INVESTIGATIONS IN ARTICULATORY SYNTHESIS

Athanassios Katsamanis [1], *Pirros Tsiakoulis* [1], *Petros Maragos*[1], *Alexandros Potamianos*[2*]

[1] National Technical University of Athens, [2] Technical University of Crete
{nkatsam,ptsiak,maragos}@cs.ntua.gr, potam@telecom.tuc.gr

ABSTRACT

Modern articulatory speech synthesizers simulate the human speech production system in an increasingly accurate manner. In this direction, we relax the simplifying assumption of zero mean flow velocity during speech production and we investigate potential effects. Further, we introduce a reduced parameter set for our 3D articulatory model which simplifies its control and does not allow humanly infeasible articulations. Vowel-Fricative-Vowel synthesis experiments using our twofold augmented synthesizer are reported.

Keywords: articulatory synthesis, mean volume velocity, 3D model, reduced parameter set

1. Introduction

Replication of the human vocal apparatus as an artificial system will have to incorporate specifics of the speech production mechanism at both the physical and the physiological level. Our intention is to incorporate knowledge and constraints related to speech aeroacoustics by also exploiting 3D geometry in a tractable scheme. In this direction, our contribution is twofold.

Firstly, we investigate the effects of non-zero mean flow velocity in the vocal tract during speech production. Typically, the opposite is assumed. However, the existence of respiratory flow and typical flow measurements during speech production [2], especially for fricatives, motivate the relaxation of this assumption. Previous approaches for articulatory synthesis that take the full flow field into consideration have been reported in [6, 4], where the solution of the related nonlinear equations is proposed. To reduce computational complexity we suppose that mean flow in the tract is not affected by acoustic disturbances. Thus, we can decouple the equations that govern the acoustic disturbances from those that govern the mean flow field. This approach has been inspired from the analysis in [16], where direct numerical simulation was applied for the estimation of the mean field and an acoustic analogy to predict sound radiation and all these in two dimensions. We investigate how similar ideas would work in a simpler computationally tractable articulatory synthesis setup.

Secondly, at the physiological level, we propose a novel control mechanism of a 3D articulatory model. Currently, there has been increasing interest in 3D articulatory modeling [3, 1, 5] augmented on the basis of new articulatory data acquisition techniques such as MRI, EPG and EMA. Such models are usually complicated 3D structures that are difficult to control. We propose a parameter set related to the phonetic attributes characterizing each phoneme class and is capable of efficiently controlling a 3D articulatory model.

Our 3D articulatory model is used to estimate geometrical vocal tract properties which are exploited by our vocal tract simulation system to synthesize Vowel-Fricative-Vowel sequences taking non-zero mean flow in the tract into consideration.

2. Wave Propagation in the Vocal Tract

Typically, sound propagation is regarded as the propagation of small-amplitude perturbations of fluid properties around an ambient state of pressure, density and velocity, $(p_0, \rho_0, \mathbf{v_0})$. By linearizing the conservation of mass and momentum equations and the pressure-density relation for the fluid around this state, equations of linear acoustics may be derived [13]. Under certain assumptions [14, pp. 34–35], including that equilibrium velocity is zero, these equations reduce to the following for volume velocity U and pressure fluctuations p in the case of sound in a nonuniform vocal tract with vibrating walls:

$$-\frac{\partial U}{\partial x} = \frac{1}{c^2\rho_0}\frac{\partial (pA_0)}{\partial t} + \frac{\partial A}{\partial t} \quad (1)$$

$$-\frac{\partial p}{\partial x} = \rho_0 \frac{\partial}{\partial t}(\frac{U}{A_0}) + RU \quad (2)$$

where ρ_0 is the density of air and c is the speed of sound. For the instantaneous vocal tract area function A and its relation to the area function A_0 at equilibrium the analysis in [9] is adopted. The flow

*This research was co-financed partially by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%) under Grant ΠΕΝΕΔ-2003ΕΔ866, and partially by the European research project ASPI under Grant FP6-021324.

resistance estimate $R = 8\pi\mu/A^2$, where μ is the air viscosity coefficient, accounts for viscosity[9] .

