PROSODIC ANNOTATION OF LARGE SPEECH CORPORA

Philippe Martin

LLF, UFRL, Université Paris Cité
philippe.martin@u-paris.fr

ABSTRACT

Recent developments in linguistic analysis and automatic recognition made the availability of large spontaneous speech corpora pivotal. The ease of use of efficient research tools is becoming essential, given the usual large amount of data considered. The software described in this paper addresses these requirements, by offering fast access to recorded data in a variety of formats, allowing efficient acoustic analysis and prosodic annotation tools of selected speech segments of interest.

Pre-compilation of transcribed text for a very large number of recordings makes the extraction of selected transcribed material extremely fast, together with the display of spectrogram, wave, intensity, and fundamental frequency curves.

Prosodic annotation uses graphic drawing tools, operating automatically or manually either with the ToBI standard notation, or an automatic categorization of melodic contours based on their glissando levels.

Annotation of large speech corpora are presently available in French, English, Italian, Portuguese, Spanish, Dutch, German, Russian, Malay, Korean and Mandarin.

Keywords: Spontaneous speech, concordancer, prosodic annotation, spectrogram, fundamental frequency.

1. INTRODUCTION

Deep learning training for TTS (text to speech) and STT (speech to text) usually requires a very large number of transcribed speech samples. This leads both private and public research institutions to develop large, dedicated speech corpora, both read and spontaneous. Next to existing speech data designed by and for linguists (e.g., [2], [3], [5]), corpora designed for AI (e.g., [11], [14], 15]) could also be advantageously used for linguistic research, especially applied to spontaneous speech whose grammatical and prosodic characteristics may not yet be completely understood.

From a more general viewpoint, to acquire a better insight in a specific linguistic domain, automatic processing of data is not always advisable, as researchers may miss important experimental events buried in a large amount of data. Popular tools such as Praat [1], even associated with dedicated scripts, may not be always suitable, especially in the prosodic domain. More efficient software should then be made available to allow detailed examination by operators possibly with less efforts, in particular pertaining to the following points:

a) Handling of various text transcription formats, from character single byte [10] to UFT-8 coding [11].
b) Fundamental frequency analysis adapted to degraded recordings frequently found in spontaneous speech recordings (e.g. [2]).
c) Fast data mining, essential for users dealing with research topics that cannot or should not use automatic extraction processes.
d) Availability of user-friendly annotation tools, pivotal if extraction of data cannot be executed automatically due to degraded speech recordings.

This paper gives the main features of a dedicated spontaneous speech analysis software WinPitch [17], whose functions are tailored for prosodic analysis and annotation of large speech data sets.

2. CORPORA

2.1 Selected speech corpora

The software described in this paper can be used as is with any corpora using similar transcription encoding (see section 3).

2.1.1. dedicated to linguistic analysis, includes mostly spontaneous speech:

Orfeo, French [2].
OFROM, French [3].
C-ORAL-ROM: French, Italian, Spanish, European Portuguese [4].
C-ORAL-ROM Brazil: Brazilian Portuguese [5].
CGN Corpus Gesproken Nederlands, Dutch and Flemish varieties [6].
Santa Barbara corpus of American English [7].
Archive for Spoken German [8].

2.1.2. dedicated to TTS, STT and voice cloning, includes mostly read speech:

Decoda, Spontaneous French conversations [9].
The LJ Speech Dataset, American English [10].
SIWIS Speech Synthesis Database, French, Italian [11].

Beijing Magic Data Technology, French, Italian, Mandarin, Korean, Spanish, … (read and spontaneous speech) [12].
CSTR VCTK: English Multi-speaker Corpus [13].
LibriSpeech ASR corpus, English [14].
Russian read Speech [15].
Zeroth-Korean corpus [16].

3. TRANSCRIPTIONS

3.1 Text formats

All the above corpora come with text transcriptions associated with segments of recordings of various lengths, from isolated sentences (e.g., [11], [15]), to sequences aligned on prosodic breaks (e.g., [2]) or on syntactic boundaries (e.g., [8]) or even of random lengths (many recordings found in [2]).

