INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque

|__|__|__|__|__|__|__|__|__|__|

T H E S E EN COTUTELLE INTERNATIONALE

pour obtenir le grade de
DOCTEUR DE L'INP Grenoble
et
DE L'INSTITUT POLYTECHNIQUE DE HANOI

Spécialité : *« Système logiciel »*

préparée au laboratoire LIG
(Laboratoire d'Informatique de Grenoble)
dans le cadre de l'Ecole Doctorale
« Mathématiques, Sciences et Technologies de l'Information, Informatique »
et au Centre de recherche international MICA
(Multimédia Information, Communication and Application)

présentée et soutenue publiquement
par
Vũ Minh Quang
le 20 Septembre 2007

TITRE

EXPLOITATION DE LA PROSODIE POUR LA SEGMENTATION ET L'ANALYSE AUTOMATIQUE DE SIGNAUX DE PAROLE

DIRECTEUR DE THESE : M. ERIC CASTELLI ET M^{me}. PHẠM THỊ NGỌC YẾN
CO-DIRECTEUR DE THESE : M. LAURENT BESACIER ET M. NGUYỄN QUỐC CƯỜNG

JURY

M. Pierre-Yves Coulon	, Président
M. Philippe Martin	, Rapporteur
M. Jean-Yves Antoine	, Rapporteur
M. Eric Castelli	, Directeur de thèse
Mme. Phạm Thị Ngọc Yến	, Directeur de thèse
M. Laurent Besacier	, Co-encadrant
M. Jean-François Bonastre	, Examinateur

A ma famille…

Remerciement

Tout d’abord, je voudrais présenter tous mes remerciements, ainsi que toute ma gratitude, à mes directeurs et co-directeur de thèse, Eric CASTELLI et Laurent BESACIER, pour m’avoir accueilli dans l’équipe GEOD, et accompagné au cours des trois années de ma thèse avec leurs conseils, leur aide, et leurs encouragements précieux.

Je tiens à remercier PHẠM Ngọc-Yến et NGUYỄN Quốc-Cường, qui m’a accueilli au centre Mica, et pour avoir accepté d’être mon directeur et co-directeur du côté vietnamien pour cette thèse en cotutelle.

Je tiens à remercier Philippe MARTIN et Jean-Yves ANTOINE, qui m’ont fait l’honneur d’être mes rapporteurs, pour leur lecture attentive et pour toutes leurs remarques constructives sur le manuscrit. Un grand merci à Pierre-Yves COULON et Jean-François BONASTRE, qui ont accepté de participer à ce jury.

Je pense à ma famille qui m’a apporté un soutien important, non seulement à l’aspect sentimental, mais également par les encouragements, dont j’avais besoin pour mener à bien ce travail.

Un grand merci à tous.

Résumé

Cette thèse se situe à la frontière des domaines du traitement automatique de la parole et de la recherche d'informations multimédia. Ces dernières années, une nouvelle tâche est apparue dans le domaine du traitement automatique de la parole : la transcription enrichie d'un document audio. Parmi les informations extra-linguistiques transportées par la parole, une meta-donnée importante pour la transcription enrichie concerne l'information sur la nature des phrases parlées (c'est-à-dire les phrases sont-elles du type interrogatif ou affirmatif ou autre).

Notre étude a principalement porté sur la différence prosodique entre les phrases de type affirmatif et de type interrogatif pour les langues française et vietnamienne, la détection et la classification automatique du type de phrase pour chacune des deux langues et la comparaison des stratégies spécifiques à chacune des deux langues.

Nous avons commencé notre travail par l'étude sur la langue française. Nous avons ainsi réalisé un système de segmentation et détection automatique de type de phrases basé à la fois sur l'information prosodique et sur l'information lexicale. Le système a été validé sur des corpus de parole spontanée de la vie courante qui sont l'enregistrement de conversations téléphoniques entre un client et une agence de tourisme, des entretiens d'embauche ou des réunions de projet.

Après cette première étude sur la langue française, nous avons élargit notre recherche en travaillant sur la langue vietnamienne, une langue où les études de base sur le système prosodique sont encore toutes préliminaires. Nous avons d'abord poursuivi une étude pour identifier la différence prosodique entre les phrases interrogatives et affirmatives à la fois sur le plan de production et sur le plan de perception. Ensuite, sur la base de ces résultats, un moteur de classification a été construit.

Mots-clés : indexation, détection de questions, modèle prosodique, modèle lexical, recherche d'informations multimédia, arbre de décision

Abstract

This thesis work is at the frontier between multimedia information retrieval and automatic speech processing. During the last years, a new task emerged in speech processing: the rich transcription of an audio document. An important *meta-data* for rich transcription is the information on sentence type (i.e. sentence of interrogative or affirmative type). The study on the prosodic differences between these two types of sentences in Vietnamese language, the detection and classification of sentence type in French language and in Vietnamese language is the main subject of this research work.

Our departure is a study on French language. We've realized a system for segmentation and automatic detection of sentence type based on both prosodic and lexical information. The system has been validated on real world spontaneous speech corpus which are recording of conversations via telephone, between a client and a tourism office staff, recruiting interview, project meeting.

After this first study on French, we've extended our research in Vietnamese language, a language where all studies until now on prosodic system are still preliminary. We've carried a study on the prosodic differences between interrogative and affirmative sentences in both production and perception levels. Next, based on these results, a classification motor has been built.

Key Words: indexation, question detection, prosodic model, lexical model, multimedia information retrieval, decision tree.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1. **Introduction**

Nous commençons par une introduction générale sur le domaine de l'annotation automatique de documents audio. Une vue d'ensemble sur l'évolution du domaine est d'abord présentée. Celle-ci nous permettra de positionner notre sujet dans son contexte, qui est la *transcription enrichie* de documents. La problématique du sujet et la méthodologie utilisée seront ensuite décrites et nous finirons par la présentation de l'organisation du manuscrit.

1.1. Contexte

Pratiquement toute information multimédia se trouve aujourd'hui sous format numérique grâce aux technologies informatiques qui permettent de numériser l'information avec beaucoup d'avantages : transfert rapide, reproduction illimitée et sans détérioration, stockage facile.

Les corpus de documents audio et vidéo ne cessent de croître à la fois en nombre et en taille. La recherche de documents multimédia basée sur le contenu est devenue ainsi, une tâche très importante. Nous nous intéressons dans cette thèse plus particulièrement aux documents audio ou à la partie audio d'un document audio-vidéo.

La condition préliminaire pour un accès selon le contenu dans de grandes bases de données multimédia, est que ces données doivent être annotées (c'est-à-dire assigner des descriptions semantiques aux données). En effet, dans le cas de la parole par exemple, une simple transcription manuelle ou automatique issue d'un moteur de reconnaissance de la parole n'est pas suffisante pour ces signaux. Le signal de parole est très *riche* en informations. Dans ce signal, en plus des informations de transcription (le contenu sémantique, c'est-à-dire ce qui est dit), on peut extraire aussi d'autres informations de haut niveau, dites extra linguistiques, comme par exemple les hésitations, les frontières de phrases, l'identité du (des) locuteur(s), l'état émotionnel ou physiologique du locuteur, etc. C'est pourquoi nous pouvons imaginer de coupler un moteur de reconnaissance de la parole classique qui fournit uniquement une transcription simple avec d'autres moteurs d'extraction de méta-données extra-linguistiques afin de fournir une « transcription enrichie » comme illustrée dans la Figure 1.

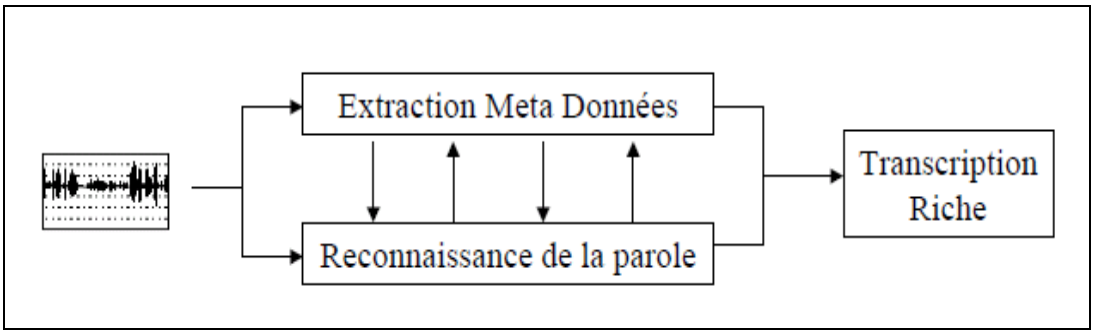


Figure 1: Architecture d'un système de transcription enrichie [Moraru, 2004]

Ce concept de transcription enrichie est fortement lié à l'évolution du domaine de la reconnaissance automatique de la parole comme cela est montré dans la Figure 2. Cette récente évolution amène de nouvelles difficultés : le flux audio est continu, hétérogène, pouvant provenir de multiples capteurs (microphones), de plusieurs médias, etc.

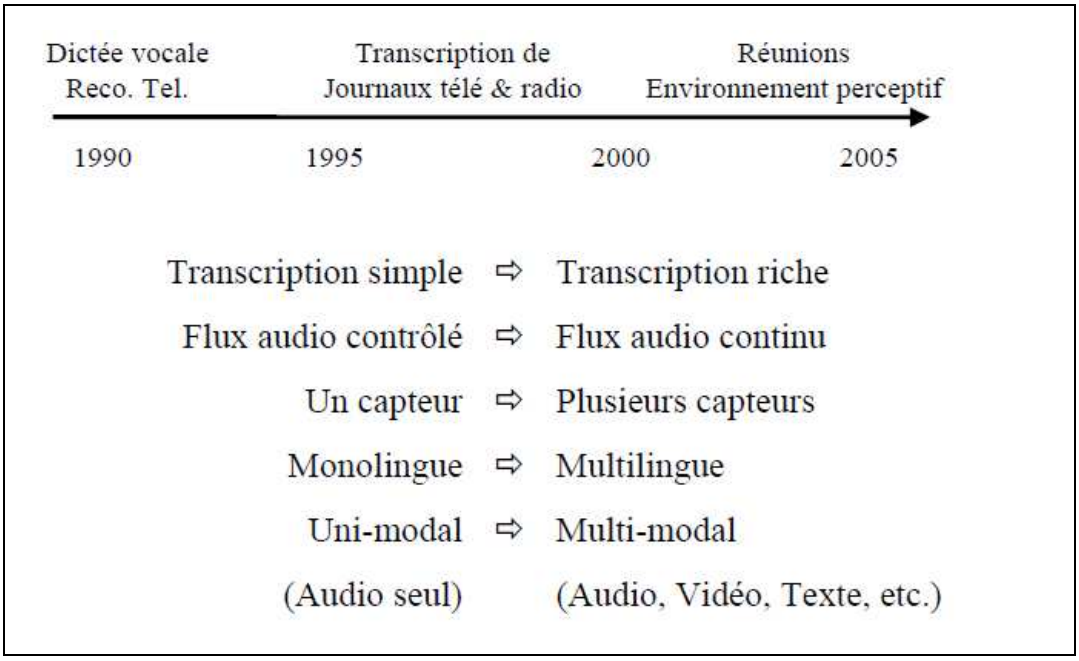


Figure 2 : Evolution du domaine en traitement automatique de la parole [Besacier, 2007]

Les signaux enregistrés dans les bases de données sont des enregistrements de longue durée et pouvant contenir plusieurs dizaines de locuteurs, différentes classes acoustiques (parole, silence, musique, etc.) et différentes qualités de parole (parole propre, parole bruitée, etc.). Le défi est de pouvoir sélectionner automatiquement les extraits ou segments d'intérêt sur les documents analysés. Plusieurs applications potentielles sont possibles grâce à cette évolution. Nous pouvons citer, par exemple, le traitement automatique des réunions à plusieurs personnes qui a récemment pris un plus grand intérêt pour les chercheurs du domaine du traitement de la parole. Ce type de réunion présente des problèmes et des défis réels qui sont liés à la communication humaine comme les répétitions, les hésitations, les émotions, les réparations, les pauses verbalisées, les ruptures, etc. Une autre application intéressante concerne les collections de journaux télévisés - notamment en raison de leur intérêt pour le grand public (retrouver un reportage à la demande, retrouver les déclarations d'une certaine personnalité politique, etc.) mais aussi en raison de leur problématique spécifique (traitement de flux audio et vidéo continus et hétérogènes, grand nombre de locuteurs, conditions audio très variables, présence de plusieurs sources d'informations différentes : audio, vidéo, sous-titrage, etc.). Pour le canal audio de ces applications, tandis qu'un meilleur taux de reconnaissance de parole est important dans ce travail, l'intérêt se porte également vers un plus haut niveau du traitement, comme la recherche d'informations et le résumé automatique du document [Fujie, 2003 ; Lin-shan, 2006 ; Chatain, 2006 ; Zhu, 2006 ; Nenkova, 2006].

Parallèlement, s'il est en effet toujours souhaitable de pouvoir éliminer certaines sources de variabilité (bruit ambiant, distorsions introduites par la prise de son, etc.), d'autres sont porteuses d'informations qu'il faut garder et exploiter : l'idéal serait de savoir décorréler les différents types d'informations qui se superposent dans le signal (audio, vidéo, transcription textuelle,

prosodie, etc.), tout en traitant leurs interactions dans l'optique de considérer l'ensemble des modalités. Si la reconnaissance de certaines de ces informations complémentaires comme l'identité du locuteur a atteint aujourd'hui un bon niveau de performance [Moraru, 2004], dans le domaine de la prosodie, en revanche, force est de constater que des recherches depuis des dizaines d'années n'ont toujours pas conduit à une utilisation efficace de la prosodie en traitement automatique de la langue française [Vaissière, 1999]. Cependant, nous proposons et défendons comme idée principale dans ce mémoire qu'il est avantageux d'exploiter et d'utiliser la prosodie dans les systèmes de traitement automatique pour extraire des informations extralinguistiques.

1.2. Objectif : Utiliser l'information prosodique en analyse automatique de parole

Utiliser la prosodie dans les systèmes de reconnaissance de la parole est un objectif de longue date, mais qui n'est toujours pas atteint. Avant d'examiner les raisons de cet état de fait, pour y apporter des propositions nouvelles, voyons ce qu'est la prosodie, quelles sont ses fonctions, et ce que l'on peut attendre de son utilisation dans un système de dialogue oral.

1.2.1. Définition de la prosodie

La prosodie désigne les phénomènes liés à l'évolution dans le temps des paramètres de hauteur, d'intensité et de durée. La perception de hauteur est essentiellement liée à la fréquence fondamentale (notée F0) qui correspond, au niveau physiologique de la production, à la fréquence de vibration des cordes vocales. La perception d'intensité est essentiellement liée à l'amplitude et à l'énergie du son, mais dépend aussi partiellement de sa durée. La perception de durée correspond à son temps d'émission, sa durée acoustique. On notera que le terme « durée » est utilisé pour désigner à la fois le paramètre perceptif et le paramètre acoustique. Le terme longueur comme synonyme de durée perçue est utile quand la distinction est importante, mais est moins souvent employé. La définition des paramètres prosodiques au niveau acoustique est illustrée Figure 3. La valeur de ces paramètres à un instant donné a peu de signification. Ce sont plutôt leurs variations qui peuvent être interprétées. La définition de la prosodie ci-dessus traduit cette préoccupation en faisant intervenir explicitement l'évolution dans le temps des paramètres. L'évolution dans le temps de la fréquence fondamentale, ou de la hauteur si l'on se place du point de vue perceptif, constitue la mélodie. L'enchaînement des durées relatives, y compris les durées des silences, constitue le rythme. Comprendre l'information contenue dans la mélodie et dans le rythme revient à en dévoiler les structures.

La notion d'accentuation se situe à un niveau d'abstraction plus élevé. On entend par accent une manifestation d'intensité, de hauteur et/ou de durée qui, portant sur une syllabe par exemple, la met en relief par rapport à ses voisines. La structure prosodique regroupe les structures accentuelle, mélodique et rythmique.

Le paramètre d'intensité a plus d'intérêt au sein de ces structures qu'en tant que paramètre autonome.

L'équation suivante résume ce qui précède :

> Prosodie = F0 + énergie + durée (grandeurs acoustiques)
>
> = hauteur + intensité + longueur (grandeurs perçues)
>
> = mélodie + rythme (structures) + accentuation

La mise sur le même plan des trois paramètres prosodiques dans cette formulation ne doit pas masquer le fait qu'ils sont de nature très différente. L'énergie est unique et bien définie pour toute portion de signal. La fréquence fondamentale F0 peut être définie dans le domaine temporel par le temps qui sépare deux impulsions glottiques, dans le domaine spectral par la structure harmonique du spectre, ou encore perceptivement comme la fréquence du son pur qui provoque la même perception de hauteur que le signal de parole.

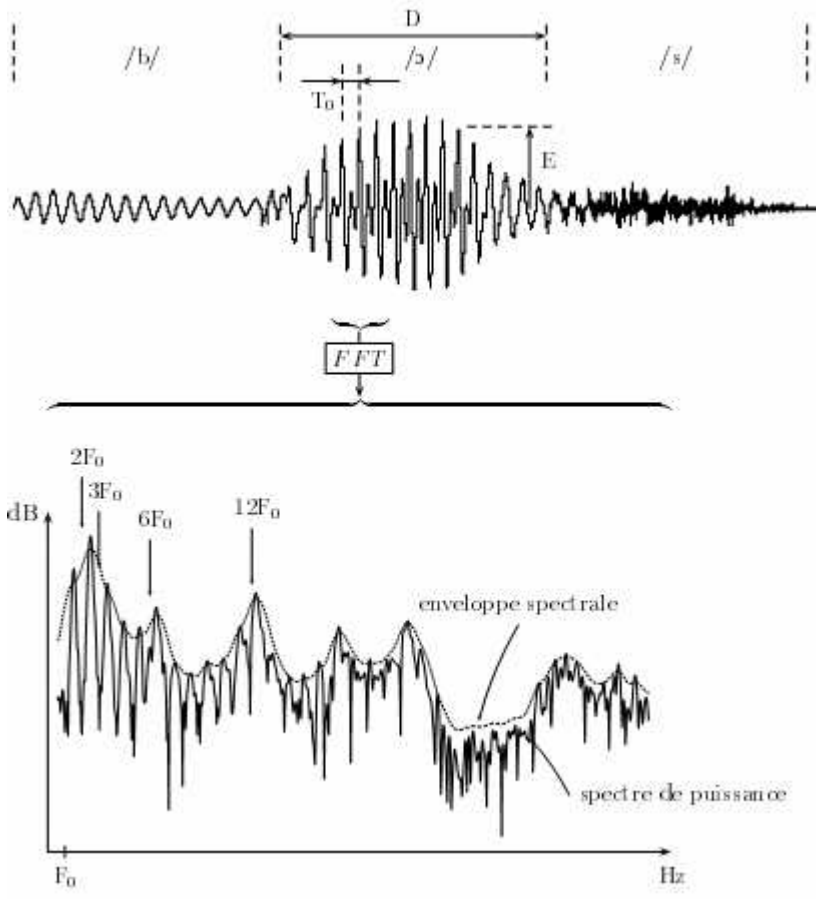


Figure 3: Définition des paramètres prosodiques au niveau acoustique

La Figure 3 illustre la définition des paramètres prosodiques au niveau acoustique: le signal correspond au son /bos/ (dans le mot « Boston »). En présupposant une segmentation, on peut définir la durée D d'un phonème. L'énergie E est liée à l'amplitude. Pour un signal supposé périodique, on peut définir la période fondamentale T0, dont l'inverse F0 et ses multiples s'observent sur la structure harmonique du spectre de puissance [Geoffrois, 1995]

Etant donné la complexité du domaine, ces définitions ne sont pas toujours consensuelles. La prosodie est le sujet d'étude de disciplines diverses (ingénierie des langues dans laquelle nous nous situons, phonétique, linguistique sur lesquelles nous nous appuyons, mais aussi psychologie, sociologie, etc.), et sa définition peut varier selon l'angle sous lequel on l'étudie. Nous allons nous positionner dans le domaine de l'ingénierie des langues dans le but d'extraire les informations prosodiques pour construire les systèmes d'analyse automatique de signaux de parole.

La prosodie est aussi définie par ses fonctions, c'est-à-dire ses liens avec la syntaxe, la sémantique et la pragmatique. Déterminer ces fonctions est un domaine de recherche à part entière, dont nous nous contentons maintenant de donner un aperçu.

1.2.2. Fonctions de la prosodie

La prosodie intervient à tous les niveaux dans le processus de production de la parole, et assure des fonctions variées. Dans les langues dites à accent libre et dans les langues à tons (comme le vietnamien, le mandarin…), elle est associée aux choix lexicaux, et dans les langues dites à accent fixe (comme l'anglais), elle est associée aux frontières de mots. Elle participe aussi à la structuration de l'énoncé, à la focalisation, à l'expression de la modalité, de l'attitude, bref, à indiquer le rôle que le locuteur attribue à la phrase dans son discours et dans la communication.

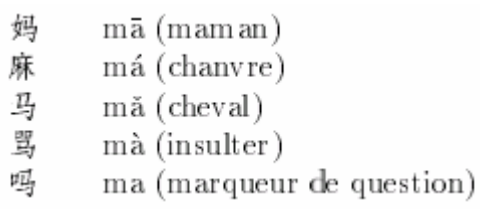
1.2.2.1 Distinction entre homonymes

Dans les langues dites à accent libre et dans les langues à tons, la prosodie est associée aux choix lexicaux. En particulier, elle peut discriminer des homonymes.

Voici quelques exemples : la position de l'accent tonique peut distinguer un nom d'un verbe (en anglais) :

permit (un permis) ←→ per**mit** (permettre)
contrast (un contraste) ←→ con**trast** (contraster)

Dans les langues tonales, l'évolution de F0 est utilisée pour coder l'information de ton. En chinois (mandarin et/ou cantonais) par exemple, le mot monosyllabique « ma » suivant peut avoir cinq significations différentes en fonction du ton qui l'accompagne :



En résumé, l'accent lexical est une information importante dans des langues comme l'anglais, et les tons sont une information quasiment indispensable dans les langues comme le chinois, le vietnamien, etc.

1.2.2.2 Structuration de l'énoncé

Dans les langues dites à accent fixe, la position de l'accent lexical permet de déterminer les frontières de mots. Cette fonction est qualifée de démarcative. Plus généralement, dans toutes les langues, la prosodie participe à la segmentation de l'énoncé en groupes plus petits. Elle indique aussi le type de relation que ces groupes entretiennent. Elle permet ainsi la hiérarchisation au sein de la phrase. Cette structure n'est pas identique à la structure syntaxique, mais est en relation avec elle.

L'exemple suivant illustre comment la prosodie participe au découpage et au regroupement de la phrase en mots à certains endroits plutôt que d'autres :

> L'instituteur, dit le proviseur, est un imbécile.
>
> L'instituteur dit : « le proviseur est un imbécile ».

Au delà de la phrase, la prosodie joue un rôle dans la structuration du discours [Shriberg, 2000]. L'indice prosodique le plus étudié est la dynamique locale de F0 (« pitch range »). Le rythme et l'intensité sont les deux autres indices importants.

1.2.2.3 Emphase et Focalisation

L'accentuation est un moyen d'insister sur tel ou tel mot :

> **Je** vais terminer (par opposition à quelqu'un d'autre)
>
> Je **vais** terminer (par opposition à une action déjà accomplie)
>
> Je vais **ter**miner (par opposition à une autre action)

Sur ce même exemple, l'insistance peut aussi servir à exprimer une volonté ferme de terminer. La prosodie peut alors éventuellement être identique à l'une des deux dernières versions ci-dessus, auquel cas c'est le contexte qui déterminera la signification.

L'accentuation peut aussi renforcer une opposition :

Non, pas le six... le **dix** août.

1.2.2.4 Modalité

La prosodie, et plus particulièrement la mélodie, est liée au mode de la phrase : affirmatif, interrogatif, impératif, ou exclamatif. En français, il est très courant de donner une intonation interrogative ou exclamative à une phrase tout en conservant sa structure grammaticale d'affirmation :

C'est fini?	Il va venir?
C'est fini.	Il va venir.
C'est fini !	Il va venir !

1.2.2.5 Attitude et Interaction

La prosodie véhicule l'attitude du locuteur vis à vis de l'énoncé ou ses intentions vis à vis de l'interlocuteur. En indiquant son adhésion plus ou moins forte envers l'énoncé, le locuteur exprime selon le contexte la conviction ou le doute, l'accord ou le désaccord, l'approbation ou la désapprobation, ou encore une invitation, une incitation. La dynamique de F0 en est une des manifestations acoustiques.

La prosodie, surtout si on y inclut les phénomènes d'hésitation, permet aussi de gérer les tours de parole dans une conversation. Un ton final, un silence laissent le champ libre à l'interlocuteur pour intervenir. Un ton continuatif, une pause verbalisée (« filled pause »), servent à conserver son tour de parole.

1.2.2.6 Fonctions non linguistiques

La prosodie au sens large traduit l'état psychologique du locuteur : calme ou énervé, triste ou gai, enthousiaste, surpris, etc. Elle caractérise aussi le locuteur en tant qu'individu ou que membre d'un groupe. L'accent régional en est un exemple.

1.2.3. Paramètres prosodiques

Comme pour les paramètres utilisés en reconnaissance (LPC, cepstre, etc.), la première étape de l'extraction de paramètres prosodiques est généralement une analyse à court terme, faisant l'hypothèse que pour une durée d'analyse assez courte le signal est quasi-stationnaire. Cette durée doit en même temps être suffisamment longue pour estimer les propriétés du signal.

1.2.3.1 La fréquence fondamentale

La mélodie de la voix se traduit sur le plan physique par l'évolution de la fréquence laryngienne - caractéristique des sons voisés - en fonction du temps. La plage de variation moyenne de cette fréquence varie d'un locuteur à l'autre en fonction principalement de son âge et de son sexe (de 80 à 160 Hz pour un homme adulte et de 150 Hz à 300 Hz pour une femme adulte), et peut enregistrer d'importantes variations chez un même locuteur.

1.2.3.2 Mesure de la durée

Les indices de durée classiques supposent généralement la disponibilité d'une segmentation, i.e. de frontières d'unités dont on désire mesurer la durée. La durée d'une unité est alors mesurée par le nombre de trames qui séparent ses frontières de début et de fin. Donner une mesure de la durée est un problème délicat qui nécessite en tout premier lieu de décider d'une unité de référence. Proposer l'alignement d'une chaîne phonétique donnée avec le signal de parole associé reste un problème ouvert. Si plusieurs techniques classiques permettent d'obtenir des résultats satisfaisants [Langlais, 1995], il faut cependant garder à l'esprit que cette même tâche demandée à plusieurs experts phonéticiens donne lieu à des variations assez sensibles dans des zones de forte co-articulation. Ceci est du à la nature continue du signal de parole.

1.2.3.3 Le paramètre d'intensité

L'énergie est à la fois le paramètre prosodique considéré comme le moins important perceptivement, et celui le plus facile à calculer. Il est d'ailleurs déjà utilisé couramment dans les systèmes de reconnaissance. L'énergie d'un signal échantillonné (x_t) est définie par :

$$E = \sum x_t^2.$$

Etant donné sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E(\mathrm{dB}) = 10 \times \log_{10}(\sum x_t^2).$$

Pour un signal échantillonné quelconque, on peut également calculer l'énergie à court terme en prenant des portions du signal par multiplication par une fenêtre glissante.

1.3. Problématique

L'étude de la prosodie est une science complexe [Langlais, 1995] mais très largement utilisée. Un système d'analyse automatique de parole peut alors être développé pour identifier et tirer

profit de la prosodie pour créer les annotations automatiques des textes, telles que des ponctuations (point, virgule…). Ceci est également important dans un contexte de compréhension automatique ou de dialogue. Parmi les informations extraites d'un flux de parole, la ponctuation indiquant si une phrase est de type affirmatif, interrogatif ou exclamatif constitue un enrichissement intéressant pour une transcription.

Le choix de la prosodie est justifié par un certain nombre de raisons. Premièrement, certains actes de dialogue sont ambigus. Par exemple, une question déclarative, « John est ici ? », présente le même ordre de mots que la phrase affirmative correspondante « John est ici. » et, par conséquent, la prosodie est le seul moyen de pouvoir les distinguer. Deuxièmement, dans une application, l'exactitude de la reconnaissance du mot peut ne pas être parfaite. En effet, les systèmes de reconnaissance récents présentent encore un taux d'erreur de reconnaissance des mots de 30% pour le discours conversationnel. Par conséquent, le fait de compter sur des identités de mot peut propager systématiquement des erreurs du système de reconnaissance de la parole aux systèmes de traitement ultérieurs qui se basent sur le texte. Troisièmement, il y a des applications potentielles pour lesquelles on ne peut pas avoir une véritable reconnaissance de la parole disponible, et où la tâche consiste plutôt à dépister sommairement ce qui se produit dans un dialogue.

Pour estimer automatiquement la nature des phrases il est possible d'analyser directement le signal de parole, sans avoir besoin nécessairement du résultat lexical d'un moteur de reconnaissance automatique de la parole mais en utilisant le contour intonatif et d'autres caractéristiques prosodiques de la phrase. Dans ce cas, l'essentiel des paramètres mesurés et analysés prennent en compte l'évolution de l'intonation pendant l'énoncé de la phrase : registre de F0, augmentation des valeurs de F0 en fin de phrase ou d'autres paramètres dérivés des valeurs de F0… De nombreux travaux ont été menés dans ce domaine, par exemple [Shriberg, 1998 ; Vu, 2005 ; Ang, 2005].

Une des questions à laquelle nous essayons de répondre dans ce travail, est de savoir si une combinaison des deux approches, qui emploient principalement l'information de mot et de la prosodie, pourrait amener à une bonne détection automatique de questions en langues française et vietnamienne. En effet, les auditeurs humains distinguent facilement de tels actes de dialogue (ADs) dans une conversation normale. Néanmoins, pour beaucoup d'autres applications, le développement de classificateurs d'actes de dialogue pour le discours conversationnel est clairement un but important. Par ailleurs, une tâche de résumé automatique de conversations (telles que l'enregistrement des réunions ou des interviews) pourrait bénéficier d'un système automatique permettant d'identifier les zones du document où ont été posées les questions, ainsi que les réponses associées.

Une autre motivation plus centrée sur le développement des technologies vocales est d'améliorer la reconnaissance du mot. Puisque le dialogue est fortement conventionnel, différents actes de dialogue tendent à impliquer différents modèles ou expressions de mot. La connaissance du type d'actes de dialogue peut donc être appliquée pour contraindre les hypothèses de mot dans un système de reconnaissance de la parole.

Par ailleurs, pour les langues non tonales occidentales comme l'anglais et/ou le français, il existe déjà dans la littérature un volume important de travaux sur la prosodie, tant sur le plan de la théorie que sur le plan de l'expérimentation. Cependant, de nombreuses autres langues n'ont pas franchi l'étape de l'informatisation : elles sont peu présentes sur Internet, les logiciels existants ne sont pas adaptés. Il s'agit de langues de pays en voie de développement, ou bien de langues régionales, ou encore de langues de minorités. Ces langues souffrent souvent de lacunes dans les travaux linguistiques qui leur sont consacrés et doivent faire face à diverses difficultés (manque de corpus, manque d'outil de traitement automatique, manque de services liés aux technologies du traitement de l'oral, etc.). Ces langues sont qualifiées de « peu dotées » : elles sont à la fois peu informatisées et peu étudiées.

Nous nous intéressons donc également à la différence prosodique entre les phrases interrogatives et affirmatives en langue vietnamienne – l'une des langues « peu dotées ». Comme le ton, ainsi que l'intonation, sont notamment véhiculés par le biais de la fréquence fondamentale (F0), comment, alors, la tonalité et l'intonation se réalisent dans le même espace F0 pour le Vietnamien ? Comment la tonalité et l'intonation interagissent l'un avec l'autre ? Par exemple, si une phrase finit avec un ton descendant, comment l'intonation d'une question, qui a normalement un contour montant dans la langue anglaise et les autres langues non tonales, peut-elle être réalisée ? Le ton descendant devient-il montant ? Est-il difficile de réaliser (et de caractériser) un contour intonatif de type interrogatif sur une telle phrase ? Pour l'intonation de question qui a une tonalité de montée à fin, comment un auditeur sait que l'expression est une question ?

Guidés par ces interrogations, nous présentons notre étude sur l'intonation du vietnamien, en nous concentrant particulièrement sur la différence entre l'intonation déclarative et interrogative.

Nous avons choisi de travailler sur une langue non tonale et une langue tonale. Ce choix est intéressant d'un point de vue de recherche parce que nous voulons vérifier si les méthodes d'analyse déjà validées sur une langue non tonale comme l'anglais et/ou le français peuvent fonctionner sur une langue tonale comme le vietnamien. Si la réponse est positive, cela ouvre alors la possibilité vers la généricité des méthodes d'analyse et leur application à un grand nombre de langues. Si la réponse est négative, cela nous permet de comparer les différences prosodiques et les différentes stratégies d'utilisation de la prosodie sur ces deux types de langues. Cela nous permettra également d'identifier s'il est possible d'envisager d'adapter un système déjà conçu pour une langue à une autre langue (problème de généricité).

1.4. Le manuscrit

Nous proposons l'organisation du manuscrit suivante.

Le deuxième chapitre est dédié à l'état de l'art. Puisque nos études s'effectuent sur deux langues : le français et le vietnamien, nous présenterons alors le domaine de recherche de la classification automatique de signaux audio et la transcription enrichie par un résumé rapide des

différents travaux publiés dans la littérature. Ensuite, nous aborderons les études actuelles sur la base de la prosodie du vietnamien.

Le troisième chapitre présentera d'abord les caractéristiques des phrases interrogatives du français, puis, il présentera notre étude sur la différence prosodique entre les phrases interrogatives et les phrases affirmatives du vietnamien.

Le quatrième chapitre présentera en détail notre système de détection automatique de question, y compris les jeux de paramètres prosodiques et les jeux de paramètres lexicaux.

Le cinquième chapitre détaillera nos expérimentations de détection de questions sur différents corpus en français et en vietnamien. Nous reporterons les taux de classification du modèle prosodique et du modèle lexical quand ils opèrent indépendamment ou conjointement. Nous présenterons aussi notre prototype de démonstration de détection de questions. Ce programme, nommé Coco, peut analyser soit le son enregistré depuis le micro, soit le son depuis un fichier de longueur variable : courte (une phrase) ou très longue (l'enregistrement d'une réunion).

Le document se terminera par la conclusion et les perspectives (le chapitre 6), suivies par les annexes.

Chapitre 2. **Etat de l’art**

La transcription enrichie d’un document audio a pour but de fournir non seulement le contenu linguistique (le résultat du module de reconnaissance automatique de la parole), mais encore la description de la nature des phrases (interrogatives, affirmatives…), du nombre de locuteurs, du regroupement en thèmes, de la détection des disfluences, etc. Avant de présenter la détection automatique de phrases interrogatives, qui est l’objet principal de cette thèse, nous allons faire un tours d’horizon sur les différents travaux qui ont été faits dans ce domaine de transcription enrichie afin de bien positionner notre travail de thèse.

2.1. Transcription enrichie de signaux de parole

La parole contient beaucoup plus d'informations que les mots en eux-mêmes, telles que l'état émotionnel du locuteur, la structuration de l'énoncé, l'accent, l'emphase, la modalité, l'attitude, etc. La transcription enrichie de signaux de parole consiste à identifier, exploiter et incorporer ces informations (que l'on appelle les informations « extra-linguistiques ») dans le résultat de l'analyse de signaux de parole. Comme nous avons montré dans le 1.2.2, ces informations sont véhiculées notamment par la prosodie de la parole. Il est donc souhaitable de pouvoir concevoir des systèmes d'analyse automatique de flux audio pour faire ressortir ces informations. En effet, dans certains cas, il serait bon de savoir ce qu'était le sujet d'un morceau de parole avant d'essayer de reconnaître les mots. Par exemple, si un système est employé pour analyser les canaux radio d'un rapport de trafic, ses performances pourraient être améliorées si le discours était préalablement catégorisé avant de le reconnaître [Tur, 2001]. Nous allons présenter dans la suite certains domaines de recherche sur l'analyse du flux de parole dans le but de fournir une transcription enrichie pour le signal de parole.

2.1.1. Détection des disfluences

Les disfluences (les pauses verbalisées, les répétitions, les réparations, les hésitations, les ruptures) sont répandues dans le discours normal et spontané. Le besoin de détecter et de corriger des disfluences est vraiment important pour une compréhension rapide du discours naturel.

En effet, les disfluences ont été étudiées en utilisant une variété d'approches. Une grande partie des linguistes et des psychologues ont considéré des disfluences d'un point de vue de production et de perception. Parallèlement, les chercheurs du domaine du traitement automatique de la parole se sont plus intéressés à identifier les disfluences dans le but d'améliorer le taux de reconnaissance automatique du discours spontané par la machine [Liu, 2004a ; Liu, 2006]. Nous allons discuter brièvement quelques études faites en psychologie et linguistique sur la théorie fondamentale de la production de disfluences et de leurs effets sur la compréhension des auditeurs. Ensuite nous allons analyser les modèles de détection automatique des disfluences.

Sur le plan de la production, les disfluences sont très communes dans le discours spontané. Quand les locuteurs ne peuvent pas immédiatement formuler une expression complète ou quand ils changent d'avis au sujet de ce qu'ils disent, ils peuvent suspendre leur discours et présenter une pause verbalisée avant de continuer, ou ajouter, supprimer, ou remplacer des mots qu'ils ont déjà produits. Les erreurs et les disfluences de la parole produites par les locuteurs normaux ont été étudiées pendant des décennies pour comprendre la production linguistique et les processus cognitifs de la planification de la parole. E. Shriberg [Shriberg, 1994] a prouvé que, dans différents types de conversations adaptées à la tâche, les longues expressions ont normalement un taux plus élevé de disfluence que les courtes. Cet effet peut être lié à la charge de

planification de l'expression, c'est-à-dire, les locuteurs ont plus de difficulté pour planifier de plus longues expressions. Une autre observation est que les disfluences se produisent plus fréquemment au début d'une expression quand celle-ci est encore dans l'étape préliminaire de préparation, fournissant l'évidence de l'impact de la planification d'expression sur les disfluences.

Sur le plan de la perception, dans une conversation, les disfluences fournissent aux personnes les moyens d'améliorer l'interaction et de contrôler le tour de parole. Les psycholinguistes croient que les disfluences jouent des rôles spécifiques dans la communication, pour envoyer des signaux à l'auditeur dans de nombreux buts tels que : attirer l'attention, aider le locuteur à trouver un mot, ou faire patienter quand le locuteur reconstitue ses pensées. Les études réalisées par Lickley [Lickley, 1995], ont prouvé que les auditeurs ont tendance à ne pas remarquer les disfluences ou rendent inexactement compte de l'occurrence des disfluences, suggérant que des disfluences aient pu avoir été filtrées afin d'assurer la compréhension de l'expression du locuteur. Cependant, les disfluences dans le discours posent des problèmes pour le traitement automatique et pour la lisibilité humaine des transcriptions de la parole. Les études récentes ont examiné l'effet des disfluences sur la lisibilité des transcriptions de la parole. Ces résultats suggèrent que le « nettoyage » du texte, en enlevant les disfluences, peut augmenter considérablement la vitesse de traitement du texte par les lecteurs [Jones, 2003].

Shriberg [Shriberg, 1994], dans sa thèse, a proposé une définition et une division des disfluences en trois composants principaux : reparandum (les mots qui sont réparés), interregnum (mots de remplissage ou pauses verbalisées) et reprise (le nouvel ensemble de mots qui répare le reparandum). D'une autre manière, les trois types de disfluences suivants sont les plus répandus et sont largement étudiés : répétitions (reparandum édité avec le même ordre des mots), réparations (reparandum édité avec un ordre différent des mots) et pauses verbalisées (mots dans la région d'interregnum). Les exemples dans le Tableau 1 illustrent ces trois types de disfluences :

Répétition	Je voudrais trois verres * trois verres de thé
Réparation	Je voudrais trois verres * non cinq tasses de thé
Pauses remplies	Je voudrais trois verres *uhm quatre verres s'il vous plaît

Tableau 1 : Exemples de disfluences

Des pauses verbalisées sont placées au point d'interruption du tour de parole du locuteur qui incluent des pauses telles que 'um', 'uh', 'bon'… des marqueurs de discours tels que 'bon',

'puis', 'vous savez'. Les pauses verbalisées peuvent servir à signaler l'hésitation ou la confusion du locuteur ou pour signifier le changement du sujet de conversation (selon le type de pause remplie qu'un locuteur emploie : par exemple 'uh' pour l'hésitation, 'ah' pour le changement du sujet de conversation…). Dans le Tableau 1, le 'umm' est une pause verbalisée placée au point d'interruption '*'.

Les répétitions sont l'un des types les plus communs de disfluences. Dans l'exemple du Tableau 1, 'trois verres' est une répétition. Des telles occurrences de répétition d'une partie d'une expression parlée sont des répétitions.

Les réparations peuvent signifier la confusion du locuteur. Dans l'exemple ci-dessus, le locuteur est confus(e) s'il/elle veut passer commande de 'trois verres' ou de 'cinq tasses' de thé. L'expression 'trois verres' est le reparandum, qui est réparé avec 'cinq tasses' après le point d'interruption. Les réparations peuvent également signifier l'hésitation du locuteur.

Il y a eu une quantité significative de travaux dans la détection automatique de disfluences [Liu, 2003 ; Liu, 2004b ; Nakatani, 1994]. La plupart des systèmes de détection de disfluence proposés utilisent une combinaison des paramètres prosodiques et des paramètres lexicaux, bien que quelques systèmes utilisent uniquement l'indice lexical et n'utilisent aucun paramètre acoustique. Par exemple, [Snover, 2004] se fonde exclusivement sur l'information lexicale des mots et il est montré dans ses travaux qu'une performance raisonnable peut être obtenue sans employer les paramètres acoustiques. En effet, les paramètres lexicaux comme les mots issus de la transcription manuelle ou du résultat de reconnaissance automatique constituent une source d'informations principales pour le traitement de disfluences. Certains mots clés sont de bons indicateurs pour les événements, par exemple « uh », « umh », « uhhuh »… [Johnson, 2004 ; Heeman, 1996].

Parallèlement, Nakatani et Hirschberg ont montré les avantages d'employer les paramètres acoustiques/prosodiques [Nakatani, 1994]. Ils ont détecté avec succès les points d'interruption (IP – interruption points) en établissant un arbre de décision avec les paramètres acoustiques. [Shriberg, 1997] a aussi proposé une méthode de détection des points d'interruption en utilisant un modèle d'arbre de décision basé seulement sur les paramètres prosodiques. [Stolcke, 1998a] a amélioré ce système en y ajoutant un modèle de langage pour modéliser et détecter les frontières et les divers types de disfluences.

L'ajout des paramètres prosodiques aux paramètres lexicaux présente quelques avantages certains. Par exemple, habituellement, l'intonation d'un locuteur est perturbée au point d'interruption qui pourrait indiquer une certaine forme de répétition. Une telle information est utile et s'est avérée significative [Shriberg, 1994]. Un avantage de l'emploi de la prosodie est aussi le suivant : même pour les langues qui manquent d'outils de traitement de langage écrit naturel, il est possible de concevoir des systèmes de détection de disfluences qui sont basés sur la prosodie et qui ne nécessitent pas une transcription automatique.

Il convient cependant de noter que les paramètres prosodiques ne sont pas toujours facilement disponibles pour quelques applications spécifiques. Les retards supplémentaires dans le traitement de son pour obtenir de divers paramètres acoustiques peuvent dégrader la performance globale du système, en particulier pour les systèmes interactifs tels que la traduction spontanée de parole qui nécessite une réponse quasi temps-réel. Par conséquent, [Maskey, 2006], dans leur étude récente, ont préféré utiliser seulement les paramètres lexicaux et non les paramètres prosodiques supplémentaires. Ils considèrent le problème de suppression des disfluences comme un processus de transformation de la transcription « bruitée » en une transcription « propre » qui pourrait être décrite en utilisant un modèle statistique de traduction.

2.1.2. Détection des frontières de phrases

À la différence des textes écrits, où les marqueurs de ponctuation indiquent clairement les frontières des clauses et des phrases, la langue parlée est construite comme une suite des mots, où les pauses (ou silence entre les mots) ne correspondent pas toujours à des segments linguistiques significatifs : quand une personne prononce une phrase ou une clause, elle peut faire une pause au milieu d'une phrase ou d'une expression ou, au contraire, ne faire aucune pause. C'est pourquoi un tour de parole peut contenir plusieurs phrases, ou au contraire, une phrase d'une personne peut se diviser en plusieurs tours. L'exemple suivant dans le Tableau 2 présente le texte à la sortie d'un moteur de reconnaissance automatique de la parole (en haut, sans aucun signe de ponctuation) et la version d'annotation manuelle correspondante (en bas, avec des signes de ponctuation). Il montre clairement que les informations structurelles des phrases (les signes de ponctuation…) sont absolument nécessaires pour une bonne interprétation par l'humain ou par l'ordinateur.

GOOD EVENING GIANNI VERSACE ONE OF THE WORLDS LEADING FASHION DESIGNERS HAS BEEN MURDERED IN MIAMI POLICE SAY IT WAS A PLANNED KILLING CARRIED OUT LIKE AN EXECUTION SCHOOLS INSPECTIONS ARE GOING TO BE TOUGHER TO FORCE BAD TEACHERS OUT AND THE FOUR THOUSAND COUPLES WHO SHARED THE QUEENS GOLDEN DAY
Good evening. Gianni Versace, one of the world's leading fashion designers, has been murdered in Miami. Police say it was a planned killing carried out like an execution. Schools inspections are going to be tougher to force bad teachers out. And the four thousand couples who shared the Queen's golden day.

Tableau 2 : Exemple de texte à la sortie du moteur de reconnaissance (en haut) et la version d'annotation manuelle correspondante (en bas) [Stevenson, 2000]

Certains des premiers travaux sur la détection de frontière de phrases se sont concentrés sur la désambiguïsation de la ponctuation dans le texte pour localiser des frontières de phrase [Reynar, 1997]. [Walker, 2001] abordent également le problème de « désambiguïsation de ponctuation » pour leur système de traduction automatique (L&H Power Translator Pro 7 English-to-French) comme illustré dans l'exemple suivant :

1. Phrase en anglais :	Is K.H. Smith here?
→ Mauvaise traduction automatique en français	Est K.H. Smith ici?
2. Phrase en anglais : enlever l'ambiguïté du signe (.) en remplaçant le nom *K.H. Smith* par *Kenvin Smith*	Is Kevin Smith here?
→ Bonne traduction automatique en français	Est-ce que Kevin Smith est ici?

Tableau 3 : Problème de l'ambiguïté du signe de ponctuation pour la traduction automatique [Walker, 2001]

Pour les applications de traitement du langage naturel, la grande majorité d'entre elles présuppose le découpage des textes en phrases. Cette tâche est depuis très longtemps automatisée, on parle alors de reconnaissance de frontières de phrases. La phrase est considérée comme l'unité centrale des processus du traitement du langage naturel. On reconnaît comme phrase la suite des mots qui se trouvent entre des signes de ponctuation dits majeurs tels que le point, le point d'exclamation, le point d'interrogation et d'autres qui précédent ou suivent ces signes. Formellement, la ponctuation désigne l'ensemble des signes qui permettent l'interprétation des textes écrits.

Le problème de la détection automatique des phrases se pose à cause de l'ambiguïté de certains signes de ponctuation. Par exemple un point peut être utilisé pour déclarer la fin d'une phrase mais aussi pour exprimer une abréviation ou un acronyme (ex. : E.D.F. ou U.S.A.), ou même un nombre décimal, par exemple : 3.14 (écriture anglosaxone).

Cette ambiguïté varie selon le type de texte ou de corpus. Des statistiques effectuées à partir d'un corpus composé d'articles du « Wall Street Journal » montrent que plus de 47% des points utilisés sont des points qui se trouvent dans les abréviations, contrairement à 10% pour le corpus Brown [Church, 1991]. Ceci démontre que si l'on coupe un texte en phrases sans tenir compte des particularités de l'ambiguïté on n'aurait que 53% des phrases correctement découpées pour le premier et 90% pour le second. Nous pouvons deviner que la plus fréquente source d'ambiguïté est le point d'une abréviation [Pappa, 2004].

