Thèse de doctorat de

UNIVERSITÉ DE RENNES 1

École Doctorale N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Aghilas SINI

Caractérisation et génération de l'expressivité en fonction des styles de parole pour la construction de livres audio

Thèse présentée et soutenue à Lannion, le 02 Octobre 2020
Unité de recherche : IRISA UMR 6074
Thèse N° :

Rapporteurs avant soutenance :

Yannick Esteve	Professeur à l'Université d'Avignon et des pays de Vaucluse
Anne-Catherine Simon	Professeure à l'Université Catholique de Louvain

Composition du Jury :

Présidente :	Sylvie Gibet	Professeure à l'Université de Bretagne Sud
Examinateurs :	Laurent Besacier	Professeur à l'Université Joseph Fourier
	Sylvie Gibet	Professeure à l'Université de Bretagne Sud
	Simon King	Professeur à l'Université d'Édimbourg
Dir. de thèse :	Damien Lolive	Maitre de Conférence-HDR à l'Université de Rennes 1,
Co-dir. de thèse :	Élisabeth Delais-Roussarie	Directrice de recherche CNRS-Univérsité de Nantes

Table of Contents

Acronyms

Synthèse en Français **1**

1 Introduction . . . 1
2 Approches proposées . . . 2
2.1 Construction de corpus . . . 2
2.2 Étude émotionnelle de corpus SynPaFlex . . . 3
2.3 Étude discursif des livres audio . . . 5
2.4 Identité prosodique d'un locuteur dans un système de synthèse vocale multilocuteurs . . . 5
3 Perspectives . . . 6
3.1 Perspective à court terme . . . 6
3.2 Perspective à long terme . . . 7
4 Discussion générale . . . 8

Introduction **1**

1 Text-to-Speech Synthesis **5**

1 Text-To-Speech Synthesis System . . . 5
1.1 Front-End . . . 5
1.2 Back-End . . . 7
2 Statistical Parametric Speech Synthesis . . . 8
2.1 Overview . . . 9
2.2 Evaluation . . . 10
3 Expressive Speech Synthesis . . . 14
3.1 What do we mean by "expressive speech synthesis"? . . . 14
3.2 Transversal questions . . . 15

2 Speech Prosody **17**

1 What is prosody? . . . 17

2 Roles of speech prosody . . . 19
2.1 Linguistic . . . 19
2.2 Para-linguistic . . . 19
2.3 Extra-linguistic . . . 20
3 Prosody Modeling for Text-to-Speech Synthesis . . . 20
3.1 Rule-based methods . . . 21
3.2 Statistical data-driven methods . . . 21
3.3 Hybrid approach . . . 21
4 What are the topics discussed in this manuscript? . . . 22

3 Audiobooks Corpora For Expressive Speech Synthesis 23
1 SynPaFlex Corpus . . . 23
1.1 Motivation . . . 24
1.2 Relation to previous work . . . 24
1.3 Data Collection and Pre-processing . . . 25
2 MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA) Corpus . . . 29
2.1 Motivation . . . 29
2.2 The novelty of this work . . . 30
2.3 General Overview . . . 30
3 Gap between Text-to-Speech (TTS) designed corpora and amateur audiobook recording . . . 32
3.1 Data and features extraction . . . 32
3.2 Results . . . 35
3.3 Discussion . . . 36
4 A Phonetic Comparison between Different French Corpora Types . . . 38
4.1 Corpus design . . . 39
4.2 Data processing . . . 40
4.3 Results and discussion . . . 40
4.4 Remarks . . . 41
5 Conclusion . . . 41

4 Annotation Protocol and Emotional Studies of SynPaFlex-Corpus 45
1 Introduction . . . 45
2 Speech annotation . . . 45

2.1 Protocol . . . 46
2.2 Intonation Patterns . . . 46
2.3 Characters . . . 49
2.4 Emotions . . . 51
2.5 Other Events . . . 53
3 Evaluation of the emotion annotation . . . 53
3.1 Data analysis . . . 54
3.2 Methodology . . . 54
3.3 Results . . . 56
4 Emotion Lexicon Study of Audiobooks . . . 57
4.1 Proposed Method . . . 58
4.2 Pre-processing stage . . . 59
4.3 Features Selection . . . 60
4.4 Clustering Stage . . . 60
4.5 Acoustic Analysis . . . 62
4.6 Experiments and Results . . . 62
4.7 Discussion and issues . . . 64
5 Conclusion . . . 64

5 Automatic Annotation of discourses in Audiobooks 67
1 Introduction . . . 67
2 Corpus and material . . . 71
3 Rule-based Approach . . . 71
3.1 Rule-based results . . . 74
4 Machine learning approach . . . 76
4.1 General Methodology . . . 77
4.2 Data used and feature extraction . . . 77
4.3 Experimental setup . . . 78
4.4 Results . . . 78
5 Conclusion . . . 80

6 Automatic prosodic analysis of discourse changes 81
1 Introduction . . . 81
2 Corpus Design . . . 83
2.1 Experimental dataset . . . 83

2.2 Preprocessing . . . 84
2.3 Text annotation . . . 84
3 Prosodic analysis . . . 87
3.1 Features Extraction . . . 87
3.2 Hypothesis . . . 89
4 Results and discussion . . . 89
5 Conclusion and perspectives . . . 91

7 Speaker Prosodic Identity 93
1 General Context . . . 93
2 Introduction . . . 93
3 Speaker Coding . . . 94
3.1 OneHot-Vector . . . 95
3.2 X-Vector . . . 95
3.3 P-Vector . . . 95
4 Analysis Methodology . . . 96
4.1 Input and Output features . . . 96
4.2 Method . . . 97
5 Experimental setup . . . 98
5.1 Dataset . . . 98
5.2 Models configuration . . . 99
6 Results . . . 99
6.1 Standard measurements . . . 99
6.2 Visualizing the first hidden-layer output . . . 100
6.3 Subjective Evaluation . . . 102
7 Conclusion . . . 103

Conclusion 107
Summary of the Contribution . . . 107
Further Issuer . . . 109
Perspectives . . . 114
General Discussion . . . 115

A Audiobooks Corpora 117
1 SynPaFlex Corpus . . . 117

2 SynPaFlex Annotated Subset . . . 119
3 MUFASA Corpus . . . 119
4 MUFASA Parallel Subcorpus . . . 133

B Data visualization and high dimension reduction 137
1 Principal Component Analysis (PCA) . . . 137

C Discourses Annotation 139
1 Speech Verbs List . . . 139

D Manual Annotation and Subjective Assessment Materials 143
1 Intonation Patterns . . . 143
1.1 Exclamation pattern . . . 143
1.2 Nopip pattern . . . 144
1.3 Nuance pattern . . . 145
1.4 Resolution pattern . . . 146
1.5 Suspense pattern . . . 147
1.6 Note pattern . . . 148
1.7 Singing pattern . . . 149
2 List of stimulis . . . 151
3 Subjective Assessment Platform . . . 154

E Futur Work 155
1 End-to-End Tacotran-2 Architecture . . . 155

Bibliography 156

List of Figures

1.1 Text-to-Speech (TTS) system pipeline 5

3.1 Overview of the Speech Segmentation process 26
3.2 the vowel trapezoids of the three cardinal vowels /u/, /i/, and /a/ 37
3.3 Pauses distribution and average duration for *"Mademoiselle Albertine est partie"* 38
3.4 Pauses distribution and average duration for *"Vingt mille lieues sous les mers Chapter 3".* 38
3.5 The vowel trapezoids of the three cardinal vowel, in the context of occlusive /p/,/t/,/k/ 42
3.6 The density distribution according to the duration of the three vowels preceded by an occlusive consonant /p/,/t/,/k/ 43

4.1 The ten fundamental intonations defined in [Delattre 1966], illustrated by a dialogue: - *Si ces oeufs étaient frais j'en prendrias. Qui les vend? C'est bien toi, ma jolie? - Évidemment, Monsieur. - Allons doc! Prouve-le-moi.* [- If these eggs were fresh, I'd take some. Who sells them? Is it you, my pretty? - Of course it is, sir. - Come on, then! Prove it to me.] 48
4.2 Nuance Intonation Pattern Example : *puis il me semblait avoir entendu sur l'escalier les pas légers de plusieurs femmes se dirigeant vers l'extrémité du corridor opposé à ma chambre.* 49
4.3 A combination of three non exclusive intonation pattern. The nuance pattern is recognized with its particular pitch contour described in Figure D.3 *Dans cette cruelle position, elle ne s'est donc pas adressée* at begining of the utterance, followed by an emotional pattern characterized by the dynamic pitch (high F0-range) *à la marquise d'Harville, sa parente,* and finishing with an explicit question pattern *sa meilleure amie ?* 50
4.4 Scheme of proposed framework 58

4.5 The data points scatter in k = 18 groups - *doc2vec* features. The right-hand side shows the result of K-means, i.e., the data points of each cluster. The left-hand side shows the silhouette coefficient of each cluster. The thickness of each cluster plot depends on the number of data points lying in the cluster. The red bar is the average of the silhouette coefficient of entire clusters. 63

4.6 The data points scatter in k = 7 groups - *doc2vecs* + emotional vector features. The right-hand side shows the result of K-means, i.e., the data points of each cluster. The left-hand side shows the silhouette coefficient of each cluster. The thickness of each cluster plot depends on the number of data points lying in the cluster. The red bar is the average of the silhouette coefficient of entire clusters. 64

4.7 Principal Component Analysis (PCA) variation coverage of 73 % with 50 components, T-distributed Stochastic Neighbor Embedding (t-SNE) with perplexity of 45 and with iteration of 250 65

5.1 This figure illustrates the workflow guiding the rule-based approach. After the phonetization and forced alignment of the chapter text with the corresponding audio file, the data are segmented into paragraphs/ pseudo-paragraphs and stored relying on roots toolkit. The segments follow two annotations process: (i) the manual annotation made by an expert (ii) the automatic annotation which has two phases, the first phase consists of labeling the segments according to typographic criteria as DD, ID, and mixed group. The mixed groups are processed in phase 2 (Figure 5.2) in order to fine-tune the annotation and label the group according to Direct Discourse (DD), Indirect Discourse (ID), and Incidental Clauses with reporting verbs (IC). The mixed groups annotated, and non-mixed groups form the automatic annotation sequence. The two annotations (manual and automatic) are fused to form the Annotated Corpus (AC). 73

5.2 Detection and annotation of incidental clauses with reporting verbs (IC) . 73

5.3 Receiver Operating Characteristic (ROC) 79

6.1 Illustration of an example of discourse passage from Direct Discourse to Incidental Clauses with reporting verbs (Direct Discourse (DD) ⇒ Incidental Clauses with reporting verbs (IC)) corresponding to one modality and data structure. The tiers correspond (from the buttom to the upper one): Articulation Rate articulation rate measured with Equation (6.2) , Fundamental frequency (F_0)-range with Equation (6.1), syllables, words, breath group and related discourse. 86

7.1 Top part represents the architecture of the proposed model, the bottom part illustrates the visualization process of the first hidden layer. 98

7.2 Principal Component Analysis (PCA) projection for the parallel data during the validation phase, the speaker identify is encoded as following (F/M: Female/Male, FR: French, ID:XXXX). 100

7.3 Principal Component Analysis (PCA) projection for the non parallel data during the validation phase. 101

7.4 Visualization of the latent representation in case of P-Vector using parallel data. We can notice the separation of the speakers representation from epoch 5 to epoch 25. 101

7.5 Result of the MUSHRA of the listening test. 103

7.6 Ranking score of two representative speakers female (ffr001) and male (mfr0008), the present results are similar for the other speakers. 104

7.7 Ranking score of all speakers 105

D.1 *Avez-vous entendu ?* 143

D.2 La voiture arrivait près de Saint-Denis, la haute flèche de l'église se voyait au loin. 144

D.3 Nuance Intonation Pattern Example *puis il me semblait avoir entendu sur l'escalier les pas légers de plusieurs femmes se dirigeant vers l'extrémité du corridor opposé à ma chambre.* 145

D.4 −− Ma cravache, s'il vous plaît 146

D.5 – *Je ne les connais pas* 147

D.6 *[Note : me tendre un piège.]* 148

D.7 *...M'en allant promener, J'ai trouvé l'eau si belle Que je me suis baigné...* . 149

D.8 Screenshot of the platform PercEval (Recently renamed FlexEval [Fayet et al. 2020]) used for collecting the subjective assessment of the participants. *Question: asked question was: " For each sample, evaluate how similar it is to the reference (0 completely different, 100 completely similar)"* 154

E.1 Block diagram of Tacotran-2 [Shen et al. 2018; Oord et al. 2016] architecture 155

List of Tables

3.1 Validation results for the segmentation step per literary genre : lengths of the validation subsets, Phoneme Error Rate (PER), and average alignment error. 27
3.3 Amounts of linguistic units in the SynPaFlex corpus 28
3.4 The main linguistic content of MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA) Corpus 30
3.5 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. PODALYDES is a male voice. The speakers FFR0001, FFR0011, FFR0020, and MFR0019 are included in the MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA) corpus. 33
3.6 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. 35
3.7 The frequency of the {/ka/,/ta/,/pa/,/ti/,/ti/,/pi/} in the considered dataset, that have been manually annotated in terms of pitch amplitude. . 36
3.8 The set of extracts for conducting a comparative study. 39

4.2 Durations and amount of annotated data according to discourse mode in the first version of the SynPaFlex-Corpus 47
4.3 Manual annotations - Total duration of intonation patterns (including combinations) in the 13h25 sub-corpus 48
4.4 Manual annotations - Total durations of emotion categories labels (including combinations) in the 13h25 sub-corpus 52
4.5 Examples of perceived impacts of emotion on the speech 52

4.6 Number of manually 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. 54

4.7 Feature set of the INTERSPEECH 2009 Emotion Challenge 384 features, (16 LLD + 16 Δ)*12 functionals 55

4.8 Unweighted Average Recall (UAR) results for binary emotion classification using the three feature subsets. In bold, UAR > 60%, which we considered as a reasonable classification rate. 57

4.9 The best K-clusters according to the silhouette average criteria and average samples per cluster 63

5.1 Composition of the corpus according to types of discourse, selected from the corpus SynPaFlex describe in Section 1 71

5.2 Results of detection and annotation of discursive changes 75

5.3 Result of classification 80

6.1 Overview of the sub-corpus content. N-utt represent the number of utterances.s 84

6.2 Discursive changes distribution sub-corpus 87

6.3 Means and standard deviations for F_0-range and articulation rate (AR) for the different types of discourse change. 90

6.4 Means and standard deviations for Inter-Breath Group Pause Duration (IBGP) according to different types of discourse change. 91

6.5 Comparing IBGP across the different discourse changes modalities (** represents p-value<0.001). 91

7.1 Objective results for multi-speaker modeling, considering five speaker code configurations. Mel-Cepstral Distortion (MCD), Band Aperiodicity Parameter (BAP), Root Mean Square Error (RMSE), Voiced/Unvoiced (VUV) and Correlation (CORR) between the predicted and the original coefficients. For the Fundamental frequency (F_0), Root Mean Square Error (RMSE) and Correlation (CORR) are computed on the voiced frames only. 100

7.2 Objective results of the acoustic model, considering the three granularity. Mel-Cepstral Distortion (MCD), Band Aperiodicity Parameter (BAP), Root Mean Square Error (RMSE), Voiced/Unvoiced (VUV) and Correlation (CORR) between the predicted and the original coefficients. For the Fundamental frequency (F_0), Root Mean Square Error (RMSE) and Correlation (CORR) are computed on the voiced frames only. 113

A.2 MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA) corpus 119

A.3 MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis (MUFASA) Parallel Subcorpus 134

Acknowledgement

I would like to thank my thesis supervisors for their trust and their unwavering support.

My thanks go to all those who contributed to this modest thesis work. I would like to express my most enormous gratitude to the jury members.

My colleagues Antoine Perquin, Betty Fabre, Cédric Fayet,David Guennec, Lily Wadoux, Clémence Metz, Soumayeh Jafaraye and Meysam Shamsi.

Special thanks to the staff members who helped me a lot during this thesis: Angelique Le Pennec and Joëlle Thepault.

My mentors and friends Aditya Arie Nugraha, Arseniy Gorin, Anastasiia Tsukanova, Emilie Doré, Gaêlle Vidal, Ilef Ben Farhat, Imran Sheikh, Raheel Qader, Sunit Sivasankara, Sébastien Lemeguer, Motaz Saad, Manuel Sam Ribeiro, and Marie Tahon, my sincerest thanks.

Great thanks to the CSTR team at the University of Edinburgh. I would like to address a big thanks for their Accueil and their support during my internship.

My sincerest thanks to my parents, my sister Sarah, and my brother-in-law Sofiane Bennai.

This work would not have been possible without the incredible support and love of my dear wife, Lynda Hadjeras.

I dedicate this work to my family, my family-in-law, and my son Juba.

Acronyms

ABC Artificial Bee Colony

BAP Band Aperiodicity Parameter

CNN Convolutional Neural Network

CORR Correlation

DD Direct Discourse

DNN Deep Neural Network

doc2vec doc2vec

E2E End-to-End

$\mathbf{F_0}$ Fundamental frequency

FF-DNN Feed-Forward DNN

HMM Hidden Markov Model

IC Incidental Clauses with reporting verbs

ID Indirect Discourse

IPU InterPausal Unit

LA LitteratureAudio.com

LTS LETTER-TO-SOUND

LV LibriVox.org

MCD Mel-Cepstral Distortion

MFCC Mel-Frequency Cepstrum Coefficient

MGC Mel-Generalized Cepstrum

MMN Min-Max Normalization

MOS Mean Opinion Score

MUFASA MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis

MUSHRA Multiple Stimuli with Hidden Reference and Anchor

MVN Mean Variance Normalization

NLP Natural Language Processing

NLTK Natural Language Toolkit

OHV OneHot-Vector

PCA Principal Component Analysis

POS Part of Speech

RCNN Recurrent CNN

RF Random Forest

RMSE Root Mean Square Error

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SFS Siwis French Speech

SGD Stochastic Gradient Descent

SHR Subharmonic-to-Harmonic Ratio

SPSS Statistical Parametric Speech Synthesis

SVD Singular Vector Decomposition

SVM Support Vector Machine

t-SNE T-distributed Stochastic Neighbor Embedding

TALN Traitement Automatique de Langues Naturelles

TTS Text-to-Speech

VUV Voiced/Unvoiced

Synthèse en Français

1 Introduction

Pour obtenir une voix de synthèse de qualitée utilisable dans des contextes particuliers, il est fondamental d'améliorer l'expressivité de la parole car elle transmet les émotions, les intentions et les états d'esprit des locuteurs. Une part importante de l'expressivité de la voix est liée au contexte d'élocution et notamment influencée par le type de texte lu. Tous ces éléments participent à ce que l'on peut nommer des styles de parole. Les poèmes, les contes, les discours politiques ou les journaux télévisés sont des textes dont l'oralisation se fait selon des styles différents; et de nombreux lecteurs, s'ils lisent des documents intégrant plusieurs types de textes ou style de parole, sont capables d'adapter leur élocution aux textes à oraliser. La caractérisation des styles de parole à partir de l'étude de différents paramètres prosodiques (rythme, intonation, etc.) et segmentaux (réalisation des segments, liaisons, etc.) est une étape fondamentale. Les résultats de ces analyses serviront de base à la construction de modèles permettant aux systèmes de synthèse de générer des styles de parole divers. L'objectif est d'améliorer le contrôle et le rendu expressif des systèmes de synthèse de la parole.

Le traitement de l'expressivité dans la parole et l'adaptation de la prosodie à des styles particuliers constituent des questions de recherche importantes à l'heure actuelle. Des études très récentes comme [Govind and Prasanna 2013], mettent en avant le manque de naturel et de qualité dans la parole synthétique expressive. Concernant les styles de parole, [Obin 2011] propose un modèle permettant la génération de quelques genres de discours. Dans [Avanzi et al. 2014] sont présentés quelques résultats d'une étude récente visant à déterminer les principaux éléments caractéristiques de quelques genres en vue de les re-synthétiser. Les changements de styles, comme lors du passage au style direct et l'expression par la parole de certaines émotions sont au cœur des travaux réalisés. Pour cela, dans un premier temps on s'intéressera à l'expression portée par des livres audio car ces types de textes permettent de regrouper certaines de ces caractéristiques.

2 Approches proposées

Dans cette thèse, nous avons exploré l'expressivité de la parole à travers des livres audio. Afin de développer des modèles autorisant un meilleur contrôle de l'expressivité en synthèse de parole, ou d'adapter la prononciation et la prosodie au type de discours (changement dans la perspective du discours, style direct/indirect, etc.), nous avons construit deux corpus de livres audio français complémentaires SynPaFlex-Corpus et MUFASA.

2.1 Construction de corpus

SynPaFlex-Corpus [Sini et al. 2018] est un corpus de livres-audios en français composé de 87 heures de parole de bonne qualité, enregistré par une unique locutrice. Il est constitué d'un ensemble de livres de différents genres. Ce corpus diffère des corpus existants, constitués généralement de quelques heures de parole mono-genre et multi-locuteurs. La motivation principale pour construire un tel corpus est l'exploration de l'expressivité à travers différents points de vue, tels que le style de discours, la prosodie, la prononciation, et en utilisant différents niveaux d'analyse (syllabe, mot prosodique ou lexical, groupe syntaxique ou prosodique, phrase, paragraphe). Le corpus a été annoté automatiquement et fournit des informations telles que les labels et frontières de phones, les syllabes, les mots et les étiquettes morpho-syntaxiques. Pour pouvoir étudier les différentes stratégies de lecture adaptées par différents locuteurs nous avons construit MUFASA qui comprend une vingtaine de locuteurs français et contient environ 600 heures de parole de bonne qualité. Dans le chapitre 3, nous avons montré que sur des données comparables, les enregistrements amateurs[1] et professionnels[2] présentent des similitudes en matière de propriétés phonétiques et prosodiques. En revanche la qualitée de la parole du corpus MUFASA est légèrement inférieure à celle des corpus professionnels, ceci est dû notamment aux conditions d'enregistrement, néanmoins la quantité et la diversité des données permettent d'explorer de nouveaux horizons de la parole expressive lue et de développer des systèmes de synthèse de la parole plus performants. Ensuite, nous avons mené une expérience dans le but de comparer des extraits du corpus MUFASA avec d'autres corpus français bien connus pour mesurer la similitude d'un point de vue phonétique. Comme nous y attendions, MUFASA présente de grande similitude avec le corpus BREF [Larnel, Gauvain,

[1]Les enregistrements amateurs sont des enregistrements audio dont la destination primaire n'était pas pour faire de la synthèse de parole et dont les conditions d'enregistrements ne sont pas connues.

[2]Les enregistrements professionnels en revanche qu'on a eux ont été conçus pour développer des voix de synthèse. Les conditions d'enregistrement sont propres.

and Eskenazi 1991], qui est également un corpus de parole lue. Pour aborder l'expressivité portée par les corpus que nous avons construits, nous avons proposé d'articuler ce travail de thèse sur trois thématiques:

- Les émotions interviennent à des moments précis du discours pour animer le discours et lui donner de la profondeur.
- Pour structurer et apporter une cohérence à l'histoire, les auteurs utilisent différents modes de discours.
- Dans les livres audio, les émotions et les discours dépendent du texte autant que du signal de la parole. Le signal de parole dépend des propriétés du locuteur, qui constitue la troisième thématique abordée dans ce travail de thèse.

2.2 Étude émotionnelle de corpus SynPaFlex

Pour étudier les caractéristiques émotionnelles des données, nous concentrons nos efforts sur la voix présente dans le corpus SynPaFlex. Pour mener les expériences, une part significative du corpus a été annoté manuellement pour encoder le style direct/indirect et des informations d'ordre émotionnel.

Pour ce faire, nous avons demandé à un annotateur expert en parole de sélectionner un extrait représentatif et d'annoter le signal de parole. L'annotation manuelle a fournit quatre transcriptions complémentaires:

- La transcription de contour intonatif: cette annotation s'appuie sur les travaux de [Delattre 1966]. De cette annotation huit patrons intonatifs principaux sont encodés: question (interrogative), note, nuance, suspense, résolution (autoritaire ou impérative), chant, et nopip (aucun patron intonatif particulier).
- La transcription du discours des personnages[3] impliqués dans les livres sélectionnés en attribuant un identifiant unique et une identité vocale en tenant compte des performances du locuteur.
- Pour étiqueter le signal de parole en ce qui concerne les émotions, l'approche catégorique des émotions a été adoptée car c'est celle qui est la plus répandue actuellement. Six émotions de base définies par [Ekman 1999] sont utilisées : colère,

[3]Le narrateur est aussi considéré comme un personnage

joie, tristesse, surprise, dégoût et peur. Deux étiquettes supplémentaires ont été ajoutées : menace et ironie.

- Transcription de phénomènes complémentaires aux patrons intonatifs et émotifs contenant des événements phonétiques et linguistiques tels que les bruits, césure (notamment liaison sans enchaînement), murmuré ou mi-voisé, langue étrangère, paraverbal et musique.

A l'issue du processus de l'annotation manuelle, nous avons voulu reproduire l'étiquetage émotif à l'aide de techniques d'apprentissages automatiques et des procédures bien établies dans la reconnaissance automatique des émotions. Pour ce faire, nous nous sommes appuyés sur la méthodologie proposée par [Schuller et al. 2013b] présentée dans le challenge paralinguistique de 2013. Cette méthodologie s'appuie sur des techniques d'apprentissage supervisées pour construire un modèle de prédiction et d'étiquetage automatique en émotions de segment de parole.

Cette technique comporte deux étapes :

1. étape d'apprentissage; le modèle est entraîné avec des données dont on connaît la réalité du terrain, à l'issue de cette étape un modèle est appris.

2. étape de test; il s'agit de tester le modèle appris lors de l'étape d'apprentissage et d'évaluer le modèle.

Pour éviter le sur-apprentissages[4], il est d'usage de faire recours à la technique de validation croisée. Cette méthode a été mise en œuvre en utilisant le sous corpus SynPaFlex manuellement annoté, les résultats des expériences ont mis en évidence la subtilité des émotions dans ce type de données. Sur la base de ce constat, nous avons proposé d'explorer les questions relatives aux émotions par l'analyse des propriétés lexicales et sémantiques des transcriptions de livres audio. Pour réaliser ces expériences, nous avons privilégié les approches non supervisées. Cette seconde expérience est basée sur les techniques d'analyse des sentiments et de traitement du langage naturel. Le processus consiste principalement à trouver une représentation numérique adéquate des textes. Pour ce faire, nous avons choisi le modèle doc2vec pour la numérisation des phrases issue du texte, puis une méthode regroupant automatiquement le texte intégré selon des affinités lexico-sémantiques en utilisant l'algorithme de K-moyennes. Une fois les clusters formés, la dernière étape consiste

[4]Le surapprentissage est notion du domain d'apprentissage automatique, qui fait référence aux modèles peu généralisable

à interpréter les clusters dans l'espace acoustique. Les résultats montrent qu'il existe une forte corrélation entre la représentation du texte et les caractéristiques acoustiques de la parole.

L'annotation émotionnelle présentée et étudiée est fortement dépendante des propriétés du vocalique et du style du locuteur.

2.3 Étude discursif des livres audio

Pour étudier le discours, nous avons d'abord construit un outil d'analyse et d'annotation des livres audio des textes considérant trois types de discours, à savoir le discours indirect, le discours direct et les incises de citation. Cet outil comprend deux approches. La première approche est basée sur des règles, qui consiste en un ensemble de règles dérivées de l'analyse des données et élaborées par des experts à l'aide des propriétés morpho-syntaxiques et typographiques du texte. Le second s'appuie sur des techniques d'apprentissage automatique; nous avons obtenu de meilleurs résultats avec les modèles d'apprentissage automatique et plus précisément les modèles réseaux de neurones récurrents. Pour mettre en évidence les propriétés prosodiques lors des changements de discours et comment les locuteurs gèrent les perspectives de changements discursif et de personnages. Nous avons proposé d'analyser ce phénomène à travers un ensemble d'indices prosodiques dérivés de l'Unité InterPausale (UIP) que nous considérons comme pertinents pour mesurer et étudier le discours. Nous avons expérimenté avec deux locutrices du corpus MUFASA lisant un seul et même texte. Les résultats confirment que les deux locuteurs marquent bien le changement de discours et que l'UIP est unité de parole adéquate pour l'étude des changements de discours; le registre (F0-range) et la durée de l'inter pause sont des indicateurs pertinents pour les changements discursifs.

2.4 Identité prosodique d'un locuteur dans un système de synthèse vocale multilocuteurs

Pour étudier les propriétés des locuteurs et l'impact de leurs styles d'élocution dans un système de synthèse vocale, il est important d'avoir une représentation couvrant les propriétés du locuteur indépendant du texte. Dans la littérature l'identité vocalique d'un locuteur donnée est souvent représenté selon des méta-information sous-forme d'encodage one-hot, qui fait souvent référence au genre et l'identité unique du locuteur, D'autres approches, consiste à dériver une représentation unique au locuteur à partir de carac-

téristiques acoustiques. X-vector est un exemple de représentation de locuteur à partir d'information acoustique, ce plongement de vecteur acoustique est dérivé d'un modèle de reconnaissance du locuteur pré-entraîné. Dans ce travail nous proposons d'avoir une nouvelle représentation du locuteur, intégrant cette fois-ci des caractéristiques prosodiques. Cette nouvelle représentation est dénommée P-Vecteur (P pour prosodique).

Pour évaluer et comparer ces trois configurations, nous avons mis en place un système de synthèse vocale multi-locuteurs basé sur des réseaux de neurones profond intégrant en entrée une des configurations représentant l'identité du locuteur et les informations linguistiques extraites à partir du texte. Pour évaluer ces différents systèmes de synthèses, nous avons effectué deux évaluations objectives : l'évaluation objective standard qui consiste à comparer les paramètres acoustique prédit et réel à l'aide de métrique adéquate, et l'évaluation objective visuelle, qui consist à projeter les sorties de la première couche cachée du réseau de neurone. En outre, nous avons mené une campagne d'évaluation subjective auprès de 30 natifs français. L'évaluation objective et subjective a montré que l'identité prosodique du vecteur P est capable de guider le système de synthèse vocale multilocuteur basé sur le DNN aussi bien que le vecteur X et le vecteur OneHot bien établis.

3 Perspectives

3.1 Perspective à court terme

Comme perspective, nous souhaitons reproduire le même schéma d'intégration et d'évaluation adapté afin d'étudier l'identité prosodique du locuteur pour les informations émotionnelles et les indices prosodiques liés au discours, qui sont encore au stade de l'analyse statistique et de l'évaluation objective. Ainsi, dans une perspective à court terme, nous visons à intégrer ces deux variables dans le cadre de la boîte à outils MERLIN [Wu, Watts, and King 2016] en nous appuyant sur la même procédure présentée dans [Malisz et al. 2017]. Concrètement, nous souhaitons insérer deux nouveaux modules basés sur les réseaux de neurones, l'un pour la construction d'un vecteur intégré de discours et l'autre pour un vecteur intégré d'émotion (EEV). Les deux modules seront insérés entre le module frontal, et les modules de durée et acoustique.

Pour évaluer les effets de ces deux modules (émotionnel et discursif), nous considérons deux modules subjectifs distincts: une évaluation pour chaque module. Pour l'évaluation du discours perceptuel, les stimuli seront des extraits issues du changement de mode de

discours (DD, ID, IC), deux questions sont prévues:

- Une question directe qui peut se formuler comme suite "remarquez-vous des changements dans l'échantillon de discours ? (oui/non), pour voir si le sujet a remarqué des changements;

- Une deuxième question "Quel type de changements percevez-vous ?

    1. la vitesse de la parole "rapide/lente"
    2. l'amplitude de la parole
    3. la durée de la pause plus courte/longue.

Un changement similaire Un processus d'évaluation sera mené pour évaluer l'impact du module émotionnel. Les stimuli seront les mêmes que ceux utilisés pour l'évaluation du module de discours, mais les questions ne seront pas les mêmes. Comme dans cette deuxième évaluation subjective, les questions seront "est-ce que vous reconnaissez une émotion dans cet échantillon de parole?" Si le sujet répond oui, une liste d'émotions sera présentée, suivie de l'intensité de l'émotion ou des émotions perçues car on suppose que le sujet peut attribuer pour un même échantillon plusieurs étiquettes d'émotion avec une intensité différente. Au-delà de l'analyse des résultats de l'effet respectif de chaque module, la combinaison des résultats est également considérée comme une perspective car elle permet de mesurer la corrélation entre le discours et l'émotion. s

3.2 Perspective à long terme

Dans une perspective concrète à long terme, nous prévoyons de changer d'environnement de développement passant ainsi de Merlin [Wu, Watts, and King 2016] à un cadre bout-à-bout, plus précisément au Tacotran2 [Wang et al. 2017b; Shen et al. 2018] disponible dans la boîte à outils ESPNET [Hayashi et al. 2020], pour obtenir une meilleure qualité de synthèse. Ensuite, nous envisageons de construire un module similaire à celui qui a été développé dans des perspectives à court terme. Toutefois, dans cette nouvelle configuration, nous fusionnerons les deux modules en un module unique reposant sur des réseaux neuronaux multitâches profonds [Liu et al. 2019]. Ce module sera formé en même temps que les modèles acoustiques.

4 Discussion générale

Dans cette thèse, nous avons abordé les caractéristiques prosodiques dans le cas de la synthèse de livres audio à travers trois dimensions :

- Émotions : l'intervention d'un orateur pour situer le contexte de l'histoire et fournir des éléments supplémentaires pour divertir l'attention de l'auditeur.
- La typographie du discours pour mettre en évidence les structures des textes et la corrélation avec des indices prosodiques.
- L'identité de locuteur avec pour objectif à long terme de mettre en évidence la stratégie de lecture.

Une lecture expressive se doit de respecter des contraintes syntaxiques, sémantiques, pragmatiques ainsi que la typologie du texte écrit. À cela s'ajoute la stratégie du locuteur lier au contrainte identitaire ainsi que la réalisation des émotions. La corrélation entre les trois paramètres explorés dans cette thèse rend difficile la mise en place d'un système de synthèse vocale expressif robuste et fiable. Démêler ces "trois paramètre" en utilisant des techniques de factorisation basées sur des algorithmes avancés d'apprentissage profond semble être intéressant, selon [Hsu et al. 2019; Mathieu et al. 2016].

Alors que [Brognaux 2015] explore l'expressivité à travers la parole spontanée, nous nous concentrons sur la lecture de textes écrits. Il sera intéressant de faire une comparaison entre parole spontanée, en particulier les commentaires sportifs et les textes lus, notamment les livres audio, pour trouver une représentation commune à la parole expressive. Les principaux résultats présentés dans cette thèse sont basés sur une perspective acoustique de la parole. Ce niveau de représentation de la prosodie est important mais pas suffisant pour caractériser le discours expressif porté par les livres audio. La représentation perceptive et linguistique de la prosodie est cruciale pour avoir une vision complète et pour valider les résultats présentés dans cette thèse.

Introduction

General Context

An expressive voice is centered on the listener; it aims to communicate precise information, particular emotions, to relate facts or events. This expressivity is achieved through a vocalic gesture with its intonational modifications of tone, pitch, timbre, the repetition of certain phonemes, the lengthening of other phonemes. These phonetic events allows encoding expressive units by taking into acount cultural habits, and they could be acommpanied by facial mimics or certains body gesture. It is possible for a non-deaf human being to become aware of the expressiveness of the message conveyed solely by the voice. A simple audio recording can be rich enough to capture an entire scene or event.

Audiobooks are a concrete example of the ability of an expressive voice to transcribe and convey emotions, the interactions between characters through dialogue, the narration of events, the description of places, and the account of time and space in which the literary work is set.

This thesis project aims to characterize the expressivity conveyed by audiobooks to improve speech synthesis systems. Text-to-Speech (TTS) systems aim to supply machines with expressiveness to facilitate human-machine interaction.

Expressive Speech Synthesis

Nowadays, speech synthesis from a text can achieve outstanding levels of quality. The use of large corpora of speech has mostly contributed to this success. Nevertheless, synthetic speech still lacks emotion, intention and style. At present, we are not able to synthesize a voice with the expressiveness needed for audiobook reading without recording a speaker to create a large corpus with this style.

Some works in the literature are interested in taking into account phenomena related to expressivity and bring interesting conclusions that partly allow us to characterize the functioning and materialization of these phenomena. Here, we intend to deal jointly with emotion, intention, and style of speech, since these notions are very closely linked in

practice. Our goal is really to integrate them into speech synthesis.

SynPaFlex-Project

The SynPaFlex project[5] mainly funded this thesis. The objective of the SynPaFlex project is to investigate the different characteristics that contribute to the expressiveness of a voice in order to build a prosody model and a pronunciation model adapted to one or several speakers. The use of these models will be explored in order to integrate expressivity into speech synthesis systems, notably through concatenation or parametric statistical models. The research work focuses on the French language.

The main challenges of the project lie in the feasibility of applications of expressive speech synthesis, applications which are still not very widespread at the moment. In particular, opportunities are to be expected in the field of video games (diversification of synthetic voices, creation of expressive voices adapted to the game situation), language learning (dictation, style of speech), and personal assistance.

Challenges

It is challenging to realize the expressiveness conveyed from a simple text. Information such as the position of the pauses and their duration according to the context, the intonation, the rhythm, and many other parameters are not encoded in the text. However, it is possible to derive this information by analyzing and modeling different prosodic descriptors responsible for a natural voice.

These prosodic descriptors vary according to the context and depend on the text to be read. For instance, poems cannot be read as a simple message. That is where the style of speech comes into play.

Texts in audiobooks have special properties compared to other spoken texts, because the texts are longer, carry the author's style, his intention, and each sentence has a particular context.

In addition to this, some parameters are speaker-dependent as based on the reading strategy of a given speaker. Furthermore, it is not easy to evaluate the quality or judge the strategy of a speaker.

[5]This project is funded by the National French Research Agency (ANR)

Document organization

Automatic characterization of prosodic descriptors responsible for expressiveness of the voice, and which can be integrated into a text-to-speech system, is still a challenge. This process requires several steps that will be addressed in this manuscript.

Chapter 1 and Chapter 2 are dedicated to state of the art of text-to-speech synthesis systems, and speech prosody modeling, respectively. The collection of audio data in sufficient quantity to highlight the properties of audiobooks, will be discussed in Chapter 3. Manual annotation as well as a quantitative study of certain aspects of the expressivity of audiobooks will be reported in Chapter 4. Chapter 5 will discuss automatic annotation of speech types in audiobooks, followed by a prosodic study of discourse changes, dialogues, and discourse markers in Chapter 6. Chapter 7 deals with the prosodic identity of a speaker in multi-speaker synthesis systems. The manuscript concludes with a general conclusion where we summarize the main contributions of this thesis as well as further issues and the perspectives in the future work.

Chapter 1

Text-to-Speech Synthesis

This chapter provides a general background of the field of TTS Synthesis, and It mainly focuses on issues relevant for this thesis. It is divided into three sections. In the first section, we give an overview of TTS systems, providing a generic technical background.In the second section, we focus on the parametric speech synthesis used in this thesis, including the evaluation methodology. And finally we present the main challenges of expressive speech synthesis.

1 Text-To-Speech Synthesis System

Speech synthesis systems aim to generate speech from a text. This process of encoding text into speech follows a path that is organized in most cases into modules. In the majority of systems, there are two main modules, back-end, front-end.

