Simulation of Human Speech Production Applied to the Study and Synthesis of European Portuguese

A. Teixeira¹,²,*, R. Martinez¹, L. Silva¹, Luis M. T. Jesus¹,³, J. C. Príncipe⁴ and F. Vaz ¹,²

¹ Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), 3810-193 AVEIRO, Portugal
² Departamento de Electrónica e Telecomunicações, Universidade de Aveiro, 3810-193 AVEIRO, Portugal
³ Escola Superior de Saúde, Universidade de Aveiro, Universidade de Aveiro, 3810-193 AVEIRO, Portugal
⁴ Computational Neuroengineering Laboratory (CNEL), University of Florida, Gainesville, FL 32611, USA

* Author to whom correspondence should be addressed; electronic mail: ajst@det.ua.pt

Abstract

A new articulatory synthesizer (SAPWindows), with a modular and flexible design, is described. A comprehensive acoustic model and a new interactive glottal source were implemented. Perceptual tests and simulations made possible by the synthesizer contributed to deepen our knowledge of one of the most important characteristics of European Portuguese, the nasal vowels. First attempts at incorporating models of frication into the articulatory synthesizer are presented, demonstrating the potential of performing fricative synthesis based on broad articulatory configurations. Synthesis of nonsense words and Portuguese words with vowels and nasal consonants are also shown. Despite not capable of competing with mainstream concatenative speech synthesis, the antropomorphic approach to speech synthesis, known as articulatory synthesis, proved as a valuable tool for Phonetics research and teaching. This was particularly true for the European Portuguese nasal vowels.

Keywords: Articulatory Synthesis, Speech Production, European Portuguese, Nasal Vowels, Fricatives.

I. INTRODUCTION

Recent technological developments are characterized by increasing physical and psychological similarity to humans (as in bio-informatics). Speech being one of the distinct characteristics of humans, it is a natural candidate to imitation by machines. Also, information can be transmitted very fast, freeing hands and eyes for other tasks.

Various designs of machines that produce and understand human speech have been available for a long time [21], [29]. The use of voice in computer systems interfaces will be an added advantage, allowing, for example, the use of information systems for people with different disabilities and the access by telephone to new information services. However, our current knowledge of the production and perception of voice is still incomplete. The quality (or lack of it) of synthetic voice of the currently available systems is a clear indication of the necessity to improve this knowledge [29].

There are two types of motivations for research in the vast domain of voice production and perception [47]. The first one aims at the deep understanding of its diverse aspects and functions, the second is the design and development of better artificial systems. When artificial systems are closely related to the way humans do things these two motivations can be synergistically merged.

We have been developing an articulatory synthesizer, since 1995, which will hopefully produce high quality synthetic European Portuguese (EP) speech. We aim at a simultaneous improvement of our synthesis quality (technological motivation) and also to expand our knowledge of Portuguese production and perception.

In this paper, we present the various models of our synthesizer and results demonstrating its capabilities and flexibility.

II. ARTICULATORY SYNTHESIS

Articulatory synthesis generates speech through modeling of physical, anatomical and physiological characteristics of the organs involved in human voice production. This is a different approach when compared with other techniques, such as formant synthesis [16]. In the articulatory approach the system is modeled instead of the signal or its acoustics characteristics. Approaches based on the signal try to reproduce natural voice as faithfully as possible with few or no concern about how it is produced. In contrast, a model based on voice production uses laws to describe the sound propagation in the vocal tract and models mechanical and aeroacoustic phenomena to describe the oscillation of the vocal folds.
A. Basic Components of an Articulatory Synthesizer

To implement an articulatory synthesizer in a digital computer a mathematical model of the vocal system is needed. Synthesizers usually include two subsystems: an anatomic-physiological model of the structures involved in voice production and a model of the production and propagation of sound in these structures.

The first model transforms the articulators positions, like the jaw, tongue body and velum, into cross-sectional areas of the vocal tract. The second model consists of a set of equations that describe the acoustic properties of the vocal tract system, and it is divided into submodels to simulate different phenomena such as the creation of a source of periodic excitation (vocal fold oscillation), sound sources caused by the turbulent flow in the case of existence of constriction zones (area sufficiently reduced along the vocal tract), propagation of the sound above and below the vocal folds, radiation at the lips and/or nostrils.

The parameters for the models can be produced by different methods. They can be obtained directly from the voice signal by a process of inversion with optimization, be defined manually by the researcher, or be the output of a linguistic processing part of a TTS (text to speech) system.

B. Motivations

Articulatory synthesis has not received as much attention in recent years as it might have because is not yet an alternative to the actual systems of synthesis currently used in TTS systems. This is due to different factors: the difficulty to get information about the vocal tract and the vocal folds during the production of voice in humans; the measurement techniques generally provide information regarding static configurations while information concerning the dynamics of the articulators is incomplete; a full and reliable inversion process for obtaining the articulatory parameters from natural voice doesn’t exist yet; this technique involves complex calculations, raising problems of stability in the numerical resolution.

