An Implementation of the Klatt Speech Synthesiser*

Luis Miguel Teixeira de Jesus, Francisco Vaz, José Carlos Principe

Resumo **- Neste trabalho descreve-se a implementação de um programa que permite explorar o sintetizador de Klatt num ambiente laboratorial com objectivos didáticos. A interface com o utilizador permite a fácil edição dos parâmetros de síntese. Desta forma o estudante pode alterar de um modo fácil e rápido as características de uma vogal podendo de imediato efectuar a sua audição ou a observação gráfica das características no tempo ou frequência.**

Abstract **-In this work it is described a software tool implementing the Klatt synthesiser to be used on a teaching laboratory. The user interface allows an easy edition of the synthesis parameters, thus the student may quick and easily change the phonetic characteristics of a vowel and listen to the result or seeing its time or frequency properties.**

I. INTRODUCTION

The aim of this project was to implement, for teaching applications, a speech synthesiser adapted to the portuguese language, based on the model proposed by Klatt [3]. The model for this synthesiser is based on the acoustic theory of speech production developed by Fant [1], [2] : the speech wave is the response of the vocal tract to one or more sound sources, as shown on figure 1.

Thus, the speech wave may be specified in terms of source and filter characteristics:

$$P(f)= U(f)\cdot T(f)= U(f).H(f).R(f)$$

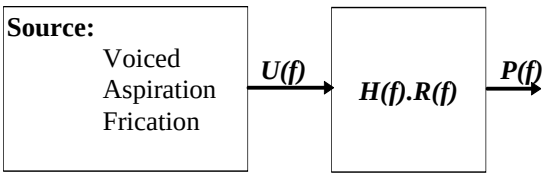


Figure 1- Acoustic model of speech production.

where $U(f)$ is the glottal source volume velocity, $T(f)=H(f)\cdot R(f)$ is the overall transfer function composed by $H(f)$ the frequency response of the vocal tract and $R(f)$ the radiation characteristic, i.e., the conversion from volume velocity to air pressure on the lips.

The source used in this synthesiser - the LF model - has been described by Fant et al [2] and it is shown on figure 2. It is as a glottal flux model with four independent parameters and it is known to ensure a smooth fit to the natural waveform with minimum number of parameters, being flexible when adjusted to turbulent phonations.

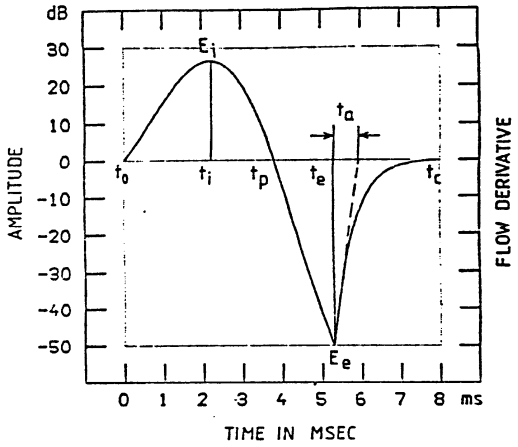


Figure 2 - LF model of the glottal flow derivative. The wave shape is determined by the parameters tp, te, ta and Ee. Adapted from [2] Fant, Liljencrants and Lin

The system that was implemented is a software package made up of three main parts: the source, the tract and the graphical interface.

The source (figure 3) is a program producing a suitable waveform for voiced (impulse train and LF model), unvoiced (noise) or mixed speech sounds.

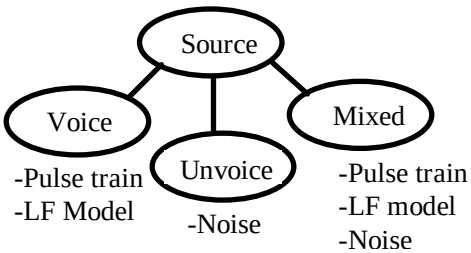


Figure 3 - The Source.

The tract is a program that implements the model of the vocal tract using a parallel or a cascade architecture, figure 4.

* Trabalho realizado no âmbito da displina de projecto.

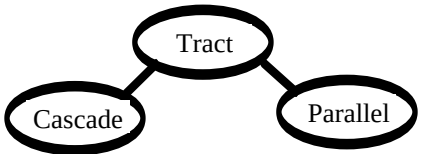


Figure 4 - The Tract.