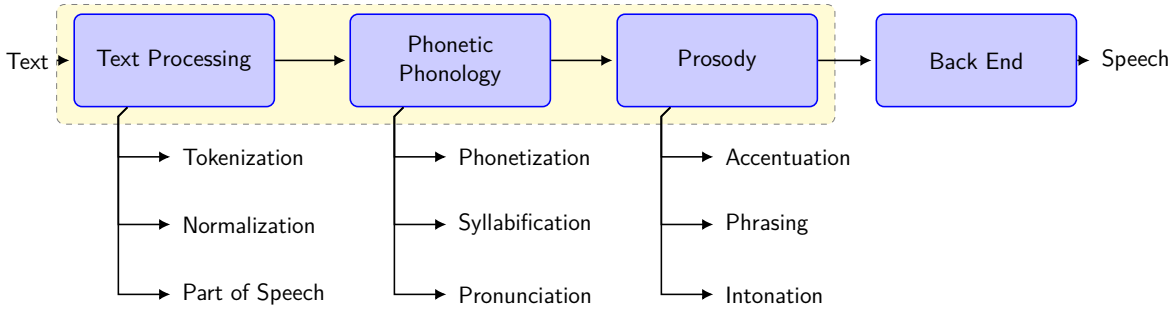


Figure 1.1: TTS system pipeline

1.1 Front-End

This first module receives the raw text as input and furnishes an output vector of the linguistic and prosodic specification.

Text Processing and decoding

In most systems, this module is used for tokenizing the raw text. Through a normalization module, each token is converted to an orthographic form, numbers, acronyms and so on being thus replaced. This stage is crucial for the following modules, it furnishes a normalized text.

Phonetic and Phonological

The sequence of words of the normalized text is phonetized either with a rule-based model or a statistical based model so-called LETTER-TO-SOUND (LTS). From the phoneme sequence, syllabification is processed, and pronunciation rules are applied to the chain according to language specification.

- Phonetization, Grapheme to Phoneme (G2P) [Novak, Minematsu, and Hirose 2016]

  Syllabification in French [Swaileh, Ait-Mohand, and Paquet 2016] we can destinc

  Pronouciation consiste of the way that

Prosody

The most challenging sub-module in the front-end module is the prediction of prosody. According to [Taylor 2009], prosody involves three phenomena:

- Accentuation or stressing is the act of emphasizing a particular speech sequence in the majority of languages, and this prominence appears at the syllable level. Prominences may assume functions such as emphases, stylistic variation. This phenomenon is language-dependent. For instance, there is no phonetic stress but only phrase stress in French, the prominence is on the last syllable of a word, unlike in English where there is lexical stress.

- Phrasing refers to the division of the speech flow into chuncks of different ranks.

- Intonation is the shape of the pitch contour at the sentence/phrasal level (to distinguish with lexical tones in tonal languages)

In [Avanzi, Simon, and Post 2016; Avanzi 2013], the authors have investigated the relation between accentuation and phrasing and the relation between intonation and rhythm was compared in the case of British English and French in [Patel, Iversen, and Rosenberg 2006].

Despite the fact that speech synthesis is now developed withing Deep Learning paradigms, prosody is still a research area in comparison to text processing and decoding modules.

For more details about prosody see Chapter 2.

1.2 Back-End

This module converts the intermediate linguistic specification into a synthetic speech waveform. In literature, this block is called the waveform generator. In most cases, this block is not language-dependent since the front-end has done most of the linguistic processing. Many approaches have been proposed to map the linguistic representation to speech.

Rule-based synthesis

This method is one of the oldest paradigms proposed for generating artificial speech. These approaches typically define a set of rules to artificially generate a waveform from a set of acoustic parameters such as formant frequencies [Klatt 1980]. The main advantage of this paradigm is that it offers a certain control over the synthesized speech. For instance, this paradigm has been used for synthesizing emotional speech [Schröder 2001]. This technique has two main disadvantages. The first one is due to the fact that systems are based on a set of rules, which are language-specific and typically require knowledge of experts, and the second one is the lack of naturalness and intelligibility.

Concatenative synthesis

This approach is one of the most common technique for waveform generation in industrial systems. This approach consists in concatenating pre-recorded units of speech to generate new waveforms. However, systems aiming for more generic synthesis focus on the concatenation of smaller units. Such units may begin, for example, at the mid-point of a phone and end at the mid-point of the following phone, thus capturing the co-articulation between the two phones. These units are diphones. Systems using a minimal database with

a single diphone sample are said to be diphone synthesizers [Moulines and Charpentier 1990]. Extensions of this idea vary the type and number of units present in the database. This generalization is referred to as unit selection [Hunt and Black 1996; Guennec 2016; Alain et al. 2016; Lolive et al. 2017]. Unit selection systems use a very large database with multiple samples of the same unit. The task of the waveform generator is then to select the optimal unit sequence given an input linguistic specification. It is hard to control the generated speech, as these techniques are not very flexible.

Parametric synthesis

This class uses parametric representations of speech waveforms, which are modeled via statistical frameworks. For this reason, these systems are often grouped under the term Statistical Parametric Speech Synthesis (SPSS)[Ze, Senior, and Schuster 2013]. Parametric systems offer several advantages over concatenative systems. For example, it is easy to see how unit selection systems can be limited by their database: larger databases allow the system to be more flexible, but also increase the number of resources needed. Parametric voices are flexible when it comes to the manipulation and control of acoustic parameters. This flexibility makes them attractive for various tasks such as speaker adaptation, multi-speaker speech, multilingual systems, voice conversion, and expressive speech. Additionally, parametric systems tend to benefit from a very small footprint when compared to standard unit selection systems. However, parametric voices suffer from various disadvantages.

Hybrid synthesis

This approach represent a class of techniques that combine unit selection and parametric methodologies. The most common hybrid approach [Tiomkin et al. 2010] uses a statistical framework to generate a sequence of acoustic parameters that are then used to guide the selection of units from the database.

2 Statistical Parametric Speech Synthesis

This thesis is mainly concerned with the statistical parametric approach for speech generation, more precisely on the Deep Neural Network based technique. Therefore, Section 2.1 will provide a further overview of this class of techniques.

2.1 Overview

In the context of speech synthesis, a vocoder, or voice encoder, extracts from the speech waveform a set of parameters that may be modeled statistically. Common approaches are based on the source-filter model of speech production. This model makes the assumption that speech is produced by first generating a source signal, which can be intuitively understood as air exiting the lungs and passing through the vocal folds. The positions of the vocal tract articulators (tongue, lips, oral, and nasal cavities) then act as a filter on the source signal. The source-filter model assumes that these two components are independents and vocoders aim to find representations that separate the effects of source and filter. Vocoders extract parameters over speech windows, referred to as a speech frame, and may, in common implementations, span 25ms. Each frame is assigned source (or excitation) parameters such as fundamental frequency and voicing information. Some vocoders include extra excitation parameters, such as band aperiodicities: this is the case of WORLD[Morise, Yokomori, and Ozawa 2016] and STRAIGHT[Kawahara, Masuda-Katsuse, and De Cheveigne 1999] vocoders.

For the filter (or spectral envelope) parameters, Mel-cepstrum coefficients [Fukada et al. 1996] are often used. Alternatively, one can use Mel-generalized cepstral coefficients [Tokuda et al. 1994] or line spectral pairs [Itakura 1975]. The speech waveform can be analyzed and reconstructed with minimal error via these speech parameters.

Recent approaches using neural networks for SPSS aim to overcome the Hidden Markov Model (HMM) based speech synthesis systems. There has been a considerable amount of earlier work using neural networks for speech synthesis [Wang et al. 2017a; Ping et al. 2017; Tachibana, Uenoyama, and Aihara 2018]. However, recent improvements in software, hardware, and data availability have caused a huge interest in these methods.

In this thesis, we use a framework such as the one described in [Ze, Senior, and Schuster 2013]. This method was implemented in the Merlin Neural Network Toolkit [Wu, Watts, and King 2016]. During the data preparation stage, we have used the JTrans [Cerisara, Mella, and Fohr 2009] software to force align the data at the phone-level, from which phone alignment can be inferred. Given this alignment between linguistic features and acoustic parameters, a Feed-Forward DNN (FF-DNN), called the acoustic model, can be trained using mini-batch Stochastic Gradient Descent (SGD). An additional neural network, called the duration model may be trained in a similar fashion to model phone durations.

Finally, a vocoder is used to synthesize the waveform. This framework is used for

DNN-based speech synthesis.

2.2 Evaluation

Two metrics are used to evaluate a synthetic speech system: objective and subjective assessments. The objective evaluation consists of measuring the acoustic and duration distance between the synthetic speech and natural one relying on different metrics, calculating the recognition rate using a speech recognition technique. These measures are not reliable for measuring the intelligibility and naturalness of synthetic speech. For this reason, a subjective assessment is somehow mandatory for validating and reinforcing objective assessment.

Objective Evaluation

In statistical parametric speech synthesis, objective evaluations compare a sequence of acoustic parameters generated from a model with a reference sequence extracted from a waveform. Most objective metrics are distance measures between the two sequences. The underlying assumption is that the distance between the sequences is meaningful in terms of the quality of the model. That is, the smaller the distance between generated and reference parameters, the better the model. However, it is not always the case that objective measures are representative of the quality of the acoustic parameters. Averaging over datasets might dilute otherwise perceptible acoustic differences between systems. Objective measures can still be useful as they are fairly easy to compute and they facilitate comparisons over a large number of systems.

In this section, we give a brief overview of the main measures used in this work. When appropriate, these are computed according to the Merlin Neural Network Toolkit [Wu, Watts, and King 2016] and the equations presented here reflect that implementation. Note that during this thesis we keep the default configuration proposed by the framework.

Objective metrics are sensitive to the vocoder used. In this thesis, we use [Morise, Yokomori, and Ozawa 2016] and these measures are computed accordingly. Mel-Cepstral Distortion (MCD) measures the distance between two sequences of Mel-Frequency Cepstrum Coefficient (MFCC). We are given a reference vector x and a generated vector $\hat{x}$ of MFCC coefficients. MCD is then computed as an extension of the standard Euclidean distance:

$$MCD = \frac{\alpha}{T}\sqrt{\sum_{d=1}^{D}(x_d(t) - \hat{x_d})^2}\alpha = \frac{10\sqrt{2}}{\ln 10}$$

with :

$$\alpha = \frac{10\sqrt{2}}{\ln 10}$$

where T is the total number of frames in the data set and D is the dimensionality of the MFCC extracted at each frame. In this thesis, we use 60 coefficients per speech frame. Following [Kominek, Schultz, and Black 2008], the constant α is included for historical reasons. Note that we exclude the first coefficient, commonly associated with the energy of a speech frame. This prevents the distance measure from being influenced by loudness, which may affect some datasets, such as non-professional audiobooks [Kominek, Schultz, and Black 2008].

Band Aperiodicity Parameter (BAP) distortion follows the same intuition (and notation) as MCD. For each frame, a D-dimensional vector of parameters is extracted to represent the source excitation signal. In this thesis, we extract 25 band aperiodicities and we compute the distortion between natural and predicted parameters.

$$BAP = \frac{1}{10T}\sum_{t=1}^{T}\sqrt{\sum_{d=1}^{D}(x_d(t) - \hat{x_d})^2}$$

In terms of objective measures related to the f0 signal, we have used the root-mean-square error and Pearson's product-moment correlation. These are standard measures in the literature, although alternatives have been suggested[Clark and Dusterhoff 1999]. For the purpose of this thesis, these measures are computed at utterance-level on voiced-frames only and the average of all utterances in the test set is reported.

For a given utterance u, the root-mean-square error of the f0 signal is determined as

$$RMSE_u = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_u(n) - \hat{x}_u(n))^2} \tag{1.1}$$

$$RMSE = \frac{1}{U}\sum_{u=1}^{U} RMSE_u \tag{1.2}$$

Similarly, the correlation of the f0 signal is determined as:

$$r_u = \frac{\sum_{n=1}^{N}(x_u(n) - \overline{x}_u)(\hat{x}_u(n) - \overline{\hat{x}}_u)}{\sqrt{\sum_{n=1}^{N}(x_u(n) - \overline{x}_u)^2}\sqrt{\sum_{n=1}^{N}}\hat{x}_u(n) - \overline{\hat{x}}_u)^2} \tag{1.3}$$

$$CORR = \frac{1}{U}\sum_{u=1}^{U} r_u \tag{1.4}$$

Where $\overline{x}_u$u and$\overline{\hat{x}}_u$ u denote the mean value of the reference and the generated $f0$ signal for utterance u, respectively. Note that the f0 correlation is here implemented as Pearson's product-moment correlation coefficient. Intuitively, this measure captures the similarity between the overall shape of generated and reference $f0$ signals, which is particularly relevant for intonation. While distance-based objective measures aim to be minimized, the signal's correlation aims to be maximized. Finally, in some chapters of this thesis, voicing error is reported as the percentage of frames that were assigned the incorrect voicing label.

Subjective Evaluation

Objective evaluation methodologies are often used as an indication of the quality of synthetic speech, especially when a large number of systems are being developed, and reference acoustic parameters are available. However, it is widely agreed that subjective listening tests still remain the standard method for the evaluation of synthetic speech.

The subjective evaluation of synthetic speech is not a simple task and still quite challenging. The majority of evaluations of systems focus on naturalness and intelligibility. In recent years, with the developments of speaker adaptation, multi-speaker modelling, and voice conversion techniques, speaker similarity has been adopted as a third dimension in the evaluation of speech synthesis systems.

Subjective evaluation methods are able to provide more accurate quality measurements than objective evaluation methods, but they also tend to be costly. Listening tests typically require a large investment in terms of time and resources, as they require well-designed experiments and listeners.

When we design perceptual listening tests several factors should also be considered for the evaluation of synthetic speech. For instance, the type of test, the question being asked, or the type and number of listeners. We briefly provide a review of well-established protocols for the evaluation of naturalness, and some methods used for the evaluation of intelligibility and comprehension.

In [Fonseca De Sam Bento Ribeiro 2018], the author grouped the protocols for the evaluation of naturalness into two main classes:

- Referenced methods, in which a synthetic sample is judged against an available natural reference.

- Non-referenced methods, in which synthetic samples do not have an available reference and are instead judged against the listener's expectations.

In both cases, naturalness means close to human voice properties with a perceptual point of view. There are many evaluation methods in literature, we will briefly give the most used for evaluating synthetic speech synthesis.

- Mean Opinion Score (MOS) is a non-referenced evaluation methodology [ITU-T and Recommend 1996]. Listeners are not given a speech reference to anchor their judgments. In a MOS evaluation, listeners are presented with one speech sample at a time. They are then asked to judge that sample on a 5-point scale in terms of quality, where 1 indicates bad and 5 indicates excellent. This methodology has variation such as DMOS (Differential MOS), which is a referenced version of the MOS test. Listeners provide their judgments for individual samples with respect to a reference sample. CMOS (Comparison MOS) presents the listeners with two randomized samples from different conditions.

- Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)[Schoeffler et al. 2015]: With this approach, listeners are presented with many conditions at once and they are asked to provide a subjective rank of the conditions with respect to each other and to an explicit reference. A copy of the explicit reference is hidden within the remaining experimental conditions, which fixes an upper bound for the listeners' judgments. In the MUSHRA paradigm, listeners provide absolute scores measuring the similarity of synthetic samples with respect to a reference. But because all conditions are rated simultaneously, multiple comparisons across conditions are also provided. This implicitly creates a ranking of systems, which might be preferable over an absolute score. Ranking scores can be interpreted as a preference judgment, while absolute scores can be interpreted as a measurement of that preference. In chapter 7, we use this methodology to evaluate a multi-speaker speech synthesis system.

- In the AB test, also called preference test, listeners are given a randomized pair of samples and are asked to express their preference with respect to some speech attributes. The most common question under this framework asks listeners to select the sample which sounds more natural.

A similarity evaluation aiming at understanding perceptual similarities across multiple systems may also be used [Mayo, Clark, and King 2005]. Similarity means the given condition and the reference are perceptually similar.

Intelligibility and comprehension evaluation: with this type of methodologies, we focus on the intelligibility and comprehension of synthetic speech as well as natural speech.

3 Expressive Speech Synthesis

In order to obtain a synthetic voice of good quality that we can use in particular contexts, it is fundamental to improve the expressiveness of speech, which is a vector of emotion, intention, or state of mind.

3.1 What do we mean by "expressive speech synthesis"?

[Brognaux 2015] defines expressive speech as *"any aspect of speech that makes it more natural-sounding and suited for a specific communicative situation other than reading non-emotional laboratory speech (that can be seen as a typical example of what we designate as 'neutral speech')."* In this work, we were somehow extending or simplifying this description by considering that human speech is always expressive as long as it is carrying emotion, intention, or state of mind of the speaker. This aspect is manifested and leveraged by the speaker according to the context "circumstances," and speaking style. Furthermore, we consider that expressive speech demands less cognitive load[1] from the listener's point of view; it makes the message among interlocutors easy in dialogue, for instance, and understandable when it is a political speech, for example. These aspects contribute to settling the definition of expressive speech as well as the primary motivation for building an expressive speech synthesis system.

[1]The reader interested in an in-depth discussion of the concepts related to cognitive load, speech production, and perception can refer to [Christodoulides 2016]

For instance, in audiobooks, the reader is constrained by the text "circumstance or context". Consequently, the produced speech has to be expressive to keep the listener's attention and give access to the various information carried by the text read, including story plot and author intention.

3.2 Transversal questions

To address the expressivity of speech in audiobooks, we need to define the speaking style and the level of granularity that we consider to study this kind of data.

Speaking Style

A significant part of the expressiveness of the voice is related to the speech context and is influenced in particular by the type of text read. All of these elements contribute to what can be called speech styles. Poems, stories, political speeches, or television news are texts vocalized according to different styles. Many readers, if they read documents integrating several types of texts or style of speech, can adapt their speech to the texts to be verbalized.

The characterization of speech styles based on the study of different prosodic parameters (such as rhythm, intonation) and segmental parameters (such as the realization of segments, links) is a fundamental step. The results of these analyses will serve as a basis for building models that allows synthesis systems to generate speech styles. The goal is to improve the control and expressive rendering of speech synthesis systems.

The treatment of expressiveness in speech and the adaptation of prosody to particular styles are important research questions at present. Recent studies like [Govind and Prasanna 2013; Jauk 2017], highlight the lack of naturalness and quality in expressive synthetic speech. In [Avanzi et al. 2014] are presented some results of a recent study which aimed at determining the main elements of characteristics of four genres, including a reading of fairy tales, dictations, political speeches, and reading of novels to re-synthesize them. Changes in style, such as changing to direct speech and the expression of certain emotions by speech, are at the heart of the work done. For that, at first, we will be interested in the expression carried by audiobooks because these types of texts allow covering some of these characteristics.

Beyond Sentence Level

To capture the style carried by a text, it is necessary to re-consider the linguistic unit used for building synthetic voice. Many information, such as expressivity and speaker attitude, can not be well characterized at a simple sentence level. For example, if we consider audiobook or dialogue, the sentence is not enough to describe the speaker strategy and to extract a consistent prosodic pattern for speech recognition or text-to-speech synthesis. Studying long text book paragraphs can be an alternative to sentence. According to [Farrus, Lai, and Moore 2016; Lai, Farrus, and Moore 2016; Doukhan 2013], prosodic cues assigned to paragraphs seem to be more relevant to study expressivity and speaker's reading strategy in audiobooks, and the authors claim that the speakers tend to reset the prosodic cues between paragraphs. In [Vaissière and Michaud 2006] the authors consider the paragraph as the largest unit defined by F0 fluctuation.

Chapter 2

Speech Prosody

This chapter aims at briefly introducing notions regarding speech prosody, its roles in communication, and at presenting the main paradigms used for modeling speech prosody in Text-To-Speech systems.

1 What is prosody?

There is no consensus on the definition of prosody; it depends on the level of analysis and representation. In this work, we have chosen the most relevant for the purpose of this thesis, which is the characterization and the generation of an adequate prosody to synthesize audiobooks. From this perspective and as defined by [Di Cristo 2013], prosody can be defined as a mechanism which supervises the management of a set of parameters that are:

- Fundamental frequency (F_0): The frequency of a sound corresponds to the number of vibrations per second, namely period: if there are few vibrations per second, we hear a low tone if there are more vibrations per second a high tone. The frequency is expressed in Hertz (Hz). This definition is valid for all types of periodical signals. However, speech is a complex signal, and it is not strictly periodic. In human speech, the principal frequency called F_0 corresponds to the frequency of the vocal folds. The F_0 is commonly referred to as pitch, which is the perceptual representation of F_0. In this work, these two denotations (F_0 and pitch) are interchangeable.

- The intensity depends on the amplitude of the vibration induced by the speech signal: the higher the amplitude, the louder the sound; the lower the amplitude, the weaker the sound. It is commonly expressed in decibels (dB).

- The duration depends on the time during which a speech unit is produced. The unit used is the second (s). Suppose we consider speech units such as phones and pauses. Their duration depends on the context in which they were generated.

From these basic parameters, prosodic elements are derived such as:

- Tone and intonation are elements of prosody that refer mainly to pitch patterns. A tone is a pitch contrast that is limited to the syllable or word, and is manifested by a relative difference in pitch between syllables or words that follow each other. This element is specific to tone languages such as Chinese. On the other hand, intonation is a pitch pattern imposed by the utterance for the purpose of expression other than the pitch difference between words or syllables. Intonation is sensitive to emotions, politeness, context, and in general, to speaking style. Unlike tones, intonation is present in all languages.

- Prominence: In [Büring 2016], prominence is defined as *"local valleys and peaks in the voice's fundamental frequency are perceived as prosodic prominence, understood as emphasis, and modeled as pitch accents."* Prominence is probably one of the broadest subjects of prosody, as many underlying notions are related to it (emphasis, accent, rhythm, and metrics). An utterance is a sequence of syllables. These syllables are not perceived as having a similar pitch level or energy. Some syllables appear more prominent because they are longer, or because they receive a particular prosody. Therefore, the terms of accent or accentuation relate to a phenomenon of prominence or local salience, which can assume in the language a metrical or pragmatic function. A distinction is thus made between metrical accentual phenomena and accentual phenomena with emphatic value. The study of metrical accentual phenomena should make it possible to account for the distribution of stressed syllables. In French, they play an essential role in the demarcation of prosodic phrasing, and therefore, also in interpreting the utterance. The study of metrical phenomena is based on a distinction between meter and rhythm. At the meter level, the study of a language's metrical functioning relies on defining the syllables that are likely to receive an accent in the language. Rhythm, on the other hand, is built from the syllables stressed in a given utterance.

- Phrasing and prosodic structure: An analysis of the speech flow highlights the establishment of chuncks composed of syllables or words. These phrases can be delimited by a pause or not. In the same way, these newly formed phrases can participate in building a larger phrase. This structure is referred to as a prosodic structure of utterance. The analysis of the prosodic structure requires to take into account syntactic and semantic information. However, in some instances, prosodic

groupings do not respect certain syntactic boundaries. Phrasing should help the interpretation and analysis of an utterance.

2 Roles of speech prosody

According to [Di Cristo 2013], many roles are attributed to prosody, including lexical, demarcative, pragmatic, behavioural, emotional, identifying, stylistic and many other roles. In [Christodoulides 2016] three prosody roles are reported: linguistic, para-linguistic, and extra-linguistic.

2.1 Linguistic

[Vaissière 1983], claim that some prosodic features fulfill similar roles across languages. For instance, pauses, some fundamental frequency features (declination tendency, resetting or baseline), durational features (final lengthening) and intensity. The author highlights the prosodic differences among languages (differences in timing, different orders of priorities, different relationships between F0, duration and intensity).

In [Cohen, Douaire, and Elsabbagh 2001], the authors have shown through two subjective assessments, including twenty subjects each, that altered prosody and punctuation similarly affect performance and seriously impair text comprehension and word recognition. The authors claim that linguistic prosody supplies redundant cues for judging sentence structure and manages attentional resources to help with the semantic encoding of lexical units and with the organization of linguistic information in long-term memory.

[Veenendaal, Groen, and Verhoeven 2014] found that speech prosody contributes significantly to the construction of the meaning of written texts. This result was founded by performing reading and language assessments over 106 subjects (Dutch fourth-grade primary school children) using storytelling task and oral text reading performance such as decoding skills, vocabulary, syntactic awareness, and reading comprehension.

2.2 Para-linguistic

Prosody is involved in several paralinguistic parameters, including speaker attitude, emotional state, affective, and cognitive states. [Liscombe 2007] has explored the primary information that are carried by prosody from three distinct speaker state-related perspectives: *a*) Paralinguistic: Pitch contour shape seems to discriminate emotions in terms of

positive and negative affect. *b*) Pragmatics: The phrase-final rising of the intonation pattern plays a crucial role in the studied case of student questions in a corpus of one-on-one tutorial dialogues. *c*) Proficiency: There is a significant correlation between intonational features, including syllable prominence, pitch accent, and boundary tones, and language proficiency assessment scores at a strength equal to that of traditional fluency metrics.

According to [Levin, Schaffer, and Snow 1982], it is possible to differentiate between story-reading and storytelling, relying on intonation prosodic features and paralinguistic features such as speakers' behavior in both situations.

2.3 Extra-linguistic

Extra-linguistic dimension is related to the speaker's physiological characteristics, and also to idiolectal, and geographical (ethnicity) information. Furthermore, all these pieces of information appear in speech communication through non-verbal vocalization.

[Labov 1970] describes "social speech registers" such as styles or modes of social speech as extra-linguistics variables that contribute to the construction of a social situation. In [Levin, Schaffer, and Snow 1982], story-telling and story-reading are seen as a concrete example of "social speech registers." This example is defined as functions of communication that can designate the speech registers, the topic, the setting, social characteristics of the listeners and the speakers (e.g., age, sex, ethnicity, education, social class) and the relationships between the speaker and the listener [Gumperz 2009]. [Maekawa 2011] showed that prosodic parameters such as prosodic label frequency information and speaking rate allow to automatically discriminate four speech registers including academic presentation, simulated public speech, dialogue, and reproduction speech in the Corpus of Spontaneous Japanese (CSJ).

In [Trouvain 2014], the authors investigate three cues considered as non-linguistic features (that we call extra-linguistic) such as laughing, audible breathing, clicking in conventional speech. The studies highlight the importance of prosody in characterizing and describing the non-linguistic speech information.

3 Prosody Modeling for Text-to-Speech Synthesis

[Rajeswari and Uma 2012] describes the prosody modelling for TTS as "the process of building computational models to produce prosodic variations in synthesized speech

automatically". The authors report three approaches that are discussed in the following.

3.1 Rule-based methods

These techniques require solid knowledges at the text processing stage as well as at the speech processing stage. Regarding text processing, techniques derive a formal description of the phonetic and phonological properties according to the context and studied language. The corresponding speech has to be consistent with text features. This approach is usually difficult to implement and expensive in terms of time because it requires manual annotation, which generates annotation conflicts between annotators in the case of several annotators. In most cases, the annotations are carried out on a small data set. Furthermore, the prosody models resulting from this method are highly dependent on the language for which they were designed as well as on the type of data studied.

3.2 Statistical data-driven methods

This technique relies exclusively on statistical analysis and modeling of the phenomena present in the data. We can cite the likelihood-based prosody model and posterior based prosody modeling. However, the most successful prosody modeling nowadays is the deep learning approach. This success is due to the availability of data and the computational power progresses of modern computers.

3.3 Hybrid approach

The hybrid model is a combination of both rule-based and statistical based approaches. The prosody modeling implemented in the Merlin toolkit[Wu, Watts, and King 2016] is an example of a hybrid model. This model requires the extraction of a set of linguistic features during the pre-processing process. These features are fed respectively to the duration model which makes predictions regarding duration, and an the acoustic model which predicts the prosodic features (mainly F0/pitch and spectral information of the speech).

For exploring and characterizing the expressivity of speech, sentence-based prosodic features are not enough.

4 What are the topics discussed in this manuscript?

It is clear that to obtain a synthetic voice of good quality that we can use in particular contexts, it is fundamental to improve the prosodic models used text-to-speech synthesis.

The French language prosody will be the focus of our investigations. Aspects related to the variation of parameters and to prosodic phenomena (such as intra-speaker and inter-speaker variability) will be addressed in chapter 4, followed by paralinguistic studies related to emotions conveyed by audiobooks. The relation between prosody and discourse will be studied in the chapters 5 and 6.

Chapter 3

Audiobooks Corpora For Expressive Speech Synthesis

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* [Sini et al. 2018].

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.

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 [Doukhan et al. 2015] 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 [Chevelu, Lecorvé, and Lolive 2014] 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

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

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 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.

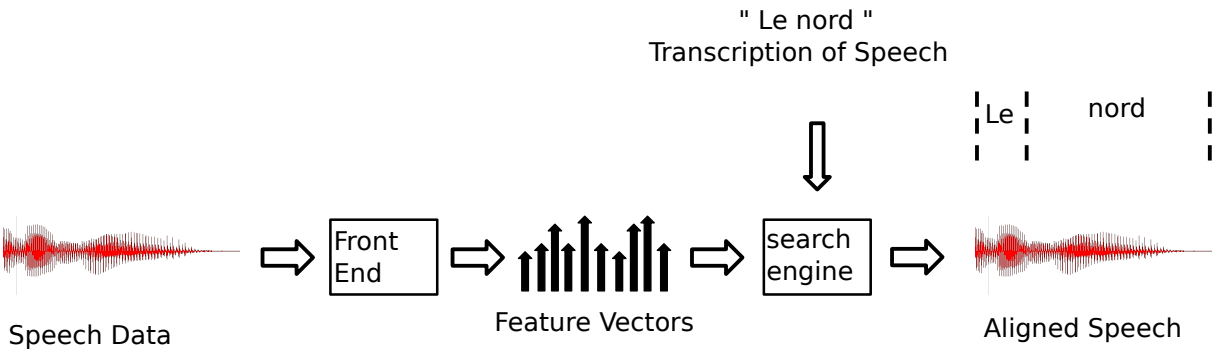


Figure 3.1: Overview of the Speech Segmentation process

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, Mella, and Fohr 2009].

To evaluate the accuracy of the phone segmentation, an expert annotator performed a manual validation using Praat [Boersma and Weenink 2016].

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 3.1. 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 3.1. 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

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

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

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 [Cerisara, Mella, and Fohr 2009] have been trained on the ESTER2 corpus [Galliano, Gravier, and Chaubard 2009] 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 [McAuliffe et al. 2017] tool which is based on Kaldi[Povey et al. 2011] 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 [Brognaux et al. 2012] 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.

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 3.3 sums up the content of the corpus in terms of linguistic units. However, we did not verify the precision of this annotation. We plan to include constituency-dependency parsing using BONSAI [4] presented in [Candito et al. 2010], which is more appropriate for long text and for French language syntactic parsing in a near future.

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 3.3: Amounts of linguistic units in the SynPaFlex corpus

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 [Gravier 2003], (ii) the instantaneous fundamental frequency (F_0) using the ESPS get_f0 method implementing the algorithm presented in [Talkin 1995], 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

[4] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html

Fundamental frequency in semi-tone[5] 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)[6] and LitteratureAudio.com (LA)[7] entirely dedicated to French audiobooks. 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[Zen et al. 2019] 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.

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

[6]https://librivox.org/

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

2.2 The novelty of this work

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

- Parallel data (Appendix 4): 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[8].

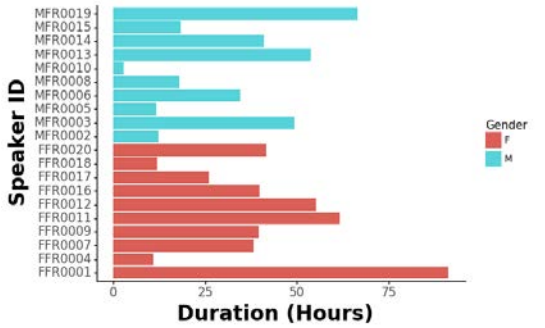
A summary of the contents of the MUFASA-Corpus is presented in the Table 3.4.

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

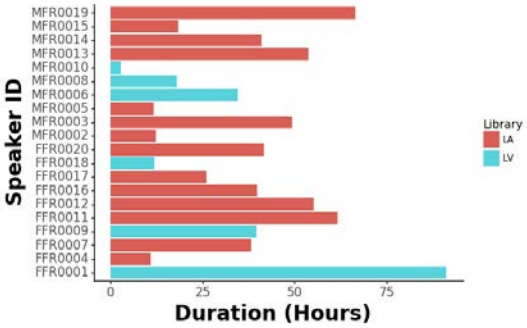
Table 3.4: The main linguistic content of MUFASA Corpus

2.3 General Overview

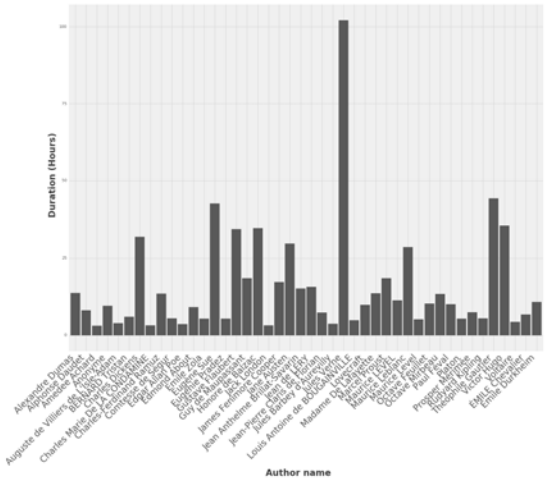
MUFASA corpus contains twenty French speakers (10 Females/10 Males). Figures 3.2a and 3.2b illustrate each speaker's duration proportion in the corpus. The speaker name is encoded as following (F/M: Female/Male, FR: French, ID:XXXX).



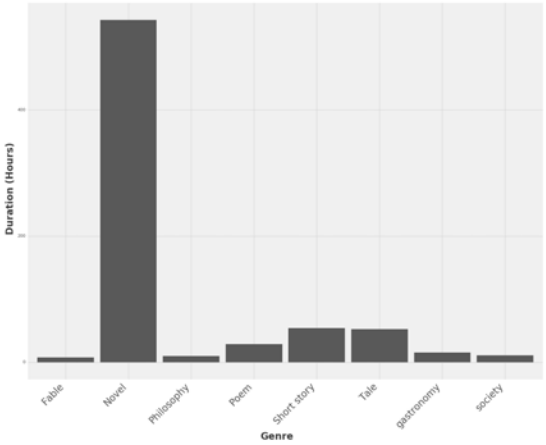
(a) MUFASA Speakers duration distribution and gender labeling



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



(a) Book author's distribution in MUFASA corpus



(b) Genre proportions in MUFASA corpus

Narrative genre such as novel, short story, and tale are the most frequent in the corpus as shown in Figure 3.2b. Some author are well represented in the corpus Figure 3.2a.

For more details about the contents of the MUFASA see the Appendix 3.

3 Gap between 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

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.

Table 3.5 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.

[8]recordings dedicated to the synthesis of speech, with favorable acoustic conditions and voice selection for this purpose unlike amateur 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 3.5: 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. PODALYDES is a male voice. The speakers FFR0001, FFR0011, FFR0020, and MFR0019 are included in the MUFASA corpus.

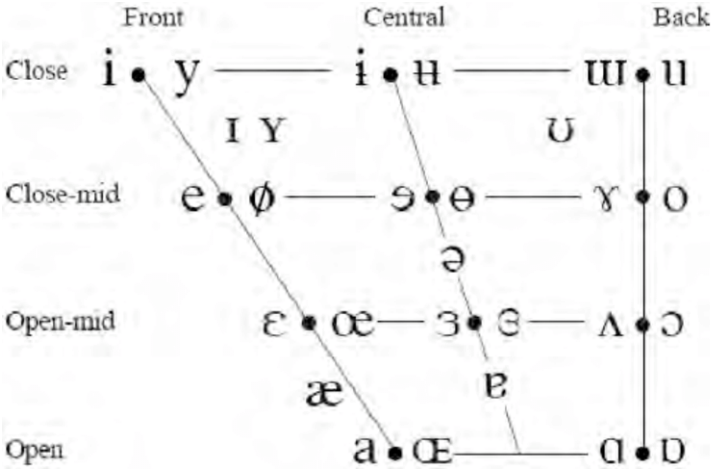
Features extraction

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.

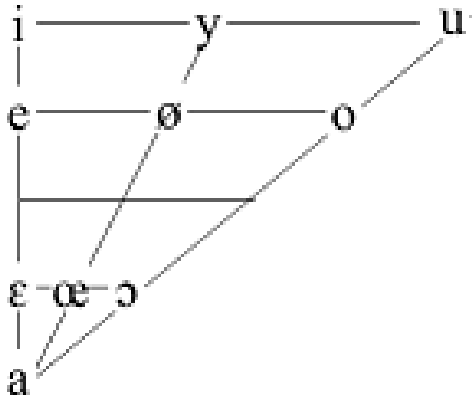
- We chose to study the pauses' duration and their distribution in an audiobook read by three different speakers to see if there is a difference between the so-called professional and amateur speakers. Because the pauses are good indicators of speech style and implicitly how the data are recorded and 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. Figure 3.2a 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. Figure 3.2b). 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/) [Audibert and Fougeron 2012] is usually studied for characterizing articulatory behaviour. The cardinal vowels represent the boundaries of the vocalic area[Lindau 1978].

- According to [Sun 2002; Sun and Xu 2002], alternate cycles[9] (alternating in amplitude or period, or both) in speech signal make the determination of pitch more difficult. [Titze 1995] 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 [Sun and Xu 2002], 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.



(a) A schematic of vowel triangle



(b) A schematic of vowel triangle (French)

[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." [Sun 2002]*

3.2 Results

Voice quality analysis

We select all the voiced frames and calculate the SHR frequency distribution (see Table 3.6 according to [Sun 2002], when SHR is in the medium range, especially (0.2, 0.4], perceived pitch becomes ambiguous. Correspondingly, in Table 3.6, 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 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. Table 3.6 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 3.6: 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 (Table 3.7). 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 3.7: 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 Figure 3.2, 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 Table 3.7 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.

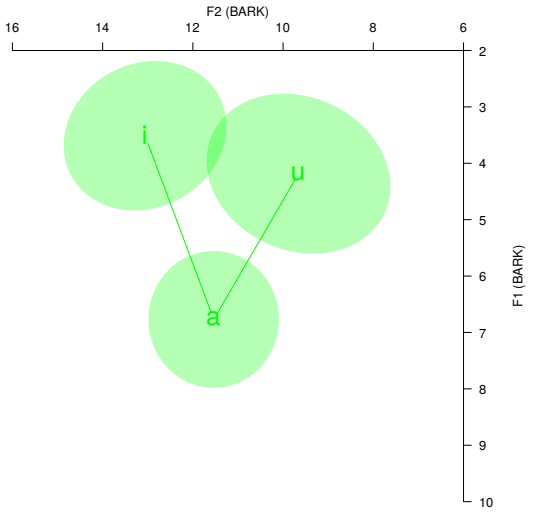
Pauses

There is no difference between the three speakers concerning the number of pauses, within utterances according to Figure 3.4a, but the Figure 3.4b 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.

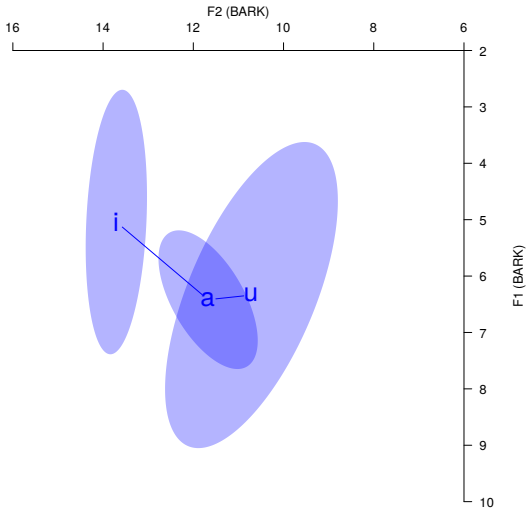
3.3 Discussion

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.

