DOI: 10.17469/O2106SLI000021

PHILIPPE MARTIN

Intonation of telephone conversations in a Customer Care service

RATP-DECODA is a Customer Care Service corpus part of the ORFEO project (2020). It includes 1988 recordings of client requests made between 2009 and 2011 to the RATP (*Régie Autonome des Transports Parisiens*) call center. While some semantic and syntactic observations have already be made available in Brechet et al. (2012), no intonation analysis has so far been conducted to show the role of prosody in these specific discourse conditions. In this paper, prosodic annotations were performed according to the dependency prosodic structure model of Martin (2018b), where melodic contours located on stressed vowels, classified according to their glissando values (perceived as a melodic change vs. as a static tone), indicate dependency relations between stress accent phrases (i.e. stress groups) and ultimately determine the prosodic structure of the sentence.

Keywords: Intonation, prosodic structure, telephone conversation, customer service.

1. *Introduction*

Sentence intonation not only conveys emotions (joy, sadness, angry…), attitudes (arrogant, submissive, cheerful…) and socio-geographical origin (French from Paris, Toulouse, Lille…), but also something quite important for speech comprehension: the prosodic structure of the sentence, prerequisite for an efficient syntactic analysis of running speech by speakers. This prosodic structure is defined by dependency relations between the sentence accent phrases, relations indicated by melodic contours located on vowels of stressed syllables.

In French, a non-lexically stressed language, stressed syllables (excluding emphatic stress) are located of the final syllable of some word (not necessarily a content word, i.e., a verb, a noun, an adjective, or an adverb) with a rhythmic constrain limiting their interval between 250 ms and some 1250-1350 ms in running speech. They determine the right boundary of the minimal prosodic units instantiated by accent phrases. It has been shown (Martin 2018) that the actual realization of stressed syllables depends on the speaker speech rate, leading to accent phrases containing 8 to 9 syllables for a fast speaker, and only 4 to 6 syllables for a slow speaker.

2. *The role of the prosodic structure in sentence comprehension*

Contrary to isolated sentences, whose acoustic image can be kept in memory for up to 20 or 30 seconds, sentences in running speech have a limited short-time memory of about 2 or 3 seconds. Since this limit does not give sufficient time for listeners to perform a syntactic analysis applied to the perceived string of words, the sentence prosodic structure is essential for language comprehension, as it provides a temporal and structural frame for syntactic decoding (cf. *syntactic bootstrapping*).

Besides, it has been experimentally shown that the perception of accent phrases is synchronized by delta brain oscillations, which explains why the duration of accent phrases in non-lexically stressed languages such as French and Korean is limited to a maximum of 1250-1350 ms (Martin 2018b), the minimal value of 250 ms referring to the time gap between consecutive stressed syllables. It explains as well why actual stress realizations (again excluding emphatic stress) is depending on speech rate (the average duration being 500-600 ms with accent phrases containing 4 to 5 syllables).

2.1 The independent prosodic structure

Contrary to the dominant Autosegmental-Metrical model which envisions prosodic events as percolating from syntax, the analysis proposed here assumes that the prosodic structure is planned and realized by the speaker before syntax in the sentence generation process at least for 3 to 4 accent phrases sequences. Hence, the phonological description of prosodic events should proceed independently from prosodic properties alone.

It is furthermore assumed here that pitch accent, i.e., the melodic movements located on stressed vowels, do interact with each other, as they indicate dependency relations between accent phrases. However, for French, a non-lexically stressed language, pitch accents and boundary tones are in syncretism and are aligned on the same stressed vowels.

2.2 Description of melodic contours

Prosodic events are described efficiently:

1. From the localization on stressed syllables vowels, excluding emphatic stress.
2. By the categorization of stressed vowels melodic changes according to the following parameters:
   a. Sentence final (reaching the lowest or highest level in the sentence).
   b. Rising or falling.
   c. Above or below the glissando threshold

The glissando threshold determines the limit above which a melodic change is perceived and under which a static tone is perceived. The glissando and the glissando threshold can be approximated from the fundamantal frequency curve F0 in Herz by the formula (Rossi 1971):

Semitone = 12 * (log(F0/100.0)) / log(2.0)
Glissando = (Semitone2-Semitone1) / (t2-t1)
Glissando threshold: between 0,16 / t^2 and 0.32 / t^2

Applying the criteria above, the retained contour categories are:

Cneu → neutralized, rising or falling below the glissando threshold
Cfal ↘ falling, above the glissando threshold
Cris ↗ rising, above the glissando threshold
Cfal# ↘# falling above the glissando threshold, before a pause longer than 250 ms
Cdec ↓ final contour déclarative (lowest frequency)
C0n ← final contour déclarative postnucleus (same as Cneu)
Cint ↑ sentence final contour interrogative (highest frequency)
Cin ↑ sentence final contour interrogative postnucleus (same as Cint).

