New functions for a multipurpose multimodal tool for phonetic and linguistic analysis of very large speech corpora

Philippe Martin
LLF UMR7110, UFRL, Université Paris Diderot
Place Paul Ricœur, 75013 Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

Abstract

The increased interest for linguistic analysis of spontaneous (i.e. non-prepared) speech from various points of view (semantic, syntactic, morphologic, phonologic and intonative) lead to the development of ever more sophisticated dedicated tools. Although the software Praat emerged as the de facto standard for the analysis of spoken data, its use for intonation studies is often felt as not optimal, notably for its limited capabilities in fundamental frequency tracking. This paper presents some of the recently implemented features of the software WinPitch, developed with the analysis of spontaneous speech in mind (and notably for the C-ORAL-ROM project 10 years ago). Among many features, WinPitch includes a set of multiple pitch tracking algorithms aimed to obtain reliable pitch curves in adverse recording conditions (echo, filtering, poor signal to noise ratio, etc.). Others functions of WinPitch incorporate an integrated concordancer, an on the fly text-sound aligner, and routines for EEG analysis.

Keywords: Spontaneous speech, fundamental frequency, concordancer

1. Introduction

A lot of interest is presently devoted to the linguistic analysis of non-prepared speech, and in particular to the prosodic correlates of syntactic and macrosyntactic units. WinPitch is a software program devoted to acoustic analysis of speech with, as its name suggests, specialized functions for research in prosody. It has been continuously developed since 1995 and runs under Windows (any flavor) on PC (and Mac with Windows emulator) personal computers. Many original functions allow effective acoustical analysis of large speech corpora, as demonstrated in its use in the C-ORAL-ROM project (2005), which assembled transcribed and aligned large spontaneous speech recordings dealing with similar topics in French, Italian, Spanish and European Portuguese (http://lablita.dit.unifi.it/coralrom/). Currently, WinPitch implements a number of new innovative functions and is intensively used in the project C-ORAL-ROM Brasil (2013).

2. Features

2.1 Sound recording made clear

Real-time spectrographic display is one of the unique features not found in other popular programs such as Transcriber (2013) or Praat [(2013). This function allows visual monitoring of speech recordings and is especially useful as speech corpora are rarely recorded by sound engineers, with the consequence that poor sound quality recordings is common (background noise, echo, wrong recording level, microphone filtering, etc.). Poorly recorded speech samples can make syllabic prominence and fundamental frequency analysis difficult or impossible. Since most of personal computers contain a sound card, it is very easy to implement a speech monitoring system by merely adding an appropriate microphone while running WinPitch, and visually check the recording quality with a minimal expertise in acoustic phonetics. This feature has proven very useful in field work.

2.2 Sound and video

WinPitch can handle directly long files either stored in the Ram or in Hard Disk memory, or with a sliding window (appropriate for a very large video files, above 6 GBytes, whose sound part exceeds the machine memory capacity). When played back, all the functions operate on the speech signal, displaying at the same time the synchronized video part. Furthermore, dedicated converters handle directly mp3 and CD sound files. Selecting a slower playback sound speed (implemented with a Psola or a phase vocoder engine) will always result in a synchronized corresponding video display.

2.3 Transcription and alignment on the fly

Aside from classical transcription tools (with automatic segmentation in short sections, automatic segmentation in syllables and phones and user defined variable playback speed), the software has a unique function allowing easy alignment of recording already transcribed but not aligned, as frequently found in on line or other corpora.

This function is especially useful in case of poorly recording examples, where automatic alignment is ineffective. It allows the user to click on any unit of text (whether on word, syntagm or the whole sentence) while the speech is played back at user selectable reduced speed (down to 7 times real-time through the use of a Psola or phase vocoder engine).

It allows an easy and close to real time alignment of already transcribed text even for data recorded in difficult conditions, since the difficult task of automatic speech recognition is transferred to the more efficient human recognition capabilities. The whole process precludes a time-consuming segment-by-segment alignment if the speech transcription is available but not aligned. Other

WinPitch modes of transcription include automatic segmentation based on silence or pause boundaries, where the user enters directly the corresponding text of predefined segments.