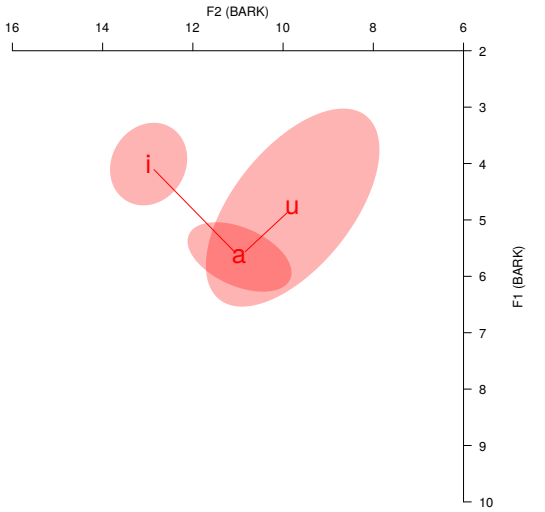
From these results, we can see that data recorded for speech synthesis has a couple of properties that distinguish them from amateur audiobooks, and professional speaker recordings that are not dedicated to speech synthesis. Despite the quality of audiobook



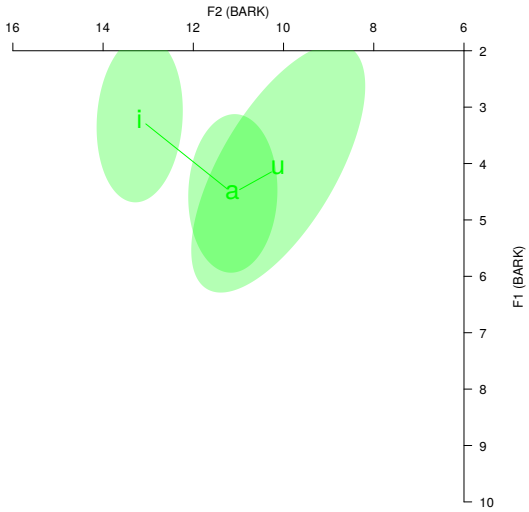
(a) SFS



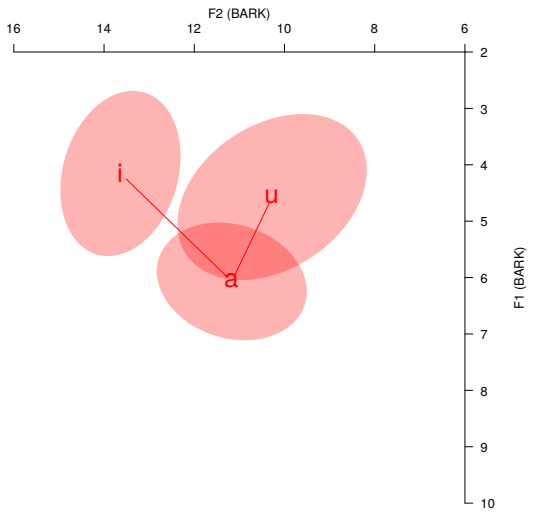
(b) FFR0001



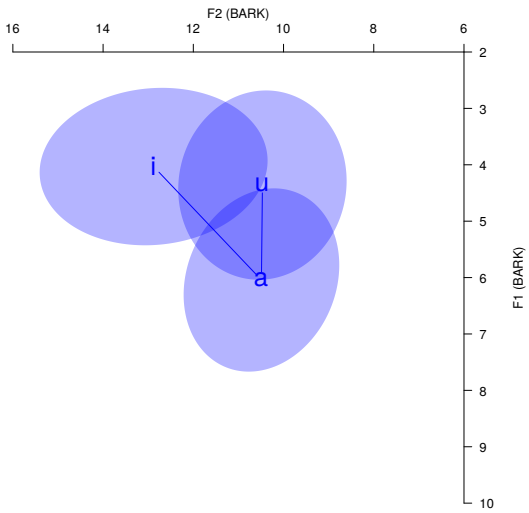
(c) MFR0019



(d) Podalydes

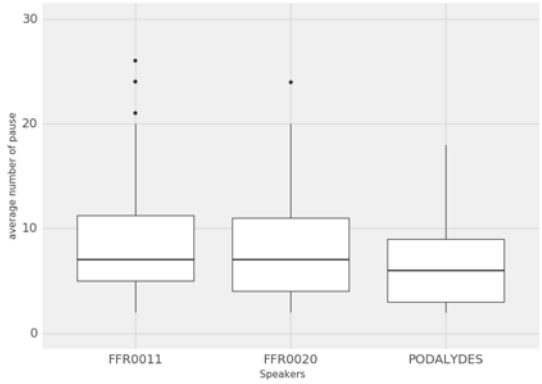


(e) FFR0011

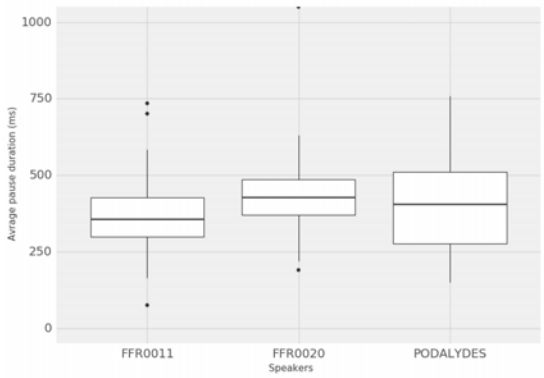


(f) FFR0020

Figure 3.2: the vowel trapezoids of the three cardinal vowels /u/, /i/, and /a/

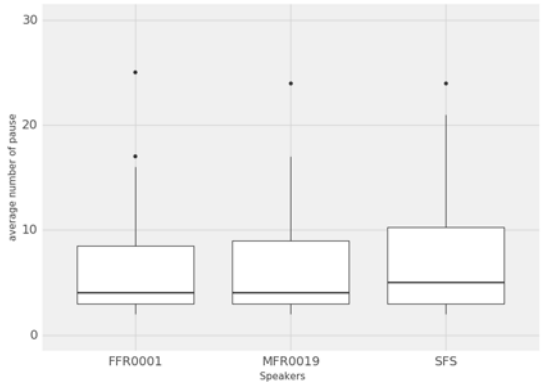


(a) Average number of pauses per utterance per speaker

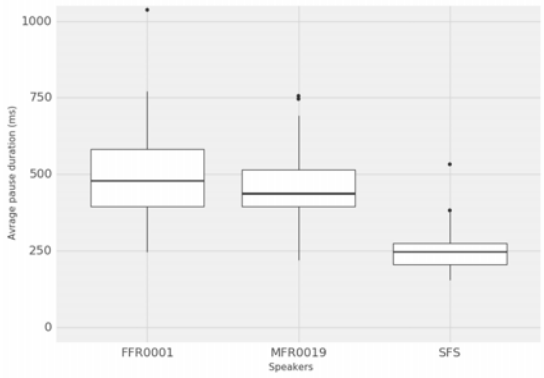


(b) Average pause duration

Figure 3.3: Pauses distribution and average duration for *"Mademoiselle Albertine est partie"*



(a) Average number of pauses per utterance per speaker



(b) Average pause duration

Figure 3.4: Pauses distribution and average duration for *"Vingt mille lieues sous les mers Chapter 3"*.

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 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 [Larnel, Gauvain, and Eskenazi 1991], ESTER [Galliano et al. 2006], RHAPSODIE [Lacheret et al. 2014], and NCCFr [Torreira, Adda-Decker, and Ernestus 2010]). 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 3.8: 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 [Larnel, Gauvain, and Eskenazi 1991]: 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 [Galliano et al. 2006]: The used data are made up of France Inter radio broadcast news recorded in 1998, covering ten speakers.

- RHAPSODIE [Lacheret et al. 2014] 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 [Avanzi et al. 2010] 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 [Torreira, Adda-Decker, and Ernestus 2010](The Nijmegen Corpus of Casual French): French speakers conversing among friends.

Table 3.8, 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 Section 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 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

[10]https://bigdataspeech.github.io/TP/tp/2018/07/10/TPPraat.html
[11]Bark is a psycho-acoustical scale closer to subjective perceptual scale
[12]http://drammock.github.io/phonR/

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 Figure 3.5 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 (Figure 3.5e), 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.)

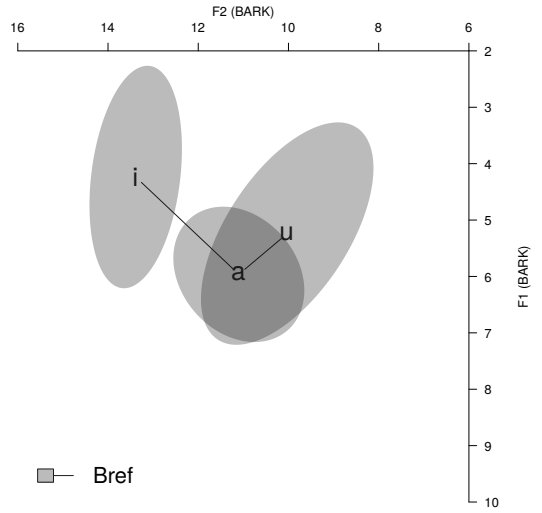
According to [Moon and Lindblom 1994; Baker and Bradlow 2009; Burdin and Clopper 2015] there is an interaction between speaking style and the duration of the vowels. The analysis of the density distribution according to the duration of the three vowels preceded by an occlusive consonant /p/,/t/,/k/ (similar context over speakers/corpus) illustrated by Figure 3.6, 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

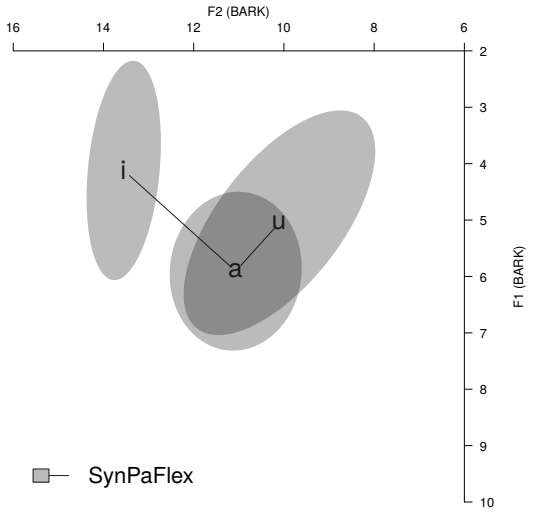
This study intended to be exploratory and attempted to provide some elements for reflection. This work raises two main reflections, which are the level of formality of the audiobook corpus in comparison with other speaking styles, and a second element, the presence of particular prosodic behavior specific 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

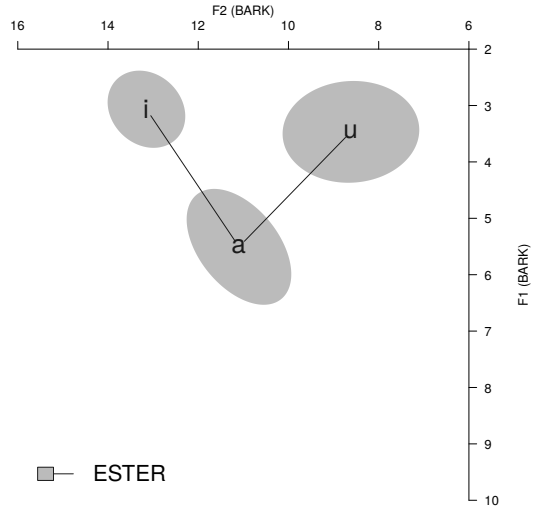
In this chapter, we presented a new audiobook corpus, the MUFASA corpus, dedicated to expressive speech synthesis but that can be used for other purposes, such as automatic speech recognition, natural language processing, second language acquisition, entity recog-



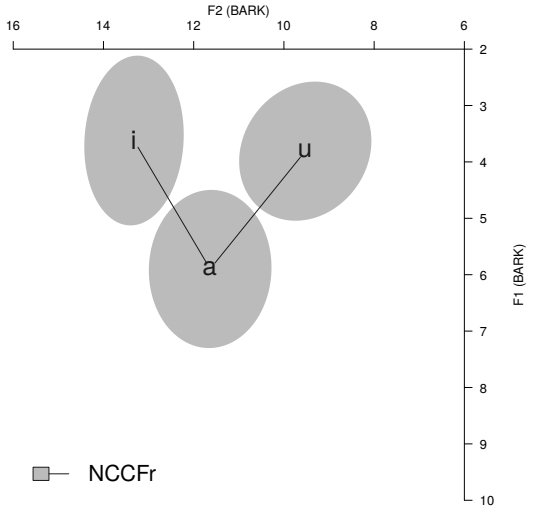
(a) BREF Corpus



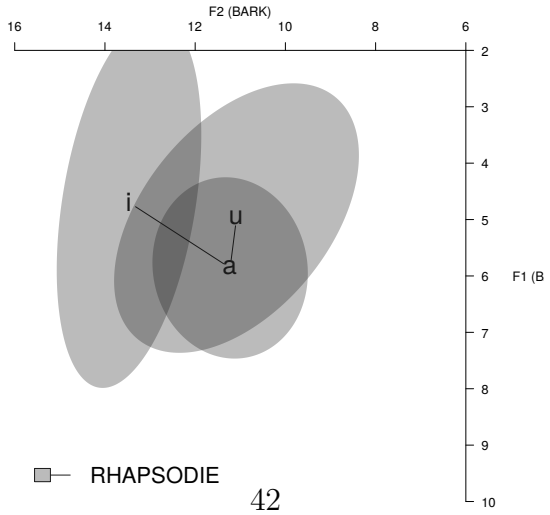
(b) MUFASA Corpus



(c) ESTER Corpus

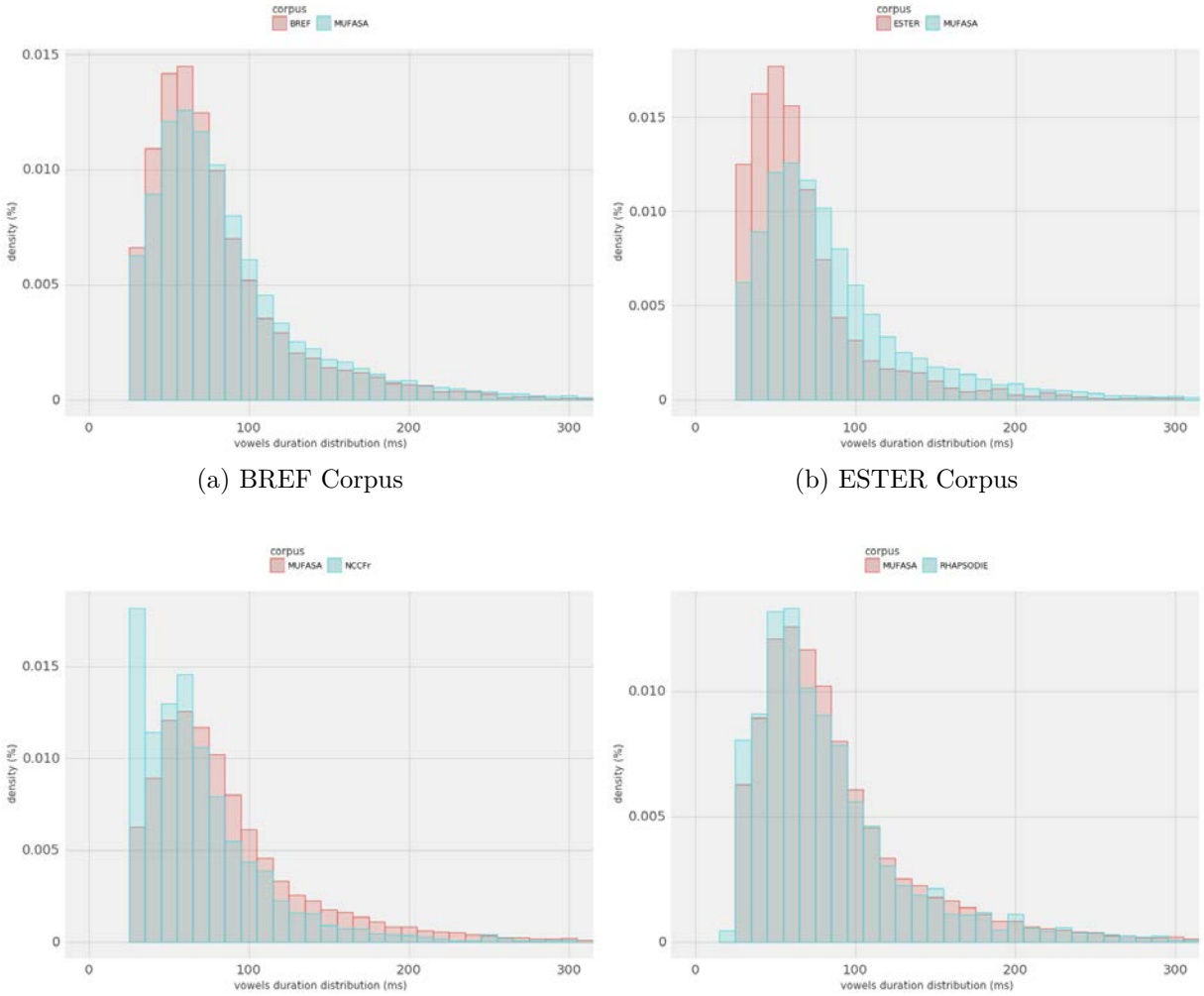


(d) NCCFr Corpus



(e) RHAPSODIE Corpus

Figure 3.5: 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

Figure 3.6: The density distribution according to the duration of the three vowels preceded by an occlusive consonant /p/,/t/,/k/

nition. Consisting of twenty speakers (ten females/ten males) 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.

Chapter 4

Annotation Protocol and Emotional Studies

The first of this work was performed in collaboration with Gaelle Vidal (IRISA) and Marie Tahon (LIUM). I would like to thank both of them: Gaëlle for collecting the initial version of SynPaFlex corpus, the annotation of the data, and providing an invaluable discussion, and Marie for her technical support in the machine learning task.

1 Introduction

In this chapter, we first report the specificity of the annotation protocol that has been realized on the SynPaFlex corpus. This annotation allows highlighting certain aspects related to intonation patterns, discourses mode properties through the fictional characters present in the narrated stories and emotion segments. At the same time, this annotation allows the exploration of the SynPaFlex corpus. Then, the focus is made on the annotation of emotions through a pattern classification experiment to evaluate the annotation process. Finally, a clustering experiment was conducted, aiming at estimating the possible correlation between text and acoustic signal properties. All these experiments were conducted using the first version of the SynPaFlex-Corpus described in [Sini et al. 2018] (Appendix 2 summarizes the proportion of manually annotated parts).

2 Speech annotation

In recent decades, many works on speech annotation protocols have been proposed [Bird and Harrington 2001]. In [Brognaux, Picart, and Drugman 2013], the authors proposed an intonation annotation protocol dedicated to living sports commentaries. The annotation is made according to two levels: local labels are assigned to all syllables and refer to accentual phenomena; and global labels allow classifying sequences of words into five

distinct sub-genres, defined in terms of valence and arousal. This approach deals with both discrete and continuous annotation strategies. Whereas [Montaño and Alías 2016] proposes an analysis methodology to annotate storytelling speech at the sentence level based on storytelling discourse modes (narrative, descriptive, and dialogue), besides introducing narrative sub-modes denoted as expressive categories. Furthermore, in [Devillers et al. 2006], the authors explore real-life emotions in French and English TV video clips aiming to a federative annotation protocol by combining continuous and discrete approaches. In this work, we propose a simple manual annotation covering four different aspects of audiobooks such as intonation patterns, characters/dialog labeling, discrete emotional labeling, and a set of labels relative to precise events.

The major challenge of expressive voice in general, and in the case of audiobooks in particular, is the lack of annotated data, as the annotation is a time-consuming step. In addition to time, the inter-annotator agreement is mostly problematic, especially when it comes to annotating emotions.

We propose unilateral annotation (One Voice - One Annotator), as we believe that this strategy will provide us with consistent and uniform annotation across all annotated data.

2.1 Protocol

Audio tracks corresponding to chapters of different books have also been annotated manually according to a set of intonation patterns, characters, emotions, and other events. This was achieved by the annotator who was listening to the audio signal using WaveSurfer[1] software. 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 4.2 shows, 38% of the whole corpus have been processed manually to provide characters annotation, and 15% - included in those 38% - to describe emotional and intonation patterns contents.

2.2 Intonation Patterns

Delattre's work is one the earliest work in French intonation modelling. In [Delattre 1966], the author defines ten fundamental intonation patterns (cf Figure 4.1) which are considered as the most frequent pitch contours in French.

[1] http://www.speech.kth.se/wavesurfer/

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 4.2: Durations and amount of annotated data according to discourse mode in the first version of the SynPaFlex-Corpus

After considering the whole speech data, eight intonation patterns were defined, then encoded and assigned by an expert annotator[2] to a large number of audio tracks corresponding to chapters of eight different books, and corresponding to 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 these labels correspond to speech showing the following types of intonation patterns: QUESTION (interrogative), NOTE, NUANCE, SUSPENSE, RESOLUTION (authority, or imperative), SINGING, and NOPIP(no particular intonation 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 2.4). Manual annotation is costly in time and redundancy is not desirable in its process. In the following analysis of the intonation manual labeling, emotion labels *surprise* will therefore be assimilated to hidden intonation labels for EXCLAMATION.

Another important point is that, when needed for a more precise description, labels were combined (e.g. Emotion+Question+Nuance illustrated in Figure 4.3).

Among the intonation parameters, the perceived pitch-curve during voice production takes an important role in assigning the labels. For instance, the NUANCE pattern, which

[2]The expert in question is Gaëlle Vidal (gaelle.vidal@irisa.fr), who collected and annotated the first version of the SynPaFlex corpus.

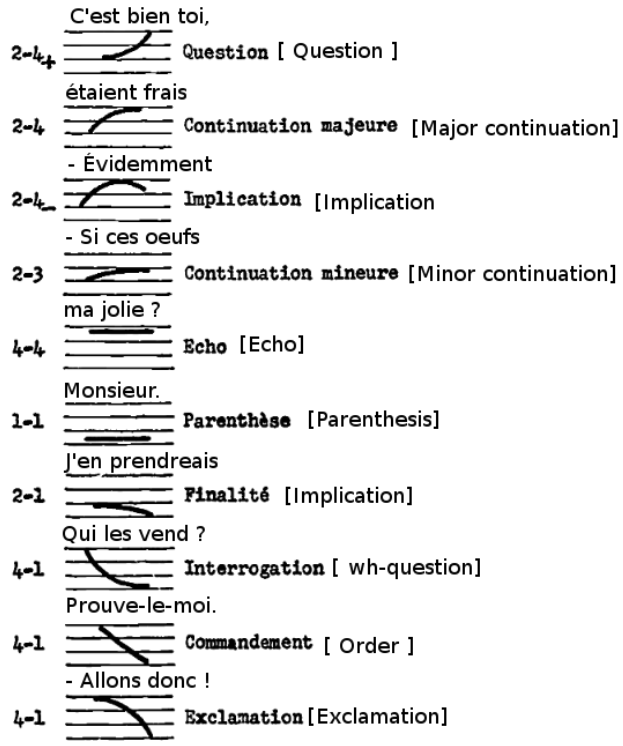


Figure 4.1: The ten fundamental intonations defined in [Delattre 1966], illustrated by a dialogue: - *Si ces oeufs étaient frais j'en prendrias. Qui les vend? C'est bien toi, ma jolie?* - *Évidemment, Monsieur. - Allons doc! Prouve-le-moi.* [- If these eggs were fresh, I'd take some. Who sells them? Is it you, my pretty? - Of course it is, sir. - Come on, then! Prove it to me.]

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 D.3).

Table 4.3 shows total duration for each manual intonation labels in the 13h25 sub-corpus.

Intonation label	Exclamation (hidden label)	NOPIP	Nuance	Resolution	Suspense	Question	Note	Singing
Duration	4h42m	4h21m	3h58m	45m	41m	38m	39m	1m
Sub-corpus %	34.8%	32.2%	29.5%	5.6%	5.1%	4.7%	4.8%	0.01%

Table 4.3: Manual annotations - Total duration of intonation patterns (including combinations) in the 13h25 sub-corpus

A non-NOPIP tag has been assigned to 68% of the speech. As shown in Table 4.3, the

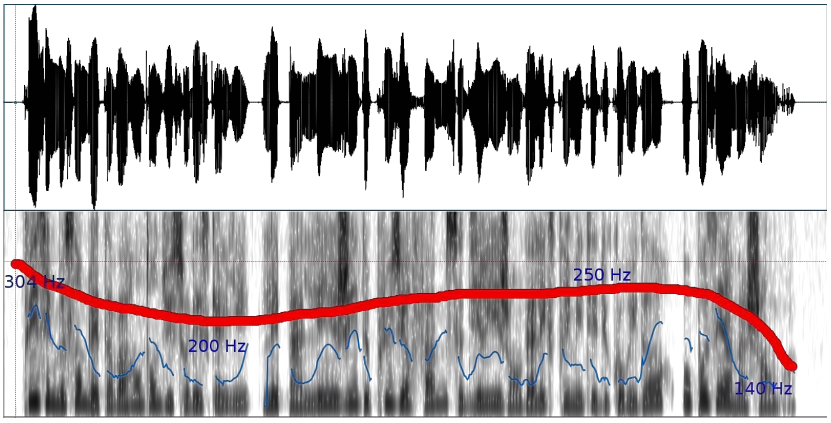


Figure 4.2: Nuance Intonation Pattern Example : *puis il me semblait avoir entendu sur l'escalier les pas légers de plusieurs femmes se dirigeant vers l'extrémité du corridor opposé à ma chambre.*

hidden Exclamation tag is very largely represented (more than 4h42), before the Idle one (4h21m). The first particular intonation pattern that comes after is Nuance (3h58m), then come all the other intonation patterns that are relatively well represented and evenly distributed (around 40m): Resolution, Suspense, Question and Note. singing was found to be exceptional and is not reported here.

More than half of the speech showing particular intonation figures is described with combined labels pointing out where prosody may be more complex (cf. Figure 4.3).

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 intonation 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 2.4.

2.3 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,

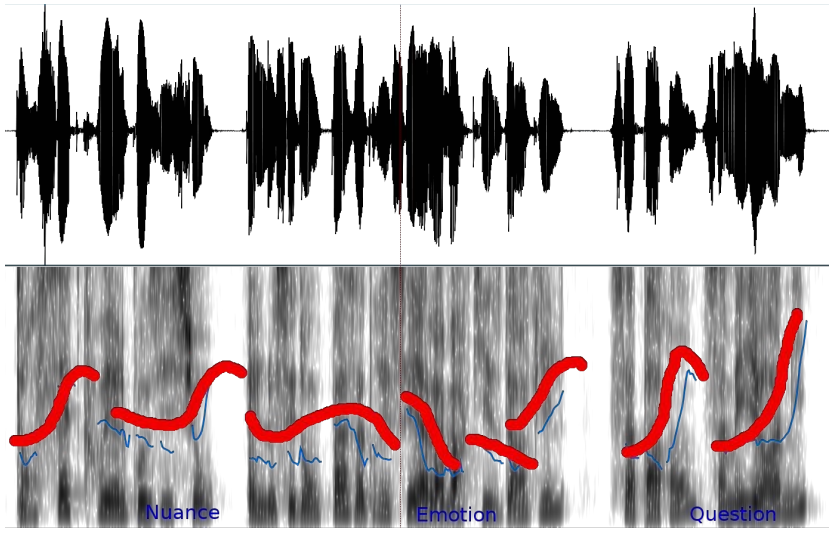


Figure 4.3: A combination of three non exclusive intonation pattern. The nuance pattern is recognized with its particular pitch contour described in Figure D.3 *Dans cette cruelle position, elle ne s'est donc pas adressée* at begining of the utterance, followed by an emotional pattern characterized by the dynamic pitch (high F0-range) *à la marquise d'Harville, sa parente,* and finishing with an explicit question pattern *sa meilleure amie ?*

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 a new character starts taking. Therefore, for such speech intervals, voice quality remains the one of speaker or corresponds to another character. This may also be an intentional choice. Reading with incessant voice changes may become painful to listen to, or artificial.

The characters labeling was annotated on more than one third of the whole corpus (33h39m) extracted from 18 different books. Dialogue tags were reported as parts of the narrator's speech.

Rough estimates indicate that one third of the speech is in direct speech style. 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 of types, 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 of them, the voice used contrasts with the speaker's natural voice.

2.4 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 selected by the Basic Emotions theory [Ekman 1999] are used: 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*: a scale from 1 to 3 was added to give a measurement of the experienced emotion intensity. For instance, one can interpret its values as follows: SLIGHTLY ANGRY (1), ANGRY (2) , and STRONGLY ANGRY (3).

- *Introversion/Extroversion*: This binary feature reflects the way the emotion is rendered through the speech (discreetly, prudently / obtrusively, ostentatiously)

The second and third features 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 sub-part of the already annotated corpus (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 4.4, and the number and average duration of labels are indicated in Table 4.6.

Emotion	Idle	Surprise	Sadness	Joy	Anger	Disgust	Fear	Irony	Threat
Duration	8h11m	4h42m	44m	32m	31m	15m	11m	10m	3m
Sub-corpus %	61.0%	34.8%	5.4 %	3.9 %	3.9%	1.9%	1.3%	1.2%	0.4%

Table 4.4: Manual annotations - Total durations of emotion categories labels (including combinations) in the 13h25 sub-corpus

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 2.2 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 in Table 4.5.

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 (suave joy)		yes	yes

Table 4.5: Examples of perceived impacts of emotion on the speech

2.5 Other Events

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

- Linguistic events such as the use of foreign languages;
- Phonetic events which are not written in the text such as 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 combining speech signal properties with linguistic information, acoustic measurements, and other descriptions. Examining how manual labels are distributed among literary genres could also be of great interest.

3 Evaluation of the emotion annotation

Among the manual annotations presented above, we decided to focus our effort on the emotion annotation in order to measure the reliably of the proposed annotation protocol and for comparing results to what was observed in previous studies.

To do this, binary emotion classification [3] experiments [Sugiyama 2015] were conducted on emotional labels of the SynPaFlex sub-corpus. Results are presented in this section.

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.

[3]The binary classification consists of assigning a given sample to one of two categories by relying on a set of attributes (features). In the case of a multiclass problem, as is the case in our case, it is possible to reduce and simplify the multiclass classification problem into a set of binary classification problems by considering two methods (i) one-versus-rest method, which makes a series of binary classifications where each model consider a class versus the others classes (ii) one-versus-one method, each binary model consider only samples from two classes at a time. In our experiments, we use the one-versus-one method because most studies have shown that this technique is more efficient than in one-versus-rest.

3.1 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 4.6. 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, which consists of all non-negative states, according to [Schuller, Steidl, and Batliner 2009]. 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. One possible solution is to analyze these samples and choose the dominant emotion. A more in-depth 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 2, 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 no 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 4.6), IDLE segments have the durations (8.76s), then comes SURPRISE segments (3.83s.) and COMBINATION labels (3.45s.).

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 4.6: Number of manually 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.

3.2 Methodology

The following experiments aim at classifying the manual annotations with binary emotional models. We know that for multi-speaker acted emotions, classification scores usually reach high performance (for example with corpora such as EMO-DB [Burkhardt et al. 2005] or JL-Corpus [James, Tian, and Watson 2018]). 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. 2009]. 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, Steidl, and Batliner 2009; Schuller et al. 2013a]. 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) + 16 Δ) $\times$ 12 functionals (Table 4.7 represent the description of the acoustic features) are extracted on emotional segments with OpenSmile toolkit and Interspeech 2009 configuration [Schuller, Steidl, and Batliner 2009].

Low-Level-Descriptors (LLD)	Functionals
zero-crossing-rate (ZCR) +Δ	mean
root mean square (RMS) Energy +Δ	standard deviation
Fundamental Frequency (F0)+Δ	kurtosis, skewness
Harmonics-to-noise ration (HNR) +Δ	min. and max. value, relative position, range
Mel-Frequency Cepstral Coefficients (MFCC) 1-12 + Δ	linear regression: offset, slope, mean square error (MSE)

Table 4.7: Feature set of the INTERSPEECH 2009 Emotion Challenge 384 features, (16 LLD + 16 Δ)*12 functionals

To avoid over fitting the data[4], 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 + mean) $\times$ 12 functionals without Δ

An informal analysis of the manual emotion segments and other emotional French speech corpora [Chateau, Maffiolo, and Blouin 2004; Scherer, Johnstone, and Klasmeyer 2003; Johnstone and Scherer 1999; Chateau, Maffiolo, and Blouin 2004; Abrilian et al. 2005; Beller and Marty 2006; Devillers et al. 2006] , we have observed that in most of cases one second is enough to recognize the emotional sample identity. For that reason and to have homogeneous segment durations, we decided to chunk manual segments every 1 s. This

[4]Over-fitting in statistic and machine learning is a model that is too close to the data is made from (training data) even the noisy ones, but can not be generalized to new coming data, for instance, test data. This model achieves very high performance in the training phase and mediocre in the phase test. Most of the time, this phenomenon is due to the size of the training data too small in comparison to the number of parameters of the model.

operation helps in increasing the amount of data available for the experiment, as reported in Table 4.6.

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.

3.3 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 4.8. 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.

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 4.8) for ANGER, SADNESS, FEAR and DISGUST. By analyzing the Table 4.8, we can distinguish two emotion groups:

- 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%)

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 4.8: Unweighted Average Recall (UAR) results for binary emotion classification using the three feature subsets. In bold, UAR > 60%, which we considered as a reasonable classification rate.

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.

4 Emotion Lexicon Study of Audiobooks

Given the difficulty encountered in classifying and recognizing emotional patterns at the acoustic analysis level, we suggest exploring the expressive properties conveyed by lexical and textual structures such as sentences in audiobooks.

In audiobooks, the written text holds a substantial portion of the expressivity. We

assume that the lexico-semantic properties and the syntax constraints have an essential impact on speech production. This assumption seems to be trivial because, beyond the speaker style and acoustic speech parameters, the lexicon-semantic and syntactic properties, which are mostly preponderant. This information is also independent of the speaker.

Therefore, the main objective of this exploratory work is to analyze if there are correlations between lexicon-semantic properties of the read text and its acoustic parameters.

To achieve this objective, we propose to rely on natural language processing and sentiment analysis techniques [Mohammad 2013; Medhat, Hassan, and Korashy 2014; Mohammad 2011; Chaffar and Inkpen 2011] to characterize and study the lexico-semantic and syntactic properties of texts corresponding to the dataset presented previously (cf. Table 4.2).

4.1 Proposed Method

In this work, we suggest to use unsupervised learning methods in order to avoid the manual annotation process of the written text. This study is articulated in three stages, as illustrated in Figure 4.4. In the pre-processing stage, We present the process that we used to make the raw text usable afterward. Then, in the clustering stage, we combine a dimension reduction technique and clustering technique in order to find the best number of clusters that meet a couple of predefined criteria, and the final stage is dedicated to acoustic speech features visualization and interpretation that correspond to the clustered text.

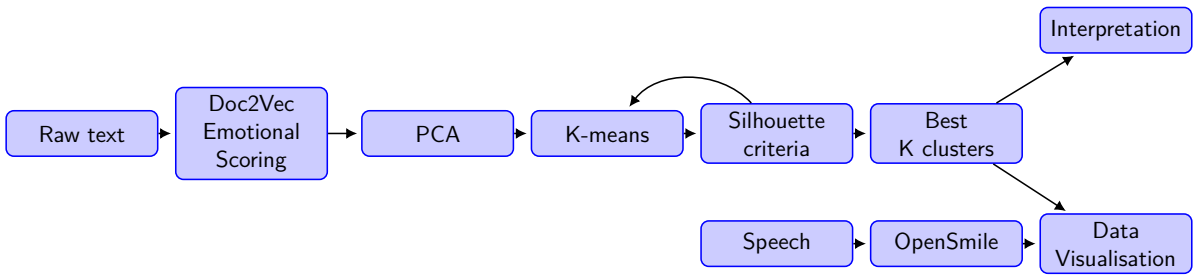


Figure 4.4: Scheme of proposed framework

4.2 Pre-processing stage

Data cleaning and representation

Each sentence is segmented into substrings (or Tokens which do not necessarily correspond to words but rather to a sequence of characters.) using whitespaces as a delimiter. This process called tokenization is provided by Natural Language Toolkit (NLTK)[Loper and Bird 2002]. This toolkit also furnishes text segmentation into sentences based on appropriate rules of targeted language, for instance, French. These rules are most of the time based on punctuation. Each sentence is in the form of a series of tokens (or terms). This form cannot be computed, so it must be converted to computational form using the *doc2vec* model. Each sentence will be represented as a set of numerical features that will be later selected and extracted.

Sentence representation

The numeric representation of the text documents, paragraph or sentence is a challenging task in Natural Language Processing (NLP). For this task, we use document embedding, which are high-dimensional continuous vector representation of document, in our case we consider sentence as document. In this vector space, sentence that have similar distributions are closer together than sentence with different distributions, given some distance measure. This type of setup has been shown to capture relevant syntactic and semantic properties of sentences, and they have been successfully applied to various tasks [Collobert and Weston 2008; Socher et al. 2011; Mikolov, Le, and Sutskever 2013].

For converting sentence to numerical vector, among the existing methods for sentence numerical representation, we can mention sparse word features, *word2vec* word vector averaging, and *doc2vec*[Mikolov et al. 2013; Le and Mikolov 2014]. We choose to use *doc2vec*, according to [Lau and Baldwin 2016]. We use the software [Rehurek and Sojka 2010] because it is easy to implement.

To build the *doc2vec* model required for generating the embedding sentence vectors, we use the entire text transcription of the SynPaFlex-Corpus. This data has been pre-processed and cleaned, and we have kept the first 5 million words. We trained the model on this dataset using an embedding size of 300. The systems use the publicly available *doc2vec*[5] implementation of the skip-gram model with negative sampling, and they were trained for 15 epochs with a window of 5 words.

[5]https://radimrehurek.com/gensim/models/doc2vec.html

Emotional Scoring Vector

For now, each sentence is represented by numerical vector of size of 300. In order to estimate the impact of emotion text labeling we propose to aggregate doc2vec (doc2vec) sentence vectors with a emotional score vector. To label each sentence in terms of emotions, we propose to use the FEEL[Abdaoui et al. 2017; Nzali et al. 2017] corpus to determine the word score (emotional score as well as polarity score) and then infer sentence vector emotional scoring. To do so we suggest to calculate the emotional scores of each sentence using the equation Equation (4.1).

$$E_{emo} = \frac{\sum_{i=1}^{n} emo_score_i}{n} \tag{4.1}$$

Where n is the number of words in sentence, and $emo \in$ { joy, fear,sadness, anger, neutral}, by aggregating the scores E_{emo}, we have an emotional score vector that represent a sentence.

4.3 Features Selection

Note that during the experiment, we investigate two configurations. For the first configuration, the feature vectors of given sentence $s^{(i)}$ are composed only with the embedded feature vector made using *doc2vec*. In the second configuration, we aggregate the embedded feature vector and the emotional scoring vector together.

The k-means clustering method is quit sensible to the dimension of the input data, for that reason, we choose to use Principal Component Analysis (PCA) as dimension reduction method. This configuration PCA+K-means [Ding and Li 2007] is widely used for sentiment and text clustering.

4.4 Clustering Stage

For clustering the embedding sentences vectors generated with *doc2vec* model, we investigate K-means [Hartigan and Wong 1979; Kanungo et al. 2002] clustering technique because it is commonly used for text clustering[Jing et al. 2005; Spangler 2008], data mining [Riaz et al. 2019] and sentiment analysis task[Orkphol and Yang 2019]. The process of the K-means clustering algorithm is simple. It consists of first to randomly generating K centroids in the data points space, where K corresponds to the number of clusters.

These centroids represent the initial positions for each cluster, and their number has to be specified. The position of these K centroids are optimized (adjusted) iteratively following these steps:

1. Calculate the sum of squared distance between data points and centroids.
2. Reassign each data point to the cluster closer than other clusters (centroid).
3. Update the position of centroids of clusters by taking the average of all data points of that cluster.

The optimization process of the centroids ends when:

- The centroids have reached their stable position; no changes in their position are possible anymore.
- Alternatively, the number of iterations is achieved.

To find the optimal K number clusters, we used silhouette analysis.