There is no common standard for text coding transcription, varying from simple 1-byte ASCII per character to the more sophisticated JSON format (with meta information imbedded) or UFT-8 found in Magic Beijing [12]).

Automatic identification, conversion, and extraction of transcription text are implemented for the following formats:

XML: C-Oral-Rom [4],[5]
Trs (Transcriber): Orfeo [2], Decoda [9]
Tag (Transcriber): Orfeo [2]
TextGrid (Praat): Orfeo [2], OFROM [3], CGN (awd) [6]
CRF-ALG Text: Orfeo [2]
Necte XML text: Necte
RTF text: Beijing Magic Data [12]
JSON: Archive for Spoken German [8]
TXT (1 and 2 bytes): SIWIS [11]
UFT-8: Beijing Magic Data [12]

The corpora transcription file extensions do not necessarily give enough information on the text format used, so that automatic selection routines have been designed to handle the text transcription correctly, sometimes imbedded in other text information pertaining to speakers, recording conditions, etc. With the automatic selection of formats, files with any of the formats given above can be mixed in the same project and accessed with all available functions of the software.

3.2 Automatic IPA segmentation

Given that the transcriptions' grain varies from phone (e.g., Decoda [9]) to long sentences reaching more than 30 seconds (e.g., Orfeo [2]), an integrated segmentation algorithm has been implemented, based on dynamic Viterbi alignment with a TTS generated speech signal (available in more than 30 languages). It is then easy to segment accurately sentences transcribed deprived from any segmentation in SIWIS [11] or CSTR VCTK [13] into phones in IPA format.

Unexpected difficulties may arise from text coding variations sometimes found in transcription files. In SIWIS [11], some transcription texts are in plain ASCII (1 byte per character), whereas some others, in the same set of sentences, uses a 2-byte format. In other instances, as in Beijing Magic Data files [12], the transcribed text uses an expected UTF-8 format for Mandarin, but the RTF format for Spanish or French.

4. ACOUSTIC ANALYSIS

Acoustic speech analysis makes it possible to perform prosodic annotation on acoustically degraded recordings, such as those found in old entertainment or documentary movies, but also in many spontaneous speech recordings (e.g., Orfeo [2]). The software described here is specially designed for this purpose. It integrates various user-friendly functions to obtain reliable pitch curves and realize prosodic annotations in adverse acoustic conditions.

Besides the use of 7 user-selected different pitch tracking algorithms (spectral comb, 4 flavours autocorrelation, spectral brush, AMDF…), displaying of an overlay narrow band spectrogram whose frequency scale is automatically aligned on the pitch scale allows the annotator (after some training…) to visually ensure that the pitch curve obtained by the selected tracking algorithms is reliable, as it should correspond to the pitch variations indicated by the first (or the second) spectrogram harmonic.

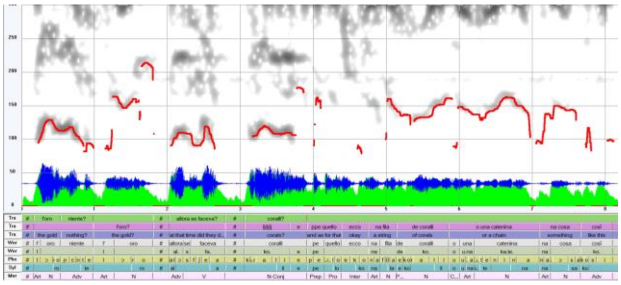


Figure 1. Superposition of the fundamental frequency curve and the first harmonic of the corresponding narrow-band spectrogram. This allows the user to visually check the validity of the fundamental frequency curve, and eventually select an alternate pitch tracking method, or draw on screen with the available drawing functions corrected pitch segments aligned on the first (or second) harmonic of the spectrogram.