Finally the graphical interface is the program that allows the user to edit the parameters, analyse the output waveform (time and frequency domains) and hear the synthesised sounds (Figure 5) . This tool permits the signal visualisation and the computation of the Fourier transform, the LPC model and the spectrogram of speech natural or synthesised segments.

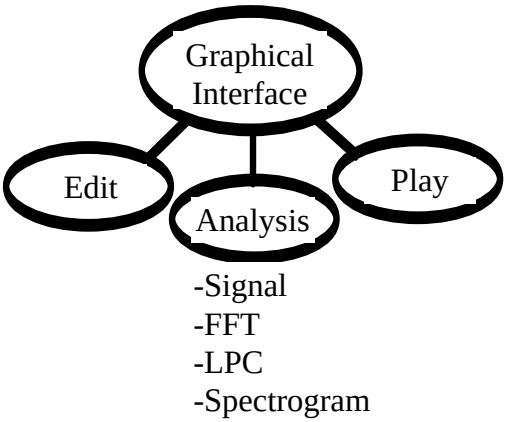


Figure 5 - The Graphical Interface.

The whole program structure is depicted in Figure 6, where we can see that The main characteristics of the waveform to synthesise may be edited through the graphical interface and stored as synthesis parameter and LF model files. With this information the vocal tract builds a waveform that can be directly listened on the loudspeakers or stored as an output file. The graphical interface provides also listening and display facilities.

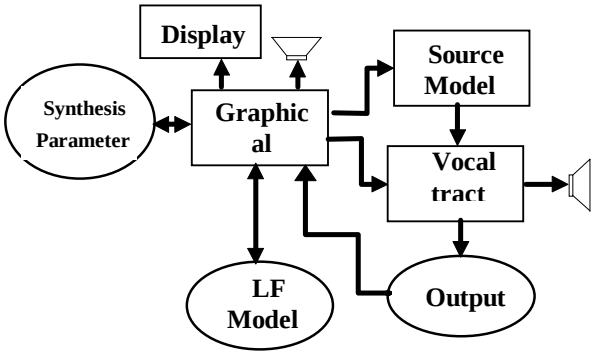


Figure 6 - The Synthesiser with all associated components.

II. THE KLATT SYNTHESISER

Speech synthesis methods can be divided into two categories:

- Articulatory synthesis: a model based on physiological knowledge tries to reproduce the acoustic properties of the vocal;
- Formant synthesis: the model is an approximation of the natural waveform through a set of rules based on acoustic studies of speech production .

The formant models are simpler and require less calculation than articulatory models, and that was the approach proposed by Klatt that we are describing. The Klatt synthesiser is basically a digital filter with several resonances modelling the speech formants. There two ways of implementation: cascade and parallel [6]:

- Parallel configuration - the resonators that model the vocal tract transfer function are connected in parallel, figure 7. Each resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak (formant) for both voiced and unvoiced sounds [7];

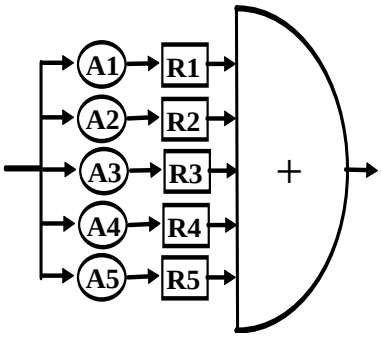


Figure 7 - Parallel Configuration. Adapted from Klatt [3]

- Cascade configuration - the voiced sounds are synthesised using a set of cascade resonators figure 8. The output of a resonator is fed into the input of the next.

Figure 8 - Cascade Configuration. Adapted from Klatt[3].

With the cascade configuration we can obtain the relative amplitudes of the desired formant peaks without individual control of each formant amplitude. Nevertheless it is still necessary to produce fricatives and plosives using the parallel configuration. Both cascade and parallel branches are used to produce speech, thus the synthesiser's overall structure is more complex, including both configurations (figure 10).

