Génération de parole expressive dans le cas des langues à tons

Dang Khoa Mac

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **EEATS / SIGNAL, IMAGE, PAROLE, TELECOMS**

Arrêté ministériel : 7 août 2006

Présentée par

Dang-Khoa MAC

Thèse dirigée par **Eric CASTELLI** et **Thi-Ngoc-Yen PHAM**
codirigée par **Véronique AUBERGE**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
et de l' **Institut de Recherche International MICA**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique & Traitement du Signal**

Génération de parole expressive dans le cas des langues à tons

Thèse soutenue publiquement le **15 Juin 2012**
devant le jury composé de :

M. Laurent BESACIER
Professeur à l'Université Joseph Fourier - Grenoble, Président
M. Christophe d'ALESSANDRO
Directeur de recherche CNRS, Laboratoire LIMSI, Université Paris XI, Rapporteur
M. Philippe MARTIN
Professeur à l'UFR de Linguistique, Université Paris Diderot, Rapporteur
M. René CARRE
Directeur de recherche CNRS émérite, DDL Lyon, Examinateur
M. Eric CASTELLI
Chargé de Recherche (HDR) - CNRS, Directeur de thèse
Mme Thi-Ngoc-Yen PHAM
Professeur à l'Institut de Polytechnique de Hanoi, Co-directrice de thèse
Mme Véronique AUBERGE
Chargé de Recherche CNRS, Laboratoire LIG, Co-encadrante de thèse

Université Joseph Fourier / Université Pierre Mendès France /
Université Stendhal / Université de Savoie / Grenoble INP

à ma famille ...

Remerciements

Tout d’abord, je voudrais présenter tous mes remerciements, ainsi que toute ma gratitude à mes directeurs de thèse, Monsieur *Eric Castelli*, Madame *Phạm Thị Ngọc Yến* et Madame *Véronique Aubergé* pour avoir accepté d'encadrer ce travail. Mes plus chaleureux remerciements à Monsieur *Eric Castelli*, qui m’a guidé depuis le master, pour toutes ses critiques constructives et ses conseils sur mes travaux de recherche. Un grand merci également à Madame *Véronique Aubergé* qui m’a encadré tout au long de ces années de thèse, pour ses conseils, ses aides et ses encouragements précieux. J'adresse mes remerciements à Madame *Phạm Thị Ngọc Yến* pour son intense participation à l'orientation de mes travaux de recherche.

Je tiens à remercier Monsieur *Laurent Besacier* pour avoir accepté d’être le président du jury. Je souhaite remercier aussi Monsieur *Christophe d'Alessandro* et Monsieur *Philippe Martin*, pour avoir accepté d'être les rapporteurs de cette thèse. Leurs remarques pertinentes sur le contenu m'ont permis d'améliorer la qualité de ce document. Un grand merci à Monsieur *René Carré* d’avoir accepté de participer à ce jury.

J’adresse mes remerciements à Monsieur *Nguyễn Trọng Giảng*, ancien directeur l’Institut MICA, à Madame *Phạm Thị Ngọc Yến*, directrice de l’Institut MICA et à Monsieur *Eric Castelli*, co-directeur de l’Institut MICA pour m’avoir accueilli dans le département SpeechCom. Un sincère remerciement également à Madame *Brigitte Plateau*, ancienne directrice du laboratoire LIG, à Monsieur *Hervé Martin*, le directeur du laboratoire LIG et à Monsieur *Laurent Besacier* pour m’avoir accueilli dans son équipe GETALP.

Je voudrais remercier également Monsieur *Albert Rilliard*, qui m’a donné beaucoup de connaissances scientifiques. En fait, il m’a guidé comme un co-encadrant non-officiel. Sans lui, je n’aurais pas pu réaliser cette thèse.

Je tiens à remercier chaleureusement Madame *Geneviève Caelen-Haumont* pour les connaissances scientifiques, les formulations et les expressions transmises pendant les nombreuses discussions que nous avons eues et pour avoir relu et corrigé soigneusement ce mémoire.

Un grand merci aux collaborateurs de mes travaux de recherche (*Trần Đỗ Đạt, Christophe Savariaux, Nicolas Audibert, Takaaki Shochi, Anne Vanpé, …*). Je remercie également tous les membres de l'équipe GETALP du LIG et l'équipe SpeechCom du MICA pour leur accueil et leur sympathie. Je remercie mes amis au MICA et mes amis à Grenoble avec qui j'ai partagé de grands moments au cours de ma thèse.

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

Un grand merci à tous.

Résumé

De plus en plus, l'interaction entre personne et machine se rapproche du naturel afin de ressembler à l'interaction entre humains, incluant l'expressivité (en particulier les émotions et les attitudes). Dans la communication parlée, les attitudes, et plus généralement les affects sociaux, sont véhiculés principalement par la prosodie. Pour les langues tonales, la prosodie est utilisée aussi pour coder l'information sémantique dans les variations de tons. Ce travail de thèse présente une étude des affects sociaux du vietnamien, une langue à tons et une langue peu dotée, afin d'appliquer les résultats obtenus à un système de synthèse de haute qualité capable de produire la parole « expressive » pour le vietnamien.

Le premier travail de cette thèse consiste en la construction du premier corpus audio-visuel des attitudes vietnamiennes, qui contient seize attitudes. Ce corpus est ensuite utilisé pour étudier la perception audio-visuelle et interculturelle des attitudes vietnamiennes. Pour cela, une série de tests perceptifs a été effectuée avec des auditeurs natifs et non-natifs (des auditeurs francophones pour les non-natifs). Les résultats de ces tests montrent que les facteurs influant sur la perception des attitudes sont l'expression de l'attitude elle-même et la modalité de présentation (audio, visuelle et audio-visuelle). Ces résultats nous ont ainsi permis de trouver des affects sociaux communs ou interculturels entre le vietnamien et le français. Puis, un autre test de perception a été réalisé sur des phrases avec tons afin d'explorer l'effet du système tonal du vietnamien sur la perception des attitudes. Les résultats montrent que les juges non-natifs peuvent traiter et séparer les indices tonals locaux et les traits saillants prosodiques de portée globale.

Après une présentation de nos études sur les affects sociaux en vietnamien, nous décrivons notre modélisation de la prosodie des attitudes en vue de la synthèse de la parole expressive en vietnamien. En nous basant sur le modèle de superposition des contours fonctionnels, nous proposons une méthode pour modéliser et générer de la prosodie expressive en vietnamien. Cette méthode est ensuite appliquée pour générer de la parole expressive en vietnamien, puis évaluée par des tests de perception sur les énoncés synthétiques. Les résultats de perception valident bien la performance de notre modèle et confirment que l'approche de superposition de contours fonctionnels peut être utilisée pour modéliser une prosodie complexe comme dans le cas de la parole expressive d'une langue à tons.

Mots clés : *parole expressive, attitude, affects sociaux, tons, contours prosodiques, vietnamien, langue tonale, synthèse de la parole, modélisation de la prosodie.*

Abstract

Today, the human-computer interaction is reaching the naturalness and is increasingly similar to the human-human interaction, including the expressiveness (especially emotions and attitudes). In spoken communication, attitudes or social affects are mainly transferred through prosody. For tonal languages, prosody is also used to encode semantic information via tones. This thesis presents a study of social affects in Vietnamese, a tonal and under-resourced language, in order to apply the results to Vietnamese expressive speech synthesis task.

The first task of this thesis concerns the construction of a first audio-visual corpus of Vietnamese attitudes which contains sixteen attitudes. This corpus is then used to study the audio-visual and intercultural perceptions of the Vietnamese attitudes. A series of perceptual tests was carried out with native and non-native listeners (French for non-native listeners). Experimental results reveal the fact that the influential factors on the perception of attitudes include the modality of presentation (audio, visual and audio-visual) and the attitudinal expression itself. These results also allow us to investigate the common specificities and cross-cultural specificities between Vietnamese and French attitudes. Another perception test was carried out using sentences with tonal variation to study the influence of Vietnamese tones on the perception of attitudes. The results show that non-native listeners can process the local prosodic cues of tones, together with the global cues of attitude patterns.

After presenting our studies on Vietnamese social affects, we describe our work on attitude modelling to apply it to Vietnamese expressive speech synthesis. Based on the concept of prosodic contour superposition, a prosodic model was proposed to encode the attitudinal function of prosody for Vietnamese attitudes. This model was applied to generate the Vietnamese expressive speech and then evaluated in a perceptual experiment with synthetic utterances. The results validate the ability of applying our proposed model in generating the prosody of attitudes for a tonal language such as Vietnamese.

Key words: *expressive speech, attitude, social affects, tones, prosodic contours, Vietnamese, tonal language, speech synthesis, prosody modeling*.

Table des matières

Liste des figures

Liste des tableaux

Introduction

"Speech is the mirror of the soul; as a man speaks, so he is"

Publilius Syrus [Lyman 1856]

Contexte général

La parole est le vecteur le plus sophistiqué pour la communication entre une personne et une autre, c'est la « communication parlée ». Cependant, dans la communication parlée, le signal acoustique de parole n'est pas seulement utilisé pour transporter les informations sémantiques (le contenu linguistique de ce qui est prononcé) ; il transporte aussi tout un ensemble d'informations désignées par certains auteurs, de façon réductrice, comme *extralinguistiques*, qui permettent à l'auditeur de reconnaître son interlocuteur, de reconnaître aussi la langue qui est parlée par cet interlocuteur, mais aussi qui lui permettent aussi d'identifier, bien au-delà de l'identité du locuteur sa personnalité, et ses états d'humeur, d'émotions, ses états mentaux, les processus cognitifs qu'il est en train d'effectuer [Loyau et al. 2006]. Très globalement ces informations véhiculées par la parole au-delà du sens déduit de la lexico-morpho-syntaxe peuvent se ramener à *l'expressivité* de la parole ou bien de la *parole expressive* dans la communication [Arnold 1960].

De plus en plus d'applications technologiques, voire industrielles, et de logiciels multimédia utilisent le son et la parole pour améliorer l'interaction entre l'homme et la machine. Quand en particulier l'interaction personne-machine est directement conduite par la parole, ces applications mettent alors en œuvre un système de *dialogue homme-machine* qui est classiquement constitué d'un pipe-line de modules avec aux extrémités un module de synthèse automatique de la parole et un module de reconnaissance automatique de la parole. Les systèmes de synthèse de la parole ont une très longue histoire [Liénard 1977 ; Calliope 1989]. Ils sont un objet d'étude de longue date dans le monde scientifique et plusieurs logiciels sont disponibles sur le marché. Les nouvelles générations de synthétiseurs (depuis [Campbell 1996 ; Iida et al. 2003b]), basés sur le « clonage » de corpus (concaténation des plus longues unités produites dans des corpus où ils sont sélectionnés en fonction et avec la variabilité prosodique et phonétique propre au contexte) sont capables de capturer le « style » du locuteur, voire son expressivité (propriété qui reste vraie dans la toute dernière avancée par l'extension des unités par HMM). Cependant, ces systèmes, même s'ils sont

capables aujourd'hui de produire de la parole naturelle, reflétant la personnalité et le rôle sociétal (le style) du locuteur selon la manière dont il les a produits dans le corpus (ou les *rushes*) de base de la synthèse, ne sont pas capables d'intégrer les valeurs interactives propres à la conduite du dialogue (valeurs illocutoires, attitudes intentions). Si la synthèse expressive est l'une des thématiques principales du domaine de recherche sur la synthèse de la parole, la prise en compte de ces « affects sociaux », pourtant valeurs fondamentales de la parole expressive est encore très lacunaire.

Dans la communication parlée, les affects (en particulier les émotions et attitudes) sont véhiculées principalement par la prosodie [Scherer 1984b ; Fónagy et al. 1991 ; Aubergé et al. 1997], et tout particulièrement les affects sociaux [Aubergé et al. 1997]. Modéliser et générer cette prosodie s'avèrent alors le point clé dans la synthèse de la parole expressive.

L'institut MICA est l'un des premiers groupes de recherche au Vietnam qui s'investit dans les thématiques du traitement de la parole. Le système de synthèse de la parole à partir du texte en vietnamien de l'institut MICA, développé initialement par Tran [2007], est considéré comme le premier système de synthèse de parole de haute qualité pour le vietnamien. Comme pour les toutes les langues, la nécessité de développer de nouvelles interfaces personne-machine pour des applications multimédia, impose de maîtriser un module de génération de parole synthétique (plus couramment appelé module de synthèse) capable de produire une parole qui transporte aussi les informations extralinguistiques.

Les travaux de thèse que nous avons poursuivis s'inscrivent naturellement dans la thématique portant sur la parole expressive pour la langue à tons qu'est le vietnamien. Les objectifs scientifiques visés peuvent se résumer en deux étapes : dans un premier temps étudier la parole expressive en vietnamien, puis, dans un deuxième temps, appliquer les résultats obtenus dans un système de synthèse de haute qualité capable de produire de la parole « expressive » pour le vietnamien.

Les problématiques principales

La première problématique que nous étudierons concerne le domaine de l'expressivité abordé selon l'angle de l'émotion, l'attitude, etc., domaine qui est très complexe, parce qu'il implique de multiples disciplines telles que la linguistique, la psychologie cognitive, la psychologie sociale, la physiologie, la neurologie, etc. Pour intégrer l'expressivité dans l'interaction homme–machine, il faut bien comprendre les fonctions, les expressions et surtout la signification des différents types d'expressivités (émotion et attitude) dans la communication. Ainsi, nous porterons tout d'abord notre intérêt vers les cadres théoriques concernant l'émotion et l'attitude, enfin de trouver la méthode appropriée pour intégrer ce type d'expressivité dans la synthèse de la parole.

Une autre difficulté dans le travail sur la parole expressive en vietnamien est le manque de données ainsi que la rareté des études antérieures sur cette langue.

Bien que le vietnamien soit parlé par environ 86,9 millions de personnes au Vietnam et environ 4 millions de personnes à l'étranger (en 2010) [1] - ce qui le place au 14ième rang mondial - les études sur le traitement de la parole n'ont commencé à être mises en œuvre que très récemment. Les travaux en traitement de la langue vietnamienne restent encore peu nombreux. Plus particulièrement, dans le domaine de la parole expressive, à notre connaissance, il n'existe que l'étude de Le [Le 1989] qui a été réalisée il y a plus de 20 ans. Un corpus pour étudier la parole expressive en vietnamien est évidemment indisponible. La langue vietnamienne peut donc être considérée pour ce domaine particulier comme une langue peu dotée.

Le deuxième verrou scientifique porte sur la modélisation et la génération de la prosodie des affects qui sont des tâches encore mal cernées, même pour des langues largement étudiées comme l'anglais ou le français. La prosodie expressive est très complexe et très variable, car l'expressivité se décline comme nous l'avons dit sur des fonctions diverses, en particulier les émotions vs. les affects sociaux. Il existe encore peu de méthodes bien établies pour représenter et modéliser les émotions dans la prosodie expressive [Banse et al. 1996 ; Bänziger 2004 ; Audibert et al. 2006b]. Concernant les affects sociaux, depuis [Martins-Baltar 1977 ; Fónagy 1983], plusieurs travaux ont été menés sur diverses langues [Fujisaki et al. 1993 ; Morlec et al. 1997b ; Wichmann 2000 ; Shochi et al. 2005 ; Moraes et al. 2010 ; Gu et al. 2011 ; Nadeu et al. 2011].

Pour les langues tonales comme le vietnamien, la prosodie n'est pas seulement utilisée pour caractériser la modalité de la phrase (interrogative ou affirmative), l'émotion du locuteur (triste, gai ou en colère) ou l'emphase qui est faite sur certains mots ou groupes de mots, mais elle est utilisée aussi pour coder l'information sémantique par les tons. La principale question posée alors peut se résumer comme suit : comment représenter et modéliser la combinaison de l'expression des émotions/attitudes et de l'expression des tons dans la prosodie ? De plus, le vietnamien est une langue tonale qui utilise phonologiquement les variations de fréquence fondamentale et de qualité de voix (avec des changements dus à la source glottique). Cela rend le système tonal en langue vietnamienne plus complexe que ceux d'autres langues tonales telles que le thaï ou le chinois. Cependant, c'est cet usage de la qualité de voix qui rend le vietnamien intéressant. En effet, si les usages les plus communs de la qualité de voix restent en général non linguistiques, en particulier en ce qui concerne l'expression des émotions et des attitudes, dans le vietnamien il s'ajoute aussi cette fonction phonologique.

Les tâches principales de la thèse

Dans le cadre de cette thèse, nous n'avons cependant pas la prétention de résoudre tous les problèmes présents dans la génération de la prosodie expressive en général et en vietnamien plus particulièrement. Plus modestement, notre

[1] Bureau de statistique générale du Vietnam http://www.gso.gov.vn

travail se concentrera sur 1) une étude préliminaire sur la parole attitudinale et ses caractéristiques dans le cadre de la langue vietnamienne, mise en perspective perceptive inter-culturelle puis 2) dans la proposition d'une approche appropriée pour intégrer ce type d'expressivité dans la synthèse de la parole en vietnamien. Ces études serviront alors de base pour construire un système de synthèse de parole expressive de haute qualité en vietnamien, et produiront, nous l'espérons, des premiers résultats qui guideront nos recherches futures sur la communication expressive entre les êtres humains et les machines, que ces dernières soient des ordinateurs, des systèmes plus complexes ou même des robots.

Nos travaux poursuivis tout au long de cette thèse vont donc consister en trois tâches principales :

La première tâche est la construction d'un corpus de parole expressive en vietnamien ; ce corpus n'est pas conçu uniquement pour la synthèse de la parole, il peut être utilisé pour d'autres études sur la parole expressive en vietnamien.

Dans un deuxième temps, les actions menées s'attachent à étudier les caractéristiques de la parole expressive en vietnamien, particulièrement au niveau de la perception audio-visuelle et à celui de la perception interculturelle. Si ces travaux nous permettent d'étudier les spécificités culturelles et la perception interculturelle de ce type de parole expressive en vietnamien dans la communication face à face, ils sont aussi nécessaires pour produire, grâce à un système de synthèse de la parole expressive, des attitudes qui sont appropriées aux différents contextes et qui peuvent être bien perçues par l'interlocuteur.

La dernière tâche consiste à proposer une méthode pour modéliser la prosodie de la parole expressive en vietnamien, puis à appliquer cette méthode en la validant dans la synthèse de la parole.

Structure de la thèse

Notre manuscrit de thèse se divise en sept chapitres principaux qui sont regroupés en deux parties : l'analyse de l'état de l'art (chapitres 1, 2, 3 et 4) et nos contributions (chapitres 5, 6 et 7).

La première partie du manuscrit présente l'état de l'art sur les problématiques et les théories proposées dans la littérature, ce qui nous permet de préciser nos objectifs de recherche et les principes appropriés pour notre travail.

Le chapitre 1 est dédié à une introduction générale sur la parole expressive, en particulier sur les concepts et la différence entre les deux termes qui sont utilisés souvent dans les champs de l'expression affective : *émotion* et *attitude*. Ce chapitre propose une discussion pour expliquer pourquoi l'objet de ce travail se concentrera sur les attitudes seulement.

Dans le chapitre 2, nous survolons les caractéristiques de la prosodie et de ses paramètres principaux. Puis, nous exposons les fonctions de la prosodie dans la communication parlée, surtout la fonction attitudinale de la prosodie sur laquelle nos recherches porteront.

Le chapitre 3 est consacré à la génération de la parole expressive. Nous présentons tout d'abord quelques approches principales permettant d'ajouter de l'expressivité dans les techniques actuelles de synthèse de la parole. Puis, nous insistons sur l'un des points-clé de la synthèse expressive : la modélisation et la génération de la prosodie. Ce chapitre se termine par des discussions sur notre travail : quelle technique de synthèse et quel modèle de prosodie sont appropriés à la génération de la parole expressive en vietnamien ?

Après la présentation succincte de l'état de l'art général sur la parole expressive et la synthèse de la parole, nous proposons au sein du chapitre 4 d'aborder le vietnamien. Nous commençons d'abord avec une présentation brève des caractéristiques phonétiques et phonologiques de la langue vietnamienne. Ensuite, nous parlons de la prosodie du vietnamien, en particulier la prosodie dans la parole expressive. Par la suite, nous résumons certains travaux de recherche pour modéliser la prosodie et pour la synthèse de la parole en vietnamien.

La deuxième partie du manuscrit se compose de trois chapitres présentant nos contributions à l'étude des attitudes vietnamiennes et à la modélisation de la prosodie des attitudes pour les appliquer à la synthèse de la parole expressive.

Le chapitre 5 présente nos travaux effectués sur la construction d'un corpus audio-visuel des attitudes vietnamiennes, qui peut être considéré comme le premier corpus pour la parole expressive en vietnamien. En nous basant sur des études précédentes sur d'autres langues, 16 attitudes ont été choisies pour le corpus d'attitudes en vietnamien. Ce corpus est alors évalué par un test perceptif.

Dans le chapitre 6, nous présentons les études sur la perception des attitudes vietnamiennes. Une série de tests perceptifs est effectuée ; premièrement pour comprendre le rôle et l'influence des deux modalités communicationnelles (l'audition et la vision) dans la perception des attitudes ; deuxièmement pour étudier les spécificités culturelles et la perception interculturelle des attitudes vietnamiennes.

Le chapitre 7 est alors consacré à la modélisation de la prosodie des attitudes comme une première étape vers la synthèse de la parole expressive en vietnamien. En nous basant sur le modèle de superposition des contours fonctionnels de la prosodie [Aubergé 2002a], une méthode est proposée pour modéliser la prosodie des attitudes en langue vietnamienne. Cette méthode est ensuite appliquée pour générer la parole expressive en vietnamien et, enfin, est validée par une autre série de tests de perception.

Le mémoire se conclut par un résumé des travaux que nous avons effectués dans le cadre de cette thèse, où nous essayons d'en faire ressortir les contributions principales, puis par les perspectives qui feront l'objet de nos recherches futures à la suite de ces travaux. Les annexes incluses à la fin du manuscrit présentent respectivement les phrases de notre corpus d'attitudes vietnamiennes, et en détail les résultats et les données que nous n'avons pas inclus dans les sept chapitres, mais qui pourront apporter des compléments d'informations au lecteur.

Partie 1 : Analyse de l’état de l’art

Chapitre 1 : La parole expressive : émotion et attitude

1.1. Introduction

Nous l'avons rappelé, dans la communication homme-homme, le signal de parole n'est pas uniquement le vecteur d'informations strictement langagières mais aussi de tout un ensemble d'informations souvent dites « extra-linguistiques », qui caractérisent la « parole expressive ».

L'importance de l'expression affective dans la communication parlée et son impact sur l'auditeur a été reconnue à travers l'histoire comme le rappelle [Scherer 2003]. D'une manière plus scientifique, l'expression des affects grâce au visage, à la parole et à la voix a été étudiée depuis le 19ème siècle suite à l'émergence de la biologie évolutionniste moderne, grâce aux contributions de Spencer, Bell, et en particulier de Darwin [Fussell 2002]. Les recherches sur l'effet de l'émotion sur la voix au début du 20ème siècle, ont conduit des psychiatres à essayer de diagnostiquer les troubles émotionnels avec les nouvelles méthodes d'analyse acoustique, par exemple, Isserlin en 1925; Ecriture en 1921 et Skinner en 1935 [Scherer 2000 ; Audibert 2008]. Toutefois, les travaux de recherche sur l'analyse en tant que telle des expressions vocales n'ont commencé que dans les années 1960 aussi bien en phonétique (phonostylistique), qu'en psychologie [Scherer 2003 ; Audibert 2008]. Dans les décades plus récentes, des scientifiques ont commencé à développer des modèles de communication expressive pour les applications de technologie de la parole en synthèse comme en reconnaissance de la parole [Schröder 2001 ; Audibert 2008].

Dans les champs d'étude de la communication affective, les deux termes « émotion » et « attitude » peuvent parfois être confondus, ces deux termes recouvrant des notions variées selon les auteurs car ces deux notions sont, en fait, assez complexes à définir et modéliser. Pour bien distinguer ces deux termes, dans les sections suivantes, nous survolerons d'abord quelques hypothèses sur l'émotion et l'attitude et leurs distinctions. Après une présentation du rôle de l'émotion et de l'attitude dans le système d'interaction personne – machine, ce chapitre se termine par une discussion destinée à expliquer pourquoi l'objet de ce travail se concentrera sur les attitudes.

1.2. Les émotions

Définir les émotions est un sujet encore controversé, bien que l'émotion soit étudiée depuis longtemps. Dans cette section, nous n'avons pas l'intention de

donner une définition précise de l'émotion. Nous voudrions seulement présenter quelques principales approches théoriques. Chaque théorie nous donne une hypothèse sur la nature de l'émotion, une façon de construire des émotions, et des suggestions sur la façon de mener des travaux de recherche sur l'émotion.

En survolant plus d'un siècle de recherches, il ressort quatre principales approches théoriques sur l'émotion.

1.2.1. L'approche évolutionniste (l'approche des émotions de base)

La théorie de l'évolution a été introduite Darwin, cette approche est donc aussi appelée « la perspective darwinienne ». Dans l'ouvrage « *The Expression of Emotion in Man and Animals*» (1872), Darwin explique que les expressions faciales et les mouvements corporels accompagnent les émotions des humains mais aussi d'autres animaux. Il a aussi présenté une théorie simple de l'évolution des expressions et des mouvements. L'idée principale de cette perspective est que les émotions constituent des phénomènes évolués qui sont apparus en adaptation à l'environnement car ils remplissent des fonctions de survie de l'espèce. Donc, nous voyons les mêmes émotions, plus ou moins, dans toutes les parties du monde. En outre, parce que les humains partagent une histoire évolutive avec d'autres mammifères, nous devrions observer des similarités dans les émotions des espèces apparentées proches.

A la suite de Darwin, nombre de chercheurs récents comme Tomkins, Izard, Ekman et Plutchik, ont adhéré à la théorie de l'émotion selon la perspective évolutionniste, en la complétant. Ils sont appelés « les néo-darwiniens » [Cornelius 2000]. Les néo-darwiniens proposent l'existence d'un ensemble d'émotions « fondamentales », « de base », ou « primaires » qui se décriraient en catégories discrètes et sont très ancrées sur les « postures faciales » (souvent décrites statiquement) de leur expressions. Selon eux, ces émotions sont considérées comme fondamentales parce qu'elles peuvent représenter notre schéma de réponse aux événements dans le monde pendant notre histoire d'évolution. Elles sont aussi considérées comme fondamentales car toutes les autres émotions pourraient dériver à partir d'elles [Cornelius 2000].

Le nombre de ces expressions/émotions « fondamentales » varie et dépend de chaque théoricien. Ekman parle de six émotions fondamentales seulement [Ekman et al. 1979], Izard en identifie dix [Izard 2000], il y en a huit selon Plutchik [1984] (qui propose également une représentation dimensionnelle des émotions). Ces dernières années, six émotions, qui sont appelées « The Big Six » [Cornelius 2000], sont utilisées couramment. Ce sont : *la joie, la colère, la tristesse, la peur, le dégoût et la surprise*.

Les expressions de ces émotions, essentiellement leurs expressions faciales, et jusqu'à récemment avant tout présentées sous forme d'images statiques sont la base de ces travaux en intra et en inter-culturel.

1.2.2. L'approche physiologique (la théorie James-Lange)

Dans l'article "*What is an Emotion ?"* (1884), William James développe une théorie qui postule que l'émotion n'est pas assimilée à la *cause* (le stimulus) de quelque chose, mais elle constituerait plutôt un *effet* (les changements physiologiques). Selon James: "*bodily changes follow directly the PERCEPTION of the exciting fact, and [...] our feeling of the same changes as they occur IS the emotion*". James a insisté sur le fait qu'il serait impossible d'avoir des émotions sans changements corporels préalables et ces changements corporels viennent toujours en premier. Il a écrit : "*we feel sorry because we cry, angry because we strike, afraid because we tremble*" [James 1884].

En 1885, Carl Lange publie, de façon indépendante, une théorie avec une idée très similaire à la proposition de James. Pour cette raison, cette approche est appelée « la théorie James-Lange » [Cannon 1931] .

James et Lange proposent un processus émotionnel séquentiel décrit par la Figure 1. Selon eux, notre corps répond automatiquement à des événements de l'environnement par des changements corporels. Puis, notre perception de ces changements constitue ce que nous appelons émotion.

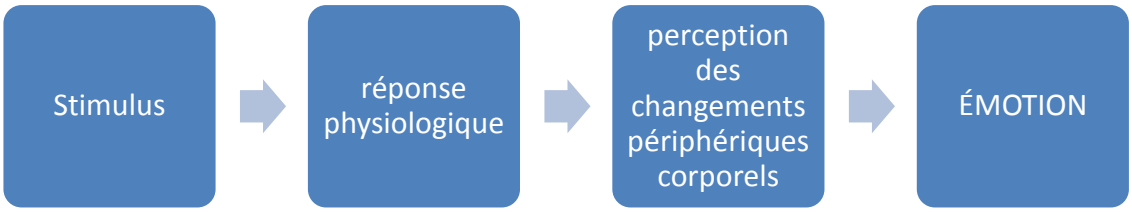


Figure 1: La séquence du processus émotionnel de la théorie James-Lange

La théorie de James-Lange concerne la nature de l'expérience émotionnelle alors que celle de Darwin concerne l'expression émotionnelle. Mais on retrouve aussi l'idée Darwinienne de l'adaptation à l'environnement extérieur par des réponses émotionnelles automatiques dans la théorie James-Lange. Comme Darwin, James-Lange considère les émotions comme une réaction aux évènements, plus ou moins automatique, qui a valeur de survie. Ces théories sont donc peu centrées sur les expressions des émotions.

1.2.3. L'approche cognitiviste de l'évaluation

L'hypothèse centrale de la perspective cognitive est que les émotions sont vues comme dépendantes de ce qui est appelé « *appraisal* » [Arnold 1960]. C'est un processus cognitif par lequel les évènements environnementaux sont jugés bons ou mauvais pour nous. Selon Arnold, une évaluation positive d'un stimulus déclenche une émotion positive, et sera exprimée par un comportement de rapprochement. Au contraire, une évaluation négative produit une émotion négative et sera exprimée par un comportement d'éloignement.

Comme James dit qu'on ne peut pas concevoir une émotion sans corps, Arnold dit que « *l'appraisal* » (évaluation) est ce dont on a besoin pour avoir une émotion. Chaque émotion est associée à un « *appraisal* » spécifique et différent.

Suivant la théorie de Arnold, Christophe [Christophe 1998] décrit les différentes étapes du processus émotionnel par la Figure 2.

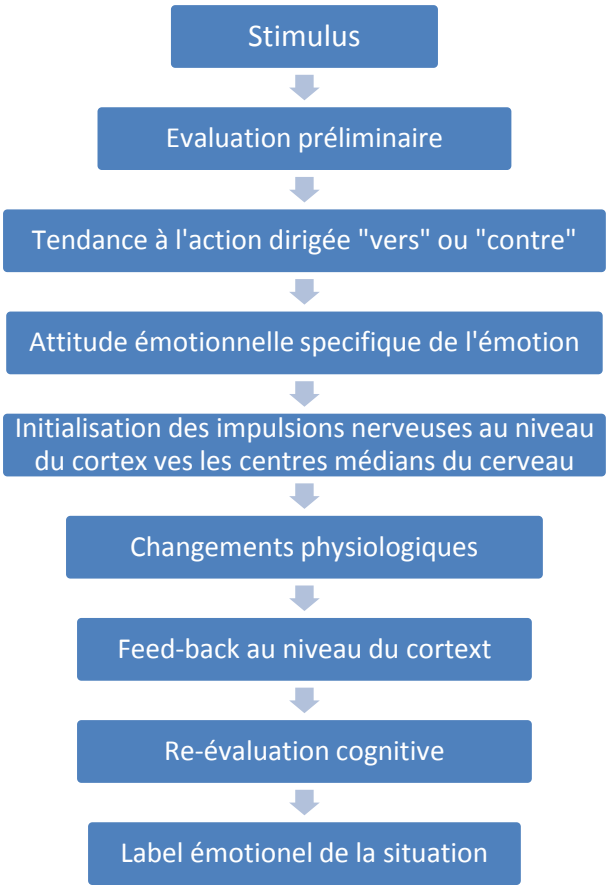


Figure 2 : Représentation de la théorie d'Arnold [Christophe 1998]

1.2.4. L'approche du constructivisme social

Dans cette approche, l'émotion est étudiée d'un point de vue social et culturel. James Averill est l'un des premiers théoriciens à penser les émotions dans le cadre du « constructivisme social ». Selon lui, les émotions seraient le produit de « constructions sociales », et dépendraient essentiellement du contexte social dans lequel elles apparaissent.

> *« emotions are not just remnants of our phylogenetic past, nor can they be explained in strictly physiological terms. Rather, they are social*

constructions, and they can be fully understood only on a social level of analysis. »[2] *[Averill 1980]*

Dans cette optique, l'émotion est le résultat de normes sociales existantes dans un contexte social précis. Donc pour le constructivisme social, la culture joue un rôle très important dans l'organisation des émotions. Averill ne sépare pas explicitement les affects sociaux des affects qui seraient innés, c'est-à-dire communs à toutes les cultures.

Conclusion

Dans cette première partie nous avons décrit brièvement les différentes approches existantes sur les émotions. A travers ces multiples points de vue théoriques, il apparaît clairement que pour étudier l'émotion, il n'existe pas de théorie standardisée qui pourrait fournir un cadre général à ces différentes approches. Finalement, pour étudier l'émotion complètement, nous devrions examiner l'émotion sous tous ses aspects : évolution, physiologie, cognition et constructivisme social.

1.3. Les attitudes en psychologie sociale

Le mot « attitude » est très ambigu et diffère dans ce qu'il désigne selon qu'il est utilisé dans le champ de la psychologie sociale ou dans le champ de la phonostylistique.

Il est largement utilisé dans la vie quotidienne pour décrire une personne et son comportement, par exemple, « elle a une bonne attitude dans son travail ». En effet, les attitudes sont très importantes et influencent ce que nous pensons et ce que nous faisons dans la vie. On peut dire que l'attitude est l'une des composantes essentielles du comportement [Maio et al. 2010]. C'est pourquoi les psychologues sociaux mettent beaucoup d'attention à comprendre comment nous formons les attitudes, comment nos attitudes influencent notre vie quotidienne, et comment nos attitudes changent. Dans cette section, nous examinerons le concept « d'attitude » au sein de la psychologie sociale.

1.3.1. Définition

L'étude des attitudes a une longue histoire au sein de la psychologie sociale. Comme l'émotion, le concept d'attitude constitue une notion sujette à des débats permanents. Gordon Allport, l'un des fondateurs de la recherche sur les attitudes, a noté que « *le concept d'attitude est probablement le concept le plus distinctif et le plus indispensable dans la psychologie sociale* » [Allport 1935]. Allport est aussi l'un des premiers chercheurs qui a défini le terme « d'attitude » :

[2] « Les émotions ne sont pas simplement les restes de notre passé phylogénétique, elles ne peuvent pas non plus être expliquées en des termes strictement physiologiques. Elles sont plutôt des constructions sociales, et peuvent être seulement comprises dans leur ensemble par une analyse située au niveau social »

> *"An attitude is a mental and neural state of readiness which exerts a directing influence upon the individual's response to all objects and situations with which it is related"*[3] *[Allport 1935]*

L'idée centrale de cette définition des attitudes est la préparation d'une réponse. Une attitude n'est pas un comportement, ni quelque chose qu'une personne fait; c'est plutôt une préparation pour un comportement, une prédisposition pour répondre particulièrement à l'objet. Cet objet peut être une chose, personne, lieux, idée, action ou situation dans la vie. Allport a aussi noté qu'il serait impossible d'expliquer un comportement quelconque sans recourir à la notion d'attitude ; dans le même temps, les attitudes ne font cependant que guider le comportement et elles ont leur origine dans l'expérience.

Dans les dernières décades, le concept d'attitude a été défini de nombreuses manières. Richard Petty et John Cacioppo [Cacioppo et al. 1981] définissent une attitude comme « un sentiment général et durable positif ou négatif sur certaines personnes, un objet, ou un problème ». Pour Mark Zanna et John Rempel, définir une attitude consiste à proposer « la catégorisation d'un objet avec une dimension évaluative » [4] [Zanna et al. 1988]. Du point de vue de la psychologie, Alice Eagly et Shelly Chaiken [Eagly et al. 1993] définissent une attitude comme « une tendance psychologique qui s'exprime par l'évaluation d'une entité particulière avec un certain degré de faveur ou défaveur»[5]. Russell Fazio [Fazio 1995] définit une attitude comme « une association en mémoire entre un objet et une évaluation de l'objet » [6].

Malgré les différences entre les définitions citées ci-dessus, elles insistent toutes sur la notion qu'une attitude implique l'expression d'un jugement évaluatif sur un objet. En effet, la plupart des théoriciens soutiennent que l'évaluation est l'aspect principal du concept d'attitude. En d'autres termes, une attitude implique une décision d'aimer, de détester, en favorisant ou défavorisant une situation particulière, un objet ou personne.

1.3.2. Les composants de l'attitude

Nous avons vu que l'attitude est considérée comme une évaluation globale (par exemple, aimer-détester) d'un objet. En se basant sur cette approche, les théoriciens proposent des modèles conceptuels des attitudes. Parmi eux, le modèle multi-composant est le plus général et le plus influent [Maio et al. 2010]. Selon ce modèle, les attitudes sont basées sur trois composantes : la cognition, l'affect et le comportement (Figure 3).

[3] « Une attitude est un état mental et neuronal de disposition à passer à l'action, exerçant une influence directive sur les réponses de l'individu à tous les objets ou situations auxquels il est confronté. »
[4] " the categorization of a stimulus object along an evaluative dimension"
[5] " a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor"
[6] "an association in memory between a given object and a given summary evaluation of the object."

- La *composante cognitive* de l'attitude concerne les croyances, les pensées et les attributs d'un objet. Normalement, l'attitude d'une personne ou son jugement évaluatif sur un objet est principalement basé sur les attributs de cet objet. Par exemple, quand on achète un ordinateur, on porte une attention particulière à ses caractéristiques (CPU, RAM, etc.) et on décide que c'est un bon ordinateur ou pas. Le jugement évaluatif sur un objet dépend aussi de la croyance et de la pensée individuelle de chaque personne et de ses connaissances. Un ordinateur est bon pour nous mais peut-être n'est pas bon pour un autre. La composante cognitive dépend aussi de la croyance. Par exemple, les enfants qui croient au père Noël ont des attitudes particulières et différentes des adultes.
- La *composante affective* des attitudes porte sur les sentiments ou sur les émotions liées à un objet. Nous avons des sentiments différents avec des objets différents. Par exemple, normalement, la musique nous rend plus heureux et le bruit nous rend aigris.
- La *composante comportementale* sur les attitudes concerne les comportements ou les expériences vécues avec un objet dans le passé. Autrement dit, le jugement évaluatif d'un objet peut se baser sur nos expériences précédentes avec cet objet ou être appris avec les attitudes des autres que l'on a observées dans la même situation. Par exemple, si on doit soigner un bébé pour la première fois, notre attitude pourra être la simulation de l'attitude d'autres personnes vues avec un bébé dans le passé.

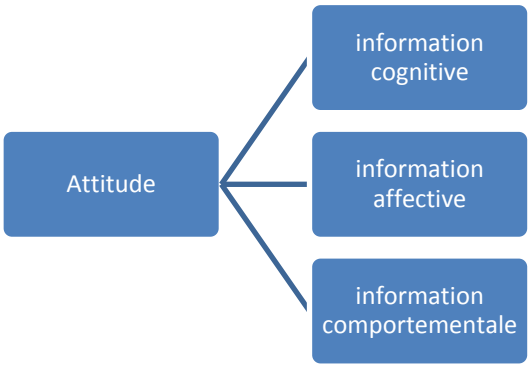


Figure 3 : Le modèle multi-composant des attitudes [Maio et al. 2010]

Trois composants de l'attitude servent à déterminer l'intention personnelle d'un individu envers une action. Ce modèle multi-composant est donc appelé model CAB (*cognition*, *affect* and *behavior*) [Maio et al. 2010].

1.4. Emotion et Attitude en phonostylistique

Comme nous l'avons mentionné ci-dessus, les deux termes « émotion » et « attitude » sont souvent confondues ou parfois utilisés comme synonymes. Après avoir présenté en bref les théories sur les émotions et les attitudes, nous

voudrions présenter quelques différences dans les caractéristiques des émotions et des attitudes, afin de distinguer ces deux termes.

1.4.1. La nature de contrôle

La parole est le vecteur privilégié de la communication humaine qui, au-delà des informations strictement linguistiques, construit le sens de la communication en la situant par son contexte, ses acteurs, par les connaissances exprimées par le locuteur sur ses états émotionnels, attitudinaux, intentionnels, mentaux et cognitifs. Les affects sont exprimés dans la parole et leurs expressions sont contrôlées à différents niveaux cognitifs [Aubergé 2002a]. A la suite de [Delattre 1966] et plus précisément de Fónagy [1983], Wichman [2000], Aubergé [2002a], de Moraes [2010] reprennent et étendent, avec quelques variantes théoriques, le concept d'attitude de Fónagy comme une valeur des affects sociaux impliquant le sujet dans son acte illocutoire, véhiculant des informations sur la situation d'énonciation, ou sur la relation sociale des sujets interagissants.

Aubergé [2002a], après Ohala [1996], a proposé que les expressions attitudinales et émotionnelles peuvent être distinguées par la nature de leur contrôle (volontaire vs. involontaire) et en conséquence par leur ancrage temporel, qui participe ou non à la construction des unités linguistiques. Selon elle :

Les émotions sont déclenchées par un contrôle involontaire et sont exprimées dans et par la voix ou la face, et sont organisées dans le temps de l'événement déclencheur de l'émotion agissant éventuellement comme une contrainte du temps énonciatif. En termes de construction ontogénétique, les émotions sont innées, et en ce sens semblables entre les langues et les cultures. Seul le contexte de déclenchement et le contrôle inhibitoire est acquis. Ce contrôle correspond à l'effet *push* proposé par Scherer [1984a].

Les attitudes, ou plus largement les affects sociaux, sont intentionnels, déclenchés par un contrôle volontaire et organisées dans le temps cognitif du langage et de la phonologie, voire même organise le temps énonciatif de l'interaction [Aubergé 2002a]. D'après Aubergé, les attitudes résulteraient de la prise de contrôle volontaire du contrôle involontaire des émotions – la simulation d'une émotion – grâce par exemple à la boucle de simulation décrite par Damasio [Damasio 1994]. Ainsi par exemple la surprise [Aubergé 2002a], peut être soit une émotion, auquel cas son expression est organisée dans la temporalité cognitive de l'événement déclencheur, soit une attitude, auquel cas l'attitude est porteuse temporellement de l'énonciation sur laquelle porte la surprise.

Les attitudes sont des codes, construits et acquis dans la dynamique du langage et de ses réalités socioculturelles. Ainsi, elles sont dépendantes de la langue et de la culture, et peuvent leur être spécifiques, dans leurs concepts comme dans leurs réalisations prosodiques. Lorsque nous grandissons, nous apprenons certaines attitudes pendant l'apprentissage de la langue [Shochi et al. 2010]. Puisque les attitudes sont liées au langage et à la culture, elles diffèrent entre les langues et

les cultures. Ainsi, une attitude présente dans une langue peut ne pas exister dans une autre langue [Shochi et al. 2006]. Même si le concept est partagé entre deux langues, l'attitude peut être exprimée différemment et ainsi ne pas être comprise en langue seconde, voire même constituer un « faux-ami », c'est-à-dire un cas de malentendu complexe à lever. Les attitudes relèvent d'un contrôle cognitif volontaire pour exprimer le point de vue (doute, autorité, confiance, politesse...) du locuteur sur ses énoncés. D'après Aubergé [1997] si le locuteur n'exprime aucune attitude (déclaration ou question « simple »), il s'agit alors bel et bien d'une attitude du sujet qui consiste à ne pas avoir, ou ne pas pouvoir ou ne pas vouloir afficher d'attitude particulière sur son discours [Danes 1994 ; Aubergé et al. 1997 ; Morlec et al. 2001 ; Aubergé 2002a ; Shochi et al. 2005]. Ce contrôle correspond à l'effet *pull* proposé par Scherer [1986].

Pour résumer, dans ce travail, les attitudes sont considérées comme des expressions conventionnellement encodées dans une culture et un langage ; elles sont construites par une société, doivent être acquises par les enfants, et de même devront être apprises par les apprenants de langue étrangère, quand ces affects sociaux ne sont pas partagés par la langue maternelle et la langue cible. Inversement, les expressions émotionnelles seraient reconnues universellement.

1.4.2. Les étiquettes des émotions et des attitudes

A cause de la confusion entre les termes « émotion » et « attitude », les étiquettes des émotions et attitudes qui sont utilisées varient selon les auteurs. En fait, les linguistes ont utilisé beaucoup d'étiquettes pour décrire les affects intentionnels ou les attitudes. Schubiger [1958], O'Connor et Arnold [1978] ont utilisé près de 300 étiquettes pour nommer les attitudes au sens psychologique, par exemple (en anglais) : '*abrupt, accusing, affable, affected, affectionate, aggressive, agreeable, airy, amused, angry, animated, annoyed, antagonistic, apologetic, appealing, appreciative, apprehensive, approving, argumentative, arrogant, authoritative, awed...'* [7]

Basé sur les distinctions entre les émotions et les attitudes décrites ci-dessus, nous considérons que la *joie*, la *tristesse*, l'*angoisse*, la *colère*, le *dégoût, etc.* sont des étiquettes réservées plus tôt à la catégorie des affects dont l'expression vocale est déclenchée automatiquement – des émotions. En revanche, l'*irritation*, l'*ironie*, la *sincérité*, le *doute*, la *politesse etc.* sont des étiquettes évaluées dans le cadre des affects sociaux – des attitudes.

En ce qui concerne la *surprise*, dans notre travail, nous adressons bien la surprise dans son affect social que nous désignerons par l'étiquette d'expression « exclamation de surprise ».

[7] *abrupt, accusant, affable, affecté, affectueux, agressif, agréable, amusé, colère, hostile, agacé, antagoniste, apologétique, émouvant, appréciatif, appréhensif, approbatif, ergoteur, arrogant, autorisé, impressionné…*'

1.5. Emotions et attitudes dans l'interaction personne-machine et objet de la thèse

Aujourd'hui, de plus en plus, les humains communiquent avec des machines, par exemples avec les systèmes de navigation, les systèmes de diagnostic automobile, des jeux vidéos, pour l'éducation à distance, les robots d'assistance ou de service.

Les récentes technologies permettent aux utilisateurs de communiquer avec des ordinateurs de plusieurs manières, des plus conventionnelles avec le clavier et la souris, aux nouvelles modalités d'interaction personne-machine par la voix, le geste, etc. De plus en plus, l'interaction entre personne et machine se rapproche du naturel afin de ressembler à l'interaction entre humains, incluant en particulier l'expressivité.

Un système d'interaction homme-machine par la parole expressive demande les deux fonctionnalités principales suivantes :

- comprendre les émotions humaines, qui sont transmises par la parole, le visage et les gestes ; c'est le but du système de reconnaissance des émotions. Si l'on souhaite inscrire le système dans une réelle interaction, les affects sociaux, les intentions illocutoires, les attitudes du sujet doivent pouvoir être identifiées ;
- donner des informations avec une « expressivité » réaliste, à la fois le « style » qui définit le rôle social du sujet (ce que la synthèse par corpus est capable aujourd'hui de reproduire), mais aussi, toujours pour s'inscrire dans un dialogue personne-machine, les informations d'attitude, d'intention, de relation sociale (hiérarchie), de contexte social (intimité, mère-enfant etc.), qui définissent aussi l'acte illocutoire.

Ce travail s'appliquera à étudier la génération de parole expressive pour appliquer les résultats obtenus à la conception de systèmes de synthèse de la parole. Précisément, l'objectif de ce travail est d'intégrer l'expressivité dans la parole synthétique. Un système de synthèse de la parole expressive utilise des informations linguistiques (à partir du texte) et des informations extralinguistiques (à partir du contexte de la communication) pour produire des extraits de parole synthétique avec un type d'expressivité approprié. Dans ce processus, le choix du type d'expressivité et la manière d'intégrer l'expressivité à la parole synthétique sont dépendants seulement de la méthode et de la fonction du système de synthèse. Autrement dit, la façon de produire des expressivités dans la parole synthétique est contrôlée volontairement par le système de synthèse de la parole. Comme nous l'avons déjà montré, la différence principale entre émotion et attitude tient dans la nature de leur contrôle (volontaire ou involontaire). Par conséquent, dans le cas de la synthèse de la parole, le type d'expressivité dans la parole synthétique se focalise plutôt sur l'attitude. Autrement dit, le système de synthèse de la parole expressive doit offrir la fonctionnalité de produire de la parole artificielle en générant simultanément des attitudes.

C’est pourquoi, afin de viser une application dans un système de synthèse de la parole expressive pour le vietnamien, et dans un premier temps, le type d’expressivité retenu pour notre travail est l’attitude. Précisément, nous allons étudier les caractéristiques des attitudes vietnamiennes, afin de modaliser et d’intégrer l’attitude dans le système de synthèse de la parole du vietnamien.

Chapitre 2 : Prosodie dans la parole expressive

Comme nous l'avons mentionné au chapitre précédent, les systèmes de synthèse de la parole vont de plus en plus vers le naturel de l'interaction. Un des aspects clés de ce naturel, en particulier pour les attitudes, est la prosodie [Fónagy 1983 ; Aubergé 1992 ; Wichmann 2000 ; Moraes et al. 2010]. Dans cette section, nous commençons par la présentation du concept de prosodie et de ses paramètres principaux. Puis, nous parlerons des fonctions de la prosodie dans la communication parlée. Nous terminerons ce chapitre par la discussion sur l'étude de la prosodie dans notre travail.

2.1. La prosodie

2.1.1. Définition

Depuis longtemps, la prosodie a été beaucoup étudiée dans les domaines de la phonétique, de la phonologie et de la stylistique. Dans le livre « Soixante et dix ans de recherches en prosodie » [1975], Di Cristo référençait déjà plus de 4000 entrées bibliographiques relatives à la prosodie. La prosodie est souvent considérée comme secondaire, parallèle ou additive à la chaîne segmentale qui délivre la double articulation du langage. Si bien que Rossi [1999] a proposé la prosodie comme 3^{e} articulation du langage. Aubergé [2002a] va jusqu'à proposer la prosodie comme « architecte » du langage.

Il est complexe de définir le concept de prosodie dans l'absolu et non par « soustraction » de ce qui n'est pas la chaîne phonémique. Le mot « prosodie » vient du grec ancien (*prosôdia*), utilisé comme une « chanson chantée avec un instrument de musique ». Puis, ce mot « prosodie » est utilisé pour désigner les règles métriques qui régissent la lecture à voix haute de la poésie. Dans le champ de la phonétique récemment, le mot « prosodie » et son adjectif « prosodique » sont le plus souvent utilisés pour se référer à des propriétés de la parole qui ne peuvent être dérivées de la séquence de phonèmes des énoncés. Dans le champ du traitement de la parole, le terme « prosodie » se réfère à certaines propriétés du signal de parole telles que les phénomènes liés à la variation dans le temps des paramètres de hauteur (liée à la fréquence fondamentale, fréquence de vibration des cordes vocales), d'intensité (liée à l'amplitude et à l'énergie) et de durée des sons [Di Cristo 1975].

Dans ce travail, nous voudrions retenir la définition de la prosodie, proposée par Di Cristo :

" La prosodie (ou la prosodologie) est une branche de la linguistique consacrée à la description (aspect phonétique) et à la représentation formelle (aspect phonologique) des éléments de l'expression orale tels les accents, les tons, l'intonation et la quantité, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (Fo), de la durée et de l'intensité (paramètres prosodiques subjectifs). Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations paralinguistiques et des informations linguistiques déterminantes pour la compréhension des énoncés et leur interprétation pragmatique dans le flux du discours." [Di Cristo 2000]

2.1.2. Les paramètres

Selon Selon la définition ci-dessus, au niveau phonétique, les paramètres de la prosodie sont principalement la fréquence fondamentale, la durée et l'intensité. Ces trois paramètres peuvent être décrits comme suit :

- **la fréquence fondamentale (notée F0) : «** *fréquence de base qui donne la périodicité d'un son périodique complexe dont les harmoniques sont les multiples entiers ; la fréquence fondamentale d'un son périodique complexe dont on connaît les harmoniques est égale au plus grand commun diviseur de ces harmoniques (PGCD), la fréquence est mesurée en Hertz (Hz), ou nombre de périodes par seconde* » [Rossi 1999, p.206] ;
- **la durée** *est vue comme l'intervalle de temps nécessaire pour émettre un signal. En d'autres termes, c'est l'organisation temporelle du message qui comprend le débit de parole (nombre de syllabes réalisées par seconde), le tempo (accélération ou ralentissement du débit dans le groupe prosodique), les pauses et la durée phonémique* [Lacheret-Dujour et al. 1999, p.12] ;
- **l'intensité** « *mesure ce qu'on appelle le volume dans le langage courant ; elle est une mesure logarithmique de l'énergie du signal. L'intensité se mesure en décibels (dB). Le seuil différentiel d'intensité pour la parole est d'environ 3 dB : augmenter une voyelle de 3 dB équivaut à augmenter son amplitude de 40% et à doubler sa puissance* » [Rossi 1999, p.206].

La fréquence fondamentale, la durée et l'intensité sont tous trois les corollaires acoustiques les plus classiques de la parole [Lacheret-Dujour et al. 1999]. Parmi les paramètres prosodiques ci-dessus, la fréquence fondamentale et la durée sont les plus étudiés, parce qu'elles sont les plus significatives et les plus faciles à percevoir [Benesty 2008].

La qualité de voix

Dans la parole expressive, et peut-être d'une manière plus générale en prosodie, les trois paramètres ci-dessus ne suffisent pas à décrire les expressions des

émotions et attitudes. Selon Campbell [2003b] la **qualité de voix** est proposée comme une 4ème dimension de la prosodie ; il a été également montré qu'il s'agit d'un paramètre fondamental pour l'expression des émotions et des attitudes [Gobl et al. 2003]. Il ne faut cependant pas non plus réduire la prosodie des émotions à ce paramètre de qualité de voix, la prosodie des émotions fait intervenir elle aussi les autres dimensions de la prosodie [Bänziger et al. 2003 ; Audibert et al. 2006b ; Mozziconaccim 2010]

La qualité de voix est définie dans le « Psychoacoustical terminology, Technical Report (American National Standard Report) » comme suit :

> *« The quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar"* [Ansi 1973]

John Laver est un des premiers à avoir donné une description de la qualité de voix. Selon lui :

> *"Voice quality is conceived here in a broad sense, as the characteristic auditory colouring of an individual speaker's voice, and not in the more narrow sense of quality deriving solely from laryngeal activity. Both laryngeal and supralaryngeal features will be seen as contributing to voice quality". [Laver 1980, p.1]*

Selon ces descriptions, la qualité de voix remplit de multiples fonctions : elle est ce qui permet d'expliquer que les voix des locuteurs sont différentes l'une de l'autre. Autrement dit, elle est donc la « couleur » du timbre « modal » d'une voix individuelle. La qualité de voix dépend aussi de l'âge, du sexe et de l'état de santé. Egalement, et c'est ce qui nous intéresse, elle varie avec les états affectifs du locuteur, que ce soit les états émotionnels ou les états attitudinaux [Aubergé 2002b ; Shochi et al. 2006]. Les changements de la qualité de voix dans l'énoncé peuvent intéresser les informations paralinguistiques comme l'état émotionnel du locuteur, l'attitude sur le message, ou les caractéristiques personnelles liées à des éléments physiques ou à la santé.

Les changements de la qualité de voix sont les résultats des changements dans la configuration du conduit vocal et de la source du larynx. Laver a développé un système pour décrire des qualités de voix différentes en fonction des changements laryngés et supra-laryngés. Ce système est utilisé pour distinguer les types différents de la qualité de voix comme : la voix modale (modal voice), la voix soufflée (breathy voice), la voix murmurée (whispery voice), la voix laryngalisée (creaky voice), la voix de fausset (falsetto voice), etc. [Gobl et al. 2003].

Contrairement à la fréquence fondamentale, la durée et l'intensité, qui sont directement inversables ou pour lequels il existe des techniques robustes de traitement du signal de la parole, il y a beaucoup de difficultés méthodologiques pour mesurer la qualité de voix, liées à la méconnaissance des structures acoustiques induites par ces changements de phonation. En fait, la qualité de voix

peut être mesurée par nombres d'approches différentes : dans la relation à la fréquence, la durée et l'intensité [Childers et al. 1991 ; Schröder 2004b] ou par une mesure articulatoire plus directe grâce à l'électroglottographie (EGG) [Gendrot et al. 2004]. Actuellement, il n'existe pas encore de méthode bien établie pour l'estimation complète de la qualité de voix en termes de paramètres acoustiques mesurables. La qualité de voix est donc un problème qui concerne le champ de la parole expressive.

2.2. La prosodie dans le système de la communication

Le rôle de la prosodie dans la communication homme-homme est largement étudié. Depuis très longtemps on connaît la fonction d'énonciation de la prosodie en cohérence avec la syntaxe (certains auteurs la disaient même il y a quelques années congruente avec la syntaxe) comme le rappelle par exemple Rossi [Caelen-Haumont et al. 1997 ; Rossi 1999]. Mais la prosodie remplit de multiples autres fonctions dans la communication, elles aussi mises en évidence et modélisées dans de nombreuses langues [Hirst et al. 1998]. Aubergé [2002a] propose même que toutes les fonctions communicatives (syntaxalisation, focalisation, modalisation, attitudinalisation etc.) seraient organisées en émergence d'une architecture modulaire coopérative (de type multi-agents), dans lequel la prosodie est le seul module à remplir toutes les fonctions (Figure 4). Dans ce système, la prosodie réalise ses fonctions en coopération interactive avec les autres agents du système, le lexique, la morpho-syntaxe etc. La cohérence entre les réalisations est assurée par des points de rendez-vous structurels entre la prosodie et les niveaux linguistiques/paralinguistiques [Aubergé 1991].

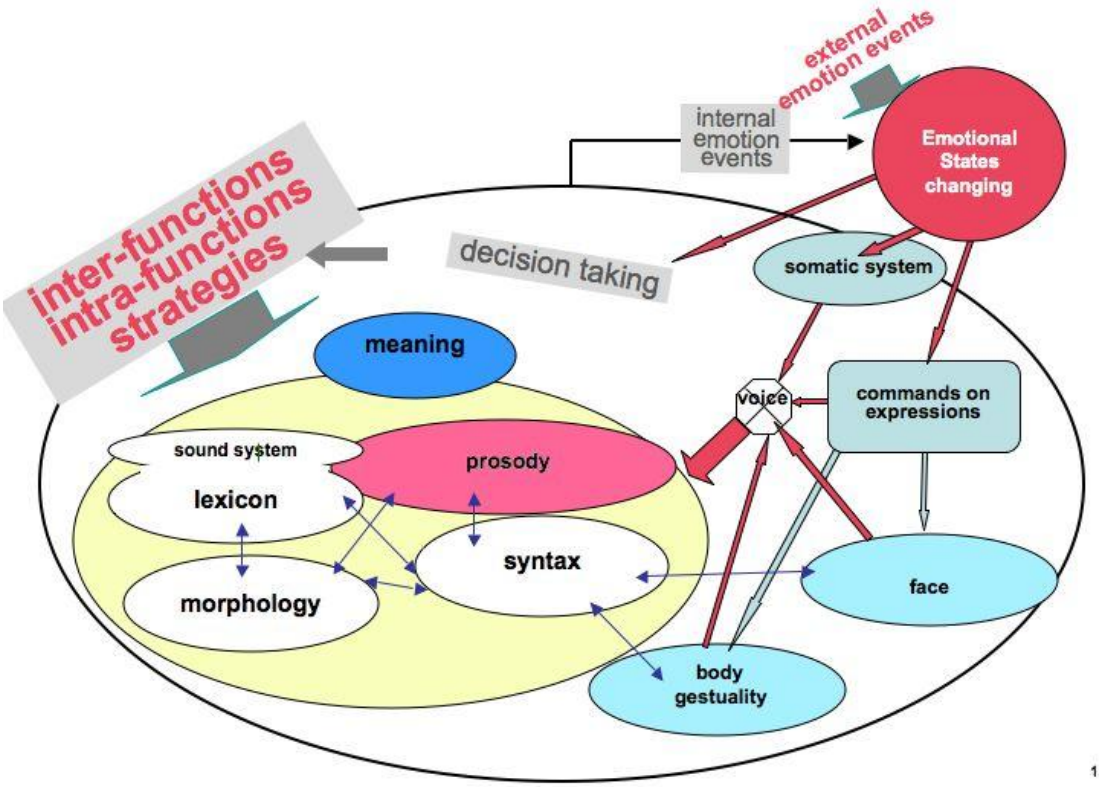


Figure 4: Le système de communication réparti sur des agents morphologiques en coopération interactive [Aubergé 2003]

2.3. Fonctions expressives de la prosodie

D'un point de vue fonctionnel, la prosodie est considérée comme un véhicule qui transporte des informations linguistiques et extralinguistiques d'un interlocuteur. Selon Fujisaki :

> "*Prosody […] serves to convey not only linguistic information, but also paralinguistic and non-linguistic information*" [Fujisaki 1997, p.28]

Aubergé [2002a] propose que les fonctions de la prosodie s'organisent selon trois niveaux de structuration des expressions : la fonction linguistique, la fonction attitudinale et la fonction émotionnelle (Figure 5). Les émotions relèvent d'un contrôle involontaire des émotions, s'expriment directement dans la voix et s'organisent dans le temps des événements émotionnels, et s'ancrent ainsi par « contrainte » dans la chaîne énonciative. Les attitudes et autres affects sociaux sont exprimés directement dans la parole dans un contrôle volontaire, intentionnel, donc langagier. Ces expressions sont ancrées dans le temps énonciatif et portent même le domaine temporel énonciatif. Enfin le choix des structures prosodiques à fonctions linguistiques relève, par cette « méta-structure », des indices sur les états affectifs du locuteur : il n'existe pas, selon [Aubergé 2003] de para-phrases prosodiques, mais le choix même d'une structure parmi des structures linguistiquement équivalentes porte des valeurs affectives.

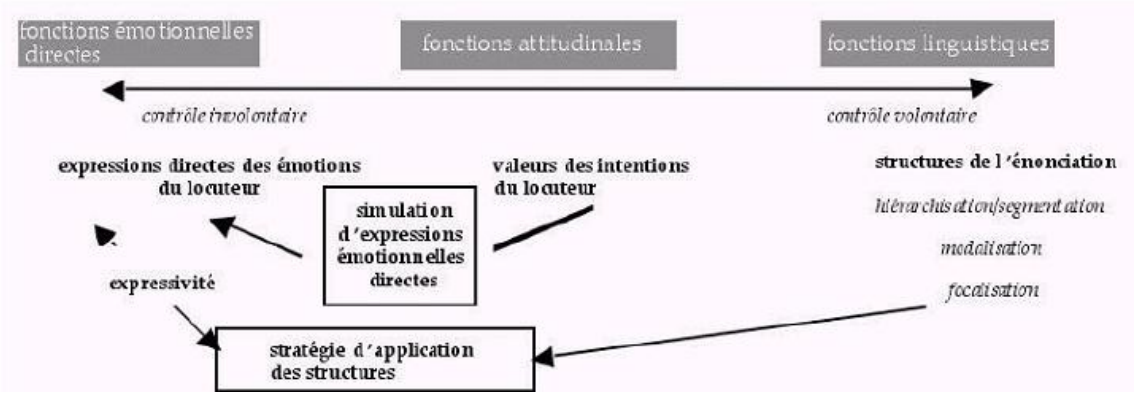


Figure 5 Les trois fonctions de la prosodie [Aubergé 2003]

2.3.1. Fonctions linguistiques

Les fonctions linguistiques de la prosodie sont contrôlées volontairement par le locuteur pour encoder des informations linguistiques dans son discours. En termes de fonctions linguistiques, la prosodie est utilisée pour véhiculer des structures d'accès lexical (c'est-à-dire du niveau de la seconde articulation du langage ou des structures de la première articulation.