The sentence nucleus is defined as a segment of a sentence that can constitue a complete well-formed sentence by itself in isolation. It ends with either a terminal conclusive declarative or interrogative contour. The nucleus is eventually followed by one (or more) segment ended by a neutralized contour C0n (declarative case) or interrogative Cin (interrogative case). C0n has the same realization that Cneu, whereas Cin is similar to Cint.

Using prosodic annotation adapted from F-ToBI (Delais et al. 2015) and integrating the glissando threshold, the retained prosodic events categories are:

H*/L* neutralized, under the glissando threshold
L*L- falling, above the glissando threshold
H*H- rising, above the glissando threshold
L*# falling > glissando threshold, before a pause > 250 ms
L*L% final conclusive declarative (lowest frequency)
H*/L* post final declarative
H*H% final conclusive interrogative (highest frequency)
H*H% post final interrogative

2.3 Dependency rules

The sentence prosodic structure is indicated by dependency relations between accent phrases (group of words with only one non-emphatic stressed syllable). These dependency relations are specified by pitch movements aligned on stressed vowels. They determine a hierarchical grouping of accent phrases which define the sentence prosodic structure. For French, dependency rules are as follows (⇨ and ⇦ indicate the direction of the dependency):

Cneu → ⇨ {Cfal ↘, Cris ↗, Cfal# ↘#, Cdec ↓, Cint ↑}
Cfal ↘ ⇨ {Cris ↗, Cfal# ↘#}
Cris ↗ ⇨ Cdec ↓
Cfal# ↘# ⇨ Cdec ↓
Cdec ↓ ⇦ C0n ← Declarative postnucleus
Cint ↑ ⇦ Cin ↑ Interrogative postnucleus

For example, the dependency rule Cfal ⇨{Cris, Cfal#} indicates that the presence of a falling contour Cfal above the glissando threshold depends on the occurrence of either a rising Cris or falling contour before pause Cfal# later in the sentence (dependency "to the right"). It indicates the regrouping of all already existing groups of accent phrases with the last one containing Cfal, with all already existing groups of accent phrases where the last one contains either Cris or Cfal#. As there is no dependency of Cfal towards Cdec, Cfal cannot immediately be followed in a sentence by Cdec.

3. *The RATP-DECODA corpus*

RATP-DECODA is a Customer Care Service corpus part of the ORFEO project (2020). It includes 1988 recordings of client requests made between 2009 and 2011 to the RATP call center (Régie Autonome des Transports Parisiens).

The total duration of these calls is about 98 hours, with an average of 177 seconds per call. Recordings were made in a mp3 format, converted in the Orfeo corpus into stereo 16 bit 8000 Hz sampling rate, and for the present research in mono 16 bit 22050 Hz. The files are originally annotated in word and phone segments, in trs (Transcriber) and also into TextGrid (Praat) format.

Conversations usually involve 2 participants: a client, the customer agent and eventually a service voice. Call types as noted in Brechet et al. (2012) were: Info Traffic 22.5%; Route planning 17.2%; Lost and Found 15.9%; Registration card 11.4; Timetable 4.6%; Ticket 4.5%; Specialized calls 4.5%; empty 3.6%; New registration 3.4%; Price info 3.0%.

