EURASIP Journal on Applied Signal Processing 2005:9, 1435–1448

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

António J. S. Teixeira
Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), 3810-193 Aveiro, Portugal
Departamento de Electrónica e Telecomunicações, Universidade de Aveiro, 3810-193 Aveiro, Portugal
Email: ajst@det.ua.pt

Roberto Martinez
Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), 3810-193 Aveiro, Portugal
Email: martinezrs@ieeta.pt

Luís Nuno Silva
Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), 3810-193 Aveiro, Portugal
Email: lnors@ieeta.pt

Luis M. T. Jesus
Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), 3810-193 Aveiro, Portugal
Escola Superior de Saúde, Universidade de Aveiro, 3810-193 Aveiro, Portugal
Email: lmtj@essua.ua.pt

Jose C. Príncipe
Computational Neuroengineering Laboratory (CNEL), University of Florida, Gainesville, FL 32611, USA
Email: principe@cnel.ufl.edu

Francisco A. C. Vaz
Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), 3810-193 Aveiro, Portugal
Departamento de Electrónica e Telecomunicações, Universidade de Aveiro, 3810-193 Aveiro, Portugal
Email: fvaz@det.ua.pt

Received 29 October 2003; Revised 31 August 2004

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 deepening 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 is also shown. Despite not being capable of competing with mainstream concatenative speech synthesis, the anthropomorphic approach to speech synthesis, known as articulatory synthesis, proved to be a valuable tool for phonetics research and teaching. This was particularly true for the European Portuguese nasal vowels.

Keywords and phrases: articulatory synthesis, speech production, European Portuguese, nasal vowels, fricatives.

1. INTRODUCTION

Recent technological developments are characterized by increasing physical and psychological similarity to humans. One example is the well-known human-like robots. Being one of the distinct characteristics of humans, speech is a natural candidate to imitation by machines. Also, information can be transmitted very fast and speech frees hands and eyes for other tasks.

Various designs of machines that produce and understand human speech have been available for a long time [1, 2]. 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 [2].

There are two types of motivations for research in the vast domain of voice production and perception [3]. The first one aims at the deep understanding of its diverse aspects and functions, the second is the design and development of artificial systems. When artificial systems are closely related to the way humans do things, these two motivations can be merged. These systems contribute to an increased knowledge of the process and this knowledge can be used to improve current systems.

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.

2. ARTICULATORY SYNTHESIS

Articulatory synthesis generates the speech signal 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 [5]. In the articulatory approach, the system is modeled instead of the signal or its acoustics characteristics. Approaches based on the signal try to reproduce the signal of a natural voice as faithfully as possible with few or no concern about how it is produced. In contrast, a model based on the production system uses physical laws to describe the sound propagation in the vocal tract and models mechanical and aeroacoustic phenomena to describe the oscillation of the vocal folds.

2.1. 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 positions of the articulators, 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. Generally 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, and 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.

2.2. Motivations

Articulatory synthesis has not received as much attention in recent years as it could have because there is not 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 does not 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 of voice [6]; this method can produce high-quality nasal consonants and nasal vowels [7]; source-tract interaction, essential for a natural sound, can be conveniently modeled when simulating the vocal folds and the tract as one system [8]; 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 [9]; small errors in the control signals do not generally produce low quality speech sounds, because the interpolated values will always be physically possible.

According to Shadle and Damper [10], 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 teaching us something and providing 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" [11]. 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 phonetics and sciences related with the processing of voice [12]. 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 [13].

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 [14] 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" [15]. 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 [16].

Recent developments presented at the ICPhS [11] show that articulatory synthesis is worth revisiting as a research tool and as a 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 [11]. Current articulatory approaches to synthesis include an open-source infrastructure that can be used to combine different models [17], recent developments in the Haskins configurable articulatory synthesizer (CASY) [18], the characterization of lip movements [19], the ICP virtual talking head that includes articulatory, aerodynamic, and acoustic models of speech [20], and the quasiarticulatory (articulatory parameters controlling a formant synthesizer) approach of Stevens and Hanson [21].

3. 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.

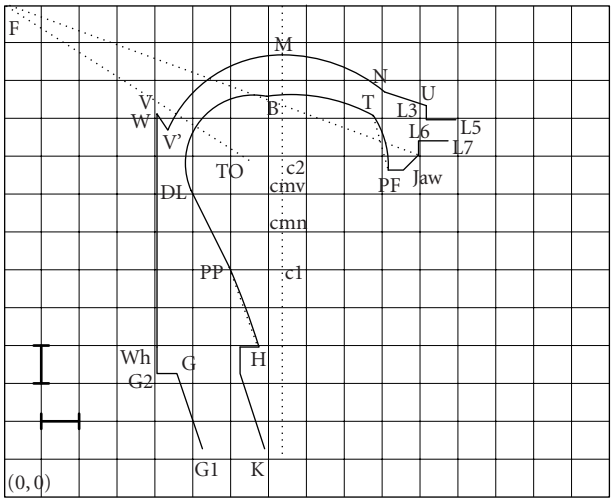
The application, developed using Microsoft Visual C++, can synthesize speech segments from parameters sequences. These sequences can be defined in a data file or edited by the user. The synthesis process is presented step by step on a graphical interface.

Presently, implemented models allow only quality synthesis of vowels (oral or nasal), nasal consonants, and fricatives.

The next sections present briefly the currently implemented models.

3.1. Anatomic models

For nonnasal sounds, we only have to consider the vocal tract, that is, 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 [23].



Jaw	: 20 deg. (0.349 rads)
Tongue tip	: (9.800; 10.040)
Tongue body	: (6.640; 8.780)
Lips opening	: 0.139082
Lips prot.	: 0.390000
Hyoid	: 0.060000
Velum position	: (4.376; 9.653)

Figure 1: Vocal tract model, based on Mermelstein's model [22].

3.1.1. Vocal tract model

The proposed anatomic model, shown in Figure 1, assumes midsagittal 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 MMIRC model [24], which in turn was a modified version of the Mermelstein's model [22]. It uses a nonregular grid to estimate section's areas and lengths.

3.1.2. 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 the most part of the nasal tract, 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 [7].

