Noname manuscript No.
(will be inserted by the editor)

MUFASA: MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis

Aghilas Sini · Damien Lolive · Elisabeth Delais-Roussarie

Received: date / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords First keyword · Second keyword · More

This chapter is an extended version of the work described in "SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis" presented at *LREC18* [**?**].

We propose a new corpus of audiobooks, containing about 600 hours of speech (silence and pauses included). We present the annotation methodology and exploratory experiments that we conducted in order to have a clear idea of the expressiveness carried by this kind of data at text level and the corresponding speech. The initial corpus called SynPaFlex-corpus contains a single female speaker, but we found that this data was not sufficient to characterize expressivity in all its complexity. So we decided to extend the corpus to multi-speakers in order to consider speakers reading strategy perspectives. This new version is named MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA).

We designed these corpora by considering three goals. The first goal consists of exploring the text related features such as morpho-syntax, semantics and phonology, discourses types, and literary genres. The second one aims to analyze and to characterize the intra-speaker prosodic patterns related to the phenomena due to reading aloud a long text, and the last goal focused on the inter-speakers variation exploration/characterization.

Aghilas Sini
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

Damien Lolive
second address

Elisabeth Delais-Roussarie
second address

We compare the MUFASA-Corpus with other existing corpora dedicated to TTS to figure out uncovered aspects of expressivity.

1 SynPaFlex Corpus

It seems impossible to describe the expressivity of speech by a finite number of rules that cover all the exceptions, factors, and contexts. Data-driven techniques seem to be a well adapted solution for this kind of challenge and the availability of data makes their use possible. Even though finding appropriate data, tidy data is, in some sense, also challenging and raises other difficulties such as the dependency between the model and the data (the data quality impact on the model performance model). In this thesis, we are especially interested in audiobook data.

1.1 Motivation

The SynPaFlex corpus is an audiobooks corpus of single female voice. The data were collected according to this criteria:

- Availability of a large quantity of data uttered by a single speaker;
- Availability of the corresponding written texts;
- Good audio signal quality and homogeneous voice;
- Various discourse styles and literary genres;
- Conveying emotions in speech.

1.2 Relation to previous work

Many corpora dedicated to the synthesis of speech already exist. Most of them are in English. For French, most corpora do not exceed ten hours of read speech by only one speaker. Most professional corpora recorded to build a synthesized voice are often sentence-by-sentence records. Except the GV-LEX [1] corpus for which the author seeks to characterize expressiveness beyond the sentence. However, this last work is entirely dedicated to a particular genre that is fantastic tales dedicated to young children.

The whole annotation pipeline were handled with the ROOTS toolkit,that allows storing various types of data in a coherent way using sequences and relations. This toolkit [2] allowed us to incrementally add new information to the corpus.

Once audio data have been selected and the corresponding texts have been collected, a few manual operations have been applied to simplify further processing. Notably, as recordings were performed in different technical and environmental conditions, loudness has been harmonized using the *FreeLCS* tool[1]. Despite of that, audio data acoustic features remain more or less heterogeneous. Therefore, analyzing the intensity of audio files is now possible.

As texts were coming from diverse sources, their formats were unified. Then the exact orthographic transcriptions of the readings were achieved by inserting the

[1] http://freelcs.sourceforge.net/

introductions and conclusions the speaker added in the recording, and by placing footnotes and end-of-book notes where they appear in the reading stream.

The next step has been to normalize the texts using rule-based techniques appropriate for the French language, and split them into paragraphs. For the rest of the process, we kept each chapter in a separate file so as to keep long term information accessible.

1.3 Data Collection and Pre-processing

Most of the texts collected are in the public domain. Two sources have been mainly used: the *Gutenberg* [2] project and the *Wikisource*[3] bookstore. Records have been collected along with the corresponding text in plain text format. Few manual adjustments were performed on the text to insure its correspondence to the audio files. The original text structure is respected. Most of the texts studied were published between the 17th and the 20th century.

In narrative or descriptive texts such as in novels, short stories and tales, the paragraph is considered as basic text unit. On the other hand, poems and fables are structured in verses. Consequently the utterance represents a verse.

Each utterance is tokenized then normalized, which consists of orthographically transcribing numbers and acronyms. This is done using rules set manually by experts. A syntactical analysis of all utterances is performed to establish the syntactic function of the words content.

The original audio files are mostly in MP3 format, with a sampling rate of 22.050 khz or 44.1 khz each of these samples being coded on 16-bit with a bit rate ranging from 64 to 128 kbps. All the recordings were converted to wav format with a sampling frequency of 22.05 khz in order to have a consistent corpus.

1.3.1 Speech Segmentation

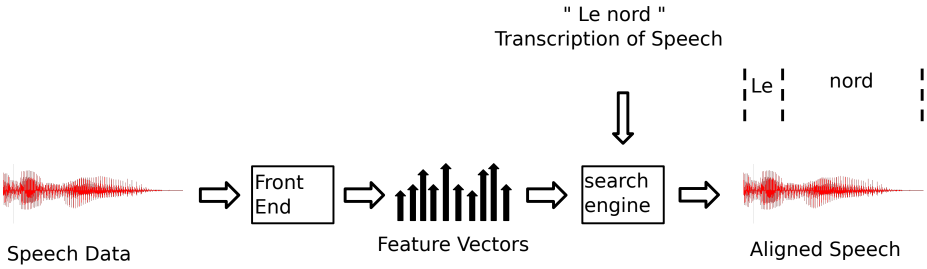


Fig. 1: Overview of the Speech Segmentation process

The broad phonetic transcription, based on the French subset of Sampa, has been extracted and aligned with the speech signal using JTrans [3].

[2] https://www.gutenberg.org/wiki/Main_Page
[3] https://fr.m.wikisource.org/wiki/Wikisource:Accueil

To evaluate the accuracy of the phone segmentation, an expert annotator performed a manual validation using Praat [4].

The evaluation process of the forced alignment consists in first generating the automatic segmentation and phonetic labels of selected dataset based on this annotation. The annotator has to add a sequence corresponding to the correction.