The basic unit of the system is the resonator: a structure able to model a spectral peak. It is characterised by the resonant (formant) frequency and respective bandwidth. A diagram of the second order resonator (with a pair of

complex conjugate poles) is shown on figure 9. The input and output signals are related by:

$y(n)=Ax(n)+By(n-1)+Cy(n-2)$

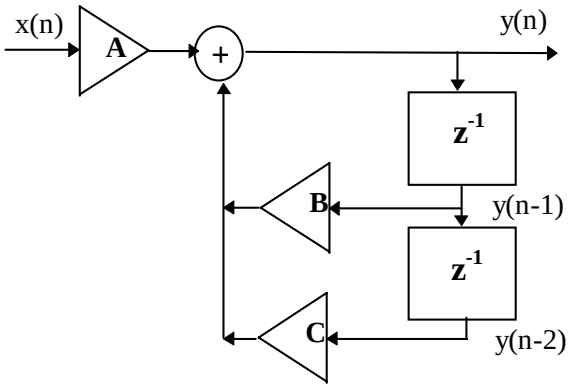


Figure 9 - Resonator block diagram.

The constants *A*, *B* and *C* are related with the parameters resonant frequency *F* and bandwidth B_w through the following set of equations:

$$C = -e^{-2\pi B_w T}$$
$$B = 2e^{-2\pi B_w T}\cos(2\pi FT)$$
$$A = 1 - C - B$$

where *T*=sampling period=1/sampling frequency

When the resonant frequency *F*=0 we obtain a low-pass filter with a -12dB/octave slope and a 3dB frequency at B_w /2. This resonator is used to model the natural glottal impulse reproduced by the synthesiser's voicing source.

Antiresonators are also used, introducing a pair of complex conjugate zeros (antiresonances or antiformants) to model the voicing source spectrum and to reproduce the effects of nasalization in the cascade configuration.

The antiresonator output *y(n)* is related to the input *x(n)* through the equation:

$y(n)=A'x(n)+B'x(n-1)+C'x(n-2)$

The constants *A'*, *B'* and *C'* are defined through the equations: *C'=-CA*, *B'=-B/A*, *A'=1/A*, where *A*, *B* and *C* are obtained substituting the antiresonance central frequency *F* and the antiresonance bandwidth *Bw* in the resonator equations.

The overall synthesiser block diagram is shown in figure 10. Each resonator is represented by the prefix r and the amplitude control is represented by the prefix a. Each resonator rn has a resonant frequency control parameter fn and a resonant bandwidth control parameter bn.

III. SYNTHESIS PARAMETERS

The synthesis parameters (symbol, name and range of values in Table1) define the output characteristics. The synthesis parameters are the characteristics of each of the 5 formants (amplitude, frequency and bandwidth) used to model the speech spectral properties, the pitch frequency, nasal characteristics, and overall gain. We have also other synthesis parameters like the LF model ones, the sampling frequency and frame length. All these parameters may vary within the range stated on figure 10 and its actual value may be edited using the graphical interface.