Despite these limitations, articulatory synthesis presents some important advantages: the parameters of the synthesizer are directly related with the human articulatory mechanisms, being very useful in studies of production and perception [28]; this method can produce high quality nasal consonants and nasal vowels [22]; source-tract interaction, essential for a natural sound, can be conveniently modeled when simulating the vocal folds and the tract as one system [17]; the parameters vary slowly in time so they can be used in efficient processes of codification; the parameters are easier to interpolate than LPC and formant synthesizers parameters [34]; small errors in the control signals do not produce sounds of low naturalness generally, because the interpolated values will always be physically possible.

According to Shadle and Damper [32] articulatory synthesis is clearly the best way to reproduce some attributes of speech we are interested in, such as to be able to sound like an extraordinary speaker (e.g. a singer, someone with disordered speech, or an alien with extra sinuses); to be able to change to another speaker type, or alter the voice quality of a given speaker, without having to go through as much effort as required for the first voice. Articulatory synthesizers have parameters that can be conceptualized, so that if a speech sample sounds wrong, intuition is useful in fixing it, always teach us something and provide opportunities to learn more as we work to produce a commercially usable system.

“Articulatory synthesis holds promise for overcoming some of the limitations and for sharpening our understanding of the production/perception link.” [48]. There is only partial knowledge about the dynamics of the speech signal, so continued research in this area is needed. The systematic study of the coarticulation effects is of special importance for the development of the Experimental Phonetic and sciences related with the processing of voice [18]. An articulatory synthesizer can be used as a versatile speaker and therefore contribute to such studies. Articulatory synthesizers can generate speech using carefully controlled conditions. This can be useful, for example, to test pitch tracking algorithms [25].

The articulatory synthesizer can be combined with a speech production evaluation tool to develop a system that can produce real-time audio-visual feedback to help people with specific articulatory disorders. For example, computer based speech therapy [35] of speakers with dysarthria tries to stabilize their production at syllable or word level, to improve the consistency of production. For severely hearing impaired persons the aim is to teach them new speech patterns and increase the intelligibility of their speech. For children with cleft lip and palate and velopharyngeal incompetence the aim is to eliminate misarticulated speech patterns so that most of these speakers can achieve highly intelligible normal speech patterns.
Also “the use of such a [articulatory] synthesizer has much to commend it in phonetic studies” (Franklin S. Cooper [7]). The audio-visual feedback could be used as an assistant for teaching phonetics to foreign students to improve their speech quality. The synthesizer can be used to help teach characteristic features of a given language such as pitch level and vowel space [49].

Recent developments presented at the ICPhS [48] show that articulatory synthesis is worth revisiting as a research tool and as part of TTS systems. Better ways of measuring vocal tract configurations, an increased research interest in the visual representation of speech and the use of simpler control structures, have renewed the interest in this research area [48]. Current articulatory approaches to synthesis include an open-source infrastructure that can be used to combine different models [9], recent developments in the Haskins Configurable Articulatory Synthesizer (CASY) [13], the characterization of lip movements [23], the ICP Virtual Talking Head that includes articulatory, aerodynamic and acoustic models of speech [3] and the quasiarticulatory (articulatory parameters controlling a formant synthesizer) approach of Stevens and Hanson [37].

III. SAPWindows Articulatory Synthesizer

Object oriented programming was used to implement the synthesizer. The Model-View-Controller concept was adopted to separate models from their controls and viewers.

The application, developed using Microsoft Visual C++, can synthesize speech segments from parameters sequences. These sequences can be defined in a data file or edited by the user. The synthesis process is presented step by step on a graphical interface.

Good quality synthesis of vowels (oral or nasal), nasal consonants and fricatives is possible with the present models. Next sections present briefly the currently implemented models.

A. Anatomic Models

For non-nasal sounds we only have to consider the vocal tract, i.e., a variable area tube between the glottis and the lips. For nasal sounds we have also to consider the nasal tract. The nasal tract area is essentially constant, with the exception of the soft palate region. The vocal tract varies continually and its form must be specified in intervals shorter than a few milliseconds [30].

1) Vocal Tract Model: The proposed anatomic model, shown in Fig. 1, assumes mid-sagittal plane symmetry to estimate the vocal tract cross-sectional area. Model articulators are tongue body, tongue tip, jaw, lips, velum and hyoid. Our model is an improved version of the University of Florida Mind-Machine Interaction Research Center (MMIRC) model [26], which in turn was a modified version of the Mermelstein’s model. It uses a non-regular grid to estimate section’s areas and lengths.

![Vocal tract model, based on Mermelstein’s model.](image)

2) Nasal Tract Model: The model of the nasal tract allows the inclusion of different nasal tract shapes and several paranasal sinuses.

The nasal cavity is modeled in a similar way to the oral tract and can be considered as a side branch of the vocal tract. The major difference is that the area function of the nasal tract is fixed for most part of the nasal tract and
for a particular speaker. The variable region, the soft palate, changes with the degree of nasal coupling. The velum parameter of the articulatory model controls this coupling. RLC shunt circuits, representing Helmholtz resonators, simulate the paranasal sinuses [22].

Our synthesizer allows the definition of different tract shapes and the inclusion of the needed sinus at any position by simply editing an ASCII file. Also, blocking of the nasal passages at any position can be simulated by defining a null area section at the point of occlusion. Implementation details were reported in [39].