Since there is only one speaker, half an hour of the SynPaFlex corpus has been taken into account to evaluate the quality of phone labels and boundaries. The set of data used for the evaluation task has been selected respecting the proportions of the different literary genres in the corpus.

Results related to the validation are presented in Table 1. We can observe that the Phoneme Error Rate (PER) is low for every literary genres, and the average PER is 6.1%. Concerning the average alignment error, results are reported in the fourth column of Table 1. Globally, on average, the error is 11ms.

As far as errors on label assignment are concerned, they mostly occur on vocalic segments. Most of the deletion observed involve /@/ (83.31%), this phoneme being generally optional in French. The majority of substitutions concerns mid vowels (37.04% for the substitution of /E/ by /e/, and 31.04% for /o/ by /O/), these realizations being the result of a specific pronunciation or simply phonetization errors.

As for boundary alignment, in 77.17% of cases, boundaries are misplaced from less than 20ms. In poems, however, alignment errors are more important: for 35% of the vowels, boundaries have been shifted by more than 20ms. It could be explained by two distinct factors. First, the speech rate is relatively slow in poems (with an average of 5 syllables/s) in comparison to other literary genres where the speech rate is of 6 syllables/s on average. Secondly, the acoustic models used to achieve the automatic segmentation [3] have been trained on the ESTER2 corpus [5] which is a French radio broadcasts corpus. The resulting models could thus be slightly not well-adapted for poem reading data.

To improve the segmentation performance we have tried two different ways:

- First way consists of adapting the default acoustic model to our speakers.
- For the second one, we have trained a new model using Montreal Forced Aligner [6] tool which is based on Kaldi[7] Speech Recognition. Although, this tool is easy to set up, it is difficult to align long speech data like the chapters. To face this problem, we had first to segment the chapters into utterances using JTrans. The utterances thus obtained are then used to learn the acoustic model and then align the corresponding transcriptions at the phoneme level.

Furthermore, we plan to use the Train& Align [8] online tool which seems to be more appropriate to our data. This tool proposes to train an acoustic model and to align the data at the same time which corresponds better to SynPaFlex corpus structure.

1.3.2 Linguistic Information

1.3.3 Acoustic and Prosodic Information

The speech signal is stored using a sampling frequency of 22.05 kHz. From the signal, we have extracted (i) the energy and 12 mel-frequency cepstral coefficients (MFCC 1-12) which we have added delta and delta-delta coefficients using [9],

	Validation subset	PER (%)	average alignment error (ms)
Novels	25m36s	5.8	11.5
Short stories	3m49s	7.1	9.4
Tales	2m47s	0.8	14.3
Fables	1m47s	6.5	12.1
Poems	1m07s	6.3	28.3
Total	35m52s	6.1	11.4

Table 1: Validation results for the segmentation step per literary genre : lengths of the validation subsets, Phoneme Error Rate (PER), and average alignment error.

Unit type	Number
Paragraphs	23 671
Sentences	54 393
Words	799 773
Orthographically distinct words	38 651
Phonemically distinct words	28 734
Non Stop Words	411 210
Syllables	1 154 714
Distinct syllables	8 910
Open	808 503
Closed	346 211
Phonemes	2 613 496
Distinct phonemes	33

Table 3: Amounts of linguistic units in the SynPaFlex corpus

(ii) the instantaneous fundamental frequency (F_0) using the ESPS get_f0 method implementing the algorithm presented in [10], and (iii) pitchmarks using our own software.

Additionally, we have added some prosody related features as the articulation rate (in syllables/s), the speech rate (in syllables/s), and F0 mean/min/max/range (in Hz) at the syllable and word levels. Since the corpus contains several speakers, we suggest to compute fundamental frequency in semi-tone[4] scale to be able to compare among speaker voices.

2 MUFASA Corpus

MUFASA is an extension of the SynPaFlex Corpus. The database has been collected and processed in a similar way of the SynPaFlex Corpus. Unlike the SynPaFlex Corpus, this corpus was collected from two different libraries i.e, LibriVox.org (LV)[5] and LitteratureAudio.com (LA)[6] entirely dedicated to French au-

[4] The logarithmic semitone scale seems to be the appropriate measure of the perceptual consequences of differences in fundamental frequency

[5] https://librivox.org/

[6] http://www.litteratureaudio.com/

diobooks. LA is not in the public domain, unlike LV, authorization is required and we have asked the administrator for authorization to use certain voices exclusively for research purposes.

2.1 Motivation

We decide to build MUFASA corpus for:

- Analyzing inter-speaker variations to have a better understanding and a more comprehensive view of the strategy adopted when reading audiobooks.
- Distinguishing the speaker-related characteristics from those related to texts.
- Finding strategies common to the various speakers, which makes it possible to extract prosodic structures appropriate to the reading of audiobooks.
- Considering the speaker's prosodic identity by characterizing the inter-speaker variability.

The MUFASA corpus is intended to be close to the LibriTTS[11] corpus, as both deal with the exponents of amateur audiobooks and contain several speakers. On the other hand, the two corpora differ in the fact that in MUFASA, each speaker is represented by at least two hours of speech, and the language of reference is French. Several recordings for the same text (parallel data) are provided, allowing an analysis of the difference between speakers without worrying about the linguistic characteristics.

2.2 The novelty of this work

This MUFASA Corpus offers the possibility to explore the expressivity in different way such as:

- Parallel data (**??**): same text recorded by different speakers.
- Certain famous French authors are well represented in the MUFASA corpus. This can be exploited to study author style.
- Enough data to characterize the style of speakers: as the first voice we favored voices that recorded more works, of different genres (poem, fable, tale, short story, and novel).
- Enough data to characterize the genre: In order to study and analyze the characteristics of genres regardless of speakers. The genre may convey a rather special expressiveness depending on the speaker and the authors.
- To compare professional and amateur recording, we also collect certain passages read by amateurs and professionals[7].

A summary of the contents of the MUFASA-Corpus is presented in the 4.

