

WINPITCH CORPUS, A SOFTWARE TOOL FOR ALIGNMENT AND ANALYSIS OF LARGE CORPORA

Philippe Martin

University Paris 7 Denis Diderot

philippe.martin@fnac.net

1. INTRODUCTION

Description of endangered languages normally starts with the collection of speech data, which are then segmented into various phonological, prosodic, morphological and syntactic units. In this process, the (phonetic) transcription is the most critical part, and user friendly tools are essential to tackle any sizeable work in a reasonable amount of time.

The software program WinPitch Corpus addresses these concerns directly, allowing two modes of operation to handle the data. In the first mode, text is not available and is generated by the user speech segment by speech segment (as it was the case when only analog tape recorders were available). In the second mode, speech has already been transcribed into text, but the text units are not aligned, i.e. a bi-univocal relationship between units of text and units of speech has not been established.

Although some existing software programs operate in the first mode, establishing implicit text and speech alignment in the process, few allow operations in commonly found (difficult) recording conditions such as voice overlapping or presence of noise. This paper introduces briefly some of the important features of WinPitch Corpus, as an efficient tool for transcription and analysis of speech data: slower speech rate for easier transcription, dynamic adjustment of segments with simultaneous display of spectrograms for precise alignment, etc.

Numerous speech analysis tools (fundamental frequency tracker, spectrogram, LPC formant analysis, etc.) are available with a quasi instantaneously display of the results. Support for the simultaneous acoustical analysis of both channels of stereo recordings is also provided.

The program has already been extensively used for analysis of large romance languages corpora of spontaneous speech (more than 1.200.000 words, C-ORAL-ROM, 2003), as well as for the phonetic and phonological description of Parkatêjê, an endangered language of the Amazon spoken by about 300 people (Araújo and Martin, 2003). WinPitch Corpus is available from the www.winpitch.com web site, under the name WinPitchPro.

1. TRANSCRIPTION AND TEXT TO SPEECH ALIGNMENT

Besides direct transcription of speech chunks, various methods based on acoustical analysis have been widely used in the past.

1.1 Spectrographic alignment

All experimental phonetic courses contain a chapter on prosodic curves and spectrogram “reading”, to train learners to position accurately the limits between speech sounds. These limits are of course approximate, as they have to segment the continuous movement from one articulatory position to another. Automatic segmentation can be made with various degrees of success, by relying on the sudden spectral transitions (Cosi, 1997). As with many automated processes in speech analysis, its liability depends on the signal properties to correspond to the implied working hypotheses of the method (the main one being to have only one source of sound, i.e. no noise). Recording made outside sound proof rooms or/and with more than one speaker are usually not very reliable.

1.2 Automatic alignment with Hidden Markov Model

Another automatic or semi-automatic alignment approach is based on speech recognition algorithms (frequently based on parameters training of statistical hidden Markov models HMM). This appears as a sub problem of the general speech recognition process, as the final result of recognition is already known (with the phonetic or orthographic transcription). The limits of speech sound are then obtained from the phonetic transcription, directly known or indirectly (Talin and Wightman 1994, Fohr, Mari, et Haton 1996).

Unfortunately, these systems generally reach an error level of 15 to 20%, and require good quality recordings as well as speech characteristics sufficiently close to the speaker’s characteristics which are in practice rarely known in advance.

1.3 Automatic alignment by synthesis

Yet another automatic text to speech alignment proceeds by comparison between the time variations of the speech signal spectra with another speech signal, generated by a text to speech synthesizer operating on the text to align. (Malfrère and Dutoit, 2000). The advantage of this method stems from the fact that it is easier to align dynamically spectra of two sentences corresponding to the same text than to segment on the base of sequential speech sound recognition.

The limits of this approach are similar to the HMM based systems: background noise and non typical voice characteristics limit its use to standard voice styles recorded in low noise conditions.

1.4 Limits of automatic alignment

Automatic alignment based on recognition algorithms or speech synthesis suffers from 3 main problems:

- a) Presence of an important noise level in the recordings;
- b) Overlapping of speakers voices;
- c) Use of old recordings, with poor frequency response (typically filtered below 300Hz).

For all these reasons, the use of a human operator seems inevitable. The problem is then to facilitate the manual segmentation work (done with a simple tape recorder in the heroic times) at the level of the phrase or syntactic group, and provide adequate tools for fine tune segmentation at the level of the syllable or the phoneme.

2. COMPUTER ASSISTED ALIGNMENT

Experimental studies showed that visual and gesture correlation between text and the sound could be made if the speech sound was slowed down by at least a factor of 30%, depending on the size of the units selected. The principle of continuous assisted alignment is based on this observation. The text to be aligned (in the second mode of transcription) is displayed on a window while the corresponding sound is played back at a slower speed, dynamically adjustable. At each identification by the operator of a sound unit (which can be a syllable, a word, a syntagms or a whole sentence), the operator clicks with the computer mouse on the corresponding part of text. The program then records the time position of the selected part of text, and continuously builds an alignment database. Various tools available in the program allow for easy step back, verification of alignment sections, etc. In the first mode, slow down speech rate is used for easier perception by the operator of the successive speech segments.