Table 1 - Synthesis parameters

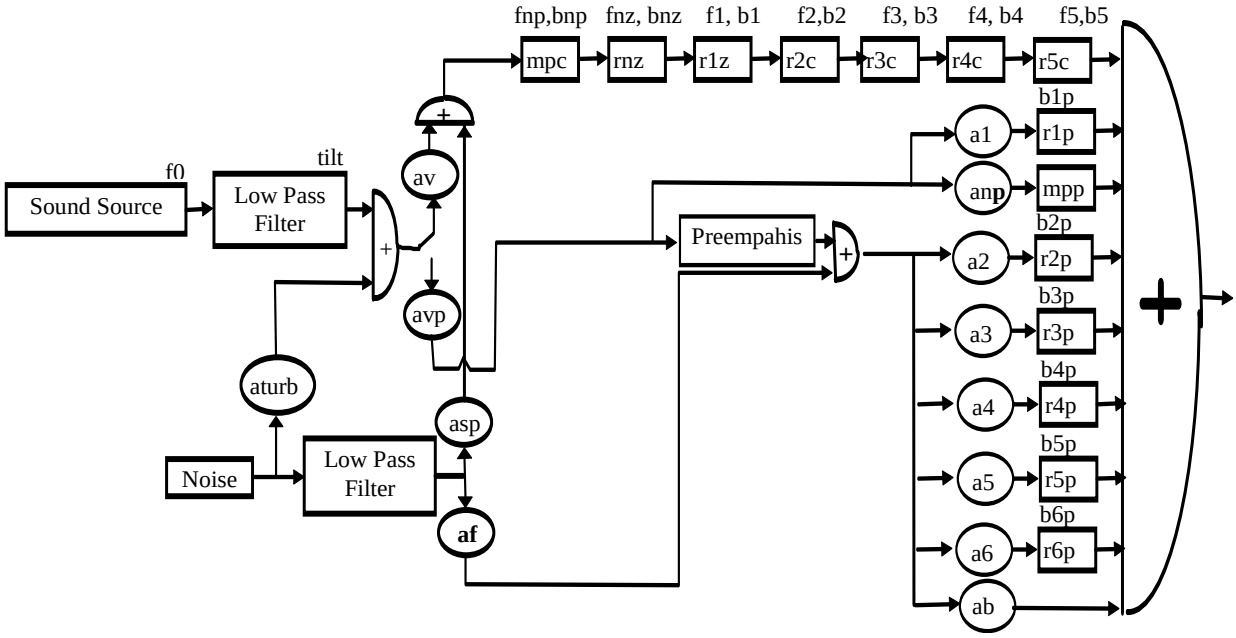


Figure 10 - Synthesiser block diagram rnpc - nasal resonator (cascade branch), rnz - nasal antiresonator (cascade branch), r1c, r2c, r3c, r4c and r5c - resonator (cascade branch), rnpp - nasal resonator (parallel branch), r1p, r2p, r3p, r4p, r5p and r6p - resonator (parallel branch)

Symbol	Name	Mín.	Máx.
f0 (Hz)	Fundamental frequency (pitch).	0	500
av (dB)	Amplitude control: voicing (cascade branch). Vowel → 60dB.	0	80
f (Hz)	1st formant frequency	150	1300
b (Hz)	1st formant bandwidth (cascade branch)	30	1000
f2 (Hz)	2nd formant frequency.	500	3000
b2 (Hz)	2nd formant bandwidth (cascade branch).	40	1000
f3 (Hz)	3rd formant frequency.	1200	4800
b3 (Hz)	3rd formant bandwidth (cascade branch).	40	1000
f4 (Hz)	4th formant frequency.	2400	4990
b4 (Hz)	4th formant bandwidth (cascade branch).	100	1000
f5 (Hz)	5th formant frequency.	3000	4990
b5 (Hz)	5th formant bandwidth (cascade branch).	100	1500
f6 (Hz)	6th formant frequency.	3000	4990
b6 (Hz)	6th formant bandwidth (cascade branch).	100	4000
fnz (Hz)	Nasal zero frequency (cascade branch).	180	000
bnz (Hz)	Nasal zero bandwidth (cascade branch).	40	1000
fnp (Hz)	Nasal pole frequency.	180	500
bnp (Hz)	Nasal pole bandwidth	40	1000
asp (dB)	Amplitude control: aspiration.	0	80
aturb (dB)	Amplitude control: turbulent. Breathy voice → 40dB	0	80
tilt (dB)	Spectral tilt (down) at 3KHz. Low frequency emphasys.	0	41
af (dB)	Amplitude control: fricative (parallel branch).	0	80
b1p (Hz)	1st formant Bandwidth (parallel branch).	40	1000
b2p (Hz)	2nd formant Bandwidth (parallel branch).	40	1000
b3p (Hz)	3rd formant Bandwidth (parallel branch).	60	1000
b4p (Hz)	4th formant Bandwidth (parallel branch).	100	1000
b5p (Hz)	5th formant Bandwidth (parallel branch).	100	1500
b6p (Hz)	6th formant Bandwidth (parallel branch).	100	4000
ai (dB)	ith formant Amplitude control (parallel branch)., i=1,...6	0	80
anp (dB)	Amplitude control: nasal (parallel branch).	0	80
ab (dB)	Amplitude control: fricative ("by-pass").	0	80
avp (dB)	Amplitude control: voiced (parallel branch).	0	80
gain (dB)	Overall gain. Unit gain → 60dB	0	80

c	parallel - cascade/parallel Configuration.		
n	Number of resonators (cascade branch).	1	6
s (Hz)	Sampling frequency.	5000	10000
f (ms)	Frame length	5	5000
v	Voiced source: impulse train - LF model.		