Our synthesizer allows the definition of different tract shapes and the inclusion of the needed sinus at any position

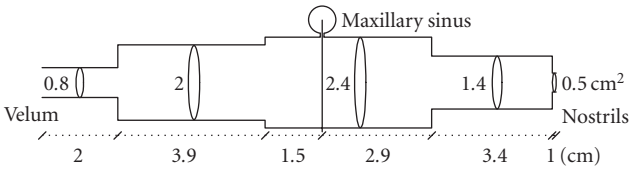


Figure 2: Default nasal model based on [26].

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 [25].

In most of our studies, we use the nasal tract dimensions from [26], as shown in Figure 2, which were based on studies by Dang and Honda [27] and Stevens [28].

3.2. Interactive glottal source model

We designed a glottal excitation model that included source-tract interaction, for oral and nasal sounds [29], 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 [30]. 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 quasiconstant 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 [31].

Several approaches were used for vocal fold modeling: self-oscillating models, parametric glottal area models, and so forth. We wanted to have a physiological model, like the two-mass model, that resulted in high-quality synthesis, but at the same time a model not too demanding computationally. Also, a direct control of parameters such as F_0 was required. We therefore chose the model proposed by Prado [24], which directly parameterizes the two glottal areas. In the model, R_g and L_g, which depend on glottal aperture, represent the vocal folds.

Systems above glottis were modeled by the tract input impedance $z_{\text{in}}(t)$ 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 Figure 3.

Pressure variation along the circuit can be represented by

$$p_l - R_l u_g(t) - \sum_{i=1}^{3} p_{sgi} - \frac{d(L_g u_g(t))}{dt} - R_g u_g(t) - p_s(t) = 0. \tag{1}$$

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 [23]. 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 1. The second type of source parameters (including lung resistance, glottis dimensions, etc.) does not vary in time. Their values can be altered by editing a configuration file.

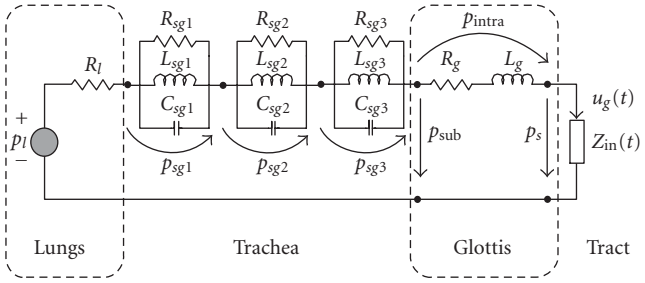


Figure 3: Electrical analogue of the implemented glottal source. Adapted from [32].

Table 1: Glottal source time-varying parameters.

Parameter	Description	Typical value	Unit
p_l	Lungs pressure	10000	dyne/cm^2
F_0	Fundamental frequency	100–200	Hz
OQ	Open quotient	60	% of T_0
SQ	Speed quotient	2	—
A_{g0}	Minimum glottal area	0	cm^2
$A_{g\max}$	Maximum glottal area	0.3	cm^2
$A_2 - A_1$	Slope	0.03	cm^2
Jitter	F_0 perturbation	2	%
Shimmer	$A_{g\max}$ perturbation	5	%
Asp	Aspiration	—	—

3.3. Acoustic model

Several techniques have been proposed for simulation of sound propagation in the oral and nasal tracts [33]: 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, models each one of these elementary tubes. Analysis of the circuit is performed in the frequency domain [9].

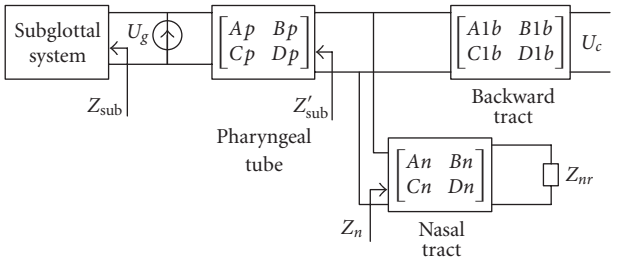


Figure 4: Matrices and impedances involved in the calculation of the transfer function H_{gn}, between glottis and a constriction point, which in turn is used in the calculation of flux at the noise source location.

Speech is generated by the acoustic model. We use a frequency-domain analysis and time-domain synthesis method—usually designated as the hybrid method [9]. The use of the convolution method avoids the problem of continuity of resonance in the faster method proposed by Lin [34]. The use of a fast implementation of the IFFT (the MIT FFTW [35]) 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.

3.4. 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_{gn} (see Figure 4).

3.4.1. 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 [36], 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 [37], and Sondhi and Schroeter [9]. 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 ($A < 0.2\,\text{cm}^2$), producing a list of tube sections that might be part of an oral constriction that generates turbulence.

The acoustic module calculates the Reynolds number (Re) at the sections selected by the articulatory module and activates noise sources at tube sections where the Reynolds number is above a critical value ($\text{Re}_{\text{crit}} = 2000$ according to [9]). Noise sources can also be inserted at any location in the vocal tract, based on additional information about the distribution and characteristics of sources [36, 38]. This is a different source placement strategy from that usually used in articulatory synthesis [9] 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

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

is activated at the correct place in the tract [9, 37]. The internal resistance of the noise pressure source is proportional to the volume velocity at the constriction: $R_{\text{noise}} = \rho|\bar{U}_c|/2A_c^2$, where ρ 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 [36] and test various experimentally derived dipole spectra.

3.4.2. 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 [9]. 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

$$H = \frac{Z_1}{Z_1 + Z_2} \frac{1}{CZ_{\text{rad}} + D}, \tag{3}$$

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 $p_{\text{radiated}}(n) = h(n) * u_{\text{noise}}(n)$, where $h(n) = \text{IFFT}(H)$. The output sound pressures 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.

4. RESULTS

In this section, we present examples of simulation experiments performed with the synthesizer and two perceptual studies regarding European Portuguese nasal vowels.

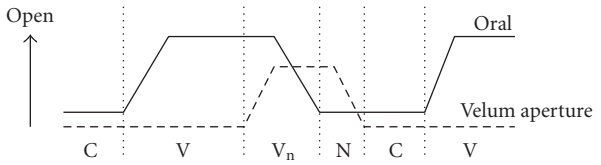


Figure 5: Movement of the velum and oral articulators for a nasal vowel between two stop consonants (CṼC context). The three phases of a nasal vowel in this context are shown.

We start by the description of the perceptual tests; then, recent results in fricative synthesis; finally, examples of produced words and quality tests are presented.

4.1. 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 CṼC contexts [39].

