TOOLS AND BASQUE LANGUAGE DATABASES DEVELOPED IN THE AHOLAB LABORATORY

Borja Etxebarria (*), Eva Navas, Ana Armenta, Imanol Madariaga, Iñaki Gaminde, Inma Hernaez

borja, eva, ana, imanol, inma@bips.bi.ehu.es
University of the Basque Country
(*) Borja Etxebarria presently at TIBCO

Abstract

This paper gives an overview of the speech material used and generated in our laboratory (AhoLab), as well as of the software tools developed for its management. The databases were created in the context of the development of a text to speech converter for Basque, in the different fields of research. The software here described is freely available and is currently being used by some educational centre and individuals.

1 Introduction

The euskara (Basque language), despite of being one of the oldest languages in Europe and being subject of several studies, is full of linguistic unknowns, including it's place of origin. Nowadays it is spoken by about 800.000 people: 600.000 of them in the Basque Country and the rest of them in many other areas around the world, most of them in America.

Having this limited number of speakers, Basque language presents a huge dialectal fragmentation. Seven main dialects and more than 20 sub-dialects are considered. In modern times, written Basque has been standardised, but the pronunciation of the standard Basque has been left aside, so when talking speakers may adapt pronunciation (including intonation) to their native variety, or to the dominant language. This poses some problems when developing technological tools for this language. The choice of the speaker becomes a crucial point as he or she should be able to apply his or her native pronunciation rules to sentences that have been syntactically and morphologically standardised.

2 Data bases

All of the data bases described below are in Basque Language, and have been recorded for many different purposes. We have named the databases according to the name of the speaker or to the dialectal variety.

The main characteristics of all the databases are summarised in table 1.

- *julen-units*

  This database was created to obtain the speech units required for concatenative synthesis. The recording was made under laboratory conditions using a Minidisk and a professional Shure microphone. It was digitised at 16 kHz with 16 bits per sample.

  The speaker was chosen mainly because of the *neutral* characteristics of his native variety: reduced number of assimilation, palatalisation, vowels and consonants deletion and insertion and so on. Besides, in this variety, the rules to place accent in words are quite close to the most recent proposal to normalise accentuation in Standard Basque [1] which has been followed by our automatic transcriptor. On the other hand the speaker was able to control the position of the accented syllable when there was no coincidence of the 'automatic' accent insertion rules and 'true' rules. This helps automatic segmentation and labelling process of the database.

The corpus consists of 1758 isolated sentences chosen to contain every possible previously defined synthesis unit (polyphones of various sizes). The sentences were selected from a great variety of texts taken from electronic publications. Most of them are declarative sentences as shown in table 2. Sentences length goes from 8 syllables up to 57, with 1, 2 or 3 intermediate pauses.

The sentences were automatically labelled at phoneme level using speech recognition, and the phones forming the unit were manually reviewed. For this reason the corpus has also been used to study phoneme duration. Besides, part of it has been used to study intonation of declarative sentences.

This database belongs to Telefónica and we are allowed to use this database only for research purposes[1].

Data Base	**MB**	**Speaker**	**Domain**	**Type**	**Record**	**#sentences**	**#words**
Julen-units	213	male	diphone	read	laboratory	1758	12351
Julen-intonation	146	male	intonation	read	laboratory	1096	11136
Ion	39,5	male	diphone	read	laboratory	-	1403
Amaia	25,7	female	diphone	read	laboratory	-	1228
Juanjo	29,5	male	diphone	read	laboratory	-	1228
Euskadi-irratia	804,8	male	news	read & spontaneous	radio	1633	9776
Bermeo	7,5	female	intonation	read	home	348	1376
Basque-varieties	240	male & female	intonation	read	home	120 * 25	-

Table 1: Basque databases of AhoLab.

- *julen-intonation*

  This is a set of sentences specifically designed for the analysis of intonation, so every thinkable syntactic structure was considered (table 2 shows some details on sentence distribution).

	julen-units	**julen-intonation**
# Declarative	1677	674
# Question	69	236
# Exclamation	12	186
# Clauses without pause	2381	2254
# Words	12351	11136

Table 2: Sentence distribution on julen-unit and julen-intonation database.