4. SYNTHESIS STRATEGY AND PARAMETERS

The glottal source is used in the synthesis of oral vowels (/i/, /e/, /E/, /a/, /6/, /o/, /O/, /u/ and /@/), nasal vowels, semivowels and diphthongs [4]. This source model generates glottal pulses that resemble the air volume velocity pulses produced by the vibration of vocal folds when the air circulates from lungs to the pharynx. The shape of these glottal pulses, controlled by the glottal source parameters, determine the voice characteristics (modal, aspired, …).

The synthesis parameter values depend on the segmental characteristics and are intimately related to the prosodic characteristics. High quality speech synthesis requires precise specification of parameters and careful control of the synthesis process.

When synthesising unvoiced sounds a white noise excitation source is used. It models the turbulent flow produced by the circulation of air through an extremely narrow constriction or even an occlusion of the vocal tract, generating a fricative source (/f/, /s/, /S/,…). The opening of a vocal tract occlusion produces an air turbulence (fricative source) followed by an uniform air flow through the open glottis (aspired source).To synthesise sounds like voiced fricatives and voiced plosives it is necessary to used a mixed source including both glottal pulses and noise modelling turbulence resulting from constrictions or occlusions of the vocal tract

In this work the to synthesise portuguese vowels we used the parameters of Table 2 that were obtained by inverse filtering method described by Teixeira [8]. For an isolated vowel f0 decreases linearly from 130Hz to 100Hz. The amplitude control av is reduced progressively close to the end of the utterance.

Table 2 - Formant frequencies for portuguese vowels

	f1 (Hz)	f2 (Hz)	f3 (Hz)
/i/ "vir"	225	2100	2750
/e/ "pêra"	390	1850	2503
/E/ "leve"	651	1629	2580
/a/ "cara"	714	1528	2425
/6/ "canto"	680	1688	2470
/o/ "dor"	453	946	2553
/O/ "corda"	588	1070	2365
/u/ "cume"	311	953	2070
/@/ "pequenina"	179	1610	2413

5. RESULTS

On figure 11 we show the graphical interface that enables the user to edit the synthesis parameters and display the synthesiser output waveform on time and frequency domains.

On figure 12 we present the result of synthesising the portuguese vowel /E/. Its audition resulted in a correct identification by the individuals submitted to perceptual tests.

The clear differences at high frequency result from the use of only 3 formants according to Table 2.

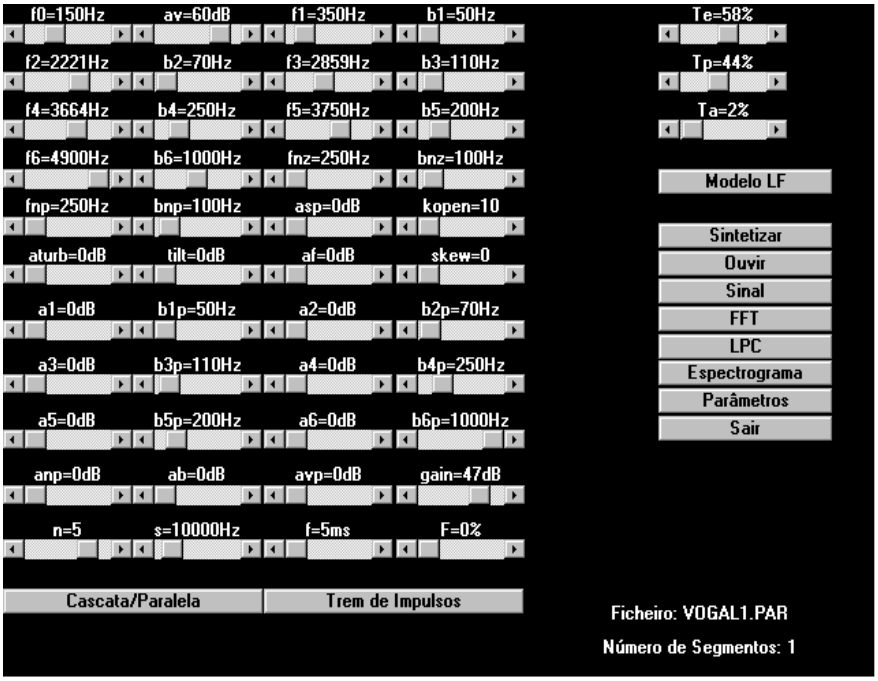


Figure 11 - Graphical interface.