Several studies point to the need of regarding speech as a dynamic phenomenon. The influence of dynamic information in oral vowel perception has been 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 CṼC context, nasal vowels between stops, the most common for nasal vowels in Portuguese.

Velum and oral passage aperture variation for a nasal vowel produced between stop consonants is represented schematically in Figure 5. During the first stop consonant, the nasal and oral passages 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 Figure 5 by a V. Velum continues its opening movement creating simultaneous sound propagation in oral and nasal tracts. This zone is represented by V_n. The oral passage must close for the following stop consonant, so the early oral closure (before the velar closure) creates a zone with only nasal radiation, represented by N. The place of articulation of this nasal consonant, created by coarticulation, is the same as 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 milliseconds, the velum stayed closed, making an opening transition in 60 milliseconds 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 milliseconds. All stimuli had a fixed duration of 300 milliseconds.

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

We used a paired comparison test [40, page 361], because we were analysing the synthesis quality, despite the demand for more decisions by each listener, which also increases test duration. The question answered by listeners was as follows: 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.

The test was divided into 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. Interstimuli interval was 600 milliseconds.

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 boxplots of Figure 6.

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

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

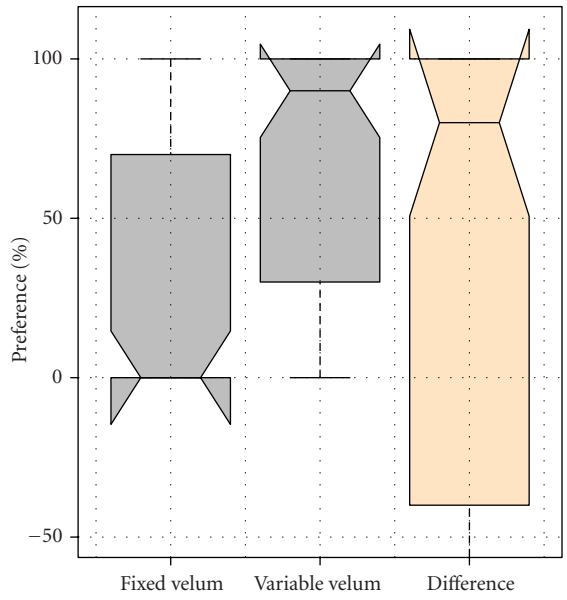


Figure 6: Boxplots of the preference scores for the first part of the perceptual test for nasal vowels in CṼC context, comparing stimuli with fixed and variable velum apertures, showing the effect of the velum aperture variation.

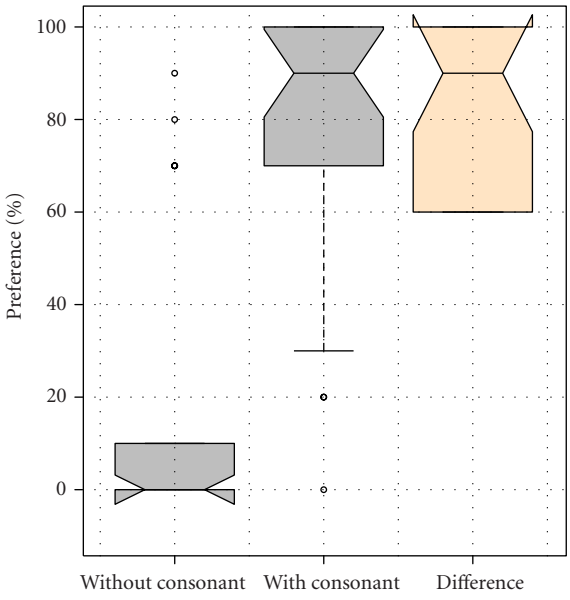


Figure 7: Boxplots of the preference scores for the second part of the perceptual test for nasal vowels in CṼC context, comparing stimuli with and without a final nasal consonant, showing the effect of the final nasal consonant.

Nasal consonant at nasal vowel end was preferred. In general, listener preferred stimuli ending in a nasal consonant. Looking at the preference scores represented graphically in Figure 7, stimuli with final nasal consonant were preferred more than stimuli without the final consonant. The confidence interval (CIp = 0.95) for the difference in preference score was between 36.1 and 87.0%, in favour of the stimuli with a final nasal consonant.

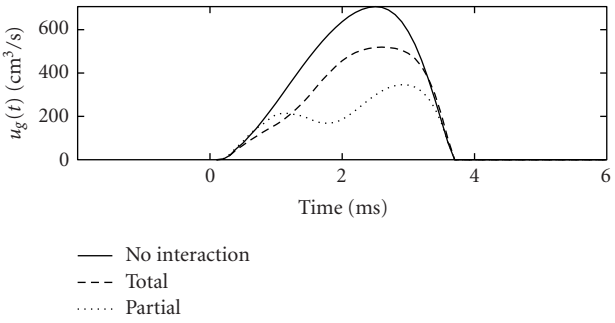


Figure 8: Glottal wave of 3 variants of vowel [ĩ]: (a) without tract load (no interaction); (b) with total tract load; (c) with tract input impedance calculated discarding nasal tract input impedance.

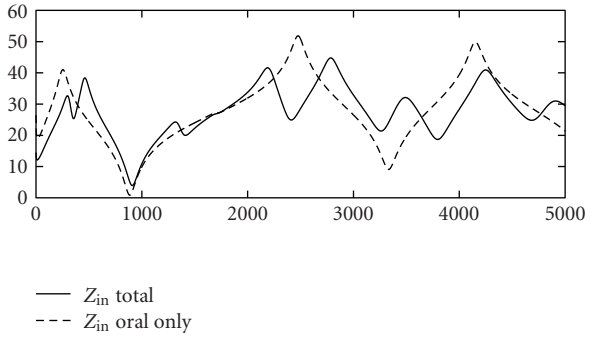


Figure 9: Input impedance for vowel [ĩ], with and without the nasal tract input impedance.

ANOVA results, with two main factors, confirmed a significant 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 [29].

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 resulted in a more natural quality synthetic speech.

Figure 8 depicts the effect of the 3 different input impedances in nasal vowel [ĩ]. The nasal tract load has a great influence on the glottal source wave, because of the noticeable difference in the input impedance calculated with or without the nasal tract input impedance, shown in Figure 9. This difference is due to the fact that for high vowels such as [ĩ] 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 Figure 10). The effect is less notorious in a low vowels, such as [ɐ̃].