- les fonctions lexicales comme l'accent lexical et le ton :
  - ✓ l'accent lexical (qui ne doit pas être confondu avec l'accent prosodique ou la proéminence prosodique qui est un fait avant tout acoustique) est réalisé selon les langues par la variation « locale », c'est-à-dire « autour » du segment phonologique concerné, de la fréquence fondamentale, de la durée (quantité vocalique) et/ou de l'intensité ;
  - ✓ les tons concernent avant tout la fréquence fondamentale et se caractérisent par non pas une simple variation (donc binaire) de la hauteur mélodique locale au segment, mais par des variations de hauteur et de contours mélodiques catégorisables en plusieurs valeurs. Certaines langues (par exemple dans le cas du vietnamien) peuvent faire intervenir conjointement la qualité de voix dans la caractérisation de la différenciation des tons. Le mandarin, le vietnamien, le thaï sont des exemples de langues à tons lexicaux. Pour ces langues, chaque syllabe est lexicalement marquée par un ton lexical. Les tons ont des contours prosodiques distinctifs. La modification du contour mélodique de tons peut changer la signification lexicale d'un mot, et peut-être le sens d'une phrase. Par exemple, en vietnamien, le mot « ba » - (le père – ton plat) est différent de « bà » (la grand-mère – ton descendant).
- les fonctions de la première articulation du discours comme la segmentation et la focalisation :
  - ✓ la fonction de segmentation est réalisée par l'interaction entre la prosodie et la syntaxe (voix Figure 4). Dans le discours, les composants syntaxiques peuvent être segmentés par des signes

prosodiques ; par exemple, la hauteur (la fréquence fondamentale) du discours est généralement haute au début et descendante en fin de la phrase [Grosz et al. 1992 ; Caelen-Haumont et al. 2000]

- ✓ la fonction de focalisation est utilisée pour faire une marque (une emphase) du locuteur dans son discours ; cette emphase peut être implémentée par des facteurs prosodiques, souvent marquée par une forte montée tonale sur la syllabe non finale dans un mot [Caelen-Haumont et al. 2000 ; Park 2000] ;

- les fonctions de modalisation sont liées aux modalités de phrase telles que la phrase assertive, la phrase interrogative ou la phrase impérative. Ces fonctions peuvent être implémentées par le changement du contour prosodique de la phrase ; par exemple, l'intonation de la phrase déclarative se termine normalement en descendant, et l'intonation d'une question oui-non montre un contour montant [Delattre 1966 ; Aubergé 1992 ; Hirst et al. 1998 ; Yuan et al. 2002 ; Vu et al. 2006b].

2.3.2. Fonctions attitudinales

La fonction attitudinale de la prosodie a été longtemps étudiée en phonostylistique, en particulier dans la motivation de l'apprentissage de langue seconde. On en voit déjà la trace dès le 19ème siècle, Pé Passy a montré que les contours mélodiques différents du mot « oui » peuvent porter des valeurs modales, illocutoires et attitudinales différentes [Léon 2000]:

Oui↓ [C'est mon avis].

Oui↓↓ [J'affirme cela].

Oui↑ [Est-ce vrai ?]

Oui↑↑ montée forte [Pas possible !]

Oui↓↑ [C'est possible mais j'en doute].

Oui↑↓ [C'est bien clair].

Oui↑↓↑ [Sans doute, au premier abord ; mais...]

Comme mentionné dans la section précédente, les attitudes sont des codes, construits et acquis dans la dynamique du langage et de ses réalités socioculturelles. Ainsi, les caractéristiques prosodiques des attitudes peuvent être différentes entre des langues et des cultures [Shochi et al. 2005 ; Rilliard et al. 2009]. Dans l'étude contrastive de l'intonation expressive en français et en vietnamien, Le Thi Xuyen [1989] a montré dans le cas de 12 émotions ou attitudes des caractéristiques communes et des différences entre le français et le vietnamien. Par exemple, pour les deux langues, la colère est caractérisée par un registre haut, une intensité forte et un débit rapide. Au contraire, les contours prosodiques de l'ironie sont différents entre les deux langues : en vietnamien,

l'ironie est caractérisée par un allongement considérable des voyelles ; tandis qu'en français, le contour prosodique de l'ironie monte assez vite et retombe lentement du départ et le changement mélodique n'est pas très marqué.

2.3.3. Les fonctions émotionnelles

Les fonctions émotionnelles de la prosodie sont contrôlées involontairement pour exprimer l'état émotionnel du locuteur [Aubergé 2002a]. La plupart des études sur l'émotion essaient de trouver les corrélats acoustiques (les corrélats des caractéristiques prosodiques) des catégories d'émotions. Par exemple, certains travaux de recherche montrent que l'excitation peut être exprimée par une intonation présentant une mélodie élevée et un rythme rapide, tandis que la tristesse est exprimée par une mélodie basse et un rythme lent ; la colère est caractérisée par plus d'articulation, un rythme rapide, et une mélodie qui monte globalement [Banse et al. 1996]. Audibert [2008] montre que aucun paramètre à lui seul ne peut véhiculer une émotion et que des caractéristiques globales ne suffisent pas, les contours semblent plus pertinents. En fait, l'étude de l'émotion dans la parole naturelle est très complexe. De plus il existe généralement des sentiments mitigés et des états d'ambiguïté (*blended emotions*), et alors les émotions ne tombent pas dans des catégories claires.

2.4. Conclusion

À travers plusieurs travaux de recherche sur le concept de la prosodie et ses fonctions dans la communication parlée, il a été montré que la prosodie joue des rôles très importants et complexes dans l'encodage du message parlé. Comme nous l'avons mentionné, notre objectif de thèse est d'étudier l'attitude en vietnamien pour appliquer nos résultats à la conception d'un système de synthèse de la parole, à entrer en interaction en dialogue ou plus généralement à exprimer les intentions du sujet dans ses énonciations. C'est pourquoi dans le cadre de notre travail, la fonction le plus importante de la prosodie qui nous intéressera sera la fonction attitudinale. Notre travail sera ainsi divisé en trois principales étapes:

- étudier les caractéristiques de la prosodie attitudinale pour la langue vietnamienne ;
- puis proposer un modèle d'encodage du lien entre prosodie et attitude ;
- finalement valider ces caractéristiques prosodiques par des tests perceptifs sur la parole expressive synthétique qui sera générée sur la base de notre modèle proposé.

À cause de la complexité dans la mesure, la modélisation et la génération de la qualité de voix, nous nous concentrerons seulement sur les variations des trois paramètres classiques : la fréquence fondamentale, la durée et l'intensité. Cependant, dans la construction de nos corpus pour étudier les attitudes en

vietnamien, la qualité de voix sera aussi considérée afin de servir à d'autres travaux de recherche dans le futur.

Chapitre 3 : Génération de la parole expressive

Comme nous l'avons présenté dans les chapitres précédents, nous avons donné à notre étude sur la parole expressive un double objectif : participer à la compréhension des phénomènes mis en jeux et appliquer nos résultats dans la conception d'un système de synthèse de la parole expressive en langue vietnamienne. Dans cette section, nous présenterons tout d'abord quelques approches principales permettant d'ajouter de l'expressivité dans les techniques de synthèse de la parole actuelles. Puis, nous insisterons sur l'un des points-clé de la synthèse expressive : la modalisation et la génération de la prosodie. Ce chapitre se termine par des discussions dans le cadre de notre travail : quelle technique de synthèse et quel modèle de prosodie sont appropriés à la génération de la parole expressive vietnamienne ?

3.1. Expressivité dans les techniques de synthèse de la parole

Faire parler des machines est un rêve que l'homme tente de réaliser depuis longtemps. Nous pouvons juste rappeler que les premières machines parlantes mécaniques ont été proposées au cours de la 2ème moitié du 18ème siècle par les travaux de Ch. G. Kratzenstein de Copenhague (production de voyelles avec des orgues à tuyaux), suivi par les travaux beaucoup plus connus de Von Kempelen en 1791. C'est avec l'arrivée de l'électronique, puis de l'informatique que les systèmes de synthèse se sont rapidement développés (Vocoder de Homer Dudley (1939), Pattern Playback de Franklin Cooper (1950), synthèse à formants de Klatt (de 1960 à 1980), puis plus récemment, la synthèse par unités stockées développée pour de nombreuses langues) [Dutoit 1997].

Un système de synthèse de la parole à partir du texte (Text-To-Speech – TTS en anglais) est un système complexe qui est constitué de plusieurs composants (Figure 6). Cependant, dans cette section, nous n'avons pas l'intention de présenter toutes les techniques et tous les composants utilisés dans un tel système. Le lecteur pourra se référer à de nombreuses références de la littérature du domaine du traitement de la parole ([Dutoit 1997 ; Xuedong et al. 2001 ; Benesty 2008], etc.). Nous insisterons sur les approches principales permettant de produire de la synthèse de la parole expressive, ou autrement dit, les différentes approches permettant d'intégrer l'expressivité aux techniques de synthèse de la parole.

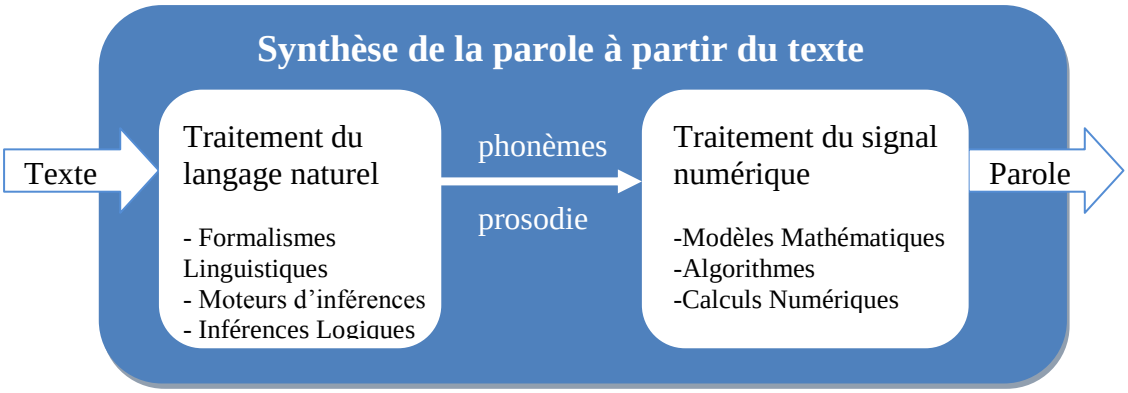


Figure 6: Diagramme fonctionnel d'un système de synthèse de la parole à partir du texte, selon Dutoit [1997]

Bien que les techniques de synthèse de la parole sont actuellement très développées et sont appliquées largement dans les systèmes d'interaction entre humain et machine, l'intégration de l'« expressivité » dans la voix synthétique est encore de nos jours un objet de recherche. La génération de voix reproduisant des émotions ou des attitudes comme l'autorité, etc. sont parmi les objectifs actuels de nombreux travaux de recherche.

Cependant, l'expressivité dans la parole synthétique a été étudiée depuis la fin des années 1980s, ce qui n'est pas récent. Les approches pour ajouter l'expressivité dans la parole synthétique sont différentes et souvent dépendantes des technologies de synthèse de la parole. Nous présentons dans les paragraphes suivants les principes d'intégration de l'expressivité dans les principales techniques de la synthèse de la parole.

3.1.1. Expressivité dans la synthèse par formants

La synthèse par formants est l'une des premières techniques de synthèse vocale qui a permis de produire de la parole intelligible. Cette méthode n'a pas besoin d'une base de sons humains enregistrés. Selon cette technique, les sons de la parole sont complètement produits par le synthétiseur, lui-même piloté par des règles basées sur les corrélats acoustiques des différents paramètres de la parole comme les fréquences centrales et les bandes passantes des formants et des antiformants. La Figure 7 présente un exemple d'un système de synthèse par formants proposé par Klatt [1980]. Parce qu'il est impossible de couvrir toute la complexité de la parole humaine par des règles, la voix synthétique ainsi produite n'est pas très naturelle.

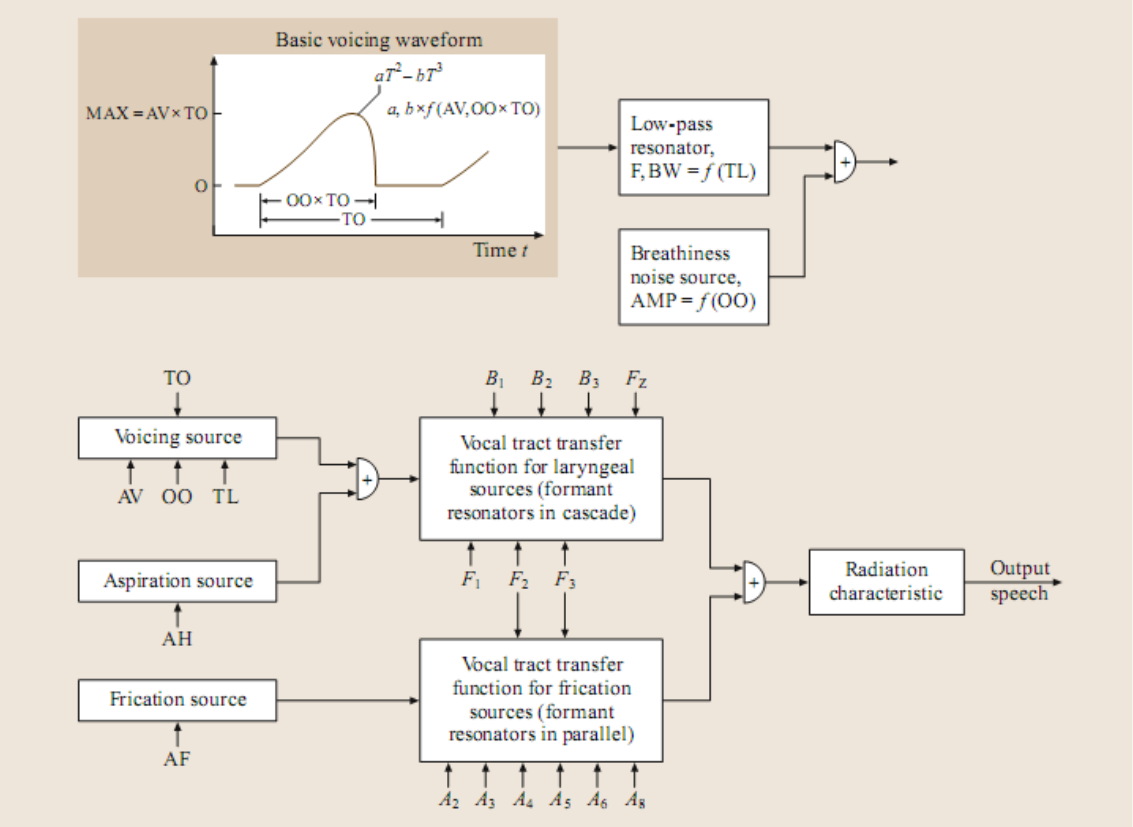


Figure 7: Structure générale d'un système de synthèse de la parole par formants proposé par Klatt [Benesty 2008]

Cependant, la synthèse par formants permet de contrôler facilement certains paramètres du processus de production de la parole, notamment les paramètres glottiques qui sont pertinents pour la génération de la parole expressive. C'est pourquoi, les premiers systèmes de synthèse vocale expressive (le système Affects Editor [Cahn 1990] et le système HAMLET de [Murray et al. 1995]) sont tous basés sur le système DECTalk [Hallahan 1995], un système commercial de synthèse de la parole par formants. Dans ces deux systèmes, pour chaque catégorie de l'émotion, les règles de modification de la sortie du synthétiseur sont obtenues à partir des analyses acoustiques sur des énoncés émotionnels. La Table 1 présente des exemples de règles pour les paramètres acoustiques pour quelques émotions dans des langues différentes. Ces règles sont appliquées dans les systèmes de synthèse par formants et puis sont évaluées par des tests perceptifs. Les résultats montrent que toutes les catégories des émotions synthétiques sont reconnues avec un score supérieur à celui du hasard [Schröder 2001].

Table 1 : Exemples des règles des paramètres acoustiques pour certains émotions [Schröder 2001]

Emotion Language Studied Percentage of recognition (chance level)	Parameter settings
Joy German 81% (1/9)	F0 mean: +50% F0 range: +100% Tempo: +30% Voice Qu.: modal or tense; "lip-spreading feature": F1 / F2 +10% Other: "wave pitch contour model": main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness American English 91% (1/6)	F0 mean: "0", reference line "-1", less final lowering "-5" F0 range: "-5", steeper accent shape "+6" Tempo: "-10", more fluent pauses "+5", hesitation pauses "+10" Loudness: "-5" Voice Qu.: breathiness "+10", brilliance "-9" Other: stress frequency "+1", precision of articulation "-5"
Anger British English	F0 mean: +10 Hz F0 range: +9 s.t. Tempo: +30 wpm Loudness: +6 dB Voice Qu.: laryngealisation +78%; F4 frequency -175 Hz Other: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear German 52% (1/9)	F0 mean: "+150%" F0 range: "+20%" Tempo: "+30%" Voice Qu.: falsetto
Surprise American English 44% (1/6)	F0 mean: "0", reference line "-8" F0 range: "+8", steeply rising contour slope "+10", steeper accent shape "+5" Tempo: "+4", less fluent pauses "-5", hesitation pauses "-10" Loudness: "+5" Voice Qu.: brilliance "-3"
Boredom Dutch 94% (1/7)	F0 mean: end frequency 65 Hz (male speech) F0 range: excursion size 4 s.t. Tempo: duration rel. to neutrality: 150% Other: final intonation pattern 3C, avoid final patterns 5&A and 12

Pour résumer, bien que la parole synthétique manque de naturel, les systèmes de synthèse par formants nous permettent de générer de la parole expressive bien contrôlé et avec une qualité suffisante pour distinguer les catégories des émotions ou attitudes synthétiques. Avec l'utilisation de la synthèse par formant, contrôlée par une évaluation perceptive, nous avons une manière flexible et simple pour étudier et vérifier les règles acoustiques proposées pour la génération des émotions et des attitudes.

3.1.2. Expressivité dans la synthèse par concaténation d'unités acoustiques

Contrairement à la synthèse par formants, la synthèse par concaténation utilise des segments de parole préenregistrés. Cette approche synthétise le signal par concaténation de ces segments, qui correspondent aux unités acoustiques. Le schéma général d'un système de synthèse par concaténation est présenté dans la Figure 8. Dans la phase de traitement de la parole, à partir de la base de données de parole, les unités acoustiques sont segmentées et stockées dans la base de données des segments de synthèse. Les informations phonétiques et prosodiques des segments sont aussi analysées pour l'entrée du synthétiseur. Dans la phase de traitement du signal, les séquences de segments sont choisies à partir de ces informations phonétiques et prosodiques. Ces segments sont ensuite envoyés au module d'adaptation de la prosodie pour ajuster fréquence fondamentale et durée à des valeurs appropriées. Puis, le module de concaténation regroupe les segments et élimine les discontinuités spectrales. Finalement, la parole synthétique est délivrée par le module de synthèse de la parole.

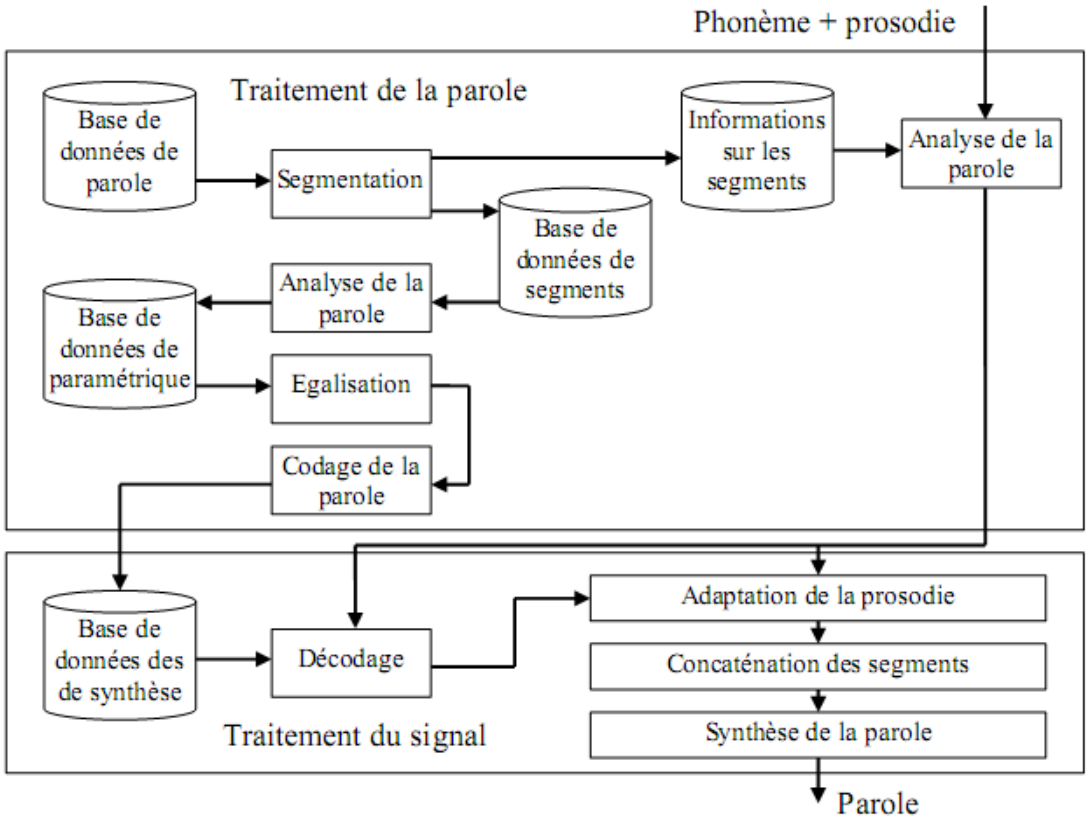


Figure 8: Schéma général d'un système de synthèse par concaténation [Dutoit 1997 ; Boite 2000]

Fondée sur des segments de parole naturelle préenregistrée, la synthèse par concaténation présente une qualité de parole synthétique plus naturelle que l'approche utilisant les formants. Selon la structure générale ci-dessus, la capacité et la qualité d'un système de synthèse par concaténation dépendent donc de nombreux facteurs, correspondant aux nombreux modules. Cependant, deux facteurs les plus importants sont le type d'unité acoustique et les fonctionnalités de l'algorithme de concaténation.

Choisir des unités acoustiques est la première tâche pour construire un système de synthèse par concaténation. Normalement, les unités longues diminuent la densité de points de concaténation ; elles tiennent mieux compte aussi de la coarticulation. En conséquence, le choix d'unités longues permet d'obtenir une meilleure qualité de synthèse.

Cependant, la longueur des unités influence le nombre des unités possibles stockées dans la base de données. La Table 2 présente un exemple des types d'unités acoustiques et leur nombre possible en anglais et vietnamien. Selon cet exemple, le nombre d'unités de type « mot » peut dépasser les centaines de milliers : pratiquement il s'avère impossible de réaliser un système de synthèse par concaténation basé uniquement sur des unités de type « mots ».

Bien que l'unité phonologique la plus courte soit le phonème, en pratique il a été démontré depuis de nombreuses années que le phonème n'est pas un bon candidat pour la synthèse car il est difficile de bien reproduire les transitions.

Dans les langues occidentales (français, anglais, etc.), l'unité minimale permettant d'obtenir une synthèse par concaténation de qualité acceptable est le diphone. Cette unité acoustique commence au milieu de la zone quasi-stable d'un phonème et se termine au milieu de la zone quasi-stable du phonème suivant.

Pour des langues isolantes telles que le thaï, le mandarin et le vietnamien, une autre unité alternative pour la synthèse par concaténation est la demi-syllabe : segment de signal qui est soit la moitié initiale d'une syllabe, soit la moitié finale. L'un des avantages de la demi-syllabe est que le nombre de demi-syllabes n'est pas beaucoup plus important que le nombre de diphones, alors que la qualité de synthèse basée sur la demi-syllabe semble meilleure pour ces langues isolantes [Kishore et al. 2003].

En fait, la plupart des systèmes de synthèse de la parole des années 1990s utilisent l'un des deux types d'unités pour la concaténation: les diphones ou les demi-syllabes [Benesty 2008].

Table 2: Les types unités acoustiques

Longueur d'unité	Type d'unité	Nombre d'unités		Qualité
		Anglais [Xuedong et al. 2001]	Vietnamien [Tran 2007]	
Courte	Phonème	42	40	Pauvre
↓	Diphone	~ 1500	620	↓
	Triphone	~ 30000		
	Demi-syllabe	~ 2000	690	
	Syllabe	~ 11000	7088	
	Mot	> 100K	>100K	
	Groupe	∞	∞	
Longue	Phrase	∞	∞	Haute

La pertinence de l'algorithme de concaténation est aussi très importante pour la synthèse par concaténation. En effet, dans la phase de la concaténation des segments acoustiques, des discontinuités peuvent apparaître du fait que les unités ont été extraites de signaux aux contextes différents. C'est pourquoi l'algorithme de concaténation doit être capable de minimiser ces discontinuités.

Parmi les techniques utilisées dans la synthèse par concaténation, la technique de synthèse dans le domaine temporel permet d'obtenir une parole de synthèse de très bonne qualité pour un temps de calcul minimal [Dutoit 1997]. Ce type de technique est basé sur l'algorithme PSOLA (Pitch-Synchronous OverLap-Add) [Charpentier et al. 1986] ou des versions dérivées de cet algorithme. PSOLA

permet de modifier les paramètres prosodiques et de concaténer les unités acoustiques. Il est fondé sur le principe de l'addition-recouvrement synchrone de la fréquence fondamentale de formes d'onde élémentaires. Cet algorithme a donné naissance à plusieurs versions dans le domaine temporel (comme TD-PSOLA), mais aussi dans le domaine fréquentiel (comme FD-PSOLA), en le combinant avec un synthétiseur LPC (LP-PSOLA) et MBROLA (Multi Band Overlap Add).

Généralement, tous ces algorithmes permettent de contrôler la fréquence fondamentale et la durée (et parfois l'intensité) dans la synthèse par concaténation. Certains travaux sur la synthèse par concaténation [Heuft et al. ; Vroomen et al. 1993 ; Rank et al. 1998 ; Montero et al. 1999] montrent qu'il est possible de simuler certains types d'émotions et attitudes en utilisant simplement des modifications de la fréquence fondamentale et de la durée. En fait, la contribution relative de ces paramètres par rapport à celle de la qualité de voix dépend du type d'émotion, selon Montero [1999] et Audibert [2006b], mais aussi du locuteur [Schröder 1999]. C'est pourquoi, avec certains types d'expressivité, les trois paramètres classiques sont suffisants pour implémenter la fonction expressive de la prosodie. L'avantage est qu'ils sont aussi contrôlés directement dans le processus de concaténation des segments acoustiques.

Pour la qualité de voix, alors que nous avons rappelé qu'elle apparaît comme le quatrième paramètre prosodique de la parole expressive [Campbell et al. 2003b], la synthèse par concaténation ne permet pas de la manipuler directement au niveau des procédures de traitement du signal. Autrement dit, dans la synthèse par concaténation, le type de qualité de la voix (par exemple la voix aspirée, murmurée, laryngalisée, etc.) est fixé : il est déterminé par le locuteur pendant l'enregistrement de la base de données.

Cependant, pour résoudre cette absence du contrôle de la qualité de voix dans la synthèse par concaténation, une approche qui consiste à enregistrer des unités acoustiques avec des qualités de voix différentes d'un même locuteur (et ce faisant, d'augmenter la base de données des unités), est proposée par Schröder and Grice [2003]. Selon cette approche, pour changer la qualité de voix, le système de synthèse va choisir les unités acoustiques préenregistrées avec la qualité de voix appropriée, dans une (ou plusieurs) bases de données comportant des signaux pour différents types de voix.

En conclusion, la synthèse par concaténation semble une méthode efficace pour étudier et générer de la parole expressive, bien qu'elle soit limitée dans le contrôle de la qualité de voix surtout si la base de données utilisée n'est pas très importante et ne présente qu'une seule qualité de voix. En fait, selon le sommaire en ligne de Felix Burkhardt[8], la plupart des systèmes de synthèse expressive actuels sont basés sur la synthèse par concaténation.

[8] http://emosamples.syntheticspeech.de/index.html

3.1.3. Expressivité dans la synthèse par sélection dynamique d'unités non uniformes à base de corpus

Au milieu des années 1990s, la méthode de synthèse par concaténation a été reconsidérée : au lieu d'utiliser un segment acoustique unique pour chaque unité de parole, produit dans des conditions d'intelligibilité segmentale optimale, mais dans un contexte prosodique non « naturel » qui sera ensuite modifié prosodiquement, les unités sont enregistrées dans plusieurs contextes prosodiques et sont produites selon des critères globaux propres au contexte de production du corpus. De plus, la taille des unités n'est pas fixée mais variable (segments de phrases, mots ou fragments de mots, syllabes, diphones ou même des sons isolés). L'idée est de sélectionner l'unité la plus longue possible existant dans le corpus de synthèse et correspondant au contexte de synthèse. C'est pourquoi, le corpus utilisé peut présenter une très grande taille (plusieurs heures) car il est nécessaire de stocker les unités acoustiques les plus diverses correspondant à un grand nombre de contextes syntaxiques, prosodiques, etc [Sagisaka 1988]. Il existe cependant de plus en plus de systèmes de synthèse basés sur ce principe et utilisant des petits corpus (synthèse par rushes), non pas explicitement produits pour la synthèse, mais recueillis en contextes naturels [Cadic 2011].

Au cours de la synthèse, ces unités sont sélectionnées et concaténées grâce à des algorithmes de sélection. Ces algorithmes utilisent généralement des coûts de calcul (pour déterminer le contexte le plus proche de la chaîne phonétique considérée) pour trouver la meilleure des unités candidates (Figure 9). Cette méthode est donc appelée la *synthèse par sélection*. Cette technique a permis de produire de la parole dont l'intelligibilité et le naturel rendent possible la confusion avec une prononciation humaine [Benesty 2008]. C'est pourquoi ce principe est à la base de la plupart des systèmes industriels actuels de synthèse de la parole, surtout pour des applications qui ont besoin d'une qualité professionnelle et qui ne sont pas limitées par la taille du corpus.

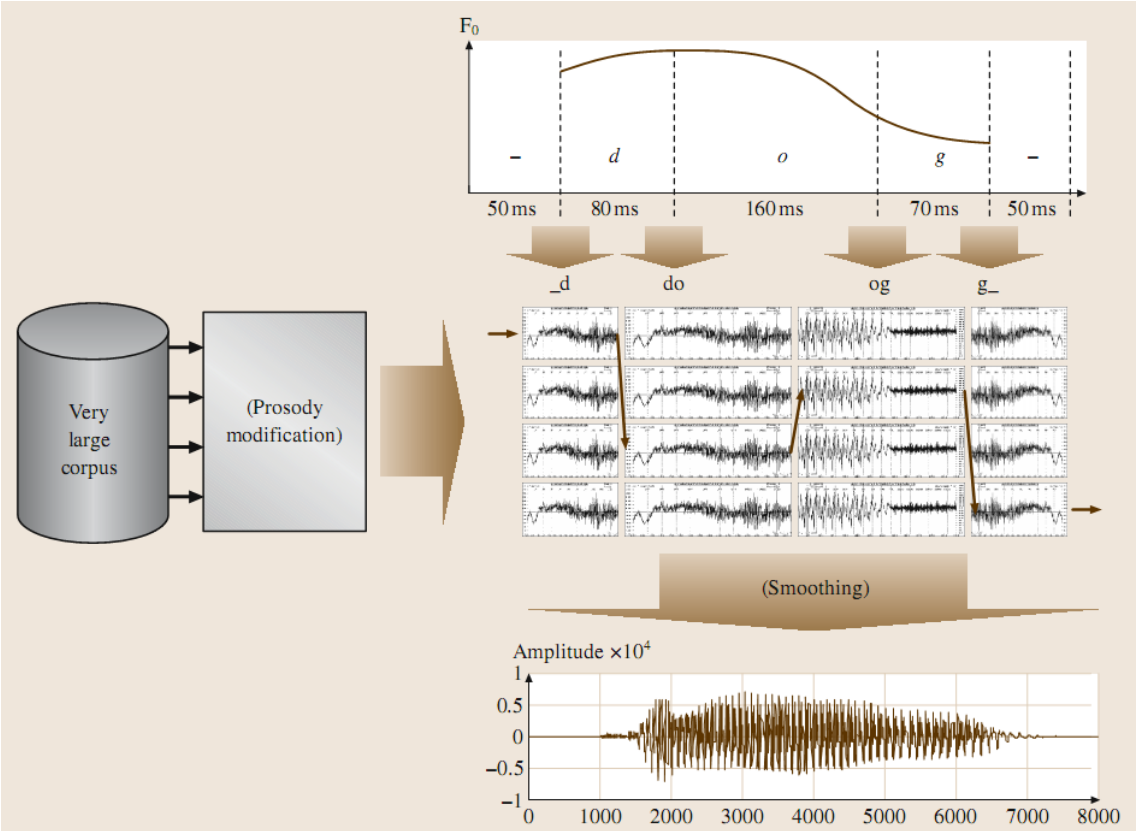


Figure 9: Un exemple de sélection et concaténation des unités dans la synthèse par sélection [Benesty 2008]

Pour ajouter de l'expressivité dans les systèmes de synthèse par sélection, il y a deux approches principales : l'approche basée sur l'utilisation de corpus étendus et l'approche basée sur la modification des signaux de parole.

Dans la première approche basée uniquement sur l'utilisation de corpus, il y a peu ou aucun contrôle sur la prosodie ou sur la qualité de voix dans les processus de synthèse. Pour ajouter de l'expressivité, le principe très simple consiste à ajouter dans le stockage des types de parole expressive. Cela peut être réalisé dans une ou plusieurs bases de données qui ont la même structure, et qui sont réalisées avec des enregistrements de types d'expressivité différents. Pour générer la parole artificielle, le système de synthèse va alors sélectionner les unités nécessaires à partir de la base de données correspondant à l'émotion ou à l'attitude choisie. Cette approche est appelée l'approche « playback ».

C'est ainsi qu'Iida [2003a] a produit un système de synthèse expressive avec trois émotions: bonheur, colère et tristesse. Les énoncés émotionnels générés par ce système ont été bien classés par les tests d'évaluation perceptive.

De la même manière, le système synthèse expressive de IBM [Hamza et al. 2004 ; Pitrelli et al. 2006] utilise des bases de données d'un locuteur avec deux types d'expressivité : «bonnes nouvelles» et « mauvaises nouvelles ». Une variante est proposée par Campbell et Marumoto [2000] qui stockent tous les types d'expressivité dans un seul corpus seulement. Cependant, dans le processus de

sélection, ces derniers utilisent des facteurs liés à la qualité de voix et à la prosodie comme critères de sélection des unités acoustiques et ont ainsi produit de la parole expressive avec trois émotions : la colère, la tristesse et la joie. Les tests perceptifs réalisés avec ces échantillons de parole synthétique expressive montrent que la colère et la tristesse sont bien reconnues (plus de 60%), et que la joie est moins bien reconnue mais avec un score restant au-dessus du niveau du hasard.

La deuxième approche, plus récente, de la synthèse de la parole expressive par sélection est la modification des signaux basée sur les caractéristiques prosodiques. Plusieurs techniques peuvent permettre de modifier les paramètres de la prosodie. Classiquement, la technique PSOLA (ou des algorithmes similaires) est utilisée pour modifier la fréquence fondamentale et la durée dans le processus de concaténation des unités [Zovato et al. 2004]. Pour la qualité de voix, de nombreuses techniques sur la transformation de voix sont proposées.

D'Alessandro et Doval [2003] ont proposé, par exemple, des algorithmes spécifiques qui permettent de modifier la qualité de voix d'un signal pour générer de la parole expressive. Vincent, Rosec, et Chonavel [2005], quant à eux, proposent une méthode d'analyse/synthèse pour l'estimation des coefficients de la source glottique et le filtre vocal. Ces méthodes ont été appliquées par Audibert [2006b] pour l'analyse du discours émotionnel et pour la modification de la qualité de voix dans la parole synthétique.

3.1.4. L'expressivité dans la synthèse statistique paramétrique

L'approche la plus récente dans la synthèse de la parole est la synthèse statistique paramétrique basée sur le modèle de Markov caché (HMM - Hidden Markov Models) [Black et al. 2007].

La Figure 10 présente le schéma général d'un système de synthèse basée sur l'utilisation des HMMs. Dans la phase d'entraînement, à partir de la base de données, des HMMs dépendant du contexte sont entrainés pour modéliser séparément les caractéristiques du spectre, la fréquence fondamentale et la durée. Un arbre de décision est utilisé pour organiser ces HMMs. Dans la phase de synthèse, à partir d'informations de contexte extraites du texte, les HMMs appropriés sont sélectionnés par l'arbre de décision. La séquence des états des HMMs sélectionnés décrivent les caractéristiques acoustiques par des paramètres acoustiques et prosodiques (valeur moyenne et écart-type). Ces paramètres sont les entrées du synthétiseur des signaux de la parole, normalement basés sur le filtre MLSA (Mel Log Spectrum Approximation) [Imai 1983].

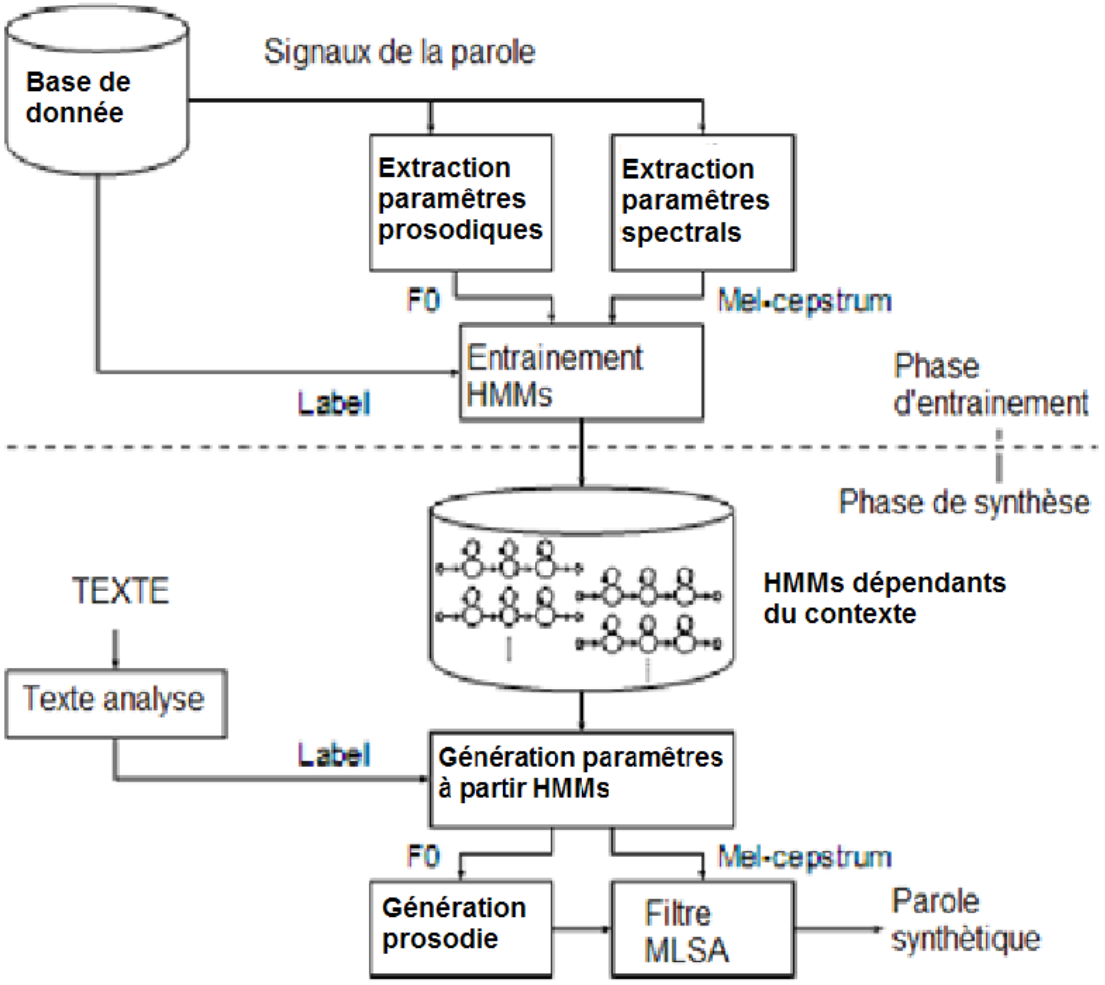


Figure 10: Schéma général du système de synthèse basé sur les HMMs, d'après [Zen et al. 2004]

Comme la synthèse par sélection, la synthèse basée sur l'utilisation d'HMMs utilise un grand corpus pour l'apprentissage. Cependant, au contraire du système de synthèse par sélection, la parole synthétique de la synthèse statistique paramétrique est obtenue à partir des paramètres stockés dans des arbres de décisions et des HMMs. Ainsi, après la phase d'apprentissage, le système de synthèse statistique paramétrique peut fonctionner sans corpus. Un avantage est alors que la synthèse statistique paramétrique peut être implémentée sur une machine moins performante que celle nécessaire à la synthèse par sélection. Par l'apprentissage du nombre des paramètres de la parole naturelle sur un grand corpus, la parole synthétique de la synthèse statistique paramétrique est lisse et assez naturelle [Black et al. 2007].

La manière la plus simple pour intégrer l'expressivité dans la synthèse statistique paramétrique est d'utiliser une grande base de données avec des types de parole expressive différents pour l'apprentissage. Suivant cette approche, [Yamagishi et al. 2003] ont entraîné des HMMs avec un corpus de quatre émotions différentes. À côté de la génération de la parole synthétique pour chaque type d'émotion, ils ont aussi produit une parole avec combinaison de deux émotions. Les bons résultats obtenus lors de leurs tests perceptifs confirment la possibilité de générer

de la parole expressive à l'aide de cette approche. Cependant, cette méthode a généralement besoin d'un très grand corpus et d'une phase d'entraînement compliquée et fastidieuse.

Une autre approche pour ajouter l'expressivité dans la synthèse statistique paramétrique est de faire une adaptation de la parole «neutre » vers la parole expressive. Selon cette approche, lors de la phase d'apprentissage, les paramètres de la parole « normale » sont adaptés à la parole expressive par un algorithme d'adaptation qui peut être réalisé avec des exemples de phrases variant de quelques dizaines à quelques centaines. Yamagishi, Onishi, Masuko, et Kobayashi [2007] montrent qu'il est possible d'adapter une voix « normale » à d'autres voix expressives avec des corpus adaptés de petite taille.

3.1.5. Conclusion

À travers les techniques principales de synthèse de la parole, il est montré qu'il y a deux approches principales pour ajouter de l'expressivité dans la parole synthétique :

- L'approche basée entièrement sur la réalisation du corpus (par exemple : la méthode « playback » de la synthèse par sélection, ou l'approche qui utilise un corpus contenant plusieurs types d'expressivité dans la synthèse par HMMs). Cette approche produit de la parole synthétique de haute qualité, mais nécessite pour chaque type d'expressivité un corpus pré-enregistré. Dans la phase de synthèse, le système utilise directement le corpus d'expressivité correspondant et aucune modification de la prosodie n'est nécessaire.
- L'approche basée sur des règles prosodiques (dans la synthèse par formants, la synthèse par concaténation, la modification des signaux dans la synthèse par sélection). Cette seconde approche permet de contrôler des paramètres de la prosodie lors du processus de synthèse. Ainsi, le système de synthèse n'a pas besoin de corpus particuliers pour la génération de l'expressivité. Cependant, le résultat de cette méthode est souvent considéré comme moins naturel par rapport à la méthode basée sur des corpus.

En fonction de l'objectif souhaité et de l'application considérée, il nous faut choisir une approche appropriée. Pour les applications industrielles, qui fonctionnent sur des machines très performantes et qui ont la possibilité d'utiliser de très grandes bases de données (par exemple : un système de central téléphonique ou un système d'annonces publiques), une approche basée principalement sur le corpus peut être utilisée. A contrario, pour des applications plus personnelles, qui peuvent être installées sur des systèmes légers ou portables (applications sur smartphone ou pour l'automobile), l'approche basée sur les règles prosodiques sera probablement un meilleur choix.

Notre objectif étant de générer une parole expressive pour le vietnamien, cette langue étant considérée comme une langue « peu dotée », un grand corpus qui contiendrait tous les formes expressives pour la synthèse par sélection ou pour synthèse statistique n'est pas encore disponible.

Notre travail va donc consister en

1) l'analyse des caractéristiques, en particulier des caractéristiques prosodiques, de la parole expressive vietnamienne ;

2) tenter de synthétiser des exemples de parole expressive en proposant une modélisation appropriée des paramètres prosodiques ;

3) évaluer cette synthèse (et sa modélisation sous-jacente) par des tests perceptifs.

Comme nous l'avons expliqué dans les paragraphes précédents, parmi les techniques disponibles pour la synthèse de la parole, la synthèse par concaténation peut générer de la parole avec une qualité tout à fait acceptable, sans avoir nécessairement besoin d'un grand corpus. En outre, cette technique nous permet de contrôler facilement les paramètres de la prosodie et, ainsi, nous permet de générer la parole pour différents types d'expressivité. En fait, à l'Institut MICA, un système de synthèse de la parole vietnamien à partir du texte a été développé par Tran [2007], qui a utilisé la technique de concaténation sur la demi-syllabe. Nous allons présenter en détail le système dans les chapitres suivants. Pour toutes ces raisons, la synthèse par concaténation sera naturellement notre choix pour générer la parole expressive en langue vietnamienne.

3.2. Les modèles de génération de la prosodie

Comme rappelé dans la section 3.1, la plupart des techniques de synthèse expressive utilisent des règles pour contrôler la prosodie. Ces règles prosodiques sont établies par un modèle de génération de la prosodie qui représente les liaisons entre les caractéristiques prosodiques du discours et les informations linguistiques (la séquence des phonèmes, les tons, le type de phrase, etc.) et, de même, avec les informations extralinguistiques (le style du discours, les modalités de la phrase, les émotions et attitudes, etc.).

Dans la section suivante, nous presenterons brièvement quelques modèles (les principaux, nous semble-t-il) de contrôle de la durée et de la fréquence fondamentale pour la génération de la parole expressive.

3.2.1. Les modèles de la durée

L'estimation précise des durées segmentales est une des principales tâches pour la génération de la parole. Le premier but de la modalisation de la durée est le choix d'une unité de base. La plupart des modèles utilisent le phonème comme unité de

base, car il est assez facile de segmenter les signaux de parole en phonèmes. Cependant, d'autres modèles [Campbell 1992 ; Barbosa et al. 1994 ; Morlec et al. 2001] sont fondés sur la syllabe ou des unités proches pour représenter les fonctions prosodiques au niveau de la phrase (la focalisation, les modalités de phrase ou les émotions et attitudes). Avec les langues isolantes telles que le thaï, le mandarin ou le vietnamien, le modèle de la durée est généralement basé sur la syllabe [Wu et al. 2001 ; Hansakunbuntheung et al. 2007 ; Tran et al. 2007].

Pour modaliser la durée, il y a ainsi deux approches principales : l'approche basée sur des règles et l'approche statistique.

3.2.1.1 Approche basée sur des règles

Cette approche a été initialement utilisée pour la génération de la prosodie dans la synthèse par formants. Le premier modèle, largement connu dans cette approche, est le modèle proposé par Klatt [Allen et al. 1987]. Dans ce modèle, Klatt a utilisé un ensemble des règles qui concernent les influences de la structure phonétique, l'accent, le stress, etc. appliquées sur la durée du segment phonétique. La durée phonétique est obtenue par l'équation suivante :

$$DUR = MINDUR + \frac{(INHDUR - MINDUR)PERC}{100}$$

où MINDUR est la durée minimale du segment, INHDUR correspond à la durée intrinsèque du phonème ; le coefficient PERC est le pourcentage de raccourcissement obtenu à partir des règles prosodiques.

Ces modèles ont été développés et appliqués pour plusieurs langues comme l'anglais américain [Allen et al. 1987] et le français [Bartkova et al. 1987].

La limite de cette approche est qu'elle utilise un nombre de règles limité pour la durée. Ces règles ne sont pas suffisantes pour couvrir toutes les variabilités et les interactions des facteurs phonétiques et prosodiques sur la durée dans plusieurs contextes, en particulier en cas de forte expressivité. C'est pourquoi avec cette méthode, même si la durée peut être prédite facilement, elle n'est pas en fait très exacte.

3.2.1.2 Approche basée sur le corpus

Loin du développement de la synthèse basée sur le corpus, de grandes bases de données ont été aussi utilisées pour former des modèles prosodiques.

Pour la modélisation de la durée, un modèle bien connu est le modèle *Sommes de produits* qui est proposé par Van Santen [1992]. Ce modèle est utilisé dans le système de synthèse multilingue de Bell Labs [Sproat 1998]. Selon ce modèle, la durée du segment phonétique est réalisée par une combinaison de phonèmes / contextes. Il est décrit par le vecteur f, qui est exprimé comme suit:

$$Dur(f) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}(f_j)$$

où les paramètres $S_{i,j}$ sont des fonctions représentant l'influence des facteurs i,j.

Pour construire un modèle de somme de produits en synthèse de la parole, un grand corpus de texte est utilisé et toutes les informations pertinentes de la durée sont calculées et encodées dans des vecteurs S. Puis, les phrases du corpus sont enregistrées et sont segmentées en phonèmes. Ensuite, les phonèmes sont regroupés en sous-classes correspondant aux contextes. Grâce aux analyses statistiques, un arbre catégorie est construit pour contenir les sous-classes des phonèmes. Pour chaque sous-classe (chaque catégorie de phonèmes), une phase d'entraînement sur des sommes de produits est réalisée pour déterminer quels niveaux peuvent être distingués par un facteur donné et l'interaction entre des facteurs. Après plusieurs répétitions, les prédictions de la durée de segments deviennent plus précises. Le modèle Somme-de-Produits est employé pour différentes langues telles que l'anglais, l'allemand [Mobius et al. 1996], le mandarin [Shih et al. 1997] et le néerlandais [Klabbers 2000].

Une autre approche basée sur corpus utilise, pour modéliser la durée, le modèle CARTs (Classification And RegressionTrees). Riley [1992] fut le premier à utiliser le modèle CARTs dans la prédiction de la durée pour la synthèse de la parole. Comme le modèle « somme de produit », le modèle avec CART utilise un grand corpus avec des phrases segmentées. À partir de ce corpus, un arbre binaire est construit par un algorithme d'apprentissage automatique sur la durée des unités dans des contextes différents. Cet arbre binaire avec ses nœuds représente la coarticulation entre les segments acoustiques : syllabes, mots, phrases, stress ou accent. Les valeurs de la durée prédite sont stockées par les feuilles de l'arbre. Dans la phase de synthèse, les valeurs prévues sont prédites en traversant l'arbre par des chemins satisfaisant les conditions aux nœuds intermédiaires, jusqu'au nœud de feuille. Ce modèle a été appliqué pour différentes langues telles que le coréen [Chung 2002],le tchèque [Batušek 2002], l'hindi [Krishna et al. 2004] et aussi en vietnamien [Tran et al. 2007].

D'autres systèmes de modélisation et de génération de la durée utilisent les réseaux de neurones. Ces méthodes sont basées sur l'hypothèse que les réseaux de neurones peuvent apprendre la durée des segments dans les interactions entre les effets contextuels. Des exemples ont été testés pour prédire la durée dans la synthèse de la parole en anglais [Campbell 1992], en français [Barbosa et al. 1994 ; Morlec et al. 2001] et en indien [Rao et al. 1997].

3.2.2. Modèles de génération de la fréquence fondamentale

Comme nous l'avons mentionné dans le chapitre qui précède, parmi les paramètres prosodiques, la fréquence fondamentale (notée F0) est le paramètre le

plus significatif. La variabilité de F0 est aussi la plus facile à percevoir. La génération du contour de la fréquence fondamentale est un des points clés dans la synthèse de la parole. Ceci peut être réalisé grâce à un modèle d'intonation, ce qui permet de décrire la variation de l'intonation du discours. Les études bibliographiques montrent qu'il existe plusieurs modèles d'intonation. Dans cette section, nous nous focalisons sur des modèles qui ont été utilisés dans les systèmes de synthèse de la parole. Chaque modèle a une façon particulière pour représenter et pour générer la fréquence fondamentale.

3.2.2.1 Approche par point-cible

L'idée centrale de ces modèles par point-cible est que la perception et la production de l'intonation du discours sont principalement basées sur certains points-cibles de la fréquence fondamentale. La transition entre ces points-cibles n'est pas très significative et peut être prédite par interpolation.

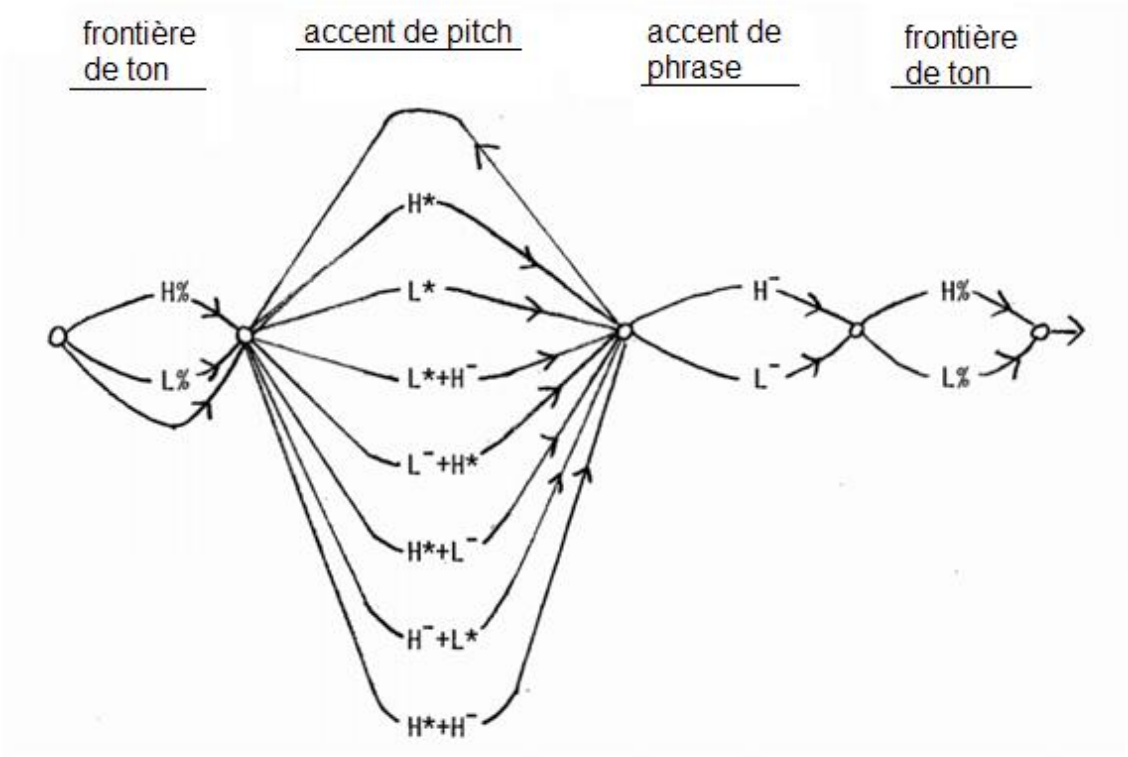


Figure 11: Grammaire intonative à état finis pour des séquences de tons [Pierrehumbert 1980]

Selon ce point de vue, les courbes mélodiques peuvent être représentées par une séquence de tons (bas/haut) [Pierrehumbert 1980], et sont transcrites par un système de symboles prosodiques. Chaque modèle présente un système de symboles différent, qui permet de représenter formellement les fonctions du contour prosodique. Pierrehumbert [1980] utilise deux catégories de tons relatifs : tons H (haut) et L (bas). En outre, elle a aussi utilisé d'autres symboles pour décrire l'accent de pitch (*), l'accent de phrase (-) et la frontière de ton (%). Les séquences de tons, qui sont représentées par les symboles ci-dessus, sont limitées par une grammaire à états finis (Figure 11). Pierrehumbert a aussi proposé un

ensemble de règles phonétiques pour la génération du contour de F0 avec le modèle d'intonation décrit ci-dessus [Pierrehumbert 1981].

Suivant les travaux de Pierrehumbert, un système de transcription de l'intonation a été proposé par Silverman [1992]. Ce système appelé ToBI (Tone and Break Indices) contient trois niveaux. Le premier niveau correspond à la présentation de la séquence des tons comme dans le modèle de Pierrehumbert. Il se compose des symboles mélodiques : H*, !H*, L*, L+H*, L*+H ou H+!H*, où « * » indique l'alignement avec la syllabe portant l'accent (stress) et « ! » signale une baisse de F0 par rapport à l'accent précédent. Le deuxième niveau utilise une échelle de 0 à 4 pour indiquer la frontière de mot (0) à la frontière de phrase (4). Le dernier niveau permet d'étiqueter diverses informations comme l'hésitation, les sons de non-parole, la pause, le bruit, etc.

Le modèle ToBI a été appliqué pour la génération de la prosodie dans les deux approches suivantes : la génération par règles [Jilka et al. 1999] et la génération par apprentissage automatique [Black et al. 1996].

De par sa construction, ToBI dépend de la langue à laquelle il est appliqué. Initialement, le modèle ToBI a été conçu pour l'anglais, mais plus tard, d'autres versions de ToBI ont vu le jour, adaptées pour la transcription de l'intonation d'autres langues telles que le néerlandais (ToDI), le grec (GRToBI), le coréen (K-ToBI), le japonais (J_ToBI), l'allemand (GToBI) et le russe (TORI) [Di Cristo 2004].

Une des limites très importante de ToBI pour la synthèse est qu'il perd les « détails » des contours mélodiques, dont il a été montré qu'ils sont pertinents pour certaines fonctions (cf. par exemple le focus [Aubergé et al. 2005]). D'une manière générale, ToBI est une modélisation qui réduit les valeurs des contours prosodiques (par la tonologie) et qui n'est pas apte à rendre compte de la prédictibilité largement démontrée des contours prosodiques, en particulier pour les fonctions expressives qui nous intéressent ici [Aubergé et al. 1997 ; Audibert et al. 2007].

3.2.2.2 Approches par apprentissage automatique

Dans ces approches, la fréquence fondamentale est généralement le paramètre considéré dans le système à apprentissage automatique. La qualité du contour mélodique en sortie va dépendre alors de la taille de la base de données et des paramètres choisis.

En utilisant un réseau de neurones avec trois niveaux, Sagisaka [1990] prédit le contour mélodique de la phrase en japonais. Pour les entrées, il utilise cinq paramètres correspondant à l'accent et la longueur du syntagme courant et du syntagme précédant/suivant. La sortie prédit trois valeurs de F0 (initiale, maximale et finale), qui présente la forme générale du segment syntagmatique

correspondant. Le résultat de cette prédiction étant assez bon, il peut être utilisé pour prédire le contour de la phrase complète, selon Sagisaka [1990].

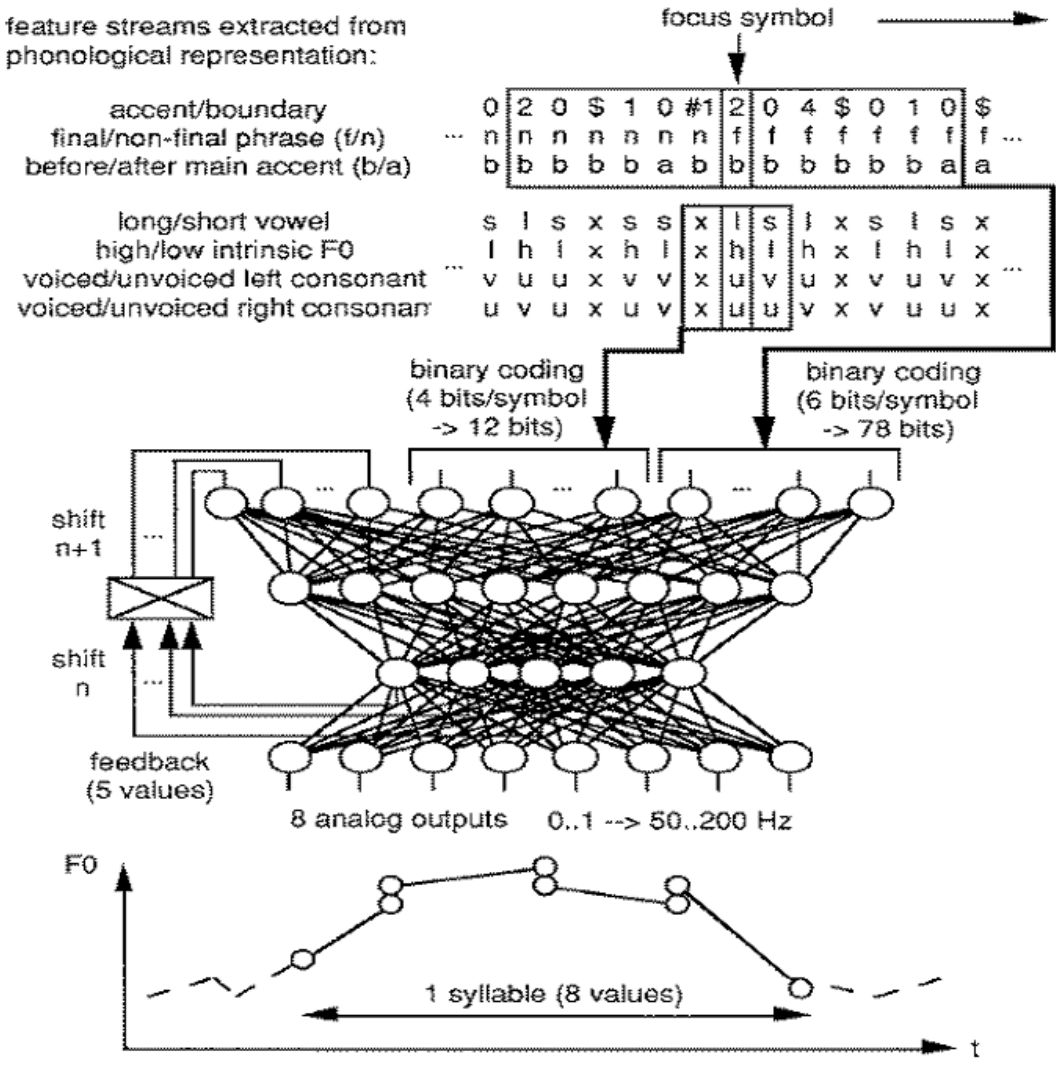


Figure 12: Génération de contour de F0 utilisant réseau des neurones [Traber 1990]

Traber [1990] a aussi utilisé un réseau de neurones pour la génération de l'intonation en allemand (Figure 12). Son réseau de neurones contient deux couches cachées avec des entrées qui comportent les informations suivantes :

- des informations d'ordre macro prosodique :
  - ✓ quatre niveaux accentuels par rapport au syntagme ;
  - ✓ le symbole de frontière de mot
  - ✓ le type de syntagme (final/non-final)
  - ✓ la position avant / après par rapport à l'accent de syntagme
- des informations d'ordre micro-prosodique :
  - ✓ voyelle longue/courte
  - ✓ voyelle haute/basse par rapport à la fréquence fondamentale intrinsèque, voisement du contexte consonantique (gauche et droit).

La sortie est le contour de F0 représenté par un contour de syllabe qui est stylisé par 8 valeurs de F0. Le contour complété de la phrase est obtenu par la concaténation simultanée des contours des syllabes. Le modèle de Trabe est implémenté dans le système de synthèse de la parole SVOX [Traber 1993].

Fondée sur de grands corpus, l'approche par apprentissage automatique nous permet de produire un système de génération de l'intonation avec plusieurs langues et plusieurs types d'expressivité différents. Récemment, cette approche a été largement appliquée dans la génération de l'intonation de plusieurs langues telles que l'anglais, le français, l'allemand [Buhmann et al. 2000], l'indien [Sreenivasa Rao et al. 2009] et le mandarin [Yu et al. ; Hwang et al. 1995 ; Chen et al. 1998]. Cependant, comme c'est essentiellement une méthode statistique qui est utilisée, cette approche ne permet pas de savoir comment se fait le passage entre les descriptions linguistiques/extralinguistiques et les paramètres prosodiques de la sortie.

3.2.2.3 Approches par superposition

Les approches basées sur la superposition considèrent que le contour intonatif est une fonction complexe qui peut être décomposée en des composants plus simples. Le premier modèle de superposition a été proposé par Öhman [1967]. Dans son modèle, la courbe mélodique contient deux composantes liées à deux niveaux linguistiques : le mot et la phrase. Cette idée initiale est reprise dans le modèle de commande mélodique de Fujisaki [1983], qui est maintenant le modèle d'intonation le plus connu dans l'approche par superposition. Dans ce modèle, le contour de F0 est considéré comme une superposition additive de deux composantes :

- la composante de phrase caractérise la tendance générale de l'intonation de l'énoncé, et est représentée par le mouvement global du contour F0 ; cette composante est modélisée par la forme prosodique globale, qui est déterminée par des impulsions.
- la composante d'accent caractérise des proéminences particulières de l'intonation (par exemple stress/accent de syllabe ou de mot) ; cette composante est modélisée par des fonctions échelons. Les deux types de commande génèrent des contours sur une échelle logarithmique, qui sont ensuite combinés par superposition pour produire le contour additif de F0 (Figure 13).

