A SOFTWARE TOOL TO STUDY PORTUGUESE VOWELS

António Teixeira, Francisco Vaz
Dep. Electrónica e Telecomunicações/INESC, Universidade de Aveiro
Campus Universitário, 3810 AVEIRO, Portugal
Tel. +351 34 370 500, FAX: +351 34 370 541, E-mail:{ajst,fvaz}@inesca.pt

José Carlos Princípe
Department of Electrical Engineering, University of Florida
CSE 444, Gainesville, FL 32611, USA, E-mail: principe@synapse.ee.ufl.edu

ABSTRACT

We are developing a software system to help the study of Portuguese Vowel Production. This tool is an articulatory synthesizer with a graphical user interface. The synthesizer is composed of a saggittal articulatory model derived from Mermelstein model and a frequency domain simulation of the electric analog of the acoustic tube. User can easily define the nasal tract configuration. System includes optimization by simulated annealing to perform acoustic-to-articulatory mapping. In this paper we present the system being developed, its current state and future perspectives. Preliminary experiments with Portuguese Vowels gave good results.

1. INTRODUCTION

The Portuguese language has a rich set of vocoids [1], central resonants nuclear in the syllable. There are 9 oral and 5 nasal monophthongs, several crescent and decrescent diphthongs, and also triphthongs. Oral vowels, using the International Phonetic Alphabet, are [i], [e], [ɛ], [a], [ɔ], [o], [u], [ə], and a "central closed unrounded [a]" represented in Portuguese phonetic literature as [α] (See [2]). Portuguese [ə] is sometimes represented as [ɨ] due to a more closed configuration. Figure 1 presents a sketch of the Portuguese vowel "triangle".

In Portuguese there are 5 nasal vowels, several nasal diphthongs and also triphthongs. The nasal vowels are [$\tilde{\alpha}$], [ẽ], [ĩ], [õ], [ũ]. Nasal sounds are a distinctive characteristic of Portuguese language. Portuguese nasal vowels are reported to have nasality contours, that is, they start in an oral configuration and make a transition to a nasal configuration. This makes Portuguese nasal vowels different from counterparts in other languages such as French.

Our purpose is study the Portuguese Vocoids, both oral and nasal, in what concerns to production and perception. Because nasals present new problems we started our work by the oral vowels. This work will, hopefully, be in the near future extended for nasal vowels, our main objective. We have choosed to study the nasals by simulation using an articulatory synthesizer, as Maeda [3] and the classical work of House & Stevens, 1956 [4]. The synthesizer is used in an analysis-synthesis task. Parameters for the articulatory synthesizer are extracted from recorded speech signal.

This work addresses only the sounds of the Portuguese language as spoken in Europe, called European Portuguese.

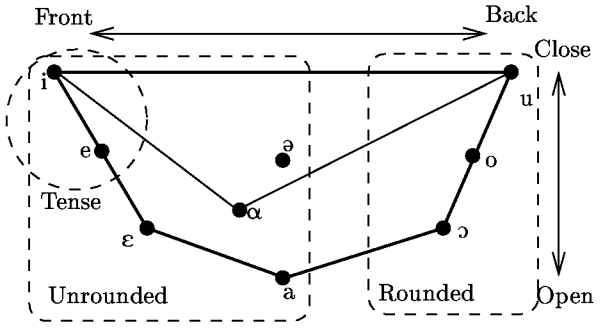


Figure 1. Portuguese Oral Monophthongs

To conduct this experiments a database of words and phrases containing all the Portuguese phonemes was collected in a sound room with an high quality microphone. In the work reported here we used only data of two male speakers.

2. TOOL: ARTICULATORY SYNTHESIS

The system is being developed in a Personal Computer with a standard sound board running the Linux operating system. Software is written in C/C++ and Tcl-Tk. This combination of Hardware with free software makes the system inexpensive. The use of Tcl-Tk has made possible fast development of a easily reconfigurable Graphical User Interface.