[1] Julen-units and Julen-intonation were created in the context of a project developed for Telefónica 1997-1998.

Most of the sentences were picked from Internet, but some of them were obtained from other non-electronic sources. They are just orthographically labelled at the sentence level, and a small part of them has been used for intonation research purposes.

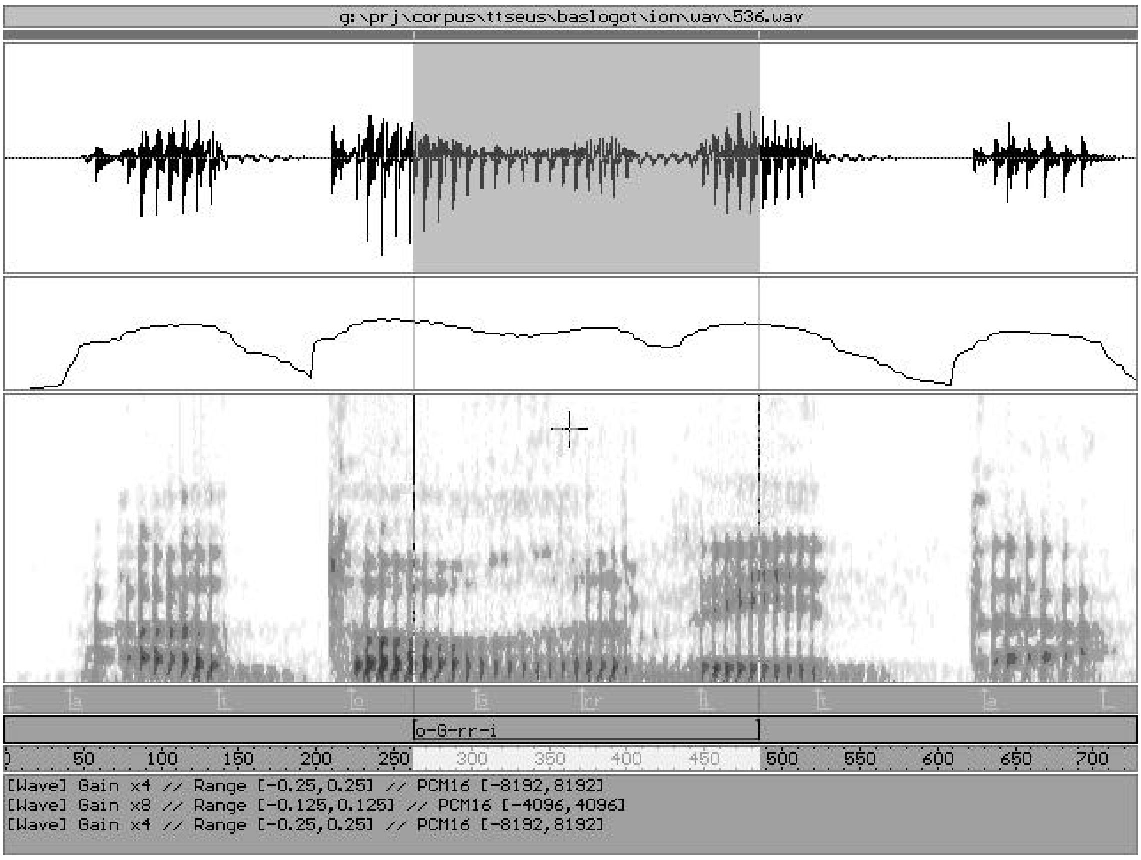
We are allowed to use it only for research purposes, for the same reason explained in the case of julen-units.

- *ion, amaia, juanjo*

  These three databases are composed of special nonsense words built to get a synthesis unit inventory in an adequate articulatory context.

  All of them were recorded under laboratory conditions, using a professional Shure microphone and a Minidisk and digitised at 16 kHz, 16 bits per sample.

