ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAÎTRISE EN GÉNIE DES TECHNOLOGIES DE L'INFORMATION
M. Sc. A.

PAR
Simon BOUTIN

UTILISATION DES CARACTÉRISTIQUES PROSODIQUES POUR OPTIMISER UN SYSTÈME DE COMPRÉHENSION DU LANGAGE NATUREL

MONTRÉAL, LE 27 JUIN 2016

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Pierre Dumouchel, directeur de mémoire
Directeur général à l'École de technologie supérieure

M. Réal Tremblay, codirecteur de mémoire
Nuance Communications

M. Patrick Cardinal, codirecteur de mémoire
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Éric Granger, président du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Najim Dehak, membre du jury
Université Johns Hopkins

Mme Sylvie Ratté, membre du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY

LE 19 MAI 2016

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

AVANT-PROPOS

Depuis mon adolescence, je rêve à la possibilité qu'un jour des machines puissent acquérir une intelligence et une conscience similaires à celles des êtres humains, voire supérieures. N'ayant pas les mêmes contraintes par rapport à nos corps biologiques, ces machines pourraient produire de nouvelles connaissances à un rythme effréné. Elles pourraient résoudre des problèmes auxquels fait face l'humanité comme la pauvreté, la dégradation de l'environnement, la maladie, le vieillissement et même la mort.

Je rêve donc depuis longtemps au potentiel qu'il serait possible d'exploiter de ces machines. N'ayant pas moi-même les capacités de créer de telles machines, j'ai progressivement pris conscience que je pouvais tout de même y apporter ma petite contribution, sachant que beaucoup d'efforts seraient nécessaires de ma part.

L'intelligence artificielle étant un domaine très vaste, je ne savais pas dans quelle branche de cette discipline il valait mieux concentrer mes efforts. Heureusement, j'ai entamé un cours de maîtrise lors de mon baccalauréat qui m'a éclairé dans ma décision. J'ai décidé de me spécialiser en traitement automatique du langage naturel, plus précisément en sa compréhension. La compréhension du langage naturel est l'une des branches les plus difficiles et les plus complexes de l'intelligence artificielle, mais est également une des facultés les plus riches de l'intelligence humaine. Les bénéfices qu'apporterait une compréhension fiable des machines par rapport à l'ensemble des connaissances humaines seraient, selon moi, sans limites.

C'est donc avec cette passion que j'ai entrepris mon projet de maîtrise. Ma contribution concerne l'utilisation des propriétés prosodiques (intonation, intensité, durée) d'un signal acoustique pour tenter d'améliorer les performances d'un système de compréhension automatique de la parole. Je considère ma contribution comme étant bien mince par rapport à l'ampleur de mon rêve. Cependant, les contributions de milliers de chercheurs à travers le

monde, échelonnées sur plusieurs décennies, pourront peut-être transformer ce rêve cher en une réalité bien concrète.

REMERCIEMENTS

Je tiens à remercier mon directeur de mémoire et directeur général de l'ÉTS, Pierre Dumouchel, pour avoir cru en moi et m'avoir offert son support et son expertise. Il m'a également initié au traitement automatique du langage naturel dans l'un de ses cours, ce qui m'a aidé à définir mon orientation professionnelle et à connaître Nuance Communications.

Je remercie mon superviseur en milieu de pratique et gestionnaire TNL, Réal Tremblay, pour avoir cru en moi et m'avoir ainsi donné la chance d'exploiter mon potentiel dans un domaine passionnant. Il m'a offert son support, son expertise, sa disponibilité, et m'a intégré au sein de son équipe.

Je remercie mon codirecteur de mémoire, Patrick Cardinal, pour son expertise, son temps et ses précieux conseils, qui ont favorisé la réussite de mon projet. Ses qualités ont même favorisé ma réussite à certains cours antérieurs à ce projet.

Je remercie l'ingénieur de recherche en milieu de pratique, Doug Peters, pour son aide, son expertise et ses conseils. Ses qualités ont eu une grande influence sur le bon déroulement de mes travaux.

Je remercie les organismes qui ont financé ma recherche, soit le CRSNG, le FQRNT, et Nuance Communications, avec leur bourse BMP-Innovation. Cette bourse m'a permis de consacrer tout mon temps et mon énergie à la réussite de ma maîtrise.

Je remercie également ma mère, Danièle Viau, pour son support, ses conseils et ses encouragements soutenus pour toute la durée de mes études, incluant la maîtrise.

Enfin, je remercie le directeur de l'ingénierie chez Google, Raymond Kurzweil, pour m'avoir fourni indirectement la motivation nécessaire pour mener à bien ma maîtrise, grâce à ses

propos soutenus et sa vision contagieuse concernant la théorie de la Singularité technologique.

UTILISATION DES CARACTÉRISTIQUES PROSODIQUES POUR OPTIMISER UN SYSTÈME DE COMPRÉHENSION DU LANGAGE NATUREL

Simon BOUTIN

RÉSUMÉ

En général, les entreprises œuvrant dans l'industrie des systèmes de dialogue homme-machine offrent plusieurs applications informatiques à leurs clients, dont la compréhension automatique du langage naturel. Les systèmes courants de dialogue humain-machine sont constitués de trois composants faiblement couplés :

-Le système de reconnaissance vocale (ASR);
-Le système de compréhension du langage naturel (NLU);
-Le système de dialogue ou agent conversationnel (CA).

Dans cette architecture, les sorties des composants précédents servent d'entrées aux composants suivants. Les caractéristiques du signal acoustique ne font pas partie de la sortie du premier composant. Il est possible qu'il y ait de l'information supplémentaire dans le signal original pouvant aider directement le système NLU à effectuer sa tâche. Cette information dite « prosodique » concerne l'intonation, l'intensité et la durée des sons, qui est évidemment absente du texte écrit. Par exemple, l'identification d'un texte libre présent dans une commande vocale est particulièrement difficile pour le système NLU actuel. La littérature n'aborde pas directement l'identification des textes libres. Le concept le plus similaire étant l'identification des citations. L'originalité de cette étude est que l'auteur de la citation et son narrateur correspondent à la même entité.

L'objectif primaire de cette étude consistait à déterminer s'il existait une corrélation entre l'information prosodique d'un signal acoustique et la présence ou non des textes libres. Trois types de caractéristiques prosodiques ont été extraits à partir d'un grand ensemble de commandes vocales. Ces extractions ont permis de générer des distributions de leurs mesures. Les distributions des textes libres ont été comparées par rapport à celles des autres concepts, grâce au test de Kolmogorov–Smirnov (test K-S) à deux échantillons. Les résultats ont indiqué qu'il existait effectivement une corrélation. L'objectif secondaire consistait à vérifier s'il était possible d'améliorer les performances d'un système NLU grâce à cette information prosodique. Un système NLU minimal a été utilisé. Pour vérifier les gains de performance, des modèles basés sur des caractéristiques lexicales seules ont été comparés par rapport à des modèles augmentés par des caractéristiques prosodiques. Le test de McNemar a été utilisé pour vérifier si les gains obtenus étaient significatifs. L'information prosodique a effectivement amélioré les performances du système.

Mots clés : prosodie, texte libre, Kolmogorov-Smirnov, compréhension du langage naturel

USE OF PROSODIC FEATURES TO OPTIMIZE A NATURAL LANGUAGE UNDERSTANDING SYSTEM

Simon BOUTIN

ABSTRACT

In general, companies working in the human-machine dialog systems industry offer several computer applications to their customers, including automatic natural language understanding. Current human-machine dialog systems are composed of three weakly coupled components:

-The automatic speech recognition system (ASR);
-The natural language understanding system (NLU)
-The dialog system or conversational agent (CA).

In this architecture, the outputs of the preceding components are the inputs of the following. The characteristics of the acoustic signal are not included in the output of the first component. But it is possible that additional information in the original signal can directly help the NLU system to perform its task. This so-called "prosodic" information concerns intonation, intensity and duration of sound, which is obviously absent from the written text. For example, the identification of free text in a voice command is particularly difficult for the current NLU system. The literature does not directly address the identification of free texts. The most similar concept is the identification of quotations. By focusing on free texts, the originality of this study is that the author of the quotation and its narrator correspond to the same entity.

The first objective of this thesis was to determine whether there is a correlation between prosodic information of an acoustic signal and the presence or absence of free texts. Three types of prosodic features were extracted from a large set of voice commands, and their sample distributions were examined. Distributions for free texts were compared to those of other concepts using the two-sample Kolmogorov-Smirnov test (K-S test). The results showed that there was indeed a correlation. The second objective was to verify whether it is possible to improve the performance of an NLU system through the integration of prosodic information. Given a minimal NLU system, the performance gains of models based on lexical features alone were compared against models augmented with prosodic features. The McNemar's test was used to verify whether the gains were significant. Prosodic information has indeed improved this system’s performance.

Keywords: prosody, free text, Kolmogorov-Smirnov, natural language understanding

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AD	Acte de Dialogue
ASR	Automatic Speech Recognition
CA	Conversational Agent
CART	Classification And Regression Tree
CDF	Cumulative Distribution Function
CRF	Conditional Random Fields
DMA	Dragon Mobile Assistant
F0	Fréquence fondamentale
HMM	Hidden Markov Model
K-S	Kolmogorov–Smirnov
LM	Language model
MLP	Multilayer Perceptron
NLP	Natural Language Processing
NLU	Natural Language Understanding
PLM	Polygram Language Models
POSH	Prosodic Ostendorf & Shattuck-Hufnagel
PPF	Percent Point Function
RBF	Radial Basis Function
SCT	Semantic Classification Tree
SVM	Support Vector Machine
ToBI	Tones and Break Indices

WHG	Word Hypotheses Graphs

LISTE DES SYMBOLES ET UNITÉS DE MESURE

UNITÉS DE MESURE

ms	milliseconde
Hz	Hertz

SYMBOLES

A	Type de modèle de référence (caractéristiques lexicales seulement)
B	Type de modèle augmenté avec la durée des pauses
C	Type de modèle augmenté avec les réinitialisations de F0
D	Type de modèle augmenté avec les continuités de F0
E	Type de modèle augmenté avec toutes les caractéristiques prosodiques
E-AC	Échantillon des autres concepts extrait de façon aléatoire
E-AT	Échantillon des textes libres extrait de façon aléatoire
E-CC	Échantillon des autres concepts extrait de façon ciblé
E-CT	Échantillon des textes libres extrait de façon ciblé
E-E	Échantillon utilisé pour l'évaluation de la qualité des outils
e_{auto}	Erreur de la mesure automatique de la durée d'une pause
F0	Fréquence fondamentale
m_{auto}	Mesure automatique de la durée d'une pause
m_{manu}	Mesure manuelle de la durée d'une pause
$t_{début}$	Position dans le temps du début d'une pause
t_{fin}	Position dans le temps de la fin d'une pause

INTRODUCTION

En général, les entreprises œuvrant dans l'industrie des systèmes de dialogue homme-machine offrent plusieurs applications informatiques à leurs clients, dont la compréhension automatique du langage naturel (NLU). Afin de satisfaire leur clientèle existante et faciliter l'acquisition de nouveaux clients, ces entreprises doivent constamment tenter d'améliorer leurs algorithmes de compréhension pour augmenter les performances globales de leurs systèmes. Ces algorithmes peuvent être intégrés à des téléphones intelligents, où les clients les utilisent pour demander des informations orales génériques à leur appareil ou déclencher certaines actions. Afin que les différentes applications soient en mesure d'utiliser cette information communiquée oralement, les bandes audio sont filtrées selon plusieurs étapes. Certaines pièces d'informations précises sont finalement étiquetées dans des champs spécifiques appelés « mentions ». C'est grâce à ces mentions que cette information orale devient utilisable par les autres applications.

Les systèmes courants de dialogue humain-machine sont constitués de trois composants faiblement couplés. Le premier composant, le système de reconnaissance vocale (ASR), a comme responsabilité la conversion d'un signal acoustique en texte, c.-à-d. en une chaîne de mots reconnus. La sortie de ce module inclue parfois les meilleures hypothèses et quelques segments de mots peuvent être enrichis avec des mesures de confiance. Les caractéristiques du signal acoustique ne font pas partie de cette sortie. Le deuxième composant, le système NLU, prend en entrée la sortie du module ASR (le texte avec les mesures de confiance) et en extrait le sens sous la forme d'une signification sémantique bien formée. Le dernier composant, le système de dialogue ou agent conversationnel (CA), prend en entrée cette signification sémantique. En considérant le contexte du dialogue entier (demander plus d'informations, accéder à une base de données et proposer des choix, etc.), ce composant décide de l'action à entreprendre et fait une mise à jour de son état interne du dialogue.

Il est possible qu'il y ait de l'information supplémentaire dans le signal acoustique original pouvant aider directement le système NLU à effectuer sa tâche. Cette information dite «

prosodique » concerne l'intonation, l'intensité et la durée des sons, qui est évidemment absente du texte écrit. Actuellement en industrie, l'information prosodique n'est généralement pas utilisée pour améliorer les performances des systèmes NLU. Dans l'objectif de fournir une preuve de cette hypothèse, il est intéressant d'étudier un cas où des corrélations peuvent être évidentes. Par exemple, l'identification d'un texte libre présent dans une commande vocale est particulièrement difficile pour le système NLU actuel.

Un « texte libre » est une séquence de mots correspondant à un sous-ensemble d'une commande vocale, ne nécessitant aucune analyse sémantique de la part des systèmes NLU actuels. Cette séquence de mots correspond à un discours dans un discours, comme un titre ou un corps de courriel. La littérature n'aborde pas directement le concept de texte libre. Le concept le plus similaire étant la citation. En s'intéressant aux textes libres, l'originalité de ce mémoire est que l'auteur de la citation et son narrateur correspondent à la même entité. Par exemple, si une commande vocale est : « Text Mary [have a nice day] », la partie entre crochets correspond à un texte libre (texto) et ne requiert aucune analyse sémantique. Le système NLU doit simplement considérer cette séquence de mots comme une séquence de caractères à accepter telle quelle. Dans cet exemple, la ponctuation « […] » permet d'identifier le texte libre. Cependant, cette ponctuation est absente lorsque la commande est exprimée oralement. L'analyse sémantique du système NLU ne permet pas d'identifier efficacement les textes libres seulement à partir des mots. En effet, il semble ne pas y avoir de motif récurent parmi les commandes vocales concernées. De plus, il existe du chevauchement entre le vocabulaire constituant les textes libres et celui constituant les autres concepts.

L'objectif primaire de cette étude consiste donc à déterminer s'il existe une corrélation entre l'information prosodique d'un signal acoustique et la présence ou non des textes libres. L'objectif secondaire est de vérifier s'il est possible d'améliorer les performances d'un système NLU grâce à cette information prosodique, notamment pour l'identification de ce concept. Par exemple, lorsque l'on réfléchit à une commande vocale contenant un texte libre, on peut s'imaginer une pause plus longue précédant immédiatement ce dernier.

Ce mémoire est divisé en cinq chapitres. Le CHAPITRE 1 présente une revue de la littérature des recherches liées à la prosodie et utiles pour l'identification des textes libres. Le CHAPITRE 2 présente la méthodologie expérimentale utilisée pour cette étude, incluant les corpus de données, les outils de mesures utilisés ainsi qu'une évaluation de leur qualité. Le CHAPITRE 3 présente la première expérimentation basée sur la durée des pauses. Une définition des textes libres est donnée, des hypothèses sont formulées et un test est présenté pour déterminer s'il existe des corrélations dans les données. Le CHAPITRE 4 présente la deuxième expérimentation basée sur des caractéristiques prosodiques issues de la fréquence fondamentale, c.à.d. F0. Des problèmes sont présentés liés à l'extraction de F0. Une méthode est présentée permettant de mieux représenter les mesures extraites. Deux caractéristiques issues de F0 sont présentées et des hypothèses sont formulées pour chacune de ces caractéristiques. Le CHAPITRE 5 présente la dernière expérimentation basée sur un système NLU. La définition du système NLU est présentée et les dernières hypothèses de cette étude sont formulées. Finalement, une conclusion confirme qu'il existe une corrélation significative entre l'information prosodique et la présence ou non des textes libres. Elle confirme également qu'il est possible d'utiliser cette information afin d'améliorer les performances d'un système NLU.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

1.1 Introduction

Dans la communication humaine, les mots sont utilisés principalement pour exprimer et comprendre un texte écrit ou exprimé oralement, mais ces mots ne sont pas toujours suffisants pour une compréhension adéquate. En effet, une même séquence de mots peut générer des phrases avec des ambiguïtés syntaxiques, c.-à-d. être sujet à différentes interprétations sémantiques. Pour le texte écrit, la compréhension du sens d'un énoncé est grandement facilitée par la ponctuation utilisée. Cependant, cette ponctuation n'est habituellement pas exprimée explicitement dans le langage parlé. Par contre, le langage parlé contient fréquemment de l'information prosodique pouvant remplir ce rôle.

Sur le plan acoustique, la prosodie peut se résumer comme l'ensemble, au cours d'un énoncé, des modifications de la fréquence fondamentale (F0), de l'intensité (force de la voix), et de la durée [12]. Le terme *suprasegmental* introduit dans [43] représente des phénomènes de la parole attribués à des segments plus larges par rapport aux phonèmes (syllabes, mots, phrases, etc.). Pour ces segments, des propriétés de perception sont attribuées comme la hauteur (intonation), le volume (intensité), le débit de parole, la qualité de la voix, la durée, la pause, le rythme, etc.

Lorsque le langage parlé est converti en une simple séquence de mots via un système ASR, les indices prosodiques du langage sont perdus. Pourtant, dans toutes les langues, la prosodie est utilisée pour fournir des informations structurelles, sémantiques et fonctionnelles [68]. Caelen-Haumont [12] affirme qu'à un niveau général, la prosodie est un élément essentiel de la parole. Sans elle la parole est grandement détériorée ou même inaudible. La prosodie serait donc le support concret de la parole. Waibel [75] affirme que de nombreuses études ont attesté cinq utilisations des indices prosodiques : pour indiquer l'émotion qu'apporte un locuteur à un énoncé, pour attirer l'attention sur des parties d'un énoncé, pour indiquer

l'intention du locuteur, pour résoudre l'ambiguïté des phrases ambigües et pour marquer des frontières et des catégories syntaxiques majeures. La littérature s'intéresse depuis plusieurs décennies à la prosodie pour étudier la communication humaine ([49], [45], [6]), de même dans le domaine du traitement automatique du langage ([19], [41], [48]). Les caractéristiques prosodiques d'un signal acoustique sont exploitées pour tenter d'améliorer les performances des systèmes ASR ([75], [40]) de même que les systèmes NLU ([67]). Comme le constate [12], l'intérêt initial du domaine de la prosodie concernant l'étude de la structuration syntaxique s'est orienté au fil des années vers la considération des facteurs sémantiques et pragmatiques. Cette orientation s'est fait au même moment où les corpus de phrases lues et isolées ont progressivement été abandonnés au profit du dialogue spontané. Cette évolution s'est également produite en parallèle dans le domaine du traitement du langage naturel (NLP).

Ce chapitre présente un sous-ensemble de la littérature concernant la prosodie appliquée au NLU. Plus précisément, l'intérêt est de déterminer s'il est possible d'améliorer l'identification des textes libres, présents dans certaines commandes vocales, à l'aide d'indices prosodiques. Pour commencer, une analyse de la littérature est effectuée concernant l'extraction et l'analyse de l'information prosodique d'un signal acoustique. Ensuite, l'utilisation de la prosodie est explorée pour tenter d'améliorer les performances des systèmes NLU. Enfin, même si la littérature ne semble pas aborder directement l'identification des textes libres, une revue des sujets similaires est présentée.

1.2 Extraction et analyse de l'information prosodique

1.2.1 Utilité des caractéristiques prosodiques

Déjà en 1972, un rapport technique [40] de l'*Advanced Research Projects Agency* (ARPA), nommée aujourd'hui la *Defense ARPA* (DARPA), s'intéresse à la prosodie pour améliorer un système ASR. Les auteurs démontrent que les caractéristiques prosodiques d'un signal acoustique peuvent aider à identifier les frontières entre les phrases. Elles permettent

également de déterminer la localisation des syllabes accentuées à l'intérieur de chaque constituant syntactique et d'effectuer une analyse plus approfondie sur celles-ci. À l'époque, les résultats sont déjà encourageants. Un programme est développé permettant d'identifier 90% de toutes les frontières entre les constituants syntactiques majeurs. L'identification des frontières permet de segmenter la parole continue en phrase ou propositions indépendantes. Cette segmentation est une étape clé pour en extraire la sémantique. Dans des travaux plus récents, les indices prosodiques montrent leur utilité pour la segmentation automatique du langage naturel ([4], [22], [51]). Les frontières entre les phrases sont normalement marquées par une pause, suivie d'une nette augmentation de F0 au début des nouvelles phrases. Cooper *et al.* [14] concluent également que les descentes suivies de montées significatives de F0 accompagnent généralement les frontières entre les clauses conjointes, ainsi que celles entre les clauses principales et intégrées. Les phrases syntaxiques (phrases contenant minimalement un groupe nominal sujet et un groupe verbal) majeures sont également rapportées comme étant signalées par des creux détectables dans le contour de F0. La hauteur pourrait donc agir en tant que facteur acoustique de la structure syntaxique. Cependant, les meilleures caractéristiques prosodiques permettant de localiser les syllabes accentuées ne peuvent être exploitées par les auteurs de ce rapport.

Une faiblesse de cette expérimentation par rapport au cadre de ce mémoire est que celle-ci est entièrement expérimentée avec des textes lus, contrairement à de l'oral spontané normalement utilisé pour les commandes vocales. L'oral spontané à l'inconvénient de contenir beaucoup plus de disfluidités par rapport aux textes lus. Ces disfluidités rendent la tâche plus difficile pour les modèles de NLU ([46]). Certains auteurs comparent le texte lu par rapport à l'oral spontané concernant la prosodie ([74]). Enfin, ce rapport n'étudie pas d'autres caractéristiques prosodiques d'intérêt comme les rythmes, les pauses et les frontières entre les mots. Les pauses représentent un intérêt particulier pour ce mémoire, puisqu'elles sont soupçonnées d'être généralement plus prononcées précédant les textes libres. Ce rapport conduit à l'article [39] publié en 1975.

1.2.2 Utilisation de la prosodie et des mots

Il peut être avantageux de tirer profit de la connaissance des mots, obtenue précédemment par un système ASR, afin de faciliter l'extraction de classes prosodiques d'un signal acoustique. Le problème peut être perçu comme un problème de classement statistique ([66]). Ainsi, certaines unités linguistiques U (par exemple des mots ou des phrases) peuvent être classées à une ou plusieurs classes prosodiques S. Le rôle de la prosodie est de fournir un ensemble de caractéristiques F pouvant aider à prédire S. Dans le cadre probabiliste, l'estimation $P(S|F)$ est recherchée. La tâche de modélisation peut être généralisée comme étant P(S|W,F) s'il est également souhaitable d'utiliser l'information contenue dans une séquence de mots W associée à U.

En 1994, Wightman *et al.* [78] décrivent un algorithme général pour étiqueter automatiquement des informations prosodiques de la parole. Bien que cette approche soit artificielle, celle-ci permet d'enrichir les connaissances du système afin d'en faciliter les traitements subséquents. Cet algorithme utilise un arbre de décision pour faire la correspondance entre une séquence d'observations (vecteurs de caractéristiques prosodiques) et une séquence d'étiquettes prosodiques. Leur système comprend un module d'extraction des caractéristiques prosodiques et un module pour leur classement. Pour cet algorithme, la séquence de mots correspondant au signal acoustique est connue. Cette séquence permet d'obtenir plusieurs caractéristiques prosodiques grâce à la disponibilité des étiquettes phonétiques et de leur durée. L'étiquetage prosodique choisi par les auteurs concerne principalement la segmentation en phrases prosodiques et la proéminence de phrase. Ces indices sont simples à étiqueter manuellement et offrent probablement la plus grande source de connaissance pour le NLU.

Les auteurs effectuent un entraînement supervisé à l'aide de deux corpus. Le premier corpus est constitué de phrases ambigües lues par des annonceurs de radio professionnels, où les frontières de phrases sont connues. Le second corpus est constitué de reportages lus, également par des annonceurs de radio professionnels, mais où les frontières de phrases sont

inconnues. L'inconvénient de leur expérimentation par rapport à ce mémoire est l'absence d'une extension pour l'oral spontané non professionnel. Les deux corpus sont manuellement étiquetés prosodiquement avec les index de rupture de *Tones and Break Indices* (ToBI [69]), ainsi qu'avec une étiquette binaire pour la proéminence.

L'avantage d'utiliser l'arbre de décision est qu'il peut être utilisé pour classer n'importe quelle séquence d'informations prosodiques. Il peut donc classer d'autres types d'étiquettes. Ces étiquettes peuvent être associées à des vecteurs de caractéristiques prosodiques hétérogènes, d'où la force d'utiliser l'arbre de décision. Cependant, Shriberg et Stolcke [66] affirment que les arbres de décision ont deux problèmes principaux. Pour diminuer leur taille, ils enveloppent un algorithme de sélection de sous-ensembles de caractéristiques autour d'un algorithme standard pour la croissance de l'arbre. Ce premier algorithme trouve fréquemment de meilleurs classifieurs, en éliminant des caractéristiques nuisibles à partir des considérations de l'arbre. Deuxièmement, ils effectuent l'entraînement sur une version rééchantillonnée de la distribution cible, dans laquelle toutes les classes ont des probabilités à priori égales. Cette version permet de sensibiliser les arbres aux caractéristiques prosodiques pour les cas où la taille des classes est très asymétrique.

L'étiquetage automatique des informations prosodiques facilite l'annotation de corpus large. En plus d'étiqueter prosodiquement des corpus à l'aide des mots, certains auteurs s'intéressent à la combinaison des mots et de la prosodie pour extraire d'autres informations. En 1999, Hakkani-tür *et al.* [18] s'intéressent à l'identification des frontières de sujets, de phrases et de noms propres en utilisant des indices prosodiques, avec et sans l'aide des mots. Par leur nature, les informations prosodiques sont peu affectées par l'identité des mots ([68]). Elles peuvent ainsi améliorer la robustesse des méthodes d'extraction d'information lexicale. Selon les auteurs, leur système est le premier à combiner l'extraction entièrement automatisée de l'information prosodique et lexicale pour le NLU. Dans leur approche, la transcription des mots est alignée automatiquement avec le signal acoustique afin d'obtenir

de l'information sur les pauses et les durées. L'étude de ce mémoire s'est inspirée de cette technique.
Concernant le modèle prosodique seul, un arbre de décision de type classement et régression (CART [9]) est utilisé avec un entraînement supervisé pour prédire le type de frontière à une position donnée. Les décisions de l'arbre ne sont que faiblement conditionnées par la séquence de mots. Cela permet aux estimations du modèle prosodique d'être robuste face aux erreurs de reconnaissance. Les caractéristiques prosodiques utilisées sont la durée (des pauses, des rythmes et des voyelles finales) et l'intonation (F0 précédent et à travers les frontières, et sa variation en relation avec le locuteur).

Concernant le modèle hybride (prosodique et lexical), un modèle de Markov caché (HMM [56]) est utilisé toujours sous un entraînement supervisé. La variable cachée représente la chaine de classement, représentant les types de frontière. Le classement est effectué en trouvant la séquence d'états la plus probable du HMM étant donné les observations (mots et prosodie). Pour ce modèle, la segmentation des phrases repose sur un modèle de langage N-gramme à évènement caché ([60]).

Les résultats montrent que le modèle prosodique seul performe mieux pour une même quantité de données d'entraînement par rapport au modèle de langage basé sur les mots. Une réduction d'erreur allant jusqu'à 25% est observée en combinant les caractéristiques prosodiques avec les mots. Pour la segmentation des phrases, quatre groupes de caractéristiques utiles se sont démarqués (durée des pauses, différence de F0 à travers les frontières, changement de locuteur et durée des rimes). Pour l'identification des textes libres, les deux premiers groupes sont d'un plus grand intérêt.

1.2.3 Utilisation de la prosodie et d'un graphe d'hypothèses de mots

En plus de bénéficier de la connaissance au préalable des mots, il est intéressant d'approfondir ce concept et d'exploiter la probabilité d'informations prosodiques liées à chacune des hypothèses de mots. Nöth *et al.* [52] présentent un système de traduction du

discours nommé VERBMOBIL. Selon eux, c'est le premier système complet utilisant avec succès l'information prosodique dans l'analyse linguistique. La contribution la plus significative de la prosodie serait liée au NLU plutôt qu'à la phase d'ASR. En effet, l'oral spontané contient divers phénomènes tels que les constructions elliptiques (lorsqu'un mot ou une expression implicite par le contexte est omis dans une phrase), les interruptions et les redémarrages. Ces phénomènes favorisent grandement les ambiguïtés.

Les fonctions de la prosodie généralement considérées comme étant les plus significatives seraient les frontières de phrases, les accents toniques et les types de phrases (déclarative, impérative, interrogative et exclamative). La plus significative étant les frontières de phrases. Un aspect intéressant du système VERBMOBIL est que l'oral spontané de la vie réelle est utilisé, malgré que ce mémoire se concentre sur des commandes vocales. Le traitement de l'information prosodique est divisé en quatre étapes : l'extraction des caractéristiques prosodiques, la description des classes à être reconnues, le classement et l'amélioration des résultats de classement. Après avoir enregistré de l'oral spontané, un graphe d'hypothèses de mot (WHG [36]) est calculé à l'aide d'un HMM standard de reconnaissance de mots. Les hypothèses de mots dans ce WHG sont ensuite enrichies avec l'information prosodique.

L'approche utilisée dans VERBMOBIL, pour l'extraction de caractéristiques prosodiques, consiste en un module de prosodie dont l'entrée est la sortie du module ASR en plus du signal acoustique. Dans ce cas, l'alignement temporel du dispositif de reconnaissance et les informations sur les classes de phonèmes sous-jacents (comme les voyelles longues) peuvent être utilisés par le module de prosodie. Ce module peut ainsi utiliser la segmentation phonétique déterminée par le module ASR comme base pour l'extraction de caractéristiques prosodiques. Cette information segmentée correspond aux segments pour lesquels l'information prosodique doit être calculée, afin de marquer prosodiquement les hypothèses de mots. Les probabilités pour les accents toniques, les frontières de phrases, et les types de phrases sont attachées à chacune de ces hypothèses de mots.

Deux types de caractéristiques prosodiques sont obtenus à l'extraction. Les caractéristiques prosodiques « de base » sont extraites du signal acoustique pur sans aucune segmentation explicite en unités prosodiques, comme l'extraction basée sur les trames de F0 et de l'énergie. Habituellement ce type de caractéristiques ne peut pas être directement utilisé pour le classement prosodique. De leur côté, les caractéristiques prosodiques « structurées » sont calculées sur de plus grandes unités de parole (syllabe, noyau de syllabe, mot). Certaines d'entre elles sont basées sur les caractéristiques prosodiques de base. D'autres sont basés sur l'information segmentaire, qui peut être fournie à partir de la sortie d'un module ASR. Par exemple, les caractéristiques décrivant les propriétés de durées de phonèmes, syllabes, noyaux de syllabes et pauses. Un point de référence fixe est choisi correspondant à la fin des mots. Ces mots sont des unités bien définies dans l'ASR, et ce point est relativement facile à définir. Le début des mots a été considéré pour ce mémoire.

Un entraînement supervisé est utilisé pour le classement des évènements prosodiques. Les résultats se basent sur le classifieur Perceptron multicouches (MLP [23]), dont les performances sont supérieures aux classifieurs de distribution de Gaussienne et aux classifieurs polynomiaux des investigations similaires ([5], [35]). Selon l'expérimentation, le meilleur taux de reconnaissance est atteint lorsque tous les ensembles de caractéristiques sont combinés (durée, F0, énergie, pause, débit de parole et indicateur marquant si une syllabe annonce la fin d'un mot ou non ou désignant quelle syllabe porte l'accent tonique du mot). Les résultats de classement sont améliorés à l'aide d'un modèle de langage stochastique nommé Polygramme (PLM [61]). Ce modèle est un type de N-gramme où les historiques sont de longueur variable selon les données d'apprentissage disponibles.

Un marquage prosodique d'un WHG est utilisé, signifiant l'annotation des hypothèses de mots dans le graphe avec les probabilités pour les différentes classes prosodiques. Ces probabilités sont utilisées par les autres modules au cours de l'analyse linguistique. Dans le cas des frontières de phrases, au lieu de calculer leur probabilité d'être située à un certain

nœud du WHG, la probabilité est calculée pour une frontière d'être après ce mot, et ce pour chacune des hypothèses de mots.

Les résultats montrent que le PLM classe mieux les frontières par rapport au MLP. Cependant, la combinaison des deux classifieurs donne les meilleurs résultats (taux de reconnaissance de 94% en utilisant des chaînes de mots). Les taux de reconnaissance sont légèrement inférieurs sur les graphes d'hypothèses de mots par rapport aux chaînes de mots. Cela est dû à la sélection sous-optimale des mots selon le contexte. Peut-être qu'une combinaison des deux approches est possible et améliorerait les performances globales, mais les auteurs n'abordent pas le sujet. Pour le classement des WHG, MLP offre un taux de reconnaissance absolu de 77,5%, le PLM de 91,9%, et la combinaison des deux de 92,2%. La segmentation est peut-être la contribution la plus significative de la prosodie pour le NLU.

1.2.4 Un apprentissage semi-supervisé

Un inconvénient de l'apprentissage supervisé est qu'une grande quantité de données étiquetées est nécessaire afin d'obtenir de bonne performance. Il peut être très couteux et laborieux de créer une quantité raisonnable de données d'entraînement manuellement annotées prosodiquement. Une solution à ce problème est d'avoir recours à un apprentissage semi-supervisé. Ce type d'apprentissage requiert une petite quantité de données étiquetées, suivi d'une grande quantité de données non étiquetées.

Hun Jeon et Liu [27] utilisent l'algorithme Co-training ([7]) au cours d'un apprentissage semi-supervisé. Cet algorithme effectue le classement automatique d'événements prosodiques représentés symboliquement comme les accents toniques, les frontières de phrases et les pauses. Deux vues différentes sont utilisées, correspondant aux informations acoustiques et syntactiques/lexicales.

Co-training utilise l'ensemble d'entraînement initial pour apprendre un (faible) classifieur dans chaque vue (acoustiques et syntactiques/lexicales). Ensuite, chaque classifieur est appliqué à tous les exemples non étiquetés. Les exemples, sur lesquels chaque classifieur génère les prévisions les plus confiantes, sont sélectionnés. Ces exemples sont ensuite étiquetés avec les étiquettes de classe estimées et ajoutés à l'ensemble d'entraînement. Un nouveau classifieur est entraîné dans chaque vue basé sur le nouvel ensemble d'entraînement. L'ensemble du processus est répété pour quelques itérations. À la fin, une hypothèse finale est créée en combinant les prédictions acquises des classifieurs dans chaque vue.

Pour être efficace, Co-training requiert que les deux vues soient compatibles (tous les exemples doivent avoir la même étiquette basée sur la prédiction des deux vues) et non corrélées (pour chaque exemple x, les deux vues $V_1(x)$ et $V_2(x)$ sont indépendantes en considérant leur étiquette). Toutefois, les données réelles respectent rarement ces deux conditions. Les auteurs proposent donc un nouveau système d'étiquetage et de sélection pour résoudre ce problème.