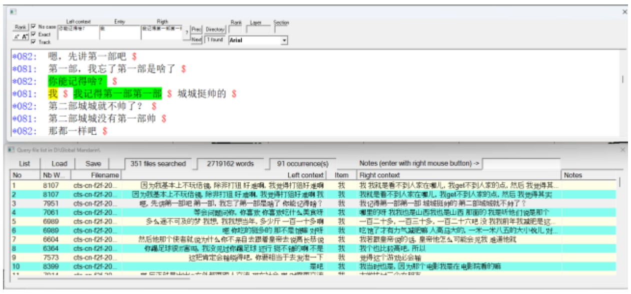
Segments of pitch errors can then easily be located and corrected using another pitch tracking algorithm or directly by the user with the available drawing functions to draw on screen pitch curve segments aligned on the first or second spectrogram harmonic.

This function allows the user to take care of the commonly found problems of spontaneous speech recordings, such as the use of unappropriated microphones (removing low pass frequencies), voice overlays, background noise or even creaky voice (relying on the second harmonic). An example is given Fig. 1.

5. DATA MINING

Dealing with a large amount of transcribed speech data requires an efficient data mining process, based on a concordancer whose input can be any sequence of words (in Unicode format, thus including Korean and Mandarin) to retrieve automatically the corresponding speech segment, together with its acoustic analysis (spectrogram, pitch and intensity curves).

To speed up the retrieval process, a precompiled file is generated for any selected group of recordings, mixing transcription formats if needed. This precompiled file consists of a database linking each word found in all the transcriptions of a user defined group of transcribed recordings to their corresponding speech segment, converted automatically in WAV format. While the compilation of all text transcriptions of a given recording set may take some time (from few seconds for corpora such as Orfeo [2] to few minutes for CSTR VCTK [13], which includes more than 88,000 sentences…). Once this file has been generated, it can be reloaded any time later for a very fast retrieval in less than 1 second of speech data from the concordancer entry. An example is given Fig. 2.



*082: 嗯，先讲第一部吧 $
*081: 第一部，我忘了第一部是啥了 $
*082: 你能记得啥？ $
*081: 我 $ 我记得第一部第一部 $ 城城挺帅的 $
*082: 第二部城城就不帅了？ $
*081: 第二部城城没有第一部帅 $
*082: 那都一样吧 $

351 files searched | 2719162 words | 91 occurrence(s) | Notes (enter with right mouse button) ->

No	Nb W...	Filename	Left context	Item	Right context	Notes
1	8107	cts-cn-f2f-20...	因为我基本上不玩信镜, 除非打诅 好难啊, 我觉得打诅好难啊	我	我 我就是看不到人家在哪儿, 我get不到人家的点, 然后 我觉得其...	
2	8107	cts-cn-f2f-20...	因为我基本上不玩信镜, 除非打诅 好难啊, 我觉得打诅好难啊 我	我	我就是看不到人家在哪儿, 我get不到人家的点, 然后 我觉得其实	
3	7951	cts-cn-f2f-20...	嗯, 先讲第一部吧 第一部, 我忘了第一部是啥了 你能记得啥？	我	我记得第一部第一部 城城挺帅的 第二部城城就不帅了？	
4	7061	cts-cn-f2f-20...	等会问题问你, 你喜欢 你喜欢吃什么美食呀	我	哪里的呀 我我也是山西我也是山西 那面的 我要听他们说是那个	
5	6989	cts-cn-f2f-20...	多么遥不可及的梦 我想, 我我想当年, 多少斤 一百一十多啊	我	一百二十多, 一百三十多, 一百二十六吧 没 我我前年我减肥是这...	
6	6989	cts-cn-f2f-20...	嗯 你吃的挺多的 那不是[illegible] 对呀	我	吃饱了才有力气减肥嘛 人高马大的, 一米一米八五的大小伙儿 对...	
7	6604	cts-cn-f2f-20...	然后他那个使者就说为什么你不亲自去跟着皇帝去说高长恭说	我	我若跟皇帝说的话, 皇帝他怎么可能会见我 通通他就	
8	6364	cts-cn-f2f-20...	你踢足球很厉害吗, 我没见过你踢足球 还行 挺不错的啊 不是	我	我个也比较高吧, 所以	
9	7573	cts-cn-f2f-20...	这把肯定会输赔得吧, 你要相当于去激一下	我	觉得这个游戏必会输	
10	8399	cts-cn-f2f-20...	是吧	我	我当时也是, 因为那个电影我是在电影院看的嘛	

Figure 2. An example of speech segment retrieval in Mandarin from the sequence entered in the concordancer (here the word 我). The context is displayed and highlighted in a dedicated window (top figure). Any surrounding speech segment can be retrieved as well by selecting any part of the text with the mouse.