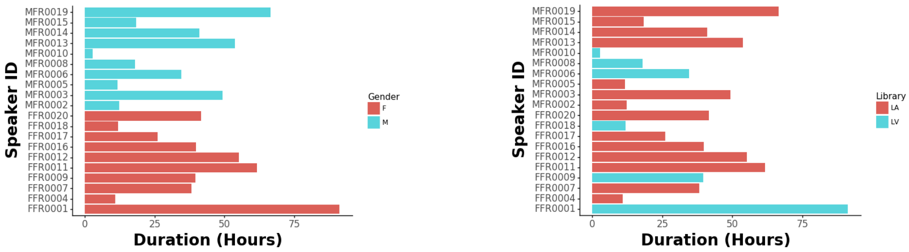
2.3 General Overview

MUFASA corpus contains twenty French speakers (10 Females/10 Males). **??** illustrate each speaker's duration proportion in the corpus. The speaker name is encoded as following (F/M: Female/Male, FR: French, ID:XXXX).

[7] recordings dedicated to the synthesis of speech, with favorable acoustic conditions and voice selection for this purpose unlike amateur recording.

Unite	Number
Utterances	79 242
Sentences	211 416
Average Sentence Length	24
Words	5 093 789
orthographically distinct	77 303

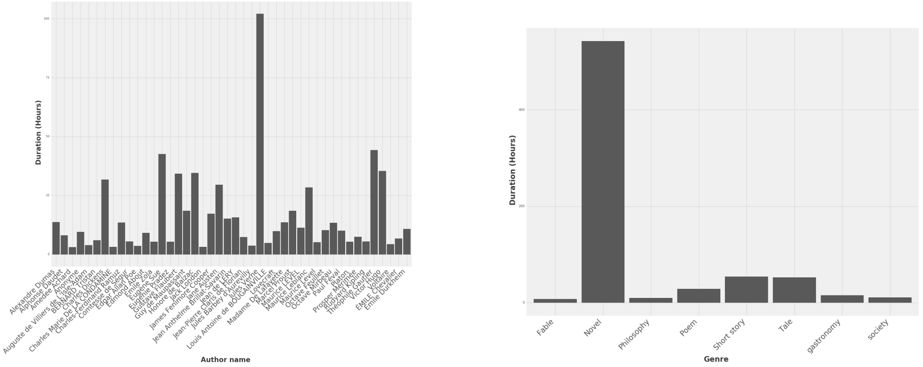
Table 4: The main linguistic content of MUFASA Corpus



(a) MUFASA Speakers duration distribution and gender labeling

(b) MUFASA Speakers duration distribution and library belong to.

Narrative genre such as novel, short story, and tale are the most frequent in the corpus as shown in 2b. Some author are well represented in the corpus 2a.



(a) Book author's distribution in MUFASA corpus

(b) Genre proportions in MUFASA corpus

For more details about the contents of the MUFASA see the **??**.

3 Gap between Text-to-Speech (TTS) designed corpora and amateur audiobook recording

Both SynPaFlex and MUFASA corpora were constructed to build and to improve TTS systems, whether it is concatenative, Statistical Parametric Speech Synthesis, or End-To-End (E2E) systems. Careful analysis of these databases that we called amateur audiobooks allows us to notice a difference between the quality of their recording and the recording quality of the databases made in a laboratory (ex. SIWIS French Speech Synthesis Database).

When we design a corpus for TTS (in a laboratory), we tend to be careful about the recording conditions, such as the microphone's position, acoustic properties of the recording room, and the reliability of recording materials. However, in amateur audiobooks, the lack of control over the recording conditions introduces an error on the measure of some prosodic parameters sensible to the noise. Beyond those signal processing gaps due to signal quality, there are differences between guided records data that we will consider as a professional recording and amateur ones at the suprasegmental level that we will try to highlight in this work.

3.1 Data and features extraction

3.1.1 Data

To figure out the differences between professional and amateur recordings, we have investigated a sub-corpus that contains two chapters from two separate novels. Both chapters have been read by three different speakers, i.e., three recordings including a single professional record at each time and two amateur recordings.

5 summarizes the linguistic contents and the duration of the subset used for conducting the experiments aiming to measure the gap between the amateur and professional recording.

Book title, author	Nbr. Utts	Nbr. Wrd	Nbr. Syl	Speaker	Recording Type	Duration
Vingt mille lieues sous les mers Chapter 3, Jules Vern	56	1830	2774	MFR0019	A	10min 56sec
				FFR0001	A	11min 42sec
				SFS	P	10min 26sec
Mademoiselle Albertine est partie Marcel Proust	74	2460	3518	FFR0011	A	17min 34sec
				FFR0020	A	14min 42sec
				PODALYDES	P	13 min 66sec

Table 5: Subcorpus contents. The first column corresponds to the title of the novel, and author's name. *Nbr. Utts* is the number of utterances(sentences), *Nbr. Wrd* is the number of words in the chapter and *Nbr. Syl* the number of syllables. The recording type (P) refers to a professional recording, whereas (A) refers to an amateur record. The Siwis French Speech (SFS) voice is the female voice of The SIWIS French Speech Synthesis Database[8]. PODALYDES is a male voice. The speakers FFR0001, FFR0011, FFR0020, and MFR0019 are included in the MUFASA corpus.

3.1.2 Features extraction

We choose two prosodic parameters to study the difference between the speakers: 1) the average length of pauses within utterances, and their distribution 2) Subharmonic-to-Harmonic Ratio (SHR) and the vowel trapezoid as voice quality features.