2.1 *Slow playback engine*

The slow down playback engine uses a modified version of the PSOLA algorithm (Moulines et Charpentier, 1990) allowing the re-synthesis of natural speech with high quality. Since the performance of this method relies heavily on the quality of F₀ detection for precise pitch period marking, the spectral comb method (Martin, 1980) used in the program allows the reduction of up to 7 times real time. The implemented PSOLA synthesizer operates in streaming mode, allowing handling of large sound files without computer memory constraints.

2.2 *Alignment fine tuning*

Once the assisted alignment is done, the program displays automatically the text under the corresponding speech segments. The user can then adjust precisely the limits of each segment by clicking on its edges. The analysis window then displays the corresponding spectrogram and prosodic curves for precise fine tuning of the segmentation. The operator can take care of overlapping voices often found in spontaneous recordings by assigning different layers to different speakers.

3. CONCLUSION

Computer assisted text to speech alignment as implemented in WinPitch Corpus generates labels attached to units of text, corresponding to labels attached to units of speech sound, so that sound can be obtained from text, and conversely. The alignment operation is done by clicking on the text units (syllables, words, syntagms, phrases), while they are perceived at playback slower speed (second mode), or by entering the text corresponding to each segment (first mode). The slowing down of speech allows for easier comprehension of segments and for the necessary psychometrical coordination for alignment (second mode). In this case, alignment can be done in one pass and does not require any expertise in phonetics.

This process appears to be much faster than traditional manual methods, where a trained phonetician has to align the sound units one by one by moving each time an analysis window along the speech signal. It appears as well more reliable as emerging

automatic methods, which require very good recording conditions and exclude large variance in speaker pronunciation characteristics.

After alignment, WinPitch Corpus allows detailed acoustical analysis of segmented units. Easy to use numerous functions - illustrated below - allow for precise segmentation, as well as the display of spectrograms, fundamental frequency, intensity and wave form. The user can edit or enter new text on the fly, in any language supported by Unicode fonts. Phonetic transcription, syntactic labeling and other information can be easily added on one of the eight available transcription tiers.

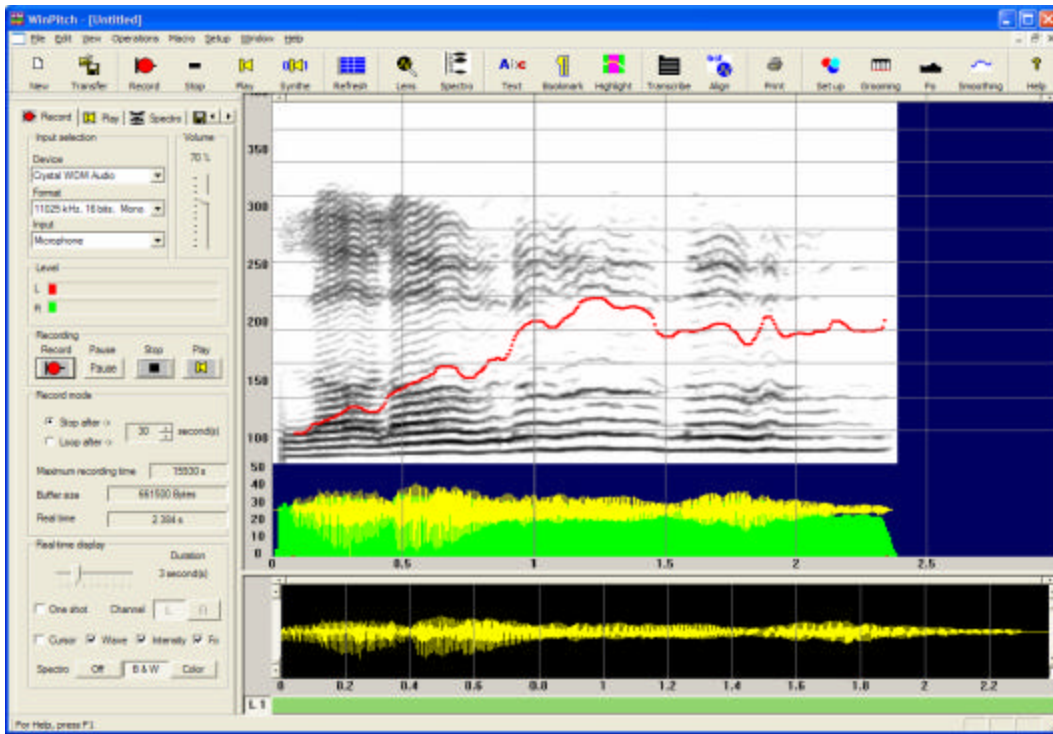


Figure 1: Real time recording and analysis.

WinPitch Corpus allows real time recording and playback with simultaneous display of a spectrogram, Fo, intensity and wave curves.