Select the number of clusters by using silhouette analysis

Silhouette analysis is a graphical tool for interpretation and validation of cluster analysis by measuring how close each data point is in a cluster compared to other data points in its neighboring clusters.

To select the number of clusters we have applied the same constraints as the one described in [Li and Liu 2014], which are:

- First, the average silhouette coefficients should be close to one as much as possible.
- Second, the plot of each cluster must be above the average of the silhouette coefficients as much as possible.
- Third, the thicknesses of all the clusters must be uniform as much as possible.
- After K-means has been run many times with different numbers of clusters (K), the best number of clusters will be selected based on previous aspects. The result of K-means with the optimal number of clusters will be interpreted and acoustically analysed in the next subsection.

4.5 Acoustic Analysis

To estimate the correlation of text with corresponding acoustic speech features, we propose analyzing and visualizing the acoustic data of the corresponding text. For this task, we follow three steps. First, we extract the acoustic features of each utterance extracted on emotional segments with OpenSmile toolkit and Interspeech 2009 configuration[95], as described in Section 3. Then, we apply PCA to reduce the dimension of features, and finally, we use T-distributed Stochastic Neighbor Embedding (t-SNE)[Maaten and Hinton 2008] for data visualization.

4.6 Experiments and Results

There are two experiments being conducted. The first experiment is to verify whether clustering k-means method is affected by the input features. The second experiment is to find the optimal number of clusters (K) using silhouette analysis. The third experiment is to visualize the acoustic data corresponding to the resulting clusters and interpret the result.

Data setup

For this work we choose to use the same audiobooks that we used in Section 3.1. Because we assume that this set of data has potential emotional contents. We split each text into sentences using NLTK[Loper and Bird 2002] toolkit, at the end of the process we obtained 13384 sentences.

Clustering results and Analysis

Table 4.9 shows two candidates (one for each configuration) produce good results. K = 18 in *doc2vec* configuration and K= 7 for the second feature vectors configuration *doc2vec* + emotional vector scoring were selected because they achieve the best trade-off in terms of Silhouette average coefficient and balanced number of samples per group.

According to this result, it seems that the emotional scoring vector reduces the number of groups needed to represent all the data. As these results are too preliminary, we cannot explain or analyze the results obtained. An in-depth analysis of the content of the groups will allow us to bring more precision and a solid explanation of the present findings.

Input Configuration	Best K clusters	Silhouette Avg	Avg samples per cluster
Doc2Vec	18	0.301	729 (±75)
Doc2Vec + Emotional Score	7	0.36	2011(±289)

Table 4.9: The best K-clusters according to the silhouette average criteria and average samples per cluster

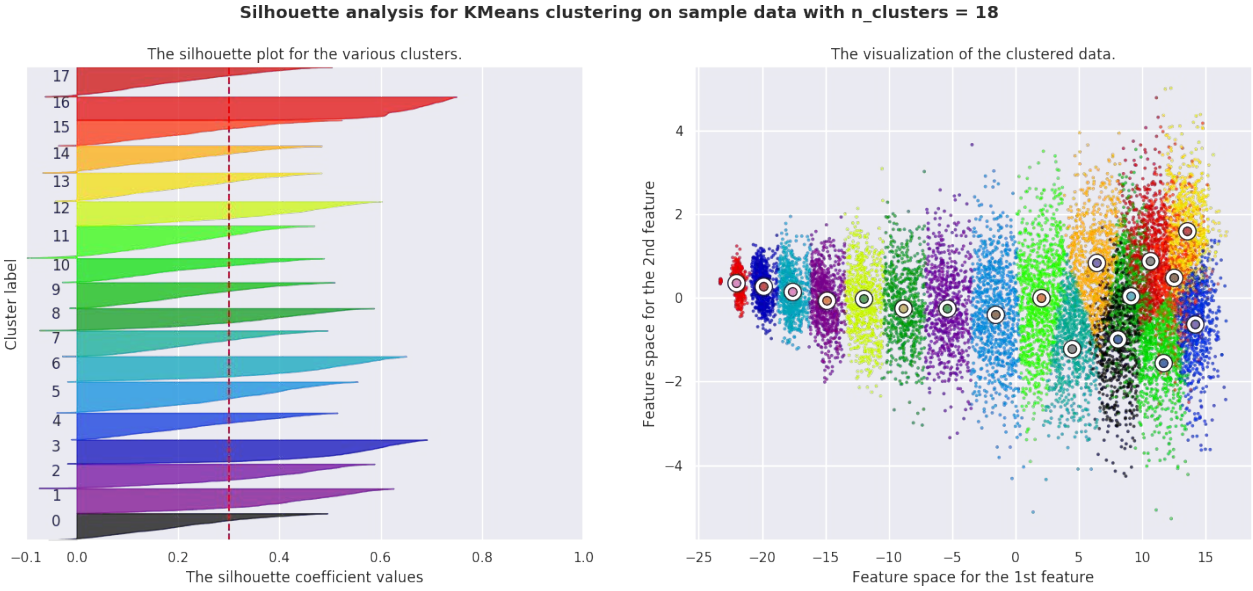


Figure 4.5: The data points scatter in k = 18 groups - *doc2vec* features. The right-hand side shows the result of K-means, i.e., the data points of each cluster. The left-hand side shows the silhouette coefficient of each cluster. The thickness of each cluster plot depends on the number of data points lying in the cluster. The red bar is the average of the silhouette coefficient of entire clusters.

Visual analysis of acoustic data

Figure 4.7 shows the projection of the acoustic data corresponding to the seven groups from the K-means analysis.

The contrast between the groups is low. It can also be observed that there is a correlation with the results obtained at the textual level, which is quite encouraging but not enough and need further experiments to understand data behaviors both in terms of acoustic and linguistic.

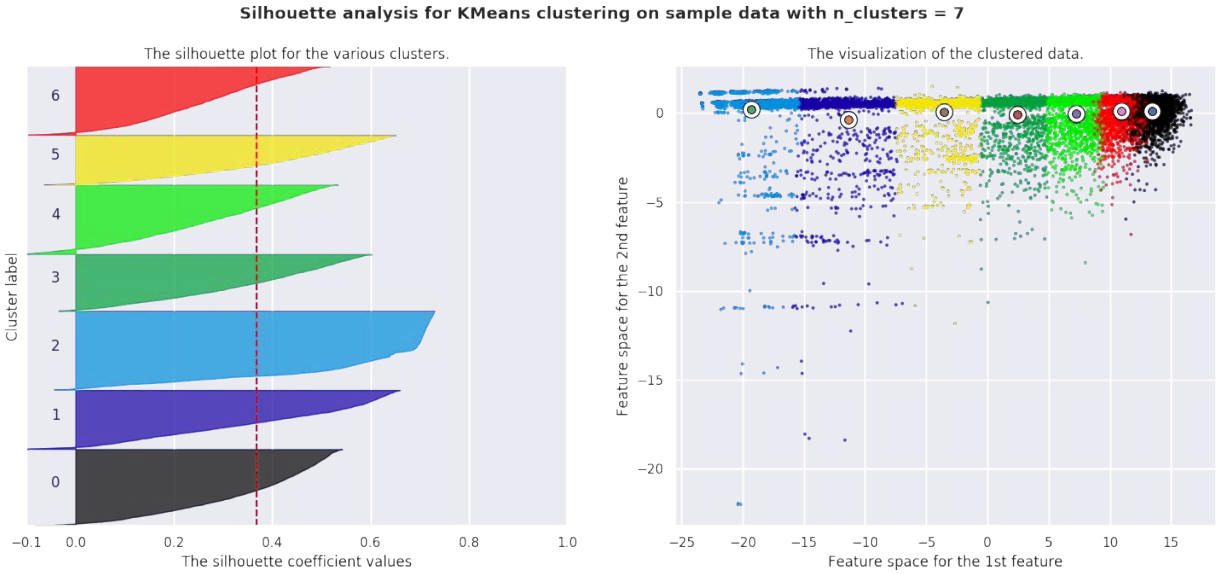


Figure 4.6: The data points scatter in k = 7 groups - *doc2vecs* + emotional vector features. The right-hand side shows the result of K-means, i.e., the data points of each cluster. The left-hand side shows the silhouette coefficient of each cluster. The thickness of each cluster plot depends on the number of data points lying in the cluster. The red bar is the average of the silhouette coefficient of entire clusters.

4.7 Discussion and issues

For clustering sentences, we used PCA + K-means. There are many issues related to this method. The coverage of PCA components does not exceed 50%, which means that we loose a lot of information from the initial features representation. The second issue is with *K-means* algorithm initialisation, where the initial clusters position are randomly assigned. This kind of initialization method is problematic regarding to the distribution of the data points. As future work, we propose to investigate Adversarial autoencoders[Makhzani et al. 2015] as dimensional reduction method instead of PCA and Artificial Bee Colony (ABC) [Krishnamoorthi and Natarajan 2013; Armano and Farmani 2014] algorithm for initializing of the K-means.

5 Conclusion

In this chapter, we have tried to make a quantitative analysis of the emotions contained in audiobooks. We have proposed a rather rich emotional annotation protocol as well as a baseline for future work.

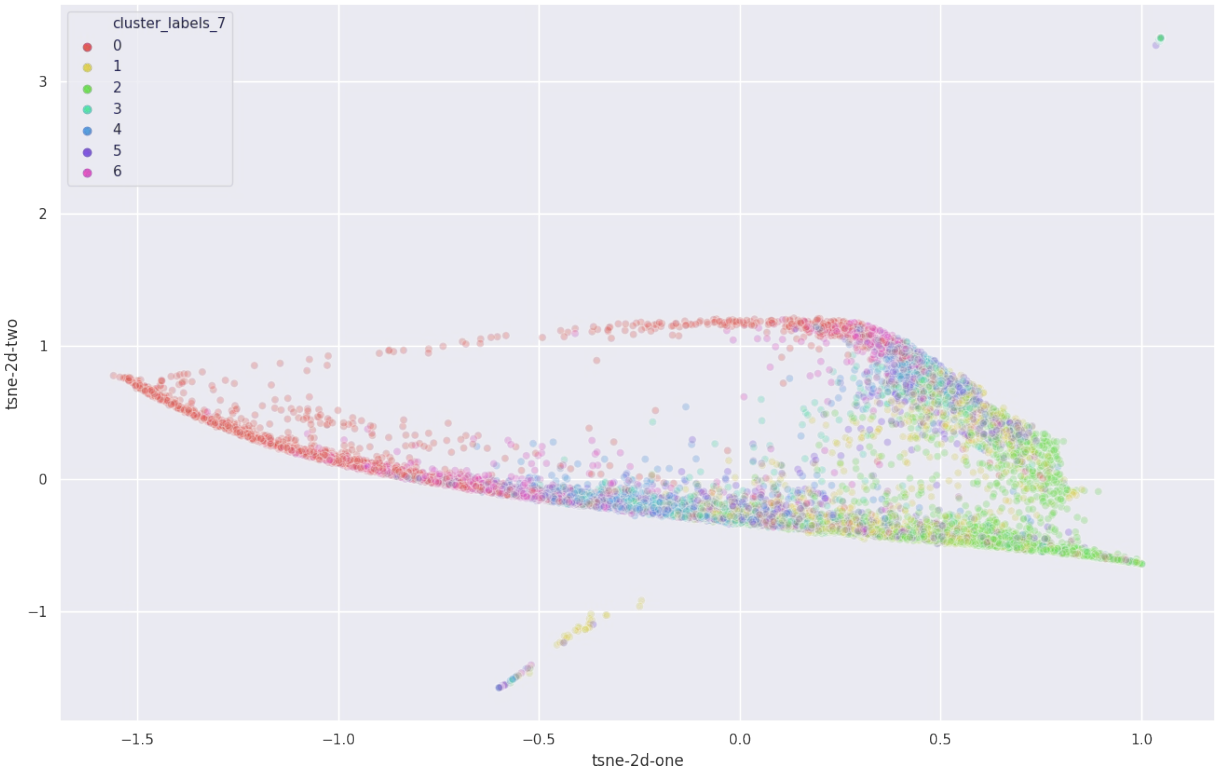


Figure 4.7: PCA variation coverage of 73 % with 50 components, t-SNE with perplexity of 45 and with iteration of 250

Indeed, the experiments carried out do not take into account the variability between speakers; we focused on a single speaker. In this work, we consider emotions as a category, thus neglecting the possibility of continuous representation with parameters such as Valence, Arousal and Dominance.

In our analysis of the text, we also omitted to take into account the phonetic and phonological transcriptions of the text, which would undoubtedly have brought possible explanations to the results. Through this preliminary work, we found that the analysis of the emotions and expressiveness carried by audiobooks is complex.

Chapter 5

Automatic Annotation of Discourses

This chapter is an extended version of the work described in "Automatic annotation of discourse types in audiobooks"[Sini, Delais-Roussarie, and Lolive 2018] presented at Traitement Automatique de Langues Naturelles (TALN) 2018.

1 Introduction

In [Durrer 1999; Dominique 1998; Perret 1994] the structure of texts such as novels, short stories, and tales, are modeled with four types of discourse: the Direct Discourse (DD), the Indirect Discourse (ID), the free indirect discourse (FID), and the narrative discourse(ND). In [Durrer 1999], the author has illustrated this model in the case of french novels.

An example, which shows the combination of the four discourse types:

Des gens qui sortaient du spectacle passèrent sur le trottoir, tout fredonnant ou braillant à plein gosier : Ô bel ange, ma Lucie ![**DD**] *Alors Léon, pour faire le dilettante, se mit à parler musique.*[**ND**] *Il avait vu Tamburini, Rubini, Persiani, Grisi ; et à côté d'eux, Lagardy, malgré ses grands éclats, ne valait rien.*[**FID**]

– *Pourtant, interrompit Charles qui mordait à petits coups son sorbet au rhum, on prétend qu'au dernier acte il est admirable tout à fait ; je regrette d'être parti avant la fin, car ça commençait à m'amuser.*[**DD**]

– *Au reste, reprit le clerc, il donnera bientôt une autre représentation.*[**DD**]

Mais Charles répondit qu'ils s'en allaient dès le lendemain.[**ID**]

– *À moins, ajouta-t-il en se tournant vers sa femme, que tu ne veuilles rester seule, mon petit chat ?*[**DD**]

Et, changeant de manœuvre devant cette occasion inattendue qui s'offrait à son espoir, le jeune homme entama l'éloge de Lagardy dans le morceau final. C'était quelque chose de superbe, de sublime ![**FID**]

[*People coming out of the theatre passed along the pavement, humming or shouting at the top of their voices, "O bel ange, ma Lucie!"*[**DD**] *Then Leon, playing the dilettante, began to talk music.* [**ND**]*He had seen Tambourini, Rubini, Persiani, Grisi, and, compared with them, Lagardy, despite his grand outbursts, was nowhere.* [**FID**]
Oh beautiful angel, my Lucie! "Yet," interrupted Charles, who was slowly sipping his rum-sherbet, "they say that he is quite admirable in the last act. I regret leaving before the end, because it was beginning to amuse me." [**DD**]
"Why," said the clerk, "he will soon give another performance."[**DD**]
But Charles replied that they were going back next day.[**ID**]
"Unless," he added, turning to his wife, "you would like to stay alone, kitten?" [**DD**]
And changing his tactics at this unexpected opportunity that presented itself to his hopes, the young man sang the praises of Lagardy in the last number. It was really superb, sublime.[**FID**]] (*Madame Bovary*, chap. 15, Part 2)

In our approach, we have decided to consider the indirect discourse, the free indirect discourse, and narrative discourse as a single entity that we call indirect discourse. This decision is motivated by the fact that these three discourse types represent words that the narrator says, even if they might express different points of view. By contrast, direct discourse corresponds to words that are said by a character involved in the story.

In [Laferrière 2018], the author pointed out the importance of the so used Incidental Clauses with reporting verbs in narrative genres texts. This particular structure is often used within DD (dialogues or monologues) to describe the characters state or to initialize the DD.

In our annotation, we consider three main categories for modeling fictional texts (novels, short stories, and tales) :

- DD: containing the dialogues and monologues as specified, (see example (2)
- ID: covering indirect discourse, free indirect discourse, and narrative discourse. Narrative discourse in (1),
- Mixed discourse involving DD, ID, and Incidental Clauses with reporting verbs
  - pairwise : in example (3), illustrate the combination of direct discourse with indirect discourse; (4.a) and (5) contains DD and IC.

– all three as in (4.b)

The analysis and annotation of the discourse in literary works are of interest to characterize the expressiveness carried by audiobooks.

To synthesize audiobooks in satisfactory and expressive manners, it is essential to be able to indicate any modification in the enunciative perspectives(characters changes in dialogue sequence) by prosodic markings comparable to those observed in real speech [Doukhan et al. 2011; Montaño, Alías, and Ferrer 2013].

For achieving a better expressivity, it is necessary to classify (paragraph means sequence separated by line breaks in the text) according to their type of discourse, and from there to have a precise idea of who speaks. All these reasons lead to classifying the paragraphs according to whether they correspond to Indirect Discourse (ID) (1), Direct Discourse (DD) (2), or mixed passages. In some paragraphs, reported discourse is inserted in the middle of narrative passages (3) or Incidental Clauses with reporting verbs (IC), which can be short (4a) or relatively long (4b). In these mixed cases, the task is to delimit precisely the types of speech present.

(1) On commença la récitation des leçons. Il les écouta de toutes ses oreilles, attentif comme au sermon, n'osant même croiser les cuisses, ni s'appuyer sur le coude, et, à deux heures, quand la cloche sonna, le maître d'études fut obligé de l'avertir, pour qu'il se mit avec nous dans les rangs.[We began reciting our lessons. He listened attentively, concentrating as though listening to a sermon, not daring even to cross his legs or lean on his elbow, and, at two o'clock, when the bell rang, the master had to tell him to line up with us all.] (*Madame Bovary*, chap. 1)

(2) – *Soit, demain à une heure.*
– *A une heure.*
– *Dans la plaine Saint-Denis?*
– *Dans la plaine Saint-Denis.*
– *Entre Saint-Ouen et le chemin de la Révolte, au bout de la route?*
– *C'est dit.*
[– Be it so; tomorrow at one.
– At one o'clock.
– In the plain of St. Denis?

– In the plain of St. Denis.
– Between St. Ouen and the road of La Revolte, at the end of the road?
– Agreed.] (*Les Mystères de Paris*, chap. 7, Tome 1)

(3) D'autre part, la mort de sa femme ne l'avait pas mal servi dans son métier, car on avait répété durant un mois : **« Ce pauvre jeune homme ! quel malheur ! »** [In any case, the death of his wife had done him no harm professionally; for a whole month people kept saying: **«That poor young man! What a terrible thing!»**] (*Madame Bovary*, chap. 3)

(4) a. – *Levez-vous,* ***reprit le professeur****, et dites-moi votre nom.*
[– Stand up, **repeated the master**, and tell me your name.] (*Madame Bovary*, chap. 1)

b. – *Débarrassez-vous donc de votre casque,* ***dit le professeur, qui était un homme d'esprit****.*
Il y eut un rire éclatant des écoliers qui décontenança le pauvre garçon, ...
[– I suggest you disencumber yourself of your helmet, **said the master, a man of wit.**
A roar of laughter came from the class and disconcerted the poor lad,...] (*Madame Bovary*, chap. 1)

(5) Puis, l'ayant considéré quelques minutes d'un œil amoureux et tout humide, **elle dit vivement:** (*Madame Bovary*, chap. 18)

While the detection of narrative passages (1), dialogues (2) and discourses related to the middle of narrative passages (3) in the mixed paragraphs may seem rather trivial, due in particular to typographical indications, the annotation of incidental citation is more complex, as evidenced by a simple comparison between cases (4a) and (4b). The presence of a comma after the citation is not enough.

The main objective of this work is to design an annotator for parsing french audiobook text in terms of discourse type.

This chapter is organized in four sections. The Section 2 present the experimental data set and the annotation protocol. Then the Section 3, present the rule based annotator, followed by Section 4 which contains the machine learning approach, and finally a conclusion in Section 5.

2 Corpus and material

For this study, we worked on a subset from the SynPaFlex-Corpus presented in Section 1. This dataset corresponds to chapters of two French novels, *Les Mystères de Paris* by *Eugène Sue* and *Madame Bovary* by *Gustave Flaubert*. An expert chose these excerpts because they contained many changes in discursive and enunciative perspectives, and, at the same time allowed arriving at a relatively coherent set in terms of sequences with direct and indirect discourses, as shown in Table 5.1. For the whole corpus and the sub-part selected for this study, we have the orthographic transcription, phonetization, and alignment to the sound signal done automatically using JTrans[Cerisara, Mella, and Fohr 2009]. Other linguistic annotations of a phonological nature (syllable division, etc.) and morpho-syntactic (grammatical categorization of words, analysis, and indication of grammatical functions) are also available through the use of automatic annotation procedures [Candito et al. 2010; Candito et al. 2009]. The entire annotation process was conducted using ROOTS [Chevelu, Lecorvé, and Lolive 2014], which allows all annotations to be maintained consistently.

Paragraph typology	Direct Discourse	Indirect Discourse	Mixed Discourse	Total
Paragraphs	1 202	844	771	2817
Sentences	4 109	2 160	2 920	9189
Words	36 722	36 622	26 001	99345
Orthographically distinct words	5399	6913	4 345	
Phonetically distinct words	5235	6764	4 248	
Syllables	49 313	55 021	35 827	140161
Different Syllables	2 692	2678	2 279	
Phonemes	111 915	124 886	80 827	304657
Distinct phonemes set	33	33	33	

Table 5.1: Composition of the corpus according to types of discourse, selected from the corpus SynPaFlex describe in Section 1

3 Rule-based Approach

Annotation procedure

The automatic annotation of discursive and changes for a given text (a chapter in our case) is done in two phases, illustrated in the following subsections:

- Classification of paragraphs according to the types of speech, in order to put aside all paragraphs with IC;

- Detection and delimitation of quotation marks (*he says*, and so on) and primers (he said: "...", and so on).

Classification of paragraphs according to types of speech:

From the text, the program classifies paragraphs, defined on a typographic basis (line break), into three distinct groups (see Figure 5.1, Phase 1). It is based on lexico-syntactic criteria (presence of speech verb, and other patterns), as well as on punctuation and typographic signs (presence of quotation marks or hyphens).

- the group DD gathers all the paragraphs which contain passages exclusively in direct speech as in the example (2);

- the group ID: contains paragraphs with only narration passages or descriptions as in example (1);

- The Mixed Discourse group is composed of paragraphs that may contain both DD and indirect speech or narration. Will be present in this group both the narrative passages in which are inserted reported speeches as in example (3) and passages to the direct speech including Incidental Clauses with reporting verbs (IC) as in example (4)

In a second phase (see Figure 5.2, Phase 2), the paragraphs of the Mixed group are analyzed to determine the boundaries of discourse. This task is performed using expert rules. This step will make it possible to identify the Direct Discourse (DD), often in quotation marks and preceded by two points (3), Incidental Clauses with reporting verbs (IC) like (4), and the sequences introducing a dialogue like (5). Among these elements, IC are essential because they make it possible to delimit changes of characters and to provide indications on the characters present and their attitudes.

Detection and annotation of Incidental Clauses with reporting verbs (IC)

At the end of the first classification phase, the Mixed paragraphs are analyzed in detail to determine the boundaries of the different types of speech.

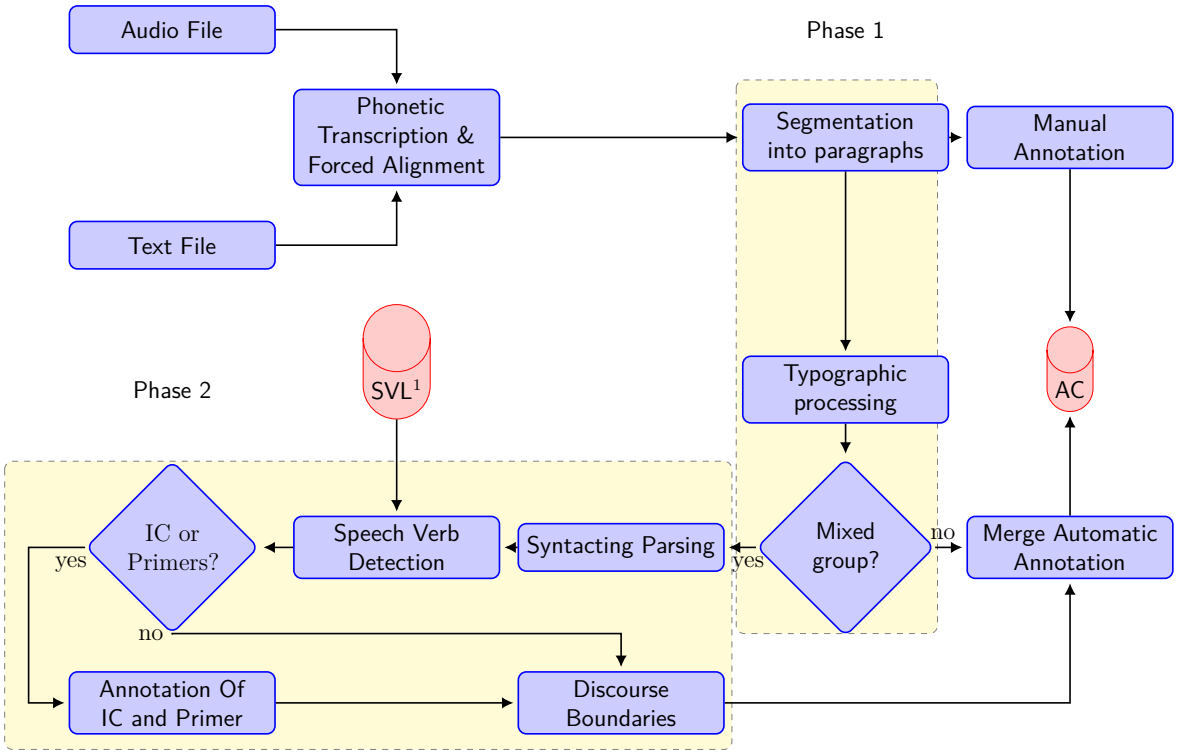


Figure 5.1: This figure illustrates the workflow guiding the rule-based approach. After the phonetization and forced alignment of the chapter text with the corresponding audio file, the data are segmented into paragraphs/ pseudo-paragraphs and stored relying on roots toolkit. The segments follow two annotations process: (i) the manual annotation made by an expert (ii) the automatic annotation which has two phases, the first phase consists of labeling the segments according to typographic criteria as DD, ID, and mixed group. The mixed groups are processed in phase 2 (Figure 5.2) in order to fine-tune the annotation and label the group according to DD, ID, and IC. The mixed groups annotated, and non-mixed groups form the automatic annotation sequence. The two annotations (manual and automatic) are fused to form the Annotated Corpus (AC).

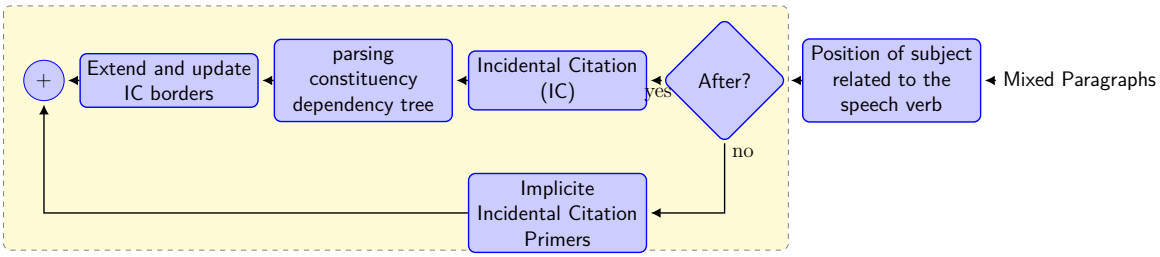


Figure 5.2: Detection and annotation of incidental clauses with reporting verbs (IC)

The method implemented to detect IC is initially based on the work of [Mareüil and Maillebau 2002], which consists of a set of regular expressions.

Based on this work, we add a set of rules that aims to detect IC in a more detailed way, and to cover more complex cases by relying on the syntactic analysis of the IC described by [Bonami and Godard 2008] and [Danlos, Sagot, and Stern 2010]. In our study, we can distinguish three configurations:

- Direct Discourse primers as in the example (5);
- Incidental Clauses with reporting verbs located in the middle of the speech of a character (4a);
- Incidental Clauses with reporting verbs placed at the end of the words of a character (4b).

We add to these three configurations, the cases where: a DD is inserted into an indirect speech in a sudden manner that is to say without a primer or other indication of a change of discursive perspective (see the example (3))

For analyzing these different configurations, it is necessary to look at other elements than just punctuation or dashes. In the proposed approach, we take into account both the result of the parser [Candito et al. 2009] and a lexicon of 327 reported verbs (affirm, repeat, exclaim, say, and so on). The entire list of reported verbs is in Appendix 1 .

Usually, when reporting verbs conjugated to the third person of the singular (in 97% of the cases), are detected, we have to have a look at their subject, in order to know its position compared to the verb. Two cases arise: if the subject is on the left of the verb (before), it is a primer; if, on the contrary, it is after the verb, we are dealing with an IC. In this case, it is crucial to establish its extension. The IC can be short (as (4a)) or relatively long (see (4b)). For doing this, we rely on punctuation, but also on parsing, especially for elements on the right of the verb that may depend on the subject as in the case of apposition or relatives (see example (4b)). The complete process is illustrated in Figure 5.2.

3.1 Rule-based results

The results obtained by this algorithm for detecting speech types are given in Table 5.2. Performance is estimated with three measurements: precision, recall, and F1-score.

	Precision	Recall	F1-score
Paragraph annotations (Phase 1)	92,6	91,2	92,19
Simplified detection of discourse types (DD, ID, IC)	87,5	85,2	86,33
Precise IC detection (with fine delimitation)	89,7	88,5	89,09

Table 5.2: Results of detection and annotation of discursive changes

The algorithm allows proper classification of the paragraphs (92.19% of good detection or F1-score) at the end of phase 1. For the analysis of the incidental citation (801 annotated manually on the corpus), performance is calculated by distinguishing two levels of annotation. Simplified detection - which is based on taking into account speech verbs and punctuation (see Figure 5.1, phase 2) does not allow to delimit with precision the extension of the IC (F1-score: 86,3%), errors appearing when a relative or an apposition depends on the subject. Taking into account the syntax and dependencies as shown in Figure 5.2 allows, on the other hand, to refine the results and improve them, so that we reach, after this precise detection, a score of 89.09% (this last one including the type of speech and the extent of the IC).

A study of the errors makes it possible to isolate two main cases:

- Those where the verb takes the form of a participle, and not of a finite verb, as in the example (6).

(6) – Cinq cents vers à toute la classe ! **exclamé d'une voix furieuse, arrêta, comme le Quos ego, une bourrasque nouvelle.** (*Madame Bovary, Chapitre 1*)

- Those where the parsing performed is erroneous as in the extract (7). "the syntactic complexity" of the Incidental Clauses with reporting verbs (IC) makes its analysis difficult

(7) – Oui... j'entends bien ; vous voulez que je vous mène à sa porte... et puis à son lit... et puis que je vous dise où frapper, et puis que je vous guide le bras, n'est-ce pas ? Vous voulez enfin me faire servir de manche à votre couteau !... vieux monstre! **reprit Tortillard avec une expression de mépris, de colère et d'horreur qui, pour la première fois de la journée, rendit sérieuse sa figure de fouine, jusqu'alors railleuse et effrontée.** On me tuerait plutôt... entendez-vous... que de me forcer à vous conduire chez votre femme. (*Les Mystères de Paris,Chapitre 7, Partie 2*)

The results obtained are not as good as those presented in direct and indirect discourse classifications based on machine learning algorithms (see [Schöch et al. 2016] who obtain an F1-score of 93.9% using a Random Forest (RF). However, this difference has to be taken with care because the objectives are not precisely the same. The procedure developed by [Schöch et al. 2016] aims to say whether each sentence is direct or indirect discourse, but does not isolate IC insights, primers, or direct speech passages in a narrative sequence. A fundamental difference of objectives explains this: whereas[Schöch et al. 2016] wants to classify the works on literary bases by relying on the presence or not of direct speech, we wish to know who speaks and at what precise moment change occurs. Moreover, the differences of results can be explained by the chosen method: whereas [Schöch et al. 2016] takes as a basic unit the sentence, we take the paragraph to indicate any discursive change in the same paragraph.

4 Machine learning approach

We extend the above work to explore procedures similar to those retained by [Schöch et al. 2016], but keeping the same objectives, namely determining precisely where discursive changes occur.

By observing more extensive data than those used Section 3, we observe that the identification of Direct Discourse (DD) , Indirect Discourse (ID),and the Incidental Clauses with reporting verbs (IC) less trivial using rule-based techniques than it seems to be since in French typographical DD is not marked with opening and closing quotation marks (example 8). By applying the above-described rule-based algorithm on new data, a different author, and written differently, we realized that rule-based algorithms have several disadvantages in other respects:

- Conflict between rules, some rules have to follow a specific order.
- When the size of the data increases, it is more likely to have additional cases not covered by the crafted rules. We need to increase the numbers of rules, which can be problematic because, at some point, the designed rules will become deprecated.
- It is difficult to generalize the procedure to other languages than the one for which the rules were designed.

To solve these problems, we propose to rely on machine learning to automatically identify the discourses in a larger collection of French-language fictional. We assume that

there is enough data to build a robust model. We also hypothesize that there are enough linguistic markers that make distinguishable the three discourse types of Direct Discourse (DD),Indirect Discourse (ID), and Incidental Clauses with reporting verbs (IC).

4.1 General Methodology

We consider the rule-based approach proposed above as a baseline for discourse classification task. We have three distinct classes DD, ID, and IC. From the manuel annotated data, we separated the text related to each class.

To carry out this study, we have mainly relied on the work presented in [Kowsari et al. 2019], which presents a survey of recent techniques dedicated to automatic text processing and classification.

4.2 Data used and feature extraction

For this work we have used the annotated extracts in Section 2 the text were preprocessed to perform supervised learning text.

In text classification paradigm, the word is the atomic element of an utterance (sentence, paragraph, and document). There are several methods for representing a word in sentence[Kowsari et al. 2019]. For this task, we use word embedding, which are high-dimensional continuous vector representation of words. In this vector space, words that have similar distributions are closer together than words with different distributions, given some distance measure. This type of setup has been shown to capture relevant syntactic and semantic properties of words, and they have been successfully applied to various tasks [Collobert and Weston 2008; Socher et al. 2011; Mikolov, Le, and Sutskever 2013].

To learn these embeddings, we have used freely available text data of SynPaFlex-Corpus. This data has been preprocessed and cleaned, and we have kept the first 5 million words. We trained the model on this dataset using an embedding size of 100. The systems use the publicly available word2vec[2] implementation of the skip-gram model with negative sampling, and they were trained for 15 epochs with a window of 5 words.

[2]https://code.google.com/p/word2vec/

4.3 Experimental setup

To perform the classification task, we first consider the model algorithm proposed in [Schöch et al. 2016], where the authors used traditional machine learning models such as Support Vector Machine (SVM) [Chang and Lin 2011], Maximum Entropy [Nigam, Lafferty, and McCallum 1999], Naïve Bayes [John and Langley 2013], Random Forest [Liaw, Wiener, et al. 2002] and JRip [Cohen 1995] to perform binary classification of direct speech vs narration. The best result were obtained with Random Forest. For this work we kept two algorithms Random Forest, Support Vector Machine, that we have fine-tuned with the grid search algorithms in order to find the optimal hyper-parameters.

We consider the deep neural approach proposed by [Tripathi, Sarkar, and Rao 2017], which aims to classify hindi sentence storytelling according to three distinct discourses: descriptive, narration, and dialogue. The authors use the Convolutional Neural Network (CNN) for extracting a robust representation of sentences from word embedding, and SVM for classification.

In this work, in addition to the CNN-SVM implementation proposed in [Tripathi, Sarkar, and Rao 2017], we propose to explore two other neural network-based architectures Recurrent Neural Network (RNN), Recurrent CNN (RCNN).

The experiment carried out on the SynPaFlex-corpus to analyze the accuracy of the discourse prediction models. To evaluate the performance of the implemented models we used a various parameters. Note that the effect of each parameter significantly alters the performance of the model. At the time of training and testing, input to the model is a sentence. Each of the sentences are labeled with one of the discourse mode(DD, ID, IC). In this work we use cross validation technique to avoid overfitting, in particular in case of models with a lot of parameters. This technique consists of splitting the data into K folds, where the blocks are sized equally. Each block turn as the validation set and the rest of (K-1) blocks is used for training the model. The final loss value is the average of K loss result. We fixed the number of blocks to K=6.

We used 2146 sentences in training and the remaining 920 sentences are used for testing.

4.4 Results

The performance of the proposed methods is evaluated using confusion matrix, Receiver Operating Characteristic (ROC) curve, precision, recall, f1-measure and accuracy. A

graphical plot of the performance is shown by ROC curve. This curve considers only true positive rate and false positive rate of the testing data. Here recall show the true positive rate, precision gives the positive predictive value, f-measure (Espindola and Ebecken, 2005) is the harmonic mean of precision and recall. The performance of the proposed methods is evaluated using confusion matrix, ROC curve, precision, recall, f1-score and accuracy. A graphical plot of the performance is shown by the ROC curves. This curve considers only true positive rate and false positive rate of the testing data. Here recall shows the true positive rate, precision gives the positive predictive value, f1-score [Espíndola and Ebecken 2005] is the harmonic mean of precision and recall. Table 5.3 presents the results of the discourse mode classification. RNN outperformed all the other models with an accuracy of 86.08%. DD mode classification is 91% because of more training data for this mode, and IC mode classification is 81%, and ID mode classification is 79% because for these classes we have fewer data to train our model. Figure 5.3 represents the ROC for tested models where class 0, class 1 and class 2 accounts for the DD, ID, and IC mode respectively. Class 0 (DD mode) has larger true positive rate than other two classes.

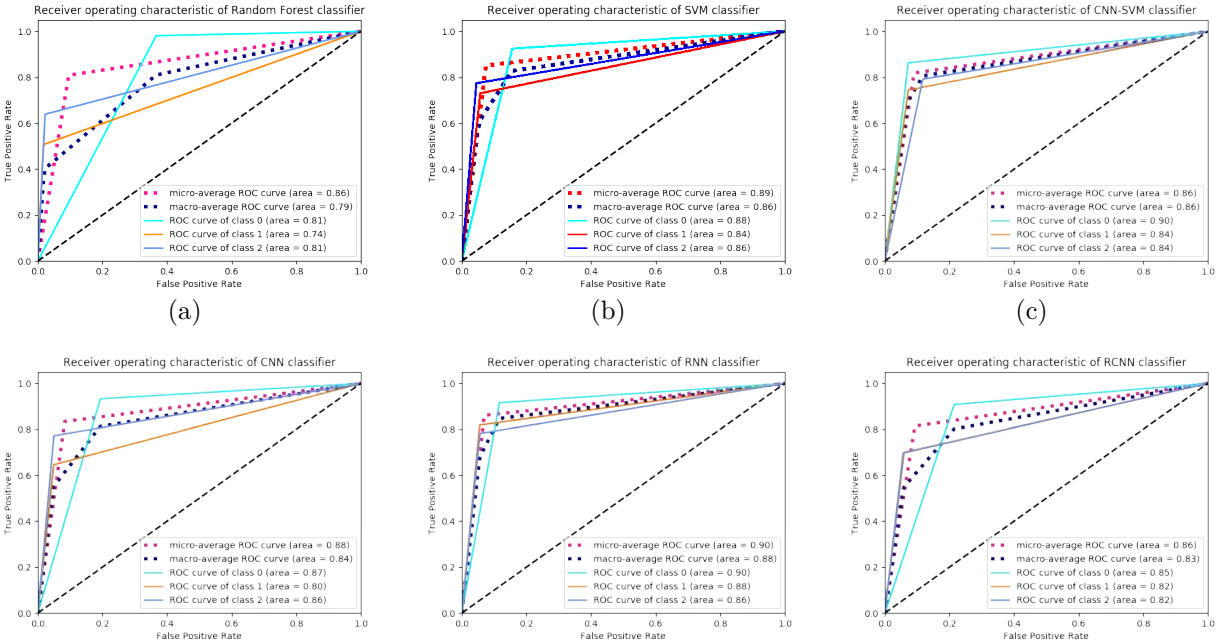


Figure 5.3: Receiver Operating Characteristic (ROC)

The results obtained with the RNN model outperforms other algorithms. The reported results indicate that almost all models show high-performance rates when it comes to

predicting direct speech, incidental citation , and low rates when it comes to the narration. These results are probably due to the fact that the narration contains a large vocabulary as well as utterances with a different organization unlike the DD and the IC, which for most of them are short, even very short for some samples. In addition, the syntactic structure of the incidental citations and the direct speech are quite consistent, which is in line with the results found in the first part based on rules.