Our synthesizer is composed of: (1) a saggittal articulatory model, (2) an acoustic model. In the following subsections we present in more detail both modules.

2.1. Saggittal Articulatory Model

The articulatory model used was developed at the *Mind-Machine Interaction Research Center*, University of Florida, by D. Childers and co-workers [5]. It's an enhanced version of Mermelstein's [6] model. Articulatory parameters are: tongue body center, tongue tip, jaw, lips, hyoid and velum. Also the dimensions of the lower pharynx can be adjusted.

User can define model parameters using sliders or typing values directly. The articulatory model window is shown in Figure 2. In the left of the window are the

sliders to adjust articulators position. Below the articulatory model are checkbuttons to enable/disable display of the grid, midline, constriction position, lips protrusion and aperture, tract length, position of maximum constriction and respective area, nasal coupling, oral area in nasal coupling section.

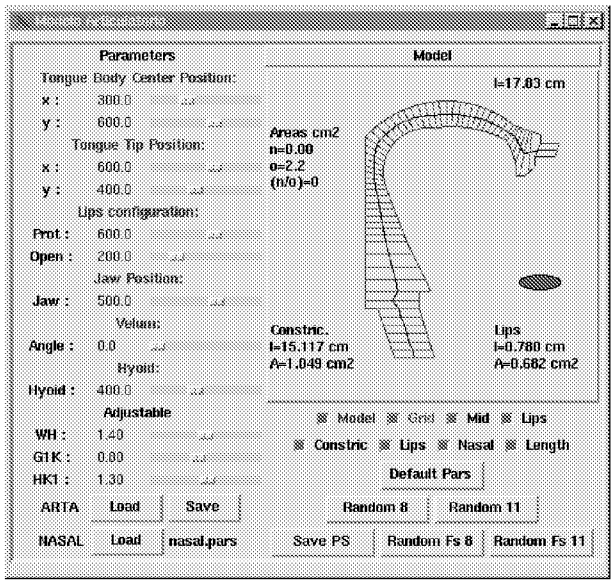


Figure 2. Articulatory Model Window

2.2. Acoustic Model

The acoustic model consists of the concatenation of elementary constant area tubes, represented by an analog electrical model. Two alternative formulations can be used for this elementary components: the one proposed by Sondhi & Schroeter 1987 [7] and the one of Flanagan an followers [8]. Viscous, conduction and yielding walls losses can be included at user choice.
3 radiation models are implemented: the Stevens, Kasowsky & Fant; Flanagan [8]; and Ideal.
Currently the glottal source model is the non-interactive parametric LF-model.

2.2.1. Articulatory synthesis of nasal sounds

Our system provides the user with flexibility in defining the nasal tract configuration. By editing a file the user can provide different dimensions of the nasal tract.
The tract can be symmetric or asymmetric. In an asymmetric model user defines area function for the common, left and right nasal paths. Using a symmetric model as the one used by Maeda in 1982 [3] or an asymmetric one using the MRI data of Dang [11] presents the user the same difficulties.
Paranasal sinus, implemented as an RLC shunt, can be include in any position. The user can simulate the blocking of the nasal passage including a zero area element.
Maeda port model [12] has been used to label formants as oral or nasal.

2.2.2. System simulation

To obtain the frequency response of the analog model an frequency domain analysis is made using a process similar to the described by Lin [9, 10]. This approach makes possible to take in account the frequency dependent losses.
To avoid contamination of the poles by the zeros and vice-versa, the pole and zero function of the transfer function are separated.
Synthetic speech can be generated in two ways: by convolution of source signal and impulse response of the analog system (obtained by Inverse Fast Fourier Transform from the frequency response); and by a parallel filter bank approximation to the system frequency response, as proposed by Lin [10]. Synthetic speech can be played directly to the soundcard or saved for further processing.