L'annotation des événements prosodiques nécessite un système de représentation approprié pouvant caractériser la prosodie de manière standardisée. L'un des systèmes les plus populaires en matière d'étiquetage est l'approche ToBI ([69]). Avant sa création, certains auteurs définissent leur propre représentation ([13], [55]). Les phénomènes prosodiques les plus significatifs capturés dans l'approche ToBI comprennent les accents toniques et les frontières de phrases. Avec ToBI, les tonalités des accents toniques sont marquées à chaque syllabe accentuée. Elles peuvent correspondre à cinq types selon la hauteur de leur contour. Les tonalités des frontières de phrases sont marquées à chaque frontière des phrases intermédiaires, ou frontières de phrases d'intonation situées à certaines frontières de mots. Il y a également des indices de pauses à chaque frontière de mot. Pour réduire les graves problèmes de parcimonie des données, les auteurs regroupent les étiquettes ToBI en catégories grossières, réduisant du même coup l'ambiguïté de la tâche. Ce niveau grossier (présence versus absence) est appliqué au niveau des syllabes et est utilisé pour les accents toniques, les frontières de phrases et les pauses. L'approche est simple pour l'identification

des accents toniques, puisque l'accentuation se produit sur les syllabes. Dans le cas des frontières de phrases et des pauses, la dernière syllabe d'un mot est utilisée, puisque ces deux événements sont associés à des frontières de mots.

Pour générer les étiquettes des échantillons, l'algorithme utilise les seuils dynamiques des deux classifieurs. Cela permet d'attribuer des étiquettes avec précision et d'y inclure le plus grand nombre d'échantillons possible pour la sélection ultérieure. Les auteurs utilisent une nouvelle méthode de classement pour la sélection d'échantillons. Cette méthode considère la confiance d'une vue et le caractère informatif de l'autre vue. L'objectif de l'algorithme est de laisser la première vue sélectionner N échantillons pour un événement (classe positive ou négative). Cet événement sera inclus dans la prochaine itération pour l'entraînement de la deuxième vue.

Pour leur méthode d'identification d'évènements prosodiques, le problème est modélisé comme une tâche de classement binaire. Ils supposent que les observations acoustiques sont conditionnellement indépendantes des caractéristiques lexicales/syntaxiques étant donné l'étiquette prosodique. Les deux modèles sont donc développés séparément et la décision finale est obtenue en les combinant.

Pour le modèle acoustique, les auteurs supposent que chaque syllabe est indépendante des autres. Ainsi, les décisions sont prises pour chaque syllabe séparément. Pour ce modèle, une machine à vecteur de support (SVM [54]) avec un noyau de fonction à base radiale (RBF [31]) est utilisée. Ce classifieur démontre de meilleures performances par rapport aux autres classifieurs des études précédentes ([26]). Semblable à la plupart des travaux antérieurs ([32], [25]), ils utilisent les caractéristiques prosodiques pour ces tâches, incluant la hauteur, l'énergie et la durée. Les mêmes caractéristiques sont utilisées pour les trois tâches d'identification. La pause fait exception, puisqu'elle est utilisée seulement pour les deux tâches d'identification des frontières. Les résultats démontrent que l'utilisation de données

non étiquetées améliore significativement l'identification des événements prosodiques (des améliorations relatives de la F-mesure allant de 1% à 3% pour le Co-training par rapport à l'entraînement supervisé). Toutefois, les données proviennent de texte lu plutôt que de l'oral spontané.

1.3 La prosodie pour la désambiguïsation syntactique

Une fois recueillis et analysés, les indices prosodiques peuvent fournir de précieux indices pour résoudre les ambiguïtés syntactiques de l'oral spontané. Une ambiguïté syntactique, ou ambiguïté structurelle, se produit quand une expression ou une phrase a plus d'une structure sous-jacente. Par exemple, dans la phrase "La fille a frappé le garçon avec un livre". Structurellement, cette phrase peut être représentée de différentes manières : « [La fille a frappé le garçon] avec un livre » et « La fille a frappé [le garçon avec un livre] » ([33]).

Selkirk [62] postule que la structure prosodique de la phrase est liée (mais pas entièrement dépendante) à la structure syntaxique de surface. En revanche, certaines théories soutiennent que la prosodie est directement régie par la structure syntaxique de surface ([29]). Cependant, les preuves montrent plutôt que la relation syntaxe-prosodie est plus difficile, d'autant plus que des niveaux inférieurs de la hiérarchie prosodique se rapprochent.

Price *et al*. [57] effectuent des expériences de désambiguïsation pour étudier la relation entre la prosodie et la syntaxe en minimisant la contribution des autres indices. Des auditeurs sont soumis à des énoncés ambigus, produits dans un des deux paragraphes en contexte, un approprié pour chaque signification. Il leur est ensuite demandé de choisir le contexte approprié à partir des deux alternatives. Dans cette expérimentation, sept classes d'ambiguïtés syntactiques sont utilisées à partir de textes lus. Chaque phrase prononcée est alors étiquetée manuellement avec des frontières de phrases prosodiques et les proéminences. La durée et l'intonation en corrélation avec les différents marquages sont analysées. Les scores de désambiguïsation perceptuelle sont analysés avec les marquages prosodiques pour

déterminer quelles structures sont désambiguïsées, et quels indices phonologiques sont le plus souvent associés à ces structures.

Comme résultats, pour les paires de phrases correctement désambiguïsées, les différences les plus fréquentes sont la localisation et la largeur relative des frontières de phrases prosodiques. Toutefois, il y a quelques structures où la proéminence de phrase est soit le seul indice ou joue un rôle de support pour distinguer les deux versions. Cependant, pour élucider complètement la relation entre la prosodie et la syntaxe, cela nécessiterait l'investigation de beaucoup plus d'exemples de constructions syntactiques bien plus poussées que ne le permet cette étude. De plus, l'étude utilise du texte lu par des annonceurs professionnels de radio. Cela permet aux auteurs d'obtenir des résultats initiaux en utilisant bien moins de locuteurs par rapport à l'utilisation de locuteurs non professionnels. Les auteurs croient toutefois que les indices prosodiques seraient similaires pour des locuteurs non professionnels, bien qu'utilisés de façon moins consistante et pas marqués aussi clairement.

Dans sa thèse [47], Lieberman propose que les locuteurs puissent utiliser la prosodie pour désambiguïser des chaînes de mots ayant plusieurs structures de surface possibles. Il constate que sur les trois classes majeures d'ambiguïté (lexicale, liée à la structure de surface et liée à la structure profonde), seules les ambiguïtés de structure de surface peuvent être résolues par la prosodie. Wales et Toner [76] vérifient cette hypothèse. Ils utilisent dix paires de phrases pour chacune des trois classes. Ils constatent que les ambiguïtés de structure de surface sont la seule classe que les locuteurs et les auditeurs peuvent résoudre par des moyens prosodiques. Ils argumentent que la prosodie ne signale pas précisément la structure syntactique. Plutôt, le moyen de désambiguïsation est un marquage prosodique qui avertit les auditeurs de porter attention sur une signification moins soupçonnée. Les textes libres peuvent correspondre à un tel marquage.

1.4 Indices pour identifier les textes libres

1.4.1 La segmentation du discours

La segmentation automatique du discours peut représenter un intérêt pour l'identification des textes libres. En effet, un texte libre peut être considéré comme un sujet indépendant du reste de la commande vocale. La segmentation du discours consiste en la division automatique d'un flux textuel ou de parole en blocs homogènes reliés à un sujet ([3]). Étant donné une séquence de mots, l'objectif est d'identifier les frontières où il y a un changement de sujet. La segmentation du discours est utilisée dans des applications variées du NLU, comme l'extraction et la recherche d'informations ou la génération de résumés.

La prosodie peut jouer un grand rôle dans la segmentation du discours lorsqu'un flux de parole est concerné. La plupart des modèles théoriques de l'intonation supposent un ou plusieurs niveaux de phrases intonationnelles (segment de discours qui se produit avec un seul contour prosodique), sous lesquels la variation des caractéristiques de tonalité est interprétée. Intuitivement, les phrases intonationnelles d'un énoncé divisent celui-ci en fragments d'information significatifs. Une variation induite à une phrase intonationnelle peut modifier le sens dont les auditeurs seront susceptibles d'attribuer à un énoncé individuel d'une phrase. Enfin, il est démontré que les propriétés prosodiques d'une phrase intonationnelle sont corrélées avec les positions structurelles du discours ([20], [73]).

Hirschberg et Nakatani [21] effectuent l'identification des frontières des phrases intonationnelles à partir d'un ensemble restreint de caractéristiques prosodiques. Ils utilisent un arbre de décision de type CART comme technique d'apprentissage. La procédure de segmentation est entraînée et testée sur un corpus de texte lu. Toutefois, un corpus d'oral spontané produit par des non professionnels est également utilisé, ce qui se rapproche de l'étude de ce mémoire. Les enregistrements d'oral spontané sont transcrits manuellement, incluant les faux départs et autres erreurs de prononciation. Les transcriptions prosodiques de ces enregistrements sont également effectuées manuellement grâce au standard ToBI. Ces

transcriptions fournissent aux auteurs une décomposition des extraits de parole en phrases intonationnelles. Plusieurs mesures prosodiques sont utilisées comme mesures prédictives. Ces mesures permettent d'identifier si une trame de 10 msec d'un signal acoustique survient à l'intérieur ou à la frontière entre deux phrases intonationnelles. Cela correspond donc à un classement binaire. Dans leur expérimentation, quatre types d'information prosodique sont utilisés pour chaque trame : un estimé de F0, un indicateur binaire estimant la probabilité de voisement (pvoice), la moyenne de la racine carrée de l'énergie (rms) et la valeur normalisée de la corrélation croisée du pic (ac-peak) afin d'obtenir une autre estimation de F0. Pour ce mémoire, les informations prosodiques ont été extraites automatiquement à partir de Praat ([8]).

Hirschberg et Nakatani développent en plusieurs étapes les modèles permettant ce classement binaire à partir du corpus d'entraînement. Premièrement, un ensemble de caractéristiques prosodiques est identifié, permettant d'obtenir les meilleures performances sur une trame. Cet ensemble se base seulement sur cette trame avec au plus une autre trame en contexte. Ensuite, des modèles pour cet ensemble des meilleures caractéristiques sont entraînés sur chaque locuteur et chaque style du corpus. Ces modèles sont alors testés sur toutes les autres partitions. Cela engendre des modèles modélisant au mieux les autres données du corpus d'entraînement. Dans la seconde étape, les partitions des données d'entraînement sont utilisées pour le modèle du locuteur/style prédisant le mieux les autres données. Cette étape permet de sélectionner un ensemble distinct de caractéristiques contextuelles multi-trames, correspondant à une fenêtre variant entre 2 et 27 trames. Ces fenêtres de trames sont alignées sur la trame courante de différentes façons (alignement à gauche, au centre et à droite). La largeur de fenêtre, obtenant les meilleures performances, est retenue à cette étape. À la troisième étape, la meilleure combinaison des caractéristiques est identifiée. Ces caractéristiques sont basées sur les fenêtres à trames simples et multiples. Finalement, ce modèle composé est testé dans une expérimentation, afin d'inférer des structures de discours

à partir des caractéristiques prosodiques. Il est ensuite testé dans une autre expérimentation utilisant les frontières prédites de phrases pour des applications de navigation audio.

Le modèle composé final est déterminé suite à l'entraînement. Celui-ci considère les modèles à trames simples et multiples offrant les meilleures performances. Il inclut deux caractéristiques à trames multiples et une caractéristique à trames simples : 15 trames d'une fenêtre centrée de la moyenne normalisée de rms, 19 trames d'une fenêtre alignée à gauche de la moyenne normalisée de F0, et ac-peak de la trame courante seule. Ayant recours à la validation croisée lors de l'entraînement, les performances de l'approche CART varient entre 80% et 93%. L'ajout de caractéristiques à trames multiples améliore la précision de classement de 2% à 5%. Ces caractéristiques représentent l'information du contexte de la trame courante. Cependant, aucune caractéristique basée sur une fenêtre alignée à droite n'est utile dans ce modèle composé final. L'expérimentation entière suggère que l'identification de phrases intonationnelles par des moyens purement automatique est possible. Cela semble prometteur pour l'identification automatique des textes libres.

D'autres auteurs s'intéressent à la combinaison de la prosodie et des mots pour la segmentation automatique du discours. Shriberg *et al.* [68] utilisent l'arbre de décision CART et un HMM pour combiner les indices prosodiques avec le modèle de langage. Selon leurs résultats, le modèle prosodique est équivalent ou supérieur au modèle de langage. De plus, ce modèle prosodique requiert moins de données d'entraînement. Un point intéressant est que ce modèle ne nécessite pas d'annotations manuelles de la prosodie. Comme pour les études discutées précédemment, les performances du système s'améliorent en combinant l'information prosodique et lexicale. Pour cette section du chapitre, l'intérêt se porte seulement sur leur modèle prosodique.

Leur travail comporte trois tâches : la segmentation de phrases sur le corpus Broadcast News, la segmentation de phrases sur le corpus Switchboard et la segmentation de sujet sur le corpus Broadcast News. L'intérêt de cette section se situe davantage sur la segmentation de sujets. Comme pour ce mémoire, le corpus Broadcast News contient principalement des

monologues. Cependant, ce corpus est constitué de textes lus contrairement au corpus Switchboard qui est constitué d'oral spontané. Toutefois, Switchboard est constitué de dialogues et contient souvent des chevauchements de locuteurs. Pour ce mémoire, les commandes vocales étaient toujours initiées par un seul locuteur. Les caractéristiques de pauses et de hauteur sont hautement informatives pour la segmentation de Broadcast News. Pour Switchboard, les durées et les indices basés sur les mots sont dominants. Par conséquent, une attention doit tout de même être portée sur la segmentation de phrases.

Ainsi, les auteurs constatent que les caractéristiques de durée et celles extraites du modèle du langage sont particulièrement utiles pour la segmentation de la conversation naturelle. D'autres indices prosodiques inclus les pauses, les changements dans l'intervalle de la hauteur et de l'amplitude, la déclinaison globale de la hauteur, la mélodie et la distribution de la tonalité aux frontières, et la variation du débit de parole. Par exemple, les frontières de phrases et de paragraphes ou les frontières de sujets sont souvent marquées. Ce marquage se caractérise par une combinaison d'une longue pause, précédée d'une frontière finale de faible tonalité, suivi d'une réinitialisation de l'intervalle de la hauteur, parmi d'autres caractéristiques ([44], [10], [11]).

Pour toutes les tâches abordées, les auteurs utilisent des caractéristiques prosodiques très locales. Les auteurs font ce choix pour des raisons pratiques. Ce choix a également été considéré pour ce mémoire. Pour chaque frontière entre deux mots, précisément le type de frontière d'intérêt, les auteurs s'intéressent aux caractéristiques prosodiques du mot précédant et suivant immédiatement la frontière. Alternativement, une fenêtre de 20 trames est utilisée précédant et suivant cette frontière. Dans le cas des frontières contenant une pause, la fenêtre s'étend vers l'arrière du début de cette pause ainsi que vers l'avant de sa fin. Une région peut-être plus efficace, non considérée par les auteurs, serait d'étendre la fenêtre vers l'arrière et vers l'avant jusqu'à ce qu'une syllabe accentuée soit atteinte. Cette idée est particulièrement intéressante pour les locuteurs anglophones, dont s'intéresse également et

exclusivement ce mémoire. Contrairement à la langue française, la langue anglaise est considérée comme une langue d'accentuation variable. Toutefois, pour des raisons pratiques, cette idée n'a pas été considérée pour ce mémoire.

Les caractéristiques extraites concernent la durée des pauses et des phonèmes ainsi que les informations sur la hauteur et sur la qualité de la voix. Les pauses sont extraites sur la frontière entre deux mots. Les autres caractéristiques sont extraites principalement à partir du mot (ou de la fenêtre) précédent la frontière (ce qui rappelle les résultats de Hirschberg et Nakatani [21] avec leur modèle composé final discuté précédemment). Les travaux antérieurs démontrent que les caractéristiques précédant les frontières détiennent plus d'information pour ces tâches par rapport au flux de parole suivant les frontières [64]. Il est intéressant de constater que les auteurs décident de ne pas s'intéresser aux caractéristiques basées sur l'énergie ou l'amplitude. En effet, ces caractéristiques sont moins fiables et largement redondantes face à celles des durées et des hauteurs. Ces deux observations ont orienté l'étude de ce mémoire.

Un aspect intéressant de la pause est que les auteurs incluent également sa durée précédant le mot avant la frontière. Cela permet de déterminer si le discours précédant la frontière débute ou s'il est la continuation du discours précédant. Comme pour les données de ce mémoire, la plupart des frontières entre les mots ne contiennent aucune pause. Dans ce cas, la valeur attribuée pour sa durée est alors de zéro. Les auteurs considèrent les pauses selon deux mesures : leur durée brute et leur durée normalisée (pour la distribution de leur durée pour un locuteur particulier). Les modèles des auteurs sélectionnent automatiquement les durées brutes plutôt que les durées normalisées. Cela est possiblement dû au manque de durée de pause suffisante par locuteur. Néanmoins, les mesures des durées brutes semblent suffisantes. Pour ce mémoire, la normalisation a été utilisée telle qu'expliquée à la section 3.3.

Concernant les durées des phonèmes, une caractéristique bien connue pour les frontières de discours est le ralentissement vers la fin des unités, soit l'allongement préfrontière. Cela affecte typiquement le noyau syllabique et la coda des syllabes. Les auteurs incluent donc

une mesure qui reflète la caractéristique de durée de la dernière rime (noyau syllabique et coda) de la syllabe précédant la frontière. Chaque phonème dans la rime est normalisé pour sa durée inhérente. Un atout de cette normalisation est que l'arbre de décision est en mesure d'utiliser certaines caractéristiques spécifiques comme indice pour l'identité des mots. Ces indices sont utilisés indirectement pour identifier les frontières du discours. Pour ce mémoire, le début des textes libres est peut-être corrélé avec certains mots le précédant. Un aspect très intéressant est que les auteurs font la distinction entre les phonèmes des pauses pleines (comme « um » et « uh ») de ceux trouvés ailleurs dans le discours. Il est démontré que la durée des phonèmes des pauses pleines est considérablement plus longue par rapport à celles trouvées dans les autres voyelles de spectre similaire [63]. Il est possible que les textes libres soient parfois précédés d'une pause pleine. La section 2.4.2 décrit un algorithme qui a permis de considérer ce problème.

Les auteurs affirment que les informations sur F0 sont typiquement moins robustes et plus difficiles à modéliser par rapport aux autres caractéristiques, comme les durées. Cela est largement attribuable à la variabilité dans la façon dont la hauteur est utilisée à travers les locuteurs et le contexte des discours, et autres facteurs. Les auteurs effectuent un post-traitement de la sortie de F0 au niveau de la trame, à partir d'un capteur standard de hauteur basé sur l'autocorrélation (fonction « get_f0 » de ESPS/Waves[1]). Ce post-traitement permet de lisser les erreurs de micro intonation et de repérage, de simplifier leur calcul de F0 et d'identifier les paramètres d'intervalle de discours pour chaque locuteur. Cela génère des estimations de F0 au niveau des trames. Un capteur de hauteur similaire est sans doute nécessaire pour l'étude de ce mémoire afin de gérer des problèmes semblables. Au total, quatre différents types de caractéristiques F0 sont calculés. Chaque type capture un aspect différent du comportement intonationnel : 1) La réinitialisation de F0 (tendance d'un locuteur à réinitialiser la hauteur au début d'une nouvelle unité majeure), 2) L'intervalle de F0

[1] http://www.speech.kth.se/software/#esps

(intervalle de la hauteur dans un mot ou fenêtre), 3) La pente de F0 (pente engendrée pour un mot ou fenêtre sur un seul côté de la frontière, et pour la continuité à travers la frontière), 4) La continuité de F0 (mesure du changement de pente à travers la frontière). Finalement, des caractéristiques de la qualité de la voix sont considérées par les auteurs. Cependant, celles-ci sont principalement dépendantes du locuteur et ne présentent donc pas d'intérêt pour ce mémoire. Toutes ces caractéristiques prosodiques servent d'entrées à l'arbre de décision. Cet arbre prédit le type approprié de frontière de segment à chaque frontière intermot.

1.4.2 Les actes de dialogue

Suite à la segmentation du discours, un domaine encore plus près de l'identification des textes libres est sans doute celui de la segmentation et du classement des actes de dialogue (AD). Un AD est une composante spécifique de la parole, marquant un morceau de dialogue en fonction de sa catégorie de signification. Des exemples comprennent la question, la requête, le remerciement et la citation (cette dernière étant la plus rapprochée du texte libre et est discutée à la section 1.4.3). Le nombre d'AD varie d'un système à l'autre. Dans certains systèmes, ce nombre peut atteindre jusqu'à plus de 40 ([65]). Plusieurs auteurs traitent ces AD indépendamment du locuteur. Cette section résume brièvement leurs travaux.

Beaucoup d'auteurs s'intéressent à la combinaison des modèles prosodiques et lexicaux pour le classement des AD. En 1996 avec le système VERBMOBIL, Mast *et al.* [50] présentent des méthodes automatiques de segmentation et de classement d'AD grâce à la prosodie. La segmentation est réalisée grâce à des PLM et à des MLP. Le classement est réalisé à l'aide de PLM et d'arbres de classements sémantiques (SCT). Les auteurs définissent 18 catégories d'AD et 42 sous-catégories. Ceux-ci croient que les SCT sont mieux adaptés au classement des AD par rapport au PLM. En effet, les SCT peuvent modéliser des dépendances de longues distances de façon plus étendue. Cependant, la quantité de données d'entraînement n'est pas suffisante pour l'affirmer avec certitude. Les meilleurs résultats de classement sont atteints avec la combinaison des deux modèles. L'erreur est réduite de plus de 19.0% par rapport au PLM seul. Les meilleurs résultats atteignent 61.7% pour le classement des AD des

segments de mots segmentés automatiquement. Ce résultat est même plus élevé par rapport au taux de 59.7% obtenu avec des données segmentées manuellement.

Warnke *et al.* [77] continuent ces travaux sur le système VERBMOBIL. Ils intègrent la segmentation et le classement dans l'algorithme A*. Cela permet de chercher la segmentation et le classement optimal des AD sur la base de WHG. Les hypothèses pour les frontières des segments sont calculées à l'aide d'un PLM. Un MLP est utilisé pour classer les caractéristiques prosodiques. Le classement des AD est effectué par un PLM basé sur des catégories pour chaque AD. Grâce à cette intégration, la reconnaissance des AD s'est considérablement améliorée. Cependant, les résultats ne peuvent être comparés à [50]. En effet, certaines catégories d'AD ne peuvent pas être modélisées, puisque le corpus d'entraînement ne contient pas suffisamment de représentations. Les auteurs concluent que cette intégration est la seule approche utile pour déterminer la séquence d'AD sur la base d'un WHG. Enfin, les auteurs prévoient démontrer que l'algorithme A* est approprié pour la reconnaissance d'AD sur des WHG reconnus automatiquement.

Stolcke *et al.* [71] présentent une approche probabiliste intégrée pour modéliser statistiquement la structure du discours pour la parole de conversation naturelle. Cette approche est utilisée pour le classement de 42 AD, combinant un modèle prosodique et lexical et une grammaire statistique du discours. Pour le classement prosodique, les arbres de décisions et les réseaux de neurones sont explorés. Les résultats démontrent une large indépendance face à la technique de modélisation. Une précision d'étiquetage de 65% des AD est atteinte basée sur la prosodie et les mots reconnus. Une précision de 72% est atteinte basée sur la transcription des mots. Ces travaux sont davantage détaillés dans [70]. Pour de futurs travaux, les auteurs prévoient explorer l'identification d'AD en utilisant des arbres de décisions multiples afin d'augmenter la robustesse. Shriber *et al.* [65] s'intéressent également à la conversation naturelle. Un arbre de décision est utilisé pour le classement des AD. Les performances sont évaluées sur le modèle prosodique seul et sur celui combiné avec

l'information lexicale. L'évaluation est effectuée à partir de la transcription, puis à partir de la sortie du module ASR. Les résultats démontrent que l'information prosodique donne une contribution significative pour le classement. L'intégration du modèle prosodique et du modèle de langage (LM) statistique spécifique à l'AD améliore les performances face au LM seul. Cette amélioration est plus prononcée pour le cas des mots reconnus.

En plus de s'intéresser aux informations prosodiques et lexicales, certains auteurs s'intéressent également à la syntaxe pour aider à l'identification des AD. Jurafsky *et al.* [28] s'intéressent à l'identification de deux types d'AD, soit les rétroactions et les remerciements, ainsi que cinq autres AD associés. Ils démontrent que les connaissances lexicales jouent un rôle dans la distinction des cinq AD, malgré la grande ambiguïté de certains mots. Les indices prosodiques jouent un rôle dans l'identification de certains types d'AD. Les indices lexicaux sont suffisants pour les autres types. L'identification de certains AD est également facilitée par les connaissances syntaxiques. Rangarajan *et al.* [59] proposent un cadre discriminatoire pour l'identification des AD utilisant la modélisation du maximum d'entropie. Ils proposent deux schémas pour l'intégration de la prosodie dans leur cadre de modélisation : 1) La prédiction de la prosodie par catégorie basée sur la syntaxe, à partir d'un étiqueteur prosodique automatique, 2) Une nouvelle méthode pour modéliser les séquences d'observations prosodiques, comme étant une séquence discrète par le biais de moyens de quantification. L'intégration des caractéristiques prosodiques engendre une amélioration de 11.8% par rapport aux caractéristiques lexicales et syntaxiques seules. La combinaison des trois types de caractéristiques engendre une précision de 84.1%, semblable au consensus humain de 84%.

Enfin, Sridhar *et al.* [58] utilisent également la modélisation du maximum d'entropie. Ils montrent que celle-ci permet d'obtenir une performance supérieure aux approches conventionnelles, de même que la séquence de valeurs prosodiques comme caractéristiques d'un N-gramme. Cette modélisation utilise une représentation grossière du contour prosodique par le biais de statistiques sommatives de ce contour. Leur schéma, incluant l'exploitation de la prosodie, résulte en une amélioration absolue de 8.7%. Ceci est par

rapport à l'utilisation de la plupart des autres représentations largement utilisées de l'acoustique corrélée à la prosodie. L'approche des auteurs est différente des systèmes d'AD traditionnels. Ceux-ci utilisent seulement des caractéristiques statiques. Ils approximent l'identification de l'AD précédent en termes d'information prosodique, lexicale et syntaxique extraite de l'énoncé précédent. Leur expérimentation obtient une précision d'identification de 72%, combinant les trois types de caractéristique.

Finalement, certains auteurs s'intéressent seulement à l'information prosodique pour identifier les AD. Fernandez et Picard [16] utilisent les SVM et des techniques d'apprentissage discriminatif pour le classement de huit AD. Ils obtiennent un taux de reconnaissance de 47.3%. Ce taux représente une amélioration par rapport aux travaux antérieurs utilisant les arbres de décisions et les MLP. Les caractéristiques prosodiques comprennent plusieurs mesures liées à l'intonation, l'intensité et la durée. Les auteurs admettent que les informations prosodiques à elles seules ne suffisent pas pour un classement robuste des AD, comme démontré précédemment. Néanmoins, les SVM représentent tout de même une alternative intéressante. Toutefois, malgré que le corpus utilisé soit constitué d'oral spontané, celui-ci est de langue espagnole. Il peut donc représenter des propriétés prosodiques différentes de la langue anglophone considérée pour ce mémoire. Ces différences peuvent survenir notamment au niveau de la localisation des accents toniques dans les phrases. Enfin, Laskowski [38] présente un système basé sur des HMM avec de longues trames acoustiques, rendant maniable la modélisation du contexte prosodique. Le système est ensuite étendu par des caractéristiques prosodiques concaténées et calculées pour des trames temporellement proches. Les caractéristiques concernent l'intensité, la qualité de la voix, le débit de parole et la variation de la hauteur. Les résultats indiquent que l'augmentation de la taille des trames, ainsi que le contexte prosodique d'un locuteur cible, améliorent les performances d'identification. Pour ce mémoire, le contexte du locuteur ne peut être considéré. En effet, les commandes vocales sont courtes et émises par des locuteurs potentiellement différents.

1.4.3 La citation

La citation, ou discours rapporté, est sans doute l'AD le plus près du texte libre. La citation est le meilleur candidat trouvé dans la table des AD originaux présentée en [65]. Elle est un mécanisme par lequel nous pouvons citer. Elle est normalement utilisée pour l'attribution des mots et des pensées précises à d'autres. Comme pour les textes libres, elle représente un texte à considérer tel quel. Si les expressions déictiques (mots ou expressions qui déterminent les conditions particulières de l'énonciation) sont ancrées en partie ou en totalité à la situation citée, elles suffisent à elles seules à signaler le discours comme étant rapporté en l'absence d'une phrase rapportée. Cependant, avec ou sans phrase rapportée et transposition déictique, le discours peut être marqué comme rapporté par la convention de guillemets. Leech et Short [42] divisent le discours rapporté en trois catégories principales : le discours direct (pour signaler que quelqu'un a dit une citation mot pour mot), le discours indirect (pour signaler qu'on exprime ce qui a été dit dans nos propres mots) et le discours direct libre (les personnages semblent nous parler plus immédiatement, sans narrateur comme intermédiaire). Ces catégories ont inspiré le classement des textes libres présenté à la section 3.2.

Klewitz et Couper-Kuhlen [34] démontrent comment les locuteurs anglophones peuvent utiliser les indices prosodiques d'un discours afin d'identifier une partie comme étant rapportée. Ils démontrent également que les changements prosodiques peuvent remplir la fonction des guillemets du texte écrit, en délimitant clairement les frontières de gauche et de droite de la citation. Il est souhaitable que ces mêmes changements puissent délimiter les frontières des textes libres. Toutefois, dans la majorité des cas, les changements prosodiques ne coïncident pas avec ces frontières et se produisent plutôt à proximité. Ces changements fonctionnent comme un cadre pour l'interprétation d'une séquence comme étant rapportée. Ils peuvent simplement correspondre à un drapeau invitant l'auditeur à reconstruire les limites correspondantes. Parmi les indices prosodiques et paralinguistiques les plus fréquemment observés, sont constatés des changements de la hauteur globale (registre), de l'intensité et du débit de parole. La première constatation des auteurs est que les frontières sont souvent accompagnées d'un changement notable du registre ou de l'intervalle de la

hauteur. Des passages de discours rapportés peuvent également être marqués par de multiples changements prosodiques se produisant simultanément dans un groupe. Ce groupe n'étant pas une collection aléatoire. Les marquages prosodiques coïncident également avec des citations non prévues. S'il n'y a pas de verbe d'introduction des citations, ce marquage peut être le seul indice extérieur de la nature citative d'un étirement de parole. Ce marquage ne se limite pas qu'au discours direct. En effet, le discours indirect peut avoir un fort marquage expressif qui est assez répandu, en particulier dans les discours à forte participation. Les guillemets sont réservés aux discours directs. Les marquages prosodiques peuvent être présents sous toutes les formes de déclaration verbale. Toutefois, ce ne sont pas tous les discours directs ou indirects qui sont marqués prosodiquement. En effet, les citations dans les discours peuvent ne recevoir aucun formatage prosodique particulier. L'absence de marquage peut être un choix de style conscient de la part du locuteur. Enfin, contrairement aux guillemets, le marquage prosodique ne se limite pas à la citation. Il se produit également ailleurs dans le discours, signalant d'autres dimensions structurelles et expressives.

Oliveira *et al.* [53] s'intéressent également au rôle de la prosodie comme marqueur de frontières pour le discours direct, mais plus spécifiquement pour identifier la fin des citations. Cet article est particulièrement intéressant pour l'identification de la fin des textes libres. Celle-ci est sans doute une tâche plus difficile par rapport à l'identification de leur début. Le début d'une citation est souvent linguistiquement marqué, généralement au moyen d'un verbe de citation. Cependant, il est plus difficile de déterminer à quel moment elle se termine. Les auteurs analysent les données grâce au programme d'analyse de signaux Praat. Leurs résultats indiquent que les pauses se produisant à la fin d'un discours direct sont en général plus longues (d'environ 100 msec) par rapport à celles se produisant ailleurs. Toutefois, la pause à elle seule ne semble pas suffisante pour marquer la frontière d'une citation. Ils constatent également que les frontières intonatives basses sont beaucoup plus utilisées à la fin d'une citation qu'ailleurs. De même, des valeurs plus élevées de la réinitialisation de la hauteur correspondent à la fin des citations. Parmi toutes les caractéristiques prosodiques

disponibles à l'auditeur, la réinitialisation de la hauteur est la plus significative. Enfin, des valeurs plus élevées de la différence d'intensité entre les unités d'intonation correspondent à la fin d'un segment de citation.

De leur côté, Jansen *et al.* [24] examinent si la prosodie établit une distinction entre le discours direct et indirect. Les discours directs sont introduits avec des verbes de citation, sans autre marquage lexical. Cependant, les discours indirects sont souvent impossibles à distinguer lexicalement des discours directs. Toutefois, les discours indirects sont optionnellement marqués par un complément précédant la clause reportée. Selon eux, puisque la distinction entre le discours direct et indirect n'est pas toujours marquée par des moyens lexicaux ou syntaxiques, cette distinction est similairement reflétée par la prosodie. Ils constatent que le discours direct est signalé par un plus grand intervalle de la hauteur globale par rapport au contenu narratif environnant. Il est également généralement précédé par des frontières de phrases intonationnelles. En revanche, la prosodie ne semble pas distinguer le discours indirect du discours narratif ordinaire. Pour vérifier s'il y a bien une distinction, les auteurs annotent et analysent les segments de parole avec Praat. L'étiquetage des caractéristiques d'accents et de frontières est effectué selon la convention du guide d'étiquetage *Prosodic Ostendorf & Shattuck-Hufnagel* (POSH). Ils établissent que toutes les citations sont contenues dans une seule phrase intonationnelle. Ils transcrivent également les phrases intonationnelles voisines lorsque celles-ci sont contenues dans le même énoncé. En effet, le comportement de nombreux indices prosodiques serait contraint au niveau de la phrase intonationnelle [37]. Trois grandes caractéristiques prosodiques sont examinées: l'intervalle de la hauteur, le niveau de hauteur globale comme reflétée par la hauteur moyenne et les pauses prosodiques. Les moyennes des intervalles de la hauteur indiquent que l'intervalle global est beaucoup plus grand pour les citations directes par rapport aux citations indirectes. Les intervalles moyens de la hauteur des phrases intonationnelles, précédent celles contenant des citations directes et indirectes, suggèrent qu'il y a également une différence considérable entre les citations directes et indirectes. Cette différence se situe dans la quantité de réinitialisations de la hauteur par rapport à une phrase précédente. Selon leur corpus, les citations directes sont deux fois plus susceptibles d'être précédées par une pause de phrase

intonationnelle par rapport aux citations indirectes. Aucune différence similaire n'est observée du côté droit des citations pour la distribution des pauses prosodiques.