  Figure 1 shows one of those nonsense words, where the segment corresponding to the synthesis unit obtained from the signal is highlighted.



```
[Wave] Gain x4 // Range [-0.25,0.25] // PCM16 [-8192,8192]
[Wave] Gain x8 // Range [-0.125,0.125] // PCM16 [-4096,4096]
[Wave] Gain x4 // Range [-0.25,0.25] // PCM16 [-8192,8192]
```

Figure 1: Tetraphoneme from Ion database.

  Each one of the databases has about 1300 isolated words uttered by the same speaker, one female and two males. All of the signals have been automatically segmented and labelled into phones using DTW (Dynamic Time Warping)[2][3] with synthetic speech, and the time marks belonging to synthesis units and their included phones have been re-examined.

- *euskadi-irratia*

  The recording consists of about 8 hours of radio news containing read and spontaneous speech. The recording was done with the purpose of investigating the positions that were preferred by the speaker to insert a pause or breath break when it is not indicated by an orthographic sign. The texts that the radio speaker was reading during the program were given to us so that orthographic transcription was fully available for the speaker of interest and partially available for the others.

This database contains recordings in standard and in non-standard Basque. Every speaker has been identified, and the part of the signal uttered by each of them has been marked (and labelled with the speaker's identifie r, but not with the text) over the signal.

The segments corresponding to one of the speakers (75' of speech) have been further studied by labelling the pauses and the morphology of the text. The morphological labels were obtained with an automatic morphological parser [4]. A total of 2662 pauses have been labelled over the sentences uttered by this speaker, without making any classification among them. The pauses insertion and deletion model is presently under development.

The distribution of sentences recorded and transcribed is showed in table 3.

	Total	**Speaker studied**
# Declarative	1470	860
# Question	113	113
# Exclamation	50	46
# Words	16875	9766

Table 3 Content of euskadi-irratia database.

- *bermeo*

It is composed by about 300 short sentences uttered by a female speaker in her native variety. The database was created for the purpose of intonation studies and specifically for the analysis of the influence of the position of the focus and the lexically stressed words in the overall intonation curve [5].

It includes 121 short declarative sentences, 129 declarative sentences including lexically accented words and 98 question sentences, composed by 2, 3 or 4 short phrases formed with words of 2,3 or 4 syllables. Focus position and the position of lexically accented and unaccented words were combined in different manners. They were recorded on a minidisk in a silent environment at the speaker's home.

This database is currently being used to prove the validity of Fujisaki's intonation model for Basque [6]. All the sentences have been manually segmented into phonemes, words and phrases. Their F0 curves have been calculated using a method based in [7] and the parameters of Fujisaki's model have been obtained both manually and automatically for all of them. The manual segmentation was done with the help of a graphic tool specifically designed for that task and which will be described later.

- *basque-varieties*

A corpus of over 120 sentences for 25 Basque dialectal varieties, with declarative, interrogative, negative, and other syntactic structures was designed with the purpose of studying the influence of focus position in the sentences over different dialectal varieties of the south region of the Basque speaking area. With this goal in mind, the speakers were selected considering mainly the fact that their speech would not be influenced by the dominant language (Spanish). The recordings were made at each speaker's home, and some of them are of low quality due to inadequacy of the room.

From the corpus, 40 declarative sentences were selected for each variety and labelled with the F0 value calculated in the centre of each syllable [8]. This time instant was automatically obtained by dynamic time warping projection.

3 Software tools

For computing purposes, the most common processing algorithms developed over the last years have been grouped into a C library for DOS and Unix environments.

To manage the speech signals a special graphic tool called AhoT has been developed. AhoFuj, is another graphic tool created to make easier the process of modelling Basque intonation using Fujisaki's model.

3.1 AhoLib

This is a C-library containing the most common signal processing algorithms, and some miscellaneous utilities that provide us with an easy way for accessing files and managing user interfaces. We will briefly describe the purpose of most relevant facilities.

- *SPL module (Signal Processing Library)*

  The sources are organised in various sub-modules, dedicated to different purposes, such as:

  - mathematics: Bessel functions, matrix management and equation solving.