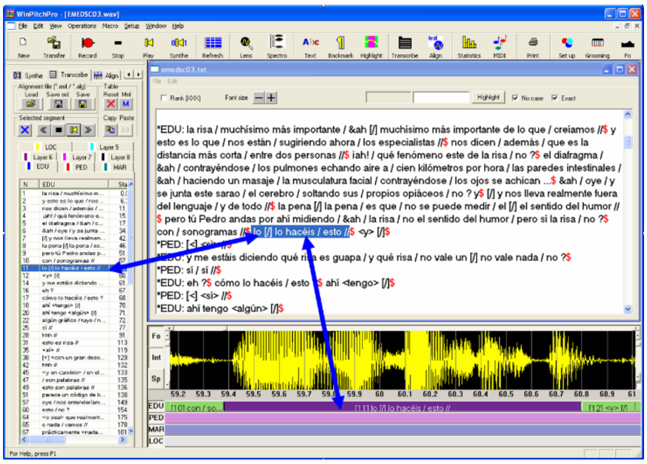

*EDU: la risa / muchísimo más importante / &ah [/] muchísimo más importante de lo que / creíamos //$ y esto es lo que / nos están / sugiriendo ahora / los especialistas //$ nos dicen / además / que es la distancia más corta / entre dos personas //$ ¡ah! / qué fenómeno este de la risa / no ?$ el diafragma / &ah / contrayéndose / los pulmones echando aire a / cien kilómetros por hora / las paredes intestinales / &ah / haciendo un masaje / la musculatura facial / contrayéndose / los ojos se achican ...$ &ah / oye / y se junta este sarao / el cerebro / soltando sus / propios opiáceos / no ? y$ [/] y nos lleva realmente fuera del lenguaje / y de todo //$ la pena [/] la pena / es que / no se puede medir / el [/] el sentido del humor //$ pero tú Pedro andas por ahí midiendo / &ah / la risa / no el sentido del humor / pero sí la risa / no ?$ con / sonogramas //$ lo [/] lo hacéis / esto //$ <y> [/]$
*PED: [<] <sí> //$
*EDU: y me estáis diciendo qué risa es guapa / y qué risa / no vale un [/] no vale nada / no ?$
*PED: sí / sí //$
*EDU: eh ?$ cómo lo hacéis / esto //$ ahí <tengo> [/]$
*PED: [<] <sí> //$
*EDU: ahí tengo <algún> [/]$

Figure 1. Assisted alignment by slowing down speech playback. At each mouse click on a unit of text perceived at slower speed (top right window), bidirectional pointers are generated automatically between the corresponding speech segment (bottom right window) and a database (left window).

The program also generates automatically an API transcription and morphological and syntactic labeling in French (Fig. 2), thanks to the use of a large lexicon, adapted from the Lexique3 (2009) project. Other dictionaries in various languages can be easily integrated in the system.

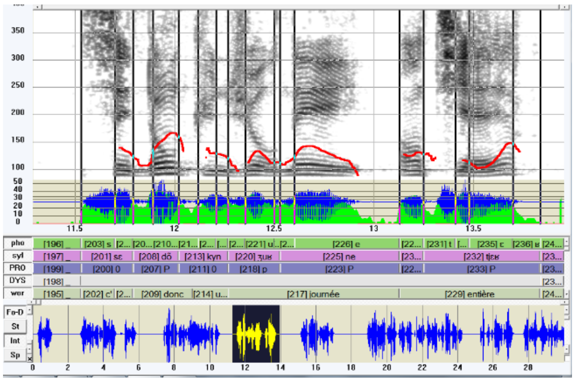
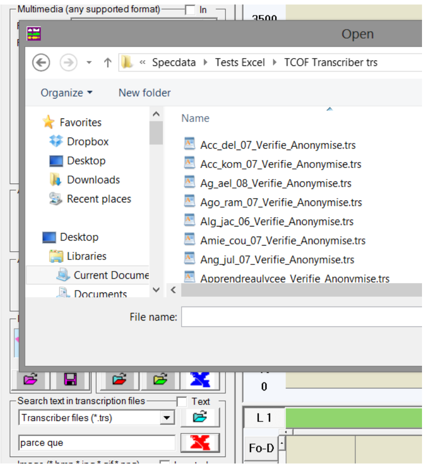


Figure 2. Automatic API transcription from orthographic text and morphological and syntactic labeling.