Finalement, Kasamir [30] étudie la corrélation entre la prosodie et les énoncés entre guillemets de façon totalement manuelle. L'auteur constate que les textes lus à voix haute, contenant des passages fermés par des guillemets, diffèrent sensiblement au niveau prosodique par rapport à leurs contreparties. Cependant, cette corrélation n'est pas suffisante pour survivre à la retraduction ultérieure dans le langage écrit. La présence ou l'absence des énoncés entre guillemets influence considérablement leur réalisation prosodique. Cependant, un tel marquage prosodique ne peut réellement être un substitut de la parole pour remplacer les guillemets à l'écrit. Toutefois, cette constatation peut être due à une erreur de performance induite par les lectures de texte à voix haute. Malgré un manque au niveau de l'analyse phonétique automatique, l'auteur suggère qu'au moins quatre stratégies peuvent être distinguées: une pause principale, une pause traînante, un changement dans la qualité de la voix (l'allongement des voyelles accentuées) et le déplacement des accents de hauteur. En fait, les locuteurs ne semblent pas compter sur une stratégie unique pour le marquage prosodique.

1.5 Conclusion

Ce chapitre a présenté les caractéristiques prosodiques comme étant susceptibles d'aider à l'identification des frontières entre les phrases, les paragraphes, les sujets et entre les noms propres. Ces caractéristiques permettent également de résoudre des ambiguïtés syntaxiques et de classer des AD, dont la citation qui est l'AD le plus près du texte libre. Pour toutes ces tâches, les études démontrent que les systèmes obtiennent de meilleures performances en combinant les modèles prosodiques et les modèles lexicaux. Pour les modèles prosodiques, le classement à l'aide d'un arbre de décision est une option très intéressante. Celui-ci peut être utilisé avec n'importe quelle séquence d'information prosodique, pouvant même être

combiné avec des informations lexicales. Les HMM, les SVM et les MLP sont également utilisés. Pour les modèles lexicaux, les HMM, les PLM, les SCT et les WHG sont utilisés.

Pour entraîner les classifieurs de mots ou de trames à partir d'un énoncé, un apprentissage supervisé ou semi-supervisé est nécessaire. Les modèles lexicaux peuvent être annotés manuellement, ou la sortie du module ASR peut être considérée. Pour les modèles prosodiques, il est recommandé d'effectuer l'annotation manuelle selon un standard connu comme ToBI ou POSH. Néanmoins, plusieurs auteurs ont extrait les caractéristiques prosodiques automatiquement à l'aide d'outil comme Praat. Cet outil a été utilisé pour ce mémoire et est présenté à la section 2.3.1.

En plus d'améliorer les performances en combinant les modèles prosodiques et lexicaux, les performances des modèles prosodiques sont nettement meilleures lorsqu'une combinaison de caractéristiques est utilisée. Il semble que même l'humain utilise cette stratégie. Les caractéristiques précédant les frontières ont en général plus de valeur par rapport à celles suivant les frontières. Parmi les différents types de caractéristiques, il semble que les informations sur F0 soient moins robustes et plus difficiles à modéliser par rapport aux autres caractéristiques, comme les pauses. Les informations intéressantes sur F0 concernent surtout son intervalle, sa pente, sa réinitialisation et sa continuité. Ces deux dernières ont été utilisées pour ce mémoire et sont présentées à la section 4.3.

Enfin, la citation se divise en trois catégories pouvant inspirer la classification des textes libres. Les indices prosodiques peuvent aider à localiser les frontières de gauche et de droite d'une citation. Les indices les plus fréquents concernent les changements de la hauteur globale, de l'intensité et du débit de parole. En particulier, concernant la localisation des frontières de droite, sont constatées des pauses plus longues, des frontières intonatives basses plus fréquentes, des valeurs plus élevées de la différence d'intensité entre les unités d'intonation, mais surtout des valeurs plus élevées pour la réinitialisation de F0. Finalement, le Tableau 1.1 résume les notions, idées ou concepts présentés dans ce chapitre et indique si ceux-ci ont été retenus pour ce mémoire.

Tableau 1.1 Notions, idées ou concepts retenus pour ce mémoire

Notion, idée ou concept	Retenu ?
Utilisation de textes lus	
Utilisation d'oral spontané	X
Utilisation de classes prosodiques	
Entraînement supervisé	X
Combinaison de la prosodie et des mots pour extraire d'autres informations	X
Alignement automatique de la transcription par rapport au signal acoustique	X
Exploitation des probabilités liées aux hypothèses de mots	
Utilisation de la sortie du module ASR en plus du signal acoustique	X
Utilisation de la segmentation phonétique déterminée par le module ASR	X
Utilisation de la fin des mots comme point de référence	X
Entraînement semi-supervisé	
Considération des faux départs et des erreurs de prononciation du locuteur	
Informations prosodiques extraites à partir de Praat	X
Sélection automatique des largeurs de fenêtre optimales	
Sélection automatique des meilleures combinaisons des caractéristiques	
Utilisation de caractéristiques à trames multiples	
Utilisation de corpus constitués de monologues	X
Utilisation de caractéristiques prosodiques locales	X
Caractéristiques prosodiques du mot précédant et suivant la frontière	X
Utilisation de fenêtres étendues jusqu'à atteindre une syllabe accentuée	
Utilisation de la durée de la pause précédant le mot avant la frontière	
Utilisation de la durée normalisée des pauses	X
Distinction entre les phonèmes des pauses pleines et les autres	

Utilisation d'un post-traitement de la sortie de F0 au niveau de la trame	
Utilisation d'une grammaire statistique du discours	
Utilisation de la syntaxe en plus de l'information prosodique et lexicale	
Utilisation du contexte du locuteur	
Utilisation des catégories de la citation pour les textes libres	X
Utilisation de la fin des citations ou des textes libres	
Distinction entre le discours direct et indirect transposée aux textes libres	X
Étude manuelle de la corrélation entre la prosodie et les textes libres	

CHAPITRE 2

MÉTHODOLOGIE EXPÉRIMENTALE

2.1 Introduction

Afin de déterminer si l'information prosodique contenue dans les signaux acoustiques peut aider le système NLU à améliorer ses performances, ce chapitre présente la méthodologie utilisée. L'identification des textes libres est au cœur de ce mémoire, puisqu'ils correspondaient à une intuition à propos d'une corrélation possible. Toutefois, le véritable intérêt concerne l'amélioration globale d'un système NLU grâce à cette information. Pour atteindre ces deux objectifs de ce mémoire, il a été essentiel d'utiliser un grand ensemble de commandes vocales afin d'y effectuer les analyses et expérimentations. En plus de ces données audio, il a été essentiel d'utiliser des outils appropriés pour en extraire l'information pertinente. La qualité des mesures recueillies a dû être évaluée. En effet, il était nécessaire de connaître la confiance pouvant être accordée à ces outils. Les mesures de l'expérimentation entière ont découlé de ceux-ci. Enfin, les conclusions finales de l'expérimentation ont découlé de ces mesures.

Ce chapitre présente tout d'abord les corpus de données utilisés pour cette étude. Ensuite, les outils de mesure sont abordés, suivis de l'évaluation de leur qualité. Cette évaluation est au cœur de ce chapitre et la présentation, l'interprétation et la discussion des résultats en font partie. Enfin, la conclusion présente les résultats de l'analyse déterminant s'il est possible de faire confiance à ces outils pour l'ensemble de l'expérimentation.

2.2 Corpus de données

Les corpus de données audio ont été divisés en deux petits échantillons extraits de façon aléatoire (échantillons E-AT et E-AC) et deux plus grands échantillons extraits de façon

ciblée (échantillons E-CT et E-CC). Ces données consistaient en des commandes vocales prononcées par des locuteurs anglophones potentiellement différents pour chaque enregistrement, hommes et femmes. Pour chaque enregistrement, la transcription textuelle ainsi que l'annotation de chaque mot étaient disponibles. Ces annotations lexicales indiquaient la mention, c.-à-d. la sémantique assignée à chaque mot d'intérêt. Par exemple, pour une commande vocale contenant un texte libre, la mention « text » ou « title » était assignée à chaque mot constituant ce dernier, correspondant principalement au corps et au titre des courriels et texto. Toutefois, les mots de moindre intérêt ne possédaient pas de mention.

Les échantillons ciblés étaient constitués de 39 584 enregistrements transcrits et annotés automatiquement par le pipeline d'un système NLU, dont les transcriptions ont été corrigées manuellement. Pour effectuer l'annotation, ce système a utilisé ses anciennes prédictions conservées dans le journal des utilisateurs. Ces échantillons étaient constitué de 10 621 mots débutant un texte libre (10 621 enregistrements) et 97 682 mots d'enregistrements ne contenant pas de texte libre (28 963 enregistrements). Ces échantillons étaient légèrement biaisés, puisqu'ils n'ont pas été extraits de la population d'une façon entièrement aléatoire. En effet, les enregistrements des requêtes de « dictation » ont été exclus. On nomme ainsi les requêtes qui n'étaient jamais envoyées au pipeline du système NLU en production, c.à.d. qu'elles étaient seulement traitées par le système ASR. Après cette exclusion, les autres enregistrements devaient satisfaire des critères très peu contraignants.

Les échantillons étaient disponibles, mais ceux-ci ne contenaient pas suffisamment d'enregistrements pour fonder cette étude. L'utilisation des échantillons ciblés ont permis de contourner cet obstacle. Malgré leur petite taille, les expérimentations ont tout de même également été effectuées sur les échantillons aléatoires. Ceux-ci étaient constitués de 14 022 enregistrements transcrits et annotés manuellement. Ils étaient constitués de 681 mots débutant un texte libre (681 enregistrements) et 40 017 mots d'enregistrements ne contenant pas de texte libre (13 341 enregistrements). La Figure 2.1 et la Figure 2.2 présentent les proportions des échantillons ciblés et aléatoires.

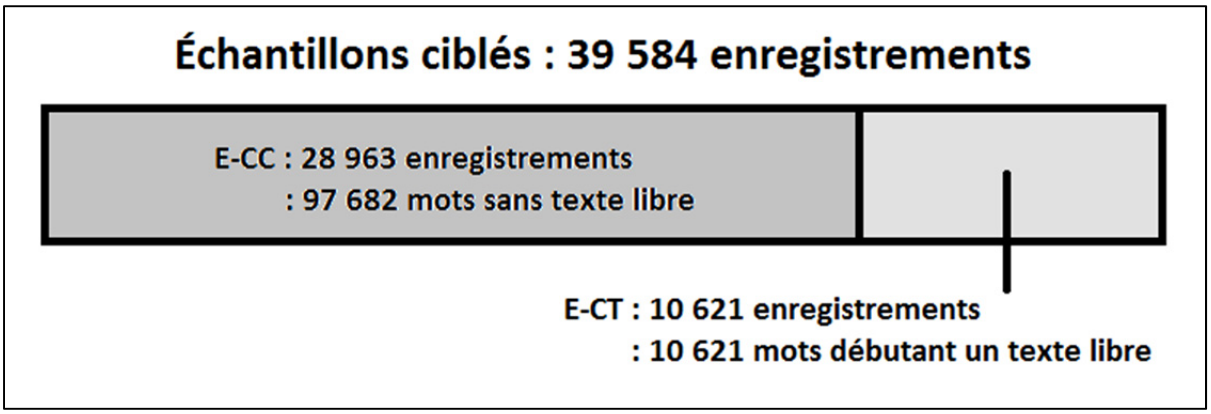

Figure 2.1 Proportions des échantillons ciblés.

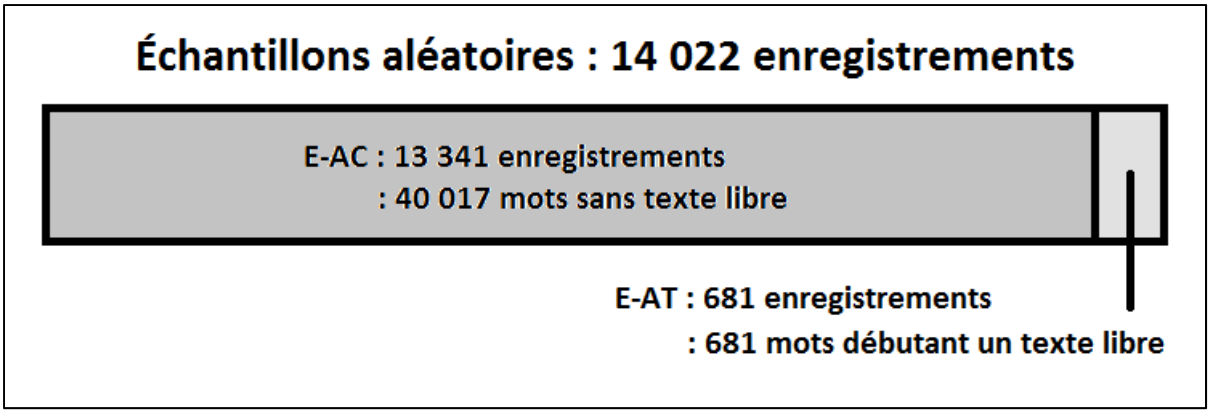

Figure 2.2 Proportions des échantillons aléatoires.

2.3 Outils de mesure automatique

Cette section présente les outils utilisés pour extraire automatiquement l'information des signaux acoustiques. En plus d'extraire les informations prosodiques, il a également été nécessaire d'extraire les informations lexicales. Celles-ci ont permis d'associer un début de texte libre avec les informations prosodiques correspondantes. Pour commencer, l'outil d'extraction prosodique est présenté, suivi de l'outil d'extraction lexicale.

2.3.1 Informations prosodiques

Les informations prosodiques ont été extraites automatiquement grâce à un outil d'extraction conçu par Paul Boersma et David Weenink de l'université d'Amsterdam nommé *Praat version 5.3.49* (Praat). Praat[2] est un logiciel libre en phonétique spécialisé dans l'analyse de la parole. Il a permis la création et l'exécution de scripts nécessaires à l'automatisation des extractions. La précision des mesures temporelles correspondait à 1 msec.

2.3.2 Informations lexicales

Les informations lexicales ont été extraites automatiquement grâce à un système ASR. Ce système a permis de générer des statistiques liées à la tâche de reconnaissance, comme les mots et les pauses dans un signal acoustique, ainsi que leur position temporelle (temps de début et de fin). La précision de ces positions correspondait à 10 msec. Pour fonctionner, ce système avait besoin d'un enregistrement audio en entrée. Il avait également besoin d'une grammaire contenant l'ensemble des mots autorisés pour générer sa transcription. Pour les besoins de cette étude, la grammaire fournie contenait seulement la même transcription automatique antérieurement conservée de l'enregistrement concerné. Cette approche a permis de maximiser l'alignement des mots et des pauses par rapport au signal.

2.4 Évaluation de la qualité des outils

Pour évaluer la qualité des outils de mesure, il a été nécessaire de comparer les mesures que nous avons extraites manuellement (exactes) avec celles extraites automatiquement (approximées). Les pauses contenues dans un enregistrement audio étaient l'information prosodique la plus simple à mesurer manuellement. Pour cette raison, la qualité des outils de cette expérimentation ne s'est basée que sur celle-ci. La confiance obtenue des mesures a donc été projetée sur les autres informations prosodiques. En effet, pour mesurer précisément

---

[2] http://www.fon.hum.uva.nl/praat/

les autres informations, les outils automatiques étaient plus fiables par rapport à la perception humaine. La durée d'une pause est relativement facile à mesurer manuellement, tandis que la hauteur ou l'intensité d'un son nécessite des instruments spécialisés. Les sections suivantes présentent la méthodologie utilisée pour évaluer ces outils.

2.4.1 Fondement théorique

Pour connaître la confiance pouvant être accordée aux outils d'extraction, l'erreur de la mesure automatique a été obtenue selon cette équation :

$$e_{auto} = m_{auto} - m_{manu} \tag{2.1}$$

où m_{auto} correspondait à la mesure automatique et m_{manu} à la mesure manuelle. Cette équation a été appliquée à deux types d'erreur. En effet, la durée d'une pause avant un mot d'une commande vocale correspondait à cette équation :

$$d = t_{fin} - t_{début} \tag{2.2}$$

où t_{fin} correspondait à la position dans le temps de la fin de la pause, exprimée en secondes, et $t_{début}$ à sa position de début. L'erreur de la mesure automatique (erreur de pause) a donc été définie pour le temps de début et le temps de fin de la pause avant chaque mot.

L'ensemble de tous les mots (y compris les redondants) contenus dans les commandes vocales correspondait à la population dont les erreurs de pause devaient être mesurées. À défaut de pouvoir mesurer toute la population, il a été nécessaire de mesurer un petit échantillon pour les fins de cette évaluation (échantillon E-E). À partir de celui-ci, des conclusions ont été déduites sur la population. Pour que E-E soit représentatif de la population, les commandes vocales ont été choisies au hasard.

Chaque mesure d'erreur de pause correspondait à une observation x_i parmi les n mesures constituant E-E. L'énumération de toutes les observations possibles, ainsi que leur fréquence respective, correspondait à une distribution. Puisque ces observations consistaient en des valeurs continues, celles-ci étaient regroupées en intervalles nommés « classes ». La Figure 2.3 présente un exemple d'une distribution des erreurs de pause.

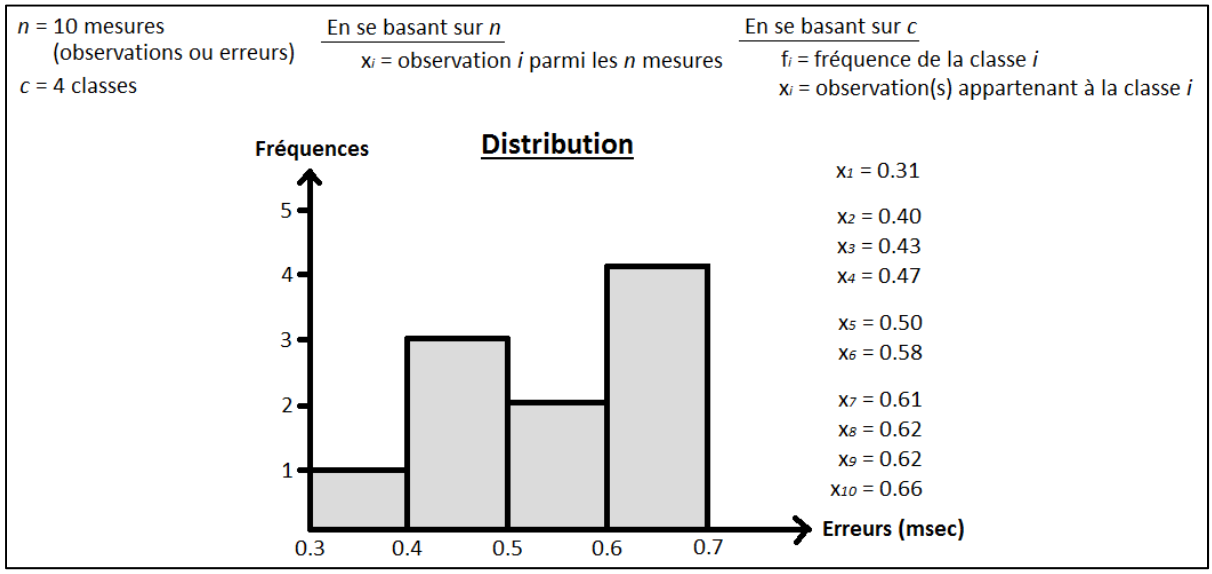

Figure 2.3 Exemple d'une distribution des erreurs de pause.

Le centre de cette distribution a été déterminé grâce à la moyenne arithmétique de cet échantillon :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{c} \frac{f_i}{n} x_i \tag{2.3}$$

où f_i correspondait à la fréquence d'une classe i et c au nombre de classes. Par exemple, toutes les erreurs contenues entre 0.5 et 0.6 msec étaient assignées à la même classe. La fréquence correspondait à leur nombre dans cette classe et la valeur était prise égale au milieu de la classe.

La dispersion de cette distribution correspondait à l'étendue des différentes valeurs pouvant prendre une observation. La dispersion a été déterminée grâce à l'écart type de cet échantillon :

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{c} f_i(x_i - \bar{x})^2}{n}} \tag{2.4}$$

À partir de la moyenne et de l'écart type d'un échantillon quelconque, il était possible d'estimer la moyenne de la population :

$$\mu = \bar{x} \pm \Delta x = \bar{x} \pm t\frac{s}{\sqrt{n}} \tag{2.5}$$

où Δx correspondait à l'incertitude attribuée à la moyenne et t au coefficient de Student, dont la table des valeurs est présentée à l'ANNEXE I. t dépendait de n et incluait la confiance qu'il était souhaitable d'attribuer au résultat. En général, les résultats des mesures sont donnés avec une confiance de 95%, c.-à-d. qu'il y a 95% des chances que μ soit compris entre $\bar{x} - \Delta x$ et $\bar{x} + \Delta x$. L'erreur moyenne de la population pour les mesures des débuts et des fins de pause (avec son incertitude) devait alors être suffisamment faible. Dans ce cas, la confiance accordée aux outils automatiques serait suffisante pour extraire les informations prosodiques nécessaires à cette étude.

2.4.2 Combinaison des deux outils automatiques

Le système ASR utilisé ne considérait pas toujours les pauses de courte durée entre les mots. Celui-ci avait tendance à empiéter sur ces pauses afin d'élargir l'étendue des mots. De son côté, Praat était très sensible aux bruits et aux pauses pleines. Souvent, il ne les considérait

pas comme des pauses. Afin de minimiser l'erreur sur l'extraction automatique des pauses, nous avons développé un algorithme combinant les forces des deux outils.

Notre algorithme débute par obtenir la pause entière avant chaque mot extrait par le système ASR, même si sa durée est souvent nulle. Ensuite, l'algorithme tente d'élargir chacune de ces pauses grâce à celles extraites par Praat. Il considère la valeur minimale entre le début de la pause de Praat et celle du système ASR. Ces pauses doivent se chevaucher, mais une marge maximale de 50 msec de non pause est permise entre la fin de la pause de Praat et le début de celle du système ASR. Ce 50 msec est une valeur *ad hoc* qui s'est avérée quasi-optimale suite aux essaies sur l'échantillon E-E. Le nouveau début de pause devient alors cette valeur minimale. La nouvelle fin de pause devient la valeur maximale entre les fins obtenue des deux outils. Les pauses doivent se chevaucher, et aucune marge de non pause n'est permise.

Lorsqu'il y a du bruit, Praat génère parfois des pauses par intermittence. L'algorithme tente ensuite d'englober celles-ci en une seule grande pause. Pour réussir, l'intervalle de bruit entre deux intermittences doit être contenu à l'intérieur de 30 msec. Ce 30 msec est également une valeur *ad hoc* quasi-optimale. Pour ces intermittences, l'algorithme procède ainsi à droite et à gauche de la pause originale. Une fois l'algorithme terminé, le début et la fin de pause obtenus avant chaque mot sont considérés comme les mesures automatiques officielles. La Figure 2.4 présente un exemple de la combinaison des pauses extraites de Praat et du système ASR. Les rectangles en gris très pâle représentent des durées de pauses ignorées. Enfin, les rectangles en gris plus foncé représentent des durées de pauses combinées pour générer des durées de pauses finales, représentées par les rectangles en noirs.

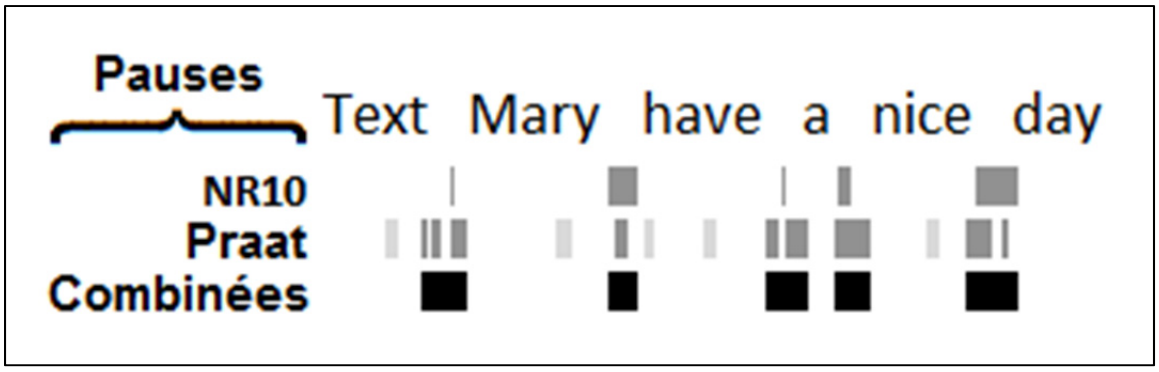

Figure 2.4 Combinaison des pauses extraites de Praat et du système ASR.

La combinaison de ces deux outils a seulement été utilisée pour l'extraction des pauses. La pause étant la seule information prosodique pouvant également être extraite à partir du système ASR. Par conséquent, les autres caractéristiques prosodiques de cette étude ont seulement été extraites à partir de Praat.

2.4.3 Validation de l'exactitude des mesures

Un test fonctionnel a été utilisé pour valider l'exactitude des mesures, issues de l'extraction automatique des durées des pauses, par rapport à notre algorithme décrit précédemment. Le corpus de données utilisé pour ce test est décrit à la section 2.4.5. Ce test de type « boîte noire » a permis de valider les sorties des classes d'équivalence des entrées, ainsi que leurs valeurs frontalières, en fonction des différentes ramifications de l'algorithme. Ce test a également été utilisé pour valider l'exactitude de nos algorithmes utilisés pour l'extraction des autres caractéristiques prosodiques (voir CHAPITRE 4). Des représentations graphiques ont également été utilisées pour cette validation.

Pour commencer, chacun de nos algorithmes a été décrit textuellement et a été imagé. Cela a permis d'en faciliter sa compréhension interne et d'en préciser ses ramifications. Ensuite, les valeurs d'entrée ont été énumérées et décrites. Des assomptions concernant ces entrées ont

été définies. Les classes d'équivalence de ces entrées ou d'autres valeurs découlant de celles-ci ont ensuite été définies avec leurs valeurs frontalières. Un tableau concernant la couverture des classes d'équivalence et des frontières a été généré. Ce tableau représentait l'ensemble des cas de test nécessaires pour effectuer la validation de l'algorithme. Chaque cas de test était représenté par ses entrées, ses sorties attendues ainsi que ses classes d'équivalence valides et invalides par rapport aux valeurs frontalières. Enfin, chaque algorithme a été exécuté isolément et validé par rapport à tous ses cas de test respectifs.

2.4.4 Outil de mesure manuelle

Les positions de début et de fin de la pause avant chaque mot ont été mesurées manuellement grâce à un éditeur audio. Cet outil a été conçu par Jonas Beskow et Kare Sjolander et se nomme *WaveSurfer 1.8.8p4* (WaveSurfer). WaveSurfer[3] est un logiciel libre largement utilisé pour les études sur la phonétique acoustique. Il a permis des opérations de manipulation de base sur des signaux acoustiques, comme la sélection précise et l'écoute d'une portion de signal. La précision des mesures temporelles correspondait à 1 msec. Les pauses ont été la seule caractéristique prosodique à avoir été mesurée manuellement.

2.4.5 Corpus de données

Pour l'évaluation de la qualité, un corpus de données a été utilisé différent de ceux utilisés pour les expérimentations. Ce corpus était représentatif (variété de locuteurs, niveau de bruits, etc.) de l'environnement dans lequel les expérimentations ont été effectuées. Il était constitué de 128 mots (15 enregistrements). Ces données consistaient en des commandes vocales prononcées par des locuteurs anglophones potentiellement différents pour chaque enregistrement, constituées de 10 hommes et 5 femmes. Celles-ci correspondaient à des requêtes simulées pour des lignes aériennes. La transcription manuelle était disponible pour chaque enregistrement.

---

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

2.4.6 Expérimentation et présentation des résultats

À partir de l'échantillon E-E, les erreurs de début et de fin de la pause avant chaque mot ont été calculées. Ces erreurs ont été obtenues en comparant les mesures automatiques de Praat et du système ASR par rapport aux mesures manuelles de WaveSurfer. Des histogrammes ont été générés pour présenter leur distribution. Des fréquences normalisées ont été utilisées pour chaque type d'erreur. Le Tableau 2.1 présente les statistiques liées à ces erreurs.

Tableau 2.1 Statistiques sur les erreurs de pause

Pause	**Quantité**	**Statistiques en msec**		
		Médiane	**Moyenne**	**Écart type**
Erreur de début	128	0.0	1.5508	44.4994
Erreur de fin	128	0.0	-2.1562	34.0404

La Figure 2.5 présente les histogrammes de ces distributions.

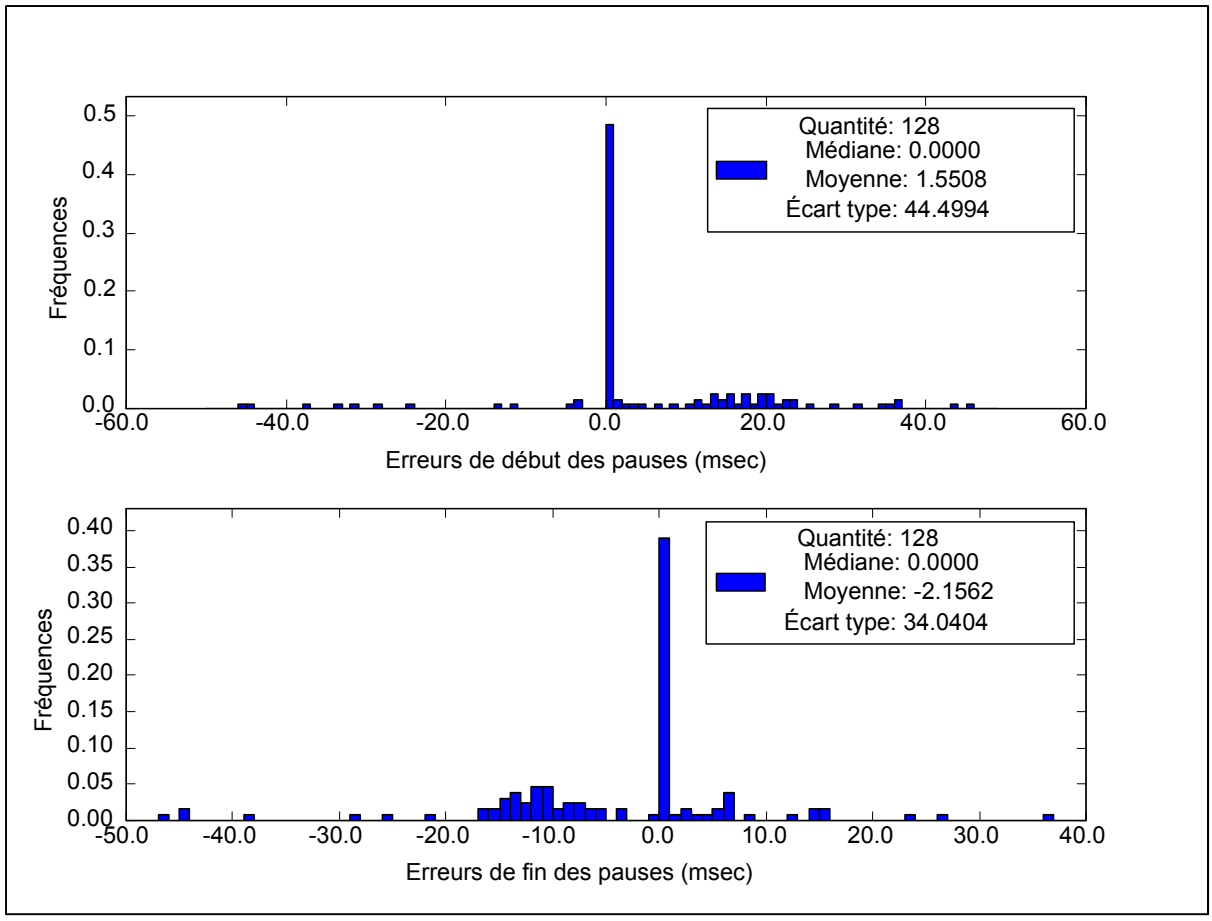

Figure 2.5 Distributions des erreurs de pause.

En se basant sur l'équation 2.5, la moyenne de la population et son incertitude, correspondant aux erreurs de début des pauses, ont été estimés à :

$$\mu = \bar{x} \pm t\frac{s}{\sqrt{n}} = 1.5508 \pm 1.98\frac{44.4994}{\sqrt{128}} = 1.5508 \pm 7.7878 \text{ msec} \qquad (2.6)$$

où $n-1$ (degrés de liberté) a été arrondi à 120 pour la correspondance de t dans la table des coefficients de Student présentée à l'ANNEXE I. Une confiance de 95% ($\alpha = 0.05$) a été choisie pour les mesures de ce mémoire. Concernant les erreurs de fin des pauses, la moyenne de la population et son incertitude ont été estimés à :

$$\mu = \bar{x} \pm t\frac{s}{\sqrt{n}} = -2.1562 \pm 1.98\frac{34.0404}{\sqrt{128}} = -2.1562 \pm 5.9574 \text{ msec} \qquad (2.7)$$

2.4.7 Interprétation et discussion

Comme l'indique le Tableau 2.1, les erreurs de début de pause étaient en général faibles et positives. De leur côté, les erreurs de fin de pause étaient en général un peu plus prononcées mais négatives. Toutefois, les écarts types des deux erreurs étaient très prononcés, étant donné la faible quantité d'observations que constituait l'échantillon E-E. Selon cette expérimentation, dans 95% des cas, l'erreur d'un début ou d'une fin de pause a varié dans un intervalle inférieur à 20 msec incluant l'erreur nulle. Cet intervalle était relativement faible par rapport à l'unité de base utilisée pour découper et analyser un signal acoustique, c.-à-d. la trame.

Dans la plupart des systèmes ASR, le signal est d'abord découpé en trame d'une durée variant entre 20 et 30 msec. La largeur du pas d'avancement de la trame correspond à 10 msec. C'est particulièrement vrai pour les systèmes basés sur les HMM, comme le système ASR utilisé. Cette segmentation s'explique du fait que le signal acoustique est non stationnaire. Toutefois, il se comporte de façon quasi stationnaire pour des durées plus courtes. [82]

L'intervalle d'erreur est donc inférieur à la durée d'une trame dans 95% des cas. Ainsi, il a été raisonnable de considérer acceptable la précision des outils de mesures automatiques. Néanmoins, cet intervalle a été supérieur au pas d'avancement de la trame. En ne considérant que les incertitudes des moyennes, c.-à-d. la moitié de cet intervalle, celles-ci ont été inférieures au pas d'avancement de la trame pour les erreurs de début et de fin de pause.

2.5 Conclusion

Pour déterminer si l'information prosodique contenue dans les signaux acoustiques peut aider le système NLU à améliorer ses performances, les caractéristiques prosodiques ont été obtenues automatiquement à l'aide d'outils d'extraction. Une évaluation de la qualité a été effectuée sur ces outils. L'évaluation a concerné seulement l'extraction des pauses, car la durée des pauses est la seule caractéristique pouvant être mesurée efficacement manuellement.

L'évaluation a démontré que dans 95% des cas, les erreurs de début et de fin de pause ont été contenues dans un intervalle inférieur à la durée d'une trame. Ce résultat a été jugé acceptable pour les besoins de l'étude de ce mémoire. Les pauses ont donc été extraites automatiquement à partir de ces outils. Toutefois, les autres informations prosodiques ont seulement été extraites à partir de Praat.