  - spectral analysis: FFT and DFT computation, LPC analysis and synthesis by various methods [9] [10], LSF/LSP analysis[12], windowing...

  - other: convolution, correlation and covariance computation, power, zero crossing, preemphasis/deemphasis filtering, random number generation, linear phase filter implementation, window and frame management...

- *Miscellaneous utilities*

  - Memory management: the purpose is to provide the programmer with a tool to easily find memory leaks and other bugs related to memory.

  - Display, keyboard and mouse management.

  - Program argument management: set of functions that will allow the final user to introduce configuration parameters by means of an interface text file.

  - Files and file header management functions adapted to our file format with mark and label handling facilities.

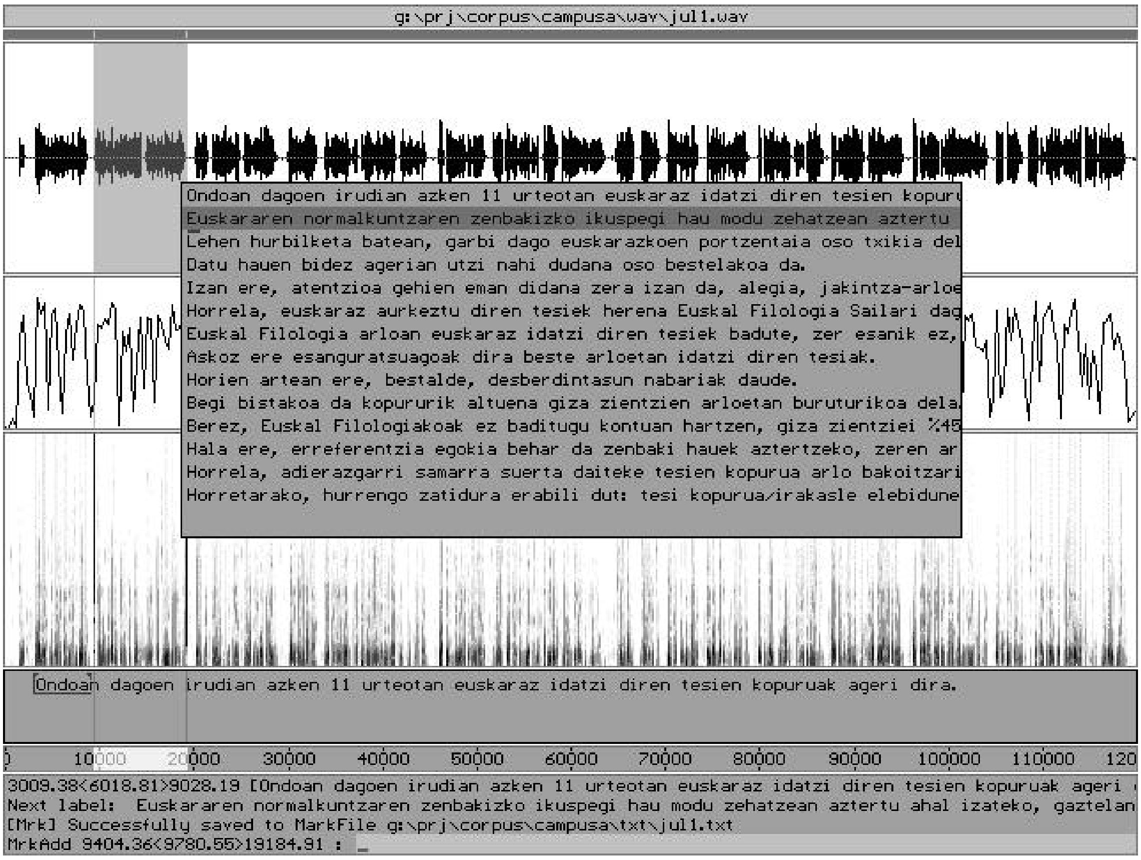
  - Unit conversion (ms, hertz, loghertz...), single and double list management.

3.2 AhoT*ools*

AhoT is a speech signal editor providing zoom and movement facilities over the displayed signals or any of their representations. Besides our own *aho* format, standard signal formats are considered. Signals and time intervals can be played over different sound cards, and the program runs under MSDOS, Windows9x and Linux.