		ID			DD			IC			Weigthed Average		
	accuracy	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
RF	80.86	86	51	64	77	98	86	91	64	75	83	81	80
SVM	85.10	72	73	72	0.88	92	90	87	77	82	85	85	85
RNN	**86.08**	**76**	**82**	**79**	**90**	**92**	**91**	**85**	**78**	**81**	**86**	**86**	**86**
CNN-DNN	83.36	75	65	70	85	93	89	86	77	81	83	83	83
CNN-SVM	81.89	62	79	70	94	86	90	78	74	76	83	82	82
CNN +RNN	81.05	72	70	71	84	91	87	82	70	75	81	81	81

Table 5.3: Result of classification

5 Conclusion

In this chapter, we have proposed an algorithm that automatically and accurately annotates discursive changes in audiobooks. The performance of the tool is relatively encouraging, but errors remain in syntactically complex cases.

The rule-based approach still useful even though it does not generalize well, and it requires knowledge. We can use this approach to validate a given rule and extract robust and valuable features for machine learning approaches.

In the next chapter, we rely on the audio signal to have a better understanding of discourse changes in general and to better delineate the Incidental Clauses with reporting verbs (IC) in complex cases.

Chapter 6

Automatic prosodic analysis of discourse changes in audiobook

Emphasizing on discourse and character changes is very important to improve expressivity in text-to-speech synthesis (TTS) systems reading audiobooks. It makes stories easily understandable by listeners with a lower cognitive effort and it enables a better access to the exact content. Understanding which prosodic cues are used to this end is thus relevant as they could then be implemented to enhance speech quality and expressiveness in TTS systems.

This work aims to investigate how discourse and character changes occur and which robust prosodic patterns are able to encode them. For this exploration, focus was given to literary fictions which contain a considerable amount of dialogues and similar narrative schemes. Moreover, each selected story was recorded by two speakers allowing a comparison between speakers and styles.

For the prosodic analysis, the data were first segmented into breath groups, then five main prosodic features were automatically analysed: (i) F0 range (in semi-tone, because semitones are more suitable for measuring temporal events and to notify relative variations in a sequence.), (ii) articulation rate (syllables/sec), (iii) breath group duration, (iv) average log energy, and (v) inter-breath group pause duration.

The results obtained from a statistical analysis show that speakers mainly employ inter-breath-group pause duration and F0 range to encode discourse and character changes.

1 Introduction

To achieve a good synthetic voice in terms of expressiveness and naturalness, a deeper understanding of natural speech is required. In recent years, although the use of data-driven techniques, together with appropriate data, enhances Text-to-Speech systems, there is still a gap between natural speech and synthesized speech. It becomes even more obvious

while reading long and coherent texts such as audiobooks aloud. However, many works have been done to tackle this problem either in traditional or end-to-end approaches. Traditional approaches explore prosodic and linguistic features and enrich feature vectors [Székely et al. 2012b; Székely et al. 2012a; Mamiya et al. 2013; Charfuelan and Steiner 2013; Vít and Matou?ek 2016]. [Sarkar and Rao 2015] for instance explore pause prediction problems using linguistic information such as discourse types. On the other hand, end-to-end paradigms[Wang et al. 2017a; Ping et al. 2017; Tachibana, Uenoyama, and Aihara 2018], which avoid complex feature engineering process and learn from raw data, are data sensitive and are affected by the quality and quantity of the data to be treated. They are also computationally expensive.

One of the main challenges of prosody modeling is that there is considerable inter- and intra-speaker variability. We make two different assumptions; however: (i) all speakers assign different prosodic properties to encode a specific style or one of the characters involved in the story; (ii) discourse and character changes are always encoded, but, because of variability, different strategies and acoustic features may be used.

These variabilities are not always integrated into the speech, some studies have shown that there are somewhat similar styles (spoken newspaper, neutral reading) and others which are not (political speech, slam); therefore, this inter-speaker variability depends on text genre.

Certain literary genres, such as storytelling [Sarkar and Rao 2015; Theune et al. 2006; Buurman 2007; Montaño, Alías, and Ferrer 2013; Harikrishna D M, Gurunath Reddy M, and Rao 2015; Ramli et al. 2016; Montaño and Alías 2016], have received much attention, most of which have analyzed discourse at the sentence level. Indeed, this granularity is not informative enough in the case of long textbooks, such as novels or short stories. It is mainly due to the fact that those texts are more complex and subtle along various dimensions such as syntactic structure, lexis, discourse patterns, but also character psychology. In this work, we thus decided to study French fictional stories addressed to adults and analyze how the different discourse types and character changes were encoded prosodically.

The current work investigates natural speech with the use of different automatic prosodic annotation procedures which can be compared to the ones used in tools such as SLAM [Obin et al. 2014], ADoReVa & ADoTeVa [De Looze and Hirst 2008], MOMEL & INTSINT [Hirst 2007]. It allowed focusing on the way certain prosodic features change over time. In order to generate adequate prosodic patterns in a TTS system, it is important to know exactly which prosodic cues come into play for indicating any change in the

discursive perspective.

To analyze discourse changes in fictions, it is necessary to clearly delineate different types of speech in the text. This allows getting a more precise idea of "who speaks". In addition, it is important to study how these changes are encoded by prosodic cues, despite inter and intra-speaker variability. We have thus investigated breath group modifications, in terms of articulation rate and pitch range, in the particular case of discursive perspectives (from direct to indirect speech), or character changes (in dialogs) based on the fact that breath group is a good unit to study continuous speech in both read and spontaneous speech according to [Wang et al. 2010].

One of the goals of this study is then to integrate the observed prosodic cues into a speech synthesis system to improve its quality and expressiveness. To reach this end, we did refer to work that has already been done on discourse properties in French. There has been a growing interest in studies focusing on the parsing of incidental clauses with reporting verbs [Buvet 2012] or on the semantic and syntactic values of such sequences in discourse [Beyssade 2012]. According to [Buvet 2012], incidental clauses are characterized, in French read speech, by syntactic features as well as by a specific prosodic behavior.

This chapitre firstly describes the corpus and methods. It further explains how the different discourse perspectives are encoded and how the prosodic features were analyzed. Finally, the results of the experiment are presented and discussed.

2 Corpus Design

2.1 Experimental dataset

This present work is concerned with highlighting prosodic cues, and the prosodic unit used for making discourse change in fictional audiobooks. This investigation was conducted on audiobook samples recorded by two female speakers (FFR0012, FFR0001 of MUFASA corpus). This subset includes extracts selected from two French novels, *les Mystères de Paris* by Eugène Sue and *Madame Bovary* by Gustave Flaubert. The selected extracts involve many discursive perspective changes, which allows getting a relatively coherent set in terms of direct and indirect discourse sequences. The *Madame Bovary* extracts are read by both speakers, where *les Mystères de Paris* extracts, is read-only by the principal speaker (FFR0001). The table Table 6.1 shows the details of the studied dataset.

Book	Nutt*	Direct Discourse (%)	Indirect Discourse (%)	Mixed Discours (%)	FFR0001	FFR0012	Duration (hours)
Madam Bovary (9 Chapters)	579	25	50	25	✓	✓	∼3h20 (X2)
Les Mystère de Paris (6 Chapters)	690	43	36	21	✓		∼2h19

Table 6.1: Overview of the sub-corpus content. N-utt represent the number of utterances.s

2.2 Preprocessing

All audio data used for this experience are in wav format, with a sampling rate of 44.1kHz, mono channel, 16 bits. To obtain aligned and consistent data, the following steps were taken during prepossessing phase:

- All the speaker dependent data such as introductions and conclusions were removed;
- For reducing the potential background noise, we considered long noisy silences as a typical profile of the noise;
- The DC offset has been removed; in some collected data, there is fixed voltage offset inserted during the recording process, this offset can affect the quality of recordings; for that reason, it is preferable to remove this offset.
- Amplitude has been normalized to avoid clipping.

2.3 Text annotation

Using a rule-based program, presented in Section 3, texts were automatically annotated in order to distinguish the paragraphs consisting of simple narration (1) and the sequences corresponding to direct speech dialogs (2). The annotation procedure also allows to delimit the extension of direct speech passages within narrative paragraphs, i.e reported speech as in (3), and also parentheticals or incidental clauses with reported verb in direct discourses, which can be short (4a) or relatively long (4b), and located in the middle of a direct speech sequence (4a) or at the end of it (4b).

(1) *On commença la récitation des leçons. Il les écouta de toutes ses oreilles, attentif comme au sermon, n'osant même croiser les cuisses, ni s'appuyer sur le coude, et, à deux heures, quand la cloche sonna, le maître d'études fut obligé de l'avertir, pour qu'il se mit avec nous dans les rangs.* [We began reciting our lessons. He listened attentively, concentrating as though listening to a sermon, not daring even to cross his legs or lean on his elbow, and, at two o'clock, when the bell rang, the master had to tell him to line up with us all.] (*Madame Bovary*, chap. 1)

(2) – *Soit, demain à une heure.*
– *A une heure.*
– *Dans la plaine Saint-Denis?*
– *Dans la plaine Saint-Denis.*
– *Entre Saint-Ouen et le chemin de la Révolte, au bout de la route?*
– *C'est dit.*
[– Be it so; tomorrow at one.
– At one o'clock.
– In the plain of St. Denis?
– In the plain of St. Denis.
– Between St. Ouen and the road of La Revolte, at the end of the road?
– Agreed.] (*Les Mystères de Paris*, chap. 7, Tome 1)

(3) *D'autre part, la mort de sa femme ne l'avait pas mal servi dans son métier, car on avait répété durant un mois :* ***« Ce pauvre jeune homme ! quel malheur ! »***
[In any case, the death of his wife had done him no harm professionally; for a whole month people kept saying: **«That poor young man! What a terrible thing!»**] (*Madame Bovary*, chap. 3)

(4) a. – *Levez-vous,* ***reprit le professeur****, et dites-moi votre nom.*
[– Stand up, **repeated the master**, and tell me your name.] (*Madame Bovary*, chap. 1)

b. – *Débarrassez-vous donc de votre casque,* ***dit le professeur, qui était un homme d'esprit****.*

Il y eut un rire éclatant des écoliers qui décontenança le pauvre garçon, ...
[– I suggest you disencumber yourself of your helmet, **said the master, a man of wit.**
A roar of laughter came from the class and disconcerted the poor lad,...] (*Madame Bovary*, chap. 1)

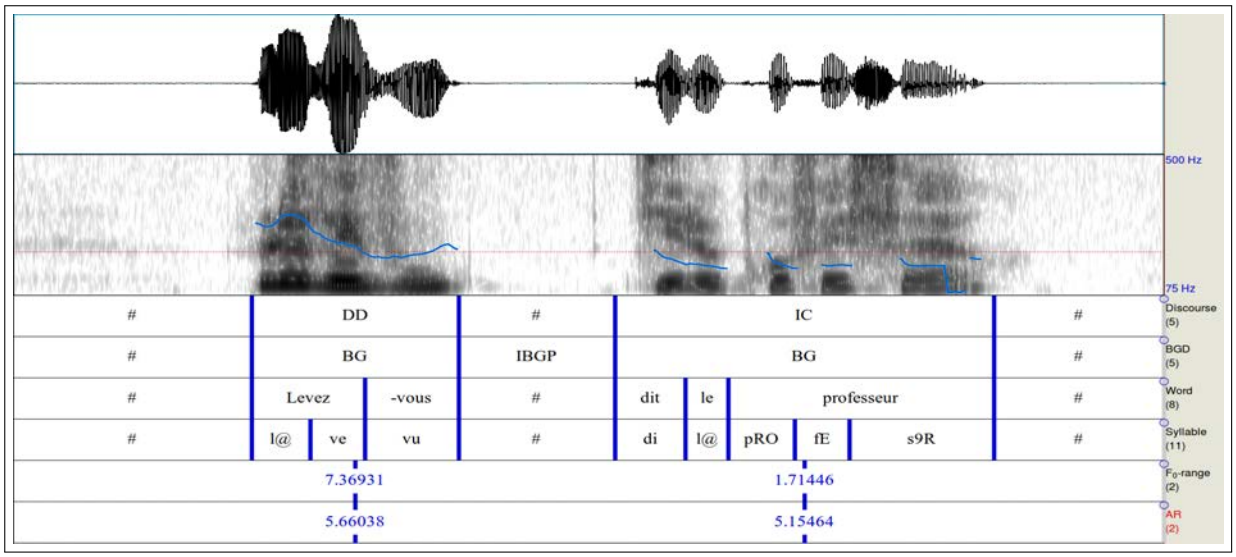


Figure 6.1: Illustration of an example of discourse passage from Direct Discourse to Incidental Clauses with reporting verbs (DD ⇒ IC) corresponding to one modality and data structure. The tiers correspond (from the buttom to the upper one): Articulation Rate articulation rate measured with Equation (6.2) , F_0-range with Equation (6.1), syllables, words, breath group and related discourse.

This work allowed to delimit with precision (87%) the passages according to their discourse type: direct discourse (DD), indirect discourse (ID) and incidental clauses with reporting verbs (IC). This allowed to distinguish six cases of discursive perspective changes:

- from indirect discourse (or narration) to direct discourse, noted DI⇒ DD;
- from direct discourse to indirect discourse (or return to narrative paragraphs), noted DD⇒ DI;
- from direct discourse to incidental clause with reporting verb, as in the transition from *levez-vous* to *dit le professeur* in (4a), noted DD⇒ IC;
- from incidental clause to direct discourse, as in *dit le professeur* to *et dites-moi votre nom* in (4a), noted IC ⇒ DD;

- from incidental clause with reporting verb to the indirect discourse (or narration), as in (4b), noted IC⇒ ID.

- from direct discourse to direct discourse, with a character change (dialog sequence), as in (2), noted DD⇒ DD.

The distribution of discursive perspective changes taken into account for the analysis of prosodic parameters are given in details in Table 6.2.

Table 6.2: Discursive changes distribution sub-corpus

Audio books	Madame Bovary		Les Mystères de Paris	
Discourse changes	items number	syl./item number	items number	syl./item number
IC ⇒ DD	88	34	554	46
IC ⇒ DI	20	49	56	55
DD ⇒ DI	73	73	172	88
DI ⇒ DD	61	61	129	54
DD ⇒ IC	107	15	528	26
DD ⇒ DD	39	29	1195	78

3 Prosodic analysis

3.1 Features Extraction

As we mention in Section 1, many tools are available to analyze prosody and extract relative features from different perspectives. We decided to build our framework to have more control and easily interpret the results. Also, we considered that the new framework corresponds better to the nature of the data that we are analyzing. The prosodic analysis is based on the breath group granularity (as a unit function) and focused on five cues: F_0-range (in semi-tone), articulation rate (syllables/sec), average vowel lengthening rate, average vowel log energy, and pause duration at the juncture between breath groups. In addition, the analysis of these prosodic cues was first carried out on a subpart of our corpus, and then validated on the whole data set. To analyze the various prosodic features, the last breath group of a given discourse type (DD, ID) and the first breath group of the targeted discourse in case of change were taken into consideration. Thus, to analyze

changes from direct discourse to incidental clause as in (4a), the values for F0-range were calculated for the last breath group of the direct discourse sequence, i.e. *Levez-vous* in (4a), and the first breath group of the clause, i.e. *dit le professeur* in (4a), according to the equation (6.1), where M represents the number of vowels within the breath group and $V_{F_{0median}}$ stands for the median vowel F_0 value within the breath group.

$$\begin{cases} F_{0min} = \arg\min(\sum_{i=0}^{M} V_{F_{0median}}) \\ F_{0max} = \arg\max(\sum_{i=0}^{M} V_{F_{0median}}) \\ F_{0range} = 12 \times log_2(\frac{F_{0min}}{F_{0max}}) \end{cases} \tag{6.1}$$

The same procedure was followed to study articulation rate AR, in syll/s, the latter being computed for the last breath group (BG) of the first discourse sequence and the first one in the second discourse sequence. The articulation rate is computed as follows:

$$AR = \frac{N}{\sum_{i=0}^{N} \text{Syllable Duration}[i]} \tag{6.2}$$

where N is the number of syllables in a given breath group.

The analysis of inter-breath groups pause duration in the studied data has shown that the speakers tend to insert breaks upper to 200 ms (0.2s) to mark the transition from discourse to another. These phenomena appear to be independent of the articulation rate of the surrounding speech segments.

For each breath group, the average log energy is computed over the set of extracted log energy of its vowels.

$$logEnergy = \frac{\sum_{i=0}^{M} V_{logEnergy}}{M}$$

Where M refer to the number of vowels in the breath group.

Since the duration of the breath group has a relative variability, we have also measured duration of each breath group.

A statistical analysis of significance has been done using a χ^2 test to analyze discourse changes impact. The six configurations have been tested for the different features, namely F_0-range, articulation rate and pause duration, at α=0.01 level.

3.2 Hypothesis

In this work, we investigate two hypotheses:

- The first hypothesis is that both speakers mark discourse and character changes by using a particular prosodic properties, which can differ from one speaker to another.
- The second hypothesis is that discourse and character changes are local phenomena ; as a consequence, the way changes are encoded may differ within an entire novel, but differences between two consecutive breath groups should occur in a clear dynamic way.

4 Results and discussion

In this section, the results of the prosodic analysis outlined in Section 2.3 are presented. Table 6.3 presents the results obtained for one speaker. Among the five cues investigated, F_0-range and pause duration play a role. By contrast, articulation rate, average log-Energy and breath group duration differences between two breath groups surrounding a discourse change are not significant for both speakers. Results even tend to show that articulation rate is quite stable among breath groups

According to the results in Table 6.3, discourse changes have a significant impact on F_0-range. Among the different cases investigated, a pitch range compression occurs on the first breath group after discourse change, except after an incidental clause (IC $\Rightarrow$ DD or IC $\Rightarrow$ ID). This could be related to the fact that incidental clauses are shorter and treated as embedded and autonomous at the syntactic and prosodic level. Concerning changes from DD to ID, F_0-range difference is less significant (statistical significance at $\alpha = 0.05$ level), but other parameters such as pause duration enter into play, as we will see later. Thus, the combination of prosodic features allows the correct encoding of discourse changes.

Average duration of inter-breath group pauses according to discourse change types are reported in Table 6.4. We can thus observe that pause duration is significantly lower when changing from an incidental clause (IC) to direct discourse (DD), and the other way around. This could result from the fact that incidental clauses are embedded in a larger group (e.g. [levez-vous (dit le professeur) et prenez...]). In addition, one can notice that longer pauses are realized when returning to indirect discourse. When changing, for instance, from direct discourse (DD) to indirect discourse (ID) the average length of a pause is 1.26s, whereas the average pause duration is only 1.02s when introducing a direct

Table 6.3: Means and standard deviations for F_0-range and articulation rate (AR) for the different types of discourse change.

Parameters	First Disc. Last BG	Second Disc. First BG	
	IC	DD	p-value
F0range,st	5.91 (5.76)	7.47 (5.91)	2.0
AR,syl./s	5.13 (0.77)	5.25 (3.17)	1.64
	DD	IC	p-value
F0range,st	**7.08 (5.91)**	**4.43 (4.94)**	**<0.001**
AR,syl./s	4.78 (1.22)	5.11 (0.80)	2.0
	DD	DD	p-value
F0range,st	**8.72 (5.64)**	**7.42 (5.78)**	**<0.001**
AR,syl./s	5.15 (0.91)	5.13 (2.70)	0.788
	DD	ID	p-value
F0range,st	**8.78 (5.66)**	**7.68 (5.25)**	**0.0235**
AR,syl./s	4.98 (0.96)	5.23 (3.06)	1.88
	ID	DD	p-value
F0range,st	**11.23 (5.35)**	**6.94 (5.67)**	**<0.001**
AR,syl./s	4.95 (0.96)	5.04 (1.99)	1.44
	IC	ID	p-value
F0range,st	6.49 (5.88)	8.11 (5.62)	1.91
AR,syl./s	4.95 (0.68)	4.98 (0.74)	1.23

discourse (ID to DD). Note however that this type of pauses, which indicates a move from Indirect Discourse to Direct one, is also among the longer ones (see Table 6.4).

Table 6.5 reports the significance level of the inter-breath group pause length difference when comparing the discourse changes two by two. It can be seen that the p-value < 0.001 is statistically significant in nearly all cases. Moreover, one can point out that if we compare IC⇒ ID to DD⇒ ID and IC⇒ ID to ID⇒ DD, the difference is less significant than in other cases.

Furthermore, we have analyzed the behavior of the same prosodic cues for consecutive breath groups when the discourse type remains the same. Concretely, it corresponds to the following transitions: IC⇒IC, ID⇒ID and DD⇒DD with no character change. The results of this analysis show that there is no significant difference for the three features in that case. This finding shows that the reader adopts a specific strategy to mark discourse changes.

The whole analysis has been done for both speakers and the results are similar. This

Table 6.4: Means and standard deviations for Inter-Breath Group Pause Duration (IBGP) according to different types of discourse change.

First Disc. ⇒ Second Disc.	IBGP (in s.)
IC ⇒ DD	0.67 (0.27)
DD ⇒ IC	0.43 (0.17)
DD ⇒ DD	0.95 (0.23)
DD ⇒ ID	1.26 (0.39)
ID ⇒ DD	1.02 (0.24)
IC ⇒ ID	1.11 (0.32)

may be related to the fact that the speakers are of the same sex. Further investigations are needed to assess how other speakers behave.

Table 6.5: Comparing IBGP across the different discourse changes modalities (** represents p-value<0.001).

	IC⇏DD	IC⇏ ID	DD⇏ IC	DD⇏ DD	DD⇏ ID	ID⇏ DD
IC⇒DD		**	**	**	**	**
IC⇒ ID	**		**	**	.004	.023
DD⇒ IC	**	**		**	**	**
DD⇒ DD	**	**	**		**	**
DD⇒ID	**	.004	**	**		**
ID⇒ DD	**	.023	**	**	**	

5 Conclusion and perspectives

This study investigated how discourse changes, in a French audiobook corpus, are prosodically characterized. This study relies on five basic prosodic cues at the breath group level: F_0-range, articulation rate, breath group duration and inter-breath groups pause durations. The results show that F_0-range and pause durations are relevant features to differentiate two distinct and consecutive discourse types. This work has also confirmed that the breath group is an interesting functional unit for studying long and expressive speech in audiobooks.

A deeper investigation on the breath group structure and its relations to the prosodic

cues needs to be done. To do so, a large set of features could be studied such as linguistic, phonetic, phonological and paralinguistic features.

We also plan to redo the experiment with the rest of the parallel data of the MUFASA-Corpus . For these future experients, we are considering other tracks for better understanding the discursive change phenomena and the prosodic properties of breath group:

- Investigating the direction, as well as the amplitude of the pitch curve on the last word during the discursive changes. These phenomena seem to be an essential and necessary measure to have more control over a micro-prosodic manifest of discourse changes.
- Analyzing the possible prosodic reset (F0 reset, for example) between the last syllable of the breath group and the first syllable of the next breath group.
- Studying the declination line, which seems to appear in the F0-range as a marker of the central prosodic unit, and then the analysis of possible declination lines within specific sequences.

Furthermore, integrating the observed results in a TTS system should allow designing perceptual tests and collecting subjective evaluations. Then studying the relation between the features and their impact on synthesized speech could lead us to build a solid knowledge on the prosodic behavior occurring in audiobooks. In addition, tested prosodic features could bring more expressiveness to speech synthesis systems and thus enable new applications.

Chapter 7

Speaker Prosodic Identity

This chapter is an extended version of the work described in "Introducing Prosodic Speaker Identity for a Better Expressive Speech Synthesis Control" presented at Speech Prosody 2020. This work was performed in collaboration with Sébastien Le Maguer (ADAPT Centre), who was an invaluable partner through-out.

1 General Context

To have more control over TTS synthesis and to improve expressivity, it is necessary to disentangle prosodic information carried by the speaker's voice identity from the one belonging to linguistic properties. In this work, we propose to analyze how information related to speaker voice identity affects a DNN based multi-speaker speech synthesis model. To do so, we feed the network with a vector encoding speaker information in addition to a set of basic linguistic features. We then compare three main speaker coding configurations: *a*) simple one-hot vector describing the speaker gender and identifier ; *b*) an embedding vector extracted from a speaker recognition pre-trained model ; *c*) a prosodic vector which summarizes information such as melody, intensity, and duration. To measure the impact of the input feature vector, we investigate the representation of the latent space at the output of the first layer of the network. The aim is to have an overview of our data representation and model behavior. Furthermore, we conducted a subjective assessment to validate the result. Results show that the prosodic identity of the speaker is captured by the model and therefore allows the user to control more precisely synthesis.

2 Introduction

The quality of speech synthesis systems has drastically increased during the last years. Thanks to the deep learning paradigm, it is now possible to generate speech, which sounds

almost like humans. In the meantime, however, the control over models remains challenging because of their complexity.

Expressive speech synthesis relies on adequate control on the prosodic parameters. These parameters depend on the linguistic features of the text to read as well as information related to the voice used for synthesizing the speech.

Therefore, disentangling the speaker characteristics from the linguistic content is a key feature to control the rendering of the synthesis.

Disentangling speaker characteristics from linguistic content is even more crucial to have proper control in multi-speaker Statistical Parametric Speech Synthesis as, by definition, the model should produce a speech corresponding to one consistent speaker. Counting on the robustness of multi-speaker modelling, studies show that expressive speech synthesis systems can benefit from such an environment [Fan et al. 2015], although it raises other challenges related to recording conditions [Hsu et al. 2019], speaker coding[Hojo, Ijima, and Mizuno 2018] and controllability [Henter, Wang, and Yamagishi 2018; Hsu et al. 2018; Lazaridis, Potard, and Garner 2015; Bian et al. 2019].

Therefore, we propose here to investigate whether a model can seperate speaker characteristics from linguistic features in a standrad DNN TTS multispeaker environnement by using a naive but fully controllable representation of prosody.

This chapter is structured as follows. The different speaker coding configurations are presented in Section 3. Section 4 gives an overview of the methodology and Section 5 details the experiments we conducted to analyze the influence of these configurations on the model. Finally, in Section 6, we go through the results of the experiments using complementary objective analysis methodologies and subjective assessment.

3 Speaker Coding

To encode the speaker voice characteristics, we are using three different configurations from the most opaque (OneHot-Vector) to the most controllable one (P-Vector). The intermediate representation (X-Vector) has been added as it is a state of the art representation for the speaker identification domain.

3.1 OneHot-Vector

This configuration to encode the speaker information for DNN based speech synthesis has been explored in [Hojo, Ijima, and Mizuno 2018]. As a first and intuitive choice for speaker encoding, we propose a simple one-hot vector of two parts: (1) gender (female/male) as this is the highest level distinction we can do, (2) identifier of the speaker to distinguish speakers intra-gender. This approach makes the control of the synthesis most complicated as we just have a discrete choice; thus it does not take into account the acoustic proximity between speakers.

3.2 X-Vector

X-Vector [Snyder et al. 2018] are the state of the art represention used in the speaker identification field. To get the X-Vectors, we extract embedded vectors independently on the text using a pre-trained model[1]. As stated before, this model was initially trained for a speaker verification task [Snyder et al. 2018; Snyder et al. 2017; Xu et al. 2018] using NIST SRE recipe supported in the Kaldi toolkit. The details about the recipe and the pretained model are available in author's github[2].

This configuration is more detailed than the OneHot-Vector but still remains difficult to control as the dimensions of the X-Vectors are difficult to interpret.

3.3 P-Vector

The last configuration we are proposing is the P-Vector. To characterize the speaker style and the specificity of an expressive voice, we propose to use the breath group as the functional unit to build a vector able to cover high-level prosodic information which are difficult to predict from the text. A P-Vector is defined by the following features:

- F0-range: for each vowel of the breath group, we are computing the median values. Then, considering $F0_{min}$ and $F0_{max}$, respectively, the minimum and the maximum median values, we computed the scaled F_0 range the following way:

$$\begin{cases} F_{0min} &= \min(V^0_{F_{0median}}, \ldots, V^M_{F_{0median}}) \\ F_{0max} &= \max(V^0_{F_{0median}}, \ldots, V^M_{F_{0median}}) \\ F_{0range} &= 12 \times log_2(\frac{F_{0min}}{F_{0max}}) \end{cases}$$

[1] https://kaldi-asr.org/models/m3

[2] https://david-ryan-snyder.github.io/2017/10/04/model_sre16_v2.html

where M represents the number of vowels within the breath group and $V^i_{F0_{median}}$ stands for the median F_0 value of the i^{th} vowel within the breath group;

- Melodic pattern: for each vowel contained in a given breath group, the $V_{F0_{median}}$ has been extracted. The resulting sequence of values has been interpolated using a cubic spline. Then, a set of five equidistant values (at each 20% of the breath group duration starting from 10%) has been selected.
- Energy pattern: the same computation as the previous one is done on $V_{logEnergy}$.
- Articulation Rate: it is the number of syllable per second computed at the breath group level ignoring pauses;
- Duration of breath group in second;
- Duration of pauses around the breath group in second.

Therefore, we obtain a fully controllable feature vector whose dimensions can be interpreted properly.

4 Analysis Methodology

The experiments and analyses presented in this work were carried out within the Merlin[Wu, Watts, and King 2016] framework. We used the default configuration proposed in the toolkit, then we integrated the speaker coding vectors to achieve a multi-speaker TTS model.

4.1 Input and Output features

The input feature vector can be viewed as two concatenated vectors corresponding to two parts: a linguistic part and a speaker coding part.The first 319 coefficients correspond to the linguistic description of the utterance. This part is based on the standard feature set for English described in [Tokuda, Zen, and Black 2002] that we have adapted for French. The main differences with the English feature set concerns the accentuation. Indeed, as the accentuation information in French is strongly correlated to the Part of Speech (POS) information, we therefore consider that the POS information, already present in the vector, is enough to encode the accentuation information.The coefficients from dimension 320 and

beyond are the speaker code. The size of this part vary according to the configuration used among the configurations under study (e.g. OneHot-Vector, X-Vector or P-Vector).

The output feature vector contains the standard coefficient vector composed by the Voiced/Unvoiced (VUV) flag, the logF_0, the Mel-Generalized Cepstrum (MGC), the BAP, and their dynamic counterparts. This leads to a vector of 265 coefficients.

Finally, the input and output vectors are normalized using, respectively, Min-Max Normalization (MMN) and Mean Variance Normalization (MVN) methods.

4.2 Method

The main goal is to see if and how the content of the input vector influences the ability to separate speaker-related information in a DNN-based TTS system. To do so, we learn several systems differing by the structure of the input vectors provided. OOnce the different systems are learned, to analyze if the various configurations are guiding the models to capture speaker specificities, we propose to measure differences at the output of the first hidden layer as well as at the output of the model.

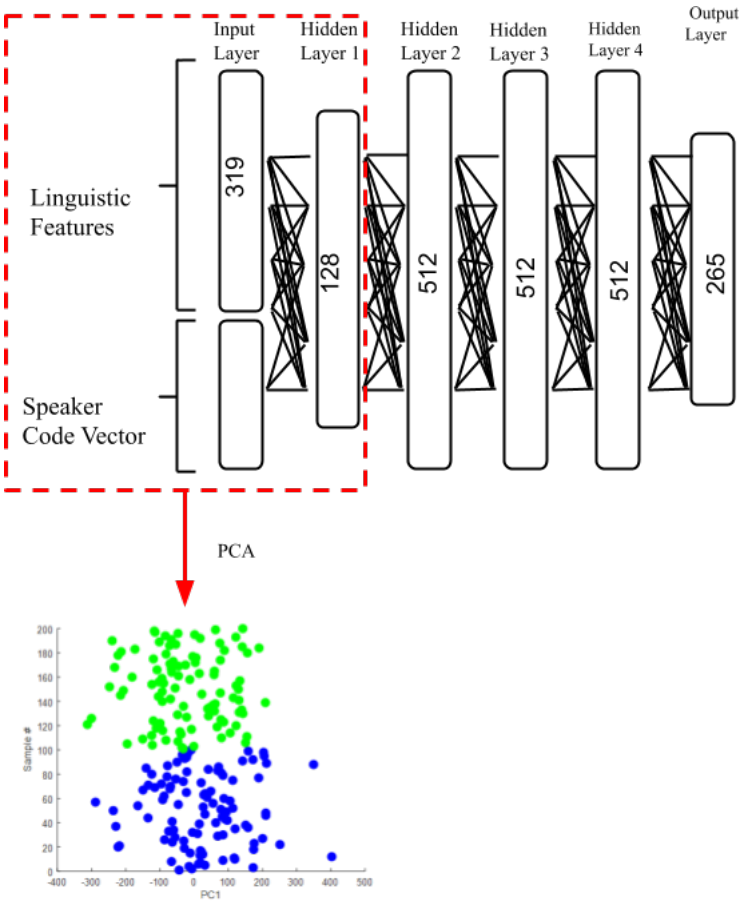
Two types of analyses are then done:

- Standard objective measures: MCD, BAP distortion, F_0 Root Mean Square Error (RMSE), F_0 correlation, VUV error rate, RMSE on the duration and duration correlation;

- A visual analysis protocol illustrated in the Figure 7.1: a PCA (see Appendix 1 for a technical prosodure of PCA) on the first hidden layer output is computed. Then, we visualize the main dimensions and analyze the results in function of the speakers to see if speaker-dependent information is captured by the model. We perform PCA at the end of each epoch on the validation dataset. We choose to do the analysis at this stage of the network because it is easier to interpret and quantify the variation brought by the input.

We also compare different epochs to see how the models are evolving. This monitoring is interesting since it enables to check quickly if the structure of the input vectors has an impact on speaker separability.

Figure 7.1: Top part represents the architecture of the proposed model, the bottom part illustrates the visualization process of the first hidden layer.



5 Experimental setup

5.1 Dataset

We have selected nine speaker from the MUFASA corpus (4 Females/ 5 Males). This subset contains fictional french audiobooks published between the 18th-20th century. We follow the same procedure as the one described in the 1. The text is splited into pseudo-paragraphs and then force-aligned to corresponding speech using JTrans[Cerisara, Mella, and Fohr 2009]. The speech signals are sampled at 48 kHz. All the meta-data information related to describe the book (speaker identifier, library name, ...) were removed. From the designed corpus, two groups of data were defined:

- *parallel data*: this group contains 5 audiobooks (for more information about the books, see the Appendix 4); each transcription have been read by at least 2 speakers.

In total, the data for 9 speakers has been collected including 4 females. Voices were selected by an informal listening test considering their recording conditions (non-audible difference), and the fact that the voice quality of the speakers are quite different.

- *non-parallel data*: for each speaker in the parallel data, 1h of extra speech has been collected with no overlap in the transcription. This set of data is used to evaluate robustness and performance of the speaker encoder input.

The procedure used to achieve the annotation process and to extract the linguistic features is described in [Sini et al. 2018].

5.2 Models configuration

To achieve training and synthesis, we used the Merlin toolkit[Wu, Watts, and King 2016]. The architecture of the model is a FF-DNN with 4 hidden layers. During the experiments, we changed first layer size to be 128, 256 or 512 neurons without any significant change. The last three layers have a fixed number of 512 neurons. The hidden layers use the *tanh* activation function and the output layer uses a linear activation function. We applied batch-training paradigm with a batch size of 256. The maximum number of epochs is set to 25 including 10 warm-up epochs. The learning rate is initially set to 0.002 for warm-up epochs and after that reduced by 50% for each epoch. Similarly, the momentum is set to 0.3 for warm-up epochs and to 0.9 otherwise. Finally, we used L2-regularization with a weight set to 10^{-5}. Models are learned considering speaker coding schemes with the following dimensions: 2 for OneHot-Vector (OHV), 32 for X-Vector and 9 for P-Vector.

6 Results

6.1 Standard measurements

In order to evaluate DNN-based TTS synthesis, the proposed method was applied to train models for each audiobook present in the parallel training set, and then on the non-parallel training set.

All the models have been evaluated using MCD, BAP distortion, RMSE on F_0 and duration, VUV rate and Correlation (CORR) on F_0 and duration, between the predicted

and the original coefficients. In this work, only the objective results concerning the non-parallel training dataset are reported as similar results have been observed in the parallel training dataset.

As shown in Table 7.1, the system involving the P-Vector outperforms the baseline system in all kinds of objective measures.

Table 7.1: Objective results for multi-speaker modeling, considering five speaker code configurations. Mel-Cepstral Distortion (MCD), Band Aperiodicity Parameter (BAP), Root Mean Square Error (RMSE), Voiced/Unvoiced (VUV) and Correlation (CORR) between the predicted and the original coefficients. For the F_0, RMSE and CORR are computed on the voiced frames only.

OHV	X-Vector	P-Vector	MCD (dB)	BAP (dB)	F0		VUV	Duration	
					RMSE (Hz)	CORR		RMSE (ms)	CORR
✓			5.833	0.301	32.597	0.807	8.950	9.232	0.558
	✓		5.935	0.303	33.018	0.801	8.971	8.889	0.601
		✓	5.748	0.296	32.203	0.811	8.851	8.883	0.604
✓		✓	5.756	0.297	32.169	0.810	8.944	8.860	0.607
	✓	✓	5.755	0.297	32.043	0.812	8.915	8.836	0.609

6.2 Visualizing the first hidden-layer output

Figure 7.2: PCA projection for the parallel data during the validation phase, the speaker identify is encoded as following (F/M: Female/Male, FR: French, ID:XXXX).

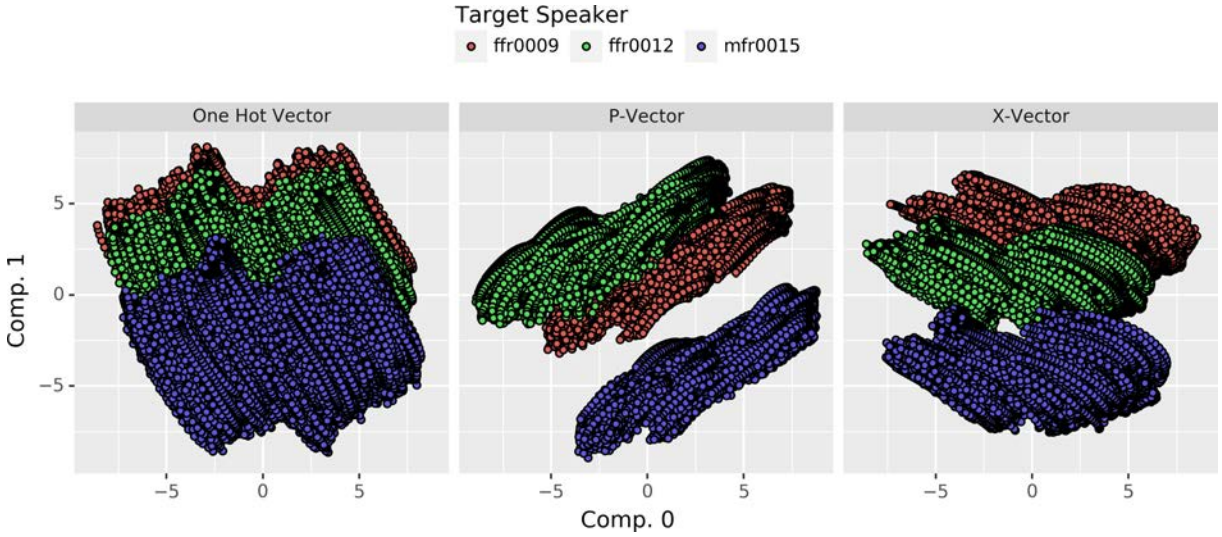


Figure 7.3: PCA projection for the non parallel data during the validation phase.