2.3. Graphical User Interface features

Several useful features are available to the user, like:

1. The user can see instantly the effect of articulatory parameter change in area function, frequency response, and formants, vocal tract total length, constriction location, lips configuration. In Figure 3 is presented the articulatory model and related information as presented in the screen;

2. The user can save to a file, for posterior analysis, frequency responses and input impedances of different subsystems, such as nasal tract alone;

3. System generates Postscript output of the articulatory model configuration directly. Articulatory configurations presented in this paper were produced automatically.

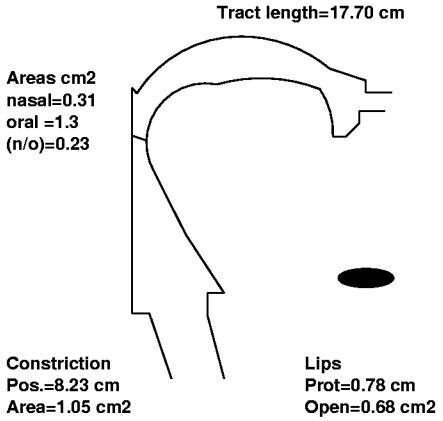


Figure 3. Articulatory Model and information about tract length, lips, constriction, and nasal coupling for a sample configuration.

3. ACOUSTIC TO ARTICULATORY MAPPING

We need the vocal tract shape for the vowels. Direct measures are difficult and for the time being we have

no access to such devices as X-rays or MRI. The alternative is determination of shape using the speech signal by an optimization process.

As optimization tool we choosed Simulated annealing [13] that gave good results in American vowels inversion [5]. In the optimization process configurations are constrained to be anatomically correct and to have an appropriate configuration for a vowel, that is minimum area can't drop below a minimum prefixed value ($0.2\ cm^2$).

Several error measures, all using formants, are available: the weighted Euclidean distance [5], Bark scale difference [14], and uniform metrics proposed by Sorokin [15]. User can choose the number of formants, between 1 and 4, to use in the error measure.

Any subset of the articulatory parameters can be selected for optimization. The others remain fixed at initial values.

As start configurations user can select between a neutral, random configuration, or user defined configuration.

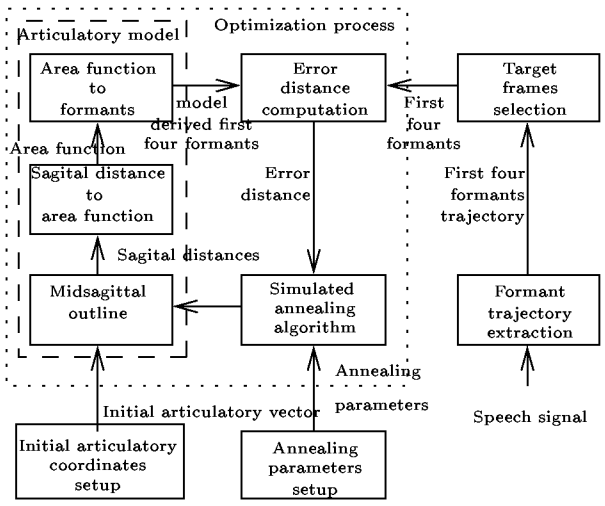


Figure 4. The block diagram of the speech inverse filtering procedure.

4. PRELIMINARY EXPERIMENTS RESULTS

We are testing the capacities of the simulated annealing in the inversion task. Due to the several degrees of freedom in the system (error measures, acoustic model, radiation, type of optimization initialization, ...) a complete evaluation of the different alternatives have not been performed.

We have made two kinds of tests. First, we tested the simulated annealing capacities with model generated formants. After, we applied the inversion process to formant data obtained from natural speech.

4.1. Tests with Model Generated Formants

Preliminary tests, with system generated configurations as targets, showed that simulated annealing is capable of finding very close configurations. With 5000 iterations system is able to obtain configurations with frequency differences below JNDs and error below 1 % in formants. Also place of maximum constriction, and constriction area, is very close. Tests with subsets of parameters also showed good results. As an example, if lips configuration is known only the other parameters are optimized.