AhoT does some basic calculations by itself. Although the program can show up to eight signal waveforms, calculations are always done over the first signal displayed. Spectral and other representations for the rest of the signals are possible through parameter files containing previously calculated data.

The program provides management of marks and text labels, which can be inserted, deleted and modified over the signal up to five levels. Labels can be any text string (see figure 3), and are specified together with a time interval or time instant in a text file. One label file always corresponds to one waveform file. To make easier the insertion of labels they can be taken from a text file, so the user has not to write them. This is particularly useful when labelling many consecutive segments of a signal from a previously known text; this can be seen in figure 2. AhoT provides with facilities to move from mark to mark in many different ways, to sort them, to select them and to search for particular labels.



```
g:\prj\corpus\campusa\wav\jul1.wav
```

```
Ondoan dagoen irudian azken 11 urteotan euskaraz idatzi diren tesien kopuru
Euskararen normalkuntzaren zenbakizko ikuspegi hau modu zehatzean aztertu
Lehen hurbilketa batean, garbi dago euskarazkoen portzentaia oso txikia del
Datu hauen bidez agerian utzi nahi dudana oso bestelakoa da.
Izan ere, atentzioa gehien eman didana zera izan da, alegia, jakintza-arloe
Horrela, euskaraz aurkeztu diren tesiek herena Euskal Filologia Sailari dag
Euskal Filologia arloan euskaraz idatzi diren tesiek badute, zer esanik ez,
Askoz ere esanguratsuagoak dira beste arloetan idatzi diren tesiak.
Horien artean ere, bestalde, desberdintasun nabariak daude.
Begi bistakoa da kopururik altuena giza zientzien arloetan buruturikoa dela.
Berez, Euskal Filologiakoak ez baditugu kontuan hartzen, giza zientziei %45
Hala ere, erreferentzia egokia behar da zenbaki hauek aztertzeko, zeren ar
Horrela, adierazgarri samarra suerta daiteke tesien kopurua arlo bakoitzari
Horretarako, hurrengo zatidura erabili dut: tesi kopurua/irakasle elebidune
```

```
Ondoan dagoen irudian azken 11 urteotan euskaraz idatzi diren tesien kopuruak ageri dira.
```