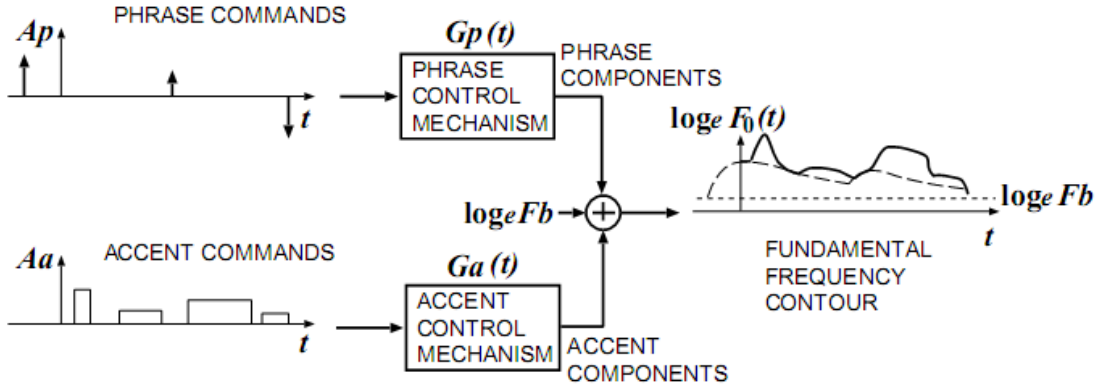


Figure 13: Schéma général du modèle de commande de Fujisaki [Fujisaki 2003]

Pour pouvoir appliquer ce modèle en synthèse de la parole, la position et l'amplitude des impulsions (de la phrase) et les fonctions échelons (de syllabe ou de mot) doivent être déterminées par les propriétés linguistiques du texte. Ces propriétés peuvent être obtenues par une analyse manuelle [Fujisaki et al. 2005] ou par un système d'apprentissage automatique [Mixdorff et al. 2001]. Le modèle de Fujisaki a été utilisé avec succès pour modéliser et générer l'intonation de nombreuses langues telles que l'anglais, l'allemand, le grec, le coréen, l'espagnol et le suédois [Fujisaki et al. 1998 ; Fujisaki 2002]. Ce modèle a aussi été modifié pour l'appliquer aux langues tonales comme le mandarin [Mixdorff et al. 2003b], le thaï [Mixdorff et al. 2002] et aussi le vietnamien [Mixdorff et al. 2003a]. Nous allons présenter en détail la manière d'utiliser ce modèle pour le vietnamien dans le chapitre suivant.

Avec la même approche que Fujisaki, Van Santen et Möbius [1999] proposent un modèle de contour F0 qui contient trois composants correspondant à la phrase, l'accent et la courbe de perturbations segmentales. Ce modèle est utilisé dans le système de synthèse multilingue de Bell Laboratories [Sproat 1998 ; van Santen et al. 1998].

Modèle de superposition des contours fonctionnels

Les modèles de Fujisaki ou de Van Santen décrits ci-dessus sont plutôt basés sur les caractéristiques acoustiques d'énoncés. Les composants du contour prosodique dans ces modèles ne sont pas construits par les fonctions prosodiques, que nous avons présentées dans la section 2.3.

Fondé sur des fonctions différentes de la prosodie (cf. section 2.3), Aubergé [1991] propose le concept de « rendez-vous » structurel entre les différents modules linguistiques : prosodie, lexique, morpho-syntaxe, discours. Ces rendez-vous opérent pour les différentes fonctions de la communication (organisation de l'énonciation, focalisation, modalisation, attitude, émotion etc.). Selon elle, chaque niveau de description linguistique sur l'axe syntagmatique (la phrase, la proposition, le groupe et éventuellement le sous-groupe) correspond à une forme particulière des contours intonatifs. Les formes des contours sont indépendantes et différentes entre les différents niveaux. La forme du contour prosodique de

chaque niveau encode aussi une valeur fonctionnelle correspondant à son niveau. Par exemple, les contours du niveau de phrase véhiculent la fonction de la modalité de phrase ou l'attitude du locuteur, le contour au niveau de mot présente la fonction de focus, le contour au niveau segmental représente l'accent/le ton. Le contour intonatif global est donc la superposition des contours de tous les niveaux linguistiques.

Pour obtenir les contours des niveaux linguistiques, Aubergé [1993] a proposé une méthode d'analyse « top-down » du niveau porteur le plus large vers le niveau porteur le plus petit. Cette analyse a besoin de la construction d'un corpus très contrôlé pour obtenir les attributs de chaque niveau linguistique. À partir de ce corpus, le contour correspondant à chaque niveau est calculé par les valeurs moyennes.

Une autre approche pour obtenir la forme des contours des niveaux dans le modèle proposé par Aubergé est l'apprentissage automatique. En application de ce modèle dans la génération de la prosodie de six attitudes françaises, Morlec [1997b] utilise un réseau de neurones récurrent (RNR) pour prédire les contours prosodiques des niveaux linguistiques. Le contour prosodique global est obtenu par la somme pondérée de ces prédictions ajoutée aux valeurs prosodiques de référence du locuteur (registre, débit moyen...), comme présenté dans la Figure 14.

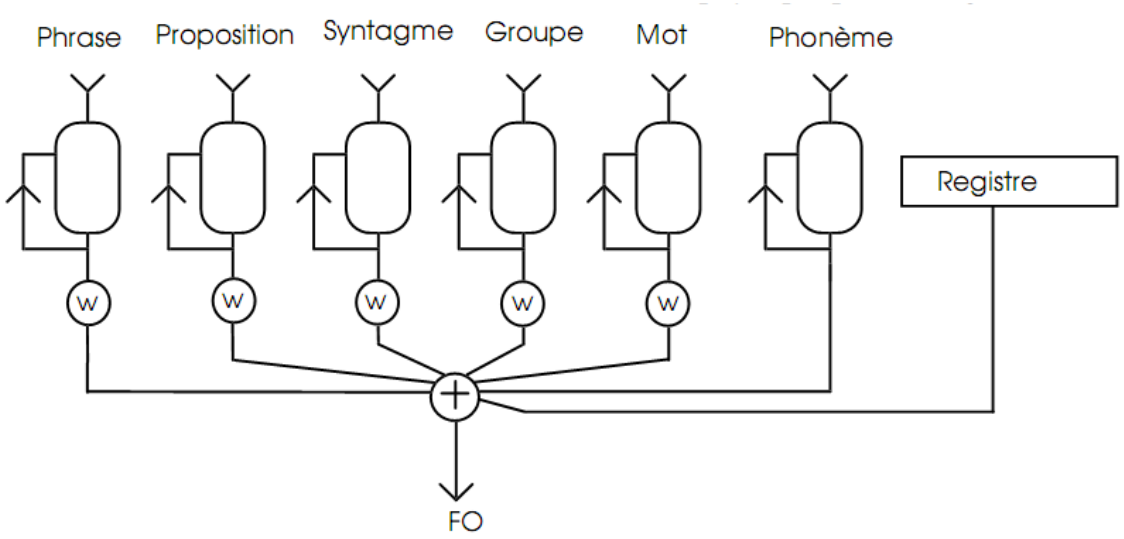


Figure 14: Architecture générale du modèle de Morlec [1997a]

À la suite de Morlec, Holm [2000] utilise ce modèle pour une application à l'énonciation de formules mathématiques. Ce modèle est donc appelé SFC (Superposition of Functional Contours). Il est ensuite appliqué pour la génération de la prosodie dans d'autres langues telle que l'allemand [Bailly et al. 2006], le mandarin [Chen et al. 2004] et l'espagnol [Bailly et al. 2008].

3.3. Conclusion

Dans cette section, nous avons présenté plusieurs modèles différents pour la génération de la durée et de l'intonation. Chaque modèle est implémenté par des stratégies et des algorithmes différents pour prédire la prosodie, qui peuvent être basés sur des règles prosodiques, sur des corpus d'apprentissage automatique (par réseau de neurones, arbre de décision, etc.) ou sur des caractéristiques acoustiques ou linguistiques. Parmi eux, le modèle de superposition des contours fonctionnels peut représenter complètement tous les attributs linguistiques et extralinguistiques. Ainsi, ce modèle nous permet de modéliser et de générer un bon nombre de fonctions de la prosodie, en particulier la fonction attitudinale.

Dans le cadre de notre travail, qui concerne la génération de la parole expressive d'une langue à tons, la prosodie est la combinaison complexe de la fonction lexicale, de la fonction locale du ton et de la fonction globale de l'attitude. Notre choix se porte alors sur l'utilisation de l'approche de superposition des contours fonctionnels, tout d'abord pour trouver les caractéristiques prosodiques correspondant aux tons et aux attitudes, puis dans un deuxième temps, pour modéliser la génération des attitudes en prosodie et appliquer cette modélisation en vietnamien. Nous allons présenter en détail notre approche dans la partie 2.

Chapitre 4 : La prosodie et la synthèse de parole en vietnamien

Après la présentation générale sur la parole expressive, la prosodie et la synthèse de parole, nous proposons au sein de ce chapitre d'aborder le vietnamien. Nous commençons d'abord avec la présentation en bref des caractéristiques phonétiques et phonologiques de la langue vietnamienne. Ensuite, nous parlerons de la prosodie du vietnamien. En ce domaine, après la présentation de certains travaux de recherche pour modéliser la prosodie du vietnamien, nous résumerons quelques travaux importants sur la synthèse de la parole vietnamienne.

4.1. La langue vietnamienne

La langue vietnamienne est la langue officielle du Vietnam qui compte environ 86,9 millions d'habitants (décembre 2010). De plus, il existe aussi 3 millions de locuteurs dans le reste du monde[9]. Le vietnamien appartient au groupe Viet-Muong, branche Mon-Khmer de la famille Austro-Asiatique [Mai et al. 2002]. La famille linguistique Mon-Khmer est largement parlée dans les pays de l'Asie du Sud-Est comme le Laos, Vietnam, Cambodge, Thaïlande. Actuellement, on dénombre 156 langues Mon-Khmers, parmi lesquelles le vietnamien est la plus populaire et la plus utilisée [Mai et al. 2002]. La langue vietnamienne moderne est aussi l'une des rares langues d'Asie qui utilise un système d'écriture latin.

Après un long processus de développement, aujourd'hui, le vietnamien est devenu une langue isolante et tonale. Ce sont les deux caractéristiques principales de la langue vietnamienne moderne.

4.1.1. Le vietnamien : une langue isolante

Les langues isolantes sont des langues qui expriment les divers rapports grammaticaux par des mots et des signes isolés. Le vietnamien et le chinois appartiennent à ce type de langues [Mai et al. 2002].

Comme langue isolante, le vietnamien est caractérisé par les phénomènes suivants :

- les mots sont morphologiquement invariables
  - ✓ il n'y a ni conjugaison des verbes, ni féminin ou pluriel des noms et des adjectifs ;

[9] www.gso.gov.vn/

- ✓ les fonctions grammaticales sont identifiées par l'ordre des mots ou par l'utilisation de mots outils ; par exemple, pour représenter le temps d'une action, le vietnamien ajoute un mot outil avant le verbe: tôi ***đã*** làm (j'ai fait) ; tôi sẽ làm (je vais faire) ; tôi đang làm (je suis en train de faire) ;

- la frontière de la syllabe et du morphème lexique est identique, une syllabe est un morphème :
  - ✓ dans l'écriture, les syllabes sont séparées par des espaces ; par exemple, le mot « đại học » (université) contient deux syllabes et deux morphèmes lexicaux
  - ✓ les mots vietnamiens sont morphologiquement mono-syllabiques et conceptuellement plutôt bisyllabiques
  - ✓ la syllabe joue le même rôle que le morphème lexical dans des langues avec des mots polysyllabiques comme l'anglais ou le français ; la syllabe se comporte donc comme une unité structurelle de base de la langue vietnamienne, appelée syllabe-morphème [Nguyen 1996] ou monosyllabisme [Nguyen 1994].

Structure de syllabe

Comme nous l'avons constaté ci-dessus, la syllabe est considérée comme l'unité structurelle de base de la langue vietnamienne. La Table 3 présente la structure phonologique d'une syllabe en vietnamien.

Table 3: La structure phonologique et le nombre de parties phonologiques de la syllabe en vietnamien, d'après [Tran et al. 2005]

<table>
<tr><td colspan="4">Syllabes avec ton (6492)</td></tr>
<tr><td colspan="4">Syllabes de base (2376)</td></tr>
<tr><td rowspan="3">INITIALE (22)
consonance</td><td colspan="3">FINALE (155)</td></tr>
<tr><td>Son pré-tonal (1)
semi-voyelle : u</td><td>Son noyau (16)
voyelle ou diphtongue</td><td>Son final (8)
consonne ou semi-voyelle u, i</td></tr>
<tr><td colspan="3">Ton (6)</td></tr>
</table>

Parmi les parties phonologiques de la syllabe ci-dessus, le son du noyau est la partie principale de la syllabe. Une syllabe contient toujours un son noyau. Le son initial, la partie pré-tonale et le son final sont optionnels dans les syllabes [Tran 2007].

Dans une syllabe vietnamienne, la consonne initiale contribue à créer la syllabe, mais elle ne participe pas à la construction du ton. Le ton vietnamien affecte seulement la partie finale de la syllabe [Tran et al. 2005].

4.1.2. Le vietnamien : une langue tonale

La langue vietnamienne est une langue tonale. Chaque syllabe a un ton. Le ton est nécessaire pour accéder à la signification d'une syllabe ou d'un mot. Le changement de ton d'une syllabe change son sens.

Nous voudrions rappeler que la famille linguistique Mon-Khmer, comporte d'autres langues tonales telles que le laotien (qui possède six tons) ou le thaï (qui possède cinq tons). En fait, le système tonal en langue vietnamienne est plus complexe, il varie en fonction des régions. Le nombre de tons peut varier de six (parlé de Hanoi) à cinq (parlé de Hô Chi Minh-ville) ou à quatre (dans les parlés du centre du Vietnam) [Do et al. 1998 ; Nguyen 2002]. Parmi eux, le système tonal de Hanoi (capitale du Vietnam) est considéré comme standard. Les six tons vietnamiens standards sont décrits dans la Table 4.

Table 4: Les six tons du vietnamien standard

Numéro ton	Signe diacritique	Nom du ton	Exemple
1		Plat (ngang)	ba (trois)
2	◣	Descendant (huyền)	bà (grand-mère)
3	~	Brisé (ngã)	bã (marc de cafe)
4	?	Interrogatif (hỏi)	bả (consonantique)
5	◤	Montant (sắc)	bá (roi)
6	■	Grave (nặng)	bạ (n'importe)

4.2. Prosodie de la langue vietnamienne

Nous présentons dans cette section les caractéristiques prosodiques du vietnamien. Cette présentation est basée sur la classification fonctionnelle de la prosodie (section 2.3).

4.2.1. Fonction linguistique de la prosodie en vietnamien

En termes de fonctions linguistiques, la prosodie de la langue vietnamienne est utilisée pour véhiculer la signification lexicale du ton (au niveau de syllabe) et la modalité au niveau de phrase.

4.2.1.1 Niveau de la syllabe : la fonction tonale

Son caractère tonal est l'un des traits les plus importants du vietnamien. Parmi trois paramètres prosodiques (la fréquence fondamentale - F0, l'intensité et la durée), la F0 est le principal paramètre servant à caractériser les tons [Doan 1997]. Les courbes de F0 des tons vietnamiens sont décrites par la Figure 16.

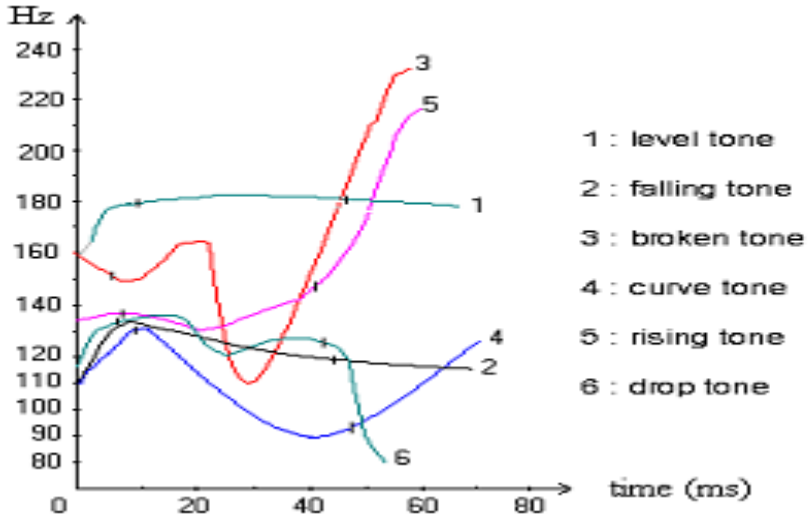


Figure 15: Les courbes classiques de fréquence fondamentale des tons du vietnamien [Doan 1997]

Les caractéristiques prosodiques des contours des six tons vietnamiens peuvent être décrites en bref comme suit [Han et al. 1974 ; Doan 1997 ; Do et al. 1998] :

- ton 1 - ton plat (ton « ngang ») : la courbe F0 est haute, plate, légèrement descendante à la fin ;
- ton 2 - ton descendant (ton « huyền ») : la forme de la courbe F0 descend à la fin du ton ;
- ton 3 - ton brisé (ton « ngã ») : sa courbe descend au début puis remonte ; il y a une glottalisation qui se produit vers le milieu du parcours ;
- ton 4 - ton interrogatif (ton « hỏi ») : présente un courbe concave qui commence par une descente au début et puis remonte vers la fin de sa réalisation ;
- ton 5 - ton montant (ton « sắc ») : la courbe de F0 est horizontale au début et monte ensuite fortement ;
- ton 6 - ton grave (ton « nặng ») : la courbe de F0 de ce ton est descendante et en général plus brève que les autres en raison d'une forte glottalisation à la fin.

Il existe une relation entre la distribution des tons dans la structure de la syllabe. Les syllabes ayant un son final /p/ ou /t/ou /k/ (syllabe fermée) ont seulement un ton montant (ton 5) ou un ton grave (ton 6). Dans ce cas-là, le ton 5 et 6 présente une durée plus courte. Ces deux représentations des tons 5 et 6 sont donc appelées les tons 5b et 6b [Nguyen 2002 ; Tran 2007]. Le Figure 16 montre un exemple des 8 représentations de tons vietnamiens.

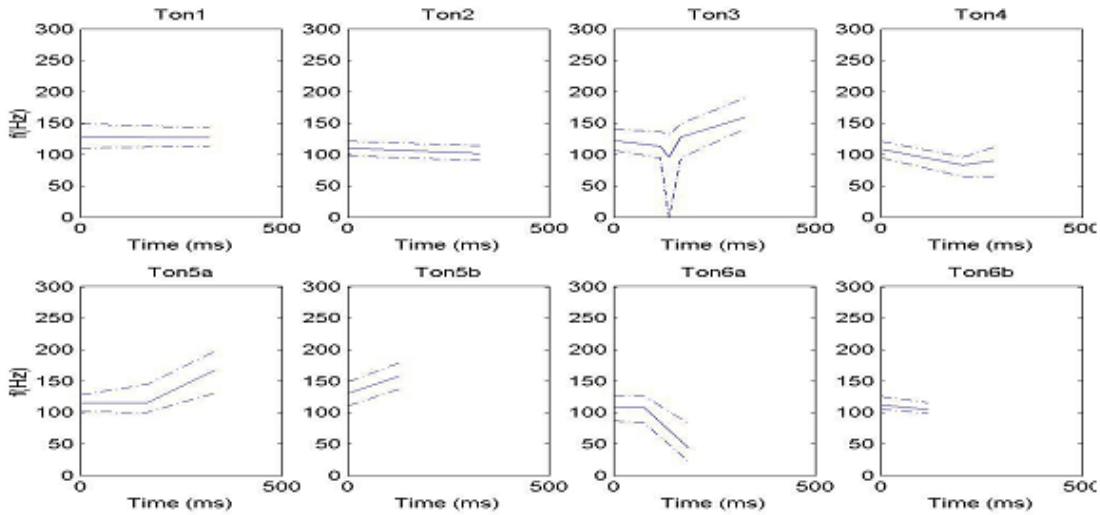


Figure 16: Exemple des contours des 8 représentations des tons vietnamiens [Pham et al. 2002]

En fonction de la hauteur de la F0, le six tons vietnamiens peuvent être divisés en deux catégories : tons de registre haut (ton 1, 3, 5, 5b) et tons de registre bas (ton 2, 4, 6 et 6b) [Do et al. 1998 ; Nguyen 2002] . La durée de la syllabe isolée en fonction des tons aussi divise les réalisations tonales en deux groupes : les tons longs (ton1, ton2, ton3, ton4, ton5a) et les tons courts (ton5b, ton6a, ton6b) [Doan 1997 ; Nguyen 2002].

Dans la parole continue, les caractéristiques prosodiques des tons varient considérablement selon leur environnement tonal immédiat. C'est le phénomène de coarticulation tonale ou « assimilation tonale » dans la parole continue [Han et al. 1974 ; Brunelle 2003].

La Figure 17 présente un exemple de la coarticulation tonale : si deux syllabes adjacentes contiennent deux tons montant, le ton montant de la deuxième syllabe montre une hauteur globale plus élevée sous l'influence de la hauteur élevée du ton montant que porte la syllabe précédente ; tandis que, si deux syllabes adjacentes contiennent deux tons descendants, le deuxième ton descendant est plus bas sous l'influence de la hauteur peu élevée du premier ton descendant.

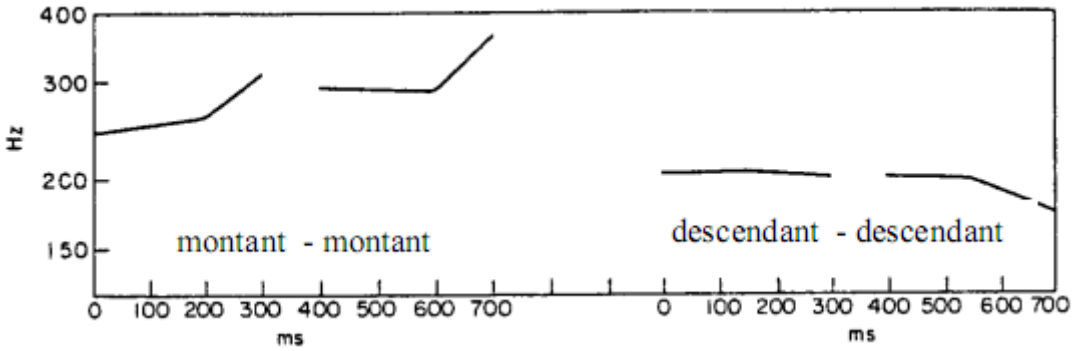


Figure 17: Variante plus élevée du ton montant dans la deuxième syllabe, dans la combinaison de tons montant – montant, et variante plus basse du ton descendant dans la deuxième syllabe, dans la combinaison de tons descendant – descendant [Han et al. 1974]

Par une étude soignée sur les caractéristiques dynamiques du ton vietnamien, Tran [2007] montre aussi que les tons vietnamiens varient considérablement selon leur environnement tonal immédiat. Selon lui, la coarticulation progressive est plus forte que la coarticulation régressive ; les tons sont modifiés sous l'influence des tons adjacents : après un ton montant (ton3 et ton5), le ton suivant commencera plus haut que sa valeur normale et après un ton descendant (ton2, ton 4 & ton6), il commencera plus bas.

4.2.1.2 Niveau de phrase : la modalité

Comme nous l'avons mentionné dans la section 2.3.1, la fonction de modalisation de la prosodie est liée aux modalités de phrase telles que la phrase assertive, la phrase interrogative ou la phrase impérative.

Pour observer les différences de prosodiques entre les différentes modalités de phrase, Nguyen T.T.H. [Nguyen et al. 1999 ; Nguyen 2004] utilise un corpus de paires d'énoncés. Pour chaque paire, deux énoncés sont basés sur une seule phrase mais leur structure morphosyntaxique respective est différente. Par conséquent, les phrases ont des sens et des modalités différentes. Par exemple:

- « Lan thích ăn **cơm không** - Lan n'aime manger **que du riz** » (phrase assertive)
- « Lan thích ăn cơm **không** ? – **Est-ce que** Lan aime manger du riz ? » (phrase interrogative)
- « Mai từ chối **đi** – Mai refuse d'(y) **aller** » (phrase assertive)
- « Mai từ chối **đi** – Mai, refuse ! (phrase impérative)

Par l'observation des patrons intonatifs, Nguyen T.T.H. montre que, comme dans la plupart des langues, la phrase déclarative est caractérisée par une déclinaison de sa fréquence fondamentale F0 globale, tandis que les phrases interrogatives mais aussi impératives ont un contour montant [Hirst et al. 1998]. Les phrases déclaratives sont prononcées avec un registre bas alors que les phrases interrogatives et les phrases imperatives le sont avec un registre haut [Do et al. 1998 ; Nguyen 2004].

Au niveau de la durée, les énoncés interrogatifs ont un débit plus rapide que les énoncés assertifs et imperatifs. La différence de durée entre ces deux derniers 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 [Nguyen 2004].

L'étude de [Vu et al. 2006a] sur la production et perception des phrases interrogatives et affirmatives en langue vietnamienne a aussi montré que les différences entre questions et affirmations sont essentiellement une différence de pente de F0 (croissante ou décroissante) en fin de la phrase (deuxième moitié de la dernière syllabe), à laquelle s'ajoutent une modification du débit. Un exemple

de différence entre des contours de F0 de phrase interrogative et phrase affirmative est présenté dans la Figure 18. Ces auteurs ont aussi montré l'effet du ton sur la perception de la modalité de phrase : des auditeurs peuvent mal classer des affirmations si les phrases produites présentent une syllabe finale avec ton montant. Des questions peuvent être mal classifiées si leur syllabe finale porte un ton descendant.

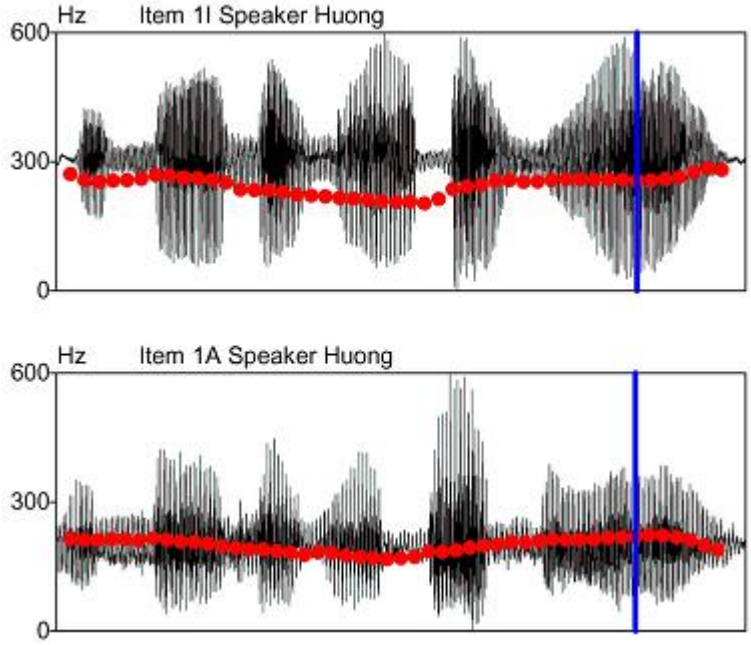


Figure 18: Exemple des contours de F0 de deux phrases à nombre de syllabes et tons identique : phrase interrogative (dessus), phrase affirmative (dessous) [Vu et al. 2006a]

4.2.2. Fonction expressive de la prosodie vietnamienne

Jusqu'à ce jour, très peu d'études ont analysé les caractéristiques prosodiques de la parole expressive en vietnamien. En fait, la thèse de Le Thi Xuyen [Le 1989] reste la seule étude sur l'expression des émotions et des attitudes pour la langue vietnamienne.

Le travail de Le Thi Xuyen présente une étude contrastive de l'intonation expressive en français et en vietnamien. Elle a utilisé des paires de phrases en français / vietnamien. Ces phrases présentent des longueurs de 1 à 6 syllabes et ont la même signification dans les deux langues Table 5.

Table 5: Les phrases utilisées dans l'étude de Le T. X. [Le 1989]

Français	**Vietnamien**
Il pleut	Mưa Trời mưa
Elle est jolie	Cô ta xinh
Il est déjà tard	Khuya rồi
Nam est rentré vers une heure et demie	Nam về lúc khoảng một rưỡi

Ces phrases sont ensuite enregistrées avec douze catégories attitudinales différentes : *neutre, déception, ennui, regret, joie, contentement, ironie, surprise, doute, colère, confirmation et conseil*.

Pour les deux langues, les tests de perception sur ces phrases montrent que les attitudes sont reconnues avec un score supérieur à celui du hasard. La colère, la tristesse, la surprise et la phrase neutre sont les mieux reconnues (avec un taux d'identification de 75 %, 52,5 %, 67,5 %), tandis que la « confirmation » et le « conseil » n'ont pas été bien perçues.

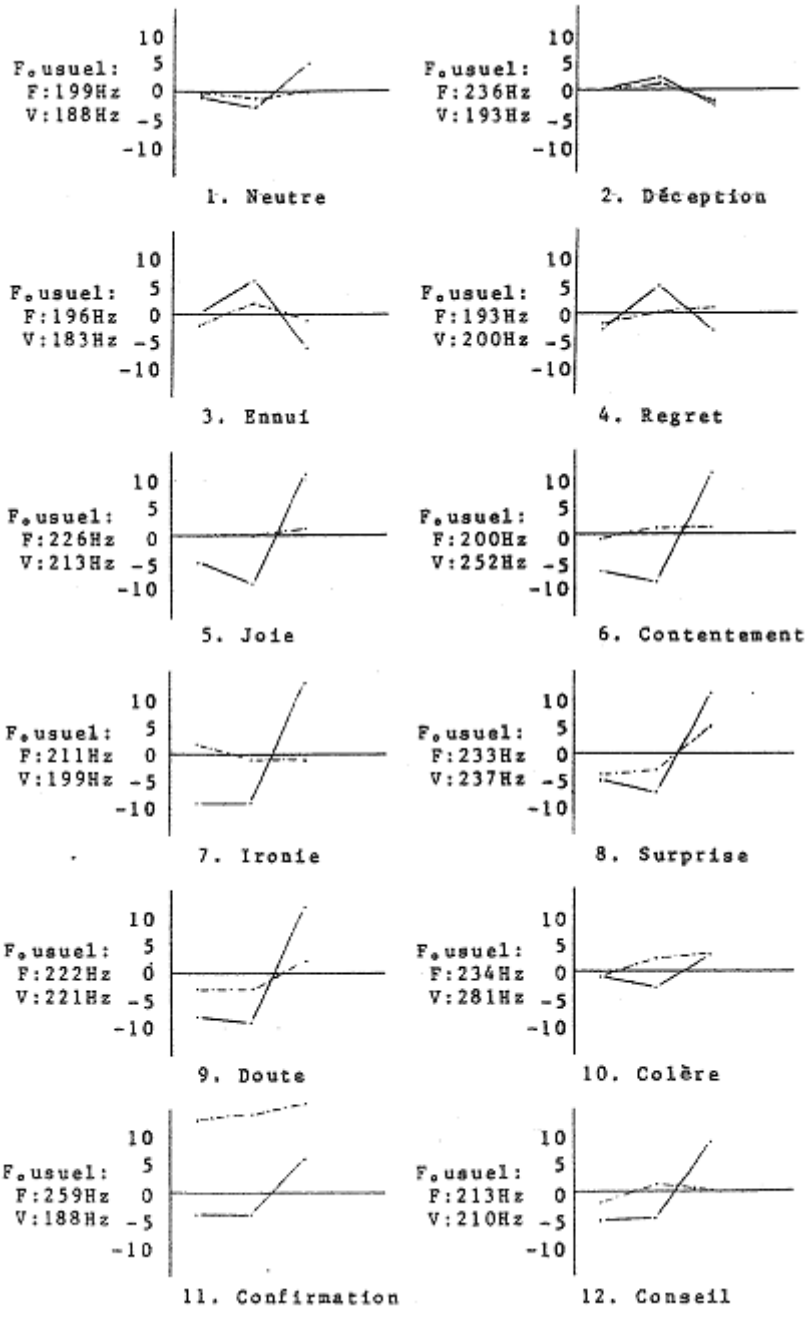


Figure 19: Un exemple de contours prosodiques des 12 attitudes de la phrase « Elle est belle » en français (-----) et « Cô ta xinh » en vietnamien (…..) [Le 1989]

Le Thi Xuyen montre aussi les caractéristiques prosodiques des attitudes. La Figure 19 présente un exemple du contour prosodique des 12 attitudes en français et vietnamien. Les analyses de fréquence fondamentale, durée et intensité, montrent des différences prosodiques entre les attitudes. Par exemple :

- la phrase neutre est caractérisée par un registre moyen et un débit moyen ;
- la colère se traduit par une élévation du registre, une accélération du débit et une forte intensité ;
- la tristesse présente des traits inverses (abaissement du registre, ralentissement du débit et faible intensité) ;
- quant à la surprise, la ligne mélodique débute dans un registre moyen puis arrive à la fin sur un registre élevé, mais l'intensité reste moins forte [Le 1989].

4.3. Génération de la parole en vietnamien

Le traitement de la parole vietnamienne a été étudié depuis plus d'une décade. Dans cette section nous présentons des travaux sur la modélisation de la prosodie et sur la synthèse de la parole en vietnamien.

4.3.1. Modélisation de l'intonation vietnamienne

Pour une langue tonale comme le vietnamien, le mandarin ou thaï, le contour prosodique de phrase se compose toujours de tons et de l'intonation globale (correspondant aux structures de plus haut niveau) [Doan 1997]. C'est pourquoi la modalisation des contours intonatifs pour les langues tonales semble plus complexe que pour des langues non tonales puisque la fréquence fondamentale est impliquée au niveau global de l'énoncé et d'une manière complexe (les configurations contrastives tonales) au niveau local du ton.

4.3.1.1 Application du modèle Fujisaki dans la langue vietnamienne

Parmi les modèles de génération d'intonation qui sont présentés dans la section 3.2, le modèle de Fujisaki a été appliqué à plusieurs langues, y compris des langues tonales telles que le mandarin, le cantonais et le thaï.

Pour modaliser la prosodie de la langue tonale, la commande accentuelle du modèle de Fujisaki devient la commande de génération des tons. A la différence du modèle original, pour modéliser le ton dans la langue tonale, le modèle de Fujisaki a besoin de deux types de commandes : commande positive et commande négative. La Figure 20 présente un exemple d'utilisation des commandes de tons pour modéliser le contour des 4 tons en mandarin.

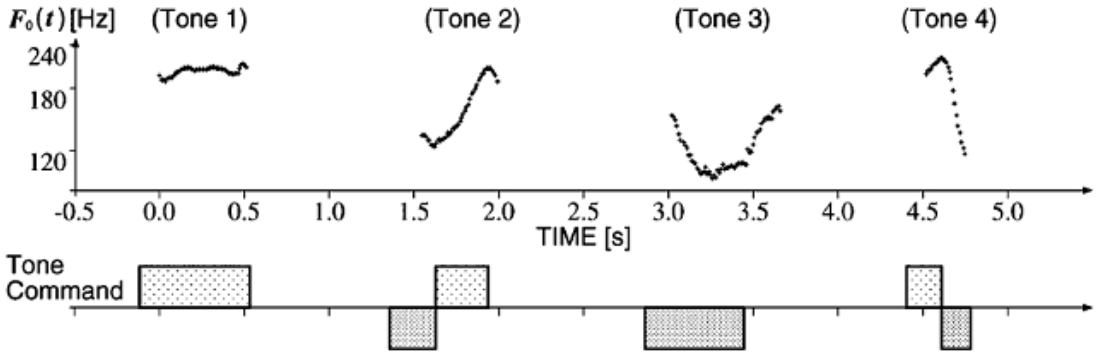


Figure 20: Commandes de tons pour modaliser le contour des 4 tons en mandarin [Fujisaki et al. 2005]

Ces contours de tons sont superposés au contour de la phrase en utilisant la superposition pour produire le contour final comme le montre la Figure 21.

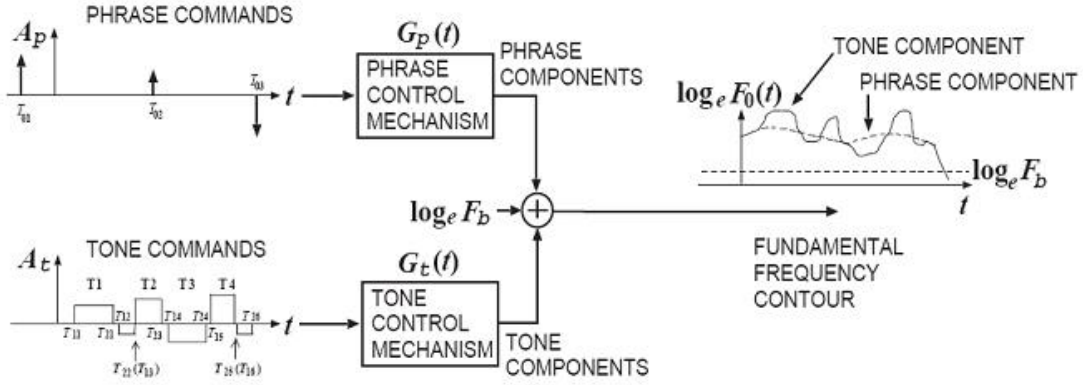


Figure 21 : Modèle de Fujisaki pour la génération du contour de F0 d'une langue tonale [Fujisaki et al. 2005]

En fait, le modèle de Fujisaki est un modèle qui a déjà été utilisé pour la génération de la prosodie de la langue vietnamienne. Dans le système de synthèse de la parole vietnamienne VnVoice, Mixdorff [2003a] et Nguyen [2004] ont appliqué le modèle de Fujisaki pour générer les contours de F0 des six tons vietnamiens. Dans ce système, une collection des règles de commande du ton du modèle de Fujisaki est obtenue par des outils d'analyse automatique sur certaines phrases en vietnamien [Mixdorff 2000]. Ces règles décrivent comment modéliser les tons vietnamiens par des commandes tonales du modèle de Fujisaki. Par exemple, selon [Nguyen et al. 2004], les tons 1 et 5 peuvent être modélisés en utilisant des commandes tonales positives ; les tons 2 et 4 sont modélisés en utilisant des commandes tonales négatives.

Cependant, comme nous l'avons présenté dans la section 4.2, le système de ton vietnamien est différent des autres langues tonales (telle que le mandarin ou le thaï) à cause de l'influence du phénomène de glottalisation au milieu du ton 3 et à la fin du ton 6. La fréquence fondamentale change très fortement dans la région de glottalisation, et ceci ne peut pas être modélisé par les commandes tonales du modèle de Fujisaki [Nguyen et al. 2004].

4.3.1.2 Modèles de génération de la prosodie vietnamienne de Tran D. D. [Tran 2007]

Dans sa thèse, Tran [2007] a proposé une nouvelle méthode permettant de produire des contours de F0 du vietnamien dans la parole continue. Cette méthode contient deux composants principaux : le modèle dynamique du ton et le modèle du registre relatif.

Le modèle dynamique du ton

Ce modèle permet de modéliser le contour de ton sur une syllabe dans la parole continue. Ce modèle est basé sur l'observation que dans la parole continue, les contours des tons sont variables et dépendent de la durée de la syllabe. Par exemple, dans la parole continue, la forme du contour de F0 du ton 3 a trois variantes dépendantes de la durée de la syllabe (Figure 22).

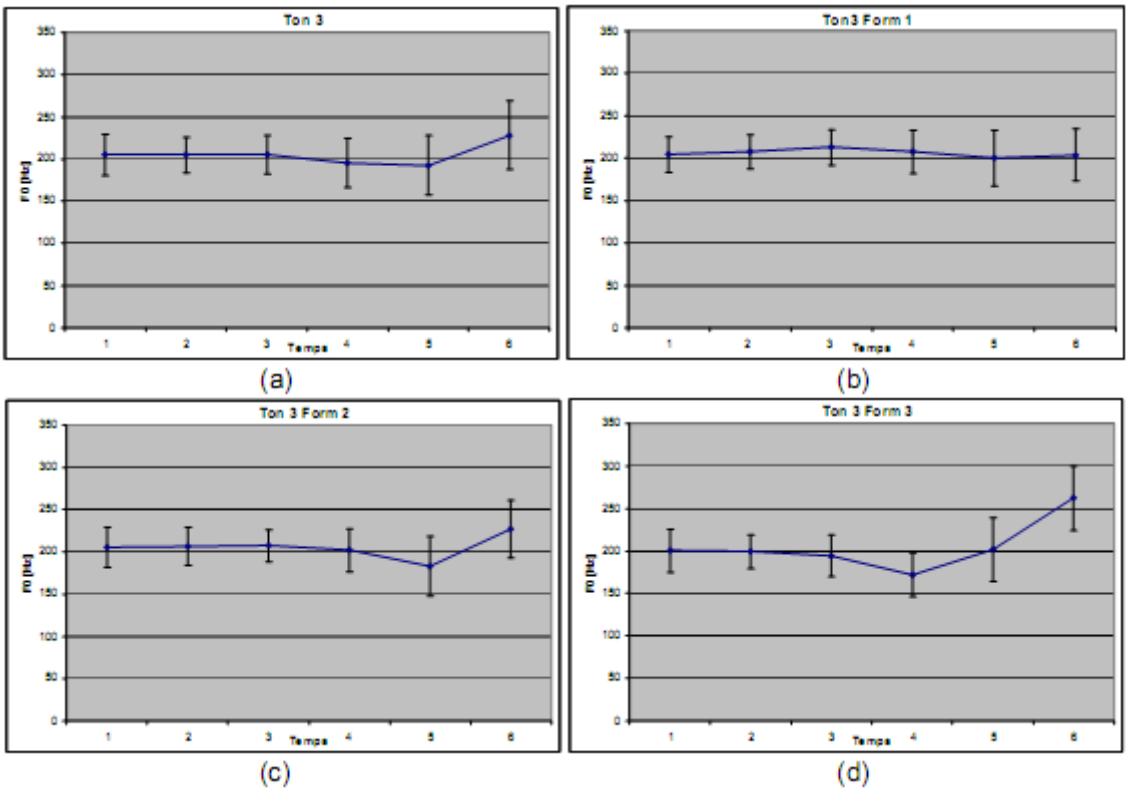


Figure 22: (a) Contours de F0 moyen du ton 3 et ses trois variantes, avec la durée de la syllabe (b) <150ms, (c) 150 – 220ms, (d) >220ms

Pour obtenir les variantes contextuelles des six tons vietnamiens, à partir d'un corpus, le contour de F0 de chaque syllabe est représenté par six points. La valeur moyenne de F0 de chaque ton est calculée à partir des syllabes avec le ton correspondant dans le corpus. Les contours de F0 des variantes des six tons pour les trois durées syllabiques différentes (<150ms, 150-220ms, >220ms) sont également normalisés par six points comme dans la Figure 23.

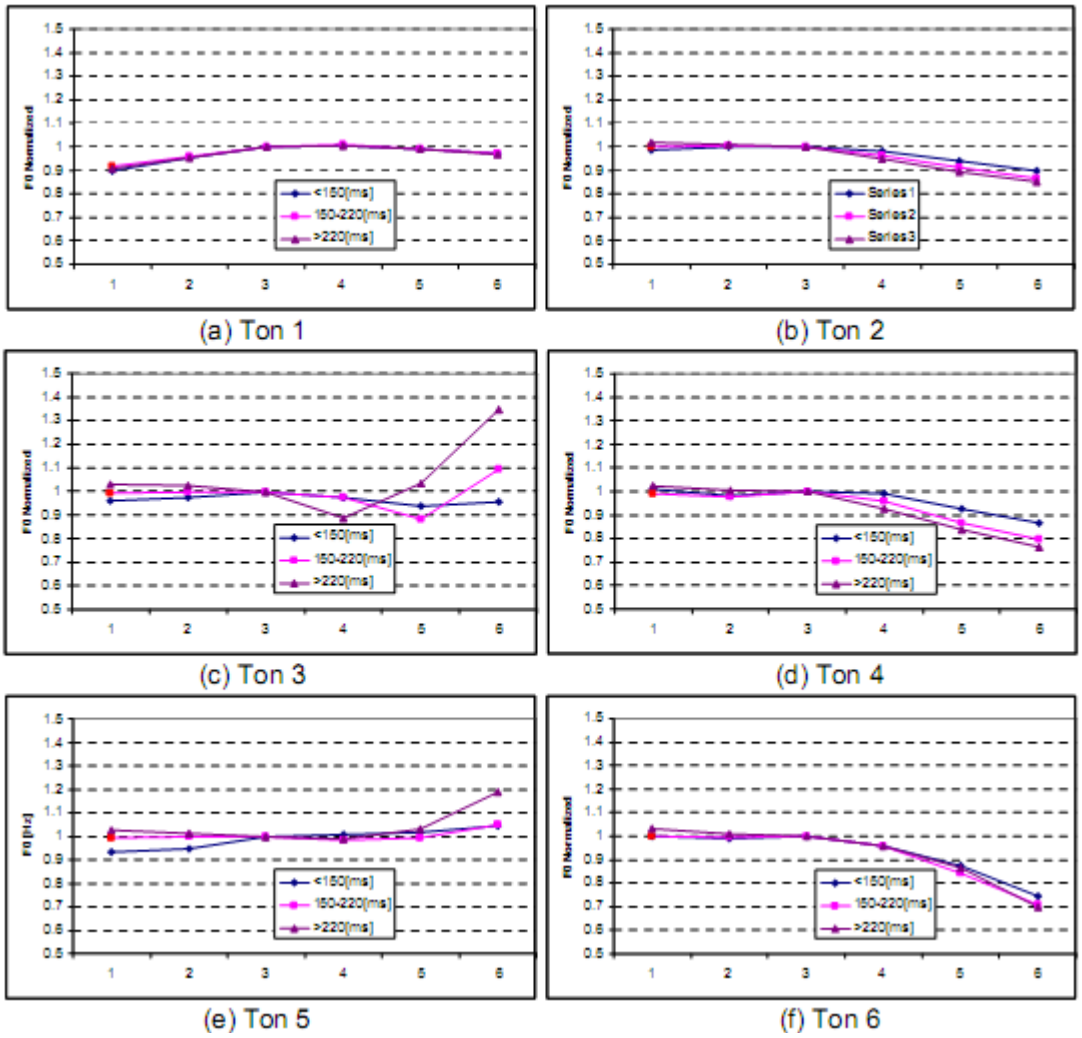


Figure 23: Les contours de F0 normalisés de six tons avec les durées de la syllabe différentes

Selon Tran [2007], les contours de F0 des tons 1, 2, et 6 sont plus stables que ceux du ton 3. Les tons 1, 2, et 6 sont stables par rapport à la durée de la syllabe et un seul gabarit représentatif peut être utilisé pour ces tons.

Le registre relatif du ton

Comme le mentionne la section 4.2, dans la parole continue en vietnamien, on observe des coarticulations tonales : les caractéristiques prosodiques d'un ton sont influencées par le ton précédent. En étudiant les contours prosodiques des tons dans la parole continue, Tran [Tran et al. 2008] trouve que cette influence dépend considérablement des tons utilisés. Il a aussi montré que la partie initiale du contour de F0 d'une syllabe est dépendante du point de terminaison et de la direction du contour de F0 de la syllabe précédente [Tran 2007 ; Tran et al. 2008].

À la différence des autres langues tonales telles que le mandarin ou le thaï, en vietnamien, les tons lexicaux devraient être considérés comme des tons de contour [Do et al. 1998], et donc la classification de ces tons en deux groupes de

registre (haut et bas) comme cela est habituellement utilisé dans certains modèles prosodiques n'est pas appropriée aux tons dynamiques de la parole continue. De plus, à cause du phénomène de coarticulation tonale, le rapport de registre entre deux tons doit être calculé dans des contextes spécifiques.

En se basant sur ces remarques, Tran [2007] a proposé d'utiliser les registres relatifs de deux tons adjacents pour présenter les valeurs moyennes des tons dans la parole continue. Selon lui, le rapport des registres des tons courants (R_i) et précédents (R_j) est calculé comme suit :

$$R_{i,j} = \frac{R_i}{R_j}$$

Pour chaque paire des tons en vietnamien, le rapport ($R_{i,j}$) est obtenu en calculant la valeur moyenne de toutes les occurrences de (i, j) dans un corpus. La Table 6 présente les rapports de toutes les combinaisons de deux tons en vietnamien.

Table 6: Les rapports de registres entre deux tons adjacents en vietnamien

Ton courant / Ton précédents	1	2	3	4	5a	6a	5b	6b
1	0.99	0.84	0.91	0.82	0.90	0.88	1.24	0.84
2	1.13	0.92	0.96	0.91	0.99	0.93	1.38	0.92
3	1.21	1.05	1.04	1.07	1.06	1.11	1.48	1.09
4	1.13	0.92	0.94	0.90	0.98	0.92	1.39	0.93
5a	1.19	1.08	1.12	1.05	1.07	1.18	1.50	1.18
6a	1.11	0.92	0.90	0.90	0.95	0.92	1.36	0.93
5b	0.84	0.73	0.77	0.68	0.78	0.80	1.10	0.70
6b	1.14	0.91	0.96	0.94	0.98	0.95	1.45	0.94

Génération du contour de F0

En utilisant le modèle dynamique et le registre relatif du ton ci-dessus, le contour de F0 d'une phrase, selon [Tran 2007], est obtenu par les 3 étapes suivantes :

- le registre de toutes les syllabes dans le syntagme est calculé en appliquant le rapport de registre relatif proposé dans la Table 6 ;
- pour chaque syllabe, le contour dynamique du ton est placé sur le registre de la syllabe ;
- pour lisser la discontinuité du contour de F0 entre deux syllabes adjacentes, une méthode d'interpolation linéaire est employée pour effectuer la transition à partir du point final du contour de F0 de la première syllabe jusqu'au premier tiers du contour de la deuxième syllabe (la position la plus stable du début).

Ce modèle est appliqué dans le système de synthèse de la parole en vietnamien. Une évaluation montre que ces méthodes peuvent fournir des contours de F0 assez proches des contours réels.

4.3.2. Les systèmes de synthèse de la parole en vietnamien

La synthèse de la parole a été étudiée depuis longtemps et les systèmes de synthèse de la parole ont été développés et largement appliqués dans plusieurs langues. Cependant, dans le cas du vietnamien, la synthèse de la parole n'a été étudiée qu'au début des années 2000. Récemment, quelques systèmes de synthèse de la parole ont été présentés [Le 2003 ; Do et al. 2004 ; Nguyen et al. 2004 ; Tran 2007 ; Vu et al. 2009b ; He et al. 2011]. Mais en fait, tous ces systèmes sont en cours de développement et aucun système n'est déployé dans des applications réelles.

Parmi les systèmes de synthèse du vietnamien, les systèmes de [Lê 2003] et [Do et al. 2004] utilisent la technique de synthèse par formants. Ils se basent sur des règles pour produire une parole vietnamienne compréhensible, mais leur qualité n'est pas très naturelle.

Les systèmes de [Nguyen et al. 2004] et [Tran 2007] utilisent, quant à eux, la méthode de synthèse par concaténation basée sur des diphones ou des demi-syllabes. La qualité de la parole synthétique de ces systèmes est bien meilleure que dans les systèmes de [Lê 2003] et [Do et al. 2004]. Le système de [Nguyen et al. 2004] utilise le modèle de Fujisaki pour générer le contour prosodique. Cependant, ce modèle ne permet pas de produire tous les tons vietnamiens [Tran 2007]. Au contraire Tran [2007] a proposé une méthode particulière pour la génération des contours de la parole continue (cf. 4.3.1.2). Cette méthode fournit des contours de F0 assez proches des contours réels.

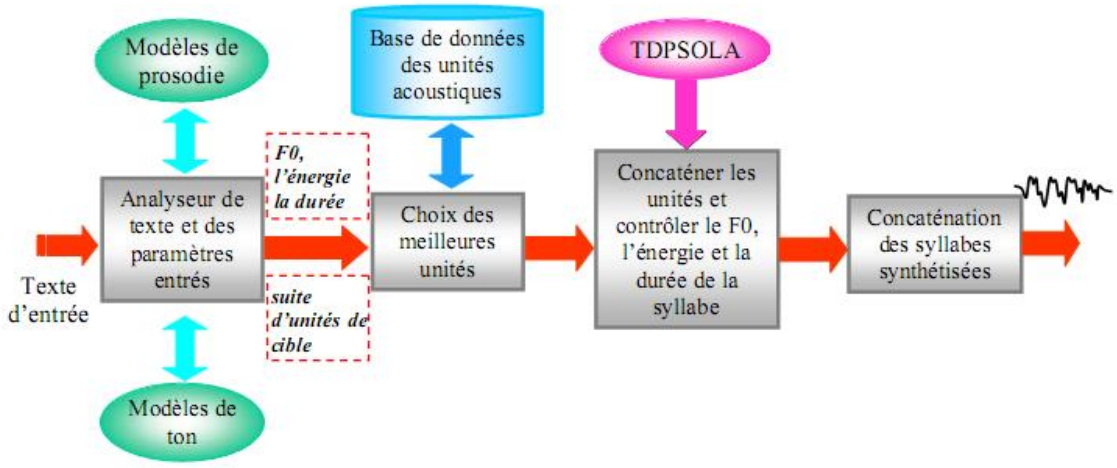


Figure 24: La structure du système de synthèse de la parole à partir du texte de [Tran 2007]

La Figure 24 présente la structure du système de synthèse de la parole à partir du texte de [Tran 2007]. Le module d'analyse du texte reçoit le texte d'entrée (des syllabes, des phrases structurées …), et il analyse ce texte en une suite de syllabes. Ces syllabes sont étiquetées par des informations contextuelles et des informations structurales. Se basant sur ces informations, le modèle de ton et le modèle de prosodie analysent les paramètres d'entrée et fournissent une suite de syllabes auxquelles sont associés des paramètres prosodiques (tels que, la F0, la

durée et l'énergie) et l'information des unités acoustiques qui sont utilisées pour construire la syllabe. A partir de la suite de syllabes, le deuxième module du système va chercher dans la base de données l'ensemble des meilleures unités candidates en se basant sur deux coûts, coût de cible et coût de concaténation. Ensuite, les syllabes sont construites en concaténant les unités choisies ; leur contour de F0 et leur durée sont contrôlés en appliquant l'algorithme TD-PSOLA [Moulines et al. 1995]. Enfin, ces syllabes synthétisées sont concaténées pour avoir la phrase synthétique désirée. Les évaluations perceptives et les analyses sur la comparaison des paramètres acoustiques entre la parole synthétique et naturelle montrent une bonne performance de ce système dans la synthèse de la parole continue en vietnamien.

Ces dernières années, certains travaux de recherche essayent d'utiliser l'approche par apprentissage automatique sur un grand corpus pour la génération de parole en vietnamien. Ces systèmes de synthèse peuvent utiliser la méthode de sélection dynamique d'unités non uniformes [Vu et al. 2009a ; Do et al. 2011] ou utiliser les HMMs [Vu et al. 2009b ; He et al. 2011]. Cependant, à cause des limitations des bases de données et de la complexité de la langue vietnamienne, la synthèse avec cette approche reste encore un sujet de recherche.

Partie 2 : Etudier les attitudes vietnamiennes pour la synthèse de la parole expressive

Chapitre 5 : Construction d'un corpus audio-visuel d'attitude vietnamienne

5.1. Introduction

Dans la première partie, nous avons présenté et analysé l'état de l'art sur la parole expressive en général et les études sur la parole expressive en vietnamien. Nous avons aussi expliqué que le type d'expressivité choisi pour notre travail est l'attitude. Plus précisément, notre travail consiste à étudier les caractéristiques des attitudes vietnamiennes, afin de proposer une première modalisation de l'attitude en vietnamien et d'intégrer ce modèle dans le système de synthèse de la parole expressive du vietnamien de l'Institut MICA.

Il y a très peu de travaux de recherche sur le traitement des émotions et attitudes en vietnamien : cette langue peut donc être considérée pour ce domaine particulier comme une langue peu dotée. Plus spécifiquement, dans le domaine de la parole expressive, nous ne trouvons que l'étude de Le [Le 1989], qui a été réalisée il y a plus de 20 ans. Un corpus pour étudier des attitudes en vietnamien est évidemment indisponible. La première tâche de notre travail de thèse porte donc sur la réalisation d'un corpus des attitudes en vietnamien.

Ce corpus est conçu d'abord pour étudier les caractéristiques des attitudes vietnamiennes, en termes de perception. Puis, des analyses acoustiques sur ce corpus nous permettent d'étudier les caractéristiques acoustiques des attitudes vietnamiennes. Ces caractéristiques vont être alors utilisées pour modéliser la prosodie d'attitude et pour appliquer cette modélisation dans la synthèse de la parole expressive en vietnamien.

De nombreuses études passées, dont certaines récentes, [Ekman et al. 1979 ; Scherer 2003 ; Rilliard et al. 2008 ; Shochi et al. 2008 ; Bänziger et al. 2009 ; Rilliard et al. 2009] ont montré que le décodage des énoncés expressifs (l'émotion et l'attitude) devait être considéré comme un processus multimodal. Autrement dit, dans la communication face à face, l'émotion et l'attitude sont représentées de plusieurs manières : la face, le geste, et la parole. Les travaux de recherche de [Rilliard et al. 2009 ; Nadeu et al. 2011] montrent aussi que les informations audio et les informations visuelles ont des rôles différents pour chaque attitude. Pour étudier des attitudes vietnamiennes sous tous ses aspects, nous avons décidé que notre corpus d'attitudes vietnamiennes serait donc un corpus audiovisuel.

Nous présentons ensuite la conception et la validation de ce corpus audio-visuel d'attitudes vietnamiennes.

5.2. Choix des attitudes du vietnamien

Comme nous l'avons mentionné dans la section 1.4.2, il existe beaucoup d'étiquettes pour décrire les attitudes. Par exemple, O'Connor et Arnold [1978] ont utilisé près de 300 étiquettes pour nommer les attitudes. Dans notre travail, nous nous concentrerons sur les attitudes qui sont régulièrement utilisées dans la communication parlée face à face. C'est pourquoi la première tâche dans la construction notre corpus est de choisir des attitudes.

Comme nous l'avons mentionné, les attitudes sont des codes construits et acquis dans la dynamique du langage et de ses réalités socioculturelles, et sont dépendantes de la langue et de la culture. Puisque les attitudes peuvent varier entre les langues et les cultures, une attitude présente dans une langue peut ne pas exister dans une autre langue. Ainsi, l'inventaire des attitudes sera différent pour des langues différentes. Nous parlons par la suite de la sélection des attitudes en français, anglais et japonais.

En 1996, Grépillat et Aubergé ont étudié la perception de six attitudes en français qui sont : *la déclaration, l'évidence, l'exclamation de surprise, le doute/incrédulité, ironie de soupçon* et *la question simple* [Grépillat 1996 ; Aubergé et al. 1997]. Ces six attitudes ont été utilisées par Morlec [1997] pour les appliquer à la génération de la parole expressive. En 2008, Rilliard [Rilliard et al. 2008] utilise ces six attitudes pour étudier la perception audio-visuelle des attitudes françaises. Puis, dans une étude sur la perception interculturelle, [Shochi et al. 2009] étendent leur travaux sur ces six attitudes accompagnées de six autres. Au total, douze attitudes ont été sélectionnées pour être étudiées en français : *la déclaration, la question simple, l'exclamation de surprise, l'évidence, le doute/incrédulité, l'autorité, l'irritation, l'ironie sarcastique, le mépris, la politesse, la séduction et l'admiration.*

De son cote, pour étudier la production et la perception d'un ensemble d'expressions attitudinales de l'anglais, Diaféria [2002] sélectionne 11 attitudes représentatives de l'anglais. Ces attitudes sont sélectionnées à partir de la littérature sur les attitudes anglaises [Trask 1996 ; Crystal et al. 1997 ; Crystal 2003]. Ces 11 attitudes sont : declaration (*la déclaration),* interrogation *(la question simple),* surprise *(l'exclamation de surprise),* evidence *(l'évidence),* doubt-incredulity *(le doute/incrédulité),* command-authority *(l'autorité),* irritation *(l'irritation),* sarcastic irony (*l'ironie sarcastique),* scorn-aloofness *(le mépris),* politeness *(la politesse),* seduction *(la séduction).*

Quant à Shochi [Shochi 2008], dans la cadre de sa thèse, il propose 12 attitudes pour le japonais. Ces attitudes sont sélectionnées à partir de deux types de littératures : des données provenant de domaines variés tels que la linguistique, la sociolinguistique et la phonétique et des données provenant de la didactique du

japonais langue étrangère. Les 12 attitudes japonaises sélectionnées par Shochi sont : *la déclaration, la question simple, l'exclamation de surprise, l'évidence, le doute/incrédulité, l'autorité, l'irritation, la politesse simple, l'admiration, la sincérité-politesse, l'arrogance-impolitesse et le kyoshuku.*

Parmi ces 12 attitudes, Shochi montre que certaines sont spécifiques à la culture japonaise. Ce sont par exemple cinq expressions reliées à la gestion des protocoles de politesse et d'impolitesse en japonais : politesse simple, sincérité-politesse, arrogance-impolitesse, kyoshuku et déclaration. L'attitude de sincérité-politesse apparaît lorsqu'un locuteur considéré comme inférieur dans la société japonaise communique avec un interlocuteur considéré comme supérieur. L'attitude arrogance-impolitesse représente une expression impolie d'un sentiment de supériorité vis-à-vis de son interlocuteur. Tandis que l'attitude de kyoshuku est utilisée quand le locuteur est dans une situation où son statut social est inférieur à celui de son interlocuteur, et quand il a de plus un avis contraire ; le locuteur doit montrer sa souffrance honteuse et son embarras afin d'atténuer l'aspect arrogant et impoli qu'engendre sa désapprobation ou sa demande [Shochi et al. 2010] L'expression de déclaration constitue une expression neutre sur ce gradient de politesse. Shochi [2009] aussi montre que certaines attitudes utilisées en anglais et en français telles que l'ironie sarcastique, le mépris ou la séduction sont des attitudes fortement marquées culturellement : ce sont des attitudes difficiles à exprimer dans un contexte culturel japonais.

Comme nous l'avons déjà souligné précédemment pour le vietnamien, la thèse de Le [Le 1989] reste la seule étude sur l'expression des attitudes. Dans ce travail, elle utilise douze attitudes pour étudier et comparer l'intonation expressive en français et en vietnamien. Ces attitudes sont : *neutre, déception, ennui, regret, joie, contentement, ironie, surprise, doute, colère, confirmation et conseil.*

Cependant, dans sa thèse, Le n'a pas expliqué le comment et le pourquoi de ce choix. Les définitions de ces attitudes ne sont pas présentées non plus. En fait, Le a réalisé une étude contrastive pour les deux langues française et vietnamienne, les attitudes choisies dans son travail sont les attitudes communes aux deux langues. Il n'y a pas d'attitude spécifique ou particulière à la culture vietnamienne.

Pour sélectionner des attitudes en vietnamien, notre étude se base sur l'étude précédente de Le mais aussi sur des études dans les autres langues ci-dessus. En effet, certaines attitudes présentées dans le travail de Le peuvent correspondre aux attitudes sélectionnées en français, anglais et japonais, par exemple : *la colère* peut correspondre à *l'irritation* ; la *surprise* semble être le correspondant de *l'exclamation de surprise,* etc. Mais en nous fondant sur la distinction entre l'émotion et l'attitude que nous avons présentée dans la section 1.4, les étiquettes des attitudes présentées dans le travail de Le Thi Xuyen ne sont pas, de notre point de vue, assez claires pour identifier ces attitudes. Il peut y avoir, en effet, confusion avec les étiquettes des émotions (par exemple : joie, colère). C'est pourquoi, nous décidons d'utiliser les étiquettes des attitudes comme celles qui

sont présentées dans les travaux de Grépillat [1996], Aubergé et al. [1997] et Shochi [2008]. Notre travail concerne 16 attitudes décrites par les définitions présentées dans la Table 7.