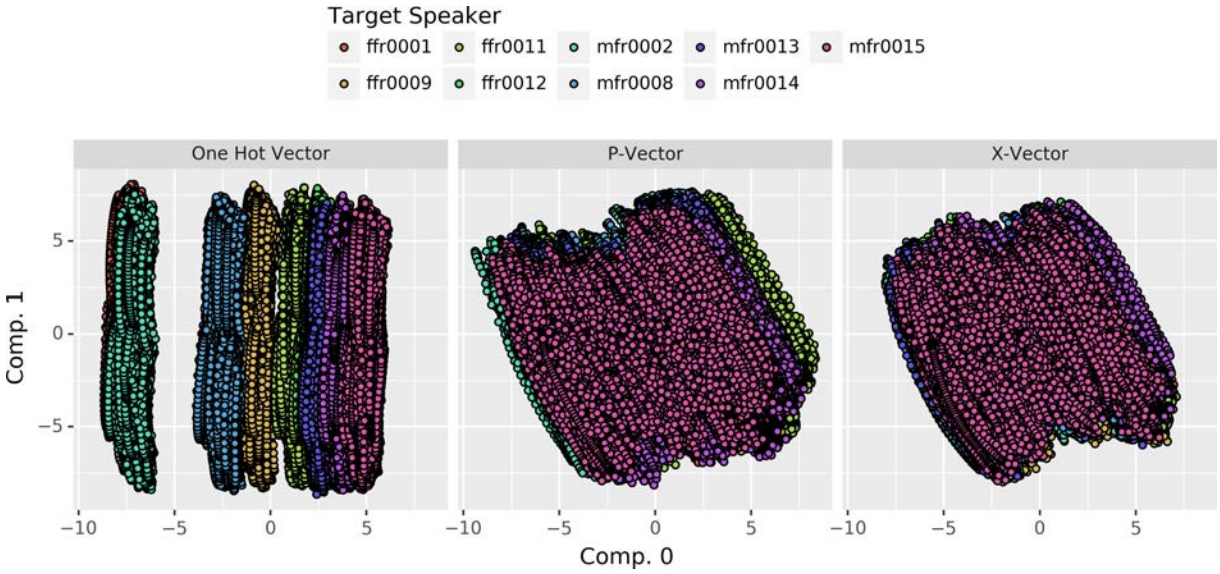
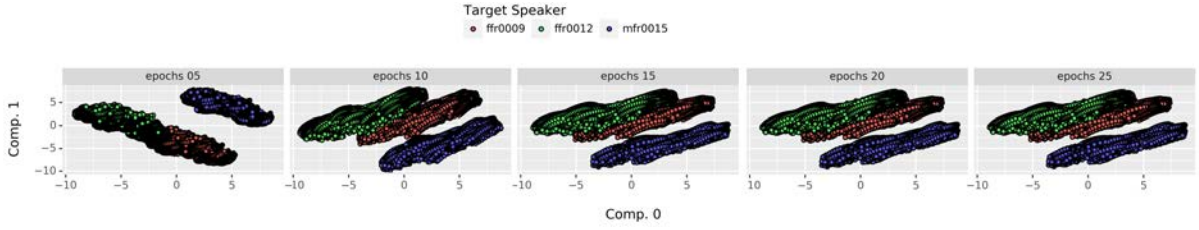


Figure 7.4: Visualization of the latent representation in case of P-Vector using parallel data. We can notice the separation of the speakers representation from epoch 5 to epoch 25.



PCA[3] has been applied on the output of the first hidden layer to reduce the number of dimensions down to the two main ones.

Figure 7.2 and Figure 7.3 illustrate respectively the parallel data and non-parallel data projections for the different configurations. While with the parallel data, it seems that the configuration involving the OneHot-Vector fails to separate the speakers, P-Vector and X-Vector achieve almost the same result and succeed to separate speakers representation. With non-parallel data, both X-Vector and P-Vector do not show a clear separation between speakers compared to OneHot-Vector.

The first explanation for this behavior is that with non-parallel data, the linguistic, prosodic and phonetic context variability are dominant and most of the variation is hold by those components. As the data are non parallel, the neural network has more difficulty to distinguish the speakers. The second possible explanation is that the size and complexity

[3]We choose PCA to find out the independent variables that hold the speaker's identity.

of X-Vector and P-Vector bring more sparsity in the latent space, which is not the case with OneHot-Vector. Finally, it seems that X-Vector and P-Vector can be used equally to bring speaker control to the system but due to the lower complexity of P-Vector, this representation might be preferable.

The visualization of the evolution of the latent space projection at different epochs is illustrated on Figure 7.4. It enables to monitor the learning process and check quickly the impact of the vector structure on the speaker separation. Here, we can notice that from epoch 10, the projected latent space is quite stable and the speakers well separated.

6.3 Subjective Evaluation

Evaluation protocol

In order to validate our proposition, we conducted a subjective evaluation based on the MUSHRA protocol [Series 2014]. The reference is the re-synthesis using **world**. We use a speaker dependent baseline (**spkdep**) as well as a speaker independent model available in [4] (**spkadapt**). Then, we evaluated the isolated configurations (**OneHot-Vector**, **X-Vector** and **P-Vector**).

The duration of each of the 54 samples presented to the listeners is comprised between 4 and 6s. The ratio of speech breaks present in the selected samples does not exceed the quarter of the total duration of the sample.The list of stimulis used for the subjective assessment are in Appendix 2

One evaluation instance is composed by 9 steps including all the models presented before (an example of step is illustated by the Appendix 3). 30 listeners completed the evaluation. They were French native speakers aged between 24 and 45. The majority of them have experience with listening tests but are not necessarily experts in the annotation of audio files. All materials are available in the dedicated repository[5].

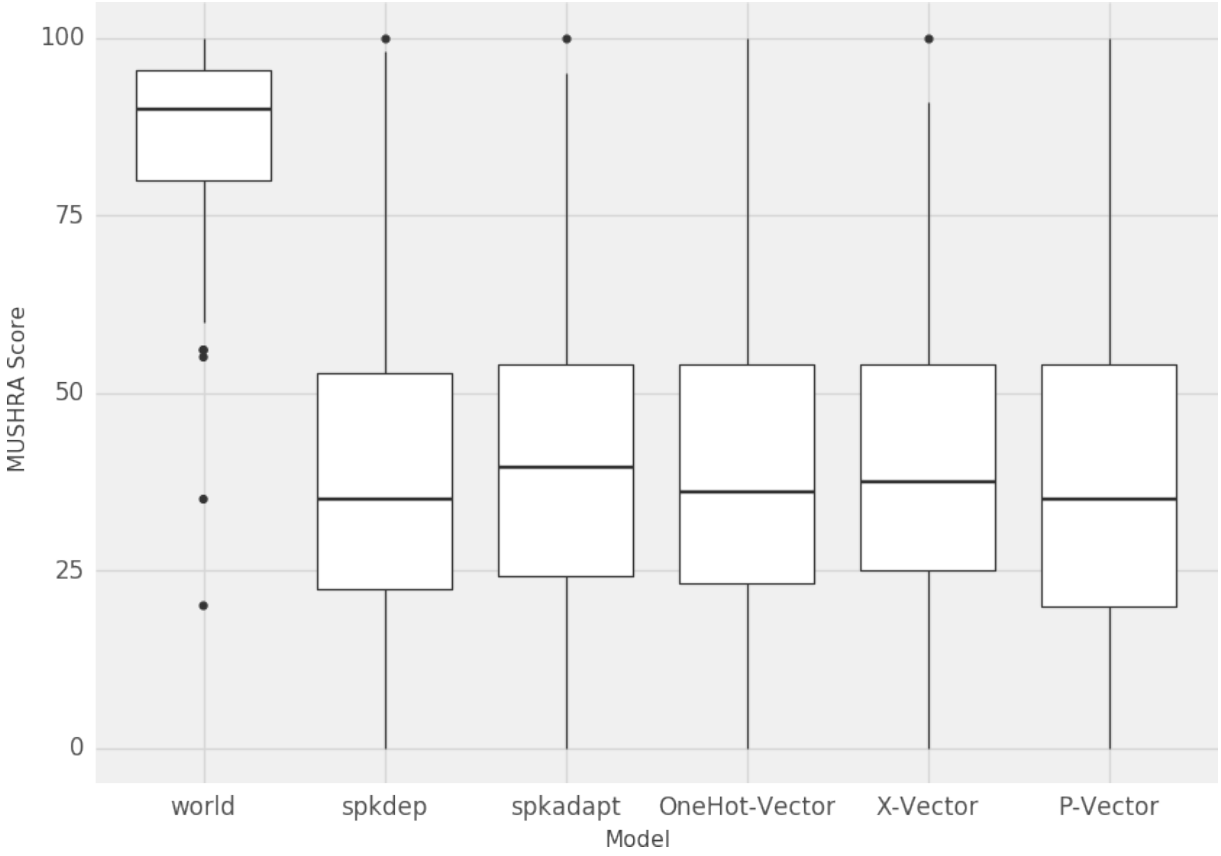
Discussion and results

The results of the evaluation are presented in Figure 7.5. From them, we can see that the reference is correctly identified which guarantees the validity of the evaluation. It seems that some annotators estimate that even the reference was not good enough for some samples which explains the fact that the reference did not achieve a score of 100. Then,

[4]https://github.com/AghilasSini/merlin/tree/master/egs/speaker_adaptation
[5]https://github.com/AghilasSini/SpeechProsody2020

considering the models evaluated, no system is outperforming the other ones. This leads us to conclude that listeners do not distinguish major differences.

Figure 7.5: Result of the MUSHRA of the listening test.



To verify that the listeners didn't perceive minor differences, we also compute the rank of each systems for each step based on its score. Results are presented in Figure 7.6 (whereas Figure 7.7 represent the result related to all speakers).

The reference is still considered in huge majority as system number one. Considering the others, the proportion are globally similar to each other with some variations. This is amplified by the fact that the other systems are often ranked in second position which indicates they are all graded ex-aequo after the reference.

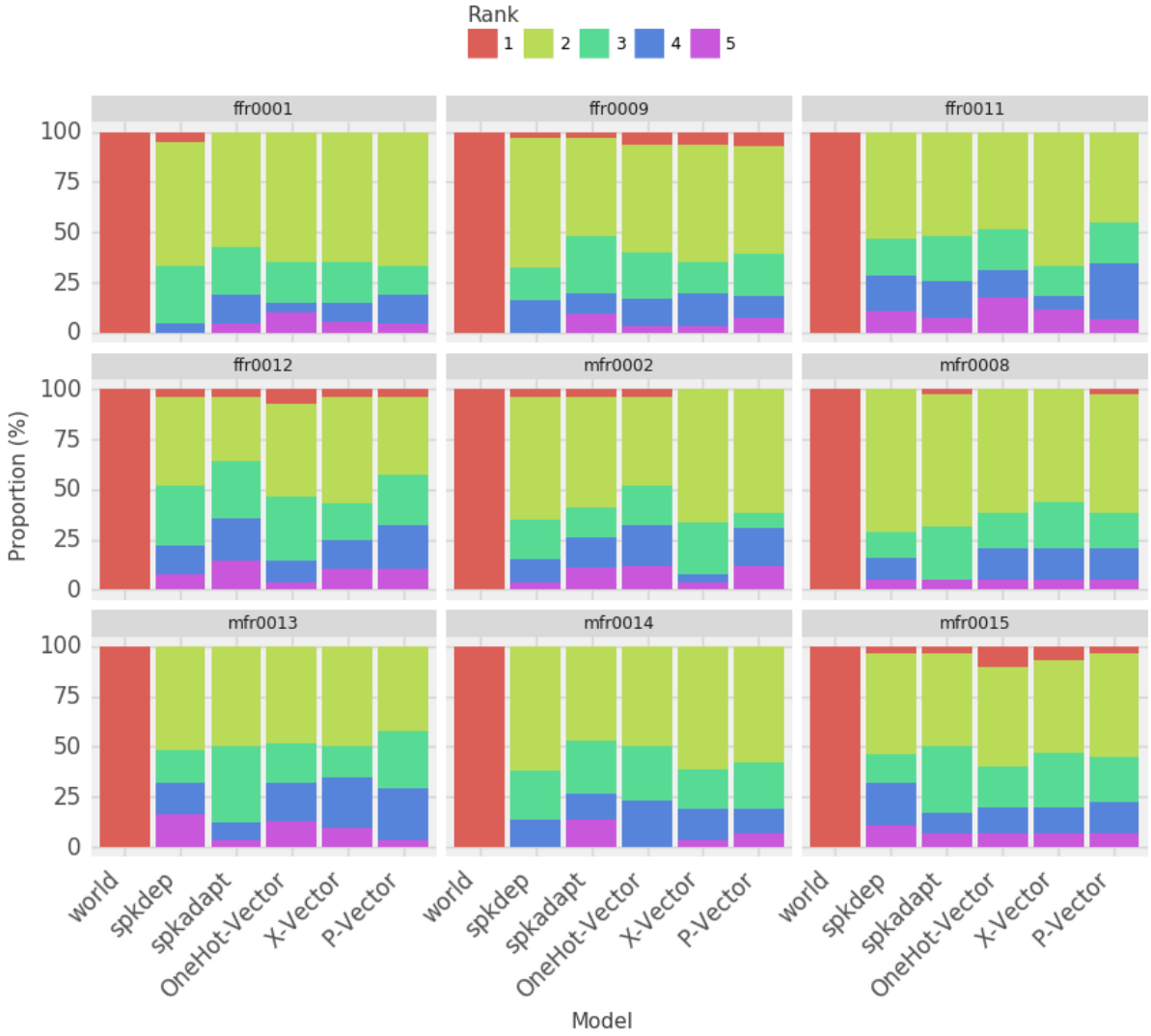
7 Conclusion

In this chapter, we have evaluated different speaker coding scheme both objectively and subjectively in a DNN-based framework. All the evaluations conducted show no difference

Figure 7.6: Ranking score of two representative speakers female (ffr001) and male (mfr0008), the present results are similar for the other speakers.



Figure 7.7: Ranking score of all speakers



in the quality of the modeling of the three different speaker coding schemes. These results are valid in both studied cases, parallel or non-parallel data for multi-speaker modeling. Moreover, the speaker coding scheme we proposed, the P-Vector, provides better control of the modeling. This investigation confirms the relevance of the prosodic parameters that we choose to build the prosodic identity of speakers. However, a close look at this representation shows that the intra-speaker prosodic variation related to discourse changes (narration, dialog) are excluded.

These results are encouraging and suggest further research work. Furthermore, we plan the evaluation of the robustness of the proposed speaker coding on a dataset that contains more speakers and investigating other factors such as language, literary genre, discourse typography, and structure.

General Conclusion

Summary of the Contribution

In this thesis, we have explored speech's expressivity through a particular speech data type, which is audiobooks. We have proposed a complete process for gathering a sizeable French audiobook corpus and annotating it manually or automatically. MUFASA corpus includes twenty French speakers and contains about 600 hours of good continuous speech quality. We have shown in chapter 3 that this collection of amateur's reading is comparable to professional recording ones in terms of prosodic properties. Even though the voice quality of speech of MUFASA corpus is lower than comparable professional corpora due to the recording conditions, the quantity and diversity of data make it possible to explore new spoken speech horizons. We have compared extracts of MUFASA corpus with other well-known French corpora to measure the similarity. As we expected, MUFASA presents high similarity with the BREF corpus, which also a read speech corpus.

In this work, we have articulated the expressivity carried by audiobooks on three pillars: emotions, discourse, speaker. The emotions intervene at specific moments of speech to animate the discourse and bring depth. To structure and bring coherence in the story, the authors use different modes of discourse. In audiobooks, emotions and discourses depend on the text as much on the speech signal. The speech signal depends on the speaker's properties, which constitutes the third axis.

To explore these three pillars, we have investigated the text properties and the prosodic properties of a set of audiobooks read by nine speakers present in the MUFASA corpus. To study the emotional characteristics of data, we focus our effort on the SynPaFlex Corpus voice, representing the initial version of the MUFASA corpus. This database contains a single female speaker. For conducting the experiments, we used a representative extract proportional to the linguistic distribution of the database.

To study the emotional characteristics of data, we proposed to focus on the SynPaFlex corpus, containing a single female speaker, representing the initial version of the MUFASA corpus. For conducting the experiments, we have asked a speech expert annotator to select representative extract and to annotate the speech signal according to two parameters:

discrete emotion labeling and discourse labeling. The annotation process ends with four complementary annotations: intonation patterns, discourse represented by characters pattern, emotions patterns, and others. With emotion patterns annotation, we build a binary classifier, the results of the experiments highlighted the subtility of emotions in such type of data. Based on this observation, we proposed to explore the questions relating to emotions through the analysis of the lexical and semantic properties of transcriptions of audiobooks. To do these experiments, we favored unsupervised approaches. This second experience is based on the techniques of sentiment analysis and natural language processing. The process mainly consists of finding an adequate numeric representation of the texts, we choose the doc2vec model, then clustering the embedded text automatically according to lexico-semantic affinities using the Kmeans algorithm. Once the clusters formed, the last phase consists of interpreting the clusters in the acoustic features space. The results show that there strong correlation between text representation and acoustic speech features. This contribution opens perspectives that we will discuss later.

For studying the discourse, we first built a tool for parsing and annotating audiobooks texts considering three discourses types, namely indirect discourse, direct discourse, and incidental clauses with speech verbs. This tool contains two approaches. The first approach is rule-based consists of a set of rules derived from data analysis and crafted by expert knowledge using morpho-syntactic and typographical properties of the text. The second approach relies on machine learning techniques; we obtained the best result with deep learning models, for highlighting the prosodic properties during discourses changes and how speakers address this phenomenon. We proposed analyzing this phenomenon through a set of prosodic cues derived from InterPausal Unit (IPU) that we consider as pertinent discourse. We experimented with two female speakers of the MUFASA corpus. The results confirmed that the IPU is an adequate speech unit for studying discourse changes; F0-range and inter-IPU pause duration are good indicators of discourse changes.

Concerning the last pilar, speaker voice properties, We explore three speakers configurations, OneHot vector, representing the speaker identity through two parameters speaker gender and identifier, X-vector, this embedded vector derived from the pre-trained speaker recognition model, P-Vector, a new vectorial prosodic representation of voices. We implemented these configurations for guiding a DNN based multi-speaker speech synthesis system. To evaluate these configurations, we conducted two objective evaluations standard objective evaluation described in chapter 3, and objective visual evaluation, consisting of projecting the first hidden layer representation. Furthermore, we investigate a subjec-

tive evaluation to support objective assessment funding. Both objective and subjective assessment has shown that P-vector's prosodic identity is capable of guiding the DNN based multi-speaker speech synthesis system as good as of the well established X-vector and OneHot Vector.

Further Issues

In this section, we briefly present work already started and preliminary results that we obtained.

Does granularity matter in speech synthesis ?

If we consider the three pillars that represent the contribution of this thesis as well as the construction of the MUFASA corpus, we can see that there is a common thread to all of them. In this work, granularity is designated as a discursive unit when it is a textual segment or speech unit when it is a segment of a speech signal.

Most speech synthesis systems are trained to process sentences. In most cases, the sentence is considered as both the discourse unit for processing the text to be analyzed, and the speech unit for training acoustic and duration models. We can easily claim that this seems relevant because most of the databases built for speech synthesis have been recorded in isolated sentences(sentence by sentence).

However, is this unit the best choice when dealing with audiobooks where the original speech records are chapters or paragraphs? Or does it matter? Some studies have looked at the optimal level of granularity to improve speech synthesis systems' expressiveness, especially when it comes to long and coherent texts such as audiobooks. This preliminary work aims to study the discourse/speech unit's effects on learning statistical models on speech synthesis.

We consider two types of units, graphical-based discourse/speech units, represented by sentence, which is the most privileged prosodic unit in speech synthesis systems and the pseudo-paragraph, which represents the largest, and prosodical-based discourse/speech unit, represented by InterPausal Unit (IPU) which is the prosodic unit between two long pauses (pause> = 200ms). To measure each of these units' impact, we rely on the standard objective measures described in chapter 1.

Data and features extraction

To evaluate the effect of the prosodic unit on parametric speech synthesis, we consider an audiobook of 90 minutes, read by an amateur female speaker available in the SynPaFlex-Corpus.

Here is an example of the considered prosodic units extracted from the short story **Boule de Suif**:

Paragraph

(P.1) La Garde nationale qui, depuis deux mois, faisait des reconnaissances très prudentes dans les bois voisins, fusillant parfois ses propres sentinelles, et se préparant au combat quand un petit lapin remuait sous des broussailles, était rentrée dans ses foyers. Ses armes, ses uniformes, tout son attirail meurtrier, dont elle épouvantait naguère les bornes des routes nationales à trois lieues à la ronde, avaient subitement disparu. [*The members of the National Guard, who for the past two months had been reconnoitering with the utmost caution in the neighboring woods, occasionally shooting their own sentinels, and making ready for fight whenever a rabbit rustled in the undergrowth, had now returned to their homes. Their arms, their uniforms, all the death-dealing paraphernalia with which they had terrified all the milestones along the highroad for eight miles round had suddenly and marvelously disappeared.*]

Sentences

(S.1) La Garde nationale qui, depuis deux mois, faisait des reconnaissances très prudentes dans les bois voisins, fusillant parfois ses propres sentinelles, et se préparant au combat quand un petit lapin remuait sous des broussailles, était rentrée dans ses foyers. [*The members of the National Guard, who for the past two months had been reconnoitering with the utmost caution in the neighboring woods, occasionally shooting their own sentinels, and making ready for fight whenever a rabbit rustled in the undergrowth, had now returned to their homes.*]

(S.2) Ses armes, ses uniformes, tout son attirail meurtrier, dont elle épouvantait naguère les bornes des routes nationales à trois lieues à la ronde, avaient subitement disparu. [*Their arms, their uniforms, all the death-dealing paraphernalia with which they had terrified all the milestones along the highroad for eight miles round, had suddenly and marvelously disappeared.*]

IPUs

(IUP.1) La Garde nationale qui, [*The members of the National Guard, who*]
(IUP.2) depuis deux mois, faisait des reconnaissances très prudentes dans les bois voisins, [*for the past two months had been reconnoitering with the utmost caution in the neighboring woods,*]
(IUP.3) fusillant parfois ses propres sentinelles, [*occasionally shooting their own sentinels,*]
(IUP.4) et se préparant au combat quand un petit lapin remuait sous des broussailles, [*and making ready for fight whenever a rabbit rustled in the undergrowth,*]
(IUP.5) était rentrée dans ses foyers. [*had now returned to their homes.*]
(IUP.6) Ses armes, ses uniformes, [*Their arms, their uniforms,*]
(IUP.7) tout son attirail meurtrier, dont elle épouvantait naguère les bornes des routes nationales à trois lieues à la ronde [*all the death-dealing paraphernalia with which they had terrified all the milestones along the highroad for eight miles round,*]
(IUP.8) avaient subitement disparu. [*had suddenly and marvelously disappeared.*]

System training configuration

To achieve training and synthesis, we used the Merlin toolkit[Wu, Watts, and King 2016]. The architecture of the model is a FF-DNN with 4 hidden layers. Each hidden layer have a fixed number of 512 neurons. The hidden layers use the *tanh* activation function and the output layer uses a linear activation function. We applied batch-training paradigm with a batch size of 256. The maximum number of epochs is set to 25 including 10 warm-up epochs. The learning rate is initially set to 0.002 for warm-up epochs and after that reduced by 50% for each epoch. Similarly, the momentum is set to 0.3 for warm-up epochs and to 0.9 otherwise. Finally, we used L2-regularization with a weight set to 10^{-5}. Models are

learned considering different granularity unit schemes.

The input feature vector contains 319 coefficients corresponding to the linguistic description of the utterance. This part is based on the standard feature set for English described in [Tokuda, Zen, and Black 2002] that we have adapted for French. The main differences with the English feature set concerns the accentuation. Indeed, as the accentuation information in French is strongly correlated to the POS information, we therefore consider that the POS information, already present in the vector, is enough to encode the accentuation information.

The output feature vector contains the standard coefficient vector composed by the VUV flag, the $\log F_0$, the MGC, the BAP, and their dynamic counterparts extracted using WORLD [Morise, Yokomori, and Ozawa 2016] vocoder. This leads to a vector of 265 coefficients.

Finally, the input and output vectors are normalized using, respectively, MMN and MVN methods.

Objective evaluation and results

In order to evaluate DNN-based TTS synthesis, the proposed method was applied to train three models using three training set and three test set.

All the models have been evaluated using MCD, BAP distortion, RMSE on F_0 and duration, VUV rate and Correlation (CORR) on F_0 and duration, between the predicted and the original coefficients.

The preliminary results reported in Table 7.2 show that the granularity of data used to build a synthetic voice is essential. According to the present results, the sentence is not always the best choice to build a synthetic voice in a Statistical Parametric Speech Synthesis (SPSS) system. A more in-depth investigation needs to be done with different voices and different audiobooks.

Modern Speech Synthesis Framework (End-to-End (E2E) Paradigm)

During this thesis, we had the opportunity to train and to test advanced techniques based on neural networks such as the WaveNet [Oord et al. 2016] Vocoder for speech generation and Tacotron-2 [Shen et al. 2018] End-to-End framework (the tacotron network architecture is illustrated in Figure E.1). The results are better in terms of quality compared to the architectures presented and used during the thesis work. Nevertheless, these techniques

test/train set granularity	IPU				
measures	MCD	BAP	F0-RMS	F0-CORR	UV
IPU	5.125	0.193	35.040	0.404	6.954
Sentence	5.155	0.193	35.033	0.400	7.188
Paragraph	5.165	0.195	35.233	0.404	7.261
test/train set granularity	Sentence				
IPU	5.102	0.194	33.358	0.511	7.348
Sentence	5.179	0.195	34.253	0.492	7.247
Paragraph	5.126	0.193	33.431	0.510	6.983
test/train set granularity	Paragraph				
IPU	5.358	0.187	33.790	0.496	6.648
Sentence	5.255	0.180	37.093	0.411	6.349
Paragraph	5.128	0.177	34.108	0.472	6.486

Table 7.2: Objective results of the acoustic model, considering the three granularity. Mel-Cepstral Distortion (MCD), Band Aperiodicity Parameter (BAP), Root Mean Square Error (RMSE), Voiced/Unvoiced (VUV) and Correlation (CORR) between the predicted and the original coefficients. For the F_0, RMSE and CORR are computed on the voiced frames only.

have a couple of constraints:

- Amateurs audiobooks are not dedicated to speech synthesis at the origin, so the recordings are not as good in term of quality as the one recorded for synthesis. During the test that we made, we found that these architectures are sensitive to the quality of the data. It thus made difficult to build a robust model with such type of data and with the difficulty to find more data.

- The majority of neural architectures rely on an attention mechanism to align the encoder part with the decoder part. Tests have shown that these mechanisms are fragile and not robust when it comes to long sentences, often present in the audiobooks of SynPaFlex corpus.

- Parameterization: Training these models, many parameters are defined empirically, which makes the training phase tricky.

- The training of the model and the synthesis phase are both highly time consuming.

- Lack of reliable objective evaluation.

For these different reasons in this thesis, we found it more judicious to focus exclusively on the improvement of statistical parametric models based on modular architectures, including distinct Front-End and Back-End parts.

Perspectives

Short-term perspective

Chapter 7 presents a whole process of integrating speaker prosodic identity from the stage of hypothesis to the stage of concrete integration in a realistic speech synthesis system and formal objective and subjective assessment. The two other works concerning emotion pattern and discourse related prosodic cues still at the statistical analytic and objective evaluation stage. So as a short term perspective, we aim to integrate these two variables in the MERLIN [Wu, Watts, and King 2016] toolkit framework relying on the same procedure presented in [Malisz et al. 2017]. Concretely, we would like to insert two new neural network-based modules, one for building a discourse embedded vector and the other one for emotion embedded vector (EEV). Both modules will be inserted after the front-end module. These two modules will be trained before the duration module and acoustic module.

To evaluate these two modules' effects, we are considering two distinct subjective assessment one for each module. To the perceptual discourse assessment, the stimuli are extracts of discourses changes mode (DD, ID, IC), two questions are planned: direct question *"do you notice any changes speech sample?" (yes/no)*, to see if the subject has noticed any changes, then a second question *"What kind of changes you perceive? a) speech rate "fast/slow" b) speech amplitude c) "pause duration shorter/longer."* .A similar evaluation process will be conducted to evaluate the emotion module impact. The stimuli will be the same as those used for assessing the discourse module, but the questions will not be the same. As in this second subjective assessment, the questions will be "do you recognize emotion in this speech sample," if the subject answer yes, a list of considering emotions will be presented followed by the intensity of the perceived emotion or emotions because we suppose that the subject can assign for same sample several emotion labels with different intensity.

Beyond the analysis of the results of each module's respective effect, the combination of the results is also considered a perspective because it allows us to measure the correlation

between discourse and emotion.

Long-term perspective

As a concrete long term perspective, we plan to migrate from the Merlin [Wu, Watts, and King 2016] -SPSS framework to E2E Framework, more precisely Tacotran2 [Wang et al. 2017a; Shen et al. 2018] available in the ESPNET [Hayashi et al. 2020] toolkit, to gain a considerable quality. Then, we plan to build a module similar to the one developed in the short term perspectives. However, in this new configuration we will merge the two modules into a single module relying on deep multi-task neural networks [Liu et al. 2019]. This module will be trained together with the acoustic models.

General Discussion

In this thesis, we addressed prosodic characterizations in case of the synthesis of audiobooks through three dimensions:

- Emotions acted by a speaker to set the story context and provided additional elements for entertaining the listener's attention.

- Discourse typography to highlight the structures of the texts and the correlation with prosodic indices.

- The speaker's identity with the long-term goal of highlighting the reading strategy. Speech has to respect syntactic, semantic, pragmatic constraints as well as related to the written-text discourse typology. In parallel, the speaker strategy and reading identity constraint emotion realization and acting.

The correlation between the three parameters explored in this thesis makes it difficult to build up a robust and reliable expressive speech synthesis system. Disentangling these "three pillars" using factorization techniques based on advanced deep learning algorithms seems to be interesting, according to [Hsu et al. 2019; Mathieu et al. 2016].

Whereas [Brognaux 2015] explores the expressivity through spontaneous speech, we focus on reading written text. It will be interesting to make a comparison between spontaneous, in particular, sports comments and read-text, in particular, audiobooks, to find common representation to expressive speech.

The major findings presented in this thesis are based on an acoustical perspective of speech. This level of representation of prosody is important but not enough to characterize the expressive speech carried by audiobooks. The perceptual representation and linguistic properties of prosody are crucial to have a complete and to validate the results presented in this thesis.

Appendix A

Audiobooks Corpora

1 SynPaFlex Corpus

Genre	**Title, author** (full reading *)		**Duration**	Lower audio quality	Abbr
Historic novels	*Les Mystères de Paris* vol.1 and 2,, Eugène Sue	*	25h 32m	22m	HNm
	Les misérables, Victor Hugo		14h 16m	2h 18m	HNm
	Madame Bovary, Gustave Flaubert	*	13h 12m	26m	HNm
	Le Novel de la momie, Théophile Gautier	*	6h 33m	-	HNrc
	Germinal, Émile Zola		1h 46m	50m	HNge
Fantastic novels and short stories	*La vampire*, Paul Féval	*	10h 02m	2h 26m	FNva
	Voyage au centre de la terre, Jules Verne		1h 52m	22m	FNvc
	La Vénus d'Ille, Prosper Mérimée	*	1h 02m	-	FSve
Adventure novel and short stories	*La fille du pirate*, Maurice Chevalier	*	6h 43m	4h 31m	ANfi
	Carmen, Prosper Mérimée	*	2h 18m	53m	ASca
Symbolism short stories	*Tales cruels*, Auguste Villiers de l'Isle-Adam		1h 42m	24m	SYco
Tales	*La malle volante*, Andersen	*	12m	-	TAan
	Le monstre Yatama, Claudius Ferrand	*	8m	-	TAfx
	Les sept chevreaux, Claudius Ferrand	*	16m	-	TAfy
	Ourashima Taro et la déesse de l'Océan, Claudius Ferrand	*	16m	-	TAfz
	La Hyène, l'Hippopotame et l'Éléphant, Franz de Zeltner	*	11m	-	TAzx

	L'histoire de Koli, Franz de Zeltner *	4m	-	TAzy
Epistolary Novel	*Lettres persanes*, Montesquieu	20m	-	ENle
Fables	*Fables de La Fontaine*	18m	5m	FAfo
Fantastic epic	*Les chants de Maldoror*, Comte de Lautréamont	18m	7m	FEch
Fantastic dramatic	*Infernaliana*, Charles Nodier	11m	-	FDin
Poems	*L'albatros*, Charles Baudelaire *	1m	-	POal
	Chanson d'automne, Paul Verlaine *	1m	-	POch
	Le Dormeur du val, Arthur Rimbaud *	1m	-	POdo
	Fiez vous y !, Charles d'Orléans *	1m	1m	POfi
	Gaudriole en six couplets, unknown *	3m	-	POga
	Un matin, Emile Verhaeren *	1m	-	POma
	Perles, Jean Courdil *	1m	-	POpe
	La veuve indienne, Eugène Fouques *	4m	-	POve
Pamphlet	*Le cerf-volant aux six têtes*, Guillaume Taillerand-Perigord *	6m	6m	PAce

2 SynPaFlex Annotated Subset

SynPaflex corpus French mono-speaker	Whole corpus audio-files duration		Manual annotation subcorpora audio-files duration		
Genres	Whole corpus	Best Quality	Characters	Emotion and Prosody	Phonetic segmentation validation
Historic novels	61h 18m	57h 21m	24h 43m	8h 41m	20m
Fantastic novels and short stories	12h 55m	10h 08m	1h 08	1h 08m	-
Adventure novels and short stories	9h 01m	3h 37m	3h 37m	3h 37m	20m
Symbolist short story	1h 42m	1h 18m	1h 42m		-
Tales	1h 06m	1h 06m	1h 06m	10m	-
Epistolary novel	20m	20m	20m	-	-
Fables	18m	13m	18m	-	1m
Fantastic epic	18m	11m	18m	-	-
Fantastic dramatic	11m	10m	10m	-	3m
Poems	12m	12m	12m	-	1m
Pamphlet	6m	-	6m	-	-
TOTAL	87h 28m	74h 35m	33h 34m	13h 36m	47m