CHAPITRE 3

EXPÉRIMENTATION AVEC LES PAUSES

3.1 Introduction

Afin de déterminer s'il existe une corrélation entre certaines caractéristiques prosodiques et la présence ou non des textes libres, il a été nécessaire d'étudier individuellement quelques caractéristiques. La pause a été la première caractéristique étudiée, puisqu'elle est à la fois simple et intuitive. Lorsqu'un locuteur est écouté, il semble y avoir une pause plus longue précédant un texte libre par rapport aux autres endroits de la commande vocale. Cette pause pourrait être volontaire de la part du locuteur pour annoncer un changement de contexte. Elle pourrait également correspondre à un moment de réflexion précédant la citation du texte libre. Ce chapitre présente donc la première expérimentation utilisant les pauses, ainsi que les notions acquises et nécessaires à son exécution. Ces notions ont été utilisées dans l'expérimentation avec d'autres caractéristiques présentées au prochain chapitre.

Ce chapitre présente tout d'abord un rappel sur la définition d'un texte libre. Il présente également les différentes catégories observées ainsi que les hypothèses de cette étude. Ensuite, un test est présenté déterminant si deux échantillons quelconques appartiennent à la même distribution. Ce test a été essentiel pour résoudre notre question de recherche. L'expérimentation est ensuite présentée suivie de la présentation, l'interprétation et la discussion des résultats. Enfin, la conclusion confirme si cette expérimentation a supporté ou non les hypothèses formulées.

3.2 Définition d'un texte libre

Comme mentionné au quatrième paragraphe de l'introduction de ce mémoire, un texte libre est une séquence de mots considérée comme une séquence de caractères à accepter telle quelle. Cette séquence correspond à un discours dans un discours. Par exemple, la section entre crochets de la commande vocale « Text Mary [have a nice day] » correspond à un texte

libre. Il est également possible qu'une commande vocale en contienne plusieurs. Par exemple, la commande vocale « Text Mary subject [meeting this morning] message [we have a meeting at nine o'clock] » contient deux textes libres.

Selon la revue de la littérature présentée au CHAPITRE 1, la citation est l'AD le plus près du texte libre. La citation a été classée sous trois catégories. Celles-ci ont inspiré le classement des textes libres. La dernière catégorie, soit le discours direct libre, se transpose bien dans le domaine des textes libres. Toutefois, cette catégorie a été ignorée pour l'étude de ce mémoire. En effet, elle aurait consisté en une commande vocale entière à considérer comme un texte libre. Par exemple, la commande « [have a nice day] » consisterait en un texte libre à part entière. Pour ce cas, il n'y aurait pas la présence de mots à la frontière gauche ou droite du texte libre. Il aurait donc été impossible d'obtenir les propriétés prosodiques nécessaires à son étude, permettant de discriminer le texte libre d'un texte ordinaire. Par conséquent, les sections suivantes présentent les deux catégories de texte libre utilisées pour ce mémoire. Celles-ci sont suivies de la formulation des hypothèses pour cette première expérimentation.

3.2.1 Texte libre direct

La première catégorie de texte libre, le texte libre « direct », se présente dans une commande vocale sans la présence immédiate de mots à sa frontière gauche annonçant son arrivée. Une rupture est observée entre la fin de la première partie de la commande et le début du texte libre. Voici deux exemples :

« Text Mary [have a nice day] »
« Send text to husband [I'm almost there] »

3.2.2 Texte libre indirect

La deuxième catégorie de texte libre, le texte libre « indirect » (ou indiqué lexicalement), se présente dans une commande vocale avec la présence immédiate d'un ou plusieurs mots à sa

frontière gauche annonçant son arrivé. Aucune rupture n'est observée entre la fin de la première partie de la commande et le début du texte libre. Voici deux exemples :

« Remind me in one hour **to** [take care of the laundry] »
« Text Mary and **tell her that** [we are just leaving now] »

3.2.3 Formulation des hypothèses

Nous avons formulé deux hypothèses pour vérifier s'il existe des corrélations entre la durée des pauses et la présence ou non des textes libres. La première hypothèse concerne les textes libres en général. La deuxième hypothèse fait une distinction entre les textes libres directs et indirects. Voici nos deux hypothèses :

Hypothèse 1 : Il existe généralement une durée de pause plus longue au début des textes libres par rapport aux autres concepts.

Hypothèse 2 : Il existe généralement une durée de pause plus longue au début des textes libres directs par rapport aux textes libres indirects.

La suite de ce chapitre présente le cheminement nécessaire qui a permis de supporter ou non ces hypothèses.

3.3 Expérimentation

L'ensemble de tout le trafic des commandes vocales correspondait à la population de cette étude. À partir de cette population, deux échantillons ont été extraits comprenant des textes libres (échantillons E-AT et E-CT). Deux autres échantillons ont été extraits comprenant d'autres concepts sans texte libre (échantillons E-AC et E-CC). Pour les textes libres, nous

avons développé un algorithme pour extraire seulement les transcriptions appartenant aux domaines « email », « sms », « reminder » et « note ». Un domaine correspond à une catégorie sémantique associée à une commande. Ces transcriptions devaient contenir au moins une fois les mentions « text » ou « title », mais dont le premier mot ne correspondait pas à celles-ci. Nous avons vérifié ces transcriptions manuellement pour éliminer les commandes ne contenant en fait aucun texte libre. Les mentions erronées ont été corrigées et les textes libres classés selon leur catégorie. Pour les autres concepts, nous avons développé un autre algorithme pour extraire seulement les transcriptions contenant au moins deux mots, mais n'appartenant pas aux domaines mentionnés. Toutefois, leurs mentions n'ont pas été corrigées.

Les durées des pauses ont été extraites à l'aide de Praat et du système ASR. Pour les textes libres, celles-ci ont été extraites avant chaque mot les débutant. Pour les autres concepts, elles ont été extraites avant chaque mot de la commande, excepté le premier mot. La première pause d'une commande (de même que la dernière) n'avait aucune valeur pour cette étude. En effet, elle était généralement beaucoup plus longue par rapport aux autres endroits de la commande. Toutes les pauses extraites ont ensuite été normalisées en les soustrayant par la pause de durée minimale (souvent nulle) de la commande concernée. Cette normalisation a réduit l'impact des débits de parole variés des différents locuteurs. Le Tableau 3.1 présente les paramètres Praat utilisés pour l'extraction de la durée des pauses.

Tableau 3.1 Paramètres Praat utilisés pour l'extraction de la durée des pauses

Minimum pitch (Hz)	100.0
Time step (s)	0.001
Subtract mean	Yes
Silence threshold (dB)	-30.0
Minimum silent interval duration (s)	0.001
Minimum sounding interval duration (s)	0.001

Pour commencer, l'échantillon E-AT constitué de textes libres a été extrait de façon aléatoire. Cependant, très peu de données ont été recueillies. L'échantillon aléatoire E-ACextrait et constitué d'autres concepts contenait tant qu'à lui suffisamment de données. Le Tableau 3.2 présente les quantités des mots extraits de ces échantillons.

Tableau 3.2 Quantités des mots extraits des échantillons E-AT et E-AC

Concept	**Quantité**
Tous les textes libres	681
Textes libres directs	384
Textes libres indirects	297
Autres mots	40 017

Comme alternative à ce manque de textes libres, deux échantillons extraits de façon ciblée (échantillons E-CT et E-CC) ont également été utilisés. Cependant, comme mentionné au CHAPITRE 2, ces échantillons ciblés étaient légèrement biaisés. Le Tableau 3.3 présente les quantités des mots extraits de ces échantillons.

Tableau 3.3 Quantités des mots extraits des échantillons E-CT et E-CC

Concept	Quantité
Tous les textes libres	10 621
Textes libres directs	8 644
Textes libres indirects	1 977
Autres mots	97 682

Pour vérifier les hypothèses formulées à la section 3.2.3, des histogrammes ont été générés afin d'obtenir les fréquences des durées extraites. Ces durées ont été classées selon des intervalles de valeurs, puisqu'elles étaient constituées de valeurs continues. Il est possible de recouvrir jusqu'à 99.7% d'une « distribution normale » en délimitant les mesures sous l'intervalle $[-3s, 3s]$ où s représente l'écart type. Cet intervalle a été choisi pour délimiter ces histogrammes, afin d'éviter l'affichage des valeurs extrêmes. Toutefois, ces distributions ne suivaient pas nécessairement une loi normale. Les échantillons à comparer étaient de grosseurs différentes. Les histogrammes ont donc été générés avec des fréquences normalisées. Les distributions pouvaient être comparées visuellement avec leur moyenne et leur écart type. Cependant, le test de Kolmogorov–Smirnov (test K-S) à deux échantillons présenté à l'ANNEXE II a été utilisé pour que les tests soient statistiquement significatifs.

3.4 Présentation des résultats

Cette section présente les résultats de cette première expérimentation avec les quatre échantillons. Le Tableau 3.4 présente les statistiques des distributions des échantillons aléatoires E-AT et E-AC concernant la durée des pauses.

Tableau 3.4 Statistiques des échantillons E-AT et E-AC pour la durée des pauses

Concept	Quantité	Statistiques en msec		
		Médiane	Moyenne	Écart type
Tous les textes libres	681	121.0	229.67	293.06
Textes libres directs	384	260.0	311.07	281.63
Textes libres indirects	297	39.0	124.42	273.38
Autres mots	40 017	0.0	214.91	681.35

La Figure 3.1 et la Figure 3.2 présentent ces distributions.

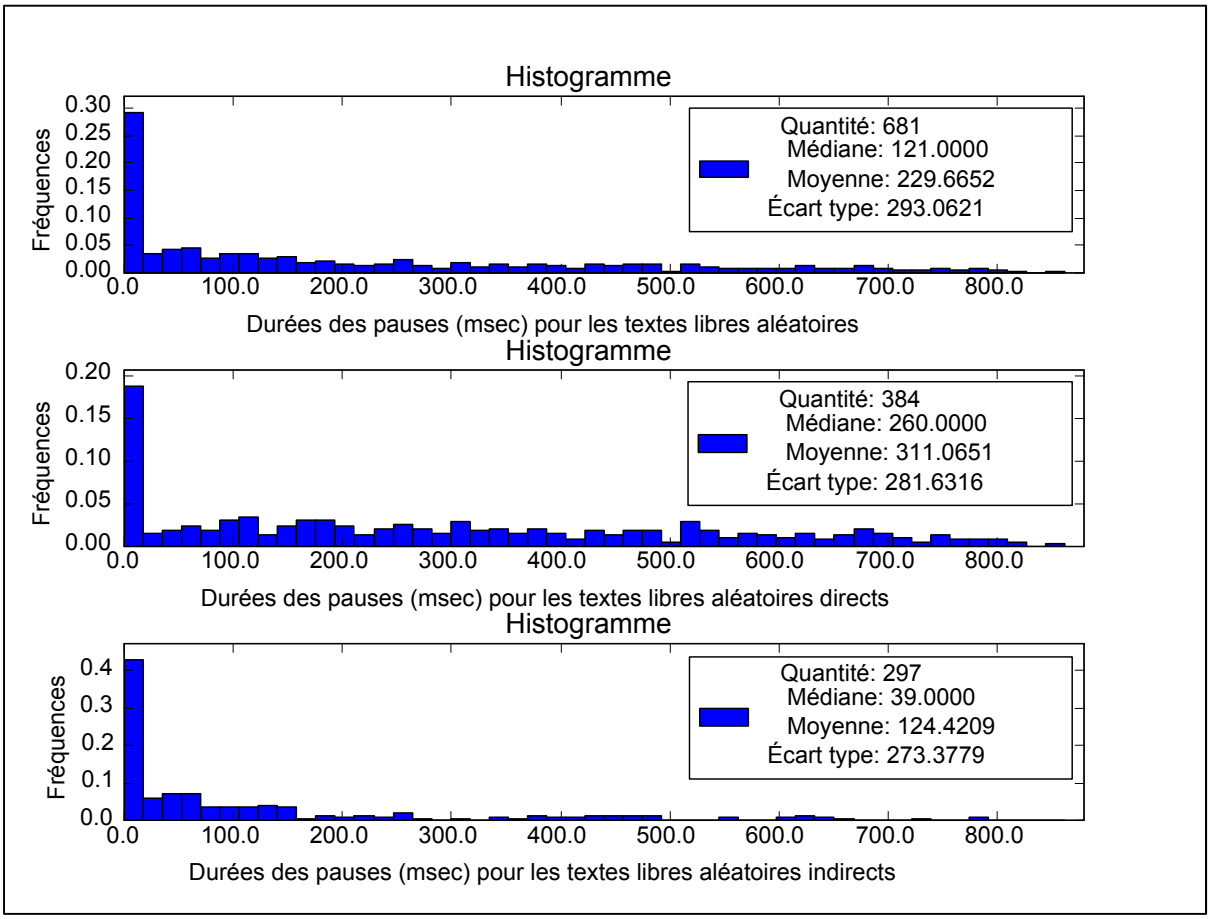

Figure 3.1 Durées des pauses de l'échantillon E-AT

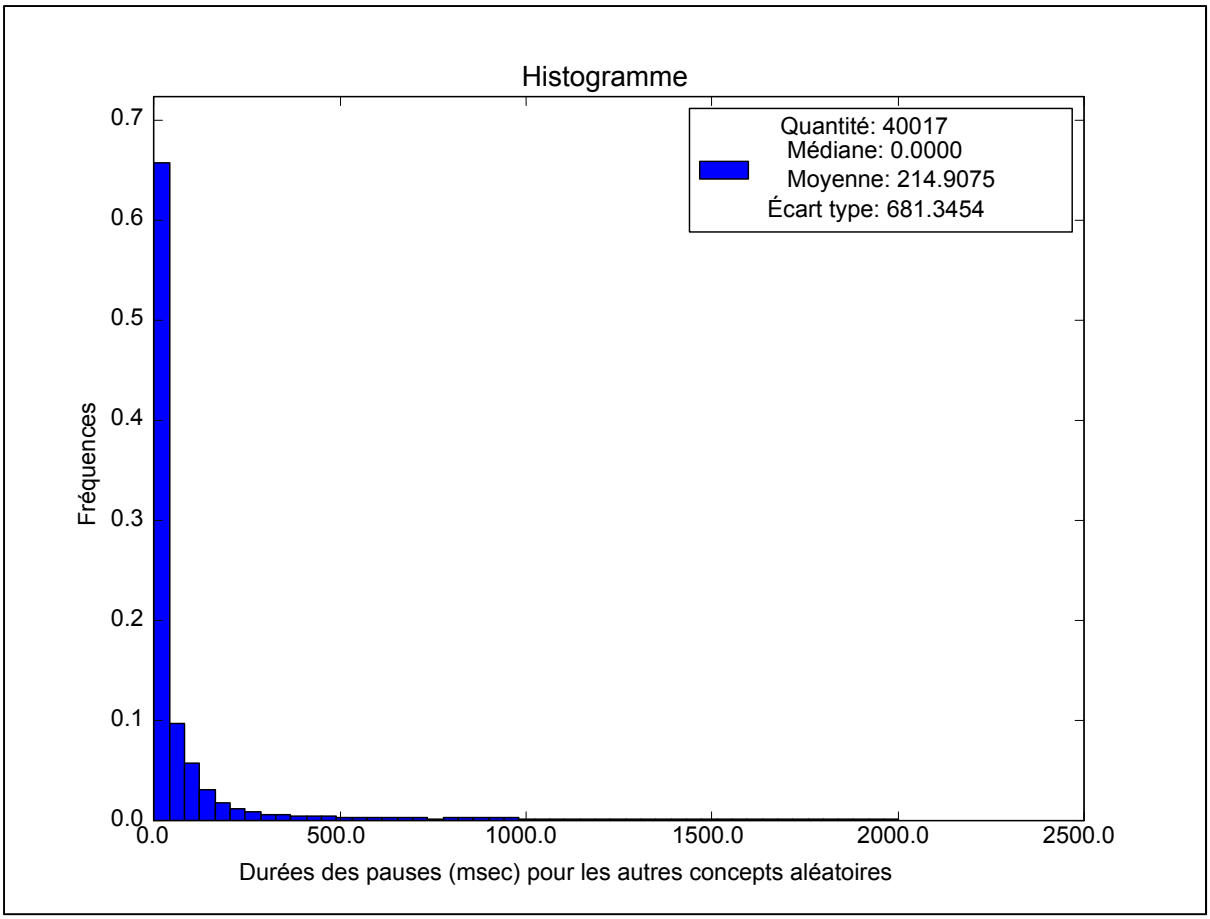

Figure 3.2 Durées des pauses de l'échantillon E-AC

Le Tableau 3.5 présente les statistiques des distributions des deux échantillons ciblés E-CT et E-CC.

Tableau 3.5 Statistiques des échantillons E-CT et E-CC pour la durée des pauses

Concept	Quantité	Statistiques en msec		
		Médiane	**Moyenne**	**Écart type**
Tous les textes libres	10 621	236.0	285.21	312.51
Textes libres directs	8 644	272.5	311.21	323.71
Textes libres indirects	1 977	60.0	171.54	224.99
Autres mots	97 682	0.0	103.51	307.81

La Figure 3.3 et la Figure 3.4 présentent ces distributions.

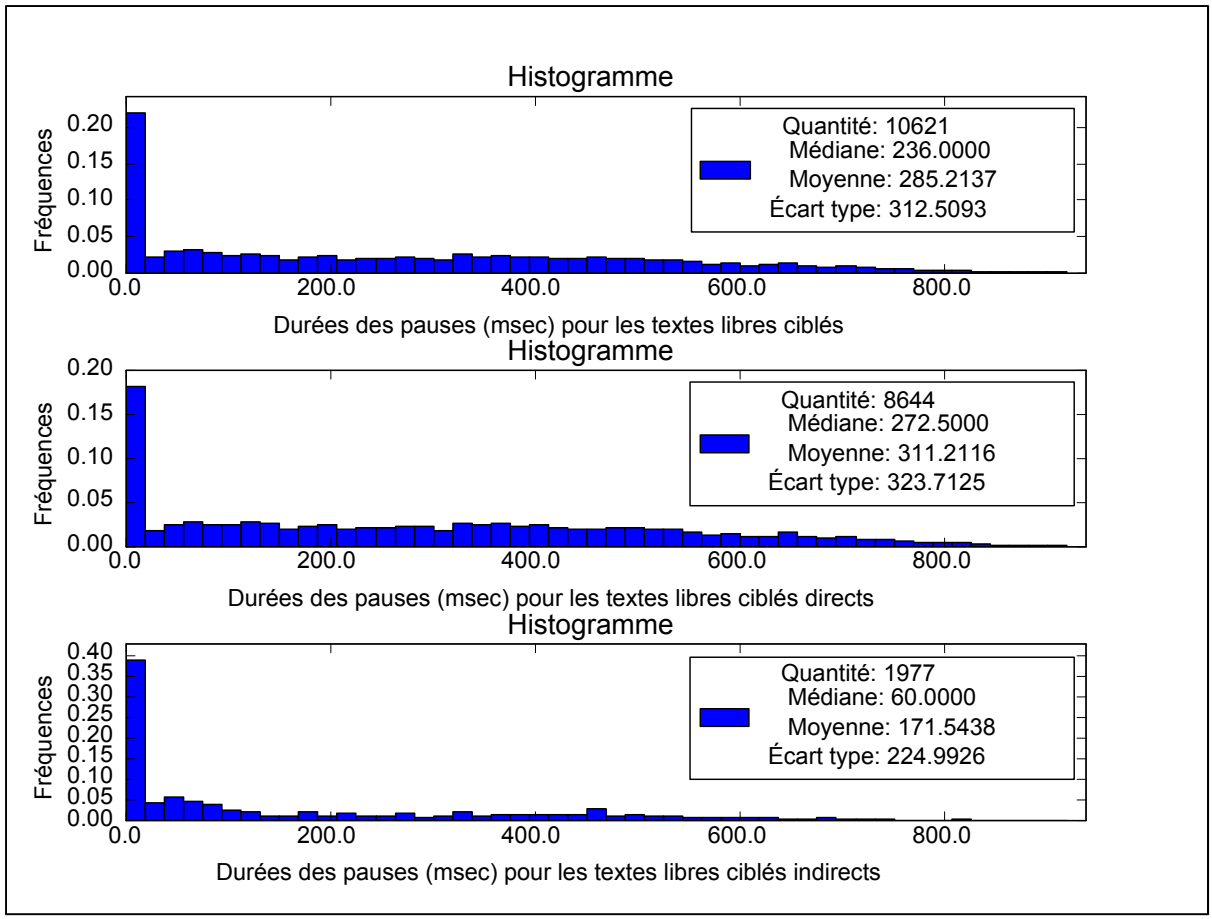

Figure 3.3 Durées des pauses de l'échantillon E-CT

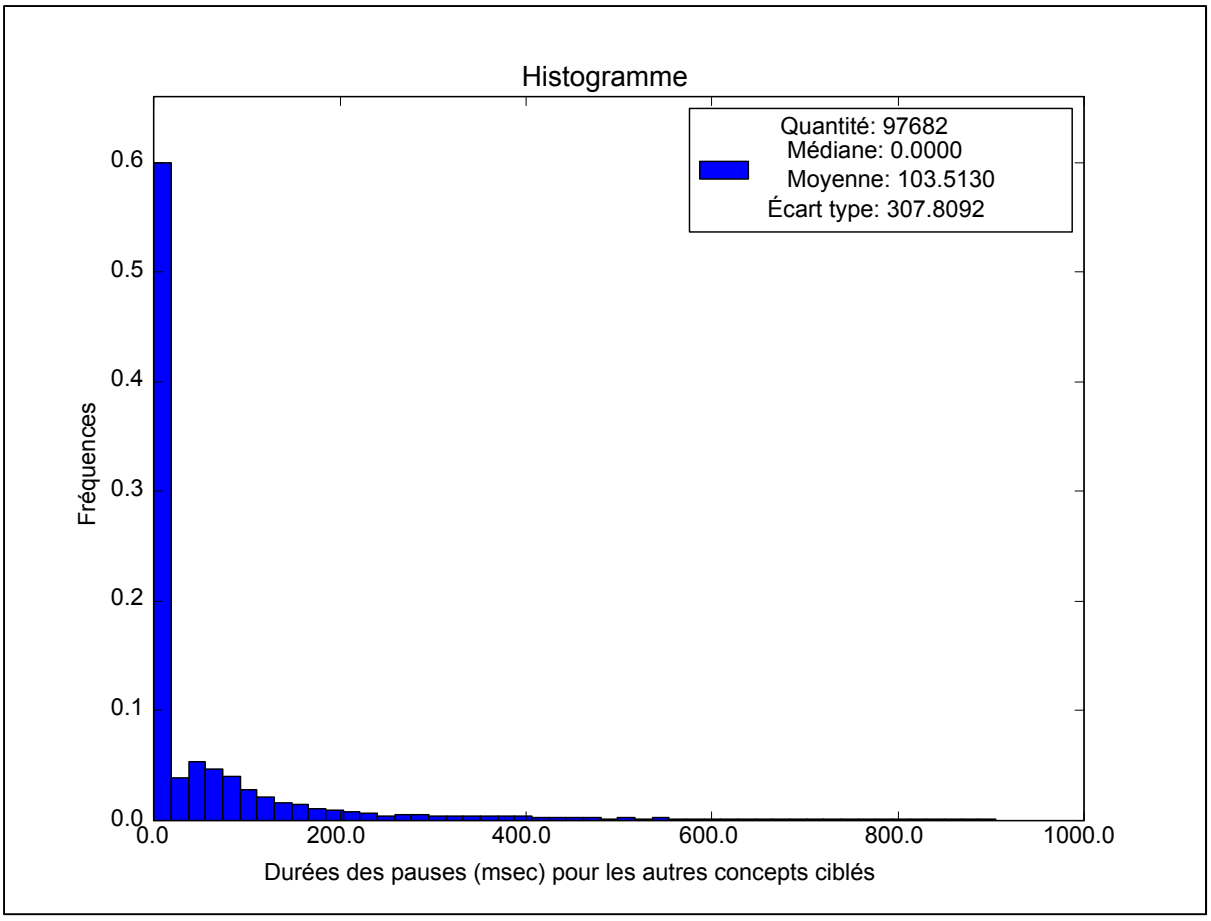

Figure 3.4 Durées des pauses de l'échantillon E-CC

3.5 Interprétation

Comme l'indique le Tableau 3.5, la moyenne de la durée des pauses des échantillons aléatoires a été plus longue avant le début des textes libres par rapport aux autres concepts. La même constatation a été observée de façon plus significative pour les échantillons ciblés. Pour les échantillons aléatoires, la moyenne pour les textes libres a été beaucoup plus longue pour le sous-échantillon des textes libres directs par rapport au sous-échantillon des textes libres indirects. La même constatation a été observée pour les échantillons ciblés.

Pour déterminer de manière statistiquement significative si ces échantillons appartenaient bien à des distributions différentes, des tests K-S ont été appliqués avec un seuil $\alpha = 0.05$. Des hypothèses nulles ont été formulées selon lesquelles les échantillons des textes libres (E-AT et E-CT) et les échantillons des autres concepts (E-AC et E-CC) appartenaient à la même distribution. Pour les échantillons aléatoires (E-AT et E-AC), la valeur-p calculée avec la fonction *ks_2samp* de Python était de 1.336e-63 et la statistique K-S était de 3.280e-1. La probabilité d'obtenir une distance de 3.280e-1 ou plus entre les fonctions de répartition empirique des deux échantillons était de 1.336e-63 s'ils appartenaient à la même distribution. Puisque la valeur-p était inférieure à α, l'hypothèse nulle a été rejetée. L'hypothèse nulle a également été rejetée pour les échantillons ciblés (E-CT et E-CC). Il est donc possible d'affirmer avec confiance que les résultats supportaient l'hypothèse 1 formulée à la section 3.2.3.

D'autres hypothèses nulles ont été formulées selon lesquelles les sous-échantillons des textes libres directs et les sous-échantillons des textes libres indirects appartenaient à la même distribution. Ces hypothèses ont également été rejetées à la fois pour l'échantillon aléatoire (E-AT) et ciblé (E-CT). Il est donc possible d'affirmer avec confiance que les résultats supportaient l'hypothèse 2 formulée à la section 3.2.3. Le Tableau 3.6 présente les résultats de ces tests K-S appliqués pour la durée des pauses.

Tableau 3.6 Résultats des tests K-S pour la durée des pauses

Échantillon	Concept 1	Concept 2	Valeur-p	Statistique K-S	Hypothèse nulle
E-AT, E-AC	Tous les textes libres	Autres mots	1.336e-63	3.280e-1	Rejetée
E-CT, E-CC	Tous les textes libres	Autres mots	0.0	4.546e-1	Rejetée
E-AT	Textes libres directs	Textes libres indirects	1.207e-30	4.513e-1	Rejetée
E-CT	Textes libres directs	Textes libres indirects	2.560e-126	3.006e-1	Rejetée

3.6 Discussion

En examinant tous les histogrammes présentés à la section précédente, des fréquences très élevées ont été observées pour les pauses de durée nulle par rapport aux autres durées. Cette

observation aurait été similaire si les intervalles des durées avaient été réduits. La raison probable est que la précision de ces durées était trop près de l'unité de base, qui était de 1 msec. Cette unité correspondait à la précision temporelle d'extraction de Praat combiné au système ASR. Le système ASR avait une précision limitée par la largeur du pas d'avancement d'une trame, correspondant à 10 msec. Pour sa part, Praat avait une précision de 1 msec. Toutefois, les pauses longues étaient plus faciles à extraire avec Praat par rapport à celles très courtes. Parfois, les durées extraites avec Praat étaient inférieures au pas d'une trame. Cependant, ces durées étaient beaucoup moins fréquentes, puisqu'elles étaient trop près de la précision temporelle d'extraction. Par conséquent, ces durées étaient plus courtes ou trop près de la précision offerte par ces outils. Une transition plus graduelle aurait peut-être été observée avec l'utilisation d'outils plus précis. Les distributions discontinues dans la nature étant très rares. Néanmoins, ces durées possiblement décalées étaient tout de même presque nulles. La Figure 3.5 présente un exemple de pauses extraites avec le système ASR et Praat. Les portions en noires correspondent aux durées considérées par ces outils, tandis que les portions en grises correspondent aux durées ignorées étant donné leur précision d'extraction.

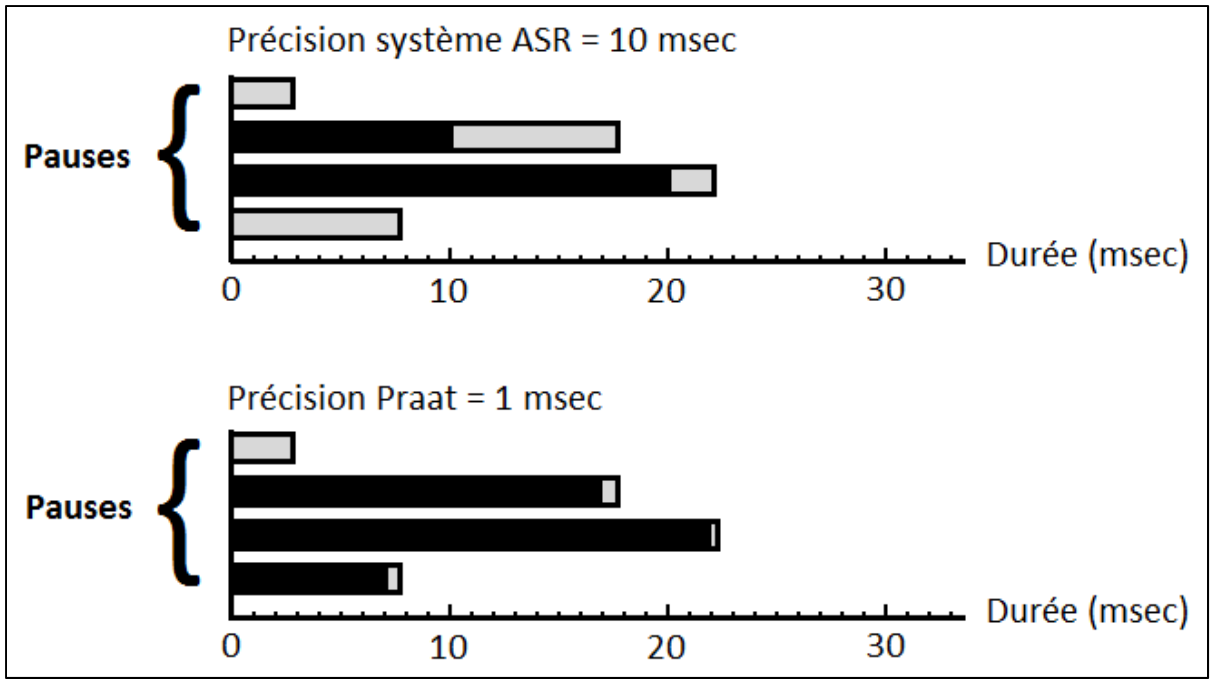

Figure 3.5 Exemple de pauses extraites avec le système ASR et Praat.

Pour les échantillons aléatoires et ciblés, conformément à l'hypothèse 1, les pauses de durée nulle étaient moins fréquentes pour les textes libres par rapport aux autres concepts. Même constatation par rapport à l'hypothèse 2. Les pauses de durée nulle étaient moins fréquentes pour les textes libres directs par rapport aux textes libres indirects. Le Tableau 3.7 présente ces proportions.

Tableau 3.7 Proportions des pauses de durée nulle pour tous les échantillons

Concept	**Proportion des pauses nulles (%)**	
	Aléatoire	**Ciblé**
Tous les textes libres	28.34	21.16
Textes libres directs	18.23	17.50
Textes libres indirects	41.41	37.25
Autres mots	59.12	58.39

Enfin, il aurait été intéressant d'effectuer une expérimentation semblable pour étudier la fin des textes libres. Cependant, la quantité de données disponible était insuffisante. Sur 10 621 textes libres du corpus ciblé, seulement 75 ne terminaient pas la commande vocale. Ce nombre était inférieur concernant les données aléatoires. Il aurait été impossible de vérifier des hypothèses avec aussi peu de données. Sélectionner seulement des textes libres ne terminant pas la commande vocale aurait été essentielle pour étudier leur fin. Autrement, aucun indice prosodique discriminant ne pourrait être recueilli. Pour une raison similaire, seulement des textes libres ne débutant pas la commande vocale ont été sélectionnés pour étudier leur début. En effet, le début ou la fin d'un texte libre avait besoin des propriétés prosodiques de la fin ou du début du mot voisin en contexte pour pouvoir éventuellement se démarquer des autres concepts. Ce mot voisin d'intérêt était absent au début et à la fin d'une commande vocale. Ainsi, les débuts et les fins des textes libres ne respectant pas cette condition ne pouvaient être étudiés.

3.7 Conclusion

Cette première expérimentation avec la durée des pauses a supporté l'hypothèse 1 selon laquelle il existe généralement une durée de pause plus longue au début des textes libres par rapport aux autres concepts. Cette expérimentation a également supporté l'hypothèse 2 selon laquelle il existe généralement une durée de pause plus longue au début des textes libres directs par rapport aux textes libres indirects. Le manque de données n'a pas permis d'étudier des corrélations pour la fin des textes libres. La question de recherche s'est alors seulement concentrée sur leur début. Le prochain chapitre présente une autre expérimentation afin de poursuivre l'étude avec d'autres caractéristiques prosodiques.

CHAPITRE 4

EXPÉRIMENTATION AVEC D'AUTRES CARACTÉRISTIQUES

4.1 Introduction

Suite aux résultats encourageants avec l'utilisation des pauses, il était intéressant d'essayer de trouver des corrélations avec d'autres caractéristiques prosodiques. Selon la littérature, une caractéristique d'intérêt pour identifier les frontières des citations est la fréquence fondamentale (F0). Selon Shriberg *et al.* [65], les caractéristiques basées sur l'énergie ou l'amplitude sont moins fiables et largement redondante par rapport à celles basées sur les durées et F0. Ces caractéristiques ont donc été ignorées pour ce mémoire.

Ce chapitre présente l'expérimentation avec deux caractéristiques issues de F0 ainsi que les notions nécessaires à leur utilisation. Pour commencer, un rappel de ce qu'est F0 est présenté, suivi de deux problèmes fréquemment rencontrés lors de son extraction automatique. Un modèle est ensuite détaillé pour représenter les mesures recueillies sur F0. Les deux caractéristiques prosodiques issues de F0 et de ce modèle sont présentées. Celles-ci sont suivies de la formulation des hypothèses pour cette expérimentation. L'expérimentation est présentée suivie de la présentation, l'interprétation et la discussion des résultats. Enfin, la conclusion confirme si cette expérimentation a supporté les hypothèses formulées.

4.2 La caractéristique F0

F0[4] correspond à la fréquence fondamentale (hauteur) d'un signal acoustique. Les niveaux de hauteur différents, ou intonations, peuvent affecter le sens d'un énoncé. L'exemple le plus évident est la manière dont les locuteurs augmentent F0 à la fin d'une question. Cette fin est indiquée par un point d'interrogation à l'écrit. Cependant, les modèles de montée et de descente peuvent également indiquer des sentiments comme l'étonnement, l'ennui ou la

[4] http://www.litnotes.co.uk/prosodicspeech.htm

perplexité. Ceux-ci peuvent être décrits à l'écrit seulement par l'emploi d'une transcription particulière. Quand un mot anglais est prononcé de façon isolée dans une intonation déclarative, F0 atteint généralement une valeur maximale sur la syllabe accentuée.