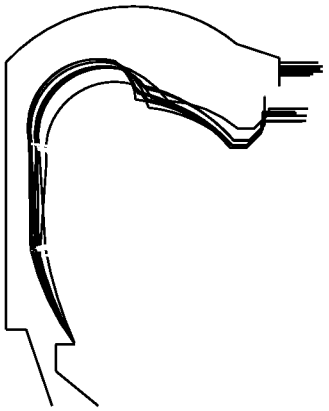


Figure 5. 5 different configurations obtained for vowel [a]. Target formants were $F_1 = 624$, $F_2 = 1316$, $F_3 = 2432$ and $F_4 = 3475$.

Simulated annealing gives similar results in several runs. As an example we present in Figure 5 10 saggittal configurations obtained for vowel [a]. Optimization used 10000 iterations and the Euclidean metric. Configuration agrees with articulatory phonetic descriptions, jaw is lowered, tongue is lowered and central.

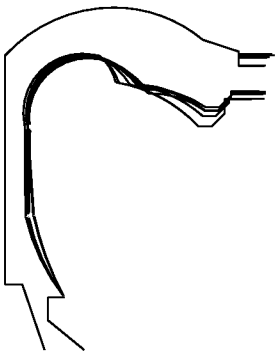


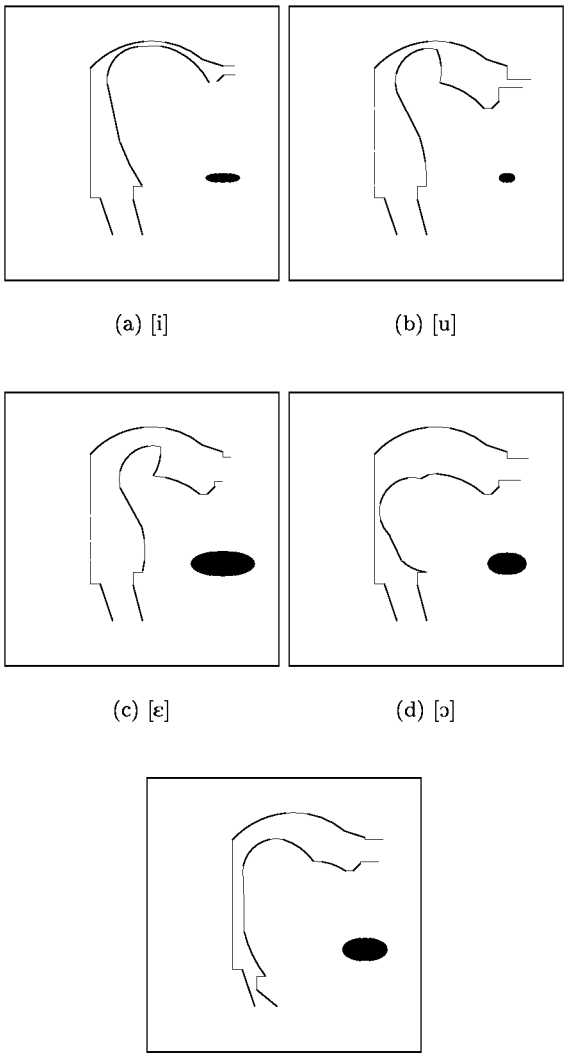
Figure 6. Configurations obtained using the three metrics for the same vowel, [a].

Results obtained using the three metrics are compared in Figure 6 for vowel [a]. Obtained configurations are very similar in tongue body position, jaw opening, an lips configuration.

4.2. Tests with natural speech

Also inversion with real speech sounds have been tested. Preliminary tests with the inversion of Portuguese oral vowels have been done with promising

results [16]. As example we present in Figure 7 vocal tract configurations obtained for 5 vowels, with the Euclidean metric.



(a) [i] (b) [u]

(c) [ɛ] (d) [ɔ]

(e) [a]

Figure 7. Configuration for 5 Portuguese vowels, obtained by inversion.