In most of our studies, we use the nasal tract dimensions from [5], as shown in Fig. 2, which were based on studies by Dang and Honda [8] and Stevens [36].

![Fig. 2. Default nasal model, based on [5].](image)

### B. Interactive Glottal Source Model

We designed a glottal excitation model that included source-tract interaction, for oral and nasal sounds [41], that allowed direct control of source parameters, such as fundamental frequency, and that was not too demanding computationally.

The interactive source model we developed was based on [1]. The model was extended to include: a two-mass parametric model of the glottal area, jitter, shimmer, aspiration, and the ability to synthesize dynamic configurations.

To calculate the glottal excitation, \( u_g(t) \), it became necessary to model the subsystems involved: the lungs, the subglottal cavities, the glottis and the supraglottal tract.

The role of the lungs is the production of a quasi-constant pressure source, modeled as a pressure source \( p_l \) in series with the resistance \( R_l \). To represent the subglottal region, including the trachea, we used three RLC resonant circuits [2].

Several approaches were used for the vocal folds modeling: self-oscillating models, parametric glottal area models, etc. We wanted a physiological model, like the two-mass model, that resulted in high quality synthesis, but at the same time not too demanding computationally. Also, a direct control of parameters such as \( F_0 \) was required. According to these requirements, we chose the model proposed by Prado [26], a model that directly parameterizes the two glottal areas of the model. In the model, \( R_g \) and \( L_g \), which depend on glottal aperture, represent the vocal chords.

Systems above glottis were modeled by the tract input impedance \( z_{in}(t) \) obtained from the acoustic model. This approach results in an accurate modeling of frequency dependent losses.

The various subsystems can be represented by the equivalent circuit shown in Fig. 3.

![Fig. 3. Electrical analogue of the implemented glottal source (adapted from [33]).](image)

Pressure variation along the circuit can be represented by:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Typical value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_l$</td>
<td>Lungs pressure</td>
<td>10000</td>
<td>dyne/cm²</td>
</tr>
<tr>
<td>$F_0$</td>
<td>fundamental frequency</td>
<td>100 – 200</td>
<td>Hz</td>
</tr>
<tr>
<td>OQ</td>
<td>open quotient</td>
<td>60</td>
<td>% of $T_0$</td>
</tr>
<tr>
<td>SQ</td>
<td>speed quotient</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>$A_{g0}$</td>
<td>Minimum glottal area</td>
<td>0</td>
<td>cm²</td>
</tr>
<tr>
<td>$A_{gmax}$</td>
<td>Maximum glottal area</td>
<td>0.3</td>
<td>cm²</td>
</tr>
<tr>
<td>$A_1 - A_4$</td>
<td>Slope</td>
<td>0.03</td>
<td>cm²</td>
</tr>
<tr>
<td>Jitter</td>
<td>$F_0$ perturbation</td>
<td>2</td>
<td>%</td>
</tr>
<tr>
<td>Shimmer</td>
<td>$A_{gmax}$ perturbation</td>
<td>5</td>
<td>%</td>
</tr>
<tr>
<td>Asp</td>
<td>Aspiration</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE I**

GLOTTAL SOURCE TIME VARYING PARAMETERS.

\[
p_l - R_t u_g(t) - \sum_{i=1}^{3} p_{sgi} - \frac{d(L_g u_g(t))}{dt} - R_g u_g(t) - p_s(t) = 0
\]

The glottal source model includes parameters needed to model $F_0$ and glottal aperture perturbations, known as jitter and shimmer. The model also takes into account the aspiration noise generation as proposed by Sondhi and Schroeter [30]. Our source model is controlled by two kinds of parameters. The first type of parameters can vary in time, having a role similar to the tract parameters. In the synthesis process, these parameters can be used to control intonation, voice quality and related phenomena. They are presented in Table I. The second type of source parameters (including lung resistance, glottis dimensions, etc) do not vary in time. Their values can be altered by editing a configuration file.

**C. Acoustic Model**

Several techniques have been proposed for simulation of sound propagation in the oral and nasal tracts [27]: direct numeric solution of the equations; time-domain simulation using wave-digital-filters (WDF), also known as Kelly-Lochbaum model; frequency domain simulation. After analyzing the pros and cons of these three approaches, we chose for our first implementation of the acoustic model the frequency domain technique. The main reason for this choice was the possibility of easily including the frequency dependent losses.

In our acoustic model, we made the following approximations: propagation is assumed planar; the tract is straight; the tube is approximated by the concatenation of elementary acoustic tubes of constant area. An equivalent circuit, represented by a transmission matrix, represents each one of these elementary tubes. Analysis of the circuit is performed in the frequency domain [34].

Speech is generated by the acoustic model. We use a frequency domain analysis and time domain synthesis method - usually designated as the hybrid method [34]. The use of the convolution method avoids the problem of continuity of resonance in the faster method proposed by Lin [20]. The use of a fast implementation of the IFFT (the MIT FFTW [11]) minimizes the convolution calculation time.

A similar procedure is applied to the input impedance $Z_{in}(\omega)$, in order to obtain $z_{in}(n)$, needed for the source-tract interaction modeling by the glottal source model.