Les travaux concernant la détection automatique des phrases représentent un premier pas pour une analyse morphologique ou syntaxique. Les techniques utilisées (souvent des listes grammaticales et de longues listes d'abréviations) visent à reconnaître les cas les plus courants. Dans leur majorité ces techniques sont orientées vers la détection des phrases pour des corpus particuliers ou dans une langue donnée, ce qui rend difficile leur adaptation à un autre type de texte ou à une autre langue sans avoir à modifier l'algorithme. De plus, puisque la détection de phrases est juste une première étape dans le traitement automatique du langage naturel, elle ne doit pas demander trop de ressources, ni de calcul.

Pour détecter des frontières ou la ponctuation de phrases dans la parole, diverses approches ont été employées. Un modèle de Markov caché (HMM) général a été conçu pour combiner l'indice lexical et l'indice prosodique pour étiqueter la parole avec différentes sortes d'informations cachées, y compris les frontières de phrase, les signes de ponctuation [Kim, 2001 ; Gotoh, 2000 ; Christensen, 2001 ; Shriberg, 2000], les disfluences, les frontières de thème, les actes de dialogue, et l'émotion [Ang, 2002 ; Stolcke, 1998a ; Stolcke, 2000].

Wang et Narayanan [Wang, 2004] ont développé une méthode qui emploie seulement des paramètres prosodiques (surtout le F0) ; cependant, ils n'ont employé aucune information lexicale qui s'est avérée très importante pour détecter des frontières de phrase. Huang et Zweig [Huang, 2002] ont développé un modèle d'entropie maximum pour ajouter automatiquement les ponctuations (points, virgules, et points d'interrogation) dans le corpus Switchboard en employant plusieurs indices lexicaux.

[Liu, 2006] proposent une grande variété de paramètres prosodiques et lexicaux pour la détection des frontières de phrases. Les études sur la détection de frontière de phrases dans la parole ont également été conduites pour d'autres langues, par exemple, le chinois [Zong, 2003] et le tchèque [Kolar, 2004]

2.1.3. Segmentation en thèmes (topic segmentation)

Dans beaucoup de cas, les dialogues parlés sont à thèmes divers. La segmentation en thèmes est le processus de division automatique d'un flux de texte ou de parole en blocs homogènes suivant le thème. C'est-à-dire, étant donnée une suite des mots (écrits ou parlés), le but de la segmentation en thème est de trouver les frontières où les thèmes changent. La segmentation en thème est une tâche importante pour différentes applications de traitement des langues, telles que l'extraction et la récupération de l'information, et le résumé automatique de textes.

Le travail sur la segmentation en thème est généralement fondé sur deux types d'indices principaux. D'une part, on peut exploiter le fait que des thèmes sont corrélés avec l'utilisation des mots spécifiques de thème, et que les variations globales dans l'utilisation de ces mots sont indicatives de changements de thèmes. Indépendamment, des indices de discours, ou les caractéristiques linguistiques telles que des marqueurs de discours, des expressions, des constructions syntaxiques, et des signaux prosodiques pourraient être utilisées par le locuteur en tant qu'indicateurs génériques pour signaler le changement de thème. Les travaux précédents sur le texte et la parole ont montré que certaines expressions clé ou les particules de discours (comme « now », « by the way »…) peuvent fournir de bons indicateurs pour le structure des unités dans le discours. Parallèlement, les paramètres prosodiques, représentés par les variations de F0, d'intensité, de durée constituent également de bons indicateurs pour le changement de thème. Beaucoup de travaux dans le domaine linguistique et les autres domaines concernés ont prouvé que la frontière de changement de thème est prosodiquement marquée de la même manière que la frontière de phrases. La plupart de changement de thème s'est manifestée par une pause plus longue, une montée (ou descente) importante de F0, une gamme de F0 et intensité plus étendue, etc.

Dans les systèmes de segmentation automatique, l'utilisation de paramètres basée sur les mots s'inspire de techniques comme le modèle statistique de langue ou par des techniques de recherche documentaire, alors que des indices de discours sont typiquement modélisés avec des approches basées sur les règles ou les techniques de type « machine learning » (telles que des arbres de décision, par exemple). Kozima [Kozima, 1993] a employé la similarité des mots dans une séquence de texte comme indicateur de structure des textes. Reynar [Reynar, 1994] a présenté une méthode qui trouve les régions de même thème dans le texte en modélisant graphiquement la distribution des répétitions de mot. Cependant, ces systèmes de segmentation automatique en thème se sont concentrés sur le texte écrit et dépendent en grande partie de l'information lexicale, cette approche est problématique en segmentant la parole, remarque Gokhan Tur [Tur, 2001]. Premièrement, le fait de compter sur des identités de mot peut

propager systématiquement des erreurs du système de reconnaissance de la parole au système de segmenteur en thème. Deuxièmement, la parole manque de signes typographiques comme de ponctuation de phrase, ou lettre en majuscule/minuscule. La parole elle-même, en revanche, fournit une source riche d'information via la prosodie.

[Tur, 2001] présentent leur travail sur la détection automatique des frontières de thème du discours en utilisant à la fois de l'information prosodique et lexicale. Des indices prosodiques sont appropriés à la structure de discours dans la parole spontanée et peuvent donc être susceptibles de jouer un rôle pour indiquer des transitions de thème. En outre, les paramètres prosodiques par leur nature sont relativement indépendants de l'identité de mot, et devraient donc améliorer la robustesse des méthodes de segmentation de thème utilisant le résultat textuel de la reconnaissance automatique de parole. Le travail de [Swerts, 1997] a utilisé des paramètres prosodiques et divers paramètres lexicaux modélisés par arbres de décision. Dans ce cas, les expressions étaient pré-segmentées, ainsi la tâche était de classifier des segments plutôt que trouver des frontières dans le discours continu.

La DARPA (U.S. Defense Advanced Research Projects Agency) a lancé le programme « détection et surveillance de thèmes » (Topic Detection and Tracking - TDT[1]) depuis 1997 (ce programme a pris fin en 2004) dédié à la détection et surveillance d'un flux de nouvelles (journal télévisé, par exemple). Une des tâches du programme TDT était de segmenter le flux en différents thèmes (ou histoires). Le programme couvre non seulement la langue anglaise, mais également une langue chinoise et une langue arabe.

Pour la campagne d'évaluation de suivi de thème TDT2002, un thème est défini par une ou plusieurs « histoires ». Ces « histoires » sont utilisées pour apprendre un modèle de classification automatique qui sera ensuite utilisé pour évaluer une nouvelle « histoire » inconnue et pour fournir une classification binaire afin de déterminer si la nouvelle « histoire » appartient ou non au thème en cours. Le plan d'évaluation de la campagne TDT2002 spécifie plusieurs conditions d'évaluation dans lesquelles certains facteurs varient, tels que le nombre d'histoires utilisées pour l'apprentissage/test de chaque thème, l'usage de la transcription automatique ou manuelle de données de test, la détermination automatique ou manuelle de la frontière des histoires, etc.

2.1.4. Segmentation en locuteurs

La segmentation et le regroupement en locuteurs consistent à identifier dans une longue conversation qui parle et quand. Dans le meilleur des cas, un système de segmentation et regroupement en locuteurs doit être capable de découvrir combien de personnes sont impliquées dans la conversation, et de comprendre quels tours de parole correspondent à quels locuteurs.

[1] Plus de détails sur le site : http://www.nist.gov/speech/tests/tdt/index.htm

En effet, le problème locuteur n'est pas spécifique à la transcription enrichie. Dans le domaine du traitement de la parole, depuis les années 70 de nombreuses recherches ont été faites en reconnaissance du locuteur avec comme but des applications de type judiciaire, serrure vocale ou authentification pour des transactions téléphoniques. Les premières tâches dans ce domaine furent l'identification du locuteur et la vérification du locuteur.

L'identification du locuteur doit établir à partir d'un échantillon de voix l'identité d'une personne parmi N personnes connues à l'avance par le système. Le système peut décider alors que l'échantillon de voix appartient à une des N personnes connues ou qu'il appartient à une personne inconnue.

La vérification du locuteur doit établir si un échantillon de voix correspond à une identité proclamée. La réponse du système dans ce cas est binaire : l'identité proclamée est l'identité réelle du locuteur ou non.

Dans ces deux tâches, le flux audio est contrôlé : le signal audio contient donc presque uniquement de la parole qui appartient à un seul locuteur et l'environnement varie peu pour un même enregistrement (niveau de bruit, type de microphone utilisé, etc.). La première évolution du domaine vers un flux audio continu est la tâche de suivi du locuteur.

Le suivi du locuteur a comme but de rechercher à partir d'un enregistrement où il y a plusieurs locuteurs qui parlent, les segments appartenant à un certain locuteur connu par le système. Cette tâche est assez proche de la segmentation en locuteurs mais il existe quelques différences : au niveau de l'évaluation, le suivi du locuteur s'intéresse uniquement à certains locuteurs et pas à tous les locuteurs, au niveau de la mise en œuvre ces locuteurs sont connus à l'avance par le système (le système dispose de modèles a priori pour les locuteurs concernés appelés locuteurs cible).

Ces dernières années, l'étude de réunions multi locuteurs est sujette à un intérêt croissant en traitement de la parole. Nous voyons cela spécialement grâce à l'apparition de plusieurs corpus volumineux d'enregistrements de réunions, et de programmes d'évaluation de transcription enrichie lancés par le NIST[2], sur des données de réunions.

La transcription totalement automatique des réunions est considérée comme un problème complexe [Jin, 2004]. Celle-ci comprend la transcription, l'extraction de méta-données, le résumé etc. La segmentation et le regroupement automatique de locuteurs sont un type d'extraction des méta-informations. Le NIST a commencé le programme d'évaluation de segmentation/regroupement en locuteur « qui a parlé quand » sur des conversations téléphoniques et sur des émissions de nouvelles en 2002. Cependant, il est plus intéressant de segmenter et de grouper des locuteurs impliqués dans les réunions où il y a du chevauchement

[2] Plus de détails sur le site : http://nist.gov/speech/tests/rt/

de locuteurs et des microphones éloignés. C'est ce type d'enregistrements qui fait l'objet de l'évaluation NIST[3] au printemps 2004

En effet, le but principal de la segmentation en locuteurs est de découper le flux audio en segments précisant pour chacun le début, la fin ainsi que l'étiquette du locuteur auquel il correspond. Les travaux sur la segmentation automatique des documents audio sont pour la plupart fondés sur des méthodes de traitement du signal et sur des approches statistiques.

Herbert Gish dans ses premiers travaux sur la segmentation en locuteurs [Gish 1991] a proposé une architecture se divisant en plusieurs étapes :

- Paramétrisation du signal de parole (calcul de paramètres acoustiques) ;
- Pré-segmentation acoustique (l'enregistrement est découpé en segments de parole, silence, musique, etc.) ;
- Détection de changements de locuteur ;
- Regroupement des segments (les segments trouvés pendant la phase précédente sont regroupés et le nombre de locuteurs est estimé).

Dans certain cas, les chercheurs ont ajouté à la fin une phase de re-segmentation [Reynolds 2002] pour augmenter la précision au niveau des frontières des segments.

Pour l'étape de paramétrisation, les coefficients cepstraux de type MFCC sont les plus utilisés et donnent les meilleurs résultats. L'utilisation de l'énergie est aussi fortement recommandée en segmentation. Les mêmes coefficients cepstraux donnent aussi les meilleurs résultats en ce qui concerne la pré-segmentation acoustique. Pourtant, d'autres paramètres issus du traitement automatique de signaux musicaux, semblent mieux fondés d'un point de vue théorique. Ces paramètres essayent d'utiliser les caractéristiques propres à chaque type de classes (silence, musique, etc.) par rapport aux paramètres cepstraux fondés uniquement sur les différences de spectres entre classes. La détection de changements de locuteur essaie de faire un compromis entre deux aspects : avoir des segments longs pour augmenter la robustesse de la phase de regroupement et en même temps, avoir des segments qui ne contiennent qu'un seul locuteur (pour le regroupement les segments sont considérés uni-locuteur). Pour la détection des changements de locuteur, l'utilisation d'une distance calculée entre deux fenêtres adjacentes le long du signal, donne les meilleurs résultats [Moraru, 2004].

2.1.5. Classification d'actes de dialogue

L'analyse et la classification automatiques de la langue parlée dans les discours structurés sont une tâche d'importance fondamentale pour les systèmes qui visent à réaliser la compréhension

[3] NIST, Rich Transcription 2004 Spring Meeting Recognition Evaluation, http://www.itl.nist.gov/iad/894.01/tests/rt/rt2004/spring/

du langage naturel. Il est souhaitable d'avoir un système qui peut comprendre non seulement l'ordre des mots que dit un locuteur, mais également le rôle qu'ils jouent dans la structure du dialogue. La théorie des actes de dialogue (AD) fournit un cadre pour analyser la structure des dialogues, l'attitude du locuteur (ou auditeur) en ce qui concerne le déroulement de la conversation, de sa compréhension, de l'intention, etc. Le développement des techniques pour identifier et classifier automatiquement les actes de dialogue est donc un but important de tout système visant à comprendre la structure de la langue parlée, que ce soit pour la communication dans un scénario d'interaction homme-machine, ou pour les applications d'analyse de dialogues homme-homme. Plusieurs études, [Finke, 1998 ; Ries, 1999 ; Shriberg, 1998 ; Stolcke, 1998b ; Stolcke, 2000], ont proposé différentes approches pour modéliser l'acte de dialogue. Ces approches combinent habituellement les modèles qui capturent des informations lexicales des actes de dialogue, leur réalisation prosodique, aussi bien que l'ordre d'apparence des actes. Les résultats reportés dans ces travaux suggèrent que la combinaison des résultats des modèles indépendants améliore les résultats obtenus avec chaque modèle pris séparément. En particulier, il est démontré que la prosodie peut aider dans la tâche de désambiguïsation entre les actes ayant une structure lexicale semblable [Shriberg, 1998].

L'étude de [Stolcke, 1998b] suggère que la contribution des modèles lexicaux à la classification globale (par exemple, les modèles de langage n-gramme) est typiquement plus importante que la contribution d'un modèle prosodique. Cependant, un modèle prosodique peut contribuer non seulement à améliorer le taux combiné de classification d'AD. Il peut également aider quand la performance d'un modèle lexical se dégrade sur une sortie bruitée d'un système de reconnaissance de la parole. L'étude de [Shriberg, 1998] se concentre sur le dialogue homme-homme sans contrainte. Dans cette étude où les auteurs cherchent à comprendre si les propriétés prosodiques des expressions elles-mêmes peuvent être employées pour détecter le type d'acte, les sources additionnelles de connaissance, qui pourraient brouiller les résultats, sont éliminées. Tout les ADs sont à probabilité égale. L'information contextuelle de la grammaire de dialogue (telle que l'AD de l'expression précédente) est également exclue dans ces travaux. Leur but est de mieux comprendre des caractéristiques prosodiques des différents ADs, qui peuvent être utilisées ensuite comme modèles généraux dans la construction des corpus de parole naturelle.

Les modèles de classification d'actes de dialogue utilisant l'indice prosodique reportés dans la littérature comprennent les arbres de décision, les réseaux neuronaux, et les modèles de machines à support de vecteurs [Fernandez, 2002]

Certaines des catégories des AD, comme des «STATEMENT[4] » ou des « ACKNOWLEDGE[5] » sont normalement plus nombreuses que les autres catégories dans un corpus. Les autres auteurs comme [Shriberg, 1998] par exemple ont préféré choisir un nombre égal d'échantillons de chaque catégorie pour l'apprentissage de modèles de classification, tandis que [Fernandez, 2002] ont essayé de modéliser également ce déséquilibre de distribution. Pour former le modèle et évaluer sa performance, le nombre d'ADs inclus dans les données d'apprentissage et les

[4] Déclaration
[5] Accusé de réception

données de test sont approximativement proportionnels à la fréquence de l'occurrence de chaque catégorie dans le corpus.

Sur le français, il y a aussi un travail de l'auteur Sophie Rosset du laboratoire LIMSI (http://www.limsi.fr/). L'auteur essaie de détecter automatiquement le type d'acte de dialogue en utilisant une méthode d'apprentissage simple (Memory based learning) et en utilisant seulement les paramètres lexicaux [Rosset, 2005]. Dans ce travail, le corpus utilisé consiste en un ensemble de dialogues homme-homme en français, enregistrés dans un centre d'appel d'un service bancaire. Ces dialogues couvrent une grande variété de thèmes comme la demande d'information, le passage d'ordres, la gestion de compte, etc. Ce corpus est annoté avec le schéma d'annotation dialogique proposé dans le cadre du projet AMITIES et fondé sur la taxionomie de DAMLS [Hardy et al, 2003]. Dans le but d'obtenir une annotation fine et sur différents niveaux résumant l'intention du locuteur, la taxonomie utilisée comprend huit classes d'actes. Pour effectuer l'annotation automatique, chaque énoncé du corpus est représenté par un vecteur dont les éléments sont : l'identité du locuteur (client ou agent), le nombre de segments dialogiques dans ce tour, les premiers mots du segment annoté et les tags des huit classes définies. Il est clair que certains de ces paramètres ne sont pas facilement calculables automatiquement (comme l'identité du locuteur, même si ceci reste possible). L'auteur a utilisé l'implémentation IB1-IG du logiciel Timbl (http://ilk.uvt.nl/timbl/) avec la méthode de mesure de distance nommé Manhatan. Dans cette métrique, la distance entre deux objets est simplement la somme de la différence entre les différents traits de ces objets. Le principe de cette méthode consiste à comparer le vecteur entrant à l'ensemble des vecteurs du modèle et à assigner à celui-ci la classe du vecteur du modèle dont il est le plus proche. Le taux d'erreur de détection d'actes de dialogue est d'environ 16% avec cette méthode.

2.1.6. Détection de questions en langue anglaise et chinoise

La classification et détection de phrases interrogatives peuvent être vues comme un cas particulier de la classification d'actes de dialogue. Généralement dans ces systèmes, des paramètres prosodiques tels que le pitch, la vitesse de parole, l'énergie, et les durées de pause sont automatiquement extraits et modélisés afin de classifier une variété d'événements de ponctuation et de dialogue. Dans le travail de [Shriberg, 2000 ; Shriberg, 2001 ; Buckow, 1999], les paramètres prosodiques se sont avérés extrêmement utiles dans la détection de ponctuations, de disfluences, et dans le traitement d'hésitations.

D. Baron [Baron, 2002] dans son étude a prolongé l'utilisation de la prosodie au domaine des réunions normales en utilisant une collection de corpus enregistrée dans un institut universitaire. Il a défini la classification de phrases interrogatives/affirmatives comme une tâche indépendante. Ce corpus présente de nouveaux défis pour le traitement de parole car : (1) les locuteurs ont accès à d'autres modalités telles que le geste ; (2) les locuteurs se connaissent l'un l'autre - ce qui fait que certaines informations sont cachées et ne sont pas exprimées par la modalité de parole ; cela rend difficile le traitement d'ambiguité/rupture/hésitation en utilisant seule la prosodie extraite de signaux de parole ; (3) les dialogues ne sont pas typiquement contraints à un thème concret ; (4) il y a beaucoup de recouvrement entre la parole des locuteurs

qui parlent en même temps. Cependant, pour la tâche de classification de phrases en interrogative/affirmative, l'auteur a choisi seulement les phrases bien formées, toutes les autres phrases incomplètes ou contenant des disfluences sont exclues.

Du côté technique, cette étude emploie les paramètres dérivés automatiquement de la prosodie, y compris le pitch (et le pitch lissé), les durées des pauses et les statistiques d'énergie, pour établir un classificateur prosodique pour différents événements d'intérêt.

Récemment, les auteurs de l'équipe « Spoken language processing - Department of Computer Science - Columbia Univeristy, New York, USA » remettent en question le problème de détection de phrases interrogatives dans le contexte des systèmes d'assistance intelligente aux cours particuliers (Intelligent Tutoring Spoken Dialog System) [Liscombe, 2006 ; Jennifer, 2006].

Les systèmes d'assistance intelligente aux cours particuliers sous forme de logiciels éducatifs sont conçus pour aider les étudiants dans leur processus d'étude par les techniques d'intelligence artificielle. Cependant, l'efficacité de l'étude réalisée avec les systèmes d'assistance intelligente aux cours particuliers actuels est toujours bien au-dessous de l'efficacité observée avec les enseignants humains. Une des raisons possibles de ceci pourrait être le fait que les systèmes d'assistance intelligente aux cours particuliers assistés par la technologie de parole actuelle ne tiennent pas compte du caractère interrogatif de certaines interventions des étudiants (c'est-à-dire lorsque les étudiants posent une question) comme le font les enseignants humains. Les recherches ont prouvé que les questions d'étudiants sont une partie importante dans l'interaction étudiant-enseignant. Etant donné cet aspect important, les chercheurs ont commencé à concevoir des systèmes d'assistance intelligente aux cours particuliers qui prennent en compte les questions des étudiants. Cependant, Liscombe [Liscombe, 2006] confirme qu'il n'existe pas encore de système qui essaie d'identifier explicitement les questions d'étudiants. C'est la raison pour laquelle il a mené une étude de prévision automatique des tours de parole d'étudiants contenant des questions. Dans son système, l'auteur utilise des paramètres lexicaux, syntaxiques, prosodiques extraits à partir d'un corpus de dialogues étudiant-enseignant, ainsi que quelques paramètres supplémentaires qui concernent l'étudiant et le cours. Ses expérimentations avec l'arbre de décision C4.5 (5-folds validation croisée – AdaBoost) ont obtenu comme résultat un taux de détection de question de 79%. L'auteur confirme que le paramètre le plus important pour la prévision automatique de question dans son système est la prosodie - spécialement le contour de F0 des derniers 200 ms. Cette étude a été faite sur la langue anglaise en 2006.

Il y a aussi un autre travail reporté dans la littérature, celui de l'auteur [Yuan, 2005]. L'auteur étudie les différences prosodiques et lexicales entre la classe des phrases interrogatives et la classe des phrases affirmatives en langue chinoise. Le chinois est une langue tonale. La différence entre l'intonation de phrases interrogatives et de phrases affirmatives dans le chinois est compliquée en raison de l'interaction entre les tons et la courbe intonative liée au style interrogatif/affirmatif de la phrase [Yuan, 2002]. Par exemple, la courbe intonative d'une phrase interrogative contenant un ton montant sur le dernier mot ressemble à la montée finale observée

en anglais, tandis que la même courbe pour une phrase interrogative contenant un ton descendant sur le dernier mot sera très différente (comme montré dans les derniers 100 ms sur la Figure 4).

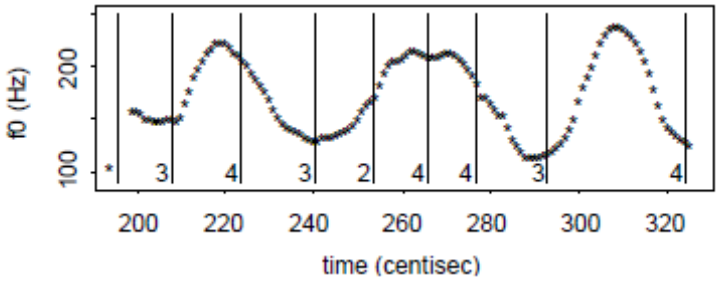


Figure 4 : L'intonation interrogative dans le chinois peut avoir en contour en chute. Les lignes verticales délimitent les syllabes. [Yuan, 2005].

En construisant des détecteurs de question pour le dialogue conversationnel chinois et anglais, et en exécutant les expérimentations de sélection des meilleurs paramètres, l'auteur veut répondre aux diverses questions telles que : quels paramètres sont utiles pour la détection de questions en chinois ? lesquels sont plus importants ? quelle différence entre le chinois et l'anglais concernant l'importance des paramètres, etc. Pour le modèle prosodique, l'auteur a choisi d'implémenter par une technique de type *machine learning* bien connue qui est l'arbre de décision. Cependant, pour le modèle lexical, l'auteur a choisi de modéliser, entre autres, le dernier mot dans une phrase. Ce choix a été motivé par une spécificité de la langue chinoise : la particule ('ma') pourrait s'ajouter à la fin de n'importe quelle phrase affirmative pour transformer celle-ci en phrase interrogative. Il constate que les paramètres textuels et prosodiques sont tous utiles pour la détection de questions en chinois comme en anglais. Parmi les paramètres textuels, l'occurrence de mots, particulièrement le mot final d'une expression, est plus utile que des probabilités de N-gramme (c'est-à-dire l'auteur s'intéresse plus particulièrement à l'identité du mot en position finale d'une phrase, il le considère comme un indice important pour la détection de question). À la différence de l'anglais, là où une courbe intonative montante en fin de phrase est une bonne indication pour des questions, l'auteur trouve pour la langue chinoise que la courbe intonative n'est pas suffisante. En plus, le paramètre prosodique le plus utile réside dans le spectre de la syllabe finale. Le système de classification final développé réalise un taux d'erreur de 14.9%.

2.2. Etudes récentes sur la prosodie en langue vietnamienne

Une grande partie de notre étude porte sur la langue vietnamienne. Nous étudions la différence prosodique entres deux classes de phrases : la classe des phrases interrogatives et la classe des phrases affirmatives.

Jusqu'à ce jour, très peu d'études ont analysé la phonologie de la langue vietnamienne en profondeur. Nous pouvons citer quelques travaux récents portant sur les tons lexicaux [Nguyen-Quoc, 2001 ; Pham-Thi, 2002 ; Michaud, 2004 ; Tran, 2005] et sur la prosodie de la phrase [Le, 1989 ; Nguyen-Thi, 1999 ; Vu, 2006].

De manière générale, les conclusions des auteurs ne vont pas toutes dans le même sens, surtout en ce qui concerne l'évolution des contours intonatifs. Il y a des doutes sur le fait que le contour descendant ou montant soit une des caractéristiques essentielles pour déterminer la modalité d'une phrase. Cependant, presque tous les auteurs sont d'accord sur le point que la distinction des types de phrases se manifeste plus par le registre (ou le niveau de hauteur globale) de la phrase que par le contour mélodique. Dans les études de [Do, 1998], les auteurs ont essayé de neutraliser l'effet de hauteur mélodique qui se manifeste déjà au niveau syllabique en travaillant sur un corpus de phrases simples. Ces phrases ont une même structure syntaxique, ayant chacune six syllabes de même ton dans le but de faire ressortir la ligne mélodique de la phrase. Cependant, ces phrases ont toutes une caractéristique plutôt artificielle parce qu'il s'agit de phrases rarement rencontrées dans la conversation courante. Les auteurs ont obtenu quelques observations intéressantes sur les patrons intonatifs suivants : la phrase déclarative est caractérisée par une déclinaison de la fréquence fondamentale en général, alors que les phrases interrogatives et impératives ont un contour montant, un niveau de hauteur global plus élevé et un débit plus rapide. De plus, les auteurs trouvent que l'évolution de la courbe d'intonation est la même dans la phrase impérative avec particule « đi » et la phrase interrogative avec « không ».

L'auteur Lê T.X [Le, 1989] dans ses études sur les différentes expressions émotionnelles dans la parole (neutre, déception, ennui, regret, joie…) a constaté que la phrase neutre est caractérisée par un registre moyen et un débit moyen. La colère se traduit par un registre plus élevé, un débit plus rapide et une intensité plus forte, alors que la tristesse se manifeste par les caractéristiques inverses : un registre plus bas, un débit plus lent, et une intensité moins forte. L'expression de surprise a normalement un contour mélodique commençant à un registre moyen, puis montant à la fin vers un registre plus élevé, mais son intensité est de niveau moyen.

Les auteurs Nguyễn Thị T. H. et Boulakia [Nguyen-Thi, 1999] affirment quant à eux que chaque type de phrase possède ses caractéristiques prosodiques particulières. Le contour général de F0 de la phrase n'est pas considéré comme un facteur essentiel en vietnamien à cause de la présence des tons. Selon eux, c'est le registre de F0, la durée et l'intensité qui constituent les facteurs discriminants des types de phrases. En effet, ils constatent qu'il existe une différence de hauteur de F0 entre les types de phrases dans leurs corpus de phrases « lues » et « spontanées ». En évaluant leur niveau, ils précisent que les assertives sont prononcées avec un registre bas alors que les questions et les injonctives le sont avec un registre haut. De plus, ils remarquent qu'au niveau de l'allure générale de l'intonation de la phrase, une pente descendante ne correspond pas toujours à une phrase déclarative. Au niveau de la durée, les énoncés interrogatifs ont un débit plus rapide que les énoncés assertifs et injonctifs. Cependant, la différence de durée entre les énoncés assertifs et injonctifs n'est pas significative. Quant à

l'intensité, elle est d'une manière générale plus forte dans la phrase interrogative, et les syllabes finales ont souvent un niveau d'intensité plus important que les autres syllabes de la phrase.

Concernant les méthodes automatiques de détection de phrases interrogatives en langue vietnamienne, aucun travail n'est reporté dans la littérature à l'heure actuelle.

2.3. Synthèse

En résumé, la transcription enrichie d'un document audio présente un potentiel important en vue d'une recherche d'informations basée sur le contenu. Parmi les informations extraites d'un flux de parole, la ponctuation indiquant si une phrase est de type affirmatif ou interrogatif constitue un enrichissement intéressant pour une transcription. Le problème d'extraction automatique de cette information extra-linguistique, qui n'est pas encore proprement traité en particulier pour la langue française et la langue vietnamienne, est l'objet principal de cette thèse. Les chapitres suivants exposent les recherches qui ont conduit à des solutions pour répondre à cette préoccupation.

Chapitre 3. **Différences prosodiques entre phrases questions et phrases nonquestions en langues française et vietnamienne**

Dans ce chapitre, nous allons aborder les points suivants :

- Une revue des résultats déjà connus sur les phrases questions du français sur le plan lexical et prosodique : les types de questions, les marques d'interrogation, etc.
- Notre étude sur la différence prosodique entre phrases questions et nonquestions en langue vietnamienne.

3.1. Caractéristiques des phrases questions en langue française

Dans cette section, nous allons résumer les caractéristiques des phrases interrogatives en langue française. Dans la littérature, nous pouvons trouver de nombreux travaux concernant l'étude sur les interrogations, citons pour exemple les travaux de [Fontaney, 1991 ; Hirst, 1998 ; Post, 2002] pour l'aspect phonétique, ou de [Bessac, 1996] pour l'aspect de traitement de dialogue en vue de la reconnaissance automatique de type d'actes de dialogue. Pour pouvoir utiliser ces résultats dans notre système de reconnaissance automatique de phrases interrogatives, nous allons faire le point sur les différentes caractéristiques essentielles des phrases interrogatives en français. Puis, ces caractéristiques seront adéquatement représentées par les techniques de modélisation : l'aspect prosodique du signal de parole sera représenté par un modèle prosodique, alors que l'aspect lexical du contenu de parole sera représenté par un modèle lexical. Ces deux types de modèles sont implémentés dans notre système sous forme d'arbres de décision.

Selon Natalie Colineau [Colineau, 1997], pour formuler une attitude d'interrogation (une question), le locuteur peut se servir de l'un des trois types de marques d'interrogation. Selon le contexte, le locuteur peut utiliser seulement une marque, ou plusieurs combinaisons de ces trois marques, qui sont :

- les marques prosodiques
- les mots ou les termes interrogatifs
- les expressions de demande

Dans les sections suivantes, nous allons faire le point sur la contribution de ces différentes marques dans la construction de phrases interrogatives.

3.1.1. Les marques prosodiques

No	Exemples de phrase
Question 1	allô oui vous m'entendez ?
Réponse 1	oui oui je vous entends.
Question 2	ça y est vous l'avez ?
Réponse 2	oui je vois le cercle oui.

Tableau 4 : Exemples de phrases interrogatives qui nécessitent la prosodie pour marquer l'interrogation

La prosodie est un marqueur pertinent pour l'analyse des actes de dialogue, et plus particulièrement pour les questions. Lorsque la question n'est pas marquée syntaxiquement (soit

avec une formule « est-ce que », soit par une inversion du sujet-verbe), la prosodie intervient alors comme seule marque possible pour distinguer la question de l'assertion. Dans ces cas, il faut utiliser des données prosodiques pour pouvoir identifier correctement que ces énoncés sont des questions (exemples de paire question-réponse dans le Tableau 4).

Pour résumer les résultats des différentes études reportées dans la littérature sur la prosodie de phrases interrogatives, il convient de citer un extrait de la thèse de N. Colineau :

« *… pour les questions à mouvement final ascendant, les phonéticiens distinguent plusieurs cas de figure selon que la montée est franche voire abrupte, selon que la montée est légère mais sur un énoncé d'un niveau général plus élevé que la moyenne, etc. Ceci montre qu'il est difficile d'établir un prototype intonatif unique qui regrouperait l'ensemble des questions à mouvement final ascendant, même si, par ailleurs, on constate que l'intonation montante est clairement dominante pour les questions oui/non (i.e. totale). Néanmoins, lorsque l'énoncé est caractérisé par un ton montant au niveau haut, on considère qu'il s'agit d'une question c'est-à-dire d'une demande d'information simple.*

De même, l'intonation descendante qui caractérise plutôt les assertions non marquées, indique assez rarement qu'il s'agit d'une question. Cependant, on peut trouver des schémas intonatifs à mouvement final descendant pour certaines questions. L. Fontaney a essayé de voir dans quelles circonstances, les questions à mouvement final descendant sont employées. Ainsi, elle montre qu'une question peut être marquée au niveau de l'intonation sans que ce soit en finale, c'est-à-dire dès le début de l'énoncé et en gardant un ton relativement élevé. »

Il est alors clair que la prosodie est en corrélation forte avec le type de phrase, surtout le type question où une intonation montante est clairement dominante. Cette caractéristique sera exploitée dans nos moteurs de détection automatique de question.

3.1.2. Les mots ou les termes interrogatifs

Ce type de marque correspond à l'usage de mots ou de termes interrogatifs. Ils peuvent assurer une fonction de pronom, d'adjectif ou d'adverbe dans la phrase. Il est possible de distinguer ici deux sous-classes de ce type :

- d'une part les termes interrogatifs qui assurent les fonctions de pronoms, d'adjectifs ou d'adverbes (*qui, quoi, que, est-ce que…*)
- d'autre part certaines expressions lexicales qui accompagnées d'une intonation montante, prennent une valeur de marque de demande (*c'est ça ? c'est bon ?….*)

Nous allons illustrer ces deux types, ainsi que leur usage dans les sections suivantes.

3.1.2.1 Les termes interrogatifs qui assurent les fonctions de pronoms, d'adjectifs ou d'adverbes

On retrouve les termes interrogatifs traditionnellement décrits par la grammaire. Nous avons relevé des pronoms interrogatifs de forme simple dans les exemples (4), (9) et (10) du Tableau 5, des pronoms interrogatifs de forme composée en (7), des adjectifs interrogatifs en (2), des adverbes interrogatifs en (1), (3) et (8), ainsi que la particule interrogative *est-ce que*.

Celle-ci peut, soit introduire une interrogation totale comme en (5), soit renforcer un terme interrogatif comme en (6).

No	Exemples de phrases
1	comment accéder aux numéros de ligne justement ?
2	quel est le prix ?
3	vous êtes combien de personnes ?
4	qu'est-ce que ça veut dire gruppi montuosi ?
5	est-ce que les couleurs vous ont choqués ?
6	où dois-je envoyer ce document hein ?
7	laquelle ?
8	jusqu'où sont-ils allés ?
9	qui est-ce qui a envie de venir ?
10	bien [ben] je ne sais pas vous en pensez quoi ?

Tableau 5 : Les exemples des phrases avec termes interrogatifs qui assurent les fonctions de pronoms, d'adjectifs ou d'adverbes

Ces marques interviennent principalement dans les questions introduites (c'est-à-dire les questions partielles marquées linguistiquement par un élément interrogatif). Elles sont généralement accompagnées d'indices prosodiques marquant la question [Colineau, 1997].

No	Exemples de phrases
11	oui ça dépend où on va parce que il peut faire froid quand même hein.
12	ouais au moins nous ça nous fera des comparaisons quoi.
13	je ne sais pas comment ça s'appelle.

Tableau 6 : Exemples des termes interrogatifs qui s'utilisent dans des énoncés affirmatifs

On peut retrouver les mêmes marques dans des emplois autres qu'interrogatifs, comme dans les énoncés du Tableau 6. Il apparaît dans ces exemples que prendre la marque isolément sans tenir compte des autres marques qui l'entourent, conduirait à une analyse erronée.

3.1.2.2 Les expressions lexicales qui sont accompagnées d'une intonation montante

Dans les conversations courantes, il apparaît que la plupart des questions possèdent à l'oral peu de marqueurs lexicaux ou morphosyntaxiques. En réalité, ces questions sont marquées : dans certains cas, il y a seulement la prosodie qui intervient pour identifier la question, dans d'autres cas la prosodie intervient en plus de certaines autres expressions spécifiques pour marquer une question comme cela est montré dans les exemples du Tableau 7 :

No	Exemples de phrases
1	c'est-à-dire on peut faire de la planche à voile des choses comme ça ?
2	je vais l'entourer en rouge vous voyez ?
3	comme ceci ?
4	oui en fait j'aurais aimé savoir si il y a des petits chalets individuels à louer ?
5	ça y est vous l'avez ?
6	okay on fait comme ça alors ?

Tableau 7 : Exemples des phrases avec expression lexicale interrogative

Les expressions les plus fréquentes sont celles que l'on rencontre dans les demandes de confirmation, de la forme « c'est ça ? », « c'est bon ? » … Elles sont en réalité des marques qui peuvent fonctionner aussi bien comme marque de confirmation que comme marque de demande de confirmation. Seule la prosodie de l'expression peut distinguer ces deux emplois.

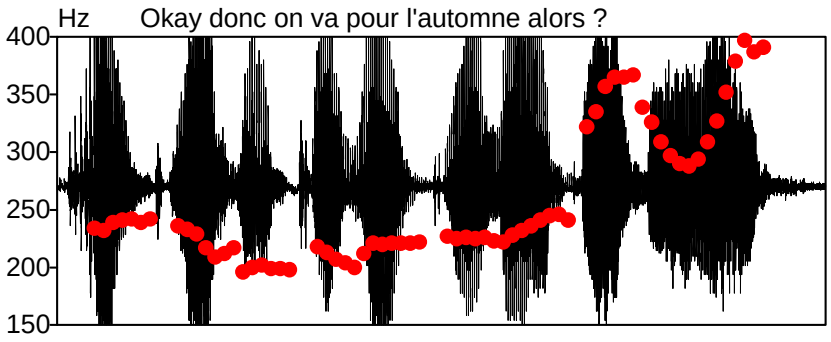


Figure 5 : Exemple d'une phrase avec expression lexicale interrogative « alors » avec le contour F0 (ligne en pointillé)

Il y a aussi d'autres marques de type marque d'insistance comme « hein », « alors », « n'est-ce pas », « non » comme illustrées dans l'exemple (6).

3.1.3. Les expressions de demande

Les expressions de demande sont des expressions qui introduisent explicitement la demande d'informations.

Lorsque la question est énoncée de manière indirecte, on relève des marques spécifiques dont le rôle est de marquer explicitement la question. C'est l'expression toute entière qui constitue la marque.

No	Exemples de phrases
1	on veut savoir qui veut faire quoi, quand, où ?
2	je voudrais savoir à qui est destiné le poster ?

Tableau 8 : Exemples des phrases interrogatives utilisant des expressions de demande

Il existe d'autres formes d'expressions de demande, ce sont les expressions performatives du type « je vous demande si ... ; je voudrais savoir… ; j'aimerais savoir… ; j'aimerais vous demander… ; je voudrais vous demander… ; j'aurais aimé savoir… ; on veut savoir…».

Ces résultats sont la base théorique de nos modèles de classification en langue française. Notre modèle prosodique ainsi que notre modèle lexical de détection de phrase interrogative prendront en compte ces descriptions.

3.2. Notre étude de la différence prosodique entre phrases questions et phrases nonquestions en langue vietnamienne

3.2.1. Introduction

Pour les langues occidentales non tonales (dont le français ou l'anglais sont des exemples) il a été montré que la prosodie de la phrase véhicule des informations extralinguistiques, comme les émotions, l'état du locuteur, ou la nature de la phrase (assertive, interrogative ou exclamative) [Rossi, 1999 ; Hirst, 1998]. Cependant, dans le cas des langues tonales (comme le mandarin ou le vietnamien), le contour mélodique de l'intonation est complexe parce qu'il est composé de macro variations correspondant à l'intonation de la phrase et de micro variations correspondant aux tons lexicaux appliqués sur chacune des syllabes des mots mono (ou bi) syllabiques.

Dans le cas de la langue vietnamienne d'ailleurs, pour différencier les phrases interrogatives des autres, l'emploi des termes interrogatifs (không, gì, ai, par exemple) est pratiquement systématique. Comme nous allons présenter dans le 3.2.1.2, la plupart des phrases interrogatives sont formées avec l'un des termes interrogatifs en langue vietnamienne.

Alors, la question que nous posons peut être résumée ainsi : existe-t-il, pour le vietnamien, langue à tons dont la prosodie est complexe, des informations extralinguistiques caractérisant le type de phrases, véhiculées par la prosodie et utilisées par l'auditeur pendant la réalisation de l'énoncé pour la classification de ces types de phrases ? La réponse, outre le fait qu'elle nous permettra d'approfondir nos connaissances de la langue, si elle est positive, nous permettra d'envisager la réalisation de classifieurs automatiques relativement indépendants des moteurs de reconnaissance.

3.2.1.1 Généralités sur la langue vietnamienne

La langue vietnamienne est généralement admise parmi les linguistes comme une langue de la famille austro-asiatique, groupe môn-khmer, sous-groupe viet-muong. Elle est utilisée par 83 millions de personnes au Vietnam et par environ 3 millions de personnes à l'étranger [Nguyen-Quoc, 2002]. Le vietnamien est une langue asiatique tonale, monosyllabique et bi-syllabique comme le mandarin, le cantonnais (un dialecte du chinois utilisé par environ 64 millions personnes dans le monde [Grimes, 1992]) ou le thaï.

Dans le vietnamien, chaque syllabe est prononcée avec un ton lexical. Le ton joue alors un rôle linguistique. Le vietnamien présente six tons lexicaux. Il y a environ 6800 syllabes avec tons parmi lesquelles environ 2400 syllabes sont indépendantes du ton (c'est-à-dire les syllabes comptées indépendamment du ton). Une syllabe peut se combiner avec un des six tons, ce qui lui donne alors six significations différentes (il est à noter cependant que toutes les combinaisons syllabe/ton ne sont pas possibles). A titre exemple, la syllabe /ma/ suivante peut avoir six sens différents comme le montre le Tableau 9 :

No	Ton	Mot en vietnamien	Mot traduit en français
1	1_NGANG	ma	fantôme
2	2_HUYEN	mà	mais
3	3_NGA	mã	cheval
4	4_HOI	mả	tombe
5	5_SAC	má	joue
6	6_NANG	mạ	semis

Tableau 9 : Différentes significations de la même syllabe /ma/ prononcée avec 6 tons

Selon les linguistes, la langue vietnamienne emprunte aux vocabulaires ainsi qu'aux grammaires des autres langues de la région. Les éléments empruntés sont : (1) les mots de base venus des langues monotoniques *mon* et *khmer* ; (2) certains éléments grammaticaux et la tonalité adoptés des langues *thaï*, (3) une grande partie du vocabulaire du vietnamien dans tous les domaines venant de la langue *chinoise* (comme par exemple le tableau 9 sur le vietnamien et la liste du

2.2.1 chapitre 1 sur le chinois avec la même syllabe « ma »), et puis (4) la langue française, puisque environ 1000 mots français, concernant essentiellement la mécanique et la cuisine, ont été adaptés en vietnamien durant le début du 20ème siècle.

Au niveau de l'écriture, la langue vietnamienne a utilisé les caractères chinois ordinaires (*chu nho*) jusqu'au XIIIe siècle, puis les vietnamiens ont inventé leur propre système d'écriture (*chu nom*) qui ressemble plus ou moins au chinois. Cependant, le vietnamien moderne s'écrit avec des caractères latins : le *quoc ngu*, inventé par le jésuite franco-portugais Alexandre De Rode vers 1700.

Après l'indépendance du pays vietnamien (2 septembre 1945), le *quocngu* a été adopté comme l'écriture officielle du pays et a été utilisé pour l'enseignement. A l'heure actuelle, seuls très peu de lettrés se souviennent encore de l'ancienne écriture comme *chu nho* ou *chu nom.*

3.2.1.2 Une vue globale sur les phrases interrogatives en vietnamien

D'une manière générale, il est possible de diviser les phrases « question » en vietnamien en trois catégories suivantes [Nguyen, 2001] :

- Les questions invitant à une réponse de type oui/non

Ce type de question peut utiliser les particules interrogatives comme « không, chưa, hay, etc. » Il n'y a que deux possibilités qui sont « oui » ou « non » pour répondre à ce type de question. Le Tableau 10 suivant présente quelques exemples :

No	Particule	Exemples de phrases en vietnamien	Traduction des phrases en français
1	Không	Anh muốn đi chơi <u>không</u> ?	Tu veux aller balader ?
2	Chưa	Anh ăn cơm <u>chưa</u> ?	Tu as mangé ?
3	Hay	Anh thích uống chè <u>hay</u> cà phê ?	Tu préfères boire du thé ou du café ?

Tableau 10 : Exemple des phrases interrogatives de type oui/non en vietnamien

- Les questions « ouvertes » qui utilisent des particules interrogatives comme « ai, gì, đâu… ». Il n'y a aucune contrainte particulière pour répondre à ce type de question. Quelques exemples sont présentés dans le Tableau 11 :

No	Particule	Exemples de phrases en vietnamien	Traduction des phrases en français
1	Ai	Anh cần gặp ai ?	Tu veux voir qui ?
2	Bao giờ	Bao giờ anh đến ?	Quand viens-tu ?
3	Bao lâu	Cuộc họp kéo dài bao lâu ?	La réunion dure combien de temps ?
4	Bao nhiêu	Cái này giá bao nhiêu ?	Ça coûte combien ?
5	Bao xa	Từ đây đến trường là bao xa ?	C'est loin, d'ici à l'école ?
6	Đâu	Anh muốn đi đâu ?	Où veux-tu aller ?
7	Gì	Anh muốn ăn gì ?	Que veux-tu manger ?
9	Mấy giờ	Anh hẹn người ta mấy giờ?	T'as rendez-vous avec lui à quelle heure ?
10	Nào	Anh muốn cái áo nào ?	Quelle chemise veux-tu ?
11	Như thế nào	anh tìm người như thế nào?	Quelle personne veux-tu chercher ?
12	Sao/tại sao/vì sao	Tại sao anh đi nhanh thế?	Pourquoi vas-tu si vite ?

Tableau 11 : Exemple des phrases interrogatives de type « ouverte » en vietnamien

- Les questions de type « suggérer une réponse attendue » : elles suggèrent une réponse de confirmation. Elles sont marquées par un terme de nuances et/ou d'attitude à la fin de phrase. La différence entre ce type et le type de question oui/non consiste dans le fait qu'il y a une suggestion d'information dans la question, et une confirmation positive à la suggestion est attendue dans la réponse.

No	Particule	Exemples de phrases en vietnamien	Traduction des phrases en français
1	À/ư	Trời mưa à ?	Il pleut ?
2	Chứ	Anh sẽ đến chứ ?	Tu vas venir ?
3	Hả/hử/hở	Anh biết tôi đợi từ lâu rồi chứ hả ?	Sais-tu que j'attendais depuis longtemps ?
4	Nhé	Anh sẽ đến đón em nhé ?	Je vais venir te chercher ?

Tableau 12 : Exemple des phrases interrogatives de type « suggérer une réponse attendue » en vietnamien

A part certains types de questions qui n'ont pas besoin de particule interrogative comme par exemple « còn anh ? » ; « thế thứ 7 ? »…, la majorité des phrases *question* en langue vietnamienne utilise l'une de ces particules interrogatives listées dans les tableaux ci-dessus (Tableau 10 ; Tableau 11 ; Tableau 12).

