SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis

Aghilas Sini, Damien Lolive, Gaëlle Vidal, Marie Tahon, Elisabeth Delais-Roussarie

SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis

Aghilas Sini[1], Damien Lolive[1], Gaëlle Vidal[1], Marie Tahon[3] and Élisabeth Delais-Roussarie[2]
[1]IRISA/Université de Rennes 1, Lannion, France;
[2]UMR 6310 - Laboratoire de Linguistique de Nantes, Nantes, France;
[3] LIUM / Le Mans Université, Le Mans, France.
elisabeth.delais-roussarie@univ-nantes.fr, marie.tahon@univ-lemans.fr
{aghilas.sini, damien.lolive, gaelle.vidal}@irisa.fr

Abstract

This paper presents an expressive French audiobooks corpus containing eighty seven hours of good audio quality speech, recorded by a single amateur speaker reading audiobooks of different literary genres. This corpus departs from existing corpora collected from audiobooks since they usually provide a few hours of mono-genre and multi-speaker speech. The motivation for setting up such a corpus is to explore expressiveness from different perspectives, such as discourse styles, prosody, and pronunciation, and using different levels of analysis (syllable, prosodic and lexical words, prosodic and syntactic phrases, utterance or paragraph). This will allow developing models to better control expressiveness in speech synthesis, and to adapt pronunciation and prosody to specific discourse settings (changes in discourse perspectives, indirect vs. direct styles, etc.). To this end, the corpus has been annotated automatically and provides information as phone labels, phone boundaries, syllables, words or morpho-syntactic tagging. Moreover, a significant part of the corpus has also been annotated manually to encode direct/indirect speech information and emotional content. The corpus is already usable for studies on prosody and TTS purposes and is available to the community.

Keywords: Speech corpus, audiobooks, emotion, expressiveness,

1. Introduction

To build an expressive Text-To-Speech (TTS) system able to read books of different literary genres, using various discourse modes and speaking styles, a corpus that covers all these specificities is required. Usually, corpora built for TTS purposes are less than ten hours long and monospeaker. In addition, the content is carefully controlled to maximize the homogeneity of the synthetic speech.

Long and coherent speech data is very interesting as it gives the possibility of studying voice expressiveness under different situations. Audiobooks are a good example of such data and are valuable for prosody modeling, especially in the field of storytelling. For instance, (Montaño et al., 2013; Montaño Aparicio, 2016) show that expressive categories might exist in the storytelling speaking style. Some works have also been done to detect the speaking style in audiobooks (Székely et al., 2012b) and to evaluate the usability of audiobooks data for the generation of conversational speech (Székely et al., 2012a).

In the last decade, many corpora were built from audiobooks. For instance, (Panayotov et al., 2015) presents a multi-speaker English corpus built for speech-text alignment purposes containing thousands of hours of speech. In (Stan et al., 2013), a multilingual corpus that contains approximately sixty hours of speech data from audiobooks in 14 languages is introduced. Since this corpus contains an average of four hours per language and only one book per language, it prevents studies on speaking styles particularly for TTS synthesis. The same analysis is true for the GV-Lex corpus (Doukhan et al., 2015) which focuses on tales analysis and contains twelve tales but only one hour of speech. In other works such as (Zhao et al., 2006; Wang et al., 2006), authors use an audiobook recorded by a professional speaker, and explore the speech recording in different spaces (expressive, acoustic and perceptual). Nevertheless, none of these works proposes a large monospeaker audiobook corpus in French language including various literary genres.

This work introduces a new corpus of read French speech, suitable to build an expressive speech synthesis or to build robust storytelling prosodic models. It is built in the context of the SynPaFlex[1] research project aiming at improving text-to-speech synthesis. The main motivation to build a corpus of audiobooks is to study a large coherent speech recording, made by a single amateur voice and containing different linguistic, acoustic, phonetic and phonological phenomena. To this end, this new corpus contains an eighty seven hours collection of good quality audiobooks, extracted from Librivox, uttered by one speaker and covering various types of literary genres (novels, short stories, tales, fables, and poems). When reading the books, the speaker has an expressive speaking style and uses personification to make the characters distinguishable. Moreover, whole books or chapters are used, thus enabling to study long term discourse strategies used by a speaker.