Fig. 3 and 4 display the pitch curve and the underlying narrow band spectrogram corresponding to the example of Fig 2.

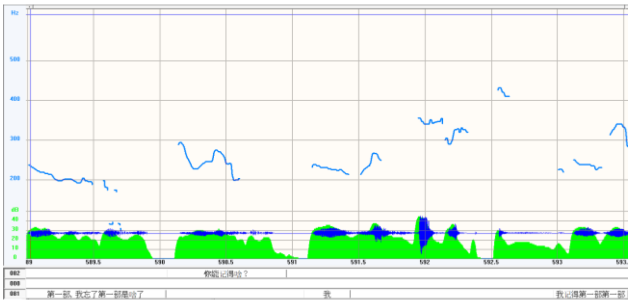


Figure 3. Display of the pitch curve corresponding to the text selected in Fig. 2, with the text transcription layers (bottom of the figure).

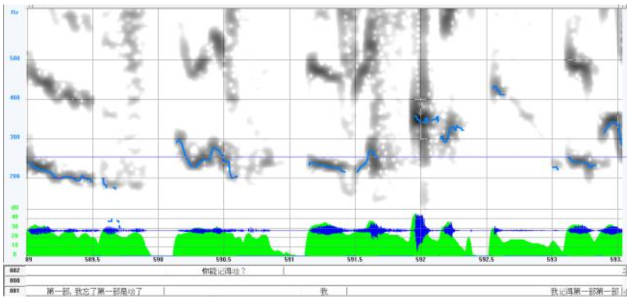


Figure 4. Display of the pitch curve with an overlaying narrow band spectrogram corresponding to the text selected in Fig. 2, with the text transcription layers (bottom of the figure). The validity of pitch tracking can then easily be checked visually.

6. PROSODIC ANNOTATION

Available prosodic annotation tools include text and graphic functions to allow the user to add any kind of information on screen, which could be saved later in text or Excel format. Segments of pitch or formant frequencies can be either automatically or manually highlighted and corrected, if necessary, with mouse-controlled graphic functions, with user-selected color and thickness.

Prosodic annotation presupposes an implied theoretical model (such as Autosegmental-Metrical) and the use of a specific annotation system (ex. ToBI). Fig. 5 shows an example of prosodic annotation by pitch contours of a French sentence. Stressed vowels, located either automatically of manually by the user, have their melodic contours automatically highlighted in various colors, according to their rising or falling pitch, above or below their pitch change perception level (glissando threshold). ToBI annotation can also be entered the same way, once melodic target properties and color coding have been predefined by the user and entered in the system.

With the implied prosodic model used in this example, prosodic annotation operates in three steps:

1) Manual or automatic localization of stressed syllables (excluding emphasis) defining accent

phrases “right” boundaries by their final position (in French);

2) Automatic classification and color coding of prosodic events occurring on stressed syllables. These events are instantiated by specific melodic contours assumed by hypothesis to indicate dependency relations between accent phrases. These user defined dependency relations define the prosodic structure.

3) The corresponding prosodic structure is then automatically displayed (Fig. 5).

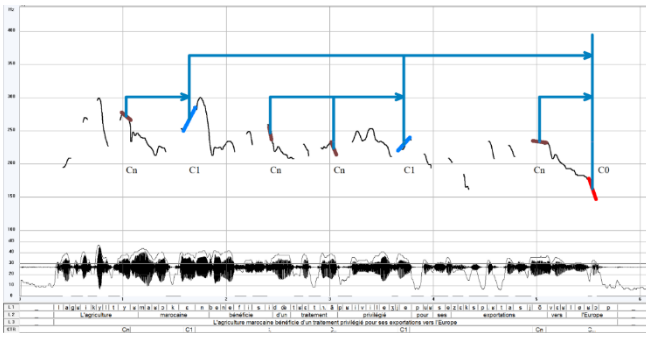


Figure 5. An example of prosodic annotation using a model relying on pitch contour on stressed syllables to indicate dependency relations between accent phrases, which in turn define the prosodic structure (in square brackets with oriented branches). (SIWIS corpus [11]).