3 MUFASA Corpus

Table A.2: A long table

MUFASA Corpus				
Genre	Title, author, date	Spkid	Spk	Dur (m
Tale	Histoire d'un chien, Alexandre Dumas,1870	FFR0017*	Cocotte	11,62
Short story	L'Enfant des eaux, Jack London,1918	MFR0005*	Alain	24,93
Tale	La Fée des eaux, Alexandre Dumas,1870	MFR0013*	DanielLuttringer	21,96
Tale	La Petite Chienne Blanche, Charles Nodier,1822	FFR0016*	Corinne	11,13
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Tale	La Reine Ysabeau, Auguste de Villiers de L'Isle-Adam,1893	FFR0017*	Cocotte	15,34
Tale	Le tailleur de Catanzaro, Alexandre Dumas,1870	FFR0017*	Cocotte	24,77
Poem	Les cailloux , Gaston Coute,1978	MFR0005*	Alain	1,12
Tale	Roland, de retour de Roncevaux, Alexandre Dumas,1870	FFR0017*	Cocotte	8,55
Novel	Voyage fait en la terre du Brésil, Jean de LÉRY,1578	MFR0019*	Damien Genevois	942,29
correspondance	la Grande Guerre, Ernst Wittefeld,1914	FFR0012*	Victoria	26,64
Short story	A quoi rêvent les pauvres filles, Emile Zola,1870	MFR0003*	Dousset	5,81
Poem	APRES VENDANGES, Gaston Coute,1978	MFR0005*	Alain	3,05
theatre	AUTREFOIS, Charle Cros,1881	MFR0003*	Dousset	6,14
Poem	Alouettes, Saint-Pol-Roux,1901	MFR0003*	Dousset	2,69
Short story	Aventure sans pareille d'un certain Hans Pfaall, Edgar Allan Poe,1835	FFR0007*	Cecile	123,68
Short story	Berthe aux grands pieds, André Rivoire,1899	FFR0011*	Pomme	52,53
Tale	Blanche-Neige, Grimm,1812	MFR0013*	DanielLuttringer	18,26
Short story	Bombard, Guy de Maupassant,1884	FFR0018	Naf	11,37
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Tale	TaleS RAPIDES, François Coppée,1890	FFR0012*	Victoria	12,74
Poem	Carmen, Théophile Gautier,1852	MFR0015*	Jean-LucFischer	1,11
Poem	Ce qu'on entend sur la montagne, Victor Hugo,1856	FFR0011*	Pomme	7,45
Poem	Chanson d'automne, Paul Verlaine,1866	FFR0001	Nadine	0,63
Short story	Claude Gueux, Victor Hugo,1834	MFR0003*	Dousset	75,68
Short story	Coco, Guy de Maupassant,1884	FFR0012*	Victoria	10,97
Short story	Coco, coco, coco frais, Guy de Maupassant,1878	FFR0004*	Julie	8,89
Poem	Complainte des ramasseux d'morts, Gaston Coute,1978	MFR0005*	Alain	4
theatre	Conclusion, Charle Cros,1873	MFR0003*	Dousset	1,31
Short story	Construire un feu, Jack London,1908	MFR0005*	Alain	48,5
Tale	Tales du Sénégal et du Niger, Zeltner,1913	FFR0001	Nadine	14,43
Short story	Tales et Short storys-Berthe, André Rivoire,1884	MFR0005*	Alain	25,47
Novel	Cousin et cousine, HENRY JAMES,1876	MFR0003*	Dousset	155,32
Novel	David Copperfield, Charles Dickens,1850	FFR0012*	Victoria	997,9
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Novel	David Copperfield, Charles Dickens,1850	FFR0012*	Victoria	49,72
Short story	Dernier vœu, Théophile Gautier,1852	MFR0015*	Jean-LucFischer	0,74
Short story	Deux acteurs pour un rôle, Théophile Gautier,1841	MFR0013*	DanielLuttringer	19,77
Short story	En voyage, Guy de Maupassant,1882	MFR0013*	DanielLuttringer	11,83
philosophie	Euthyphron, Platon,399 av. J-C	MFR0003*	Dousset	57,68
Fable	Fables de La Fontaine, De la Fontaine,1668	FFR0001	Nadine	18,09
Tale	Fables et légendes du Japon, Claudius Ferrand ,1903	FFR0001	Nadine	39,37
Short story	Facino Cane, Honore de Balzac,1836	MFR0013*	DanielLuttringer	31,48
Poem	Fiez-Vous-Y, Charles d'Orléan,1450	FFR0001	Nadine	0,6
Novel	Filles, lorettes et courtisanes, Alexandre Dumas,1843	MFR0005*	Alain	163,69
Short story	Gustave Flaubert, Guy de Maupassant,1884	MFR0013*	DanielLuttringer	13,81
Poem	Géorgiques, Virgile,30 av. J.-C.	MFR0005*	Alain	152,5
Short story	Infernaliana, Charles Nodier,1822	MFR0014*	ReneDepasse	8,84
theatre	Inscription, Charle Cros,1908	MFR0003*	Dousset	2,78
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
article	J'accuse, Emile Zola,1898	FFR0009	Ezwa	28,51
Poem	Jean-Luc persécuté, Charles-Ferdinand Ramuz,1908	FFR0011*	Pomme	245,92
Novel	Kéraban-le-Têtu, Jules Verne,1883	FFR0009	Ezwa	698,82
Short story	L' Eau Qui Dort, Amedee Achard,1860	MFR0013*	DanielLuttringer	184,76
Fiction	L' Épouvante, Maurice LEVEL,1908	FFR0009*	Ezwa	308,84
Fiction	L' Épouvante, Maurice LEVEL,1908	MFR0014*	ReneDepasse	370,93
Novel	L'Affaire Charles Dexter Ward, Lovecraft,1941	MFR0015*	Jean-LucFischer	304,13
Poem	L'Albatros, Charles BAUDELAIRE,1861	FFR0001	Nadine	1,12
Novel	L'Appel de Cthulhu, Lovecraft,1926	MFR0015*	Jean-LucFischer	86,83
Poem	L'Art d'être grand-père, Victor Hugo,1877	MFR0006	Bernard	269,26
Short story	L'Enfant, Guy de Maupassant,1882	FFR0012*	Victoria	14,06
theatre	L'Homme propre, Charle Cros,1883	MFR0003*	Dousset	6,79
Poem	L'Homme qui marche, Alain Degandt,2011	MFR0005*	Alain	1,59
Short story	L'Infirme, Guy de Maupassant,1888	MFR0013*	DanielLuttringer	11,93
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Novel	L'art De Payer Ses Dettes, Émile Marco de Saint-Hilaire,1911	FFR0009	Ezwa	143,43
Tale	L'expiation du roi Rodrigue, Alexandre Dumas,1870	FFR0017*	Cocotte	17,72
Tale	LA PORTE DES CENT MILLE PEINES, Anonyme,1918	MFR0013*	DanielLuttringer	14,96
Tale	La Belle au bois dormant, Grimm,1812	FFR0017*	Cocotte	8,37
Short story	La Chambre 11, Guy de Maupassant,1884	FFR0012*	Victoria	15,61
Tale	La Chèvre de Monsieur Seguin, Alphonse Daudet,1887	FFR0017*	Cocotte	14,87
Novel	La Comtesse de Cagliostro, Maurice Leblanc,1924	MFR0002*	Menager	451,03
Short story	La Confession, Guy de Maupassant,1883	FFR0012*	Victoria	14,43
Novel	La Cousine Bette, Honore de Balzac,1846	FFR0007*	Cecile	1006,16
Novel	La Demoiselle aux yeux verts, Maurice Leblanc,1927	MFR0013*	DanielLuttringer	410,27
Novel	La Fille Du Pirate, ÉMILE Chevalier,1878	FFR0001	Nadine	4,64
Short story	La Fille aux yeux d'or, Honore de Balzac,1833	MFR0010	Graigolin	118,92
Short story	La Folie de John Harned, Jack London,1912	MFR0005*	Alain	47,56
Short story	La Main , Guy de Maupassant,1883	FFR0012*	Victoria	15,73
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Short story	La Petite Roque, Guy de Maupassant,1885	FFR0012*	Victoria	88,96
Novel	La Princesse de Montpensier, Madame De Lafayette,1662	FFR0012*	Victoria	72,76
Poem	La Revanche du Passé, Eugénie Pradez,1900	FFR0011*	Pomme	321,69
Tale	La Tarasque, Alexandre Dumas,1870	FFR0017*	Cocotte	21,13
Novel	La Tulipe noire, Alexandre Dumas,1850	FFR0009	Ezwa	21,94
Tale	La Vision du Juge de Colmar, Alphonse Daudet,1880	MFR0013*	DanielLuttringer	9,21
Tale	La chèvre, le tailleur et ses trois fils, Alexandre Dumas,1838	FFR0017*	Cocotte	36,28
Novel	La fille du pirate, ÉMILE Chevalier,1878	FFR0001	Nadine	398,68
Tale	La fée des eaux, Alexandre Dumas,1870	FFR0017*	Cocotte	25,11
Fable	La jeune veuve, De la Fontaine,1668	FFR0017*	Cocotte	3,1
Tale	La jeunesse de pierrot, Alexandre Dumas,1854	FFR0017*	Cocotte	143,93
Short story	La jeunesse de pierrot, Alexandre Dumas,1854	FFR0017*	Cocotte	7,78
theatre	La jeunesse de pierrot, Alexandre Dumas,1854	FFR0017*	Cocotte	4,51
Novel	La maison à vapeur, Jules Verne,1880	FFR0020*	Orangeno	841,35
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Short story	La moustache, Guy de Maupassant,1883	FFR0017*	Cocotte	12,29
Tale	La petite sirene, Alexandre Dumas,1860	FFR0017*	Cocotte	83,81
Tale	La reine des neiges, Alexandre Dumas,1860	FFR0017*	Cocotte	98,22
Tale	La reine des poissons, Gerard de Nerval,1850	FFR0016*	Corinne	7,71
Tale	La sirène du Rhin, Alexandre Dumas,1870	FFR0017*	Cocotte	29,88
Novel	La vampire, Paul Féval,1865	FFR0001	Nadine	602,38
Short story	La vengeance d'une femme, Jules Barbey d'Aurevilly,1883	MFR0002*	Menager	88,03
Short story	Le Bifteck, Jack London,1911	MFR0005*	Alain	48,8
Novel	Le Capitaine Fracasse, Théophile Gautier,1863	FFR0016*	Corinne	1307,65
Novel	Le Cauchemar d'Innsmouth, Lovecraft,1936	MFR0015*	Jean-LucFischer	198,07
Short story	Le Chat noir, Edgar Allan Poe,1843	FFR0012*	Victoria	28,39
Fable	Le Chat, la Bellette, & le petit Lapin, De la Fontaine,1678	FFR0017*	Cocotte	3,04
Novel	Le Dernier des Mohicans, James Fenimore Cooper,1826	MFR0006	Bernard	1036,84
Poem	Le Dormeur du val, Arthur Rimbaud,1870	FFR0001	Nadine	1,15
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Short story	Le Horla, Guy de Maupassant,1868	FFR0018	Naf	223,07
Novel	Le Journal d'une femme de chambre, Octave Mirbeau,1900	FFR0012*	Victoria	804,86
Tale	Le Livre de la jungle, Rudyard Kipling,1894	FFR0017*	Cocotte	329,63
Short story	Le Loup, Guy de Maupassant,1882	FFR0012*	Victoria	12,74
Poem	Le Luneux (Chanson de Colporteur), Anonyme,19éme	MFR0005*	Alain	2,41
Short story	Le Masque, Guy de Maupassant,1889	MFR0013*	DanielLuttringer	18,94
Tale	Le Merle blanc, Henri Carnoy,1879	FFR0016*	Corinne	9,27
Novel	Le Mystère de la chambre jaune, Gaston LEROUX,1907	FFR0018	Naf	98,1
Short story	Le Port, Guy de Maupassant,1889	MFR0013*	DanielLuttringer	15,64
Short story	Le Père Mongilet, Guy de Maupassant,1885	MFR0013*	DanielLuttringer	11,4
Novel	Le Tour du monde en 80 jours, Jules VERNE,1872	MFR0019*	Damien Genevois	402,19
theatre	Le capitaliste, Charle Cros,1884	MFR0003*	Dousset	10,85
Tale	Le cigare de don Juan, Alexandre Dumas,1870	FFR0017*	Cocotte	5,58
Tale	Le dragon des chevaliers de Saint-Jean, Alexandre Dumas,1870	FFR0017*	Cocotte	13,36
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Poem	Le déraillement, Gaston Coute,1978	MFR0005*	Alain	0,68
Short story	Le fifre rouge, Paul_Arene,1887	FFR0017*	Cocotte	10,12
Tale	Le grand mouton noir, Abbe Baudiau,1854	FFR0016*	Corinne	5,82
Novel	Le nez d'un notaire, Edmond About,1862	MFR0008	Didier	166,8
Novel	Le nez d'un notaire, Edmond About,1862	MFR0014*	ReneDepasse	190,61
Short story	Le pont du diable, Alexandre Dumas,1844	FFR0016*	Corinne	14,25
Novel	Le Novel de la momie, Théophile Gautier,1857	FFR0001	Nadine	388,89
Novel	Le Novel de la momie, Théophile Gautier,1857	MFR0014*	ReneDepasse	468,49
Tale	Le vilain petit Canard, Hans Christian Andersen,1876	MFR0013*	DanielLuttringer	20,85
Tale	Le_Conseiller_Krespel, ETA Hoffmann,1967	MFR0005*	Alain	68,85
Poem	Les Amoueuses Trois jours de vendange, Alphonse Daudet,1908	FFR0017*	Cocotte	1,36
Poem	Les Amoureuses Le Rouge Gorge, Alphonse Daudet,1908	FFR0017*	Cocotte	3,64
Poem	Les Amoureuses Le croup, Alphonse Daudet,1908	FFR0017*	Cocotte	3,07
Short story	Les Bords du Sacramento, Jack London,1922	MFR0005*	Alain	22,77
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Novel	Les Chouans, Honore de Balzac,1829	FFR0016*	Corinne	919,58
Tale	Les Deux Chemises, Alexandre Dumas,1870	FFR0017*	Cocotte	11,22
societe	Les Formes élémentaires de la vie religieuse, Émile Durkheim,1912	FFR0004*	Julie	648,09
Poem	Les Mangeux d'terre, Gaston Coute,1978	MFR0005*	Alain	2,73
biographie	Les Rustiques-Un Point d'Histoire, Louis Pergaud,1921	MFR0005*	Alain	18,54
Novel	Les Temps difficiles, Charles Dickens,1854	MFR0003*	Dousset	858,9
Short story	Les Trois Dames de la Kasbah, Pierre_Loti,1884	FFR0011*	Pomme	61,3
Poem	Les accroche-cœurs, Théophile Gautier,1852	MFR0015*	Jean-LucFischer	0,73
Tale	Les aventures du chardon, Hans Christian Andersen,1873	FFR0017*	Cocotte	11,75
Tale	Les deux bossus, Alexandre Dumas,1870	FFR0017*	Cocotte	9,32
Tale	Les onze mille vierges, Alexandre Dumas,1870	FFR0017*	Cocotte	6,93
Short story	Les présents des gnomes, Grimm,1864	MFR0015*	Jean-LucFischer	4,51
Poem	Les tâches, Gaston Coute,1978	MFR0005*	Alain	2,21
Tale	Les voleurs et l'âne, Emile Zola,1864	FFR0017*	Cocotte	39
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Tale	Lettres de mon moulin, Alphonse Daudet,1968	FFR0018	Naf	262,96
Novel	Lettres persanes, Montesquieu,1721	FFR0001	Nadine	20,45
Fable	Livre VI des Fables de La Fontaine, Jean-Pierre Claris de Florian,1668	FFR0018	Naf	1,02
Fable	Livre VIII des Fables de La Fontaine, Jean-Pierre Claris de Florian,1678	FFR0018	Naf	2,52
Novel	Lord of the world, Robert Hugh Benson,1907	FFR0001	Nadine	50,06
Tale	L'Eau de la vie, Grimm,1815	MFR0015*	Jean-LucFischer	11,76
Tale	L'Expérience du docteur Heidegger, Nathaniel Hawthorne,1837	MFR0003*	Dousset	44,07
theatre	L'Homme qui a réussi, Charle Cros,1882	MFR0003*	Dousset	12,94
Tale	L'Oiseau bleu , Madame d'Aulnoy,1697	MFR0003*	Dousset	112,41
Short story	Magnétisme, Guy de Maupassant,1882	MFR0013*	DanielLuttringer	9,55
Poem	Marizibill, Apollinaire,1913	MFR0003*	Dousset	1,51
Novel	Maître du monde, Jules Verne,1904	FFR0020*	Orangeno	337,15
Poem	Message au poète adolescent, Saint-Pol-Roux,1892	MFR0003*	Dousset	1,59
Poem	Miserere de l'amour, Alphonse Daudet,1908	FFR0017*	Cocotte	2,9
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Novel	Monsieur Lecoq, Emile Gaboriau,1869	FFR0001	Nadine	40,1
Novel	Mémoires d'un jeune homme rangé, BERNARD Tristan,1899	FFR0011*	Pomme	359,9
Short story	Notre Dame de la Mort, Arthur Conan Doyle,1910	FFR0016*	Corinne	102,13
Tale	Nouveaux Tales de Fées Pour les Petits Enfants,Comtesse de Ségur,1857	FFR0009	Ezwa	328,39
Short story	Novembre, Gustave Flaubert,1842	FFR0012*	Victoria	49,97
Short story	Novembre, Gustave Flaubert,1842	FFR0012*	Victoria	146,12
Short story	Noël, Théophile Gautier,1872	MFR0015*	Jean-LucFischer	0,74
Poem	Passage du poète, Charles-Ferdinand Ramuz,1923	FFR0011*	Pomme	256,29
Poem	Petit Poucet, Gaston_Coute,1978	MFR0005*	Alain	1,54
Short story	Petite discussion avec une momie, Edgar Allan Poe,1845	FFR0007*	Cecile	38,22
autre	Physiologie du goût, Jean Anthelme Brillat-Savarin,1825	MFR0003*	Dousset	907
histoire	Principes et motifs du plan de Constitution, Nicolas de Condorcet,1793	MFR0003*	Dousset	41,28
Short story	Promenade, Guy de Maupassant,1884	FFR0012*	Victoria	16,07
Novel	Robur le conquérant, Jules Verne,1886	FFR0020*	Orangeno	422,7
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Short story	Révélation magnétique, Edgar Allan Poe,1844	MFR0013*	DanielLuttringer	27,74
Poem	Saoul mais logique, Gaston Coute,1978	MFR0005*	Alain	2,4
Poem	Si le soleil ne revenait pas, Charles-Ferdinand Ramuz,1937	FFR0011*	Pomme	310,56
Short story	Solitude, Guy de Maupassant,1884	FFR0012*	Victoria	13,87
Short story	Sur les chats, Guy de Maupassant,1886	FFR0012*	Victoria	17,85
Poem	Sur un air de reproche, Gaston Coute,1978	MFR0005*	Alain	1,54
Poem	Sur_le_Pressoir, Gaston Coute,1978	MFR0005*	Alain	1,15
Novel	Tartarin de Tarascon, Alphonse Daudet,1872	FFR0009	Ezwa	181,84
Novel	The Guilty River, Wilkie Collins,1886	FFR0001	Nadine	68,59
philosophie	Traité sur la tolérance, Voltaire,1763	MFR0003*	Dousset	259,38
Short story	Un aristocrate célibataire, Arthur Conan Doyle,1892	MFR0013*	DanielLuttringer	45,59
Short story	Un drame dans les airs, Jules Verne,1874	FFR0007*	Cecile	48,4
Poem	Un matin, Emile Verhaeren,III ème siecle	FFR0001	Nadine	1,23
Novel	Une femme, Maurice Leblanc,1893	FFR0011*	Pomme	512,67
Genre	Title, author, date	Spkid	Spk	Dur

Table A.2: (continued)

MUFASA Corpus				
Genre	**Title, author, date**	Spkid ()	Spk	**Dur**
Poem	Vert-Vert ou les Voyages du perroquet de la Visitation de Nevers,Jean Baptiste Gresset,1734	FFR0011*	Pomme	43,25
Novel	Voyage autour du monde , Louis Antoine de BOUGAINVILLE,1769	MFR0019*	Damien Genevois	290
Voyages	Voyage sur l'Amazone, Charles Marie De LA CONDAMINE ,1744	MFR0019*	Damien Genevois	129,23
Voyages	Voyage sur l'Amazone, Charles Marie De LA CONDAMINE ,1744	MFR0019*	Damien Genevois	61,98
Voyages	Voyage à la cime du Mont-Blanc, Horace-Bénédict de SAUSSURE,1787	MFR0019*	Damien Genevois	47,32
Novel	A stange goldfield, Guy Boothby,1904	FFR0018	Naf	15,3
Tale	Tales d'Andersen, Hans Christian Andersen,1835	FFR0001	Nadine	11,98
Novel	La fille, de la fontaine,1678	FFR0017*	Cocotte	3,27
Poem	Le champ de naviots, Gaston Coute,1978	MFR0005*	Alain	2,12
Novel	Mysteries of paris, Eugène Sue,1843	FFR0001	Nadine	25,13
Novel	The vicomte de Bragelonne, Alexandre Dumas,1847	FFR0001	Nadine	40,54
Tale	Un bain,Emile Zola,1893	FFR0017*	Cocotte	21,23
Genre	Title, author, date	Spkid	Spk	Dur

4 MUFASA Parallel Subcorpus

Table A.3: MUFASA Parallel Subcorpus

MUFASA Parallel Subset				
Genre	Title, author, date	Dur	Spkid	Spk
Novel	Albertine Disparue, Marcel Proust, 1925	325,19	FFR0020	Orangeno
		393,26	FFR0011	Pomme
Short story	Boule de Suif, Guy de Maupassant, 1879	90,89	FFR0009	Ezwa
		97,33	FFR0012	Victoria
		82,32	MFR0015	Jean-Luc Fischer
Short story	Carmen, Prosper Mérimée, 1845	137,84	FFR0001	Nadine
		150,57	MFR0014	Rene Depasse
Novel	Cinq Semaines en Ballon, Jules Verne, 1863	576,1	MFR0006	Bernard
		564,57	FFR0020	Orangeno
Tale	Tales Cruels, Auguste de villiers de L'isle Adam, 1883	102,11	FFR0001	Nadine
		113,72	MFR0014	Rene Depasse
Tale	Tales de la Bécasse, Guy de Maupassant, 1883	188,64	MFR0006	Bernard
		65,31	MFR0008	Didier
Fable	Fables, Jean Pierre Clarisde Florian, 1792	180,72	FFR0017	Cocotte
		259,88	FFR0009	Ezwa
Novel	Germinal, Emile Zola, 1885	105,83	FFR0001	Nadine
		123,2	FFR0011	Pomme
Tale	Infernalia, Charles Nodier, 1822	48,98	MFR0010	Graigolin
		10,58	FFR0001	Nadine
Fiction	L'épouvante, Maurice Level, 1908	308,84	FFR0009	Ezwa
		370,93	MFR0014	Rene Depasse
Novel	La comtesse Cagliostro, Maurice leblanc, 1924	166,29	MFR0013	Daniel Luttringer
		159,91	MFR0003	Menager
Novel	La Princesse de Clèves, Madame Lafayette, 1678	336,33	FFR0017	Cocotte
		407,85	FFR0011	Pomme
Short story	La vengeance d'une Femme, Jules Barbey d'Aurevilly, 1874	32,56	MFR0003	Menager
		44,6	MFR0014	Rene Depasse
Short story	La Vénus D'Ille, Posper Mérimée, 1837	61,74	FFR0001	Nadine
		74,76	MFR0014	Rene Depasse
Genre	Title, author, date	Dur	Spkid	Spk

Table A.3: (continued)

MUFASA Parallel Subset				
Genre	**Title, author, date**	**Dur (min)**	**Spk**	**Spkid**
Fiction	La petite comtesse, Octave Feuillet, 1857	186,26	MFR0013	Daniel Luttringer
		205,81	FFR0011	Pomme
		224,64	MFR0014	Rene Depasse
Novel	Le nez d'un Notaire, Edmond About, 1862	166,8	MFR0008	Didier
		190,61	MFR0014	Rene Depasse
Novel	Le Novel de la Momie, Théophile Gautier, 1857	388,89	FFR0001	Nadine
		468,49	MFR0014	Rene Depasse
Poem	Les chants de Maldoror, Comte de Lautréamont, 1869	23,13	FFR0018	Naf
		97,61	FFR0001	Nadine
Tale	Les Milles et une nuit, Anonyme, X siècle	398,96	MFR0015	Jean-Luc Fischer
		153,8	FFR0007	Cecile
Novel	Les Misérables, Victor Hugo, 1862	849,49	MFR0008	Didier
		856,21	FFR0001	Nadine
		68,77	FFR0018	Naf
Novel	Les mystères de Paris, Eugène Sue, 1843	998,42	MFR0013	Daniel Luttringer
		1531,55	FFR0001	Nadine
Novel	Madame Bovary, Gustave Flaubert, 1857	791,69	FFR0001	Nadine
		784,18	FFR0012	Victoria
Novel	Raison et sensibilité, Jane Austen, 1857	918,54	FFR0007	Cecile
		849,85	MFR0013	Daniel Luttringer
Short story	Un coeur simple, Gustave Flaubert, 1877	95,51	MFR0003	Dousset
		87,65	MFR0014	Rene Depasse
Novel	Vingt mille lieues sous les mers, Jules Verne, 1870	111,71	FFR0001	Nadine
		901,52	MFR0019	Damien Genevois
Novel	Voyage au centre de la terre, Jules Verne, 1864	111,64	FFR0001	Nadine
		1209,74	MFR0019	Damien Genevois
Genre	Title, author, date	Dur	Spkid	Spk

Appendix B

Data visualization and high dimension reduction

1 Principal Component Analysis (PCA)

In this section, a brief procedural description of PCA is provided. More detailed theoretical description is directed to [**bishop2006pattern**; **hastie2009elements**; **rogers2016first**]. Assume that we are given by a m-by-n data matrix X consists of n number of m-dim vectors $\overrightarrow{x_i} \in \mathbb{R}$.

Step 1: Compute mean and covariance of data matrix

The covariance matrix of X is called $S \in \mathbb{R}^{m \times m}$ and defined by

$$S = \frac{1}{n} \sum_{i=1}^{n} (\overrightarrow{x_i} - \overline{x})(\overrightarrow{x_i} - \overline{x})^T$$

where $\overline{x} \in \mathbb{R}^m$ is the mean of each row in X and defined by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} \overrightarrow{\mathrm{x_i}}.$$

Step 2: Singular Vector Decomposition (SVD)

SVD of S is implemented to extract principal components and directions:

$$S = U \Sigma V^T$$

where $U \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{n \times m}$, and $V \in \mathbb{R}^{m \times m}$. In the implementation, we use the matrix $V = [u_1, u_2 \dots u_m]$ where a vector represents a principal component direction.

Step 3: Projection

The data matrix X can be projected into a new matrix $Y \in \mathbb{R}^{k \times m}$ by multiplying a matrix P^T

$$Y = P^T X$$

where $P = [u_1 u_2 \dots u_k]$, $k \leqslant m$. Proper number of principal components k should be selected in prior to perform projection of data matrix.

Appendix C

Discourses Annotation

1 Speech Verbs List

#Neutre	# Argumentation	# Désaccord
affirmer	alléguer	accuser
affranchir (apprendre à qqun)	apprendre (à quelqu'un)	combattre
apprendre	arguer	contester
assurer	argumenter	contredire
aviser	assener	critiquer
commenter	assurer	démentir
considérer	avancer	dénoncer
conter	(se) dédouaner	discuter
déclarer	(se) défendre	douter
décrire	détailler	huer
dire	distinguer	infirmer
émettre (un son)	égrener	(s')insurger
exprimer	émettre(une opinion)	nier
formuler	énumérer	(s')offusquer
narrer	exagérer	protester
observer	exposer	rectifier
parler	faire miroiter	remettre en question
penser tout haut	faire remarquer	renâcler
préciser	garantir	reprendre (contredire)
raconter	glisser	rétorquer
remarquer	indiquer	riposter
rappeler	innocenter	tempérer
(se) souvenir	insinuer	# Enigme

Echange

bavarder
confier
converser
deviser
dialoguer
discourir
papoter
parler
saluer

Question

demander
interroger
questionner
(s')enquérir
(s')instruire

Réponse

éluder
expliquer
indiquer
répliquer
répondre

#Promesse

jurer
mentir
promettre

Déroulement du dialogue

achever
(s')adresser
ajouter
arrêter
compléter
conclure
couper
insister
intercéder
inventorier
juger
lister
mettre en garde
minimiser
plaider
présenter
rajouter
rappeler
rapporter
récapituler
requérir
résumer
révéler
signaler
souligner
soutenir
tenter de convaincre

Accord

accorder
acquiescer
adhérer
admettre
approuver
capituler
composer
concéder
confirmer
croire
choisir
féliciter
flatter
avancer
deviner
énoncer
estimer
examiner
imaginer
hasarder
jauger
proposer
supputer

Réticence / regret

admettre
(se) décider
lâcher
regretter

Demanander une faveur

adjurer
demander
exhorter
implorer
négocier
parlementer
pleurer
prier
quémander
réclamer
revendiquer
solliciter
suggérer
supplier

Permission

accepter
encourager
permettre

entamer
entrer en matière
finir
interrompre
intervenir
poursuivre
répéter
répondre
reprendre la parole
terminer

Moquerie

ironiser
(se) moquer
narguer
persifler
railler

Honte

avouer
confesser
(s')excuser
marmonner
souffler

#Tristesse / douleur

compatir
geindre
gémir
(s')inquiéter
(se) plaindre
rassurer

Surprise

Surprise
(s')etonner
(s')étouffer
(s')exclamer
(s')incliner
louer (faire un compliment)
obtempérer
opiner
préférer
réaliser
reconnaître
renchérir
réviser (son opinion)
souscrire

#Humour

Humour
badiner
blaguer
éclater de rire
(s')esclaffer
glousser
gouailler
plaisanter
pouffer
(se) réjouir
rire
sourire

#Volume

acclamer
appeler
beugler
brailler
bramer
clamer
crier
(s')égosiller
héler
hurler
proposer

Interdiction

interdire
prohiber
refuser
résister

#Exigence

ger
intimer (quelqu'un de parler)
obliger
ordonner
sommer

#Façon de parler

ânonner
articuler
babiller
bafouiller
balbutier
balbutier
baragouiner
bégayer
bredouiller
cafouiller
chantonner
couiner
crachoter
crépiter
débiter
déclamer
dégoiser
entonner
épeler
éternuer
faire

manquer de ..

#Colère

Colère
aboyer
apostropher
bougonner
cracher
(s')enflammer
(s')emporter
(s')étrangler
enguirlander
exploser
grincer
grogner
grommeler
gronder
(s')impatienter
injurier
insulter
piaffer (d'impatience)
proférer (des menaces)
râler
réprimander
ronchonner
siffler
s'offusquer
tempêter
tonner
vilipender
vitupérer
vociférer
vomir des injures

rugir
chuchoter
murmurer

#Tentative

essayer
(se) lancer
risquer
tenter
vérifier

#Hésitation

Hésitation
décider
hésiter
risquer

haleter
miauler
minauder
postillonner
psalmodier
susurrer

#Pédant

annoter
commenter
disserter
fanfaronner
(se) gargariser
(se) glorifier
gloser
monologuer
palabrer
pérorer
philosopher
plastronner
pontifier
prophétiser
rabâcher
réciter
seriner
soliloquer
traduire

Appendix D

Manual Annotation and Subjective Assessment Materials

1 Intonation Patterns

1.1 Exclamation pattern

The Figure D.1 illustrate a typical example of the exclamation intonation pattern, the pitch contour of this pattern is similar to the one define in Figure 4.1 work labeled as question pattern.

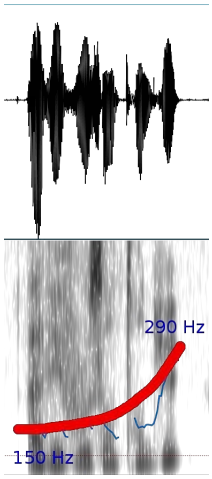


Figure D.1: *Avez-vous entendu ?*

1.2 Nopip pattern

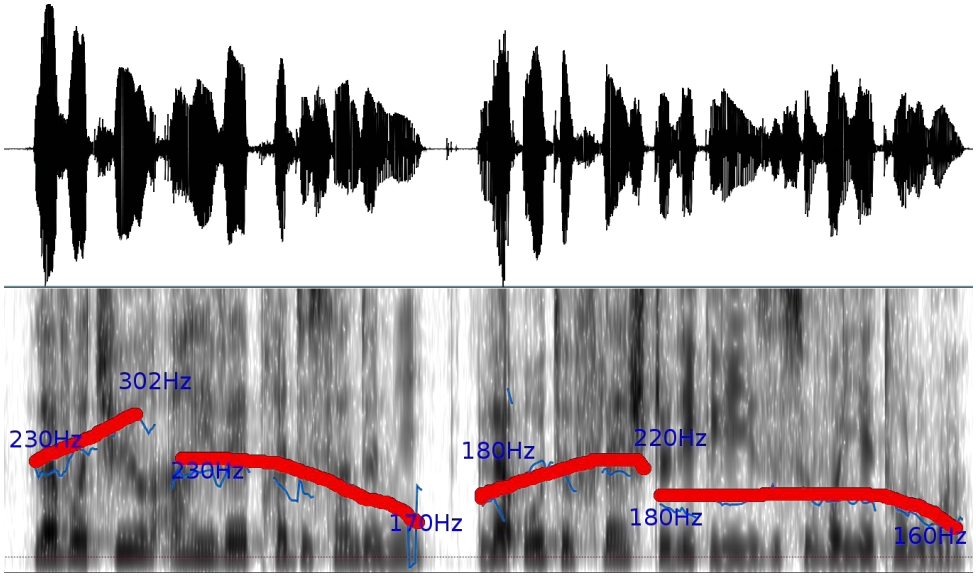


Figure D.2: La voiture arrivait près de Saint-Denis, la haute flèche de l'église se voyait au loin.

1.3 Nuance pattern

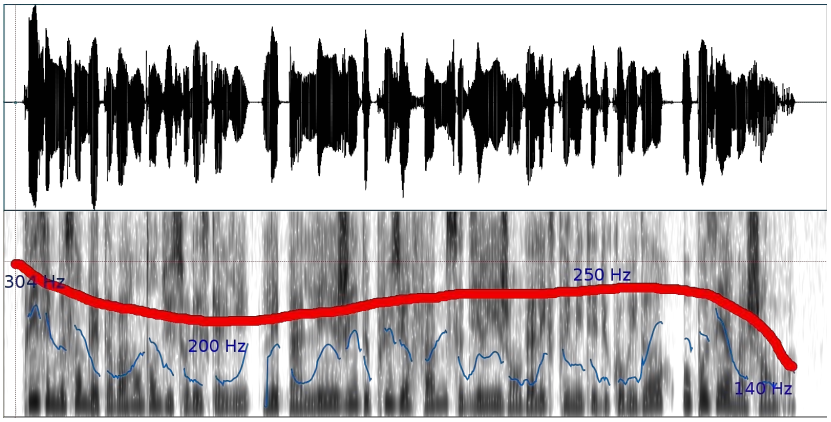


Figure D.3: Nuance Intonation Pattern Example *puis il me semblait avoir entendu sur l'escalier les pas légers de plusieurs femmes se dirigeant vers l'extrémité du corridor opposé à ma chambre.*

1.4 Resolution pattern

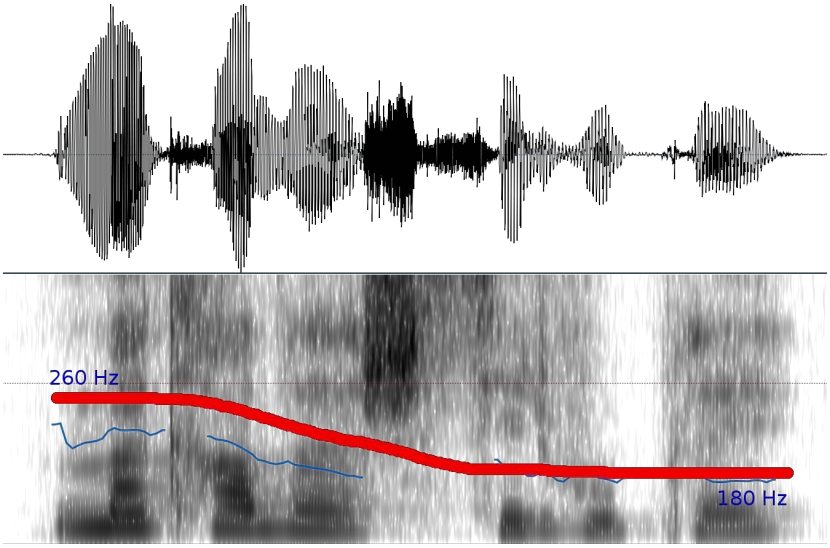


Figure D.4: −− Ma cravache, s'il vous plaît

1.5 Suspense pattern

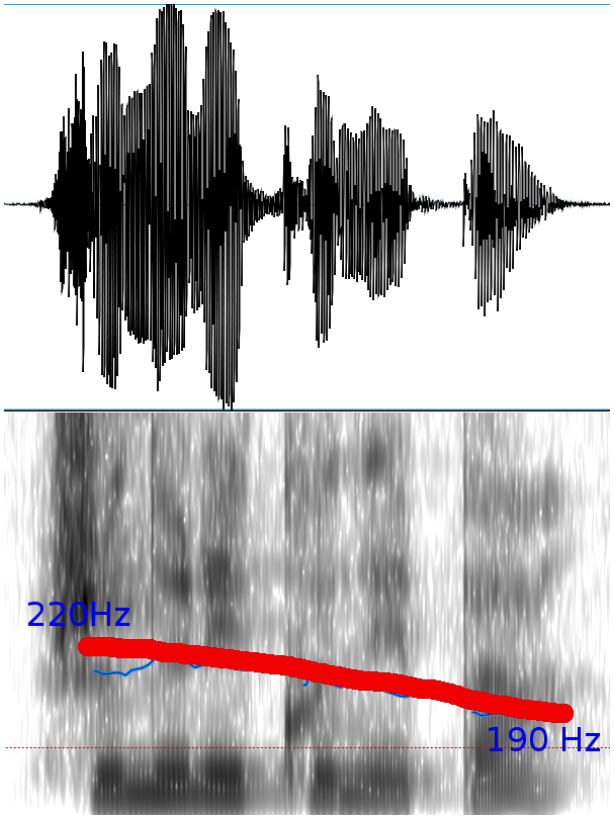


Figure D.5: – *Je ne les connais pas*

1.6 Note pattern

Note intonation pattern which correspond to note, chapter introduction and conclusion is often assigned with flat with quasi null slope pattern as shown in the Figure D.6, this pattern can be assimilated to neutral pattern.

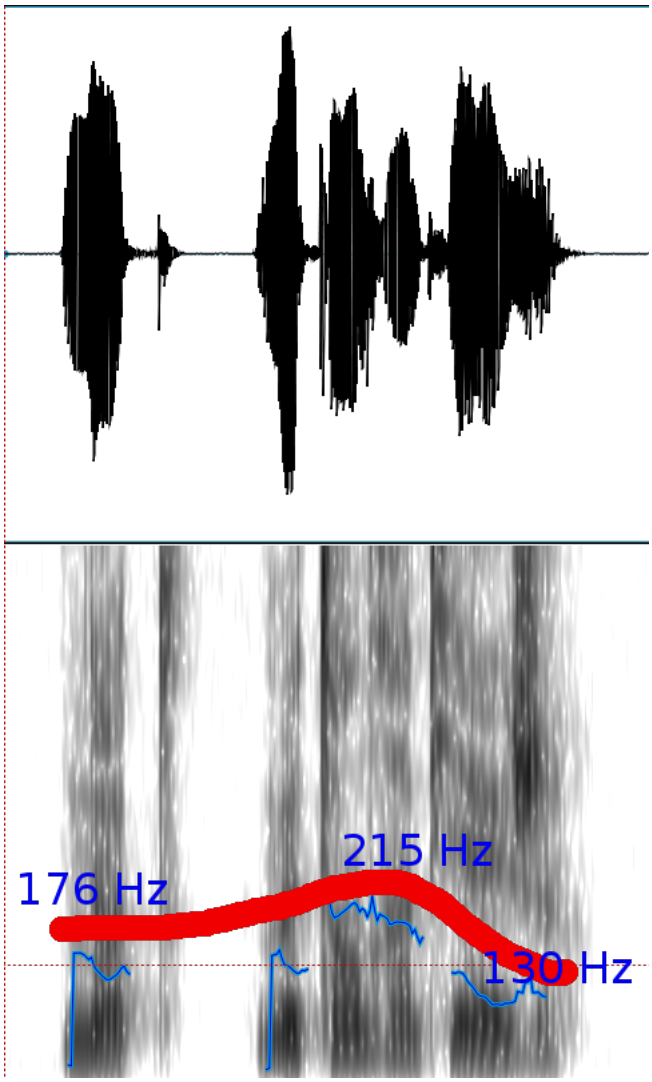


Figure D.6: *[Note : me tendre un piège.]*

1.7 Singing pattern

Whereas, singing pattern distinguishable from the rest with a cyclic pitch contour as shown in the Figure D.7

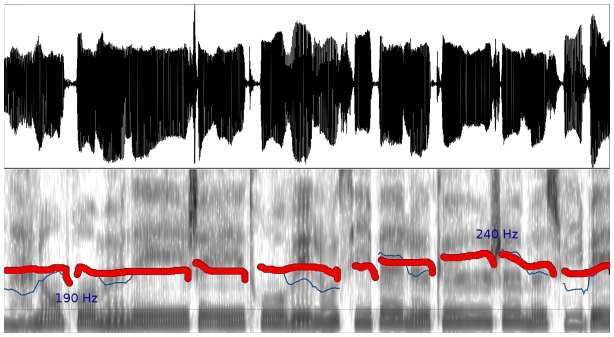


Figure D.7: *...M'en allant promener, J'ai trouvé l'eau si belle Que je me suis baigné...*

2 List of stimulis

(1) Cela venait du fond, et s'arrêta court dans les bas-côtés de l'église. [*It came from the back, and stopped short in the aisles of the church.*]

(2) Les six hommes, trois de chaque côté, marchaient au petit pas et en haletant un peu. [*The six men, three on each side, walked at a slow pace and panting a little.*]

(3) Les hommes continuèrent jusqu'en bas, à une place dans le gazon où la fosse était creusée. [*The men continued all the way down to a spot in the grass where the pit was dug.*]

(4) Enfin on entendit un choc ; les cordes en grinçant remontèrent. [*At last a shock was heard; the squeaking ropes went up.*]

(5) Les Français ne sortaient guère encore, mais les soldats prussiens grouillaient dans les rues. [*The French were hardly out yet, but the Prussian soldiers were swarming in the streets.*]

(6) Les habitants payaient toujours ; ils étaient riches d'ailleurs. [*The inhabitants were still paying; they were rich by the way.*]

(7) Les quatre femmes marchaient devant, les trois hommes suivaient, un peu derrière. [*The four women marched in front, the three men followed, a little behind.*]

(8) Alors on parla de lui, de sa tournure, de son visage. [*Then they talked about him, about his appearance, about his face.*]

(9) Ce jeune homme était à cheval : deux amis et deux dames l'accompagnaient. [*This young man was on horseback: two friends and two ladies accompanied him.*]

(10) Le choix des armes appartenait, sans aucun doute possible, à notre adversaire. [*The choice of arms was, without a doubt, the opponent's.*]

(11) Il fut effectivement résolu, et la rencontre fut fixée au lendemain neuf heures. [*It was indeed resolved, and the meeting was set for the following day at nine o'clock.*]

(12) À dix heures, il se retira, et je vis encore de la lumière chez lui deux heures plus tard.[*At ten o'clock he withdrew, and I saw the light at his house two hours later.*]