Il y a une autre caractéristique des phrases interrogatives en vietnamien, c'est le fait qu'une question peut utiliser certains termes de politesse à la fin de la phrase. Ces termes de politesse sont utilisés surtout dans les questions qui sont destinées à une personne plus âgée que le locuteur. Ces termes ne changent pas la signification de la phrase. Ils servent à altérer la nuance de l'interrogation dans la phrase et, de ce fait, la phrase interrogative devient moins « lourde » à percevoir pour son interlocuteur. On rencontre également l'usage de la forme de politesse dans les conversations plus formelles où un certain respect vis-à-vis de l'interlocuteur est nécessaire. Un exemple de question sous les deux formes « directe » et « politesse » est présenté dans le Tableau 13. La particule interrogative utilisée est « bao giờ ». La particule « politesse » utilisée est « **ạ** » marquée en gras.

No	Forme de phrase	Exemples de phrases en vietnamien	Traduction des phrases en français
1	directe	Bao giờ thì bác đến chơi ?	Quand venez-vous ?
2	politesse	Bao giờ thì bác đến chơi **ạ** ?	Quand venez-vous ?

Tableau 13 : Exemple d'une question sous deux formes : "directe" et "politesse"

Par conséquent, une phrase interrogative peut avoir plusieurs variations possibles. L'exemple suivant est une même question, mais sous 17 formes différentes dues à l'utilisation de particules de « politesse » et de différents pronoms personnels en fin de phrase (voir Tableau 14). Dans ce tableau, la colonne (1) est la racine de la phrase ; la colonne (2) donne les différentes particules de politesse accompagnées avec des pronoms personnels ; la colonne (3) est une simple traduction possible qui n'est pas vraiment équivalente significativement de ces termes en français.

(1)	(2)	(3)
Có mấy phòng tất cả + Combien de chambre y-a-t-il +	?	?
	ạ ?	s'il vous plaît ?
	rồi ?	déjà ?
	vậy ?	…
	hả bác ?	monsieur ?
	hả chú ?	monsieur ?
	hả ông ?	monsieur ?
	hả cô ?	madame ?
	hả anh ?	monsieur ?
	hả em ?	sœur/frère ?
	hả bà ?	madame ?
	hả dì ?	madame ?
	hả dượng ?	monsieur ?
	hả cậu ?	monsieur ?
	hả mợ ?	madame ?
	hả chị ?	sœur ?
	hả bạn ?	ami(e) ?

Tableau 14 : Exemple d'une phrase « question » avec 17 différentes variations tenant compte de l'interlocuteur et de la forme de politesse

3.2.1.3 Notre approche pour étudier la différence prosodique entre les phrases questions/nonquestions du vietnamien

Nous allons enregistrer des paires de phrases. Chaque paire se compose d'une phrase question et d'une phrase nonquestion. Puis, nous allons analyser ces paires de phrases dans le but de mettre en évidence les différences entre ces deux classes de phrases au niveau des trois aspects classiquement utilisés pour caractériser la prosodie : la fréquence fondamentale, l'intensité, le débit de parole. La deuxième étape qui vient ensuite consiste à vérifier, au niveau de la perception, si ces différences détectées dans la partie d'analyse sont effectivement perçues

et utilisées par les auditeurs comme moyen pour distinguer les phrases questions des phrases nonquestions. Cette vérification est faite par un test de perception

La superposition entre la macro-variation et la micro-variation (ou le ton) rend complexe la prosodie globale d'une phrase, et ainsi rend l'étude de la langue vietnamienne difficile. Une des difficultés demeure dans le fait que l'on ne peut pas extraire facilement la macro-prosodie de la prosodie globale. C'est la raison pour laquelle nous avons décidé d'adopter une nouvelle approche qui ne nécessite pas la séparation de ces composantes. Avec cette approche, nous allons choisir des paires de phrases qui se composent chacune d'une phrase question et d'une phrase nonquestion. Ces deux phrases comprennent un même nombre de syllabes et les tons de syllabes situées à la même position dans les deux phrases sont identiques.

Dans notre étude, nous avons porté une attention toute particulière au naturel des phrases, pour les phrases nonquestions comme pour les phrases questions. Pour cela, nous les avons extraites de corpus reproduisant des situations de la vie courante. Quand il y a des particules dans la phrase interrogative, nous avons, dans la mesure du possible, gardé aussi les mêmes mots, ou bien nous avons utilisé des mots à la prononciation peu différente afin que ces deux phrases soient les plus ressemblantes que possible au niveau de la prononciation. De cette façon, nous éliminons ainsi tous les phénomènes de co-articulation qui pourraient interférer avec notre analyse prosodique. Le fait de choisir les mêmes tons nous permet d'éliminer l'influence des tons des syllabes sur l'intonation générale de la phrase.

Après avoir sélectionné les phrases, nous procédons ensuite à l'enregistrement et à l'analyse de la différence prosodique entre phrases questions et nonquestions.

3.2.2. Analyse de la prosodie

3.2.2.1 Recueil du corpus

Il est nécessaire de rappeler qu'en vietnamien, pour la construction de phrases interrogatives, en plus de l'utilisation pratiquement systématique de mots du type « classifieurs interrogatifs », le locuteur peut ajouter en fin de phrase certains mots qui sont normalement facultatifs mais dont l'usage éventuel dépend fortement de l'habitude, de la façon de parler du locuteur, du contexte dans lequel se produit le dialogue, d'une manifestation de respect et/ou de politesse avec l'interlocuteur, etc. Cependant, la signification sémantique de la phrase n'est pas changée. Ces mots affectent seulement la modalité (voire l'attitude) liée à la phrase.

Comme ces mots terminaux facultatifs peuvent porter n'importe lequel des six tons de la langue vietnamienne, alors la portion finale du contour intonatif de la phrase peut être fortement modifiée par le contour du ton du mot terminal. C'est pourquoi, pour chaque phrase interrogative sélectionnée, nous avons choisi d'incorporer au corpus un certain nombre de variations possédant des mots terminaux de différents tons afin d'étudier plusieurs formes de contours intonatifs possibles de la fin de la phrase.

Les phrases questions et nonquestions sont ensuite incorporées dans des dialogues significatifs afin que leur prononciation soit la plus naturelle possible. Nous avons enregistré la totalité des dialogues, puis extrait les phrases choisies pour l'analyse. Chaque dialogue scripté est répété cinq fois par six locuteurs (3 hommes et 3 femmes) originaires du Nord du Vietnam dont le dialecte est considéré comme standard de la langue vietnamienne. Les phrases sélectionnées sont présentées dans le Tableau 15 suivant :

No de paire	Phrases en vietnamien	Phrases traduite en français
1	Hôm nay là ngày bao nhiêu ?	Quel jour sommes-nous aujourd'hui ?
	Hôm nay là ngày ba mươi.	Aujourd'hui nous sommes le trente.
2	Hôm nay là ngày bao nhiêu rồi ?	Quel jour sommes-nous aujourd'hui ?
	Hôm nay là ngày ba mươi rồi.	Aujourd'hui nous sommes le trente.
3	Hôm nay là ngày bao nhiêu vậy ?	Quel jour sommes-nous aujourd'hui ?
	Hôm nay là ngày đi chơi vậy.	Aujourd'hui nous nous baladons.
4	Hôm nay là ngày bao nhiêu thế ?	Quel jour sommes-nous aujourd'hui ?
	Hôm nay là ngày ba mươi đấy.	Aujourd'hui nous sommes le trente.
5	Hôm nay là ngày bao nhiêu hả ?	Quel jour sommes-nous aujourd'hui ?
	Hôm nay là ngày ba mươi bảy.	Aujourd'hui nous sommes le trente sept.
6	Tên anh ta là gì ?	Il s'appelle comment ?
	Tên anh ta là Trì.	Il s'appelle Trì.
7	Tên anh ta là gì rồi ?	Il s'appelle comment ?
	Tên anh ta là Trì rồi.	Il s'appelle Trì.
8	Tên anh ta là gì vậy ?	Il s'appelle comment ?
	Tên anh ta là Kì Cậy.	Il s'appelle Kì Cậy.
9	Tên anh ta là gì thế ?	Il s'appelle comment ?
	Tên anh ta là Kỳ Thế.	Il s'appelle Kỳ Thế.
10	Anh ăn cơm không ?	Tu manges du riz ?
	Anh ăn cơm không.	Tu manges du riz seulement.
	Anh ăn cơm không vậy ?	Tu manges du riz ?

11	Anh ăn cơm không vậy	Tu manges du riz seulement
12	Anh ăn cơm không thế ?	Tu manges du riz ?
	Anh ăn cơm Không Thế.	Tu manges du riz Không Thế.
13	Em ăn bánh nhé ?	Tu manges du gâteau ?
	Em ăn bánh ché.	Je mange du gâteau Ché.
14	Bao giờ chị gặp anh Nghĩa ?	Quand rencontres-tu Mr. Nghĩa ?
	Ba giờ chị gặp anh Nghĩa.	A trois heures je rencontre Mr.Nghĩa.

Tableau 15 : Les 14 paires de phrases nonquestions et questions du corpus

Notre corpus se compose alors au total de 840 phrases dont 420 questions et 420 nonquestions.

3.2.2.2 Méthodologie de l'analyse

Pour cette partie d'analyse, nous allons analyser la différence entre phrases questions et nonquestions pour plusieurs paramètres prosodiques : F0, intensité, durée. Ces paramètres seront extraits pour chaque phrase. La comparaison de ces paramètres se fait entre les deux phrases d'une même paire. Nous allons également comparer ces paramètres entre les hommes et les femmes participant à l'enregistrement du corpus. Les résultats des comparaisons et des statistiques seront présentés en détail dans le 3.2.2.3.

Afin d'avoir des données pour l'analyse, nous devons tout d'abord calculer la fréquence fondamentale F0 et l'intensité de chaque phrase. Nous avons utilisé le logiciel Praat[6], et le calcul a été achevé automatiquement par un programme script[7]. Pour avoir l'information sur le débit de syllabe, les signaux des phrases sont tout d'abord transcrits manuellement : cette transcription est faite aussi par le logiciel Praat avec son objet de type TextGrid.

[6] Praat : Doing phonetic by computer. http://www.praat.org

[7] script Praat : c'est un programme écrit en langage propriétaire du logiciel Praat pour automatiser certains traitements.

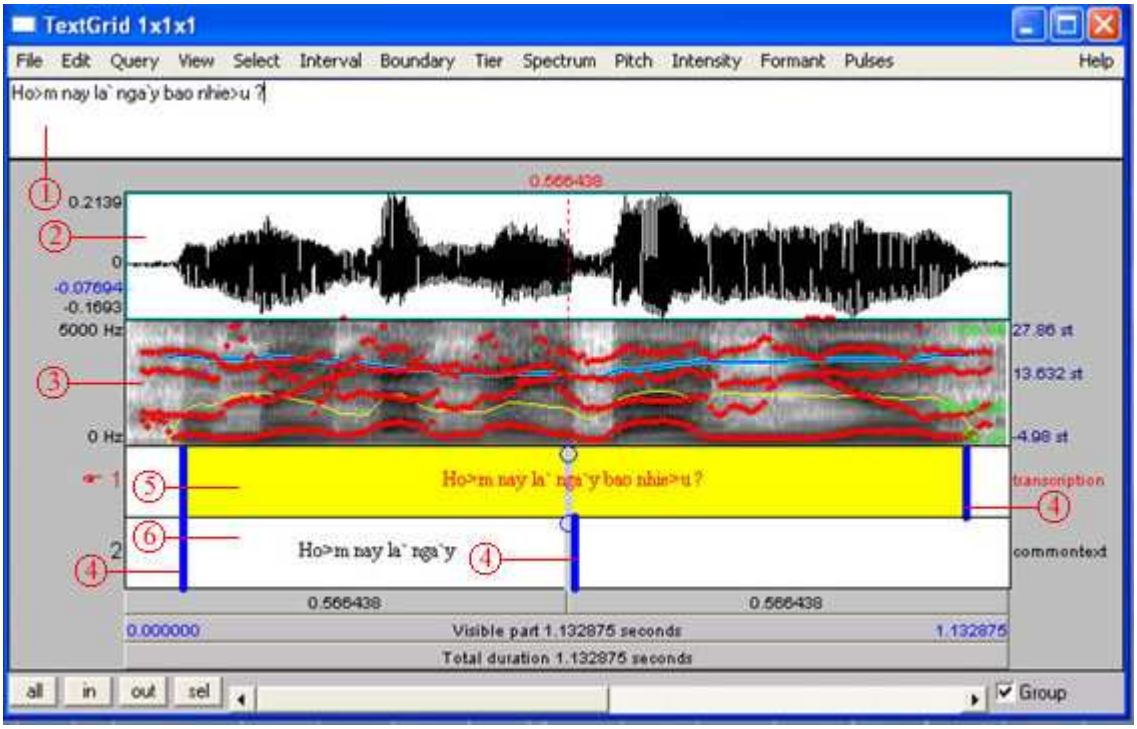


Figure 6 : Exemple de transcription d'une phrase dans l'environnement Praat

Dans la Figure 6, nous avons un exemple d'utilisation du logiciel Praat pour la transcription d'un fichier son. Les éléments dans cette figure sont :

1. Le texte de la couche de transcription selectionnée
   -dans ce cas, il peut être la couche 5 ou 6
2. Le signal
3. Le spectrogramme
4. Les frontières dans les couches de transcription
5. La premère couche de transcription : le texte de la phrase
6. La deuxième couche de transcription : le texte en commun de phrase interrogative et phrase affirmative dans la même paire

Nous avons transcrit l'ensemble des 840 phrases de notre corpus. Ensuite, l'information de la durée de la partie de texte en commun entre phrase question et nonquestion d'une même paire est extraite pour l'analyse.

De la même manière, nous avons extrait aussi l'information de F0 et d'intensité de chaque phrase. Comme nous voulons pouvoir observer le contour de F0 de l'ensemble des phrases, nous avons superposé le signal de parole avec son contour F0 comme montré dans la Figure 7.

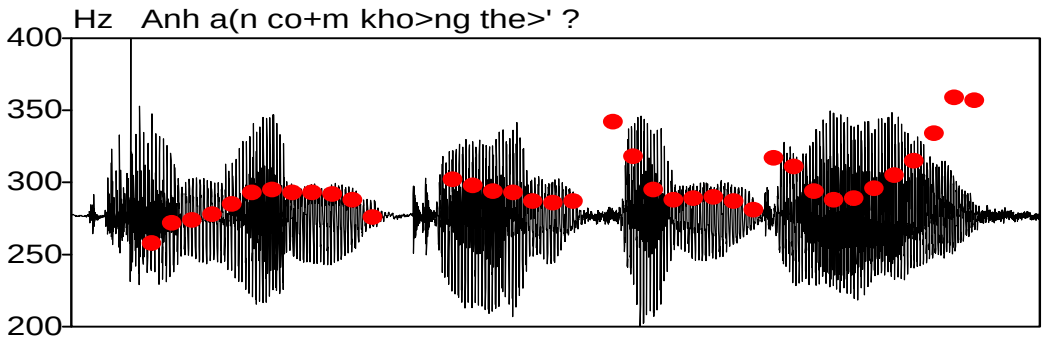


Figure 7 : Contour de Fo superposé avec le signal de parole correspondant

C'est en observant l'ensemble des images de courbe F0 des phrases questions et celles des phrases nonquestions que nous pouvons faire une première remarque importante sur le contour F0 : la plupart des phrases interrogatives possèdent un contour F0 montant dans la zone correspondant à la dernière moitié de la dernière syllabe. Cette constatation, qui a été validée ensuite par une étude statistique, sera présentée dans la section suivante. Cela constitue un des traits discriminants de la plupart des phrases interrogatives dans notre corpus.

3.2.2.3 Résultats d'analyse

3.2.2.3.1 Fréquence fondamentale F0

En étudiant chaque paire de phrases présentées dans le tableau 1, nous remarquons que l'essentiel des différences d'intonation se situe à la fin de la phrase (dans le Figure 8, c'est la zone située après la barre verticale) : le contour de la dernière syllabe ou de la deuxième moitié de celle-ci semble être croissant pour les phrases interrogatives. En effet, si on fait l'hypothèse que la durée moyenne d'une syllabe pour le vietnamien est environ 200 ms [Nguyen-Quoc, 2002], la dernière partie de la dernière syllabe sera la zone des dernières 100 ms en fin de phrase (après la barre verticale).

Nous trouvons que 357 phrases, ce qui représente un pourcentage de 85% des phrases questions, possèdent cette pente montante, alors que dans l'ensemble des phrases nonquestions, il n'y a que 190 phrases, soit 45% des phrases ayant cette pente montante. Il semble donc confirmé aussi pour la langue vietnamienne que la macro-prosodie influence la prosodie globale de phrase. Cela peut expliquer le fait que les courbes de F0 montantes sont beaucoup plus nombreuses pour les phrases questions que pour les phrases nonquestions. Si nous faisons l'hypothèse que la pente, de la dernière partie de la dernière syllabe, est montante dans les phrases questions et descendante dans les phrases nonquestions, il y a alors 70% des phrases dans le corpus satisfaisant cette hypothèse.

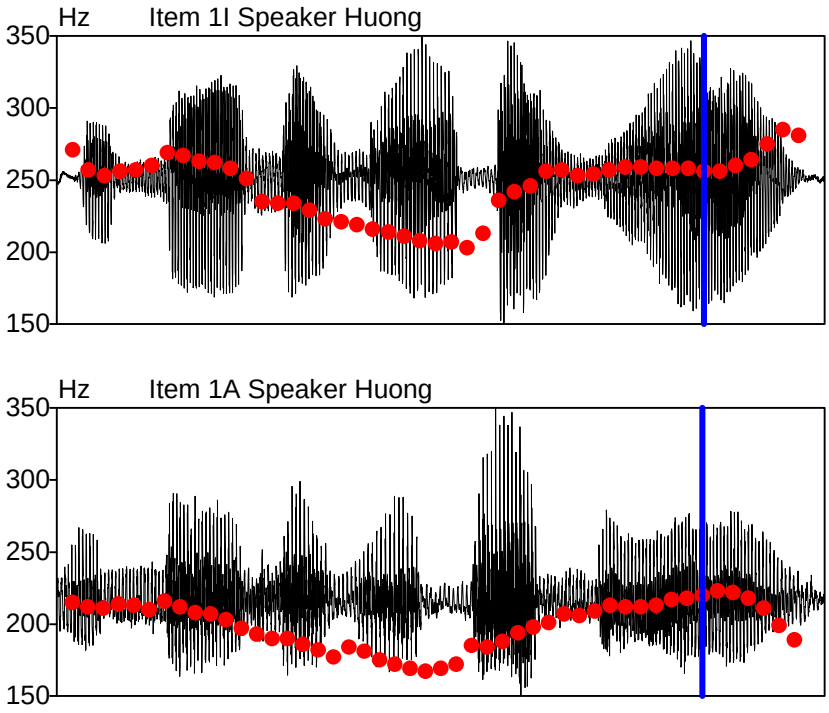


Figure 8: Deux phrases à nombre de syllabes et tons identiques. Contour de Fo superposé avec le signal de parole.Figure du dessus : phrase question, figure du dessous : phrase nonquestion.

Type de phrase	Direction de la pente	Nombre et pourcentage
Question	montant	357 (85%)
Question	descendant	63 (15%)
nonquestion	montant	190 (45%)
nonquestion	descendant	230 (55%)

Tableau 16 : Direction de la pente de la dernière partie de la dernière syllabe : Nombre et pourcentage de pente montante/descendante en fonction du type de phrase

Si nous regardons plus en détail la répartition de cette pente montante/descendante selon les tons de la dernière syllabe, nous nous apercevons que dans les phrases questions, la pente montante demeure pour tous les tons. Alors que dans les phrases nonquestions, cette pente reste descendante chez presque tous les tons, sauf le ton3 (NGA, BRISE) et le ton5 (SAC, AIGU). Pour ces deux tons, la pente est montante quel que soit le type de phrase porteuse. En effet, ces deux tons sont à registre haut, et leur propre contour F0 est de forme montante comme illustré sur la Figure 9 :

Type	Pente\Ton	1NGANG	2HUYEN	3NGA	4HOI	5SAC	6NANG
Inter-rogative	Montante	36 (60%)	69 (77%)	30 (100%)	24 (80%)	116 (97%)	82 (91%)
	Descendante	24 (40%)	21 (23%)	0 (0%)	6 (20%)	4 (3%)	8 (9%)
Affir-mative	Montante	1 (2%)	9 (10%)	30 (100%)	13 (43%)	119 (99%)	18 (20%)
	Descendante	59 (98%)	81 (90%)	0 (0%)	17 (57%)	1 (1%)	72 (80%)

Tableau 17 : Direction de la pente de la dernière partie de la dernière syllabe : Nombre et pourcentage de pente montante/descendante en fonction de type de phrase et de tons

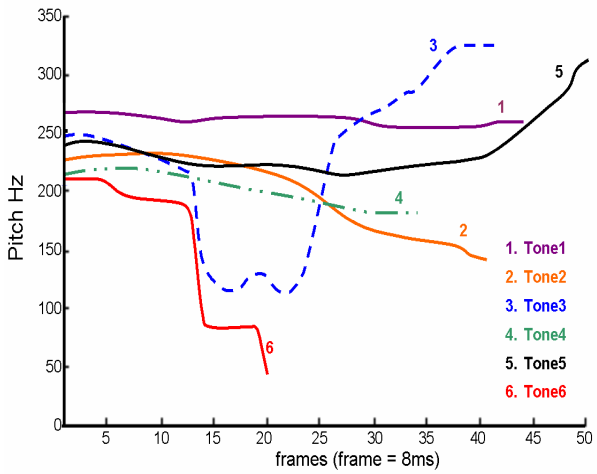


Figure 9: Contour de F0 des 6 tons vietnamienne [Nguyen-Quoc, 2002]

La Figure 10 présente un exemple de cas de phrases de type affirmatif (A) dont le contour de la dernière moitié de la dernière syllabe est aussi croissant (c'est le cas pour 29 enregistrements sur 30) : cette phrase « Em ăn bánh Ché » présente en fin de phrase deux mots au ton croissant qui influencent le contour intonatif global de la phrase.

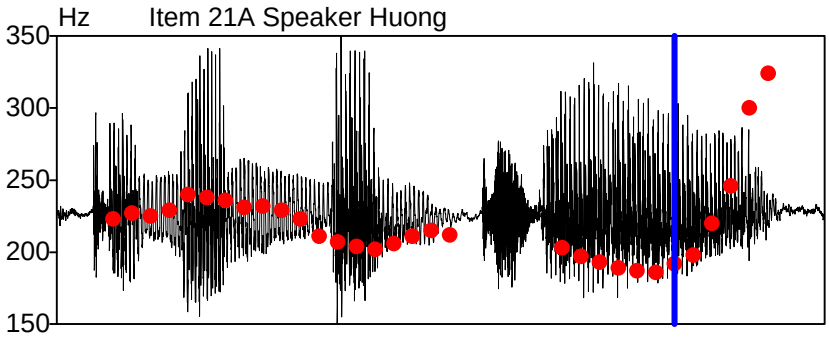


Figure 10 : Le signal et le contour F0 (en rouge) de la phrase nonquestion "Em ăn bánh Ché" avec deux derniers mots au ton croissant qui influence le contour intonatif global de la phrase.

[Le, 1989] et [Nguyen-Thi, 1999] ont suggéré que les phrases de type interrogatif sont prononcées avec un registre plus haut. Pour ce point, l'étude statistique de notre corpus montre qu'effectivement elles semblent montrer une valeur moyenne de F0 plus importante que celle des phrases de type affirmatif. Ici dans notre étude, nous faisons la comparaison par la méthode mathématique ANOVA [Ramousse, 1996] (voir le Tableau 18) :

Locuteur :	DIEP (femme)	HUONG (femme)	LAN (femme)	KHOA (homme)	THANH (homme)	PHUONG (homme)
Moyenne F0 - Question	262	260	254	144	144	125
Moyenne F0 - NonQuestion	226	218	246	127	137	119
F	108	150	13	95	23	20
Valeur critique pour F	3.91	3.91	3.91	3.91	3.91	3.91

Tableau 18 : Résumé de résultat d'analyse ANOVA sur F0 moyen (ou registre) pour les six locuteurs

Le Tableau 18 présente un résumé du résultat de l'analyse ANOVA pour les six locuteurs. Nous présentons seulement les valeurs du rapport « F » et le seuil critique pour F (qui est le seuil calculé en fonction du seuil de signification). Il convient de noter que pour toutes ces mesures ANOVA, le seuil de signification est fixé à 95%. La valeur de « F » doit être supérieuse à celle de la « Valeur critique pour F » pour pouvoir conclure que les jeux de données sont significativement différents.

ANOVA est une méthode pour étudier les différences de moyenne entre populations (par exemple, trois populations ont-elles la même moyenne? ou autrement dit, les différences de moyenne entre les trois populations sont-elles significatives ?) Cette méthode utilise des mesures de variance afin de déterminer le caractère significatif, ou non, des différences de moyenne mesurées sur les populations [Ramousse, 1996]

L'analyse de variance à un facteur (one-way analysis of variance) consiste à chercher le rapport F entre la variance entre les groupes (V. inter-groupe) et la variance à l'intérieur des groupes (V. intra-groupe) et à comparer ce rapport avec son seuil de critique afin d'avoir une conclusion sur la similarité ou la différence des jeux de données.

Après cette vérification par ANOVA, nous pouvons donc bien dire que l'effet de registre est significatif chez tous les locuteurs

3.2.2.3.2 Intensité

Certaines hypothèses trouvées dans la littérature indiquent qu'il existerait une différence d'intensité entre phrases de modalité interrogative et phrases des autres modalités [Nguyen-Thi, 1999 ; Nguyen-Thi, 2004]. Nous voulons nous aussi vérifier cette hypothèse sur notre corpus. Le tableau suivant présente les statistiques sur l'intensité moyenne des phrases questions et nonquestions. Ici, l'intensité est en unité décibel. Nous calculons d'abord l'intensité moyenne pour chaque phrase, puis l'intensité moyenne de chaque locuteur dans le tableau est alors calculée comme la moyenne de l'intensité de toutes les phrases interprétées par ce locuteur.

Nous trouvons que dans 87% des cas l'intensité des phrases questions est plus forte que celle de phrases nonquestions. Cette différence est généralement de l'ordre de 2 décibels.

Nous avons mesuré aussi l'intensité moyenne de la partie texte commune entre phrase question et nonquestion d'une même paire. Par exemple, pour une paire telle que :

- Phrase question : **Hôm nay là ngày** bao nhiêu ?
- Phrase nonquestion : **Hôm nay là ngày** ba mươi.

La partie commune est définie comme étant le texte **en gras.**

Nous retrouvons le même tendance que pour l'intensité de la phrase complète : dans 87% des cas, l'intensité moyenne d'une phrase question est plus forte que celle d'une phrase nonquestion. Dans les cas les plus forts, cette différence est en moyenne de 2 décibels, et dans les 13% des cas plus faibles restants, cette différence est en moyenne de 1,3 décibels.

Nous avons vérifié les statistiques par un test ANOVA (voir le Tableau 19) :

Locuteur :	DIEP (femme)	HUONG (femme)	LAN (femme)	KHOA (homme)	THANH (homme)	PHUONG (homme)
Moyen d'intensité - Question	62	58	60	65	61	59
Moyen d'intensité - NonQuestion	60	55	60	63	60	58
F	39.15	27.86	2.56	13.59	9.37	11.90
Valeur critique pour F	3.91	3.91	3.91	3.91	3.91	3.91

Tableau 19 : Résumé de résultat d'analyse ANOVA sur intensité moyenne pour les six locuteurs

Il semble confirmé alors que l'intensité des phrases questions est significativement supérieure à celle des phrases nonquestions.

Ensuite, par une analyse croisée avec le registre de phrase, nous trouvons que, parallèlement, un registre plus haut s'accompagne généralement d'une intensité plus forte dans 81% des phrases de notre corpus. Ce phénomène peut être dû à deux possibilités : (1) d'une part, il est lié à la mécanique du système de production de parole humaine : quand on veut produire un son à une fréquence supérieure du registre haut, on est obligé d'utiliser plus d'énergie que si c'était un son à une basse fréquence du registre bas et ceci implique qu'une intensité plus forte est obtenue avec un registre haut. (2) D'autre part, il est lié à une caractéristique particulière de la langue vietnamienne : la F0 étant déjà utilisée pour coder le ton de syllabe, il est donc plus difficile d'utiliser encore la F0 pour véhiculer un autre type d'information telle que le type de phrase. Le locuteur semble alors obligé d'utiliser un autre canal pour coder cette information. C'est là qu'une intensité plus forte intervient.

3.2.2.3.3 Analyse de la durée

Certain travaux de la littérature [Le, 1989 ; Nguyen-Thi, 2004] ont suggéré un débit plus rapide pour les phrases interrogatives. Nous avons mesuré et fait une étude statistique sur la durée de la partie commune du texte. Une durée plus courte de cette partie commune équivaut à un débit plus rapide. Nous trouvons que, dans une même paire, la durée de la partie commune pour une phrase question est plus courte que cette même partie commune dans le cas de la phrase nonquestion dans 63% des cas. Dans les cas où la phrase question est prononcée plus rapidement que la phrase nonquestion, la différence de durée est en moyenne de 7%.

Cependant, l'analyse ANOVA sur les deux ensembles de données (données des durées des phrases questions et des phrases nonquestions) montre que ces deux ensembles ne sont pas statistiquement différents : la valeur du rapport « F » est bien inférieure à celle de la « Valeur critique pour F » comme illustré dans le Tableau 20 :

Locuteur :	DIEP	HUONG	LAN	KHOA	THANH	PHUONG
Moyen de durée de texte commun - Question	616	559	531	628	584	508
Moyen de durée de texte commun - NonQuestion	646	565	530	656	604	532
F	1.76	0.06	0.01	1.07	1.12	1.74
Valeur critique pour F	3.91	3.91	3.91	3.91	3.91	3.91

Tableau 20 : Résumé de résultat d'analyse ANOVA sur la durée moyen de la partie de texte en commune pour les six locuteurs

Ainsi, nous pouvons conclure que l'effet de la durée (ou du débit de parole) n'est pas significatif entre les phrases nonquestions et questions pour les locuteurs de notre corpus.

3.2.3. Perception de la prosodie des phrases questions et nonquestions

3.2.3.1 Organisation du test de perception

Nous souhaitons vérifier que les différences détectées dans notre analyse sont effectivement perçues comme un moyen pour l'auditeur de différencier les phrases questions des phrases nonquestions, ou, en d'autres termes, que la prosodie de la phrase, malgré sa complexité due à la présence des tons, véhicule des informations permettant à l'auditeur de faire cette classification sans avoir besoin de comprendre la signification lexicale et sémantique de la phrase.

A cette fin, notre méthodologie est la suivante :

- nous utilisons le même corpus décrit ci-dessus ;
- pour chaque phrase, après avoir extrait les paramètres prosodiques, nous utilisons ces paramètres pour synthétiser une pseudo-phrase dans laquelle toutes les syllabes sont remplacées par une voyelle unique /a/ ; nous éliminons ainsi la possibilité pour l'auditeur de reconnaître une question uniquement par la présence d'un mot ou d'une particule « interrogatif » ; nous reproduisons le plus fidèlement possible, non seulement le contour de l'intonation, mais encore la durée des segments voisés/non voisés et le contour de l'intensité ;
- après avoir construit correctement ces paires de pseudo-phrases, nous les faisons écouter aux auditeurs, puis nous leur demandons de déterminer si la pseudo-phrase synthétisée entendue est question ou nonquestion ; afin de garantir un équilibre du test, nous avons synthétisé les phrases avec une voix d'homme et une voix de femme.

3.2.3.2 Préparation du corpus : synthèse des pseudo-phrases

Nous préparons ainsi 13 paires de pseudo-phrases. Pour chaque phrase, deux versions sont disponibles : la voix synthétisée d'un homme et la voix synthétisée d'une femme.

La méthode de synthèse est la suivante :

- à partir de 13 paires de phrases source du corpus, l'extraction des paramètres prosodiques tels que F0, intensité, durée est faite par Praat pour chaque phrase ;
- pour F0, la fenêtre d'analyse est fixée à 20ms, alors que pour l'intensité, la taille de fenêtre est de 5ms (afin d'obtenir un contour d'intensité plus fin de la phrase source) ;
- enfin, pour la durée, nous respectons l'équivalence de durée entre phrase source et phrase synthétisée par le fait que dans la phrase synthétisée, les zones correspondant à un signal non-voisé (consonne) sont remplacés par du silence, alors que les zones correspondant à un signal voisé (voyelle) sont remplacées par la voyelle unique /a/.

La synthèse de voix d'homme et de voix de femme se fait de la même manière. Pour la voix de femme par exemple, nous avons extrait deux périodes de signal de la voyelle /a/ de l'une des phrases prononcées par une locutrice de notre précédente étude.

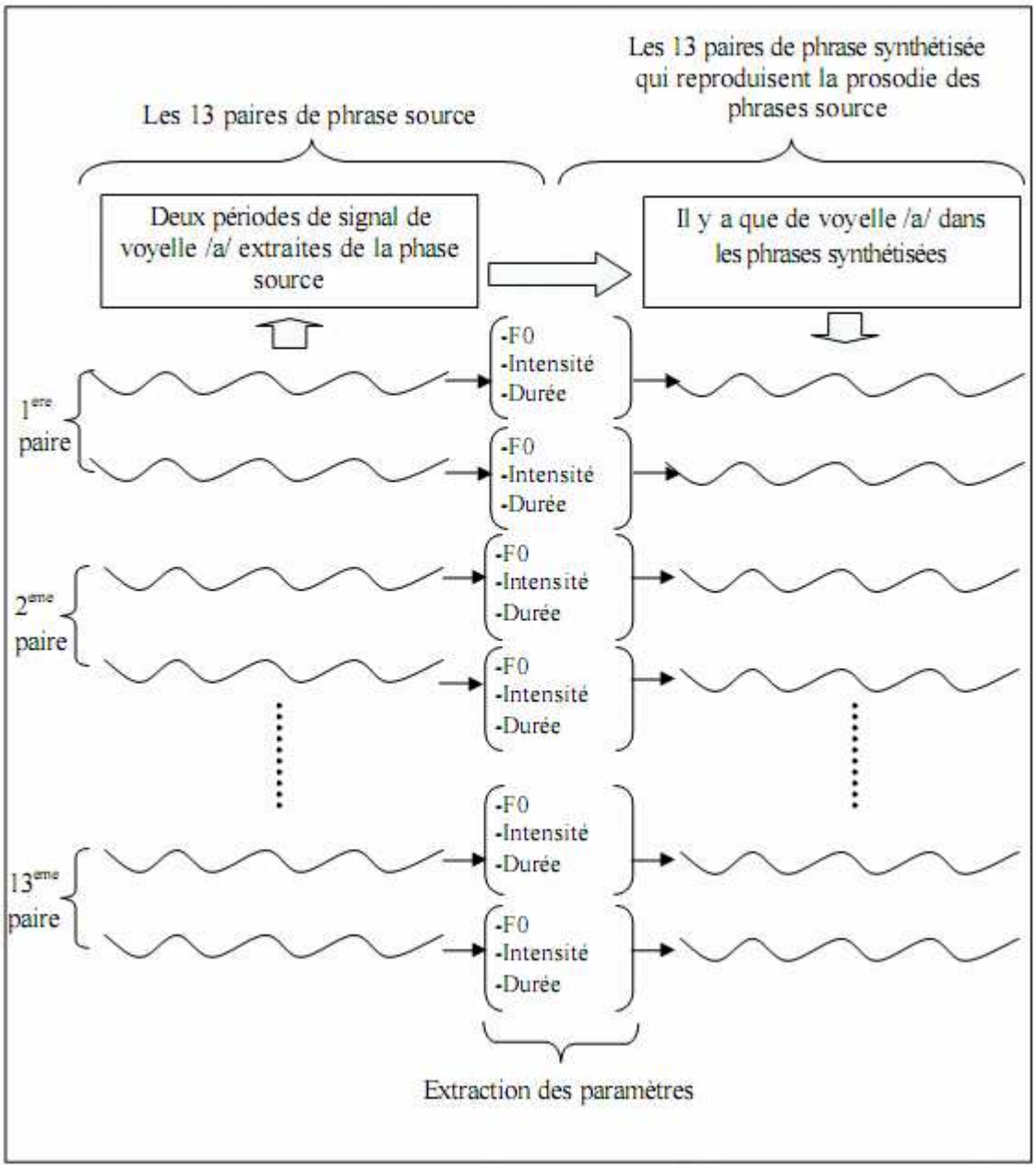


Figure 11: Schéma de méthode de reproduction des pseudo-phrases par synthèse

Pour la synthèse, la compatibilité entre l'échantillon de signal et les paramètres acoustiques utilisés pendant la synthèse joue un rôle très important sur la qualité de la parole synthétisée. En utilisant le signal de synthèse (deux périodes de la voyelle /a/) et les paramètres prosodiques (F0, intensité, durée) d'une même personne, nous pouvons obtenir une bonne qualité des phrases synthétisées qui seront utilisées dans les tests de perception ultérieurs.

L'algorithme TD-PSOLA est utilisé pour concaténer ces extraits de signal, tout en contrôlant le pitch (F0), l'énergie et la durée de chaque syllabe. L'image suivante montre un exemple de la

similarité obtenue entre la phrase source et sa version synthétisée correspondante (au niveau du spectrogramme et des courbes de pitch et d'intensité)

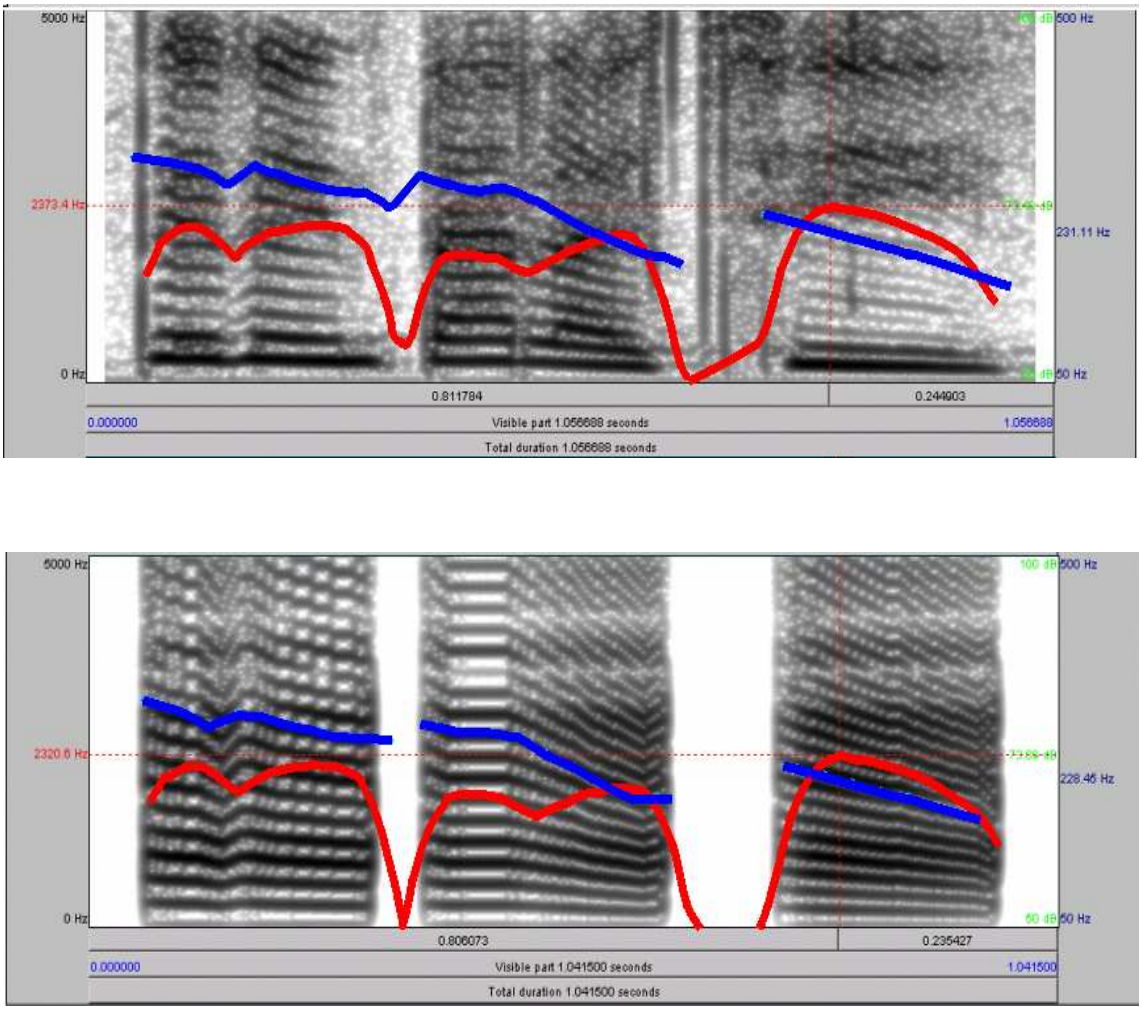


Figure 12 : Spectrogramme, contour de F0 (bleu) et contour d'énergie (rouge) du signal source (en haut) et du signal synthétisé correspondant (en bas). Phrase : «Tên anh ta là Trì » (en français : « Il s'appelle Trì »)

Ce corpus de 13 paires de pseudo-phrases synthétiques est ensuite utilisé dans le test de perception dont le résultat est présenté dans le paragraphe suivant.

3.2.3.3 Résultats du test de perception

Six auditeurs (3 hommes et 3 femmes) participent à notre test de perception. Ils doivent choisir entre deux réponses « question » ou « nonquestion ». Chaque auditeur fait le test 10 fois (5 fois pour la voix d'homme, 5 fois pour la voix de femme), et pour chaque session, l'ordre des phrases qui sont proposées est aléatoire.

Ici, le taux de bonne reconnaissance est calculé de manière très classique :

$$Taux\_de\_reconnaissance\_correcte = \frac{Nombre\_de\_bonne\_réponses}{Nombre\_total\_d'exemples}$$

Équation 1: Formule de calcul de taux de reconnaissance correcte

Exemple : pour le taux de reconnaissance correcte de voix femme, nous avons au total 13paires x 2 (deux phrases par paire) x 5 (chaque auditeur fait le test cinq fois) x 6 (six auditeurs) = 780 exemples. Le nombre de réponses correctes (c'est-à-dire phrase question bien reconnue comme étant une phrase question, phrase nonquestion bien reconnue comme étant une phrase nonquestion) est de 538. Alors le taux de reconnaissance correcte globale des phrases questions et nonquestions sera 538/780= 69%.

Le résultat du test est présenté dans le Tableau 21. Le taux de bonne reconnaissance sur l'ensemble des phrases questions et nonquestions est d'environ 70% pour la voix de femme et d'environ 58% pour la voix d'homme (ligne 1 du tableau).

Taux / Voix synthétisée	Voix femme	Voix homme
1. Taux de reconnaissance globale	69%	58%
2. Taux de reconnaissance des phrases questions	74%	61%
3. Taux de reconnaissance des phrases nonquestions	63%	55%

Tableau 21 : Taux de reconnaissance des phrases synthétisées

Les lignes 2 et 3 présentent respectivement les taux de bonne classification des phrases questions et des phrases nonquestions : nous pouvons remarquer que les phrases questions semblent mieux reconnues (environ 74 % de bonnes réponses) que les phrases nonquestions (seulement 63%) pour la voix de femme. Les phrases de la voix d'homme semblent bien moins reconnues que voix de femme (61% de bonnes réponses pour phrases questions et 55% de bonnes réponses pour phases nonquestions).

Les figures suivantes détaillent les résultats pour les 13 paires de phrase question/ nonquestion pour les voix de femme et d'homme respectivement.

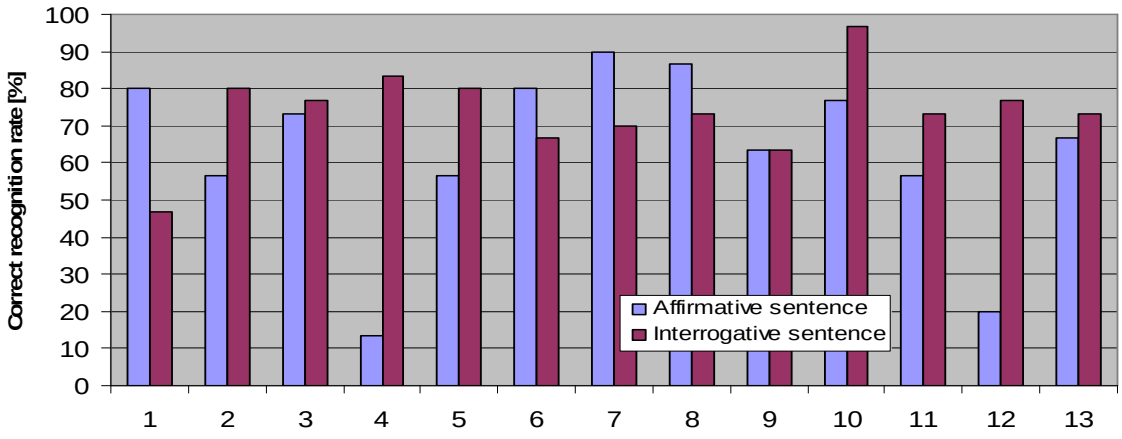


Figure 13: Taux de reconnaissance correcte des phrases synthétisées de voix de femme

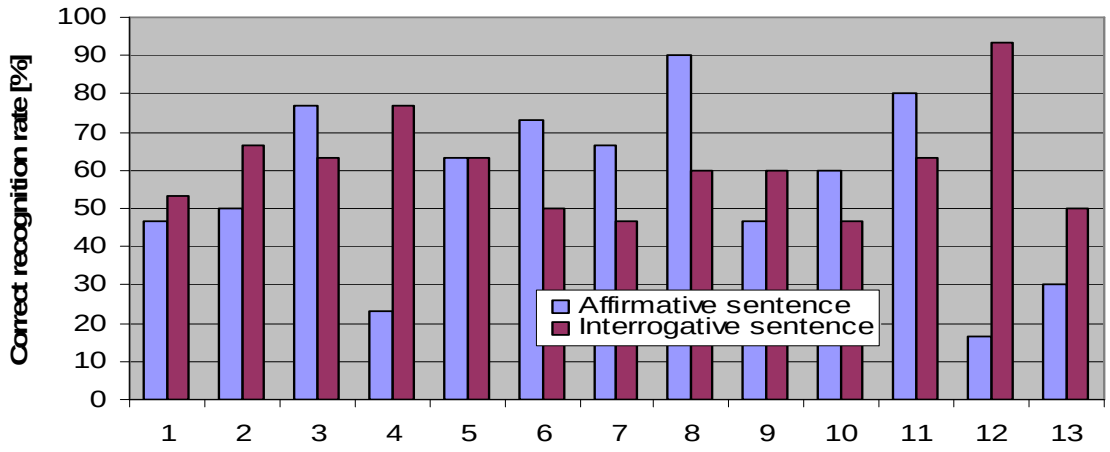


Figure 14 : Taux de reconnaissance correcte des phrases synthétisées de voix d'homme

Une analyse commune des deux tableaux montre que pour les paires 4 et 12, la phrase nonquestion est très mal reconnue (respectivement 12% et 20%), et c'est le cas pour les deux types de voix (voir la Figure 13 et la Figure 14).

Nous trouvons que ce résultat de test de perception est corrélé avec le résultat d'analyse prosodique. Nous remarquons que :

- l'auditeur semble juger une phrase comme étant question, si elle présente les caractéristiques suivantes :

- Une intonation croissante en fin de phrase
- Un registre plus haut
- Le dernier ton en registre haut : ton 1NGANG, 3NGA, 5SAC
- Une intensité plus forte

- l'auditeur semble juger une phrase comme étant nonquestion, si elle présente les caractéristiques suivantes :

- Une intonation descendante en fin de phrase
- Un registre plus bas
- Le dernier ton en registre bas : ton 2HUYEN, 4HOI, 6NANG
- Une intensité plus faible

Cette hypothèse semble être valable pour expliquer le cas des paires 4, 12 et 13 où le taux de reconnaissance des phrases type question est beaucoup plus élevé que celui des phrases type nonquestion. Pour ces paires, les phrases présentent toutes une dernière syllabe possédant le ton 5 montant, qui fait croître la partie finale du contour intonatif de la phrase, tant pour les interrogations que pour les affirmations. L'effet d'intensité y intervient également : ces phrases possèdent les derniers mots à ton en registre haut (tons 1NGANG, 3NGA, 5SAC), elles sont donc accompagnées par une intensité plus forte que les autres phrases (comme justifié dans la section d'analyse).

