WinPitchPro
A Tool for Text to Speech Alignment and Prosodic Analysis

Philippe Martin

ARP - UFRL
Université Paris 7 Denis Diderot
philippe.martin@linguist.jussieu.fr

Abstract

Traditional experimental phonetics laboratories are made somewhat obsolete by the use of popular software tools such as Praat [7]. Indeed, these tools provide most of the acoustic analysis engines needed for prosodic research, in particular fundamental frequency trackers and speech prosodic morphing synthesizer. Still, their usage is not always totally intuitive, and considerable training must sometimes be provided in order to ensure a reasonable degree of success and efficiency when used in a research project. In this perspective, new generation acoustical analysis software such as WinPitchPro will put emphasis on reliability of measurements and ease of use.

1. Introduction

There are well known and widely used software tools such as Praat [7], which somewhat make traditional experimental phonetics laboratories obsolete. Indeed, these tools provide most acoustic analysis engines needed for prosodic research, in particular fundamental frequency trackers and synthesizers for speech prosodic morphing. Still, their use is not always intuitive, and considerable training must be provided in order to ensure a high degree of success and efficiency when used in a research project. New generation of acoustical analysis software should put emphasis on reliability of measurements and ease of use, as a good ergonomic design is not a vain luxury.

WinPitchPro is one of these newly designed and ergonomic tools, completely redesigned after a first version appeared in 1996 [4]. Using very reliable speech analysis engines, it was designed with ease of use in mind, so that the reduced number of manual operations needed to perform a common tasks will make the difference between a feasible and a non feasible too time consuming research project. Indeed in the software industry, emphasis has been put for some times on ergonomics.

2. Recording and real time analysis

Speech recording with WinPitchPro allows for real time analysis and display of the prosodic curves (Fo and intensity) together with the corresponding spectrogram. This allows to a very precise monitoring of recordings, letting the user not only to adjust the input level to an optimal value, but also to better position the sound capture devices (microphones, etc.) as the recorded sound spectrogram is displayed in real time. With real time spectrographic display, the presence of echoes and various noise sources can be easily detected and corrected. As an added facility, a loop recording mode is also provided. When the user stops the recording, the last n seconds of recorded sound (n being adjustable) are stored in a single time synchronized file.

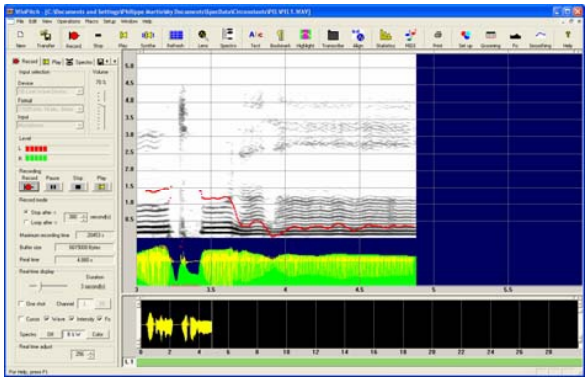

Figure 1: *Real time Fo and spectrographic analysis allows for precise monitoring of speech recordings.*

3. Speech transcription and text to speech assisted alignment

Once speech data have been captured, text transcription and alignment are generally made, tasks which are executed easily and efficiently with various tools provided by the program.

3.1. Text transcription

In speech transcription mode, only the sound file is available. A set of integrated functions allow for very fast operations for

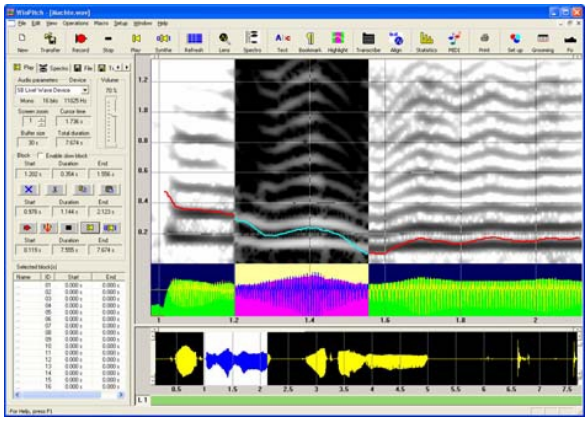

Figure 2: *Text transcription is performed by defining continuous segments of speech played back at reduced speed to facilitate their auditory perception.*

transcription in any Unicode available font. As the user progresses through the sound file by defining segments of speech played back at reduced speed to facilitate their auditory perception and transcription, a database is simultaneously built, containing the text and time positions of each segment. This database can be saved in XML and Excel formats, for easy interface with other data processing programs. For each defined speech segment, the user can enter the corresponding text on one of 8 available layers, devoted to different speakers or different levels of transcription (such as syntagms, word, syllable or speech sound).