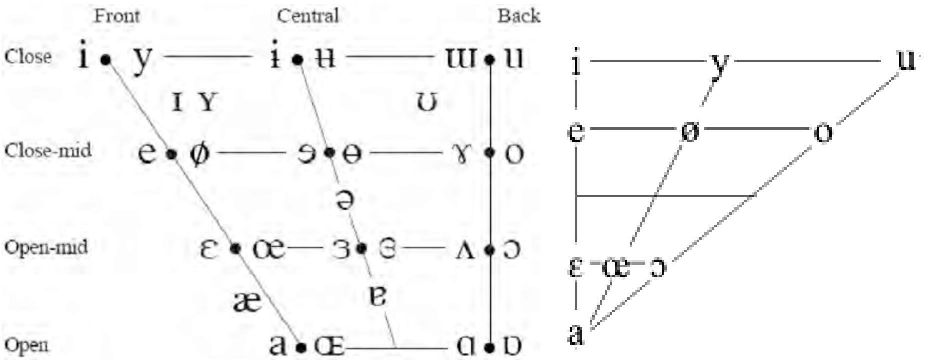
- We chose to study the pauses' duration and their distribution in an audiobook read by three different speakers to see if there is a difference between the so-called professional and amateur speakers. Because the pauses are good indicators of speech style and implicitly how the data are recorded and prepared, we hypothesize that in a professional recording, the pauses' position, frequency, and duration are controlled and regularized. In contrast, in the amateur record, the preparation level is lower, thus a broader range of variation, and this tends to influence the other prosodic parameters such as articulation rate and F0-range.
- The vowel trapezoid is an articulatory schema that represents all possible vocalic timbers of the human vocal tract. 2a describes all the timbres of the oral vowels of the world's language. This space is divided according to language in functional units. In French, there are ten functional timbers of oral vowels (cf. 2b). This schema space is formed along two first formants, F1/F2. In a prepared speech, such as read speech, the contrast between the three cardinal vowels (/i/, /a/, and /u/) [12] is usually studied for characterizing articulatory behaviour. The cardinal vowels represent the boundaries of the vocalic area[13].
- According to [14,15], alternate cycles[9] (alternating in amplitude or period, or both) in speech signal make the determination of pitch more difficult. [16] claims that the alternating cycles in the time domain are manifested by the presence of subharmonics in the frequency domain. Furthermore, the magnitude of subharmonics with respect to harmonics reflects the degree of deviation from modal voice. From that, a new parameter called Subharmonic-to-Harmonic Ratio (SHR) was introduced to describe the amplitude ratio between subharmonics and harmonics. The [15], present a pitch perception study. This experiment consisted of asking participants (expert listeners) to determine the pitch of synthesized vowels with alternate cycles through amplitude and frequency modulation. The results show that pitch perception is closely related to SHR. This experiment aims to find the relationship between the perceived pitch and SHR. This parameter is used for describing voice quality as well as for classifying voice production mode for one speaker or comparing voice quality for different speakers reliable for perceptual quality of voice.

3.2 Results

3.2.1 Voice quality analysis

We select all the voiced frames and calculate the SHR frequency distribution (see 6 according to [14], when SHR is in the medium range, especially (0.2, 0.4], perceived pitch becomes ambiguous. Correspondingly, in 6, MFR0019 and FFR0001 have higher SHR percentage in the range of (0.2, 0.4] among three speakers, whereas Siwis French Speech (SFS) female speaker has the lowest. Visual inspection and

[9] *"For normal speech, alternate cycles usually appear in creaky voice or voice with laryngealization, which are often characterized as perceptually rough voices. In pathological voice, alternate cycles can be found even in normal mode of production." [14]*



(a) A schematic of vowel triangle

(b) A schematic of vowel triangle (French)

listening to the speech waveform confirm that MFR0019 has indeed more ”irregular” speech cycles and appears to have low and rough voice, whereas SFS's speech seems to be much more “regular”. Similarly FFR0001 has more creaky voice than SFS despite the average pitch of FFR0001 is much higher. 6 also show that the professional speaker have greater number of SHRs in the range of (-0.2, 0.0] compared with amateur speakers. This indicates that professional speaker (SFS) speech might have greater amount of small amplitude or period fluctuations, which, however, are not significant enough to affect pitch perception.

Speaker	Gender	Types of recording	(-0.2, 0.0] (%)	(0.0, 0.2](%)	(0.2, 0.4](%)	(0.4, 0.6](%)	(0.6, 0.8](%)	(0.8, 1.0](%)
Marcel Proust , À la Recherche du Temps perdu, Albertine disparue								
FFR0011	Female	Amateur	78.33	1.90	3.15	4.78	5.96	5.87
FFR0020	Female	Amateur	78.14	1.83	2.26	4.73	6.59	6.45
podalydes	Male	Professional	82.45	2.64	3.49	3.90	3.89	3.63
Jules Verne, Vingt Mille Lieues sous les mers.								
FFR0001	Female	Amateur	60.94	3.97	8.12	8.42	8.83	9.68
MFR0019	Male	Amateur	61.11	5.31	10.24	10.66	7.25	5.41
SFS	Female	Professional	87.46	1.20	1.40	2.07	2.37	5.48

Table 6: Subharmonic-to-Harmonic Ratio distribution of the subcorpus speakers . For each speaker, we select all the voiced frames and calculate the Subharmonic-to-Harmonic Ratio frequency distribution.

The table shows that the SFS female voice has a high percentage of SHR close to 0 and the very low percentage of medium values (0.2 - 0.4] which means that the pitch of this voice is quite easy to perceive by annotators if we consider the study [1,25]. In comparison with other voices FF0019/FFR0001 (reading the same text), the percentage of medium values is high and the percentage of values of SHR close to 0.0 is lower. In the second example considered in this study, we did not find a significant result between the professional voice (Podalydes) and amateurs' voices (FFR0020, FFR0011). We have conducted an informal perceptual test, where we asked an expert annotator to determine the pitch of the vowels /i/ and /a/ preceded by /p/,/t/,/k/ present in the subcorpus (7). This perceptual test has confirmed that among the six speakers, SFS voice is the easiest to determine the pitch.

book title, author	/ta/	/ka/	/pa/	/ti/	/ki/	/pi/
Vingt mille lieues sous les mers Chapter 3, Jules Vern	21	33	51	28	27	9
Mademoiselle Albertine est partie Marcel Proust	30	22	87	58	34	2

Table 7: The frequency of the {/ka/,/ta/,/pa/,/ti/,/ti/,/pi/} in the considered dataset, that have been manually annotated in terms of pitch amplitude.