	Phrases	Taux. voix femme	Taux. voix homme
Paire 4	Hôm nay là ngày bao nhiêu thế ?	83%	67%
	Hôm nay là ngày ba mươi đấy	13%	50%
Paire 12	Anh ăn cơm không thế ?	77%	93%
	Anh ăn cơm Không Thế	20%	17%
Paire 13	Em ăn bánh nhé ?	73%	50%
	Em ăn bánh ché	67%	30%

Le cas inverse se retrouve pour les paires 6, 7, 8 où le taux « nonquestion » est plus élevé que le taux « question ». Ces paires possèdent les derniers mots à tons en registre bas (2HUYEN et 6NANG) qui sont de contour descendant, ce qui fait descendre la partie finale du contour intonatif de la phrase, tant pour les interrogations que pour les affirmations. Parce que ces tons sont en registre bas, ils ont une intensité plus faible comme prouvée dans le paragraphe 3.2.2.3.2.

	Phrases	Taux. voix femme	Taux. Voix homme
Paire 6	Tên anh ta là gì ?	67%	50%
	Tên anh ta là Trì	80%	73%
Paire 7	Tên anh ta là gì rồi ?	70%	47%
	Tên anh ta là Trì rồi	90%	67%
Paire 8	Tên anh ta là gì vậy ?	73%	60%
	Tên anh ta là Kì Cậy	87%	90%

Nous avons aussi remarqué que le taux de bonne reconnaissance est supérieur pour la voix de femme que pour la voix d'homme. Dans la mesure où notre test est fait sur seulement deux locuteurs (un homme et une femme), les raisons que nous pouvons formuler sont les suivantes : premièrement, pour la voix synthétisée féminine, la variation mélodique est plus importante que pour la voix synthétisée masculine La locutrice féminine joue plus sur l'intonation quand elle parle, l'information intonative est donc plus nettement introduite dans les signaux de parole. Pour le locuteur masculin qui utilise moins l'intonation lorsqu'il parle, le signal de parole contient moins d'information intonative - ce qui cause plus de difficultés lors de la perception et détermination de type de phrase. Pour comparer la variation, nous avons mesuré la moyenne des écart-types de F0 des phrases synthétisées en demi-tons[8]. Cette valeur moyenne est de 2,4 demi-tons pour la voix féminine et seulement 1,7 demi-tons pour la voix masculine.

Type de phrase	Type de voix synthétisée	Moyenne des écart-types de F0 des phrases synthétisées en demi-tons
Question	Voix féminine	2,6
	Voix masculine	1,7
Non question	Voix féminine	2,3
	Voix masculine	1,7

Tableau 22 : Moyenne par type de phrase des écart-types de F0 des phrases synthétisées en demi-ton

[8] Transformer de fréquence acoustique (Hz) en demi-ton relative à 100Hz par la formule :
demi-ton = 12 ln (x / 100) / ln 2

Le Tableau 22 présente en détail la moyenne par type de phrase : il montre que la variation mélodique de la voix féminine est plus importante que celle de la voix masculine dans tous les cas.

Deuxièmement, comme le mécanisme de fonctionnement de l'oreille humaine est plus sensible en valeur relative qu'en valeur absolue, une même variation de 10% autour de 100Hz (F0 moyen d'homme) sera moins perçue que cette même variation autour de 250Hz (F0 moyen de femme).

3.2.4. Conclusion sur l'étude de la prosodie du vietnamien

Le fait que le taux de bonne reconnaissance global des phrases nonquestions et questions soit d'environ 70% (et que pour certaines d'entre elles, elles sont même reconnues à plus de 90%) montre que les paramètres prosodiques de la phrase vietnamienne transportent des informations extralinguistiques qui peuvent permettre à l'auditeur de discriminer le type de phrase. Nous avons limité notre étude à la discrimination « question/nonquestion », mais il serait intéressant à l'avenir de généraliser notre étude à tous les types de phrases (exclamatives, ordres, etc.).

Au niveau de la production, notre étude a permis de caractériser la prosodie des phrases simples de la langue vietnamienne (dialogue), en éliminant l'influence des tons : les différences entre questions et affirmations sont essentiellement une différence de pente de F0 (croissante ou décroissante) en fin de phrase (deuxième moitié de la dernière syllabe), à laquelle s'ajoutent l'effet de registre et de l'intensité. Cependant, pour notre étude, le changement de débit de parole ne semble pas effectif, ni significatif.

Au niveau de la perception, nous avons montré que, comme pour les langues non tonales, la prosodie de la phrase transporte des informations extralinguistiques sur la nature de la phrase, bien que celle-ci, à cause de la présence des tons lexicaux, ne soit pas toujours suffisante. En effet, ces informations peuvent être brouillées par la modulation du contour prosodique par les tons lexicaux. L'utilisation de mots interrogatifs pour lever les ambiguïtés est donc nécessaire et logique.

Dans le chapitre suivant, nous allons essayer d'exploiter ces résultats d'analyse en proposant des paramètres permettant une détection automatique de questions pour des langues tonales et non tonales (vietnamien et français).

Chapitre 4. **Système de détection automatique de questions**

Dans ce chapitre, nous allons présenter notre système de détection de question avec les détails des paramètres du modèle prosodique et du modèle lexical. La technique d'arbre de décision est choisie pour le moteur de détection. Enfin, nous allons discuter la formule de mesure de la performance du système, ainsi que la méthode de sélection du meilleur jeu de paramètres.

4.1. Présentation du système

La Figure 15 présente une vue d'ensemble de notre système de détection automatique de questions :

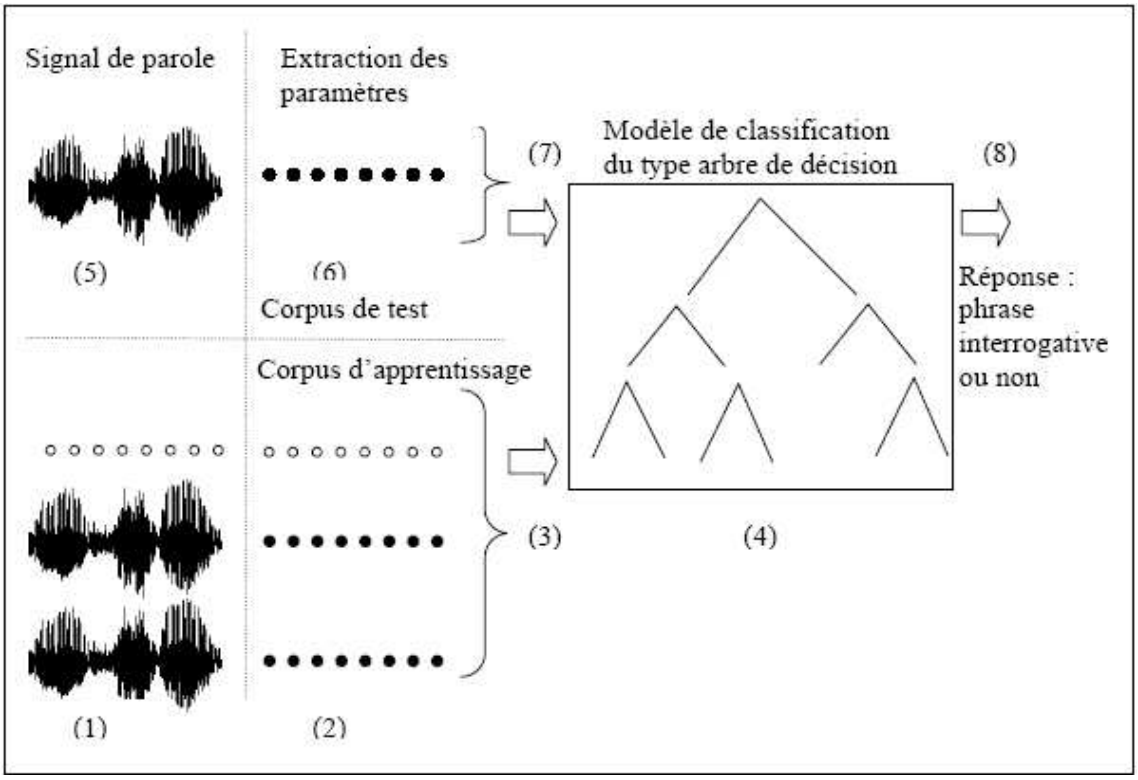


Figure 15 : Principe de classification

La première étape est l'apprentissage du système. Cela se fait grâce à un corpus d'apprentissage qui comprend un certain nombre d'enregistrements de phrases étiquetées (questions et non questions) (1). Ces signaux sont ensuite analysés et transformés en vecteurs de paramètres calculés et extraits à partir du signal de parole (2).Nous utilisons ces vecteurs de paramètres de phrases de corpus d'apprentissage pour entraîner un modèle de classification (3). Ce modèle est, dans notre cas, sous forme d'un arbre de décision (4) et sera décrit en détail dans le 4.4. Après avoir obtenu le modèle de classification, nous l'utilisons pour évaluer et classifier les phrases inconnues (5). Pour cette étape de test, la phrase en entrée est aussi transformée en un vecteur de paramètres (6). Celui-ci est alors évalué par le modèle de classification (7), le modèle va enfin classifier la phrase en classe interrogative ou non (8).

Il est important dans ce processus de trouver les bons paramètres pour représenter une phrase. Si les paramètres sont pertinents, les phrases de deux classes différentes (question et nonquestion dans notre cas) seront mieux distinguées. Dans nos travaux de recherche, nous nous

intéresserons tout particulièrement aux paramètres prosodiques issus du signal de parole : la fréquence fondamentale (F0 ou intonation), l'intensité. En effet, comme montré dans le chapitre 3, la variation de l'intonation semble être l'une des caractéristiques principales permettant de différencier les phrases interrogatives des autres types de phrase.

Notre but est essentiellement de détecter le type de phrase en analysant directement le signal de parole en terme d'utilisation des paramètres prosodiques. Cependant, nous sommes conscients que la prosodie n'est pas le seul indice pour distinguer la nature de la phrase (comme discuté dans le 3.1). C'est pourquoi nous nous sommes aussi investis dans l'étude d'un autre indice qui est l'indice lexical, c'est-à-dire la représentation lexicale (textuelle) de la phrase. Dans ce cas là, nous supposons que le signal de parole a été transcrit par un moteur de reconnaissance automatique de parole (RAP) et que le résultat de la reconnaissance est disponible. Alors, pour exploiter l'indice lexical, nous pouvons utiliser le même schéma et la même architecture du système illustré dans la Figure 15, sauf que les paramètres ne sont plus ceux extraits de la prosodie, mais seront ceux extraits à partir du texte de la phrase. L'exploitation de l'indice lexical a été beaucoup utilisée pour la tâche de segmentation en thème et la classification d'actes de dialogue. Les résultats du projet de recherche sur le dialogue homme-homme adapté à la tâche en langue allemande [Mast, 1996] indiquent qu'une source primaire de la connaissance pour la classification d'acte de dialogue est constituée par les mots (mots vrais, ou résultat d'un système de reconnaissance de la parole). Beaucoup d'actes de dialogue peuvent être distingués largement en utilisant un modèle statistique de langue qui enregistre des distributions de probabilité de mot pour chaque type d'acte.

Nous allons détailler les paramètres, qui sont l'essentiels de notre étude, plus loin dans ce manuscrit.

4.2. Modèle prosodique

Le modèle prosodique correspond au block (4) de la Figure 15. C'est le modèle de classification qui a pour but de modéliser l'évolution de la prosodie d'une phrase pour la tâche de détection automatique de question. La prosodie est représentée par un ensemble de paramètres. Nous allons d'abord faire le point sur les paramètres déjà proposés dans les travaux reportés dans la littérature, ensuite nous allons présenter notre propre jeu de paramètres.

4.2.1. Paramètres proposés dans la littérature

Certains travaux dans la littérature ont proposé un grand nombre de paramètres prosodiques. Ces travaux portent essentiellement sur le domaine de la modélisation et de la classification des actes de dialogue [Shriberg, 1998 ; Stolcke, 2000 ; Zechner, 2001]. Cependant tous ces travaux sont dédiés à la langue anglaise. Ils exploitent à la fois l'indice lexical et l'indice prosodique.

Les paramètres (ou « features » en anglais) prosodiques utilisés dans ces travaux sont très nombreux et très complexes. Ils représentent, comme dans [Shriberg, 1998], un nombre total de

59 paramètres issus de F0 (29 paramètres), de l'intensité (10 paramètres), de la durée (9 paramètres), des pauses (5 paramètres), du débit de parole (5 paramètres) et du genre du locuteur (1 paramètre).

D'autres travaux dans le domaine de la segmentation et de la classification de l'audio en signaux de parole, de musique, de sons en provenance de l'environnement et/ou du silence [Lu, 2001], de la détection de fin d'un énoncé [Ferrer, 2003], et de la segmentation en phrases [Wang, 2003] utilisent eux aussi des vecteurs sophistiqués de paramètres acoustiques tels que le taux de passage par zéro (zero-crossing rate ZCR), le rapport d'énergie à court terme (low short-time energy ratio LSTER) et/ou le flux spectral.

Comme ces travaux reportés dans la littérature, nous avons également proposé et utilisé plusieurs paramètres dont certains d'entre eux sont identiques à quelques paramètres déjà proposés, mais d'autres sont des paramètres originaux. En effet, en étudiant ces paramètres, nous nous intéressons - au delà de la détection automatique des questions - à relever le phénomène du brouillage de l'information prosodique pour une langue à ton comme le vietnamien.

4.2.2. Nos paramètres

Dans la langue française, la forme interrogative d'une phrase est en corrélation avec le contour d'intonation [[Fontaney, 1991 ; Hirst, 1998 ; Bessac, 1996]. C'est pourquoi nous avons décidé d'utiliser l'évolution de la fréquence fondamentale (F0) pour la tâche de détection de question dans un flux de signaux de parole. A partir de ce contour F0, un ensemble de nouveaux paramètres sont dérivés dans le but de modéliser la forme du contour F0. Ces paramètres sont présentés dans la section 4.2.2.1 suivante. Il est important de noter ici que, contrairement aux classiques paramètres à courte-terme utilisés dans la reconnaissance de parole, un seul vecteur des paramètres à long-terme est calculé automatiquement pour chaque phrase du corpus.

Après cette première étude sur la langue française, notre recherche élargie sur la langue vietnamienne a montré que la différence entre phrase question et phrase du vietnamien consiste principalement en : une différence de contour F0 à la fin de phrase, un registre plus haut et une intensité plus forte (présenté en détail dans le 3.2). Ces résultats nous ont permis de proposer un nouveau jeu de paramètres qui est plus approprié pour modéliser ces différences en vietnamien que les paramètres initialement proposés pour le français. Ces nouveaux paramètres sont détaillés dans la section 4.2.2.2 suivante.

La relation entre ces deux jeux de paramètres peut être illustrée dans la Figure 16 suivante :

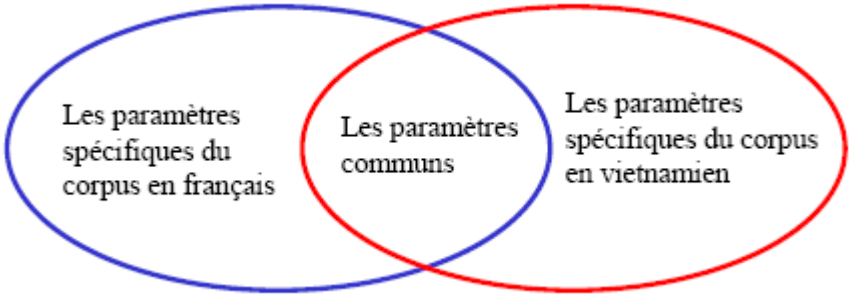


Figure 16 : Relation entre les deux jeux de paramètres

Les paramètres en commun des deux jeux sont des statistiques sur la distribution de F0 tels que min, max, moyen… alors que les autres paramètres sont spécifiques à chaque langue. Ces paramètres spécifiques ont pour but de mieux modéliser les caractéristiques spécifiques de chaque langue.

4.2.2.1 Le premier jeu de paramètres développé pour le corpus en langue française

Nous utilisons ici uniquement la courbe d'intonation F0 qui est calculée directement à partir du signal de chaque phrase, en découpant celui-ci en fenêtres de 20ms. Ensuite, à partir de cette courbe de F0, d'autres paramètres peuvent être dérivés et nous proposons un ensemble de 12 paramètres listés dans le Tableau 23.

No	Paramètre	Description
1	Min	Valeur minimale de F0
2	Max	Valeur maximale de F0
3	Range	Gamme de F0 pour la phrase entière (Max-Min)
4	Mean	Moyenne des valeurs de F0 d'une phrase
5	Median	Médiane des valeurs F0 d'une phrase
6	HighGreaterThanLow	Est-ce que la somme des valeurs F0 dans la première moitié de la phrase est supérieure à celle des valeurs F0 dans la dernière moitié ?
7	RaisingSum	Somme des $F0_{i+1} - F0_i$ si $F0_{i+1} > F0_i$

8	RaisingCount	Nombre de $F0_{i+1} > F0_i$
9	FallingSum	Somme des $F0_{i+1} - F0_i$ si $F0_{i+1} < F0_i$
10	FallingCount	Nombre de $F0_{i+1} < F0_i$
11	IsRaising	Est-ce que la forme F0 est montante ? (oui/non) Teste si RaisingSum > FallingSum
12	NonZeroFrameCount	Nombre de valeurs de F0 qui sont non nulles ?

Tableau 23 : Les 12 paramètres dérivés de F0

Nous pouvons remarquer que ces paramètres peuvent se diviser en deux catégories distinctes : les 5 premiers paramètres sont des statistiques sur la valeur de F0 (min, max, range…), alors que les paramètres restants caractérisent le contour (la forme) de l'évolution de F0 (contour montant ou descendant).

Les images suivantes montrent comment ces paramètres sont calculés à partir du contour intonatif d'une phrase (les points rouges représentent ce contour) :

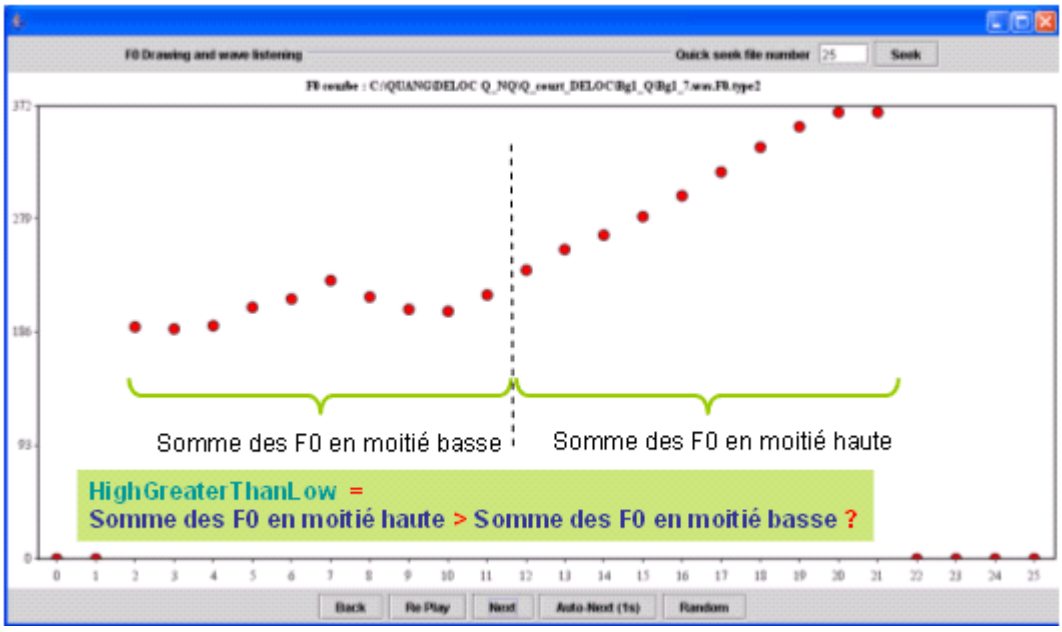


Figure 17 : Explication du paramètre HighGreaterThanLow

La Figure 17 explique la mesure du paramètre *HighGreaterThanLow* : il est la somme des valeurs de F0 de la moitié en fin *moins* la somme des valeurs de F0 de la moitié au début.

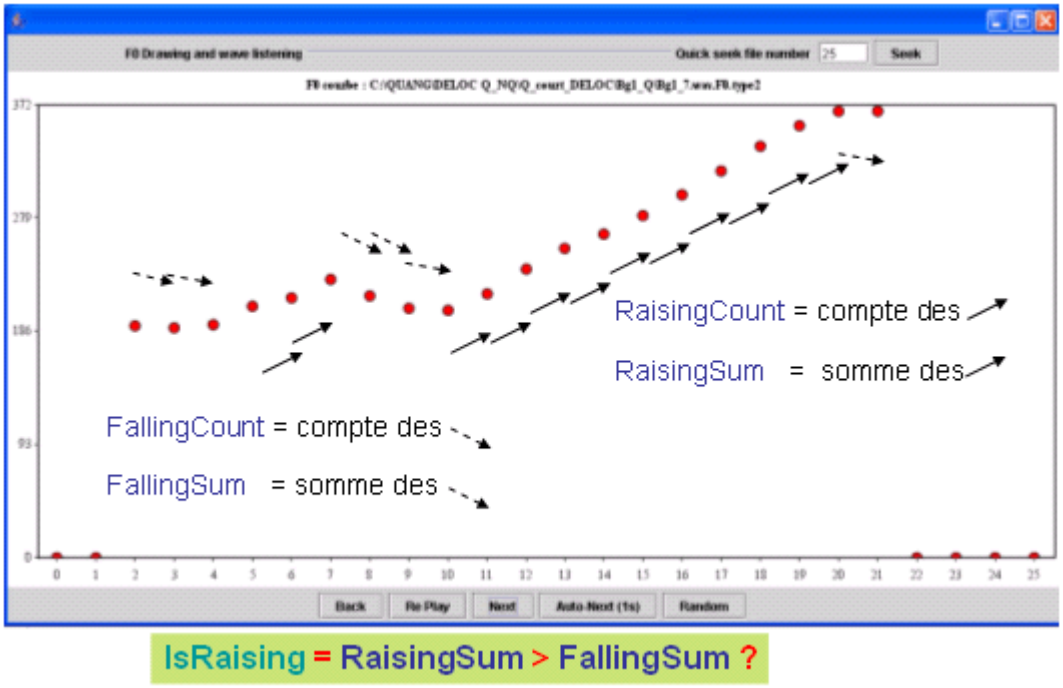


Figure 18 : Explication des paramètres IsRaising, FallingCount, FallingSum, RaisingCount, RaisingSum

Sur la Figure 18, nous voyons que l'extraction du paramètre *FallingCount* consiste à compter dans l'évolution du contour combien de fois le contour descend (valeur actuelle de F0 inférieure à la valeur F0 précédente), il correspond au nombre de flèches en pointillés. Le paramètre *FallingSum* a pour but de quantifier la grandeur des flèches en pointillé en terme de la différence (en hertz) entre le point du début et de la fin de chaque flèche. Ensuite, ce paramètre *FallingSum* fait la *somme mathématique* des grandeurs de toutes ces flèches. De cette façon, il peut fournir un nombre représentant combien de hertz le contour de F0 descend. Le même principe est appliqué pour les paramètres *RaisingCount* et *RaisingSum* qui sont représentés par les flèches en trait plein. Enfin, le paramètre *IsRaising* va comparer la valeur des deux paramètres *RaisingSum* et *FallingSum*. *IsRaising* aura la valeur « true » si la valeur de *RaisingSum* est supérieure à la valeur de *FallingSum*, et la valeur « false » dans le cas inverse. Par conséquent, il est clair que la valeur « true » signifie que le contour présente une allure montante, alors que la valeur « false » signifie que le contour est plutôt descendant.

Ces 6 derniers paramètres représentant la forme du contour de F0 sont originaux ; nous verrons plus tard dans la partie expérimentale leur efficacité pour une tâche de détection de questions.

4.2.2.2 Le deuxième jeu de paramètres développé pour le corpus en langue vietnamienne

Le résultat de l'analyse de la différence entre phrases questions et phrases nonquestions en langue vietnamienne dans le chapitre 3 nous a permis de proposer un nombre important de paramètres qui visent à capturer le mouvement du contour de F0. En plus, la condition

d'enregistrement dans laquelle notre corpus en langue vietnamienne a été réalisé (studio calme, distance microphone-bouche assurée constante) nous permet d'exploiter non seulement le F0, mais aussi d'autres paramètres tels que l'intensité et la durée.

Il y a au total 12 paramètres développés pour le corpus VietP en vietnamien. Le tableau suivant est la liste des ces paramètres.

No	Nom du paramètre	Description du paramètre
1	lastDemiSyllable-HighLevel	Niveau de la hauteur de la pente de la dernière moitié de la dernière syllabe
2	minF0	Valeur minimale de F0 de la phrase
3	maxF0	Valeur maximale de F0 de la phrase
4	moyenF0	F0 moyen de la phrase
5	rangeF0	Gamme de F0 de la phrase
6	moyenF0OfDebutSentence	F0 moyen de la première moitié du temps de la phrase
7	moyenF0OfFinSentence	F0 moyen de la dernière moitié du temps de la phrase
8	moyenF0OfFinMinusDebutHighLevel	[F0 moyen de la dernière moitié du temps de la phrase] moins [F0 moyen de la première moitié du temps de la phrase]
9	minIntensity	Valeur minimale de l'intensité de la phrase
10	maxIntensity	Valeur maximale de l'intensité de la phrase
11	moyenIntensity	Intensité moyenne de la phrase
12	rangeIntensity	Gamme de l'intensité de la phrase

Tableau 24 : Les 12 paramètres développés pour le corpus en langue vietnamienne

Par rapport à la Figure 16, les paramètres en commun avec le jeu de paramètres du français sont les 2, 3, 4 et 5, alors que les paramètres différents sont les restants. Les paramètres 9, 10, 11, 12 sont relatifs à l'intensité et ne sont pas présents dans le jeu de paramètres du français. Nous n'avons pas utilisé les paramètres comme raisingSum, raisingCount… pour le vietnamien parce que dans le résultat d'analyse de la production de la prosodie du vietnamien (présenté dans le

3.2.2), nous avons identifié que, au niveau de la fréquence fondamentale, la différence entre phrase question et phrase nonquestion du vietnamien consiste plutôt à la dernière partie de la dernière syllabe. Dans cette optique, les paramètres comme raisingSum, rasingCount... du français qui visent à capturer l'évolution du contour F0 en entier, ne sont pas adéquats pour le vietnamien. Cette présuposition est encore consolidée dans les expérimentations croisées ultérieures (qui seront présentées dans le 5.2.3).

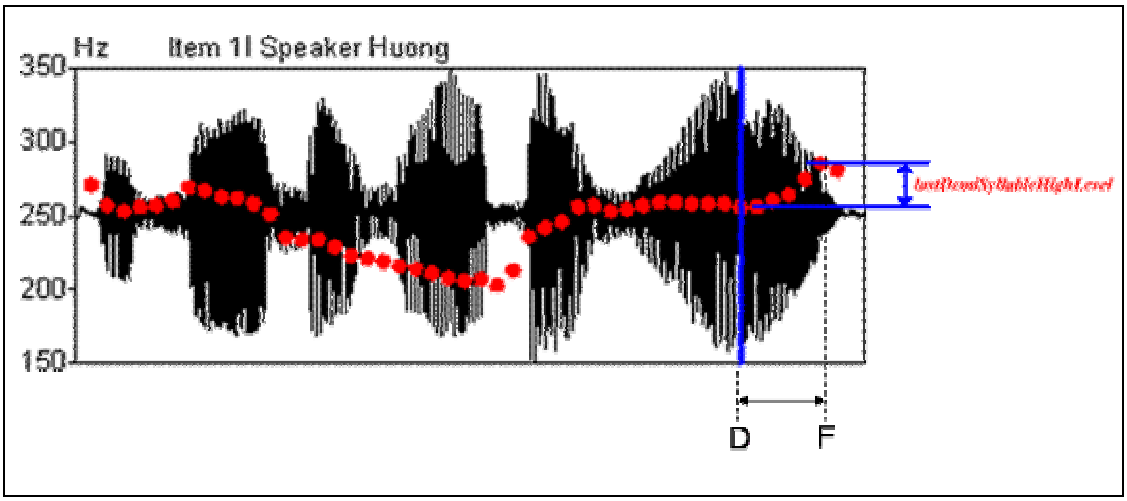


Figure 19 : Explication du paramètre lastDemiSyllableHighLevel

La Figure 19 explique la mesure du paramètre « lastDemiSyllableHighLevel » : ce paramètre vise à capturer le niveau de hauteur du contour F0 de la zone après la barre bleue verticale (qui est la zone de la dernière moitié de la dernière syllabe). En le mesurant en Hz, nous voulons quantifier par un chiffre cette hauteur : si ce contour de F0 est montant, ce paramètre a une valeur positive, alors que si le contour F0 est descendant, ce paramètre va avoir une valeur négative. En plus, la grandeur mathematique de ce paramètre est aussi proportionnelle à la pente du contour : plus ce contour est montant, plus cette grandeur est élevée et inversement, plus ce contour est descendant, plus cette grandeur sera basse.

Ce paramètre « lastDemiSyllableHighLevel » peut être extrait automatiquement sans avoir besoin d'une intervention manuelle. En effet, nous considérons que la longueur moyenne d'une syllabe est de l'ordre de 200 ms pour le vietnamien [Nguyen-Quoc, 2002], une demie-syllable est par conséquent de 100ms de longueur. La localisation de la dernière partie de la dernière syllabe d'une phrase revient alors à la localisation de la zone sonore de 100 ms à la fin de cette phrase. Afin de déterminer les frontières du début (point D) et de la fin (point F) de cette zone, nous allons commencer par déterminer le point F, le point D sera ensuite déterminé en utilisant l'écart de 100 ms avant ce point F. Parce que nous avons mesuré F0 par fenêtre de 20ms, cette zone D-F contient 5 valeurs de F0 et correspond à la dernière partie de la dernière syllabe de la phrase. Les 5 valeurs F0 seront utilisées pour calculer le paramètre « lastDemiSyllableHighLevel » qui est la différence entre le F0 du point F et le F0 du point D.

Le point F doit satisfaire ces deux conditions :

- La valeur de F0 de ce point n'est pas égale à zéro
- La valeur d'intensité de ce point doit être supérieure ou égale à 90% de la moyenne des valeurs d'intensité de 5 points qui le précèdent. Cette condition permet de garantir que la zone D-F trouvée est effectivement la zone sonore de 100ms à la fin de la phrase et n'est pas du silence.

Un algorithme pour déterminer automatiquement cette zone de la dernière partie de la dernière syllabe d'une phrase a été implémenté et incorporé dans notre système.

4.3. Modèle lexical

Dans l'optique de la recherche d'information et l'indexation de document audio, nous avons voulu tout d'abord développer un système de détection automatique de questions en utilisant seulement la prosodie de la parole. C'est pourquoi les modèles prosodiques ont été créés et présentés dans la section 4.2 précédente. Cependant, pour la tâche de détection de question, il est possible d'exploiter non seulement la prosodie, mais encore un autre indice qui est l'information lexicale vehiculée dans la parole. La transcription textuelle d'une phrase peut être obtenue automatiquement à la sortie d'un moteur de reconnaissance de parole. Elle est ensuite examinée pour chercher la présence éventuelle des termes interrogatifs et/ou expressions de demande dans le but de déterminer le type de phrase. Bien que la liste des termes et expressions interrogatives sont spécifiques à chaque langue, le même principe peut être appliqué pour les deux langues française et vietnamienne comme cela est présenté en détail dans les sections 4.3.1 et 4.3.2 suivantes

4.3.1. Modèle lexical développé pour le corpus en français

Le modèle lexical a pour but de modéliser la présence de termes ou d'expressions interrogatifs. La position de ces termes interrogatifs est aussi importante pour pouvoir déterminer si la phrase est de type interrogatif ou non. La présence et la position de termes interrogatifs sont des bons indices lexicaux pour la détection de question. Le terme « qui » par exemple, quand il se trouve au début de la phrase comme « Qui est le Président de la France ? », signale que la phrase est une question ; quand il se trouve au milieu de la phrase comme « C'est Mr Jacques Chirac qui est le président de la France », il n'est plus un terme interrogatif. C'est pourquoi au niveau des paramètres du modèle, nous avons créé les paramètres qui visent à capturer la présence ainsi que la position des mots interrogatifs dans la phrase. Les paramètres peuvent se diviser en 3 groupes en fonction de leur signification :

- 1er groupe : les paramètres dans ce groupe visent à détecter si certains termes interrogatifs suivants sont présentés au début de la phrase : "pourquoi" ; "qui" ; "quand" ; "pour quand" ; "comment" ; "combien" ; "pour combien" ; "de combien" ; "où" ; "quel" ; "quelle" ; "quels" ; "quelles" ; "de quel" ; "lequel" ; "laquelle" ; "lesquels" ;

"lesquelles" ; "jusqu'où". Il est fort probable que la phrase est une question si ces termes interrogatifs sont présents au début de la phrase.

- 2ème groupe : les paramètres dans ce groupe visent à détecter si certains termes et expressions de demande suivants sont présents dans la phrase : "je voudrais savoir" ; "j'aimerais savoir" ; "j'voudrais savoir" ; "j'aimerais vous demander" ; "je voudrais vous demander" ; "j'voudrais vous demander" ; "je voudrais vous d'mander" ; "j'voudrais vous d'mander" ; "est-ce que" ; "est-ce qu'il" ; "est-ce qu'elle" ; "est-ce qu'ils" ; "est-ce qu'elles" ; "qu'est-ce que" ; "qu'est-ce qu'il" ; "qu'est-ce qu'elle" ; "qu'est-ce qu'ils" ; "qu'est-ce qu'elles" ; "qu'est qui" ; si ces expressions de demande se présentent dans la phrase, alors la phrase a une forte probabilité d'être une question.

- 3ème groupe : les paramètres dans ce groupe visent à détecter si certains termes interrogatifs suivants sont présentés à la fin de la phrase : "n'est-ce pas" ; "pardon" ; "ah bon" ; "qui" ; "quand" ; "pour quand" ; "comment" ; "combien" ; "pour combien" ; "de" ; "combien" ; "où" ; "en quoi" ; "pourquoi" ; "allô" ; "c'est clair" ; "ça va" ; "c'est bon" ; "c'est tout bon" ; "c'est ça" ; "c'est bien ça" ; "non" ; "ou pas" ; "mhm" ; "hein" ; "d'accord" ; "alors" ; Lors de la présence de ces termes interrogatifs à la fin de la phrase, la phrase a une forte probabilité d'être une question.

No	Nom du paramètre	Description
1	OneWordBefore_pourquoi	Le mot avant le terme *pourquoi* dans la phrase
2	TwoWordBefore_pourquoi	Les deux mots avant le terme *pourquoi* dans la phrase
3	OneWordBefore_qui	Le mot avant le terme *qui* dans la phrase
…		

Tableau 25 : Exemple des paramètres lexicaux pour le corpus en langue française

Ces termes interrogatifs et ces expressions de demande ont été présentés en détail dans le 3.1. Ils comprennent les termes comme par exemple "Allô" ; "Mhm" ; "Hein"… qui sont trouvés dans notre corpus. Ensuite, la liste de ces termes a été complétée et enrichie avec des termes présentés dans la thèse de Natalie Colineau [Colineau, 1997]. Cette thèse, réalisée aussi au sein du laboratoire CLIPS, porte sur l'étude des marqueurs discursifs (dont les marques interrogatives)

dans le dialogue finalisé. Dans cette thèse, une liste exhaustive des termes interrogatifs est présentée.

Nous avons alors au total 71 paramètres lexicaux associés chacun à un terme interrogatif dans les 3 groupes. Quelques paramètres sont présentés dans le Tableau 25, la liste complète se trouve en annexe C.1.

Pour représenter la position du début ou de la fin de phrase, la phrase est mise entre deux balises spéciales : START et END. Cette étape de balisage est appliquée sur toutes les phrases avant l'entrée du système de classification.

Afin de clarifier la signification des paramètres, examinons les exemples suivants. Supposons qu'il y a un exemple d'une phrase « qui va jouer le rôle ? » qui est une question. Si la phrase est obtenue à la sortie d'un simple système de reconnaissance de parole, il y a donc pas de point d'interrogation « ? ». Après le balisage, nous avons :

START qui va jouer le rôle END

Dans l'étape d'extraction des paramètres pour ce cas, le paramètre « OneWordBefore_qui » aura la valeur de « START » qui signifie que le terme interrogatif « qui » est présent et se trouve au début de la phrase, les autres paramètres auront tous une valeur spéciale « N/A » abbréviation de « NotAvaillable » qui signifient que les autres termes interrogatifs correspondant à ces paramètres ne sont pas présentés dans cette phrase. De cette façon, les paramètres peuvent encoder à la fois la présence et la position d'un terme interrogatif dans la phrase.

Regardons maintenant un autre exemple d'une phrase nonquestion : « oui c'est vrai il y avait ça aussi », après le balisage, la phrase sera :

START oui c'est vrai il y avait ça aussi END

Pour cette phrase, aucun des 71 paramètres ci-dessus ne peut donner une valeur. Ils sont alors tous assignés à « N/A ». Cela veut dire qu'il y a aucun terme interrogatif, ni expression de demande présentés dans la phrase.

De cette manière, les paramètres du modèle lexical discutés ci-dessus peuvent satisfaire les objectifs fixés : modéliser les indices pertinents pour la reconnaissance d'une phrase question en capturant la présence et la position des termes interrogatifs et les expressions de demande si ces derniers sont présents dans la phrase. En raison d'un grand nombre de termes interrogatifs et/ou expressions de demande dont la langue française dispose, les paramètres lexicaux sont proportionellement nombreux (jusqu'à 71 paramètres) afin de pouvoir couvrir le maximum possible des formes de phrases de type question. La liste complète se trouve en annexe C.1.

Nous avons calculé ces paramètres lexicaux pour chaque phrase dans le corpus. Ces paramètres sont ensuite utilisés pour construire le modèle lexical qui est aussi sous forme d'un arbre de décision. Dans le 5.2.4, section 5.2.4.1 nous allons voir en détail l'efficacité de ce modèle lexical.

4.3.2. Modèle lexical développé pour le corpus en vietnamien

Le modèle lexical pour le corpus en vietnamien est construit suivant le même principe que le modèle lexical développé pour le corpus en français. La différence entre eux concerne seulement la liste des termes interrogatifs et des expressions de demande utilisées dans le modèle. Dans le cas de la langue française, ce sont les termes interrogatifs du français, dans le cas du modèle lexical de la langue vietnamienne, ces termes sont bien évidemment spécifiques à la langue vietnamienne.

Les paramètres lexicaux du vietnamien peuvent se diviser en 3 groupes en fonction de leur signification :

- 1er groupe : les paramètres dans ce groupe visent à détecter si certains termes interrogatifs suivants sont présentés au début de la phrase (entre parenthèse est donnée la traduction en français) : "sao (pourquoi)" ; "ai (qui)" ; "bao giờ (quand)" ; "thế nào (comment)" ; "bao nhiêu (combien)" ; "ở đâu (où)"… Il est fort probable que la phrase est une question si ces termes interrogatifs sont présents au début de la phrase.

- 2ème groupe : les paramètres dans ce groupe visent à détecter si certains termes et expressions de démande suivants sont présents dans la phrase : "tôi muốn biết (je voudrais savoir)" ; "tôi muốn hỏi (je voudrais demander)" ; "Bác có biết (savez-vous que) " ; "sao không (pourquoi pas) "… Si ces expressions de demande se présentent dans la phrase, alors la phrase est une question.

- 3ème groupe : les paramètres dans ce groupe visent à détecter si certains termes interrogatifs suivants sont présentés à la fin de la phrase : "không (n'est-ce pas)" ; "phải không (n'est-ce pas)" ; "gì (quoi)" ; "thế sao (ah bon)" ; "thế nào (ah bon)" ; "thì sao (et alors)" ; "à, á, hả, nhé, chi, nhỉ, chưa[9]"…; Lors de la présence de ces termes interrogatifs à la fin de la phrase, la phrase a une forte probabilité d'être une question

Dû à un grand nombre de termes interrogatifs et/ou expressions de demande dans la langue vietnamienne, les paramètres lexicaux sont proportionellement nombreux (jusqu'à 83 paramètres) afin de pouvoir couvrir le maximum possible des formes de phrases de type question. La liste complète se trouve en annexe C.2. Dans le 5.2.4, section 5.2.4.2 nous allons voir en détail l'efficacité de ce modèle lexical.

[9] Ces termes interrogatifs du vietnamien n'ont pas une traduction correspondante en français

4.4. Arbre de décision

Généralement, les techniques utilisées dans le domaine de Traitement Automatique de parole comprennent deux catégories principales : (1) les méthodes statistiques telles que les modèles de Markov cachés (HMM) ou les modèles de mélanges de Gaussiennes (GMM) et leurs variantes, et (2) les méthodes d'apprentissage automatique (machine learning). L'arbre de décision est une méthode classique d'apprentissage automatique [Breiman, 1984 ; Witten, 1999] et peut être vu comme la représentation graphique d'une procédure de classification. Cette procédure de classification est interprêtée en terme de règles de décision.

Pour certains domaines d'application, il est souhaitable de produire des procédures de classification compréhensibles par l'utilisateur. Les arbres de décision répondent à cette contrainte car ils représentent graphiquement un ensemble de règles et sont aisément interprétables. Dans notre application de détection automatique de question par exemple, en utilisant un arbre de décision comme classificateur, il est possible de comprendre, à travers les nœuds de l'arbre, les caractéristiques prosodiques qui influent sur le processus de classification. Pour les arbres de grande taille, la procédure globale peut être difficile à appréhender, cependant, la classification d'un élément particulier est toujours compréhensible. Les algorithmes d'apprentissage par arbres de décision sont efficaces, et disponibles dans la plupart des environnements de fouille de données.

La technique des arbres de décision est fondée sur le concept « *diviser-et-conquérir* ». Un nœud dans l'arbre consiste à tester une condition particulière qui, en général, compare la valeur d'un attribut avec une constante, ou compare ensemble deux attributs, ou on utilise des fonctions mathématiques d'un ou plusieurs attributs. La feuille de l'arbre donne une classification des éléments satisfaisant toutes les conditions menant à cette feuille.

Un arbre de décision parfait est un arbre de décision dans lequel tous les exemples de l'ensemble des données d'apprentissage sont correctement classifiés. Un tel arbre n'existe pas toujours (s'il existe deux exemples tels que à deux descriptions identiques correspondent deux classes différentes). L'objectif est de construire un arbre présentant une erreur de classification la plus petite possible. L'idée centrale pour la construction d'arbres est la suivante : diviser récursivement et le plus efficacement possible les éléments de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant tous à une même classe.

Les méthodes d'apprentissage d'arbre vont différer par les choix effectués pour ces différents opérateurs, c'est-à-dire sur le choix d'un test et le critère d'arrêt (c'est-à-dire quand arrêter la croissance de l'arbre, quand décider si un noeud est terminal). Néanmoins, dans toutes les méthodes, on trouve les trois opérateurs suivants :

- sélectionner un test à associer à un nœud.

- décider si un noeud est terminal : c'est-à-dire décider si un noeud doit être étiqueté comme une feuille ; par exemple les critères de décision pourront être : tous les exemples sont dans la même classe, il y a moins d'un certain nombre d'erreurs, ...
- affecter une classe à une feuille.

Pendant la construction d'arbres, l'idéal serait de trouver un critère qui permette d'arrêter la croissance de l'arbre au bon moment - ce qui n'est pas toujours évident. De plus, le risque d'arrêter trop tôt la croissance de l'arbre est plus important que de l'arrêter trop tard. Par conséquent, les méthodes utilisées procèdent souvent en deux phases. La première phase correspond à l'expansion de l'arbre ; dans une seconde phase, on *élague* l'arbre obtenu (élaguer un arbre consiste à en supprimer certains sous-arbres) afin que l'arbre ne soit pas trop profondément divisé (on appelle cela « overfitting » en anglais). Les méthodes se distinguent donc les unes des autres par les choix des opérateurs, mais aussi par les méthodes d'élagage utilisées.

Pour classifier un élément inconnu, l'algorithme teste les attributs dans les nœuds jusqu'à ce qu'il atteigne une feuille. Là, cet élément est classifié selon la classe attribuée à la feuille. Nous avons utilisé l'algorithme de construction d'arbre C4.5 [Quinlan, 1993] dont l'implémentation provient du logiciel *open-source* Weka[10] qui comprend les algorithmes de *classification*, *régression*, *clustering*, *règles d'association* écrits en Java. Plus d'informations sur Weka se trouvent en annexe D.2.

4.5. Mesures de performance

Notre système de détection de phrases interrogatives peut également être vu comme un système de recherche d'informations : la base de données est le corpus, alors que les objets de recherche sont des phrases de type question. Rappelons qu'un système de recherche d'informations (SRI) est défini comme un système permettant l'accès par le contenu à des documents satisfaisant les besoins d'informations des utilisateurs. Un système de recherche d'informations réalise une liaison informatisée entre un utilisateur humain, ayant le besoin d'une information et un ensemble de documents susceptibles de contenir cette information.

Dans la théorie, on utilise généralement les deux critères suivants pour évaluer un système de recherche d'informations : le rappel (*recall* en anglais) et la précision (*precision* en anglais) :

[10] http://www.cs.waikato.ac.nz/~ml/weka/

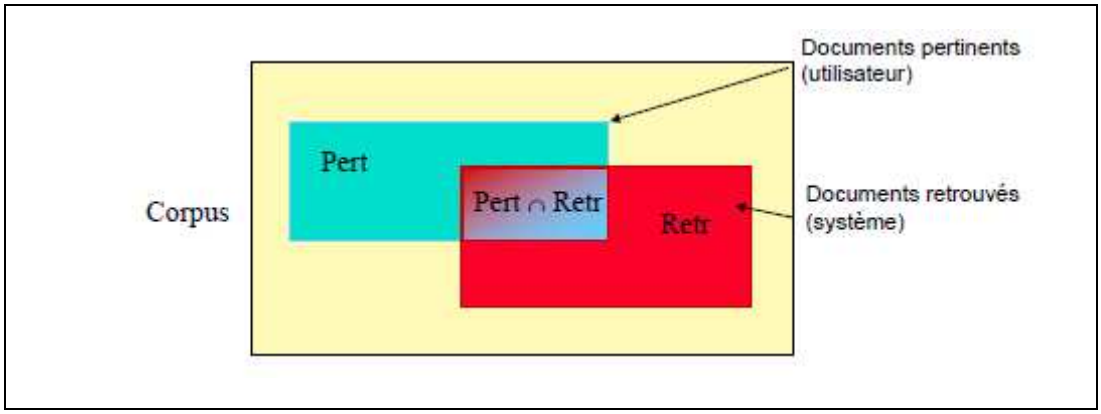


Figure 20 : Résultat d'un système de recherche d'informations : le jeu des documents pertinents pour l'utilisateur et le jeu des documents retrouvés par le système.