Stimuli

Stimuli were produced for the EP nasal vowels varying only one factor: the input impedance of the tract used by the

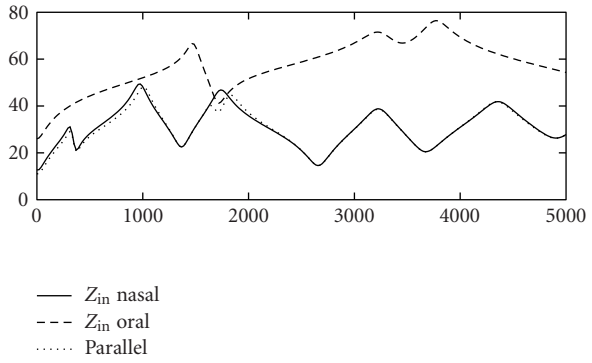


Figure 10: Input impedances in the velum region for nasal vowel [ĩ]. The Figure presents the oral input impedance (Z_{in} oral), the nasal tract input impedance (Z_{in} nasal), and the equivalent parallel impedance (parallel). The parallel impedance is, for this vowel, approximately equal to the nasal tract input impedance.

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, [ɐ̃], [ĩ], and [ũ], were considered to reduce test realization time.

The same timing was used for all vowels. In the first 100 milliseconds, the velum stayed closed, making an opening transition in 60 milliseconds to the maximum value. The velum remained at this maximum until the end of the vowel. The stimuli ended with a nasal consonant, a bilabial [m], produced by closing the lips. Closing movement of the lips started at 200 milliseconds and ended 50 milliseconds later. Stimulus duration was fixed at 300 milliseconds for all vowels. These choices were based on the results of the Experiment 1, where dynamic velum stimuli were preferred.

The interactive source model was used with variable F_0. F_0 starts around 100 Hz, raises to 120 Hz in the first 100 milliseconds, and then gradually goes back down to 100 Hz. The open quotient was 60% and the speed quotient 2. Jitter and shimmer were added to improve naturalness.

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, because better discrimination results have been reported with this type of perceptual test [4].

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 has different stimuli.

Table 2: Results of the 4IAX test.

Listener	Sex	[ɐ̃]	[ĩ]	[ũ]	Average
1	M	50.0	33.3	41.7	41.7
2	M	58.3	100.0	50.0	69.4
3	F	50.0	41.7	50.0	47.2
4	F	33.3	83.3	66.7	61.0
5	M	16.7	58.3	33.3	36.1
6	M	66.7	66.7	66.7	66.7
7	M	50.0	50.0	41.7	47.2
8	M	58.3	58.3	41.7	52.8
9	F	41.7	50.0	66.7	52.7
10	M	58.3	50.0	58.3	55.6
11	M	33.3	83.3	58.3	58.3
12	M	75.0	58.3	58.3	63.9
13	M	50.0	41.7	33.3	41.4
14	M	83.3	50.0	58.3	63.8
Average	—	51.8	58.9	51.8	54.1
Std.	—	17.3	18.6	11.9	10.3

Signals were presented over headphones in rooms with low ambient noise. Each of the 4 combinations (ABAA, ABBB, AAAB, and 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. Interstimuli interval was 400 milliseconds and interpairs interval was 700 milliseconds.

Results

Table 2 shows 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 Figure 11.

From the table and the boxplots, it is clear that listeners' correct answers were close to 50%, being a little higher for the nasal vowel [ĩ]. 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 [ĩ]. For this vowel, the 95% confidence interval for the mean was between 50.1 a 67.7. For [ɐ̃], we obtained $p = 0.36$ and for [ũ], $p = 0.29$. For the 3 vowels considered together, the average was 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 were perceived by listeners, supported the idea that these changes are hardly perceptible. These results agree with results reported in [41]. In their work, Titze and Story reported that "An open nasal port ... showed no measurable effect on oscillation threshold pressure or glottal flow."

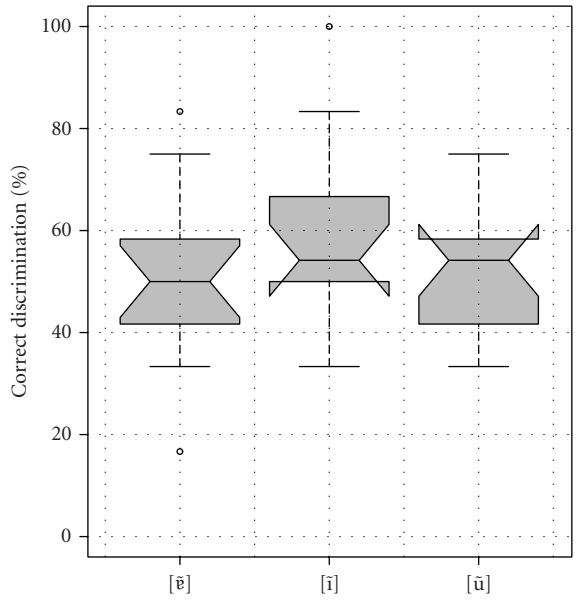


Figure 11: Boxplot of the 4IAX discrimination test results for evaluation of the listeners ability to perceive the effects of source-tract interaction on nasal vowels.

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

4.2. Fricatives

In a first experiment the synthesizer was used to produce, sustained unvoiced fricatives [42]. 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 [42]. Articulatory configurations for vowels were obtained by inversion [43]. 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 [ifi], 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 /ifi/are shown in Figure 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 $A_{g\max}$ rises to 2 cm^2 at the fricative middle point and at the end of the fricative returns to the value used during vowel production.

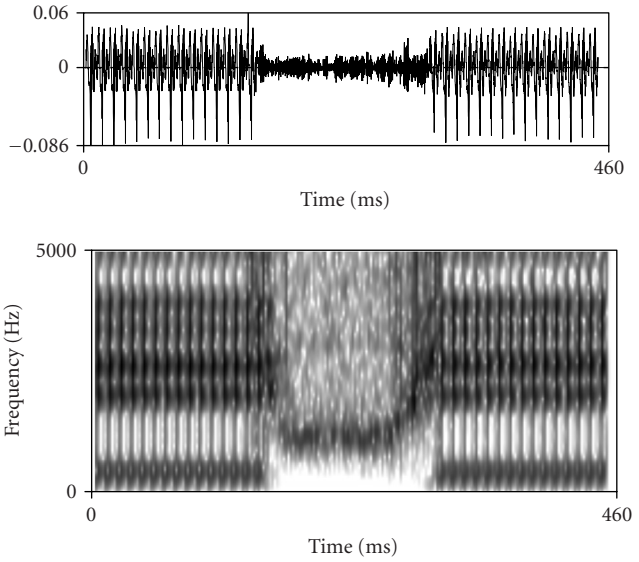


Figure 12: Synthetic [ifi], showing speech signal and spectrogram.