The analysis of the vocalic trapeze in 2, shows that SFS voice makes an important contrast among the vowels with minor variation, whereas, for the other speakers the contrast is not clear. The subjective assessment of the samples present in the 7 has also confirmed that among the six speakers, SFS is the speaker that tend to over-articulate the cardinal vowels. For instance there is a strong variation of */u/* and */i/* along F2 which implies a significant overlap between considered vowels.

3.2.2 Pauses

There is no difference between the three speakers concerning the number of pauses, within utterances according to 4a, but the 4b shows that the professional speaker SFS makes short pauses (average of 250 ms) and in constant manner. Whereas the two other speakers seem to produce long pauses ($\tilde{5}$00 ms for FFR0001 and $\tilde{4}$80 ms for MFR0019), with important variation.

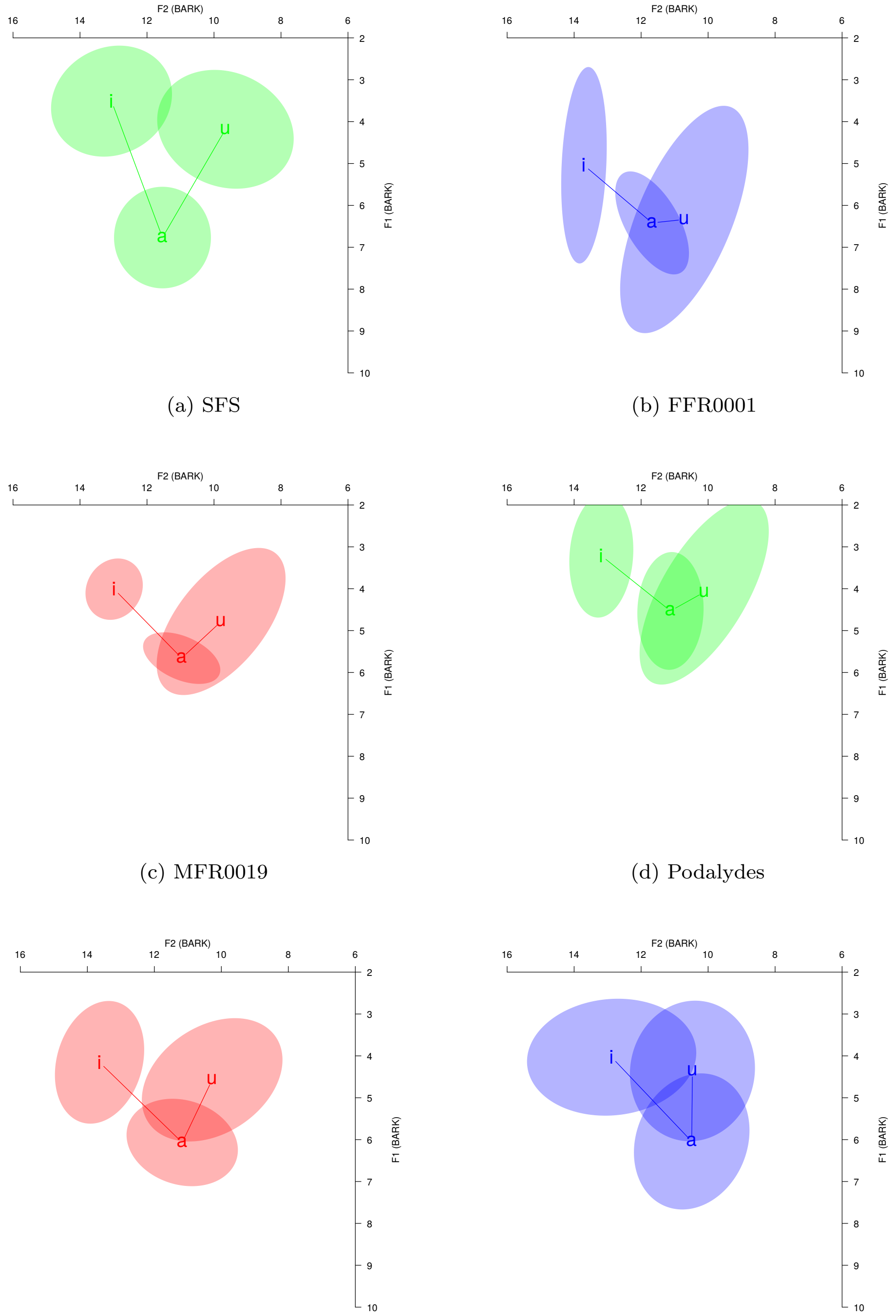
3.3 Discussion

In this study, we compared extracts of MUFASA corpus considering amateurs recordings with professional recordings through three prosodic parameters: SHR, vocal trapezoid, and pauses. The results show that professional data have stable pause durations and good voice quality. In contrast, amateur recordings tend to have inconsistent pause durations with considerable variation, and low recording condition compared to professional.

From these results, we can see that data recorded for speech synthesis has a couple of properties that distinguish them from amateur audiobooks, and professional speaker recordings that are not dedicated to speech synthesis. Despite the quality of audiobook data, professional data dedicated to speech synthesis is better to build a model of good quality.