Table 7: Les attitudes pour le vietnamien et leur définition

No	Attitude	Abréviation	Définitions
1	Déclaration	DEC	Le locuteur fait part d'une simple information, sans exprimer aucun point de vue
2	Question-simple	QUE	Le locuteur demande une information, sans exprimer de point de vue, et sans attendre autre chose qu'une simple réponse
3	Exclamation de surprise neutre	EXo	Le locuteur manifeste son étonnement concernant l'information qu'il donne à son interlocuteur, sans préciser si cette information surprenante le dérange ou lui plait
4	Exclamation de surprise positive	EXp	Le locuteur manifeste son étonnement concernant l'information qu'il donne à son interlocuteur : cette information est une très **bonne** nouvelle pour le locuteur
5	Exclamation de surprise négative	EXn	Le locuteur manifeste son étonnement concernant l'information qu'il donne à son interlocuteur : cette information est une **mauvaise** nouvelle pour le locuteur
6	Evidence	EVI	Le locuteur parle de quelque chose dont il est certain et manifeste cette certitude
7	Doute-incrédulité	DOU	Le locuteur veut exprimer son incertitude, ou son manque de conviction, concernant une information que vient de lui donner son interlocuteur : il répète cette information tout en manifestant son doute
8	Autorité	AUT	Le locuteur veut imposer son avis à son interlocuteur, ou au moins l'influencer fortement
9	Irritation	IRR	Le locuteur est fortement mécontent de ce qui vient d'être dit, cela le dérange et il le manifeste.
10	Ironie sarcastique	SAR	Son interlocuteur vient d'affirmer une information avec laquelle le locuteur n'est pas d'accord, il le manifeste mais par le biais de l'ironie, ce qui pourrait ressembler à prononcer par ex « Oui, c'est exactement ça… », mais en exprimant clairement **qu'il pense le contraire…**
11	Mépris	MEP	Le locuteur manifeste de l'arrogance, du mépris, manifestant qu'il considère ce qui vient d'être dit n'est pas digne d'intérêt, voire pire....
12	Politesse	POL	Le locuteur souhaite exprimer sa courtoisie et de la politesse vis-à-vis de son interlocuteur
13	Admiration	ADM	Le locuteur est admiratif et le manifeste.
14	Maternel	MAT	Le locuteur bienveillant s'adresse à un petit enfant, le met en confiance affectueusement.
15	Séduction	SED	Le locuteur veut plaire à son interlocuteur, gagner son estime et sa confiance, peut-être même la séduire amoureusement
16	Familier-Intime	FAM	Le locuteur partage une intimité avec son interlocuteur et s'adresse à lui avec naturel

La Table 8 présente la liste des attitudes sélectionnées en français, en anglais, en japonais et pour notre corpus d'attitudes vietnamiennes. Parmi les attitudes sélectionnées pour le vietnamien, les attitudes maternel et familier-intime sont

deux nouvelles attitudes, qui n'ont pas été étudiées dans les autres langues. Ces deux attitudes sont utilisées par le locuteur quand il parle avec des interlocuteurs très spécifiques : des petits enfants, des familiers ou des intimes. L'utilisation et les caractéristiques acoustiques de ces deux attitudes dans la communication parlée sont étudiés dans plusieurs langues [Kenyon 1948 ; Fisher et al. 1996 ; Katz et al. 1996 ; Thiessen et al. 2005]. Dans notre travail, nous avons choisi ces deux attitudes pour examiner si elles peuvent être véhiculées par la parole expressive en vietnamien.

Table 8: La sélection des attitudes en francais, anglais, japonais et vietnamien

Français *Grépillat T. [1996]* *Aubergé [1997]* *Shochi [2008]*	**Anglais** *Diaféria [2002]*	**Japonais** *Shochi [2008]*	**Vietnamien**
1. Déclaration	1. Declaration	1. Déclaration	1. Déclaration
2. Question simple	2. Interrogation	2. Question-simple	2. Question-simple
3. Exclamation de surprise	3. Surprise	3. Exclamation de surprise	3. Exclamation de surprise neutre 4. Exclamation de surprise positif 5. Exclamation de surprise négatif
4. Evidence	4. Evidence	4. Evidence	6. Evidence
5. Doute/Incrédulité	5. Doubt / Incredulity	5. Doute-incrédulité	7. Doute-incrédulité
6. Autorité	6. Command / Authority	6. Autorité	8. Autorité
7. Irritation	7. Irritation	7. Irritation	9. Irritation
8. Ironie sarcastique	8. Sarcastic irony		10. Ironie sarcastique
9. Mépris	9. Scorn / Aloofness		11. Mépris
10. Politesse	10. Politeness	8. Politesse-simple	12. Politesse
11. Séduction	11. Seduction		15. Séduction
12. Admiration		9. Admiration	13. Admiration 14. Maternel 16. Familier-Intime
		10. Arrogance-impolitesse 11. Sincérité-politesse 12. Kyoshuku	

Pour être un peu plus précis, on peut séparer les attitudes ci-dessus entre les groupes suivants (voir [Fónagy et al. 1984 ; Wichmann 2000 ; Moraes et al. 2010] à propos de ces regroupements) :

- des attitudes proprement dites, ce qui inclut une modalité illocutoire : *déclaration, question, surprise, évidence, doute, admiration, mépris, ironie, irritation ;*
- des attitudes qui véhiculent des valeurs sociales hiérarchiques pendant l'interaction : *autorité, politesse* ;
- des attitudes qui définissent les rôles sociétaux pendant l'interaction *familier-intime, maternel, séduction*.

5.3. Composition du corpus

5.3.1. Méthodologie

Comme nous l'avons décrit dans le Chapitre 3, notre approche pour la génération de la prosodie expressive est basée sur le concept de la superposition des contours prosodiques de niveaux linguistiques différents. Ces contours peuvent être obtenus par l'analyse de corpus dont la structure doit révéler la nature des rendez-vous structurels entre les niveaux de description linguistique (et éventuellement para-linguistiques) et les fonctions de la prosodie [Aubergé 1991].

Dans le cadre de notre étude, nous nous intéresserons à deux fonctions de la prosodie : la fonction attitudinale (portée par le contour prosodique de la phrase), et la fonction tonale (portée par le contour prosodique de la syllabe).

À cause de la complexité de la syntaxe de la langue vietnamienne, la fonction syntaxique sera considérée et étudiée dans des travaux de recherche futurs. C'est pourquoi, dans ce corpus, nous essayons de nous affranchir de la fonction de segmentation syntaxique, en construisant des énoncés avec une segmentation syntaxique la plus simple possible. Plus précisément dans notre corpus, nous utilisons des phrases « balancées », distribuées dans leurs frontières et valeurs de segments syntaxiques, afin que les contours des attitudes ne soient pas perturbés, en moyenne ou dans le gabarit de l'énoncé (sans segmentation syntaxique).

5.3.2. La structure phono-linguistique du corpus

Selon la méthodologie ci-dessus, nous avons choisi 125 phrases isolées dont la longueur varie entre 1 et 8 syllabes. Le corpus contient donc :

- 9 phrases d'une seule syllabe, qui correspondent aux 8 représentations des tons vietnamiens ;
- 72 phrases de 2 syllabes, qui correspondent à toutes les combinaisons de tons. Ces compositions ont pour but de couvrir toutes les possibilités du phénomène de coarticulation tonale entre deux tons adjacents en vietnamien, ce que nous avons présenté dans la section 4.2.1 ;

- le reste du corpus concerne 45 phrases de 3 à 8 syllabes, variables dans leur structure syntaxique : mono-mot, groupe nominal isolé (GN), groupe verbal isolé (GV) ou structure simple « sujet-verbe-objet » (S-V-O), une structure courante en vietnamien.

Pour chaque longueur de phrase, notre corpus présente des phrases avec toutes les syllabes en ton 1 (ton plat). Nous les appelons les phrases « sans tons ». Ces phrases nous permettent d'étudier les contours prosodiques qui transportent la fonction attitudinale (au niveau de phrase) sans influence prosodique des tons (au niveau de syllabe). Dans les autres phrases, des tons différents (tons 2, 3, 4, 5, 6, 5b, 6b) sont positionnés sur les syllabes au début, au milieu et en fin de la phrase. Avec cette distribution, nous pouvons étudier l'influence des tons pour des positions différentes dans la phrase.

Puisque le corpus doit être produit selon 16 attitudes sélectionnées comme expliqué ci-dessus, les phrases dans ce corpus doivent être sans connotation particulière, ou avec un sens neutre. Autrement dit, nous avons exclu toute phrase portant un contenu sémantique significatif « attitudinalement » afin de ne pas interférer dans la production et la perception des attitudes véhiculées. Nous avons aussi utilisé comme énoncés des nombres, qui sont considérés comme des phrases sans segmentation syntaxique. Ayant un sens neutre, ils expriment facilement toutes les attitudes.

La Table 9 présente un exemple des phrases de notre corpus. Le corpus texte complet est décrit dans l'annexe 1.

Table 9: Exemple des phrases du corpus d'attitudes vietnamiennes

Nombre de syllabes	Séquence des tons	Phrases		Syntaxe
		Vietnamien	Traduction en français	
1	1	Ta	nous	mot
2	1_1	Anh ta	lui	mot
3	1_1_1	Hai mươi ba	vingt trois	Numéro
3	1_4_6	Em bảo chị	Tu me dis	SOV
4	1_1_2_1	Găng tay bằng da	Le gant en cuir	GN
5	1_1_1_1_1	Hai em đi theo anh	Vous me suivez tous les deux	SVO
6	5b_4_3_1_1_6	Tất cả đã đi theo chị	Ils t'ont suivi	SVO
7	1_1_1_1_1_2_1	Hai đôi găng tay da màu nâu	Deux paires de gants bruns en cuir	GN
8	1_1_2_1_1_1_1_1	Hai mươi ngàn hai trăm năm mươi ba	20253	Numéro

5.4. Protocole d'enregistrement du corpus

5.4.1. Locuteur

La recherche sur le corpus de parole expressive menée par Campbell [2000] montre que la parole expressive produite par un acteur risquerait d'être en décalage avec la réalité. Audibert et al [2008] montrent que la parole actée est identifiable perceptivement par les sujets naïfs. Notre but n'est pas de produire des attitudes vraiment naturelles, mais de produire des stéréotypes, plus exactement des prototypes, d'attitudes facilement identifiables par les sujets, sans pour autant tomber dans la caricature. Dans les travaux sur les attitudes d'autres langues dont nous nous inspirons, et comme le précise Aubergé [2002a], à la suite de Fónagy [1983], le locuteur idéal est l'enseignant de langue L2 qui a l'expérience de produire des énoncés réalistes pour l'usage courant. C'est pourquoi, nous n'utilisons pas un acteur professionnel comme locuteur. Pour enregistrer notre corpus, nous utilisons un locuteur masculin, originaire de Hanoi (prononciation standard du vietnamien), qui n'est pas un enseignant de langue L2 mais est par contre est un jeune enseignant universitaire et un doctorant qui travaille sur la parole expressive et qui connaît le but de cette recherche.

Pour s'assurer que le locuteur a une bonne connaissance de la stratégie expressive qui lui permettra de véhiculer naturellement les 16 attitudes sélectionnées, une phase d'entraînement est mene sous la direction d'une spécialiste de la prosodie et la parole expressive. Un pré-corpus avec dix phrases est enregistré. Pendant l'enregistrement, la spécialiste explique au locuteur le concept de chaque attitude et le guide de manière à ce qu'il exprime l'attitude le plus naturellement possible.

Le pré-corpus est ensuite validé par des auditeurs vietnamiens. Ils écoutent les phrases des 16 attitudes et donnent des commentaires sur la nature de l'attitude et la manière de véhiculer ces attitudes par le locuteur. Sur la base de ces commentaires, le locuteur peut s'entraîner pour améliorer sa capacité à exprimer des attitudes.

5.4.2. Enregistrement

La Figure 25 montre une photo de l'enregistrement du corpus audio-visuel des attitudes vietnamiennes. Le corpus audio-visuel a été enregistré dans une chambre sourde. De l'extérieur de la chambre sourde, les phrases à prononcer ainsi que l'attitude à reproduire sont affichées sur un écran faisant face au locuteur. Les 125 énoncés à oraliser sont affichés dans un ordre aléatoire pour chaque attitude. Le locuteur est debout devant la caméra, avec un micro AKG C1000S placé à quarante centimètres de sa bouche. Le microphone est connecté à un dispositif audio qui numérise le son (à 44,1kHz, 16 bits) et le transmet à un ordinateur situé à l'extérieur de la chambre sourde, sur lequel sont stockés les enregistrements sonores. La gestuelle faciale des locuteurs (la face et le haut du buste sont cadrés par la caméra) est enregistrée grâce une caméra numérique

(Sony DXC990). Les signaux vidéos sont numérisés grâce au codec « CinePack » avec une résolution vidéo de 784 x 576 pixels.

Comme nous l'avons mentionné, dans le cadre de ce travail nous n'étudions pas la qualité de voix des attitudes vietnamiennes. Cependant, le corpus est aussi enregistré afin de pouvoir servir à d'autres travaux de recherche dans le futur, en particulier sur la qualité de voix. Pendant l'enregistrement du corpus, un électroglottographe est utilisé pour mesurer directement les vibrations des cordes vocales du locuteur.

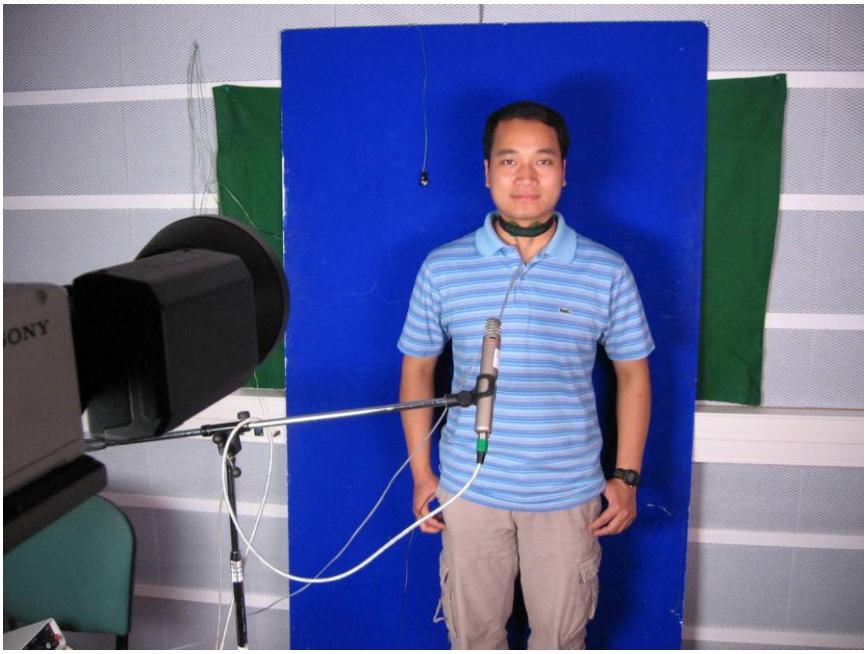


Figure 25: Enregistrement du corpus audio-visuel

Pour contrôler les performances du locuteur, une spécialiste de la parole expressive et un locuteur natif vietnamien ont observé le processus d'enregistrement de l'extérieur de la salle. Pendant que le locuteur produit son énoncé, la spécialiste et le locuteur natif peuvent entendre et voir le locuteur directement grâce à un système vidéo. S'ils trouvent qu'un énoncé n'est pas naturel, ils peuvent demander au locuteur de le reproduire.

L'ensemble du corpus a été enregistré 3 fois et représente environ 10 heures de signaux audiovisuels. Le corpus final est donc composé de 6000 énoncés au total (3 fois * 16 attitudes * 125 phrases).

Les signaux audio-visuels du corpus ont été découpés semi-automatiquement au niveau de la phrase grâce aux scripts des logiciels VirtualDub[10] et Praat[11]. Les phrases sont ensuite annotées au niveau phonétique avec le logiciel Praat.

[10] http://www.virtualdub.org/

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

5.5. Validation perceptive

Après la construction du corpus, un test de perception a été effectué pour évaluer la capacité du locuteur à transmettre les 16 attitudes représentées en audio seul. Les autres expérimentations sur la perception audio-visuelle des attitudes vietnamiennes sont présentées dans le chapitre suivant.

5.5.1. Protocole

Dans cette validation perceptive, afin de limiter l'influence du ton sur la perception des attitudes, nous avons choisi des énoncés sans variation tonale, qui contiennent toutes les syllabes basées sur le ton 1 (le ton plat). Parmi les énoncés du corpus, 48 énoncés avec une longueur différente ont été choisis pour le test de validation.

Vingt auditeurs vietnamiens (10 hommes et 10 femmes d'un âge moyen de 25 ans) ont participé à cette expérience. Les sujets sont tous de langue maternelle vietnamienne et parlent de même dialecte que le locuteur. Aucun sujet n'a fait état de trouble de l'audition.

Les tests de perception ont été effectués dans une pièce calme. Lors du test, le sujet est assis en face d'un écran, et écoute les stimuli selon un ordre de présentation aléatoire différent pour chaque sujet. L'interface de test (Figure 26) donne l'étiquette et la définition des 16 attitudes pour les auditeurs. Tous les sujets ont écouté chaque énoncé une seule fois. Après chaque énoncé, on leur a demandé d'indiquer l'attitude supposée parmi les 16 attitudes et d'indiquer une intensité allant de « à peine perceptible » (codé comme 1) à « très forte » (codé 100). Un score de 0 a été attribué aux 15 autres attitudes non sélectionnées.

Pour assurer que les sujets comprennent bien les concepts des 16 attitudes, avant le test, ils sont informés oralement des définitions des 16 attitudes. A la fin du test perceptif, nous avons aussi interrogé les sujets à propos des attitudes afin de savoir quelles attitudes étaient difficiles à distinguer ou s'ils avaient perçu d'autres attitudes qui n'étaient pas proposées.

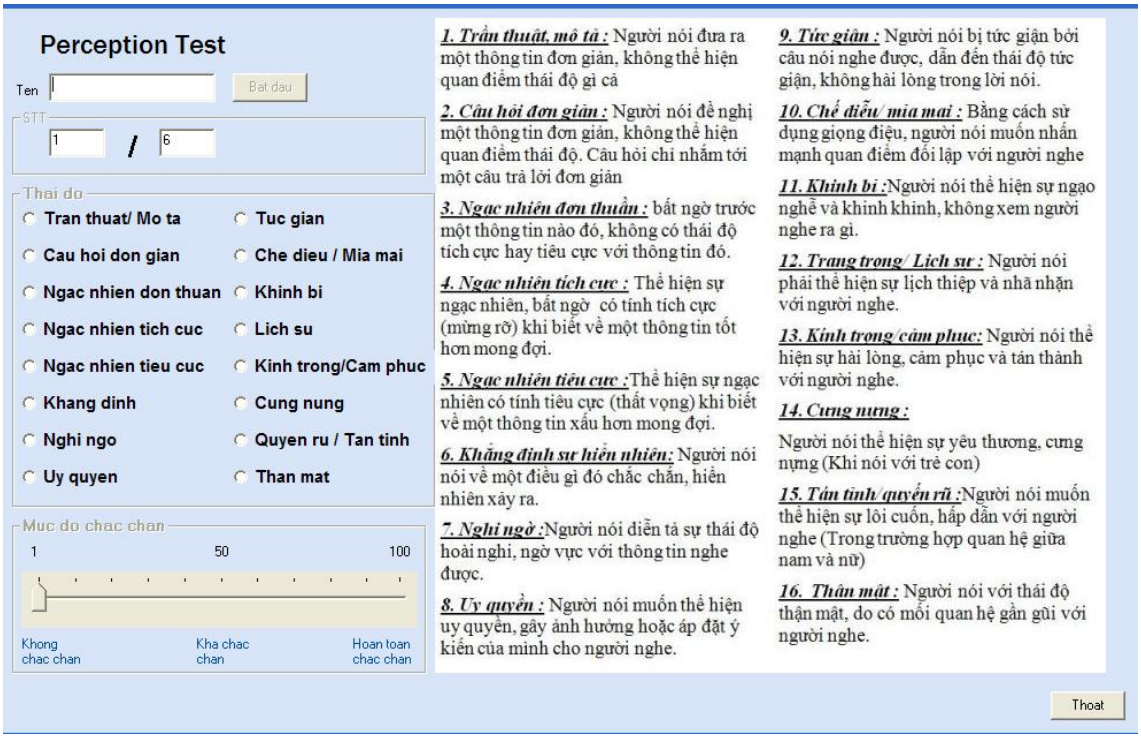


Figure 26: Interface du test de perception pour les sujets vietnamiens

5.5.2. Résultats

Les résultats de la validation perceptive pour les 16 attitudes vietnamiennes sont présentés Figure 27. Globalement, la plupart des attitudes sont reconnues au-dessus du niveau du hasard. Les expressions de déclaration, irritation, autorité, et ironie sarcastique reçoivent les meilleurs scores de reconnaissance (plus de 50%). Par contraste, les taux d'identification obtenus par la question simple et le doute-incrédulité restent faibles. Les attitudes « exclamation de surprise négatif » et « admiration », selon cette figure, sont difficiles à percevoir. Dans une première explication rapide, il est possible que ces deux attitudes ne puisse pas être véhiculées par les signaux de parole seuls, ou bien ces deux attitudes ne puisse pas être produites sans une cohérence morpho-lexicale, c'est-à-dire qu'il serait nécessaire d'utiliser des mots spécifiques pour exprimer ces deux attitudes ; ce point-là est aussi mentionné dans les travaux de Le [Le 1989]. Nous vérifierons cette hypothèse ci-dessous, en traitant de la perception audio-visuelle et de la perception interculturelle.

5.5.3. Conclusion

En comparant les résultats de la validation perceptive sur les attitudes en français, anglais et japonais (Figure 28), nous trouvons que la plupart des attitudes en vietnamien sont bien identifiées, d'une manière similaire aux français, anglais et japonais. Nous en concluons que la capacité du locuteur à transmettre les attitudes vietnamiennes dans notre corpus est assez bonne, en comparaison avec d'autres corpus en français, anglais et japonais [Shochi 2008]. Le résultat de cette

validation perceptive nous permet d'utiliser ce corpus pour la suite de nos travaux sur les attitudes vietnamiennes, travaux présentés dans les chapitres suivants.

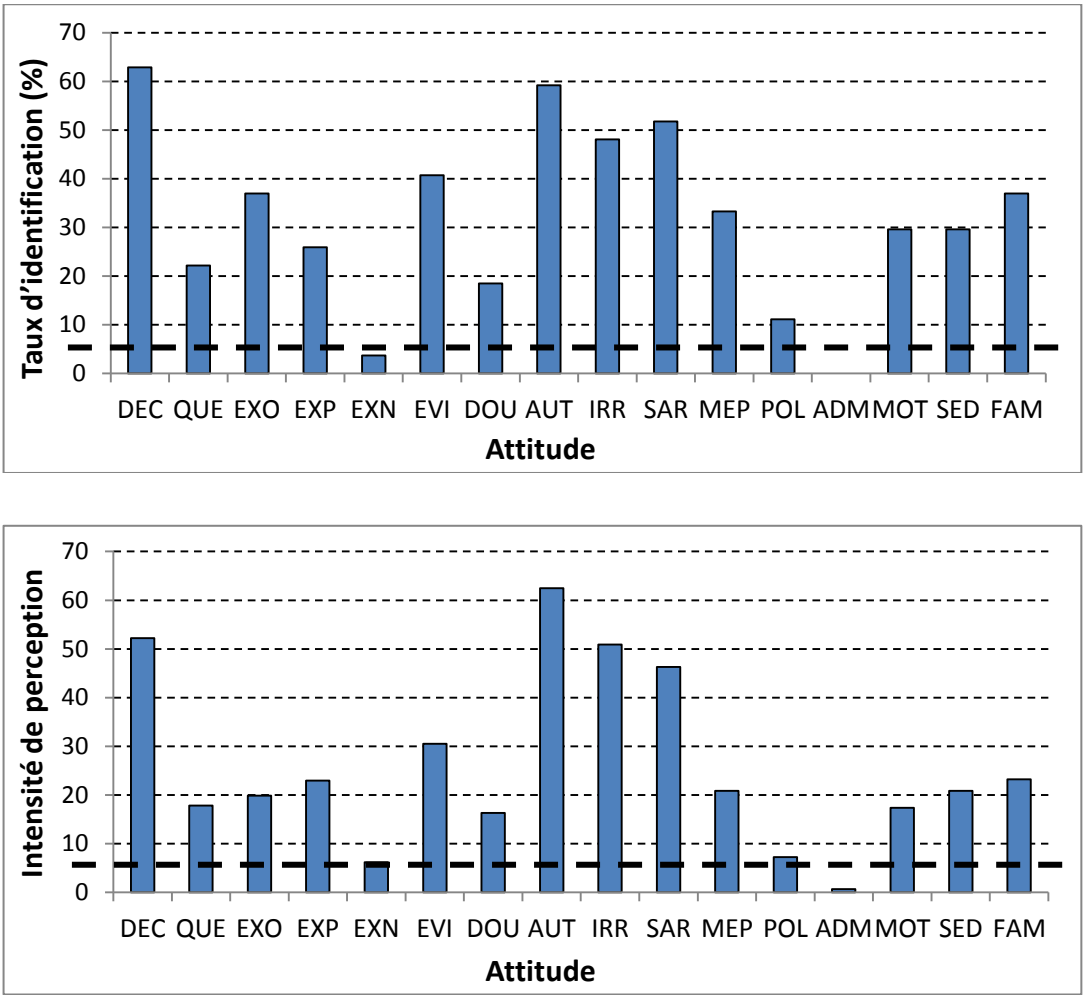
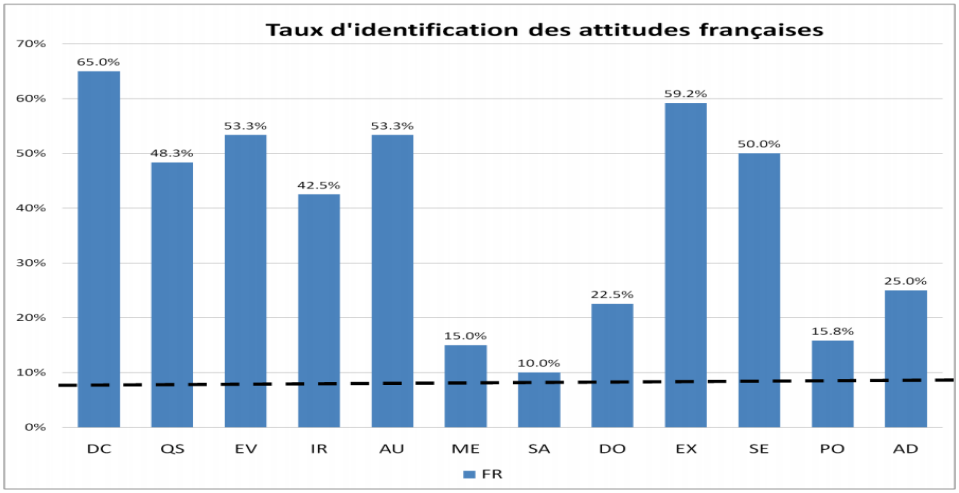
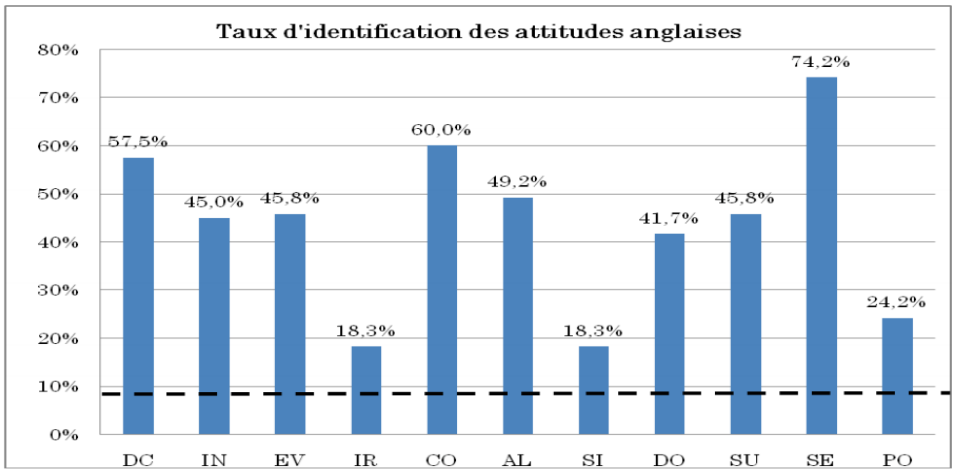


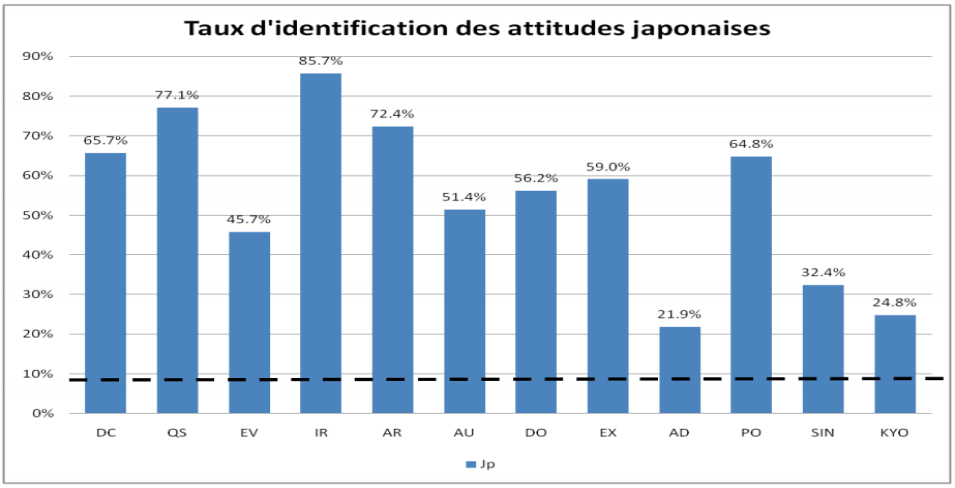
Figure 27: Résultat de la perception des 16 attitudes vietnamiennes en pourcentages d'identification (en haut) et en intensité moyenne de la perception (en bas). La ligne pointillée indique le seuil du hasard (6,25%)



(a)



(b)



(c)

Figure 28: Taux d'identification (pourcentages) des attitudes en français (a), anglais (b) et japonais (c). La ligne pointillée indique le seuil du hasard (8,4%). Signification des étiquettes : DC (déclaration), QS (question-simple), EV (évidence), IR (irritation), AU (autorité), ME (mépris), SA (ironie-sarcastique), DO (doute-incrédulité), EX (exclamation de surprise), SE (séduction), PO (politesse) et AD (admiration) [Shochi 2008]

Chapitre 6 : Etude de la perception des attitudes vietnamiennes

Comme nous l'avons mentionné dans notre introduction, l'un des objectifs de notre étude sur les attitudes vietnamiennes consiste à appliquer nos résultats dans la communication interactive entre l'humain et la machine. Dans le processus de cette communication, les attitudes produites par le système de synthèse de parole, sont véhiculées par de la parole synthétique, puis sont perçues par l'interlocuteur. Ce processus est un succès si l'interlocuteur perçoit exactement les attitudes souhaitées au moment de leur production par la machine. La perception des attitudes par l'interlocuteur dépend de plusieurs facteurs tels que le type d'attitudes présenté, la manière de la véhiculer et aussi l'état physique et les attributs individuels de l'interlocuteur (par exemple : sa culture, son sexe, son âge, etc.). C'est pourquoi, des études sur la perception des attitudes dans la communication face à face sont nécessaires afin que le système de synthèse de la parole expressive puisse produire des attitudes qui sont appropriées aux différents contextes et qui peuvent être bien perçues par l'interlocuteur.

Après la présentation dans le chapitre précédent de la construction du premier corpus audio-visuel des attitudes vietnamiennes, nous souhaitons présenter dans ce chapitre nos travaux d'une part sur la perception audio-visuelle et interculturelle des attitudes vietnamiennes et d'autre part sur l'étude de l'influence des tons vietnamiens sur la perception de ces attitudes.

6.1. Perception audio-visuelle des attitudes vietnamiennes

6.1.1. Introduction

Dans la communication face à face, les attitudes sont véhiculées d'un interlocuteur à l'autre à travers plusieurs modalités : les expressions du visage, les gestes et la parole. De nombreuses études [Ekman et al. 1979 ; Scherer 2003 ; Bänziger et al. 2009 ; Rilliard et al. 2009] ont montré que le décodage des expressions des émotions et des attitudes doit être considéré comme un processus multimodal.

Dans cette section, nous étudierons la perception audio-visuelle des attitudes vietnamiennes :

- premièrement pour connaître les effets des différents facteurs sur la perception des attitudes ;

- deuxièmement pour étudier la contribution relative des informations auditives, visuelles et audio-visuelles sur la perception de chaque attitude. Cette étude nous permet de choisir les attitudes qui sont bien véhiculées par les informations auditives, pour appliquer les résultats de notre analyse au système de synthèse de la parole expressive.

6.1.2. Méthodologie

Les tests de perception sont réalisés pour évaluer la contribution relative des facteurs suivants sur la perception des 16 attitudes vietnamiennes:

- la longueur de la phrase (en nombre de syllabes) ;
- les trois modalités (audio seul, vidéo seul et audio-visuel) ;
- l'ordre de présentation des modalités (audio en premier ou vidéo en premier).

Pour évaluer l'influence des informations auditives et des informations visuelles sur la perception des attitudes, nous avons effectué ces tests perceptifs avec des stimuli en trois modalités particulières : audio seule, visuelle seule et audio-visuelle :

- dans la modalité « audio seule », les auditeurs doivent juger des attitudes exprimées à l'aide de la seule information auditive, qui est véhiculée par la prosodie ;
- dans la modalité « visuelle seule », les auditeurs jugent des attitudes exprimées grâce à la seule information visuelle, qui est véhiculée par la gestualité faciale et du haut du buste ;
- dans la modalité « audio-visuelle », les auditeurs formulent leur jugement sur la base de ces deux informations combinées.

Pour examiner l'effet de l'ordre de présentation des modalités, les sujets sont soumis aux tests dans les deux ordres suivants :

- la modalité « audio seule » en premier, puis la modalité « visuelle seule », et pour finir la modalité « audio-visuelle » (A-V-AV) ;
- la modalité « visuelle seule » en premier, puis « audio seule » et pour finir la modalité « audiovisuelle » (V-A-AV).

Afin de limiter la complexité de ces tests, l'influence du ton n'a pas été étudiée dans cette expérience qui contient uniquement des énoncés sans variation tonale (l'influence du ton sera étudiée dans une expérience ultérieure). C'est pourquoi toutes les syllabes des énoncés sont prononcées avec le ton 1 (le ton plat).

Parmi les 15 phrases sans ton de notre corpus, trois phrases d'une, deux et cinq syllabes ont été choisies pour examiner l'influence de la longueur des phrases. Nous soulignons de nouveau que la plupart des mots vietnamiens est mono- ou

bi-syllabique [Mai et al. 2002] et la moitié des phrases de notre corpus sont composées de deux syllabes (72/125).

Les trois phrases choisies pour les tests sont présentées la Table 10. Les trois phrases choisies pour leurs longueurs différentes sont prononcées dans les 16 attitudes et présentées avec les trois modalités A, V et AV. Les tests de perception sont donc constitués de 3 * 16 * 3 = 144 stimuli.

Table 10: Les phrases choisies pour le test perceptif

Nombre de syllabes	Séquence des tons	Phrases	
		Vietnamien	Français
1	1	Ta	nous
2	1_1	Anh ta	lui
5	1_1_1_1_1	Hai trăm hai mươi ba	223

6.1.3. Protocole expérimental

Vingt auditeurs (10 hommes et 10 femmes d'un âge moyen de 25 ans), qui parlent le même dialecte que le locuteur, ont participé à cette expérience. Aucun participant n'a signalé de trouble de l'audition ni de la vision. Ces participants ont été séparés en deux groupes. Le premier groupe écoute les stimuli en audio seul d'abord, puis regarde les stimuli en vidéo seule, et enfin les stimuli audio-visuels. Le deuxième groupe a commencé avec les stimuli en vidéo seule, ensuite en audio seul et termine avec les stimuli audio-visuels.

Les tests de perception ont été effectués dans une pièce calme. L'interface donne l'étiquette et l'explication des 16 attitudes. Tous les sujets ont écouté (et/ou regardé) chaque stimulus une seule fois. Après chaque stimulus, on leur a demandé d'indiquer l'attitude ressentie parmi les 16 attitudes et d'indiquer une intensité allant de « à peine perceptible » (codé comme 1) à « très forte » (codé 100). Un score de 0 a été attribué aux 15 autres attitudes non sélectionnées.

6.1.4. Analyse des résultats

Comme nous l'avons déjà présenté, dans les tests, les sujets donnent leurs réponses en donnant deux informations différentes : un simple choix catégoriel sur l'attitude perçue ou l'intensité perçue pour l'expression de cette attitude. C'est pourquoi, notre analyse des résultats se base sur les pourcentages de bonnes réponses concernant le choix catégoriel et/ou l'intensité moyenne obtenue par ces bonnes réponses.

6.1.4.1 Effets des facteurs sur la perception des attitudes

Tout d'abord, l'influence des différents facteurs des tests perceptifs sur les réponses des sujets sont mesurés grâce à une analyse de la variance (ANOVA). Dans ces ANOVA, le taux d'identification des attitudes et l'intensité moyenne des bonnes réponses ont été choisis comme variables dépendantes de l'analyse de

variance. Les facteurs inter-sujets sont les 16 attitudes, la longueur (3 niveaux), et les modalités (3 niveaux).

La Table 11 montre les résultats d'analyses de la variance. Selon cette table, aucune différence n'a été observée entre les deux types de mesure : sur le taux d'identification et sur l'intensité moyenne. Donc, dans la suite de cette partie, seuls les résultats exprimés à partir des réponses catégorielles (ou le taux d'identification) seront utilisés pour l'analyse des résultats.

Table 11: Résultats d'ANOVA sur le taux d'identification et d'intensité moyenne. Des effets significatifs au niveau de 1 % sont en gras. Att : attitude; Mod : Modalité ; Ord : ordre de présentation des modalités ; Len: longueur de la phrase

		% identification		Intensité	
	ddl	F	p	F	p
Att	**15**	**42.304**	**0.000**	**47.804**	**0.000**
Mod	**2**	**36.830**	**0.000**	**45.373**	**0.000**
Ord	**1**	**8.950**	**0.003**	.022	0.882
Len	2	3.246	0.039	3.735	0.024
Att*Mod	**30**	**5.022**	**0.000**	**6.096**	**0.000**
Att*Ord	15	1.559	0.077	1.527	0.087
Att*Len	**30**	**3.471**	**0.000**	**3.542**	**0.000**
Mod*Ord	2	0.257	0.773	0.749	0.473
Mod*Len	4	2.572	0.036	1.822	0.122
Ord*Len	2	0.327	0.721	.238	0.788
Att*Mod*Ord	30	1.283	0.139	1.175	0.235
Att*Mod*Len	**60**	**2.153**	**0.000**	**2.104**	**0.000**
Att*Ord*Len	30	0.845	0.707	0.806	0.763
Mod*Ord*Len	4	0.989	0.412	0.547	0.701
Att*Mod*Ord*Len	60	0.615	0.991	0.644	0.985

Effet de l 'ordre de présentation des modalités

Les résultats de l'ANOVA montrent que l'effet de l'ordre de présentation des modalités n'est pas significatif ($p > 0.01$). C'est-à-dire que les deux groupes de sujets, qui passent les tests dans deux ordres différents (A-V-AV et V-A-AV), obtiennent des résultats comparables.

Effet de la longueur

Dans la table ci-dessus, le facteur de longueur de phrase (Len) ne montre pas d'influence significative sur la perception des attitudes ($p>0.01$). Mais l'interaction entre attitude, modalité et la longueur de la phrase (Att*Mod*Len) a des effets significatifs sur la perception des attitudes. Cela veut dire que dans

notre test, le nombre de syllabes des phrases n'a pas d'influence significative sur la perception de l'attitude de la phrase. Inversement la combinaison des trois facteurs *Attitude, Modalité* et *Longueur* semble, quant à elle avoir de l'influence sur la perception des attitudes.

Effet de la modalité

La modalité est un facteur important pour la perception de l'attitude. Les résultats de l'ANOVA montrent une influence significative du facteur modalité (« audio seul », « vidéo seule » et « audio-vidéo ») sur la perception des attitudes ($p < 0,01$).

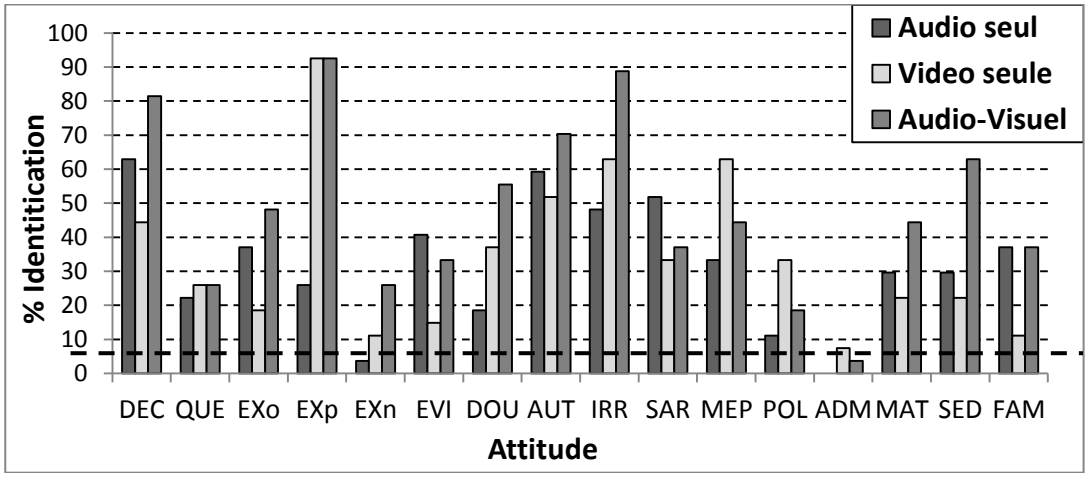


Figure 29: Taux d'identification (%) pour chaque attitude dans chaque modalité. La ligne pointillée indique le seuil du hasard (6.25 %)

La Figure 29 illustre le taux de reconnaissance moyen obtenu par chacune des attitudes pour les trois modalités différentes. Comme attendu, pour la plupart des attitudes, le score moyen en modalité audio-visuelle est meilleur que celui en audio seul ou visuel seule.

Cette figure montre que les taux d'identification pour la condition audio seule sont supérieurs à ceux pour la condition vidéo seule, et cela dans les cas des attitudes : *déclaration, exclamation de surprise neutre, évidence, autorité, ironie sarcastique* et *familier*. Cela signifie que pour ces attitudes, l'information auditive semble plus importante que l'information visuelle.

Au contraire, les informations visuelles sont nécessaires dans la perception des attitudes : *exclamation de surprise positive, doute-incrédulité, irritation, mépris* et *politesse*. En particulière, le taux de reconnaissance en modalité vidéo seule de *l'exclamation de surprise positive* est bien meilleur que celui pour la modalité audio seule, et est presque identique à celui obtenu en modalité audio-visuelle : les informations visuelles semblent être la manière principale de véhiculer ces attitudes.

Cette figure montre aussi des résultats intéressants avec les attitudes *mépris* et *politesse*. Le taux de reconnaissance de ces attitudes en modalité vidéo seule est

même plus important que dans le cas de l'utilisation de la modalité audio-visuelle. Ceci suppose que ces deux attitudes sont bien reconnues avec les informations visuelles seulement, mais aussi que les informations auditives entrainent des confusions avec d'autres types d'attitudes.

6.1.4.2 Les confusions de perception des attitudes

L'analyse sur les confusions de perception des attitudes est basée sur des matrices de confusion, qui sont présentées dans la Table 12.

Selon cette table, nous trouvons qu'en général, la modalité audio-visuelle montre le moins de confusion, tandis que la modalité audio seule montre le plus de confusion.

Dans la condition audio seule, les auditeurs ont tendance à percevoir de manière similaire la *déclaration* et la *politesse* ; *l'ironie sarcastique* et le *mépris*. La modalité vidéo seule montre de la confusion entre la *déclaration* et *l'exclamation de surprise neutre*, l'*autorité* et l'*irritation* ; *l'ironie sarcastique* et le *mépris* ; la *séduction* et le *familier*. Par contre, en utilisant les deux informations auditives et visuelles, les auditeurs ont bien distingué la plupart des attitudes.

Pour les trois modalités, *ironie sarcastique* et *mépris* sont confondus : il semble que la manière d'exprimer ces deux attitudes soient similaires pour toutes les conditions (audio, vidéo et audio-visuel). C'est pourquoi il est difficile de distinguer ces deux attitudes d'une manière indépendante du contexte. Autrement dit, ces deux attitudes ne sont bien distinguées que dans une situation déterminée.

Table 12: Matrices de confusion en taux d'identification **pour trois modalités : (a) audio seul ; (b) vidéo seul et (c) audio-visuel**

(a)

Att. Pre.	Attitudes perçues (audio seul)															
	DEC	QUE	EXo	EXp	EXn	EVI	DOU	AUT	IRR	SAR	MEP	POL	ADM	MAT	SED	FAM
DEC	63	0	0	0	0	11	0	4	0	0	0	19	0	0	0	4
QUE	26	22	11	0	0	7	19	4	0	0	0	7	4	0	0	0
EXo	0	7	37	4	15	11	22	4	0	0	0	0	0	0	0	0
EXp	0	0	7	26	30	11	7	4	15	0	0	0	0	0	0	0
EXn	15	7	19	7	4	7	19	4	0	0	0	4	4	0	7	4
EVI	4	7	7	4	0	41	4	22	0	4	0	0	4	0	0	4
DOU	4	0	26	4	30	11	19	0	0	4	0	0	0	0	4	0
AUT	0	0	0	0	0	33	0	59	7	0	0	0	0	0	0	0
IRR	0	0	11	0	11	0	4	26	48	0	0	0	0	0	0	0
SAR	0	0	0	0	0	0	0	0	0	52	44	0	0	0	4	0
MEP	0	0	0	0	0	0	0	0	0	48	33	0	0	7	4	7
POL	63	0	0	0	0	11	0	0	0	0	0	11	11	0	0	4
ADM	7	4	19	4	0	7	7	0	0	0	0	11	0	7	11	22
MAT	0	0	0	0	0	11	0	7	0	33	11	0	0	30	4	4
SED	7	0	4	0	0	4	4	0	0	0	0	7	0	11	30	33
FAM	15	0	0	0	0	15	0	0	0	0	0	7	4	7	15	37

(b)

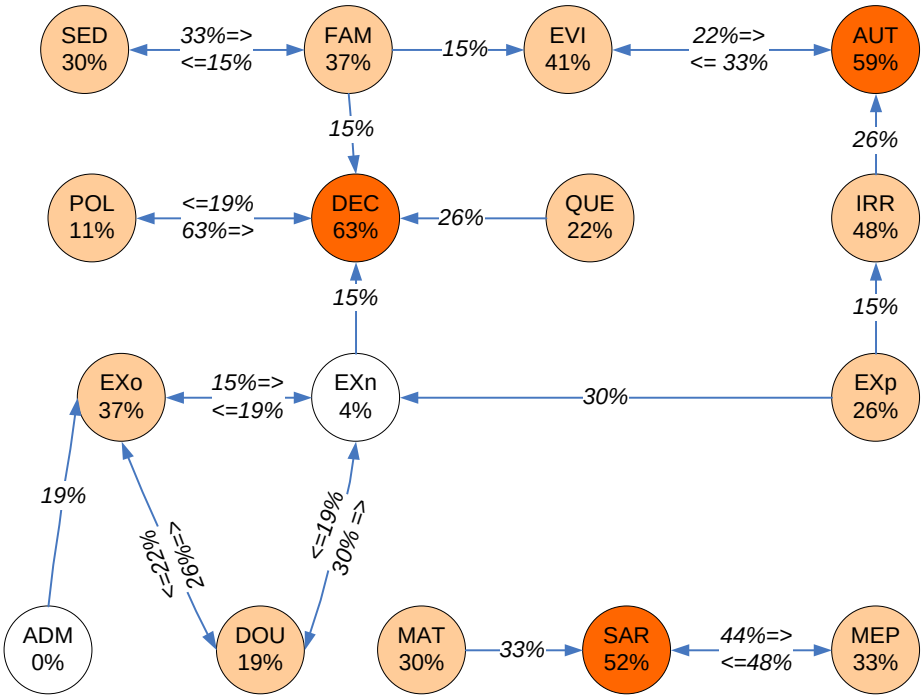
Att. Pre.	Attitudes perçues (visuel seul)															
	DEC	QUE	EXo	EXp	EXn	EVI	DOU	AUT	IRR	SAR	MEP	POL	ADM	MAT	SED	FAM
DEC	44	11	0	0	0	26	4	7	0	0	0	7	0	0	0	0
QUE	33	26	11	0	0	7	4	0	0	4	0	7	0	0	0	7
EXo	30	19	19	4	4	11	7	0	4	0	0	0	0	0	0	4
EXp	0	0	4	93	0	0	4	0	0	0	0	0	0	0	0	0
EXn	0	7	33	4	11	4	30	0	4	4	4	0	0	0	0	0
EVI	22	0	15	4	4	15	7	0	0	4	0	19	7	0	0	4
DOU	0	4	7	0	26	4	37	0	15	7	0	0	0	0	0	0
AUT	15	4	0	0	0	15	4	52	0	0	0	7	4	0	0	0
IRR	0	0	0	0	0	0	4	30	63	0	4	0	0	0	0	0
SAR	0	4	0	0	0	0	0	0	0	33	63	0	0	0	0	0
MEP	0	0	0	0	4	0	0	0	0	33	63	0	0	0	0	0
POL	7	4	4	0	0	11	7	0	0	0	0	33	33	0	0	0
ADM	4	4	26	30	0	0	4	0	0	0	0	4	7	4	0	19
MAT	4	0	11	0	0	7	7	4	4	30	4	4	0	22	4	0
SED	7	0	4	7	0	0	4	0	0	19	4	0	11	0	22	22
FAM	7	0	7	0	0	4	7	0	0	11	0	11	4	11	26	11

(c)

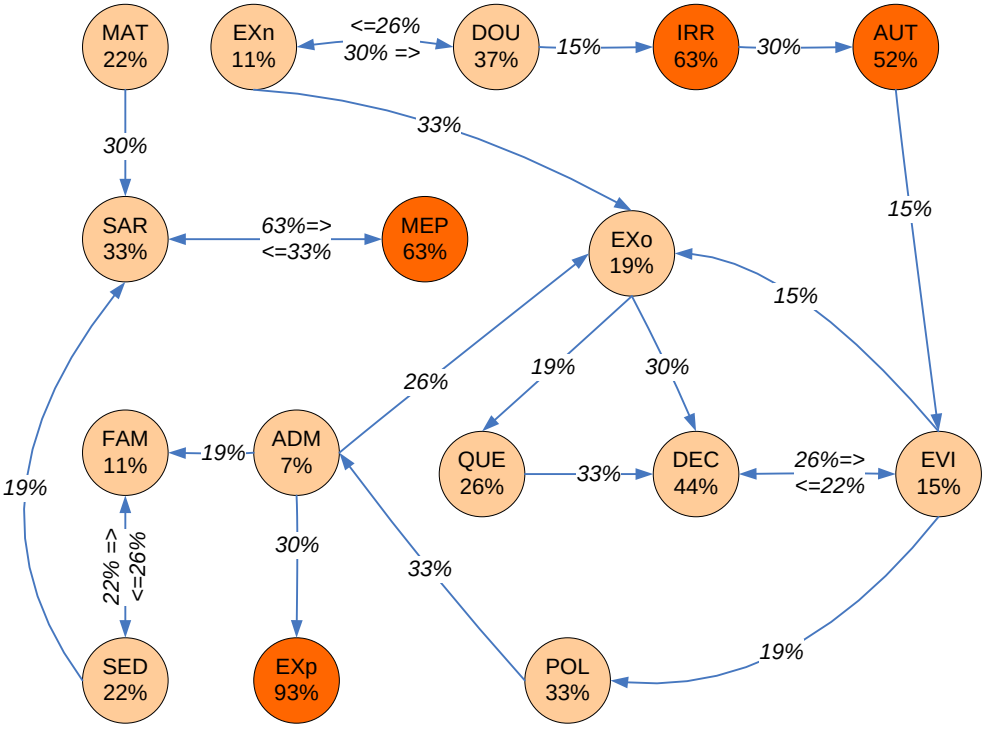
Att. Pre.	Attitudes perçues (audio-visuel)															
	DEC	QUE	EXo	EXp	EXn	EVI	DOU	AUT	IRR	SAR	MEP	POL	ADM	MAT	SED	FAM
DEC	81	0	0	0	0	11	0	4	0	0	0	0	0	0	0	4
QUE	33	26	0	0	0	22	15	0	0	0	0	4	0	0	0	0
EXo	4	7	48	15	0	7	11	0	0	0	0	0	0	0	0	7
EXp	0	0	4	93	4	0	0	0	0	0	0	0	0	0	0	0
EXn	11	4	22	7	26	4	26	0	0	0	0	0	0	0	0	0
EVI	4	4	30	11	0	33	7	0	0	0	0	11	0	0	0	0
DOU	0	4	15	0	22	4	56	0	0	0	0	0	0	0	0	0
AUT	0	0	0	0	0	22	0	70	7	0	0	0	0	0	0	0
IRR	0	0	0	0	0	0	0	11	89	0	0	0	0	0	0	0
SAR	0	0	0	0	0	0	0	0	0	37	63	0	0	0	0	0
MEP	0	0	0	0	0	0	0	0	0	56	44	0	0	0	0	0
POL	11	0	0	0	0	15	0	0	0	0	0	19	56	0	0	0
ADM	0	4	22	19	0	0	7	0	0	0	0	0	4	4	7	33
MAT	0	0	0	0	0	4	0	4	0	37	0	0	0	44	11	0
SED	0	0	0	0	0	0	0	0	0	4	0	4	0	19	63	11
FAM	0	0	4	4	0	4	0	0	0	0	0	4	0	7	41	37

Pour les trois modalités, l'*admiration* est mal reconnue, la *politesse* est aussi fortement confondue avec l'*admiration* et la *déclaration*. Ceci suppose qu'en vietnamien, ces attitudes ne peuvent être reconnues sans une cohérence sémantique ou une situation explicite. C'est-à-dire qu'en vietnamien, ces deux attitudes sont besoin de mots spécifiques pour être correctement exprimées. Donc une prosodie d'*Admiration* et de *Politesse* sur un énoncé « neutre » n'est pas écologique.

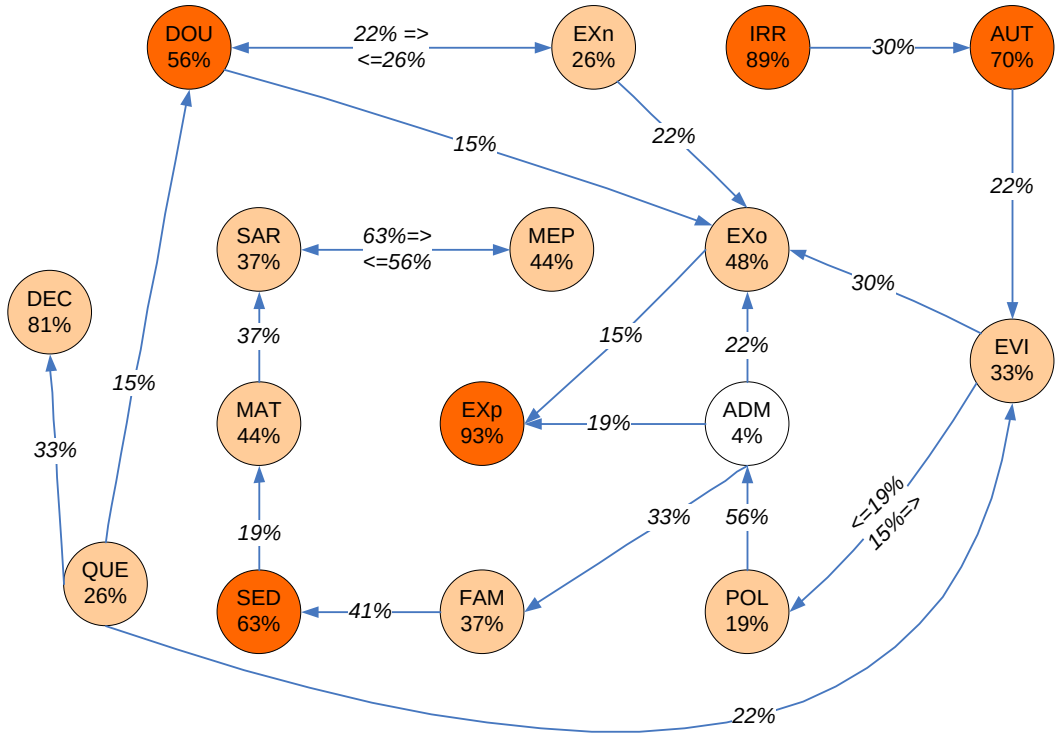
Les confusions principales entre les 16 attitudes sont schématisées par les graphes de confusions dans la Figure 30 suivante :



(a) audio seul



(b) vidéo seule



(c) audio-visuel

Figure 30: Les graphes de confusion entre les 16 attitudes vietnamiennes dans trois modalités :(a) audio seul ; (b) vidéo seule et (c) audio-visuel

6.1.4.3 Le classement des attitudes

Pour classer ces 16 attitudes vietnamiennes en termes de perception, les matrices de confusion sont analysées grâce à une méthode de classification hiérarchique, ce qui permet de regrouper les stimuli perçus pour chacune des attitudes proposées selon la proximité des distributions des réponses. La Figure 31 présente la classification hiérarchique pour les auditeurs vietnamiens dans chacune des 3 modalités, obtenue en utilisant la distance de Ward pour calculer les distances interclasses. En utilisant un seuil de 75 (proche de la moitié de la distance maximale obtenue), les 16 attitudes peuvent être ainsi réparties en groupes qui varient en fonction de la modalité considérée.

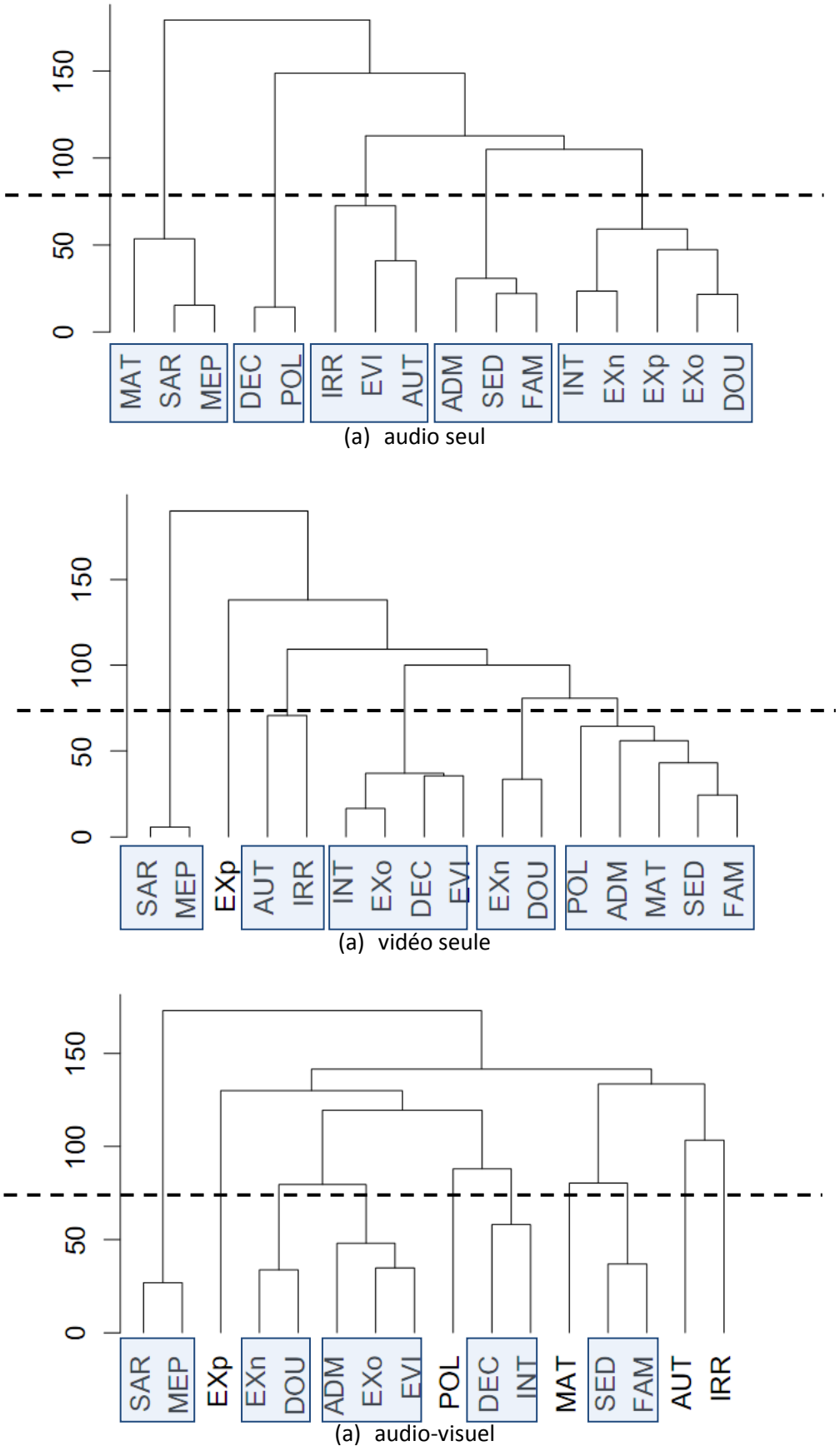


Figure 31: Dendrogrammes des 16 attitudes pour les auditeurs vietnamiens en trois modalités : (a) audio seul ; (b) vidéo seule et (c) audio-visuel

Dans la condition audio seule (a), le regroupement montre cinq groupes principaux. Avec l'information de vidéo seule (b), les résultats de classification hiérarchique donnent 5 clusters des attitudes, plus *l'exclamation de surprise positive* reconnue sans confusion. Comme prévu, la modalité AV (c) permet une meilleure identification. Le résultat de classification hiérarchique dans cette modalité regroupe les 16 expressions attitudinales en 5 clusters principaux et 5 attitudes sans confusion (*l'exclamation de surprise positive*, la *politesse*, *l'attitude maternelle*, l'*autorité* et l'*irritation*).

6.2. Perception interculturelle des attitudes vietnamiennes

6.2.1. Introduction

Comme nous l'avons mentionné dans le Chapitre 1, les attitudes sont des codes, construits et acquis dans la dynamique du langage et de ses réalités socioculturelles. Puisque les attitudes sont liées au langage et à la culture, elles peuvent différer entre les langues et les cultures. Ainsi, une attitude présente dans une langue peut ne pas exister dans une autre langue. Les études interculturelles sont un moyen pour découvrir les similarités et les décalages de perception des attitudes.

En raison de la distance linguistique (grandes différences morpho-syntaxiques, phonologiques et prosodiques) entre français et vietnamien et de la distance géoculturelle, le français est particulièrement intéressant à croiser pour cette étude interculturelle des attitudes vietnamiennes.

6.2.2. Méthodologie expérimentale

Cette expérience est réalisée selon la même procédure expérimentale que les tests précédents mais avec des sujets francophones natifs.

Vingt auditeurs francophones (10 hommes et 10 femmes avec un âge moyen de 35 ans), qui n'ont aucune expérience de la langue ni de la culture vietnamienne, ont participé à cette expérience. Ils ont aussi été séparés en deux groupes selon l'ordre de présentation des différentes modalités (A-V-AV et V-A-AV).