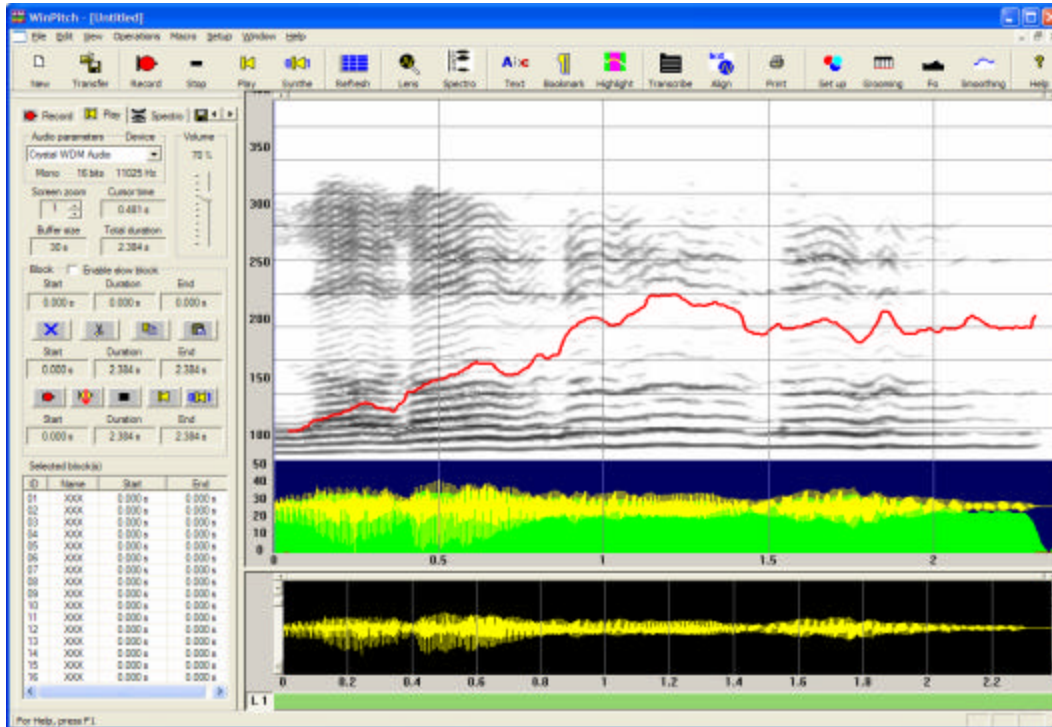


Figure 2: Signal analysis: spectrogram, waveform, intensity and fundamental frequency curves .

Playback analysis mode allows the quasi instantaneous display of spectrogram, F_0 , intensity and wave form curves. Three zooming modes are available, by expanding a navigation window (bottom right), selecting a block in the navigation window, or by creating a virtual expanding analysis screen (up to 10 times the physical horizontal size of the display).

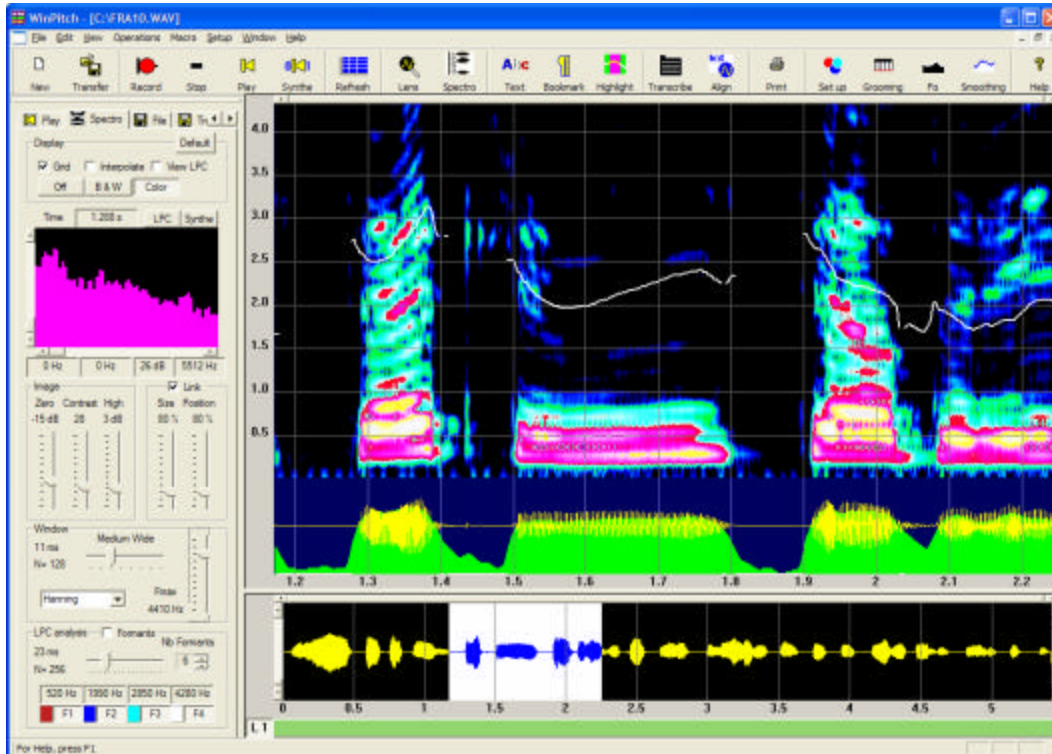


Figure 3: Wide band color spectrogram.