To build such large corpora from audiobooks, many techniques have been proposed such as in (Braunschweiler et al., 2010; Boeffard et al., 2012; Mamiya et al., 2013; Charfuelan and Steiner, 2013). In this paper, we use an automatic large-scale speech-to-text aligner for the French language (Cerisara et al., 2009) to perform the segmentation into phones.

The remainder of the paper is organized as follows. Section 2 explains how data used to build the corpus were chosen. In sections 3 and 4 , the different manual and automatic an-

[1]https://synpaflex.irisa.fr/

Literary genre	Duration	Discourses annotation	Expressivity annotation
Novels	80h12m	27h21m	10h59m
Short stories	5h01m	4h08m	2h26m
Tales	1h22m	1h22m	10m
Fables	18m	18m	/
Poems	29m	29m	/
Total	87h23m	33h39m	13h25m

Table 1: Collected data durations and amount of annotated data according to speaking style.

notations included in the corpus are detailed. Finally, Section 5 gives first results about the proposed emotion annotation scheme.

2. Corpus Construction

2.1. Corpus Constraints

Designing a speech corpus requires the definition of some criteria for data selection. As this corpus is built in the framework of the SynPaFlex project, it will be used to study prosody, pronunciation, and also to build expressive speech synthesis models for French. Considering this, we have made the following requirements:

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

2.2. Data Selection

Investigations were conducted using two main types of sources: audio CD and on-line servers. Because of the lesser accessibility to CD audiobooks, online servers were found to be the most appropriate search areas, even if they sometimes provides data patchy in quality.

After several trials, we concentrated our efforts on the recordings made by a female speaker and available on Librivox, one of the public domain projects that provide audio recordings of book readings on a Web server. In total, eighty seven hours of audio files and the corresponding texts have been collected.

As shown in Table 1, the novel genre represents about 90% of the corpus. Among the list of selected books for this literary genre, there are "Les Misérables" (Victor Hugo), "Madame Bovary" (Gustave Flaubert) and "Les Mystères de Paris" (Eugène Sue).

As the selected audiobooks are read by a non-professional speaker, the acoustic conditions might be different between chapters and books. Some listening evaluations have been done on audio signal in order to identify those of lesser quality, allowing to potentially exclude them from further processes.

3. Automatic Annotations

3.1. Data Preparation

The whole annotation process has been conducted relying on the ROOTS toolkit (Chevelu et al., 2014), that allows storing various types of data in a coherent way using sequences and relations. This toolkit 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[2]. Despite of that, audio data acoustic features remain more or less heterogeneous.

As texts were coming from diverse sources, their formats were unified. Then the exact orthographic transcriptions of the readings were achieved by inserting 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 keep each chapter in a separate file so as to keep long term information accessible.

3.2. Speech Segmentation

The broad phonetic transcription, based on the French subset of Sampa, has been extracted and aligned with the speech signal using JTrans (Cerisara et al., 2009).

To evaluate the accuracy of the phone segmentation, an expert annotator performed a manual validation using Praat (Boersma and Weenink, 2016). Since there is only one speaker, half an hour of the SynPaFlex Corpus was taken into account to evaluate the quality of the phone labels and boundaries. The set of data used for the evaluation task has been selected respecting the proportion of the different literary genres in the corpus.

Results related to the validation are presented in Table 2. 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 2. 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 concern 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, errors in alignment are more important: in 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

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

	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 2: 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 corpus

is of 6 syllables/s on average. Secondly, the acoustic models used to achieve the automatic segmentation (Cerisara et al., 2009) have been trained on the ESTER2 corpus (Galliano et al., 2009) which is a French radio broadcasts corpus. The resulting models could thus be slightly unadapted for poem reading data.

3.3. Linguistic Information

Additional linguistic information has been added to the corpus, such as syllables and Part-Of-Speech tags using the Stanford parser (Green et al., 2011). Table 3 sums up the content of the corpus in terms of linguistic units. We also plan to include syntactic information in a near future.