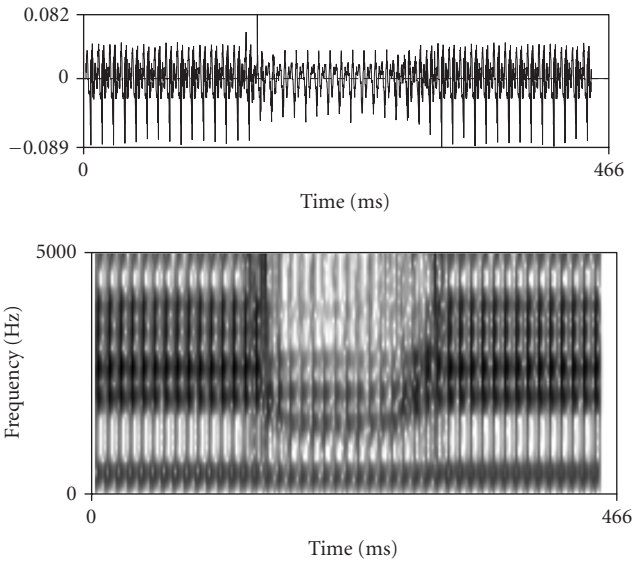


Figure 13: Synthetic [ivi], showing speech signal and spectrogram.

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

4.3. 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 durations for each phonetic segment entering the word composition (presently by direct analysis

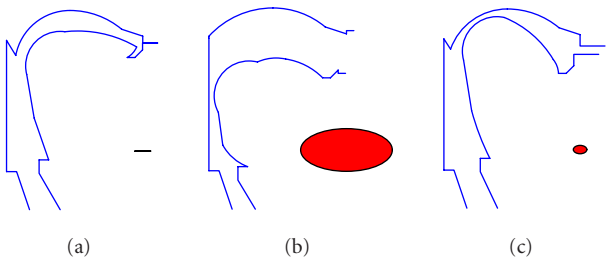


Figure 14: Tract configurations used to synthesize the word *mão* (hand): (a) [m], (b) [a], and (c) [u].

of natural speech although an automatic process, such as a CART tree, can be used in the future);

(2) obtaining oral articulators' configurations for each of the phones. For vowels we used configurations obtained by an inversion process based on the natural vowels' first four formants [43, 44]. These configurations were already available from previous work [39, 43]. For the consonants, for which we do not 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 the Portuguese language [45]. We now present three examples of synthetic words: *mão*, *mãe*, and *António*.

Example 1 (word *mão* (hand)). First, from natural speech analysis, we measured durations of 100 milliseconds for the [m] and 465 milliseconds 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 [43, 46]. The three configurations are presented in Figure 14.

A velum trajectory was defined, based on articulatory descriptions of the intervening sounds. As shown in Figure 15, the velum starts closed, in a preproduction 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 to 130 Hz at the end of the nasal consonant, then to 150 Hz to stress the initial part of the diphthong, and finally decreases to 80 Hz at the end of the word. This variation in time was based, partially, on 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.

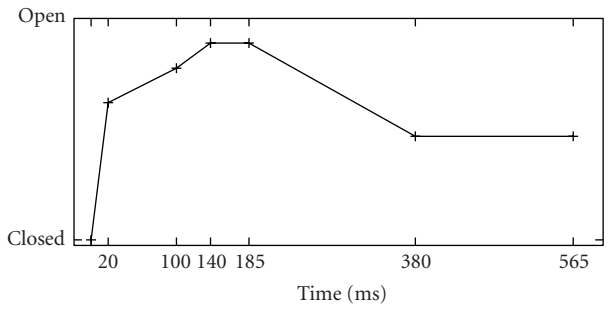


Figure 15: Velum trajectory used to synthesize the word *mão* (hand).

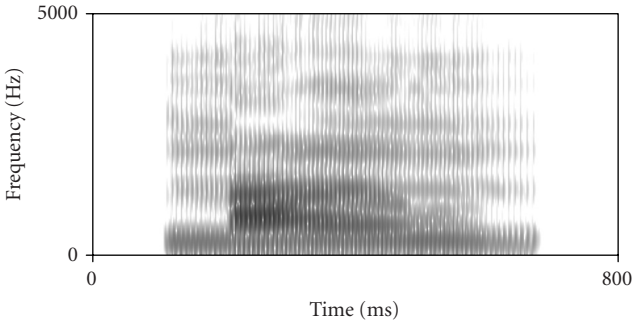


Figure 16: Spectrogram of the word *mão* produced by the articulatory synthesizer.

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 Figure 16.

Example 2 (word *mãe* (mother)). A possible phonetic transcription for the word *mãe* (mother) is ['mɐ̃ĩɲ], including a palatal nasal consonant at the end [45, page 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 [ɐ̃ĩ], we produced variants differing in the configuration used for the nasal vowel [ɐ̃]. One version was produced using the configuration of oral vowel [a], another, with a higher tongue position, using the configuration of vowel [ɐ]. 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 nonnatural F_0, source-tract interaction, configuration of [a] for nasal vowel [ɐ̃], and final palatal occlusion.

Example 3 (word *António*). The first name of the first author, *António* [ɐ̃'tɔnju], 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

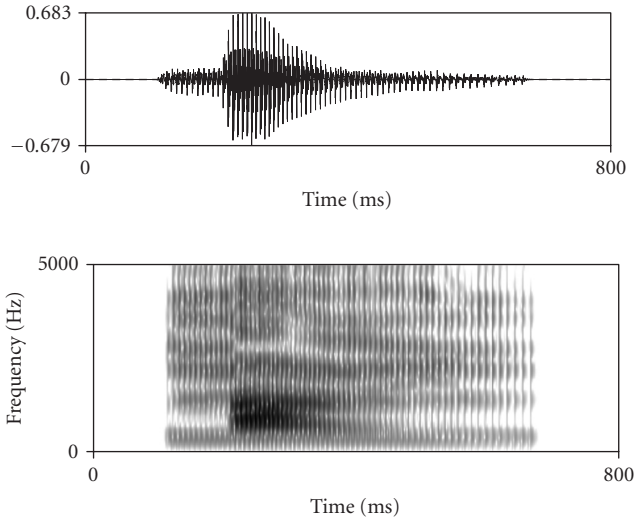


Figure 17: Speech signal and spectrogram of one of the versions of the word *mãe* synthesized using an [a] configuration at the beginning of the nasal diphthong, oral occlusion at the end, source-tract interaction, and synthetic values for F_0.