Dans le schéma ci-dessus, pour une requête donnée Q:

- "Pert" représente le sous-ensemble des documents effectivement pertinents
- "Ret" représente le sous-ensemble des documents retrouvés par le système

On définit les notions « rappel » et « précision » de la façon suivante:

- Le « rappel » pour une requête Q est défini par le rapport entre le nombre de documents pertinents retrouvés, et le nombre de documents pertinents. Le « rappel » est donc un réel entre 0 et 1: si le rappel est proche de 0, une très faible proportion de documents pertinents a été retrouvée; s'il est proche de 1, pratiquement tous les documents pertinents ont été retrouvés. Le rappel mesure donc la capacité du système à retrouver TOUS les documents pertinents
- La « précision » pour une requête Q est définie par le rapport entre le nombre de documents pertinents retrouvés, et le nombre de documents retrouvés. La précision est donc un réel entre 0 et 1: si la précision est proche de 0, une très faible proportion de documents pertinents figure dans la réponse; s'il est proche de 1, pratiquement tous les documents retrouvés sont pertinents. La précision mesure la capacité du système à ne retrouver QUE les documents pertinents

Dans notre système, nous considérons la classification en question/non question comme une requête de recherche des phrases question. Par conséquent, nous définissons les notions « précision » et « rappel » comme :

$$précision\_de\_classe\_Question = \frac{Nombre\_de\_Questions\_correctes}{Nombre\_de\_Questions\_trouvées}$$

$$rappel\_de\_classe\_Question = \frac{Nombre\_de\_Questions\_correctes}{Nombre\_total\_de\_Questions}$$

L'indice F_mesure est un compromis entre la « précision » et le « rappel ». :

$$F\_mesure = \frac{2*précision\_Question*rappel\_Question}{précision\_Question + rappel\_Question}$$

Le F_mesure est aussi un réel entre 0 et 1. Il est largement utilisé comme un indice sur la performance d'un système de recherche d'information. Si le F_ratio est proche de 0, le système est faible en performance ; et contrairement, s'il est proche de 1, le système a une bonne performance dans le sens où pratiquement toutes les phrases « question » ont été retrouvés et qu'elles sont des vraies phrases interrogatives.

4.6. Sélection du meilleur jeu de paramètres par la méthode « leave-one-out »

Dans notre travail, nous cherchons également à comprendre quels paramètres sont plus importants que d'autres dans la classification. Pour trouver l'ordre d'importance des paramètres en langue française et en langue vietnamienne, nous avons appliqué la méthode « leave-one-out » qui est bien connue dans le domaine de la classification automatique. Il s'agit d'un algorithme qui permet, après son exécution, de classer les paramètres par ordre d'importance en fonction de leur contribution au processus de classification (comme fait dans le [Besacier, 1998]). Le schéma suivant montre le principe de cet algorithme :

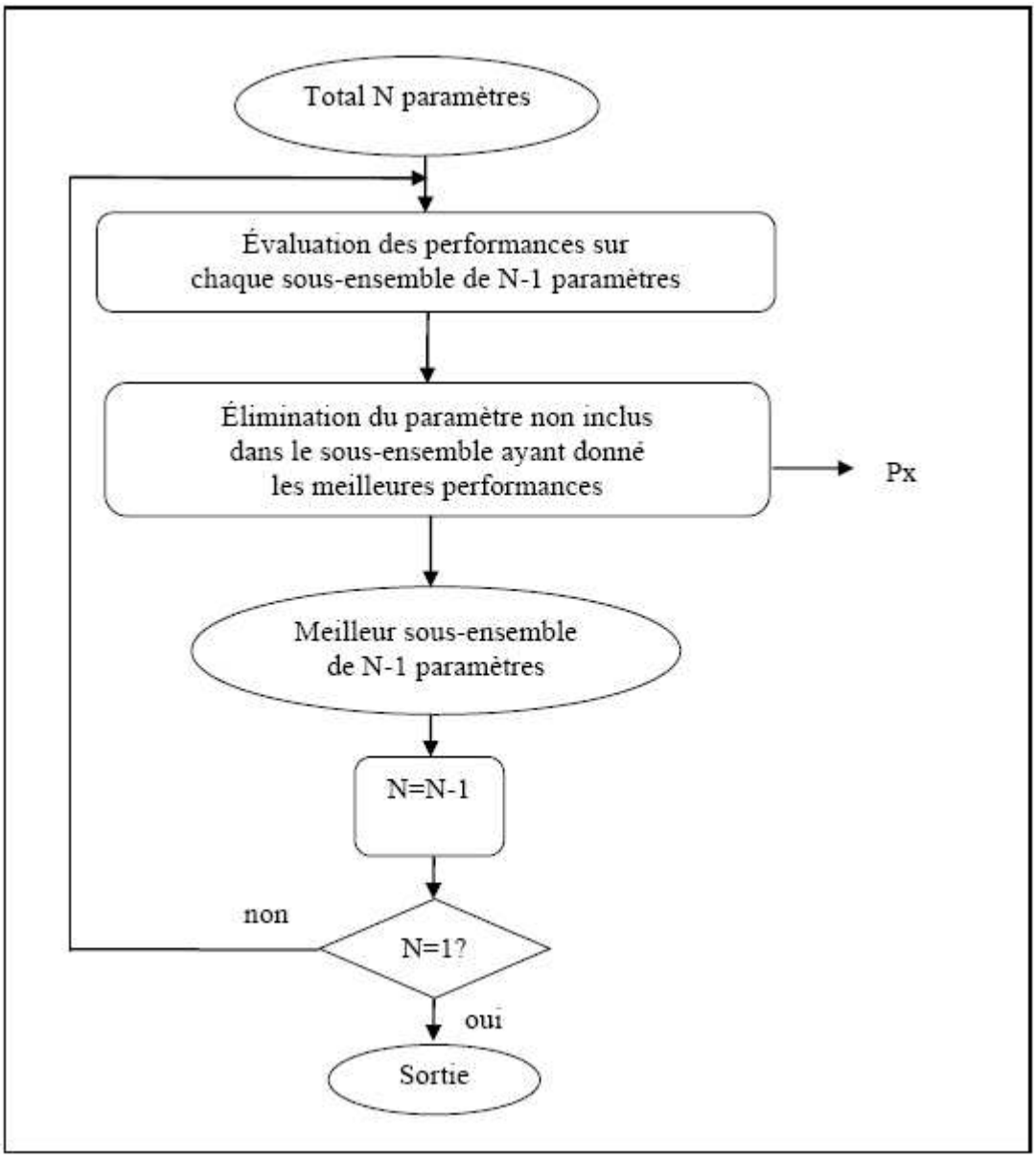


Figure 21 : Principe de l'algorithme « Leave-one-out »

La procédure commence par évaluer les performances d'identification de chacun des N sous-ensembles composés de N-1 paramètres. Le meilleur sous-ensemble (en terme de taux de classification) est alors déterminé et le paramètre qui n'est pas inclus dans ce sous-ensemble est défini comme le moins bon paramètre. Ce paramètre est alors mis de côté et la procédure recommence avec les N-1 sous-ensembles de N-2 paramètres restant. L'algorithme continue jusqu'à ce que tous les paramètres soient éliminés. La séquence inverse des paramètres supprimés nous donne alors une liste des paramètres classés du plus efficace au moins efficace pour cette tâche de classification

Chapitre 5.
Expérimentations de classification

Dans ce chapitre, nous allons présenter en détail les résultats de nos expérimentations sur les corpus en langue française ainsi que sur les corpus en langue vietnamienne. Les expérimentations présentées incluent :

- détection de questions par le modèle prosodique.
- détection de questions par le modèle lexical.
- combinaison des résultats de ces deux modèles.
- Recherche du meilleur jeu des paramètres pour la détection de questions

La première partie est une présentation des corpus utilisés, alors que la deuxième partie présente nos résultats d'expérimentation. Enfin, nous présentons un prototype de démonstration développé pour la reconnaissance de phrases interrogatives (Coco).

5.1.Corpus

5.1.1. DELOC

Nous utilisons le corpus du projet DELOC mené dans notre laboratoire, dont le but consiste à étudier différents types de réunions, ainsi que les différentes façons de s'exprimer (comportements langagiers selon les types de réunions). Le but du projet était de proposer des outils « collaboratifs » associés à la visioconférence, ou à n'importe quel contexte de réunions délocalisées, c'est-à-dire des outils d'aide à la rédaction du compte rendu en fin de réunion, ou d'aide à la transcription.

Ce corpus se compose de différents types de réunions délocalisées réalisées par téléphone : 1) « brainstorming » ou remue-méninges ; 2) (pré-)entretien d'embauche ; 3) réunion de projet. Les conversations étaient enregistrées au format A-law, 8 kHz (cf. qualité téléphonique), 16 bits, en mono, 21 locuteurs, ce qui représente environ 7h00 de parole. Ces enregistrements ont été segmentés manuellement en phrases qui correspondent chacune à une *question* ou une *non question.* Un sous-ensemble du corpus de 852 phrases dont 295 phrases *question* et 557 phrases *non question* est utilisé dans notre expérimentation. Les phrases de courte durée correspondent à : « *Allo?* », « *D'accord* »…alors que celles de longue durée correspondent par exemple à : « *parce que chez Multicom, j'imagine qu'il y a quand même...il y a quand même des gens qui pourraient peut être compléter ?* ».

5.1.2. NESPOLE!

Le projet NESPOLE[11] [The Nespole Project Consortium, 2002], co-financé par l'Union Européenne et la NSF (USA), adressait la problématique de la traduction automatique de parole et ses éventuelles applications dans le domaine du commerce électronique et des services. Les langues impliquées étaient : l'Italien, le Français, l'Allemand et l'Anglais. Les partenaires du projet étaient : ITC/IRST de Trento (Italie), ISL Labs. de Karlsruhe (Allemagne), CMU (Pittsburgh, USA), Aethra (une société italienne spécialisée dans le domaine de la vidéoconférence), APT (une agence de tourisme dans la région du Trentin en Italie) et le laboratoire CLIPS (Grenoble, France).

Le scénario NESPOLE consiste à mettre en jeu un agent parlant italien, présent dans une agence de tourisme en Italie, et un client qui peut être n'importe où (parlant anglais, français ou allemand) et utilisant un terminal de communication le plus simple possible (PC équipé d'une carte son et d'un logiciel de vidéoconférence type NetMeeting™). Le client veut organiser un voyage dans la région du Trentin en Italie, et navigue sur le site Web de APT (l'agence de tourisme) pour obtenir des informations. Si le client veut en savoir plus, sur un sujet particulier,

[11] NESPOLE!- Negotiating through SPOken Language in E-commerce
(voir http://nespole.itc.it)

ou préfère avoir un contact plus direct, un service de traduction de parole en ligne lui permet de dialoguer, dans sa propre langue, avec un agent italien de APT. Une connexion, via NetMeeting™, est alors ouverte entre le client et l'agent, et la conversation médiatisée (avec service de traduction de parole) entre les deux personnes peut alors démarrer. Les conversations étaient enregistrées au format PCM, 16 kHz, 16 bits, en stereo, 9 locuteurs, ce qui représente environ 3h22 de parole. Nous utilisons dans notre expérimentation un sous-ensemble en langue française de ce corpus qui comprend 650 phrases questions et 650 phrases nonquestions.

5.1.3. Assimil

Assimil est à l'origine une compilation de CDROM destinés à l'apprentissage de la langue vietnamienne pour les étrangers. Il comprend de la parole lue par des locuteurs professionnels, dans des conditions non bruitées. Comme ce cdrom a pour but l'apprentissage d'une langue pour les étrangers, le débit de parole est plus lent que celui de conversations normales. Les conversations étaient enregistrées au format PCM, 16 kHz, 16 bits, en mono, 5 locuteurs, ce qui représente environ 1h40 de parole. Nous avons extrait de ce corpus 168 phrases question et 168 phrases nonquestion pour nos expérimentations.

5.1.4. VietP

VietP est le nom donné au corpus que nous avons enregistré et utilisé pour l'étude de la différence entre phrases questions et phrases nonquestions en langue vietnamienne (présenté dans le Chapitre 3). Ce corpus était enregistré au format PCM, 16 kHz, 16 bits, en mono, et représente environ 2h20 de parole. Il contient 14 paires de phrases question/nonquestion différentes. Chaque paire est répétée cinq fois par six locuteurs, ce qui donne à ce corpus un nombre total de 420 phrases questions et 420 phrases nonquestions

5.2.Expérimentations

Dans cette partie, nous allons présenter et discuter des expérimentations de détection de questions en langue française ainsi qu'en langue vietnamienne. Une comparaison croisée entre les systèmes de classification pour une langue non tonale (le français) et pour une langue tonale (le vietnamien) est aussi discutée.

5.2.1. Le premier jeu de paramètres développé pour le corpus en langue française

Avec ce jeu de paramètres, nous avons mené plusieurs expérimentations. Cependant, une première étape à achever avant les expérimentations est l'estimation de la fréquence fondamentale à partir de signaux de parole. A partir de cette courbe mélodique de parole, les autres paramètres seront dérivés et ensuite utilisés pour construire les modèles de classification. Dans la littérature, il existe beaucoup de méthodes de calcul de F0. Nous pouvons citer ici brièvement le principe de quelques méthodes [Hess, 1983] :

- Méthode d'auto-corrélation

L'utilisation de l'autocorrélation pour la détection de la fréquence fondamentale est très classique en traitement de la parole. Pour un signal x(n), la fonction d'autocorrélation r(k) est définie par :

$$r_x(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$$

La cross-corrélation entre deux fonctions x(n) et y(n) est calculée par :

$$r_{xy}(k) = \sum_{m=-\infty}^{\infty} x(m)y(m+k)$$

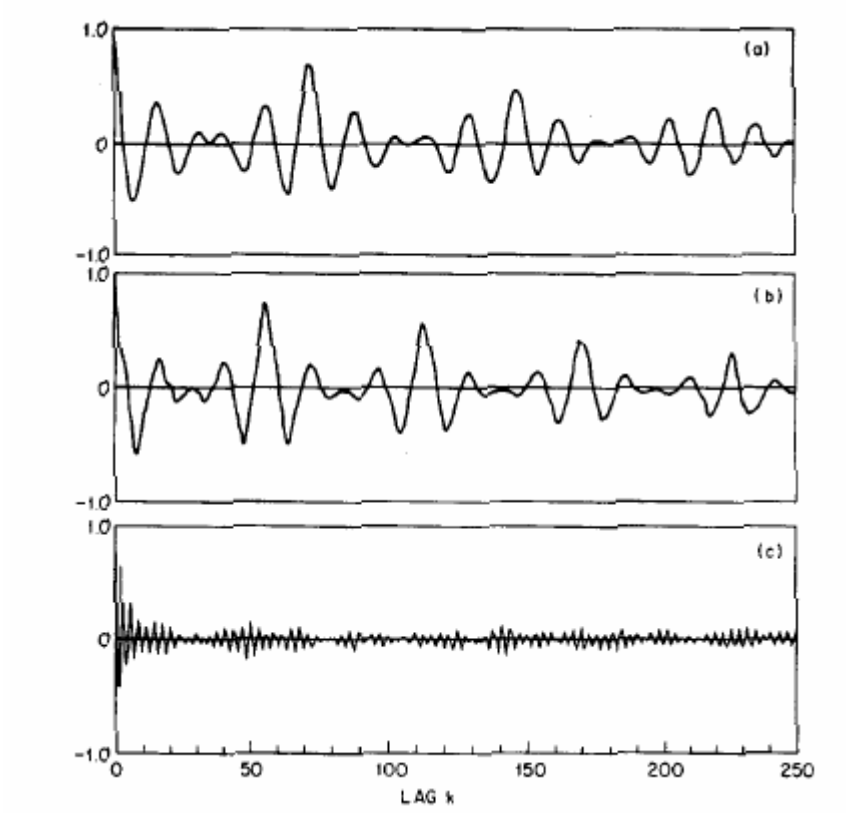


Figure 22 : Fonction d'Autocorrelation pour (a) et (b) parole voisée,et (c) parole non-voisée

La fonction d'autocorrélation ci-dessus vérifie les propriétés :

- Si xn est périodique, r(k) l'est aussi, avec la même période.

- r(k) atteint un maximum pour k=0

Ces deux propriétés impliquent que pour un signal périodique, la fonction d'autocorrélation possède des pics aux multiples de T0. De plus, la fonction d'autocorrélation normalisée donne

une estimation du « dégré de périodicité » du signal : une valeur $\max_k r(k)$ proche de 1 indique que le signal est très périodique.

L'estimation de la fréquence fondamentale du signal peut donc être faite par le calcul de l'autocorrélation, suivi d'une recherche de maximum.

Un problème fréquemment rencontré est celui du doublement de période pour lequel le pic situé à 2*T0 possède une amplitude supérieure à celui situé à T0, ce qui conduit à estimer une fréquence fondamentale égale à la moitié de la fréquence fondamentale réelle (erreur d'octave). C'est un type d'erreur qui affecte pratiquement toutes les méthodes d'estimation du fondamental.

Pour résoudre ce problème, on peut pré-traiter le signal par des transformations non-linéaires. Une des techniques est le clippage : on applique au signal une transformation non linéaire où seuls les échantillons du signal d'amplitude suffisante sont conservés, les autres sont mis brutalement à zéro.

Cette transformation a pour effet de diminuer considérablement l'influence des formants, sans altérer l'estimation de la fréquence du signal.

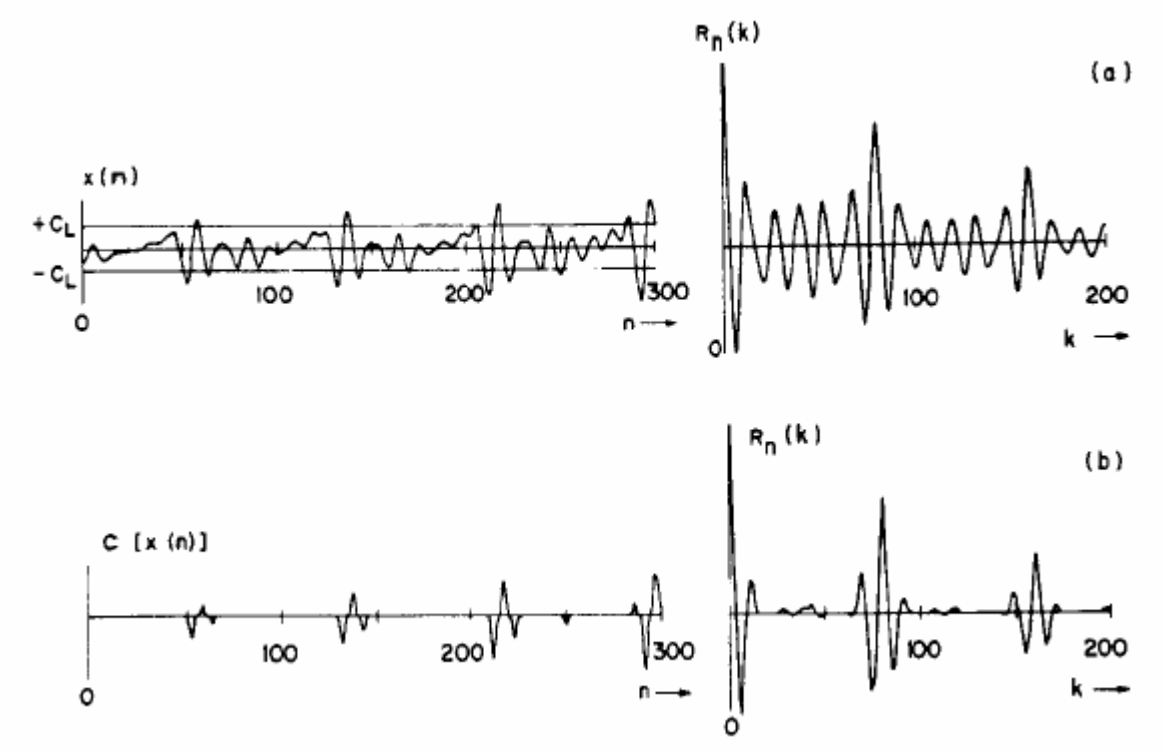


Figure 23 : Exemple du signal de parole et de sa fonction d'autocorrélation : (a) pas de clippage, (b) avec clippage

- Méthode AMDF (Average Magnitude Difference Function)

Pour mesurer la similarité entre deux périodes de signal, on peut utiliser la formule suivante qui définit l'AMDF:

$$AMDF(k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} |x_n - x_{n+k}|$$

Si le signal est parfaitement périodique de période T0, AMDF(i x T0) est bien sûr nul.

En filtrant le signal par les filtres correspondant à différentes valeurs de k, et en évaluant l'amplitude en sortie, on peut déterminer la valeur de k qui fournit la sortie la plus faible, et la retenir comme valeur de la période.

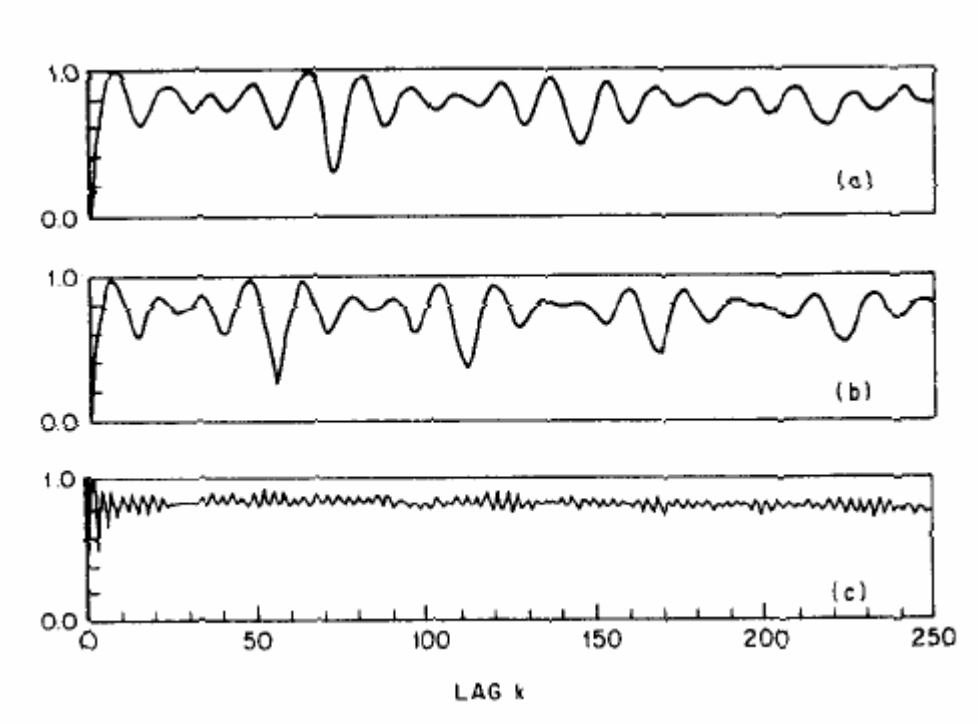


Figure 24 : Fonction d'AMDF pour (a) et (b) parole voisée,et (c)parole non-voisée

Cette méthode a surtout été utilisée pour sa simplicité numérique (pas de multiplication) lorsque les processeurs de traitement du signal savaient surtout faire des additions! Mais elle se révèle très sensible au bruit. C'est l'une des plus anciennes méthodes de détection de la fréquence fondamentale.

- Méthode SIFT (Simplified Inverse Filtering)

La méthode SIFT se base sur une estimation du maximum de l'autocorrélation d'un signal filtré au préalable par le filtre LPC inverse dont le but est d'atténuer l'influence des formants. En effet,

grâce à une analyse LPC on peut retrouver les coefficients du filtre correspondant aux formants du signal, si l'on inverse ce filtre on annule donc l'effet des formants sur le signal.

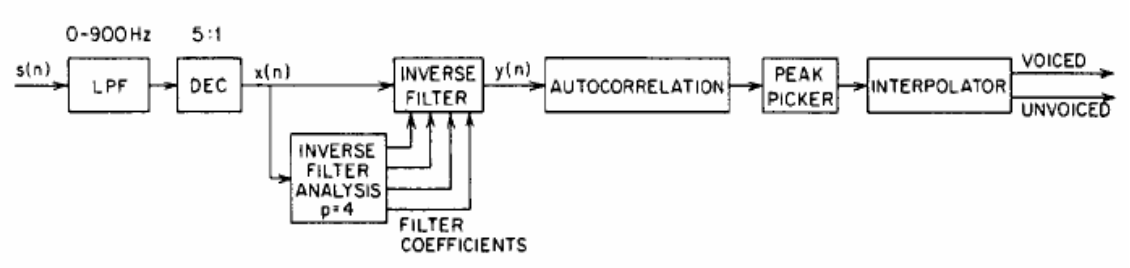


Figure 25 : Schéma de l'algorithm SIFT

Le processus commence avec un filtre passe-bas, dont le but est d'atténuer l'influence des fréquences autres que la fondamentale. Le filtre coupe à 1000Hz étant donné qu'une valeur de pitch est comprise entre environ 80 et 450 Hz. Mais son rôle le plus important reste d'éviter les recouvrements de spectre lors du sous-échantillonnage qui suit le filtre. Le but de ce sous-échantillonneur est simplement de diminuer la charge de calcul, en diminuant le nombre de points sur lesquels porte l'analyse.

L'étape suivante est une analyse LPC avec une pré-accentuation et une pondération de la fenêtre par une fenêtre de Hamming. Cette pondération ayant pour but d'amoindrir les fortes variations du signal sur les bords de la fenêtre, variations qui entraînent une mauvaise estimation des coefficients du filtre si elles n'étaient pas atténuées.

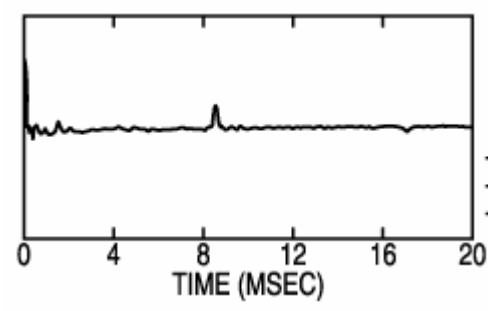
Ensuite, le vecteur d'autocorrélation de la fenêtre est calculé afin d'obtenir un maximum pour un décalage équivalent à la période du signal.

Enfin, un seuil de décision variant avec la fréquence est ajouté permetant de déterminer, avec plus de sécurité, la valeur du pitch. Une fois cette valeur obtenue, on effectue un dernier test qui consiste à vérifier qu'il est compris dans des valeurs possibles (entre 80 et 400Hz), s'il passe ce test, la valeur du pitch est sortie et une fenêtre voisée (V) est signalée; sinon la valeur de F0 est mise à zéro et la fenêtre non-voisée (NV) est signalée.

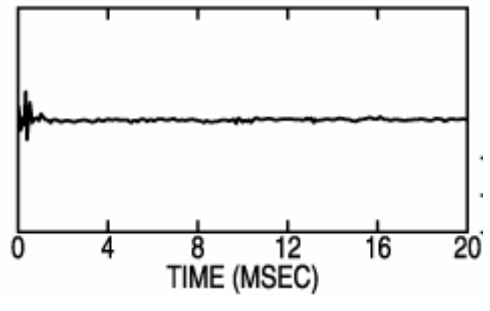
- Méthode fondée sur le cepstre

Le cepstre d'un signal discret x(n) est défini par :

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X(e^{j\omega})\right| e^{j\omega n} d\omega$$



(a)



(b)

Figure 26 : Cepstre d'un segment de parole : (a) voisé, (b) non-voisé

La Figure 26 suggère une bonne méthode pour l'estimation de pitch fondée sur le cepstre. On voit que pour l'exemple de parole voisée, il existe un pic dans le cepstre à la position de la période fondamentale du segment de parole. Il n'en existe pas pour le segment de parole non-voisée. Ces caractéristiques du cepstre peuvent être utilisées pour déterminer si un segment de parole est voisé ou non-voisé et pour estimer la période fondamentale du segment de parole voisée.

Le principe de la procédure de calcul de pitch fondée sur le cepstre est plutôt simple. On recherche dans le cepstre un pic dans la région autour de la période du pitch. Si le pic est supérieur à un seuil fixé, le segment de parole en entrée est probablement voisé, et la position autour du pic est la zone dans laquelle on peut estimer le pitch. Si le pic n'est pas supérieur au seuil, il est alors probable que le segment de parole en entrée est non-voisé.

- Les autres méthodes :

Il existe de nombreuse autres méthodes pour l'estimation de F0 : ondelette, estimateur des passages par zéro, analyse structurelle et beaucoup de variantes ou combinaisons de ces méthodes…

Nous nous intéressons dans notre étude aux représentations de la courbe mélodique de parole. Nous avons choisi d'utiliser le logiciel Praat pour l'extraction de F0 qui utilise la méthode d'autocorrélation. L'algorithme d'estimation de F0 de Praat est simple, flexible et robuste, facile à mettre en oeuvre. Le pitch dans une large gamme (voix d'homme, voix de femme, voix de jeunes enfants…) peuvent être extrait avec l'exactitude attestée dans [Boersma, 1993]. Par ailleurs, les résultats de l'évaluation comparative des algorithmes d'estimation de F0 - une étude réalisée par les auteurs [Cheveigné, 2001] - nous aident également à consolider le choix. Dans cette étude, dix méthodes d'estimation de F0 ont été évaluées sur quatre bases de données de parole différentes. Au total, ces quatre bases de données représentent 1,75 heure de parole par 38 locuteurs (19 hommes + 19 femmes ; 30 japonnais + 4 anglais + 4 français). Pour chaque locuteur, le signal laryngographique est enregistré en parallèle avec le signal de parole. En

analysant la comparaison de F0 obtenue par chaque méthode avec la F0 obtenue à partir du signal laryngographique, la méthode autocorrélation - une des méthodes expérimentées - montre une bonne performance sur des langues différentes et des locuteurs hommes/femmes (de 28 locuteurs). Cette performance est encore améliorée sur toutes les données avec la méthode YIN [Cheveigné, 2002] qui est aussi fondée sur l'autocorrélation.

5.2.1.1Sélection de la meilleure taille de fenêtre de calcul de F0

Pour la tache de classification du type de phrases, nous voulons étudier si la taille de la fenêtre de mesure de chaque méthode pourrait influencer également la performance de classification. C'est la raison pour laquelle nous avons calculé la courbe de fréquence fondamentale pour différentes fenêtres d'analyse : 20ms ; 40ms ; 60ms. Ensuite, pour chaque taille de fenêtre d'estimation de F0, les paramètres dérivés de la courbe d'intonation sont calculés, et les modèles de classification par arbre de décision sont construits. Enfin, les performances de classification sont mesurées et comparées entre elles afin d'identifier la taille de fenêtre qui semble optimale. Le Tableau 26 présente les performances obtenues. Les données utilisées dans ces expérimentations sont une version préliminaire du corpus Deloc comprenant 295 phrases question et 557 phrases nonquestion. Nous appliquons la méthode « validation croisée à 50 blocs » - c'est-à-dire nous répétons 50 fois le processus de division aléatoire du corpus en deux parties : une pour l'apprentissage (200 *questions* et 200 *non questions*), une pour le test (le reste : 95 *questions* et 357 *non questions*). 50 arbres de décision sont alors obtenus, chacun présentant une performance différente. La performance correspondant à chaque taille de fenêtre de calcul F0 est alors définie comme la valeur moyenne de performance pour les 50 configurations « apprentissage / test » différentes. La performance est calculée partout en utilisant la F_mesure (la F_mesure est présentée dans le 4.5) :

No	Méthode d'extraction de F0 et taille de fenêtre	F_mesure sur les données d'apprentissage	F_mesure sur les données de test
1	Autocorrélation 20ms	84%	56%
2	Autocorrélation 40ms	83%	53%
3	Autocorrélation 60ms	81%	55%

Tableau 26 : Liste des résultats obtenus par différentes tailles de fenêtre de calcul de F0 en utilisant les paramètres développés pour le corpus en français

Dans ce tableau, nous voyons bien que la méthode « Autocorrélation » à fenêtre de 20 milisecondes est la meilleure méthode pour cette tache de classification. La performance atteint son maximum avec cette méthode : les performances (F_mesure) sur les données de test sont de 56%. A partir de ce résultat, nous avons adopté uniquement cette méthode de calcul de F0 « autocorrélation – 20ms » pour les expérimentations ultérieures.

Pour mieux comprendre ces chiffres, nous avons comparé la performance de notre système de classification avec celle de systèmes simplistes ou répondant « au hasard ». Nous avons considéré les systèmes suivants :

- Un système qui répond toujours *question* quelque soit la phrase en entrée (1)
- Un système qui répond toujours *nonquestion* quelque soit la phrase en entrée (2)
- Un système qui répond au hasard question ou nonquestion à la proportion 50%-50% quelque soit la phrase en entrée (3)

Alors, sur le même corpus, le système (1) peut avoir une F_mesure sur les données de test de 34%[12]; ce taux pour le système (2) serait de 0%[13] et ce taux pour le système (3) serait de 28%[14]. Il est clair que notre système avec la performance de 56% est un système qui est significativement au dessus des peformances de ces systèmes répondant au hasard.

5.2.1.2Sélection du meilleur jeu de paramètres

No	Paramètres à l'ordre décroissant d'importance	F_mesure sur les données d'apprentissage
1	isRaising	75,96
2	min	77,08
3	highGreaterthanLow	78,24
4	range	79,57
5	fallingCount	80,68
6	raisingCount	81,77
7	mean	82,25
8	fallingSum	82,48
9	raisingSum	82,77
10	nonZeroFramesCount	83,01
11	max	83,10
12	median	83,98

Tableau 27 : L'ordre décroissant d'importance des paramètres du français

Après avoir identifié la meilleure méthode de calcul F0 qui est la méthode « auto corrélation – 20ms », nous nous interrogeons maintenant sur le niveau d'importance des paramètres. Est-t-il possible que les paramètres n'aient pas la même contribution dans le processus de classification ? Certains paramètres ont ils plus d'influence que d'autres ? Pour répondre à ces

[12] Système répondant toujours question : précisionQ=95/(95+357)=0,21 ; rappelQ=95/(95+0)=1 ; F_mesureQ=(2*0,21*1)/(0,21+1)=0,34 = 34%

[13] Système répondant toujours nonquestion (NQ) : précisionQ=0/(0+0)=0 ; rappelQ=0/(0+95)=0 ; F_mesureQ=(2*0*0)/(0+0)= 0 = 0%

[14] Système répondant question au hasard à la probabilité 50% : précisionQ=47/(47+178)=0,21 ; rappelQ=47/(47+48)=0,5 ; F_mesureQ=(2*0,21*0,5)/(0,21+0,5)=0,28 = 28%

questions, la méthode de sélection du meilleur jeu des paramètres « leave-one-out » présentée dans le Chapitre 4 section 4.6 est mise en œuvre. Pendant l'exécution de cette procédure, nous avons fait la sélection des paramètres à chaque itération en fonction des performances sur les données d'apprentissage. La liste des paramètres suivante, classés par ordre d'importance, est donc obtenue dans le Tableau 27

A la lecture du Tableau 27, il apparaît que le paramètre le plus important est « IsRaising », avec le résultat 75% pour la « F_mesure ». Le fait que ce paramètre est le plus important, ou autrement dit, est un bon indice pour détecter une phrase question, peut s'expliquer de manière assez logique : le paramètre « IsRaising » décrit globalement l'évolution de la forme du contour de F0, si ce contour est montant, le paramètre « IsRaising » aura une valeur positive, alors que si ce contour est descendant, ce paramètre aura une valeur négative. Dans cette optique, ce paramètre est pertinent pour la discrimination du type de phrase.

5.2.1.3Bilan des performances sur les corpus en français

Nous avons alors expérimenté ce jeu des paramètres sur les corpus en français DELOC (234 phrases question + 234 phrase nonquestion) et NESPOLE (650 phrases question + 650 phrases nonquestion). Les résultats sont résumés dans la Figure 27.

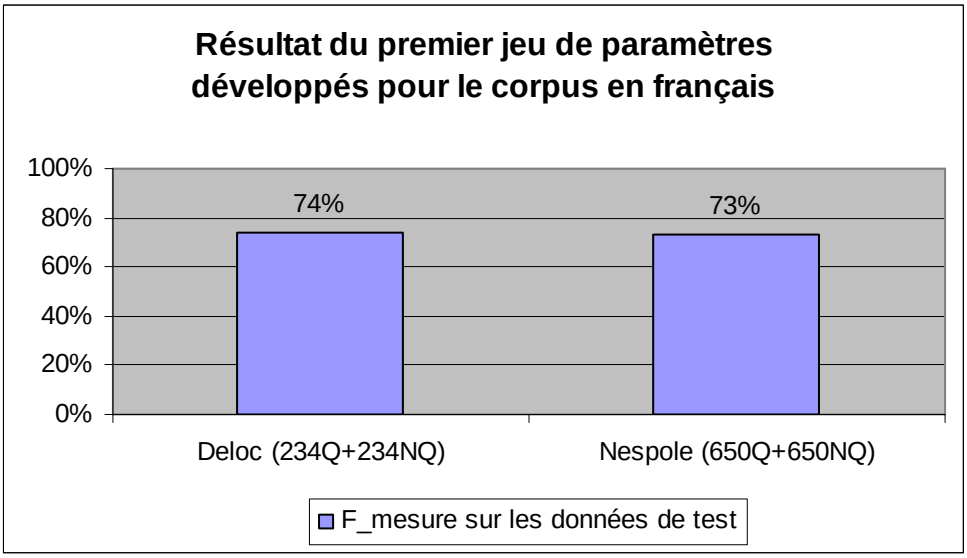


Figure 27 : Résultat du premier jeu des paramètres expérimenté sur les corpus en français

Pour mieux comprendre les propriétés de la prosodie des phrases question du français, nous nous sommes intéressés aux arbres de décision construits. En analysant ces arbres, nous trouvons que les paramètres « isRaising », « raisingSum », « raisingCount » sont ceux qui apparaîssent le plus souvent dans le premier rang (la racine) ou le deuxième rang (les nœuds suivants la racine) des arbres. Cependant, les autres paramètres comme « min », « max », « mean » n'apparaissent jamais en premier ou deuxième rangs. Ce phénomène nous suggère que les paramètres qui capturent le contour de F0 contiennent plus d'informations (l'entropie est

plus grande) ; ils sont donc plus pertinents que le deuxième groupe des paramètres classiques (« min », « max », « mean »..) qui sont des statistiques de distribution de F0. Pour la tâche de détermination du type de phrase, c'est donc plutôt l'évolution du contour de F0 qui compte, les statistiques de F0 étant moins pertinentes, pour la langue française. Nous remarquons le paramètre « isRaising » en racine dans la Figure 28 qui est une partie d'un arbre de décision construit sur le corpus en français. Plus d'exemples d'arbre sont donnés dans l'annexe B.1.

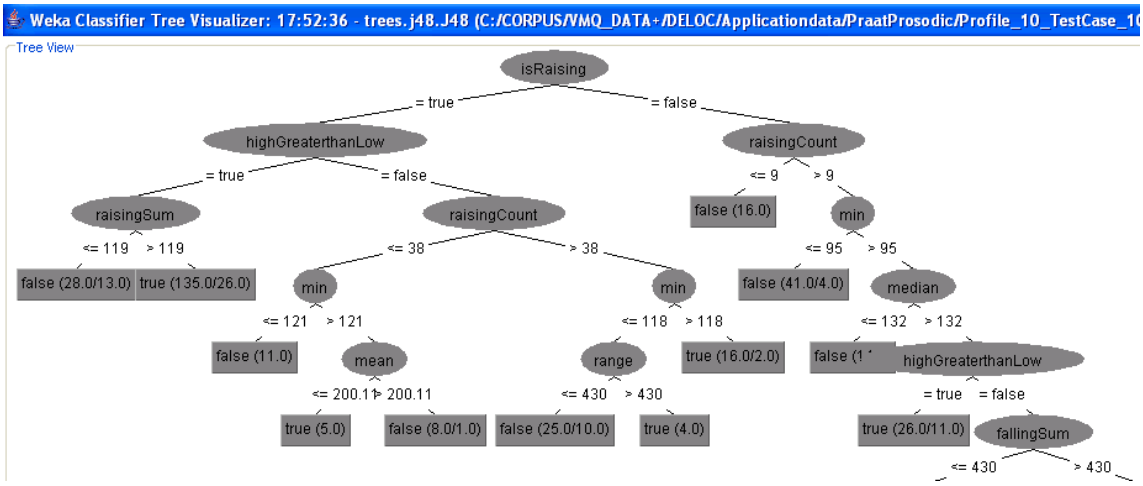


Figure 28 : Exemple d'arbre de décision obtenu pour le français : le paramètre "isRaising" est très souvent en racine.

5.2.2. Le deuxième jeu de paramètres développé pour le corpus en langue vietnamienne

5.2.2.1 Expérimentation sur les corpus en vietnamien

Nous avons expérimenté ce jeu des paramètres sur les corpus en vietnamien qui sont le corpus VietP (420 phrases question + 420 phrases nonquestion) et le corpus Assimil (168 phrases question + 168 phrases nonquestion). Les résultats sont présentés dans la Figure 29 :

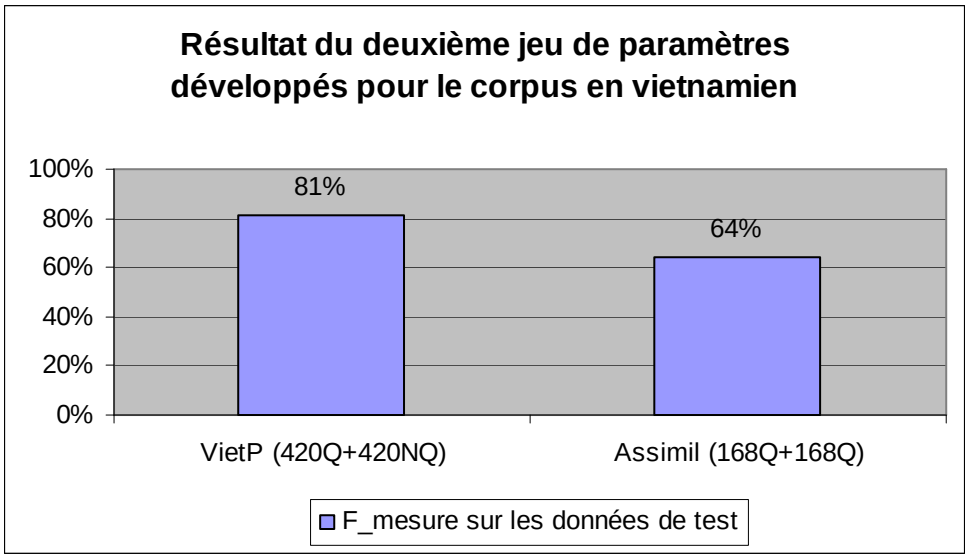


Figure 29 : Résultat du deuxième jeu des paramètres expérimenté sur le corpus en vietnamien

Si nous regardons en détail les arbres de décision construits par l'application de ce jeu de paramètres développés pour les corpus en langue vietnamienne, nous trouvons que le paramètre « lastDemiSyllableHighLevel » apparait très souvent en racine, vient ensuite le paramètre « moyenF0OfDebutSentence ». Cette observation nous permet de formuler une hypothèse que ces deux paramètres sont importants pour la classification de type de phrase en vietnamien. En effet, cette hypothèse est validée dans les prochaines expérimentations sur la sélection des meilleurs paramètres du vietnamien. En plus, cette hypothèse est également en corrélation forte avec le résultat d'analyse de la production prosodique en vietnamien présenté dans le 3.2.2. Du point de vue théorique, nous y avons démontré que parmi les autres facteurs qui distinguent les deux types de phrase question et nonquestion en vietnamien, il y a le contour de la dernière partie de la dernière syllabe. Du point de vue expérimental, le paramètre « lastDemiSyllableHighLevel » qui capture l'évolution de ce contour, s'est avéré être un des paramètres les plus pertinents dans l'arbre de décision (voir l'exemple d'un arbre construit sur le corpus VietP en vietnamien dans la Figure 30)

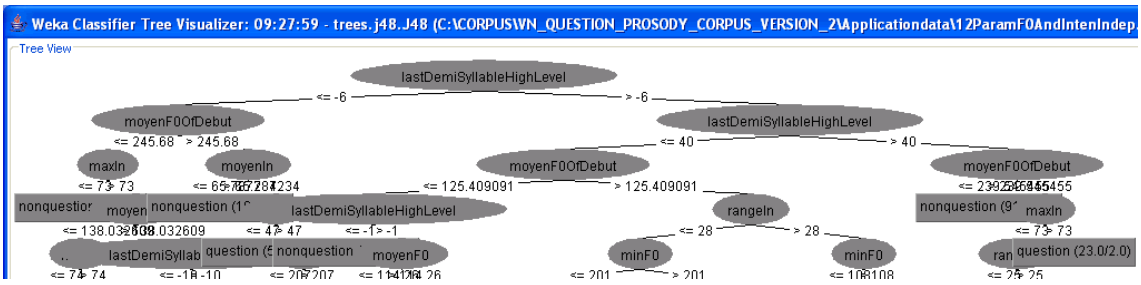


Figure 30 : Exemple d'arbre de décision obtenu pour le vietnamien : le paramètre "lastDemiSyllableHighLevel " est très souvent en racine.

5.2.2.2 Sélection du meilleur jeu des paramètres

Nous avons expérimenté la méthode de sélection du meilleur jeu des paramètres « leave-one-out » présentée dans le Chapitre 4 section 4.6, et voici la liste des paramètres dans l'ordre d'importance obtenu :

No	Paramètres à l'ordre décroissant d'importance	F_mesure sur les données d'apprentissage
1	lastDemiSyllableHighLevel	77,4
2	moyenF0OfDebutSentence	86,0
3	maxIntensity	87,3
4	minIntensity	89,0
5	moyenF0OfFinMinusDebutHighLevel	91,1
6	moyenIntensity	91,7
7	moyenF0	92,9
8	maxF0	93,2
9	moyenF0OfFinSentence	93,5
10	minF0	93,8
11	rangeIntensity	93,9
12	rangeF0	94

Tableau 28 : L'ordre décroissant d'importance des paramètres du vietnamien

Le taux de classification dans le cas où il y a seulement un paramètre « lastDemiSyllableHighLevel» est de 76% sur les données d'apprentissage. Le fait que ce paramètre est le plus important pour détecter une phrase question en langue vietnamienne peut s'expliquer de manière assez logique : ce paramètre caractérise l'évolution en terme de direction et de grandeur de la forme du contour F0 dans la zone de la dernière partie de la dernière syllabe. Plus la valeur de ce paramètre augmente (valeur positive), plus le contour F0 est montant, et inversement, plus la valeur de ce paramètre diminue (valeur négative), plus le contour F0 est descendant - ce qui est en corrélation avec les résultats d'analyse de la production de la prosodie en vietnamien présentée dans le 3.2.2.

5.2.3. Expérimentations croisées des jeux de paramètres prosodiques sur les corpus en français et en vietnamien

Cette section a pour but de valider nos hypothèses formulées pour le vietnamien en vérifiant que le jeu de paramètres conçu spécifiquement pour cette langue est plus efficace qu'un jeu de paramètre standard validé pour une langue non tonale comme le français. C'est le but principal des expérimentations suivantes que nous appelons « expérimentations croisées ». Les résultats sont résumés dans la Figure 31 :

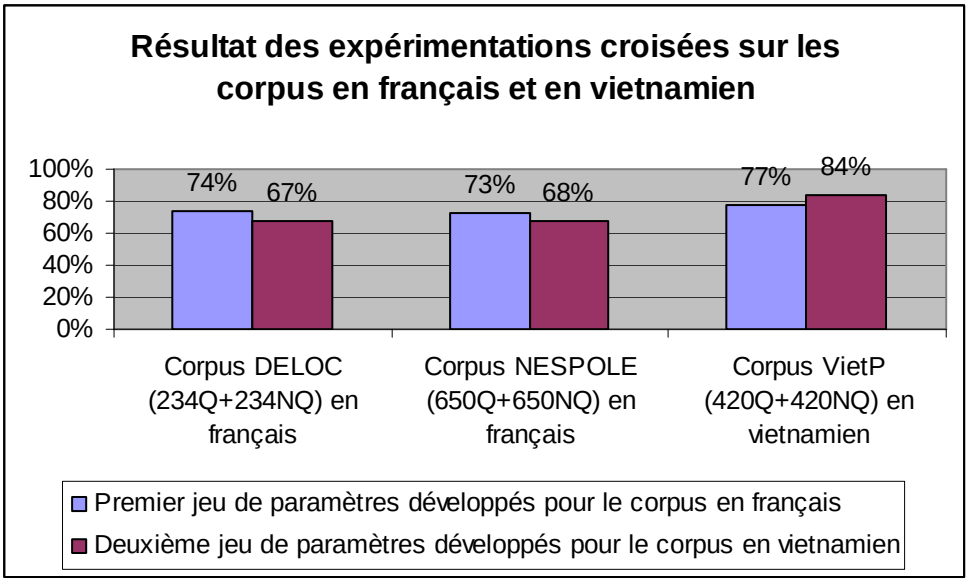


Figure 31 : Résultat des expérimentations croisées sur les corpus en français et en vietnamien

Nous trouvons que les paramètres initialement conçus pour les corpus en français obtiennent une meilleuse performance sur les corpus en français, alors que les paramètres initialement conçus pour les corpus en vietnamien obtiennent une meilleuse performance sur les corpus en vietnamien.