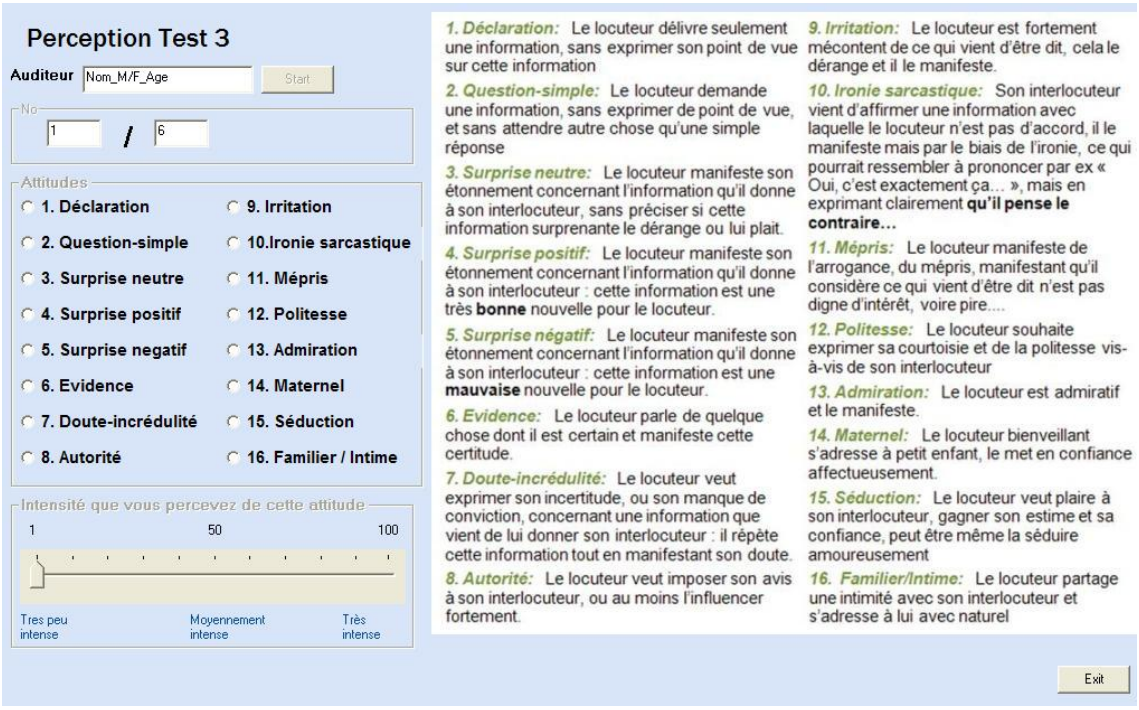


Figure 32: Interface du test de perception pour les sujets français

Les tests de perception ont été effectués dans une pièce calme. L'interface donne l'étiquette et l'explication des 16 attitudes en français (cf. Figure 32). Tous les sujets ont écouté (et/ou regardé) chaque stimulus une seule fois. Après chaque stimulus, il leur est demandé d'indiquer l'attitude supposée parmi les 16 attitudes et d'indiquer une intensité allant de « à peine perceptible » (codé comme 1) à « très forte » (codé 100). Un score de 0 a été attribué aux 15 autres attitudes non sélectionnées.

6.2.3. Analyse des résultats

6.2.3.1 Effet des facteurs

De la même manière que l'analyse des résultats avec les sujets vietnamiens, une ANOVA est réalisée pour mesurer l'influence des différents facteurs sur la perception des attitudes des sujets français. Les résultats de l'ANOVA sont présentés dans la Table 13.

Table 13: Résultats d'ANOVA sur le taux d'identification **et l'intensité moyenne avec les sujets français. Des effets significatifs au niveau de 1% sont en gras. Att: attitude; Mod: Modalité; Ord: ordre de présentation des modalités; Len: longueur de la phrase**

		% Reco.		Intensity	
	ddl	F	p	F	p
Att	**15**	**27.383**	**.000**	**33.100**	**.000**
Mod	**2**	**65.632**	**.000**	**74.767**	**.000**
Ord	1	.470	.493	.001	.975
Len	2	1.531	.217	1.655	.191
Att*Mod	**30**	**8.016**	**.000**	**9.104**	**.000**
Att*Ord	15	1.994	.013	2.971	.000
Att*Len	**30**	**2.612**	**.000**	**3.007**	**.000**
Mod*Ord	2	3.214	.040	4.955	.007
Mod*Len	**4**	**6.644**	**.000**	**6.061**	**.000**
Ord*Len	2	.570	.565	.564	.569
Att*Mod*Ord	30	.975	.504	.872	.666
Att*Mod*Len	**60**	**1.569**	**.004**	**1.721**	**.001**
Att*Ord*Len	30	1.138	.276	1.138	.277
Mod*Ord*Len	4	.965	.426	.913	.455
Att*Mod*Ord*Len	60	1.079	.318	1.122	.244

En comparant avec les résultats de l'ANOVA de l'expérience précédente, nous trouvons que les effets des facteurs sur la perception des attitudes par les sujets français sont similaires à ceux des sujets vietnamiens : l'attitude et la modalité ont un effet significatif sur la perception ; au contraire, la longueur de phrase et l'ordre de présentation ne montrent pas d'influence significative.

La Figure 33 présente le taux d'identification (%) pour chaque attitude dans chaque modalité avec les sujets français.

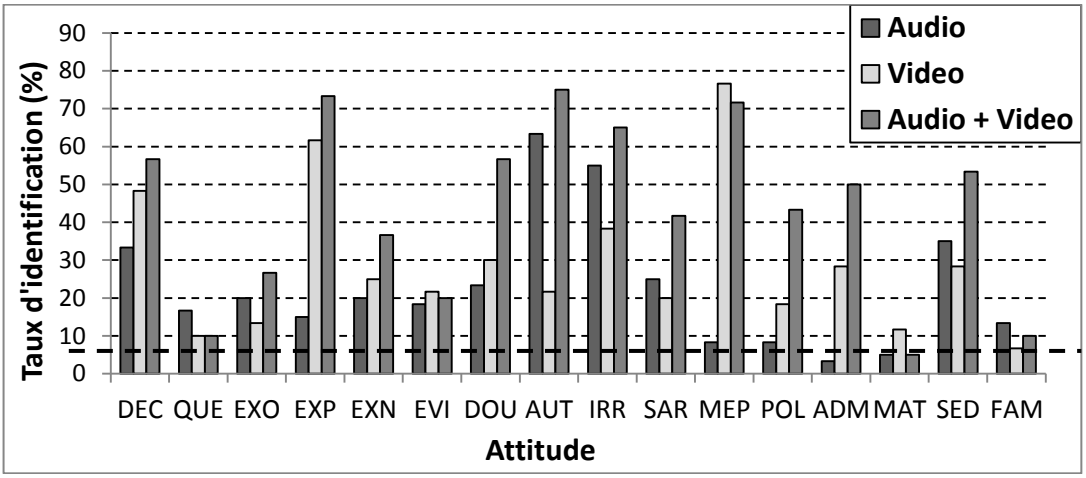
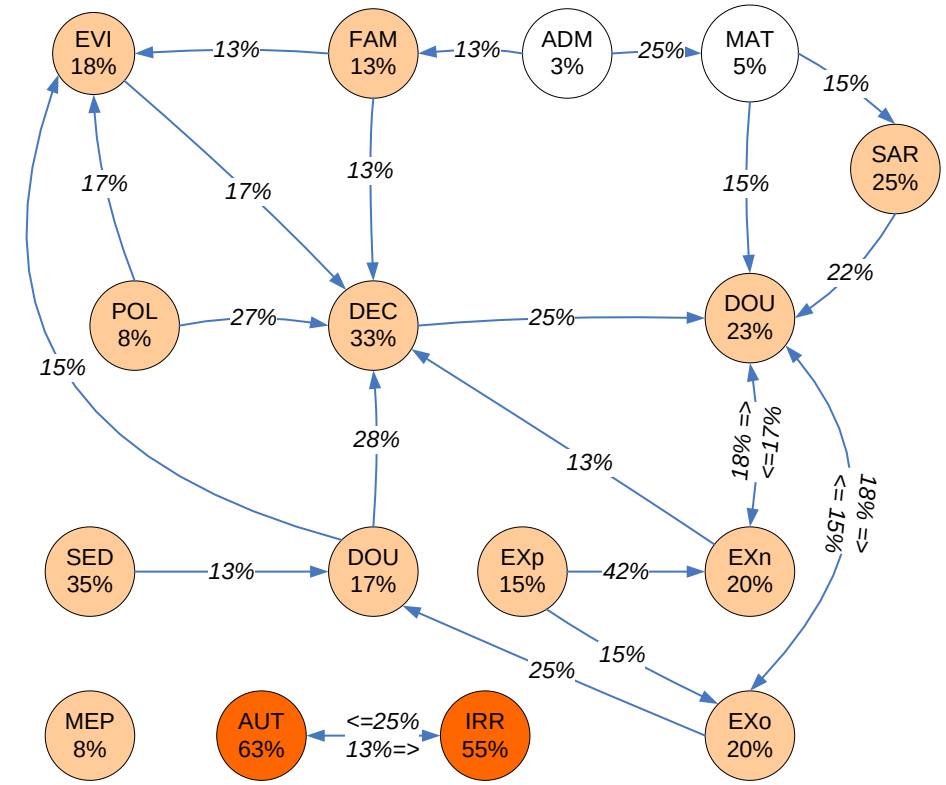


Figure 33: Taux d'identification (%) pour chaque attitude dans chaque modalité avec les sujets français. La ligne pointillée indique le seuil du hasard (6.25 %)

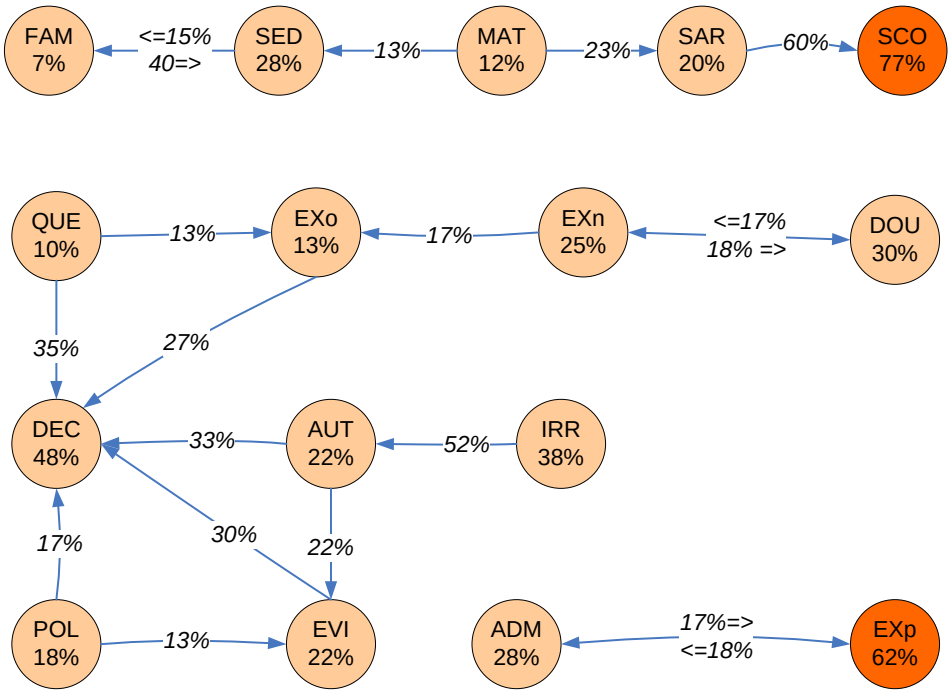
La figure montre que, globalement, la plupart des attitudes sont reconnues au-dessus du niveau du hasard (6,25 %) sauf dans le cas de l'attitude *maternel.* Avec les sujets français, la modalité audio est essentiellement informative pour l'*autorité* et l'*irritation*. Les informations visuelles sont importantes, quant à elles, dans le cas des attitudes : *déclaration, exclamation de surprise positive, mépris, politesse* et *admiration*.

6.2.3.2 La confusion

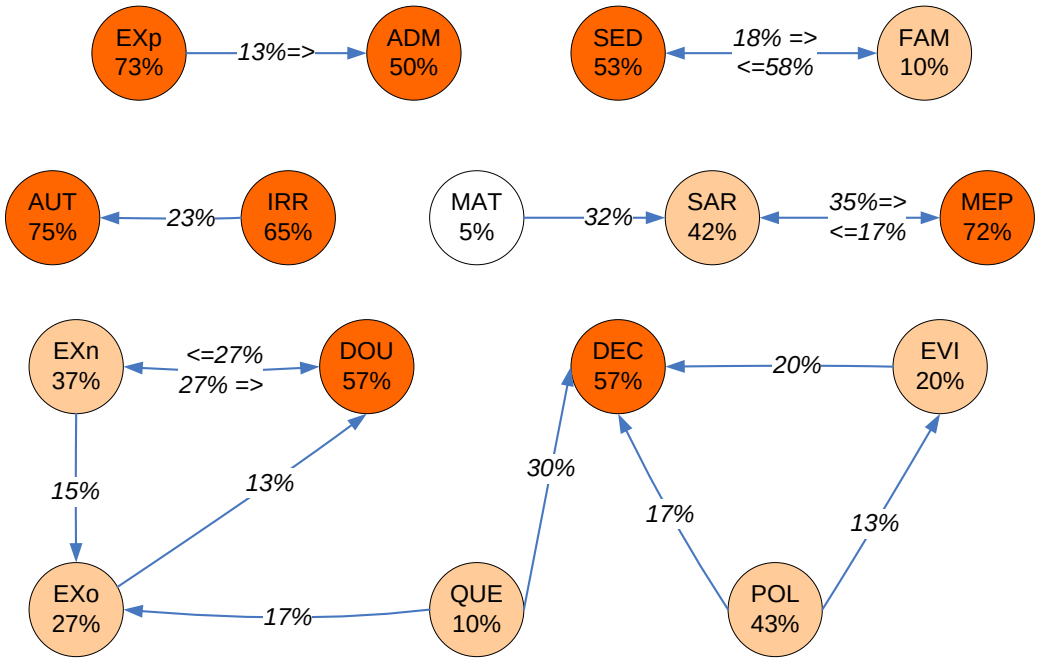
La Figure 34 montre les confusions principales dans la perception des 16 attitudes vietnamiennes par les sujets français.



(a) audio seul



(b) vidéo seule



(c) audio-visuel

Figure 34: Les graphes de confusion avec les sujets français dans trois modalités : (a) audio seul ; (b) vidéo seule et (c) audio-visuel

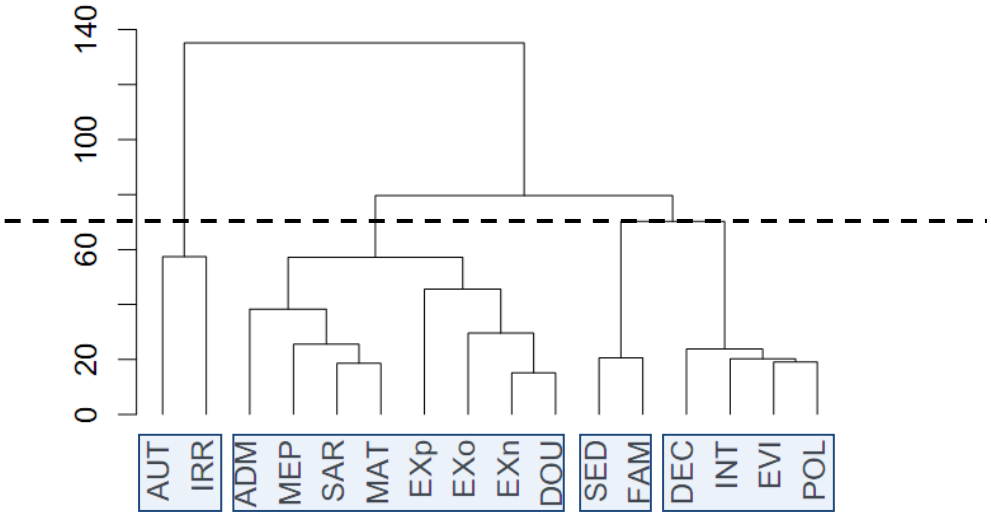
Dans les trois modalités, les sujets français ont toujours confondu l'*autorité* avec l'*irritation*. L'attitude *maternel* est aussi mal reconnue dans les trois modalités et elle est confondue avec *l'ironie sarcastique*. L'*admiration* est mal reconnue dans la condition audio seule, mais avec les informations visuelles, elle est très bien reconnue.

Comme prévu, et conformément aux résultats de l'expérience précédente avec les sujets vietnamiens, la modalité audio-visuelle permet une meilleure identification et montre le moins de confusions. Certaines confusions principales dans cette condition se trouvent entre : la *séduction* et le *familier*, l'*autorité* et l'*irritation*, *l'ironie sarcastique* et le *mépris*, *l'exclamation de surprise neutre* et le *doute-incrédulité*.

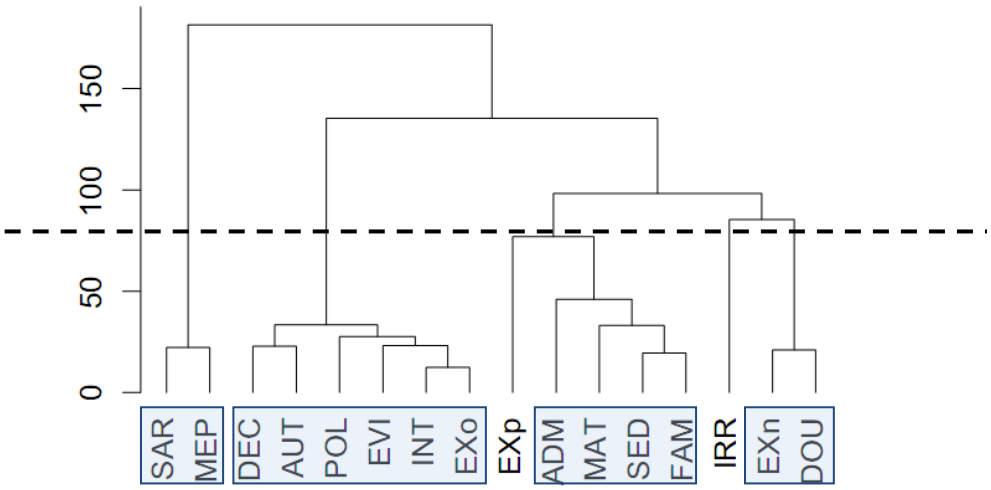
6.2.3.3 Le classement

La Figure 35 présente la classification hiérarchique des 16 attitudes vietnamiennes pour les auditeurs français dans chacune des 3 modalités, obtenue en utilisant la distance de Ward pour calculer les distances interclasses.

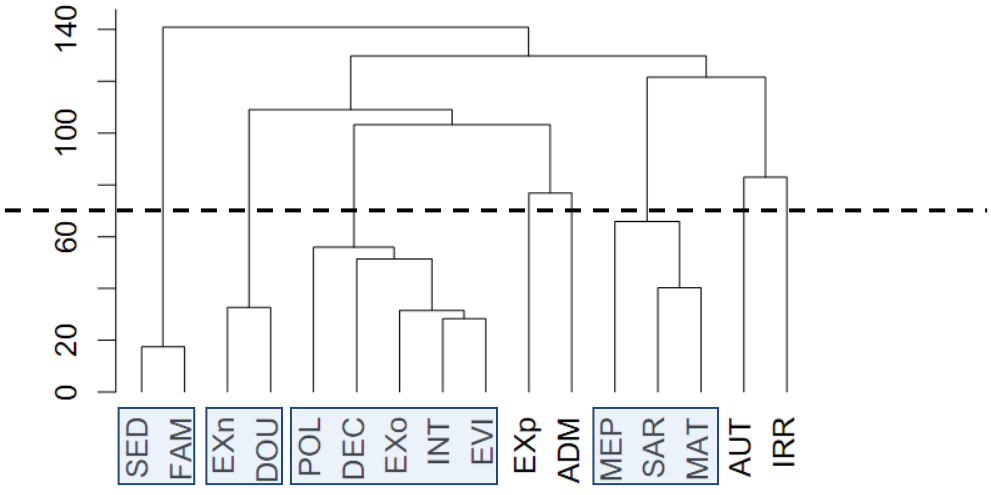
Dans la condition audio seule (a), le regroupement montre 4 groupes principaux. Les auditeurs français confondent *autorité* avec *irritation*, et *séduction* avec *familier*.



(a)



(b)



(c)

Figure 35: Le classement de la perception 16 attitudes vietnamien pour les sujets français en trois modalités : (a) audio seul ; (b) vidéo seule et (c) audio-visuel

Avec l'information de vidéo seule (b), les résultats pour les auditeurs français montrent 4 clusters, avec deux attitudes reconnues sans confusion (*exclamation de surprise positive* et *irritation*). Les éléments de chaque groupe dans la condition de vidéo seule (b) sont différents de ceux obtenus en audio seul. Dans cette modalité, les attitudes *ironie sarcastique* et *mépris* ont été regroupées dans un cluster, qui peut être appelé le groupe des expressions « impolies ».

Dans la modalité audio-visuelle, le regroupement des auditeurs français donne 4 groupes et 4 attitudes bien reconnues (*exclamation de surprise positive*, *admiration*, *autorité* et *irritation*). Dans cette figure, un cluster regroupant *l'exclamation de surprise négative* avec la *doute-incrédulité* et un autre regroupant la *séduction* et *l'exclamation de surprise neutre* sont aussi observés.

6.2.4. Comparaison des résultats entre les sujets français et vietnamiens

En comparant les résultats de la perception des attitudes entre les auditeurs vietnamiens et les auditeurs français, nous trouvons bien que les scores d'intensité moyenne obtenus par les auditeurs français sont inférieurs à ceux des auditeurs vietnamiens. Cependant, les taux d'identification des sujets français sont cohérents avec les résultats des auditeurs vietnamiens.

Pour les deux groupes d'auditeurs, certaines attitudes sont bien reconnues : la *déclaration*, *l'exclamation de surprise positive*, le *doute-incrédulité*, l'*autorité*, l'*irritation* et la *séduction*. Ceci suppose que les concepts et les réalisations de ces attitudes sont similaires dans les deux langues et deux cultures. Nous pourrons donc les considérer comme des affects sociaux interculturels entre le vietnamien et le français.

Par ailleurs, il y a des attitudes bien reconnues par les auditeurs natifs, mais qui ne sont pas reconnues par les francophones : ce sont le *maternel* et le *familier*. Ceci suppose que les réalisations prosodiques vietnamiennes de ces concepts ne sont pas partagées avec les Français et qu'elles ont besoin d'être acquises par des apprenants de langue étrangère. Des conclusions semblables ont déjà été tirées pour certaines attitudes en japonais, qui ne sont pas reconnues par les français ou les anglophones américains [Shochi et al. 2009].

Un cas très intéressant est l'expression de l'*admiration*, qui est mal reconnue par les auditeurs natifs, mais mieux reconnue par les non-natifs (dans les modalités visuelle et audio-visuelle). On peut, peut-être, relier ce résultat au fait que pour les vietnamiens, cette attitude ne peut pas se produire sans une cohérence sémantique [Le 1989]. C'est-à-dire qu'il y a impossibilité à séparer énoncé et prosodie : une prosodie d'*admiration* sur un énoncé « neutre » n'est pas naturelle, ce qui n'est pas le cas en français.

6.3. Effets des tons sur la perception des attitudes

6.3.1. Introduction

Dans les sections précédentes, nous avons présenté la perception des attitudes vietnamiennes par des sujets natifs et non natifs. Pour les deux groupes d'auditeurs (natif et non-natif), les résultats expérimentaux montrent les facteurs influant sur la perception des attitudes : la modalité de présentation et l'expression de l'attitude elle-même. Les résultats ne montrent aucun effet significatif de la longueur de phrase (en nombre de syllabes) sur la perception des attitudes. Cependant, pour limiter la complexité, les tests perception ci-dessus ne testent pas l'effet des tons vietnamien sur la perception des attitudes. Toutes les syllabes des phrases dans ces tests sont basées sur le ton 1 (le ton plat ou ton neutre).

Comme nous l'avons mentionné dans le Chapitre 2, pour une langue tonale telle que le vietnamien, la prosodie a deux fonctions importantes :

- la fonction locale (et de l'accès lexical) du ton
- la fonction globale d'attitude.

Les tons lexicaux des langues tonales ne doivent surtout pas être confondus avec les tons des théories tonales de la prosodie (modèles où l'on décompose la morphologie de la prosodie en éléments prosodiques supposés minimaux qui sont réduits à des valeurs symboliques sur la seule hauteur de la fréquence fondamentale dans ces segments prosodiques minimaux). Précisons que dans d'autres approches théoriques, comme les modèles globaux (par exemple dans [Aubergé 1992]), ces segments tonals n'existent pas et sont même des artefacts. Les tons des langues tonales comme le vietnamien sont des attributs de fréquence fondamentale perçus sur les segments de la deuxièmes articulation, les phonèmes, c'est à dire des éléments qui dans leur réalisation phonétique constituent la fonction de l'accès lexical.

Nous avons présenté aussi les caractéristiques du système tonal du vietnamien (Chapitre 4), qui utilise phonologiquement des variations de fréquence fondamentale et de qualité de voix (avec des changements dus à la source glottique). En comparant les contours prosodiques des tons vietnamiens et les contours prosodiques des attitudes vietnamiennes qui sont étudiées par [Le 1989] (voix Chapitre 4 : Figure 16 et Figure 19), nous trouvons que certains contours des tons vietnamiens et des attitudes sont similaires. Par exemple, la courbe de F0 est montante à la fin pour les tons 5, 5b et aussi pour certains attitudes telle que la joie, la doute et la surprise. De plus, l'usage de la qualité de voix rend le vietnamien intéressant, parce que des phénomènes de variation de qualité de voix ont été observés dans la morphologie de certaines attitudes (et émotions) dans d'autres langues [Campbell et al. 2003a ; Shochi et al. 2006].

Donc, la question est de savoir s'il y a des effets du ton sur la perception des attitudes en vietnamien. Pour les auditeurs vietnamiens natifs, les tons lexicaux modifient localement la courbe mélodique des contours globaux d'attitudes, sans pourtant que cette perturbation locale ait une influence sur le décodage attitudinal, parce que les auditeurs natifs ont l'expérience cognitive de la perception des tons et que de plus ils comprennent la représentation lexicale de l'énoncé. Mais une question se pose pour des auditeurs non-natifs, qui n'ont aucune expérience de la langue ni des tons vietnamiens : comment les auditeurs non-natifs filtrent ou au contraire sont mystifiés par les tons lors du décodage des contours globaux des attitudes ?

Dans cette section, nous présenterons notre expérience de perception avec des variations tonales, afin d'explorer l'effet d'un tel système tonal sur la perception des attitudes pour les auditeurs étrangers sans pratique d'une langue tonale : sont-ils capables de séparer l'information tonale locale de l'information globale d'attitude, sans confondre des deux niveaux informatifs ?

6.3.2. Méthodologie expérimentale

6.3.2.1 Sélection des phrases

Ce test de perception est réalisé pour évaluer la contribution relative des facteurs suivants sur la perception des 16 attitudes vietnamiennes:

- 8 représentations des tons vietnamiens (1, 2, 3, 4, 5, 6, 5b, 6b) ;
- les positions du ton dans la phrase (sur la première syllabe ou sur a dernière syllabe).

Parmi les 125 phrases dans notre corpus, 19 phrases ont été choisies et sont présentées Table 14. Nous choisissons des phrases avec une longueur de deux et trois syllabes pour limiter la complexité syntaxique. Nous remarquons que les tests perceptifs précédents (avec les sujets natifs et non natifs) montrent qu'il n'y a pas d'influence de la longueur de phrase (en nombre de syllabes) sur la perception des attitudes vietnamiennes.

Pour examiner l'influence du ton sur la perception des attitudes, pour chaque phrase de deux syllabes, une des 8 représentations des tons (1, 2, 3, 4, 5, 6, 5b, 6b) est positionnée sur la première ou la dernière syllabe. Pour les phrases de trois syllabes, les tons sont placés ailleurs.

Les 19 phrases choisies sont prononcées dans les 16 attitudes. Les tests de perception sont donc constitués de 19 * 16 = 304 stimuli.

Table 14: Phrases choisies pour le test perceptif avec tons

N0.	Nombre de syllabes	Séquence des tons	Phrases	
			Vietnamien	Français
1	2	1_1	Anh ta	lui
2	2	2_1	Người ta	eux
3	2	3_1	Đã xong	fini
4	2	4_1	Thủy tinh	verre
5	2	5_1	Chúng ta	nous
6	2	6_1	Chị ta	elle
7	2	5b_1	Héc ta	hectare
8	2	6b_1	Tốp ca	chant choral
9	2	1_2	Rau cần	oenanthe
10	2	1_3	Dây kẽm	fil d'acier
11	2	1_4	Cây cảnh	plante d'agrément
12	2	1_5	Y tá	infirmier
13	2	1_6	Danh bạ	annuaire
14	2	1_5b	Công tác	mission
15	2	1_6b	Sa mạc	désert
16	3	4_1_1	Bảy mươi ba	73
17	3	1_5_1	Hai chúng ta	nous deux
18	3	6_5b_3	Hợp tác xã	coopération
19	3	1_4_6	Em bảo chị	tu me dis

6.3.2.2 Sujets

Quinze auditeurs francophones natifs (8 sujets masculins et 7 sujets féminins) ont passé le test. L'âge moyen de ces 15 sujets est de 26,5 années. Ils n'ont aucune expérience de la langue ni de la culture vietnamienne..

6.3.2.3 Protocole expérimental

De la même manière que celle des expériences précédentes, le test de perception des attitudes vietnamiennes avec tons a été effectué dans une pièce calme. L'interface donne l'étiquette et l'explication des 16 attitudes en français. Tous les sujets français ont écouté chaque stimulus une seule fois. Après chaque stimulus, on leur a demandé d'indiquer l'attitude pressentie parmi les 16 attitudes et d'indiquer une intensité allant de « à peine perceptible » (codé comme 1) à « très forte » (codé 100). Un score de 0 a été attribué aux 15 autres attitudes non sélectionnées.

6.3.3. Analyse des résultats

Les résultats de ce test sont analysés et sont comparés avec les résultats des tests précédents (en modalité audio seul) pour trouver les différences entre la perception de phrases sans tons et de phrases avec tons.

6.3.3.1 Effet des facteurs

Tout d'abord, une ANOVA est réalisée pour mesurer l'influence des différents facteurs (Attitude, Ton et Position du ton) sur la perception des attitudes. Les résultats de l'ANOVA sur le taux d'identification et l'intensité moyenne sont présentés dans la Table 15.

Comme dans les expériences précédentes, les résultats de l'ANOVA ne montrent pas de différence entre les deux mesures (sur le taux d'identification et l'intensité moyenne). Évidemment, le facteur attitude a un effet significatif sur la perception. La position du ton a aussi une influence significative sur la perception des attitudes. Il y a aussi un effet de la combinaison de deux facteurs : attitude et position du ton.

Table 15: Résultats de l'ANOVA sur le taux d'identification et l'intensité moyenne pour la perception des attitudes avec ton avec sujets francophones. Des effets significatifs au niveau de 1 % sont en gras

		% identification		**Intensité**	
	ddl	F	p	F	p
Attitude	**15**	**13.493**	**.000**	**13.739**	**.000**
Ton	7	1.070	.380	1.091	.366
TonPosition	**2**	**5.433**	**.004**	**6.450**	**.002**
Attitude * Ton	**105**	**1.480**	**.001**	**1.607**	**.000**
Attitude * TonPosition	**30**	**2.121**	**.000**	**2.184**	**.000**
Tone * TonPosition	6	2.412	.025	2.076	.053
Attitude * Ton * TonPosition	**90**	**2.901**	**.000**	**2.750**	**.000**

Le facteur ton seul ne montre pas d'effet significatif, mais la combinaison du ton avec l'attitude, celle de ton, l'attitude, et la position du ton dans la phrase, montrent une influence significative sur la perception des attitudes.

La Figure 36 présente les taux de reconnaissance de toutes les attitudes pour les 8 représentations du ton vietnamien. Nous trouvons que les scores obtenus pour des tons différents sont assez similaires (environ 20 %). C'est-à-dire, globalement le ton n'a pas d'influence sur la perception des attitudes. Autrement dit, il semblerait que les sujets non-natifs peuvent séparer l'information locale du ton et l'information globale de l'attitude.

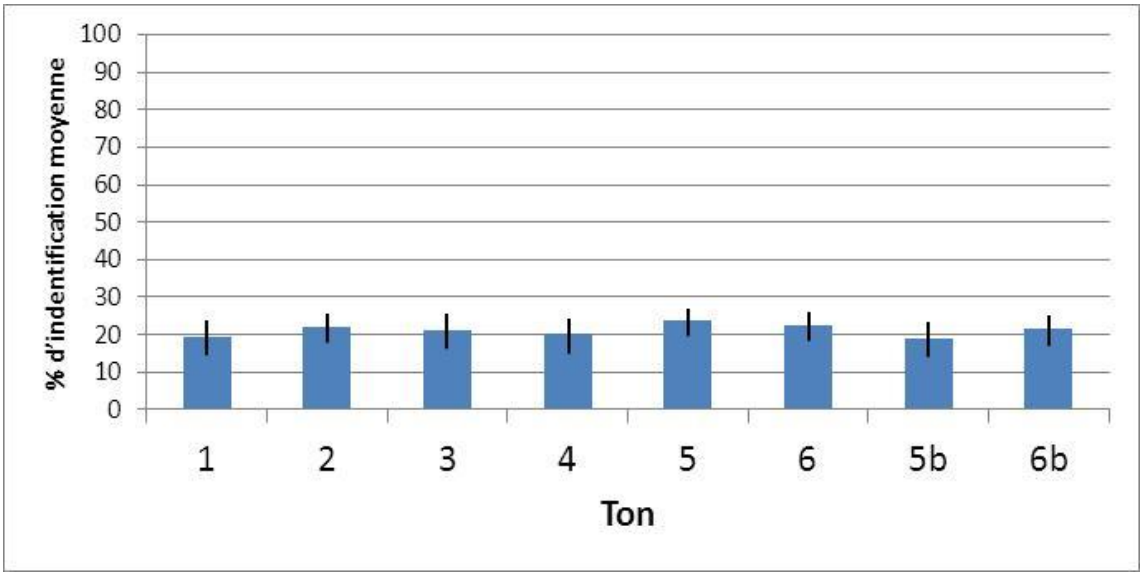


Figure 36: Taux d'identification des attitudes pour les 8 représentations du ton vietnamiens

6.3.3.2 Effet de la combinaison du ton et de la position du ton

Le résultat de l'ANOVA détaillé ci-dessus a montré qu'il y a un effet de la combinaison entre attitude, ton et position du ton. La

Figure 37 représente le taux de reconnaissance des 16 attitudes selon les 8 représentations du ton sur la première et la dernière syllabe. Nous pouvons observer dans cette figure que pour chaque attitude, les scores sont très différents en fonction des tons et de leurs positions.

Selon cette figure, certaines attitudes sont mal reconnues si un ton est positionné sur la première syllabe, mais sont bien reconnues si ce ton est positionné sur la dernière syllabe. Par exemple, la déclaration est mal reconnue avec le ton 4 et ton 6b sur la première syllabe, mais beaucoup mieux reconnue si ces tons sont positionnés sur la dernière syllabe. Les mêmes cas peuvent être observés avec le ton 6 pour l'exclamation de surprise négative, le ton 6b pour l'irritation et le mépris, les tons 5b et 6b pour la séduction.

Au contraire, certaines attitudes sont mieux reconnues si un certain ton est positionné sur la première syllabe. Ces sont les cas de l'exclamation de surprise neutre avec les tons 5, 6, 5b, 6b ; de l'évidence avec le ton 6b, de l'autorité avec le ton 3 et le ton 5, du mépris avec le ton 5b et de la séduction avec le ton 6.

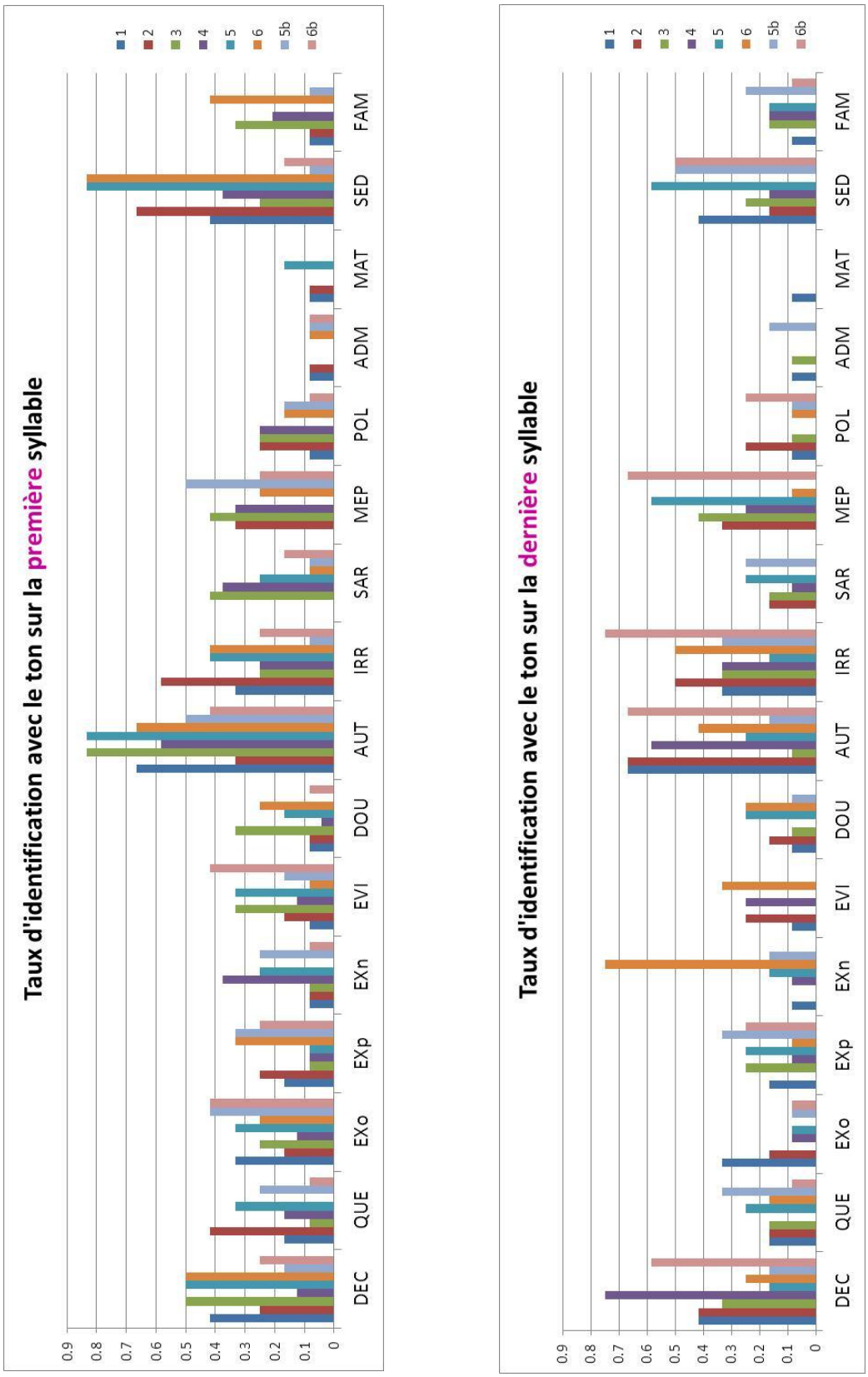


Figure 37: Taux d'identification des attitudes pour 8 représentations en première et dernière syllabes

6.3.3.3 Comparaison des résultats entre les sujets français et vietnamiens, sur des phrases avec et sans ton

La Figure 38 montre les taux d'identification des 16 attitudes pour les sujets vietnamiens et les sujets français (sur les phrases sans ton et avec tons).

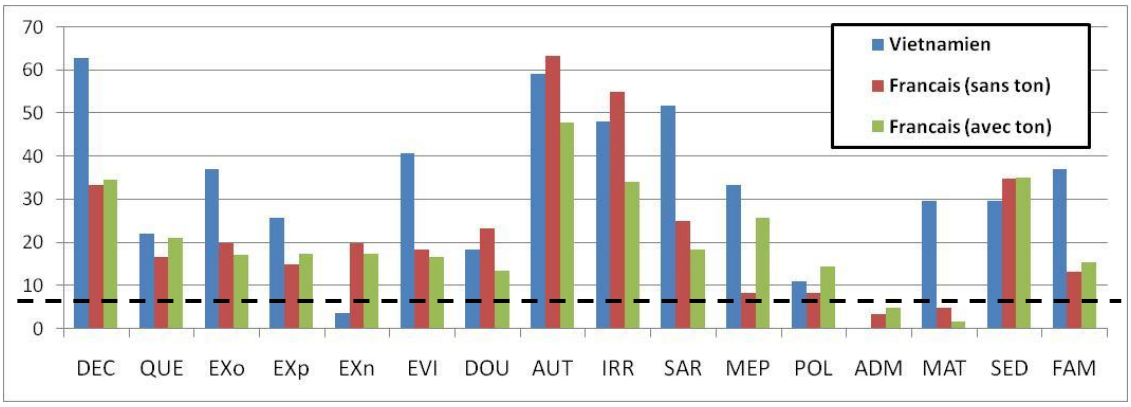


Figure 38 : Les taux d'identification des 16 attitudes pour les sujet vietnamiens et les sujets française (sur les phrase sans ton et avec ton). La ligne pointillée indique le seuil du hasard (6.25 %)

Cette figure montre que la plupart des attitudes (sauf l'*admiration*) ont été reconnues au-dessus du seuil du hasard (6,25 %). L'attitude « *maternelle* » est bien reconnue par les auditeurs natifs, mais elle n'est pas reconnue par les auditeurs non natifs. Ceci suppose que les réalisations prosodiques vietnamiennes de ce concept ne sont pas partagées avec les français et elles ont besoin d'être acquises par des apprenants de langue étrangère. Un cas intéressant est *l'exclamation de surprise négative*, qui est mal reconnue par les auditeurs natifs, mais mieux reconnue par les non-natifs dans les deux cas : avec et sans ton. Comme nous l'avons montré dans la section 6.1, pour les vietnamiens, les informations visuelles sont nécessaires dans la perception de cette attitude. C'est-à-dire qu'il y a impossibilité pour le sujet vietnamien de distinguer l'attitude d'*exclamation de surprise négative* avec l'information audio seulement, ce qui n'est pas le cas avec les sujets français.

Pour la plupart des attitudes, les résultats de la perception des auditeurs français sur les phrases avec ton sont presque égaux à ceux obtenus pour les phrases sans ton. Ce résultat de la perception vérifie donc ce que nous avons montré dans l'ANOVA ci-dessus : globalement, le ton n'a pas d'influence sur la perception des attitudes pour les sujets francophones.

Cependant, dans les cas d'*autorité* et d'*irritation*, les scores avec des phrases sans tons sont meilleurs que les scores des phrases avec tons (plus de 10 %). Ceci suppose que les caractéristiques prosodiques des tons vietnamiens provoquent des confusions entre ces deux types d'attitudes.

Au contraire, l'attitude de *mépris* est beaucoup mieux reconnue avec des phrases avec tons. Peut-être, pour les sujets français, le *mépris* est difficile à distinguer

avec des phrases sans tons, mais avec des tons, les différences entre cette attitude et les autres est plus clair. Autrement dit, les caractéristiques prosodiques de cette attitude sont d'autant plus saillantes quand elles sont combinées aux caractéristiques prosodiques du ton : les tons vietnamiens reproduisaient artificiellement une « prosodie française » de cette attitude.

6.3.3.4 Etude de la confusion entre attitudes pour les phrases sans ton et pour les phrases avec tons

La Table 16 présente les matrices de confusion des 16 attitudes avec les sujets français sur les phrases avec et sans tons. Comme nous l'avons mentionné, globalement, le ton n'a pas d'influence sur la perception des attitudes. Et dans cette table, nous trouvons aussi qu'il n'y a pas beaucoup de différences entre les deux matrices de confusion : globalement, le ton vietnamien ne change pas les confusions entre les attitudes.

Cependant, pour chaque attitude, nous pouvons trouver quelques différences entre les deux cas. Il y a de nouvelles confusions qui apparaissent pour la perception sur des phrases avec ton. Par exemple dans ce cas *l'exclamation de surprise positive* est bien confondue avec l'*autorité*, ce qui n'existe pas dans la perception pour les phrases sans ton. Dans le cas avec ton, les attitudes d'*évidence* et de *doute* montrent une confusion plus forte avec la *question-simple* que dans le cas de phrases sans ton. L'attitude « *familier* » présente aussi une nouvelle confusion avec l'attitude « *maternel* » dans la perception sur des phrases avec ton.

Table 16: les matrices de confusion des 16 attitudes avec les sujets français avec les phrases sans ton (a) et avec ton (b)

(a) sans ton

Att. Prés_entées	Attitudes perçues															
	DEC	QUE	EXo	EXp	EXn	EVI	DOU	AUT	IRR	SAR	MEP	POL	ADM	MAT	SED	FAM
DEC	33	5	7	0	7	25	2	7	0	2	7	3	0	2	0	2
QUE	28	17	2	0	2	15	8	3	3	3	3	7	2	0	0	7
EXo	7	25	20	10	7	7	15	0	0	2	2	3	2	0	0	2
EXp	0	0	15	15	42	2	10	2	10	0	0	0	5	0	0	0
EXn	13	3	13	5	20	3	18	0	0	3	0	5	3	7	2	3
EVI	17	10	12	7	5	18	10	0	5	3	2	0	3	0	0	8
DOU	5	10	18	3	17	2	23	0	2	7	2	2	2	3	2	3
AUT	7	0	2	0	0	3	0	63	13	3	3	3	0	0	0	2
IRR	2	3	2	0	3	2	2	25	55	0	7	0	0	0	0	0
SAR	5	3	3	5	3	8	22	0	8	25	5	3	2	2	0	5
MEP	10	0	10	0	3	17	8	2	0	10	8	10	2	7	7	7
POL	27	5	12	3	0	17	5	0	0	0	0	8	5	8	0	10
ADM	12	7	7	3	2	3	12	0	0	0	2	7	3	25	5	13
MAT	5	5	10	7	7	13	15	3	0	15	3	2	5	5	0	5
SED	8	13	5	0	0	5	2	0	0	0	0	10	8	8	35	5
FAM	13	3	8	0	0	13	0	0	0	2	2	8	7	7	23	13

(b) avec ton

Att. prés_entées	Attitudes perçues															
	DEC	QUE	EXo	EXp	EXn	EVI	DOU	AUT	IRR	SAR	MEP	POL	ADM	MAT	SED	FAM
DEC	35	9	6	0	0	19	5	4	2	4	4	5	0	1	0	5
QUE	31	21	6	1	0	9	8	2	1	0	2	7	1	2	1	8
EXo	7	22	17	4	9	7	11	3	1	0	1	6	3	2	0	6
EXp	0	0	4	18	39	2	3	16	10	0	2	0	3	1	0	1
EXn	7	8	8	3	18	4	10	0	0	0	0	9	12	10	5	5
EVI	14	18	11	2	1	17	7	4	2	2	1	8	2	1	1	7
DOU	4	23	8	6	6	2	14	1	2	1	0	10	8	7	1	6
AUT	8	8	2	0	0	8	1	48	14	1	6	1	0	0	0	3
IRR	1	2	3	2	11	0	2	32	34	1	12	0	0	0	0	0
SAR	3	1	5	6	3	5	14	0	16	18	7	2	12	1	2	4
MEP	7	4	4	2	0	15	8	3	6	10	26	3	3	4	3	3
POL	29	7	5	0	0	10	3	0	0	0	0	14	3	11	5	11
ADM	10	3	4	0	0	5	5	0	0	1	2	9	5	21	18	16
MAT	3	6	3	1	2	7	10	4	8	31	7	5	6	2	0	6
SED	11	5	2	0	0	4	3	0	0	1	0	8	5	14	35	12
FAM	13	2	1	1	0	6	1	1	0	1	2	11	4	21	22	15

6.4. Conclusion

Dans ce chapitre, nous avons présenté une série de tests perceptifs sur les attitudes vietnamiennes :

Le premier test vise à l'évaluation des attitudes vietnamiennes par les auditeurs natifs, en modalité audiovisuelle.

Tout d'abord la performance des réalisations des attitudes par le locuteur vietnamien a été bien évaluée. Le résultat montre que les distributions des réponses données pour la plupart des attitudes sont significativement différentes du hasard (sauf dans le cas de l'admiration). Les résultats de ce test montrent aussi les facteurs influençant la perception des attitudes : l'expression de l'attitude elle-même et la modalité de présentation (audio, visuelle et audio-visuelle). Un point important montré par ce test est que la longueur de la phrase (en nombre de syllabes) n'a pas d'influence notable sur la perception des attitudes.

Ce test montre aussi le rôle comparé de l'information auditive et de l'information visuelle sur la perception des 16 attitudes vietnamiennes. Les informations audio jouent un rôle important dans les cas des attitudes : *déclaration, exclamation de surprise neutre, évidence, autorité, ironie sarcastique* et *familier*. Au contraire, les informations visuelles sont nécessaires pour : *exclamation de surprise positive, doute-incrédulité, irritation, mépris* et *politesse*.

Par ailleurs, l'analyse sur la confusion nous permet de connaître les similarités dans la perception entre les 16 attitudes dans les trois modalités. Ensuite, les matrices de confusion ont été analysées grâce à une méthode de classification hiérarchique, qui permet de regrouper les stimuli perçus selon la proximité des distributions des réponses pour chacune des attitudes proposées. Ces regroupements ont permis de mettre en évidence les expressions les mieux reconnues et celles qui montrent le plus de confusions. Ces résultats seront utilisés dans le chapitre suivant pour choisir des attitudes bien distinctes pour la modélisation de la prosodie d'attitudes et pour leur application à la synthèse de la parole.

Le deuxième test vise à l'évaluation interculturelle des attitudes vietnamiennes, par les sujets français.

Ce test nous permet de trouver les spécificités culturelles des attitudes vietnamiennes. Les résultats de ce test montrent que certaines attitudes sont perçues de manière similaire dans les deux langues et les deux cultures (*la déclaration, l'exclamation de surprise positive, le doute-incrédulité, l'autorité, l'irritation et la séduction*), mais certaines attitudes reflètent des spécificités culturelles du vietnamien et ne sont pas partagées avec les auditeurs non natifs (les attitudes « *maternel* » et « *familier* »).

Ces résultats sont issus des premières études sur la perception audio-visuelle et interculturelle des attitudes vietnamiennes. Ils posent des questions intéressantes

pour des recherches futures, par exemple dans les domaines de la communication multilingue ou dans le domaine de l'enseignement des langues étrangères.

Le troisième test est réalisé avec les auditeurs non-natifs d'une langue à tons sur des phrases avec tons enfin d'explorer l'effet d'un tel système tonal sur la perception des attitudes.

La conclusion importante de ce test est que le ton n'a pas d'influence significative globale sur le comportement perceptif des auditeurs. Cette conclusion peut confirmer notre hypothèse sur la proposition des contours fonctionnels de la prosodie (voir la section 3.2.2.3). Selon cette hypothèse, le contour intonatif d'un énoncé peut être décomposé en des composants plus simples, correspondant aux niveaux linguistiques/paralinguistiques différents. Chaque niveau à une forme particulière de contours intonatifs. Le test perceptif sur la phrase avec tons montre que les auditeurs non natifs peuvent séparer l'information locale du ton et l'information globale d'attitude. Autrement dit, il semblerait que ces deux fonctions sont indépendantes dans la production de la prosodie d'énoncé et aussi sont indépendantes dans la perception.

Ce point est très important pour notre travail sur la modélisation et la génération de la prosodie de la parole expressive dans le cadre d'une langue à tons. L'idée est que la génération de la prosodie complexe d'un énoncé expressif avec tons peut être obtenue par la combinaison de deux modèles prosodiques indépendants : un modèle pour générer les contours des tons et un modèle pour générer les contours prosodiques globaux d'attitude. Nous présenterons cette méthode en détail dans le chapitre suivant.

Chapitre 7 : Modélisation de la prosodie de la parole expressive en vietnamien

Après avoir présenté nos travaux sur la construction et la perception d'affects sociaux prosodiques en vietnamien, nous nous tournons maintenant vers notre second objectif de thèse : l'intégration de l'expressivité dans la parole synthétique. Nous présentons dans ce chapitre la génération de la prosodie d'attitudes vietnamiennes et son application à la synthèse de la parole expressive. Ce chapitre commence d'abord par l'explication de notre sélection des attitudes pour application à la synthèse, fondée sur les résultats de la perception des attitudes vietnamiennes décrits dans le chapitre précédent. Puis, en nous basant sur le modèle de superposition des contours fonctionnels (voir la section 3.2), nous proposons une méthode pour modéliser et générer de la prosodie expressive en vietnamien. Cette méthode est ensuite appliquée pour générer de la parole expressive en vietnamien, puis évaluée par des tests de perception sur des énoncés synthétiques.

7.1. Sélection des attitudes pour la modélisation

Comme nous l'avons mentionné dans le Chapitre 5, notre corpus est certainement le premier corpus audio-visuel permettant d'étudier la prosodie des attitudes en vietnamien. Le corpus contient donc 16 affects sociaux, mais les évaluations perceptives ont permis de dégager celles parmi les 16 qui sont, dans notre corpus, les plus performantes dans la modalité acoustique.

Rappelons que les étiquettes que nous avons posées a priori et à partir desquelles ont été produits les énoncés du corpus, même si elles ont déjà été discutées expérimentalement en perception intra et inter-culturelle, depuis [Fónagy 1983] jusqu'à [Lu et al. 2012], restent encore des hypothèses empiriques et que même le fonctionnement cognitif catégoriel (vs. dimensionnel) de ces objets est toujours une question ouverte [Aubergé 2012]. C'est pourquoi, si bien même ces étiquettes répondent à un fonctionnement cognitif catégoriel, leurs frontières, leurs valeurs, restent encore à établir expérimentalement. Ainsi, à partir du test de perception réalisé avec les auditeurs vietnamiens en audio seul, et de la classification hiérarchique (voir la section 6.1.4), nous pouvons proposer de regrouper ces 16 attitudes en 5 classes comme dans la Figure 39, les attitudes d'un même groupe, ou plutôt la réalisation de ces attitudes par notre locuteur, étant perceptivement proches.

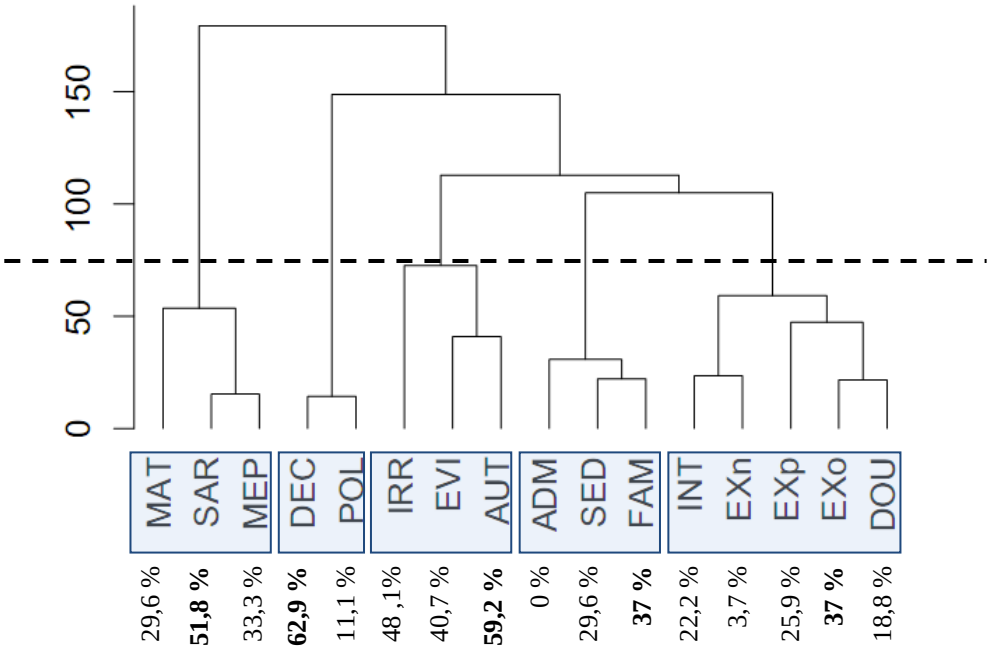


Figure 39 : Classement et taux d'identification **des 16 attitudes pour les auditeurs vietnamiens en modalité audio seul. Les meilleurs résultats pour chaque groupe sont en gras.**

Si nous essayons de rassembler dans une pseudo-cohérence conceptuelle les étiquettes de chaque groupe, nous pouvons proposer :

1. *Les expressions de position sociale négative* : l'ironie sarcastique (SAR), le mépris (MEP), par contre ici l'attitude maternel (MAT) est difficile à expliquer dans ce regroupement ; cette attitude a été assez bien reconnue par les vietnamiens, mais très mal par les français. Peut-être que le fait que le locuteur soit masculin introduit un biais « situationnel » sur la perception de cette attitude, et la ramènerait ainsi à un concept de « second degré » propre à l'ironie et au mépris.

2. *Les expressions sans valeur illocutoire* : la déclaration (DEC), la politesse (POL) ; notons que la politesse, quand elle n'est pas extrême, est souvent confondue avec la déclaration [Grépillat 1996 ; Shochi et al. 2007 ; Rilliard et al. 2009 ; Lu et al. 2012]

3. *Les expressions impliquant une dominance relationnelle*: l'irritation (IRR), l'évidence (EVI) et l'autorité (AUT) ;

4. *Les expressions de relation sociale positive* : l'admiration (ADM), la séduction (SED), le familier (FAM) ;

5. *Les expressions impliquant la nouveauté informationnelle* : l'exclamation de surprise neutre (EXo), l'exclamation de surprise positive (EXp), l'exclamation de surprise négative (EXn), la question-simple (QUE) et le doute-incrédulité (DOU).

Pour l'application à la synthèse de la parole, dans chaque groupe, nous choisissons l'attitude qui présente le meilleur résultat dans le test perceptif en audio seul (voir la Figure 40). De cette façon, nous retenons 5 attitudes : déclaration (DEC), exclamation de surprise neutre (EXo), autorité (AUT), ironie sarcastique (SAR) et familier (FAM).

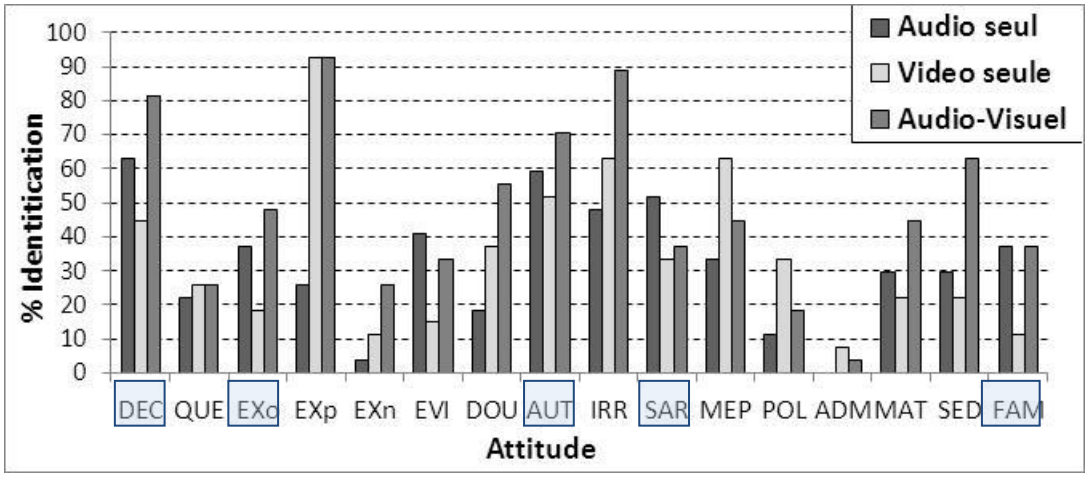


Figure 40 : Taux d'identification **(%) pour chaque attitude dans chaque modalité avec les auditeurs vietnamiens**

Comme rappelé dans le Chapitre 2, la qualité de voix a été proposée comme le quatrième paramètre de la prosodie, à côté des trois paramètres classiques (la fréquence fondamentale, la durée et l'intensité). Mais actuellement, il n'existe pas de méthodes bien établies pour l'estimation complète de la qualité de voix. La modélisation et la génération de la qualité de voix restent des problèmes à étudier dans les champs de la parole expressive [Audibert 2008]. De plus, le système de synthèse de la parole utilisé à ce jour ne peut pas encore générer de la parole avec différentes qualités de voix : seules quelques techniques de synthèse sont aujourd'hui capables de générer des paramètres de qualité de voix ; elles sont presque toutes basées sur le modèle source-filtre [Fant 1970], par exemple celui développé à Orange Labs, et évalué comme pertinent par [Audibert et al. 2006a]. Sinon une solution par défaut consisterait, dans une méthode de synthèse par corpus, à enregistrer des segments non modaux, et à étiqueter leur qualité de voix en vue de la sélection pour la synthèse. Pour ces raisons, dans notre première étude sur la modélisation de la prosodie des attitudes en vietnamien, et pour limiter dans cette première approche la complexité du problème, nous nous concentrons seulement sur les trois paramètres classiques : la fréquence fondamentale et l'intensité organisées sur la durée. Cependant, la qualité de voix sera prise en compte dans le futur pour d'autres travaux visant à la modéliser et à la générer.

Parmi les 5 attitudes retenues ci-dessus, nous constatons que la qualité de voix de l'attitude « familier » (FAM) est modifiée. Cette attitude est souvent, et c'est le cas dans notre corpus, fortement caractérisée par une voix soufflée (« breathy voice », cf. [Wichmann 2000 ; Campbell et al. 2003b]), voir Figure 41. Donc, nous décidons de ne pas choisir l'attitude « familier » pour la modélisation de la

prosodie de la parole expressive en synthèse dans son état actuel, puisque seuls les paramètres « classiques » y sont manipulables.

Finalement, parmi les 16 attitudes vietnamiennes dans notre corpus, quatre attitudes sont validées pour la modélisation de la prosodie (sur 3 paramètres classiques) : la déclaration (DEC), l'exclamation de surprise neutre (EXo), l'autorité (AUT) et l'ironie sarcastique (SAR).

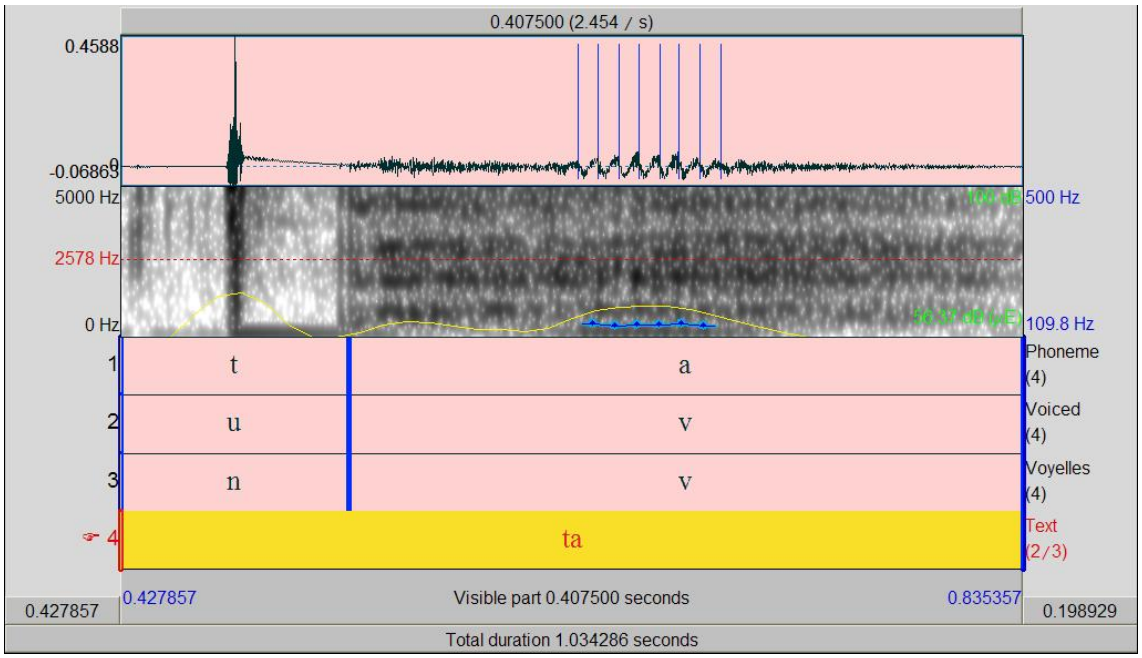


Figure 41: Un exemple de phrase d'une syllabe avec l'attitude « familier »

7.2. Méthodologie de modélisation de la prosodie des attitudes

Comme nous l'avons présenté dans le Chapitre 2, dans la communication parlée, les fonctions de la prosodie sont multiples, elles sont partagées avec d'autres modules structurels du langage et de la parole, et toutes concourent à la construction du sens.

La morphologie en partie autonome de la substance prosodique par rapport à la morphologie des autres substances de la parole (l'intégration prosodie/continuum des phonèmes étant un problème ouvert malgré les travaux multiples par exemple sur « l'alignement » tonal) a inspiré de très nombreux débats scientifiques toujours non résolus quant au fonctionnement phonologique et cognitif de la prosodie.