The spectrogram dialog box allows easy expansion of the frequency and time scales, narrow or wide band selection, instantaneous display of spectral cross section at any cursor position. Formant frequencies obtained through LPC analysis are also displayed.

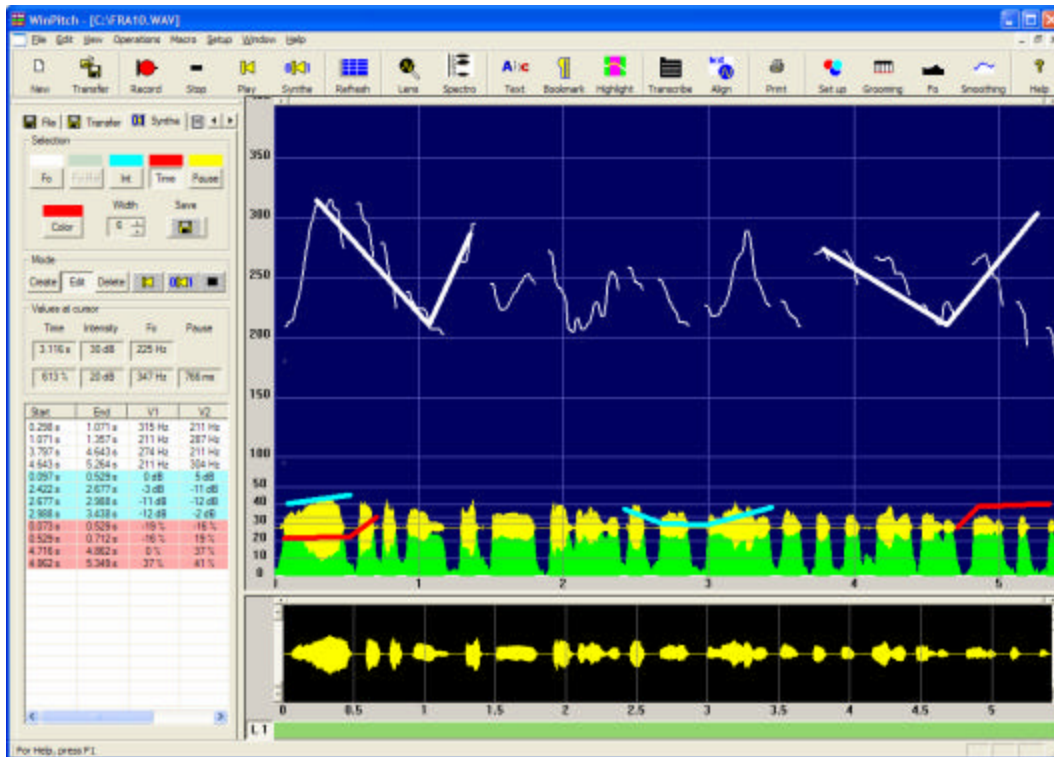


Figure 4: Prosodic morphing: piecewise linear segments define Fo, intensity, duration and pause.

Intuitive graphic interface allows precise layout on screen of linear piecewise segments defining the evolution of Fo, intensity, time and pause values for prosodic morphing of recorded speech through PSOLA synthesis. Synthetic prosodic values are displayed while new curves are positioned by the user.