Pour les arbres construits par les paramètres conçus pour les corpus en français sur les corpus en vietnamien, ce sont les paramètres « median », « mean » qui sont les plus souvent au premiers rang. Les paramètres comme « isRaising », « raisingSum » qui étaient importants pour le corpus en français se révèlent moins pertinents pour le corpus en vietnamien : ils apparaissent plus en bas de l'arbre. Une explication que nous pouvons formuler pour ce phénomène est qu'il existe une différence entre la prosodie des phrases question en langue française et celle des phrases question en langue vietnamienne : les paramètres « isRaising », « raisingSum », « raisingCount » sont capables de capturer une évolution du contour F0 pour une durée plus longue, ils ne sont en revanche pas capables de capturer une évolution montante sur une courte durée comme celle de la dernière demi-syllabe de la phrase [Vu, 2006].

Dans les expérimentations sur les corpus en vietnamien (qui sont présentées dans le 5.2.2), nous démontrons que c'est le registre de phrase qui est pertinent pour la discrimination de phrase question/nonquestion. C'est la raison pour laquelle les paramètres « median », « mean » qui visent à capturer le registre de phrase deviennent plus importants quand le jeu de paramètres est appliqué sur les corpus en langue vietnamienne.

5.2.4. Expérimentations des modèles lexicaux

5.2.4.1 Sur les corpus en français

Nous avons utilisé les 71 paramètres lexicaux pour modéliser un arbre de décision. La méthode de test est toujours la validation croisée à 10 blocs. L'application de l'arbre sur les corpus donne les résultats présentés dans la Figure 32 suivante :

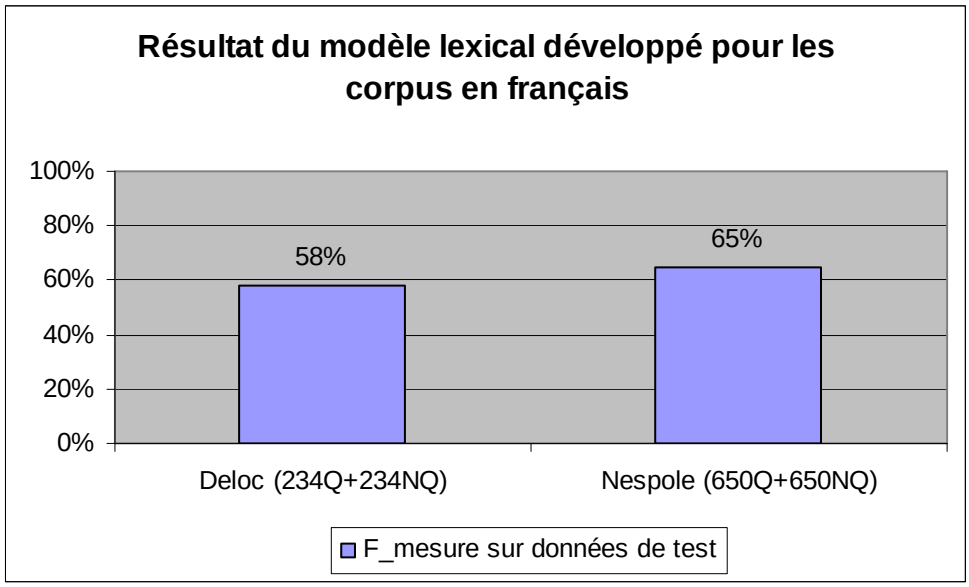


Figure 32 : Résultat modèle lexical du français expérimenté sur le corpus en français

Pour certaines phrases interrogatives n'ayant aucun terme interrogatif comme par exemple « parce que tu penses toi qu'on aurait intérêt à diversifier les magiciens ? », le modèle lexical ne peut pas reconnaître cette phase comme étant une question. Ce type de phrase interrogative a en effet une forme lexicale ressemblant à une phrase affirmative. L'aspect interrogatif n'est pas codé au niveau lexical, mais à un autre niveau qui est le niveau prosodique. Ce sont des cas de phrases interrogatives non reconnues par ce modèle lexical.

Les 71 paramètres lexicaux peuvent identifier correctement les phrases interrogatives qui possèdent au moins un des termes interrogatifs. Le nombre des phrases interrogatives possédant au moins un des termes interrogatifs pour le corpus DELOC est de 135 phrases, représentant 135/234=58% du total des phrases question. Pour le corpus NESPOLE ce chiffre est de 618/650=95%. Cependant, le nombre de phrases nonquestion ayant également un terme interrogatif est de 357 phrases, représentant 55% (6% pour le DELOC) – ce qui fait que certaines phrases nonquestion ont été inexactement classifiées et la performance finale atteint 65%.

5.2.4.2 Sur les corpus en vietnamien

Le modèle lexical du vietnamien présenté dans le 4.3.2 section 4.3.2 appliqué sur les corpus de texte en vietnamien par la validation croisée à 10 blocs donne les résutats suivants :

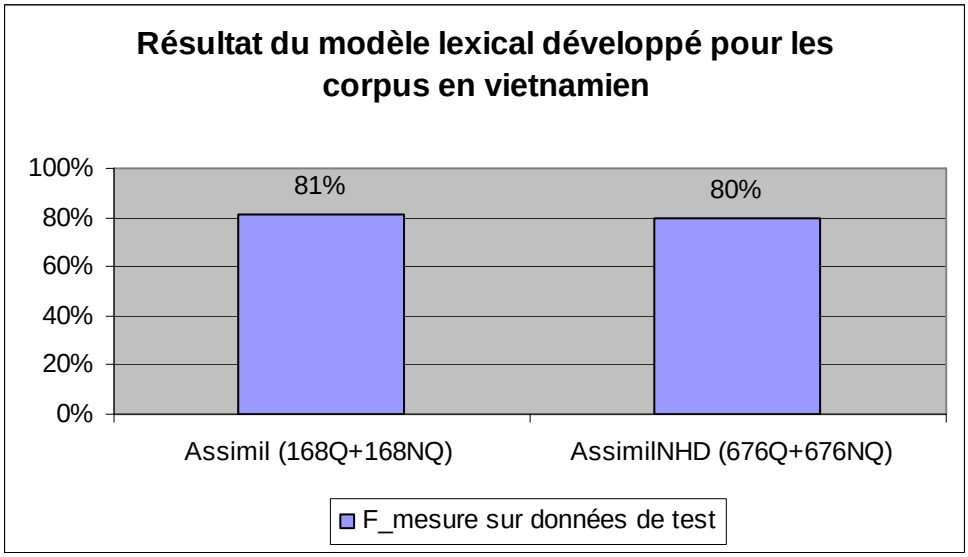


Figure 33 : Résultat du modèle lexical du vietnamien expérimenté sur le corpus en vietnamien

Le premier corpus est Assimil (168 phrases question + 168 phrases nonquestion) qui a été manuellement transcrit. Le deuxième corpus AssimilNHD est un nouveau corpus lexical composant des dialogues du corpus Assimil et des nouveaux dialogues collectés depuis le journal électronique vietnamien « Lao Động online »[15]. Ces dialogues sont des dicussions portant principalement sur les problèmes économiques/sociaux au Vietnam. Ce corpus AssimilNHD comprend 676 phrases question et 676 phrases nonquestion.

Le modèle lexical atteint un bon Fratio (plus de 80%) sur la détection de questions du vietnamien. Une explication possible est que : comme le vietnamien est une langue à ton, la prosodie contribue à former les tons, elle est par conséquent moins utilisée pour coder d'autres informations telles que le type de phrase. La modalité (ou le type de phrase) est plutôt codée par l'utilisation des mots, des termes interrogatifs et des expressions de demande. Il faut cependant noter que seuls les corpus français (DELOC et NESPOLE) sont issus de conversations réelles, tandis que les corpus vietnamiens sont plus « artificiels » et donc les phrases y sont mieux construites gramaticalement (le corpus ASSIMIL est issu d'un CDROM de l'apprentissage de la langue vietnamienne pour les étrangers). Ceci explique probablement en partie, aussi, les meilleurs performances du modèle lexical pour le vietnamien par rapport au français.

Cependant, en analysant les résultats, certains types de phrase interrogative sont difficiles à reconnaître. Ce sont les questions de type « choix alternatif » avec le terme « hay ». Ce mot peut

[15] « Lao Động » est un des grands journaux au Vietnam. Version électronique à l'adresse : http://www.laodong.com.vn/

avoir deux traductions équivalentes en français : (1) « ou » quand il joue le rôle de terme interrogatif, et (2) « bon », « bien » quand il est un adjectif et/ou adverbe. L'exemple suivant montre les deux rôles de ce terme:

- Etant un terme interrogatif : Anh thích uống chè hay cà phê ? (Tu préfères boire du thé ou du café ?)
- Etant un adjectif : Phim này rất hay nên tôi xem từ đầu đến cuối (ce film est si bon que je regarde depuis le début jusqu'à la fin)

La présence et la position de ce terme dans une phrase ne suffisent pas pour déterminer si la phrase est interrogative ou non. Dans ce but, une solution possible peut consister à mettre en œuvre une analyse grammaticale de la phrase. Actuellement, une telle analyse n'est pas réalisée pour notre jeu de paramètres lexicaux, nous la considérons comme une perspective possible.

5.2.5. Combinaison du modèle prosodique et du modèle lexical

Nous voulons tester ici si une combinaison de deux modèles prosodique et lexical peut donner une meilleure performance par rapport au cas où chaque modèle opère séparément. Dans cette expérimentation, nous allons combiner les deux modèles par deux méthodes d'intégration différentes qui sont la « combinaison précoce » et la « combinaison tardive » [Naturel, 2005].

5.2.5.1 Principe des méthodes de combinaison « précoce » et « tardive »

- Combinaison précoce

Le principe d'intégration de cette méthode est illustré dans le schéma de la Figure 34.

D'une manière générale, s'il y a deux modèles A et B à combiner, les paramètres du modèle A et les paramètres du modèle B seront mélangés dans le but de former un nouveau jeu de paramètres plus grand. Ce nouveau jeu de paramètres sera utilisé pour construire un nouvel arbre de décision, et c'est ce nouveau modèle qui sera utilisé pour évaluer la performance.

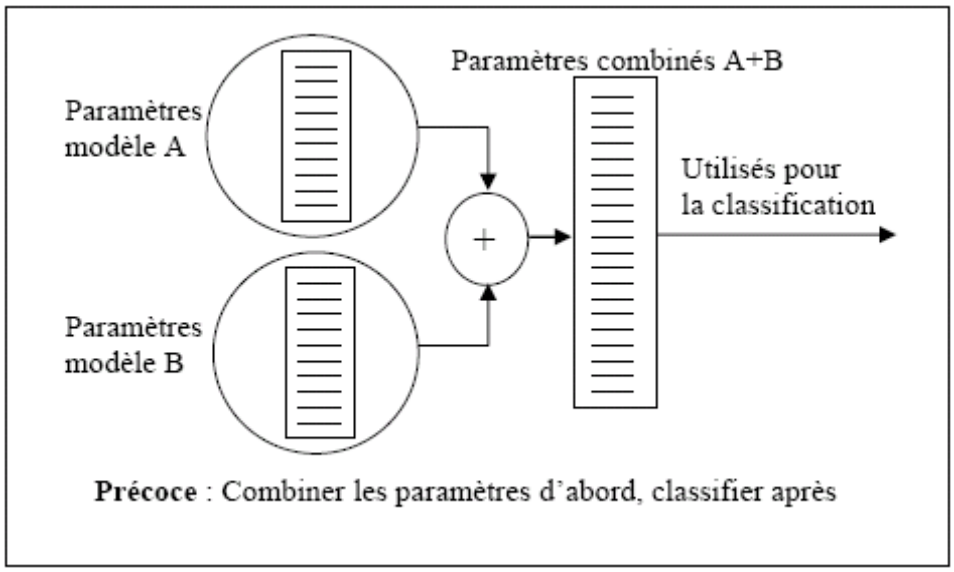


Figure 34 : Principe de combinaison "précoce"

- Combinaison tardive

Le principe d'intégration de cette méthode est illustré dans le schéma de la Figure 35 :

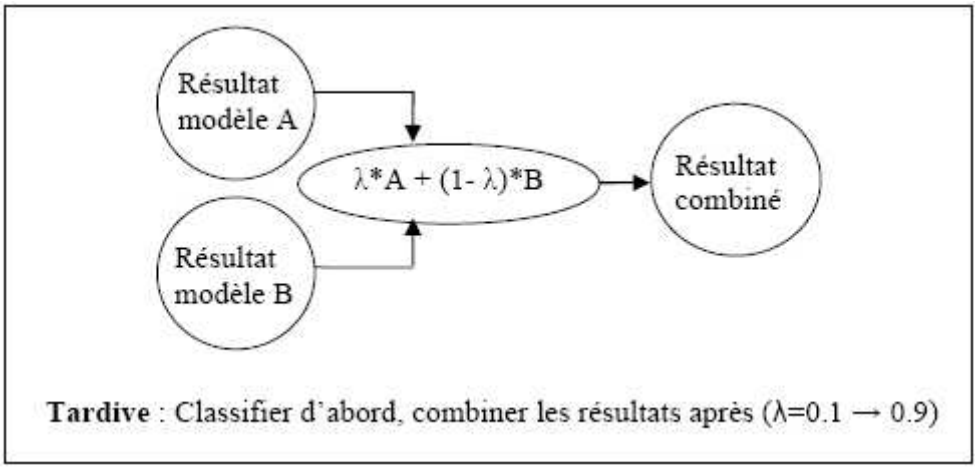
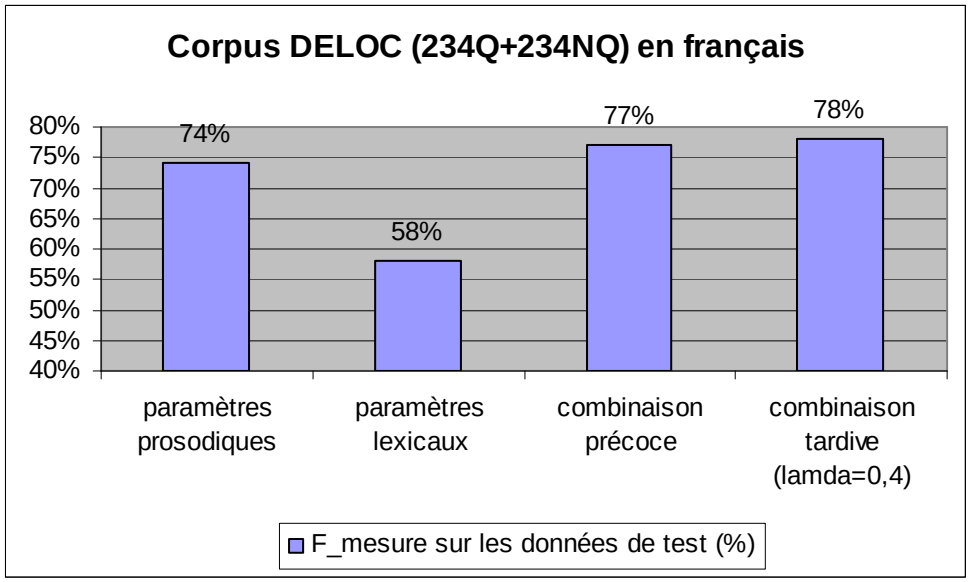


Figure 35 : Principe de combinaison de la méthode « tardive »

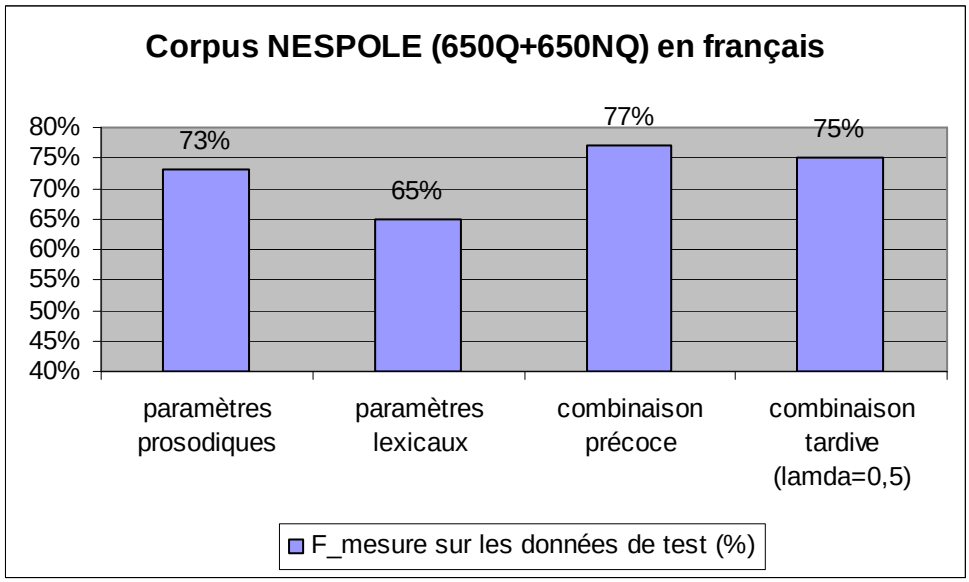
Dans cette méthode d'intégration, les deux modèles A et B opèrent indépendamment, c'est-à-dire chaque modèle propose son propre résultat. Ensuite, ces deux résultats seront combinés linairement afin d'obtenir un résultat final. En faisant varier la valeur de lamda de 0.1 à 0.9 sur des données de développement, nous obtiendrons 9 résultats combinés différents. La valeur lamda pour laquelle le résultat combiné atteint un maximum sera choisie comme la meilleure valeur pondérée.

5.2.5.2 Résultats de combinaison sur les corpus en français et en vietnamien

- Sur le corpus Deloc234 :



- Sur le corpus Nespole650 :



- Sur le corpus Assimil

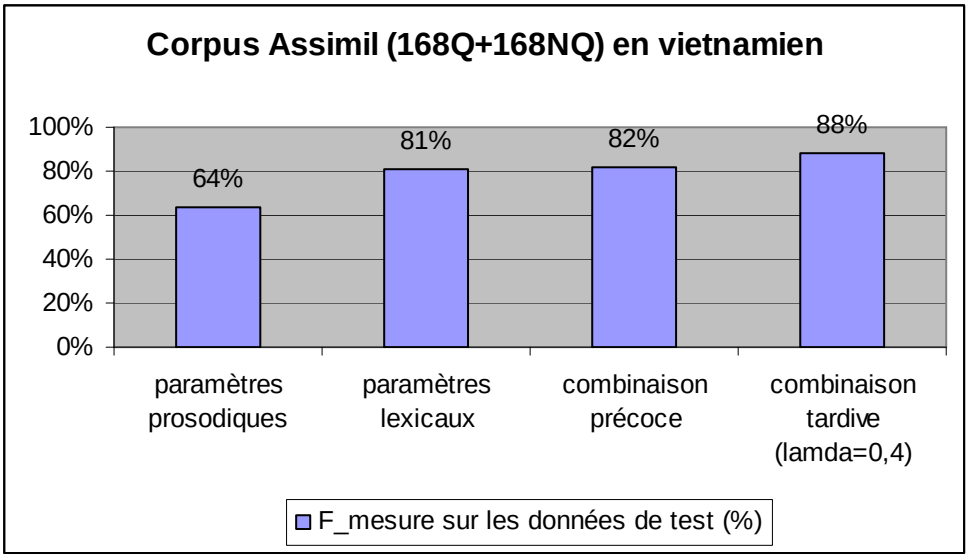


Figure 36 : Résultat de combinaison par méthodes « précoce » et « tardive » sur les corpus Deloc (en haut) ; Nespole (au milieu) et Assimil (en bas)

Avec la méthode d'intégration « précoce », nous avons un vecteur caractéristique de très grande taille pour chaque phrase dans le corpus. Ce vecteur est composé de paramètres prosodiques qui sont extraits à partir du signal de parole et de paramètres lexicaux qui sont extraits à partir de la transcription textuelle de la phrase. Le taux de classification s'est amélioré de 74% à 77% pour le corpus Deloc, de 73% à 77% pour le corpus Nespole, et de 64% à 82% pour le corpus Assimil.

Phrase	Vrai type	Classifiée par modèle prosodique	Classifiée par modèle lexical	Classifiée par combinaison de deux modèles
qu'est-ce que tu donnes comme exemple ?	Q	NQ (87%)	Q(100%)	Q(100%)
quelqu'un qui a dû dire ça.	NQ	Q (63%)	NQ (100%)	NQ (100%)
vous êtes à l'aise avec les outils informatiques ?	Q	Q (96%)	NQ (56%)	Q (96%)
d'accord okay pardon	NQ	NQ (100%)	Q (100%)	NQ (94%)

Tableau 29 : Exemples d'erreurs avec 1 seul modèle qui sont réparées par la combinaison des 2 modèles (Q=question ; NQ= nonquestion).

Bien que ce soit une amélioration légère, cette intégration montre quand même l'intérêt de combiner plusieurs sources d'informations de différentes modalités. Le Tableau 29 démontre quelques exemples de phrases, où l'erreur de classification existe avec 1 seul modèle, mais est corrigée par la combinaison des 2 modèles (les pourcentages entre les parenthèses sont les probabilités de classification).

Avec la méthode d'intégration « tardive », chaque modèle prosodique et modèle lexical opère séparément sur la modalité audio et la modalité textuelle respectivement. Ensuite, les résultats de classification des deux modèles ont été combinés en utilisant un coefficient de pondération lamda. En faisant varier le lamda de 0.1 à 0.9, nous avons trouvé que le taux de classification global atteint un maximum pour le corpus Deloc à la valeur lamda=0.4 ; pour le corpus Nespole à la valeur lamda=0.5, et pour le corpus Assimil à la valeur lamda=0.4. En plus, le taux de classification global a également augmenté pour tous ces corpus : le taux augmente pour Deloc de 74% à 78% ; pour Nespole de 73% à 75% ; et le taux augmente pour Assimil de 64% à 88%.

5.2.6. Conclusion sur les expérimentations

Nous avons expérimenté plusieurs jeux de paramètres, plusieurs taux de classification ont été enregistrés.

Tout d'abord, nous voulions savoir si la taille de fenêtre de calcul de F0 (la taille de fenêtre fixée à 20ms signifie qu'une valeur F0 est extraite tous les 20ms sur le signal de parole) pourrait avoir une influence éventuelle sur la performance de classification. C'est pourquoi nous avons évalué différentes tailles de fenêtre d'estimation de F0. Nous avons trouvé que la taille de fenêtre de 20ms est la meilleure.

Ensuite, nous avons construit plusieurs modèles de classification sur plusieurs jeux de paramètres prosodiques. Le meilleur résultat sur les corpus en langue française est de 74%, et sur les corpus en langue vietnamienne est de 81%. Nous avons trouvé que les phrases question en français sont bien représentées par les paramètres capturant la forme de la courbe mélodique, alors que les phrases question en vietnamien sont mieux représentées par les paramètres capturant la forme d'une toute dernière demie syllabe de la phrase. Les performances obtenues sur ces modèles montrent qu'un système de détection automatique de questions exploitant la prosodie de parole ayant une bonne performance est réalisable pour les deux langues française et vietnamienne.

Enfin, en essayant d'améliorer les performances de notre système, nous avons cherché à exploiter l'indice lexical qui est la représentation textuelle d'une phrase. Nos modèles lexicaux atteignent une performance de 65% sur les corpus en français et 81% sur les corpus en vietnamien. Lorsqu'ils sont combinés avec les modèles prosodiques, ils conduisent à un gain intéressant sur la performance globale du système – ce qui conduit à une performance globale du système de 77% pour le français et de 88% pour le vietnamien. Ce résultat montre l'intérêt d'exploiter et de combiner plusieurs modalités dans un système final.

Il est également démontré dans ces expérimentations que les paramètres conçus pour le corpus en français obtiennent une meilleure performance de classification sur les corpus en français que sur les corpus en vietnamien. De la même manière, les paramètres conçus spécifiquement pour le corpus en vietnamien obtiennent une meilleure performance de classification sur les corpus en vietnamien que sur les corpus en français, ce qui valide notre hypothèse consistant à dire que le jeu de paramètres conçu spécifiquement pour une langue tonale est plus efficace qu'un jeu de paramètre standard validé pour une langue non tonale.

5.3. Coco – une application pour la recherche d'informations: détection de phrases question

Dans cette partie, nous présentons un prototype d'application de notre méthode à la recherche d'informations. Ce programme est développé dans le but spécifique de détecter des phrases interrogatives dans un document audio.

Ce programme nommé Coco est développé dans le but de détecter si une phrase prononcée est de type interrogatif ou affirmatif dans un document audio. Il peut fonctionner de 3 manières différentes :

- L'utilisateur prononce une phrase devant le microphone, cette phrase est alors directement et immédiatement analysée par le programme Coco qui répond si la phrase dite est interrogative ou affirmative.
- L'utilisateur sélectionne un certain nombre de phrases préalablement enregistrées sur le disque dur ; le programme Coco analyse alors l'ensemble des phrases et répond pour chaque phrase si elle est interrogative ou affirmative.
- L'utilisateur sélectionne un fichier d'enregistrement de longue durée (par exemple, l'enregistrement d'une réunion…), le programme Coco va segmenter ce grand fichier en petits segments grâce à la présence des silences entres les segments. Chaque segment est ensuite analysé pour identifier son type : interrogatif ou affirmatif

Voici une image de l'interface du programme :

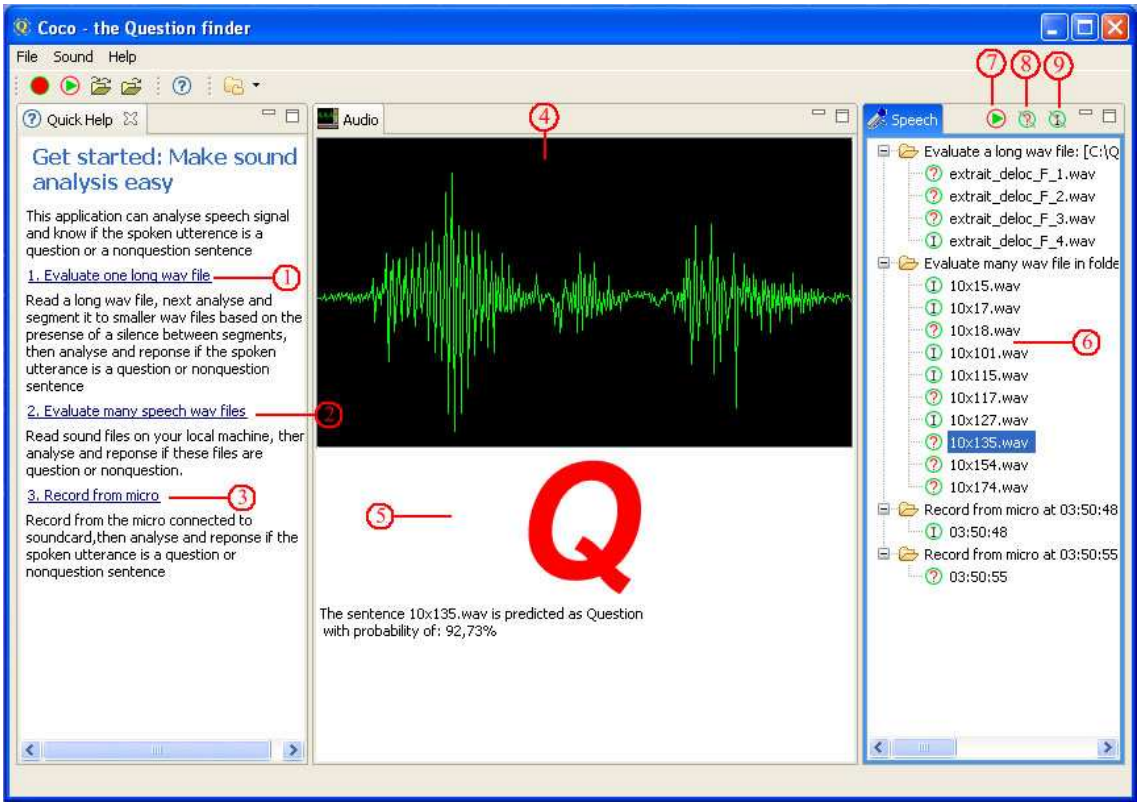


Figure 37 : Interface du programme Coco pour la reconnaissance de phrases interrogatives

Au niveau de l'interface du programme Coco, il y a plusieurs boutons et menus qui sont :

1. Le menu pour choisir l'évaluation d'un fichier son de longue durée
2. Le menu pour choisir l'évaluation de plusieurs petits fichiers en même temps
3. Le menu pour choisir la fonction d'enregistrement depuis le micro, puis de l'évaluer
4. Le zone d'affichage du signal sélectionné
5. Le zone d'affichage du résultat de l'évaluation : le caractère Q (en rouge) signifie que le signal est une question, alors que le NQ (en noir) signifie que le signal n'est pas
6. Le zone d'affichage de la liste des fichiers ouverts (ou enregistrés) ainsi que leur résultat d'évaluation : question ou nonquestion
7. Le bouton pour jouer le fichier son actuellement sélectionné
8. Le bouton pour masquer les phrases question dans la liste, c'est pour but de ne réviser que les phrases nonquestion
9. Le bouton pour masquer les phrases nonquestion dans la liste, c'est pour but de ne réviser que les phrases question

Du point de vue technique, ce programme Coco a été construit sur la base des composants suivants :

- Le modèle de classification est le modèle appris sur le corpus Deloc (mais il peut être changé très facilement pour être appliqué sur des signaux en vietnamien).
- L'estimation de la fréquence fondamentale est réalisée par le logiciel Praat.
- L'interface graphique a été développée avec l'environnement de travail Eclipse RCP (Riche Client Platform).
- Le moteur de classification utilise la boite à outils Weka
- L'enregistrement et la segmentation du signal de parole utilise la boite à outil Sphinx-4

Plus de détails sur ces composantes se trouvent en Annexe D.

Chapitre 6. **Conclusions et perspectives**

6.1. Conclusions

Dans cette thèse, nous avons abordé le problème de détermination automatique du type de phrase dans un document de parole. Il y a une quantité de plus en plus importante de données numériques qui sont disponibles pour le grand public sur le web, sur les chaînes média numériques (télévisions et radio). L'annotation selon le contenu d'une telle quantité de données est indispensable pour un accès facile à ce type de corpus. La reconnaissance de type de phrase peut contribuer à cette annotation structurée d'un document multimédia.

Nous avons présenté une nouvelle méthode de détection de questions par l'utilisation d'un arbre de décision. Les paramètres utilisés dans l'arbre comprennent des paramètres prosodiques qui sont extraits directement à partir du signal de parole ; des paramètres lexicaux qui sont liés à des propriétés spécifiques de la langue comme les expressions et les termes de demande. A partir de ces paramètres prosodiques de base (F0, intensité, durée), les autres paramètres dérivés sont calculés puis utilisés pour construire l'arbre.

Pour la langue vietnamienne, nos recherches ont contribué à mieux comprendre la prosodie des phrases *question* et *nonquestion* en vietnamien. Au niveau production, notre étude a permis de caractériser la prosodie des phrases simples de la langue vietnamienne (dialogue), en éliminant l'influence des tons : les différences entre *questions* et *nonquestions* sont codées essentiellement par trois facteurs principaux : une différence de pente de F0 (croissante ou décroissante) en fin de phrase (deuxième moitié de la dernière syllabe), une différence de registre (haut ou bas) et une différence du niveau d'intensité de la phrase. Au niveau perceptif, nous avons montré que, comme pour les langues non tonales, la prosodie de la phrase transporte des informations extralinguistiques sur la nature de la phrase. Cependant, à cause de la présence des tons appliqués sur les syllabes lexicales, ces informations ne sont pas toujours discriminantes.

Pour les deux langues française et vietnamienne, nous avons construit un modèle de classification fondé sur la prosodie avec un taux de classification correcte supérieur à 70%. Nous avons également investi dans le modèle lexical et dans les méthodes de combinaison de différentes sources de connaissances pour une tâche de classification globale. Nous avons montré qu'une combinaison de deux modèles prosodique et lexical peut amener une meilleure performance que chaque modèle opérant séparément. Paralèllement, nous avons démontré qu'un jeu de paramètres conçu spécifiquement pour une langue tonale (le vietnamien) est plus efficace qu'un jeu de paramètre standard validé pour une langue non tonale (le français).

Les résultats de nos recherches pourraient être utilisés dans d'autres applications en parole telles que le résumé automatique, la navigation ou la recherche d'information, car les zones autour d'une question contiennent souvent des informations importantes à identifier. De plus, les résultats de l'analyse de la production de la prosodie du vietnamien se révèlent fort utiles pour

les applications de synthèse de parole du vietnamien. En effet, ils ont été utilisés dans le moteur de synthèse du vietnamien développé au sein du centre de recherche MICA[16].

6.2. Perspectives

Pour la poursuite de ce travail, nous voulons augmenter la performance du système en étudiant davantage de paramètres. Par exemple, la durée n'est pas encore extensivement employée. La construction et l'expérimentation de modèles prosodiques dépendant du locuteur sont intéressantes et importantes à étudier car nous pouvons espérer obtenir une meilleure performance avec des systèmes qui pouraient modéliser les spécificités prosodiques de chaque locuteur.

Une deuxième perspective consisterait à explorer la pertinence des informations contextuelles autour d'un tour de parole. Par exemple, étant donné que le tour précédent était une question, on peut calculer, pour le tour de parole en cours, la probabilité qu'il soit encore une autre question ou qu'il soit plutôt une réponse pour la question posée dans le tour précédent.

Nous voulons expérimenter le système sur d'autres langues tonales pour voir comment le nouveau jeu de paramètres prosodiques peut y généraliser ; tester le modèle lexical sur les résultats du moteur de reconnaissance automatique de parole pour voir si la performance du modèle lexical est éventuellement dégradée.

Par ailleurs, beaucoup d'interrogations demeurent au sujet de la meilleure manière d'intégrer les diverses sources de connaissance. Nous avons expérimenté les deux méthodes « précoce » et « tardive », cependant il reste encore bien d'autres méthodes de fusion à expérimenter (voir [Naturel, 2005]. par exemple). Paralèllement, nous avons réalisé jusqu'à maintenant des traitements indépendants sur la modalité audio et la modalité textuelle, il est envisageable de les intégrer en plusieurs niveaux d'intéraction (niveau de paramètres, niveau de structure ou niveau de sémantique) comme proposé dans [Martin, 2005]

Les perspectives à plus long terme consisteraient à étudier si le résultat du système de détection de question pourrait bénéficier à un moteur de reconnaissance automatique de parole comme proposé dans un processus à deux phrases dans la Figure 38 :

[16] Centre de recherche international MICA – CNRS/UMI-2954 : http://mica.edu.vn/Home/index.jsp

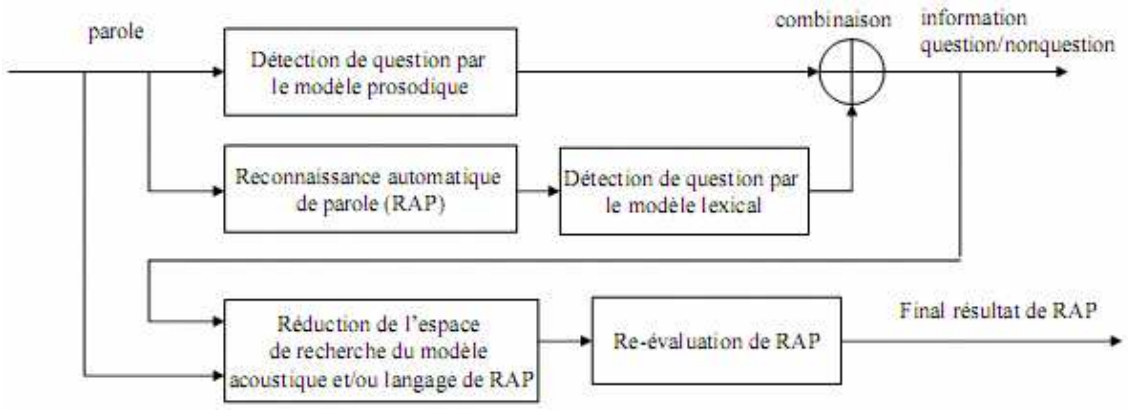


Figure 38 : Possible utilisation de l'information question/nonquestion pour la re-évaluation de la reconnaissance automatique de parole

Le taux de reconnaissance pourrait être amélioré pour un signal de parole : ce signal serait d'abord analysé par le système de détection de question. Une fois le résultat obtenu, le type de phrase serait utilisé pour réduire l'espace de recherche des candidats dans le modèle acoustique/langage du système de reconnaissance. Finallement, le modèle acoustique/langage réduit serait utilisé pour un deuxième tour de reconnaissance. Il s'agit ici d'une simple approche de couplage, une nouvelle approche de couplage plus « intégrée » pourrait amener une meilleure performance.

Une autre perspective consiste à appliquer l'approche utilisée dans cette étude pour d'autre tâches, comme cela est illustré dans les travaux suivants : détection d'événements structurels dans un dialogue (plusieurs niveau d'informations structurelles : frontière de phrase, pause verbalisée, réparation, disfluence) [Liu, 2004a], reconnaissance d'attitudes dans un système de dialogue [Fujie, 2003], détection de zones d'insistance (emphasis) sur des documents de réunions [Kennedy, 2003], détection de dialecte et d'émotions dans le système HMIHY[17] d'AT&T [Shafran, 2003], segmentation et classification automatique d'acte de dialogue [Ang, 2005], et puis généraliser le système pour la détection d'autres types de phrase comme exclamation, impératif…

Comme présenté dans le 5.2.4.2, certains types de question en vietnamien ne sont pas facilement identifiables par une simple détection de la présence et la position du terme interrogatif dans la phrase. Ils nécessitent un étiquetage en classes grammaticales de la phrase pour pouvoir déterminer si, dans le cas du mot « hay » par exemple, le mot joue le rôle d'un terme interrogatif ou d'un adjectif/adverbe. Nous voulons donc développer l'aspect multidisciplinaire de cette étude en étudiant des solutions intégrant une analyse sémantique de la langue.

[17] How May I Help You

Chapitre 7. **Bibliographie**

[Ang, 2002] J. Ang, R. Dhilon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody based automatic detection of annoyance and frustration in human computer dialog” In *proceedings of International Conference of Spoken Language Processing (ICSLP)*, Denver, Colorado, USA, September 16-20, 2002, pp. 2037–2040.

[Ang, 2005] J. Ang, Y. Liu, and E. Shriberg, “Automatic Dialog Act Segmentation and Classification in Multiparty Meetings” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, 2005, pp. 1061-1064

[Baron, 2002] D. Baron "*Prosody-Based Automatic Detection of Punctuation and Interruption Events in the ICSI Meeting Recorder Corpus*" M.S. Thesis, University of California at Berkeley, May 2002

[Besacier, 1998] L. Besacier, "*Un modèle parallèle pour la reconnaissance automatique du locuteur*", thèse de doctorat, University of Avignon, April 1998

[Besacier, 2007] L. Besacier, "*Transcription enrichie de documents dans un monde multilingue et multimodal*" HDR thesis, University Grenoble I, Jan 2007

[Bessac, 1996] M. Bessac, N. Colineau, "Patrons prosodiques et détermination d’actes de dialogue". In : *Actes du colloque Jeunes Chercheurs en Sciences Cognitives*, Giens, France, 5-6-7 juin, 1996. pp. 259-262

[Boersma, 1993] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." *Proceedings of the Institute of Phonetic Sciences* 17: 97–110. University of Amsterdam.1993

[Boersma, 2005] P. Boersma & D. Weenink, “*Praat: doing phonetics by computer (Version 4.3.14)*” [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/

[Breiman, 1984] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. “*Classification and regression trees*”. Technical report, Wadsworth International, Monterey, CA, 1984.

[Buckow, 1999] J. Buckow, V. Warnke, R. Huber, A. Batliner, E. Noth, and H. Niemann. "Fast and Robust Features for Prosodic Classification." In V. Matousek, P. Mautner, J. Ocelikova, and P. Sojka, editors, *Proc.Workshop on TEXT, SPEECH and DIALOG (TSD 99)*, volume 1692 of Lecture Notes for Artificial Intelligence, pages 193 198, Berlin, September 1999. Springer Verlag.

[Chatain, 2006] P. Chatain, E. Whittaker, J. Mrozinski, S. Furui, “Perplexity Based Linguistic Model Adaptation for Speech Summarisation”. *Proceedings of Interspeech*, Pittsburgh, USA, 2006, pp.1535-1538

[Cheveigné, 2001] A. d. Cheveigné, H. Kawahara, « Comparative evaluation of F0 estimation algorithms », *Proceeding of Eurospeech*, Scandinavia, 2001, pp. 2451-2454

[Cheveigné, 2002] A. d. Cheveigné, H. Kawahara, "Yin: A fundamental frequency estimator for speech and music", *Journal of the Acoustical Society of America.* 111, 1917-1930. 2002

[Christensen, 2001] H. Christensen, Y. Gotoh, and S. Renal, "Punctuation annotation using statistical prosody models" in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001.

[Church, 1991] K. Church, M. Liberman, Rapport sur ACL/DCI. In *Proc. of the 7 th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, 1991, pp. 80 - 91.

[Colineau, 1997] N. Colineau, "*Etude des marqueurs discursifs dans le dialogue finalisé*" Thèse de 3ème cycle : Sciences Cognitives, Université Joseph Fourier – Grenoble 1, 1997

[Do, 1998] T.D. Đỗ, T.H. Trần et G. Boulakia, "Intonation in Vietnamese" ; in Hirst & Di Cristo (ed.) *Intonation Systems : A Survey of 22 languages* (chap. 22) Cambridge U.P. 1998, ISBN: 0521395135

[Fernandez, 2002] R. Fernandez and R. Picard, "Dialog act classification from prosodic features using support vector machines," in *Proc. of Speech Prosody*, Aix-en-Provence, 2002.

[Ferrer, 2003] L. Ferrer, E. Shriberg, A. Stolcke, "A Prosody-Based Approach to End-of-Utterance Detection That Does Not Require Speech Recognition", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, Hong Kong, 2003, pp. 608-611.

[Finke, 1998]. M. Finke, M. Lapata, "CLARITY: Inferring Discourse Structure from Speech". In *Proc. AAAI '98 Spring Symposium on Applying Machine Learning to Discourse Processing.* Stanford, California, March 1998.

[Fontaney, 1991] L. Fontaney, 1991. « A la lumière de l'intonation ». In : *La question*. Kerbrat-Orecchioni (éd), Lyon : Presses Universitaires de Lyon.

[Fujie, 2003] S. Fujie, Y. Ejiri, Y. Matsusaka, H. Kikuchi, T. Kobayashi "Recognition of paralinguistic information and its application to spoken dialogue system". in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[Geoffrois, 1995] E. Geoffrois, *"Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole"*, thèse de doctorat, Université Paris XI Orsay, 20 décembre 1995.

[Gish 1991] H. Gish, M. H. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification",In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Mai 1991

[Gotoh, 2000] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000, pp. 228–235.

[Grimes, 1992] B. F. Grimes, "*Ethnologue: Languages of the World*". Dallas, Texas: Summer Institute of Linguistics, 1992

[Hardy et al, 2003] H. Hardy, K. Baker, H. Bonneau-Maynard, L. Devillers, S. Rosset, and T. Strzalkowski. "Semantic and Dialogic Annotation for Automated Multilingual Customer Service". *In ISCA Eurospeech*, Geneva, September 2003.

[Heeman, 1996] P.A. Heeman and J.F. Allen, "Combining the Detection and Correction of Speech Repairs", In *proceedings of International Conference of Spoken Language Processing (ICSLP)*, Philadelphia, PA, USA, October 3-6, 1996, pp. 363-365.

[Hess, 1983] W. Hess, "*Pitch determination of speech signal: algorithms and devices*", Springer, Berlin, 1983, ISBN : 0387119337

[Hirst, 1998] D.J. Hirst, & A. Di Cristo (Eds.) "*Intonation Systems. A Survey of 20 Languages*" Cambridge: Cambridge University Press. 1998

[Huang, 2002] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech", In *proceedings of International Conference of Spoken Language Processing (ICSLP)*, Denver, Colorado, USA, September 16-20, 2002, pp. 917–920.

[Jennifer, 2006] Jennifer J. Venditti, Julia Hirschberg, Jackson Liscombe, « Intonational Cues to Student Questions in Tutoring Dialogs », *Proceedings of Interspeech*, Pittsburgh, USA, 2006

[Jin, 2004] Q. Jin and T. Schultz. "Speaker Segmentation and Clustering in Meetings". In *proceedings of International Conference of Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, October 2004.

[Johnson, 2004] M. Johnson, E. Charniak, M. Lease, "An Improved Model for Recognizing Disfluencies in Conversational Speech", In *Proc. Rich Transcription 2004 Fall Workshop (RT-04F)*, 2004.

[Jones, 2003] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts", in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, pp. 1585-1588. 2003

[Kennedy, 2003] L. Kennedy and D. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003

[Kim, 2001] J. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. Of Eurospeech*, 2001, pp. 2757–2760.

[Kolar, 2004] J. Kolar, J. Svec, and J. Psutka, "Automatic punctuation annotation in czech broadcast news speech," in *Proc. of the 9th conference Speech and Computer*, 2004.

[Kozima, 1993] H. Kozima, "Text segmentation based on similarity between words". *In Proceedings of the 31st Annual Meeting*, 1993, pages 286-288, Ohio State University, Columbus, Ohio, June. Association for Computational Linguistics.

[Langlais, 1995] P. Langlais, « *Traitement de la prosodie en reconnaissance automatique de la parole* », thèse de doctorat : Spécialité « Informatique », Université d'Avignon et des Pays de Vaucluse, 1995

[Le, 1989] T-X. Lê, "*Etude contrastive de l'intonation expressive en français et en vietnamien*". Thèse de doctorat : Spécialité « linguistique », Université Paris 7, 1989

[Lickley, 1995] R. Lickley. "Missing disfluencies". *In Proceedings of International Congress of Phonetics Sciences*, pp. 192-195, 1995.

[Lin-shan, 2006] L-S. Lee, S-Y. Kong, Y-C. Pan, Y-S. Fu, Y-T. Huang, « Multi-layered Summarization of Spoken Document Archives by Information Extraction and Semantic Structuring ». *Proceedings of Interspeech*, Pittsburgh, USA, 2006

[Liscombe, 2006] J. Liscombe, J. Venditti, J. Hirschberg, « Detecting Question-Bearing Turns in Spoken Tutorial Dialogues », *Proceedings of Interspeech*, Pittsburgh, USA, 2006

[Liu, 2003] Y. Liu, E. Shriberg, A. Stolcke, "Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources", in *Proc. EuroSpeech*, Geneva, Switzerland, 2003.

[Liu, 2004a] Y. Liu, "*Structural Event Detection for Rich Transcription of Speech"*, PhD thesis, Purdue University, Indiana. 2004

[Liu, 2004b] Y. Liu, E. Shriberg, A. Stolcke, D. Hilliard, M. Ostendorf, B. Peskin, M. Harper, "The ICSI-SRI-UW Meta-Data Extraction System", in *Proc. of the NIST RT04 Workshop, November 2004.*

[Liu, 2006] Y. Liu, "Using SVM and Error-correcting Codes for Multiclass Dialog Act Classification in Meeting Corpus", *Proceedings of Interspeech*, Pittsburgh, USA, 2006

[Liu, 2006] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, & M. Harper, "Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies". *IEEE Trans. Audio, Speech and Language Processing* 14(5), 1526-1540, 2006

[Lu, 2001] L. Lu, H. Jiang, H.J. Zhang, "A Robust Audio Classification and Segmentation Method", *9th ACM Int. Conf. on Multimedia*, 2001, pp.203-211.

[Martin, 2005] T. Martin, A. Boucher & J-M. Ogier. « Multimodal Analysis of Recorded Video for E-Learning". *ACM Multimedia (MM), Doctorate Symposium*, November 2005, Singapore.