2.4 Integrated concordancer

WinPitch functions include an integrated concordancer. Figures 3 to 6 illustrate the details of the operations involved. In Fig. 3 the user enters the key word "*parce que*" taken as an example, selects an appropriate alignment source format (Transcriber *.trs in this example), and clicks on any of the file names stored in the same directory. This directory should contain all the alignment files of interest in the same format, together with their corresponding sound files (Six formats are currently available: Transcriber, Praat, CRF, Necte, WinPitch, XML). In the case of Praat textgrid files, the corresponding sound files must have the same name as their textgrid counterpart, as Praat textgrid files do not contain any reference to their corresponding speech file.


Open

« Specdata › Tests Excel › TCOF Transcriber trs

Acc_del_07_Verifie_Anonymise.trs
Acc_kom_07_Verifie_Anonymise.trs
Ag_ael_08_Verifie_Anonymise.trs
Ago_ram_07_Verifie_Anonymise.trs
Alg_jac_06_Verifie_Anonymise.trs
Amie_cou_07_Verifie_Anonymise.trs
Ang_jul_07_Verifie_Anonymise.trs
Apprendreaulycee_Verifie_Anonymise.trs

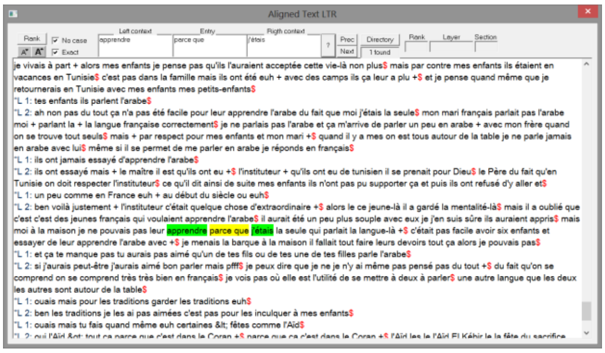
Search text in transcription files: Transcriber files (*.trs)
parce que

Figure 3. Entering the key word "*parce que*" and selecting a Transcriber files.

An Excel table listing all found occurrences of the key word is immediately generated (Fig. 4). This operation is very fast, in the example of *parce que*, the completion takes less than one second to scan 104 files giving 1194 occurrences.

Figure 4. Table generated automatically listing the occurrences of the entered keyword ("*parce que*"). The whole process takes less than 1 second for a list of 104 files and 1194 occurrences found.

When the user clicks on any line of the excel table, a specific occurrence of the keyword is selected together with its left and right contexts. The corresponding text and speech segments are then automatically displayed together with its corresponding spectrogram and fundamental frequency, intensity and duration curves, as shown in Fig 5 and Fig. 6.