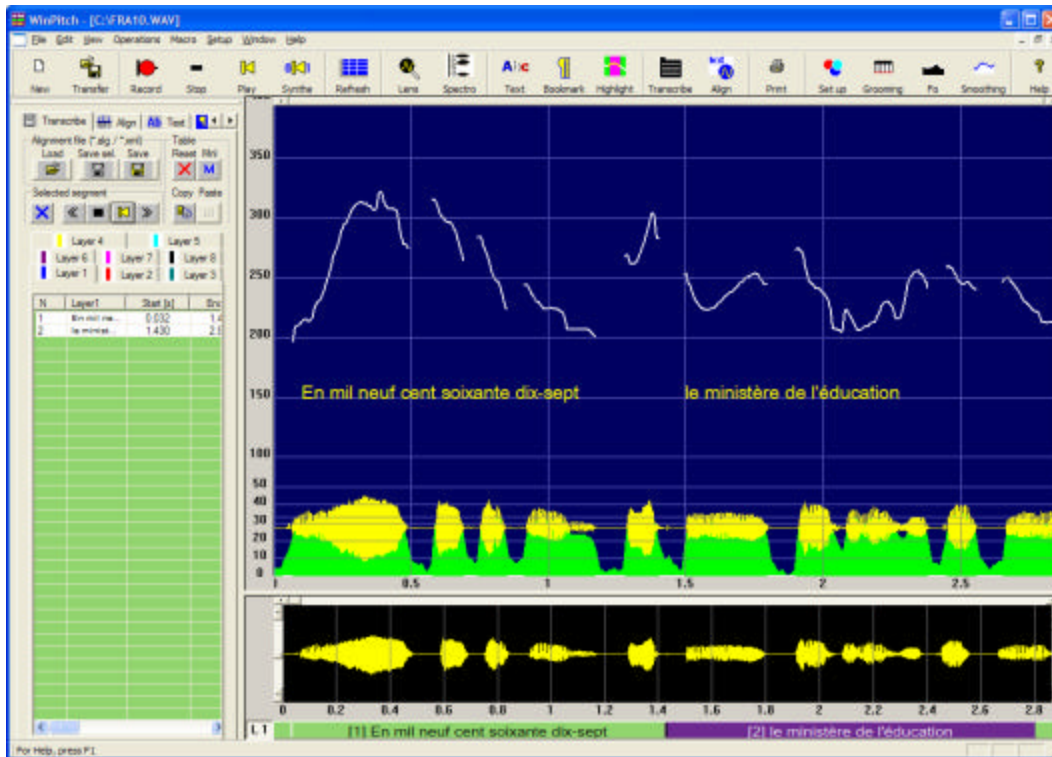


Figure 5: Speech direct labeling and segmentation.

Text to speech alignment can be done in two modes. In the first mode, text does not exist, and the user selects blocks of speech (which can be slowed down for playback), and enters the corresponding text (any UNICODE font can be used directly). In this process, a database is automatically built, which can be later saved in XML or Excel® formats.

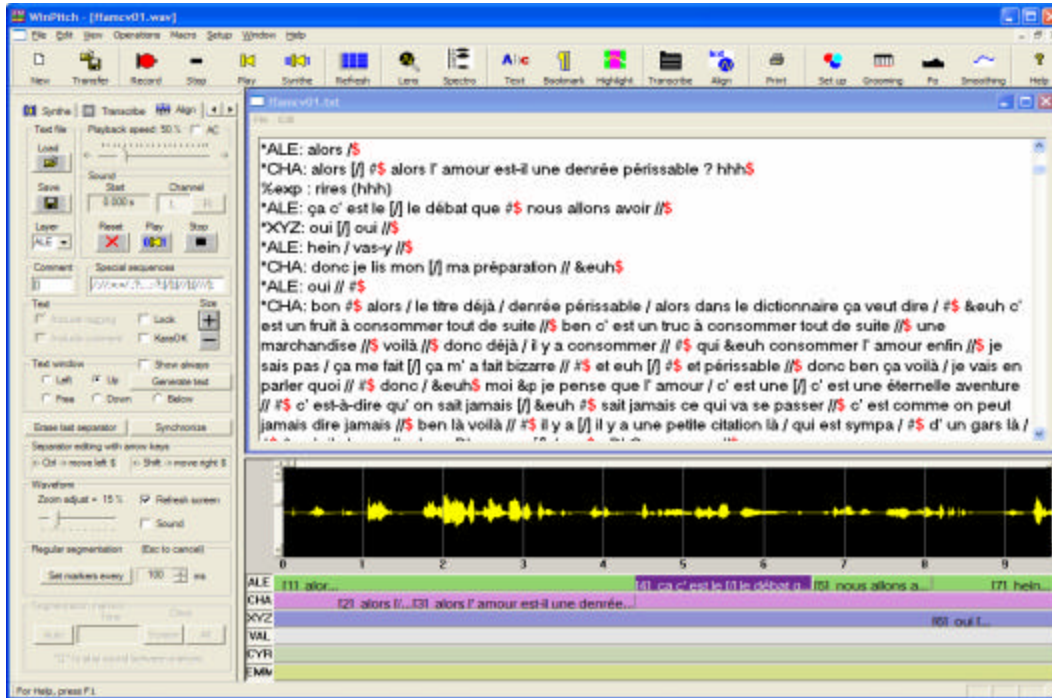


Figure 6: Assisted alignment.

The second mode of text to speech alignment implies a preexisting text. The speech sound is then played back at a reduced speed (dynamically programmable) while the user clicks on the part of text corresponding to the perceived sound unit. A database of the dynamically defined segments is automatically built (table in the dialog box on the left).

```

<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE Alignment SYSTEM "alignment.dtd">
<alignment>
<TimeStamp Value="Tuesday, June 11, 2002 time 01h 25m 11s"/>
<MetaPitch Program="Aligner" Version="1.0"/>
<Track version="1.0" creationDate="Tuesday, June 11, 2002 time 01h 25m 11s" audioFileName="C:\ESTR.MOV" textFileName="Colonna1.txt"/>
<Layer1 Name="EST" ID="EST" Short="EST" Color="RGB(145,233,120)"/>
<Layer2 Name="CLA" ID="CLA" Short="CLA" Color="RGB(213,145,220)"/>
<Layer3 Name="Text" ID="Text" Short="Text" Color="RGB(145,145,233)"/>
<Layer4 Name="Title" ID="speaker" Short="Tit" Color="RGB(228,228,228)"/>
<Layer5 Name="Text line" ID="type" Short="Spe" Color="RGB(200,213,180)"/>
<Layer6 Name="Background" ID="background" Short="Bak" Color="RGB(123,222,140)"/>
<Layer7 Name="Episode" ID="episode" Short="Epi" Color="RGB(120,195,200)"/>
<Layer8 Name="Comment" ID="comment" Short="Com" Color="RGB(255,228,255)"/>
<UNIT speaker="EST" startTime="0.000" endTime="1.366" channel="R">o viera / da /</UNIT>
<UNIT speaker="CLA" startTime="1.366" endTime="3.307" channel="R">a patre /</UNIT>
<UNIT speaker="EST" startTime="3.307" endTime="4.838" channel="R">no /</UNIT>
<UNIT speaker="EST" startTime="4.838" endTime="5.976" channel="R">ascolta / qui sopra /</UNIT>
<UNIT speaker="EST" startTime="5.976" endTime="6.951" channel="R">si /</UNIT>
<UNIT speaker="CLA" startTime="6.951" endTime="8.400" channel="R">qui ? si /</UNIT>
<UNIT speaker="EST" startTime="8.400" endTime="9.821" channel="R">sopra /</UNIT>
<UNIT speaker="CLA" startTime="9.821" endTime="11.880" channel="R">si /</UNIT>
<UNIT speaker="EST" startTime="11.880" endTime="13.221" channel="R">si / leva leva / vai /</UNIT>
<UNIT speaker="EST" startTime="13.221" endTime="14.627" channel="R">stenta di sportarti /</UNIT>
<UNIT speaker="CLA" startTime="14.627" endTime="17.333" channel="R">vai /</UNIT>
<UNIT speaker="EST" startTime="17.333" endTime="18.117" channel="R">si /</UNIT>
<UNIT speaker="EST" startTime="18.117" endTime="20.845" channel="R">ora ti fo un po' di male /</UNIT>
</Alignment>