**D. Acoustic Model for Fricatives**

The volume velocity at a constriction is obtained by the convolution of the glottal flow with the impulse response calculated, using an IFFT, from the transfer function between the glottis and the constriction point $H_{gm}$ (see Fig. 4).

1) Noise sources: Fluctuations in the velocity of airflow emerging from a constriction (at an abrupt termination of a tube) create monopole sources and fluctuations of forces exerted by an obstacle (e.g. teeth, lips) or surface (e.g. palate) oriented normal to the flow generate dipole sources. Since dipole sources have been shown to be the most influential in the fricative spectra [24], the noise source of the fricatives has only been approximated by equivalent pressure voltage (dipole) sources in the transmission-line model. Nevertheless, it is also possible to
insert the appropriate monopole sources, which contribute to the low-frequency amplitude and can be modeled by an equivalent current volume velocity source.

Frication noise is generated at the vocal tract according to the suggestions of Flanagan [10], and Sondhi and Schroeter [34]. A noise source can be introduced automatically at any T-section of the vocal tract network, between the velum and the lips. The synthesizer’s articulatory module registers which vocal tract tube cross sectional areas are below a certain threshold ($A < 0.2 cm^2$), producing a list of tube sections that might be part of an oral constriction that generates turbulence.

The acoustic module calculates the Reynolds number at the sections selected by the articulatory module and activates noise sources at tube sections where the Reynolds number is above a critical value ($Re_{crit} = 2000$ according to [34]). Noise sources can also be inserted at any location in the vocal tract, based on additional information about the distribution and characteristics of sources [31], [24]. This is a different source placement strategy from that usually used in articulatory synthesis [34] where the sources are primarily located in the vicinity of the constriction. The distributed nature of some noise sources can be modeled by inserting several sources located in consecutive vocal tract sections. This will allow us to try combinations of the canonical source types (monopole, dipole and quadrupole).

A pressure source with amplitude proportional to the squared Reynolds number

$$P_{noise} = \begin{cases} 2 \times 10^{-6} \times \text{rand}(Re^2 - Re_{crit}^2) & Re > Re_{crit} \\ 0 & Re \leq Re_{crit} \end{cases}$$

is activated at the correct place in the tract [10], [34]. The internal resistance of the noise pressure source is proportional to the volume velocity at the constriction: $R_{noise} = \rho U_c / A_c$, where $\rho$ is the density of the air, $U_c$ is the flow at the constriction, and $A_c$ is the constriction cross-sectional area. The turbulent flow can be calculated by dividing the noise pressure by the source resistance. This noise flow could also be filtered in the time domain to shape the noise spectrum [24] and test various experimentally derived dipole spectra.

2) Propagation and radiation: The general problem associated with having $N$ noise sources is decomposed in $N$ simple problems by using the superposition principle. In order to calculate the radiated pressure at the lips due to each noise source, the vocal tract is divided into the following three sections: pharyngeal, region between velum coupling point and noise source and region after the source. Data structures based on the area function of each section are defined and ABCD matrices calculated [34]. The ABCD matrices were then used to calculate downstream ($Z_1$) and upstream ($Z_2$) input impedances, as well as the transfer function, $H$, given by:

$$H = Z_1 Z_2 \frac{1}{CZ_{rad} + D},$$

where $C$ and $D$ are parameters from the ABCD matrix (from noise source to lips), and $Z_{rad}$ is the lip radiation impedance. The radiated pressure at the lips due to a specific source is given by: $p_{radiated}(n) = h(n) * u_{noise}(n)$, where $h(n) = IFFT(H)$. The output sound pressure due to the different noise sources are added together. The output sound pressure resulting from the excitation of the vocal tract by a glottal source is also added when there is voicing.
IV. RESULTS

In this section, we present examples of simulation experiments performed with the synthesizer and two perceptual studies regarding European Portuguese nasal vowels. We start by the description of the perceptual tests; then, recent results in fricative synthesis; finally, examples of produced words and their respective quality, as evaluated in a quality test, are presented.

A. Nasal vowels studies

The synthesizer was used to produce stimuli for several perceptual tests, most of them for studies of nasal vowels. Next, we present two representative studies: the first, investigating the effect of velum, and other oral articulators variation over time; the second, addressing the source-tract interaction effects in nasal vowels.

Experiment 1 - Study of the influence of velum variation in the perception of nasal vowels on CVC contexts [42]

Several studies point to the need of regarding speech as a dynamic phenomenon. The influence of dynamic information in oral vowel perception is a subject of study for many years. In addition, some researchers also see nasal vowels as dynamic. To produce high quality synthetic nasal vowels, would be useful to know in what measure we need to include dynamic information.

We investigated if it is enough, to produce a good quality Portuguese nasal vowel, to couple the nasal tract or the degree of coupling variation in time improves quality. The null hypothesis is that static and dynamic velum will produce stimuli of similar quality.

Our first tests addressed the CVC context, nasal vowels between stops, the most common for nasal vowels in Portuguese.