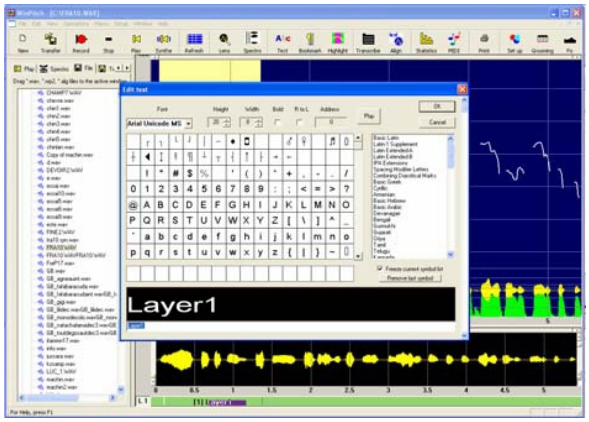

Figure 3: *Text transcription can use any Unicode available font and keyboard. A table of frequently used symbols can also be built on line.*

3.2. Text to speech alignment

In text to speech alignment mode, text has already been transcribed or is otherwise available in electronic form, but has to be aligned with the sound file by defining a set of bidirectional pointers linking segments of text with segments of sound. A unique tool, based on speech playback at a user programmable (reduced) speed, allow the operator to simply click on the last element of the currently played back segment of text displayed while it is perceived (the slower playback speed rate allows for simultaneous and synchronic sound perception and screen cursor positioning). The program automatically stores a bidirectional pointer between text and sound to establish the alignment.

This procedure has the tremendous advantages over automatic methods based on the use of speech recognition emerging engines (see for ex. [2]): they are insensitive to the quality of the sound recording, and of course they do not require any speaker training (which is most of the time impossible to execute as large corpus speakers are generally not available to train the recognizer).

Indeed the problems inherent to automatic recognition are passed to the human operator, while speech recognition based tools must adapt to individual speakers to be efficient. Furthermore, problems arising from the presence of background noise, or of simultaneous speakers segments (very common in spontaneous discourse corpora) are as well handled by this approach.

As in transcription mode, the program establishes automatically an alignment database in XML and Excel formats.

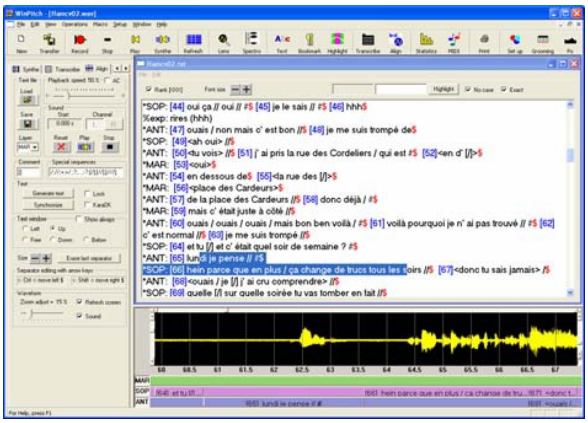

Figure 4: *Computer assisted text to speech alignment. While listening to speech played back at reduced speed, the operator clicks on text units corresponding to the elements of speech perceived.*

4. Navigation

Once the alignment between text and speech has been established, the navigation through the sound file is extremely easy. Merely by clicking on a word or a sequence of words, the user gets automatically and immediately the surrounding segments played back and the corresponding acoustical analysis is instantaneously displayed, with spectrogram and oscillographic, intensity and fundamental frequent curves.