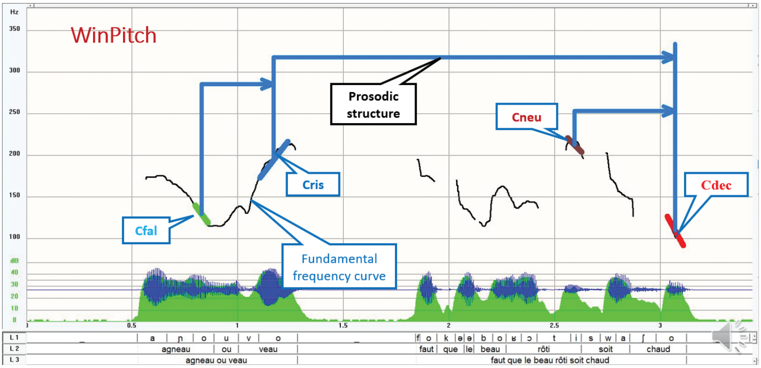
4. *Acoustic analysis*

The telephone landline also used by callers and the mp3 compressed recording format limit the actual sound bandwidth to about 120-3900 Hz. This may possibly affect the reliability of pitch trackers and automatic segmentation programs used for word and phone annotations, especially for male voices.

Nevertheless, using dedicated graphic functions of the speech analysis software WinPitch (2020), errors of segmentation were easily corrected manually, and pitch detection errors have been monitored thanks to the simultaneous display of a narrow band spectrogram allowing more reliable annotations (Martin 2018a).

An example of acoustic analysis using WinPitch is given Fig. 1, with the automatic display of the prosodic structure defined by pitch movements located on stressed vowels. The segmentation into words and API phones is done automatically by comparison with a TTS voice generating the text to analyze.

Figure 1



[*Agn***EAU**] Cfal↘ [*ou v***EAU**]↗ Cris [*faut que le beau rôt***I**] Cneu→ [*soit ch***AU***d*] Cdec↓ "Lamb or veal the beautiful roast must be hot". (stressed vowels are in bold capital), accent phrases are in brackets)

5. *Predicted configurations of the sentence prosodic structure*

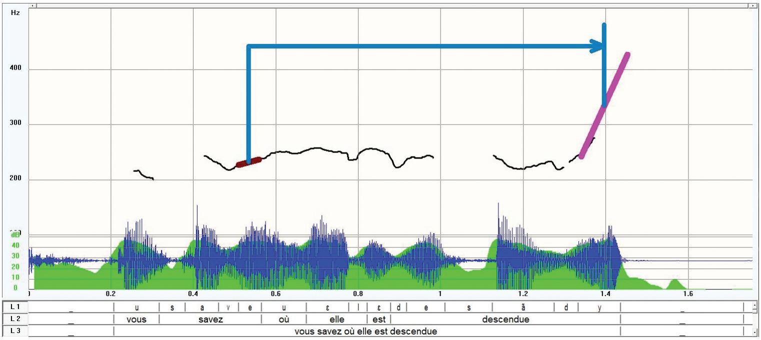
The four possible configurations of an interrogative sentence are:

a. Interrogative questions with or without a morphosyntactic marker ending with a rising terminal conclusive contour.
b. Declarative questions with a morphosyntactic marker falling terminal conclusive contour.
c. Interrogative postnucleus rising contour after the nucleus terminal conclusive interrogative.
d. Declarative postnucleus flat contour after the nucleus terminal conclusive declarative.

5.1 Interrogative questions rising terminal conclusive contour

Fig. 2 and 3 show an interrogative modality mode indicated by a rising terminal interrogative contour only. The first accent phrase of both examples is terminated by a neutralized contour Cneu necessary and sufficient in the configuration of the prosodic structure. The realization of an Cfal above the glissando threshold, while possibly attested, would be redundant.

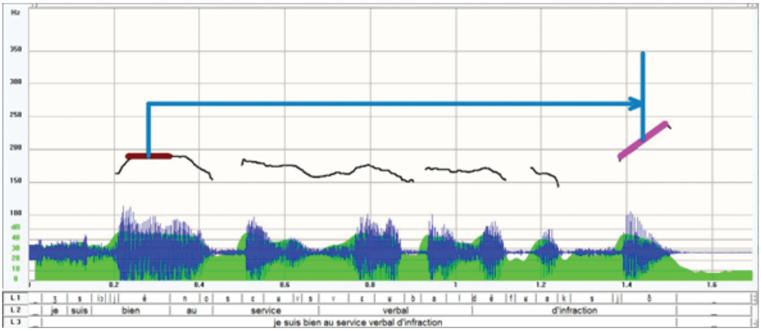
Figure 2



[*Vous sav**Ez***] Cneu → [*où elle est descend**Ue** ?*] Cint ↑ “Do you know where she got off?” (20091112_RATP_SCD_0353 25.852 s 27.609 s).