Aligned Text LTR

Left context: apprendre — Entry: parce que — Rigth context: j'étais — 1 found

je vivais à part + alors mes enfants je pense pas qu'ils l'auraient acceptée cette vie-là non plus$ mais par contre mes enfants ils étaient en vacances en Tunisie$ c'est pas dans la famille mais ils ont été euh + avec des camps ils ça leur a plu +$ et je pense quand même que je retournerais en Tunisie avec mes enfants mes petits-enfants$
*L 1: tes enfants ils parlent l'arabe$
*L 2: ah non pas du tout ça n'a pas été facile pour leur apprendre l'arabe du fait que moi j'étais la seule$ mon mari français parlait pas l'arabe moi + parlant la + la langue française correctement$ je ne parlais pas l'arabe et ça m'arrive de parler un peu en arabe + avec mon frère quand on se trouve tout seuls$ mais + par respect pour mes enfants et mon mari +$ quand il y a mes on est tous autour de la table je ne parle jamais en arabe avec lui$ même si il se permet de me parler en arabe je réponds en français$
*L 1: ils ont jamais essayé d'apprendre l'arabe$
*L 2: ils ont essayé mais + le maître il est qu'ils ont eu +$ l'instituteur + qu'ils ont eu de tunisien il se prenait pour Dieu$ le Père du fait qu'en Tunisie on doit respecter l'instituteur$ ce qu'il dit ainsi de suite mes enfants ils n'ont pas pu supporter ça et puis ils ont refusé d'y aller et$
*L 1: un peu comme en France euh + au début du siècle ou euh$
*L 2: ben voilà justement + l'instituteur c'était quelque chose d'extraordinaire +$ alors le ce jeune-là il a gardé la mentalité-là$ mais il a oublié que c'est c'est des jeunes français qui voulaient apprendre l'arabe$ il aurait été un peu plus souple avec eux je j'en suis sûre ils auraient appris$ mais moi à la maison je ne pouvais pas leur apprendre parce que j'étais la seule qui parlait la langue-là +$ c'était pas facile avoir six enfants et essayer de leur apprendre l'arabe avec +$ je menais la barque à la maison il fallait tout faire leurs devoirs tout ça alors je pouvais pas$
*L 1: et ça te manque pas tu aurais pas aimé qu'un de tes fils ou de tes une de tes filles parle l'arabe$
*L 2: si j'aurais peut-être j'aurais aimé bon parler mais pfff$ je peux dire que je ne je n'y ai même pas pensé pas du tout +$ du fait qu'on se comprend on se comprend très très bien en français$ je vois pas où elle est l'utilité de se mettre à deux à parler$ une autre langue que les deux les autres sont autour de la table$
*L 1: ouais mais pour les traditions garder les traditions euh$
*L 2: ben les traditions je les ai pas aimées c'est pas pour les inculquer à mes enfants$
*L 1: ouais mais tu fais quand même euh certaines &lt; fêtes comme l'Aïd$

Figure 5. Automatic generation of text from alignment files and selection of the entered key word ("*parce que*"), highlighted with its immediate context.

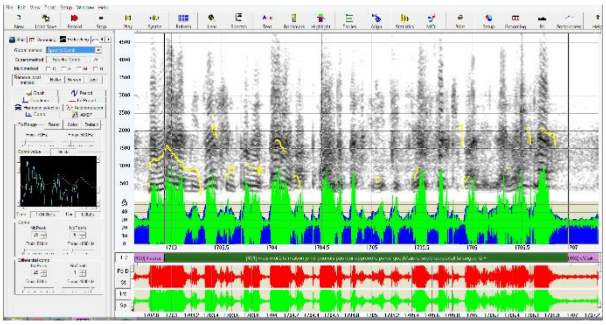


Figure 6. Resulting display of the spectrogram, intensity and pitch curves corresponding to the segment automaticity retrieved from the Excel table.

Integrating this function in one single software package makes possible specific research topics on prosody that would have been seen as too time consuming previously.

3. Acoustic analysis

The measurement of fundamental frequency is particularly sensitive to recorded speech signal distortions due to
1) Poor signal to noise ratio;
2) Filtering of low frequencies, eliminating low harmonics for male voices;
3) Various spurious components due to room echo in the recording places;
4) Encoding in formats such as mp3 or wma with excessive compression levels;
5) External sound sources (car engine, overlapping speech segments, etc.);
6) Presence of creaky segments where the fundamental frequency is not really defined.
Since pitch tracking algorithms are so far prone to errors in adverse recording conditions, and given that for a particular speech segment some algorithms are less prone to errors than others are, WinPitch includes seven different pitch tracking routines to evaluate the fundamental frequency (spectral comb, spectral brush, autocorrelation, AMDF, spectral fit, harmonic selection, Cepstrum and more on the way).
These algorithms and their related parameters can be independently applied on user defined segments of the speech wave, in order to use the most appropriate scheme in a given speech section of the recording. The spectral comb and spectral brush are especially resistant to noise and absence of some harmonics in the spectrum, so that even creaky segments can be adequately analyzed with appropriate parameters. WinPitch includes also a scanning feature allowing a quality analysis of the recording in terms of fundamental frequency coherence, transition and presence of creak (Martin, 2012).