![Fig. 5. Movement of the velum and oral articulators for a nasal vowel between two stop consonants (CVC context). The three phases of a nasal vowel in this context are shown.](image)

Velum and oral passage aperture variation for a nasal vowel produced between stop consonants is represented schematically in FIG. 5. During the first stop consonant velum and oral passage are closed. The beginning of the nasal vowel coincides with the release of the oral occlusion. To produce the nasal vowel both the oral passage and the velum must be open. Possibly due to the slow speed of velum movements, in European Portuguese, there is a period of time where oral passage is open and velum is in a closed, or almost closed position, producing a sound with oral vowel characteristics, represented in Fig. 5 by a V, for oral Vowel. Velum continues its opening movement creating simultaneous sound propagation in oral and nasal tracts. This zone is represented by Vn, standing for nasal vowel. The oral passage must close for the following stop consonant, and by closing before the velum, creates a zone with only nasal radiation, represented by N. Place of articulation of this nasal consonant, created by coarticulation, is the same of the following stop.

**Stimuli:** For this experiment, 3 variants of each of the 5 EP nasal vowels were produced differing in the way velum movement was modeled. For the first variant, called static, the velum was open at a fixed value during all vowel production. The other two variants used time varying velum opening. In the first 100 ms the velum stayed closed, making an opening transition in 60 ms to the maximum aperture, and then remaining open. In one of these variants a final bilabial stop consonant, [m], was created at the end by lip closure at 250 ms. All stimuli had a fixed duration of 300 ms.

**Listeners:** A total of 11, 9 male and 2 female, European Portuguese native speakers participated in the test. They had no history of speech, hearing or language impairments.
Procedure: Being interested in quality, we decide for a paired comparison test [15, p. 361], despite the demand for more decisions by each listener, which also increase test duration. The question answered by listeners was: Which of the two stimuli do you prefer as a European Portuguese nasal vowel? In preparing the test, we noticed that listeners had, in some cases, difficulty in choosing the preferred stimulus. The causes were traced to either good or poor quality of both stimuli. To handle this situation, we added two new possibilities, for a total of four possible answers: First, Second, Both, and None.

Test was divided in two parts. In the first part we compared static versus velum dynamic stimuli. In the second part comparison was made between dynamic stimuli with and without a final bilabial nasal consonant. Stimuli were presented 5 times in both AB and BA order. Inter Stimuli Interval was 600 ms.

The results for each possible pair of stimuli in the test were checked for listener consistency. They were retained if the listener preferred one stimulus in more than 70% of the presentations. Only clear choices of one stimulus against others were analyzed.

Results

Variable velum preferred to static velum: Preference scores (percentage of the designated stimuli chosen as the preferred one) for fixed velum aperture, variable velum aperture and the difference between the two are presented in the boxplot of Fig. 6.

Clearly, listeners preferred stimuli with time variable velum aperture. Average preference, including all vowels and listeners, is as high as 71.8%. Confidence interval (Clp=0.95) for the difference in preference score is between 24.2 and 65.6%, in favour of the variable velum case.

Repeated measures ANOVA showed as significant velum variation effect \[ F(1,10) = 5.67, p < 0.05 \] and as non-significant \( (p > 0.05) \) vowel and interaction between the two main factors (vowel and velum variation).

Nasal consonant at nasal vowel end is preferred: In general, listener preferred stimuli ending in a nasal consonant. Looking at the preference scores represented graphically in Fig. 7, stimuli with final nasal consonant are preferred more than stimuli without the final consonant. Confidence interval (Clp=0.95) for the difference in preference score is between 36.1 and 87.1%, in favour of the stimuli with a final nasal consonant.
ANOVA results, with two main factors, confirm as significant the effect of the final nasal consonant \( F(1, 8) = 9.5, p < 0.05 \) and nonsignificant \( p > 0.05 \) vowel effect and interaction between main factors.

**Experiment 2 - Study of source-tract interaction for nasal vowels [41]**

We investigated if the extra coupling of the nasal tract in nasal vowels produced identifiable alterations in the glottal source due to source-tract interaction, and if modeling of such effects results in a more natural quality synthetic speech.

Fig. 8 depicts the effect of the 3 different input impedances in nasal vowel [i]. The nasal tract load has a great influence on the glottal source wave. This influence comes from the noticeable difference in the input impedance calculated with or without the nasal tract input impedance, Fig 9. This difference is due to the fact that for high vowels such as [i] the impedance load for the pharyngeal region, which is equal to the parallel of the oral cavity and nasal tract input impedances, is almost equal to the nasal input impedance (see Fig. 10). The effect is less notorious in a low vowel, such as [ë].

**Stimuli:** Stimuli were produced for the EP nasal vowels varying only one factor: the input impedance of the tract used by the interactive source model. This factor had 3 values: (1) input impedance including the effect of all supraglottal cavities; (2) input impedance calculated without taking into account the nasal tract coupling; or (3) no tract load. Only 3 vowels, [ë], [i] and [u], were considered to reduce test realization time.

Same timing was used for all vowels. In the first 100 ms velum stays closed, making an opening transition in 60 ms to the maximum value. The velum remains at this maximum to the end of the vowel. Stimuli end with a nasal consonant, a bilabial ([m]), produced by closing the lips. Closing movement of the lips starts at 200 ms ending 50 ms later. Stimulus duration was fixed at 300 ms for all vowels. These choices were based on the results of the Experiment 1, where dynamic velum stimuli were preferred regarding naturalness.