[Maskey, 2006] S. R. Maskey, B. Zhou, and Y. Gao. "A phrase-level machine translation approach for disfluency detection using weighted finite state transducers". *Proceedings of Interspeech*, Pittsburgh, USA, 2006.

[Mast, 1996] M. Mast, R. Kompe, S. Harbeck, A. Kiebling, H. Niemann, Noth, E.G. Schukat-Talamazzini, and V. Warnke. « Dialog act classification with the help of prosody. » In *Fourth International Conference on Spoken Language Processing*, volume 3, pages 1732–1735, 1996.

[Michaud, 2004] A. Michaud & N-T. Vu, "Glottalised and non glottalised tones under emphasis: open quotient curves remain stable, F0 curve is modified", in Proc. *Speech Prosody, Nara, Japan*. 745-748, 2004

[Moraru, 2004] D. Moraru, « *Segmentation en locuteurs de documents audios et audiovisuels: application à la recherche d'information multimédia* », Thèse de doctorat : Spécialité : « Signal Image Parole Télécoms », Institut National Polytechnique de Grenoble, 2004

[Nakatani, 1994] C. Nakatani and J. Hirschberg, "A Corpus Based Study of Repair Cues on Spontaneous Speech", *Journal of the Acoustical Society of America*, pp 1603 - 1616. 1994

[Naturel, 2005] X. Naturel, G. Gravier, and P. Gros, « *Etiquetage automatique de programmes de télévision* », In: Compression et Représentation des Signaux Audiovisuels (CORESA), edited 2005.

[Nenkova, 2006] A. Nenkova « Summarization Evaluation for Text and Speech: Issues and Approaches ». *Proceedings of Interspeech,* Pittsburgh, USA, 2006

[Nguyen, 2001] H-Q. Nguyen, "*Ngữ pháp tiếng Việt*", Nhà xuất bản từ điển Bách Khoa. 2001

[Nguyen-Quoc, 2001] Q-C. Nguyen**,** N-Y. Pham Thi and E. Castelli, "Shape vector characterization of Vietnamese tones and application to automatic recognition", in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Madonna di Campiglio, 2001.

[Nguyen-Quoc, 2002] Q-C. Nguyen, « *Reconnaisance de la parole en langue vietnamienne* ». Thèse de doctorat : Spécialité : « Signal Image Parole Télécoms », Institut National Polytechnique de Grenoble, 2002

[Nguyen-Thi, 1999] T-H. Nguyen Thi & G. Boulakia, "Another look at vietnamese intonation" *14th International Congress of Phonetic Sciences, San Francisco, California*, pp. 2399–2402, 1999

[Nguyen-Thi, 2004] T-H. Nguyen Thi, « *Contribution à l'étude de la prosodie du vietnamien. Variations de l'intonation dans les modalités: assertive, interrogative et impérative* » Thèse de doctorat : Spécialité « linguistique », Paris, Université Paris 7 : 2004

[Pappa, 2004] A. Pappa, G. Bernard, H. Oukerradi, « Détection automatique de frontières des phrases – Un système adaptatif multi-langues ». *ISDM n°13 février 2004*. Permanent online Journal of Information and Communication Technologies. ISDM (Informations, Savoirs, Décisions et Médiations)

[Pham-Thi, 2002] N. Y. Pham Thi**,** E. Castelli & Q-C. Nguyen, "Gabarits des tons vietnamiens" *JEP 2002 Nancy*, pp 23-26, juin 2002

[Post, 2002] B. Post, "French tonal structures". In: *Proc. Speech Prosody 2002 Conf.*, Aix-en-Provence, pp. 11-13.

[Quinlan, 1993] R. Quinlan. *"C4.5: Programs for Machine Learning"*, Morgan Kaufmann Publishers, San Mateo, CA, 1993, ISBN:1-55860-238-0.

[Ramousse, 1996] R. Ramousse, M. Le Berre & L. Le Guelte, « *Introduction aux statistiques* », 1996.

[Reynar, 1994] J.C. Reynar 1994. « An automatic method of finding topic boundaries". *In Proceedings of the 32nd Annual Meeting*, pages 331-333, New Mexico State University, Las Cruces, NM, June. Association for Computational Linguistics.

[Reynar, 1997] J. Reynar, & A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries", in *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., pp. 16-19. 1997

[Reynolds 2002] D. Reynolds, J. Campbell, B. Dunn, D. Jones, D. Sturim, T. Quatieri, "MIT Lincoln Laboratory System: 1sp, 2sp and Segmentation", *Proc of NIST SpRec 2002 Workshop*, Vienna, VA, Mai 2002

[Ries, 1999] K. Ries, "HMM and Neural Network Based Speech Act Detection". In *Proc. International Conf. Acoustics and Signal Processing (ICASSP)*, Arizona, March 15-19, 1999.

[Rosset, 2005] S. Rosset and D. Tribout, "Multi-Level Information and Automatic Dialog Acts Detection in Human-Human Spoken Dialogs", *Proceeding of InterSpeech*, Lisbon, September 2005

[Rossi, 1999] M. Rossi, « *L'intonation, le système du français : description et modélisation* » Editions Ophrys, 1999, ISBN : 2-7080-0912-5

[Shafran, 2003] I. Shafran, M. Riley, M. Mohri, "Voice signatures" in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[Shriberg, 1994] E. Shriberg, "*Preliminaries to a theory of speech disfluencies"*. PhD thesis, University of California, Berkeley. 1994

[Shriberg, 1997] E. Shriberg, R. Bates, A. Stolcke, "A Prosody only Decision Tree Model for Disfluency Detection", *Proc. of Eurospeech* 1997.

[Shriberg, 1998] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Cocarro, R. Martin, M. Meteer, and C. Van Ess-Dykema, "*Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?*", Language and Speech 41, pp 439-487, 1998.

[Shriberg, 2000] E. Shriberg, A. Stolcke, D. Hakkani−Tur, and G. Tur, "Prosody−based automatic segmentation of speech into sentences and topics." *Speech Communication*, 32(1−2): pp. 127-154, 2000. Special Issue on Accessing Information in Spoken Audio.

[Shriberg, 2001] E. Shriberg, A. Stolcke, and D. Baron. "Can Prosody Aid the Automatic Processing of Multi−Party Meetings? Evidence from Predicting Punctuation, Disfluencies, and Overlapping Speech." *Proc.ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, 2001.

[Snover, 2004] M. Snover, B. Dorr, S. Richard, « A Lexically Driven Algorithm for Disfluency Detection », Short Paper *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLTNAACL)*, May 2-7, Boston, USA, 2004.

[Stevenson, 2000] M. Stevenson and R. Gaizauskas, "Experiments on Sentence Boundary Detection". 2000. In *Proceedings of the North American Chapter of the Association for Computational Linguistics annual meeting (NAACL-2001)*, pages 24-30, Seattle, 2000

[Stolcke, 1998a] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, "Automatic Detection of Sentence Boundaries and Disfluencies Based on Recognized Words", *Proc. of ICSLP* 1998.

[Stolcke, 1998b] A. Stolcke, E. Shriberg, "Dialog Act Modeling for Conversational Speech". In *Proc. AAAI '98 Spring Symposium on Applying Machine Learning to Discourse Processing.* 1998

[Stolcke, 2000] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[Swerts, 1997] M. Swerts and M. Ostendorf.. "Prosodic and lexical indications of discourse structure in human-machine interactions". *Speech Communication,* 22(1), pp. 25-41. 1997

[The Nespole Project Consortium, 2002] The Nespole Project Consortium, "The NESPOLE! Speech-to-Speech Translation System", *Proc HLT (Human Language Technologies)*, San-Diego, CA, 2002

[Tran, 2005] D-D. Tran, E. Castelli**,** J.F. Serignat, V-L. Trinh & X-H. Le, « Influence of F0 on Vietnamese syllable perception", in *Proc. Insterspeech - Eurospeech 2005*, Lisbon, Portugal, September 4-8, 2005

[Tur, 2001] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg.. "Integrating prosodic and lexical cues for automatic topic segmentation". *Computational Linguistics*, 27(1):31–57. 2001

[Vaissière, 1999] J. Vaissière, "*Utilisation de la prosodie dans les systèmes automatiques : un problème d'intégration des différentes composantes."*, Faits de Langues, Oral-écrit: Formes et théories, Ophys, pp. 9-16, 1999

[Vu, 2005] M-Q. Vu, E. Castelli, A. Boucher & L. Besacier, "Classification de parole en Question et Non-Question par arbre de décision" *SFC 05, 12èmes Rencontres de la Société Francophone de Classification* - Montréal, 2005

[Vu, 2006] M-Q. Vu, D-D. Tran & E. Castelli, « Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results ». *Interspeech2006 ICSLP, International Conference on Spoken Language Processing*. 17-21 September 2006, Pittsburgh PA, USA

[Walker, 2001] D.J. Walker, D.E. Clements, M. Darwin and W. Amtrup, "Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality". *In Proceedings of the 8th Machine Translation Summit.* Spain. 2001

[Walker, 2004] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, J. Woelfel. « *Sphinx-4: A. Flexible Open Source Framework for Speech Recognition* » Technical Report TR-2004-0811, Sun Microsystem Inc.

[Wang, 2003] D. Wang, L. Lu, H.J. Zhang, "Speech Segmentation Without Speech Recognition", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, vol I, april 2003, pp. 468-471.

[Wang, 2004] D. Wang and S. S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, 2004.

[Witten, 1999] I.H. Witten, E. Frank, "*Data mining: Pratical machine learning tools and techniques with Java implementations"*, Morgan Kaufmann, 1999.

[Yuan, 2002] J. Yuan, C. Shih, G. P. Kochanski, "Comparison of Declarative and Interrogative Intonation in Chinese". *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, pp. 711-714

[Yuan, 2005] J. Yuan and D. Jurafsky, "Detection of questions in. Chinese conversational speech", in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Puerto Rico, November 28 - December 1, 2005.

[Zechner, 2001] K. Zechner, "*Automatic Summarization of Spoken Dialogues in Unrestricted Domains"*. PhD thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh. 2001

[Zhu, 2006] X. Zhu & G. Penn, « Summarization of Spontaneous Conversations ». *Proceedings of Interspeech*, Pittsburgh, USA, 2006

[Zong, 2003] C. Zong and F. Ren, “Chinese utterance segmentation in spoken language translation,” in *The 4th International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, 2003, pp. 516–525.

Annexe A. Les articles joints

- **VU M.Q**., TRAN D.D. & CASTELLI E. (2006) *Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results.* Interspeech2006 ICSLP, International Conference on Spoken Language Processing. 17-21 September 2006, Pittsburgh PA, USA

- **VU M.Q**., BESACIER L., CASTELLI E. (2007) « *Automatic question detection: prosodic-lexical features and crosslingual experiments* ». Interspeech2007 ICSLP, International Conference on Spoken Language Processing. 27-31 August 2007, Antwerp, Belgium

Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results

Vũ Minh Quang - Trần Đỗ Đạt - Eric Castelli

International research center MICA
IP Hanoi – CNRS/UMI-2954 – INP Grenoble
1-Dai Co Viet - Hanoi, Vietnam
{Minh-Quang.Vu ; Do-Dat.Tran ; Eric.Castelli}@mica.edu.vn

Abstract

This paper presents a new study on the prosody of Vietnamese language. Sentence pairs containing one interrogative sentence and one affirmative sentence, which have the same tones and the same number of syllables to avoid the effects of lexical tones and of co-articulation, are recorded in order to analyze their prosody evolution. Comparisons allow us to characterize differences between interrogative and affirmative sentences at sentence prosody level. Our work is completed by a perceptual study on re-synthesized sound where all syllables of the sentence are replaced by the vowel /a/ to hide lexical meaning, while the prosody of the sentence is kept unchanged. Our goal is to see if sentence prosody carries any information about sentence nature characteristics, and then whether it enables listeners to classify sentence type (in this case interrogative and affirmative), despite the complex form of this prosody in tonal languages. The obtained results show that information on sentence type is present at the end of the second half of the last syllable and that about 70% of sentences are properly classified for female synthesis voice.

Index terms: prosody, intonation, analysis, perceptive

1. Introduction

For the non-tonal Western languages (French or English) it was validated that sentence prosody carries extra linguistic information, such as the emotions or the state of the speaker, or the sentence type (affirmative, interrogative or exclamative [1, 2]). To automatically evaluate the type of sentences for detection or classification purposes, it is possible to analyze the sentence signal directly by using the prosodic characteristics of the sentence, without any need for lexical information from, for example, an Automatic Speech Recognition (ASR) engine. In this case, the measured and analyzed parameters take into account the evolution of the intonation during sentence statement: register of F0, increase of F0 at the end of the sentence or other parameters derived from the values of F0, for example [3, 4]. However, in the case of tonal languages (like Mandarin or Vietnamese), the melody contour of the intonation is complex. It is composed of macro-variations corresponding to the intonation of the sentence and of micro-variations corresponding to the lexical tone applied to each syllable of mono (or bi) syllabic words. This is why the direct application to tonal languages of analysis methods developed for non-tonal languages is very likely to fail, because tonal micro-variations tend to scramble the extra-lexical information coded on the sentence prosody. In the case of Vietnamese language, in order to differentiate the interrogative sentences from other sentence types, the use of specific words called "interrogative classifiers" (không, gì, chưa, for example) is practically systematic. Therefore, the main goal of our study here can be summarized as follows: is there in Vietnamese, a tonal language whose prosody is complex, any extra-linguistic information characterizing the sentence type conveyed by prosody and used during acts of dialogue? The answer will on one hand enable us to go further into our knowledge of the language, and on the other hand (in case of positive answer), will allow us to consider the realization of an automatic classification of sentence type that is independent of ASR system.

2. Prosodic analysis

2.1. Methodology and corpus preparation

Up to now, very few studies have deeply analyzed the phonology of the Vietnamese language. We can cite some recent works relating to the lexical tone [5, 6, 7] and to the prosody of the sentence [8, 9]. After analyzing sentences of "read" and "spontaneous" corpora, Lê T. X. [8] and Nguyễn Thị T. H. & Boulakia [9] noted that there is a difference in height of F0 between sentence types. By evaluating their register level, [8 & 9] presented that the assertive sentence is marked with a low register whereas the interrogative and the injunctive sentence have a high register. Moreover, [9] made the report that, on the intonation contour level, a descending slope does not always correspond to a declarative sentence. On the duration level, the interrogative statements have a faster rhythm than the assertive and injunctive ones, although the difference in duration between the two last is not significant [9]. As for the intensity level, it is generally stronger in the interrogative sentence, and the intensity of the final syllables is often more significant than the other syllables of the sentence [9]. Based on these reports, we wish to further determine the prosodic differences between interrogative sentences and affirmative sentences. For this purpose, we built a specific corpus made up of pairs of interrogative/affirmative sentences. The two sentences of one pair have the same tonal context and the same number of syllables. The choice of identical tones enables us to eliminate the influence from the syllable tones on the general intonation of the sentence, and also to control the micro-variations of the intonation. Furthermore, to also eliminate all the phenomena of co-articulation, which could interfere with our prosodic analysis, sentences had the same word structure, or we used words with little pronunciation difference. All these sentences were integrated in significant dialogues; so that their pronunciation is the most natural possible (we recorded the totality of the dialogues, and then extracted chosen sentences for analysis). Each dialogue is repeated five times by six native speakers (3 men and 3 women) from Hanoi (the North

region, considered to be official pronunciation of the Vietnamese language). It is noted that, in Vietnamese, for the construction of interrogative sentences, besides using practically and systematically the "interrogative classifier" words, speakers can add at the end of the sentence certain words which are normally optional. However, this addition possibility, which does not make change to sentence's meaning, depends strongly on the habitude, on the way the person speaks, on the context in which the dialogue occurs, on the expression of respect and/or courtesy towards the interlocutor, etc. Because these optional final words can carry any of the six tones of Vietnamese, the final portion of the intonation contour can be modified by the contour of the tone of these final words. This is why, for each selected interrogative sentence, we decided to incorporate into the corpus a certain number of variants which have final words with different tones, in order to study as much as possible the forms of sentence intonation contours. The Table 1 presents the 14 selected sentence pairs investigated. A complete sentence is formed by the root part (underlined text) followed by one of words in ending part (words in bracket "[ ]" and separated by "|"). We noted that one root can combine with one of many terminals while the sentence's meaning is still unchanged.

Table 1: 14 pairs of affirmative (A) and interrogative (I) sentences in corpus.

Affirm 1 to 5	Hôm nay là ngày + [ba mươi | ba mươi rồi | ba mươi vậy | ba mươi đấy | ba mươi bảy*] *Today is thirty (today is thirty seven*)*
Interro 1 to 5	Hôm nay là ngày + [bao nhiêu | bao nhiêu rồi | bao nhiêu vậy | bao nhiêu thế | bao nhiêu hả] ? *What is the date today?*
Affirm 6 to 9	Tên anh ta là + [Trì | Trì rồi | Kỳ Cậy* | Kỳ Thế**] *His name is +[Trì | Trì | Kỳ Cậy* | Kỳ Thế**]*
Interro 6 to 9	Tên anh ta là [gì | rồi | vậy | thế] ? *What is his name?*
Affirm 10 to12	Anh ăn cơm + [không | vậy | Không Thế **] *He eats rice only (He eats the rice Không Thế**)*
Interro 10 to12	Anh ăn cơm + [không | không vậy | không thế]? *Do you eat rice?*
Affirm 13	Em ăn bánh Ché *I eat Ché cake.*
Interro 13	Em ăn bánh nhé? *You will eat cake?*
Affirm 14	Ba giờ chi gặp anh Nghĩa *I meet Mr Nghĩa at three o'clock*
Interro 14	Bao giờ chi gặp anh Nghĩa? *When do you meet Mr Nghĩa ?*

2.2. Analysis results

For each signal, we have analyzed the contour of the fundamental frequency F0 using the software Praat (example presented in Figure 1).

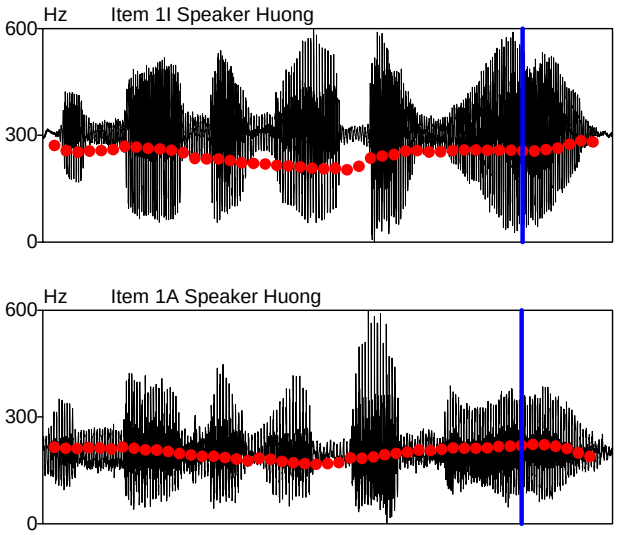


Figure 1: Two sentences (1I and 1A of Table 1) having the same number of syllables and the same tone. F0 contour is in red; on top: interrogative sentence, on bottom: affirmative sentence

By studying each pair of sentences presented in Table 1, we found that the main part of differences in intonation is at the end of the sentence (zone located on Figure 1 after the vertical bar): the contour of the last syllable or of its second half tends to increase for the interrogative sentences. A statistical study presented in Table 2 confirms this tendency: 85% of interrogative (I) sentences have an F0 contour increasing at the end of sentence. We recover here a well-known tendency for the non-tonal languages like French.

Table 2: F0 contours of the last half of the last syllable: Count (and percent) of contour rising/falling

Sentence type \ F0 contour	Rising contour	Falling contour
Interrogative	357(85%)	63(15%)
Affirmative	190(45%)	230(55%)

However, there's special case of 3 affirmative sentences (A) with the contour of the last half of the last syllable also increased for 29 of 30 recordings (even 30 of 30). These sentences are: "*Em ăn bánh Ché*"; "*Tên anh ta là Kỳ Thế*"; "*Anh ăn cơm Không Thế*". All of them contain terminal word which is a proper noun of ton5 (rising tone). Due to this fact, speakers have tendency to pronounce them very clearly to avoid misunderstanding which makes sentence intonation contour on this region very closed to syllable contour. The first sentence contains even two words of rising tone at the end. Both of these facts influence strongly the global intonation contour of the sentence as illustrated in the following Figure 2.

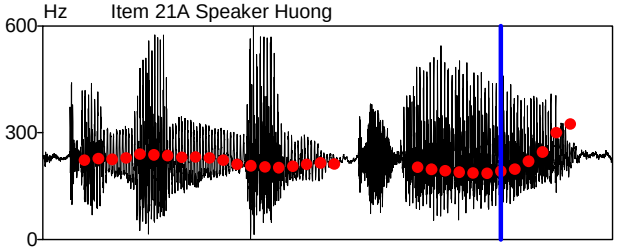


Figure 2: The sentence "Em ăn bánh Ché" with both two final words of rising tone

If we look closer to the form of sentence final part in case of different final tons, we can remark that: in case of I sentences, the contour of sentence final part increases regardless of the tone of the final syllable, or, increases in all cases of final tons (from 1 to 6). But it's not the same in case of A sentences where the slope of this region varies with final tons: it increases in case of ton3 (ngã-broken) or ton5 (sắc-rising), but decreases with remaining tons as illustrated on the following Table 3:

Table 3: Contours of the last-half of the last syllable in interrogative sentence with terminal words of different tones: count (and percentage %) contours rising/falling.

	ton1	ton2	ton3	ton4	ton5	ton6
Interrogative Rising	36 (60)	69 (77)	30 (100)	24 (80)	116 (97)	82 (91)
Interrogative Falling	24 (40)	21 (23)	0 (0)	6 (20)	4 (3)	8 (9)
Affirmative Rising	1 (2)	9 (10)	30 (100)	13 (43)	119 (99)	18 (20)
Affirmative Falling	59 (98)	81 (90)	0 (0)	17 (57)	1 (1)	72 (80)

[8 and 9] suggested that the sentences of the type I are marked with a higher register. For this point, statistical study on our corpus shows that sentences of type I have an average value of F0 higher than that of sentences of type A (Table 4). However, while the difference is significant for female speakers, that of male speakers are weak and smaller than the values of the corresponding standard deviation. Thus, unlike work of [8 and 9], the effect of register is not very significant in our corpus. For the duration, we find the same tendency noted in [9]: the duration of the interrogative sentences is on average smaller 10% than that of the affirmative ones (12% in [9]).

Table 4: F0 average (and standard deviation) in Hz of I and A sentences of six speakers. (M = Male, F = Female).

Speaker	Diệp F	Hương F	Lan F	Thành M	Khoa M	Phương M
Interrogative	261 (22)	259 (22)	253 (12)	143 (10)	144 (11)	124 (8)
Affirmative	226 (19)	217 (18)	245 (14)	136 (7)	127 (10)	119 (6)

3. Perception of I/A sentences

3.1. Methodology and corpus preparation

We wish to verify whether the differences detected in our analysis are actually perceived as a factor for listeners to classify interrogative sentences and affirmative sentences, or, in other words, that prosody of the sentence, in spite of its complexity due to the presence of tones, carries information which allows listeners to make this classification. We used sentence pairs in the corpus described above. For each sentence, after extraction of the prosodic contour, we used that contour to synthesize a pseudo sentence in which all syllables are replaced by the vowel /a/. Because the lexical signification of the words does not exist anymore, we thus eliminate the possibility for the listener to recognize a question only by the presence of an "interrogative" word. We have reproduced as accurately as possible, not only the intonation contour, but also the duration of the voiced/unvoiced segments, along with the intensity contour. Then, listeners heard these pseudo-sentences and were asked to determine whether the perceived synthetic sentences were interrogative or affirmative.

3.2. Synthesis for perception test

The extraction of this prosodic information is carried out by the software Praat with an analyzing windows of 20 ms for F0 and 5ms for the intensity. For the synthesis, we extracted two periods from signal of a vowel /a/ in one sentence uttered by the same speaker of our preceding study. A male voice and a female voice were synthesized.

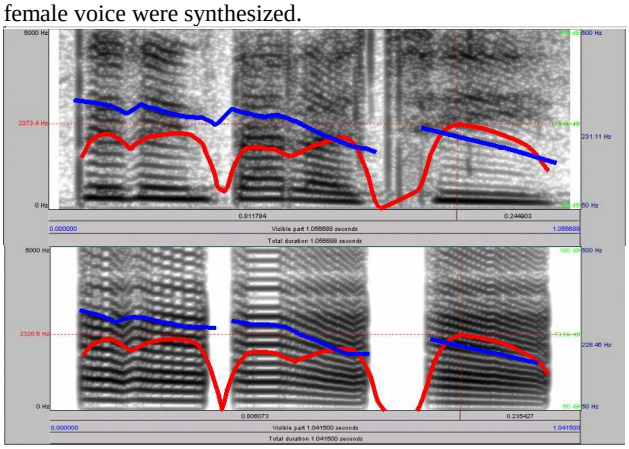


Figure 3: Spectrogram, F0 contour (blue) and energy contour (red) of source signal (a) and corresponding synthesis signal (b). Sentence 6a of Table 1

The TD-PSOLA algorithm was used to concatenate these extracts of signal, and to control the pitch (F0) of synthesized sentences, and zones corresponding to an unvoiced signal (consonants) in these sentences were replaced by silence. We thus obtained a corpus made up of 13 synthetic "interrogative" pseudo-sentences and of 13 synthetic "affirmative" pseudo-sentences, without any semantic information. Six listeners (3 men and 3 women) participated in our perception test. They had to choose between two answers "I" or "A". Each listener did the test 5 times for female voice and 5 times for male voice, and for each time, the order of the sentences proposed to him was random.

3.3. Perception results

The perception test results are presented in Table 5. For female synthesis voice, the global correct recognition rate on the whole of I and A is approximately 70%. We can see that the interrogative sentences were better recognized (approximately 74 % of good answers) than affirmative sentences (only 63%). For male synthesis voice, the global recognition rate is only 60% while other rate for I/A sentences are also weaker than those of female voice.

Table 5: Correct recognition rate (in percent) of sentence types

Rates \ Synthesis voice	Female	Male
Global Correct Recognition rate	69%	58%
Correct Recognition rate of I sentences	74%	61%
Correct Recognition rate of A sentences	63%	55%

We can see that with female synthesis voice, listeners can do the recognition and classification better than with male voice. Based on analyzed results, we found that the F0 variations of this male speaker are less significant than that of female speaker.

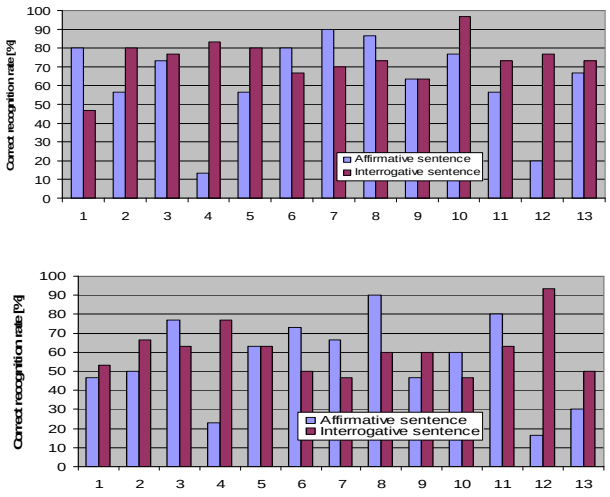


Figure 4: Correct recognition rate of 13 pairs of sentences I/A.
Synthesis of: on top female voice, on bottom male voice.

Figure 4 details the results for the 13 pairs of I/A sentences. With female voice, for 10 of these 13 pairs, the interrogative sentence is well recognized with a rate higher than 70%, and with the 10th pair, this rate even reaches 95%. However for pairs 4 and 12, the affirmative sentence is very poorly recognized (12% and 20% respectively).

3.4. Discussion

While trying to correlate these perceptive results with those of our analysis on the intonation contour production, we found that listeners had a tendency to consider a sentence as being interrogative if it presents a rising intonation at the end, and to consider the sentence as being affirmative in the opposite case. This assumption seems to be valid to explain the case of pairs 4 and 12 where recognition rate of I-type sentences is much higher than that of A-type sentences. For these two pairs, all the sentences present a last syllable having tone 5 (rising tone), which makes the final part of the intonation contour of the sentence raise, both for interrogations and assertions. The fact that the global correct recognition rate of sentences A and I is approximately 70% (and some of them have correct recognition rate higher than 90%) shows that the prosodic parameters of Vietnamese sentence transport extra-linguistic information which can allow listeners to discriminate sentence types. Beside other factors of intensity and duration, as for the non-tonal languages, this information is coded by the fact that the intonation goes up or not at the end of the sentence. However, this information can be scrambled by the modulation of prosodic contour by the lexical tone: listeners can badly classify assertions in case that produced sentences present a final syllable with the rising tone. Questions can be badly classified if their final syllable carries a falling tone. The use of interrogative words to eliminate ambiguities is thus necessary and logical.

4. Conclusions

In the production level, our study help us to characterize the prosody of simple sentences of Vietnamese language (dialogue), by eliminating the influence from tone: the differences between interrogative and affirmative sentences are characterized primarily by a difference in F0 contour (increasing or decreasing) at the end of the sentence (second half of the last syllable), and by a modification of speaking rate. However, for our study, the change of register seems weaker than for [8 and 9]. At the perceptive level, we showed that, as for the non-tonal languages, the prosody of the sentence transports extra-linguistic information of the type of the sentence, but they are not always discriminative, due to the presence of the lexical tone.

5. References

[1] Rossi M. "*L'intonation, le système du français : description et modélisation*" Editions Ophrys, 1999, ISBN : 2-7080-0912-5.

[2] Hirst, D.J. &Di Cristo, A. (Eds.) "*Intonation Systems. A Survey of 20 Languages*" Cambridge: Cambridge University Press.

[3] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M. & Van Ess-Dykema, C. "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech 41*, pp. 439-487, 1998.

[4] Vu M.Q., Castelli E., Boucher A. & Besacier L. "Classification de parole en Question et Non-Question par arbre de décision" *SFC 05, 12èmes Rencontres de la Société Francophone de Classification* - Montréal, 2005

[5] Nguyen Q.C.**,** Pham Thi N.Y. & Castelli E. "Shape vector characterization of Vietnamese tones and application to automatic recognition" *ASRU 2001 Madonna di Campiglio*, cdrom

[6] Pham Thi N. Y.**,** Castelli E. & Nguyen Q.C. "Gabarits des tons vietnamiens" *JEP 2002 Nancy*, pp 23-26, juin 2002

[7] Michaud A. & Vu N.T. "Glottalised and non glottalised tones under emphasis: open quotient curves remain stable, F0 curve is modified" *Speech Prosody, Nara, Japan*. 745-748, 2004

[8] Lê Thị X., "Etude contrastive de l'intonation expressive en français et en vietnamien". *Thèse en linguistique : Paris, Université Paris 7*, 1989

[9] Nguyễn Thị T.H. & Boulakia, G. "Another look at vietnamese intonation" *14th International Congress of Phonetic Sciences, San Francisco, California*, pp. 2399–2402, 1999

[10] Yuan, J.; Shih, C; Kochanski, G. P., Comparison of Declarative and Interrogative Intonation in Chinese. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, pp. 711-714.

Automatic question detection: prosodic-lexical features and crosslingual experiments

Vũ Minh Quang – Laurent Besacier** - Eric Castelli**

*International research center MICA, CNRS/UMI-2954 – 1, Dai Co Viet - Hanoi, Vietnam
**LIG Laboratory, CNRS/UMR-5217 - 681 rue de la passerelle - BP 72 - 38402 Saint Martin d'Hères , France
{Minh-Quang.Vu ; Eric.Castelli}@mica.edu.vn ; Laurent.Besacier@imag.fr

Abstract

In this paper, we present our work on automatic question detection from the speech signal. We are interested in developing automatic detection system and investigate the portability of such system to a new language. The first goal of this paper is to propose and evaluate a combined approach for automatic question detection where prosodic features are augmented by the use of lexical features. It is shown that both early and late integration of theses features in a decision tree-based classifier improves the question detection performance compared to a baseline system using prosodic features only. The second goal of this paper is to conduct a crosslingual (French / Vietnamese) evaluation concerning the use of prosodic features.
It is shown that our first system developed for French which uses an initial prosodic feature set can be improved using a new feature set that takes into account some specific prosodic characteristics of the Vietnamese tonal language. Both Vietnamese and French question detection systems obtain F-ratio performance around 80% on pre-segmented meeting and dialog utterances.

Index terms: prosodic / lexical features, automatic question detection, crosslingual experiments

1. Introduction

Automatically detecting questions from the speech signal is a challenging task that may be interesting for document summarization or to enrich a transcription by punctuation marks as well as in the objective of dialog act detection [1, 2]. For the non-tonal Western languages (French or English) it was validated that sentence prosody carries extra linguistic information, such as emotions, state of the speaker, or sentence type (affirmative, interrogative or exclamative [3, 4]). To automatically evaluate the type of sentences for detection or classification purposes, it is possible to analyze the speech signal directly by using its prosodic characteristics, as done in [1]. Using only prosodic information leads to acceptable performance but could be probably improved by using other cues. Thus, the first goal of this paper is to propose and evaluate a combined approach for automatic question detection where prosodic features are augmented by the use of lexical features that might be obtained by an Automatic Speech Recognition (ASR) engine for instance. The second goal of this paper is to conduct a crosslingual (French / Vietnamese) evaluation concerning the use of prosodic features. In fact, typical prosodic features take into account the evolution of the intonation during sentence statement: range of F0, increase of F0 at the end of the sentence or other parameters derived from the values of F0 [1, 5]. However, in the case of tonal languages (like Mandarin or Vietnamese), the melody contour of the intonation is complex [6]. It is composed of macro-variations corresponding to the intonation of the sentence and of micro-variations corresponding to the lexical tone applied to each syllable of mono (or bi) syllabic words. This is why the direct application to tonal languages of prosodic feature sets developed for non-tonal languages is very likely to be suboptimal since tonal micro-variations tend to scramble the extra-lexical information coded in the sentence prosody.
A former study from the authors of this paper [7] has shown that, as for non-tonal languages, the sentence type (question or not) information is mainly coded by the fact that the intonation goes up or not at the end of the sentence. However, this information can be scrambled by the modulation of prosodic contour due to the lexical tone. For instance, it was shown in a perception test that listeners (and so probably an automatic system) can badly classify assertions when produced sentences have a final syllable with the Vietnamese *rising* tone[1]. Similarly, questions can be badly classified if their final syllable carries a *falling*[2] tone. Thus, the second contribution of this paper consists in proposing an optimized prosodic feature set for Vietnamese that takes into account these problems, and to evaluate it through crosslingual experiments.
Section 2 of this paper presents the baseline features initially designed for question detection in French as well as our optimized feature set for Vietnamese. Section 3 shows briefly our lexical features while sections 4 and 5 present our crosslingual and prosodic-lexical experiments respectively. Finally section 6 concludes this work.

2. Prosodic features

2.1. Baseline features for French

In French language, the interrogative form of a sentence is strongly related to its intonation curve. Therefore, we decided to use the evolution of the fundamental frequency (F0) to automatically detect questions in an audio input.
From this F0 curve, we derive a set of features which aim at describing the shape of the intonation curve. The parameters defined for our work are listed in Table 1. It is important to

[1] The *rising* tone is one of the 6 different tones of the vietnamese language
[2] The *falling* tone is one of the 6 different tones of the vietnamese language

note here that, contrarily to classical short term feature extraction used in speech recognition, a unique long term feature vector is automatically extracted for each utterance of the database.
[3] software.

No	Parameter	Description
1	Min	Minimal value of F0
2	Max	Maximal value of F0
3	Range	Range of F0-values of the whole sentence (Max-Min)
4	Mean	Mean value of F0
5	Median	Median value of F0
6	HighGreater-ThanLow	Is sum of F0 values in first half-length smaller than sum of F0 values in last half-length of utterance?
7	RaisingSum	Sum of F0i+1− F0i if F0i+1 > F0i
8	RaisingCount	How many F0i+1 > F0i
9	FallingSum	Sum of F0i+1− F0i if F0i+1 < F0i
10	FallingCount	How many F0i+1 < F0i
11	IsRaising	Is F0 contour rising? *(yes/no)*. Test whether RaisingSum > FallingSum
12	NonZero-FrameCount	How many non-zero F0 values?

Table 1: 12-dimensional feature vector derived from the F0-curve for each utterance

2.2. Optimized features for Vietnamese

In [7], it was shown that for Vietnamese, the differences between interrogative and affirmative sentences are characterized primarily by: a difference in F0 contour (increasing or decreasing) at the very end of the sentence; a higher register and a stronger intensity. These findings allow us to propose a new set of features which are more appropriate to describe these differences than those initially developed for French corpus in table 1. We do not detail all these features here but the main added parameters are related to the intensity (for instance minimum intensity of the current sentence), to the syllable durations, and above all, an important parameter was added which represents the F0 range in the last demi-syllable of the sentence. More precisely, it is the difference between F0 values of ending and beginning points of the last demi syllable. This parameter is particularly important to tackle the problem of the final tone (*rising* or *falling*) described at the end of section 1. All these features can be measured automatically; the parameters that need syllable begin / end labels were extracted after applying a forced alignment between the speech signal and its transcription[4].

3. Lexical features

The goal of our lexical features is to represent interrogative terms or expressions. However, for many languages, the presence / absence of such terms is not sufficient, since their position in the sentence is also very important. For instance, These features can be divided into 2 main categories: the first 5 features are the statistics on F0 values, and the 7 next features describe the contour of F0 (raising or falling). The F0 contour was extracted using the Praat
French interrogative words like *pourquoi*[5] or *comment*[6] will probably indicate an interrogative utterance if they are the first word of it, while this might be different if they are not at the beginning of the sentence. Thus, bag-of-words techniques, like those used in information retrieval for instance, may be not adequate in our case. We need also to represent, in our lexical features, the position information of the words into a sentence.
Consequently, we defined parameters that describe both presence and position of interrogative words into the sentence. Our lexical features can be classified into three sub-categories :

- unigrams or bigrams present before a group of interrogative terms[7]
- the presence or absence of some interrogative terms in the utterance
- the unigrams present after a group of interrogative terms

These interrogative terms were chosen differently in each category and they are all specific for each language (French or Vietnamese). Moreover, to capture their positions (begin, middle or end in the sentence), two special tags “BEGIN” and “END” were added surrounding each sentence. In this way, both the presence and the position of interrogative words are correctly modeled by our lexical features.

4. Crosslingual Experiments

4.1. Experimental framework

4.1.1. French telephone meeting corpus: DELOC

Our telephone meeting corpus (called Deloc for “delocalized meetings”) is made up of 13 meetings of 15 to 60 minutes, involving 3 to 5 speakers (spontaneous speech). The total duration is around 7 hours and the language is French.
Different types of meetings were collected which correspond to three categories: recruitment interviews; project discussions in a research team; and brainstorming-style talking.
From this corpus, we have manually extracted a subset composed of 468 utterances: 234 question (Q) utterances and 234 non-question (NQ) utterances.

4.1.2. French client / agent dialog corpus : NESPOLE

The NESPOLE project was a common EU NSF funded project exploring future applications of automatic speech-to-speech translation in e-commerce and e-service sectors. The scenario of NESPOLE involves an Italian speaking agent, located in a tourism agency in Italy discussing with a client

[3] http://www.fon.hum.uva.nl/praat/
[4] for the experiments reported here, the forced alignment was done using the verbatim transcriptions.

[5] why
[6] how
[7] among these terms we find for instance for French : *pourquoi, qui, quand, comment, combien, où…*

(English, German or French speaking) located anywhere via Internet and using audio-conferencing tools like NetMeeting. The client wants to organize a trip in the Trentino (Italia) area, and asks the agent for information concerning his trip. More information on this database can be found in [8]. We use in this experimentation a subset of the French-speaking part of this corpus; it consists in 650 Q- and 650 NQ-sentences

4.1.3. Vietnamese databases : VietP and Assimil

The *VietP* corpus was used in our early study on Vietnamese prosody [7]. This corpus is made up of 14 pairs of Q/NQ sentences which were extracted from significant dialogues. Each dialogue is repeated five times by six native speakers (3 men and 3 women) from Hanoi (the North region, considered to be official pronunciation of the Vietnamese language). The corpus is finally composed of 420 Q- and 420 NQ-sentences.
The *Assimil* corpus was extracted from the "Assimil language learning CDrom" for Vietnamese and contains 168 Q- and 168 NQ-sentences. It is used in the prosodic-lexical experiments for Vietnamese since the *VietP* corpus transcriptions are too poor (only 14 different Q/NQ pairs) to be used in experiments where lexical information is taken into account.

4.1.4. Decision tree-based classifier

Traditionally, statistical-based methods such as Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM) can be used to solve classification problems in speech processing. These statistical methods generally apply on short term features, extracted for instance at a 10ms frame rate. However, in our case, statistical methods are hard to use since we do not use short term feature vectors, as explained in section 2.1: one feature vector only is extracted for the whole utterance to be classified, which excludes the use of conventional statistical classifiers.
Thus, decision trees, which correspond to another classical machine learning (ML) method [9,10] are a good alternative.
Decision tree is a divide-and-conquer approach to the problem of learning from a set of independent examples (a concrete example is called instance). Nodes in a decision tree involve testing a particular condition, which usually compares an attribute value with a constant. Some other trees compare two attributes with each other, or utilize some functions of one or more attributes. Leaf nodes give a classification for all instances that satisfy all conditions leading to this leaf, or a set of classifications, or a probability distribution over all possible classifications. To classify an unknown instance, it is routed down the tree according to the values of attributes tested in successive nodes, until it reaches a leaf. The instance is classified according to the class assigned to this leaf.
For this work, we have used an implementation of decision-tree algorithms that is included in the open-source toolkit Weka[8] which is a collection of algorithm implementations written in Java for data mining tasks such as classification, regression, clustering, and association rules.

4.1.5. Protocol and evaluation measure

In all of these experiments and for each corpus, we use 10-folds cross validation procedure: the whole corpus is equally and randomly divided in 10 parts. In the first repetition, the first part is taken apart and used for test data while the 9 remaining parts are used for training data. In each of these next repetitions (from 2nd to 10th repetition), the test data is changed to the 2nd part, 3rd part…10th part while all remaining 9 parts are for training data. The performance given for each experiment is then the mean of the scores obtained for each training / test configuration.

For the training data, a decision tree is constructed (the decision-tree algorithm used in our experiments is called "C4.5") and the obtained classifier is evaluated on the remaining test data. The evaluation is based on measures coming from the information retrieval domain such as recall (R), precision (P) and F-ratio, where :

$$R=\frac{N_{\text{correctly detected questions}}}{N_{\text{total questions in the test set}}} \qquad P=\frac{N_{\text{correctly detected questions}}}{N_{\text{total questions detected}}}$$

$$FRatio=\frac{2P\cdot R}{P+R}$$

The final detection performance is thus measured as the mean of F-ratio on test data for all 10-folds cross validation.

4.1.6. Crosslingual experiments results

Table 2: Crosslingual experiments for automatic question detection using prosodic features

	Feature set FR (Fratio)	Feature set VN (Fratio)
Deloc database (FR)	**74%**	67%
Nespole database (FR)	**73%**	68%
VietP database (VN)	77%	**81%**

Table 2 shows our crosslingual experiments results where both baseline (French) and adapted (Vietnamese) feature sets are evaluated on french and vietnamese test sets. The results show that the use of the matched feature set is important to optimize performance: while baseline feature sets are optimal for non tonal languages, it is important to use optimized features for the Vietnamese tonal language (81% Fratio instead of 77% for Vietnamese on *VietP* database). It may be found surprising that the overall performance is higher on the VietP database than on the other corpora. The main explanation is that both French corpora (FR) are made of utterances extracted from real dialogs or meetings while the VietP data (VN) was recorded in a much more controlled situation.

5. Combining lexical and prosodic features

The combination of prosodic and lexical features was done both in early and late integration fashion:
-in the early integration, both prosodic and lexical feature sets where merged into a single vector during training and a new decision tree was built. The same process was applied to the feature set during testing before questioning the decision tree.
-in the late integration, both lexical and prosodic features were used separately to train two different decision trees during training. During the test, each prosodic and lexical feature set was sent to the corresponding decision tree and the

[8] http://www.cs.waikato.ac.nz/~ml/weka/

final decision was based on a composite score which was a linear combination between prosodic and lexical scores.

Table 3 summarizes the results obtained for the features used separately or combined. For this experiment, in order to test the potential of the approach without taking into account the quality of the ASR recognizer that would be used to decode the utterances, we used the verbatim transcription of the test utterances for the lexical approach. We are aware that a noisy ASR output may have a bad influence on the lexical part of the combined approach.

The results show that the combined approach may be interesting to slightly improve the overall question detection procedure. However, further investigation is needed to 1) have conclusive idea of the better combination to use (differences between early and late integration are not always significant) 2) check if the improvement remains when using a noisy ASR output to get the lexical features. It is also interesting to note that the lexical features seem more important, compared to the prosodic features, for the tonal language. One explanation may be that, for tonal languages, the prosody also contributes to the tone applied to each syllable, and is consequently less used to encode other informations like sentence type.

Table 3: Combining prosodic and lexical features for automatic question detection (%Fratio)

	Prosodic feat.	Lexical feat.	Combined (early integration)	Combined (late integration)
Deloc database (FR)	74%	58%	77%	**78%**
Nespole database (FR)	73%	65%	**77%**	75%
Assimil database (VN)	64%	81%	82%	**88%**

6. Conclusion

This paper presented our ongoing work concerning automatic question detection. First, we proposed to use a new prosodic feature set which takes into account the specific prosodic characteristics of tonal language (like Vietnamese) and have shown that it can lead to better results than a feature set originally developed for a non-tonal language (French). We have also shown that, despite the fact that in a tonal language, the prosody is complex due to the presence of lexical tones, an automatic question sentence detection from the speech signal can be obtained with satisfying performance (81%). Secondly, it was shown that a better performance can be obtained by combining prosodic and lexical features. Future works must include : identifying the best way to combine these two models, seeing how our new prosodic features generalize to other tonal languages, using lexical features extracted from an ASR output instead of verbatim transcriptions and evaluate our question detection system on a continuous audio flaw using an automatic utterance segmenter.

7. References

[1] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M. & Van Ess-Dykema, C. “Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?” *Language and Speech 41*, pp. 439-487, 1998.

[2] Vu M.Q., Castelli E., Boucher A. & Besacier L. “Classification de parole en Question et Non-Question par arbre de décision” *SFC 05, 12èmes Rencontres de la Société Francophone de Classification* - Montréal, 2005

[3] Rossi M. “*L’intonation, le système du français : description et modélisation*” Editions Ophrys, 1999, ISBN : 2-7080-0912-5.

[4] Hirst, D.J. &Di Cristo, A. (Eds.) “*Intonation Systems. A Survey of 20 Languages*” Cambridge: Cambridge University Press.

[5] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” Computational Linguistics, vol. 26, pp. 339–373, 2000.

[6] Yuan, J.; Shih, C; Kochanski, G. P., Comparison of Declarative and Interrogative Intonation in Chinese. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, pp. 711-714.

[7] Vu M.Q., Tran D.D. & Castelli E. (2006) *Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results.* Interspeech2006 ICSLP, International Conference on Spoken Language Processing. 17-21 September 2006, Pittsburgh PA, USA

[8] Mana N., Burger S., Cattoni R., Besacier L., Maclaren V., Macdonough J., Metze F., "The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains" Eurospeech 2003, Geneva, 1-4 Sept. 2003.

[9] Marquez L., “Machine learning and Natural Language processing”, Technical Report LSI-00-45-R, Universitat Politechnica de Catalunya, 2000.

[10] Witten I.H., Frank E., Data mining: Pratical machine learning tools and techniques with Java implementations, Morgan Kaufmann, 1999.