To address these potential problems en to ensure the generation of reliable F0 data, WinPitch features a catalog of methods applicable independently on user-selected speech segments:

1. Spectral comb (Martin, 1981), obtained by correlation of the signal spectrum with a spectral comb with variable teeth intervals. Harmonics frequency range retained in the computation are user selectable;
2. Spectral brush (Martin, 2008), obtained by aligning signal harmonics on a selectable time window followed by a spectral comb analysis;
3. Cepstrum (Noll, 1967), evaluation of the periodicity of the log spectrum;
4. Swipep, developed by IRCAM, derived from the Swipe algorithm (Camacho, 2007) based on harmonic detection followed by a Viterbi smoothing process;
5. Harmonic selection followed by spectral comb, with the retained harmonics selected by the user from a visual inspection on a simultaneously displayed narrow band spectrogram;
6. Autocorrelation, operating directly on the speech waveform, available in three flavors, standard, normed Praat (Boersma, 1993) and Yin (de Cheveigné and Kawahara, 2002), with adjustable window duration;
7. AMDF: average magnitude difference function, with the window length and the clipping percentage user adjustable;
8. Period analysis: F0 values are obtained from period's measurements from pitch markers placed automatically in a first pass and later manually corrected by the user;
These various methods give globally comparable results on good quality recordings. However, for lower quality recordings, the main problems of analysis are:

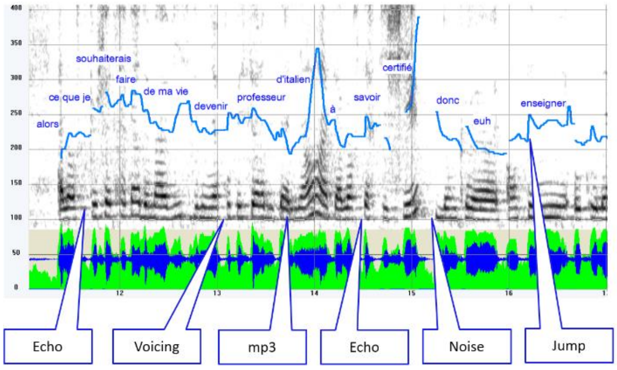
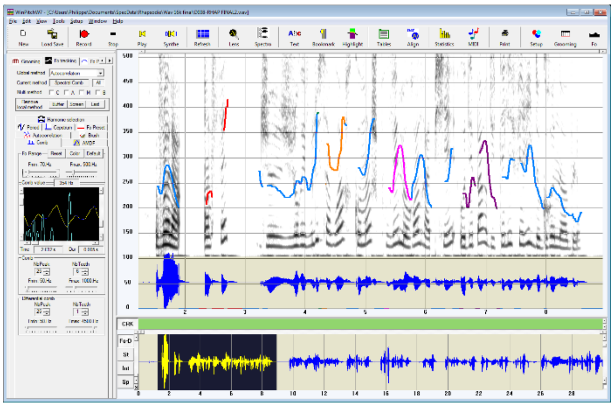


Figure 7. Most common sources of errors for F0 tracking (Rhap-D0003, PFC)

To apply one of these methods, the user first selects a F0 tracking method in the command window (Fig. 6). Then a time window is selected on screen with the mouse guided by visual inspection of an underlying narrow band spectrogram. By releasing the mouse left button, the corresponding segment of the signal is automatically reanalyzed with the selected method, replacing F0 data with the new obtained values. The new F0 curve segment is displayed in a color specific to the tracking method chosen, so that the user can identify visually on the overall F0 curve the tracking method pertaining to a specific time segment. Furthermore, by moving the cursor on screen, the corresponding command box corresponding to the F0 tracking method used for the wave segment defined by the cursor is displayed dynamically in the command box, together with all parameters values used for the chosen tracking method (Fig. 8).

A file containing all the information about corrections made can be saved in text format, as well as a .pitch file describing the corrected pitch curve to be exported to Praat. Applied locally on user's specified speech segments, this method is unique as preventing the use of algorithms applied globally, assuming that a single method would reveal appropriate in all cases.