We relax the assumption of zero equilibrium velocity and accept that there exists non-zero mean volume velocity U_0 in the tract for which we may assume that $\iint_{A(x)} p\nu_{0x} dA \simeq pU_0$, $\nu_{0x} \simeq U_0/A(x)$ and $\partial U_0/\partial x = 0$. The total volume velocity is $U_{tot} = U_0 + U$. In this case, it can be shown that the sound transmission equations become:

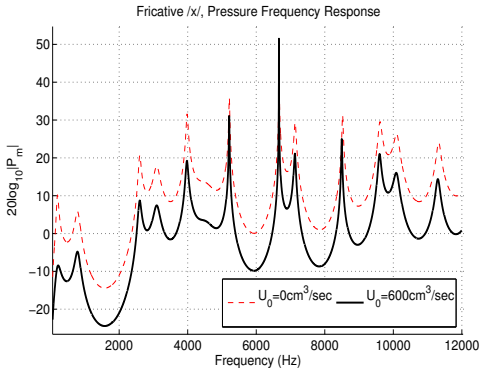
$$(3) \quad -\frac{\partial U}{\partial x} = \frac{1}{c^2\rho_0}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} + \frac{U_0}{\rho_0 c^2}\frac{\partial p}{\partial x}$$

$$(4) \quad -\frac{\partial p}{\partial x} = \rho_0 \frac{\partial}{\partial t}\left(\frac{U}{A}\right) + \rho_0 \frac{U_0}{A}\frac{\partial}{\partial x}\left(\frac{U}{A}\right) + \left[R + \rho_0 \frac{U_0}{A}\frac{\partial}{\partial x}\left(\frac{1}{A}\right)\right] U + \frac{1}{c^2}\left[\frac{\partial}{\partial t}\left(\frac{U_0}{A}\right) + \frac{U_0^2}{A}\frac{\partial}{\partial x}\left(\frac{1}{A}\right)\right] p$$

To investigate the importance of the additional terms, the system was initially simulated in the frequency domain. Its frequency response was estimated for various geometries and various steady mean flow velocities in the range $[0, 1000\ cm^3/sec]$, which are relevant for speech production [2]. For the simulation, the approach in [14, Ch.4] was followed. Indicative results are shown in Fig. 1 for a vocal tract configuration, i.e. area function, corresponding to fricative /x/. The spectrum at $U_0 = 600cm^3/sec$ has been downshifted by 10dB for better visualization. The most prominent effects of nonzero mean volume velocity can be observed at low frequencies where we have a considerable decrease of the first formant peak. For vowel configurations, the frequency response is only slightly affected. This could be justified by the relative importance of the space derivative $\partial/\partial x(1/A)$.

Time domain simulation was implemented basically using the numerical scheme described in [14]. Voiced excitation is achieved either by the Ishizaka-Flanagan two-mass model for the glottis or by defining a proper glottal area function as in [9]. The latter is mostly preferred as it is more easily controlled. The mean flow is considered time varying and for its estimation at each moment we apply a so-called low frequency model as in [10]. The mean volume velocity is given as a function of the subglottal pressure and the cross-sectional areas of the glottis and the maximum supraglottal constriction (minimum vocal tract area function). The Bernoulli law is used and viscous losses are also taken into consideration. The model is simple and computationally efficient. However, it probably does not accurately capture all the relevant flow characteristics. Alternatively, more sophisticated approaches could be followed [2].

Figure 1: Frequency response for non-zero mean volume velocities. The area function used corresponds to the fricative /x/. The spectrum for $U_0 = 600cm^3/sec$ has been downshifted by $10dB$ for better visualization. Effects at the lower frequencies may be observed, i.e. the relative magnitude of the first two peaks has changed.



3. 3D Articulatory Model Control

The used articulatory model is based on the 3D articulatory model in [3]. This model is actually an extension to the Mermelstein's model. The 2D configuration of the vocal tract is extended into a 3D grid by attaching lateral ribs to it. The 3D grid representing the vocal tract simplifies the extraction of area function and perimeter to be used in the acoustic wave propagation numerical simulation. In our implementation, 15 parameters control the model's configuration in the midsagittal plane and 8 extra parameters the lateral shape of the tongue. The resulting 3D articulatory model is a pure geometric model, that is not based on data such as MRI [5].

The articulatory parameter set, totaling 23 articulatory parameters, gives a good control mechanism for the three-dimensional geometry of the vocal tract, but at the same time complicates the control process, due to the set's large dimension. The actual speech motor control in humans is highly unlikely to use such a large set of articulatory commands. Instead it is believed that the the commands are rather related to acoustic parameters, such as formants [12]. This implies the existence of various correlations between the articulatory parameters. We devised a new set of parameters for our articulatory model's control process. The new set consists of direct coordinates in the phonetic space, resulting in a parameter set with minimal dimensionality. The parameters for the vowels are derived directly from the phonetic vowel space, whereas for

Table 1: $Height$ and $backness$ of vowels

	Front	**Central**	**Back**
Close	$(1, 0)$	$(1, 0.5)$	$(1, 1)$
Close-mid	$(0.7, 0.1)$	$(0.7, 0.55)$	$(0.7, 1)$
Open-mid	$(0.3, 0.3)$	$(0.3, 0.6)$	$(0.3, 1)$
Open	$(0, 0.5)$	$(0, 0.7)$	$(0, 1)$

the consonants, vowel context is taken into account.