An interrogative questions marked by a terminal rising interrogative contour Cint on the last accent phrase, the first [*Vous sav**Ez***] being ended by a neutralized contour Cneu, as any contour phonologically contrasting with Cdec or Cint would adequately indicate a simple prosodic structure with only two accent phrases.

Figure 3



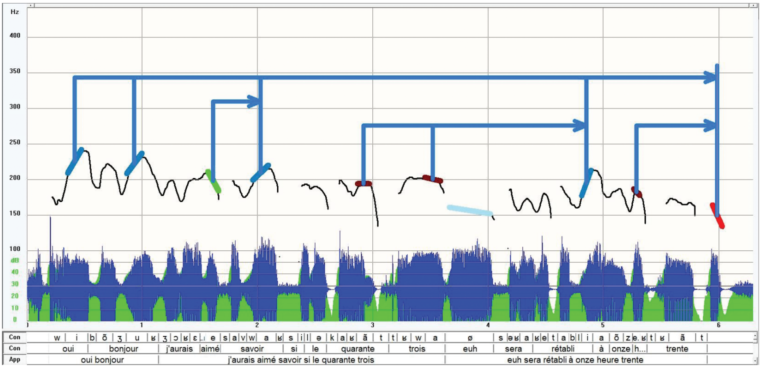
[*Je suis bi**EN***] Cneu → [*au service verbal d'infracti**ON** ?*] ↑Cint “Am I on verbal offense service?” (20091112_RATP_SCD_0215 8.057 s 9.648 s).

Another interrogative questions marked by a terminal rising interrogative contour Cint and a neutralized Cneu on the first accent phrase [*Je suis bi**EN***].

5.2 Declarative questions falling terminal conclusive contour

Examples of Fig. 4 and 5 are declarative sentences whose interrogative modality is only referred by the text (*j'aurais aimé savoir* "I would have liked to know" and *je voudrais savoir* "I would like to know".

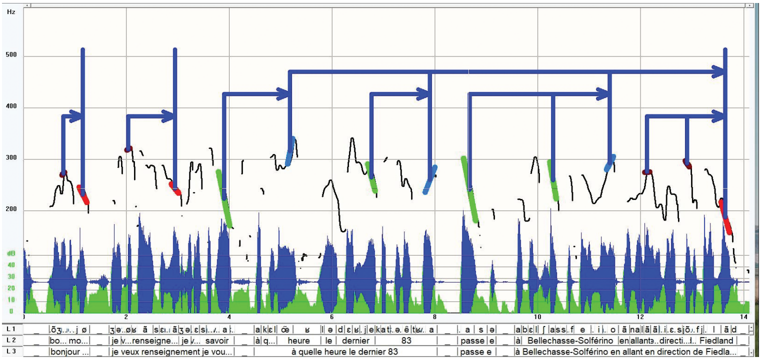
Figure 4



[*OuI*] Cris ↗ [*bonjOUr*] Cris ↗ [[*j'aurais aimÉ*] Cfal ↘ *savOIr*] Cris ↗ [*si le quarANte trOIs euh sera rétablI*] Cris ↗ [*à onze hEUres trENte*] Cdec ↓ "Yes hello I would have liked to know if the forty three uh will be restored to eleven thirty" (20091112_RATP_SCD_0357 6.236 s 12.339 s).

This is an example of a declarative prosodic structure associated with a text with an indirect interrogation instantiated by *j'aurais aimé savoir*.

Figure 5



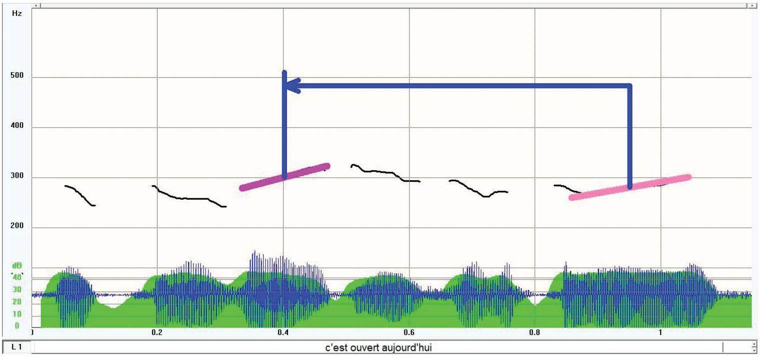
[[*Je voudrais savOIr*] Cfal ↘ *à quelle hEUre*] Cris ↗ [[*le derniEr*] Cfal ↘ *quatre vingt-trOIs*] Cris ↗ [[*pAsse*] Cfal ↘ [*à BellechAsse*] Cfal ↘ *SolférinO*] Cris ↗ [*en allANt en directiON de FriedlANd*] Cdec ↓ "I would like to know at what time the last 83 passes Bellechasse-Solférino heading towards Friedland". (20091112_RATP_SCD_0336 6.686 s 20.790 s).