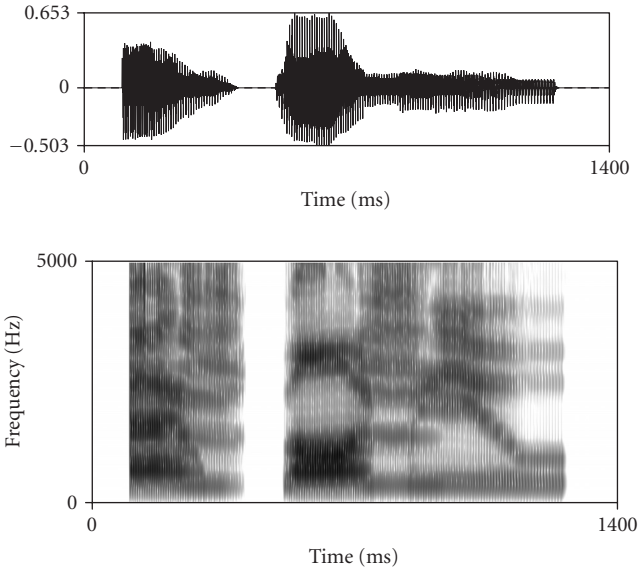


Figure 18: Speech signal and spectrogram for the synthetic word *António* [ɐ̃'tɔnju] produced using F_0 extracted from a natural pronunciation.

consonant, and a final oral diphthong. Two versions were produced: one with natural F_0, and another with synthetic F_0. The signal and its spectrogram obtained for the first version are presented in Figure 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 aspiration at the release of closure.

As part of a mean opinion score (MOS) quality test, this and many other stimuli produced by our synthesizer, were evaluated. To document the quality level achieved by our models, Table 3 shows the ratings of the various versions of the 3 examples presented above. The normalized (to 5) results varied between the values 3 and 4 (from fair to good). The top-rated word obtained 3.7 (3.4 without normalization).

5. CONCLUSION

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 and it allows 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 comprise 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 fair to good.

The synthesizer has been used, mainly, in the production of stimuli for perceptual tests of Portuguese nasal vowels (e.g., [39, 47, 48]). The two studies on nasal vowels reported in this paper were only possible with the use of the articulatory 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 the following.

(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 nasal vowels were word-final, the lips or tongue movement, even without occlusion, improved the quality of the synthesized 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 due to extra coupling of the nasal tract is not easily perceived. Discrimination was significantly above chance level only for the high vowel [ĩ], which can possibly be explained by the relation of nasal and oral input impedances at the nasal tract coupling point.

Table 3: Quality ratings for several words produced by the synthesizer. For each word, the table includes the mean opinion score (MOS), its respective 95% confidence interval, and the normalized value resulting from scaling natural speech scores to 5.

Word	F_0	Interac.	Observ.	MOS	CI 95%	Norm
mão	Synthetic	Yes	no [m] at end	3.4	[3.0–3.7]	3.7
	Synthetic	Yes	[m] at end	3.0	[2.7–3.4]	3.3
mãe	Natural	yes	[a], [ɲ] at end	2.9	[2.6–3.3]	3.2
	Synthetic	Yes	[a], [ɲ] at end	3.1	[2.7–3.4]	3.3
	Synthetic	Yes	[ɐ], [ɲ] at end	2.9	[2.6–3.3]	3.2
	Synthetic	Yes	[a], no [ɲ]	3.0	[2.6–3.4]	3.3
	Synthetic	No	[ɐ], [ɲ] at end	2.9	[2.5–3.3]	3.1
	Synthetic	No	[a], [ɲ] at end	2.8	[2.5–3.2]	3.1
António	Natural	Yes	—	3.0	[2.8–3.2]	3.3
	Synthetic	Yes	—	2.7	[2.4–2.9]	2.9
Natural speech	—	—	—	4.6	—	5.0

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. Namely, 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 judged in informal listening tests as highly intelligible. Our model of fricatives is comprehensive and flexible, making the new version of SAPWindows a valuable tool for trying out new or improved source models, and running production and perceptual studies of European Portuguese fricatives [49]. The possibility of automatically inserting and removing noise sources along the oral tract is a feature we regard as having great potential.

SAPWindows articulatory synthesizer is useful in phonetics research and teaching. We explored the first area for several years with very interesting results, as shown in this paper. Recently, we started exploring the second area, aiming at using the synthesizer in phonetics teaching at our University's Languages and Cultures Department. Articulatory synthesis is also of interest in the field of speech therapy because of its potential to model different speech pathologies.

Development of this synthesizer is an unfinished task. The addition of new models for other Portuguese sounds, the use of a combined data (MRI, EMA, EPG, etc.) for a detailed description of the vocal tract configurations and an optimal match between the synthesized and the Portuguese natural spectra [49], and the integration of the synthesizer in a text-to-speech system are planned as future work.

ACKNOWLEDGMENTS

This work was partially funded by the first author's Ph.D. Scholarship BD/3495/94 and the project "Articulatory Synthesis of Portuguese" P/PLP/11222/1998, both from the Portuguese Research Foundation (FCT) PRAXIS XXI program. We also have to thank the University of Florida's MMIRC, headed by Professor D. G. Childers, where this work started.

REFERENCES

[1] R. Linggard, *Electronic Synthesis of Speech*, Cambridge University Press, Cambridge, UK, 1985.

[2] M. R. Schroeder, *Computer Speech: Recognition, Compression, Synthesis*, vol. 35 of *Springer Series in Information Sciences*, Springer Verlag, New York, NY, USA, 1999.

[3] J.-P. Tubach, "Présentation Générale," Fondements et Perspectives en Traitment Automatique de la Parole, H. Méloni, Ed., Universités Francophones, 1996.

[4] G. J. Borden, K. S. Harris, and L. J. Raphael, *Speech Science Primer—Physiology, Acoustics, and Perception of Speech*, LWW, 4nd edition, 2003.