Vowels The phonetic space for vowels, also called vocoid space [8], has roughly 3 dimensions. The dimensions of the vocoid space correspond to the common features *height*, *backness* and *roundedness*. The first two vowel features (height and backness) are related to the tongue, with height describing the vertical position of the tongue and with backness being the horizontal position relative to the pharyngeal wall. Roundedness determines whether the lips are rounded or not, and is also related to the lip protrusion; rounded lips are usually also protruded.

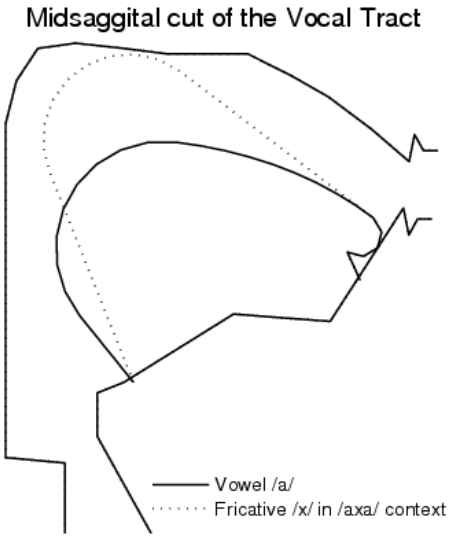
For our purposes, we quantify the above vowel features in order to be used for the control of the articulatory model. The usual quantification in phonetics is a discrete one, with various levels of discretization. We rather apply a continuous quantification with each feature described by a variable with a range of $[0, 1]$. A rough correspondence between the discrete values for height/backness used in the IPA vowel chart and our continuous height/backness variables is shown in Table 1. Each cell of the table is a pair in the form (height, backness) and corresponds to the values in our quantization scheme.

Roundedness is quantized in the same way resulting in a third variable; a value of 0 corresponds to no rounding and a value of 1 signifies maximum rounding and protrusion. The value of this variable is estimated automatically, following the case of most languages, where front and open vowels tend to be unrounded, whereas back and close ones tend to be rounded. In cases where this rule does not apply a value for the roundedness can be explicitly specified. Similar assumptions are made for the hyoid bone position. Finally other parameters may also be specified, such as the velum opening for nasalization, and the degree of jaw opening.

The resulting parameter set for the vowels has two basic parameters ($height$ and $backness$) and three optional parameters ($roundedness$, $nasality$ and $opening$). We have heuristically tuned a set of mapping functions between the phonetic parameters and the geometric articulatory parameters. An example for the vowel /a/ is shown in Fig. 2 (solid line), produced with parameters: $height = 0$, $backness = 0.5$, $roundedness = 0$ and $opening = 0.5$.

Figure 2: The midsagittal cut of the vocal tract for the vowel /a/ (solid line) and for the fricative /x/ in the context of /axa/ (dotted line)

.



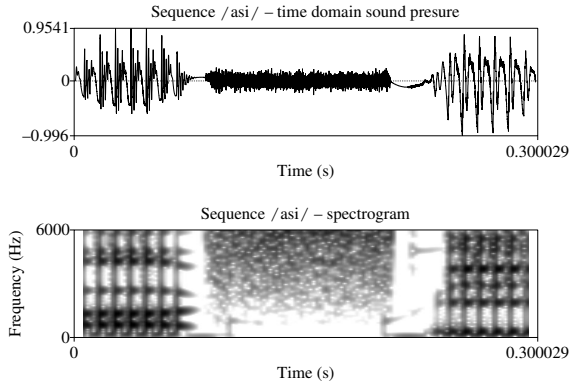
Consonants, unlike vowels, do not have a uniform way of production. Each consonant is identified by its place of articulation (place of constriction) and by the type of articulation(s), such as closure, nasalization, frication etc (this can be seen in full detail in the consonants chart of the IPA [7]). This proliferation of the way consonants are articulated makes the use of continuous variables, such as height and backness, impossible. Therefore we use the place and type of articulation, as our consonant parameter set. For each combination of place and type, if it corresponds to a real consonant, we determine which of the articulators are involved. We then tune the articulatory parameters to accomplish the desired articulation. The articulatory parameters that are not involved in this process are left undefined and their value is derived from the interpolation of the same parameters of the context vowels. If no such vowel is available the neutral articulation is considered. In Fig. 2 the fricative /x/ in /axa/ context is shown with the dotted line. We have assumed that only the tongue is involved.