Another example where both the prosodic structure and the associated text carry a declarative modality. The acoustic analysis shows the melodic slope contrast involving the dependency rule Cfal → Cris characteristic of French marking the dependency between accent phrases: [[*Je voudrais sav**OI**r*] *à quelle h**EU**re*], [[*le derni**E**r*] *quatre vingt-tr**OI**s*], [[*p**A**sse*] [*à Bellech**A**sse*] *Solférin**O***] Cris ↗.

5.3 Interrogative postnucleus rising contour after the nucleus terminal conclusive interrogative

Examples of Fig. 6 and 7 are interrogative sentences, deprived from morphosyntactic interrogative markers, and whose modality is indicated by a terminal conclusive interrogative contour Cint, followed by a postnucleus ending with a copy of Cint. This is an example of a *focus-topic* configuration.

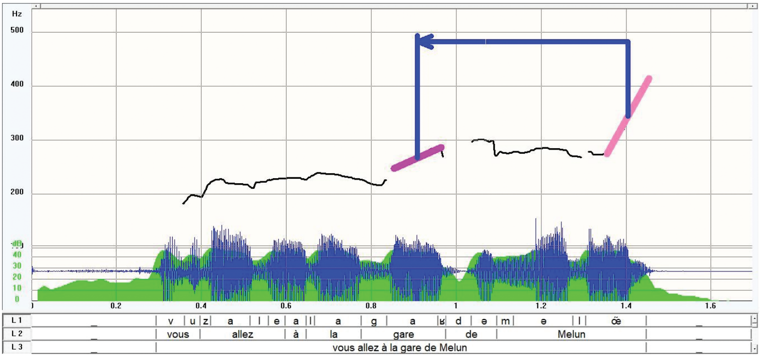
Figure 6



[*C'est ouvErt*] Cint ↑[*aujourd'huI ?*] Cin ↑ "Are you open today?" (20091112_RATP_SCD_0060 18.045 s 21.860 s 2.560 s 3.705 s).

In this example, [*C'est ouv**E**rt*] is the nucleus, and [*aujourd'hu**I** ?*] the interrogative postnucleus, ended by a copy of the first interrogative Cint of the first accent phrase [*C'est ouv**E**rt*].

Figure 7



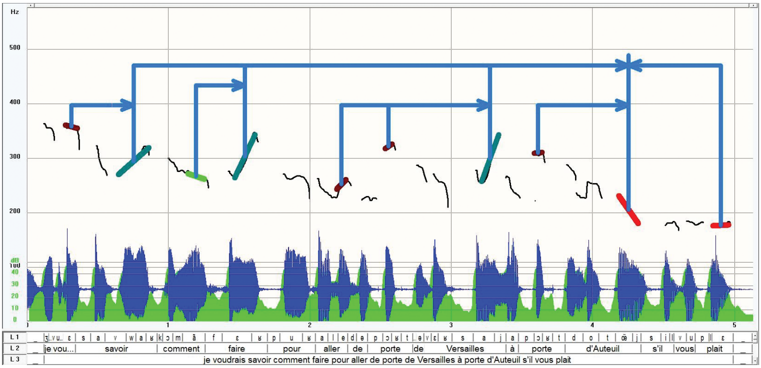
[*Vous allez à la gAre*] Cint ↑[*de MelUN ?*] Cin ↑ "Are you going to Melun station?" (20091112_RATP_SCD_0023 16.420 s 18.118 s).

The nucleus is [*Vous allez à la gAre*] and the postnucleus [*de MelUN ?*].

5.4 Declarative postnucleus flat contour after the nucleus terminal conclusive declarative

Fig. 8 and 9 show examples of the declarative counterpart of preceding cases of Fig. 6 and 7. The postnucleus *s'il vous plait* "please" are ended with a neutralized contour Cneu in a *focus-topic* declarative configuration.