Melodic contours, as well as ToBI melodic targets, can be dynamically edited with simple mouse commands, automatically updating the resulting prosodic structure, as well as the actual sentence prosody, modified by the imbedded Psola based melodic morphing routine.

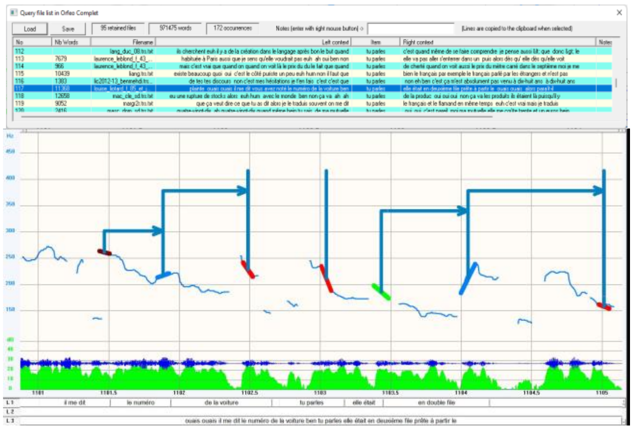
7. EXAMPLES OF APPLICATIONS

7.1 Multi speaker realizations

The SIWIS French corpus [11] consists of a set of relatively short sentences read by 8 to 14 speakers, with a total of some 5340 isolated sentences. Analyzing the variations of prosodic realizations between speakers for a particular sentence is a breeze, requiring only a few minutes from the alphabetic ranking of text transcriptions. Clicking on text entries on the concordancer displays the corresponding speech samples, together with their pitch, intensity, and spectrographic data. One can then quickly in a few seconds observe how different speakers use different prosodic coding strategies even if they are reading the same text.

7.2 Phraseology

Analysis of prosodic realization of connectors, such as *enfin*, or *en fait* in French, requires with the commonly available tools many manipulations to extract the pertinent information from one of available corpora such as Orfeo [2] or OFROM [3]. However, with WinPitch [13], the software described in this paper, the 179 occurrences of *tu parles* found in the Orfeo corpus of more than 3,438,438 words are displayed in less than one second, and their pitch curves can be retrieved and displayed with a single mouse click (Fig. 6).


Query file list in Orfeo Complet
Load
Save
95 retained files
971475 words
172 occurences
Notes (enter with right mouse button)
(Lines are copied to the clipboard when selected)
Hz
450
400
350
300
250
200
150

No	Nb Words	Filename	Left context	Item	Right context	Notes
112		lang_duc_08.trs.txt	ils cherchent euh il y a de la création dans le langage après bon le but quand	tu parles	c'est quand même de se faire comprendre je pense aussi tût que donc tût le	
113	7679	laurence_leblond_I_43_...	habituée à Paris aussi que je sens qu'elle voudrait pas euh ah oui ben non	tu parles	elle va pas aller s'enterrer dans un puis alors dès qu' elle dès qu'elle voit	
114	966	laurence_leblond_I_43_...	mais c'est vrai que quand on quand on voit là le prix du du le fait que quand	tu parles	de cherté quand on voit aussi le prix du mètre carré dans le septième moi je me	
115	10439	lang.trs.txt	existe beaucoup quoi oui c'est le côté puriste un peu euh hum non il faut que	tu parles	bien le français par exemple le français parlé par les étrangers et n'est pas	
116	1383	lic2012-13_benmehdi.trs...	de tes tes discours non c'est mes hésitations je t'en fais c'est c'est que	tu parles	non eh ben c'est ça n'est absolument pas venu à dix-huit ans à dix-huit ans	
117	11368	louise_liotard_f_85_et_j...	plainte ouais ouais il me dit vous avez noté le numéro de la voiture ben	tu parles	elle était en deuxième file prête à partir le ouais ouais alors pass'il	
118	12658	mac_cle_sd.trs.txt	eu une rupture de stocks alors euh hum avec le monde ben non ça va ah ah	tu parles	de la produc oui oui oui non ça va les produits ils étaient là puisqu'il y	
119	9052	magi2t.trs.txt	que ça veut dire ce que tu as dit alors je le traduis souvent on me dit	tu parles	le français et le flamand en même temps euh c'est vrai mais je traduis	
120	2416	marc_don_sd.trs.txt	quatre-vingt-dix ah quatre-vingt-dix quand même hein tu sais de ma mutuelle	tu parles	oui oui c'est pareil moi ma mutuelle elle me coûte trente et un euros hein	