Regarding the source, the interactive source model was used with variable \( F_0 \). \( F_0 \) starts around 100 Hz raises 20 Hz in the first 100 ms and then gradually goes to 100 Hz. The open quotient was 60 % and the speed quotient 2. Jitter and shimmer were added to improve naturalness.
**Fig. 10.** Input impedances in the velum region for nasal vowel [ɨ]. Figure presents the oral input impedance ($Z_{\text{in\ oral}}$), the nasal tract input impedance ($Z_{\text{in\ nasal}}$) and the equivalent parallel impedance (parallel). The parallel impedance is, for this vowel, approximately equal to the nasal tract input impedance.

### Table II

**RESULTS OF 4IAX DISCRIMINATION TEST**

<table>
<thead>
<tr>
<th>Listener</th>
<th>Sex</th>
<th>[ɨ]</th>
<th>[ɨ]</th>
<th>[ɨ]</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>50.0</td>
<td>33.3</td>
<td>41.7</td>
<td>41.7</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>58.3</td>
<td>100.0</td>
<td>50.0</td>
<td>69.4</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>50.0</td>
<td>41.7</td>
<td>50.0</td>
<td>47.2</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>33.3</td>
<td>83.3</td>
<td>66.7</td>
<td>63.9</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>16.7</td>
<td>58.3</td>
<td>33.3</td>
<td>36.1</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>66.7</td>
<td>66.7</td>
<td>66.7</td>
<td>66.7</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>50.0</td>
<td>50.0</td>
<td>41.7</td>
<td>47.2</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>58.3</td>
<td>58.3</td>
<td>41.7</td>
<td>52.8</td>
</tr>
<tr>
<td>9</td>
<td>F</td>
<td>41.7</td>
<td>50.0</td>
<td>66.7</td>
<td>52.9</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>58.3</td>
<td>50.0</td>
<td>58.3</td>
<td>55.6</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>33.3</td>
<td>83.3</td>
<td>58.3</td>
<td>58.3</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>75.0</td>
<td>58.3</td>
<td>58.3</td>
<td>63.9</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>50.0</td>
<td>41.7</td>
<td>33.3</td>
<td>41.7</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>83.3</td>
<td>50.0</td>
<td>58.3</td>
<td>63.9</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>51.8</td>
<td>58.9</td>
<td>52.4</td>
<td>54.4</td>
</tr>
<tr>
<td><strong>std</strong></td>
<td></td>
<td>17.4</td>
<td>18.6</td>
<td>12.8</td>
<td>16.4</td>
</tr>
</tbody>
</table>

**Listeners:** A total of 14, 11 males and 3 females, European Portuguese native speakers participated in the test. They had no history of speech, hearing or language impairments.

**Procedure:** A 4IAX (four-interval forced choice) discrimination test was performed to investigate if listeners were able to perceive changes in the glottal excitation caused by the additional coupling of the nasal tract.

The 4IAX test was chosen, instead of the more commonly used ABX test, due to reported better discrimination results with this type of perceptual test [4, p. 213].

In the 4IAX paradigm listeners hear two pairs of stimuli, with a small interval in between. The members of one pair are the same (AA); the members of the other pair are different (AB). Listeners have to decide which of the two pairs is made of different stimuli.

Signals were presented over headphones in rooms with low ambient noise. Each of the 4 combinations (ABAA, AABB, AAAB, BBAB) was presented 3 times in a random order. With this arrangement each pair to be tested appears 12 times. The order was different for each listener. Inter Stimuli Interval was 400 ms and Inter Pairs Interval was 700 ms.

**Results**

Table II presents the percentage of correct answers for the 4IAX test. The table presents results for each listener and vowel. Also, the statistics (mean and standard deviation) for each vowel, and for the 3 vowels, are presented at the bottom of the table. Results are condensed, in graphical form, in Fig. 11.

From the table and the boxplot it is clear that listeners correct answers are close to 50 %, being a little higher for the nasal vowel [ɨ]. These results indicate that stimuli differences are of difficult perception by the listeners.
Statistical tests, having as null hypothesis $H_0 : \mu = 50$ and alternative $H_1 : \mu > 50$, were only significant, at a 5% level of significance, for [i]. For this vowel the 95% confidence interval for the mean is between 50.1 a 67.7. For [ɛ] we obtained $p = 0.36$ and for [u] $p = 0.29$. For the 3 vowels considered together, average is also not significantly superior to 50% ($p = 0.08$).

Discussion

Simulations showed some, small, effects of the nasal tract load in the glottal wave time and frequency properties. Results of perceptual tests, conducted to study to what extent these alterations are perceived by listeners, support the idea that these changes are hardly perceptible. These results are in concordance with results reported in [46]. In their work Titze and Story reported that “An open nasal port ... showed no measurable effect on oscillation threshold pressure or glottal flow”.

There is however a tendency for the effect of interaction being more perceptible for the high vowel [i], produced with reduced vocal cavity. Our simulations results suggest as an explanation for this difference the relation between the nasal tract input impedance and the impedance of the vocal cavity, at the nasal tract coupling point.