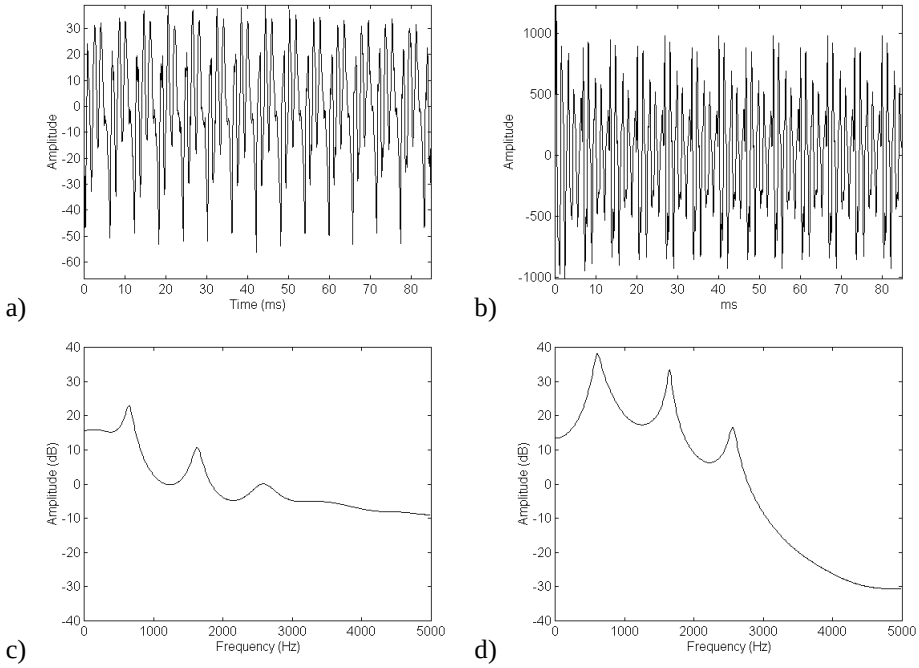


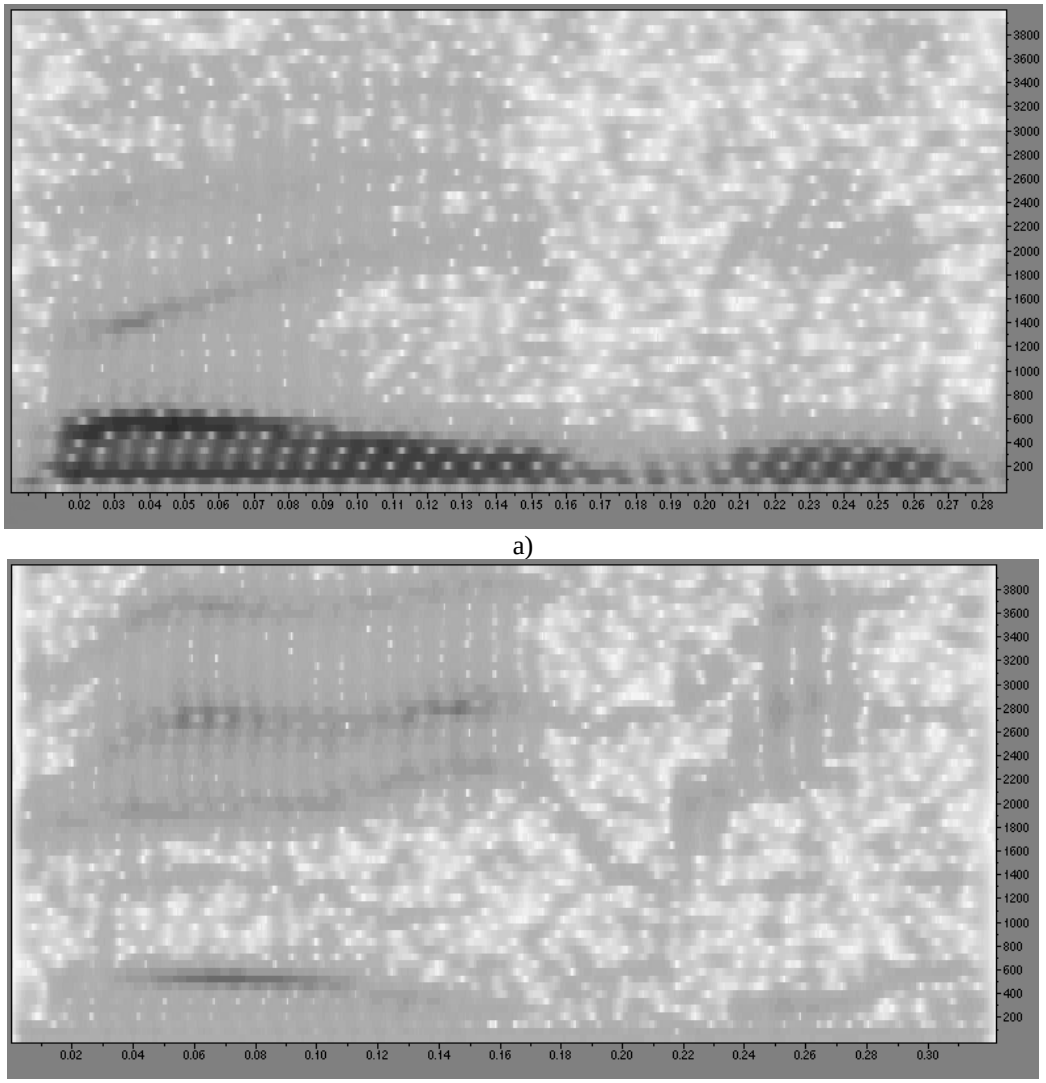
Figure 12 - /E/ “leve”, a) natural b)synthesised. (duration = 85 ms). Power spectrum estimation with an 16 order LPC c)natural d)synthesised