4. VFV Synthesis Experiments

Considering that the investigated directions would be more relative for the synthesis of fricatives we have mainly worked on the synthesis of phoneme sequences of vowels and fricatives, V_1FV_2. The fricative in the middle could be one of /f/,/s/,/x/.

In the articulatory synthesis framework, natural sound generation mechanisms involved in the production of fricatives are commonly modeled phe-

Figure 3: Sequence /asi/ as generated from our synthesizer. Mean flow in the tract is considered non-zero and the area function and tract perimeter are given from our 3D articulatory model.



nomenologically. Noise pressure or volume velocity sources are simply introduced in the simulation and they excite the vocal tract resulting in the synthesis of sounds with the desired spectral content and pressure level. For this purpose, the properties of these sources, i.e. type, spectrum, position and power, should be properly determined. Many different approaches have been proposed in this direction [11, 10, 15]. Having experimented with various combinations, we currently apply two noise sources, a pressure source in series at the glottis to model aspiration noise and a volume velocity source in parallel at the point of maximum supraglottal constriction. Their attributes are determined as described in [15]. For the estimation of their power we used the mean volume velocity of the flow as given by the low-frequency model mentioned in Sec. 2.

To control the transitions between phonemes we basically adopt and extend the methodology proposed in [10]. For each phoneme three regions are identified, i.e. left transition, steady state, right transition. At the boundaries, we define control structures which contain the values of the articulatory model parameters, i.e. the phonetic parameters, the pitch and the glottal area properties at that moment. In between, the desired properties are obtained by linear or cosine interpolation. A synthesized sequence /asi/ is shown in Fig. 3.

5. Discussion

We have investigated an approach to take non-zero mean flow in the vocal tract into consideration for articulatory synthesis of VFVs, Fig. 3. The sound field may be decoupled and solved independently. At the first-order approximation that has been explored no dramatic effects of the non-zero mean flow on the radiated sound have been observed. This may be in part due to the simple mean flow model we have used or because a higher order approximation is necessary. Further analysis and investigation of this observation is in progress.

At the physiological level, we propose a new articulatory control mechanism based on a parameter set of minimal dimensionality. This set reduces the complexity of the model's control process. Though this dimension reduction results in a reduction of the model degrees of freedom, the model is still capable of producing all the phoneme articulations avoiding those that are humanly infeasible. Currenlty, we are working on properly extending our 3D articulatory model in order to allow prediction of 3D speech production phenomena, which appear to be relevant for the improvement of our mean flow model and in general our vocal tract simulation system.

6. REFERENCES

[1] Artisynth. http://www.artisynth.org/home.xml.

[2] P. Badin. Acoustics of voiceless fricatives: Production theory and data. *QPSR*, 3:033–055, 1989.

[3] P. Birkholz and D. Jackel. A three-dimensional model of the vocal tract for speech synthesis. In *Proc. 15th ICPhS*, pages 2597–2600, 2003.

[4] P. Birkholz and D. Jackel. Simulation of flow and acoustics in the vocal tract. In *Proc. CFA/DAGA*, pages 895–896, 2004.

[5] O. Engwall. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Sp. Comm.*, 41(2-3):303–329, October 2003.

[6] J. Huang, S. Levinson, D. Davis, and S. Slimon. Articulatory speech synthesis based upon fluid dynamic principles. In *Proc., ICASSP*, 2002.

[7] IPA. http://www.arts.gla.ac.uk/IPA/ipachart.html.

[8] J. Laver. *Principles of Phonetics*. CU Press, 1994.

[9] S. Maeda. A digital simulation method of the vocal-tract system. *Sp. Comm.*, pages 199–229, 1982.

[10] S. Maeda. Phonemes as concatenable units: VCV synthesis using a vocal tract synthesizer. In *Sound Patt. of Conn. Sp. Desc., Models and Explanation, Proc. of the Symp., Kiel Univ.*, 1996.

[11] S. Narayanan and A. Alwan. Noise source models for fricative consonants. *IEEE TSAP*, 8:328–344, 2000.

[12] J. Perkell, et al. Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Sp. Comm.*, 22:227–250, 1997.

[13] A. D. Pierce. *Acoustics*. ASA, 1989.

[14] M. Portnoff. A quasi-one dimensional digital simulation for the time-varying vocal tract. MIT, 1973.

[15] M. M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE TASSP*, 35(7):955–967, July 1987.

[16] W. Zhao, et al. Computational aeroacoustics of phonation, part I: Computational methods and sound generation mechanisms. *JASA*, 112(5):2134–2143, November 2002.