[5] D. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustic Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[6] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *Journal of the Acoustical Society of America*, vol. 70, no. 2, pp. 321–328, 1981.

[7] S. Maeda, "The role of the sinus cavities in the production of nasal vowels," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '82)*, vol. 2, pp. 911–914, Paris, France, May 1982.

[8] T. Koizumi, S. Tanigushi, and S. Hiromitsu, "Glottal source-vocal tract interaction," *Journal of the Acoustic Society of America*, vol. 78, no. 5, pp. 1541–1547, 1985.

[9] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 7, pp. 955–967, 1987.

[10] C. H. Shadle and R. Damper, "Prospects for articulatory synthesis: A position paper," in *Proc. 4th ISCA Tutorial and Research Workshop (ITRW '01)*, Perthshire, Scotland, August–September 2001.

[11] D. H. Whalen, "Articulatory synthesis: Advances and prospects," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, pp. 175–177, Barcelona, Spain, August 2003.

[12] B. Kühnert and F. Nolan, "The origin of coarticulation," *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, vol. 35, pp. 61–75, 1997, and also in [50]. Online. Available: http://www.phonetik.uni-muenchen.de/FIPKM/index.html.

[13] A. Pinto and A. M. Tomé, "Automatic pitch detection and midi conversion for the singing voice," in *Proc. WSES International Conferences: AITA '01, AMTA '01, MCBE '01, MCBC '01*, pp. 312–317, Greece, 2001.

[14] A. M. Öster, D. House, A. Protopapas, and A. Hatzis, "Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia)," in *Proc. TMH-QPSR, Fonetik 2002*, vol. 44, pp. 45–48, Stockholm, Sweden, May 2002.

[15] F. S. Cooper," Speech synthesizers," in *Proc. 4th International Congress of Phonetic Sciences (ICPhS '61)*, A. Sovijärvi and P. Aalto, Eds., pp. 3–13, The Hague: Mouton, Helsinki, Finland, September 1961.

[16] M. Wrembel, "Innovative approaches to the teaching of practical phonetics," in *Proc. Phonetics Teaching & Learning Conference (PTLC '01)*, London, UK, April 2002.

[17] S. S. Fels, F. Vogt, B. Gick, C. Jaeger, and I. Wilson, "User-centred design for an open source 3-D articulatory synthesizer," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, vol. 1, pp. 179–183, Barcelona, Spain, August 2003.

[18] K. Iskarous, L. Goldstein, D. H. Whalen, M. K. Tiede, and P. E. Rubin, "CASY: The Haskins configurable articulatory synthesizer," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, vol. 1, pp. 185–188, Barcelona, Spain, 2003.

[19] S. Maeda and M. Toda, "Mechanical properties of lip movements: How to characterize different speaking styles?" in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, vol. 1, pp. 189–192, Barcelona, Spain, August 2003.

[20] P. Badin, G. Bailly, F. Elisei, and M. Odisio, "Virtual talking heads and audiovisual articulatory synthesis," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, vol. 1, pp. 193–197, Barcelona, Spain, August 2003.

[21] K. N. Stevens and H. M. Hanson, "Production of consonants with a quasi-articulatory synthesizer," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, vol. 1, pp. 199–202, Barcelona, Spain, August 2003.

[22] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[23] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Siganl Processing*, Marcel Dekker, New York, NY, USA, 1992.

[24] P. P. L. Prado, *A target-based articulatory synthesizer*, Ph.D. dissertation, University of Florida, Gianesville, Fla, USA, 1991.

[25] A. Teixeira, F. Vaz, and J. C. Príncipe, "A comprehensive nasal model for a frequency domain articulatory synthesizer," in *Proc. 10th Portuguese Conference on Pattern Recognition (RecPad '98)*, Lisbon, Portugal, March 1998.

[26] M. Chen, "Acoustic correlates of English and French nasalized vowels," *Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360–2370, 1997.

[27] J. Dang and K. Honda, "MRI measurements and acoustic investigation of the nasal and paranasal cavities," *Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1765–1765, 1993.

[28] K. N. Stevens, *Acoustic Phonetics*, Current Studies in Linguistics, MIT Press, Cambridge, Mass, USA, 1998.

[29] A. Teixeira, F. Vaz, and J. C. Príncipe, "Effects of source-tract interaction in perception of nasality," in *Proc. 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, vol. 1, pp. 161–164, Budapest, Hungary, September 1999.

[30] D. Allen and W. Strong, "A model for the synthesis of natural sounding vowels," *Journal of the Acoustical Society of America*, vol. 78, no. 1, pp. 58–69, 1985.

[31] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, no. 3-4, pp. 167–184, 1982.

[32] L. Silva, A. Teixeira, and F. Vaz, "An object oriented articulatory synthesizer for Windows," *Revista do Departamento de Electrónica e Telecomunicações, Universidade de Aveiro*, vol. 3, no. 5, pp. 483–492, 2002.

[33] E. L. Riegelsberger, *The acoustic-to-articulatory mapping of voiced and fricated speech*, Ph.D. dissertation, The Ohio State University, Columbus, Ohio, USA, 1997.

[34] Q. Lin, "A fast algorithm for computing the vocal-tract impulse response from the transfer function," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 6, pp. 449–457, 1995.

[35] M. Frigo and S. Johnson, "FFTW: an adaptive software architecture for the FFT," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 3, pp. 1381–1384, Seattle, Wash, USA, 1998.

[36] S. S. Narayanan and A. A. H. Alwan, "Noise source models for fricative consonants," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 328–344, 2000.

[37] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, NY, USA, 2nd edition, 1972.

[38] C. H. Shadle, "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., pp. 187–209, Kluwer Academic, Dordrecht, The Netherlands, 1990.

[39] A. Teixeira, F. Vaz, and J. C. Príncipe, "Influence of dynamics in the perceived naturalness of portuguese nasal vowels," in *Proc. 14th International Congress of Phonetic Sciences (ICPhS '99)*, San Francisco, Calif, USA, August 1999.

[40] N. Kitawaki, "Quality assessment of coded speech," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds., chapter 12, pp. 357–385, Marcel Dekker, New York, NY, USA, 1992.

[41] I. R. Titze and B. H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2234–2243, 1997.