Analyzing the obtained configurations they are as predicted by articulatory phonetic descriptions. Vowel [i] has a high and front tongue, [u] has also high but backed tongue and is noticeable the rounding of lips, vowel [a] as anticipated presents a pharynx narrowing and jaw is at maximum lowering permitted by the model. Additionally we have permitted optimization of lower pharynx in vowel [a] to attain the very high value of F_1 needed. Vowel [ɔ] is clearly backed.

5. CONCLUSION

Our experiments in acoustic-to-articulatory mapping by an optimization process using simulated annealing gave plausible configurations for Portuguese oral vowels. Nevertheless, our work has only started. There is many improvements to be made.
We are extending this work to obtain also the Velum parameter in nasal vowels. Also to improve optimization we intend to use a codebook, neural network or fuzzy system to get initial configuration. Different metrics, like the LPC poles metric [17], could also improve performance.

6. ACKNOWLEDGMENTS

The first author is grateful to JNICT (Praxis XXI) for the PhD scholarship BD/3495/94 that made possible this work. We also thank University of Florida for the stay of the first author at the Mind-Machine Interaction Research Center were this work was started.

7. REFERENCES

[1] John Laver, "Principles of Phonetics", Cambridge University Press, 1994.

[2] Maria Raquel Delgado Martins, "Ouvir Falar - Introdução à Fonética do Português", Caminho, 1988.

[3] Shinji Maeda, *"The role of the sinus cavities in the production of nasal vowels"*, Proc. ICASSP, pp. 911–914, 1982.

[4] Arthur S. House and Kenneth S. Stevens, "Analog studies of the nasalization of vowels", *Journal of Speech and Hearing Disorders*, 21(2), pp. 218–232, June 1956.

[5] Yu-Fu Hsieh, "A Flexible and High Quality Articulatory Speech Synthesizer". PhD thesis, University of Florida, 1994.

[6] P. Mermelstein, "Articulatory model for the study of speech production". *J. Acoust. Soc. Am.*, 53(4), pp. 1070–1082, 1973.

[7] Man Mohan Sondhi and Juergen Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-35(7), pp. 955–967, July 1987.

[8] James L. Flanagan, "Speech Analysis, Synthesis and Perception", Springer-Verlag, New York, 1972.

[9] Qiguang Lin, "Speech production theory and Articulatory Speech Synthesis". PhD thesis, Dept. of Speech Comm. & Music Acoustics, Royal Institute of Technology (KTH), Stockolm, Sweden, 1990.

[10] Qiguang Lin, "A fast algorithm for computing the vocal-tract impulse response from the transfer function", *IEEE Trans. Speech Audio Proc.*, 3(6), pp. 449–457, Nov. 1995.

[11] Jianwu Dang and Kiyoshi Honda, "MRI measurements and acoustic of the nasal and paranasal cavities", *J. Acoust. Soc. Am.*, 94(3), Pt. 2, pp. 1765, Sept. 1994.

[12] Shinji Maeda, *"Acoustics of vowel nasalization and articulatory shifts in french nasal vowels"*. In Marie K. Huffman and Rena A. Krakow, editors, "Nasals, Nasalization, and the Velum", pp. 147–167, Academic Press, 1993.

[13] A. Corana, M. Marchesi, C. Martini, and S. Ridella, "Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm", *ACM Transactions on Mathematical Software*, 13(13), Sept. 1987.

[14] Mats Båvegård and Gunnar Fant, "From formant frequencies to VT area fuction parameters", *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4, pp. 55–66, 1995.

[15] Victor N. Sorokin, "Inverse problem for fricatives", *Speech Communication*, 14(3), pp. 249–262, June 1994.

[16] A. Teixeira, F. Vaz, J. C. Princípe, and D. G. Childers, "Articulatory synthesis of portuguese vocoids", 9th Portuguese Conference on Pattern Recognition (RecPad'97), pp. 219–224, Coimbra, March 1997.

[17] Frédéric Zussa, "A new design for articulatory parametrization of speech: Application to low-bit rate coding and recognition", Industrial thesis report, CAIP, Rutgers University, August 1995.