```

Figure 7: Assisted alignment. Output.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Speaker	Text	Start	End	Channel										
2															
3	EST	o viera / da /	0	1.366	M										
4	CLA	a patre /	1.366	3.307	M										
5	EST	no /	3.307	4.838	M										
6	EST	ascolta / qui sopra /	4.838	5.976	M										
7	EST	si /	5.976	6.951	M										
8	CLA	qui ? si /	6.951	8.4	M										
9	EST	sopra /	8.4	9.821	M										
10	CLA	si / vai / aspetta / rite lo levo il cabinò ?	9.821	11.88	M										
11	EST	si / leva leva / vai / te lo sporti /	11.88	13.221	M										
12	EST	stenta di sportarti /	13.221	14.627	M										
13	CLA	vai / oodi basta /	14.627	17.333	M										
14	EST	si /	17.333	18.117	M										
15	EST	ora ti fo un po' di male	18.117	20.845	M										
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															
28															
29															
30															
31															
32															
33															
34															
35															
36															
37															
38															
39															

Figure 8: Assisted alignment: direct output in Excel®.

Aligned text is saved either under a proprietary format (alg), or standard XML format, allowing easy interchange with other programs. Data are also directly transfer to Excel® in a one step operation.

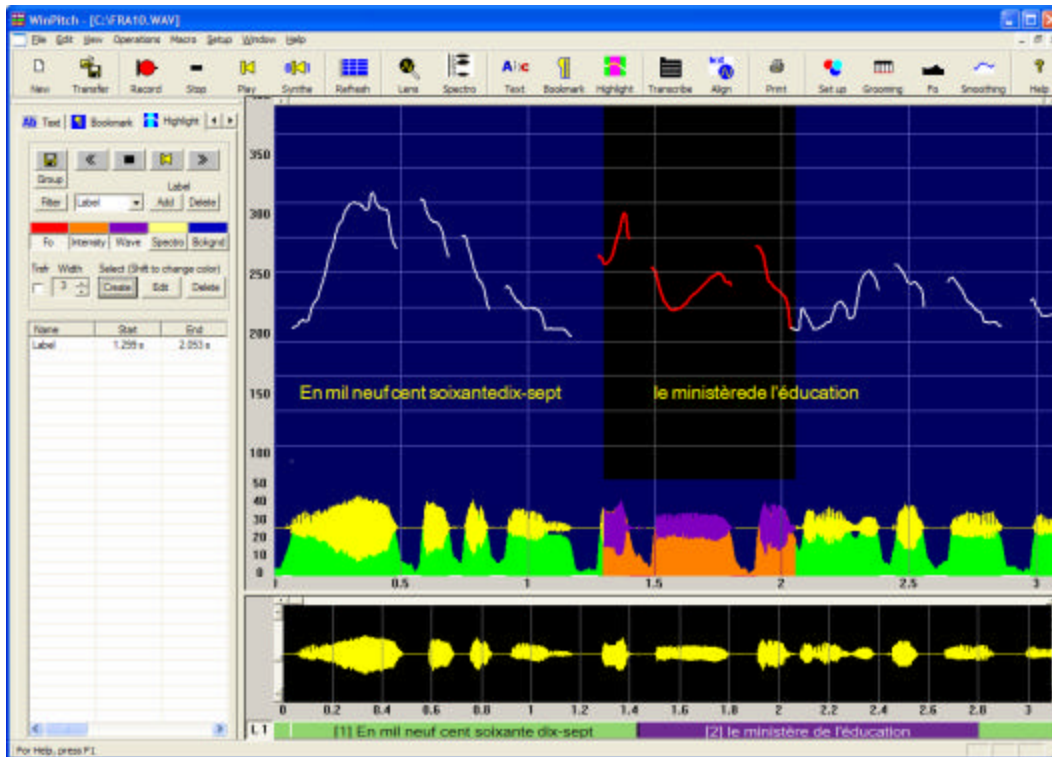


Figure 9: Speech segment highlighting.