3.4. Acoustic and Prosodic Information

The speech signal is stored using a sampling frequency of 44.1kHz. 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 (Gravier, 2003), (ii) the instantaneous fundamental frequency (F_0) using the ESPS get_f0 method implementing the algorithm presented in (Talkin, 1995), and (iii) pitch-marks 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.

4. Manual Annotations

Audio tracks corresponding to chapters of different books have also been annotated manually according to prosodic units, characters, emotions, and other events that were heard. The annotation method had first been defined on a small subset of readings, and then tested on audiobook recordings completed by other readers. It was found to be generic enough to render a global perceptive description of the speech. As Table 1 shows, 38% of the whole corpus have been processed manually to provide characters annotation, and 15% - included in those 38% - to describe emotional and prosodic patterns contents.

4.1. Prosodic Patterns

After considering the whole speech data, eight prosodic descriptors were defined, then encoded and assigned by an expert annotator to a large number of audio tracks corresponding to chapters of 18 different books, and defining a 13h25m sub-corpus. As far as possible, labels were assigned according to the perceived prosody, without taking into account the linguistic content. They characterize units which could range in length from a word to several sentences. Seven of them correspond to speech showing the following types of prosodic patterns: QUESTION (interrogative), NOTE, NUANCE, SUSPENSE, RESOLUTION (authority, or imperative), SINGING, and IDLE (no particular prosodic pattern, or declarative). The eighth label, EMOTION, was used to report - but without describing it - the presence of any perceived emotional content.

Let's notice as of now that the tag EXCLAMATION is not listed above. This is because this information can be simply deduced from another level of description: in this corpus, the *Exclamation* pattern was found strictly correlated with the emotional content of *surprise*, which is reported in the emotion labeling level (presented in Section 4.3.). Manual annotation is costly in time and redundancy is not desirable in its process. In the following analysis of the prosodic manual labeling, emotion labels *surprise* will therefore be assimilated to hidden prosodic labels for EXCLAMATION.

Another important point is that, when needed for a more precise description, labels were combined (e.g. Emotion+Question+Nuance). Hence, simply summing the labels duration for each type of prosodic pattern gives a value which exceeds the duration of the sub-corpus.

Among the prosodic parameters, the perceived pitch-curve during voice production takes an important role in assigning the labels. For instance, the NUANCE pattern, which is one of the reading strategy of the speaker, maintains listener's attention. This pattern is characterized melodically by a high pitch at the beginning, then a decrease with modulations, and finally a slight increase when it doesn't end the sentence (see Figure 1).

Table 4 shows total duration for each manual prosodic labels in the 13h25 sub-corpus.

A non-IDLE prosodic tag has been assigned to 68% of the speech. As shown in Table 4, the hidden EXCLAMATION tag is very largely represented (more than 4h42), before the IDLE one (4h21m). The first particular prosodic pattern that comes after is NUANCE (3h58m), then come all the other prosodic patterns that are relatively well represented