L'étude de la morphologie prosodique est depuis de nombreuses années, en soi, un large champ d'études : modèles globaux vs. modèles morphologiques vs. modèles tonals se complètent ou se contredisent en bien des points. Une proposition des plus « provocatrices » a même été de donner à la prosodie le

statut de troisième articulation du langage [Rossi 1999]. Récemment, Aubergé [2012] propose que la construction du sens ne commence pas aux mots, c'est-à-dire à la première articulation mais directement par un pointage direct du sens par la substance prosodique, puis indissociablement se complexifierait en une double direction, par un pointage indirect passant par les pointeurs concaténés des phonèmes (les traces de cette « prosodie pure » se retrouvant dans les micro-expressions décrites par Loyau et al [2007], Signorello et al [2010], Vanpé [2011], de Biasi et al. [2012]).

Nous avons aussi présenté dans la section 3.2 le modèle de superposition des contours fonctionnels de la prosodie et le concept de « rendez-vous » structurels entre l'intonation et les niveaux linguistiques/paralinguistiques, proposé par Aubergé [1991].

Selon cette approche, la prosodie se décompose en contours spécifiques de segments de chaque niveau hiérarchique, le plus long étant l'énoncé ou l'acte de langage. La fonction de modalisation porte uniquement sur le segment le plus haut. La fonction de focalisation porte quant à elle sur le segment le bas de la fonction de segmentation / hiérarchisation. La fonction segmentation / hiérarchisation définit la syntaxe, c'est pourquoi, souvent on la nomme fonction syntaxique, alors qu'il serait plus juste de souligner que la prosodie et la syntaxe coopèrent pour mettre en œuvre cette fonction.

Rappelons que la fonction des affects sociaux (ou des attitudes) apporte des informations sur:

- l'attitude ou le point de vue du sujet sur son énonciation, une attitude particulière étant de ne pas indiquer de point de vue ou d'attitude sur cette énonciation ou acte de parole (ce qui est communément appelé « neutre » et qui est à notre sens l'expression de ne pas avoir/vouloir/pouvoir donner une attitude, ce qui est bien en soi une attitude : c'est pourquoi ce « neutre » figure dans nos énoncés comme la modalité déclaration ;
- la situation sociale de l'interaction, en particulier reliée à la hiérarchie sociale des interlocuteurs au moment de leur interaction ; il peut s'agir de valeurs variant selon les cultures, avec par exemple des expressions de politesse ou d'autorité ;
- le contexte socioculturel de l'énonciation, ceci est typiquement relié à l'expression de l'intimité, à la situation du parent s'adressant au jeune enfant, à la séduction, etc.

Fondé sur l'approche proposé par Aubergé [1991], notre modèle de superposition des contours fonctionnels de la prosodie pour la langue tonale vietnamienne est présenté dans le Figure 42.

Dans ce modèle :

- la fonction des affects sociaux (ou des attitudes) porte sur le segment le plus haut, soit le niveau de phrase ;
- la fonction syntaxique porte au niveau du groupe et sous-groupe démarqués ;
- la fonction des tons porte sur le segment de niveau le plus bas, au niveau de la syllabe.

Nous supposons que chaque niveau (la phrase, le groupe / sous-groupe et la syllabe) correspond à une forme particulière des contours intonatifs. Les formes des contours sont indépendantes et différentes entre les différents niveaux. La forme du contour de chaque niveau encode aussi une fonction prosodique correspondant à son niveau. Par exemple, le contour au niveau de phrase véhicule la fonction de la modalité de phrase ou l'attitude du locuteur, le contour au niveau de groupe / sous-groupe présente la fonction de segmentation / hiérarchisation. Les formes prosodiques au niveau de la syllabe correspondent aux contours des tons.

Segment	**Fonction prosodique**
Phrase	Modalité Attitude
Groupe	Syntaxe
Sous-groupe	
Syllabe	Tons

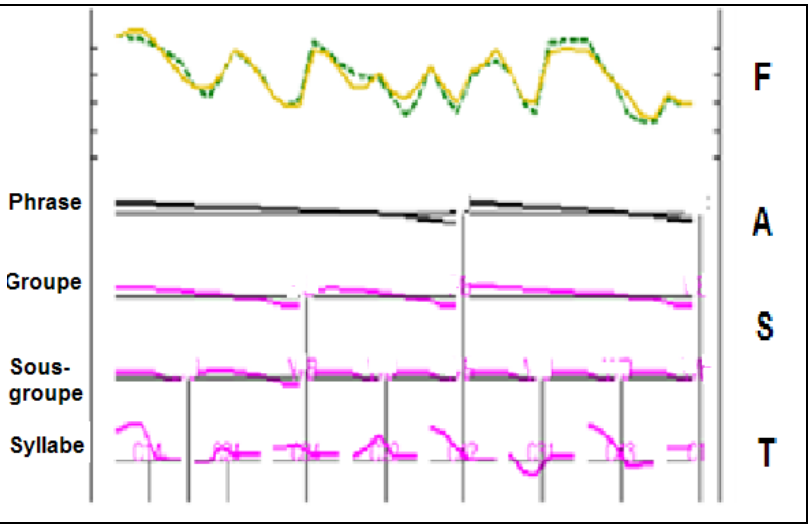


Figure 42 : Modèle de superposition des contours prosodiques pour la langue tonale, d'après [Aubergé 1991] et [Chen et al. 2004]

Supposons que F(X) représente le contour prosodique observé sur une phrase portant l'attitude X. Selon notre proposition ci-dessus, F(X) se décompose selon les contours spécifiques suivants:

- le contour au niveau de la phrase, qui véhicule la fonction prosodique de l'attitude X, est représenté par A(X) ;
- le contour au niveau du groupe et du sous-groupe, qui véhicule la fonction prosodique segmentation / hiérarchisation, est représenté par S ;
- le contour au niveau de la syllabe, ou la séquence des contours des tons dans la phrase, est représenté par T ;

Le contour prosodique final F(X) est donc la superposition des contours de tous ces niveaux. Nous proposons que cette superposition soit implémentée par une fonction additive simple comme suit :

$$F(X) = A(X) + S + T \qquad (1)$$

Supposons que R (référence) = S+T, soit la superposition des contours au niveau de la syllabe (T) et du niveau de groupe /sous-groupe (S). R véhicule donc la fonction du ton et la fonction syntaxique. Autrement dit, R présente le contour prosodique de la phrase, qui contient les informations syntaxiques et tonales, mais pas de fonctions attitudinale. Par suite nous pouvons considérer R comme la représentation du contour prosodique de la phrase avec attitude « neutre ».

Selon [Aubergé 2002a], l'attitude est utilisée pour donner le point de vue du locuteur pendant son énonciation, son acte de parole. Un énoncé avec une attitude « neutre » est une attitude qui correspond au fait de ne pas avoir/vouloir/pouvoir donner un point de vue personnel de la part du locuteur. Selon notre définition des attitudes vietnamiennes dans la Table 7, la *déclaration* est une attitude quand le locuteur fait part d'une simple information, sans exprimer aucun point de vue. Nous pouvons considérer la *déclaration* comme une attitude « neutre ». Donc, R peut être considéré comme le contour prosodique de la phrase pour l'attitude de déclaration.

Dans l'équation (1) ci-dessus, en remplaçant S+T par R, nous obtenons :

$$F(X) = A(X) + R \qquad (2)$$

Le contour fonctionnel de l'attitude X peut être obtenu comme suit :

$$A_X = F(X) - R \qquad (3)$$

En conséquence, le contour fonctionnel au niveau de la phrase, véhicule la fonction prosodique d'attitude X, représenté par A(X), et peut être considéré comme la différence entre le contour prosodique final avec l'attitude X, représenté par F(X), et le contour prosodique de déclaration, représenté par R. Parce que F(X) et R représentent les contours prosodiques sur la même structure syntaxique et tonale, les informations syntaxiques et tonales sont éliminées dans le résultat A(X). Autrement dit, A(X) est indépendant de la syntaxe et les tons, il dépend seulement de l'attitude X et de la longueur de la phrase.

En résumé, le contour fonctionnel de l'attitude X peut être considéré comme la différence entre la prosodie la phrase portant l'attitude X et la prosodie de la phrase portant l'attitude « neutre » (ou la phrase de déclaration). Nous supposons que, pour une attitude, cette différence peut être représentée par une famille de contours prosodiques particuliers, correspondant à des longueurs différentes de la phrase. Les contours dans une famille présentent des formes similaires et ils représentent la caractéristique prosodique de l'attitude correspondante. Selon le modèle de superposition des contours fonctionnels proposé ci-dessus, la famille

des contours prosodiques fonctionnels correspondant à l'attitude X peut être obtenue par l'écart ou le rapport entre les contours prosodiques des phrases avec une attitude X et des phrases avec l'attitude de déclaration. Nous avons:

Contour fonctionnel de l'attitude X A(X)	=	Contour prosodiques de la phrase avec attitude F (X)	-	Contour prosodique de la phrase de déclaration R

Après avoir obtenu les contours fonctionnels de l'attitude X, représenté par A(X), nous pouvons générer la prosodie finale de la phrase avec l'attitude X, représenté par F(X). F(X) peut être obtenu par une superposition additive entre R et le contour fonctionnel A (X) comme suit :

Contour prosodique de la phrase avec attitude X F(X)	=	Contour prosodique de la phrase de déclaration (R)	+	Contour fonctionnel de l'attitude X A(X)

Nous allons présenter dans les paragraphes suivants l'extraction des paramètres prosodiques et la modélisation du contour prosodique fonctionnel à propos des 4 attitudes sélectionnées : *la déclaration* (DEC), *l'exclamation de surprise neutre* (EXo), *l'autorité* (AUT) et *l'ironie sarcastique* (SAR).

7.3. Extraction et modélisation la prosodie fonctionnelle d'attitude

7.3.1. Méthodologie générale

Comme nous l'avons mentionné, notre tâche principale pour la modélisation de la prosodie attitudinale consiste à trouver la différence entre la prosodie de la phrase avec une attitude X et la prosodie de la phrase avec l'attitude « neutre » (ou la déclaration). Nous supposons que cette différence peut être représentée par une famille de contours prosodiques particuliers, qui peuvent être appelés une famille de « contours fonctionnels d'attitude ».

Dans les deux équations ci-dessus :

- F(X) et R présente le contour prosodique sur la même structure syntaxique et tonale ;
- A(X) est indépendant de la syntaxe et des tons.

A(X) ne dépend pas de la structure syntaxique et tonale choisie pour F(X) et R. Donc, pour calculer A(X), et afin de limiter la complexité des tons, nous

choisissons F(X) et R possédant des contours prosodiques sur des phrases sans tons (c'est-à-dire où toutes les syllabes dans la phrase ont le ton 1, ou ton neutre).

Supposons que $F(X)_{L,}$ et $R_{L,}$ représentent les contours prosodiques des phrases sans ton avec une longueur L (en nombre de syllabes), alors le contour fonctionnel de l'attitude X avec la longueur L ($A(X)_L$) est calculé comme suit :

$$A(X)_L = F(X)_{L,} - R_L \qquad (4)$$

Si nous avons *n* phrases d'attitude X avec la même longueur L, la moyenne du contour fonctionnel d'attitude X pour la longueur L est calculée :

$$A_{moyen}(X)_L = \frac{\sum_{i=1}^{n}(F_i(X)_L - R_{i,L})}{n}$$

$$A_{moyen}(X)_L = \frac{\sum_{i=1}^{n} F_i(X)_L}{n} - \frac{\sum_{i=1}^{n} R_{i,L}}{n}$$

Nous obtenons alors :

$$A_{moyen}(X)_L = F_{moyen}(X)_L - R_{moyen,L}$$

où :

- $A_{moyen}(X)_L$ est la moyenne du contour fonctionnel de l'attitude X pour la longueur L ;
- $F_{moyen}(X)_L$ est la moyenne des tous les contours prosodiques de l'attitude X sur les phrases sans tons pour la longueur L ;
- $R_{moyen,L}$ est la moyenne de tous les contours prosodiques de l'attitude *déclaration* sur les phrases sans tons pour la longueur L.

La collection des $A_{moyen}(X)_L$ avec des longueurs L différentes constitue la famille des contours fonctionnels de l'attitude X, qui représente alors les caractéristiques prosodiques de l'attitude X.

Notre objectif consiste maintenant à calculer et modéliser ces familles de contours pour les 4 attitudes choisies, en nous basant sur les paramètres F0 fréquence fondamentale, durée et intensité.

7.3.2. Contour de la fréquence fondamentale F0

Dans notre modèle de la superposition de contours fonctionnels de la prosodie, la variation de la fréquence fondamentale est considérée globalement comme un contour dans l'axe temporel (organisé par exemple en évolution de la durée de segments), même si nous traitons séparément, par facilité, la génération de la durée. Nous allons présenter dans cette section notre travail sur l'extraction de la modélisation du contour F0 selon la méthode générale détaillée ci-dessus.

7.3.2.1 Extraction les contours moyens de F0

Pour les 4 attitudes choisies, nous allons calculer les 8 contours moyens de F0, correspondant aux 8 longueurs de phrase (de 1 à 8 syllabes, la syllabe étant notre base segmentale phonologique pour établir le compteur de longueur).

L'extraction des valeurs de F0 est réalisée semi-automatiquement avec des scripts Praat. Le choix de Praat [Boersma et al. 2011] plutôt que d'autres éditeurs plus performants comme WinPitch [Martin 2000], s'est justifié par son accessibilité et par sa cohérence avec les autres modules développées en synthèse, même si pour des études prosodiques ultérieures plus fines et précises, un éditeur comme WinPitch sera bien plus indiqué.

Tout d'abord, toutes les phrases sans tons dans le corpus sont segmentées en phonèmes. Puis les valeurs de F0 (en demi-tons et Hertz) sont extraites avec des fenêtres d'analyse de 10 ms pour chaque segment vocalique délimité par les frontières phonémiques. Ensuite, les contours de F0 sont normalisés sur 10 points pour chaque syllabe afin de calculer les valeurs moyennes syllabe par syllabe. Un exemple des contours moyens de F0 des phrases de 5 syllabes, pour les 4 attitudes, est présenté dans la Figure 43. Les contours avec d'autres longueurs peuvent être consultés dans l'Annexe 2.

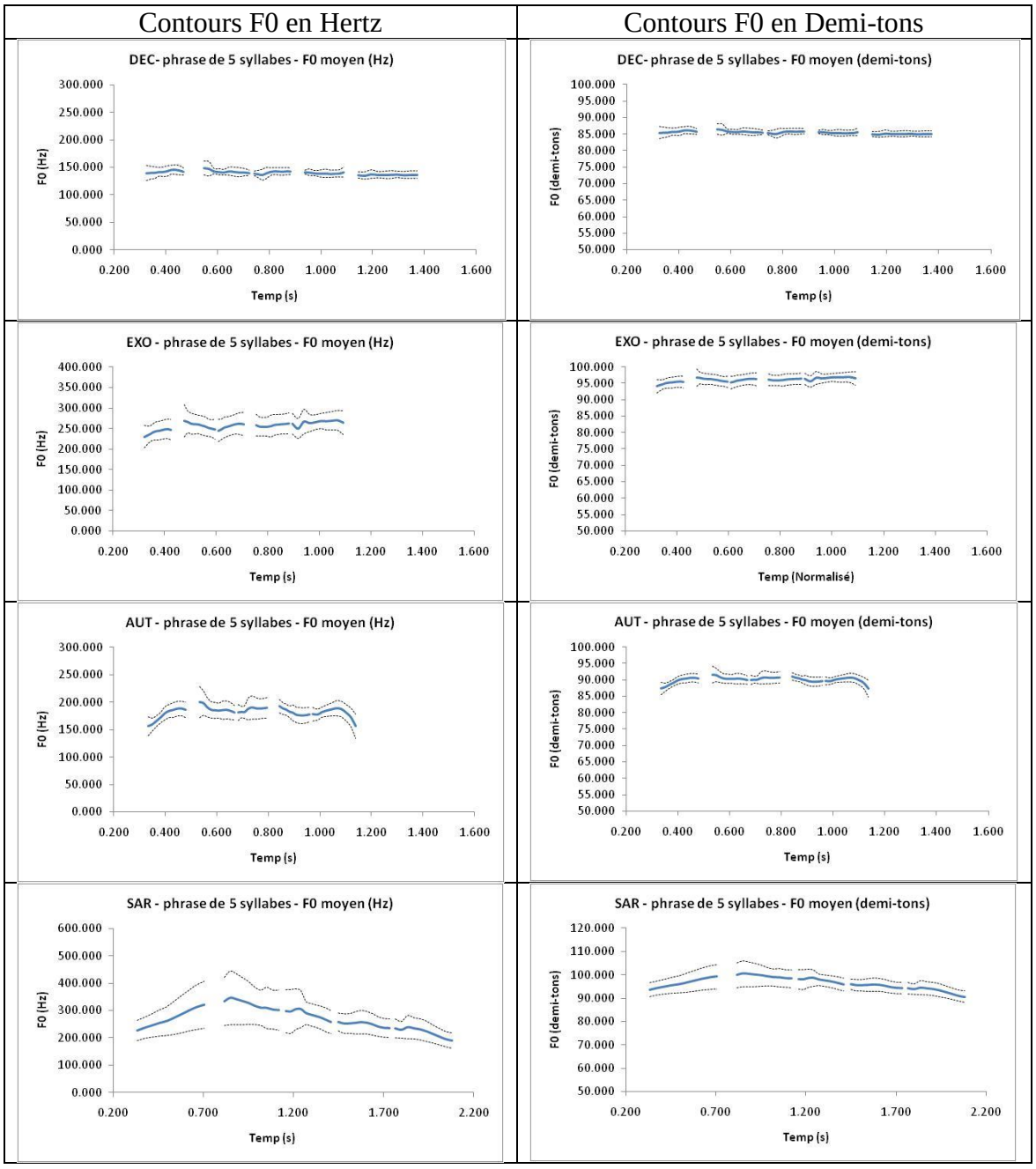


Figure 43: Un exemple des contours moyens de F0 des phrases de 5 syllabes avec 4 attitudes : déclaration (DEC), surprise neutre (EXo), autorité (AUT), et sarcastique (SAR)

7.3.2.2 Contours fonctionnels de F0 pour l'attitude

Comme nous l'avons présenté dans la méthodologie, pour chaque longueur de la phrase, le contour fonctionnel de F0 de l'attitude X est calculé comme la différence entre le contour F0 moyen de l'attitude X et celui de la déclaration. Nous avons trois familles de contours fonctionnels de F0 correspondant aux trois attitudes, *surprise neutre* (EXo), *autorité* (AUT), et *sarcastique* (SAR), qui sont présentées dans les figures suivantes (Figure 44 à Figure 46).

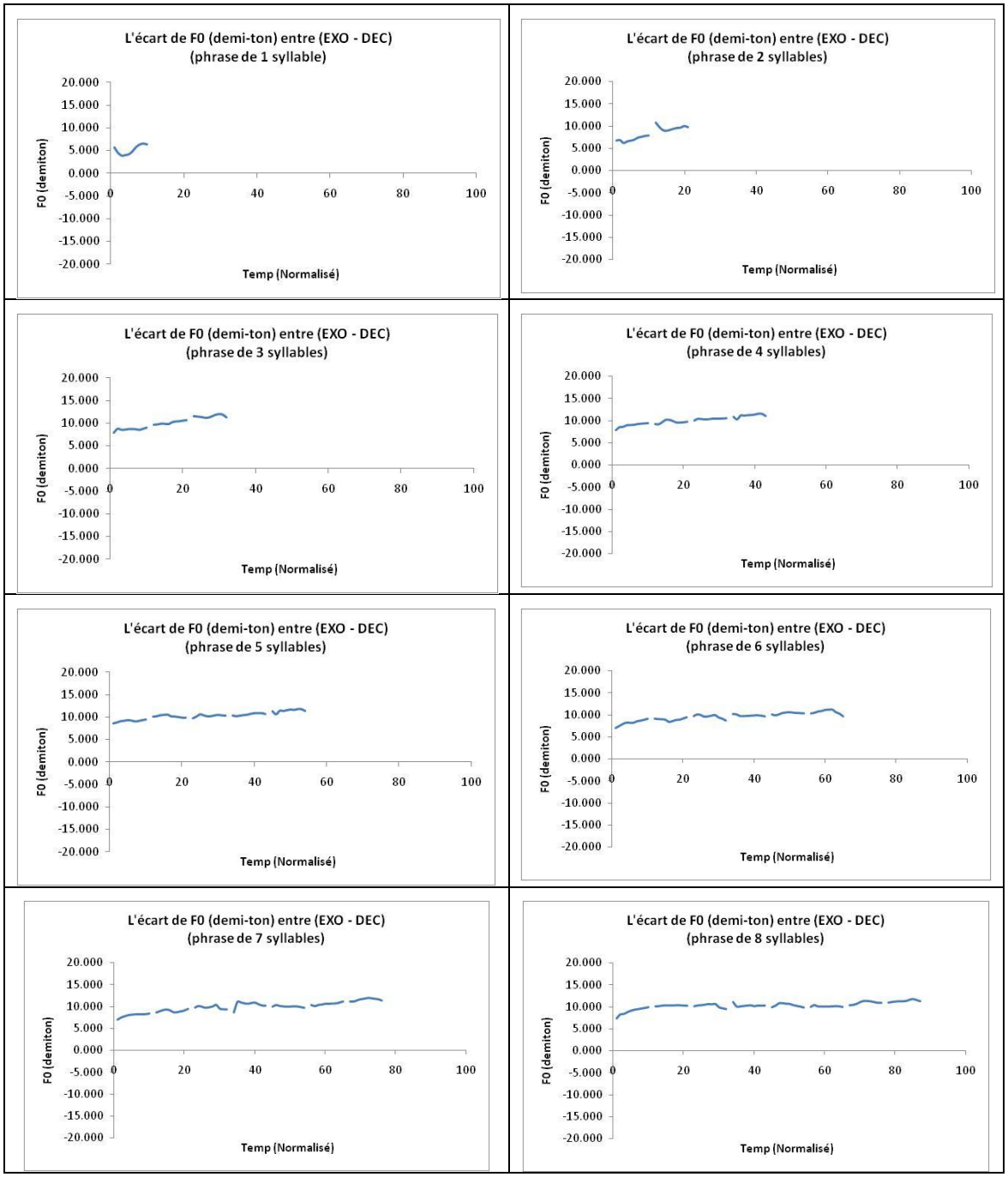


Figure 44: Famille des contours fonctionnels de F0 de l'attitude surprise neutre (EXo)

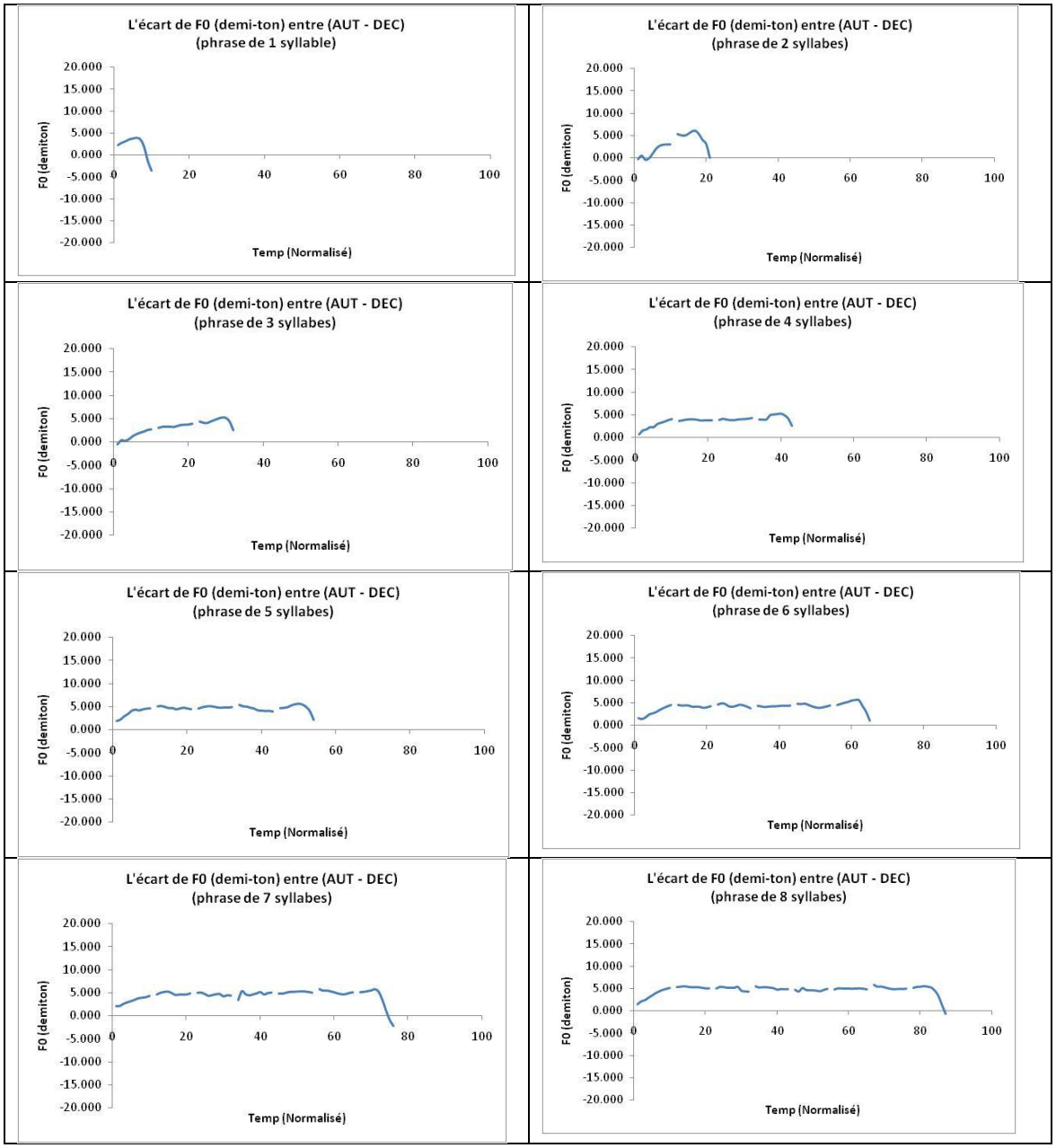


Figure 45: Famille des contours fonctionnels de F0 de l'attitude autorité (AUT)

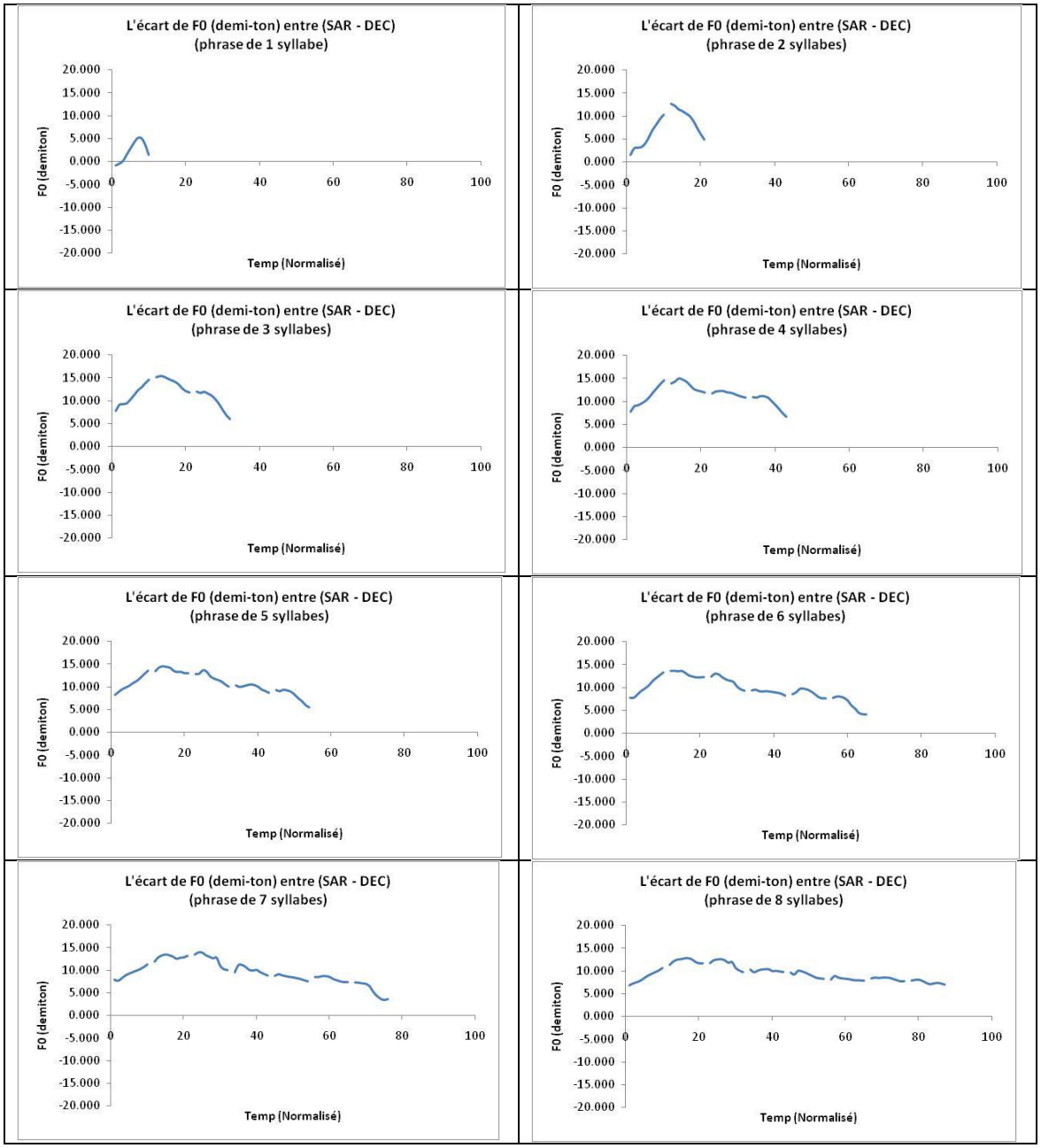


Figure 46: Famille des contours fonctionnels de F0 de l'attitude sarcastique ironie (SAR)

En nous basant sur l'observation des contours fonctionnels de F0 des trois attitudes ci-dessus, nous trouvons que pour chaque attitude, les contours fonctionnels de F0, quelle que soit la longueur, ont une forme générique commune. Cette forme commune diffère d'une attitudes à l'autre. Ces formes spécifiques permettent donc de caractériser et de distinguer des attitudes. Les formes communes des trois attitudes peuvent être décrites qualitativement comme suit :

- la forme prosodique commune de la *surprise neutre* (Figure 44) est caractérisée par une courbe linéaire globalement montante, partant d'une valeur initiale de 7 demi-tons, montant progressivement jusqu'à une valeur finale de 11 demi-tons ;
- la forme prosodique commune de l'attitude *autorité* (Figure 45) débute par une valeur basse (de 0 à 2 demi-tons) et puis monte très rapidement ; Cependant, après la première syllabe, elle reste stable et est caractérisée par un plateau à 5 demi-tons ; dans la dernière syllabe, le contour remonte à 7 demi-tons pour ensuite tomber à 1 ou 0 demi-tons ;
- la forme prosodique commune de l'attitude *sarcastique ironie* (Figure 46) commence avec une valeur initiale de 7 demi-tons, puis monte rapidement à la valeur maximum (15 demi-tons) dans la deuxième syllabe ; Il s'ensuit ensuite une descente graduelle pour finalement décroitre rapidement dans la dernière syllabe.

Ces descriptions nous permettent de confirmer notre hypothèse sur l'existence d'une forme de contour prosodique global correspondant à chaque attitude. Nous fondant sur l'observation de l'évolution de ces contours de F0, en fonction du nombre de syllabes qui les portent, nous établissons qu'il existe une expansion du mouvement du contour de F0 correspondant à la longueur de la phrase (cf. [Morlec et al. 2001]). Ceci nous donne la capacité de styliser et de modéliser le contour fonctionnel de F0 des attitudes en fonction des différentes longueurs des phrases. C'est ce que nous présentons dans la section suivante.

7.3.2.3 Stylisation et modélisation des contours fonctionnels de F0 des attitudes

Comme dans la présentation ci-dessus, nous trouvons que chaque attitude peut être caractérisée par une forme prosodique commune. Cette forme s'allonge en fonction de la longueur de la phrase (c'est-à-dire en fonction du nombre de syllabes). Notre objectif consiste à trouver une méthode pour styliser cette forme prosodique commune d'attitude, afin d'obtenir la capacité de prédire le contour de F0 de l'attitude pour une longueur de phrase donnée.

Observons l'évolution des contours fonctionnels de F0 des trois attitudes (Figure 44, Figure 45 et Figure 46), nous trouvons que les contours de F0 des trois attitudes sont caractérisés par un mouvement initial et un mouvement final. Autrement dit, en comparant différentes attitudes, il semble que le contour de F0 est plus variable en début et en fin de phrase. Au contraire, le mouvement du contour de F0 au milieu de la phase est quasiment invariant : pour toutes les attitudes décrites ici, la partie correspondant au milieu du contour de F0 peut être caractérisée comme une courbe linéaire. De plus, quand la longueur de la phrase augmente, les parties initiale et finale ne sont pas allongées : l'expansion du contour de F0 est principalement réalisée par un allongement linéaire de la partie au milieu du contour.

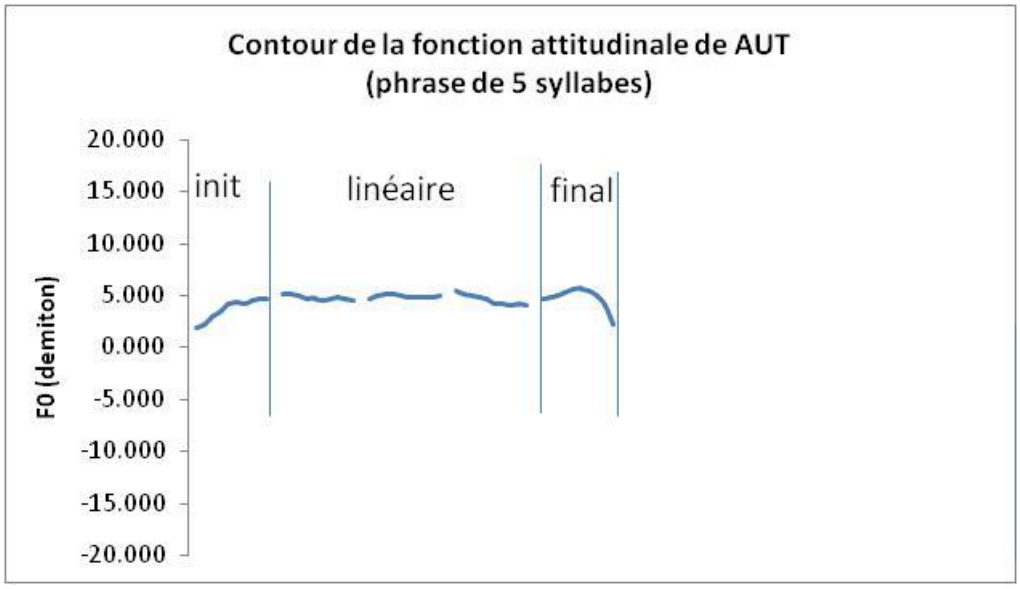


Figure 47: Un exemple des trois parties du contour fonctionnel de F0

En nous basant sur ces dernières remarques, les contours fonctionnels de F0 des attitudes paraissent donc se diviser en trois parties comme dans la Figure 47 :

- la partie initiale commence au début du contour et se termine dans la première ou la deuxième syllabe (dans le cas de l'attitude d'ironie sarcastique, voir la Figure 46) ;
- la partie linéaire, commence dans la première ou la deuxième syllabe et se termine sur la dernière syllabe ; dans cette partie, le contour du F0 varie linéairement et est quasi-stable ;
- la partie finale commence dès que le contour du F0 n'est plus linéaire et se termine en fin de phrase.

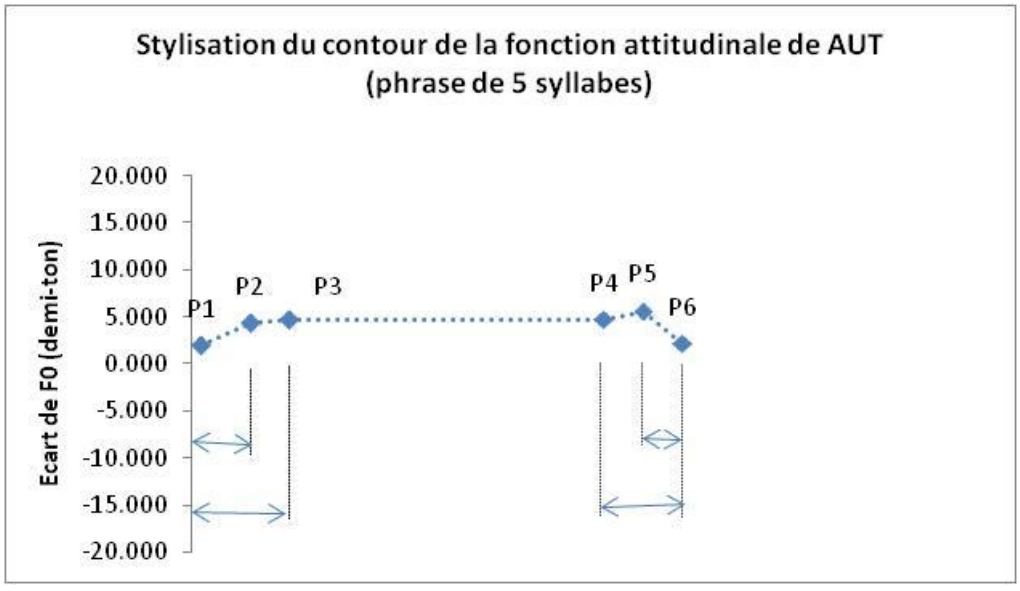


Figure 48: Un exemple de la stylisation du contour fonctionnel de F0 d'attitude

Nous observons aussi que la forme des parties initiale et finale peut être simplement constituée par 2 segments linéaires : un segment montant et un segment descendant. Le contour F0 de la fonction attitudinale peut donc être stylisée par 6 points comme exposé à la Figure 48 :

- le point 1 constitue le point initial du contour ;
- le point 2 constitue le point extrême de la partie initiale. La direction du contour de F0 subit sa plus grande variation (montant / descendant) après ce point ;
- le point 3 est le point final de la partie initiale, après ce point, le contour de F0 est linéaire ;
- le point 4 est le point final de la partie linéaire – ou le point initial de la partie finale. Après ce point, le contour du F0 n'est plus linéaire ;
- le point 5 est le point extrême de la partie finale ; la direction du contour de F0 de la partie finale est le plus fortement modifié (montant / descendant) après ce point ;
- le point 6 est le point final du contour.

Supposons que les valeurs de F0 des 6 points décrits ci-dessus sont représentées par P1, P2, P3, P4, P5 et P6 ; les positions relatives entre ces points sont alors représentées par 4 valeurs : T12, T13, T46, T56.

- T12 est le rapport de la durée entre points 1 et 2 sur la durée moyenne de la première syllabe ;

$$T12 = \frac{\text{durée } (Point1, Point\ 2)}{\text{durée } de\ la\ première\ syllabe}$$

  par exemple, T12= 0.5 quand le point 2 est le milieu de la première syllabe.

- T13 est le rapport de la distance entre les points 1 et point 3 sur la durée moyenne de la première syllabe ;

$$T13 = \frac{durée\ (Point1, Point\ 3)}{durée\ de\ la\ première\ syllabe}$$

  par exemple, T12= 1 quand le point 3 est en fin de la première syllabe

- T46 est le rapport de la distance entre les points 4 et 6 sur la durée moyenne de la dernière syllabe ;

$$T46 = \frac{durée\ (Point1, Point\ 2)}{durée\ de\ la\ dernière\ syllabe}$$

  par exemple, T5= 1 quand le point 4 est au début de la dernière syllabe

- T56 est le rapport de la distance entre les points 5 et 6 sur la durée moyenne de la dernière syllabe ;

$$T56 = \frac{durée\ (Point1, Point\ 2)}{durée\ de\ la\ dernière\ syllabe}$$

La Figure 49 Figure 49 présente un exemple d'utilisation des six points présentés ci-dessus pour styliser les contours fonctionnels de F0 pour l'attitude d'*autorité*.

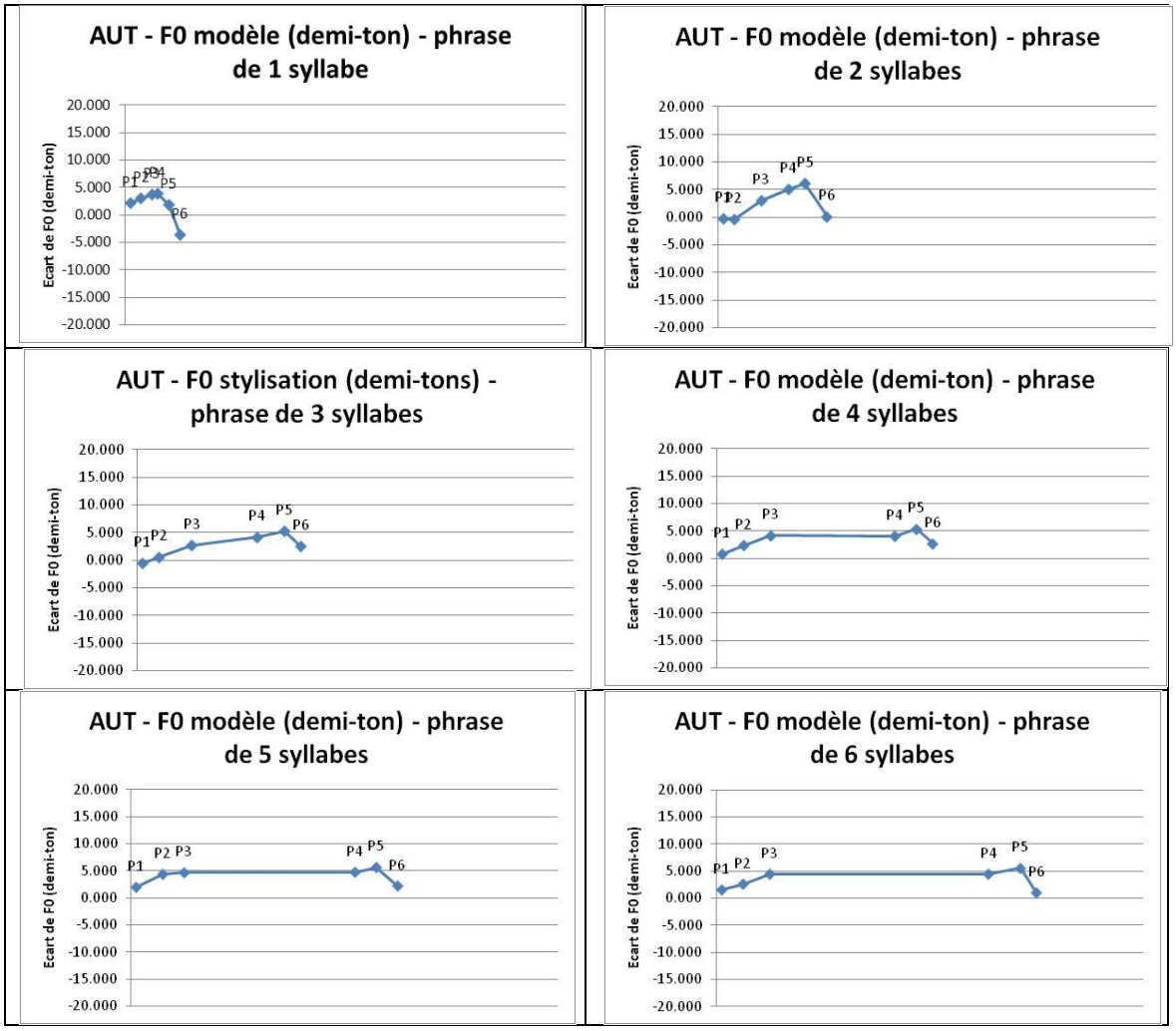


Figure 49: La stylisation des contours de F0 de la fonction attitudinale avec l'attitude d'*autorité*

De l'observation de la variation des 6 points stylisés en fonction de la longueur de phrase, nous trouvons que pour chaque attitude, les valeurs P1, P2, P3, P4, P5, P6 et leurs valeurs de distances relatives T12, T13, T46, T56 sont assez invariantes. Autrement dit, dans l'évolution du contour de F0, en fonction du nombre de syllabes, les formes de la partie initiale et de la partie finale ne sont pas beaucoup modifiées. C'est pourquoi, pour représenter la forme du contour

global de F0 d'une attitude, nous proposons d'utiliser la valeur moyenne pour les six points stylisés comme suit :

$$\overline{\mathrm{P}}(\mathrm{i}) = \frac{\sum_{l=3}^{8} P_l(i)}{6}$$

$$\overline{\mathrm{T}}(\mathrm{ij}) = \frac{\sum_{l=3}^{8} T_l(ij)}{6}$$

où

- $\overline{\mathrm{P}}(\mathrm{i})$ est la valeur F0 moyenne du point i (i = 1 à 6) ;
- $P_{l,i}$ est la valeur de F0 du point i (i = 1 à 6) pour les phrases avec une longueur de l syllabe (l = 3 à 8) ;
- $\overline{\mathrm{T}}(\mathrm{ij})$ est la valeur moyenne de la distance relative entre le point i et le point j ;
- $T_{l,}(ij)$ est la valeur de la distance relative entre le point i et le point j dans la phrase avec une longueur de l syllabe (l = 3 à 8).

Les valeurs moyennes $\overline{\mathrm{P}}$ et $\overline{\mathrm{T}}$ des six points stylisés pour les trois attitudes choisies sont présentées dans la Table 17.

Table 17: Valeurs des points stylisés pour les contours fonctionnels de F0 des trois attitudes

	Longueur de phrase	Valeur moyens des points stylisés									
		F0 en demi-tons (écart type ≤ 0.2)						Distances relatifs (écart type ≤ 0.05			
		$\overline{\mathrm{P}}_1$	$\overline{\mathrm{P}}_2$	$\overline{\mathrm{P}}_3$	$\overline{\mathrm{P}}_4$	$\overline{\mathrm{P}}_5$	$\overline{\mathrm{P}}_6$	$\overline{\mathrm{T}}_{1\text{-}2}$	$\overline{\mathrm{T}}_{1\text{-}3}$	$\overline{\mathrm{T}}_{4\text{-}6}$	$\overline{\mathrm{T}}_{5\text{-}6}$
EXo	1 syllabe	5.6	3.9	4.2	4.8	6.3	6.3	0.2	0.4	0.4	0.2
	2 syllabes	6.8	6.6	7.9	9.1	10.1	9.8	0.3	0.9	0.7	0.1
	> 3 syllabes	7.6	8.8	9.2	11.0	11.7	11.0	0.3	0.9	0.9	0.2
AUT	1 syllabe	2.2	3.0	3.7	3.8	1.8	-3.6	0.2	0.2	0.4	0.2
	2 syllabes	-0.3	-0.4	3.0	5.0	6.0	0.0	0.2	0.7	0.7	0.4
	> 3 syllabes	1.2	2.7	4.2	4.6	5.5	0.9	0.4	0.9	0.8	0.4
SAR	1 syllabe	-0.8	1.5	5.1	5.0	3.7	1.5	0.3	0.6	0.2	0.1
	2 syllabes	1.5	3.2	10.3	12.7	10.6	4.9	0.3	0.9	0.9	0.5
	> 3 syllabes	7.7	9.1	14.1	9.1	9.1	5.5	0.3	1.2	0.9	0.6

Les valeurs moyennes ci-dessus constituent les paramètres de notre modèle pour générer les contours fonctionnels de F0 des attitudes.

7.3.3. Modéliser la durée syllabique

Comme nous l'avons mentionné dans la section 3.2, la première tâche dans la modélisation de la durée est le choix d'une unité de base. Avec les langues isolantes telles que le thaï, le mandarin ou le vietnamien, le modèle de la durée est généralement basé sur la syllabe [Wu et al. 2001 ; Hansakunbuntheung et al. 2007 ; Tran et al. 2007]. En fait, dans notre système de synthèse de la parole en vietnamien [Tran et al. 2007], l'unité de base pour la durée est aussi la syllabe.

C'est pour cette raison, que nous modélisons la durée syllabique des attitudes choisies. Cette tâche est réalisée selon les étapes suivantes :

- pour chaque attitude et pour chaque longueur de phrase, calculer la durée syllabique moyenne, syllabe par syllabe ;
- calculer le rapport de la durée moyenne entre les trois attitudes choisies et l'attitude de *déclaration* (la référence). Ainsi les variations intrinsèques et co-intrinsèques des phonèmes sont implicitement normalisées ;
- modéliser les rapports obtenus.

7.3.3.1 Durée syllabique moyenne

Pour les quatre attitudes choisies et pour chaque longueur de phrase, nous calculons la durée syllabique moyenne, syllabe par syllabe. Un exemple des durées syllabiques moyennes d'une phrase de 5 syllabes est présenté dans la Figure 50. Les résultats obtenus avec les autres longueurs peuvent être consultés dans l'Annexe 3.

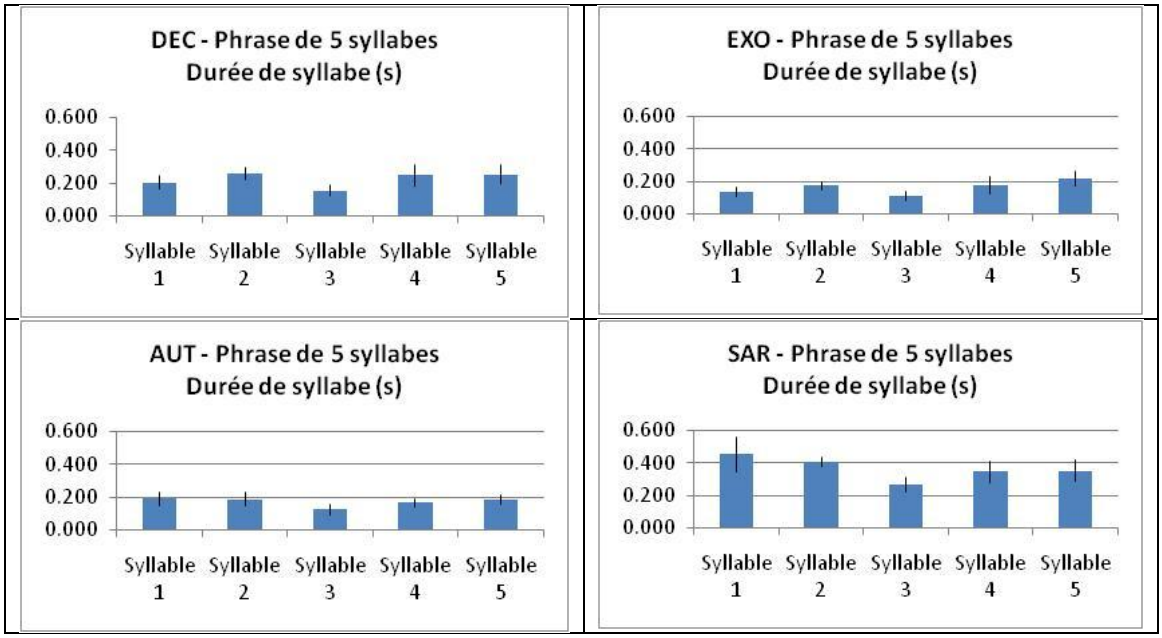


Figure 50: Durées syllabiques moyennes de la phrase de 5 syllabes avec 4 attitudes : déclaration (DEC), surprise neutre (EXo), autorité (AUT), et sarcastique (SAR)

7.3.3.2 Modélisation du rapport de la durée syllabique moyenne

Comme nous l'avons présenté dans les paragraphes précédents, notre objectif pour modéliser la fonction attitudinale de la prosodie consiste à trouver la différence entre la prosodie de la phrase portant l'attitude considérée et celle de la phrase d'attitude « neutre » (ou la phrase de déclaration) qui sert de référence. Pour la durée, cette différence est estimée comme le rapport de la durée syllabique moyenne entre l'une des trois attitudes choisies et la déclaration, comme présenté dans les figures suivantes.

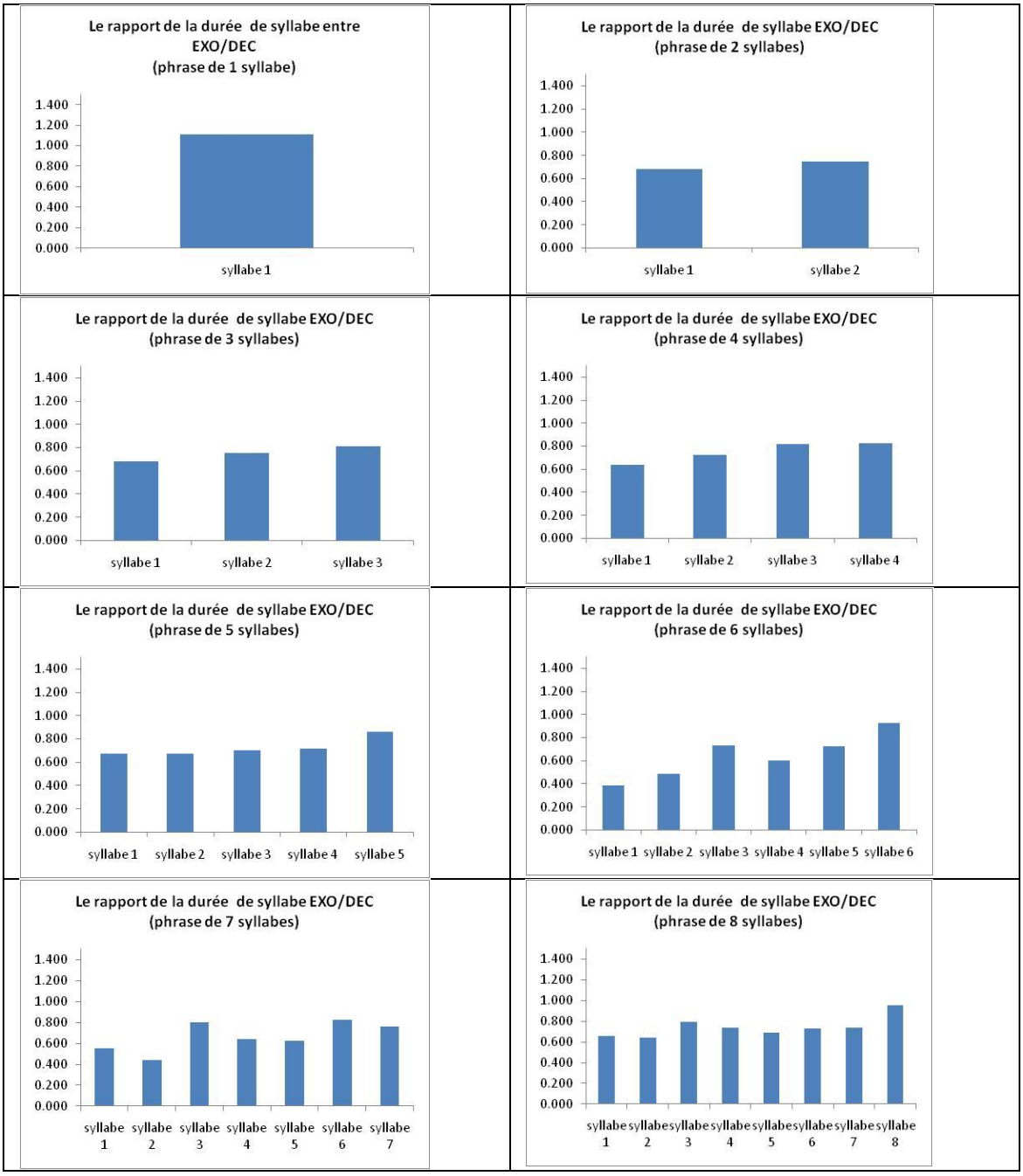


Figure 51: Rapports de la durée syllabique entre la surprise neutre (EXo) et la déclaration (DEC)

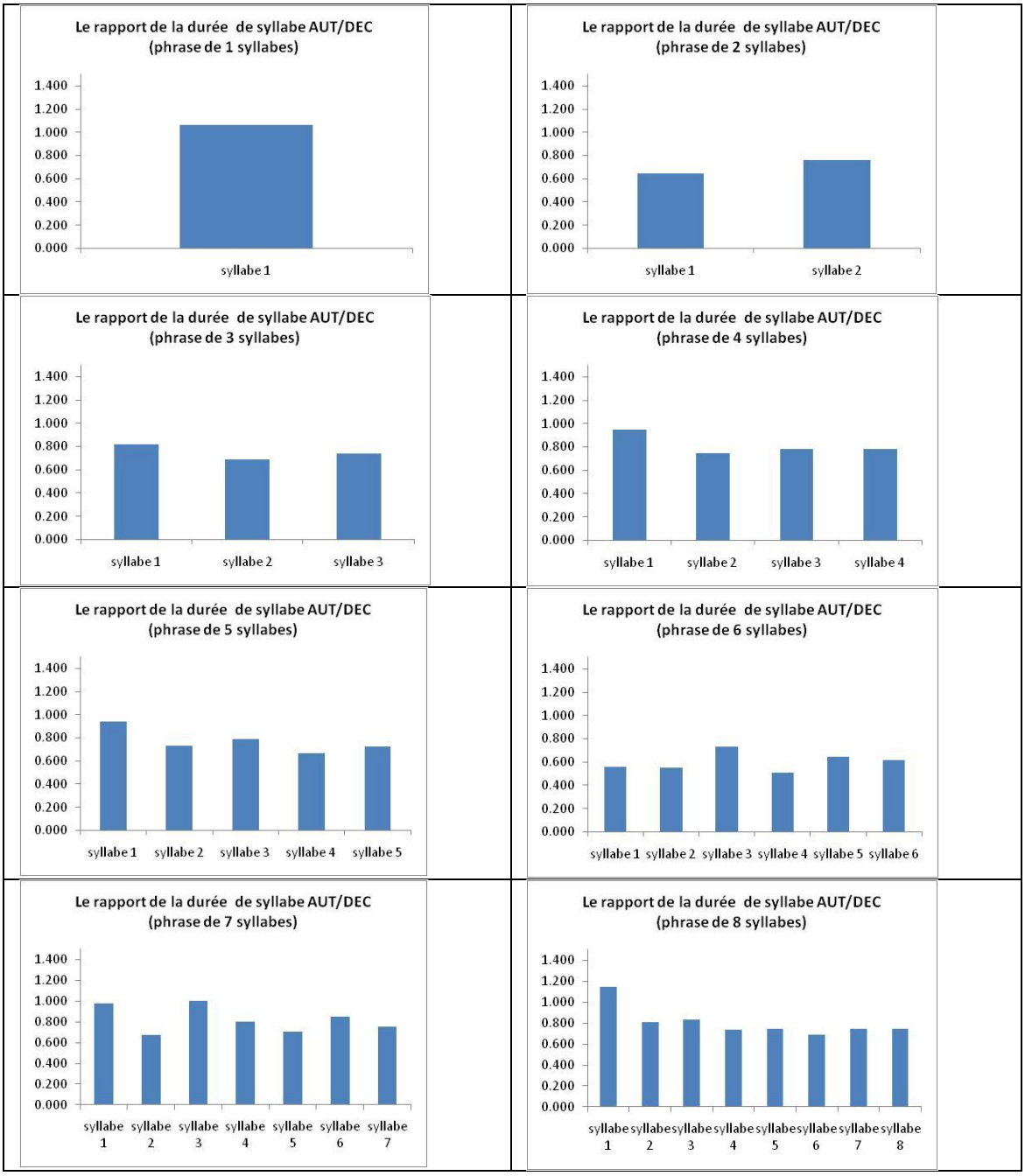


Figure 52: Rapports de la durée syllabique entre autorité (AUT) et la déclaration (DEC)

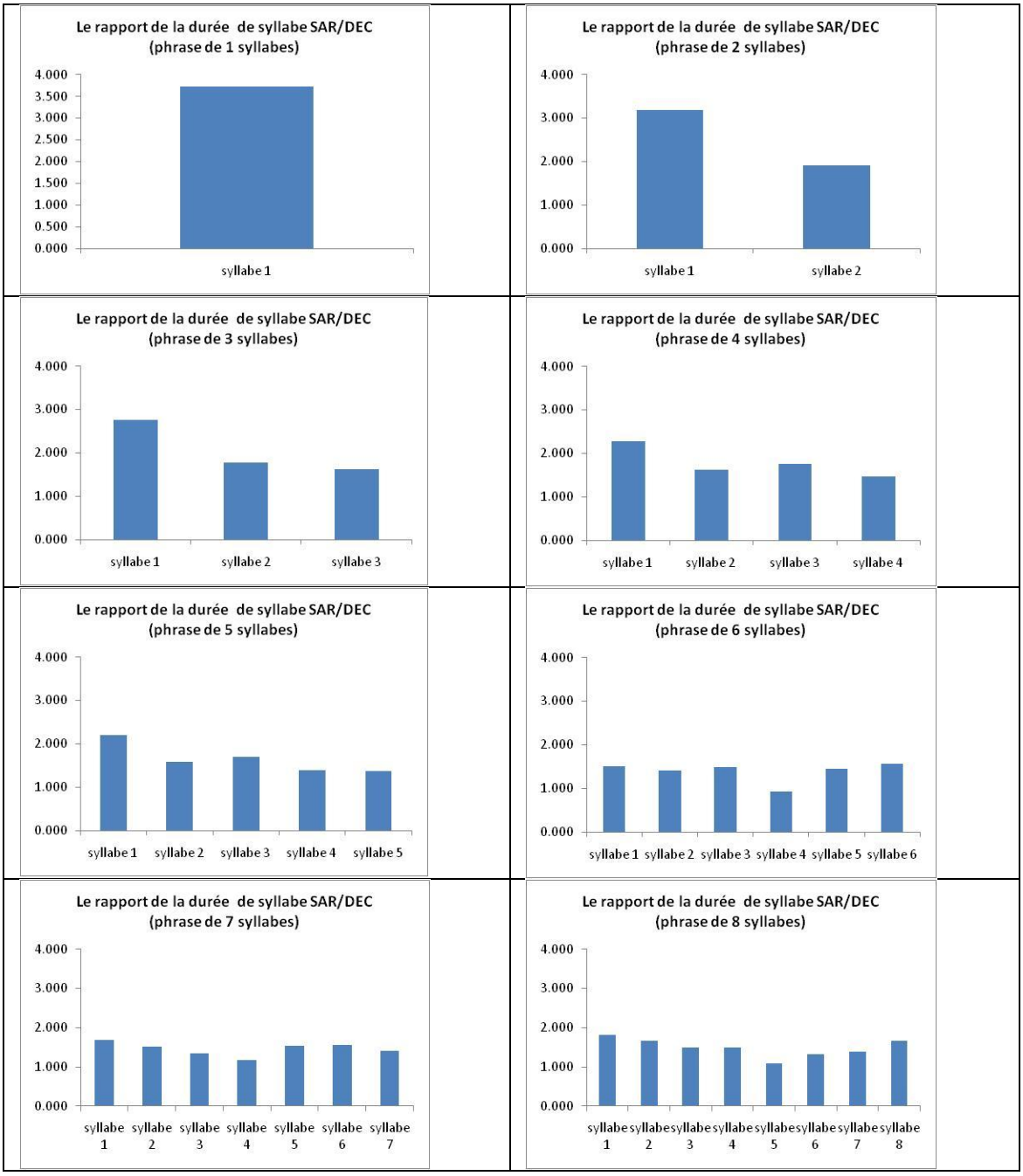


Figure 53: Rapports de la durée syllabique entre l'ironie sarcastique (SAR) et la déclaration (DEC)

En nous basant sur l'observation des rapports de la durée syllabique entre les trois attitudes et la référence (la déclaration), nous pouvons faire les remarques suivantes :

- globalement, la surprise neutre et l'autorité ont une durée syllabique plus courte que la référence (rapport < 1) ; au contraire, la durée syllabique de l'attitude d'*ironie sarcastique* est beaucoup plus longue que celle de la référence (rapport > 1) ;
- la *surprise neutre* est caractérisée par un allongement de la dernière syllabe, tandis que l'*autorité* et l'attitude d'*ironie sarcastique* sont caractérisées par un allongement de la première syllabe ;

- pour les phrases dont la longueur est supérieure ou égale à trois syllabes, les rapports de la durée des syllabes du milieu de la phrase sont assez invariants. Mais pour une longueur de plus de six syllabes, le changement des rapports de la durée syllabique est plus complexe : nous pouvons remarquer des syllabes avec une durée très courte (par rapport aux autres syllabes dans la phrase, comme dans le cas de l'*autorité* pour des phrases de 6 et 7 syllabes), ou très longue (dans le cas de l'attitude *sarcastique ironie*, avec des phrases de 6, 7 et 8 syllabes). Nous supposons que cette variabilité provient de la structure syntaxique de la phrase longue. Nous rappelons que dans notre corpus, la longueur maximale de la phrase sans tons et mono-mot est de 5 syllabes, les autres phrases présentant une structure S-O-V, la variation syntaxique pouvant justifier cette variabilité dans la variation de la durée.