4 A Phonetic Comparison between Different French Corpora Types

The main purpose of this section is to highlight two representative properties of speech style carried by the audiobook corpus, which are the duration of the vowels and the values of the two first formants (F1, F2) of the cardinal vowels (/a/, /i/ and /u/). This second parameter has been chosen as the structure of



(a) SFS

(b) FFR0001

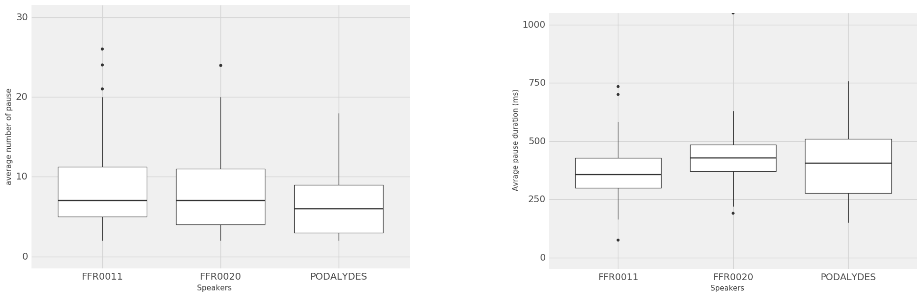
(c) MFR0019

(d) Podalydes

(e) FFR0011

(f) FFR0020

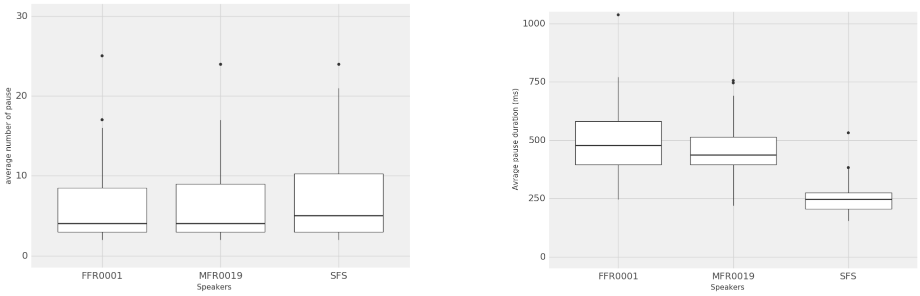
Fig. 2: the vowel trapezoids of the three cardinal vowels /u/, /i/, and /a/



(a) Average number of pauses per utterance per speaker

(b) Average pause duration

Fig. 3: Pauses distribution and average duration for *"Mademoiselle Albertine est partie"*



(a) Average number of pauses per utterance per speaker

(b) Average pause duration

Fig. 4: Pauses distribution and average duration for *" Vingt mille lieues sous les mers Chapter 3"*

the vowel trapezoids follows a logical organization, linked to the contrast and the stability of articulation. To highlight these properties, we conducted a comparative graphical and statistical analysis between a representative subset of MUFASA corpus and four other different types of French corpora(BREF [17], ESTER [18], RHAPSODIE [19], and NCCFr [20]). While each of the considered corpora have been recorded for a different purpose and in varying conditions, using them allows us to evaluate the MUFASA-Corpus.

4.1 Corpus design

To be able to compare the data of the MUFASA Corpus with that of the other corpora, the vowels studied must appear with similar proportions in each of the extracts from the studied corpus. Ideally, even the context has to be similar. So the targeted vowels were therefore placed in an open syllable of CV structure and preceded by the consonants /p/, /t/, or /k/. The choice of consonants /p/, /t/, and /k/ is justified by the fact that this type of consonants facilitates the segmentation of vowels since their limits do not merge with those of vowels. Thus, three syllabic contexts were chosen for the three vowels studied.

	Type of data	Number Of Speakers	number of /a/	number of /i/	number of /u/	Duration (sec)
MUFASA	Audiobook	9 (4F/5M)	1662	821	321	4130
BREF	Newspaper	9 (5F/4M)	1045	634	419	4167
ESTER	Radio broadcast news	10 (1F/9M)	1021	1036	260	4201
RHAPSODIE	Monologues (Various Style)	30 (13/17)	1353	878	673	4046
NCCFr	Casual Conversion	10 (5F/5M)	1372	1042	103	4369

Table 8: The set of extracts for conducting a comparative study.

We will briefly describe the used dataset :

- The MUFASA extract contains nine different speakers reading distinct novels. We have selected the speakers with varying strategies of narration based on two criteria vowel duration and the average F0 amplitude.
- Bref [17]: This read speech corpus designed for speech recognition (speaker-dependent and independent case), and it consists of texts selected from French newspapers, *Le Monde*. The extract contains nine speakers.
- Ester [18]: The used data are made up of France Inter radio broadcast news recorded in 1998, covering ten speakers.
- RHAPSODIE [19] is a spoken french corpus annotated in terms of prosody and syntax. From this corpus, we extract only the monologues and the private domain. This subset contains short clips ($\tilde{5}$ min per clips). Each clip was spoken by a single and unique speaker. The samples of these extracts have been mainly derived from C-PROM [21] corpus, which is also a French spoken corpus containing seven speaking styles: radio broadcast news, aloud reading, political speech, university conference, radio interview, route prescription, narrative-life story. Unlike the other corpora chosen for conducting this study, RHAPSODIE clips cover diverse speaking styles and contain 30 speakers.
- NCCFr [20](The Nijmegen Corpus of Casual French): French speakers conversing among friends.

8, summarize the contents of the designed corpus.

4.2 Data processing

The forced alignment at the phone level was performed using JTrans, then annotated according to the procedure described in the 1. All the information was stored in the TextGrid format. A Praat[10] software script was used to collect the formant values F1 and F2 for each of the three vowels. All these data were compiled in a CSV data file where they were sorted and manipulated with a script written in the R language. The outliers, identified following the analysis of the formed vowel trapezoids, were removed from the report. The formants (F1 and F2) values in Hertz were then converted to Bark[11] to be able to compare the data

[10] https://bigdataspeech.github.io/TP/tp/2018/07/10/TPPraat.html

[11] Bark is a psycho-acoustical scale closer to subjective perceptual scale

of the different corpora. The graphical analysis of the vowel trapezoids was done using the phonR[12] package.

4.3 Results and discussion

After a graphical analysis of our data, we were able to observe a different dispersion in the vowel trapezoid for each corpus. Indeed, we can also notice that there is certain similarity among the read speech corpora (MUFASA and BREF), where there is a strong overlap between the vowels /u/ and /a/. We can group RHAPSODIE corpora that represent the diverse speaking styles. We can notice a large contrast on F2 in the ESTER and NCCFr corpora, which is not the case in the other corpora, this could be explained by the fact that these two last corpora have been recorded in good conditions.

The 5 illustrates the dispersion obtained for each corpus. As expected, for most of the corpora studied, the /i/ is articulated on average in the closed anterior position, the /u/ in the closed posterior position and the /a/ in the open middle position.