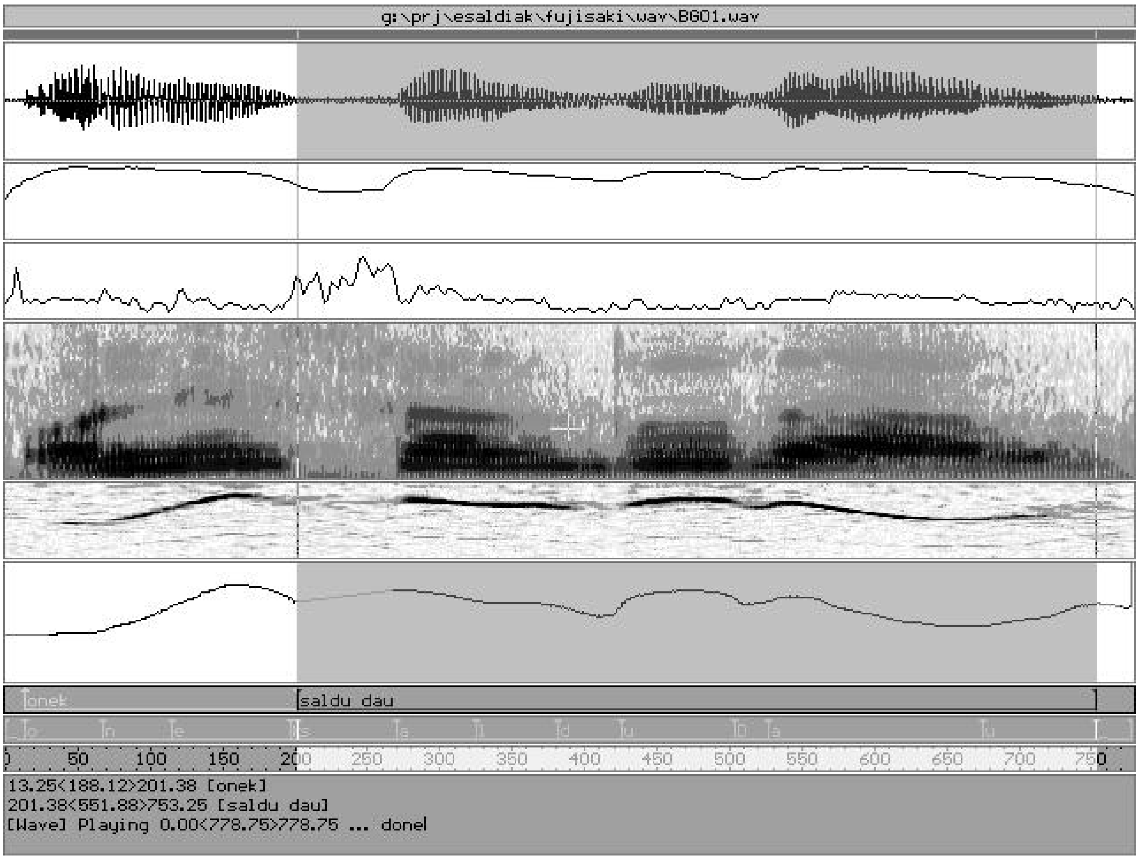
```
3009.38<6018.81>9028.19 [Ondoan dagoen irudian azken 11 urteotan euskaraz idatzi diren tesien kopuruak ageri
Next label:  Euskararen normalkuntzaren zenbakizko ikuspegi hau modu zehatzean aztertu ahal izateko, gaztelan
[Mrk] Successfully saved to MarkFile g:\prj\corpus\campusa\txt\jul1.txt
MrkAdd 9404.36<9780.55>19184.91 : _
```

Figure 2: Label insertion from text file.

The program can be fully customised using a configuration text file, where interface and label files used, as well as all the parameters for the calculations, are defined. Names for the label files are assumed to be the same as the ones of the signal file if no other name is provided, so just the path and file extension must be indicated.

For every desired calculation a set of parameters can be configured. In general, some displaying options for every window can be selected, as well as analysis window length and type, frame length, number of points when displaying FFTs, order of the LPC analysis, and some specific options for the cepstral calculations. The calculations performed by AhoT are: signal spectrogram, pitch representation by means of cepstral coefficients, power and number of zero crossing by time unit.

Figure 3 shows a typical AhoT display appearance with the waveform, the power of the signal, the zero crossings, the spectrogram, the pitch obtained with the cepstral coefficients and the pitch calculated outside AhoT with the Griffin algorithm.



```
13.25<188.12>201.38 [onek]
201.38<551.88>753.25 [saldu dau]
[Wave] Playing 0.00<778.75>778.75 ... done
```

Figure 3: AhoT displaying a speech signal from Bermeo database. From top to bottom, waveform, power, zero crossings, spectrogram, cepstrum-pitch and Griffin pitch, phrase labels, phone labels, time axis. The lower lines are the user dialogue area.

3.3 AhoFuj labeller

The AhoFuj is a graphic tool designed to dynamically represent the F0 curve calculated from the Fujisaki's model parameters [15] and acoustical evaluation of the results (see Figure 4).