Comme mentionné au CHAPITRE 1, Shriberg *et al.* [68] affirment que les informations sur F0 sont typiquement moins robustes et plus difficiles à modéliser par rapport à d'autres caractéristiques prosodiques, comme les pauses. Ceci est largement attribuable à la variabilité dans la façon dont F0 est utilisé parmi les locuteurs et les contextes de discours. Également, la complexité à représenter des modèles de F0, les effets segmentaires et les discontinuités lors de son extraction automatique en font une caractéristique difficile. La section suivante présente deux problèmes de discontinuité fréquemment observés lors de son extraction. Ces problèmes ont également été rencontrés au cours de cette expérimentation.

4.2.1 Problèmes lors de l'extraction de F0

Un premier problème rencontré lors de l'extraction automatique de F0 était sa réduction de moitié à certains endroits dans le signal, souvent appelé le *pitch halving*. Dans ce problème, l'outil d'extraction considère la période fondamentale du signal acoustique comme étant deux fois plus longue par rapport à ce qu'elle est vraiment. Par conséquent, l'outil estime F0 à la moitié de sa valeur réelle [1], faussant ainsi les mesures recueillies.

Un second problème rencontré était à l'opposé du premier. Il consistait au doublement de F0 à certains endroits, souvent appelé le *pitch doubling.* Dans ce problème, l'outil d'extraction considère la période fondamentale comme étant la moitié par rapport à ce qu'elle est vraiment. Par conséquent, l'outil estime F0 au double de sa valeur réelle [1], faussant également les mesures recueillies.

Une des raisons causant ces deux problèmes est l'apparition de cycles d'impulsions alternées dans un signal de parole, ce qui reflète l'instabilité à court terme du système de cordes vocales [72]. Il fallait éviter au mieux de fausser les résultats avec ces discontinuités. Des

stratégies ont donc été appliquées sur l'extraction des mesures. Ainsi, la plupart de ces discontinuités ont été ignorées pour cette étude. Ces stratégies sont présentées à la section 4.4 décrivant cette expérimentation.

4.3 Caractéristiques issues de F0

La littérature suggère plusieurs types de caractéristiques pouvant être issus des mesures recueillies sur F0. Le modèle de régression linéaire simple présenté à l'ANNEXE III a permis de représenter grossièrement ces mesures à l'aide d'une droite. Grâce à ce modèle appliqué sur une fenêtre à gauche et à droite d'une pause entre deux mots, une droite pour chaque côté de la frontière a été obtenue. La Figure 4.1 présente un exemple de droites utilisées pour représenter les mesures de F0.

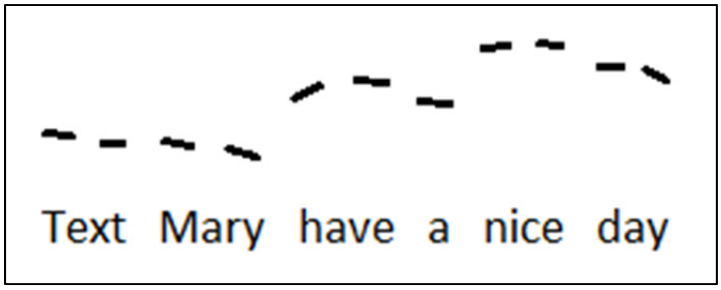

Figure 4.1 Droites représentant les mesures de F0.

Les deux sections suivantes présentent les deux caractéristiques prosodiques utilisées issues de ces droites. Ces sections sont suivies par la formulation des hypothèses pour cette expérimentation.

4.3.1 La réinitialisation

Comme première caractéristique issue de F0, la réinitialisation consiste en la différence de hauteur entre le début et la fin de la pause entre deux mots. Cette caractéristique est utilisée

pour capturer la tendance bien connue des locuteurs à réinitialiser la hauteur au début d'une nouvelle unité majeure, comme une frontière de sujet ou de phrase, par rapport à l'endroit où ils l'avaient laissé [68]. Typiquement, la réinitialisation est précédée par une descente finale en hauteur associée avec la fin d'une telle unité. Ainsi, une plus grande réinitialisation devrait être présente aux frontières de ces unités par rapport aux autres endroits du discours. Par exemple, la Figure 4.1 présente une réinitialisation plus prononcée à la frontière entre les mots « Mary » et « have » ainsi que les mots « a » et « nice ». Cependant, la réinitialisation est négligeable ou inexistante à la frontière entre les mots « Text » et « Mary ».

4.3.2 La continuité

Comme deuxième caractéristique issue de F0, la continuité consiste en la différence des pentes des droites à la frontière entre deux mots. Une trajectoire continue devrait être corrélée avec une absence de frontière. Une trajectoire brisée signalerait sa présence, sans considérer la différence de hauteur à travers les mots. Par exemple, la Figure 4.1 présente une continuité plus prononcée (brisée) à la frontière entre les mots « Mary » et « have ». Cependant, la continuité est négligeable ou inexistante (continue) à la frontière entre les mots « have » et « a ».

4.3.3 Formulation des hypothèses

Nous avons formulé quatre hypothèses pour vérifier s'il existe des corrélations entre les caractéristiques issues de F0 et la présence ou non des textes libres. Les deux hypothèses suivantes concernent la réinitialisation de F0.

Hypothèse 3 : Il existe généralement une réinitialisation de F0 plus prononcée au début des textes libres par rapport aux autres concepts.

Hypothèse 4 : Il existe généralement une réinitialisation de F0 plus prononcée au début des textes libres directs par rapport aux textes libres indirects.

Enfin, les deux hypothèses suivantes concernent la continuité de F0.

Hypothèse 5 : Il existe généralement une continuité de F0 plus prononcée au début des textes libres par rapport aux autres concepts.

Hypothèse 6 : Il existe généralement une continuité de F0 plus prononcée au début des textes libres directs par rapport aux textes libres indirects.

La suite de ce chapitre présente le cheminement qui a permis de vérifier ces hypothèses.

4.4 Expérimentation

Pour cette expérimentation basée sur F0, les mêmes échantillons que l'expérimentation sur les pauses ont été utilisés. Les mesures de F0 ont été extraites automatiquement avec Praat à intervalle de 10 msec. Toutefois, les caractéristiques issues de F0 étaient plus difficiles à modéliser par rapport aux pauses. Il existait beaucoup de plages dans les signaux où aucune valeur n'était mesurée. Ces plages contenaient surtout des pauses ou des fricatives. Une fenêtre a été appliquée à gauche et à droite de chaque pause entre deux mots. La fenêtre de gauche s'étendait jusqu'à 50 msec avant le début de la pause. Si cette fenêtre ne contenait pas minimalement deux valeurs de F0, celle-ci s'étendait de 50 msec supplémentaires vers la gauche. Cette extension se répétait tant qu'il n'y avait pas minimalement deux valeurs à l'intérieur de cette fenêtre. Toutefois, une extension maximale était atteinte à 400 msec. De son côté, la fenêtre de droite s'étendait jusqu'à 50 msec après la fin de la pause. Nous avons développé un algorithme d'extension identique pour le côté droit. Le modèle de régression linéaire simple était ensuite appliqué sur chacune de ces fenêtres. La droite générée débutait à la première valeur située à l'intérieur de la fenêtre. Elle se terminait à sa dernière valeur. La Figure 4.2 présente des exemples de droites représentant les mesures de F0. Plusieurs droites sont superposées verticalement uniquement pour les besoins de cette explication.

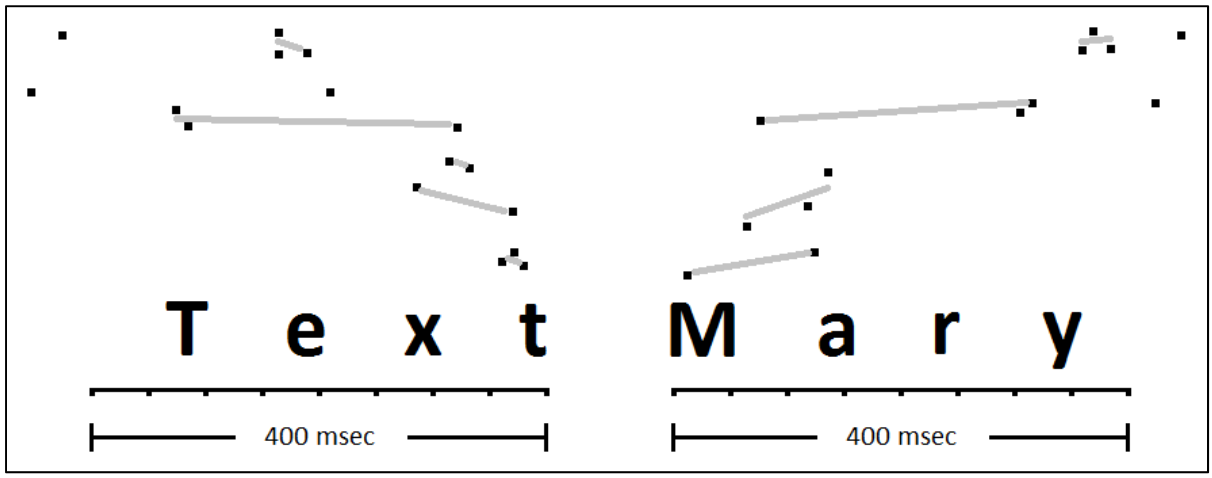

Figure 4.2 Exemples de droites représentant les mesures de F0.

Même après l'atteinte des extensions maximales, certaines fenêtres ne contenaient pas le nombre minimal de valeurs. Dans ces cas, ces frontières étaient ignorées de l'expérimentation. Nous avons développé un algorithme pour tenter également d'ignorer les frontières où se manifestaient les erreurs de réduction de moitié et de doublement de F0. L'algorithme mesurait la différence maximale de hauteur entre deux valeurs voisines utilisées pour la génération d'une même droite. Si cette différence était plus de 75% supérieure par rapport à la hauteur minimale des deux valeurs, cette droite était ignorée. Ce seuil de 75% a été choisi suite aux observations sur des représentations graphiques des mesures de dizaines de commandes vocales. Ce seuil a permis d'obtenir un bon compromis entre les droites faussement acceptées et celles faussement ignorées. De même, pour considérer la réinitialisation de F0, celle-ci devait être inférieure à la hauteur minimale des deux valeurs utilisées pour son calcul.

Si un élément nécessaire au calcul d'une caractéristique était ignoré, la caractéristique elle-même était ignorée pour cette frontière. Puisque toutes ces conditions devaient être respectées, la quantité finale de mesures issues de F0 était significativement réduite par rapport aux pauses. Le Tableau 4.1 présente les paramètres Praat utilisés pour l'extraction de F0.

Tableau 4.1 Paramètres Praat utilisés pour l'extraction de F0

Time step (s)	0.01
Minimum pitch (Hz)	75.0
Maximum pitch (Hz)	600.0

Pour vérifier les hypothèses formulées à la section 4.3.3, des histogrammes ont été générés afin d'obtenir les fréquences des caractéristiques extraites. Comme pour la première expérimentation avec les pauses, les mesures ont été classées selon des intervalles. Les histogrammes étaient compris dans l'intervalle $[-3s, 3s]$ où s représente l'écart type. Les histogrammes ont également été générés avec des fréquences normalisées. Enfin, le test K-S a été utilisé pour comparer les distributions d'une façon statistiquement significative.

4.5 Présentation des résultats

Cette section présente les résultats de cette deuxième expérimentation avec tous les échantillons. Le Tableau 4.2 présente les statistiques des distributions des deux échantillons aléatoires concernant la réinitialisation de F0.

Tableau 4.2 Statistiques des échantillons E-AT et E-AC pour la réinitialisation de F0

Concept	**Quantité**	**Statistiques en hertz**		
		Médiane	**Moyenne**	**Écart type**
Tous les textes libres	554	8.27	9.28	30.62
Textes libres directs	316	9.85	9.43	32.04
Textes libres indirects	238	5.30	9.07	28.64
Autres mots	32 177	0.73	4.00	26.13

La Figure 4.3 et la Figure 4.4 présentent ces distributions.

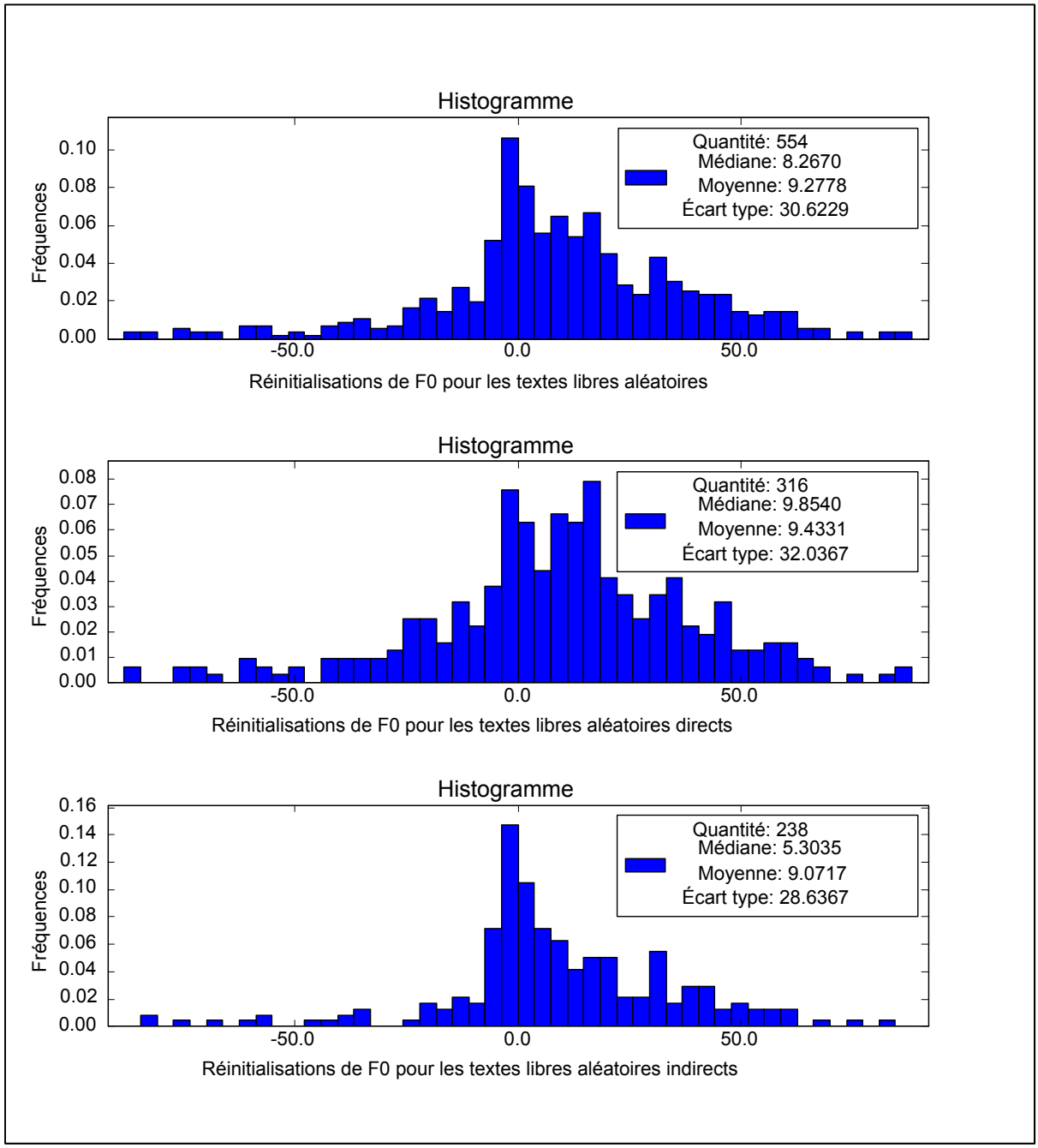

Figure 4.3 Réinitialisation de F0 de l'échantillon E-AT.

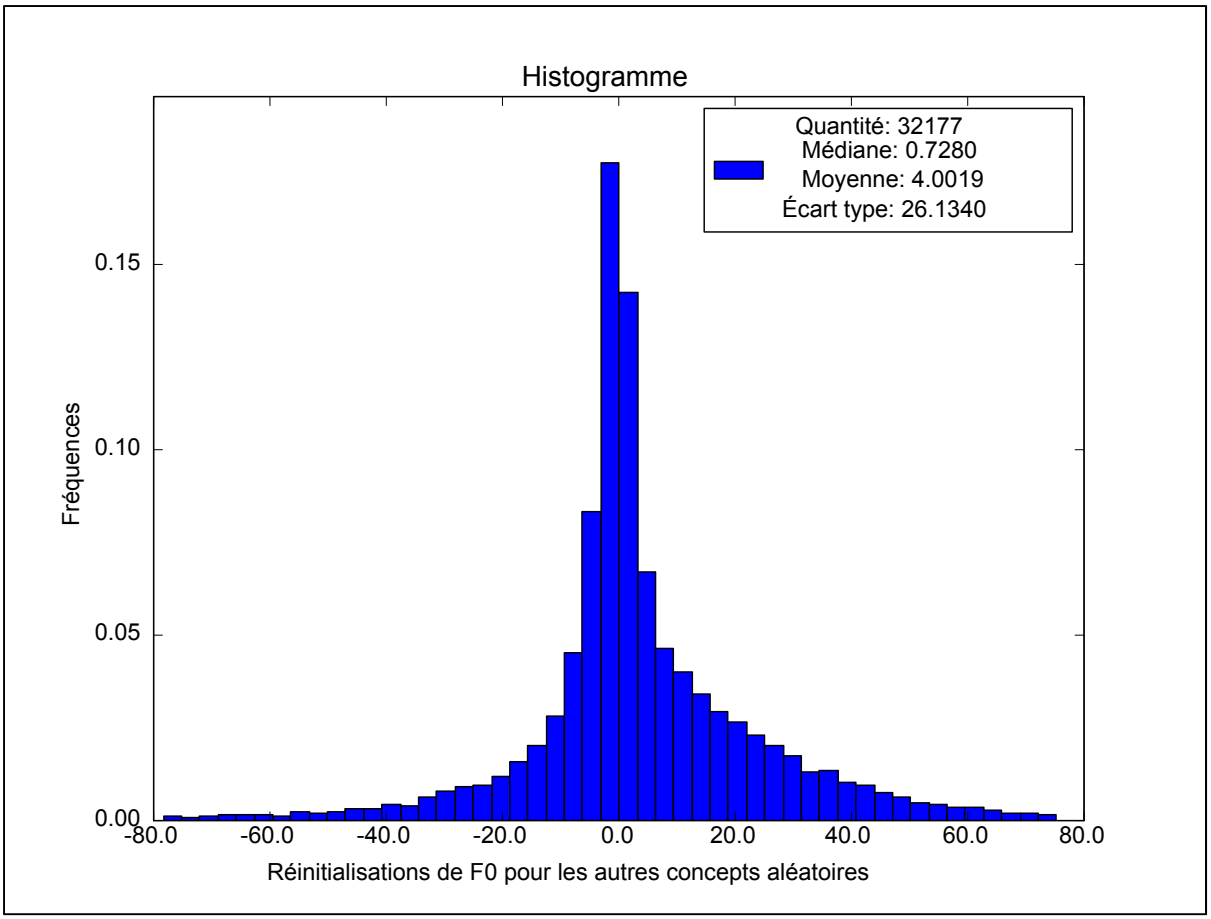

Figure 4.4 Réinitialisation de F0 de l'échantillon E-AC.

Le Tableau 4.3 présente les statistiques des distributions des deux échantillons ciblés concernant la réinitialisation de F0.

Tableau 4.3 Statistiques des échantillons E-CT et E-CC pour la réinitialisation de F0

Concept	Quantité	Statistiques en hertz		
		Médiane	**Moyenne**	**Écart type**
Tous les textes libres	8 806	8.01	10.38	32.59
Textes libres directs	7 216	8.48	10.09	32.52
Textes libres indirects	1 590	5.29	11.71	32.87
Autres mots	80 084	0.84	3.65	26.44

La Figure 4.5 et la Figure 4.6 présentent ces distributions.

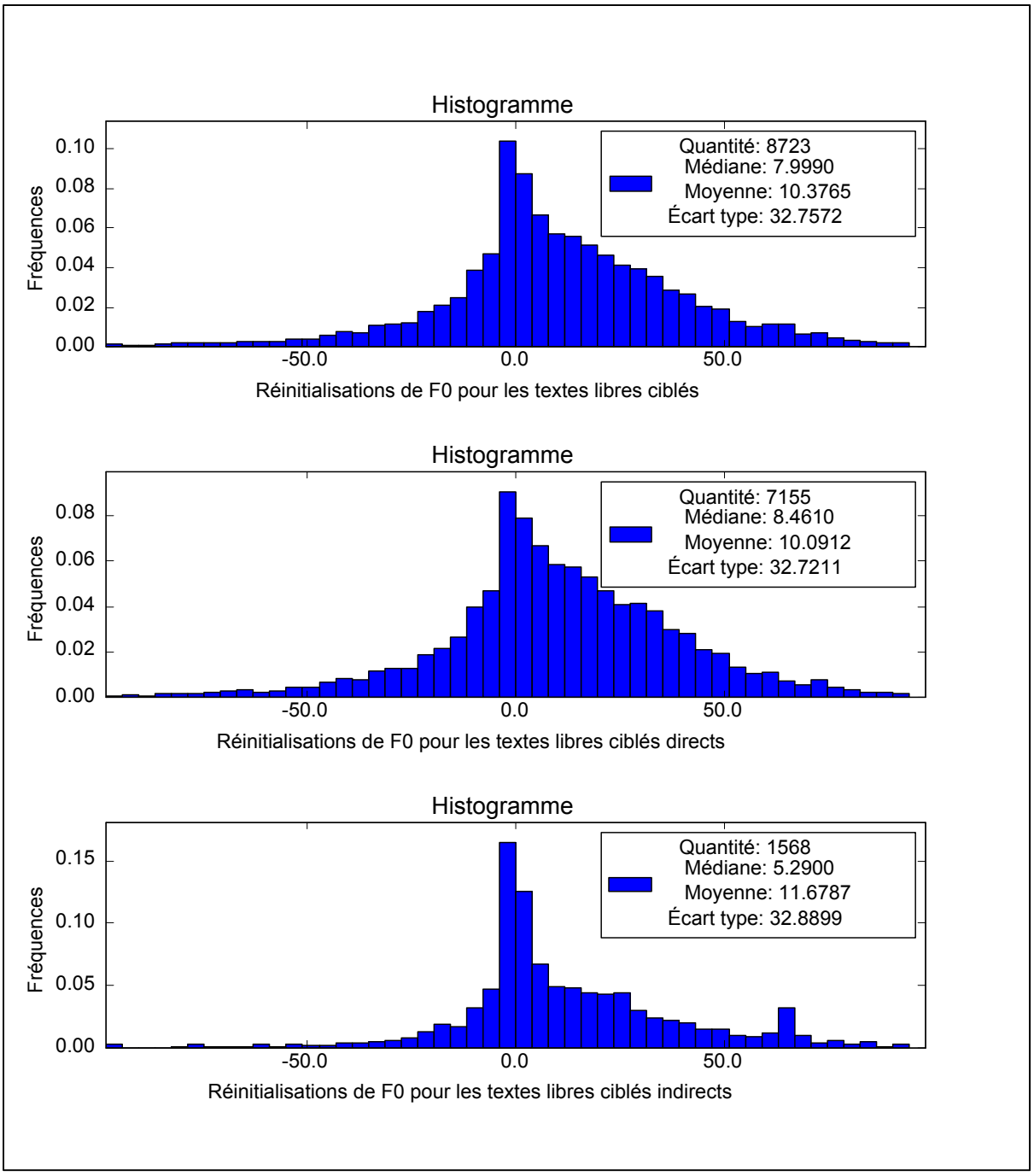

Figure 4.5 Réinitialisation de F0 de l'échantillon E-CT.

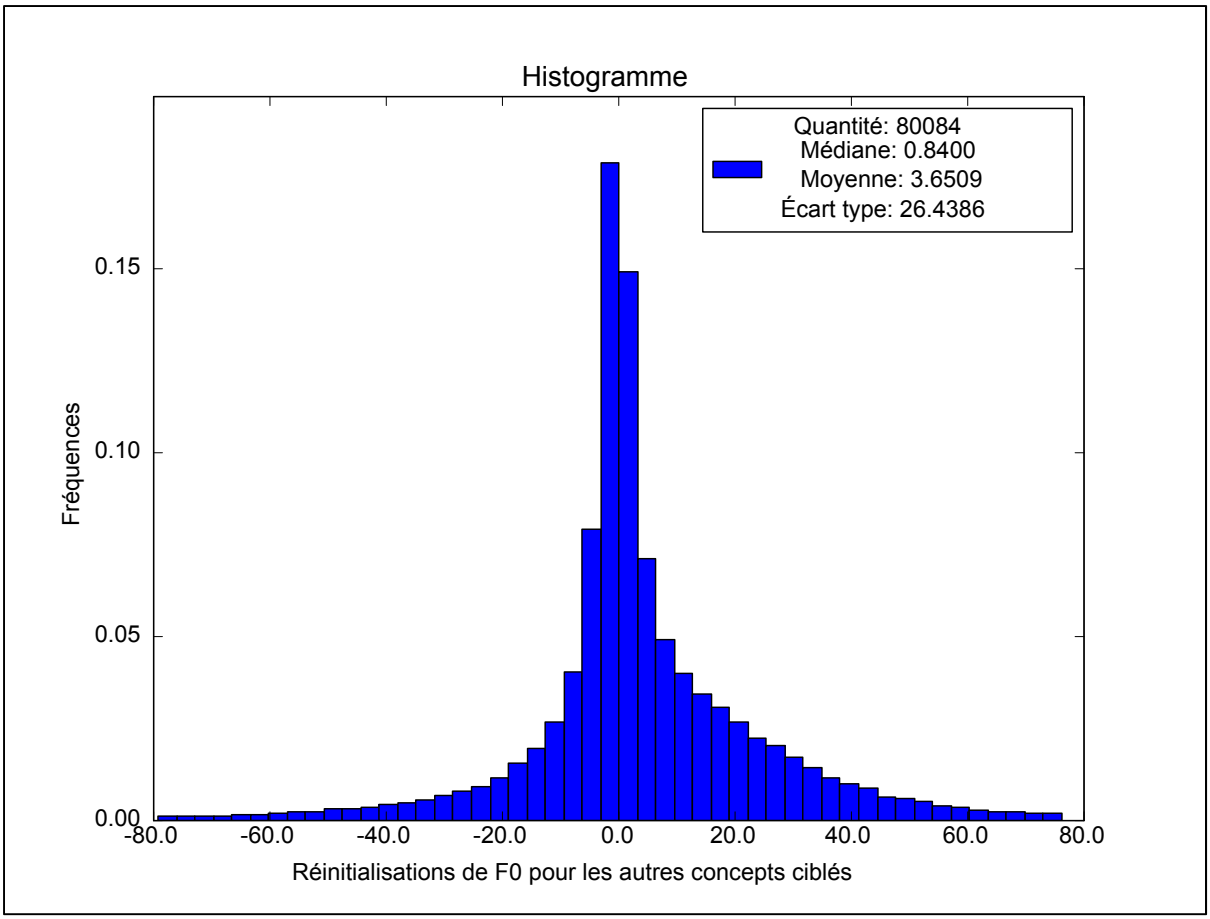

Figure 4.6 Réinitialisation de F0 de l'échantillon E-CC.

Le Tableau 4.4 présente les statistiques des distributions des deux échantillons aléatoires concernant la continuité de F0.

Tableau 4.4 Statistiques des échantillons E-AT et E-AC pour la continuité de F0

Concept	Quantité	Statistiques		
		Médiane	**Moyenne**	**Écart type**
Tous les textes libres	623	104.36	124.79	732.18
Textes libres directs	360	73.10	77.73	722.10
Textes libres indirects	263	164.03	189.20	740.93
Autres mots	34 286	35.58	50.80	668.95

La Figure 4.7 et la Figure 4.8 présentent ces distributions.

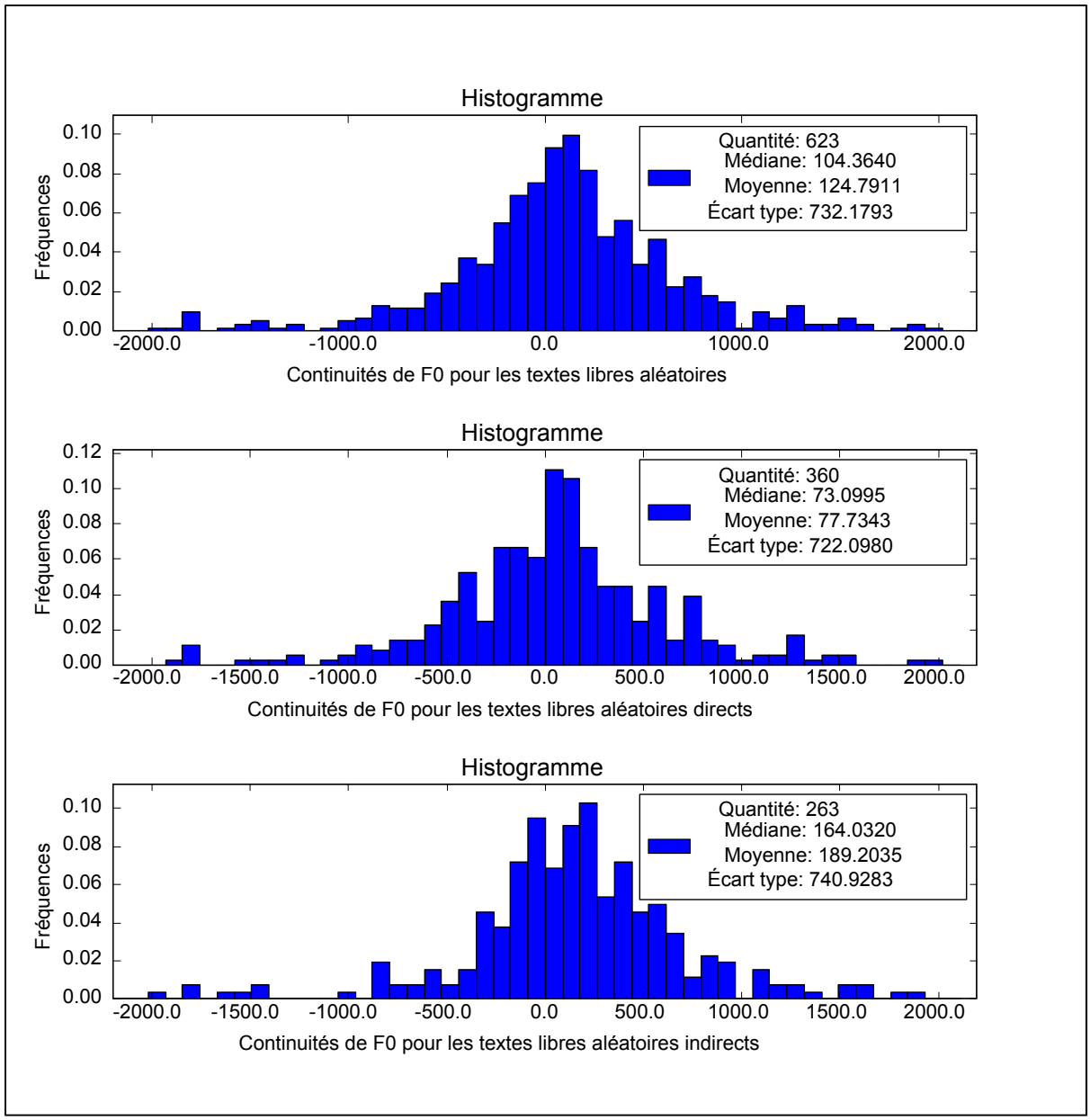

Figure 4.7 Continuité de F0 de l'échantillon E-AT.

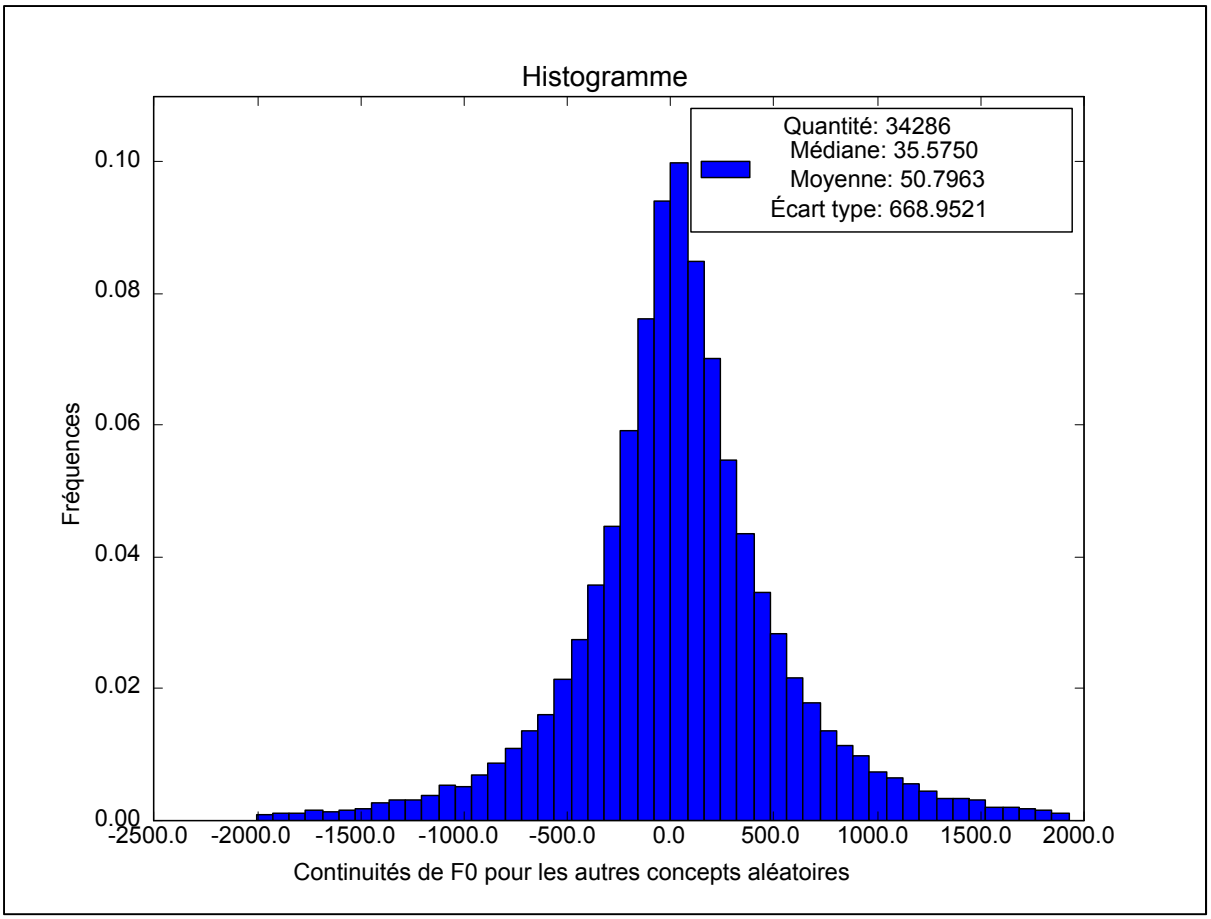

Figure 4.8 Continuité de F0 de l'échantillon E-AC.