Sections of the speech wave can be highlighted and tagged, allowing the definition of specific sections (such as stressed syllables, unvoiced consonants, etc.) for automatic statistical analysis.

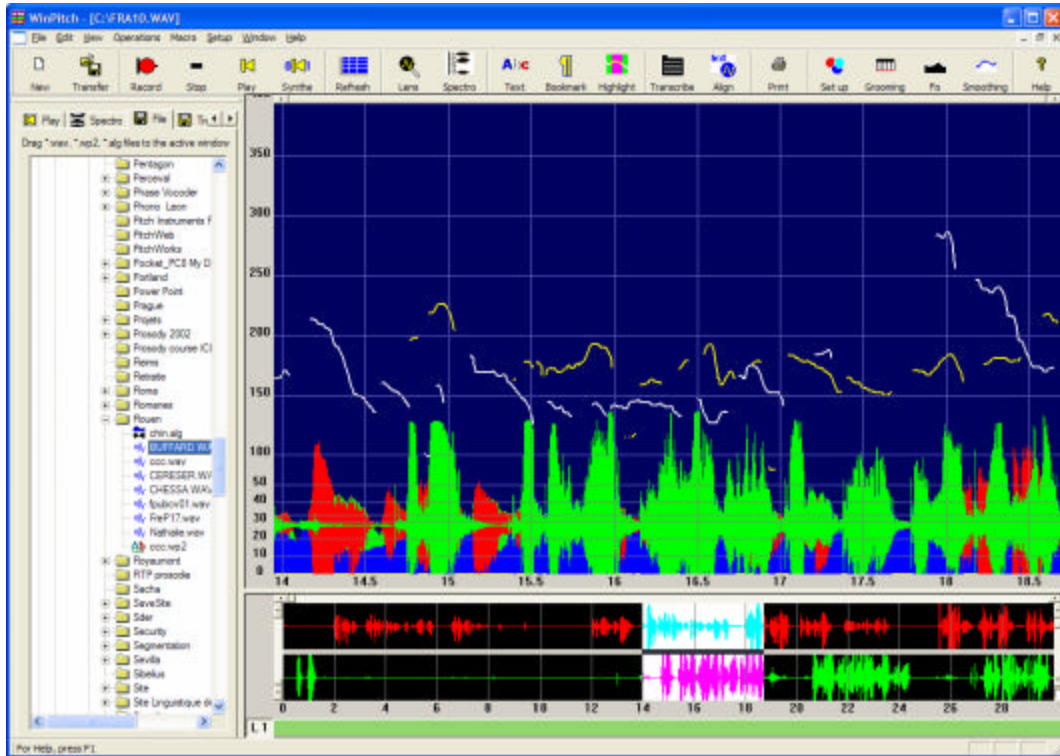


Figure 10: Two channels analysis, with simultaneous display of waveforms, Fo and intensity curves.

Prosodic analysis of both channels of stereo signals can be simultaneously displayed on the analysis window (top right): left channel gives a Fo curve in yellow, and the right channel in white.