(13) Elle était là, devant lui, étendue sur le dos, au milieu de la route. [*She was there, in front of him, lying on his back in the middle of the road.*]
(14) Il imaginait qu'elle était partie en voyage, bien loin, depuis longtemps. [*He imagined that she had been away on a journey, far away, for a long time.*]
(15) Les porteurs glissèrent leurs trois bâtons sous la bière, et l'on sortit de l'église. [*The porters slipped their three sticks under the beer, and we left the church.*]
(16) Les six hommes, trois de chaque côté, marchaient au petit pas et en haletant un peu. [*The six men, three on each side, walked at a slow pace and panting a little.*]
(17) En une seconde il se débarrassa de la robe et du chapeau, et les jeta au milieu des fourrés. [*In a second he got rid of the robe and hat, and threw them into the thickets.*]
(18) Ce fut le sourire qui le premier apparut, hésitant, timide comme un rayon de soleil hivernal. [*It was the smile that first appeared, hesitant, shy as a winter sunbeam.*]
(19) Les bords en étaient guillochés, la plaque d'or par derrière toute meurtrie de coups. [*The edges were guilloché, the gold plate from behind any bruises.*]
(20) Ils arrivèrent ainsi sur un terre-plein, et tout près de la péniche que masquait encore un rideau de saules. [*They thus arrived on a terrace, and very close to the barge that was still hidden by a curtain of willows.*]
(21) Il lui donna le logement de son propre valet de chambre, pour l'avoir plus près de lui. [*He gave it the lodging of his own valet, to have it closer to him.*]
(22) Durant un mois, il remplit les fonctions de garde-malade et passa même plusieurs nuits. [*For a month he acted as a nurse's warden and even spent several nights.*]
(23) Il voulut s'expliquer ; la parole lui mourut dans la gorge. [*He wanted to explain himself; the word died in his throat.*]
(24) Mais son nez n'était plus là, et le mouchoir de batiste ne rencontra que le vide. [*But his nose was no longer there, and the batiste handkerchief met only emptiness.*]

(25) Deux heures se passèrent dans l'agitation, le désordre et le bruit. [*Two hours went by in agitation, disorder and noise.*]
(26) La mer très calme, sans la moindre vague, en baignait les quilles. [*The sea was very calm, with no waves at all, and the keels were bathed in it.*]
(27) Cela parut leur rendre un peu de courage, car ils se relevèrent brusquement, avides d'en finir. [*This seemed to give them back some courage, for they rose up suddenly, eager to finish it all off.*]
(28) Il courut jusqu'au volet et l'attira vers lui, emplissant ainsi le grenier de lumière. [*He ran to the shutter and drew it towards him, filling the attic with light.*]
(29) Il descendit, chercha dans le verger, fouilla la plaine voisine et le chemin. [*He went downstairs, searched the orchard, searched the nearby plain and the path.*]
(30) Il lui donna le logement de son propre valet de chambre, pour l'avoir plus près de lui. [*He gave him the lodging of his own valet, to have him closer to him.*]
(31) Deux heures se passèrent dans l'agitation, le désordre et le bruit. [*Two hours passed in the bustle, disorder and noise.*]
(32) Il arriva pourtant, et comprit à première vue que Romagné était mort. [*He arrived, however, and understood at first sight that Romagna was dead.*]
(33) Quelques amis, bons vivants, égayèrent sa retraite. [*A few friends, bon vivants, brightened his retreat.*]
(34) ... les chevaux restaient à l'écurie, le cocher demeurait invisible. [*... the horses stayed in the stable, and the coachman was invisible.*]
(35) La conversation fut vive, enjouée, pleine de traits. [*The conversation was lively, cheerful, full of features.*]
(36) Le lendemain, un clair soleil d'hiver rendait la neige éblouissante. [*The next day, a clear winter sun made the snow dazzling.*]
(37) Elle restait droite, le regard fixe, la face rigide et pâle, espérant qu'on ne la verrait pas. [*She remained upright, her gaze fixed, her face rigid and pale, hoping she would not be seen.*]

3 Subjective Assessment Platform

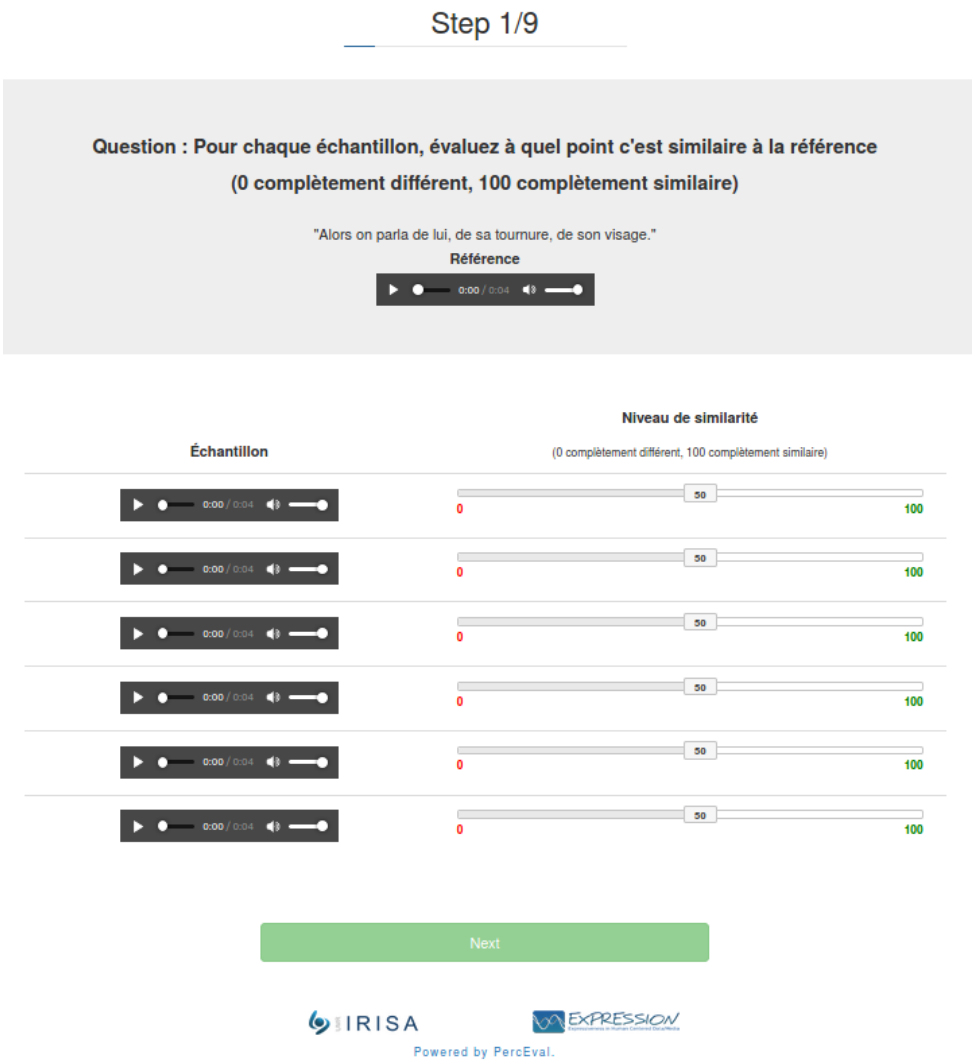


Figure D.8: Screenshot of the platform PercEval (Recently renamed FlexEval [Fayet et al. 2020]) used for collecting the subjective assessment of the participants. *Question: asked question was: " For each sample, evaluate how similar it is to the reference (0 completely different, 100 completely similar)"*

Appendix E

Futur Work

1 End-to-End Tacotran-2 Architecture

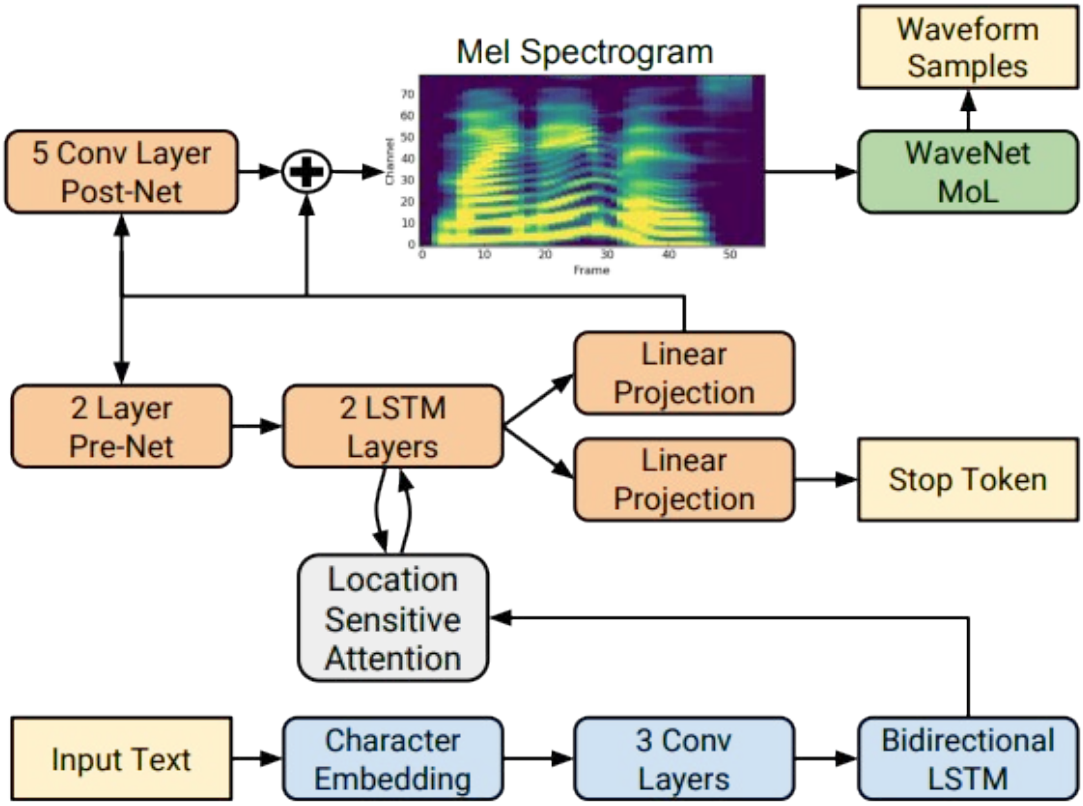


Figure E.1: Block diagram of Tacotran-2 [Shen et al. 2018; Oord et al. 2016] architecture

Bibliography

Abdaoui, Amine et al. (2017). "Feel: a french expanded emotion lexicon". In: *Language Resources and Evaluation* 51.3, pp. 833–855.

Abrilian, Sarkis et al. (2005). "Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces". In: *HCI International*. Vol. 401, pp. 407–408.

Alain, Pierre et al. (Sept. 2016). "The IRISA Text-To-Speech System for the Blizzard Challenge 2016". In: *Blizzard Challenge 2016 workshop*. Cupertino, United States.

Armano, Giuliano and Mohammad Reza Farmani (2014). "Clustering Analysis with Combination of Artificial Bee Colony Algorithm and k-Means Technique". In: *International Journal of Computer Theory and Engineering* 6, pp. 141–145.

Audibert, Nicolas and Cécile Fougeron (2012). "Distorsions de l'espace vocalique: quelles mesures? Application à la dysarthrie (Distortions of vocalic space: which measurements? An application to dysarthria.)[in French]". In: *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP*, pp. 217–224.

Avanzi, Mathieu (2013). "Notes de recheche sur l'accentuation et le phrasé prosodique à la lumière des corpus de français". In: *Revue Tranel (Travaux neuchâtelois de linguistique)* 59, pp. 5–24.

Avanzi, Mathieu, Anne Catherine Simon, and Brechtje Post (2016). "La prosodie du français: accentuation et phrasé". In: *Langue française* 3, pp. 5–10.

Avanzi, Mathieu et al. (2010). "C-prom: An annotated corpus for french prominence study". In: *Speech Prosody 2010-Fifth International Conference*.

Avanzi, Mathieu et al. (2014). "Towards the adaptation of prosodic models for expressive text-to-speech synthesis". In: *Fifteenth Annual Conference of the International Speech Communication Association*.

Baker, Rachel E and Ann R Bradlow (2009). "Variability in word duration as a function of probability, speech style, and prosody". In: *Language and speech* 52.4, pp. 391–413.

Beller, Grégory and Aurélien Marty (2006). "Talkapillar: outil d'analyse de corpus oraux". In: *Rencontres Jeunes Chercheurs de L'Ecole Doctorale* 268, pp. 97–100.

Beyssade, Claire (2012). "Le statut sémantique des incises et des incidentes du français". In: *Langages* 2, pp. 115–130.

Bian, Yanyao et al. (2019). “Multi-reference Tacotron by Intercross Training for Style Disentangling, Transfer and Control in Speech Synthesis”. In: *arXiv preprint arXiv:1904.02373*.

Bird, Steven and Jonathan Harrington (2001). “Speech annotation and corpus tools”. In: *Speech Communication* 33.1, pp. 1–4.

Boersma, Paul and David Weenink (2016). *Praat: Doing phonetics by computer.[Computer program]. Version 6.0. 19*.

Bonami, Olivier and Danièle Godard (2008). “Syntaxe des incises de citation”. In: *Actes du premier Congrès Mondial de Linguistique Française*. France, pp. 2395–2408.

Brognaux, Sandrine (2015). “Expressive speech synthesis : research and system design with hidden Markov models”. PhD thesis. Louvain School of Engineering (UCL) - Faculty of Engineering (UMons).

Brognaux, Sandrine, Benjamin Picart, and Thomas Drugman (2013). “A new prosody annotation protocol for live sports commentaries.” In: *INTERSPEECH*, pp. 1554–1558.

Brognaux, Sandrine et al. (2012). “Train&Align: A new online tool for automatic phonetic alignment”. In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 416–421.

Burdin, Rachel Steindel and Cynthia G. Clopper (2015). “Phonetic reduction, vowel duration, and prosodic structure.” In: Glasgow, Scotland.

Büring, Daniel (2016). *Intonation and meaning*. Oxford University Press.

Burkhardt, Felix et al. (2005). “A database of German emotional speech”. In: *Ninth European Conference on Speech Communication and Technology*.

Buurman, H. A. (2007). “Virtual Storytelling: Emotions for the narrator”. PhD thesis.

Buvet, Pierre-André (2012). “Traitement automatique du discours rapporté”. In: *Actes du colloque JADT 2012*.

Campbell, Nick (2006). “Conversational speech synthesis and the need for some laughter”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4, pp. 1171–1178.

Candito, Marie et al. (June 2009). “Analyse syntaxique du français : des constituants aux dépendances”. In: *16e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2009*. Senlis, France.

Candito, Marie et al. (2010). “Benchmarking of statistical dependency parsers for french”. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pp. 108–116.

Cerisara, Christophe, Odile Mella, and Dominique Fohr (Sept. 2009). “JTrans, an open-source software for semi-automatic text-to-speech alignment”. In: *Proceedings of the*

10th Annual Conference of the International Speech Communication Association - Interspeech 2009. Brighton, United Kingdom.

Chaffar, Soumaya and Diana Inkpen (2011). "Using a heterogeneous dataset for emotion analysis in text". In: *Canadian conference on artificial intelligence*. Springer, pp. 62–67.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3, pp. 1–27.

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

Chateau, Noel, Valerie Maffiolo, and Christophe Blouin (2004). "Analysis of emotional speech in voice mail messages: The influence of speakers' gender". In: *Eighth International Conference on Spoken Language Processing*.

Chevelu, Jonathan, Gwénolé Lecorvé, and Damien Lolive (2014). "ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections". en. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykavik, Iceland.

Christodoulides, George (2016). "Effects of cognitive load on speech production and perception". PhD thesis. UCL-Université Catholique de Louvain.

Clark, Robert AJ and Kurt E Dusterhoff (1999). "Objective methods for evaluating synthetic intonation." In: *International Speech Communication Association*.

Cohen, Henri, Josée Douaire, and Mayada Elsabbagh (2001). "The role of prosody in discourse processing". In: *Brain and cognition* 46.1-2, pp. 73–82.

Cohen, William W. (1995). "Fast Effective Rule Induction". In: *Twelfth International Conference on Machine Learning*. Morgan Kaufmann, pp. 115–123.

Collobert, Ronan and Jason Weston (2008). "A unified architecture for natural language processing: Deep neural networks with multitask learning". In: *Proceedings of the 25th international conference on Machine learning*, pp. 160–167.

Cowen, Alan S. and Dacher Keltner (2017). "Self-report captures 27 distinct categories of emotion bridged by continuous gradients". In: *Proceedings of the National Academy of Sciences* 114.38, E7900–E7909. ISSN: 0027-8424. DOI: 10.1073/pnas.1702247114.

Danlos, Laurence, Benoît Sagot, and Rosa Stern (July 2010). "Analyse discursive des incises de citation". In: *2ème Congrès Mondial de Linguistique Française - CMLF 2010*. Institut de Linguistique Française. La Nouvelle Orléans, United States.

De Looze, Céline and D. J. Hirst (2008). "Detecting changes in key and range for the automatic modelling and coding of intonation". In.

Delattre, Pierre (1966). "Les dix intonations de base du français". In: *French review*, pp. 1–14.

Devillers, Laurence et al. (2006). "Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches." In: *LREC*, pp. 1105–1110.

Di Cristo, Albert (2013). *La prosodie de la parole*. De Boeck Superieur.

Ding, Chris and Tao Li (2007). "Adaptive dimension reduction using discriminant analysis and k-means clustering". In: *Proceedings of the 24th international conference on Machine learning*, pp. 521–528.

Dominique, Maingueneau (1998). "Analyser les textes de communication". In: *Paris, Dunod*.

Doukhan, David (2013). "Synthèse de parole expressive au delà du niveau de la phrase: le cas du conte pour enfant: conception et analyse de corpus de contes pour la synthèse de parole expressive". PhD thesis. Paris 11.

Doukhan, David et al. (Aug. 2011). "Prosodic Analysis of a Corpus of Tales". In: Florence, Italy: International Speech Communication Association (ISCA), pp. 3129–3132.

Doukhan, David et al. (2015). "The GV-LEx corpus of tales in French". In: *Language Resources and Evaluation* 49.3, pp. 521–547.

Durrer, Sylvie (1999). "Le modelage de la séquence dialoguée". In: *Le dialogue dans le roman*. Nathan.

Ekman, Paul (1999). "Basic Emotions". In: *Handbook of Cognition and Emotion, 1999*. Ed. by Dalgleish T. and Power M. Chap. 3.

Espíndola, RP and NFF Ebecken (2005). "On extending f-measure and g-mean metrics to multi-class problems". In: *WIT Transactions on Information and Communication Technologies* 35.

Fan, Y. et al. (2015). "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4475–4479. DOI: 10.1109/ICASSP.2015.7178817.

Farrus, Mireia, Catherine Lai, and Johanna D. Moore (2016). "Paragraph-based prosodic cues for speech synthesis applications". In: *Proceedings of Speech Prosody 2016*. Boston, MA, USA, pp. 1143–1147. DOI: 10.21437/SpeechProsody.2016-235.

Fayet, Cédric et al. (2020). "FlexEval, creation of light websites for multimedia perceptual test campaigns." In: *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*. Ed. by Christophe Benzitoun et al. Nancy, France: ATALA, pp. 22–25.

Fonseca De Sam Bento Ribeiro, Manuel (2018). "Suprasegmental representations for the modeling of fundamental frequency in statistical parametric speech synthesis". In.

Fukada, Toshiaki et al. (1996). "Speech recognition based on acoustically derived segment units". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Vol. 2. IEEE, pp. 1077–1080.

Galliano, Sylvain, Guillaume Gravier, and Laura Chaubard (2009). "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts". In: *Tenth Annual Conference of the International Speech Communication Association*. Brighton, UK, pp. 2583–2586.

Galliano, Sylvain et al. (2006). "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News." In: *LREC*, pp. 139–142.

Govind, D and SR Mahadeva Prasanna (2013). "Expressive speech synthesis: a review". In: *International Journal of Speech Technology* 16.2, pp. 237–260.

Gravier, Guillaume (2003). *SPro: "Speech Signal Processing Toolkit"*.

Green, Spence et al. (2011). "Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French". In: Conference on Empirical Methods in Natural Language Processing (EMNLP'11). Edinburgh, United Kingdom.

Guennec, David (2016). "Study of unit selection text-to-speech synthesis algorithms.(Étude des algorithmes de sélection d'unités pour la synthèse de la parole à partir du texte)." PhD thesis. University of Rennes 1, France.

Gumperz, John J (2009). "The speech community". In: *Linguistic anthropology: A reader* 1, p. 66.

Harikrishna D M, Gurunath Reddy M, and K. S. Rao (2015). "Multi-stage children story speech synthesis for Hindi". In: *2015 Eighth International Conference on Contemporary Computing (IC3)*, pp. 220–224.

Hartigan, John A and Manchek A Wong (1979). "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, pp. 100–108.

Hayashi, Tomoki et al. (2020). "Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7654–7658.

Henter, Gustav Eje, Xin Wang, and Junichi Yamagishi (2018). "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis". In: *arXiv preprint arXiv:1807.11470*.

Hirst, Daniel J (2007). "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation". In: *Proceedings of the XVIth International Conference of Phonetic Sciences*. Vol. 12331236.

Hojo, Nobukatsu, Yusuke Ijima, and Hideyuki Mizuno (2018). "DNN-based speech synthesis using speaker codes". In: *IEICE TRANSACTIONS on Information and Systems* 101.2, pp. 462–472.

Hsu, W. et al. (2019). "Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5901–5905. DOI: 10.1109/ICASSP.2019.8683561.

Hsu, Wei-Ning et al. (2018). "Hierarchical generative modeling for controllable speech synthesis". In: *arXiv preprint arXiv:1810.07217*.

Hsu, Wei-Ning et al. (2019). "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5901–5905.

Hunt, Andrew J and Alan W Black (1996). "Unit selection in a concatenative speech synthesis system using a large speech database". In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 1. IEEE, pp. 373–376.

Itakura, Fumitada (1975). "Line spectrum representation of linear predictor coefficients of speech signals". In: *The Journal of the Acoustical Society of America* 57.S1, S35–S35.

ITU-T, Recommendation and I Recommend (1996). "P. 800". In: *Methods for subjective determination of transmission quality*.

James, Jesin, Li Tian, and Catherine Inez Watson (2018). "An Open Source Emotional Speech Corpus for Human Robot Interaction Applications." In: *Interspeech*, pp. 2768–2772.

Jauk, Igor (2017). "Unsupervised learning for expressive speech synthesis". PhD thesis. Universitat Politècnica de Catalunya.

Jing, Liping et al. (2005). "Subspace clustering of text documents with feature weighting k-means algorithm". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 802–812.

John, George H and Pat Langley (2013). "Estimating continuous distributions in Bayesian classifiers". In: *arXiv preprint arXiv:1302.4964*.

Johnstone, Tom and Klaus R Scherer (1999). "The effects of emotions on voice quality". In: *Proceedings of the XIVth international congress of phonetic sciences*. Citeseer, pp. 2029–2032.

Kanungo, Tapas et al. (2002). "An efficient k-means clustering algorithm: Analysis and implementation". In: *IEEE transactions on pattern analysis and machine intelligence* 24.7, pp. 881–892.

Kawahara, Hideki, Ikuyo Masuda-Katsuse, and Alain De Cheveigne (1999). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". In: *Speech communication* 27.3-4, pp. 187–207.

Klatt, Dennis H (1980). "Software for a cascade/parallel formant synthesizer". In: *the Journal of the Acoustical Society of America* 67.3, pp. 971–995.

Kominek, John, Tanja Schultz, and Alan W Black (2008). "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion". In: *Spoken Languages Technologies for Under-Resourced Languages*.

Kowsari, Kamran et al. (2019). "Text classification algorithms: A survey". In: *Information* 10.4, p. 150.

Krishnamoorthi, M and AM Natarajan (2013). "ABK-means: an algorithm for data clustering using ABC and K-means algorithm". In: *International Journal of Computational Science and Engineering* 8.4, pp. 383–391.

Labov, William (1970). "The study of language in its social context". In.

Lacheret, Anne et al. (2014). "Rhapsodie: a prosodic-syntactic treebank for spoken french". In: *Language Resources and Evaluation Conference*.

Laferrière, Aude (2018). *Les Incises dans les genres narratifs.«Certaines formules des plus prometteuses»*. Classiques Garnier.

Lai, Catherine, Mireia Farrus, and Johanna Moore (2016). "Automatic Paragraph Segmentation with Lexical and Prosodic Features". In: *Proceedings of Interspeech 2016*. San Francisco, CA, USA.

Larnel, Lori F, Jean-Luc Gauvain, and Maxine Eskenazi (1991). "BREF, a large vocabulary spoken corpus for French". In: *Second european conference on speech communication and technology*.

Lau, Jey Han and Timothy Baldwin (2016). "An empirical evaluation of doc2vec with practical insights into document embedding generation". In: *arXiv preprint arXiv:1607.05368*.

Lazaridis, Alexandros, Blaise Potard, and Philip N Garner (2015). "DNN-based speech synthesis: Importance of input features and training data". In: *International Conference on Speech and Computer*. Springer, pp. 193–200.

Le, Quoc and Tomas Mikolov (2014). "Distributed representations of sentences and documents". In: *International conference on machine learning*, pp. 1188–1196.

Levin, Harry, Carole A Schaffer, and Catherine Snow (1982). "The prosodic and paralinguistic features of reading and telling stories". In: *Language and speech* 25.1, pp. 43–54.

Li, Gang and Fei Liu (2014). "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions". In: *Applied intelligence* 40.3, pp. 441–452.

Liaw, Andy, Matthew Wiener, et al. (2002). "Classification and regression by randomForest". In: *R news* 2.3, pp. 18–22.

Lindau, Mona (1978). "Vowel features". In: *Language* 54.3, pp. 541–563.

Liscombe, Jackson J (2007). *Prosody and speaker state: paralinguistics, pragmatics, and proficiency*. Citeseer.

Liu, Xiaodong et al. (2019). "Multi-task deep neural networks for natural language understanding". In: *arXiv preprint arXiv:1901.11504*.

Lolive, Damien et al. (2017). "The IRISA text-to-speech system for the Blizzard challenge 2017". In.

Loper, Edward and Steven Bird (2002). "NLTK: the natural language toolkit". In: *arXiv preprint cs/0205028*.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov, pp. 2579–2605.

Maekawa, Kikuo (2011). "Discrimination of speech registers by prosody". In: *Dialogue* 9.9, pp. 3–7.

Makhzani, Alireza et al. (2015). "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644*.

Malisz, Zofia et al. (2017). "Controlling Prominence Realisation in Parametric DNN-Based Speech Synthesis." In: *INTERSPEECH*, pp. 1079–1083.

Mamiya, Yoshitaka et al. (2013). "Lightly supervised GMM VAD to use audiobook for speech synthesiser". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*. New Orleans, U.S.A.: IEEE, pp. 7987–7991.

Mareüil, Philippe Boula de and Estelle Maillebau (2002). "Traitement des incises en français: capture automatique et modèle prosodique". In: *XXIVèmes Journées d'Étude sur la Parole, Nancy*.

Mathieu, Michael F et al. (2016). "Disentangling factors of variation in deep representation using adversarial training". In: *Advances in neural information processing systems*, pp. 5040–5048.

Mayo, Catherine, Robert AJ Clark, and Simon King (2005). "Multidimensional scaling of listener responses to synthetic speech". In: *International Speech Communication Association*.

McAuliffe, Michael et al. (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech*, pp. 498–502.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy (2014). "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams engineering journal* 5.4, pp. 1093–1113.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). "Exploiting similarities among languages for machine translation". In: *arXiv preprint arXiv:1309.4168*.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mohammad, Saif (2013). "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales". In: *arXiv preprint arXiv:1309.5909*.

Mohammad, Saif M (2011). *Sentiment Analysis of Mail and Books*. Tech. rep. Technical report, National Research Council Canada.

Montaño, Raúl and Francesc Alías (Dec. 2016). "The Role of Prosody and Voice Quality in Indirect Storytelling Speech: Annotation Methodology and Expressive Categories". In: *Speech Commun.* 85.C, pp. 8–18. ISSN: 0167-6393. DOI: 10.1016/j.specom.2016.10.006.

Montaño, Raúl, Francesc Alías, and Josep Ferrer (2013). "Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis". In: *Eighth ISCA Workshop on Speech Synthesis*.

Moon, Seung-Jae and Björn Lindblom (1994). "Interaction between duration, context, and speaking style in English stressed vowels". In: *The Journal of the Acoustical society of America* 96.1, pp. 40–55.

Morise, Masanori, Fumiya Yokomori, and Kenji Ozawa (2016). "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE TRANSACTIONS on Information and Systems* 99.7, pp. 1877–1884.

Moulines, Eric and Francis Charpentier (1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". In: *Speech communication* 9.5-6, pp. 453–467.

Nigam, Kamal, John Lafferty, and Andrew McCallum (1999). "Using maximum entropy for text classification". In: *IJCAI-99 workshop on machine learning for information filtering*. Vol. 1. 1. Stockholom, Sweden, pp. 61–67.

Novak, Josef Robert, Nobuaki Minematsu, and Keikichi Hirose (2016). "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework". In: *Natural Language Engineering* 22.6, pp. 907–938.

Nzali, Mike Donald Tapi et al. (2017). "FrenchSentiClass: un Système Automatisé pour la Classification de Sentiments en Français". In.

Obin, Nicolas (June 2011). "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style". Theses. Université Pierre et Marie Curie - Paris VI.

Obin, Nicolas et al. (May 2014). "SLAM: Automatic Stylization and Labelling of Speech Melody". en. In: pp. 246 –250.

Oord, Aaron van den et al. (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*.

Orkphol, Korawit and Wu Yang (2019). "Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony". In: *International Journal of Computational Intelligence and Applications* 18.03, p. 1950017.

Patel, Aniruddh D, John R Iversen, and Jason C Rosenberg (2006). "Comparing the rhythm and melody of speech and music: The case of British English and French". In: *The Journal of the Acoustical Society of America* 119.5, pp. 3034–3047.

Perret, Michèle (1994). *L'énonciation en grammaire du texte*. Vol. 45. Nathan.

Ping, Wei et al. (2017). “Deep voice 3: Scaling text-to-speech with convolutional sequence learning”. In: *arXiv preprint arXiv:1710.07654*.

Povey, Daniel et al. (2011). “The Kaldi speech recognition toolkit”. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society.

Rajeswari, KC and MP Uma (2012). “Prosody modeling techniques for text-to-speech synthesis systems–a survey”. In: *International Journal of Computer Applications* 39.16, pp. 8–11.

Ramli, Izzad et al. (2016). “Prosody Analysis of Malay Language Storytelling Corpus”. In: Springer, pp. 563–570.

Rehurek, Radim and Petr Sojka (2010). “Software framework for topic modelling with large corpora”. In: *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Riaz, Sumbal et al. (2019). “Opinion mining on large scale data using sentiment analysis and k-means clustering”. In: *Cluster Computing* 22.3, pp. 7149–7164.

Sarkar, P. and K. S. Rao (Aug. 2015). “Analysis and modeling pauses for synthesis of storytelling speech based on discourse modes”. In: *2015 Eighth International Conference on Contemporary Computing (IC3)*, pp. 225–230. DOI: 10.1109/IC3.2015.7346683.

Scherer, Klaus R, Tom Johnstone, and Gundrun Klasmeyer (2003). *Vocal expression of emotion.* Oxford University Press.

Schöch, Christof et al. (2016). “Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels”. In: *Digital Humanities 2016: Conference Abstracts*, pp. 346–353.

Schoeffler, Michael et al. (2015). “Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)”. In: *1st Web Audio Conference*, pp. 1–6.

Schröder, Marc (2001). “Emotional speech synthesis: A review”. In: *Seventh European Conference on Speech Communication and Technology*.

Schuller, Björn, Stefan Steidl, and Anton Batliner (2009). “The interspeech 2009 emotion challenge”. In: *10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*. Brighton, U.K.

Schuller, Björn et al. (2009). “Acoustic emotion recognition: A benchmark comparison of performances”. In: *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, pp. 552–557.

Schuller, Björn et al. (2013a). “Paralinguistics in speech and language—state-of-the-art and the challenge”. In: *Computer Speech & Language* 27.1, pp. 4–39.

Schuller, Björn et al. (2013b). “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism”. In: *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Series, B (2014). “Method for the subjective assessment of intermediate quality level of audio systems”. In: *International Telecommunication Union Radiocommunication Assembly*.

Shen, Jonathan et al. (2018). “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4779–4783.

Sini, Aghilas, Elisabeth Delais-Roussarie, and Damien Lolive (May 2018). “Annotation automatique des types de discours dans des livres audio en vue d’une oralisation par un système de synthèse”. In: *TALN-RECITAL 2018*. Rennes, France.

Sini, Aghilas et al. (2018). “SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan.

Snyder, David et al. (2017). “Deep Neural Network Embeddings for Text-Independent Speaker Verification.” In: *Interspeech*, pp. 999–1003.

Snyder, David et al. (2018). “X-vectors: Robust DNN embeddings for speaker recognition”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5329–5333.

Socher, Richard et al. (2011). “Semi-supervised recursive autoencoders for predicting sentiment distributions”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 151–161.

Spangler, William Scott (2008). *Method for adapting a K-means text clustering to emerging data*. US Patent 7,430,717.

Sugiyama, Masashi (2015). *Introduction to statistical machine learning*. Morgan Kaufmann.

Sun, Xuejing (2002). “Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio”. In: *2002 IEEE international conference on acoustics, speech, and signal processing*. Vol. 1. IEEE, pp. I–333.

Sun, Xuejing and Yi Xu (2002). “Perceived pitch of synthesized voice with alternate cycles”. In: *Journal of Voice* 16.4, pp. 443–459.

Swaileh, Wassim, Kamel Ait-Mohand, and Thierry Paquet (2016). "Un modèle syllabique pour la reconnaissance de l'écriture." In: *CORIA-CIFED*, pp. 23–37.

Székely, Eva et al. (2012a). "Detecting a targeted voice style in an audiobook using voice quality features". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*. IEEE, pp. 4593–4596.

Székely, Eva et al. (2012b). "Evaluating expressive speech synthesis from audiobooks in conversational phrases". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, pp. 3335–3339.

Tachibana, Hideyuki, Katsuya Uenoyama, and Shunsuke Aihara (2018). "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4784–4788.

Tahon, Marie and Laurence Devillers (2016). "Towards a small set of robust acoustic features for emotion recognition: challenges". In: *IEEE/ACM transactions on audio, speech, and language processing* 24.1, pp. 16–28.

Talkin, David (1995). "A robust algorithm for pitch tracking (RAPT)". In: *Speech coding and synthesis*, pp. 495–518.

Taylor, Paul (2009). *Text-to-speech synthesis*. Cambridge university press.

Theune, M. et al. (July 2006). "Generating expressive speech for storytelling applications". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4, pp. 1137–1144. ISSN: 1558-7916. DOI: 10.1109/TASL.2006.876129.

Tiomkin, Stas et al. (2010). "A hybrid text-to-speech system that combines concatenative and statistical synthesis units". In: *IEEE transactions on audio, speech, and language processing* 19.5, pp. 1278–1288.

Titze, Ingo R (1995). *Workshop on acoustic voice analysis: Summary statement*. National Center for Voice and Speech.

Tokuda, Keiichi, Heiga Zen, and Alan W Black (2002). "An HMM-based speech synthesis system applied to English". In: *IEEE Speech Synthesis Workshop*, pp. 227–230.

Tokuda, Keiichi et al. (1994). "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation". In: *Third International Conference on Spoken Language Processing*.

Torreira, Francisco, Martine Adda-Decker, and Mirjam Ernestus (2010). "The Nijmegen corpus of casual French". In: *Speech Communication* 52.3, pp. 201–212.

Tripathi, Kumud, Parakrant Sarkar, and Sreenivasa Rao (2017). "Sentence Based Discourse Classification for Hindi Story Text-to-Speech (TTS) System". In.

Trouvain, Jürgen (2014). "Laughing, breathing, clicking-The prosody of nonverbal vocalisations". In: *Proceedings of Speech Prosody*. Vol. 7, pp. 598–602.

Vaissière, Jacqueline (1983). "Language-independent prosodic features". In: *Prosody: Models and measurements*. Springer, pp. 53–66.

Vaissière, Jacqueline and Alexis Michaud (2006). *Prosodic constituents in French: a data-driven approach*.

Veenendaal, Nathalie J, Margriet A Groen, and Ludo Verhoeven (2014). "The role of speech prosody and text reading prosody in children's reading comprehension". In: *British Journal of Educational Psychology* 84.4, pp. 521–536.

Vít, Jakub and Jind?ich Matou?ek (2016). "Unit-Selection Speech Synthesis Adjustments for Audiobook-Based Voices". In: *Text, Speech, and Dialogue*. Ed. by Petr Sojka et al. Vol. 9924. DOI: 10.1007/978-3-319-45510-5_38. Cham: Springer International Publishing, pp. 335–342. ISBN: 978-3-319-45509-9 978-3-319-45510-5.

Wang, Yu-Tsai et al. (2010). "Breath group analysis for reading and spontaneous speech in healthy adults". In: *Folia Phoniatrica et Logopaedica* 62.6, pp. 297–302.

Wang, Yuxuan et al. (2017a). "Tacotron: Towards end-to-end speech synthesis". In: *arXiv preprint arXiv:1703.10135*.

Wang, Yuxuan et al. (2017b). "Uncovering latent style factors for expressive speech synthesis". In: *arXiv preprint arXiv:1711.00520*.

Wu, Zhizheng, Oliver Watts, and Simon King (2016). "Merlin: An Open Source Neural Network Speech Synthesis System." In: *SSW*, pp. 202–207.

Xu, Longting et al. (2018). "Generative x-vectors for text-independent speaker verification". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 1014–1020.

Ze, Heiga, Andrew Senior, and Mike Schuster (2013). "Statistical parametric speech synthesis using deep neural networks". In: *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, pp. 7962–7966.

Zen, Heiga et al. (2019). "LibriTTS: A corpus derived from librispeech for text-to-speech". In: *arXiv preprint arXiv:1904.02882*.

Titre : Caractérisation et génération de l'expressivité en fonction des styles de parole pour la construction de livres audio

Mot clés : Informatique, Prosodie de la parole, Livres audio, Synthèse de la parole expressive, Apprentissage Automatique

Résumé : Dans ces travaux de thèse nous abordons l'expressivité de la parole lue avec un type de données particulier qui sont les livres audio. Les livres audio sont des enregistrements audio d'œuvres littéraires fait par des professionnels (des acteurs, des chanteurs, des narrateurs professionnels) ou par des amateurs. Ces enregistrements peuvent être destinés à un public particulier (aveugles ou personnes mal voyantes). La disponibilité de ce genre de données en grande quantité avec une assez bonne qualité a attiré l'attention de la communauté scientifique en traitement automatique du langage et de la parole en général, ainsi que des chercheurs spécialisés dans la synthèse de parole expressive. Pour explorer ce vaste champ d'investigation qui est l'expressivité, nous proposons dans cette thèse d'étudier trois entités élémentaires de l'expressivité qui sont véhiculées par les livres audio: l'émotion, les variations liées aux changements discursifs et les propriétés du locuteur. Nous traitons ces patrons d'un point de vue prosodique. Les principales contributions de cette thèse sont la construction d'un corpus de livres audio comportant un nombre important d'enregistrements partiellement annotés par un expert, une étude quantitative caractérisant les émotions dans ce type de données, la construction de modèles basés sur des techniques d'apprentissage automatique pour l'annotation automatique de types de discours et enfin nous proposons une représentation vectorielle de l'identité prosodique d'un locuteur dans le cadre de la synthèse statistique paramétrique de la parole.

Title: Characterisation and generation of expressivity in function of speaking styles for audiobook synthesis

Keywords: Computer Science,Speech Prosody, Audiobook, Expressive Speech Synthesis,Machine Learning

Abstract: In this thesis, we study the expressivity of read speech with a particular type of data, which are audiobooks. Audiobooks are audio recordings of literary works made by professionals (actors, singers, professional narrators) or by amateurs. These recordings may be intended for a particular audience (blind or visually impaired people). The availability of this kind of data in large quantities with a good enough quality has attracted the attention of the research community in automatic speech and language processing in general and of researchers specialized in expressive speech synthesis systems. We propose in this thesis to study three elementary entities of expressivity that are conveyed by audiobooks: emotion, variations related to discursive changes, and speaker properties. We treat these patterns from a prosodic point of view. The main contributions of this thesis are: the construction of a corpus of audiobooks with a large number of recordings partially annotated by an expert, a quantitative study characterizing the emotions in this type of data, the construction of a model based on automatic learning techniques for the automatic annotation of discourse types and finally we propose a vector representation of the prosodic identity of a speaker in the framework of parametric statistical speech synthesis.