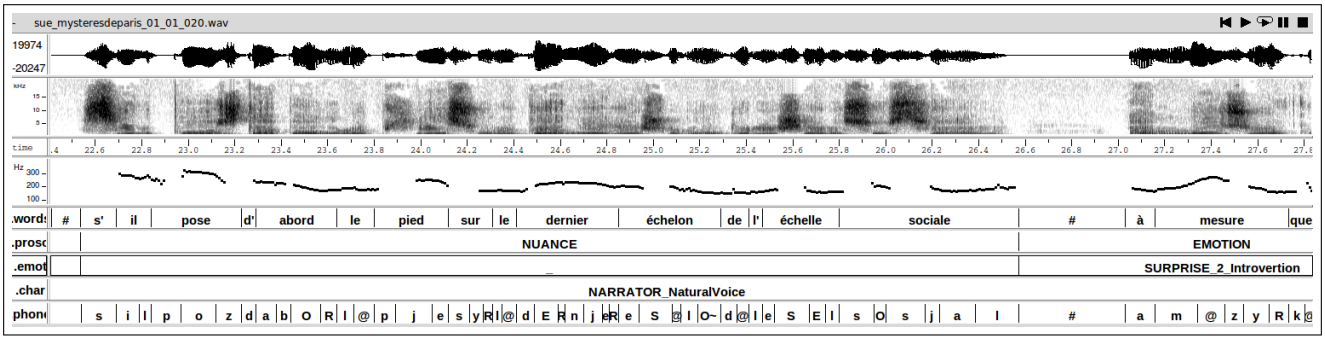

Figure 1: Manual annotations - NUANCE prosodic pattern example

Prosodic label	EXCLAMATION (hidden label)	IDLE	NUANCE	RESOLUTION	SUSPENSE	QUESTION	NOTE
Duration	4h42m	4h21m	3h58m	45m	41m	38m	39m
Sub-corpus %	34.8%	32.2%	29.5%	5.6%	5.1%	4.7%	4.8%

Table 4: Manual annotations - Total duration of prosodic patterns (including combinations) in the 13h25 sub-corpus

and evenly distributed (around 40m): RESOLUTION, SUSPENSE, QUESTION and NOTE. SINGING was found to be exceptional and is not reported here.
More than a half of the speech showing particular prosodic figures is described with combined labels, pointing out where prosody may be more complex.
Most of all, it was found that the EXCLAMATION pattern happens very frequently, especially in narration. In a way, it is an inherent part of the speaker's style.
The generic EMOTION prosodic indicator is assigned to 39% of the whole sub-corpus (5h18m), showing a large amount of emotional data. Its manual description is presented in Section 4.3.

4.2. Characters

The speaker, who is the same for the whole corpus, can personify the different characters of the book by changing her voice or her way of speaking. The character's tags were identified from the text and any turn of speech has been labeled according to the following annotation scheme:

- CHARACTER ID: indicates which character is talking according to the text, and refers to Meta-data where each character is summarily described (name, age, gender, prosody and timbre features). For instance, to personify a gloomy man, the speaker uses a low pitch, low energy and devoiced voice.

- VOCAL PERSONALITY ID: indicates which character is talking according to the vocal personality. Indeed, even if the speaker is very talented and coherent along the books, she can for example forget to change her voice when comes a new character. Therefore, for such speech intervals, the timbre remains the own speaker's timbre or corresponds to another character.

The characters labeling was annotated on more than one third of the whole corpus (33h39m) mined from 18 different books. Dialogue tags were reported as parts of the narrator's speech.
First estimates indicate that one third of the speech content is personified. The average duration for speech turns being of 7s, against 29s for the narrator. In some chapters, direct speech segments can also be very long, typically when a character becomes a narrator who tells his own story.
370 characters were identified, and the full data of their vocal personality labeling indicates a not negligible amount of prosody and vocal tone personification. Covering a wide range and typology, the speaker's voice is thus more or less radically far from her natural style (males, children and elderly people embodiments, psychological singularization, imaginary figures). These vocal personality changes often happen: around 20% of the speech is concerned and, for the half, in stark contrast with the natural speaker's voice.

4.3. Emotions

Different theoretical backgrounds are classically used to identify emotional states, principally based on either distinct emotion categories or affective dimensions (Cowen and Keltner, 2017). Usually, choosing the emotion categories and their number, or the emotion dimensions is an issue.
In the present study, the basic scheme used to manually encode emotions has three items:

- *Emotion category*: Six categories are available, those selected by the Basic Emotions theory (Ekman, 1999): SADNESS, ANGER, FEAR, HAPPINESS, SURPRISE, DISGUST. Two other categories were added to better represent the content of the different books: IRONY and THREAT.

- *Intensity level*: This item, on a scale from 1 to 3, is meant to give a measurement of the experienced emotion intensity according to the speech. For instance, one can interpret its values as follows: SLIGHTLY ANGRY (1), ANGRY (2) , and STRONGLY ANGRY (3).