il me dit
le numéro
de la voiture
tu parles
elle était
en double file
ouais ouais il me dit le numéro de la voiture ben tu parles elle était en deuxième file prête à partir le

Figure 6. An example of analysis of the occurrence of the French locution *tu parles* "you speak" with its context *il me dit le numéro de la voiture tu parles elle était en double file* (Corpus ORFEO [2]). Pitch contours of stressed vowels are highlighted, and the corresponding prosodic structure (with square brackets) displayed automatically

8. CONCLUSION

The described user-friendly speech analysis software allows efficient prosodic annotations for prosodic research in general on a large variety of speech corpora, both read and spontaneous, making advantage of recent quasi-industrial corpora production designed for other purposes, such as speech recognition and synthesis based on deep learning algorithms, requiring a large amount of data.

The interoperability routines integrated in the software as well as the precompiled transcription data makes research of prosodic and other phonetic characteristics user selected sequences or words extremely fast and easy, making data collection that would normally take months executed in a few hours.

This interoperability allows the use of the program on any collection of transcribed recordings sharing one of the formats listed above, the precompiled transcription being generated by one mouse click operating on data simply collected in the same computer directory.

Developed by the author, the software can be downloaded from www.winpitch.com [17], and used freely for academic purposes.

REFERENCES

[1] Praat, doing phonetics by computers, https://www.fon.hum.uva.nl/praat/

[2] Orfeo, Le Corpus d'Etude du Français Contemporain - Le projet ORFEO www.projet-orfeo.fr/corpus/le-corpus-d-etude-du-francais-contemporain/14-orfeo

[3] Avanzi Mathieu, Béguelin Marie-José, Corminboeuf Gilles, †Diémoz Federica & Johnsen Laure Anne (2012-2023). Corpus OFROM – Corpus oral de français de Suisse romande. Université de Neuchâtel, http://ofrom.unine.ch

[4] C-ORAL-ROM French, Italian, Spanish, European Portuguese, http://www.elda.org/en/proj/coral/fr/coralrom.html

[5] C-ORAL-ROM Brazil, http://www.c-oral-brasil.org/english-site/index.html

[6] CGN Corpus Gesproken Nederlands, https://ivdnt.org/images/stories/producten/documentatie/cgn_website/doc_English/topics/index.htm

[7] Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

[8] Archive for Spoken German, https://agd.ids-mannheim.de/index_en.shtml

[9] Bechet F. et al. (2012) DECODA: a call-centre human-human spoken conversation corpus, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 1343-1347.

[10] The LJ Speech Dataset, Version 1.0 July 5, 2017, https://keithito.com/LJ-Speech-Dataset

[11] The SIWIS French Speech Synthesis Database, https://datashare.ed.ac.uk/handle/10283/2353.

[12] Beijing Magic Data Technology Co., Ltd., High quality Training Datasets. https://www.magicdatatech.com/

[13] CSTR VCTK Corpus: English Multi-speaker Corpus, https://datashare.ed.ac.uk/handle/10283/3443.

[14] LibriSpeech ASR corpus, Large-scale corpus of read English speech, https://www.openslr.org/12/

[15] Pavel Skrelin et al., A Fully Annotated Corpus of Russian Speech, https://aclanthology.org/L10-1188/

[16] Zeroth-Korean corpus SLR40, https://www.openslr.org/40/

[17] WinPitch, Speech analysis software, https:// www.winpitch.com