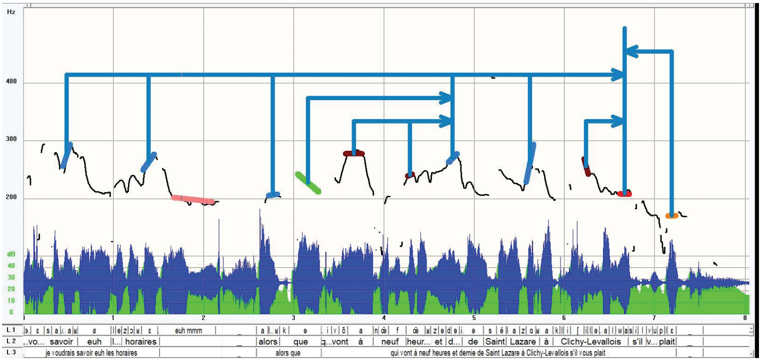
Figure 8



[*Je voudrAIs savOIr*] Cris ↗ [[*commENt*] Cfal↘ [*fAIre*] Cris ↗ [*pour allER de POrte de VerAilles*] Cris ↗ [*à POrte d'AutEUil*] Cdec ↓ [*s'il vous plAIt*] C0n ↓ "I would like to know how to get from Porte de Versailles to Porte d'Auteuil please" (20091112_RATP_SCD_0344 8.067 s 13.197 s).

Here the nucleus is [*Je voudrAIs savOIr commENt fAIre pour allER de POrte de VerAilles à POrte d'AutEUil*] and the postnucleus [*s'il vous plAIt*].

Figure 9



[*Je voudrais savOIr*] Cris ↗ euh [*les horAIres*] Cris ↗ *euh* [*alOrs*] Cris ↗ [[*que*] Cfal ↘ *qui vONt à neuf hEUres et demIe*] Cris ↗ [*de Saint Lazarre*] Cris ↗ [*à ClichY-Levallois*] Cdec ↓ [*s'il vous plAIt*]
C0n ← "I would like to know uh the hours uh mmm whereas which go at half past nine from Saint-Lazare to Clichy-Levallois please" (20091112_RATP_SCD_0070 8.794 s 16.839 s).

In this example, [*Je voudrais savOIr* euh *les horAIres euh alOrs que qui vONt à neuf hEUres et demIe de Saint Lazarre à ClichY-Levallois*] is the nucleus and [*s'il vous plAIt*] the postnucleus.

6. *In summary*

These few examples aim to show that sentence intonation is not "the cherry on the syntactic tree", but on the contrary shapes speech production both on the paratactic and syntactic axis. Its importance in actual speech perception stems essentially from the fact that listeners must quickly analyze the content of a sentence, given the allowed short time of speech memory given in continuous speech (2 to 3 seconds).

References

Bechet, Frederic & Maza, Benjamin & Bigouroux, Nicolas & Bazillon, Thierry & El-Bèze, Marc & De Mori, Renato & Arbillot, Eric. 2012. DECODA: a call-centre human-human spoken conversation corpus. In

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 1343-1347.

Delais-Roussarie, Elisabeth & Post, Brechtje & Avanzi, Mathieu & Buthke, Carolin & Di Cristo, Albert & Feldhausen, Ingo & Jun, Sun-Ah & Martin, Philippe & Meisenburg, Trudel & Rialland, Annie & Sichel-Bazin, Rafèu & Yoo, Hi-Yon. 2015. Intonational Phonology of French: Developing a ToBI system for French. In Sónia Frota & Pilar Prieto (eds.), *Intonation in Romance*, 63-100. Oxford: Oxford University Press.

Martin, Philippe. 2018a. Prosodic annotation in adverse recording conditions. In De Dominicis, Amedeo (a cura di), *International Workshop. Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis*, Accademia Nazionale dei Lincei.

Martin, Philippe. 2018b. *Intonation, structure prosodique et ondes cérébrales*, London: ISTE.

ORFEO. 2020. Corpus d'étude pour le français contemporain https://www.projet-orfeo.fr/

Rossi, Mario. 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica* 23, 1–33.

WinPitch. 2020. Computer program, www.winpitch.com