- *Introversion/Extroversion*: This binary item reflects

Emotion	SURPRISE	SADNESS	JOY	ANGER	DISGUST	FEAR
Effects on the first syllable of focus word(s)	accentuation	disappearance		accentuation	accentuation	
Pitch median	high	low	according to joy type	low	low	low
Pitch curve		flat	flat (suave joy)	flat or top-down	flat or top-down	flat
Rate		slow	according to joy type	fast	fast on focus words	varying with fear intensity
Loudness		low	loud (intense joy)		low	
Timbre changes		breath during the speech	breath during the speech		yes	yes

Table 5: Examples of perceived impacts of emotion on the speech

the way the emotion is rendered through the speech (discreetly, prudently / obtrusively, ostentatiously)

The second and third items may have strong correlations with some of the widely used affective dimensions, as activation and arousal. Furthermore, an important feature of the manual emotion annotation used for the corpus is that the three items labels can be mixed together to provide a more precise description of the perceived emotion. For instance, speech can continuously convey strong and very expressive SADNESS as well as FEAR through some words, which could be tagged as [sadness-3-E + fear-1-E].

Manual emotion labeling was done on the same sub-corpus as for prosody (13h25m). A large amount of emotional content was reported (39% of the speech, including 13% with combined tags). Duration of tagged speech for each category of emotion is given in Table 6, and the number and average duration of labels are indicated in Table 7.

Significant observations have emerged during the annotation. A challenging one is that two radically different types of JOY can be conveyed by the speech, whereas none of the three items could take over their differentiation: on the one hand suave joy, and on other hand elation or gladness. Also, it is suggested that labels should be interpreted in context, notably in conjunction with the discourse mode. In particular, the expressive strategy implemented in the corpus narration is very specific, conveying almost continuously positive valence but in a subtle way, through pitch modulation and with focus words. The SURPRISE label was widely assigned to those recurrent patterns showing (i) a sudden pitch shifting upwards (ii) at least one accentuation onto the first syllable of a focus word (iii) a phonetic elongation or a short silence before this first syllable. Thus, as introduced in section (see 4 1)SURPRISE describes a recurrent emotional attitude of the reader, attracting the listener attention by regularly emphasizing the text.

Other types of variation occur when the speech conveys emotion, some examples are related Table 5 .

4.4. Other Events

Besides acoustic indications of loud noises or music, different unexpected speech events were also reported:

- Linguistic events: for example, the use of foreign languages;
- Phonetic events which are not written in the text: phoneme substitutions, elisions and insertions, high elongations, breaks and pauses, specific voice quality (e.g. whispered voice).

All these features can be of high interest for rendering a more human synthetic voice (Campbell, 2006).

The manual data-sets could provide valuable guidance for further analysis, especially by linking with linguistic information, acoustic measurements, and other descriptions. Examining how manual labels are distributed among literary genres could also be of great interest.

5. Emotion classification

This section presents binary emotion classification experiments conducted on emotional labels of the SynPaFlex corpus. The use of a state of the art methodology aims at positioning our mono-speaker read expressive speech corpus among existing multi-speaker acted or spontaneous emotional speech corpora.

5.1. Data analysis

The manual segmentation and labeling of emotion – which concerns 15% of the whole corpus – results in a total number of 8 751 segments as shown in Table 7. Among them, 5 387 convey an emotional content, while 3 364 do not. To get around the issue of a "neutral" emotion, we decided to label these segments as Idle. As mentioned previously, label combinations were used during the annotation phase to better characterize some expressive content. Consequently, these annotations are considered as new emotional labels which can not be merged with single labels easily. A deeper

Emotion	IDLE	SURPRISE	SADNESS	JOY	ANGER	DISGUST	FEAR	IRONY	THREAT
Duration	8h11m	4h42m	44m	32m	31m	15m	11m	10m	3m
Sub-corpus %	61.0%	35.0%	5.4 %	3.9 %	3.9%	1.9%	1.3%	1.2%	0.4%

Table 6: Manual annotations - Total durations of emotion categories labels (including combinations) in the 13h25 sub-corpus