Except for the data from RHAPSODIE (5e), this differs from the others, especially from the two vowels /i/ and /u/, however the vowel /a/ is in the same position (in the open middle position.)

According to [22–24] there is an interaction between speaking style and the duration of the vowels. The analysis of the density distribution according to the duration of the three vowels preceded by an occlusive consonant /p/,/t/,/k/ (similar context over speakers/corpus) illustrated by 6, which consists of comparing each of the excerpts from the different corpora to the MUFASA corpus. It can be observed that the duration of the vowels is quite long, which is quite logical since speakers tend to take their time when it comes to read loudly or even when it is a prepared or partially prepared speech such as a radio diary.

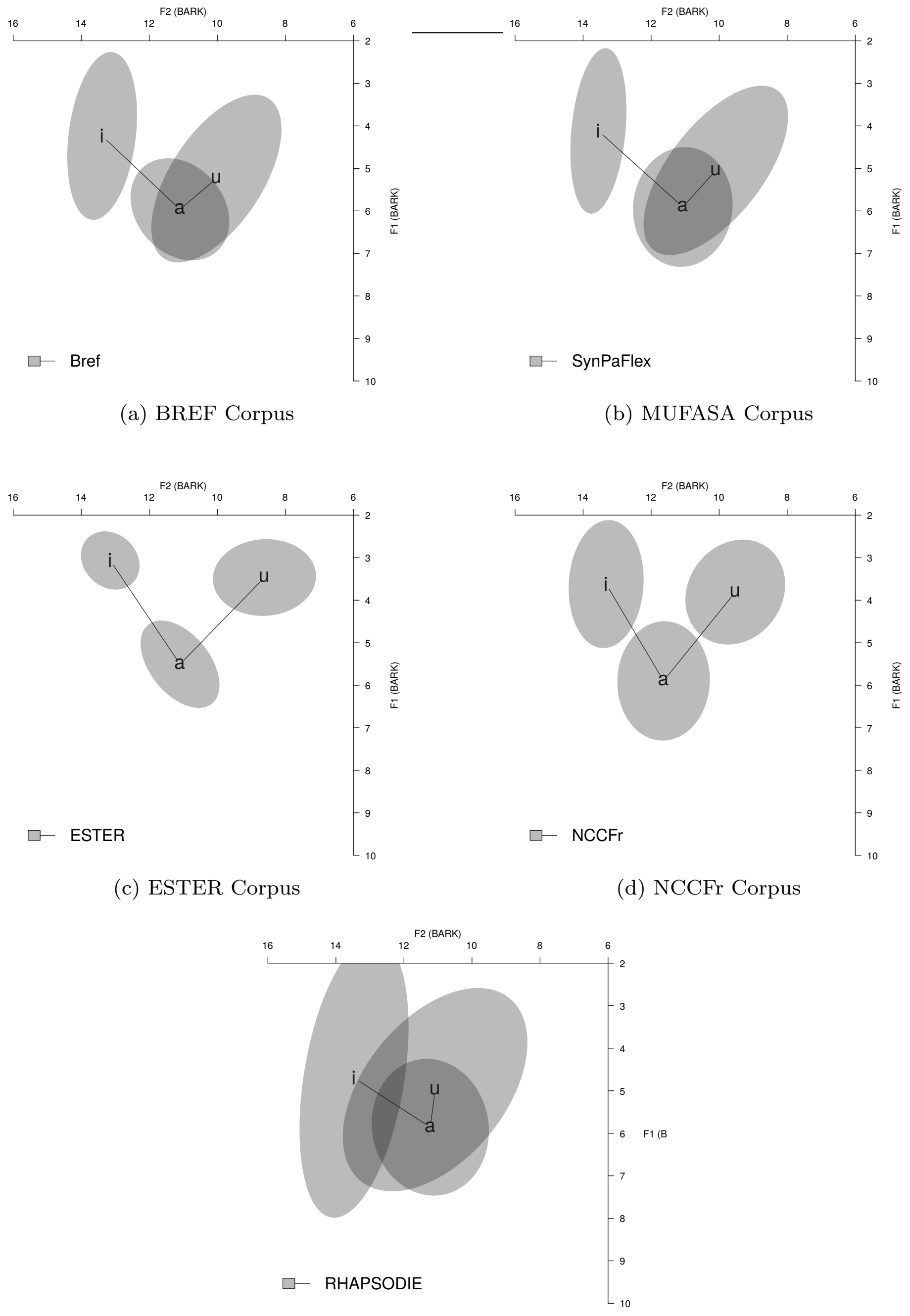
4.4 Remarks

This study intended to be exploratory and attempted to provide some elements for reflection. This work raises two main reflections, which are the level of formality of the audiobook corpus in comparison with other speaking styles, and a second element, the presence of particular prosodic behavior specific to audiobooks data. In this work we try to compare five corpora designed for a different task with different sized speech inventories. Certain factors, mainly acoustic factors, may have influenced our results.

5 Conclusion

In this chapter, we presented a new audiobook corpus, the MUFASA corpus, dedicated to expressive speech synthesis but that can be used for other purposes, such as automatic speech recognition, natural language processing, second language acquisition, entity recognition. Consisting of twenty speakers (ten females/ten males)