B. Fricatives

In a first experiment the synthesizer was used to produced sustained unvoiced fricatives [44]. The vocal tract configuration derived from a natural high vowel was adjusted by raising the tongue tip in order to produce a sequence of reduced vocal tract cross-sectional areas. The lung pressure was linearly increased and decreased at the beginning and end of the utterance, to produce a gradual onset and offset of the glottal flow.

The second goal was to synthesize fricatives in VCV sequences [44]. Articulatory configurations for vowels were obtained by inversion [38]. The fricative segment was obtained by manual adjustment of articulatory parameters. For example, to define a palato-alveolar fricative configuration for the fricative in [iʃi] we used the configuration of vowel [i] and only changed the tongue tip articulator to a raised position ensuring a cross-sectional area small enough to activate noise sources.

For [iʃi], besides raising the tongue tip, described for [iʃi], we used lip opening to create the necessary small area passage at the lips. Synthesis results for the nonsense word /iʃi/ is shown in Fig. 12.

An $f_0$ value of 100 Hz and a maximum glottal opening of 0.3 cm$^2$ were used to synthesize the vowels. The time trajectory of the glottal source parameter $A_{\text{max}}$ rises to 2 cm$^2$ at the fricative middle point and at the end of the fricative returns value used during vowel production.

Nonsense words with voiced fricatives were also produced, keeping the glottal folds vibration throughout the fricative. Results for the [ivi] sequence are presented in Fig. 13.

C. Words

The synthesizer is also capable of producing words containing vowels (oral or nasal), nasal consonants, and (lower quality) stops.

To produce such words, and since the synthesizer is not connected to the linguistic and prosodic components of a text-to-speech system, we used the following manual process:
1) Obtaining duration for each phonetic segment entering the word composition (presently by direct analysis of natural speech although an automatic process, such as a CART tree, can be used in the future);

2) Obtain oral articulators configurations for each of the phones. For vowels we used configurations obtained by an inversion process based on the natural vowels first formants [38], [40]. These configurations were already available from previous work [42], [38]. For the consonants, for which we don’t have, yet, an inversion process, configurations were obtained manually, based on the Articulatory Phonetics description and published X-ray and MRI images;

3) Velum trajectory definition, using adequate values for each vowel and consonant;

4) Setting glottal source parameters, in particular the fundamental frequency (\(F_0\)).

We first attempted to synthesize words containing nasal sounds due to their relevance in Portuguese language [19]. We now present three examples of synthetic words: mão, mãe and António.

**Example 1 - word mão (hand):** First, from natural speech analysis, we obtained durations of 100 ms for the [m] and 465 ms for the nasal diphthong.

In this case the [m] configuration was obtained manually and configurations for [a] and [u] were obtained by an inversion process [38], [6]. The three configurations are presented in Fig. 14.

A velum trajectory was defined, based on articulatory descriptions of the intervening sounds. As shown in Fig. 15, the velum starts closed, in a pre-production position, opens for the nasal consonant, opens more during the first vowel in the diphthong, and finally raises towards closure in the second part of the diphthong.
Fundamental frequency, $F_0$, and other source parameters were also defined. $F_0$ starts at 120 Hz, increases 10 Hz at the end of the nasal consonant, and later to 150 Hz to stress the initial part of the diphthong, decreasing to 80 Hz at the end. This variation in time was based, partially, in the $F_0$ contour of natural speech. Values of 60 % for the open quotient (OQ) and 2 for speed quotient (SQ) were used. Jitter, shimmer and source-tract interaction were also used.

Two versions were produced: with and without lip closure at the end of the word. Due to the open state of the velum, this final oral closure results in the final nasal consonant [m]. The spectrogram of this last version is presented in Fig. 16.

**Example 2 - word mãe (mother):** A possible phonetic transcription for the word mãe (mother) is [’mẽ̞ĩ], including a palatal nasal consonant at the end [19, p. 292]. Keeping the oral passage open at the end of the word produced a variant. Due to the lack of precise information regarding oral tract configuration during production of [ẽ], we produced variants differing in the configuration used for the nasal vowel [ẽ]. One version was produced using the configuration of oral vowel [a], other, with a higher tongue position, using the configuration of vowel [e]. Another parameter varied was $F_0$: versions with values obtained by analysis of a natural speech, and versions with synthetic $F_0$. For the synthetic case, a further variation was used: the inclusion or not of source-tract interaction. Figure 17 shows the speech signal and respective spectrogram for non-natural $F_0$, source-tract interaction, configuration of [a] for nasal vowel [ẽ], and final palatal occlusion.

**Example 3 - word António (Anthony):** The first name of the first author, António [’t’ỗ’nju], was also synthesized using the same process as in the two previous examples. This word has a nasal vowel at the beginning, a stop, an oral vowel, a nasal consonant, and a final oral diphthong. Two versions were produced: one with natural $F_0$, other with synthetic $F_0$. The signal and respective spectrogram obtained for the first version are presented in Fig. 18. The stop consonant [t] was obtained closing and opening the oral passage without modeling important phenomena for the perception of a natural quality stop such as the voice onset time (VOT) and the frication noise at the release of closure.