Le Tableau 4.5 présente les statistiques des distributions des deux échantillons ciblés concernant la continuité de F0.

Tableau 4.5 Statistiques des échantillons E-CT et E-CC pour la continuité de F0

Concept	Quantité	Statistiques		
		Médiane	**Moyenne**	**Écart type**
Tous les textes libres	9 715	95.87	104.59	718.35
Textes libres directs	7 938	95.86	96.65	718.47
Textes libres indirects	1 777	96.52	140.09	716.76
Autres mots	85 663	45.45	64.96	673.83

La Figure 4.9 et la Figure 4.10 présentent ces distributions.

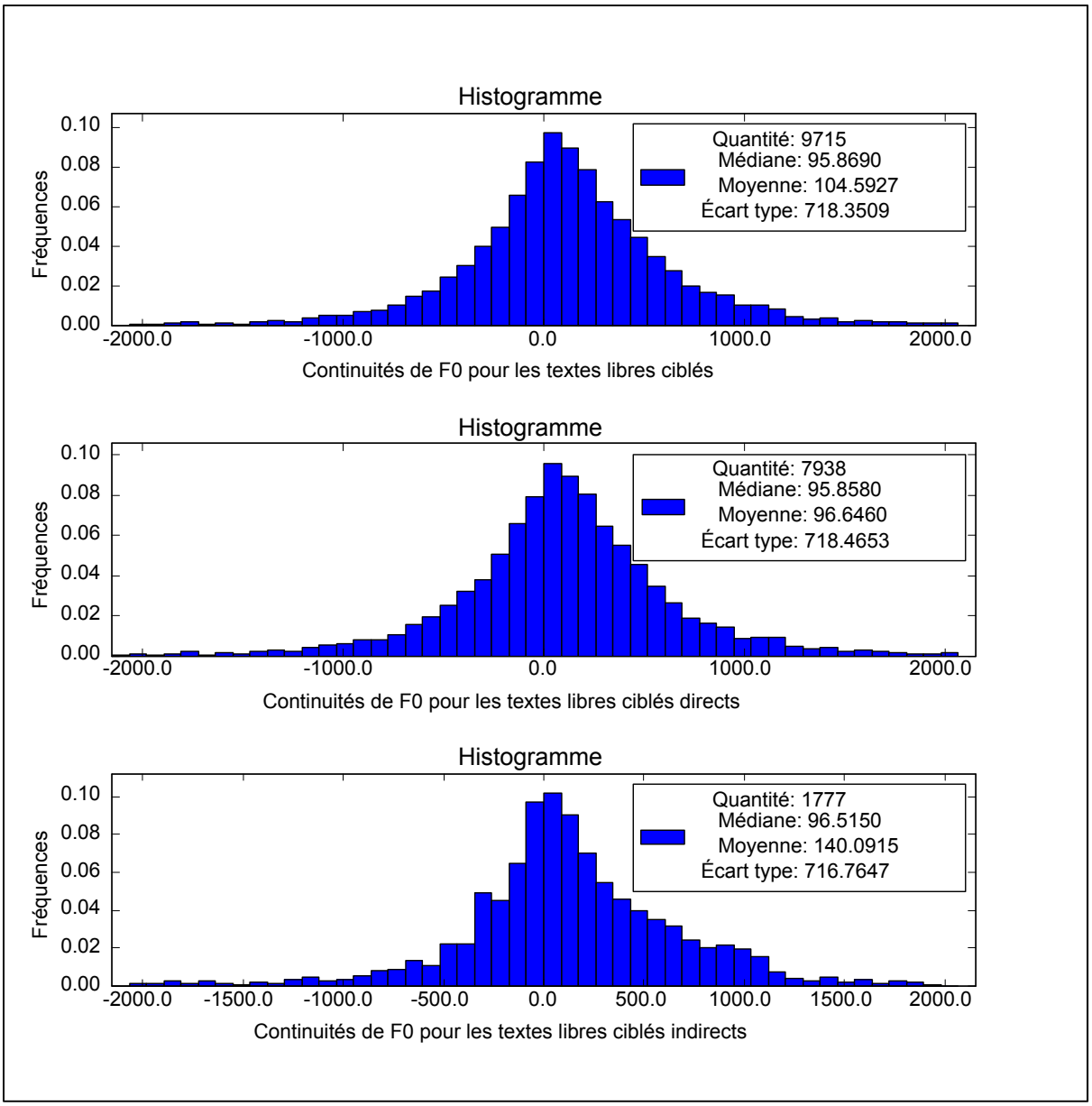

Figure 4.9 Continuité de F0 de l'échantillon E-CT.

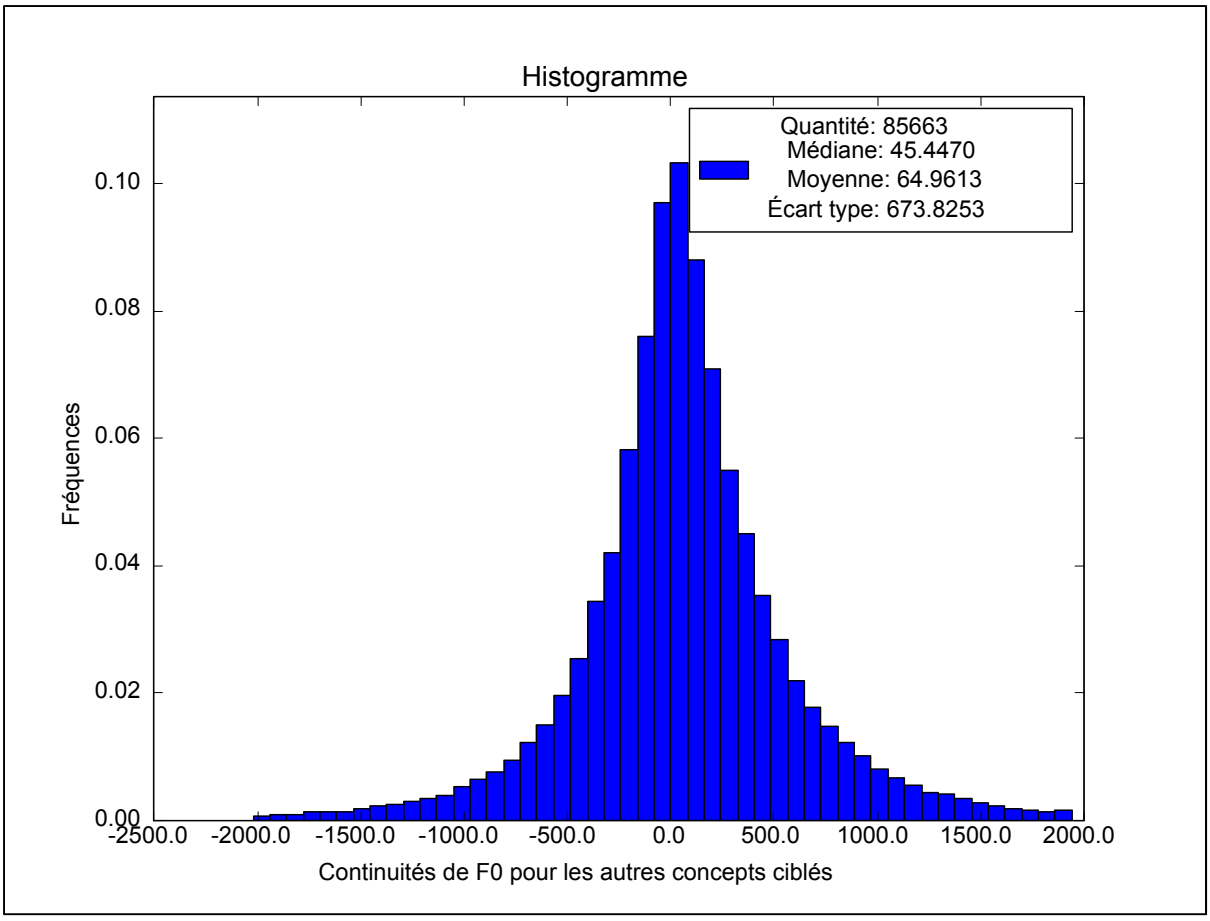

Figure 4.10 Continuité de F0 de l'échantillon E-CC.

4.6 Interprétation de F0

4.6.1 La réinitialisation

Comme l'indique le Tableau 4.2, la moyenne de la réinitialisation de F0 des échantillons aléatoires a été plus élevée avant le début des textes libres par rapport aux autres concepts. Le Tableau 4.3 indique que la même constatation a été observée pour les échantillons ciblés. Pour l'échantillon E-AT, la moyenne pour les textes libres a également été plus élevée pour le sous-échantillon des textes libres directs par rapport au sous-échantillon des textes libres indirects. Cependant, l'inverse a été observé pour l'échantillon E-CT.

Pour déterminer de manière statistiquement significative si ces échantillons appartenaient bien à des distributions différentes, des tests K-S ont été appliqués avec un seuil α=0.05. Pour les besoins du test, des hypothèses nulles ont été formulées selon lesquelles les échantillons E-AT et E-AC et les échantillons E-CT et E-CC appartenaient à la même distribution. Puisque les valeurs-p étaient inférieures à α pour les échantillons aléatoires et ciblés, ces hypothèses ont été rejetées. Il est donc possible d'affirmer avec confiance que les résultats supportaient l'hypothèse 3 formulée à la section 4.3.3.

D'autres hypothèses nulles ont été formulées selon lesquelles les sous-échantillons des textes libres directs et les sous-échantillons des textes libres indirects appartenaient à la même distribution. L'hypothèse nulle a été acceptée pour l'échantillon E-AT, tandis qu'elle a été rejetée pour l'échantillon E-CT. Concernant E-CT, il est possible d'affirmer avec confiance que les sous-échantillons appartenaient à des distributions différentes. Cependant, les résultats ne supportaient pas l'hypothèse 4 formulée à la section 4.3.3. Le Tableau 4.6 présente les résultats de ces tests K-S appliqués pour la réinitialisation de F0.

Tableau 4.6 Résultats des tests K-S pour la réinitialisation de F0

Échantillon	Concept 1	Concept 2	Valeur-p	Statistique K-S	Hypothèse nulle
E-AT, E-AC	Tous les textes libres	Autres mots	1.722e-19	1.997e-1	Rejetée
E-CT, E-CC	Tous les textes libres	Autres mots	4.252e-259	1.943e-1	Rejetée
E-AT	Textes libres directs	Textes libres indirects	1.744e-1	9.374e-2	Acceptée
E-CT	Textes libres directs	Textes libres indirects	2.472e-7	7.837e-2	Rejetée

4.6.2 La continuité

Comme l'indique le Tableau 4.4, la moyenne de la continuité de F0 des échantillons aléatoires était beaucoup plus élevée avant le début des textes libres par rapport aux autres concepts. Le Tableau 4.5 indique que la même constatation a été observée pour les

échantillons ciblés. Pour l'échantillon E-AT, la moyenne pour les textes libres a toutefois été beaucoup moins élevée avant le début des textes libres directs par rapport aux textes libres indirects. La même constatation a été observée pour l'échantillon E-CT.

Pour déterminer de manière statistiquement significative si ces échantillons appartenaient bien à des distributions différentes, des tests K-S ont été appliqués avec un seuil α=0.05. Des hypothèses nulles ont été formulées selon lesquelles les échantillons E-AT et E-AC et les échantillons E-CT et E-CC appartenaient à la même distribution. Puisque les valeurs-p étaient inférieures à α pour les échantillons aléatoires et ciblés, ces hypothèses ont été rejetées. Il est donc possible d'affirmer avec confiance que les résultats supportaient l'hypothèse 5 formulée à la section 4.3.3.

D'autres hypothèses nulles ont été formulées selon lesquelles les sous-échantillons des textes libres directs et les sous-échantillons des textes libres indirects appartenaient à la même distribution. Ces hypothèses ont également été rejetées à la fois pour les échantillons aléatoires et ciblés. Il est donc possible d'affirmer avec confiance que les sous-échantillons appartenaient à des distributions différentes. Cependant, les résultats ne supportaient pas l'hypothèse 6 formulée à la section 4.3.3. Le Tableau 4.7 présente les résultats de ces tests K-S appliqués pour la continuité de F0.

Tableau 4.7 Résultats des tests K-S pour la continuité de F0

Échantillon	Concept 1	Concept 2	Valeur-p	Statistique K-S	Hypothèse nulle
E-AT, E-AC	Tous les textes libres	Autres mots	2.304e-4	8.565e-2	Rejetée
E-CT, E-CC	Tous les textes libres	Autres mots	8.188e-32	6.427e-2	Rejetée
E-AT	Textes libres directs	Textes libres indirects	7.289e-3	1.345e-1	Rejetée
E-CT	Textes libres directs	Textes libres indirects	2.745e-2	3.831e-2	Rejetée

4.7 Discussion

Concernant la réinitialisation de F0, il n'y avait pas suffisamment de données dans l'échantillon E-AT pour bien identifier la forme des distributions des textes libres. Pour

l'échantillon E-CT, il y avait suffisamment de données. Les distributions de celui-ci avaient l'apparence d'une asymétrie à droite. Toutefois, les valeurs extrêmes pouvaient contredire cette observation. L'analyse de ces distributions et de leur moyenne suggère que les locuteurs augmentaient généralement la hauteur lorsqu'ils débutaient à dicter un texte libre, qu'il soit direct ou indirect. Cependant, cette constatation a également été observée à un degré moindre en examinant les autres concepts. Cela suggère que les locuteurs avaient tendance à augmenter la hauteur en débutant un nouveau mot, débutant un texte libre ou non.

Étonnamment, la réinitialisation de F0 semblait moins prononcée au début des textes libres direct par rapport aux textes libres indirects. Cela était peut-être attribué à la hauteur souvent basse de la fenêtre droite du mot précédent le début du texte libre indirect. Par exemple, dans le texte libre « Remind me at nine **to** [watch the tv] », le mot « to » se termine par une fenêtre de hauteur basse. La gamme des mots précédents les textes libres indirects était très restreinte et ce mot était souvent présent. Cette hauteur basse précédant les textes libres indirects pourrait donc justifier cette observation contre-intuitive.

Concernant la continuité de F0, il n'y avait également pas suffisamment de données dans l'échantillon E-AT pour analyser la distribution des textes libres. Pour l'échantillon E-CT, contrairement à celle observée pour la réinitialisation, aucune asymétrie ne semblait exister au niveau des distributions. La moyenne était positive et plus prononcée pour les textes libres par rapport aux autres concepts. Cela suggère que les locuteurs avaient tendance à diminuer la hauteur à la fin du mot précédent un texte libre, pour ensuite remonter au début de celui-ci. Cependant, cette constatation a également été observée à un degré moindre en examinant les autres concepts. Cela suggère que les locuteurs avaient tendance à briser la trajectoire de la hauteur en débutant un nouveau mot, débutant un texte libre ou non.

Enfin, la continuité de F0 semblait en moyenne moins prononcée pour les textes libres directs par rapport aux textes libres indirects. Cela pourrait encore s'expliquer par la hauteur de la

fenêtre droite du mot « to » précédant majoritairement les textes libres indirects. Une analyse visuelle a été effectuée concernant les mesures de F0 extraites de quelques commandes vocales. Il a été observé que la fenêtre droite du mot « to » possédait majoritairement une droite de pente négative relativement prononcée. De leur côté, l'ensemble des autres mots faisait varier cette pente plus ou moins négativement ou positivement. La moyenne de toutes ces variations atténuait leur importance. Toutefois, étant donné que le mot « to » était majoritaire pour les textes libres indirects, la moyenne des pentes était renforcée négativement. Une pente positive (montée de F0 au début d'un texte libre), soustraite par une pente négative prononcée (descente de F0 à la fin d'un mot précédent un texte libre indirect), génère une continuité de F0 positive et prononcée. Celle-ci est supérieure par rapport à la soustraction avec une pente négative moins prononcée, généralement rencontrée avant le début des textes libres directs.

4.8 Conclusion

Cette expérimentation basée sur F0 a supporté l'hypothèse 3 selon laquelle il existe généralement une réinitialisation de F0 plus prononcée au début des textes libres par rapport aux autres concepts. Cependant, l'expérimentation n'a pas supporté l'hypothèse 4. En effet, les résultats ont démontré généralement une réinitialisation de F0 moins prononcée au début des textes libres directs par rapport aux textes libres indirects. Même si cette hypothèse n'a pas été supportée, le test K-S suggère que ces deux sous-échantillons appartenaient à des distributions différentes.

L'expérimentation a supporté l'hypothèse 5 selon laquelle il existe généralement une continuité de F0 plus prononcée au début des textes libres par rapport aux autres concepts. Enfin, celle-ci n'a pas supporté l'hypothèse 6. En effet, les résultats ont démontré généralement une continuité de F0 moins prononcée au début des textes libres directs par rapport aux textes libres indirects. Encore une fois, le test K-S suggère que ces deux sous-échantillons appartenaient à des distributions différentes.

Grâce aux deux expérimentations, il a été démontré qu'il existe des corrélations, entre les caractéristiques prosodiques basées sur les pauses et sur F0, et la présence ou non des textes libres dans une commande vocale. Le prochain chapitre présente la dernière expérimentation basée sur un système NLU. Le premier objectif était de vérifier s'il était possible d'améliorer ses performances pour l'identification du début des textes libres, grâce à l'ajout de caractéristiques prosodiques. Enfin, le second objectif concernait l'identification des mentions en général.

CHAPITRE 5

EXPÉRIMENTATION SUR UN SYSTÈME NLU

5.1 Introduction

Ce chapitre présente la dernière expérimentation de ce mémoire. Celle-ci consiste à déterminer si l'ajout des informations prosodiques à un système NLU améliore ses performances d'identification des mentions. Comme mentionné dans l'introduction du CHAPITRE 2, l'identification des textes libres est au cœur de ce mémoire. Toutefois, l'intérêt principal concerne l'amélioration globale des performances du système. Pour cette expérimentation, il a été nécessaire d'utiliser un grand ensemble de commandes vocales. Celles-ci ont été représentées grâce à leurs caractéristiques prosodiques et lexicales. Pour commencer, un modèle du système a été appris à partir d'un ensemble d'entraînement constitué seulement des caractéristiques lexicales. Sa performance a été évaluée à partir d'un ensemble de test, différent de celui d'entraînement. Ensuite, d'autres modèles ont été appris à partir de ce même ensemble d'entraînement, mais augmenté avec les caractéristiques prosodiques. Leur performance a été évaluée sur le même ensemble de test, également augmenté. Pour ces derniers modèles, si un gain significatif de performance était observé par rapport au premier, cela confirme que l'ajout de leurs informations prosodiques a amélioré les performances du système.

Ce chapitre présente d'abord le système NLU utilisé ainsi que la formulation de nos hypothèses pour cette expérimentation. L'expérimentation est présentée et est suivie de la présentation et de la discussion des résultats.

5.2 Définition du système NLU

Le système NLU utilisé pour cette expérimentation a été un classifieur de mentions. Pour les classifieurs, un entraînement supervisé est nécessaire, c.-à-d. qu'une étiquette catégorielle (mention) est associée au vecteur de caractéristiques pour chaque exemple. Dans cette

expérimentation, un vecteur de caractéristiques a été utilisé pour représenter chaque mot d'une commande vocale. Ces vecteurs ont été produits à partir des informations prosodiques et lexicales utilisées lors des expérimentations précédentes. Les commandes devaient également être considérées dans leur ensemble, c.-à-d. qu'une relation était assumée entre les mots d'une même commande. Les classifieurs classiques assument que tous les exemples sont distribués de façon identique et indépendante, ce qui ne s'appliquait pas pour cette expérimentation. Néanmoins, l'indépendance de toutes les commandes pouvait être assumée. Par conséquent, un classifieur du type CRF (Conditional Random Fields) a été utilisé. Ce classifieur de séquence, décrit à la section 5.2.2, permet de tenir compte de l'aspect temporel du problème. L'étiquetage de séquence consiste à traiter une séquence de données en entrée, afin de produire en sortie une séquence d'étiquettes discrètes qui la caractérisent.[17]

Dans une commande, la même mention était souvent assignée à une séquence de mots consécutifs. Pour un mot, l'identification de sa mention était plus difficile si une mention différente était assignée à son mot précédent. Pour aider le classifieur, une étiquette différente a été assignée au premier mot d'une séquence par rapport aux autres mots de celle-ci. Par exemple, pour une séquence assignée à la mention « text », l'étiquette « B-text » était assignée au premier mot, tandis que l'étiquette « I-text » était assignée aux autres mots. Cette règle a été appliquée pour toutes les mentions. La section suivante présente un classifieur couramment utilisé, suivie du classifieur utilisé pour cette expérimentation.

5.2.1 Classifieur Wapiti

Wapiti[5] est une boîte à outils développée par LIMSI-CNRS pour la segmentation et l'étiquetage de séquences avec des modèles discriminatifs. Wapiti est basé sur des modèles d'entropie maximale (MaxEnt), de Markov d'entropie maximale et de CRF à chaîne linéaire. Il propose diverses méthodes d'optimisation et de régularisation pour améliorer à la fois la

[5] http://wapiti.limsi.fr/

complexité de calcul et de la performance de prédiction des modèles standards. Cependant, Wapiti possède un inconvénient majeur présenté à la section suivante.

5.2.1.1 Problèmes de la discrétisation

Wapiti ne gère pas les caractéristiques à valeur continue telles que les mesures prosodiques recueillies lors des expérimentations précédentes. Pour être exploitables, ces mesures devaient nécessairement être discrétisées. Cependant, la discrétisation diminue drastiquement le nombre de valeurs possibles que peut prendre une mesure. Plus grave encore, elle fait croître très rapidement la dimension du vecteur de caractéristiques fourni en entrée au classifieur. En effet, le nombre de valeurs possibles que peut prendre une mesure discrétisée se traduit par l'ajout d'un nombre similaire de caractéristiques booléennes dans le vecteur. Cette représentation booléenne multiple est nécessaire afin de conserver une notion d'écart entre les valeurs, comme le font naturellement les valeurs continues. Une croissance significative de ce vecteur peut alors engendrer la malédiction de la dimension[6] lors de l'entraînement du classifieur, un problème bien connu en apprentissage machine. Ce problème peut réduire les performances du classifieur. La section suivante présente un autre classifieur qui a permis d'éviter cette discrétisation.

5.2.2 Classifieur CRFSuite

CRFSuite[7] est une implémentation de CRF à chaîne linéaire pour la segmentation et l'étiquetage de séquences. Il a été conçu par Naoaki Okazaki de l'université de Tohoku au Japon. La priorité première de ce logiciel est de générer et d'utiliser des modèles CRF le plus rapidement possible, même au détriment de son espace mémoire et de la généralité du code source. Contrairement à Wapiti, CRFSuite gère facilement les caractéristiques à valeur continue. La discrétisation des mesures prosodiques n'était donc pas nécessaire. Par

[6] http://fr.wikipedia.org/wiki/Fl%C3%A9au_de_la_dimension
[7] http://www.chokkan.org/software/crfsuite/

conséquent, CRFSuite a été choisi comme système NLU pour cette expérimentation. La section suivante présente la formulation des hypothèses.

5.2.3 Formulation des hypothèses

Deux hypothèses ont été définies pour vérifier si les informations prosodiques pouvaient aider un système NLU à identifier correctement les mentions. La première hypothèse concerne le début des textes libres (mentions « B-text » et « B-title »). La deuxième hypothèse concerne toutes les mentions. Voici ces deux hypothèses :

Hypothèse 7 : L'ajout des informations prosodiques au système NLU améliore ses performances d'identification du début des textes libres.

Hypothèse 8 : L'ajout des informations prosodiques au système NLU améliore ses performances globales d'identification des mentions.

La suite de ce chapitre présente le cheminement nécessaire qui a permis de supporter ou non ces hypothèses.

5.3 Expérimentation

Cette expérimentation a utilisé les mêmes échantillons de données que ceux utilisés lors des expérimentations précédentes. Afin de favoriser le fonctionnement du système NLU, une adaptation a été nécessaire concernant les caractéristiques prosodiques et lexicales recueillies précédemment. La section suivante présente comment ces caractéristiques ont été représentées dans ce système, suivi de la méthodologie utilisée pour cette expérimentation.

5.3.1 Représentation des caractéristiques

Différentes combinaisons de types de caractéristique prosodique, ainsi que les caractéristiques lexicales, ont été utilisées pour ce système. Le Tableau 5.1 présente les cinq types de modèle étudiés.

Tableau 5.1 Cinq types de modèle étudiés pour le système NLU

Type de modèle	Caractéristiques utilisées
A	Lexicales seules
B	Lexicales et les durées des pauses
C	Lexicales et les réinitialisations de F0
D	Lexicales et les continuités de F0
E	Lexicales et les trois types de caractéristiques prosodiques

Le type de modèle A a servi de référence pour déterminer les gains de performance des quatre autres types. La section suivante présente comment les caractéristiques prosodiques ont été représentées dans ce système, suivi de celle concernant les caractéristiques lexicales.

5.3.1.1 Caractéristiques prosodiques

Les mesures prosodiques, issues des expérimentations précédentes, ont été normalisées pour favoriser l'apprentissage des modèles. Cette normalisation était nécessaire afin que les trois types de caractéristiques puissent détenir des poids du même ordre. Par exemple, les durées de pause originales possédaient un intervalle de valeurs beaucoup plus restreint que celui des réinitialisations de F0 ou des continuités de F0. La normalisation a permis de contraindre ces trois types sous des intervalles similaires, variant grossièrement entre zéro et un. Ces nouveaux intervalles ont été générés à partir de la totalité des mesures recueillies sur l'ensemble de données ciblées.

Pour la durée des pauses, la valeur zéro du nouvel intervalle correspondait aux pauses de durées nulles, tandis que la valeur un correspondait à trois fois son écart type. Suite à la normalisation, des valeurs supérieures à un pouvaient avoir été assignées pour les valeurs extrêmes. Pour les deux autres types de caractéristiques, la valeur zéro du nouvel intervalle correspondait à trois fois l'écart type dans le sens négatif, et la valeur une à trois fois l'écart type dans le sens positif. Des valeurs à l'extérieur du nouvel intervalle pouvaient également avoir été assignées pour les valeurs extrêmes. Pour les trois types de caractéristiques, les mesures manquantes ont simplement été ignorées.

5.3.1.2 Caractéristiques lexicales

Afin d'être comparable aux systèmes NLU courants, d'autres caractéristiques lexicales ont été utilisées en plus des mots. Celles-ci ont été représentées sous forme de N-grammes. Dans ce contexte, un N-gramme est une succession de N mots extraits à partir d'une commande vocale. Un maximum de neuf caractéristiques a été généré à partir de ces N-grammes. Ces caractéristiques ont été représentées pour chaque mot de position i d'une commande. Celles-ci ont concerné tous les uni-grammes et les bi-grammes constitués des mots situés entre les positions i-2 et i+2 inclusivement. Les N-grammes impossibles à extraire, dus aux frontières de la commande, ont simplement été ignorés.

5.3.2 Méthodologie

Pour le système NLU à l'étude, il était préférable d'obtenir les performances les plus élevées possibles. Pour les atteindre, son algorithme d'apprentissage devait être configuré avec des valeurs de paramètre précises. Le terme « configuration optimale » désigne ces valeurs. Un choix parmi cinq algorithmes était offert avec CRFSuite. La quantité et les types de paramètre offerts étaient propres à chaque algorithme. Pour des performances similaires, l'algorithme possédant le moins de paramètres était un meilleur choix. Cette caractéristique facilitait grandement la recherche de sa configuration optimale. Parmi les algorithmes offerts,

« Averaged Perceptron »[8] semblait être le meilleur, car celui-ci ne possédait que deux paramètres. Un premier paramètre correspondait au nombre d'itérations requis pour l'apprentissage des modèles. Tous les algorithmes offraient ce paramètre. Afin d'optimiser l'apprentissage, il a été décidé que ce paramètre serait toujours fixé à 200, peu importe l'algorithme. Le second et dernier paramètre était « Epsilon » et correspondait à une valeur continue. La configuration optimale dépendait donc seulement de ce dernier. Malgré la simplicité de cet algorithme, des tests préliminaires ont démontré des performances similaires par rapport aux autres algorithmes. Par conséquent, l'algorithme « Averaged Perceptron » a été choisi pour cette expérimentation.

L'expérimentation s'est effectuée essentiellement en trois phases. La section 5.3.2.1 présente les corpus de données utilisés pour chacune de ces phases. La phase de développement a permis d'estimer la fiabilité des performances des modèles appris en fonction de leur configuration. Une configuration optimale a été obtenue pour chacun des cinq types de modèle étudiés. Pour ce, la section 5.3.2.2 présente la technique de validation croisée utilisée. La phase d'entraînement a généré des modèles en fonction des configurations optimales obtenues précédemment. Grâce à ces modèles, la phase de test a généré les performances finales. Les hypothèses de cette expérimentation ont été vérifiées grâce à ces performances. La section 5.3.2.3 présente l'approche utilisée pour vérifier ces hypothèses.

5.3.2.1 Corpus de données

Au cours des phases de développement et d'entraînement, les modèles ont été appris à partir d'un ensemble de données ciblé. Cet ensemble correspondait à un sous-ensemble des données ciblé utilisé lors des expérimentations précédentes. Une quantité significative de commandes vocales ont été ignorées par un algorithme de conversion de format de fichier, faute d'obtenir une liaison adéquate avec certaines informations d'origine. Ce sous-ensemble contenait 30 905 commandes. Il était constitué de 12 655 commandes possédant des textes

[8] http://cseweb.ucsd.edu/~yfreund/papers/LargeMarginsUsingPerceptron.pdf

libres, et 18 250 commandes ne possédant aucun texte libre. Pour la phase de développement, ce sous-ensemble a été divisé alternativement en des ensembles d'apprentissage et de validation. Toutefois, celui-ci a été considéré entièrement pour la phase d'entraînement. Pour la phase de test, les modèles générés à la phase d'entraînement ont été évalués à partir d'un ensemble de données aléatoire. Cet ensemble correspondait à un sous-ensemble des données aléatoire utilisé lors des expérimentations précédentes. Celui-ci a été réduit afin de conserver la même proportion des commandes possédant des textes libres par rapport au sous-ensemble ciblé. Ce sous-ensemble contenait 1 587 commandes. Il était constitué de 650 commandes possédant des textes libres, et 937 commandes ne possédant aucun texte libre.

5.3.2.2 La validation croisée

La technique de validation croisée, utilisée lors de la phase de développement, a été la « *k*-fold cross-validation »[9]. Pour cette technique, les données ont été divisées en k sous-ensembles. Un des k sous-ensembles a été sélectionné comme ensemble de validation. Les $k-1$ autres sous-ensembles ont été sélectionnés comme ensemble d'apprentissage. Un modèle a alors été appris sur l'ensemble d'apprentissage, puis a été évalué par rapport à l'ensemble de validation. L'opération a été répétée en sélectionnant un autre ensemble de validation. Cet ensemble a été choisi parmi les $k-1$ sous-ensembles n'ayant pas encore été utilisés pour l'évaluation du modèle. L'opération s'est répété ainsi k fois afin que chaque sous-ensemble de validation soit utilisé exactement une fois pour l'évaluation. La moyenne des k performances obtenues a finalement été calculée pour estimer la performance du modèle. Afin de réduire les temps de traitement, k a été fixé à 5.

5.3.2.3 Vérification des hypothèses

Pour confirmer les hypothèses de cette expérimentation, des gains de performance devaient être observés. Ces gains correspondaient aux comparaisons des performances, concernant les

---

[9] http://fr.wikipedia.org/wiki/Validation_crois%C3%A9e

types de modèle ajoutés de la prosodie, par rapport à la référence sans prosodie. Le test de McNemar, décrit à l'ANNEXE IV, a été utilisé pour s'assurer que ces gains étaient statistiquement significatifs. Cependant, les modèles appris à partir du corpus ciblé, lors de la phase d'entraînement et selon les configurations optimales associées, ont été appliqués sur le corpus aléatoire lors de la phase de test. Ces deux corpus étaient de nature différente, ce qui était problématique. En effet, ces modèles n'étaient pas optimisés par rapport à la nature du second corpus. Par conséquent, certains gains déclarés non significatifs selon le test de McNemar pouvaient masquer des résultats prometteurs. Avec des données de nature identique, peut-être que ces gains auraient été plus prononcés et ainsi être significatifs.

Une approche arbitraire a été utilisée pour accepter des gains même s'ils échouaient ce test. Lors de la phase d'entraînement, les modèles ont été appris par rapport à leur configuration optimale. L'évaluation de ces modèles, lors de la phase de test, a généré une performance pour chacun d'eux. Toutefois, des modèles ont également été appris et évalués par rapport aux configurations se situant autour des configurations optimales. Ces configurations ont été obtenues en ne variant que les valeurs du paramètre « Epsilon » de l'algorithme d'apprentissage. Ainsi, pour chaque type de modèle, une courbe représentant ses performances a été générée. Si un gain était déclaré non significatif, mais qu'une analyse visuelle des courbes suggérait vraisemblablement un gain, la signifiance de celui-ci était acceptée. Pour ce, ces gains devaient persister en général autour de la configuration optimale. Ces courbes ont également été générées lors de la phase de développement, mais n'étaient pas considérées pour cette analyse. À cette phase, la performance maximale d'un type de modèle correspondait à sa configuration optimale. L'ANNEXE V présente ces courbes lors de la phase de développement, ainsi que celles lors de la phase de test.

5.4 Types de classification

L'évaluation finale a été effectuée par rapport à deux types de classification. La classification « textes libres » a concerné l'identification des textes libres exclusivement. La classification

« multi-mentions » a concerné l'identification des textes libres et des mentions en général. Les sections suivantes présentent ces deux types et les résultats obtenus.

5.4.1 Classification texte libre

La classification texte libre a permis de faciliter la tâche de classification en diminuant considérablement la diversité des mentions disponibles. À partir des 270 mentions distinctes de l'ensemble d'entraînement, une régression a été appliquée pour les restreindre à seulement cinq mentions. Cette régression a permis de distinguer exclusivement les textes libres des autres concepts. Le Tableau 5.2 présente ces nouvelles mentions et la façon dont elles ont été obtenues. Ce tableau est suivi des résultats pour ce type de classification.

Tableau 5.2 Mentions utilisées pour la classification texte libre

Mention	**Détail**
B-FreeText	Fusion des mentions « B-text » et « B-title »
I-FreeText	Fusion des mentions « I-text » et « I-title »
B-Others	Fusion de toutes les autres mentions débutant par « B- »
I-Others	Fusion de toutes les autres mentions débutant par « I- »
O	Mention assignée aux mots de moindre intérêt

5.4.1.1 Présentation des résultats

Le Tableau 5.3 présente les performances globales d'identification pour chaque type de modèle présenté à la section 5.3.1. Ce tableau présente également les gains de performance des quatre derniers types (avec prosodie) par rapport à la référence (sans prosodie).