Emotion	Idle	Anger	Joy	Sadness	Fear	Surprise	Disgust	Other	Comb.	Total
# Seg. manual	3 364	147	115	295	76	2 895	47	23	1 699	8751
Avg. dur (s)	8.76	2.62	2.99	2.67	2.20	3.83	2.26	2.30	3.45	5.55
# Seg. 1 s. max	30 989	447	397	929	199	12 794	125	0	0	45 880

Table 7: Number of manual annotated emotional segments and segments resulting from a 1 s. max chunking. The latest are used in the classification experiments. Other includes IRONY and THREAT labels.

investigation of these label combinations is needed in order to manage them in a speech synthesis system.
Interestingly, the SURPRISE label is highly represented among other single emotional labels. Actually, as described in Section 4 , SURPRISE better corresponds to an emotional attitude of the reader to keep the listener's attention, than an emotion conveyed from the text.
Emotional segments are defined as segments consisting of an homogeneous emotion, be it characterized by single or combined labels. Therefore, there is not constraint on segments' duration. As a consequence, some segments can be very long. For example, one IDLE segment lasts more than 43s. On average (cf Table 7), IDLE segments have the highest durations (8.76s), then comes SURPRISE segments (3.83s.) and COMBINATION labels (3.45s.).

5.2. Methodology

The following experiments aim at classifying the manual annotations with binary emotional models. We know that for multi-speaker acted emotions, classification rates usually reach high performance (for example with corpora such as EMO-DB). However, with multi-speaker spontaneous speech, the classification rates are much lower, thus reflecting the difficulty to discriminate emotions in such a context (Schuller et al., 2009b). The present corpus gives the opportunity to bring a new benchmark of performances on mono-speaker read speech.
To do so, our experimental set up follows a standard classification methodology (Schuller et al., 2009a; Schuller et al., 2013). By this way, our results are comparable with those obtained on other existing emotion corpora. In other words, emotional models are trained in cross-validation conditions (here 5 folds to keep enough data) on acoustic features. 384 acoustic features – 16 Low-Level Descriptors (LLD) $\times$ 12 functionals + Δ – are extracted on emotional segments with OpenSmile toolkit and Interspeech 2009 configuration (Schuller et al., 2009a). To avoid over fitting the data, different subsets of features are tested:

- OS192: 16 LLD $\times$ 12 functionals without Δ
- Δ OS192: 16 LLD $\times$ 12 functionals with Δ only
- OS24: 2 LLD (range + amean) $\times$ 12 functionals without Δ

In order to have homogeneous segment durations, we decided to chunk manual segments every 1 s keeping the remaining part. This operation helps in increasing the amount of data available for the experiment, as reported in Table 7. As aforementioned, COMBINATION labels are not taken into account because merging them with single labels is clearly not obvious. Also, IRONY and THREAT segments are discarded regarding to the small number of labels. To better identify the pairs of labels that can be easily discriminated from those which can not, only binary models are trained thus resulting in an emotion confusion matrix. The number of segments is equally balanced among the two classes.

5.3. Results

Models are trained with Random Forests and entropy criterion. Similar performances were obtained with optimized Support Vector Machines (polynomial kernel, C=1, $\gamma = 0.01$) and normalized features. The results are given as a confusion matrix between emotions as shown in Table 8. On average, performances obtained with the smaller set are the best: 59.9% with OS24, 59.5% with OS192 and 58.8% with ΔOS192. This first observation underlines the importance of selecting features when classifying emotions in such corpora in order to avoid over fitting the data (Tahon and Devillers, 2016).
As we were expecting, the binary emotion classification UAR results range from 43.6% to 81.8%, a typical range for induced and spontaneous speech emotion recognition. These performances also reflect the high diversity of vocal personifications during direct speech as well as different recording conditions. The most impressive classification rates are reached with ΔOS192 for IDLE/ANGER (77.7%) and ANGER/DISGUST (81.8%) emotion pairs. Is seems that the acoustic dynamics captured by this feature subset is very relevant for these two emotion pairs. With Δ features, classification rates drop compared to non-Δ features on other pairs of emotions.