As part of a Mean Opinion Score (MOS) quality test these, and many other stimuli produced by our synthesizer, were evaluated. To document the quality level achieved by our models, Table III presents the ratings of the various versions of the 3 examples presented above. Looking at the values after normalizing the rating of natural speech samples to 5, quality is between 3 and 4, fair to good. Top rated word obtained 3.7 (3.4 without normalization).
Fig. 17. Spectrogram of one of the versions of the word *mãe* produced, using an [a] configuration at the beginning of the nasal diphthong, oral occlusion at the end, source-tract interaction, and synthetic values for $F_0$.

Fig. 18. Speech signal and spectrogram for the synthetic word *António* [ět̪onju] produced using $F_0$ extracted from a natural pronunciation.

V. CONCLUSIONS

From the experience with simulations and perceptual tests using stimuli generated by our articulatory synthesizer, we believe that articulatory synthesis is a powerful approach to speech synthesis because of its anthropomorphic origin and by allowing us to address questions regarding human speech production and perception.

We developed a modular articulatory synthesizer architecture for Portuguese, using object oriented programming. Separation of control, model and viewer allows the addition of new models without major changes to the user interface. Implemented models include a glottal interactive source model, a flexible nasal tract area model, and a hybrid acoustic model capable of dealing with asymmetric nasal tract configurations and frication noise sources. Synthesized speech has a quality ranging from reasonable to good.

The synthesizer has been used, mainly, in the production of stimuli for Perceptual tests regarding Portuguese nasal vowels (for example [45], [42], [43]). The two studies on nasal vowels reported in this paper were only made possible by the use of the articularatory approach to speech synthesis, allowing the creation of stimuli by direct and precise control of the articulators and the glottal source. They illustrate the potential of articulatory synthesis in production and perception studies and the flexibility of our synthesizer.

Perceptual tests and simulations contributed to improve our knowledge regarding EP nasal sounds, namely:

1) It is necessary to include the time variation of velum aperture, combined with the time variation of articulators controlling the oral passage, in order to synthesize high quality nasal vowels;
2) Nasality is not controlled solely by the velum movement. Oral passage reduction, or occlusion, can be also used to improve nasal vowel quality. When situated at the end of a word, the lips or tongue movement, even without occlusion, contributes to a better nasal vowel by increasing the predominance of nasal radiation. Oral occlusion, due to coarticulation, before stops also contributes to nasal quality improvement;
3) Source-tract interaction effect of extra coupling of the nasal tract is not easily perceived. Discrimination was significantly above chance level only for the high vowel [i], what can possibly be explained by the relation of nasal and oral input impedances at the nasal tract coupling point.
A nasal vowel, at least in European Portuguese, is not a sound obtained only by lowering the velum. The way this aperture, and other articulators, vary in time is important. Including how the velum and the oral articulators vary in the various contexts improves quality.

With the addition of noise source models and modifications to the acoustic model, our articulatory synthesizer is capable of producing sustained fricatives and fricatives in VCV sequences. First results were presented, and removing noise sources along the oral tract is a feature we regard as having great potential.

The possibility of automatically inserting and removing noise sources along the oral tract is a feature we regard as having great potential.

Table III includes the Mean Opinion Score (MOS), its respective 95% confidence interval, and the normalized value resulting from scaling natural speech scores to 5.

<table>
<thead>
<tr>
<th>Word</th>
<th>$F_0$</th>
<th>int.</th>
<th>observ.</th>
<th>MOS</th>
<th>CI 95%</th>
<th>norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>mão</td>
<td>synthetic</td>
<td>yes</td>
<td>no [m] at end</td>
<td>3.4</td>
<td>[3.0..3.7]</td>
<td>3.7</td>
</tr>
<tr>
<td>mae</td>
<td>natural</td>
<td>yes</td>
<td>[a], [n] at end</td>
<td>2.9</td>
<td>[2.6..3.3]</td>
<td>3.2</td>
</tr>
<tr>
<td>synthetic</td>
<td>yes</td>
<td>[a], [n] at end</td>
<td>3.1</td>
<td>[2.7..3.4]</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>synthetic</td>
<td>yes</td>
<td>[a], no [n]</td>
<td>3.0</td>
<td>[2.6..3.4]</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>synthetic</td>
<td>no</td>
<td>[a], [n] at end</td>
<td>2.9</td>
<td>[2.5..3.3]</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>synthetic</td>
<td>no</td>
<td>[a], [n] at end</td>
<td>2.8</td>
<td>[2.5..3.2]</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>António</td>
<td>natural</td>
<td>yes</td>
<td>—</td>
<td>3.0</td>
<td>[2.8..3.2]</td>
<td>3.3</td>
</tr>
<tr>
<td>synthetic</td>
<td>yes</td>
<td>—</td>
<td>2.7</td>
<td>[2.4..2.9]</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>natural speech</td>
<td>4.6</td>
<td>—</td>
<td>—</td>
<td>5.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. ACKNOWLEDGMENTS

This work was made possible by the first author Ph.D. scholarship BD/3495/94 and project “Articulatory Synthesis of Portuguese” P/PLP/1122/1998, both from Portuguese research foundation (FCT) PRAXIS XXI program. We also have to thank University of Florida MMIRC, headed by Professor D. G. Childers, where this work started. The authors would like to thank the two anonymous reviewers for their comments and suggestions.

REFERENCES