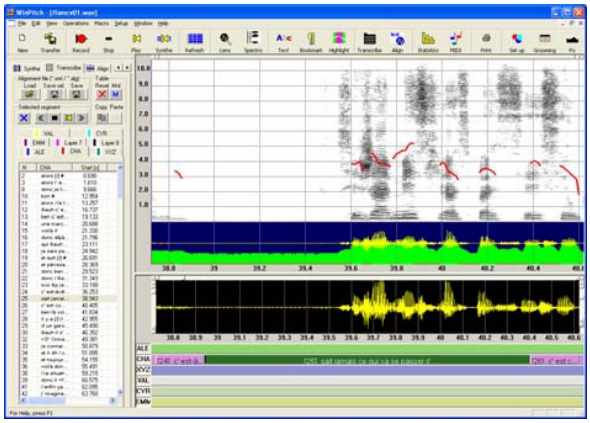

Figure 5: *After text to speech alignment, a database is automatically built, allowing the user to quickly retrieve and analyze segments of speech by clicking on the text.*

The same result is obtained when the user selects a whole word, syntagms, or one or more than one sentence: the corresponding sound segments get analyzed and played back automatically.

Provision is made in the alignment process for discourse analysis (spontaneous speech when speaker's voices overlap). In the case of multiple simultaneous speakers a separate transcription layer is automatically allocated to each speaker

(provided the text follows a predefined format, see [1] for details).

Other navigation tools include automatic pan and zoom from simple mouse commands inside a navigation window displayed the sound wave at a user defined time scale.

5. Fundamental frequency analysis

Prosodic analysis requires reliable fundamental frequency tracking algorithms, ensuring accurate reading and display of Fo values in a wide range of laryngeal frequencies and resistant to the presence of background noise. For this purpose, WinPitchPro includes 5 separate fundamental frequency analysis engines that can be activated globally on the whole sound file, or locally on any number of user defined time segments: AMDF, Spectral Comb, Spectral brush, Autocorrelation and Selected Harmonics Comb.

The spectral comb method [3] is particularly robust in the case of noise, even when other sources such as musical instruments are present with the speech sound. Its robustness stems essentially from the use of all harmonic information (frequency and intensity) brought in the spectrum. As the other WinPitchPro Fo tracking methods, parameters such as range of harmonics retained, number of comb's teeth, etc. are user configurable on the whole sound buffer or on selected segments.

A variant of the spectral comb allows the user to select specific harmonics for Fo evaluation. For instance low frequency noise harmonics can be isolated by selecting higher speech harmonics with a simple graphic command (an enclosing rectangle defined on the narrow band spectrogram).

The spectral brush method is an experimental which uses the non stationary property of the human fundamental frequency during voicing to separate them from the harmonics of musical instruments to determine Fo. This feature is particularly useful for research on singing voice from currently available recordings [5]

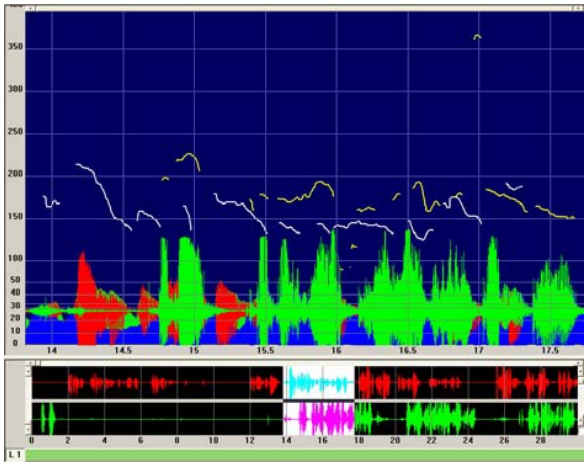

Figure 6: *Analysis of stereo data: Fo and intensity curve are displayed simultaneously with different colors.*

6. Prosodic morphing

New values of prosodic parameters (fundamental frequency, intensity, segment duration and pauses) can be entered graphically on screen with mouse commands defining piecewise linear functions. The exact values of segment vertices are monitored on a table while placed on screen, and graphic segments can be edited easily by dragging, creating or deleting their vertices. The morphing is executed by a PSOLA like method, relying on precise automatic placement of pitch periods in the signal [6]. These pitch markers can also be edited by the user for further improvement of the quality of re synthesized speech.