Annexe B. Les arbres de décision construits par Weka

B.1. L'arbre de décision du corpus Deloc

La figure suivante est un arbre de décision construit par le logiciel WEKA sur les données du corpus DELOC en français :

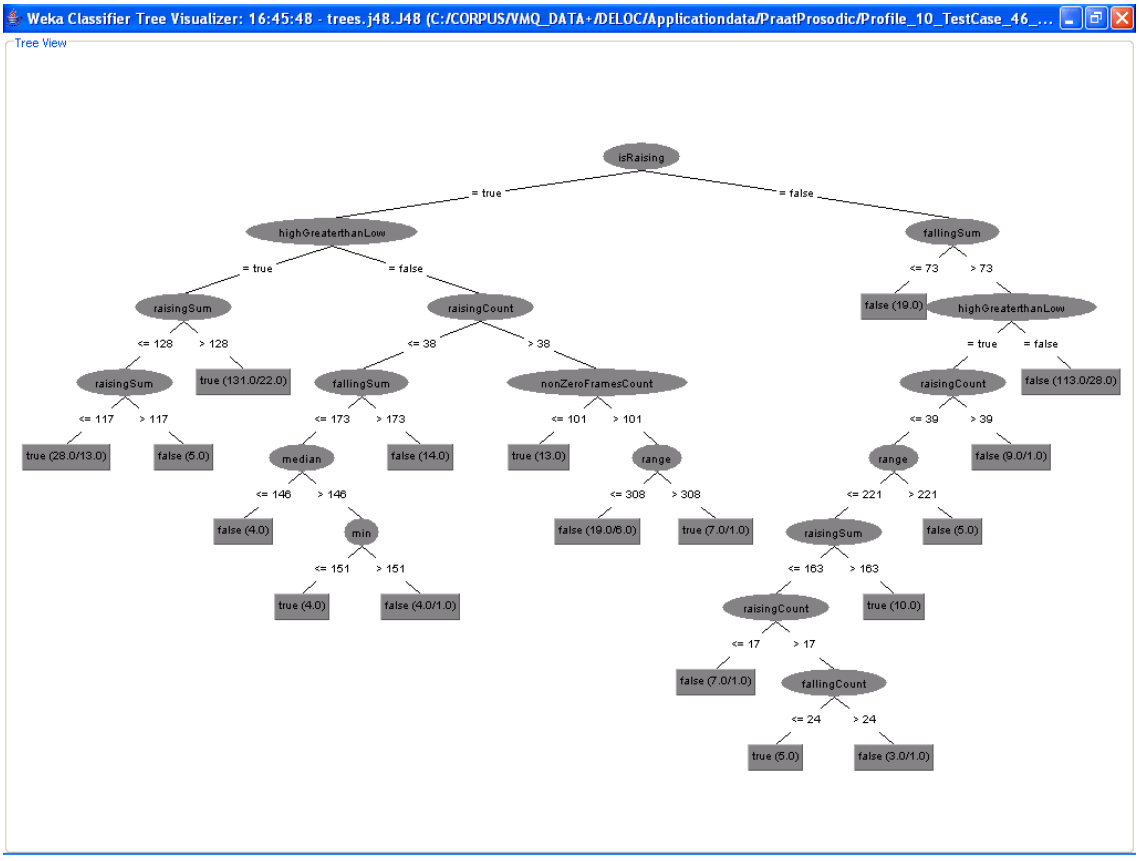


Figure 39 : Exemple d'image de l'arbre de décision du corpus Deloc

Voici la version textuelle de cet arbre :

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     C:/CORPUS/VMQ_DATA+/DELOC/Applicationdata/PraatProsodic/
Profile_10_TestCase_46_For_TRAINING.arff
Instances:    400
Attributes:   13
              min
              max
              mean
              median
              range
              highGreaterthanLow
              raisingSum
              raisingCount
              fallingSum
              fallingCount
              isRaising
              nonZeroFramesCount
              isQuestion
Test mode:    user supplied test set: 438 instances

=== Classifier model (full training set) ===

J48 pruned tree
------------------

isRaising = true
|   highGreaterthanLow = true
|   |   raisingSum <= 128
|   |   |   raisingSum <= 117: true (28.0/13.0)
|   |   |   raisingSum > 117: false (5.0)
|   |   raisingSum > 128: true (131.0/22.0)
|   highGreaterthanLow = false
|   |   raisingCount <= 38
|   |   |   fallingSum <= 173
|   |   |   |   median <= 146: false (4.0)
|   |   |   |   median > 146
|   |   |   |   |   min <= 151: true (4.0)
|   |   |   |   |   min > 151: false (4.0/1.0)
|   |   |   fallingSum > 173: false (14.0)
|   |   raisingCount > 38
|   |   |   nonZeroFramesCount <= 101: true (13.0)
|   |   |   nonZeroFramesCount > 101
|   |   |   |   range <= 308: false (19.0/6.0)
|   |   |   |   range > 308: true (7.0/1.0)
isRaising = false
|   fallingSum <= 73: false (19.0)
|   fallingSum > 73
```

```
|  |  highGreaterthanLow = true
|  |  |  raisingCount <= 39
|  |  |  |  range <= 221
|  |  |  |  |  raisingSum <= 163
|  |  |  |  |  |  raisingCount <= 17: false (7.0/1.0)
|  |  |  |  |  |  raisingCount > 17
|  |  |  |  |  |  |  fallingCount <= 24: true (5.0)
|  |  |  |  |  |  |  fallingCount > 24: false (3.0/1.0)
|  |  |  |  |  raisingSum > 163: true (10.0)
|  |  |  |  range > 221: false (5.0)
|  |  |  raisingCount > 39: false (9.0/1.0)
|  |  highGreaterthanLow = false: false (113.0/28.0)

Number of Leaves  : 18

Size of the tree : 35


Time taken to build model: 0.06 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         312               71.2329 %
Incorrectly Classified Instances       126               28.7671 %
Kappa statistic                          0.2043
Mean absolute error                      0.3677
Root mean squared error                  0.48
Relative absolute error                 73.533  %
Root relative squared error             96.0022 %
Total Number of Instances              438

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.623    0.275     0.237     0.623     0.344    true
  0.725    0.377     0.933     0.725     0.816    false

=== Confusion Matrix ===

   a   b   <-- classified as
  33  20 |   a = true
 106 279 |   b = false
```

B.2. L'arbre de décision du corpus VietP

Voici un exemple d'un arbre construit sur le corpus VietP. Ici n'est présentée que la version textuelle de cet arbre (car il y a un grand nombre de nœud dans la version graphique que la taille de l'arbre dépasse la taille de ce manuscrit).

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:
C:\CORPUS\VN_QUESTION_PROSODY_CORPUS_VERSION_2\Applicationdata\
12ParamF0AndIntenIndependentLocuteur/
OnlyFOAndIntensityFeatures_CVFold_9_For_TRAINING.arff
Instances:    756
Attributes:   13
              minF0
              maxF0
              rangeF0
              moyenF0
              moyenF0OfDS
              moyenF0OfFS
              moyenF0OfFMDHL
              lastDemiSylHL
              minInten
              maxInten
              moyenInten
              rangeInten
              isQuestion
Test mode:    user supplied test set: 84 instances

=== Classifier model (full training set) ===

J48 pruned tree
------------------

lastDemiSylHL <= -6
|   moyenF0OfDS <= 245.461538: nonquestion (166.0/10.0)
|   moyenF0OfDS > 245.461538
|   |   moyenF0 <= 234.7375: nonquestion (7.0)
|   |   moyenF0 > 234.7375
|   |   |   moyenInten <= 65.9
|   |   |   |   rangeF0 <= 71: question (2.0)
|   |   |   |   rangeF0 > 71
|   |   |   |   |   minF0 <= 203
|   |   |   |   |   |   moyenF0OfFS <= 214.704545: nonquestion (3.0)
|   |   |   |   |   |   moyenF0OfFS > 214.704545: question (4.0/1.0)
|   |   |   |   |   minF0 > 203: nonquestion (6.0)
|   |   |   moyenInten > 65.9: question (11.0/2.0)
```

```
lastDemiSylHL > -6
|  lastDemiSylHL <= 40
|  |  moyenF0 <= 116.431818
|  |  |  maxInten <= 71: nonquestion (11.0)
|  |  |  maxInten > 71
|  |  |  |  maxInten <= 72: question (3.0)
|  |  |  |  maxInten > 72: nonquestion (6.0)
|  |  moyenF0 > 116.431818
|  |  |  moyenInten <= 61.181818
|  |  |  |  minF0 <= 196: nonquestion (13.0/1.0)
|  |  |  |  minF0 > 196: question (6.0)
|  |  |  moyenInten > 61.181818
|  |  |  |  lastDemiSylHL <= -3
|  |  |  |  |  moyenF0OfDS <= 123.697368: nonquestion (4.0)
|  |  |  |  |  moyenF0OfDS > 123.697368
|  |  |  |  |  |  moyenF0OfFMDHL <= 0.217949
|  |  |  |  |  |  |  maxInten <= 70: question (2.0)
|  |  |  |  |  |  |  maxInten > 70
|  |  |  |  |  |  |  |  maxInten <= 74
|  |  |  |  |  |  |  |  |  moyenF0OfFMDHL <= -27.333333
|  |  |  |  |  |  |  |  |  |  minF0 <= 190: nonquestion (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  minF0 > 190: question (3.0)
|  |  |  |  |  |  |  |  |  moyenF0OfFMDHL > -27.333333: nonquestion (9.0/1.0)
|  |  |  |  |  |  |  |  maxInten > 74: question (2.0)
|  |  |  |  |  |  moyenF0OfFMDHL > 0.217949: question (13.0)
|  |  |  |  lastDemiSylHL > -3
|  |  |  |  |  moyenF0OfDS <= 137.04
|  |  |  |  |  |  maxInten <= 72: question (32.0/1.0)
|  |  |  |  |  |  maxInten > 72
|  |  |  |  |  |  |  moyenF0OfFMDHL <= -1.554762: question (11.0/1.0)
|  |  |  |  |  |  |  moyenF0OfFMDHL > -1.554762
|  |  |  |  |  |  |  |  minF0 <= 110: nonquestion (27.0/8.0)
|  |  |  |  |  |  |  |  minF0 > 110: question (4.0)
|  |  |  |  |  moyenF0OfDS > 137.04
|  |  |  |  |  |  rangeInten <= 23
|  |  |  |  |  |  |  minF0 <= 201
|  |  |  |  |  |  |  |  rangeInten <= 19: nonquestion (2.0)
|  |  |  |  |  |  |  |  rangeInten > 19
|  |  |  |  |  |  |  |  |  minInten <= 49
|  |  |  |  |  |  |  |  |  |  moyenF0OfFMDHL <= -24.03337: question (3.0)
|  |  |  |  |  |  |  |  |  |  moyenF0OfFMDHL > -24.03337
|  |  |  |  |  |  |  |  |  |  |  rangeF0 <= 71: question (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  rangeF0 > 71: nonquestion (5.0)
|  |  |  |  |  |  |  |  |  minInten > 49
|  |  |  |  |  |  |  |  |  |  rangeInten <= 22: question (5.0)
|  |  |  |  |  |  |  |  |  |  rangeInten > 22
|  |  |  |  |  |  |  |  |  |  |  moyenInten <= 65.04918: nonquestion (2.0)
|  |  |  |  |  |  |  |  |  |  |  moyenInten > 65.04918: question (4.0)
|  |  |  |  |  |  |  minF0 > 201: question (18.0/1.0)
|  |  |  |  |  |  rangeInten > 23: question (185.0/10.0)
```

```
|  lastDemiSylHL > 40
|  |  moyenF0OfDS <= 239.545455: nonquestion (94.0/10.0)
|  |  moyenF0OfDS > 239.545455
|  |  |  minInten <= 51
|  |  |  |  maxInten <= 73
|  |  |  |  |  rangeInten <= 25
|  |  |  |  |  |  rangeInten <= 24
|  |  |  |  |  |  |  maxInten <= 69: question (8.0/1.0)
|  |  |  |  |  |  |  maxInten > 69
|  |  |  |  |  |  |  |  moyenInten <= 64.040816: nonquestion (9.0)
|  |  |  |  |  |  |  |  moyenInten > 64.040816
|  |  |  |  |  |  |  |  |  rangeInten <= 22: question (4.0)
|  |  |  |  |  |  |  |  |  rangeInten > 22
|  |  |  |  |  |  |  |  |  |  moyenInten <= 66.461538
|  |  |  |  |  |  |  |  |  |  |  rangeInten <= 23: nonquestion (2.0)
|  |  |  |  |  |  |  |  |  |  |  rangeInten > 23
|  |  |  |  |  |  |  |  |  |  |  |  rangeF0 <= 102: question (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  rangeF0 > 102: nonquestion (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  moyenInten > 66.461538: question (5.0)
|  |  |  |  |  |  rangeInten > 24: question (13.0/2.0)
|  |  |  |  |  rangeInten > 25: nonquestion (8.0/1.0)
|  |  |  |  maxInten > 73: question (23.0/2.0)
|  |  |  minInten > 51: nonquestion (9.0/1.0)

Number of Leaves  : 44

Size of the tree : 87


Time taken to build model: 0.15 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          77               91.6667 %
Incorrectly Classified Instances         7                8.3333 %
Kappa statistic                          0.8333
Mean absolute error                      0.1462
Root mean squared error                  0.2938
Relative absolute error                 29.2376 %
Root relative squared error             58.7574 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   Class
                 0.905    0.071      0.927     0.905     0.916    question
                 0.929    0.095      0.907     0.929     0.918    nonquestion

=== Confusion Matrix ===
```

```
 a  b   <-- classified as
38  4 |  a = question
 3 39 |  b = nonquestion
```

Annexe C. Liste complète des paramètres lexicaux

C.1. Les paramètres lexicaux pour le corpus en langue française

Nous avons au total 72 paramètres lexicaux pour le français :

No	Nom du paramètre	Description
1	OneWordBefore_pourquoi	Le mot avant le terme *pourquoi*
2	TwoWordBefore_pourquoi	Les deux mots avant le terme *pourquoi*
3	OneWordBefore_qui	Le mot avant le terme *qui*
4	TwoWordBefore_qui	Les deux mots avant le terme *qui*
5	OneWordBefore_quand	Le mot avant le terme *quand*
6	TwoWordBefore_quand	Les deux mots avant le terme *quand*
7	OneWordBefore_pour_quand	Le mot avant le terme *pour quand*
8	TwoWordBefore_pour_quand	Les deux mots avant le terme *pour quand*
9	OneWordBefore_comment	Le mot avant le terme *comment*
10	TwoWordBefore_comment	Les deux mots avant le terme *comment*
11	OneWordBefore_combien	Le mot avant le terme *combien*
12	TwoWordBefore_combien	Les deux mots avant le terme *combien*
13	OneWordBefore_pour_combien	Le mot avant le terme pour *combien*

14	TwoWordBefore_pour_combien	Les deux mots avant le terme *pour combien*
15	OneWordBefore_de_combien	Le mot avant le terme *de combien*
16	TwoWordBefore_de_combien	Les deux mots avant le terme *de combien*
17	OneWordBefore_où	Le mot avant le terme *où*
18	TwoWordBefore_où	Les deux mots avant le terme *où*
19	OneWordBefore_quel	Le mot avant le terme *quel*
20	TwoWordBefore_quel	Les deux mots avant le terme *quel*
21	OneWordBefore_quelle	Le mot avant le terme *quelle*
22	TwoWordBefore_quelle	Les deux mots avant le terme *quelle*
23	OneWordBefore_quels	Le mot avant le terme *quels*
24	TwoWordBefore_quels	Les deux mots avant le terme *quels*
25	OneWordBefore_quelles	Le mot avant le terme *quelles*
26	TwoWordBefore_quelles	Les deux mots avant le terme *quelles*
27	OneWordBefore_de_quel	Le mot avant le terme *de quel*
28	TwoWordBefore_de_quel	Les deux mots avant le terme *de quel*
29	OneWordBefore_lequel	Le mot avant le terme *lequel*
30	TwoWordBefore_lequel	Les deux mots avant le terme *lequel*
31	OneWordBefore_laquelle	Le mot avant le terme *laquelle*
32	TwoWordBefore_laquelle	Les deux mots avant le terme *laquelle*
33	OneWordBefore_lesquels	Le mot avant le terme *lesquels*
34	TwoWordBefore_lesquels	Les deux mots avant le terme *lesquels*
35	OneWordBefore_lesquelles	Le mot avant le terme *lesquelles*
36	TwoWordBefore_lesquelles	Les deux mots avant le terme *lesquelles*
37	OneWordBefore_jusqu_où	Le mot avant le terme *jusqu où*
38	TwoWordBefore_jusqu_où	Les deux mots avant le *terme jusqu' où*
39	Present_je_voudrais_savoir	y-a-t-il dans la phrase ce terme *je voudrais savoir*
40	Present_j_aimerais_savoir	y-a-t-il dans la phrase ce terme *j'aimerais savoir*
41	Present_j_voudrais_savoir	y-a-t-il dans la phrase ce terme *j'voudrais savoir*

42	Present_j_aimerais_vous_demander	y-a-t-il dans la phrase ce terme *j'aimerais vous demander*
43	Present_je_voudrais_vous_demander	y-a-t-il dans la phrase ce terme *je voudrais vous demander*
44	Present_j_voudrais_vous_demander	y-a-t-il dans la phrase ce terme *j'voudrais vous demander*
45	Present_je_voudrais_vous_d_mander	y-a-t-il dans la phrase ce terme *je voudrais vous d'mander*
46	Present_j_voudrais_vous_d_mander	y-a-t-il dans la phrase ce terme *j'voudrais vous d'mander*
47	Present_est-ce_que	y-a-t-il dans la phrase ce terme *est-ce que*
48	Present_est-ce_qu_il	y-a-t-il dans la phrase ce terme *est-ce qu'il*
49	Present_est-ce_qu_elle	y-a-t-il dans la phrase ce terme *est-ce qu'elle*
50	Present_est-ce_qu_ils	y-a-t-il dans la phrase ce terme *est-ce qu'ils*
51	Present_est-ce_qu_elles	y-a-t-il dans la phrase ce terme *est-ce qu'elles*
52	Present_qu_est-ce_que	y-a-t-il dans la phrase ce terme *qu'est-ce que*
53	Present_qu_est-ce_qu_il	y-a-t-il dans la phrase ce terme *qu'est-ce qu'il*
54	Present_qu_est-ce_qu_elle	y-a-t-il dans la phrase ce terme *qu'est-ce qu'elle*
55	Present_qu_est-ce_qu_ils	y-a-t-il dans la phrase ce terme *qu'est-ce qu'ils*
56	Present_qu_est-ce_qu_elles	y-a-t-il dans la phrase ce terme *qu'est-ce qu'elles*
57	Present_qu_est_qui	y-a-t-il dans la phrase ce terme *qu'est qui*
58	OneWordAfter_n_est-ce_pas	Le mot après le terme *n'est-ce pas*
59	OneWordAfter_pardon	Le mot après le terme *pardon*
60	OneWordAfter_ah_bon	Le mot après le terme *ah bon*
61	OneWordAfter_qui	Le mot après le terme *qui*
62	OneWordAfter_quand	Le mot après le terme *quand*
63	OneWordAfter_pour_quand	Le mot après le terme *pour quand*
64	OneWordAfter_comment	Le mot après le terme *comment*
65	OneWordAfter_combien	Le mot après le terme *combien*
66	OneWordAfter_pour_combien	Le mot après le terme *pour combien*

67	OneWordAfter_de_combien	Le mot après le terme *de combien*
68	OneWordAfter_où	Le mot après le terme *où*
69	OneWordAfter_en_quoi	Le mot après le terme *en quoi*
70	OneWordAfter_pourquoi	Le mot après le terme *pourquoi*
71	OneWordAfter_allô	Le mot après le terme *allô*

C.2. Les paramètres lexicaux pour le corpus en langue vietnamienne

Nous avons au total 83 paramètres lexicaux pour le vietnamien :

No	Nom du paramètre	Description
1	OneWordBefore_ai	Le mot avant le terme *ai*
2	TwoWordBefore_ai	Les deux mots avant le terme *ai*
3	OneWordBefore_còn	Le mot avant le terme *còn*
4	TwoWordBefore_còn	Les deux mots avant le terme *còn*
5	OneWordBefore_mấy	Le mot avant le terme *mấy*
6	TwoWordBefore_mấy	Les deux mots avant le terme *mấy*
7	OneWordBefore_sao	Le mot avant le terme *sao*
8	TwoWordBefore_sao	Les deux mots avant le terme *sao*
9	OneWordBefore_có_tin	Le mot avant le terme *có tin*
10	TwoWordBefore_có_tin	Les deux mots avant le terme *có tin*
11	OneWordBefore_thế_nào	Le mot avant le terme *thế nào*
12	TwoWordBefore_thế_nào	Les deux mots avant le terme *thế nào*
13	OneWordBefore_thế	Le mot avant le terme *thế*
14	TwoWordBefore_thế	Les deux mots avant le terme *thế*
15	OneWordBefore_đâu	Le mot avant le terme *đâu*
16	TwoWordBefore_đâu	Les deux mots avant le terme *đâu*
17	OneWordBefore_hay	Le mot avant le terme *hay*
18	TwoWordBefore_hay	Les deux mots avant le terme *hay*
19	OneWordBefore_tại_sao	Le mot avant le terme *tại sao*
20	TwoWordBefore_tại_sao	Les deux mots avant le terme *tại sao*

21	Present_Bác_có_biết	y-a-t-il dans la phrase ce terme *Bác có biết*
22	Present_phải_không	y-a-t-il dans la phrase ce terme *phải không*
23	Present_có_đúng_là	y-a-t-il dans la phrase ce terme *có đúng là*
24	Present_cái_gì	y-a-t-il dans la phrase ce terme *cái gì*
25	Present_bác_có_biết	y-a-t-il dans la phrase ce terme *bác có biết*
26	Present_bác_có_biết_không	y-a-t-il dans la phrase ce terme *bác có biết không*
27	Present_hay_là	y-a-t-il dans la phrase ce terme *hay là*
28	Present_sao_không	y-a-t-il dans la phrase ce terme *sao không*
29	Present_sao_lại	y-a-t-il dans la phrase ce terme *sao lại*
30	Present_gì	y-a-t-il dans la phrase ce terme *gì*
31	Present_chi	y-a-t-il dans la phrase ce terme *chi*
32	Present_thế_nào	y-a-t-il dans la phrase ce terme *thế nào*
33	Present_bao_nhiêu	y-a-t-il dans la phrase ce terme *bao nhiêu*
34	Present_bao_xa	y-a-t-il dans la phrase ce terme *bao xa*
35	Present_bao_lâu	y-a-t-il dans la phrase ce terme *bao lâu*
36	Present_bao_giờ	y-a-t-il dans la phrase ce terme *bao giờ*
37	Present_khi_nào	y-a-t-il dans la phrase ce terme *khi nào*
38	Present_mấy	y-a-t-il dans la phrase ce terme *mấy*
39	Present_nào	y-a-t-il dans la phrase ce terme *nào*
40	Present_có_nghĩa_là	y-a-t-il dans la phrase ce terme *có nghĩa là*
41	Present_đâu	y-a-t-il dans la phrase ce terme *đâu*
42	Present_lúc_nào	y-a-t-il dans la phrase ce terme *lúc nào*
43	Present_ai	y-a-t-il dans la phrase ce terme *ai*
44	Present_làm_sao	y-a-t-il dans la phrase ce terme *làm sao*
45	Present_ở_đâu	y-a-t-il dans la phrase ce terme *ở đâu*
46	OneWordAfter_không	Le mot après le terme *không*
47	OneWordAfter_gì	Le mot après le terme *gì*
48	OneWordAfter_gì_nhỉ	Le mot après le terme *gì nhỉ*

49	OneWordAfter_làm_sao	Le mot après le terme *làm sao*
50	OneWordAfter_làm_sao_nhỉ	Le mot après le terme *làm sao nhỉ*
51	OneWordAfter_thế	Le mot après le terme *thế*
52	OneWordAfter_thế_nào	Le mot après le terme *thế nào*
53	OneWordAfter_thế_nào_nhỉ	Le mot après le terme *thế nào nhỉ*
54	OneWordAfter_thế_sao	Le mot après le terme *thế sao*
55	OneWordAfter_không_đấy	Le mot après le terme *không đấy*
56	OneWordAfter_sao	Le mot après le terme *sao*
57	OneWordAfter_nào	Le mot après le terme *nào*
58	OneWordAfter_thì_sao	Le mot après le terme *thì sao*
59	OneWordAfter_nhỉ	Le mot après le terme *nhỉ*
60	OneWordAfter_là_gì	Le mot après le terme *là gì*
61	OneWordAfter_đâu	Le mot après le terme *đâu*
62	OneWordAfter_à	Le mot après le terme *à*
63	OneWordAfter_bao_nhiêu	Le mot après le terme *bao nhiêu*
64	OneWordAfter_cái_gì	Le mot après le terme *cái gì*
65	OneWordAfter_chưa	Le mot après le terme *chưa*
66	OneWordAfter_vậy	Le mot après le terme *vậy*
67	OneWordAfter_chứ	Le mot après le terme *chứ*
68	OneWordAfter_phải_không	Le mot après le terme *phải không*
69	OneWordAfter_phải_không_bác	Le mot après le terme *phải không bác*
70	OneWordAfter_phải_không_anh	Le mot après le terme *phải không anh*
71	OneWordAfter_rồi_sao	Le mot après le terme *rồi sao*
72	OneWordAfter_thì_sao	Le mot après le terme *thì sao*
73	OneWordAfter_chắc	Le mot après le terme *chắc*
74	OneWordAfter_mà	Le mot après le terme *mà*
75	OneWordAfter_á	Le mot après le terme *á*
76	OneWordAfter_hả	Le mot après le terme *hả*
77	OneWordAfter_hả_bác	Le mot après le terme *hả bác*
78	OneWordAfter_hả_anh	Le mot après le terme *hả anh*

79	OneWordAfter_sao_vậy	Le mot après le terme *sao vậy*
80	OneWordAfter_mấy	Le mot après le terme *mấy*
81	OneWordAfter_vậy_sao	Le mot après le terme *vậy sao*
82	OneWordAfter_nhé	Le mot après le terme *nhé*
83	OneWordAfter_chi	Le mot après le terme *chi*

Annexe D. Les boites à outil Praat, Weka, Sphinx4, Eclipse RCP

Dans cette annexe, nous présentons les différents boites à outil qui ont été utilisées dans notre recherche, particulièrement dans les expérimentations et dans le programme de reconnaissance des phrases interogatives Coco présenté dans le 5.3.

D.1. Praat

Dans cette annexe, nous présentons les différents scripts Praat que nous avons utilisés extensivement dans notre étude. Le logiciel Praat est un outil très performant pour le traitement de parole [Boersma, 2005]. Il est équipé d'un langage script qui permet l'utilisateur de simuler les activités choisir-activer de la souris sur l'interface graphique du programme. De ce fait, nous pouvons automatiser, sauvegader, paramétrer les manipulations répétitives pour un usage facile et rapide. Un fichier script est un fichier textuel qui contient les instructions pour commander le logiciel Praat. Pour exécuter un fichier script *calculF0script.txt* :

Sous Linux :

```
> praat calculF0script.txt
```

Sous Window :

```
C:\> praat.exe calculF0script.txt
```

Ou :

```
C:\> praatcon.exe calculF0script.txt
```

L'exécutable *praatcon.exe* est une version « sans interface » de Praat, conçu spécialement pour exécuter les scripts.

Scripts pour calculer F0, intensité

Ce script est pour but de calculer F0 et intensité pour tous les fichiers .wav dans un répertoire. Après le calcul, chaque fichier .wav va avoir un fichier .F0.AC et un fichier .Intensity qui contiennent sur chaque ligne une valeur de temps et une valeur de F0 (ou d'intensité) correspondant à ce temps. Les valeurs F0 (ou intensité) peuvent être corrigées facilement dans ce fichier texte (corriger les sautes octaves de F0, par exemple) avant d'entrer dans les calculs des prochaines étapes.

```
#
# "I want a list of pitch and intensity values at the same times."
# Calculate for all wav files in source directory, write results to destination directory
#-------------------------------------------------------------------------
# BEGIN MAIN PROGRAM
#-------------------------------------------------------------------------

#Change this param to meet your need
wavDirectory$ = "C:\TEMP\data"

#from here: intact
name_of_the_destination_pitch_directory$ = wavDirectory$
Create Strings as file list... list 'wavDirectory$'\*.wav
numberOfFiles = Get number of strings
for ifile to numberOfFiles
select Strings list
wavFileName$ = Get string... ifile
printline ['ifile'] Calculating for file: 'wavFileName$'

outF0filename$ = "'name_of_the_destination_pitch_directory$'\'wavFileName$'.F0.AC"
outIntensityfilename$ =
"'name_of_the_destination_pitch_directory$'\'wavFileName$'.Intensity"
Read from file... 'wavDirectory$'\'wavFileName$'

sound = selected ("Sound")
tmin = Get starting time
tmax = Get finishing time
To Pitch (ac)... 0.001 75 15 no 0.03 0.45 0.01 0.35 0.14 600
Rename... pitch
select sound
To Intensity... 75 0.001
Rename... intensity
for i to (tmax-tmin)/0.02
   time = tmin + i * 0.02
   #read pitch
   select Pitch pitch
   pt = Get value at time... time Hertz Linear
   if pt = undefined
```

```
pt = 0
   endif
   #read intensity
   select Intensity intensity
   inten = Get value at time... time Cubic
   if inten = undefined
inten = 0
   endif
   fileappend    '"outF0filename$"' 'time:2' 'pt:0"newline$'
   fileappend    '"outIntensityfilename$"' 'time:2' 'inten:0"newline$'
endfor
#remove objects
select sound
Remove
select Pitch pitch
Remove
select Intensity intensity
Remove
endfor
printline DONE.

#------------------------------------------------------------
# END MAIN PROGRAM
#------------------------------------------------------------
```

Scripts pour dessiner la courbe F0 en image

Ce script est pour but de dessiner la forme du son en paralelle avec le contour F0. Il lit un fichier .wav et un fichier .F0.AC qui contient les valeurs F0. Le script produit pour chaque fichier .wav un fichier .emf qui est un fichier image lisible par WindowXP ou certains autres programmes gratuits d'affichage d'image comme IrfanView par exemple.

```
#####################################################################
#Read wav file and pitch file to draw in a same image
#Used to create image in :
C:\CORPUS\VN_QUESTION_PROSODY_CORPUS\all_sentences
# BEGIN MAIN PROGRAM
###############

#Change this param to meet your need
name_of_the_source_wav_directory$="C:\TEMP\data"
name_of_the_transcription_directory$="C:\TEMP\data"
name_of_the_destination_image_directory$=name_of_the_source_wav_directory$

minpitch=100
maxpitch=400
```

```
#from here: intact
wavDirectory$ = name_of_the_source_wav_directory$
txtDirectory$ = name_of_the_transcription_directory$
Create Strings as file list... list 'wavDirectory$'\*.wav
numberOfFiles = Get number of strings
for ifile to numberOfFiles
select Strings list
wavFileName$ = Get string... ifile
printline ['ifile'] drawing file: ['wavDirectory$'\'wavFileName$']
#read transcription file
text$ < 'txtDirectory$'\'wavFileName$'.txt
#read pitch file
pitchFileName$ = wavFileName$+".F0.AC"
Read Strings from raw text file... 'wavDirectory$'\'pitchFileName$'
numberOfFrames = Get number of strings
nameOfPitchFileNameObject$ = selected$ ("Strings")
#we extract the pitch value here
for iPitch to numberOfFrames
select Strings 'nameOfPitchFileNameObject$'
# a line is as : timeValue pitchValue
line$ = Get string... iPitch
pitchValueIndex = rindex (line$," ")
#printline pitchValueIndex : 'pitchValueIndex'
timeValue$ = left$ (line$,pitchValueIndex)
time'iPitch' = 'timeValue$'
#printline timeValue: ['timeValue$']
lineLength = length (line$)
#printline length : 'lineLength'
pitchValueString$ = right$ (line$,lineLength - pitchValueIndex)
#printline pitchValue : ['pitchValueString$']
pitch'iPitch' = 'pitchValueString$'
printline ['ifile'-'iPitch'] Found pitch value: ['pitchValueString$']
endfor
#read wav file to draw
Read from file... 'wavDirectory$'\'wavFileName$'
duration = Get total duration
nameOfWavObject$ = selected$ ("Sound")
# Create top box for plotting speech waveform, measured and smoothed pitch contours
# coordinates of the first box (in inches)
Erase all
Black
lengthUnit = 1
xLeftOfBox1 = 1
xRightOfBox1 = 10
#xRightOfBox1 = lengthUnit * numberOfFrames
yTopOfBox1 = 1
yBottomOfBox1 = 5

Select inner viewport... xLeftOfBox1 xRightOfBox1 yTopOfBox1 yBottomOfBox1
Draw inner box
Text left... yes Pitch (Hz)
```

```
# Draw wave signal
Plain line
Draw... 0 0 0 0 no curve

Axes... 0 duration 'minpitch' 'maxpitch'
Marks left every... 1 100 yes yes no
#Marks bottom every... 1 0.2 yes yes no
One mark bottom... duration yes yes no

# plot pitch values
for iPitch to  numberOfFrames
y = pitch'iPitch'
x = time'iPitch'
Paint circle (mm)... Red  x y 2
endfor

##### draw texts into 2 top corners of the drawing
Blue
Select outer viewport... xLeftOfBox1 xRightOfBox1 yTopOfBox1-0.5 yTopOfBox1
# draw file name and transcription text on the left-top  corner
Viewport text... Left Bottom 0 'text$'
Viewport text... Right Bottom 0 'wavFileName$'

#Remove read objects
select Strings 'nameOfPitchFileNameObject$'
Remove
select Sound  'nameOfWavObject$'
Remove

# select both boxes to write to image file
Select inner viewport... xLeftOfBox1 xRightOfBox1 yTopOfBox1 yBottomOfBox1

# Write to Windows image file
Write to Windows metafile...
'name_of_the_destination_image_directory$'\'wavFileName$'.emf

endfor

printline DONE.

#------------------------------------------------------------
# END MAIN PROGRAM
#--------
```

D.2. Weka

Le logiciel libre Weka a été écrit en Java à l'Université de Waikato en Nouvelle Zelande par un groupe de chercheurs issus des domaines de l'apprentissage, la reconnaissance des formes et du data mining. Weka est un ensemble d'algorithmes d'apprentissage, de reconnaissance des formes et de data mining recouvrant les méthodes de classification supervisées et non supervisées. Une structure de tableau de données normalisé (le format ARFF) et des outils d'interrogation de bases de données permettent d'avoir une entrée unique. Actuellement Weka contient des modules de pré analyse et de visualisation, de classification, de régression et des algorithmes construisant des règles d'association.

Weka est un "open source software" sous licence publique GNU. En tant qu'un programme écrit entièrement en Java, il nous offre donc deux possibilités d'utilisation : soit nous pouvons exécuter Weka comme un programme indépendant « standard » et le programme présente son interface graphique comme dans la Figure 40, soit nous pouvons utiliser Weka en mode « embarqué ». Dans le mode « embarqué », nous pouvons incorporer Weka dans notre propre système, puis appeler et utiliser les fonctions de Weka. C'est le mode « embarqué » que nous avons utilisé tout au long de notre recherche, ainsi que dans le programme Coco. Il nous offre la possibilité de mener des expérimentations complexes sur un grand volume de données de manière tout automatique. Si les mêmes expérimentations devaient se réaliser avec son interface graphique, il serait fatigant de passer beaucoup de temps seulement pour cliquer et manipuler manuellement l'interface.

Figure 40 : Interface du programme Weka

Weka est composé de 4 outils d'analyse de données accessibles au lancement :

- l'utilisation des modules de calcul en ligne de commande: SIMPLE CLI. Pas très pratique au vu des nombreux paramètres à fournir pour l'utilisation des classes implémentées (bouton en haut à gauche).
- l'explorer qui permet l'analyse d'un jeu de données avec interface graphique de paramètrage et de visualisation : EXPLORER (bouton en haut à droite).
- l'expérimenter qui permet de configurer des expériences complètes et complexes d'analyse de plusieurs jeux de données par plusieurs méthodes de traitement différentes : EXPERIMENTER (bouton en bas à gauche).
- le knowledge flow qui permet d'effectuer des analyses via une interface graphique de gestion de composants graphiques associés aux différents traitement : KNOWLEDGEFLOW (bouton en bas à droite).

Nous nous intéressons tout particulièrement aux bases de fonctionnement de Weka et de l'analyse de données en utilisant l'EXPLORER. Il permet de pré-traiter des données, de les analyser à l'aide d'une méthode d'apprentissage et d'afficher le modèle résultant et ses performances.

WEKA traite des données au format ARFF (Attribute Relation Format File). Un certain nombre de jeux de données, au format ARFF, issus du site de l'UCI (UCI Machine Learning, University of California) peuvent être téléchargés à partir du même site de diffusion. Par exemple, le fichier weather.arff est listé ci-dessous :

```
@relation weather
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Ce fichier contient deux parties : les « en-tête » qui sont les lignes commencées par le caractère « @ », le reste du fichier constitue les données. Dans la partie « en tête », il faut noter que :

- @relation : défini le nom de ce jeu de données
- @attribute : défini le nom et le type des attributes dans ce jeu de données
- @data : signifi que l'entête est terminé, la section suivante contient les données.
- Les lignes après @data : chaque ligne représente un « instance ». Chaque « instance » est définie par les valeurs des attributes. Les valeurs manquantes sont représentées par une signe question « ? »

Dans WEKA, chaque méthode de transformation, de sélection d'attributs, d'apprentissage, de prédiction numérique, de clustering ou de découverte d'associations est implémentée par une classe Java.

D.3. Sphinx-4

Sphinx-4[18] est un système de reconnaissance vocale entièrement écrit dans le langage de programmation Java. Il a été créé conjointement par le groupe Sphinx à l'université Carnegie Mellon, les laboratoires Sun Microsystems et Hewlett-Packard [Walker, 2004].

Les buts de Sphinx sont d'avoir une reconnaissance vocale hautement flexible, d'égaler les autres produits commerciaux et de mettre en collaboration les centres de recherche de diverses universités, des laboratoires Sun, des laboratoires HP et du MIT.

Sphinx-4 est hautement configurable. La reconnaissance de Sphinx-4 supporte notamment les mots isolés et les phrases (utilisation de grammaires). L'architecture de Sphinx-4 est modulable pour permettre de nouvelles recherches et pour tester de nouveaux algorithmes. Le Figure 41 présente les principaux modules dans l'architecture générale de Sphinx-4 :

[18] http://cmusphinx.sourceforge.net/sphinx4

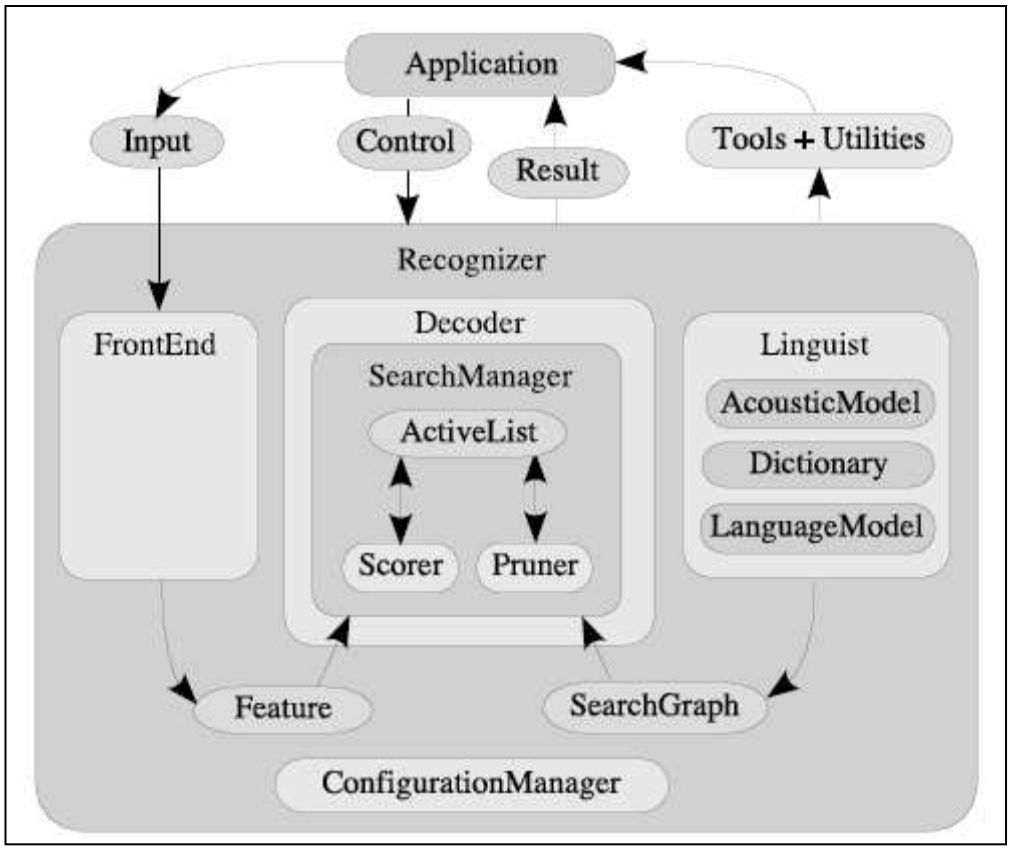


Figure 41 : Architecture générale de Sphinx-4

Front-End : Le Front-End découpe la voix enregistrée en différentes parties et les prépare pour le décodeur.

Linguist : Le linguist est la base de connaissance qui contient l'information qu'utilise le décodeur pour déterminer les mots et les phrases prononcées. Cette base de connaissance est composée :

- D'un dictionnaire :

  – Classification des mots.

  – Prononciation des mots (un mot peut avoir plusieurs prononciations).

  – Prononciation représentée comme des sons ou dans d'autres unités.

  – Peut varier en taille, de quelques mots à plusieurs centaines de milliers.

- D'un modèles acoustique.

- D'un modèle de langage :

– Décrit ce qui peut être dit dans un contexte bien spécial.

– Aide à rétrécir l'espace de recherche.

Il y a trois sortes de modèle de langage : le plus simple est utilisé pour les mots isolés, le deuxième pour les applications basées sur des commandes et des contrôles et le dernier pour le langage courant (modèle N-gram).

Décodeur : Le décodeur est le coeur de Sphinx-4. C'est lui qui traite les informations reçues depuis le Front-End, les analyse et les compare avec la base de connaissances pour donner un résultat à l'application.

Sphinx-4 a été compilé et testé sur Solaris, Mac OS X, Linux et Windows. L'exécution, la compilation et les tests de Sphinx-4 demandent des logiciels supplémentaires. Les logiciels suivants doivent être installés sur la machine :

- Java 2 SDK, Standard Edition 5.0. (http://java.sun.com.)
- Les différentes librairies qui composent Sphinx-4 (sphinx4-1.0beta-bin.rar). (http://cmusphinx.sourceforge.net/sphinx4/). Cette archive contient les différentes libraires (lib), la javadoc de Sphinx-4 et des démonstrations (sources et exécutables).

D.4. Eclipse RCP

Lors de sa création en 2001, le but du projet Eclipse était de fournir un socle pour la création d'environnements de développement. Eclipse a une structure modulable, basée sur des plug-in, qui l'a rendu très populaire au sein de la communauté des développeurs Java.

Avec le lancement d'Eclipse RCP en 2004, l'objectif du projet Eclipse a été étendu en prenant en compte l'utilisation du framework Eclipse pour tous les types d'applications clientes. Cette évolution a été demandée par une partie de la communauté des utilisateurs d'Eclipse qui avait commencé à réutiliser des portions d'Eclipse pour le développement d'applications clientes.

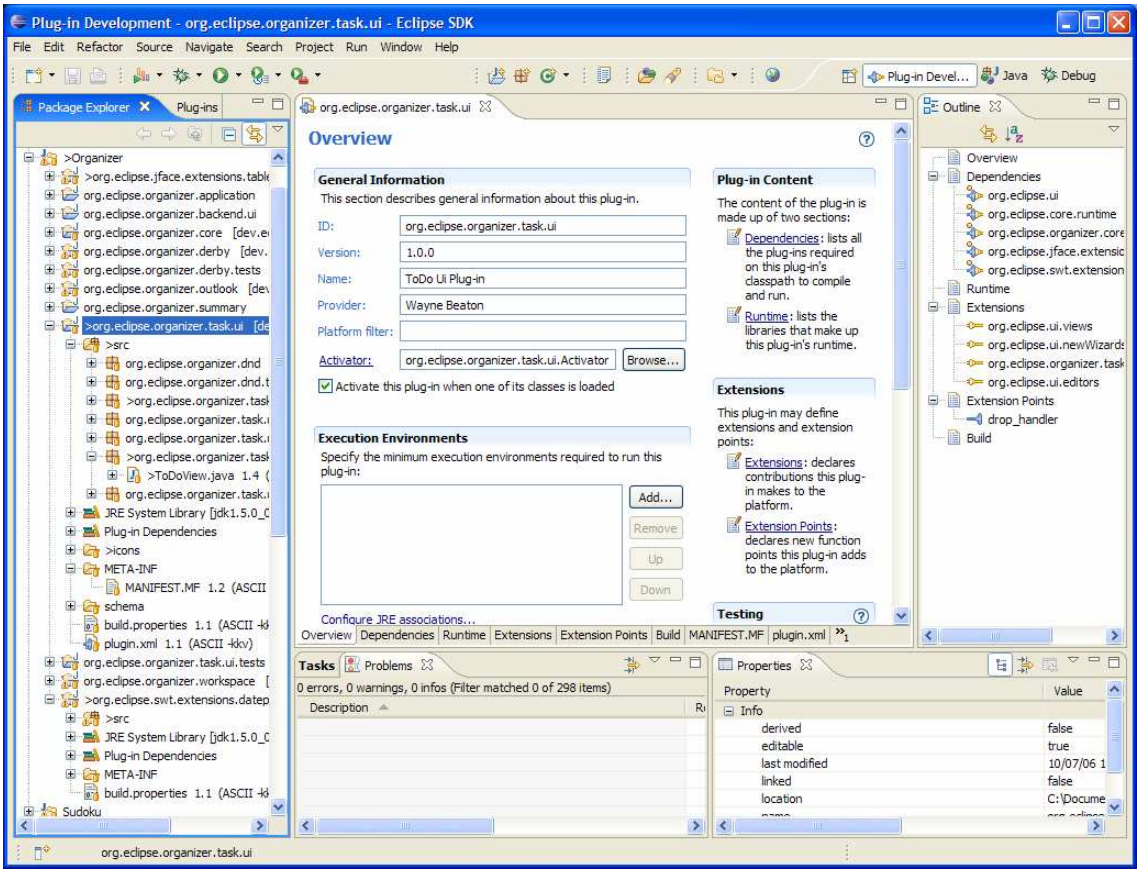


Figure 42 : Interface du "framework" Eclipse pour développer les applications Java et RCP

Eclipse RCP est, en quelque sorte, une version simplifiée de l'IDE Eclipse dont on a enlevé les modules couvrant les besoins spécifiques à un environnement de développement. Il est donc possible de réutiliser cette base pour développer des applications clientes dîtes riches (Le terme 'applications riches', apparu pendant les années 90, marque l'évolution par rapport aux applications accessibles par des terminaux passifs. Ces applications riches n'ont pas de contraintes d'ergonomie, de rapidité, de complexité ou encore d'intégration aux outils bureautiques comme les applications de terminaux).

Eclipse RCP apporte des solutions aux deux problèmes principaux de ces architectures : la difficulté de distribution de l'application sur les postes utilisateurs et la forte dépendance à des technologies propriétaires. En plus, il propose des avantages évidents pour l'utilisateur : réactivé, richesse et qualité des interfaces graphiques, souplesse avec la possibilité de fonctionner en mode déconnecté, intégration avec les autres applications installées sur le poste client…

C'est la qualité d'Eclipse qui est la raison principale qui a amené des utilisateurs d'Eclipse à le considérer comme socle pour des applications clientes. L'environnement de développement Eclipse s'est notamment imposé par sa fiabilité et la qualité de ses interfaces graphiques. L'utilisation massive de l'environnement de développement Eclipse par la communauté des développeurs Java ainsi que son utilisation comme base de produits commerciaux (WebSphere

Studio, SAP NetWeaver Studio...) ont permis d'éprouver le framework Eclipse. Le framework Eclipse est disponible en open-source.

De nombreux exemples d'applications basées sur Eclipse RCP sont disponibles sur le site Eclipse, y compris les applications open-sources et les applications commerciales. Des entreprises françaises comme l'AFP et le CNCC, ou américaines comme la banque JP Morgan et la NASA, l'utilisent déjà comme base de leurs clients riches Java.