UAR		Ang.	Sad.	Joy	Fea.	Dis.	Sur
OS192	Idl.	**.640**	**.618**	.572	**.638**	.592	.571
	Ang.		.550	**.677**	.563	.572	**.637**
	Sad.			**.610**	.475	.524	**.616**
	Joy				**.636**	**.620**	.573
	Fea.					.584	**.636**
	Dis.						**.600**
ΔOS192	Idl.	**.777**	**.601**	.557	**.650**	.524	.555
	Ang.		.544	.594	.523	**.818**	.584
	Sad.			.621	.525	.436	.566
	Joy				**.638**	.548	.544
	Fea.					.588	**.631**
	Dis.						.532
OS24	Idl.	**.624**	**.621**	.580	**.628**	**.612**	.563
	Ang.		.567	**.671**	.578	.580	**.623**
	Sad.			**.616**	.530	.548	**.631**
	Joy				**.638**	.596	.567
	Fea.					.580	**.633**
	Dis.						.584

Table 8: Unweighted Average Recall (UAR) results for binary emotion classification using the three feature subsets. In bold, UAR > 60%

Regarding the results obtained with the small OS24 feature subset, classification between non emotional (IDLE) and emotional segments is over 60% (bold font in Table 8) for ANGER, SADNESS, FEAR and DISGUST. Two emotion groups emerge from the results:

- IDLE/JOY (58.0%), IDLE/SURPRISE (56.3%) and JOY/SURPRISE (56.7%)
- SADNESS/FEAR (53.0%), SADNESS/DISGUST (54.8%), SADNESS/ANGER (56.7%), FEAR/DISGUST (58.0%), FEAR/ANGER (57.8%) and ANGER/DISGUST (58.0%)

The second group clearly contains negative emotions with different arousal levels.
Further experiments are needed to deeper investigate these groups such as unsupervised clustering, feature selection, etc. For example, Δ features are clearly relevant for ANGER/DISGUST classification. Moreover, emotions are likely to be strongly correlated with direct/indirect speech and also with characters. Additional analyses are required to confirm this observation. The addition of phonological and linguistic information could also help in understanding the emotional distribution of the SynPaFlex corpus.

6. Conclusion and Perspectives

This paper describes a new large speech corpus composed of eighty seven hours of audiobooks from several literary genres read by a single speaker. By being mono-speaker, this corpus can be used to study the strategy of a speaker over entire books. Annotations and speech segmentation into phones are also provided. Among them, we can mention that a manual annotation of emotional contents and characters has been done for respectively 15% and 38% of the whole corpus.
Preliminary emotion classification experiments show that the expressive read speech contained in the SynPaFlex corpus is much closer from spontaneous speech than acted speech. From binary classification results, two emotional groups emerge, one clearly containing negative content. Deeper investigations and analyses of the corpus are planned in a future work: correlation between direct and indirect phrases, emotional speech and characters. Further experiments on feature selection and clustering could also help in investigating the emotional content of this corpus. Moreover, the full corpus has already been used to build a speech synthesis voice, and informal evaluations show that the output of the unit-selection speech synthesis system is relatively good. The corpus is available on our website[3].
This corpus could also be used to study prosodic and phonological aspects of expressive read speech (character personification, speaking styles) and to develop expressive synthesized speech of good audio quality. The amount of data available with the SynPaFlex corpus is consequent enough to allow carrying a deeper analysis of the linguistic, acoustic and prosodic features associated with some of these aspects. Machine Learning techniques such as deep learning can also be used to build prosodic models.

7. Acknowledgements

This study has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015 and the LABEX EFL (Empirical Foundations in Linguistics) ANR-10-LABEX-0083.

8. Bibliographical References

Boeffard, O., Charonnat, L., Le Maguer, S., Lolive, D., and Vidal, G. (2012). Towards Fully Automatic Annotation of Audio Books for TTS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 975–980, Istanbul, Turkey.

Boersma, P. and Weenink, D. (2016). Praat: Doing phonetics by computer.[computer program]. version 6.0. 19.