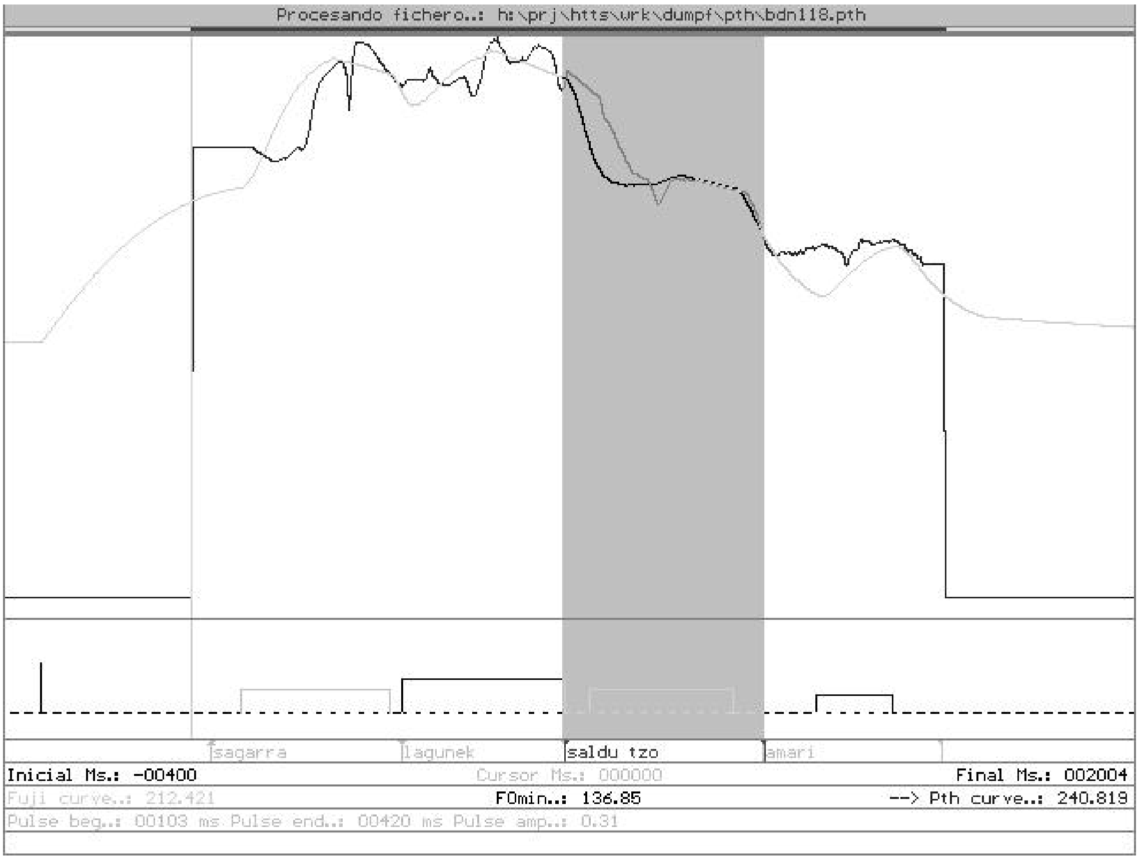
The parameters of the model can be modified graphically using the keyboard and saved to a text file (from where they can be loaded later). The effects of the modifications of the parameter values on the final synthetic pitch curve are seen on the fly, which permits a very fast manual adjustment of the model to the real pitch curve.

A vocoder has been incorporated to the program, so that original and synthetic pitch can be exchanged in the coded signal and both coded signals can be played. This permits an immediate perceptual evaluation of the parameters being edited.

Labels and time marks can be displayed in the same way as in AhoT, to help the user to fulfil linguistic constraints if convenient (i.e. it may happen that some Fujisaki parameters must correspond to particular linguistic items). Also like in AhoT, configuration of the program (status and display options, file location, Fujisaki's model constants, frame rate of F0 synthetic curve) is made trough a text file interface.

Figure 4 shows a typical AhoFuj session. In the upper window synthetic and natural pitch curves are shown. In the lower window, from top to bottom we can see the pulses and impulses of the Fujisaki model, phrase labels, position of the cursor in time axis, the same for F0 axis (plus selection of natural or synthetic curve and F0min value), selected pulse or impulse information. Finally one line to dialogue with the user.

The program presently runs only under MSDOS and it is being ported to Windows9x and Linux.



```
Procesando fichero..: h:\prj\htts\wrk\dumpf\pth\bdn118.pth
```

```
Inicial Ms.: -00400          Cursor Ms.: 000000              Final Ms.: 002004
Fuji curve..: 212.421        F0min..: 136.85           --> Pth curve..: 240.819
Pulse beg..: 00103 ms Pulse end..: 00420 ms Pulse amp..: 0.31
```

Figure 4: Natural (brown) and synthetic (light) pitch curves of one of the signals from Bermeo database. See text for an explanation.