Global method Autocorrelation
Current method Spectral Comb
Fmin 70 Hz Fmax 500 Hz
Comb value 354 Hz
NbPeak 25 NbTeeth 6
Fmin 50 Hz Fmax 1000 Hz

Figure 8. F0 curve sections are displayed in different colors according to the F0 tracking method used. The corresponding command box selected automatically on the left side (Rhap-D1001)

4. Interface with other programs

WinPitch can import Transcriber, PFC, Necte files among others, and read and save Praat files (old and new TextGrid format). All data can be exported in Ascii (with Unicode extension) directly as a text file or into Excel®. It can load wav and many other sound or video files directly, with direct resampling into any user selected sampling frequency. This is especially important to avoid wasting storage space and computing power by using too high sampling frequency, whereas 16,000 Hz or 22,050 Hz are sufficient for speech recordings.

Sound files can be edited (segment deletion, copy and paste), and can be concatenated or “glued” together to form a stereo file from 2 mono files (in case where a same event recorded into two independent files must be analyzed together). Text can be added (in any color and font) on the analysis window for illustration purposes. The resulting augmented analysis window can then be exported in a picture format in a text editor such as Word for example. Segments of the acoustic analysis (Fo, intensity, waveform, spectrogram) can be highlighted and independently labeled, for paper illustration and for later selection in Excel (or other program) for further statistical analysis.

5. Examples of applications

An example of current research projects pertains to the prosodic properties of conjunctions in French such as *parce que*, *alors que*, *bien que*, etc. The preliminary concordance analysis extracted from the text more than 8,500 occurrences of these conjunctions. Without some efficient data mining tool, the acoustical analysis of all these occurrences appeared almost impossible to be carried out, as the retrieval of one single occurrence implies the loading of the appropriate sound and aligned text file, and the manual search for the conjunction with its left and right contexts, a time consuming task in itself. If a manual retrieval of one occurrence would take 2 minutes (for a trained user), to total retrieval of all examples would have taken some 17,000 minutes, or 233 hours…

In the Rhapsodie project [11], where the pitch curves related to typical spontaneous speech examples in French are going to be displayed with Praat, the actual fundamental frequency analysis is performed by WinPitch, which provides much more satisfactory pitch curves thanks to its multi method pitch tracking capability.

WinPitch can be downloaded from www.winpitch.com and is free for the asking.

6. References

Boersma, Paul (1993) Accurate short time analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound, Proc. Institute of Phonetic Sciences, 17. Univ. Amsterdam, 97-110.

Camacho, Arturo (2007) Swipe: a sawtooth waveform inspired pitch estimator for speech and music, PhD thesis, University of Florida, 116 p.

C-ORAL-ROM (2005) Integrated Reference Corpora for Spoken Romance Languages, Edited by Emanuela Cresti and Massimo Moneglia, Studies in Corpus Linguistics 15, John Benjamins, Amsterdam.

C-ORAL-ROM Brasil (2013) http://www.c-oral-brasil.org/

de Cheveigné, Alain and Hideki Kawahara (2002) Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, 111(4).

Lexique3 (2009) http://www.lexique.org/outils/Manuel_Lexique.htm

Martin, Ph. (1981) Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne, 12e Journées d'Etude sur la Parole, SFA, Montréal, 1981.

Martin, Ph. (2008) Crosscorrelation of adjacent spectra enhances fundamental frequency estimation Proc. Interspeech, Brisbane, 22 – 26 September 2008.

Martin, Ph. (2012) Automatic detection of voice creak, Proc. Speech Prosody, Shanghai, September 26-28.

Noll, A. Michael (1967) Cepstrum Pitch Determination, Journal of the Acoustical Society of America, Vol. 41, No. 2, (February 1967), 293-309.

Praat, www.praat.org.

Rhapsodie (2010) Corpus prosodique de référence en français parlé, http://rhapsodie.risc.cnrs.fr/en/archives.html

Transcriber, a tool for segmenting, labeling and transcribing speech, http://trans.sourceforge.net/en/presentation.php

WinPitch (2013) www.winpitch.com