was used to produce the voice source signal. This model approximates the glottal flow $u_g(t)$ and its time derivative $u'_g(t)$ in terms of four parameters ($T_p$, $T_e$, $T_s$, $E_e$) that describe its time-domain properties (see Fig. 3). The control of this model can be simplified with the single glottal shape parameter $R_d$ [16]. This is defined as

$$R_d = \frac{T_d}{T_0} = \frac{1}{10} = \frac{U_0}{E_e},$$

where $T_d$ is the declination time, $T_0$ the period, and $E_0$ the fundamental frequency. The declination time $T_d$ corresponds to the quotient between the glottal flow peak $U_0$ and the negative amplitude of the differentiated glottal flow $E_e$.

In this work, we used the Kawahara’s implementation of the LF model [20], which generates a free-aliasing excitation source signal. We adapted this model to our purposes, modifying the sampling frequency from its original value of 44100 Hz to 24 kHz. Moreover, we introduced the $R_d$ glottal shape parameter. This allows one to easily control the voice source with a single parameter, which runs from $R_d = 0.3$ for a very abducted phonation, to $R_d = 2.7$ for a very abducted phonation (see [16]). From the $R_d$ range [0.3, 2.7] two extreme values plus a middle one were chosen. We used $R_d = 0.3$ to generate a tense phonation, $R_d = 2.7$ for a lax production, and $R_d = 1$ for a normal (modal) voice quality. With regard to $F_0$, a pitch curve was obtained from a real sustained vowel lasting 4.4 seconds. This pitch contour was placed around 120 Hz to generate all the source signals. Figure 4a shows four periods of the three simulated voice source waveforms. Moreover, the LTAS of the glottal source signals are represented in Fig. 4b. As observed, the phonation type obviously changes the glottal pulse shape, thus modifying the spectral energy distribution of the source signal.

### 3. Results

Six versions of vowel [a] (see Fig. 1) have been generated using the three glottal source signals corresponding to a tense, a modal and a lax phonation, and the two impulse responses obtained from the 3D FEM simulations of the realistic and simplified vocal tract geometries. The six synthesised vowels are normalised with the same scaling factor to obtain reasonable sound pressure levels. This factor has been selected so as to produce 70 dB SPL in the realistic geometry with a modal phonation ($R_d = 1$). The LTAS have then been computed for each audio.

Figure 5 shows the obtained LTAS for the six generated vowels. As also appreciated in the vocal tract transfer functions (see Fig. 2), small differences between geometries are produced for frequencies below 5 kHz, whereas beyond this range higher order modes propagate in the realistic case, thus inducing larger deviations. This behaviour can be observed for the three phonation types. Essentially the glottal source modifies the overall energy level and also introduces an energy decay in frequency (compare Fig. 2 with Fig. 5). This decay, known as the spectral tilt, strongly depends on the phonation type. The laxer the phonation the larger the spectral tilt [16]. Furthermore, the source signal also affects the energy balance of the first harmonics (below ~ 500 Hz). For instance, the lax phonation has the lowest overall energy values among all phonation types. However, one can see that the first harmonic (close to 120 Hz) has larger amplitude levels than the rest of the spectrum, in contrast to what occurs for the other phonations.

HFE levels have been computed by integrating the power spectral density in the 8 kHz octave band, as in [11]. In addition, the overall energy levels have been calculated following the same procedure but for the whole examined frequency range.

The obtained results are listed in Table 1. Note first that in the realistic case with a modal phonation ($R_d = 1$) the overall level is 70 dB SPL. Remember that this value was fixed to
compute the scaling factor used to normalise the audio files. The overall level variations for the other configurations will thus correspond to modifications either introduced by the vocal tract geometry or by the glottal source. As expected, the larger the $R_d$ value (laxer phonation) the smaller the overall levels.

Far more interesting is to compare the results between geometries. The HFE levels decay between 5.6 dB and 5.9 dB for the realistic vocal tract depending on the phonation type, which only manifests as an overall level difference of 1.2 dB and 1.4 dB. The higher order modes tend to reduce the levels in the HFE content. According to [11], minimum difference limen scores of about 1 dB are given for normal-hearing listeners in the 8 kHz octave band, so one may hypothesise that the higher order modes may be perceptually relevant. However, depending on the phonation type the HFE could be too small to notice any difference. This seems to be the case of the lax phonation ($R_d = 2.7$), which gives HFE levels of 1.0 dB and 6.6 dB, depending on the geometry. We may then conjecture, that for this phonation type no differences in the outputs from the two geometries will be perceived. In other words, we would not notice the influence of higher order modes.

4. Conclusions

In this work we have studied the influence of tense, modal and lax phonation on the 3D finite element synthesis of vowel [a], considering a realistic and a simplified vocal tract geometry. The 3D simulations behave very similarly for both geometries below 5 kHz, but significant differences appear beyond this frequency because of the rising of higher order propagation modes. It is worth mentioning that these modes only appear when using the realistic vocal tract. They induce a reduction of the HFE levels at the 8 kHz octave band from 5.6 to 5.9 dB, depending on the phonation type. These differences may be perceptually relevant, according to previous works in the literature. Specifically, a realistic 3D vocal tract geometry would be required for an accurate synthesis of vowel [a] through 3D FEM, when trying to simulate a modal and a tense voice production. Conversely, when a lax phonation is considered, the influence of higher order propagation may be imperceptible, since the HFE levels are very small. Therefore, a simpler 1D simulation would suffice in this case.

Future work will consider other $R_d$ values and geometry simplifications as well as other vowels to complete the study. Finally, we also plan to include aspiration noise in the LF model to evaluate its impact on the HFE content of the numerical simulations.

5. Acknowledgements

The authors are grateful to Saeed Dabbaghchian for the design of the vocal tract geometry simplifications. This research has been supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. The fourth author acknowledges the support from the Obra Social “La Caixa” under grant ref. 2018-URL-IR1rQ-021.

6. References