Tableau 5.3 Performances globales pour la classification texte libre

Type de modèle	Description	Performance (%)	Gain
A	Lexical	73.21	(Référence)
B	Lex. + durée des pauses	73.86	+0.65
C	Lex. + réinit. de F0	73.43	+0.22
D	Lex + cont. de F0	73.52	+0.31
E	Toutes	74.18	+0.97

L'intérêt concerne surtout les performances d'identification du début des textes libres. Pour ce, le Tableau 5.4 présente la précision, le rappel et la F-mesure concernant la mention « B-FreeText ». Il présente également le ratio des cellules b et c du test de McNemar selon les statistiques de l'ANNEXE VI. Pour ce ratio, une valeur supérieure à un correspond à un gain.

Tableau 5.4 Performances pour « B-FreeText » pour la classification texte libre

Type de modèle	Performance (%)			Ratio b/c McNemar
	Précision	Rappel	F-mesure	
A	60.32	57.73	59.00	(Référence)
B	63.68	60.95	62.29	1.71
C	60.03	58.19	59.10	0.97
D	60.76	58.81	59.77	1.09
E	62.10	59.72	60.89	1.35

5.4.1.2 Discussion

Les résultats du Tableau 5.3 ont suggéré que des gains de performances pouvaient être observés lorsque des caractéristiques prosodiques étaient combinées aux caractéristiques lexicales. Ces gains étaient prononcés avec le type de modèle B (+0.65% pour les durées des pauses) et davantage prononcés avec le type E (+0.97% pour les trois types de caractéristiques prosodiques). Pour vérifier si les gains étaient statistiquement significatifs, le

test de McNemar a été utilisé. Selon ce test, seuls les types B et E ont obtenu des gains significatifs. Toutefois, une analyse des courbes des performances de la Figure-A V-3 de l'ANNEXE V a suggéré que les gains du type C (réinitialisations de F0) pouvaient être significatifs.

Pour le début des textes libres, les résultats du Tableau 5.4 ont suggéré que des gains, au niveau du ratio b/c du test de McNemar, pouvaient être observés concernant la mention « B-FreeText ». Ces gains étaient prononcés avec le type E (1.35), mais davantage prononcés avec le type B (1.71). Cependant, une perte a été observée avec le type C (0.97). Pour appliquer le test de McNemar, toutes les mentions ne correspondant pas à « B-FreeText » ont été converties sous la mention « Others ». Selon ce test, seul le type B a obtenu des gains significatifs. Toutefois, l'analyse des courbes des performances (F-mesure) de la Figure-A V-6 a suggéré que les gains du type E pouvaient être significatifs. Cela confirme, pour ce type de classification et ces deux types de modèles seulement, l'hypothèse 7 formulée à la section 5.2.3.

La section suivante présente la classification multi-mentions. L'hypothèse 8 est vérifiée, ainsi que l'hypothèse 7 en distinguant les mentions « B-text » et « B-title ».

5.4.2 Classification multi-mentions

Pour la classification multi-mentions, la diversité entière des 270 mentions originales de l'ensemble d'entraînement a été utilisée. Ce type de classification était plus représentatif des systèmes NLU courants. La section suivante présente les résultats pour ce type de classification.

5.4.2.1 Présentation des résultats

Le Tableau 5.5 présente les performances globales d'identification pour chaque type de modèle. Les gains de performance sont également présentés par rapport à la référence.

Tableau 5.5 Performances globales pour la classification multi-mentions

Type de modèle	**Description**	**Performance (%)**	**Gain**
A	Lexical	67.68	(Référence)
B	Lex. + durée des pauses	67.79	+0.11
C	Lex. + réinit. de F0	67.70	+0.02
D	Lex + cont. de F0	68.12	+0.44
E	Toutes	68.55	+0.87

Il est également intéressant d'observer exclusivement les performances pour le début des textes libres. Pour ce, le Tableau 5.6 présente la précision, le rappel et la F-mesure concernant la mention « B-text ». Il présente également le ratio des cellules b et c discuté précédemment. Ce tableau est suivi du Tableau 5.7 concernant la mention « B-title ».

Tableau 5.6 Performances pour « B-text » pour la classification multi-mentions

Type de modèle	**Performance (%)**			**Ratio b/c McNemar**
	Précision	**Rappel**	**F-mesure**	
A	54.74	62.50	58.36	(Référence)
B	53.11	63.96	58.03	0.82
C	52.56	64.17	57.79	0.75
D	53.21	63.96	58.09	0.82
E	55.77	65.42	60.21	1.16

Tableau 5.7 Performances pour « B-title » pour la classification multi-mentions

Type de modèle	Performance (%)			Ratio b/c McNemar
	Précision	Rappel	F-mesure	
A	80.67	55.49	65.75	(Référence)
B	77.36	47.40	58.78	0.46
C	78.57	50.87	61.76	0.67
D	78.26	52.02	62.50	0.65
E	83.65	50.29	62.82	0.89

5.4.2.2 Discussion

Les résultats du Tableau 5.5 ont suggéré que des gains de performances pouvaient être observés lorsque des caractéristiques prosodiques étaient ajoutées aux caractéristiques lexicales. Ces gains étaient prononcés avec le type de modèle D (+0.44% pour les continuités de F0) et davantage prononcés avec le type E (+0.87%). Selon le test de McNemar, seuls les types D et E ont obtenu des gains significatifs. Toutefois, une analyse des courbes des performances de la Figure-A V-7 a suggéré que les gains du type B pouvaient être significatifs. Cela confirme, pour ces trois types de modèle seulement, l'hypothèse 8 formulée à la section 5.2.3.

Ce type de classification a semblé défavorable concernant l'identification du début des textes libres. Les résultats du Tableau 5.6 ont indiqué un gain, au niveau du ratio b/c du test de McNemar, seulement avec le type de modèle E (1.16) concernant la mention « B-text ». Pour ce qui est de la mention « B-title », les résultats du Tableau 5.7 n'ont indiqué que des pertes. Le test de McNemar a été appliqué concernant la mention « B-text ». Pour ce test, toutes les mentions ne correspondant pas à « B-text » ont été converties sous la mention « Others ». Selon ce test, aucun gain significatif n'a été observé. Toutefois, l'analyse des courbes des performances (F-mesure) de la Figure-A V-10 a suggéré que les gains du type E pouvaient être significatifs. Cela confirme, pour ce type de classification et ce type de modèle seulement, l'hypothèse 7 formulée à la section 5.2.3.

La vraisemblance de l'hypothèse 7 pour le type B contredit celle obtenue lors de la classification texte libre. Pour la classification texte libre, les mentions « B-text » et « B-title » ont été fusionnées sous la mention « B-FreeText ». Ces mentions étaient distinctes lors de la classification multi-mentions, mais celles-ci ont souvent été confondues. Les matrices de confusion ont été examinées pour tous les types de modèle. Ces matrices ont indiqué que les exemples étiquetés sous « B-title » ont souvent été prédits sous « B-text » lors de l'évaluation finale. Toutefois, l'inverse n'a jamais été observé. Les écarts entre les mesures de précision et de rappel présentées dans le Tableau 5.6 et le Tableau 5.7 sont compatibles avec cette observation, dont la cause est simple. Dans l'ensemble d'entraînement, la fréquence des exemples étiquetés sous « B-text » (11 934 exemples) était plus de 16 fois supérieure à ceux étiquetés sous « B-title » (732 exemples). Avec ce ratio considérable, l'apprentissage du système a été favorable pour les exemples étiquetés sous « B-text ». Ce débalancement à la phase d'entraînement a favorisé les prédictions sous « B-text » à la phase de test lorsque des débuts de texte libre étaient identifiés. La fusion des mentions « B-text » et « B-title » ainsi que des mentions « I-text » et « I-title » aurait été préférable pour ce type de classification.

Il aurait été possible d'obtenir de meilleurs gains concernant les deux types de classification. Étant donné la quantité importante de mesures manquantes dans les corpus de données, principalement dû aux problèmes de réduction de moitié et de doublement de F0, plusieurs mentions n'ont pas bénéficié pleinement de l'ajout de l'information prosodique. De plus, certaines mentions étaient beaucoup plus fréquentes au début des commandes vocales. Toutefois, aucune mesure prosodique n'était associée au premier mot d'une commande. Par conséquent, cette lacune a défavorisé l'identification de ces mentions. Enfin, la différence de nature des données a également été défavorable.

5.5 Conclusion

Cette expérimentation a permis de vérifier que l'ajout des caractéristiques prosodiques aux caractéristiques lexicales pouvait améliorer les performances d'un système NLU. Pour ce,

l'expérimentation s'est divisée en deux types de classification. Pour la classification texte libre, l'ajout des durées des pauses et surtout l'ajout des trois types de caractéristiques prosodiques, ont amélioré significativement les performances d'identification du début des textes libres. Cependant, pour la classification multi-mentions, les trois types devaient être combinés afin d'améliorer ces mêmes performances (mention « B-text » seulement). Par contre, à l'exception des réinitialisations de F0, l'ajout des caractéristiques prosodiques a amélioré les performances d'identification des mentions en général. Ces performances étaient meilleures lorsque les trois types étaient combinés. Cela démontre que les informations prosodiques peuvent également bénéficier à l'identification des autres mentions. Enfin, sans doute que les résultats auraient été meilleurs avec l'utilisation de données de nature identique.

CONCLUSION GÉNÉRALE

L'information prosodique, présente dans le signal acoustique original des commandes vocales, n'est généralement pas exploitée par les systèmes NLU courants. Toutefois, il est possible que cette information supplémentaire puisse bénéficier ces systèmes. Actuellement, l'identification automatique des textes libres est une tâche particulièrement difficile. La problématique de ce mémoire était donc de déterminer si l'information prosodique pouvait aider directement ces systèmes. Toutefois, l'intérêt principal concernait l'impact de la prosodie sur leurs performances globales. La littérature n'aborde pas directement l'identification des textes libres. Néanmoins, plusieurs auteurs s'intéressent à l'identification de différents types d'AD. Parmi tous les AD étudiés, la citation est la plus similaire au texte libre. Contrairement à la citation, l'auteur d'un texte libre et son narrateur correspondent à la même entité. Cette caractéristique distingue cette étude des recherches précédentes.

Pour aborder la problématique, l'étude consistait à vérifier s'il existait une corrélation entre la prosodie et la présence ou non des textes libres. Si une corrélation était observée, il était intéressant de vérifier si cette information prosodique pouvait améliorer les performances d'un système NLU. La résolution de la problématique s'est effectuée en plusieurs étapes. La revue de la littérature a apporté des connaissances sur l'identification des citations grâce à la prosodie. Elle a inspiré la classification des textes libres utilisée pour ce mémoire, soit les textes libres directs et indirects. Différentes caractéristiques prosodiques, susceptibles d'engendrer les meilleurs résultats, ont également été proposées. Comme méthodologie, des outils ont permis d'extraire automatiquement des mesures prosodiques à partir des signaux acoustiques. Une évaluation de la qualité de ces outils a confirmé la fiabilité de leurs mesures. Pour obtenir la durée des pauses, les mesures de Praat et du système ASR ont été combinées. Les autres caractéristiques prosodiques issues de F0 ont été générées à partir des mesures de Praat seulement. Une première expérimentation a confirmé qu'il existait une corrélation entre la durée des pauses et la présence ou non d'un début de texte libre. En général, il a été démontré qu'il existait une durée de pause plus longue au début des textes libres par rapport aux autres concepts. De plus, cette durée était plus longue pour les textes

libres directs par rapport aux textes libres indirects. Une deuxième expérimentation a confirmé qu'il existait des corrélations par rapport à d'autres caractéristiques prosodiques, à savoir la réinitialisation de F0 et sa continuité. En général, il a été démontré qu'il existait une réinitialisation et une continuité de F0 plus prononcées au début des textes libres par rapport aux autres concepts. Par contre, ces deux caractéristiques étaient moins prononcées au début des textes libres directs par rapport aux textes libres indirects.

Une dernière expérimentation a été effectuée sur un système NLU sous forme d'un classifieur de mentions. L'objectif était de vérifier si l'ajout de ces trois caractéristiques prosodiques améliorait l'identification du début des textes libres et des autres concepts en général. Deux types de classification ont été étudiés. Pour la classification texte libre, l'ajout des durées des pauses a amélioré l'identification du début des textes libres. Une amélioration plus faible a également été observée lorsque les trois types de caractéristiques étaient combinés. En effet, puisque que les deux caractéristiques issues de F0 n'amélioraient pas le classifieur, leurs ajouts ont plutôt défavorisé l'apprentissage. À l'exception des continuités de F0, l'ajout des caractéristiques a amélioré les performances générales, surtout lorsque les trois types étaient combinés. Pour la classification multi-mentions, l'ajout des durées des pauses a amélioré l'identification de la mention « B-text ». À l'exception des réinitialisations de F0, l'ajout des caractéristiques a amélioré les performances générales, surtout lorsque les trois types étaient combinés.

Certains facteurs ont contraint cette étude. Les corpus de données principaux utilisés pour ces trois expérimentations étaient biaisés, puisque les échantillons n'avaient pas été extraits de la population d'une façon entièrement aléatoire, ce qui a influencé les résultats. Avec suffisamment de données extraites de façon aléatoire, la dernière expérimentation aurait bénéficié d'une phase d'entraînement et de test sur des données de même nature. Cette similarité des données aurait certainement favorisé les résultats et renforcé la validité des conclusions. De plus, la proportion significative de mesures prosodiques manquantes dans les ensembles d'entraînement et de test a défavorisé cette expérimentation. Ce manque concernait surtout les caractéristiques issues de F0 ignorées dû aux problèmes de réduction

de moitié et de doublement de F0. Enfin, la corrélation entre la prosodie et la fin des textes libres n'a pas été étudiée, puisque la quantité de textes libres ne terminant pas la commande vocale était insuffisante.

Néanmoins, cette étude a démontré que les systèmes NLU peuvent bénéficier de l'information prosodique en plus de l'information lexicale. Ce savoir pourrait bénéficier l'industrie afin d'améliorer la qualité des applications informatiques vendues à leurs clients. Suite à cette étude, d'autres expérimentations pourraient étudier des corrélations entre la prosodie et d'autres concepts. Par exemple, il serait intéressant d'étudier les concepts les plus fréquents, car ils pourraient bénéficier d'une phase d'apprentissage plus favorable. D'autres caractéristiques prosodiques pourraient également être étudiées, comme le contour de F0, l'intensité et la durée des sons. Les caractéristiques pourraient être obtenues à partir d'autres traitements et algorithmes. Par exemple, développer des solutions pour réduire les problèmes de réduction de moitié et de doublement de F0 responsables de la diminution significative des caractéristiques issues de F0. Par conséquent, d'autres résultats seraient observés. Tous les corpus de données pourraient être de nature identique et être plus représentatifs de la population. En général, les entreprises accumulent des commandes vocales extraites aléatoirement et les annotent manuellement. Ces données pourraient être en quantité suffisante dans les années à venir. Finalement, en plus des informations prosodiques et lexicales, toutes autres informations paralinguistiques pourraient bénéfique à ces systèmes.

ANNEXE I

COEFFICIENTS DE STUDENT

Le Tableau-A I-1 donne la probabilité (α) pour que le coefficient de Student (t) égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (ddl). ddl correspond à la taille de l'échantillon moins le nombre de paramètres.

Tableau-A I-1 Coefficients de Student
Tiré de Youcef Elmeddah (2013) [15]

α / ddl	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	**12,706**	31,821	**63,656**	636,578
2	0,142	0,816	1,386	1,886	2,920	**4,303**	6,965	**9,925**	31,600
3	0,137	0,765	1,250	1,638	2,353	**3,182**	4,541	**5,841**	12,924
4	0,134	0,741	1,190	1,533	2,132	**2,776**	3,747	**4,604**	8,610
5	0,132	0,727	1,156	1,476	2,015	**2,571**	3,365	**4,032**	6,869
6	0,131	0,718	1,134	1,440	1,943	**2,447**	3,143	**3,707**	5,959
7	0,130	0,711	1,119	1,415	1,895	**2,365**	2,998	**3,499**	5,408
8	0,130	0,706	1,108	1,397	1,860	**2,306**	2,896	**3,355**	5,041
9	0,129	0,703	1,100	1,383	1,833	**2,262**	2,821	**3,250**	4,781
10	0,129	0,700	1,093	1,372	1,812	**2,228**	2,764	**3,169**	4,587
11	0,129	0,697	1,088	1,363	1,796	**2,201**	2,718	**3,106**	4,437
12	0,128	0,695	1,083	1,356	1,782	**2,179**	2,681	**3,055**	4,318
13	0,128	0,694	1,079	1,350	1,771	**2,160**	2,650	**3,012**	4,221
14	0,128	0,692	1,076	1,345	1,761	**2,145**	2,624	**2,977**	4,140
15	0,128	0,691	1,074	1,341	1,753	**2,131**	2,602	**2,947**	4,073
16	0,128	0,690	1,071	1,337	1,746	**2,120**	2,583	**2,921**	4,015
17	0,128	0,689	1,069	1,333	1,740	**2,110**	2,567	**2,898**	3,965
18	0,127	0,688	1,067	1,330	1,734	**2,101**	2,552	**2,878**	3,922
19	0,127	0,688	1,066	1,328	1,729	**2,093**	2,539	**2,861**	3,883
20	0,127	0,687	1,064	1,325	1,725	**2,086**	2,528	**2,845**	3,850
21	0,127	0,686	1,063	1,323	1,721	**2,080**	2,518	**2,831**	3,819
22	0,127	0,686	1,061	1,321	1,717	**2,074**	2,508	**2,819**	3,792
23	0,127	0,685	1,060	1,319	1,714	**2,069**	2,500	**2,807**	3,768
24	0,127	0,685	1,059	1,318	1,711	**2,064**	2,492	**2,797**	3,745
25	0,127	0,684	1,058	1,316	1,708	**2,060**	2,485	**2,787**	3,725
26	0,127	0,684	1,058	1,315	1,706	**2,056**	2,479	**2,779**	3,707
27	0,127	0,684	1,057	1,314	1,703	**2,052**	2,473	**2,771**	3,689
28	0,127	0,683	1,056	1,313	1,701	**2,048**	2,467	**2,763**	3,674
29	0,127	0,683	1,055	1,311	1,699	**2,045**	2,462	**2,756**	3,660
30	0,127	0,683	1,055	1,310	1,697	**2,042**	2,457	**2,750**	3,646
40	0,126	0,681	1,050	1,303	1,684	**2,021**	2,423	**2,704**	3,551
80	0,126	0,678	1,043	1,292	1,664	**1,990**	2,374	**2,639**	3,416
120	0,126	0,677	1,041	1,289	1,658	**1,980**	2,358	**2,617**	3,373
∞	0,126	0,675	1,037	1,282	1,645	**1,960**	2,327	**2,577**	3,293

ANNEXE II

TEST DE KOLMOGOROV–SMIRNOV À DEUX ÉCHANTILLONS

Le test de Kolmogorov–Smirnov (test K-S) [79] à deux échantillons est un test non paramétrique vérifiant l'égalité des distributions de probabilité continue à une dimension. Ce test est utilisé pour vérifier si deux échantillons appartiennent à une même distribution. Cette vérification est effectuée en quantifiant une distance entre leur fonction de répartition empirique (statistique K-S). La

Figure-A II-1 présente deux exemples de fonction de répartition empirique et une flèche correspondant à la statistique K-S.

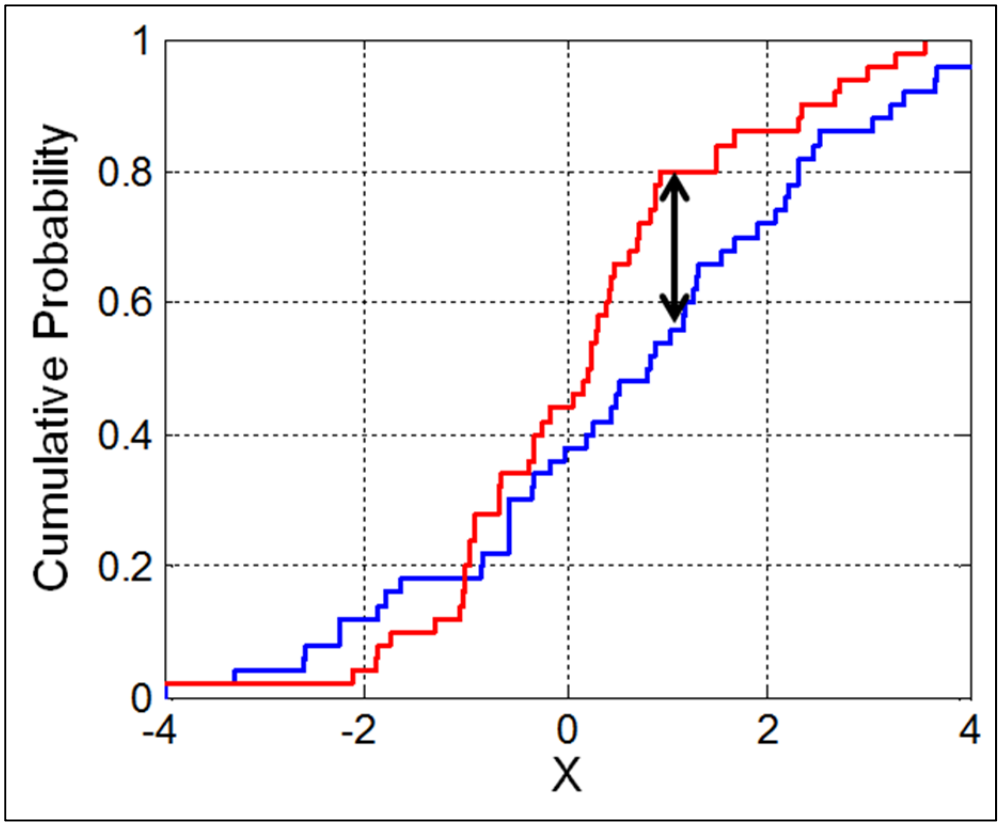

Figure-A II-1 Fonctions de répartition empirique et la statistique K-S.
Tirée de Wikipédia (2015) [79]

Ce test est l'une des méthodes non paramétriques les plus générales et utiles pour la comparaison de deux échantillons. Il est sensible à des différences à la fois pour l'emplacement et la forme des fonctions de répartition empiriques des deux échantillons. Pour ce test, une hypothèse nulle est définie spécifiant que les échantillons sont prélevés à partir de la même distribution. Pour vérifier si une même distribution est concernée, la fonction de répartition empirique est calculée pour chaque échantillon. Ces échantillons concernent des observations X_i de variables aléatoires distribuées de façon identique et indépendante :

$$F_m(x) = \frac{1}{m}\sum_{i=1}^{m} I_{X_i \leq x} \qquad F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x} \tag{A II-1}$$

où $F_m(x)$ et $F_n(x)$ sont les fonctions de répartition empirique pour le premier et deuxième échantillon de grosseur m et n respectivement. $I_{X_i \leq x}$ est la fonction caractéristique et vaut 1 si $X_i \leq x$ et vaut 0 sinon. La statistique K-S est ensuite calculée comme suit :

$$D_{m,n} = sup_x |F_m(x) - F_n(x)| \tag{A II-2}$$

où sup est la fonction supremum (borne supérieure). L'hypothèse nulle est rejetée selon un niveau de confiance arbitraire α si :

$$D_{m,n} > c(\alpha)\sqrt{\frac{m+n}{m\,n}} \tag{A II-3}$$

où α est généralement assigné à 0.05 ou 0.01. Les valeurs attribuées à $c(\alpha)$ sont données dans le Tableau-A II-1 pour chaque niveau de α :

Tableau-A II-1 Correspondances entre le niveau de confiance α et c(α)
Tiré de Wikipédia (2015) [79]

$\boldsymbol{\alpha}$	0.10	0.05	0.025	0.01	0.005	0.001
$\boldsymbol{c(\alpha)}$	1.22	1.36	1.48	1.63	1.73	1.95

Une autre façon de déterminer si l'hypothèse nulle est rejetée consiste à calculer la valeur-p[10]. La valeur-p est la probabilité d'obtenir la même statistique K-S (ou une valeur encore plus extrême) si l'hypothèse nulle était vraie. La valeur-p est comparée directement au niveau de confiance α. Si la valeur-p est inférieure à ce seuil, l'hypothèse nulle est rejetée et le résultat du test est déclaré « statistiquement significatif ». Si la valeur-p est supérieure à ce seuil, l'hypothèse nulle est acceptée.

Le calcul de la valeur-p nécessite une hypothèse nulle, une statistique de test (statistique K-S) et des données. Le calcul de la statistique de test sur les données est relativement simple. Cependant, le calcul de la distribution d'échantillonnage sous l'hypothèse nulle, puis le calcul de sa fonction de répartition empirique est souvent un calcul difficile. Aujourd'hui, ce calcul est effectué en utilisant des logiciels statistiques. Des méthodes numériques sont souvent utilisées plutôt que des formules exactes.

Pour ce mémoire, la valeur-p a été choisie pour rejeter ou non les hypothèses nulles. La valeur-p, de même que la statistique K-S, ont été calculées dans le langage de programmation objet Python à l'aide de la fonction *ks_2samp* de la librairie *scipy.stats*.

[10] http://en.wikipedia.org/wiki/P-value

ANNEXE III

MODÈLE DE RÉGRESSION LINÉAIRE SIMPLE

Afin de simplifier la manipulation des mesures recueillies sur F0, la stratégie utilisée a été de représenter un sous-ensemble de valeurs (de points) à l'aide d'une droite. Les propriétés de ces droites pouvaient ensuite être exploitées comme caractéristiques issues de F0. Pour générer ces droites, le modèle de régression linéaire simple [81] a été utilisé, car ce modèle était simple à implémenter et était satisfaisant pour les besoins de cette étude.

La régression linéaire simple est l'estimateur des moindres carrés d'un modèle de régression linéaire avec une seule variable explicative. Elle fait correspondre une ligne droite à travers un ensemble de n points. Cette correspondance est telle que la somme des carrés des résidus du modèle (c.-à-d. les distances verticales entre les points de l'ensemble des données et la droite ajustée) est la plus petite possible. Cette régression est l'une des plus simples en statistique. La pente de la droite ajustée est égale à la corrélation entre y et x corrigée par le rapport entre les écarts-types de ces variables. L'ordonnée à l'origine de la droite de régression est telle qu'elle passe par le centre de masse $(\bar{x}, \bar{y})$ des points. En supposant un ensemble de n points $\{(x_i, y_i),\ i = 1, \dots, n\}$, la fonction décrivant x et y est donnée par :

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{(A III-1)}$$

L'objectif est de trouver l'équation pour la ligne droite :

$$f = \alpha + \beta x \qquad \text{(A III-2)}$$

permettant le meilleur ajustement pour les points, où α et β correspondent à l'ordonnée à l'origine et à la pente respectivement. Cela consiste en une ligne qui minimise la somme des carrés des résidus du modèle de régression linéaire, c.-à-d. α et β résolvent le problème de minimisation suivant :

$$\text{Trouver } \min_{\alpha,\beta} Q(\alpha,\beta) \text{ pour } Q(\alpha,\beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \qquad \text{(A III-3)}$$

Les valeurs α et β minimisant la fonction objective Q sont données par :

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{Cov[x,y]}{Var[x]} = r_{xy}\frac{s_y}{s_x} \qquad \text{(A III-4)}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{(A III-5)}$$

où r_{xy} correspond au coefficient de corrélation de l'échantillon entre x et y. s_x et s_y correspondent à l'écart type de x et y respectivement. Une barre horizontale sur une quantité signifie sa moyenne au niveau de l'échantillon. La valeur r_{xy} est obtenue à l'aide de cette équation :

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right)\left(\overline{y^2} - \bar{y}^2\right)}} \qquad \text{(A III-6)}$$

Pour ce mémoire, α et β ont été calculées dans le langage de programmation objet Python à l'aide de la fonction *polyfit* de la librairie *numpy*.

ANNEXE IV

TEST DE MCNEMAR

Le test de McNemar [80] est un test non paramétrique utilisé sur des données nominales appariées. Il est appliqué sur des paires appariées de sujets à partir d'un tableau de contingence de dimension 2 x 2 avec un trait dichotomique. L'objectif est de déterminer si les fréquences marginales des lignes et de colonnes sont égales (s'il y a une « homogénéité marginale »). Le Tableau-A IV-1 est utilisé pour ce test, en compilant les résultats de deux tests à partir d'un échantillon de n sujets.

Tableau-A IV-1 Tableau utilisé pour le test de McNemar

	Posttest positif	**Posttest négatif**	**Total**
Prétest positif	a	b	a + b
Prétest négatif	c	d	c + d
Total	a + c	b + d	n

L'hypothèse nulle d'homogénéité marginale spécifie que les deux probabilités marginales sont les mêmes pour chaque résultat, c.-à-d. :

$$p_a + p_b = p_a + p_c \quad \text{et} \quad p_c + p_d = p_b + p_d \tag{A IV-1}$$

où p_a, etc. désigne la probabilité théorique d'occurrences dans les cellules avec l'étiquette correspondante. Étant donné que p_a et p_d s'annulent sur les deux côtés des équations, $p_b = p_c$ et constitue la base de ce test. La statistique est donnée par :

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{A IV-2}$$

Sous l'hypothèse nulle, avec un nombre suffisamment grand de discordance (cellules b et c), χ^2 a une distribution de chi-carré[11] avec un degré de liberté. La Figure-A IV-1 présente cette distribution.

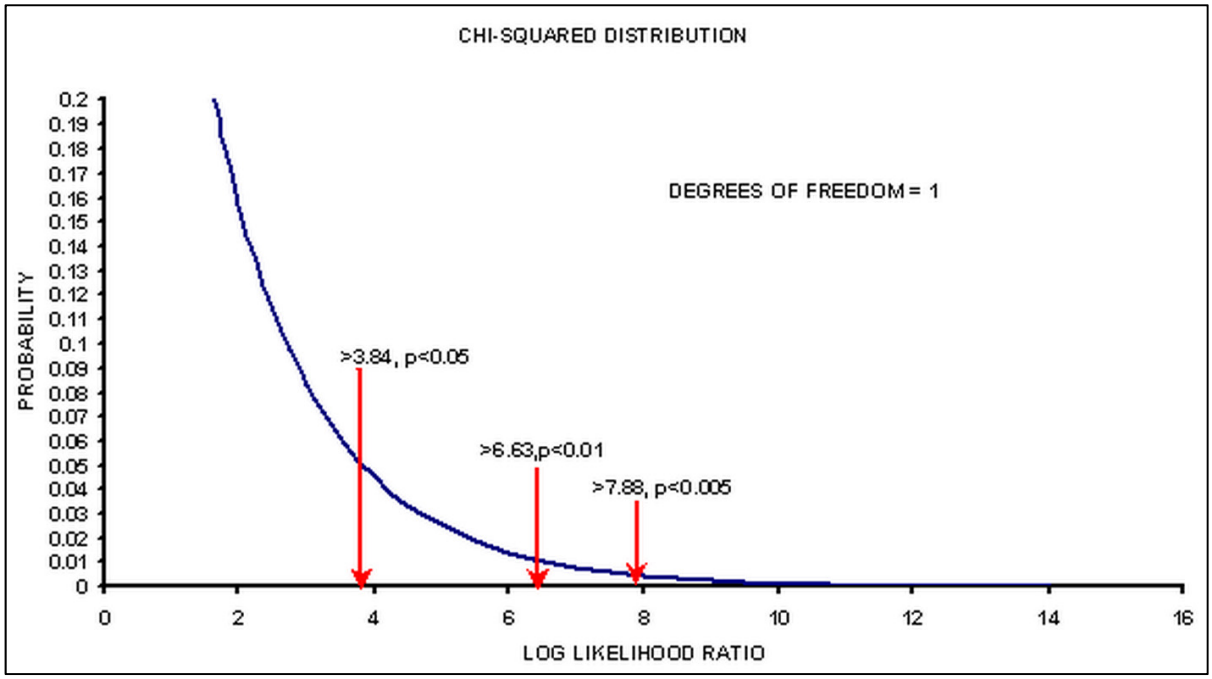

Figure-A IV-1 Distribution de chi-carré avec un degré de liberté [2].

La signification statistique est déterminée en évaluant χ^2 par rapport à une table de probabilités cumulatives de cette distribution ou une fonction informatique équivalente. Pour ce mémoire, la fonction de point de pourcentage (PPF) a été utilisée. Celle-ci correspond à l'inverse de sa fonction de distribution cumulative (CDF). La Figure-A IV-2 présente la PPF de la distribution de chi-carré avec un degré de liberté.

[11] https://en.wikipedia.org/wiki/Chi-squared_distribution

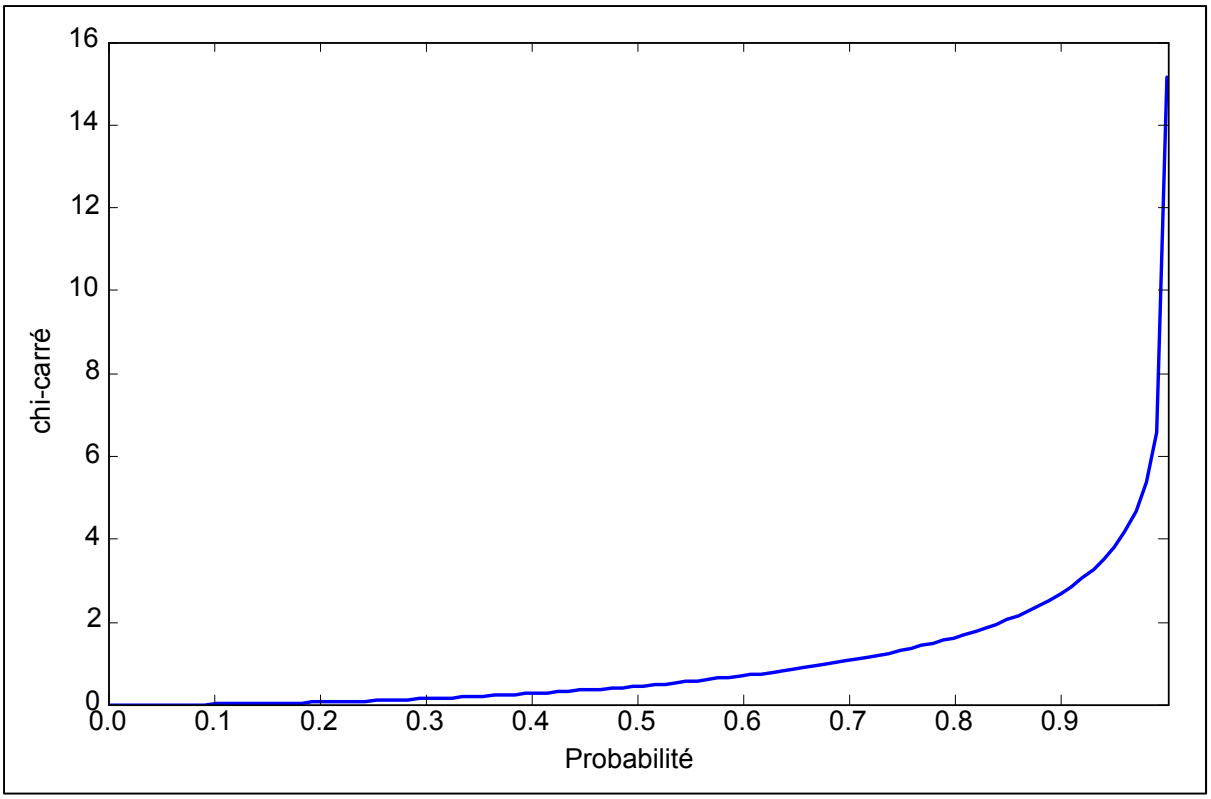

Figure-A IV-2 PPF de la distribution de chi-carré avec un degré de liberté.