Braunschweiler, N., Gales, M. J., and Buchholz, S. (2010). Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Makuhari, Chiba, Japan.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1171–1178.

Cerisara, C., Mella, O., and Fohr, D. (2009). JTrans, an open-source software for semi-automatic text-to-speech alignment. In *10th Annual Conference of the International Speech Communication Association(Interspeech 2009)*, Brighton, U.K.

Charfuelan, M. and Steiner, I. (2013). Expressive speech synthesis in MARY TTS using audiobook data and emotionML. In *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pages 1564–1568, Lyon, France.

Chevelu, J., Lecorvé, G., and Lolive, D. (2014). ROOTS: a toolkit for easy, fast and consistent processing of large

[3]https://synpaflex.irisa.fr/corpus/

sequential annotated data collections. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykavik, Iceland.

Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Doukhan, D., Rosset, S., Rilliard, A., d'Alessandro, C., and Adda-Decker, M. (2015). The GV-LEx corpus of tales in French: Text and speech corpora enriched with lexical, discourse, structural, phonemic and prosodic annotations. *Language Resources and Evaluation*, 49(3):521–547.

Ekman, P. (1999). Basic emotions. In Dalgleish T. et al., editors, *Handbook of Cognition and Emotion, 1999*.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*, pages 2583–2586, Brighton, UK.

Gravier, G. (2003). Spro: "speech signal processing toolkit". *Software available at http://gforge.inria.fr/projects/spro*.

Green, S., De Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. Conference on Empirical Methods in Natural Language Processing (EMNLP'11), Edinburgh, United Kingdom.

Mamiya, Y., Yamagishi, J., Watts, O., Clark, R. A., King, S., and Stan, A. (2013). Lightly supervised GMM VAD to use audiobook for speech synthesiser. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pages 7987–7991, New Orleans, U.S.A. IEEE.

Montaño Aparicio, R. (2016). Prosodic and Voice Quality Cross-Language Analysis of Storytelling Expressive Categories Oriented to Text-To-Speech Synthesis. *TDX (Tesis Doctorals en Xarxa)*.

Montaño, R., Alías, F., and Ferrer, J. (2013). Prosodic analysis of storytelling discourse modes and narrative situations oriented to text-to-speech synthesis. In *8th ISCA Speech Synthesis Workshop*, pages 171–176, Barcelona, Spain.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, pages 5206–5210. IEEE.

Schuller, B., Steidl, S., and Batliner, A. (2009a). The interspeech 2009 emotion challenge. In *10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, U.K.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009b). Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 552–557. IEEE.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Stan, A., Watts, O., Mamiya, Y., Giurgiu, M., Clark, R. A., Yamagishi, J., and King, S. (2013). TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. In *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pages 2331–2335, Lyon, France.

Székely, E., Cabral, J. P., Abou-Zleikha, M., Cahill, P., and Carson-Berndsen, J. (2012a). Evaluating expressive speech synthesis from audiobooks in conversational phrases. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3335–3339, Istanbul, Turkey.

Székely, E., Kane, J., Scherer, S., Gobl, C., and Carson-Berndsen, J. (2012b). Detecting a targeted voice style in an audiobook using voice quality features. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, pages 4593–4596. IEEE.

Tahon, M. and Devillers, L. (2016). Towards a small set of robust acoustic features for emotion recognition: challenges. *IEEE/ACM transactions on audio, speech, and language processing*, 24(1):16–28.

Talkin, D. (1995). A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, pages 495–518.

Wang, L., Zhao, Y., Chu, M., Chen, Y., Soong, F., and Cao, Z. (2006). Exploring expressive speech space in an audio-book. *3th International Conference on Speech Prosody*, page 182.

Zhao, Y., Peng, D., Wang, L., Chu, M., Chen, Y., Yu, P., and Guo, J. (2006). Constructing stylistic synthesis databases from audio books. In *7th Annual Conference of the International Speech Communication Association (Interspeech 2006)*, pages 1750–1753, Pittsburgh, Pennsylvania, U.S.A.