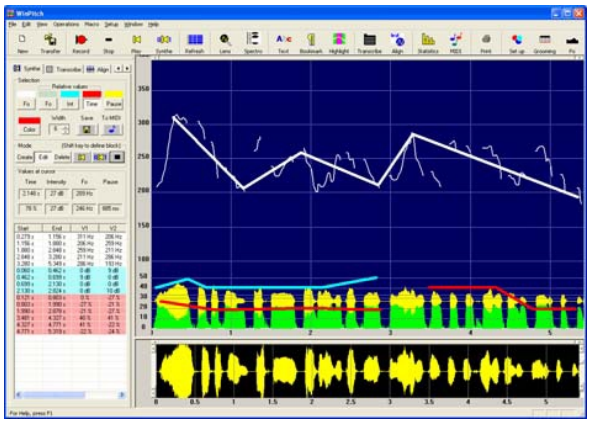

Figure 7: *Prosodic morphing: relative or absolute new values of Fo, intensity, duration and pause is entered graphically with direct control of values defined (table on left side of the screen).*

7. Sampling of data and statistics

As mentioned earlier, text to speech alignment can be saved in both XML and Excel ® formats. Other data such as text placed on screen, local values of analysis parameters, highlighted sections, etc. can also be saved in a text format for easy editing by the user of by other programs.

Fundamental frequency and intensity values pertaining to various scopes (screen selected speaker, selected highlight time segments, whole buffer) can be directly saved into Excel (with one single mouse click) for further statistical or other processing.

The highlight function is especially useful as it allows the user to define and label specific sound segments (such as stressed syllables, diphthongs, syntagms, etc.). With this approach, the elaboration of script files is unnecessary, as highlighted segments define a set of sound segments upon which further processing can be executed (filtering, statistics, etc.).

8. Multimedia processing

WinPitchPro can read multimedia files in most formats (wav, au, aiff, mp3, mpg, mpeg2, mpeg4, avi, …) In case of formats containing video information, the images or film are played back synchronously with the sound, even in the case of slow rate playback.

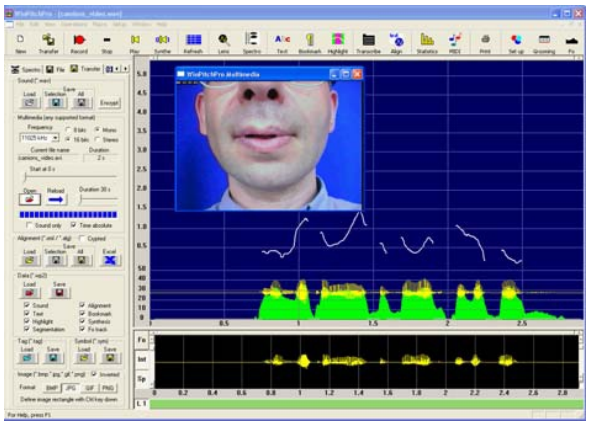

Figure 8: *Simultaneous display of speech and video data (ICP data, Grenoble).*

9. Conclusion

While many speech analysis tools are now commonly available, many beginners are puzzled as how to acquire the necessary knowledge to efficiently use the software. The lack of appropriate experimental phonetic training may, on the other hand, induce blind trust in the acoustical analysis provided (this may have severe consequences especially for fundamental frequency, and let building on models and theories based on wrongly analyzed data).

Whereas any reasonable designed software tool may be fit for limited scope experimental analysis of speech signals, carefully designed interface can make the difference between a feasible and an impossible project when the amount of data is very large. The number of mouse clicks and keystrokes necessary to perform a specific function becomes then extremely important.

WinPitchPro was designed with these considerations in mind, reducing the number of operations to a minimum for each function, and providing at the same time an intuitive approach for the overall set of operations.

WinPitchPro [8] operates under Windows 98, Me, 2k and XP, and can be downloaded form the www.winpitch.com web site. It can read and playback most formats of multimedia files.

10. References

[1] C-Oral-Rom, 2000. http://lablita.dit.unifi.it/coralrom

[2] 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, pp. 541-552.

[3] Martin, Ph., 1981. Mesure de la fréquence fondamentale par intercorrélation avec une fonction peigne, *Actes des XIIèmes Journées d'Etude sur la Parole*, Montréal, juin 1981.

[4] Martin, Ph., 1996. WinPitch: un logiciel d'analyse temps réel de la fréquence fondamentale fonctionnant sous Windows, *Actes des XXI Journées d'Etude sur la Parole*, Avignon, pp. 224-227

[5] Martin, Ph., 2000. Peigne et brosse pour Fo : Mesure de la fréquence fondamentale par alignement de spectres séquentiels, *Actes des XXIIIèmes XXI Journées d'Etude sur la Parole*, Aussois, France, juin 2000, pp. 245-248.

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

[7] Praat, 2003. http://www.praat.org

[8] WinPitchPro, 2003. http://www.winpitch.com