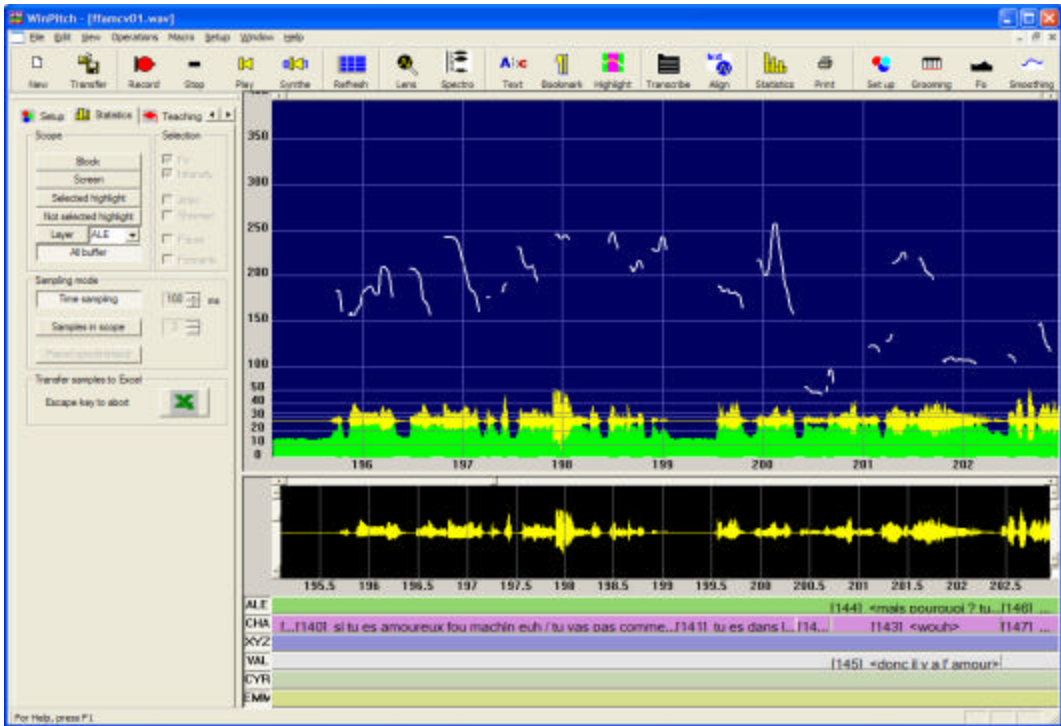


Figure 11: Statistical sampling, with direct output to Excel®

A table in Excel® format can be easily created containing values of F_0 , intensity, jitter, shimmer, and formant frequencies and amplitudes, allowing the use of all functions available under Excel to evaluate statistical parameters. Selection of data include block, screen, highlighted sections, not highlighted sections, segmentation layers and whole buffer.

	A	B	C	D	E	F	G	H	I	J	K	L
	[Rank]	Sample	Time [s]	Type	F Zero [Hz]	Intensity [dB]	Jitter [%]	Shimmer [%]	Formant1 [Hz]	Ampl F 1 [dB]	Formant2 [Hz]	Ampl F 2 [dB]
3	0	2207	0.1	ALE	0	21	0	0	0	0	0	0
4	0	4412	0.2	ALE	0	20	0	0	0	0	0	0
5	0	6617	0.3	ALE	0	20	0	0	0	0	0	0
6	0	8822	0.4	ALE	89	29	0	0	0	0	0	0
7	0	11027	0.5	ALE	81	21	0	0	0	0	0	0
8	0	13232	0.6	ALE	0	20	0	0	0	0	0	0
9	1	99227	4.5	ALE	130	30	0	0	0	0	0	0
10	1	101432	4.6	ALE	0	21	0	0	0	0	0	0
11	1	103637	4.7	ALE	105	26	0	0	0	0	0	0
12	1	105842	4.8	ALE	99	27	0	0	0	0	0	0
13	1	108047	4.9	ALE	90	27	0	0	0	0	0	0
14	1	110252	5	ALE	89	23	0	0	0	0	0	0
15	1	112457	5.1	ALE	0	20	0	0	0	0	0	0
16	1	114662	5.2	ALE	104	27	0	0	0	0	0	0
17	1	116867	5.3	ALE	116	30	0	0	0	0	0	0
18	1	119072	5.4	ALE	108	23	0	0	0	0	0	0
19	1	121277	5.5	ALE	134	30	0	0	0	0	0	0
20	1	123482	5.6	ALE	0	19	0	0	0	0	0	0
21	1	125687	5.7	ALE	0	24	0	0	0	0	0	0
22	1	127892	5.8	ALE	102	28	0	0	0	0	0	0
23	1	130097	5.9	ALE	98	24	0	0	0	0	0	0
24	1	132302	6	ALE	94	23	0	0	0	0	0	0
25	1	134507	6.1	ALE	93	22	0	0	0	0	0	0
26	1	136712	6.2	ALE	0	20	0	0	0	0	0	0
27	1	138917	6.3	ALE	0	17	0	0	0	0	0	0
28	1	141122	6.4	ALE	0	16	0	0	0	0	0	0
29	1	143327	6.5	ALE	0	15	0	0	0	0	0	0
30	1	145532	6.6	ALE	0	14	0	0	0	0	0	0

Figure 12: Statistical sampling, with direct output to Excel®

BIBLIOGRAPHY

Araújo, L. and Martin, Ph. (2003) "Accent de mot et intonation en Parkatêjê, une langue timbira de l'Amazonie Brésilienne", *Actes du colloque Interfaces Prosodiques*, Nantes 27-29 mars 2003.

C-ORAL-ROM (2003) <http://lablita.dit.unifi.it/coralrom>

Cosi, P. (1997) "SLAM v1.0 for Windows : a Simple PC-Based Tool for Segmentation and Labelling", *Proc. Of ICSPAT-97, Int. Conf. On Signal processing Applications and Technology*, San Diego, CA, Sept. 1997, 1714-1718.

Fohr, D., Mari, J.-F. et Haton, J.-P. (1996) "Utilisation des modèles de Markov pour l'étiquetage automatique et la reconnaissance de BREF80", *Actes des XXIèmes Journées d'Etude sur la Parole*, Avignon, 339-342.

Malfrère, F. et Dutoit, T. (2000) "Alignement automatique du texte sur la parole et extraction de caractéristiques prosodiques", in *Ressources et évaluation en ingénierie des langues*, Chibout, Mariani, Masson, Néel ed., De Boeck et Larcier, Paris, 541-552.

Moulines, E. & Charpentier, M. (1990) "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol 9, 453-467.

Talin, D. and Wightman, C.W. (1994) "The Aligner: Text-to-Speech Alignment using Markov Models and a Pronunciation Dictionary", *Second ESCA/IEEE Workshop on Speech Synthesis*, 89-92.

WinPitch (2003) <http://www.winpitch.com>