A partir de ces remarques, nous supposons que pour une phrase présentant une syntaxe simple, les variations de la durée syllabique sont observées principalement dans les première et dernière syllabes. Des observations similaires sont faites dans d'autres travaux sur les attitudes en français [Aubergé et al. 1997 ; Morlec et al. 1997a] et en japonais [Shochi 2008]. C'est pourquoi, nous proposons de modéliser la durée syllabique de l'attitude par les valeurs moyennes du rapport de la durée syllabique entre la première syllabe, la dernière syllabe et une syllabe du milieu de la phrase.

Table 18: Valeurs moyennes du rapport de la durée syllabique entre les trois attitudes choisie et la déclaration

	Longueur de phrase	Valeur moyennes du rapport de la durée syllabique (écarts types ≤ 0.1)		
		première syllabe	Syllabe du milieu	dernière syllabe
EXo	1 syllabe	1.1		
	2 syllabes	0.7		0.8
	≥ 3 syllabes	0.6	0.7	0.9
AUT	1 syllabe	1.1		
	2 syllabes	0.7		0.8
	≥ 3 syllabes	0.9	0.7	0.7
SAR	1 syllabe	3.7		
	2 syllabes	3.2		1.9
	≥ 3 syllabes	2.0	1.5	1.5

La Table 18 ci-dessus présente les valeurs moyennes des rapports de la durée syllabique pour les attitudes considérées. Ces valeurs seront les paramètres de notre modèle pour générer la durée des attitudes dans la section suivante.

7.3.4. Modéliser l'intensité moyenne

Comme nous l'avons mentionné dans le Chapitre 2, parmi les trois paramètres classiques de la prosodie, l'intensité est la moins étudiée. Dans les signaux de parole continue, l'intensité peut être représentée par un contour continu comme la fréquence fondamentale. Cependant, les variations du contour d'intensité sont complexes, car elles dépendent de la structure phonétique (voisé ou non voisé,

consonne ou voyelle, etc.) mais aussi d'autres paramètres prosodiques tels que la qualité de voix et la fréquence fondamentale [Strik et al. 1992] et aussi de la distance au micro d'enregistrement. C'est pourquoi, dans la synthèse de la parole, l'intensité ne peut pas être modélisée et générée par un contour simple et indépendant comme celui de la fréquence fondamentale [Benesty 2008]. Dans la synthèse de la parole expressive, l'intensité d'une attitude ou d'une émotion est habituellement obtenue par modification de la valeur d'intensité moyenne de la phrase ou de l'unité de base. Par exemple, dans la Table 1, l'intensité de la phrase pour la colère peut s'obtenir en augmentant l'intensité de 6 dB [Schröder 2004a ; Schröder et al. 2011]. De plus, pour une langue isolante comme le vietnamien, l'unité de base de durée est la syllabe [Doan 1997 ; Do et al. 1998]. Nous faisons donc le choix de modéliser l'intensité des attitudes vietnamiennes par un contrôle de la valeur d'intensité moyenne syllabique.

Selon notre méthodologie proposée dans la section 7.3.1, nous calculons et modélisons la différence d'intensité syllabique entre les trois attitudes (EX0, AUT et SAR) et la référence (la déclaration). Tout d'abord, afin de calculer la valeur de l'intensité moyenne pour chaque syllabe, pour chaque attitude choisie, et pour chaque longueur de phrase, les valeurs d'intensités sont extraites sur la portion de signal restreinte à la voyelle et exprimées en décibels. Puis, les différences d'intensité moyenne entre les trois attitudes choisies et l'attitude déclaration (la référence) sont calculées pour chaque longueur de phrase.

La Figure 54 présente un exemple des différences d'intensité syllabique moyenne entre les trois attitudes et la déclaration, pour une phrase de 5 syllabes. Nos résultats obtenus avec les autres longueurs de phrase peuvent être consultés dans l'Annexe 4.

A partir de l'observation des différences de l'intensité syllabique moyenne entre les trois attitudes considérées et la déclaration, nous trouvons que :

- globalement, l'intensité moyenne des trois attitudes choisies est supérieure à celle de la référence ;

- De même que dans l'observation de la durée syllabique, nous constatons que pour les phrases présentant une longueur supérieure ou égale à trois syllabes, l'intensité syllabique moyenne du milieu de la phrase est assez invariante ; autrement dit, l'intensité des trois attitudes peut se caractériser principalement dans les première et dernière syllabes.

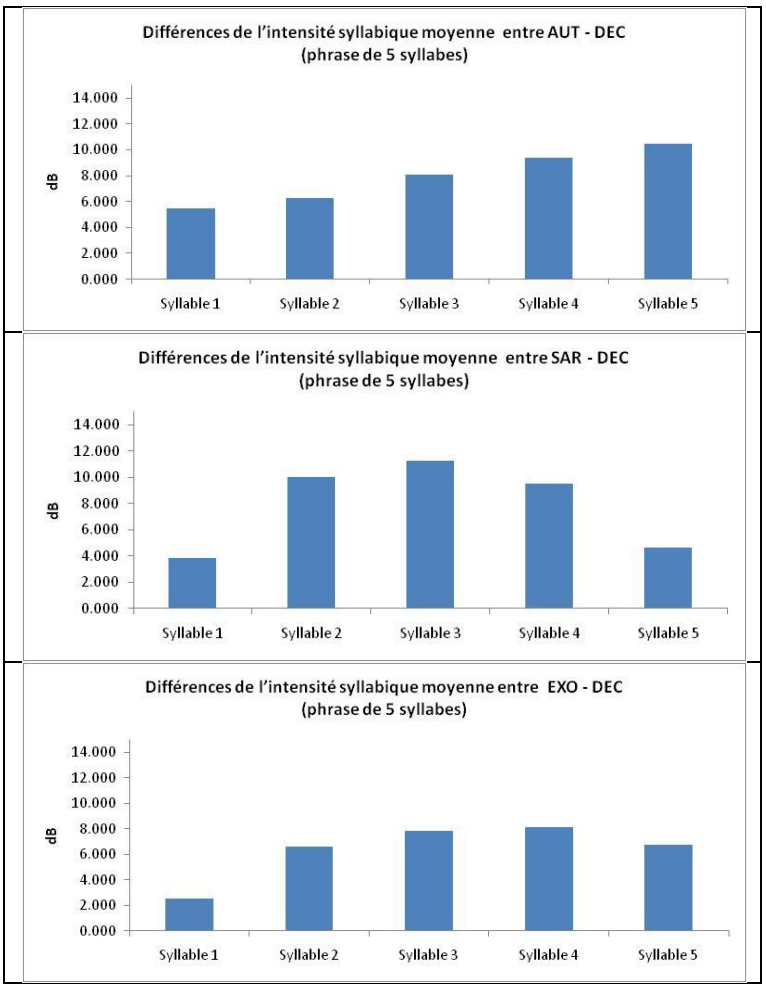


Figure 54: Différences de l'intensité syllabique moyenne entre les trois attitudes et la déclaration pour une phrase de 5 syllabes

C'est pourquoi, de la même manière qu'avec la durée, nous proposons de modéliser l'intensité syllabique par trois valeurs moyennes définies pour la première syllabe, la dernière syllabe et une syllabe du milieu de la phrase, comme résumé dans la Table 19. Ces valeurs sont utilisées pour les paramètres de notre modèle pour contrôler l'intensité dans la synthèse de la parole.

Table 19: Les valeurs du modèle de l'intensité syllabique moyenne

	Longueur de phrase	Les valeurs de la différence de l'intensité syllabique moyenne (dB) par rapport à la déclaration (écarts types ≤ 0.06)		
		première syllabe	syllabe du milieu	dernière syllabe
EXo	1 syllabe	-1.4		
	2 syllabes	2.0		2.8
	≥ 3 syllabes	1.6	6.9	6.3
AUT	1 syllabe	7.1		
	2 syllabes	1.9		6.8
	≥ 3 syllabes	4.3	7.0	9.6
SAR	1 syllabe	-1.2		
	2 syllabes	1.3		3.0
	≥ 3 syllabes	3.2	8.0	4.6

7.4. Application à la synthèse de la parole expressive

7.4.1. Génération de la prosodie expressive

Comme nous l'avons mentionné, les systèmes de synthèse de la parole (y compris le système de l'Institut MICA) génèrent habituellement une parole synthétique fondée sur l'attitude de déclaration. Notre approche va consister à modifier cette prosodie de *déclaration* pour atteindre une prosodie proche de celle de l'attitude considérée.

Dans la section précédente, nous avons présenté notre modèle fonctionnel de la prosodie des attitudes. Ce modèle sera utilisé pour la génération de notre parole synthétique. Pour transformer une phrase d'attitude déclarative vers l'attitude considérée A, nous pouvons procéder comme suit :

- extraire les paramètres prosodiques de la phrase de déclaration (contour de F0, la durée syllabique et l'intensité syllabique moyenne), qui sont générés par le système de synthèse de la parole
- sur la base du modèle proposé, calculer la différence entre la prosodie de la *déclaration* et la prosodie d'attitude A
    - ✓ le contour fonctionnel de F0 de l'attitude A (ou la différence entre le contour F0 de la *déclaration* et de l'attitude A) peut être prédit en appliquant les points stylisés appropriés extraits de la Table 17. Une interpolation linéaire est employée pour réaliser la transition entre les points considérés
    - ✓ le rapport de la durée syllabique entre la *déclaration* et l'attitude A peut être obtenu à partir de la Table 18
    - ✓ la différence de l'intensité syllabique moyenne entre la déclaration et l'attitude A peut être calculée en utilisant la Table 19
- la prosodie finale de la phrase pour l'attitude A (contour de F0, la durée syllabique, l'intensité syllabique moyenne) est alors obtenue par superposition additive de la prosodie de la *déclaration* et des patrons de différences prosodiques ainsi obtenus.

Pour valider notre approche, nous allons appliquer les étapes présentées dans le paragraphe précédent dans deux types de synthèse :

- pour la resynthèse : l'idée consiste à simplement modifier la prosodie d'une phrase de parole naturelle enregistrée. Cette méthode nous permet d'évaluer la prosodie générée d'une manière indépendante de la qualité de la voix
- puis, en utilisant le système de synthèse de la parole en vietnamien de l'Institut MICA [Tran 2007] pour générer la parole avec des attitudes fondée sur notre modèle proposé. Cette seconde étape permettra d'évaluer l'intégralité de la synthèse de la parole expressive en vietnamien dans notre système.

7.4.2. Première étape de re-synthèse

La resynthèse nous permet de générer les stimuli synthétiques à partir d'énoncés de parole naturelle enregistrée. Ceci est réalisé à l'aide du module de manipulation du logiciel Praat [Boersma et al. 2011]. Ce module permet de modifier les contours de fréquence fondamentale ainsi que la durée de tout ou partie d'énoncés naturels en s'appuyant sur l'algorithme TD-PSOLA, ainsi que de modifier leurs contours d'intensité avant de générer les fichiers sons ainsi modifiés.

Tout d'abord, les paramètres prosodiques des énoncés enregistrés sont extraits semi automatiquement avec des scripts sous Praat :

- le contour de F0 est extrait en demi-tons, puis normalisé sur 10 points pour chaque syllabe
- la durée syllabique est calculée en millisecondes, syllabe par syllabe
- l'intensité est calculée en dB.

Puis, les valeurs du contour fonctionnel des attitudes sont calculées, et sont superposées avec les paramètres prosodiques des énoncés source (énoncés enregistrés) pour obtenir la prosodie de l'énoncé cible avec l'attitude choisie :

- le contour F0 de l'énoncé source est superposé au contour fonctionnel de F0 de l'attitude par une simple addition ;
- la durée syllabique de l'énoncé source est multipliée par le rapport de la durée syllabique, qui est calculé à partir de la Table 18 ;
- l'intensité de l'énoncé source est superposé à la différence d'intensité syllabique moyenne, qui est calculée à partir de la Table 19.

Au final, la prosodie de l'énoncé cible obtenu est utilisée pour remplacer la prosodie de l'énoncé source dans l'algorithme de resynthèse (qui s'appuie sur l'algorithme TD-PSOLA), afin d'obtenir l'énoncé synthétique avec la prosodie de l'attitude correspondante.

Un exemple de la génération de la prosodie de l'attitude d'*ironie sarcastique* pour une phrase avec tons est illustré par la Figure 55.

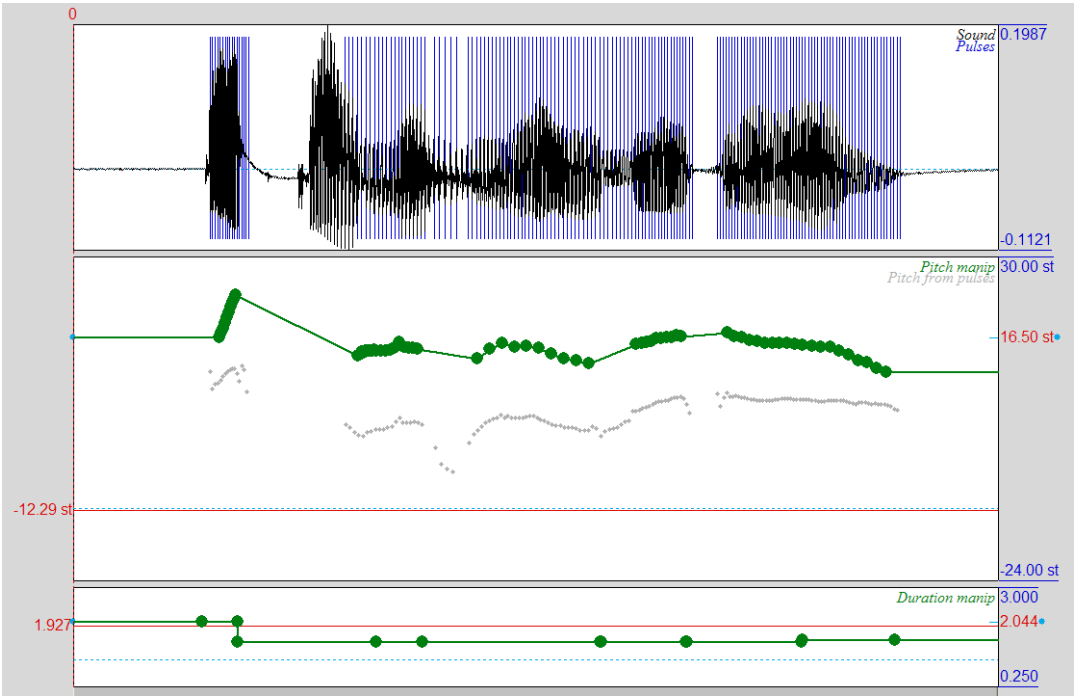


Figure 55 : Génération de la prosodie de l'attitude d'*ironie sarcastique* pour la phrase « Tất cả mọi người đi theo anh » (la séquence des tons : 5b-4-3-2-1-1-1) ; le contour original de F0 de l'expression neutre apparaît en gris, le contour final est dessiné en vert.

La Figure 56 montre un exemple du contours prosodique correspondant à l'attitude d'*ironie sarcastique* de l'énoncé réel et l'énoncé de resynthèse sur la phrase avec tons «Tất cả mọi người đi theo anh ». Nous pouvons observer dans cette figure que le contour de F0 prédit est assez proche du contour réel. Ce contour prédit est ensuite évalué grâce au test de perception présenté dans la section suivante.

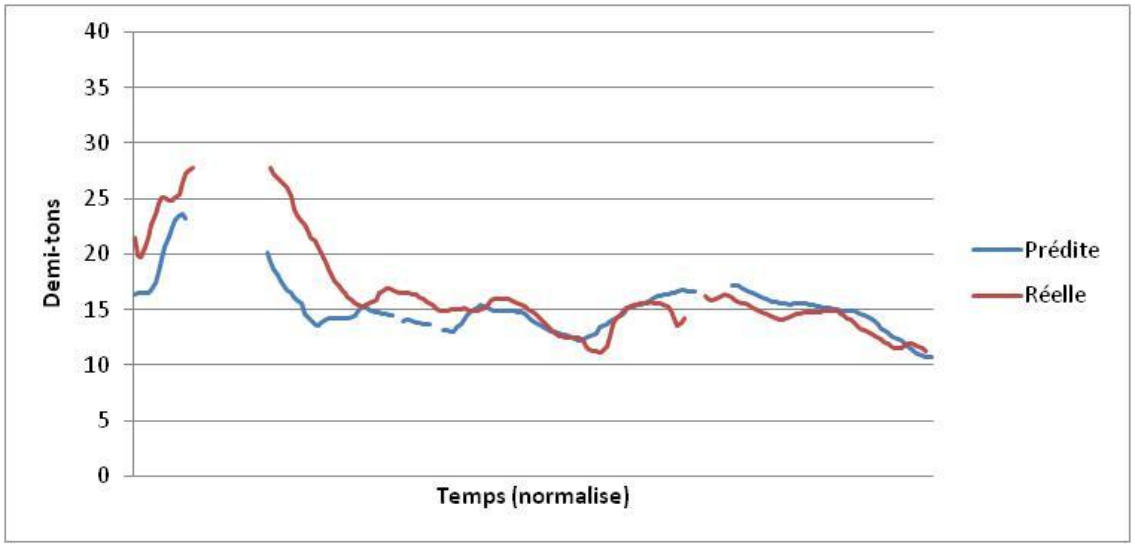


Figure 56: Contours prosodiques réel et prédit de la phrase «Tất cả mọi người đi theo anh » - «Tout le monde te suit »

7.4.3. Application au système de la synthèse de MICA

Comme nous l'avons présenté, le système de synthèse de la parole de l'Institut MICA, développé par Tran [Tran 2007], est considéré comme un des premiers système de synthèse de la parole de haute qualité en vietnamien. Ce système utilise la méthode de synthèse par concaténation basée sur la demi-syllabe. Dans ce système, Tran a proposé aussi une nouvelle méthode permettant de modéliser et de produire la prosodie du vietnamien en parole continue. Cette méthode est constituée par trois modèles : le modèle dynamique de ton, le modèle de durée et le modèle de registre relatif de l'intonation. Une évaluation montre que ces méthodes peuvent fournir des contours de F0 assez proches des contours réels [Tran 2007].

La Figure 57 présente l'intégration de notre modèle dans le système de synthèse de l'Institut MICA pour générer la parole en tenant compte de l'attitude. Nous rappelons que la description détaillée de la structure du système de synthèse de MICA peut être trouvée à la section 4.3.2.

Au niveau de la sortie du module de génération de la prosodie, nous trouvons les paramètres prosodiques correspondant à l'attitude *déclaration*, la référence dans notre modèle. Notre approche consiste à ajouter un module pour modifier la prosodie de la déclaration et la transformer en prosodie de l'attitude correspondante. À partir des informations sur la longueur de la phrase (en nombre de syllabes), les valeurs de la prosodie fonctionnelle de l'attitude (ou la différence entre la prosodie de la déclaration et celle de l'attitude) sont estimées en utilisant les Table 17, Table 18 et Table 19. Puis, la prosodie fonctionnelle d'attitude est superposée avec la prosodie de la *déclaration* pour obtenir les paramètres de la prosodie finale correspondant à l'attitude choisie. Ces paramètres sont utilisés dans le module de la concaténation (en appliquant l'algorithme TD-PSOLA) pour produire l'énoncé synthétisé expressif.

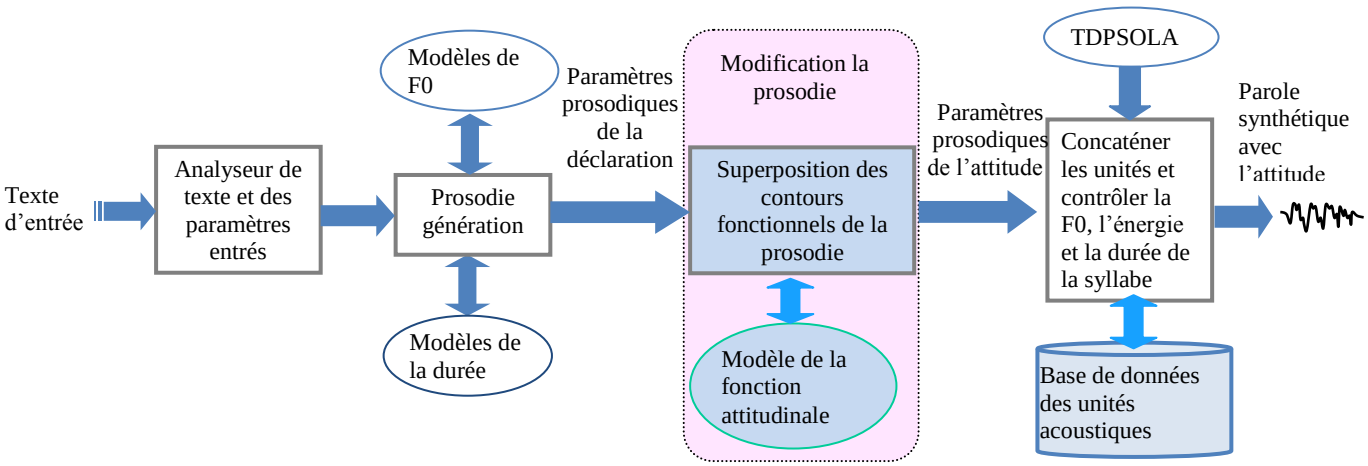


Figure 57: Application du modèle de superposition des contours fonctionnels des attitudes dans système de synthèse de MICA pour générer la parole expressive

La Figure 58 présente l'interface du système de synthèse de la parole à partir du texte de l'Institut MICA. Cette interface nous permet d'entrer directement le texte à synthétiser ainsi que de modifier les paramètres prosodiques de l'énoncé. Dans la première phase d'application de notre modèle dans ce système, la phase de modification de la prosodie est réalisée semi-automatiquement à l'aide de cette interface. Après la validation des performances de notre modèle, notre prochain travail consistera à intégrer le module de modification de la prosodie dans ce système, pour permettre de générer automatiquement la parole expressive à partir du texte.

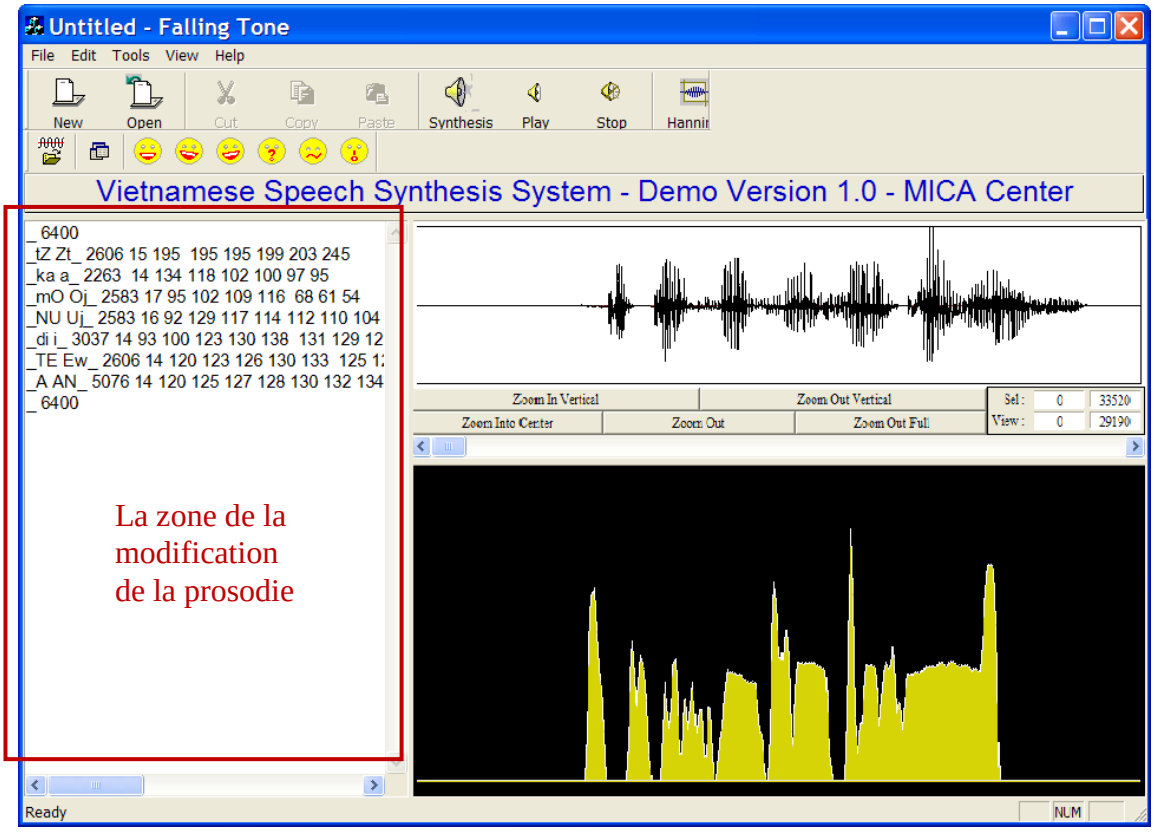


Figure 58: l'interface du système de synthèse de la parole de l'Institut MICA. La zone rouge (à gauche) concerne la modification de la prosodie

7.4.4. Validation par tests de perception

Dans les sections précédentes, nous avons présenté notre modèle de la fonction attitudinale de la prosodie et son application à la génération de parole expressive. Nous présentons maintenant la validation de notre modèle par des tests de perception.

7.4.4.1 Méthodologie expérimentale

Les tests de perception sont réalisés afin de valider notre modèle avec les objectifs suivants :

- évaluation de la qualité de la prosodie de l'attitude prédite : est-ce que les prédictions ont capté suffisamment d'informations prosodiques pour permettre aux auditeurs d'identifier l'attitude du locuteur ?
- évaluation de notre hypothèse concernant l'expansion du mouvement prosodique : est-ce que le modèle permet de générer correctement le contour prosodique en fonction des différentes longueurs de phrase (en nombre de syllabes) ?
- évaluation de notre hypothèse sur la superposition des contours fonctionnels et des contours du ton : est-ce que notre modèle peut être appliqué pour générer le contour prosodique de l'attitude sur une phrase avec des variations de tons ?

Pour atteindre ces objectifs de validation, nous choisissons 4 phrases (cf. la Table 20). Parmi ces phrases, on trouve :

- trois phrases sans ton, avec une longueur de 3, 5 et 8 syllabes pour valider le modèle avec des longueurs différentes
- une phrase de 7 syllabes avec tons pour examiner notre hypothèse de superposition des contours du ton et des contours fonctionnels des attitudes.

Table 20: Les phrases choisies pour le test perceptif de validation du modèle

No.	Nombre de syllabes	Série de tons	Vietnamien	Français
1	3	1_1_1	Anh em ta	Nous deux (toi et moi)
2	5	1_1_1_1_1	Em đang đi theo anh	Tu es en train de me suivre
3	8	1_1_1_1_1_1_1_1	Hai anh em kia đang đi theo anh	Les deux sont en train de me (te) suivre
4	7	5b_4_3_2_1_1_1	Tất cả mọi người đi theo anh	Tout le monde te suit

Les phrases ci-dessus sont générées de deux manières différentes : en utilisant la resynthèse à partir de la phrase naturelle déclarative, et en utilisant la synthèse du système de l'Institut MICA, avec 4 attitudes : *déclaration* (DEC), *exclamation de surprise neutre* (EXo), *autorité* (AUT), *ironie sarcastique* (SAR). Nous utilisons aussi des énoncés naturels comme référence dans ce test afin de comparer les résultats obtenus entre des énoncés synthétiques et des énoncés naturels.

Les tests de perception sont donc constitués de 4 phrases * (2 types de synthèse + 1 naturel) * 4 attitudes = 48 stimuli.

Vingt auditeurs vietnamiens (10 hommes et 10 femmes avec un âge moyen de 24 ans) ont participé à cette expérience. Les tests de perception ont été effectués dans une pièce calme. L'interface donne l'étiquette et l'explication des 4 attitudes. Tous les auditeurs ont écouté chaque stimulus une seule fois. Après l'écoute, on leur demande d'indiquer l'attitude qu'ils ont perçu parmi les 4 attitudes et d'indiquer un score de la qualité de la prosodie allant de « très mauvais » (codé

comme 1) à « très naturel » (codé 100). Un score de 0 a été attribué automatiquement aux 3 autres attitudes non sélectionnées.

7.4.4.2 Les résultats

La Figure 59 présente les scores moyens des phrases naturelles, des phrases sans tons et des phrases avec tons, qui sont générées par resynthèse et par le système de l'Institut MICA. Les scores moyens des phrases sans tons sont meilleurs que ceux obtenus par les phrases avec tons (d'environ 10 à 15 %). Nous supposons que c'est à cause de la complexité du système des tons vietnamiens dans la parole continue. Cependant, la qualité de la prosodie d'attitudes pour les phrases avec tons est quand même bien évaluée (le score moyen est d'environ 63 sur 100). Ceci montre que notre méthode peut être utilisée pour générer la prosodie des phases expressives avec tons. Ceci nous permet aussi de valider notre hypothèse de superposition des contours de ton et des contours fonctionnels d'attitude.

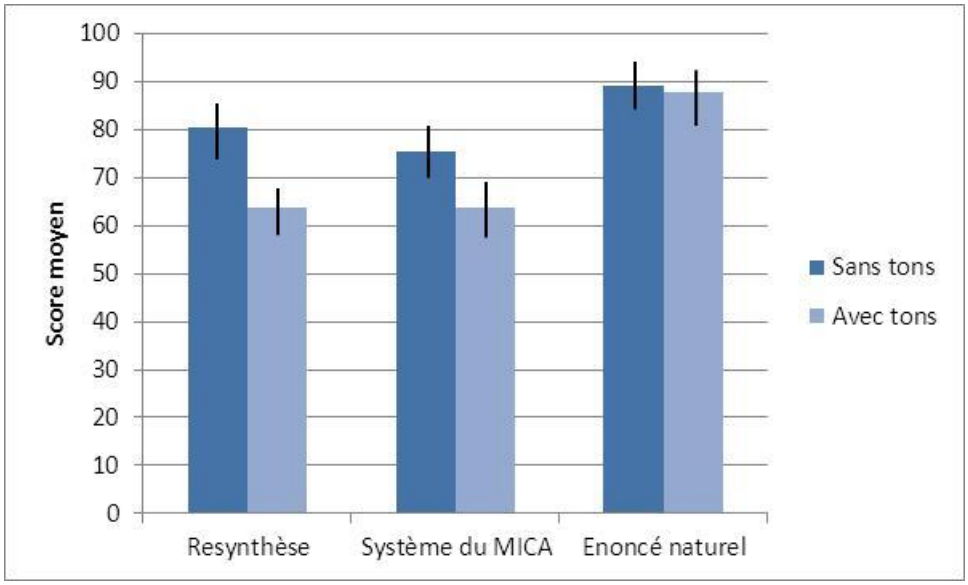


Figure 59 : Les scores moyens de l'énoncé sans tons et des énoncés avec tons, qui sont générés par resynthèse et par le système de l'Institut MICA

La figure ci-dessus montre aussi que les scores moyens des énoncés générés par le système de synthèse de l'Institut MICA ne sont pas très différents des scores des phrases produites par resynthèse. Ce résultat valide la capacité de synthèse de la parole expressive de notre système, en appliquant le modèle proposé pour la génération de la prosodie des attitudes.

La Figure 60 illustre le score moyen obtenu pour les quatre attitudes avec des longueurs différentes de phrases. Globalement, la qualité de la prosodie prédite pour les quatre attitudes est bien évaluée par les auditeurs. L'attitude *d'ironie sarcastique* présente le meilleur résultat (environ 90), tandis que l'*autorité* montre le moins bon score. L'autorité est aussi l'attitude qui reçoit les résultats les plus variables entre les phrases de longueurs différentes. Il est possible ce type

d'attitude soit « sensible » au nombre de syllabes. Un énoncé plus long pourra donc avoir tendance à être moins bien identifié.

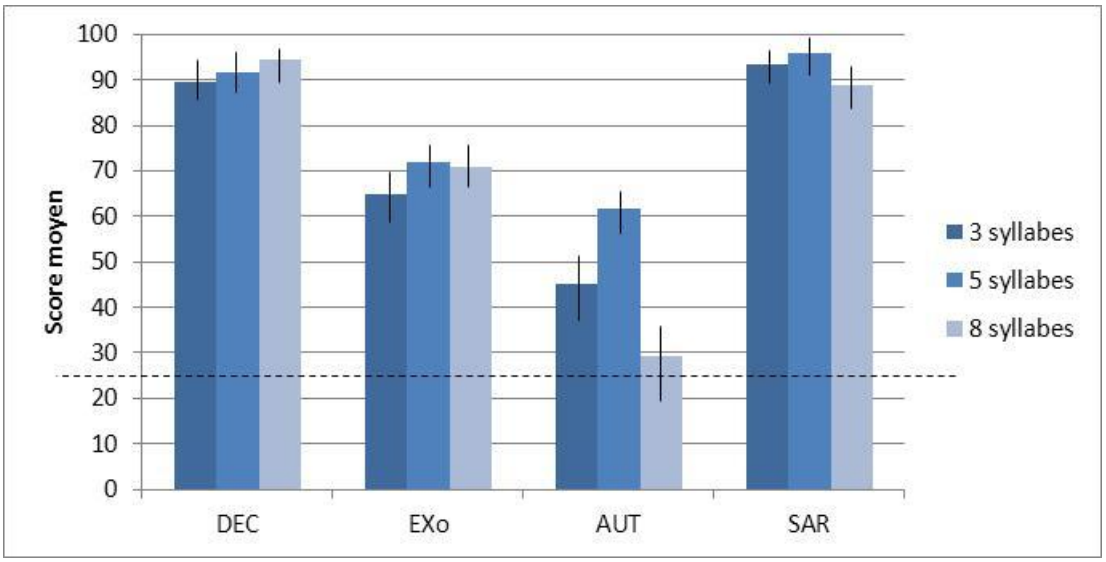


Figure 60: Les scores pour les 4 attitudes avec les longueurs différentes. La ligne pointillée indique le seuil du hasard (25)

Selon cette figure, nous constatons que les scores des phrases de longueurs différentes (en nombre de syllabes) ne sont généralement pas très différents (< 10 %) – sauf dans le cas de l'autorité, dont le cas méritera une analyse plus détaillée. Ceci confirme que notre modèle peut être appliqué pour générer des contours prosodiques de phrase avec des longueurs différentes. Autrement dit, ce résultat vérifie l'existence de l'expansion du mouvement du contour prosodique, ce qui est aussi montré dans d'autres travaux sur la génération de la prosodie en français [Morlec 1997 ; Holm 2003].

7.4.5. Conclusion

Dans ce chapitre, nous avons présenté notre méthode de modélisation de la prosodie attitudinale, basée sur le concept des « rendez-vous » structurels entre l'intonation et les niveaux linguistiques/paralinguistiques, et inspiré du modèle de superposition des contours fonctionnels proposé par Aubergé [1991]. Ce modèle a été appliqué avec succès dans la synthèse de la parole expressive pour le français [Morlec 1997]. Dans notre travail, cette approche théorique est appliquée pour générer de la prosodie de la parole expressive en vietnamien, une langue tonale.

Les tests de perception valident bien la performance de notre modèle, ils nous permettent aussi d'affirmer que le modèle de superposition du contour fonctionnel peut aussi être utilisé pour modéliser la prosodie de la parole expressive et cela dans des langues différentes.

Conclusions et perspectives

Tout au long de notre manuscrit nous avons présenté nos travaux sur l'étude de la prosodie des affects sociaux en vietnamien dans le but de pouvoir générer de la parole expressive synthétique en langue vietnamienne. Nous profitons de cette dernière partie pour résumer les résultats et les contributions les plus importantes de notre travail, et aussi pour proposer un certain nombre de pistes pour des recherches à venir.

Conclusions

La première partie de ce manuscrit a présenté les études théoriques sur la parole expressive, la prosodie et la synthèse parole. Dans le but de générer la parole synthétique expressive en langue vietnamienne, nous avons aussi survolé les connaissances qui nous ont paru les plus judicieuses sur la phonologie et la phonétique vietnamiennes (présentés dans le chapitre 4). Après avoir regroupé et analysé l'état de l'art général sur la parole expressive, rappelé les caractéristiques principales de la langue vietnamienne, et détaillé les approches actuelles utilisées en génération de la prosodie, nous avons alors précisé nos objectifs de recherche et les principes appropriés pour notre travail :

- afin de viser une application de synthèse de la parole expressive pour le vietnamien, le type d'expressivité étudié dans notre travail est l'**attitude**, qui est contrôlée volontairement dans l'interaction humain-humain et aussi dans l'interaction homme-machine ;
- pour tenter de résoudre la complexité en génération de la parole expressive d'une langue à tons, nous avons suivi l'approche de la superposition des contours fonctionnels de la prosodie, approche proposée par Aubergé [1993] ; cette approche nous a permis de décomposer la prosodie en unités plus simples, ce qui nous a permis alors de modéliser séparément les attributs linguistiques (les tons) et les attributs extralinguistiques (les attitudes).

Après avoir précisé les objets et les principes de nos travaux de recherche, nous avons présenté dans la deuxième partie du manuscrit les trois principales contributions de notre thèse.

Notre premier travail a consisté en la construction du premier corpus audio-visuel des attitudes vietnamiennes. Seize attitudes ont ainsi été sélectionnées en nous fondant sur la seule étude disponible dans la littérature sur la parole expressive en vietnamien [Le 1989] mais aussi sur d'autres études plus nombreuses concernant les attitudes pour d'autres langues. Le corpus est composé de 125 phrases,

chaque phrase a été produite pour chacune des 16 attitudes et a été enregistrée avec les deux modalités audio-visuelles. Le corpus complet contient donc 2000 stimuli, correspondant après traitement à plus de 90 minutes de signal audio-visuel. Un test de perception a été effectué pour évaluer la capacité du locuteur à transmettre les 16 attitudes. En comparant les résultats de la validation perceptive sur ce corpus vietnamien avec d'autres corpus d'attitudes en français, anglais et japonais, nous avons montré que la plupart des attitudes en vietnamien sont bien identifiées, et en tout cas d'une manière similaire aux autres corpus en d'autres langues.

En utilisant le corpus ainsi enregistré, nous avons ensuite étudié la perception audio-visuelle et interculturelle des attitudes vietnamiennes. Pour cela, une série de tests perceptifs a été effectuée.

Les deux premiers tests portent sur la perception des attitudes vietnamiennes par les auditeurs natifs et non-natifs (les non-natifs étant des auditeurs francophones), en modalité audiovisuelle. Les résultats de ces tests montrent que les facteurs influant sur la perception des attitudes sont l'expression de l'attitude elle-même et la modalité de présentation (audio, visuelle et audio-visuelle). Un point important soulevé par ces tests est que la longueur de la phrase (en nombre de syllabes) n'influence pas d'une manière notable la perception des attitudes. Ces résultats nous ont aussi permis de trouver des affects sociaux communs (ou interculturels) entre le vietnamien et le français : la *déclaration*, *l'exclamation de surprise positive*, le *doute-incrédulité*, l'*autorité*, l'*irritation* et la *séduction*. Il y a cependant des attitudes montrant des réalisations très spécifiques qui ne sont pas partagées entre les deux langues, comme l'attitude *maternelle* et l'attitude *familière*.

Un troisième test de perception a été réalisé sur des phrases avec tons afin d'explorer l'effet du système tonal vietnamien sur la perception des attitudes par des auditeurs non-natifs. Nous avons noté comme résultat important de ce test que les tons n'ont pas directement d'influence significative sur le comportement perceptif, même avec des auditeurs non-natifs. Nous supposons pour l'expliquer que les auditeurs peuvent séparer l'information locale du ton et l'information globale de l'attitude : ces deux fonctions de la prosodie (fonction tonale et fonction attitudinale) seraient indépendantes dans la production et dans la perception.

Cette hypothèse nous a alors permis de modéliser la prosodie complexe de la parole expressive en vietnamien par la combinaison de deux modèles prosodiques indépendants : un modèle pour générer les contours des tons et un modèle pour générer les contours prosodiques globaux des attitudes.

La troisième contribution principale de notre thèse concerne la modélisation de la prosodie des attitudes pour la synthèse de la parole expressive en vietnamien. En nous basant sur l'approche de la superposition des contours fonctionnels de la prosodie [Aubergé 1993], nous avons proposé que la prosodie de la parole

expressive en vietnamien puisse être obtenue par la superposition de deux composants : le contour prosodique de la fonction tonale et le contour prosodique de la fonction attitudinale.

Le contour de la fonction tonale peut être modélisé par un modèle dynamique des tons et par le modèle de registres relatifs des tons dans la parole continue, proposés tous les deux par [Tran 2007]. Pour modéliser le contour de la fonction attitudinale, nous avons analysé les différences prosodiques entre les attitudes proposées et l'attitude neutre (ou l'attitude de *déclaration*). Ces différences sont caractérisées par des valeurs prosodiques moyennes et sont utilisées comme paramètres de notre modèle pour générer la prosodie des attitudes. Ce modèle est ensuite utilisé dans notre système de synthèse de la parole pour générer la parole avec des attitudes en vietnamien.

Les tests de perception réalisés sur de la parole synthétique valident bien la performance de notre modèle (les taux d'identification de toutes les attitudes sont supérieurs à 60%). Ce résultat permet aussi d'avancer que l'approche de superposition du contour fonctionnel peut être utilisée pour modéliser des variations prosodiques complexes comme dans le cas de la parole expressive d'une langue à tons.

Perspectives

Cependant, ce premier ensemble de travaux sur les attitudes pour la génération de parole synthétique expressive en vietnamien, ouvre d'autres pistes à explorer dans le cadre de travaux futurs.

Premièrement, notre corpus audio-visuel des attitudes vietnamiennes doit être étendu. En effet, le présent corpus n'est constitué que de 125 phrases isolées produites avec une structure syntaxique simple. Il nous paraît important d'ajouter à ce corpus d'autres types de phrases avec d'autres longueurs et d'autres structures syntaxiques afin d'étendre notre modèle prosodique à la fonction syntaxique.

L'un des défauts de notre corpus est qu'il n'a été enregistré que par un seul locuteur, d'un seul genre (homme). Ce corpus doit donc aussi être complété par l'enregistrement de plusieurs autres locuteurs, afin d'une part sur le plan de la modélisation de pouvoir généraliser nos résultats, en particulier en les validant sur des locutrices, et d'autre part de pouvoir produire des variantes attitudinales. Ceci nous permettra de modéliser et d'ajouter les « caractéristiques du locuteur » dans la synthèse de la parole expressive. L'extension de ce corpus, qui présentera alors une grande taille avec plusieurs locuteurs, nous permettra aussi d'étudier les attitudes vietnamienne par des méthodes statistiques (par exemple : pour la synthèse statistique paramétrique (voir la section 3.1), ou nous permettra de l'utiliser dans la reconnaissance automatique de la parole expressive.

Pour les études sur la perception interculturelle des attitudes vietnamiennes, nous voulons vérifier les résultats obtenus avec les auditeurs d'autres langues, afin de caractériser les attitudes universelles ou communes à plusieurs langues. Il est aussi intéressant d'effectuer des tests perceptifs des attitudes vietnamiennes sur les phrases avec tons avec les auditeurs d'autres langues tonales telles que le chinois, le thaï ou le laotien, afin de voir si les auditeurs étrangers de langue tonale peuvent aussi séparer les fonctions prosodiques du ton et des attitudes en vietnamien. Les études interculturelles que nous avons présentées ici, et celles que nous nous proposons de faire par la suite, pourront permettre, au-delà des connaissances théoriques apportées, de proposer des méthodes, et des outils grâce à la synthèse, pour l'enseignement de la prosodie langue seconde spécifiquement à la langue maternelle (dans les couples croisés vietnamien et langue comparée).

Pour la modélisation de la prosodie des attitudes vietnamiennes, comme nous l'avons présenté, notre étude actuelle se concentre seulement sur les trois paramètres classiques de la prosodie : la fréquence fondamentale, la durée et l'intensité alors que, certaines études de la littérature proposent la qualité de voix comme une quatrième dimension de la prosodie : cette qualité de voix a été également proposée comme étant un paramètre fondamental pour l'expression des émotions et des attitudes [Campbell et al. 2003b ; Gobl et al. 2003 ; Audibert et al. 2006b]. Il nous semble donc important que nos travaux de recherche futurs s'intéressent à l'analyse et à la modélisation de la qualité de voix des attitudes vietnamiennes, enfin de générer plus complètement la prosodie de la parole expressive.

Pour obtenir à terme une parole expressive englobant toutes les aspects, il sera aussi nécessaire d'étudier la mise en relief des mots jugés comme le plus importants du point de vue du locuteur et qui jouent un rôle de premier ordre dans l'intelligibilité et le naturel de la prosodie. En effet le phénomène se réalisant régulièrement (en français en moyenne tous les sept mots) [Caelen-Haumont 2012], et ciblant à chaque occurrence du processus, un seul mot, ces mots constituent donc les mots-clés structurant le discours, facilitent de manière naturelle le décodage, et améliorent la qualité de la synthèse. De plus la modélisation des attitudes qui concernent l'ensemble de la phrase se combine parfaitement bien avec une modélisation des mots proéminents, processus macro lexical [Caelen-Haumont 2008].

Dans la section 6.1 les résultats de la perception audio-visuelle des attitudes vietnamiennes montrent que certaines attitudes sont principalement reconnues grâce aux informations visuelles (par exemple : *exclamation de surprise positive, doute-incrédulité, irritation, mépris* et *politesse*). C'est pourquoi, nous souhaitons étendre nos travaux sur l'analyse et la modélisation des attitudes vietnamiennes en prenant en compte le canal visuel. Notre objectif sera alors de concevoir un système complet de synthèse audio-visuelle des attitudes vietnamiennes, pour disposer d'un outil performant d'interaction multimodal personne-machine capable d'utiliser avec efficacité les attitudes.

Finalement, avec ce travail, l'approche de la superposition des contours fonctionnels de la prosodie, qui a déjà été appliquée dans la modélisation de la prosodie de certaines langues telle que le français ou l'anglais, a été validée avec le vietnamien, une langue asiatique tonale. Dans le futur, nous souhaitons utiliser cette approche théorique et adapter notre modèle pour générer la prosodie des autres langues asiatiques de notre région comme le khmer (langue non tonale) et le laotien (langue tonale). Elles font partie elles aussi des langues peu dotées, objet de recherche principal de notre Institut MICA, dans le domaine de traitement de la parole.

Publications

1. **Dang-Khoa Mac,** Véronique Aubergé, Albert Rilliard, Eric Castelli **(**2009). "*Audio-Visual prosody of social attitudes in Vietnamese: building and evaluating a tones balanced corpus*", InterSpeech 2009, Brighton, UK, pp 2263-2266

2. **Dang-Khoa Mac,** Véronique Aubergé, Albert Rilliard, Eric Castelli **(**2010). "*Cross-cultural perception of Vietnamese Audio-Visual prosodic attitudes"*, Speech Prosody 2010, Chicago, USA

3. **Dang-Khoa Mac,** Véronique Aubergé, Albert Rilliard, Eric Castelli **(**2010). "*Vietnamese multimodal social affects: how the prosodic attitudes can be recognized and confused"*, International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU 2010), Penang, Malaysia, pp 24-28

4. **Dang-Khoa Mac,** Véronique Aubergé, Albert Rilliard, Eric Castelli **(**2010). "*Perception interculturelle des attitudes audio-visuelles vietnamiennes"*, JEP 2010, Mons, Belgium.

5. **Dang-Khoa Mac,** Eric Castelli, Véronique Aubergé, Albert Rilliard**,** (2011). "*How Vietnamese attitudes can be recognized and confused: Cross-cultural perception and prosodic analysis"*, International Conference on Asian language processing (IALP 2011), Penang, Malaysia, pp 220-223

6. **Dang-Khoa Mac**, Véronique Aubergé, Eric Castelli, Albert Rilliard (2012). *"Can the tones (local function) influence the acoustic perception of the Vietnamese attitudes (global function) for French listeners (non tonal)?"*. International Conference on Speech and Copora (GSCP 2012), Belo Horizonte, Brazil

7. **Dang-Khoa Mac**, Véronique Aubergé, Eric Castelli, Albert Rilliard (2012). *"Local vs. global prosodic cues: effect of tones on attitudinal prosody in cross-perception of Vietnamese by French"*. SpeechProsody 2012, Shanghai, China, pp 222-225

8. **Dang-Khoa Mac**, Eric Castelli, Véronique Aubergé (2012). *"Modeling the prosody of Vietnamese attitudes for expressive speech synthesis"*. Workshop of Spoken Languages Technologies for Under-resourced Languages (SLTU 2012), Cape Town, South Africa, pp 114-118

Bibliographie

Allen J., Hunnicutt M. S. and Klatt D. (1987). *From text to speech: The MITalk system*. 216 pp.

Allport G. W. (1935). *Attitude*. Handbook of Social Psychology. Murchison. Worcester, Clark University Press.

Ansi P. (1973). *Psychoacoustical Terminology*, American National Standards Institute, New York.

Arnold M. B. (1960). *Emotion and personality*, Columbia University Press.

Aubergé V. (1991). *La synthèse de la parole : "des règles aux lexiques"*. Grenoble, Université Pierre Mendès-France. PhD Thesis.

Aubergé V. (1992). *Developing a structured lexicon for synthesis of prosody*. Talking Machines: Theories, Models and Designs. G. Bailly and C. Benoît, 307-321.

Aubergé V. (1993). *Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis*. In proceedings of ESCA Workshop on Prosody, 62-66.

Aubergé V. (2002a). *A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step by Step Model Developed at ICP*. In proceedings of Speech Prosody.

Aubergé V. (2002b). *Prosodie et émotion*. In proceedings of 2e assises nationales du GDR I3 (Information Interaction Intelligence), 263-274.

Aubergé V. (2003). *Expressions, attitudes et expressivité: une architecture cognitive distribuée pour les voies parlées des émotions*. Interfaces Prosodiques.

Aubergé V. (2012). *Attitude vs. emotion: a question of voluntary vs. involuntary control*. In proceedings of International Conference on Speech and Copora (GSCP2012), Belo Horizonte, Brazil.

Aubergé V., Grépillat T. and Rilliard A. (1997). *Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours.* 5th Eurospeech: 871-874.

Aubergé V. and Rilliard A. (2005). *The focus prosody: more than a simple binary function*. In proceedings of Speech Prosody, Lisbon, Portugal, 1373-1376.

Audibert N. (2008). *Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés*. Ingénierie de la Cognition, de la Création et des Apprentissages, Grenoble INP. PhD Thesis.

Audibert N., Aubergé V. and Rilliard A. (2007). *When is the emotional information? A gating experiment for gradient and contours cues*. In proceedings of ICPhS 2007, Saarbrucken, Germany, 2137-2140.

Audibert N., Aubergé V. and Rilliard A. (2008). *Emotions actées vs spontanées: variabilité des compétences perceptives*. Actes des 27e Journées d'Etudes sur la Parole 20: 4p.

Audibert N., Vincent D., Aubergé V. and Rosec O. (2006a). *Evaluation of expressive speech resynthesis*. In proceedings of LREC, Genoa, Italy, 37.

Audibert N., Vincent D., Aubergé V. and Rosec O. (2006b). *Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions*. In proceedings of SpeechProsody, Dresden, Germany, 525–528.

Averill J. R. (1980). *A constructivist view of emotion*. Emotion: Theory, research and experience R. Plutchik and H. Kellerman. New York, Academic Press, 305-339.

Bailly G. and Bartroli A. (2008). *Generating Spanish intonation with a trainable prosodic model*. In proceedings of Speech Prosody, Campinas - Brazil.

Bailly G. and Gorisch J. (2006). *Generating German Intonation with a Trainable Prosodic Model*. In proceedings of INTERSPEECH 2006, Pittsburgh, PA, USA, 2017-Thu2011FoP.2013.

Banse R. and Scherer K. R. (1996). *Acoustic profiles in vocal emotion expression*. Journal of Personality and Social Psychology 70(3): 614.

Bänziger T. (2004). *Communication vocale des émotions: perception de l'expression vocale et attributions émotionnelles*, University of Geneva. PhD Thesis.

Bänziger T., Grandjean D. and Scherer K. R. (2009). *Emotion recognition from expressions in face, voice, and body. The Multimodal Emotion Recognition Test (MERT)*. Emotion 9(5): 691-704.

Bänziger T. and Scherer K. R. (2003). *Relations entre caractéristiques vocales perçues et émotions attribuées*. Actes des Journées Prosodie 2001 10-11: 119-124.

Barbosa P. and Bailly G. (1994). *Characterisation of rhythmic patterns for text-to-speech synthesis*. Speech Communication 15: 127-137.

Bartkova K. and Sorin C. (1987). *A model of segmental duration for speech synthesis in French* 1*. Speech Communication 6(3): 245-260.

Batušek R. (2002). *A duration model for Czech text-to-speech synthesis*. In proceedings of Speech Prosody 2002, Aix-en-Provence, France, 167-170.

Benesty J. (2008). *Springer handbook of speech processing*, Springer Verlag.

Black A. W. and Hunt A. J. (1996). *Generating F0 contours from ToBI labels using linear regression*. In proceedings of ICSLP 96, 1385-1388.

Black A. W., Zen H. and Tokuda K. (2007). *Statistical parametric speech synthesis*. In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007., Honolulu, Hawaii, U.S.A, 1229-1232.

Boersma P. and Weenink D. (2011). *Praat : doing phonetics by computer*. from http://www.praat.org/. .

Boite R. (2000). *Traitement de la parole*, PPUR presses polytechniques.

Brunelle M. (2003). *Coarticulation effects in Northern Vietnamese tones*. In proceedings of 15th ICPhS, Barcelona, 2673 – 2676.

Buhmann J., Vereecken H., Fackrell J., Martens J. P. and Coile B. (2000). *Data driven intonation modelling of 6 languages*. In proceedings of ICSLP 2000, Beijing, China, 179-182.

Cacioppo J. T. and Petty R. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Brows, Dubuque.

Cadic D. (2011). *Optimisation du procede de creation de voix en synthese par selection*. Paris, Université Paris-Sud 11. PhD Thesis.

Caelen-Haumont G. (2008). *Prosodie et sens: une approche expérimentale*, l'Harmattan.

Caelen-Haumont G. (2012). *Une méthode prosodique et linguistique pour générer une parole subjective en synthèse*. CORELA.

Caelen-Haumont G. and Bel B. (2000). *Le caractère spontané dans la parole et le chant improvisés: de la structure intonative au mélisme*. revue PArole 15(16): 251-302.

Caelen-Haumont G. and Keller E. (1997). *La prosodie, de la parole à la synthèse: l'apport de la sémantique et de la pragmatique*. Etudes de lettres(3): 103-130.

Cahn J. E. (1990). *The generation of affect in synthesized speech*. Journal of the American Voice IO Society 8(1): 1-19.

Calliope (1989). *La parole et son traitement automatique*, Masson.

Campbell N. (2000). *Databases of emotional speech*. In proceedings of Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, 34-38.

Campbell N. and Marumoto T. (2000). *Automatic labelling of voice-quality in speech databases for synthesis*. In proceedings of Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 468-471.

Campbell N. and Mokhtari P. (2003a). *Voice quality: the 4th prosodic dimension*. In proceedings of, 2417-2420.

Campbell N. and Mokhtari P. (2003b). *Voice Quality: the 4th Prosodic Dimension*. In proceedings of 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2417-2420.

Campbell W. N. (1992). *Syllable-based segmental duration*. Talking machines: Theories, models, and designs: 211–224.

Campbell W. N. (1996). *CHATR: A high-definition speech re-sequencing system*. In proceedings of Joint Meeting of Acoustical Society of America and Acoustical Society of Japan, 1223-1228.

Cannon W. B. (1931). *Again the James-Lange and the thalamic theories of emotion*. Psychological Review 38(4): 281-295.

Charpentier F. and Stella M. (1986). *Diphone synthesis using an overlap-add technique for speech waveforms concatenation*. In proceedings of IEEE International Conference on Acoustics Speech and Signal Processing- ICASSP 86, 2015-2018.

Chen G.-P., Bailly G., Liu Q.-F. and Wang R.-H. (2004). *A superposed prosodic model for Chinese text-to-speech synthesis*. In proceedings of International Conference of Chinese Spoken Language Processing.

Chen S.-H., Hwang S.-H. and Wang Y.-R. (1998). *An RNN-based prosodic information synthesizer for Mandarin text-to-speech*. Speech and Audio Processing, IEEE Transactions on 6(3): 226-239.

Childers D. G. and Lee C. (1991). *Vocal quality factors: Analysis, synthesis, and perception*. Journal of the Acoustical Society of America 90(5): 2394-2410.

Christophe V. (1998). *Les Emotions. Tour D'Horizon Des Principales Theories*, Presses Universitaires du Septentrion.

Chung H. (2002). *Duration models and the perceptual evaluation of spoken Korean*. Proceedings of Speech Prosody, Aix-en-Provence, France: 219–222.

Cornelius R. R. (2000). *Theoretical Approaches to Emotion*. In proceedings of ISCA 2000, Northern, Ireland, 3-10.

Crystal D. (2003). *A dictionary of linguistics & phonetics*, Wiley-Blackwell.

Crystal D. and WangFengxin (1997). *English as a global language*, Cambridge University Press Cambridge.

d'Alessandro C. and Doval B. (2003). *Voice quality modification for emotional speech synthesis*. In proceedings of InterSpeech 2003, Geneva, Switzerland, 1653-1656.

Damasio A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. New York, Putnam Press.

Danes F. (1994). *Involvement with language and in language*. Journal of pragmatics 22(3-4): 251-264.

De Biasi G., Aubergé V., Granjon L. and Vanpé A. (2012). *Perception of social affects from non lexical sounds*. In proceedings of International Conference on Speech and Copora (GSCP 2012), Belo Horizonte, Brazil.

Delattre P. (1966). *Les dix intonations de base du français*. The French Review 40(1): 1-14.

Di Cristo A. (1975). *Soixante et dix ans de recherches en prosodie*, Publ. de l'Université de Provence.

Di Cristo A. (2000). *La problématique de la prosodie dans l'étude de la parole dite spontanée*. revue PArole 15: 189-250.

Di Cristo A. (2004). *La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions*. Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence: 67-211.

Diaféria M.-L. (2002). *Les Attitudes de l'Anglais : Premiers Indices Prosodiques.* Science Cognitives. Grenoble, France, Institut National Polytechnique de Grenoble Master.

Do T. D., Tran T. H. and Boulakia G. (1998). *« Intonation in vietnamese »*. Intonation systems: A survey of 22 languages. D. Hirst and A. Di Cristo, Cambridge University Press, 395-416.

Do T. T. and Takara T. (2004). *Vietnamese Text-To-Speech system with precise tone generation*. Acoustical Science and Technology 25(5): 347-353.

Do V. T., Tran D. D. and Nguyen T. T. T. (2011). *Non-uniform unit selection in Vietnamese Speech Synthesis*. In proceedings of 2nd International Symposium on Information and Communication Technology - SoICT 2011, Hanoi.

Doan T. T. (1997). *Ngữ âm tiếng Việt (Vietnamese phonetics)*. Hanoi, Nhà xuất bản Đại học quốc gia Hà Nội.

Dutoit T. (1997). *An introduction to text-to-speech synthesis*, Kluwer Academic Publishers.

Eagly A. H. and Chaiken S. (1993). *The psychology of attitudes*, Harcourt Brace Jovanovich College Publishers.

Ekman P. and Oster H. (1979). *Facial Expressions of Emotion*. Annual Review of Psychology 30(1): 527-554.

Fant G. (1970). *Acoustic theory of speech production*, Walter de Gruyter.

Fazio R. H. (1995). *Attitudes as object-evaluation associations: determinants, consequences, and correlates of attitude accessibility*. Attitude Strength Antecedents and Consequences: 247-282.

Fisher C. and Tokura H. (1996). *Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure*. Signal to syntax: Bootstrapping from speech to grammar in early acquisition: 343–363.

Fónagy I. (1983). *La vive voix. Essais de psycho-phonétique*, Paris, Payot.

Fónagy I., Bérard E. and Fónagy J. (1984). *Clichés mélodiques*. Folia Linguistica 17: 153-185.

Fónagy I. and Jakobson R. (1991). *La vive voix*, Payot.

Fujisaki H. (1983). *Dynamic characteristics of voice fundamental frequency in speech and singing*. The production of speech: 39–55.

Fujisaki H. (1997). *Prosody, models, and spontaneous speech*. Computing Prosody. Y. Sagisaka, N. Campbell and N. Higuchi. New-York Springer-Verlag.

Fujisaki H. (2002). *Modeling in the study of tonal feature of speech with application to multilingual speech synthesis*. In proceedings of Joint Conference of SNLP and Oriental COCOSDA, Hua Hin Prachuapkirikhan.

Fujisaki H. (2003). *Prosody, information, and modeling - with emphasis on tonal features of speech*. In proceedings of WSLP-2003, 5-14.

Fujisaki H. and Hirose K. (1993). *Analysis and perception of intonation expressing paralinguistic information in spoken Japanese*. In proceedings of ESCA Workshop on Prosody, Lund, Sweden, 254-257.

Fujisaki H., Ohno S. and Wang C. (1998). *A command-response model for F0 contour generation in multilingual speech synthesis*. In proceedings of The Third ESCA/COCOSDA Workshop on Speech Synthesis, Australia.

Fujisaki H., Wang C., Ohno S. and Gu W. (2005). *Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model*. Speech Communication 47(1-2): 59-70.

Fussell S. R. (2002). *The Verbal Communication of Emotions: Interdisciplinary Perspectives*, Psychology Press.

Gendrot C., Henrich N., Sshade G., Muller F. and Expert R. (2004). *Vocal folds vibratory patterns of laryngeal mechanism M0 as investigated with high speed cinematography and electroglottography*. In proceedings of International Conference on Voice Physiology and Biomechanics, Marseille, France.

Gobl C. and Ni Chasaide A. (2003). *The role of voice quality in communicating emotion, mood and attitude*. Speech Communication 40(1-2): 189-212.

Grépillat T. (1996). *Perçoit-on, par l'intonation, l'attitude d'un locuteur avant la fin de l'énoncé*, Université Stendhal Grenoble III. T.E.R. de Maîtrise d'Anglais.

Grosz B. and Hirschberg J. (1992). *Some intonational characteristics of discourse structure*. In proceedings of Second International Conference on Spoken Language Processing (ICSLP'92), Banff, Alberta, Canada, 429-432.

Gu W., Zhang T. and Fujisaki H. (2011). *Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes*. In proceedings of Interspeech 2011, Firenze, Italy, 1069-1072.

Hallahan W. I. (1995). *DECtalk software: Text-to-speech technology and implementation*. Digital Technical Journal 7(4): 5-19.

Hamza W., Bakis R., Eide E. M., Picheny M. A. and Pitrelli J. F. (2004). *The IBM expressive speech synthesis system*. In proceedings of InterSpeech 2004, Jeju Island, Korea, 2577–2580.

Han M. S. and Kim K. O. (1974). *Phonetic variation of Vietnamese tones in disyllabic utterances*. Journal of Phonetics 2: 223-232.

Hansakunbuntheung C., Kato H. and Sagisaka Y. (2007). *Syllable-Based Thai Duration Model using Multi-Level Linear Regression and Syllable Accommodation*. In proceedings of 6th ISCA Workshop on Speech Synthesis, Bonn, Germany.

He L., Yang J., Zuo L. and Kui L. (2011). *A trainable Vietnamese speech synthesis system based on HMM*. In proceedings of International Conference on Electric Information and Control Engineering (ICEICE 2011) 3910-3913.

Heuft B., Portele T. and Rauth M. *Emotions in time domain synthesis*. In proceedings of Fourth International Conference on Spoken Language (ICSLP 96), 1974-1977.

Hirst D. and Di Cristo A. (1998). *Intonation systems: a survey of twenty languages*, Cambridge Univ Pr.

Holm B. (2003). *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie – Apprentissage automatique et application à l'énonciation de formules mathématiques*, Institut National Polytechnique, Grenoble - France.

Holm B. and Bailly G. (2000). *Generating Prosody by Superposing Multi-Parametric Overlapping Contours*. In proceedings of ICSLP 2000, Beijing, China, 203-206.

Hwang S. H. and Chen S. H. (1995). *A prosodic model of Mandarin speech and its application to pitch level generation for text-to-speech*. In proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), 616-619.

Iida A. and Campbell N. (2003a). *Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders*. International Journal of Speech Technology 6(4): 379-392.