Un niveau de confiance $\alpha = 0.05$ est habituellement choisi pour ce test. La valeur-p est obtenue en considérant la PPF de $1 - \alpha$. Lorsque celle-ci est inférieure à χ^2, l'hypothèse nulle est rejetée et le résultat est déclaré statistiquement significatif. Cela signifie que les fréquences marginales (ou proportions) sont significativement différentes l'une de l'autre. Pour ce mémoire, la valeur-p a été calculée dans le langage de programmation objet Python à l'aide de la fonction *chi2.ppf* de la librairie *scipy.stats*.

ANNEXE V

COURBES DES PERFORMANCES

Ce qui suit présente les courbes des performances, concernant les deux types de classification présentés à la section 5.4, générées lors de l'expérimentation sur un système NLU présenté au CHAPITRE 5. La Figure-A V-1 et la Figure-A V-2 présentent ces courbes pour la phase de développement. Sur ces courbes, le point étiqueté correspond à la configuration optimale du type de modèle concerné.

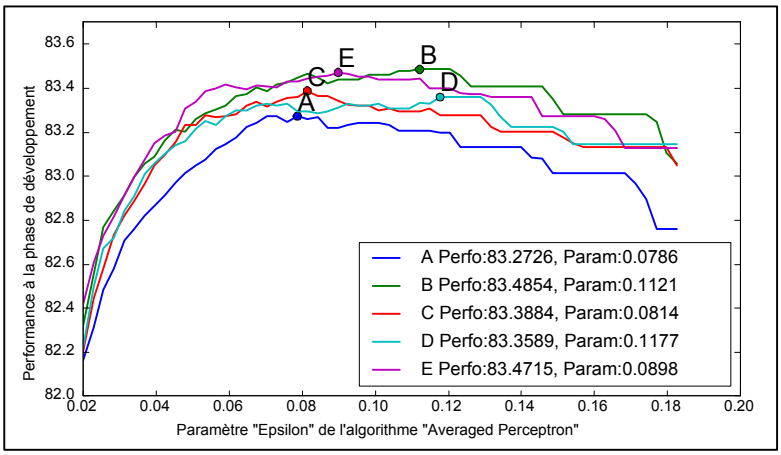

Figure-A V-1 Performances à la phase
de développement pour la classification texte libre.

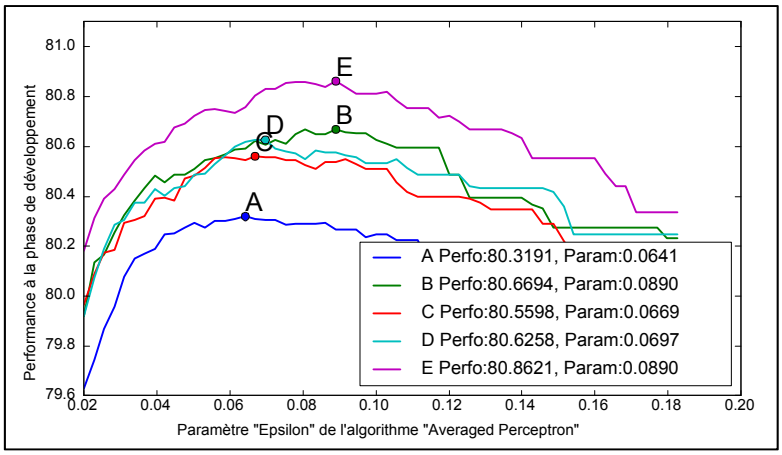

Figure-A V-2 Performances à la phase
de développement pour la classification multi-mentions.

Les figures suivantes présentent ces courbes pour la phase de test. Les performances officielles correspondent aux points étiquetés, représentant les configurations optimales discutées précédemment. Les hypothèses de cette expérimentation ont été vérifiées à partir de ces performances.

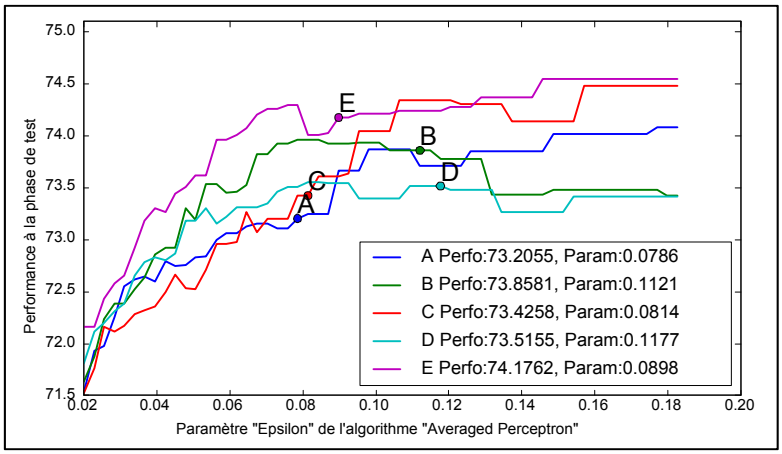

Figure-A V-3 Performances à la phase
de test pour la classification texte libre.

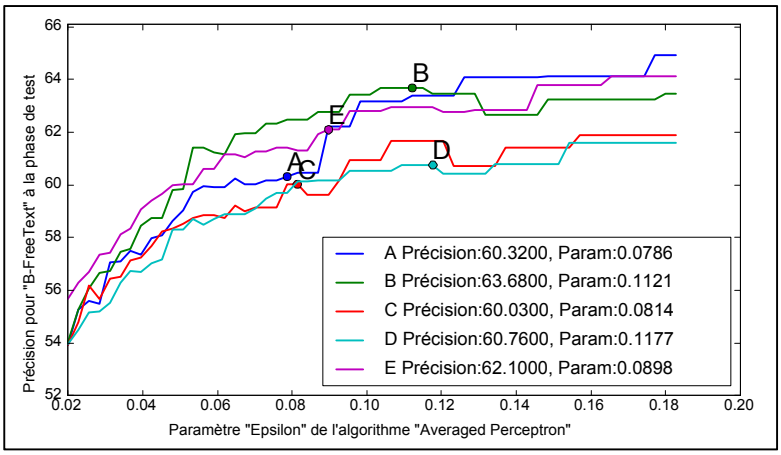

Figure-A V-4 Précisions pour "B-FreeText" à la phase de test pour la classification texte libre.

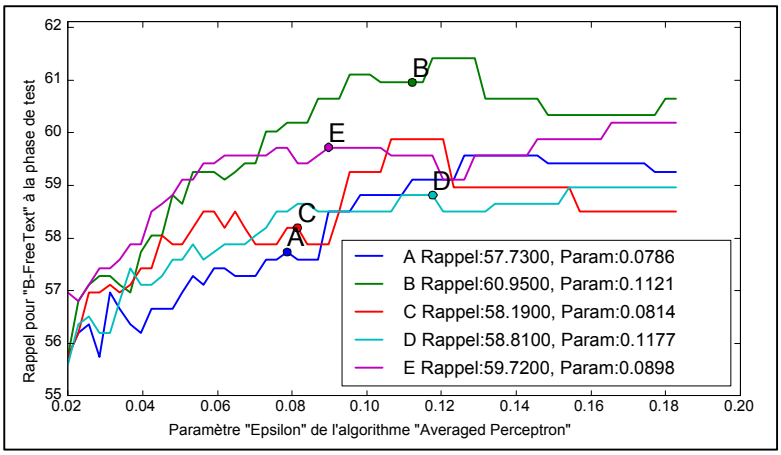

Figure-A V-5 Rappels pour "B-FreeText" à la phase de test pour la classification texte libre.

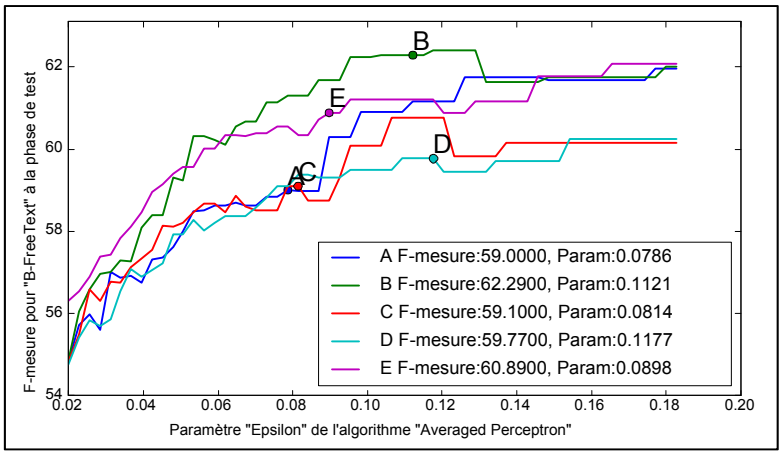

Figure-A V-6 F-mesures pour "B-FreeText" à la phase de test pour la classification texte libre.

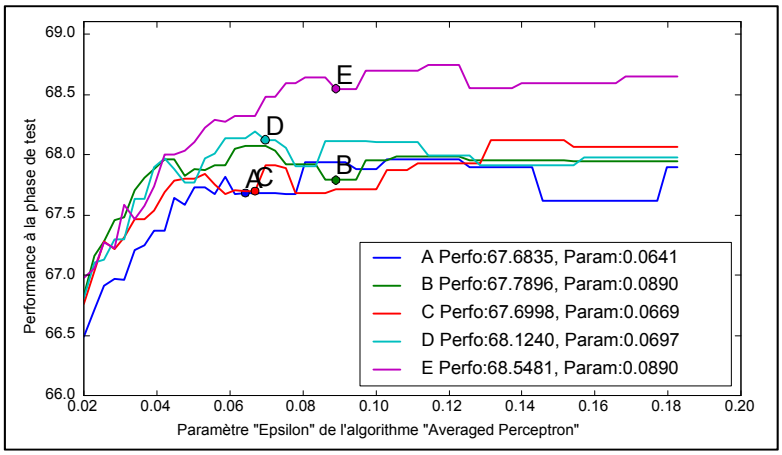

Figure-A V-7 Performances à la phase de test pour la classification multi-mentions.

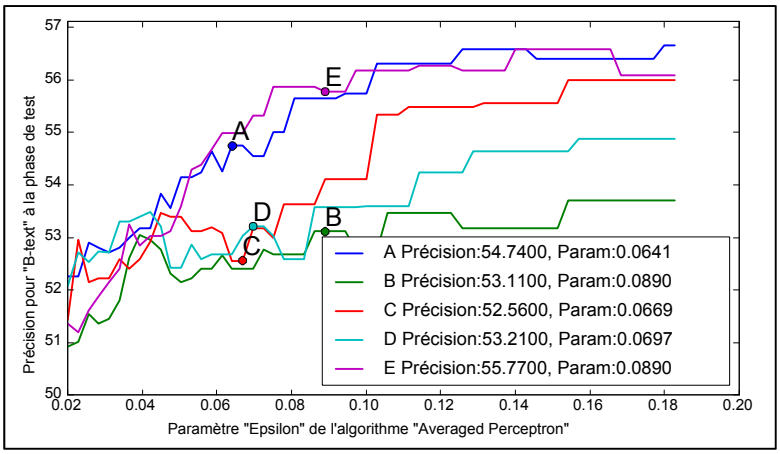

Figure-A V-8 Précisions pour "B-text" à la phase de test pour la classification multi-mentions.

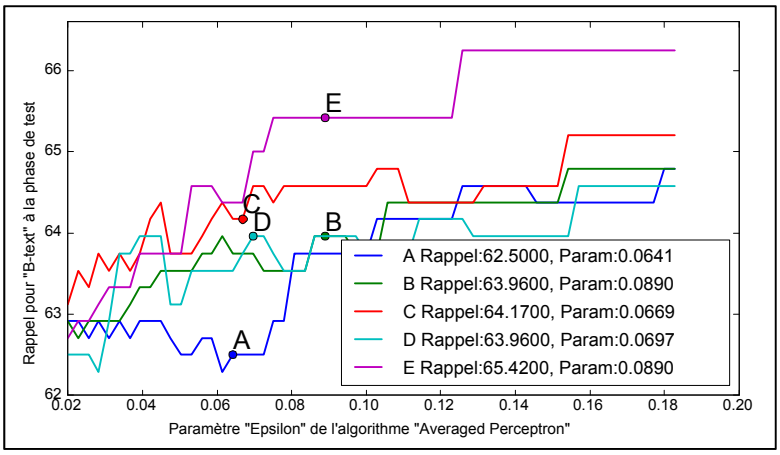

Figure-A V-9 Rappels pour "B-text" à la phase de test pour la classification multi-mentions.

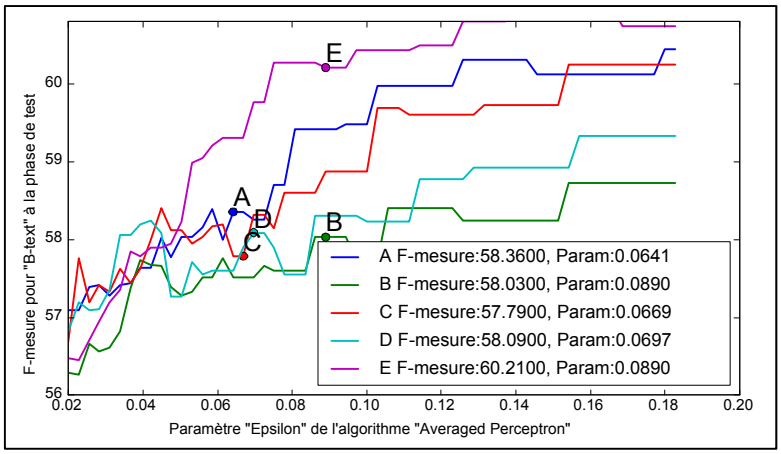

Figure-A V-10 F-mesures pour "B-text" à la phase de test pour la classification multi-mentions.

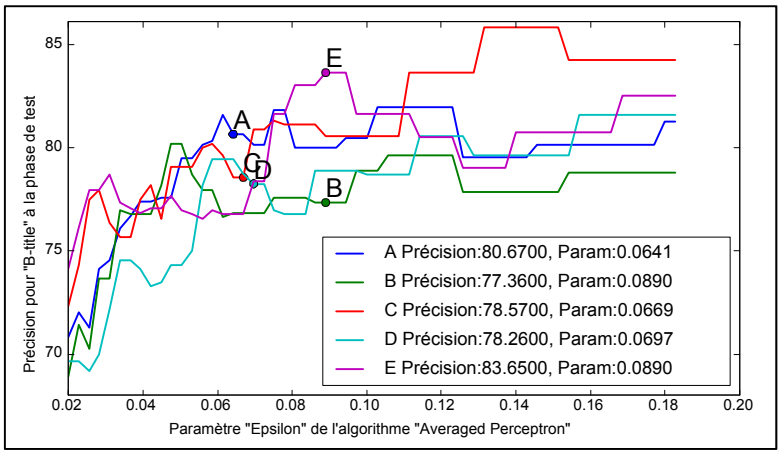

Figure-A V-11 Précisions pour "B-title" à la phase de test pour la classification multi-mentions.

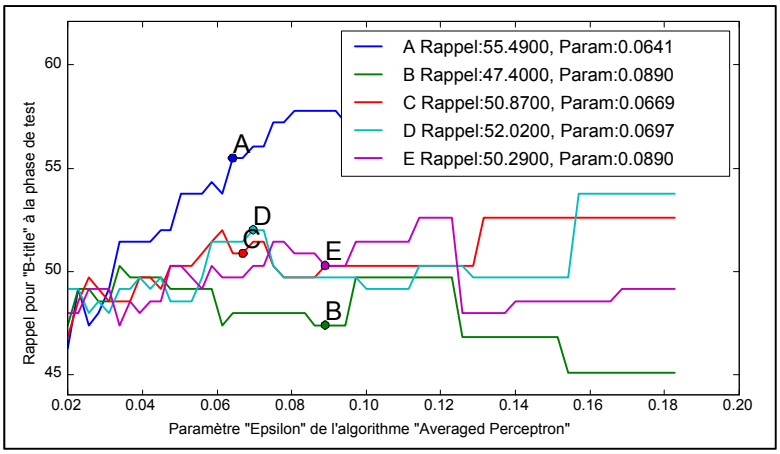

Figure-A V-12 Rappels pour "B-title" à la phase de test pour la classification multi-mentions.

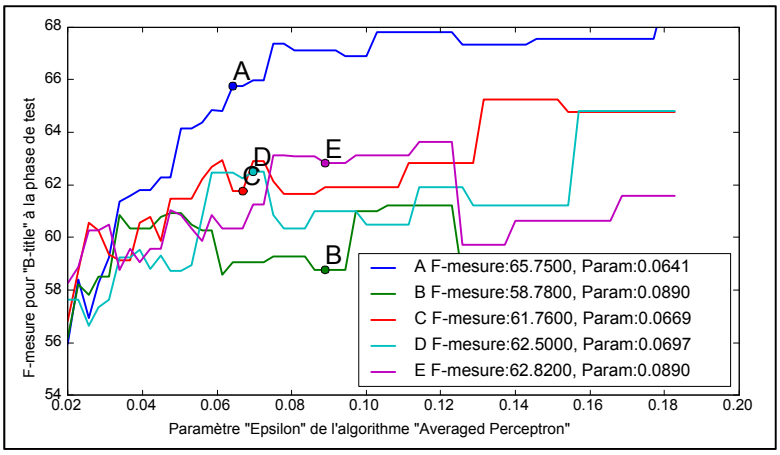

Figure-A V-13 F-mesures pour "B-title" à la phase de test pour la classification multi-mentions.

ANNEXE VI

STATISTIQUES ET DÉCISIONS DU TEST DE MCNEMAR

Ce qui suit présente les statistiques utilisées et les décisions obtenues du test de McNemar nécessaire lors de l'expérimentation sur un système NLU présentée au CHAPITRE 5. Ce test présenté à l'ANNEXE IV a permis de vérifier si les gains de performance obtenus lors des deux types de classification étaient statistiquement significatifs.

Tableau-A VI-1 Test de McNemar pour la classification texte libre

Témoins	**Cas**	**Type de modèle (comparé par rapport à A)**			
		B	**C**	**D**	**E**
+	+	8 680	8 672	8 687	8 741
-	+	375	330	326	353
+	-	295	303	288	234
-	-	2 910	2 955	2 959	2 932
Décision hypothèse nulle		Rejetée	Acceptée	Acceptée	Rejetée

Tableau-A VI-2 Test de McNemar pour « B-FreeText » pour la classification texte libre

Témoins	**Cas**	**Type de modèle (comparé par rapport à A)**			
		B	**C**	**D**	**E**
+	+	11 677	11 663	11 662	11 671
-	+	101	71	81	88
+	-	59	73	74	65
-	-	423	453	443	436
Décision hypothèse nulle		Rejetée	Acceptée	Acceptée	Acceptée

Tableau-A VI-3 Test de McNemar pour la classification multi-mentions

Témoins	Cas	Type de modèle (comparé par rapport à A)			
		B	**C**	**D**	**E**
+	+	8 045	8 051	8 096	8 048
-	+	266	249	256	356
+	-	253	247	202	250
-	-	3 696	3 713	3 706	3 606
Décision hypothèse nulle		Acceptée	Acceptée	Rejetée	Rejetée

Tableau-A VI-4 Test de McNemar pour « B-text » pour la classification multi-mentions

Témoins	Cas	Type de modèle (comparé par rapport à A)			
		B	**C**	**D**	**E**
+	+	11 742	11 745	11 750	11 750
-	+	74	65	67	95
+	-	90	87	82	82
-	-	354	363	361	333
Décision hypothèse nulle		Acceptée	Acceptée	Acceptée	Acceptée

Tableau-A VI-5 Test de McNemar pour « B-title » pour la classification multi-mentions

Témoins	Cas	Type de modèle (comparé par rapport à A)			
		B	**C**	**D**	**E**
+	+	12 132	12 133	12 137	12 133
-	+	13	18	15	24
+	-	28	27	23	27
-	-	87	82	85	76
Décision hypothèse nulle		Rejetée	Acceptée	Acceptée	Acceptée

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

1. *Pitch Tracking + Prosody*. 2012 [cited 2015 April 6]; Available from: http://www.basesproduced.com/441/notes/2-Pitch-Prosody.pdf.

2. ACCP. *Pharmacometrics*. 2011 [cited 2015 August 2]; Available from: http://accp1.org/pharmacometrics/theory_gmp_model5.htm.

3. Allan, J., et al., *Topic Detection and Tracking Pilot Study: Final Report*, in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 1998: Lansdowne, VA, USA. p. 194-218.

4. Baron, D., E. Shriberg, and A. Stolcke, *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues*, in *Proceedings of the International Conference on Spoken Language Processing*. 2002: Denver. p. 4.

5. Batliner, A., et al., *The Prosodic Marking of Phrase Boundaries: Expectations and Results*, in *Speech Recognition and Coding: New Advances and Trends*, S.B. Heidelberg, Editor. 1995, Springer: Berlin, Germany. p. 325-28.

6. Blesser, B.A., *Inadequacy of a Spectral Description in Relationship to Speech Perception.* The Journal of the Acoustical Society of America, 1970. **47**(1A): p. 19.

7. Blum, A. and T. Mitchell, *Combining labeled and unlabeled data with co-training*, in *Proceedings of the 1998 11th Annual Conference on Computational Learning Theory*. 1998, ACM: Madison, WI, USA. p. 92-100.

8. Boersma, P. and D. Weenink, *Praat: doing phonetics by computer [Computer program]*. 2013.

9. Breiman, L., et al., *Classification and Regression Trees*. 1984, Pacific Grove, CA: Wadsworth and Brooks.

10. Brown, G., K.L. Currie, and J. Kenworthy, *Questions of Intonation*. Croom Helm linguistics series. 1980: Croom Helm.

11. Bruce, G., *Textual aspects of prosody in Swedish.* Phonetica, 1982. **39**(4-5): p. 274-87.

12. Caelen-Haumont, G., *Prosodie et sens: Une approche expérimentale*. Marges linguistiques. Vol. 1. 2009: Editions L'Harmattan. 216.

13. Collier, R., J. Roelof de Pijper, and A. Sanderman, *Perceived Prosodic Boundaries and their Phonetic Correlates*, in *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro*. 1993: New Jersey. p. 341-45.

14. Cooper, W. and J. Sorensen, *Fundamental frequency contours at syntactic boundaries.* The Journal of the Acoustical Society of America, 1977. **62**(3): p. 683-692.

15. Elmeddah, Y. *Distribution t de Student*. 2013 [cited 2015 July 31]; Available from: http://fr.slideshare.net/Youcef63000/tables-statistiques.

16. Fernandez, R. and R.W. Picard, *Dialog Act Classification from Prosodic Features Using Support Vector Machines.* Proceedings of speech prosody 2002, 2002.

17. Flamary, R., et al., *Variational sequence labeling*, in *Machine Learning for Signal Processing XIX - 2009 IEEE Signal Processing Society Workshop, MLSP 2009*. 2009, IEEE Computer Society, 445 Hoes Lane - P.O.Box 1331, Piscataway, NJ 08855-1331, United States: Grenoble, France.

18. Hakkani-tür, D., et al., *Combining Words and Prosody for Information Extraction from Speech.* Proc. Eurospeech, 1999: p. 1991-94.

19. Hashimoto, S. and A. Saito, *Prosodic rules for speech synthesis*, in *7th International congress on acoustics*. 1971, Akademiai Kiado: Budapest, Hungary. p. 129-32.

20. Hirschberg, J. and C. Nakatani, *A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues*, in *34th Annual Meeting of the Association for Computational Linguistics*. 1996, Association for Computational Linguistics: Santa Cruz. p. 286-93.

21. Hirschberg, J. and C. Nakatani, *Acoustic indicators of topic segmentation*, in *Proceedings of the International Conference on Speech and Language Processing*. 1998: Sydney, Australia.

22. Hirschberg, J. and C. Nakatani, *Acoustic indicators of Topic segmentation*, in *Proc. Intl. Conf. on Spoken Language Processing*. 1998: Philadelphia. p. 1255–58.

23. Huang, W.Y. and R.P. Lippman, *Comparisons Between Neural Net and Conventional Classifiers*. 1987, Lincoln Laboratory, MIT.

24. Jansen, W., M.L. Gregory, and J.M. Brenier, *Prosodic correlates of directly reported speech: Evidence from conversational speech* Prosody in Speech Recognition and Understanding. Molly Pitcher Inn, 2001.

25. Je Hun, J. and L. Yang. *Automatic prosodic events detection using syllable-based acoustic and syntactic features*. in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. 2009.

26. Jeon, J.H. and Y. Liu, *Syllable-level prominence detection with acoustic evidence*, in *11th Annual Conference of the International Speech Communication Association:*

Spoken Language Processing for All, INTERSPEECH 2010. 2010, International Speech Communication Association: Makuhari, Chiba, Japan. p. 1772-75.

27. Jeon, J.H. and Y. Liu, *Automatic prosodic event detection using a novel labeling and selection method in co-training.* Speech Communication, 2012. **54**(3): p. 445-58.

28. Jurafsky, D., et al., *Lexical, Prosodic, and Syntactic Cues for Dialog Acts*, in *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*. 1998. p. 114-20.

29. Kaisse, E.M., *Connected Speech: The Interaction of Syntax and Phonology*. 1985, San Diego: Academic Press.

30. Kasimir, E., *Prosodic correlates of subclausal quotation marks.* ZAS Papers in Linguistics, 2008. **49**: p. 67-78.

31. Kaylani, T. and S. DasGupta. *A new method for initializing radial basis function classifiers*. in *Systems, Man, and Cybernetics, 1994. Humans, Information and Technology., 1994 IEEE International Conference on*. 1994.

32. Ken, C., M. Hasegawa-Johnson, and A. Cohen. *An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model*. in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*. 2004.

33. Khawalda, M. and E. Al-Saidat, *Structural Ambiguity Interpretation: A Case Study of Arab Learners of English.* Global Journal of Human Social Science, 2012. **7**(6).

34. Klewitz, G. and E. Couper-Kuhlen, *Quote - Unquote? The Role of Prosody in the Contextualization of Reported Speech Sequences.* Pragmatics, 1999. **9**(4): p. 459-85.

35. Kompe, R., et al. *Automatic classification of prosodically marked phrase boundaries in German*. in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. 1994. Adelaide, SA: IEEE.

36. Kompe, R., et al., *Prosodic Scoring Of Word Hypotheses Graphs.* 1995.

37. Ladd, D.R., *Intonational Phonology*. 1996, Cambridge: Cambridge University Press.

38. Laskowski, K., *A Frame-Synchronous Prosodic Decoder for Text-Independent Dialog Act Recognition*, in *Proc. 5th International Conference on Speech Prosody*. 2010, International Speech Communication Association: Chicago IL, USA.

39. Lea, W., M. Medress, and T. Skinner, *A prosodically guided speech understanding strategy.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1975. **23**(1): p. 30-38.

40. Lea, W., M. Medress, and T. Sklnner, *Prosodic Aids to Speech Recognition.* 1972: p. 68.

41. Lea, W.A., M.F. Medress, and T.E. Skinner, *A prosodically guided speech understanding strategy*, in *IEEE Symposium on Speech Recognition*. 1975, IEEE: Pittsburgh, PA, USA. p. 30-8.

42. Leech, G.N. and M. Short, *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2007: Pearson Longman. 404.

43. Lehiste, I., *Suprasegmentals*. réimprimée ed. 1970, Cambridge, MA: M.I.T. Press. 194.

44. Lehiste, I., *Perception of sentence and paragraph boundaries.* Frontiers of speech communication research, 1979: p. 191–201.

45. Levitt, H., *Acoustic and perceptual characteristics of the speech of deaf children*, in *Program of the 83rd meeting of the Acoustical Society of America*. 1972, Acoust. Soc. America: Buffalo, New York, NY, USA. p. 65-6.

46. Lickley, R.J., et al., *Processing disfluent speech: How and when are disfluencies found?*, in *Proceedings of EUROSPEECH-1991.* 1991, Istituto Internazionale delle Comunicazioni. p. 1499-1502.

47. Lieberman, P., *Intonation, perception, and language*. 1967, Cambridge, MA: M.I.T. Press.

48. Lindblom, B. and S.-G. Svensson, *Interaction between segmental and nonsegmental factors in speech recognition.* 1973. **AU-21**(6): p. 536-45.

49. Majumder, D.D., A.K. Dutta, and N.R. Ganguli, *Some studies on acoustic features of human speech in relation to Hindi speech sounds.* 1973. **47**(10): p. 598-613.

50. Mast, M., et al. *Dialog act classification with the help of prosody*. in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. 1996.

51. Niemann, H., et al., *Using Prosodic Cues In Spoken Dialog Systems*, in *International Workshop "Speech and Computer"*. 1998, Proc. Intl. Conf. on Spoken Language Processing: St. Petersburg. p. 17-28.

52. Noth, E., et al., *VERBMOBIL: the use of prosody in the linguistic components of a speech understanding system.* Speech and Audio Processing, IEEE Transactions on, 2000. **8**(5): p. 519-32.

53. Oliveira, M.J. and D.A.C. Cunha, *Prosody As Marker of Direct Reported Speech Boundary*, in *International Conference*. 2004.

54. Peng, J.X., *On Issues and Applications for Support Vector Machine*. 2008, Shanghai University (People's Republic of China): Ann Arbor.

55. Pijper, J.R.d. and A.A. Sanderman, *On the perceptual strength of prosodic boundaries and its relation to supresagmental cues.* Journal of the Acoustical Society of America, 1994. **96**: p. 2037-47.

56. Poritz, A. *Hidden Markov models: a guided tour*. in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. 1988. New York, NY: IEEE.

57. Price, P.J., et al., *The Use of Prosody in Syntactic Disambiguation.* Journal of the Acoustical Society of America, 1991. **90**(6): p. 2956-70.

58. Rangarajan, S., et al., *Combining lexical, syntactic and prosodic cues for improved online dialog act tagging.* Computer Speech & Language, 2009. **23**(4): p. 407-22.

59. Rangarajan, V., S. Bangalore, and S. Narayanan, *Exploiting prosodic features for dialog act tagging in a discriminative modeling framework*, in *8th Annual Conference of the International Speech Communication Association, Interspeech 2007*. 2007: Antwerp, Belgium. p. 2460-63.

60. Rosenfeld, R., *Two decades of statistical language modeling: where do we go from here?* Proceedings of the IEEE, 2000. **88**(8): p. 1270-78.

61. Schukat-Talamazzini, E.G., T. Kuhn, and H. Niemann, *Progress and Prospects of Speech Research and Technology*, in *Proc. of the CRIM/FORWISS Workshop*, H. Niemann, R. de Mori, and G. Hanrieder, Editors. 1994, Infix: St. Augustin, CA. p. 110-20.

62. Selkirk, E., *The Syntax-Phonology Interface*, in *International Encyclopaedia of the Social and Behavioural Sciences*, N.J.S.a.P.B.B. (Eds), Editor. 2001: Oxford: Pergamon. p. 15407–12.

63. Shriberg, E., *Phonetic Consequences Of Speech Disfluency*, in *Proceedings of the International Congress of Phonetic Sciences*. 1999: San Francisco. p. 619-22.

64. Shriberg, E., R. Bates, and A. Stolcke, *A prosody-only decision-tree model for disfluency detection*, in *Proc. EUROSPEECH*. 1997. p. 2383-86.

65. Shriberg, E., et al., *Can prosody aid the automatic classification of dialog acts in conversational speech.* Language and Speech, 1998. **41**(3-4): p. 439-87.

66. Shriberg, E. and A. Stolcke, *Prosody modeling for automatic speech understanding: an overview of recent research at SRI*, in *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. 2001. p. 13-16.

67. Shriberg, E. and A. Stolcke, *Prosody modeling for automatic speech recognition and understanding*, in *in Proc. Workshop on Mathematical Foundations of Natural Language Modeling*. 2002, Springer. p. 105-14.

68. Shriberg, E., et al., *Prosody-based automatic segmentation of speech into sentences and topics.* 2000. **32**(1-2): p. 127-54.

69. Silverman, K., et al., *TOBI: a standard for labeling English prosody*. 1992, Proceedings of ICSLP. p. 867–870.

70. Stolcke, A., et al., *Dialogue act modeling for automatic tagging and recognition of conversational speech.* Computational Linguistics, 2000. **26**: p. 339-73.

71. Stolcke, A., et al., *Dialog Act Modeling for Conversational Speech*, in *AAAI Spring Symposium on applying Machine Learning to Discourse Processing*. 1998, AAAI Press. p. 98-105.

72. Sun, X., *A pitch determination algorithm based on subharmonic-to-harmonic ratio*. the 6th International Conference of Spoken Language Processing. 2000.

73. Swerts, M., G. R., and T. J., *Prosodic correlates of discourse units in spontaneous speech*, in *Proceedings of the International Conference on Spoken Language Processing*. 1992: Banff. p. 421-28.

74. Swerts, M., E. Strangert, and M. Heldner. *F0 declination in read-aloud and spontaneous speech*. in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. 1996.

75. Waibel, A., *Prosody and Speech Recognition (Artificial Intelligence)*. 1986, Carnegie Mellon University: Ann Arbor. p. 225.

76. Wales, R. and H. Toner, *Intonation and ambiguity*, in *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garret*, W.E.C.E.C. Walker, Editor. 1979, Halsted Press: Erlbaum, Hills-dale, NJ. p. 135-58.

77. Warnke, V., et al., *Integrated Dialog Act Segmentation And Classification Using Prosodic Features And Language Models.* 1997.

78. Wightman, C.W. and M. Ostendorf, *Automatic labeling of prosodic patterns.* Speech and Audio Processing, IEEE Transactions on, 1994. **2**(4): p. 469-481.

79. Wikipedia. *Kolmogorov–Smirnov test*. 2015 [cited 2015 April 6]; Available from: https://en.wikipedia.org/w/index.php?title=Kolmogorov%E2%80%93Smirnov_test&oldid=653097412.

80. Wikipedia. *McNemar's test*. 2015 [cited 2015 August 2]; Available from: https://en.wikipedia.org/w/index.php?title=McNemar%27s_test&oldid=65715987.

81. Wikipedia. *Simple linear regression*. 2015 [cited 2015 August 2]; Available from: https://en.wikipedia.org/w/index.php?title=Simple_linear_regression&oldid=673677013.

82. Zhu, Q. and A. Alwan, *On the use of variable frame rate analysis in speech recognition*, in *2000 IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*. 2000, IEEE, Piscataway, NJ, United States: Istanbul, Turkey. p. 1783-1786.

BIBLIOGRAPHIE

Rouaud, M. (2013). Calcul d'incertitudes, Mathieu Rouaud.