[42] A. Teixeira, L. M. T. Jesus, and R. Martinez, "Adding fricatives to the Portuguese articulatory synthesizer," in *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 2949–2952, Geneva, Switzerland, September 2003.

[43] A. Teixeira, F. Vaz, and J. C. Príncipe, "A software tool to study Portuguese vowels," in *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)* , G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds., vol. 5, pp. 2543–2546, Rhodes, Greece, September 1997.

[44] A. Teixeira, F. Vaz, and J. C. Príncipe, "Some studies of European Portuguese nasal vowels using an articulatory synthesizer," in *Proc. 5th IEEE International Conference on Electronics, Circuits and Systems (ICECS '98)*, vol. 3, pp. 507–510, Lisbon, Portugal, September 1998.

[45] J. Laver, *Principles of Phonetics*, Cambridge Textbooks in Linguistics, Cambridge University Press, Cambridge, UK, 1st edition, 1994.

[46] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, New York, NY, USA, 2000.

[47] A. Teixeira, L. C. Moutinho, and R. L. Coimbra, "Production, acoustic and perceptual studies on European Portuguese nasal vowels height," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)*, Barcelona, Spain, August 2003.

[48] A. Teixeira, F. Vaz, and J. C. Príncipe, "Nasal vowels following a nasal consonant," in *Proc. 5th Seminar on Speech Production: Models and Data*, pp. 285–288, Bavaria, Germany, May 2000.

[49] L. M. T. Jesus and C. H. Shadle, "A parametric study of the spectral characteristics of European Portuguese fricatives," *Journal of Phonetics*, vol. 30, no. 3, pp. 437–464, 2002.

[50] W. J. Hardcastle and N. Hewlett, Eds., *Coarticulation: Theoretical and Empirical Perspectives*, Cambridge University Press, Cambridge, Mass, USA, 1999.

António J. S. Teixeira was born in Paredes, Portugal, in 1968. He received his first degree in electronic and telecommunications engineering in 1991, the M.S. degree in electronic and telecommunications engineering in 1993, and the Ph.D. degree in electrical engineering in 2000, all from the University of Aveiro, Aveiro, Portugal. His Ph.D. dissertation was on articulatory synthesis of the Portuguese nasals. Since 1997, he has been teaching in the Department of Electronics and Telecommunications Engineering at the University of Aveiro, a "Professor Auxiliar" since 2000, and has been a Researcher, since its creation in 1999, in the Signal Processing Laboratory at the Institute of Electronics and Telematics Engineering of Aveiro (IEETA), Aveiro, Portugal. His research interests include digital processing of speech signals, particularly (articulatory) speech synthesis; Portuguese phonetics; speaker verification; spoken language understanding; dialogue systems; and man-machine interaction. He is also involved in a new Master's program in the area of speech sciences and hearing, as the Coordinator. He is a Member of The Institute of Electrical and Electronics Engineers, International Speech Communication Association, and the International Phonetic Association.

Roberto Martinez was born in Cuba in 1961. He received his first degree in physics in 1986 from the Moscow State University M. V. Lomonosov, former USSR. He is a Microsoft Certified Engineer since 1998. From 1986 to 1994, he was an Assistant Professor of mathematics and physics at the Havana University, Cuba, doing research in computer aided molecular design. From 1996 to 1998, he was with SIME Ltd., Cuba, as an Intranet Developer and System Administrator. From 1999 to 2001, he was with DISAIC Consulting Services, Cuba, doing training of Network Administrators and consulting in Microsoft BackOffice Systems Integration and Network Security. He is currently working toward the Doctoral degree in articulatory synthesis of Portuguese at the University of Aveiro, Portugal.

Luís Nuno Silva received his first degree in electronics and telecommunications engineering in 1997 and the M.S. degree in electronics and telecommunications engineering in 2001, both from the Universidade de Aveiro, Aveiro, Portugal. From 1997 till 2002, he worked in research and development at the Instituto de Engenharia Electrónica e Telemática de Aveiro, Aveiro, Portugal (former Instituto de Engenharia de Sistemas e Computadores of Aveiro, Aveiro, Portugal) as a Research Associate. Since 2002, he has been working as a Software Engineer at the Research and Development Department of NEC Portugal, Aveiro, Portugal. His (research) interests include digital processing of speech signals and speech synthesis. He is a Member of The Institute of Electrical and Electronics Engineers.

Luis M. T. Jesus received his first degree in electronic and telecommunications engineering in 1996 from the Universidade de Aveiro, Aveiro, Portugal, the M.S. degree in electronics in 1997 from the University of East Anglia, Norwich, UK, and the Ph.D. degree in electronics in 2001 from the University of Southampton, UK. Since 2001, he has been a Reader in the Escola Superior de Saúde da Universidade de Aveiro, Aveiro, Portugal, and has been a member of the Signal Processing Laboratory at the Instituto de Engenharia Electrónica e Telemática de Aveiro, Aveiro, Portugal. His research interests include acoustic phonetics, digital processing of speech signals, and speech synthesis. He is a Member of The Acoustical Society of America, Associação Portuguesa de Linguística, International Phonetic Association, International Speech Communication Association, and The Institute of Electrical and Electronics Engineers.

Jose C. Príncipe is a Distinguished Professor of electrical and biomedical engineering at the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is a BellSouth Professor and Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). He has been involved in biomedical signal processing, brain machine interfaces, nonlinear dynamics, and adaptive systems theory (information theoretic learning). He is the Editor-in-Chief of IEEE Transactions on Biomedical Engineering, President of the International Neural Network Society, and formal Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He is also a Member of the Scientific Board of the Food and Drug Administration, and a Member of the Advisory Board of the University of Florida Brain Institute. He has more than 100 publications in refereed journals, 10 book chapters, and over 200 conference papers. He has directed 42 Ph.D. degree dissertations and 57 M.S. degree theses.

Francisco A. C. Vaz was born in Oporto, Portugal, in 1945. He received the Electrical Engineering degree from University of Oporto, Portugal, in 1968, and the Ph.D. degree in electrical engineering from the University of Aveiro, Portugal, in 1987. His Ph.D. dissertation was on automatic EEG processing. From 1969 to 1973, he worked for the Portuguese Nuclear Committee. After several years working in the industry, he joined, in 1978, the staff of the Department of Electronics Engineering and Telecommunications, the University of Aveiro, where he is currently a Full Professor. His research interests have centred on the digital processing of biological signals, and since 1995 on digital speech processing.