[12] http://drammock.github.io/phonR/



(a) BREF Corpus

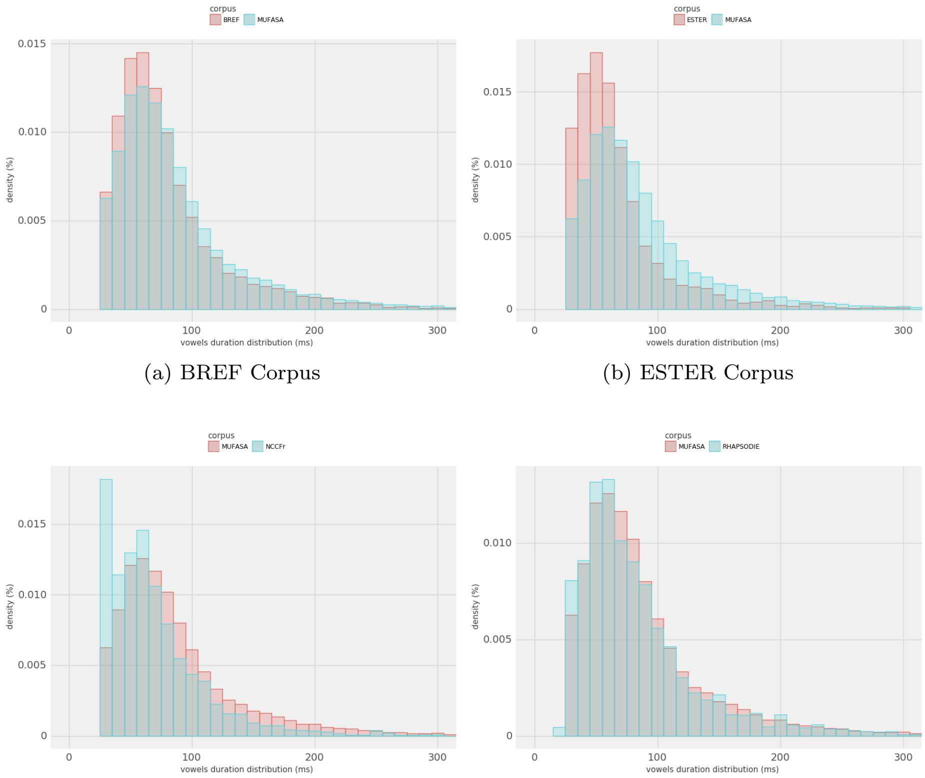
(b) MUFASA Corpus

(c) ESTER Corpus

(d) NCCFr Corpus

(e) RHAPSODIE Corpus

Fig. 5: The vowel trapezoids of the three cardinal vowel, in the context of occlusive /p/,/t/,/k/



(a) BREF Corpus

(b) ESTER Corpus

(c) NCCFr Corpus

(d) RHAPSODIE Corpus

Fig. 6: The density distribution according to the duration of the three vowels preceded by an occlusive consonant /p/,/t/,/k/

and included around 600 hours of audiobook. The majority of the data is in French, and a few hours are in English. Furthermore, we analyzed some aspects of expressivity of speech covered by the MUFASA corpus. We have shown that the recording of audiobooks differs between professional and amateurs in terms of voice quality. Nevertheless, we did not treat two important aspects of expressivity in this chapter, which are the emotion and discourses. Emotional speech is the main topic of the coming chapter. The second aspect not treated in this chapter is the discourse. Audiobooks cover an extensive variety of discourse encoded in the text, like dialogues amongst characters in a given novel, which contribute a lot to the expressivity of the audiobooks. We addressed the discourse typology in audiobooks in two chapters. In Chapter 5, we present the automatic detection and classification of the discourses types present in the audiobooks. Then, in chapter 6, we discuss the prosodic characteristics of discourse in audiobooks.

References

1. D. Doukhan, S. Rosset, A. Rilliard, C. d'Alessandro, M. Adda-Decker, Language Resources and Evaluation **49**(3), 521 (2015)
2. J. Chevelu, G. Lecorvé, D. Lolive, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (Reykavik, Iceland, 2014)
3. C. Cerisara, O. Mella, D. Fohr, in *Proceedings of the 10th Annual Conference of the International Speech Communication Association - Interspeech 2009* (Brighton, United Kingdom, 2009)
4. P. Boersma, D. Weenink. Praat: Doing phonetics by computer.[computer program]. version 6.0. 19 (2016)
5. S. Galliano, G. Gravier, L. Chaubard, in *Tenth Annual Conference of the International Speech Communication Association* (Brighton, UK, 2009), pp. 2583–2586
6. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, in *Interspeech* (2017), pp. 498–502
7. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., in *IEEE 2011 workshop on automatic speech recognition and understanding* (IEEE Signal Processing Society, 2011), CONF
8. S. Brognaux, S. Roekhaut, T. Drugman, R. Beaufort, in *2012 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2012), pp. 416–421
9. G. Gravier, (2003)
10. D. Talkin, W.B. Kleijn, Speech coding and synthesis **495**, 518 (1995)
11. H. Zen, V. Dang, R. Clark, Y. Zhang, R.J. Weiss, Y. Jia, Z. Chen, Y. Wu, arXiv preprint arXiv:1904.02882 (2019)
12. N. Audibert, C. Fougeron, in *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP* (2012), pp. 217–224
13. M. Lindau, Language **54**(3), 541 (1978)
14. X. Sun, in *2002 IEEE international conference on acoustics, speech, and signal processing*, vol. 1 (IEEE, 2002), vol. 1, pp. I–333
15. X. Sun, Y. Xu, Journal of Voice **16**(4), 443 (2002)
16. I.R. Titze, *Workshop on acoustic voice analysis: Summary statement* (National Center for Voice and Speech, 1995)
17. L.F. Larnel, J.L. Gauvain, M. Eskenazi, in *Second european conference on speech communication and technology* (1991)
18. S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, K. Choukri, in *LREC* (2006), pp. 139–142
19. A. Lacheret, S. Kahane, J. Beliao, A. Dister, K. Gerdes, J.P. Goldman, N. Obin, P. Pietrandrea, A. Tchobanov, in *Language Resources and Evaluation Conference* (2014)
20. F. Torreira, M. Adda-Decker, M. Ernestus, Speech Communication **52**(3), 201 (2010)
21. M. Avanzi, A.C. Simon, J.P. Goldman, A. Auchlin, in *Speech Prosody 2010-Fifth International Conference* (2010)
22. S.J. Moon, B. Lindblom, The Journal of the Acoustical society of America **96**(1), 40 (1994)
23. R.E. Baker, A.R. Bradlow, Language and speech **52**(4), 391 (2009)
24. R.S. Burdin, C.G. Clopper, (Glasgow, Scotland, 2015)