On figure 13 we show the natural and synthesised spectrogram of the english word “baby” pronounced by an english speaking male. The synthesis parameters were generated using a high level package for english text-to-speech conversion.

IV. CONCLUSIONS

The implemented system is easy to use and it showed to be an helpful tool when used on a basic experimental phonetics course, allowing students to observe the differences on the waveform or on the sound if the formant characteristics were changed.

The package was developed using MATLAB. Although this option seems to be the right one for the prototype



a)

b)

Figure 13 - /beIbi/ “Baby”,
a) Spectrotrogram of the natural word. Duration 286ms
b) Spectrogram of the synthesised word. Duration 322ms Sampling Frequency = 8KHz.

development, due to the intrinsic features of MATLAB the system is slow. Further work should be done to speed up the system, using a compiled version of MATLAB or another programming language.

V. REFERENCES

[1] Fant, G.: Acoustic Theory of Speech Production, Mouton, 1960.

[2] Fant, G., Liljencrants, J., Lin, Q. : A Four-Parameter Model of Glottal Flow, Speech Transmission Laboratory - Quarterly Progress and Status Report - Royal Institute of Technology - Stockholm - Sweden, 4, 1985, 1-13, 1985

[3] Klatt, D. H.: Software for a Cascade/Parallel Formant Synthesiser, The Journal of the Acoustical Society of America, 67(3), Mar. 1980, 971-995, 1980.

[4] Klatt, D.H., Klatt, L . C.: Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers, The Journal of the Acoustical Society of America, 87(2), February 1990, 820-857, 1990

[5] Holmes, J. N.: The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesiser, IEEE Transactions on Audio and Electroacoustics, AU21(3), June 1973, 298-305, 1973

[6] Holmes, J. N.: 1983, Research Report - Formant Synthesisers: Cascade or Parallel?, Speech Communication, 2(4), 251-273., 1983

[7] Holmes, W. J., Holmes, J. N., Judd, M.W.: Extension of the Bandwidth of the JSRU Parallel-Formant Synthesizer for High Quality Synthesis of Male and Female Speech, IEEE International Conference on Acoustics, Speech, and Signal Processing 90 Proceedings - Albuquerque - New Mexico - USA, 1, 313-316, 1990

[8] Teixeira, A. J.: 1995, Current Research (Internal), Universidade de Aveiro.