4 Conclusions and future work

Most of the material describe in this work has been created in the laboratory to make possible the development of technological applications for the Basque, and specifically for text to speech conversion. Part of this material is not available to the public, as it was obtained with private company financial help, and the companies are reluctant to share it.

This work does not present all the material available for Basque, in any way. We know of many recordings which have been made with non-technical interests and that we consider very valuable, but that have not been processed (labelled to any level, sometimes in analog support) thus they can not be used efficiently. Our final goal referring Basque language oral databases is to develop a standard and homogeneous system able to accept material of non-homogenous characteristics and sources. To be more precise, we would like that system to be used by anyone interested in Basque language, being that interest social, linguistic or technical. With that in mind, we have been compiling every available recording no matter the quality, and now we are focussing efforts on the Bizkaian variety, after having got the interest of some official organisation of the region.

Part of the work presented here in this paper has received project financial help fro m the UPV/EHU, which we would like to acknowledge (UPVEHU-147.345-TA066/98).

Finally, we would like to thank all the students that are digitising, labelling and segmenting unselfishnessly, to Iñaki Gaminde for the heaps of hours of recordings, and to all the speakers for the boring recording sessions.

5 Bibliography

[1] Hualde, J. I.
Euskal azentuak eta euskara batua
Euskaltzaindia. Euskera XXXIX, 1549-1568, 1994 (in Basque)

[2] H. Sakoe, S. Chiba
A dynamic programming approach to continuous speech recognition
Proc. Int. Congr. Acoust. Budapest, Hungary, Rep 20-C-13. 1971

[3] L.R. Rabiner / R.W Schafer
Digital Processing of Speech Signals
Prentice Hall, 1978

[4] http://ixa.si.ehu.es/ingeles/dokument/MORFEUS.html

[5] G. Elordieta, I. Gaminde, I. Hernáez, J. Salaberria, I. Matín de Vidales
Another step in the modelling of Basque intonation: Bermeo
Lecture Notes in Computer Science; Vol 1692: Lecture Notes in Artificial Intelligence pp 361-364, 1999

[6] E. Navas, I. Hernáez, A. Armenta, B. Etxebarria, J. Salaberria
Modelling Basque intonation using Fujisaki's model and CARTs
State of the Art in Speech Synthesis, London, April 2000 [to appear]

[7] D. Griffin, J. S. Lim
Multiband excitation vocoder
IEEE Trans. ASSP. Vol 36, N 8. August 1988

[8] I. Hernáez, I. Gaminde, B. Etxebarria, P. Etxeberria
Intonation modelling for the southern dialects of the Basque language
Eurospeech'97. Rhodes, Greece. pp 807-810. 1997

[9] J. Makhoul
Linear Prediction: A Tutorial Review
Proc. IEEE vol 63, pp. 561-580, Apr. 1975

[10] Delsarte, Y. V. Genin
The Split Levinson Algorithm
IEEE Trans. ASSP, Vol 34, N. 3, Jun. 1986

[11] K.K Paliwal, B.S. Atal
Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame
IEEE Trans. Speech, and Audio Processing, Vol 1, N.1 Jan 1993

[12] F.K. Soong B.H. Juang
Optimal Quantization of LSP Parameters
IEEE Trans. Speech, and Audio Processing, Vol 1, N.1 Jan 1993

[13] B. S. Atal, M. R. Schroeder
Predictive Coding of Speech Signals and Subjective error criteria
IEEE Trans. ASSP, vol 27, N.3, Jun. 1979

[14] B. S. Atal
Predictive Coding of Speech at Low bit Rates
IEEE Trans. Comm. Vol COM-30, N.4, April 1982

[15] H. Fujisaki, K. Hirose
Analysis of voice fundamental frequency contours for declarative sentences of Japanese
Journal of Acoustic Society. Jpn. (E) 5, 4. 1984