Iida A., Campbell N., Higuchi F. and Yasumura M. (2003b). *A corpus-based speech synthesis system with emotion*. Speech Communication 40(1): 161-187.

Imai S. (1983). *Cepstral analysis synthesis on the mel frequency scale*. In proceedings of ICASSP 1983, 93-96.

Izard C. E. (2000). *Organizational and Motivational Functions of Discrete Emotions*. Handbook of Emotions. M. Lewis and J. Haviland. New York, Guilford, 631-641.

James W. (1884). *What is an emotion?* Mind 9: 188-205.

Jilka M., Mohler G. and Dogil G. (1999). *Rules for the generation of ToBI-based American English intonation1*. Speech Communication 28(2): 83-108.

Katz G. S., Cohn J. F. and Moore C. A. (1996). *A Combination of Vocal f0 Dynamic and Summary Features Discriminates between Three Pragmatic Categories of Infant Directed Speech*. Child Development 67(1): 205-217.

Kenyon J. S. (1948). *Levels of Speech and Colloquial English*. The English Journal 37(1): 25-31.

Kishore S. and Black A. W. (2003). *Unit size in unit selection speech synthesis*. In proceedings of InterSpeech 2003, Geneva, Switzerland, 1317–1320.

Klabbers E. (2000). *Segmental and prosodic improvements to speech generation*. Eindhoven, Eindhoven University of Technology. PhD Thesis.

Klatt D. H. (1980). *Software for a cascade/parallel formant synthesizer*. Journal of the Acoustical Society of America 67(3): 971-995.

Krishna N. S. and Murthy H. A. (2004). *Duration modeling of Indian languages Hindi and Telugu*. In proceedings of Fifth ISCA ITRW on Speech Synthesis, Pittsburgh, PA, USA, 197–202.

Lacheret-Dujour A. and Beaugendre F. (1999). *La prosodie du français*, CNRS EDITIONS.

Laver J. (1980). *The phonetic description of voice quality*, Cambridge Studies in Linguistics

Le H. M. (2003). *Một số kết quả nghiên cứu và phát triển hệ phần mềm chuyển văn bản thành tiếng nói cho tiếng Việt bằng tổng hợp Formant*. In proceedings of Hội thảo Khoa học Quốc gia lần thứ nhất - Nghiên cứu Phát triển và Ứng dụng Công nghệ Thông tin và Truyền thông, Hanoi.

Le T. X. (1989). *Etude contrastive de l'intonation expressive en français et en vietnamien*. Linguistic and Phonetic, Université Paris 3. PhD.

Léon J.-P. (2000). *Phonétisme et prononciation du français*. Paris, Cursus.

Liénard J. (1977). *Les processus de la communication parlée, introduction à l'analyse et à la synthèse de la parole, édition*, Masson.

Loyau F. (2007). *Expressions des états mentaux et émotionnels de l'humain en intéraction: ébauches du «Feeling of Thinking»*, Grenoble-INP. PhD Thesis.

Loyau F., Aubergé V. and Vanpé A. (2006). *Expressions hors des tours de parole: éthogrammes du «feeling of thinking»*. In proceedings of JEP 2006, Dinard.

Lu Y. and Aubergé V. (2012). *A corpus devoted to the cross-perception of mandarin Chinese vs. French social affects*. In proceedings of GSCP, Belo Horizonte, Brazil.

Lyman D. (1856). *The moral sayings of Publius Syrus, a Roman slave: from the Latin*, LE Bernard & Company.

Mai N. C., Vũ Đ. N. and Hoàng T. P. (2002). *Cơ sở ngôn ngữ học và Tiếng Việt*, Nhà xuất bản Giáo dục.

Maio G., Maio G. R. and Haddock G. (2010). *The psychology of attitudes and attitude change*, Sage Publications Ltd.

Martin P. (2000). *WinPitch 2000: a tool for experimental phonology and intonation research*. W: Pros: 149-155.

Martins-Baltar M. (1977). *De l'énoncé à l'énonciation: une approche des fonctions intonatives*. Paris, Didier.

Mixdorff H. (2000). *A novel approach to the fully automatic extraction of Fujisaki model parameters*. In proceedings of, 1281-1284 vol. 1283.

Mixdorff H., Bach N. H., Fujisaki H. and Luong M. C. (2003a). *Quantitative analysis and synthesis of syllabic tones in Vietnamese*. In proceedings of InterSpeech 2003, Geneva, Switzerland, 177-180.

Mixdorff H., Fujisaki H., Chen G. P. and Hu Y. (2003b). *Towards the automatic extraction of Fujisaki model parameters for Mandarin*. In proceedings of InterSpeech 2003, Geneva, Switzerland, 873-876.

Mixdorff H. and Jokisch O. (2001). *Building An Integrated Prosodic Model of German*. In proceedings of EUROSPEECH 2001, Aalborg, Denmark.

Mixdorff H., Luksaneeyanawin S., Fujisaki H. and Charnvivit P. (2002). *Perception of tone and vowel quantity in Thai*. In proceedings of 7th International Conference on Spoken Language Processing, Denver, Colorado, USA, 753-756.

Mobius B. and Von Santen J. (1996). *Modeling segmental duration in German text-to-speech synthesis*. In proceedings of ICSLP 96, Philadelphia, USA, 2395-2398.

Montero J., Gutiérrez-Arriola J., Colás J., Enríquez E. and Pardo J. (1999). *Analysis and modelling of emotional speech in Spanish*. In proceedings of ICPhS99, 957-960.

Moraes J. A., Rilliard A., Mota B. A. O. and Shochi T. (2010). *Multimodal perception and production of attitudinal meaning in Brazilian Portuguese*. In proceedings of Speech Prosody 2010, Chicago, IL, USA, 340.

Morlec Y. (1997). *Génération multiparamétrique de la prosodie du Français par apprentissage automatique*, Institut National Polytechnique de Grenoble, France. PhD Thesis.

Morlec Y., Bailly G. and Aubergé V. (1997a). *Apprentissage automatique d'un module de génération multistyle de l'intonation*. In proceedings of 1ères JST FRANCIL, Avignon, France, 407-412.

Morlec Y., Bailly G. and Aubergé V. (1997b). *Generating the prosody of attitudes*. In proceedings of ETRW Workshop on Prosody, Athens, Greece, 251-254.

Morlec Y., Bailly G. and Aubergé V. (2001). *Generating prosodic attitudes in French: data, model and evaluation*. Speech Communication: 357-371.

Moulines E. and Laroche J. (1995). *Non-parametric techniques for pitch-scale and time-scale modification of speech*. Speech Communication 16(2): 175-205.

Mozziconaccim S. (2010). *Emotion and attitude conveyed in speech by means of prosody*. In proceedings of 2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction, Sonthofen, Germany.

Murray I. R. and Arnott J. L. (1995). *Implementation and testing of a system for producing emotion-by-rule in synthetic speech*. Speech Communication 16(4): 369-390.

Nadeu M. and Prieto P. (2011). *Pitch range, gestural information, and perceived politeness in Catalan*. Journal of pragmatics 43(3): 841-854.

Nguyen Q. C. (2002). *Reconnaissance de la parole en langue Vietnamienne*, INP Grenoble. PhD Thesis.

Nguyen Q. H. (1994). *Âm tiết và loại hình ngôn ngữ*, Khoa học xã hội.

Nguyen T. C. (1996). *Ngữ pháp tiếng Việt (Vietnamese Grammar)*. Hanoi, National University in Hanoi.

Nguyen T. D., Mixdorff H., Luong C. M., Ngo H. H. and Vu K. B. (2004). *Fujisaki Model based F0 contours in Vietnamese TTS*. In proceedings of ICSLP2004, Korea, 1429-1432.

Nguyen T. T. H. (2004). *Contribution à l'étude de la prosodie du vietnamien. Variations de l'intonation dans les modalités: assertive, interrogative et impérative*. Lettres, sciences sociales et humaines, Université Paris Diderot - Paris 7. PhD Thesis.

Nguyen T. T. H. and Boulakia G. (1999). *Another look at Vietnamese intonation*. In proceedings of ICPhS, San Fancisco.

O'Connor J. D. and Arnold G. F. (1978). *Intonation of Colloquial English*, Longman.

Ohala J. J. (1996). *Ethological theory and the expression of emotion in the voice*. In proceedings of, 1812-1815 vol. 1813.

Ohman S. (1967). *Word and sentence intonation: A quantitative model*. Speech Transmission Laboratory Quarterly Progress and Status Report 2(3): 20-54.

Park M. (2000). *Les réalisations prosodiques de la focalisation en coréen spontané*, Actes des XXIIIèmes Journées d'Etudes de la Parole, Aussois.

Pham T. N. Y., Castelli E. and Nguyen Q. C. (2002). *Gabarits des tons vietnamiens*. In proceedings of JEP, Nancy, France, 23-26.

Pierrehumbert J. (1981). *Synthesizing intonation*. Journal of the Acoustical Society of America 70(4): 985-995.

Pierrehumbert J. B. (1980). *The phonology and phonetics of English intonation*, MIT Cambridge, MA. PhD.

Pitrelli J. F., Bakis R., Eide E. M., Fernández R., Hamza W. and Picheny M. A. (2006). *The IBM expressive text-to-speech synthesis system for American English*. Audio, Speech, and Language Processing, IEEE Transactions on 14(4): 1099-1108.

Plutchik R. (1984). *Emotion : a general psychoevolutionary theory*. In Approach to emotion. K. R. SCHERER and P. Ekman, Hillsdale, Laurence Erlbaum Associates.

Rank E. and Pirker H. (1998). *Generating emotional speech with a concatenative synthesizer*. In proceedings of ICSLP 98, 671-674.

Rao K. S. and Yegnanarayana B. (1997). *Modeling syllable duration in Indian languages using neural networks*. In proceedings of ICASSP 2004, 313.

Riley M. D. (1992). *Tree-based modeling for speech synthesis*. Talking Machines: Theories, Models, and Designs. G. Bailly, C. Benoit and T. Sawallis. Amsterdam, 265–273.

Rilliard A., Martin J.-C., Aubergé V. and Shochi T. (2008). *Perception of French Audio-Visual Prosodic Attitudes* In proceedings of Speech Prosody 2008, Campinas, Brazil, 685-688.

Rilliard A., Shochi T., Martin J.-C., Erickson D. and Aubergé V. (2009). *Multimodal indices to Japenese and French: Prosodicaly expressed social affects*. Language and Speech: 223-243.

Rossi M. (1999). *L'intonation: le système du français: description et modélisation*, Editions Ophrys.

Sagisaka Y. (1988). *Speech synthesis by rule using an optimal selection of non-uniform synthesis units*. In proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 679-682 vol. 671.

Sagisaka Y. (1990). *On the prediction of global F0 shape for Japanese text-to-speech*. In proceedings of ICASSP 1990, 325-328.

Scherer K. R. (1984a). *On the nature and function of emotion: A component process approach*. Approaches to emotion. K. R. Scherer and P. Ekman. Hillsdale, NJ: Erlbaum, 293-317.

Scherer K. R. (1984b). *The state of the art in vocal communication: A partial view*. Nonverbal behavior: Perspectives, applications, intercultural insights. A. Wolfgang. New York, Hogrefe, 41-74.

Scherer K. R. (1986). *Studying emotion empirically: Issues and a paradigm for research*. Experiencing emotion: A cross-cultural study. H. G. W. K. R. Scherer and A. B. Summerfield. Cambridge, Cambridge University Press, 3-27.

Scherer K. R. (2000). *Emotion*. Introduction to Social Psychology: A European perspective (3rd. ed). M. Hewstone and W. Stroebe. Oxford, Blackwell, 151-191.

Scherer K. R. (2003). *Vocal communication of emotion: A review of research paradigms*. Speech Communication 40(1-2): 227-256.

Schröder M. (1999). *Can emotions be synthesized without controlling voice quality*. Phonus 4: 37-55.

Schröder M. (2001). *Emotional Speech Synthesis: A Review*. In proceedings of EUROSPEECH-2001, Aalborg, Denmark, 561-564.

Schröder M. (2004a). *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions*. Affective Dialogue Systems, Springer Berlin / Heidelberg.

Schröder M. (2004b). *Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. Institute of Phonetics, Saarland University.

Schröder M. and Grice M. (2003). *Expressing vocal effort in concatenative synthesis*. In proceedings of The 15th International Conference of Phonetic Sciences, Barcelona, 2589–2592.

Schröder M., Pirker H., Lamolle M., Burkhardt F., Peter C. and Zovato E. (2011). *Representing Emotions and Related States in Technological Systems*. Emotion-Oriented Systems, Springer Berlin Heidelberg.

Schubiger M. (1958). *English Intonation: Its Form and Function*, Max Niemeyer Verlag GmbH & Co KG.

Shih C. and Ao B. (1997). *Duration study for the Bell Laboratories Mandarin text-to-speech system*, Springer, New York.

Shochi T. (2008). *Prosodie des affects socioculturels en japonais, français et anglais : à la recherche des vrais et faux-amis pour le parcours de l'apprenant* Science du Langage. Grenoble, Universites Grenoble III PhD Thesis.

Shochi T., Aubergé V. and Rilliard A. (2005). *Because Attitudes Are Social Affects, They Can Be False Friends*. Affective Computing and Intelligent Interaction, 482-489.

Shochi T., Aubergé V. and Rilliard A. (2006). *How prosodic attitudes can be false friends: Japanese vs. French social affects.* In proceedings of Speech Prosody, Dresden, 692-696.

Shochi T., Aubergé V. and Rilliard A. (2007). *Cross-listening of japanese, english and french social affect: about universals, false friends and unknown attitudes*. In proceedings of 16th ICPhS, Saarbrücken.

Shochi T., Erickson D., Rilliard A. and Aubergé V. (2008). *Recognition of Japanese attitudes in Audio-Visual speech*. In proceedings of Speech Prosody, Campinas, Bresil, 689-692.

Shochi T., Gagnié G., Rilliard A., Erickson D. and Aubergé V. (2010). *Learning effect of prosodic social affects for Japanese learners of French language*. In proceedings of Speech Posody, Chicago, USA.

Shochi T., Rilliard A., Auberge V. and Erickson D. (2009). *Intercultural Perception of English, French and Japanese Social Affective Prosody*. The Role of Prosody in Affective Speech. D. Erickson, Peter Lang. 97, 31-59.

Signorello R., Aubergé V., Vanpé A., Granjon L. and Audibert N. (2010). *A la recherche d'indices de culture et/ou de langue dans les micro-e ve nements audio-visuels de l'interaction face a face*. In proceedings of WACA 2010, Lille, France.

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J. and Hirschberg J. (1992). *ToBI: A standard for labeling English prosody*. In proceedings of Spoken Language Processing, Banff, 867-870.

Sproat R. W. (1998). *Multilingual text-to-speech synthesis: the Bell Labs approach*, KLUWER academic publishers.

Sreenivasa Rao K. and Yegnanarayana B. (2009). *Intonation modeling for Indian languages*. Computer Speech & Language 23(2): 240-256.

Strik H. and Boves L. (1992). *Control of fundamental frequency, intensity and voice quality in speech*. Journal of Phonetics 20(1): 15-25.

Thiessen E. D., Hill E. A. and Saffran J. R. (2005). *Infant-directed speech facilitates word segmentation*. Infancy 7(1): 53-71.

Traber C. (1990). *F0 generation with a data base of natural F0 patterns and with a neural network*. In proceedings of The ESCA Workshop on Speech Synthesis, Autrans, France.

Traber C. (1993). *Syntactic processing and prosody control in the SVOX TTS system for German*. In proceedings of EUROSPEECH '93, Berlin, Germany, 2099-2102.

Tran D. D. (2007). *Synthèse de la parole à partir du texte en langue vietnamienne*, INP Grenoble. PhD Thesis, 227.

Tran D. D. and Castelli E. (2008). *Registres des tons vietnamiens dans la parole continue*. In proceedings of Journées d'Etudes sur la Parole, Avignon, France.

Tran D. D., Castelli E., Serignat J.-F. and Le V. B. (2007). *Analysis and Modeling of Syllable Duration for Vietnamese Speech Synthesis*. In proceedings of International Conference on Speech Databases and Assessment, Oriental COCOSDA, Hanoi, Vietnam.

Tran D. D., Castelli E., Serignat J.-F., Trinh V. L. and Le X. H. (2005). *Influence of F0 on Vietnamese syllable perception*. In proceedings of Interspeech, Lisbon, 1697-1700.

Trask R. L. (1996). *A dictionary of phonetics and phonology*, Burns & Oates.

van Santen J. (1992). *Contextual effects on vowel duration*. Speech Communication 11(6): 513-546.

van Santen J. P. H. and Möbius B. (1999). *A quantitative model of F0 generation and alignment*. Intonation: Analysis, Modelling and Technology: 269–288.

van Santen J. P. H., Möbius B., Venditti J. J. and Shih C. (1998). *Description of the Bell Labs intonation system*. In proceedings of 3rd ESCA Speech Synthesis Workshop, Jenolan Caves, 293-298.

Vanpé A. (2011). *Expressions et micro-expressions spontanées de la face et de la voix en Interaction Homme-Machine : esquisse d'un modèle du "Feeling of Thinking"*. Grenoble, Université Stendhal - Grenoble III. PhD.

Vincent D., Rosec O. and Chonavel T. (2005). *Estimation of LF glottal source parameters based on an ARX model*. In proceedings of Interspeech'2005, Lisbon, Portugal, 333-336.

Vroomen J., Collier R. and Mozziconacci S. (1993). *Duration and intonation in emotional speech*. In proceedings of EUROSPEECH '93, Berlin, Germany, 577-580.

Vu H. Q. and Cao X. N. (2009a). *Tổng hợp tiếng nói tiếng Việt, theo phương pháp ghép nối cụm từ*. The Journal on Information Technologies and Communications V.

Vu M. Q., Tran D. D. and Castelli E. (2006a). *Prosody of Interrogative and Affirmative Sentences in Vietnamese Language:Analysis and Perceptive Results*. In proceedings of Interspeech, Pittsburgh, Pennsylvania, USA.

Vu M. Q., Trân Đ. Đ. and Castelli E. (2006b). *Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results*. In proceedings of Interspeech 2006, Pittsburgh, PA, USA.

Vu T. T., Luong M. C. and Nakamura S. (2009b). *An HMM-based Vietnamese speech synthesis system*. In proceedings of Oriental COCOSDA International Conference on Speech Database and Assessments, 2009 Urumqi 116-121.

Wichmann A. (2000). *The attitudinal effects of prosody, and how they relate to emotion*. In proceedings of ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK.

Wu C. H. and Chen J. H. (2001). *Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis*. Speech Communication 35(3-4): 219-237.

Xuedong H., Alex A. and Hsiao-Wuen H. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR.

Yamagishi J., Kobayashi T., Tachibana M., Ogata K. and Nakano Y. (2007). *Model adaptation approach to speech synthesis with diverse voices and styles*. In proceedings of, IV-1233-IV-1236.

Yamagishi J., Onishi K., Masuko T. and Kobayashi T. (2003). *Modeling of various speaking styles and emotions for HMM-based speech synthesis*. In proceedings of INTERSPEECH 2003, Geneva, Switzerland, 2461-2464.

Yu J., Zhang W. and Tao J. *A New Pitch Generation Model Based on Internal Dependence of Pitch Contour for Manadrin TTS System*. In proceedings of ICASSP 2006 Toulouse, France.

Yuan J., Shih C. and Kochanski G. P. (2002). *Comparison of declarative and interrogative intonation in Chinese*. In proceedings of Speech Prosody 2002, Aix-en-Provence, France, 711-714.

Zanna M. P. and Rempel J. K. (1988). *Attitudes: A new look at an old concept*. The social psychology of knowledge. Cambridge, UK, Cambridge University Press, 315–334.

Zen H., Tokuda K. and Kitamura T. (2004). *An introduction of trajectory model into HMM-based speech synthesis*. In proceedings of Fifth ISCA ITRW on Speech Synthesis, Pittsburgh, PA, USA, 191–196.

Zovato E., Pacchiotti A., Quazza S. and Sandri S. (2004). *Towards emotional speech synthesis: A rule based approach*. In proceedings of ISCA Speech SynthesisWorkshop, Pittsburgh, PA, 219–220.

Annexe 1 : Les phrases du corpus d'attitude vietnamienne

Le corpus d'attitudes vietnamiennes contient au total 125 phrases isolées dont la longueur varie entre 1 et 8 syllabes comme suivant :

1 syllabe

N0	Ton	Phrases			Commentaire
		Vietnamien	IPA (pas de ton)	Francais	
1	1	Ta	/ tɐ:/	nous	
2	2	Cần	/kɜn/	besoin\ la flèche	
3	3	Đã	/ɗɐ:/	Déjà	
4	4	Trả	/ cɐ:/	Rendre	
5	5i	Cá	/kɐ:/	le poisson	
6	5ii	Tám	/tɐ:m/	8	
7	6	Tạm	/tɐ:m/	Temporairement	
8	5b	Hát	/hɐ:t/	chanter	
9	6b	Một	/moʇ /	1	

2 syllabes

N0	Ton	Phrases			Commentaire
		Vietnamien	IPA	Français	
10	1_1	Anh ta	/ɐ:ʲŋ tɐ:/	lui	
11	2_1	Người ta	/ŋɨɜj tɐ:/	autrui	
12	3_1	Đã xong	/ɗə̰; sɔŋ/	déjà fini	
13	4_1	Thủy tinh	/tʰwi tiʲŋ /	verre	
14	5_1	Chúng ta	/cuŋ tɐ:/	nous	
15	6_1	Chị ta	/cị tɐ:/	elle	
16	5b_1	Héc ta	/hɛk tɐ:/	hectare	
17	6b_1	Tốp ca	/top kɐ:/	chant choral	
18	1_2	Rau cần	/zɐw: kɜn/	oenanthe de Java	
19	2_2	Cần cù	/ kɜn kᴜ /	laborieux	
20	3_2	Mũ nồi	/ mᴚ nɔj/	Le béret	

21	4_2	Quả cà	/kwɐ: tɐ:/	fruit de la morelle	
22	5_2	Mái nhà	/mɐ:j ɲɐ:/	La toiture	
23	6_2	Dọn nhà	/zɔ̰n ɲɐ:/	déménager	
24	5b_2	Góc nhà	/ɣɔk ɲɐ:/	réduit	
25	6b_2	Cột nhà	/koʈ ɲɐ:/	pilier	
26	1_3	Dây kẽm	/zɜj kɛm/	fil d'acier	
27	2_3	Hà mã	/ hɐ: mɐ:/	hippopotame	
28	3_3	Mẫu mã	/mɜ̰w mɐ:/	type	
29	4_3	Triển lãm	/ciɜn lɐ̰:m/	l'exposition	
30	5_3	Nước lã	/nɨɜk lɐ̰:/	eau plate	
31	6_3	Thị xã	/tʰḭ sɐ̰:/	cité municipale	
32	5b_3	Cái tã	/kɐ:j tɐ:/	le lange	
33	6b_3	Mật mã	/mɜ̰t mɐ:/	code secret	
34	1_4	Cây cảnh	/kɜj kɐ:ʲŋ/	plante d'agrément	
35	2_4	Đường kẻ	/ɗɨɜŋ kɛ/	la ligne	
36	3_4	Đũa cả	/ɗuɜ̰35 kɐ/	grandes baguettes	
37	4_4	Của cải	/kuɜ kɐ:j/	richesses	
38	5_4	Thước kẻ	/tʰɨɜk kɛ/	la règle	
39	6_4	Thợ cả	/tʰə̰: kɐ/	ouvrie chef d'équipe	
40	5b_4	Tất cả	/tɜt kɐ:/	tout	
41	6b_4	Dịch tả	/zḭk tɐ:/	choléra	
42	1_5	Y tá	/i tɐ:/	l'infirmier	
43	2_5	Gò má	/ɣɔ mɐ:/	pommette	
44	3_5	Bãi cá	/ɓɐ̰:j kɐ:/	terrain de poisson	
45	4_5	Chả cá	/cɐ: kɐ:/	hachis frit de poisson	
46	5_5	Trứng cá	/cɨŋ kɐ/	frai (de poissons)	
47	6_5	Phụ tá	/fụ̰ tɐ:/	assistant auxiliaire	
48	5b_5	Diếp cá	/ziɜp kɐ:/	houttuynia	
49	6b_5	Bột cá	/ɓoʈ̰ kɐ:/	farine de poisson	
50	1_6	Danh bạ	/zɐ:ʲŋ ɓɐ:/	annuaire	
51	2_6	Dừa cạn	/zɨɜ kɐ:n/	pervenche	
52	3_6	Bãi cạn	/ɓɐ̰ :j kɐ:n/	sédentaire	
53	4_6	Quả tạ	/kwɐ tɐ:/	haltère	
54	5_6	Búa tạ	/ɓuɜ tɐ:/	la masse	
55	6_6	Ruộng cạn	/zuɜ̰ŋ kɐ:n/	rizière à sec	
56	5b_6	Mắc cạn	/mɐk kɐ:n/	condamner	

57	6b_6	Học bạ	/hɔ̰k ɓɐ̰:/	livret scolaire	
58	1_5b	Công tác	/koŋ tɐ:k/	mission	
59	2_5b	Xà kép	/sɐ: kɛp/	barres parallèles	
60	3_5b	Bãi rác	/ɓɐ̰:j zɐ:k/	champ d'ordures	
61	4_5b	Tuổi tác	/tuɜj tɐ:k/	âge	
62	5_5b	Sáng tác	/sɐ:ŋ tɐ:k/	composer	
63	6_5b	Cộng tác	/koŋ tɐ:k/	collaborer	
64	5b_5b	Xúc tác	/suk tɐ:k/	catalyser	
65	6b_5b	Hợp tác	/hə̰p tɐ:k/	coopérer	
66	1_6b	Sa mạc	/sɐ: mɐ̰:k/	désert	
67	2_6b	Dàn nhạc	/zɐ:n ɲɐ̰:k31/	orchestre	
68	3_6b	Võng mạc	/vɔ̰ŋ mɐ̰:k/	rétine	
69	4_6b	Đoản mạch	/ɗwan mɐ̰:k/	court-circuit	
70	5_6b	Bế mạc	/ɓe mɐ̰:k/	clôture	
71	6_6b	Động mạch	/ɗo̰ŋ mɐ̰:k/	artère	
72	5b_6b	Giác mạc	/zɐ:k mɐ̰k/	immonde	
73	6b_6b	Chập mạch	/cɜp mɐ̰:k/	court-circuit	

3 syllabes

N0	Ton	Phrases		Commentaire
		Vietnamien	Francais	
74	1_1_1_i	Hai mươi ba	23	Numéro
75	1_1_1_ii	Anh em ta	Nous deux (toi et moi)	GN
76	1_1_1_iii	Găng tay da	Gant en cuir	NG
77	1_1_1_iv	Em theo anh	Tu me suis	SVO
78	4_1_1	Bảy mươi ba	73	Numéro
79	1_5_1	Hai chúng ta	Nous deux	NG
80	1_1_6	Găng tay dạ	Le gant en cuir	NG
81	1_1_6	Em theo chị	Tu me suis	SVO
82	6_2_3	Một ngàn rưỡi	1500	Mot
83	6_5b_3	Hợp tác xã	coopération	Mot
84	1_4_6	Em bảo chị	Tu me dis	SOV

4 syllabes

	Ton	Phrases		Commentaire
		Vietnamien	Francais	
85	1_1_1_1_i	Hai trăm hai ba	223	Numéro
86	1_1_1_1_ii	Hai anh em ta	Nous deux (toi et moi)	GN
87	1_1_1_1_iii	Đôi găng tay da	Paire de gants en cuir	GN

88	1_1_1_1_iv	Em đi theo anh	Tu me suis	SVO
89	4_1_1_1	Bảy trăm hai ba	723	Numéro
90	1_2_1_1	Hai ngàn ba trăm	2300	Numéro
91	1_1_4_1	Hai trăm bảy ba	273	Numéro
92	1_1_1_4	Ba trăm hai bảy	327	Numéro
93	1_1_2_1	Găng tay bằng da	Le gant en cuir	GN
94	1_1_1_6	Em đi theo chị	Tu me suis	SOV
95	2_6_2_3	Mười một ngàn rưỡi	11500	Numéro
96	1_5b_1_6	Đôi tất xanh lục	Les chaussettes vertes	GN
97	5b_4_1_6	Tất cả theo chị	Tous me suivent	SOV

5 syllabes

	Ton	Phrases		Commentaire
		Vietnamien	Francais	
98	1_1_1_1_1_i	Hai trăm hai mươi ba	223	Numéro
99	1_1_1_1_1_ii	Em đang đi theo anh	Te me suis encore	SVO
100	4_1_1_1_1	Bảy trăm hai mươi ba	723	Numéro
101	1_4_2_1_1	Ba bảy ngàn hai trăm	37200	Numéro
102	1_1_2_6_1	Hai mươi ngàn một trăm	20100	Numéro
103	1_1_4_1_1	Hai trăm bảy mươi ba	273	Numéro
104	1_1_1_1_4	Ba trăm hai mươi bảy	327	Numéro
105	1_1_1_2_1	Găng tay da màu nâu	Les gents bruns en cuir	GN
106	1_1_1_1_6	Hai em đi theo chị	Les deux me suivent	SOV
107	1_1_5b_2_3	Hai mươi mốt ngàn rưỡi	21500	Numéro
108	5b_4_3_1_6	Tất cả đã theo chị	Tout le monde m'a (t'a) suivi	

6 syllabes

	Ton	Phrases		Commentaire
		Vietnamien	Francais	
109	1_1_1_1_1_1	Hai anh em đi theo anh	Tous les deux me suivent	SVO
110	1_2_1_1_1_1	Hai ngàn hai trăm năm mươi	2250	Numéro
111	1_1_1_1_2_1	Đôi găng tay da màu nâu	Paire de gants bruns en cuir	GN
112	1_1_1_1_1_6	Hai anh em đi theo chị	Tous les deux me suivent	SOV

113	5_1_6_4_1_2	Sáu mươi triệu bảy trăm ngàn	60 700 000	Numéro
114	5b_4_3_1_1_6	Tất cả đã đi theo chị	Tout le monde m'a (t'a) suivi	SVO

7 s1yllabes

	Ton	Phrases		Commentaire
		Vietnamien	Francais	
115	1_1_1_1_1_1_1	Hai anh em đang đi theo anh	Les deux me suivent maintenant	SVO
116	1_2_1_1_1_1_1	Hai ngàn hai trăm năm mươi ba	2253	Numéro
117	1_1_1_1_1_2_1	Hai đôi găng tay da màu nâu	Deux paires de gants bruns en cuir	GN
118	1_1_1_1_1_1_6	Hai anh em đang đi theo chị	Les deux elle (te) suivent encore	SOV
119	5_1_1_6_4_1_2	Sáu mươi ba triệu bảy trăm ngàn	63 700 000	Numéro
120	5b_4_3_2_1_1_1	Tất cả mọi người đi theo anh	Tout le monde te suit	SVO

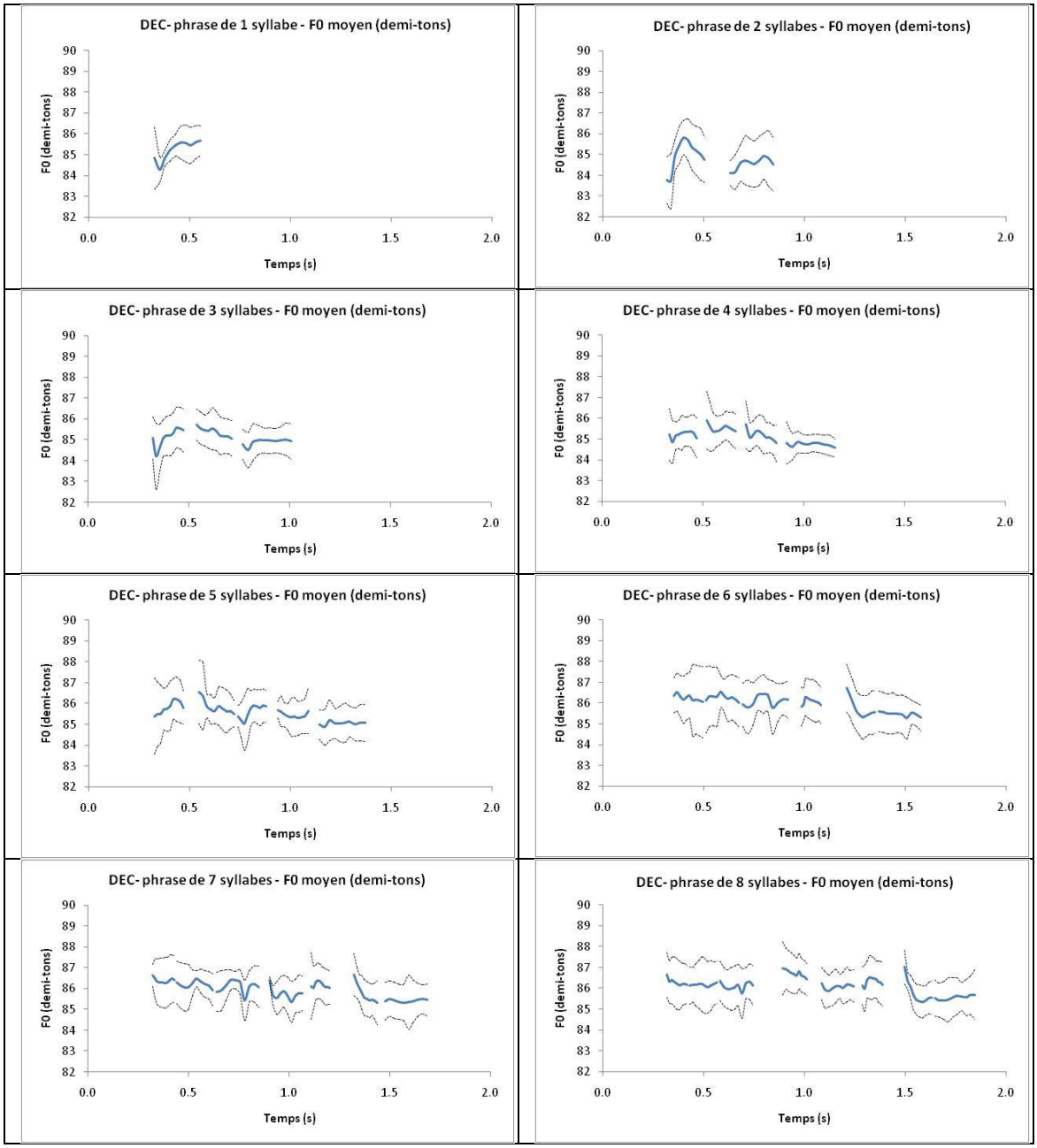
8 syllabes

	Ton	Phrases		Commentaire
		Vietnamien	Francais	
121	1_1_1_1_1_1_1_1	Hai anh em kia đang đi theo anh	Les deux me (te) suivent encore	SVO
122	1_1_2_1_1_1_1_1	Hai mươi ngàn hai trăm năm mươi ba	2253	Numéro
123	1_1_1_1_1_1_1_6	Hai anh em kia đang đi theo chị	Les deux gens me (tu) suivent encore	SOV
124	2_6b_2_4_1_1_1_6	Mười một ngàn bảy trăm năm mươi tám	11758	Numéro
125	5b_4_6_2_3_1_1_6	Tất cả mọi người đã đi theo chị	Tout le monde t'a (m'a) suivi	SVO

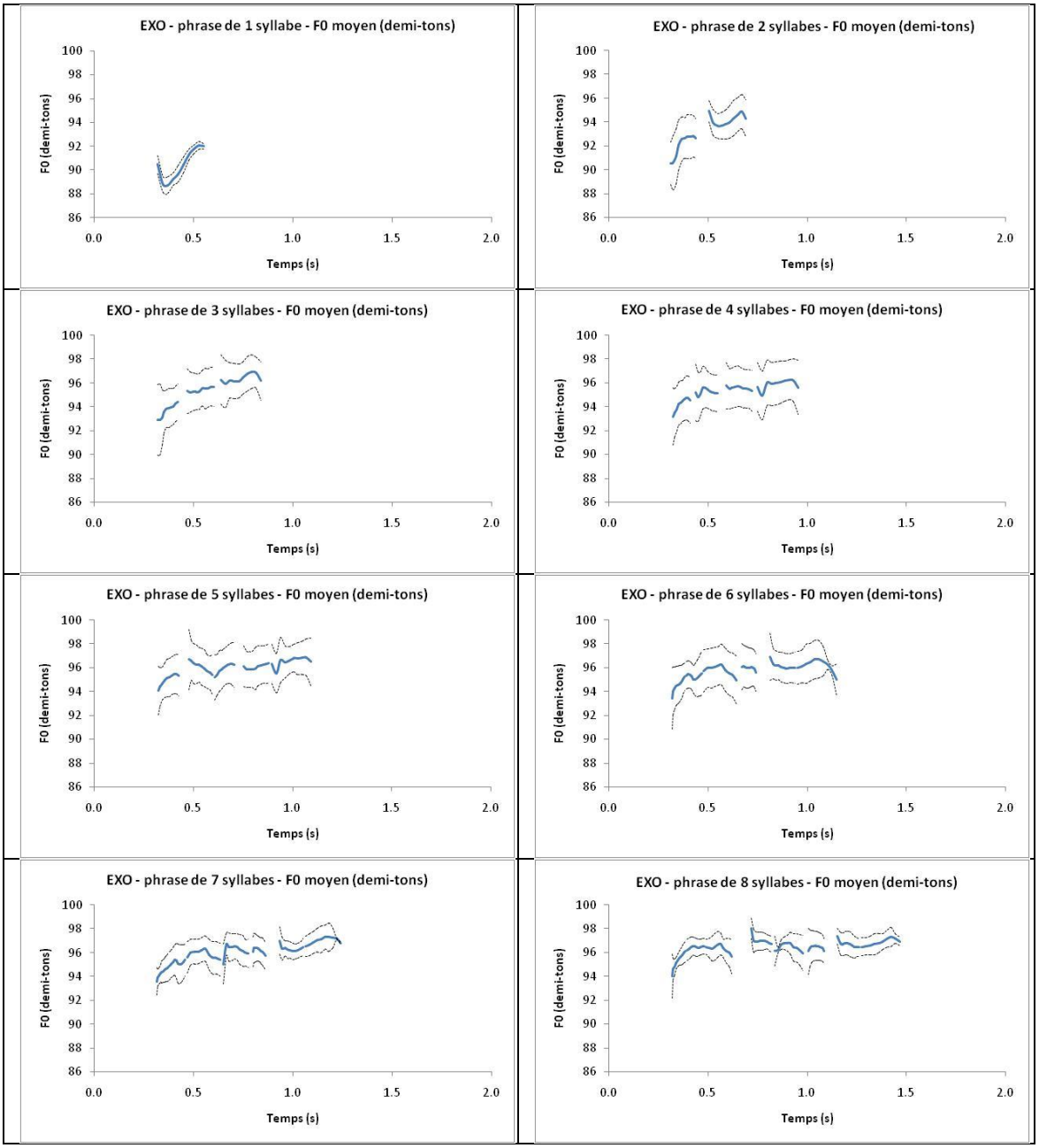
Annexe 2: Contours prosodiques moyens

Les 8 contours moyens de F0 (en demi-tons), correspondant aux 8 longueurs de phrase (de 1 à 8 syllabes)

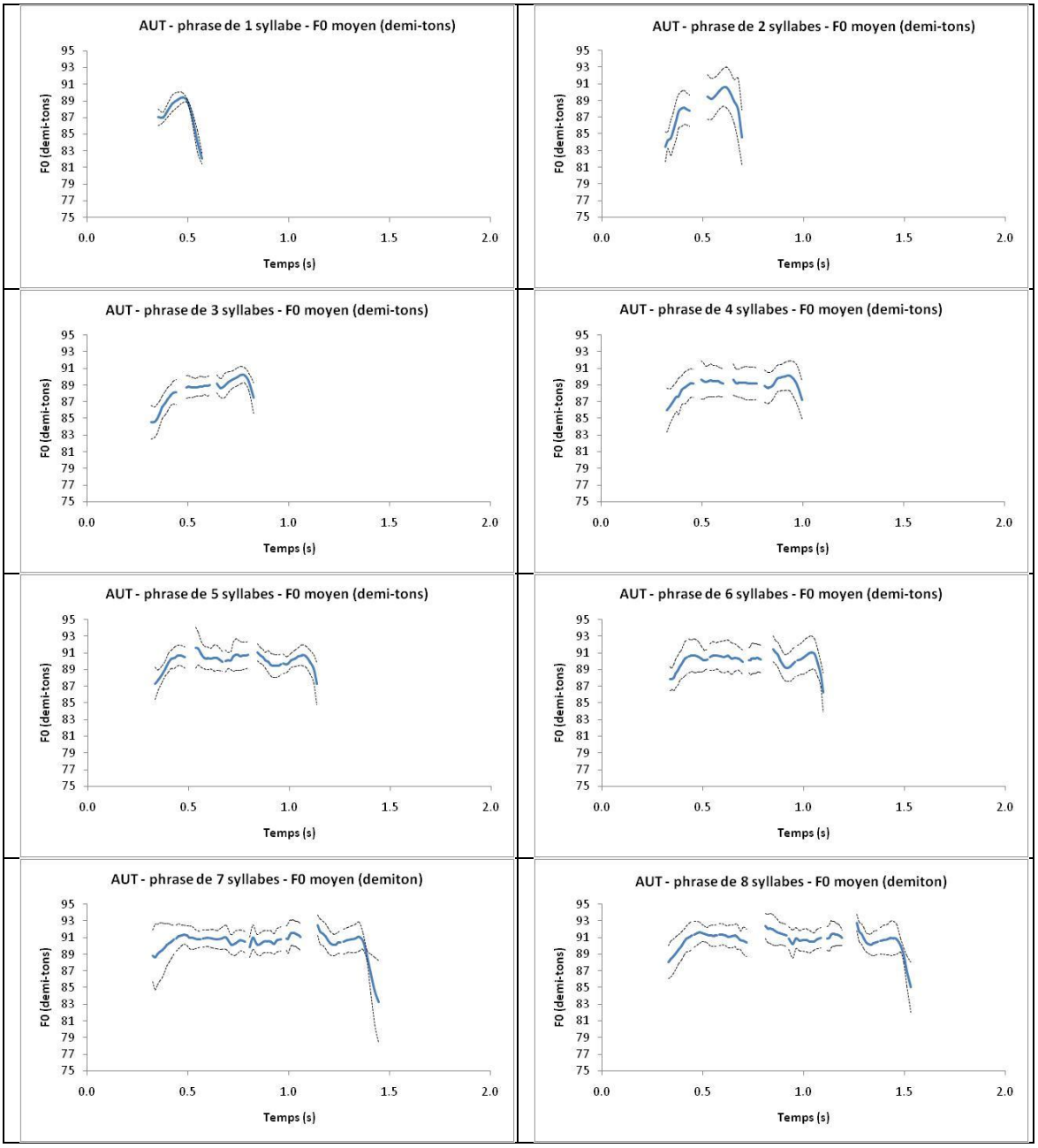
Déclaration (DEC)



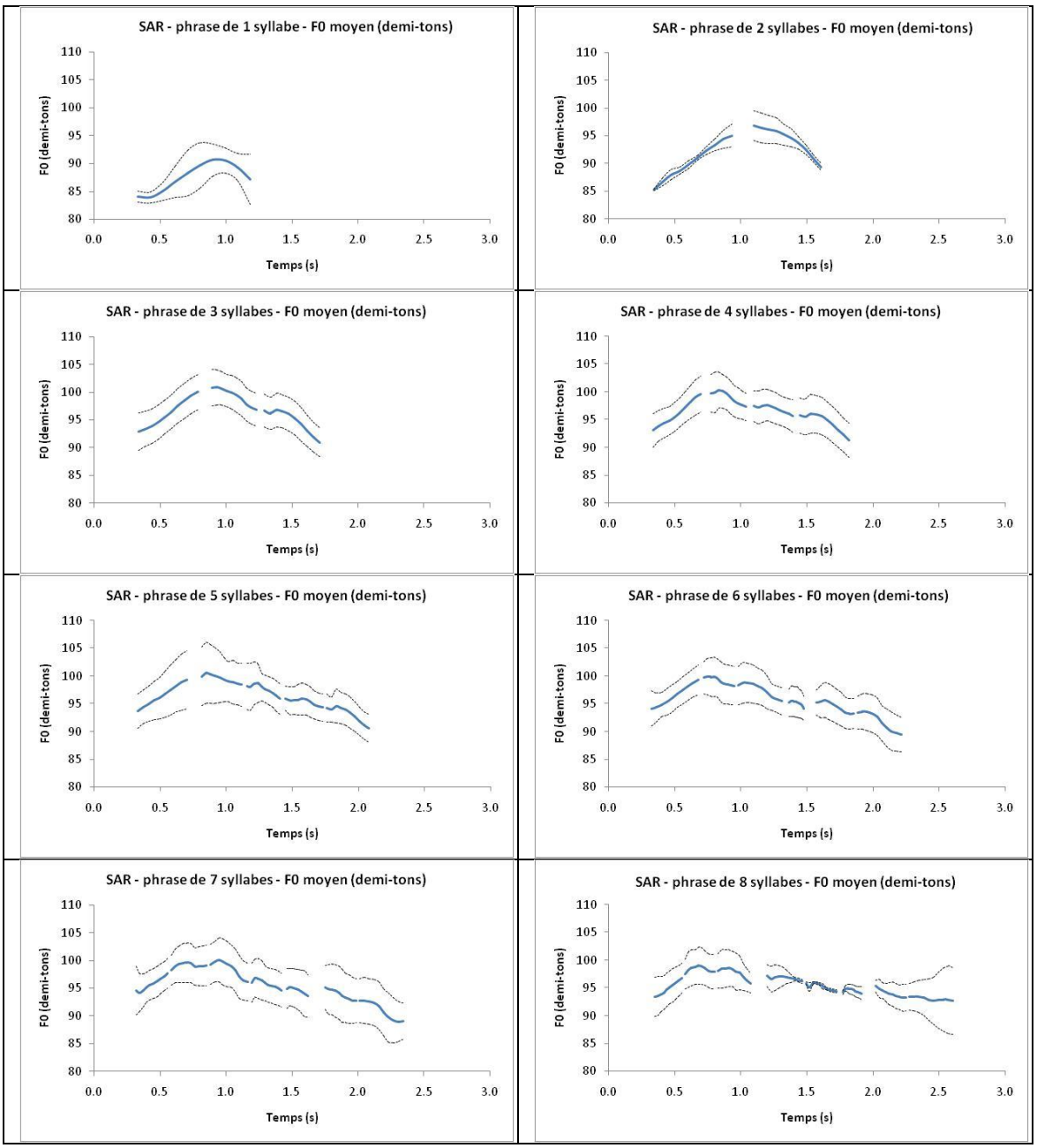
Surprise neutre (EXo)



Autorité (AUT)



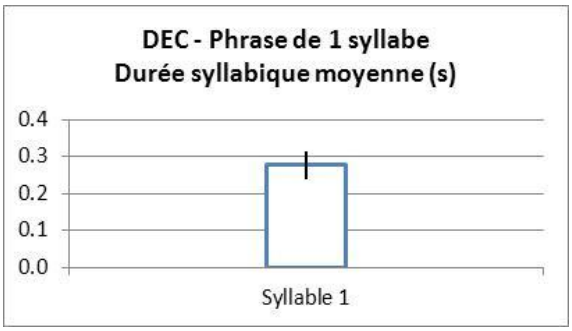
Ironie sarcastique (SAR)

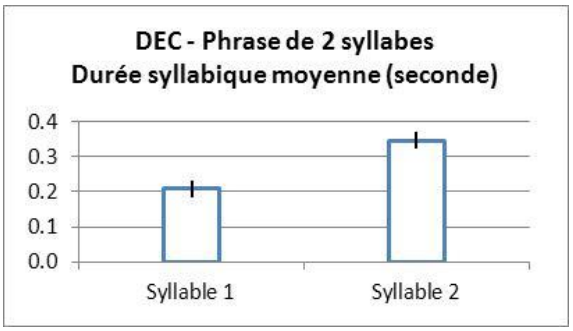


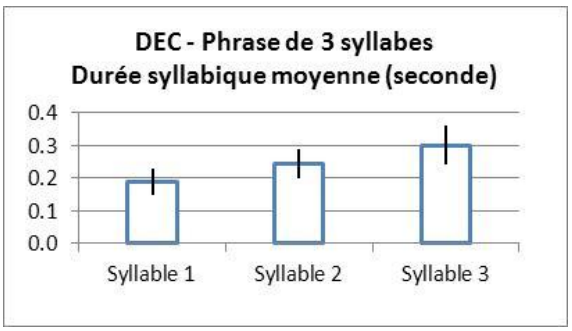
Annexe 3: Durée syllabique moyenne

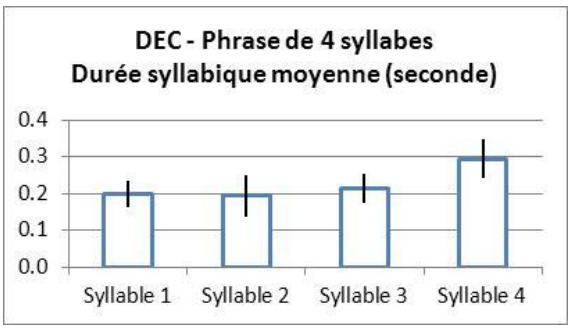
La durée syllabique moyenne (en seconde) des attitudes

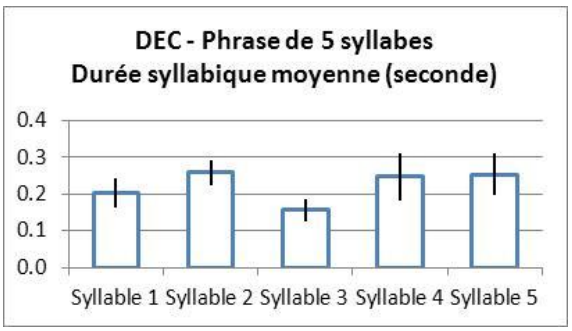
Déclaration (DEC)

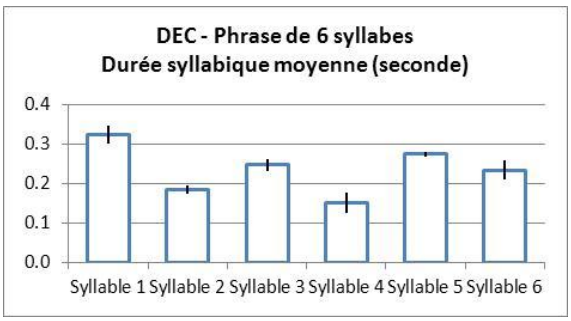


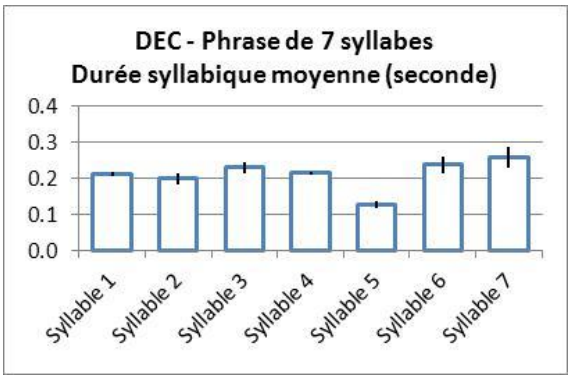


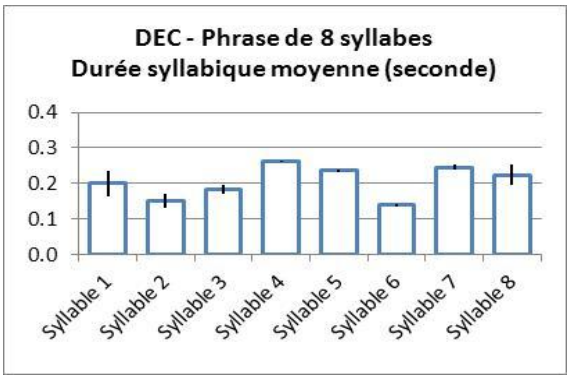




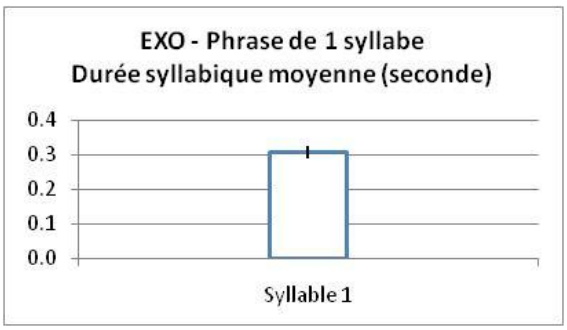


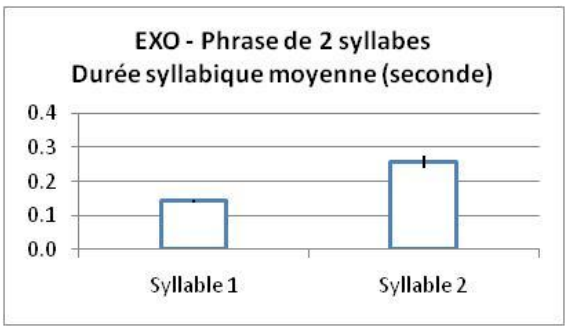


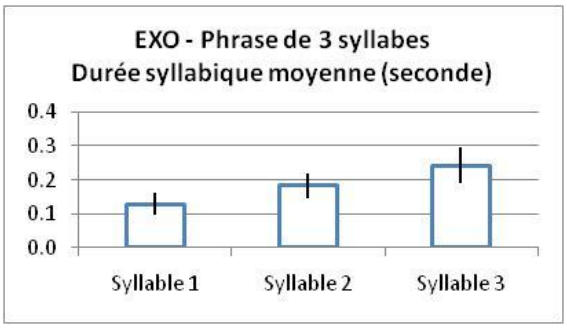


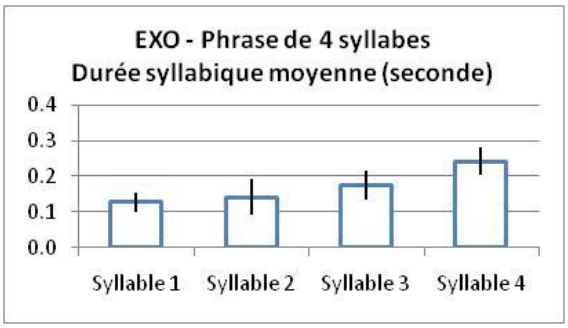


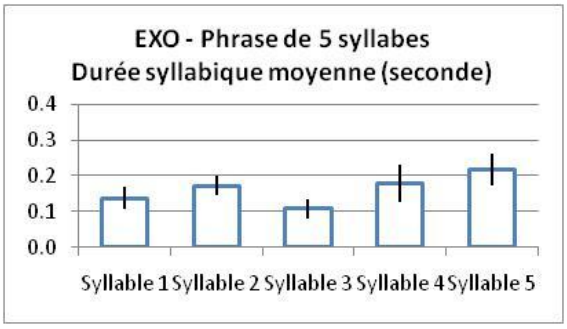
Surprise neutre (EXo)

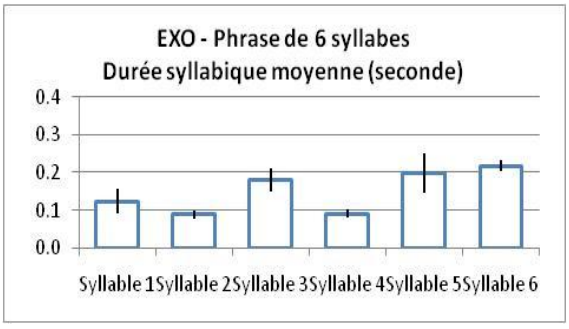


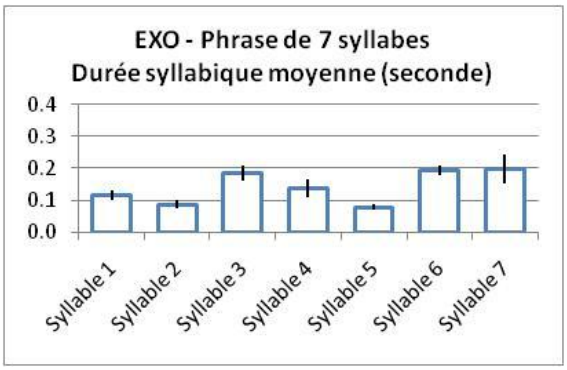


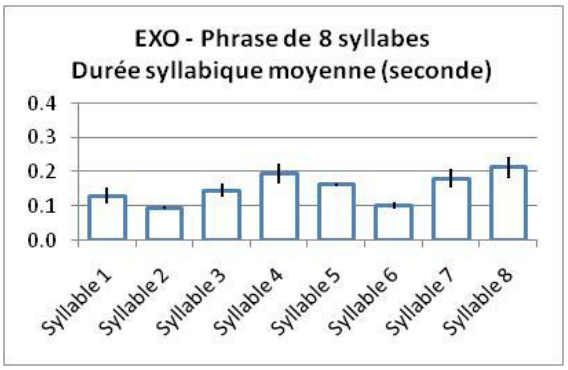




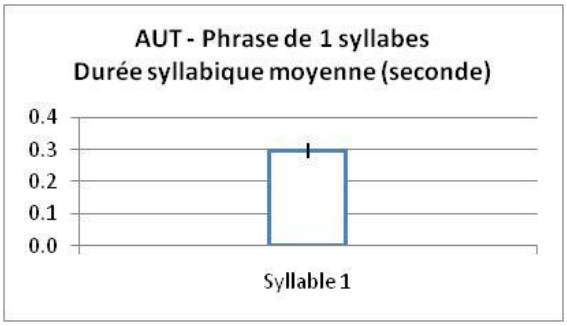


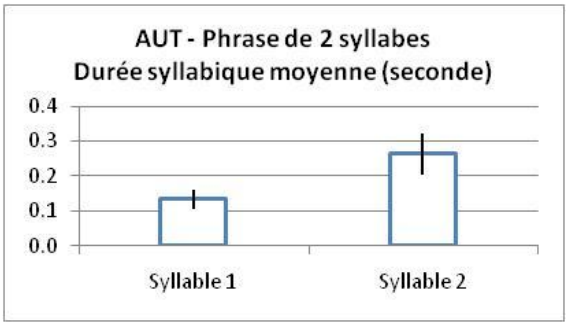


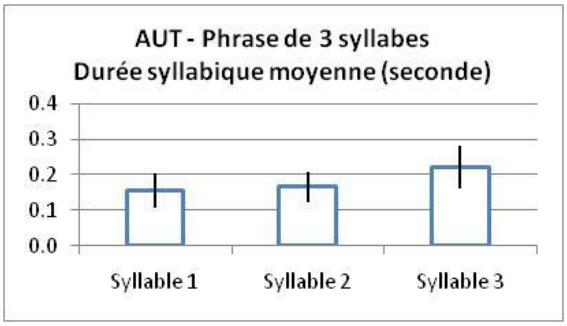


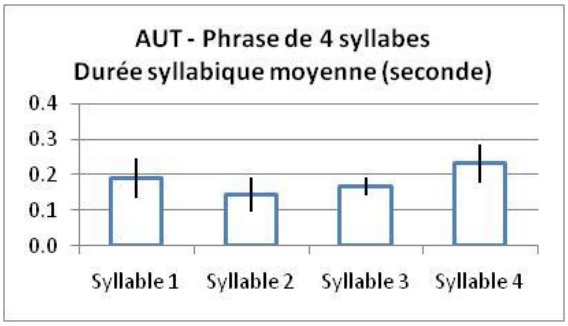


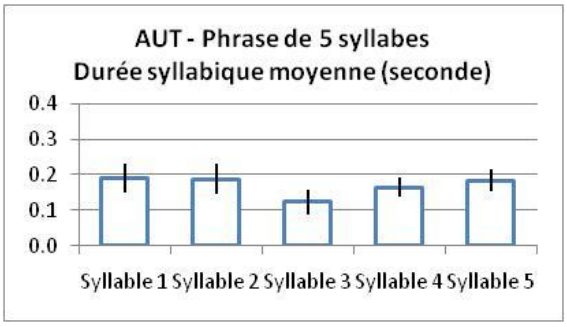
Autorité (AUT)

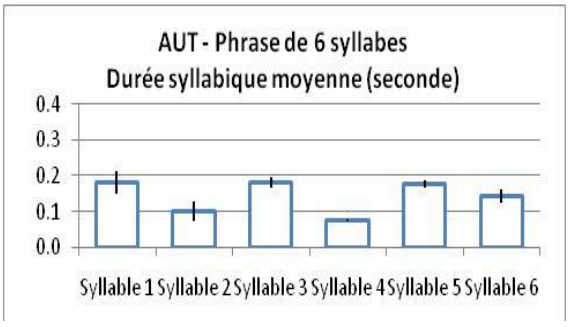


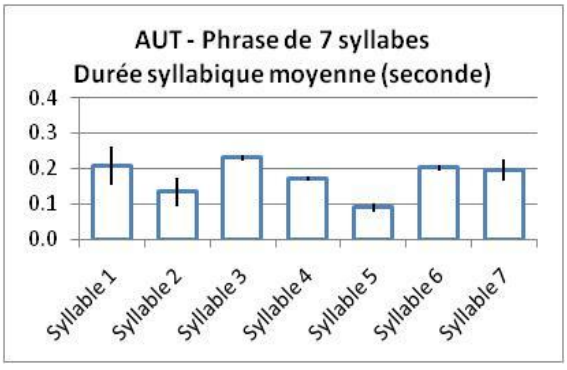


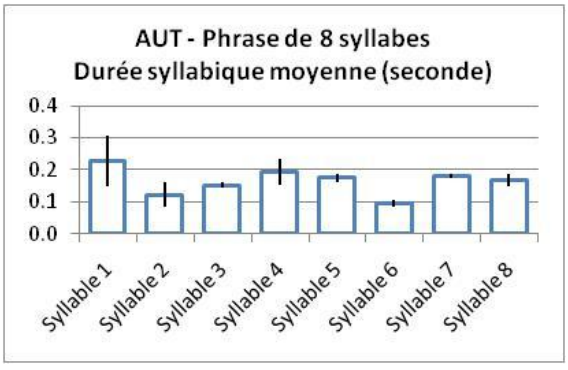




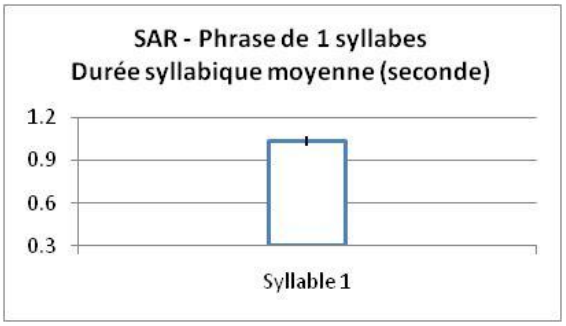


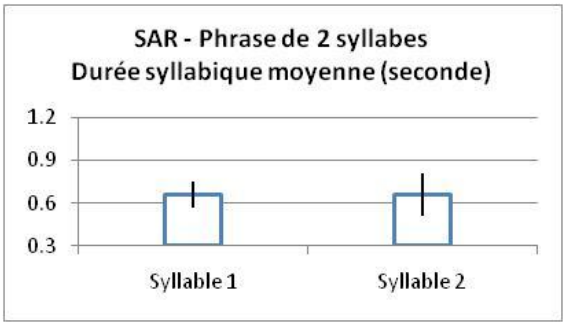


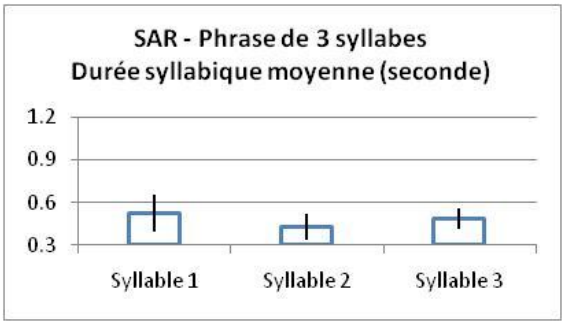


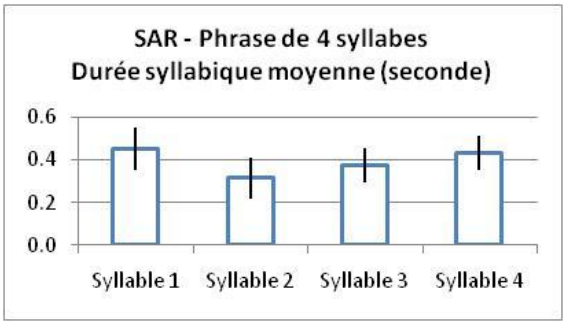


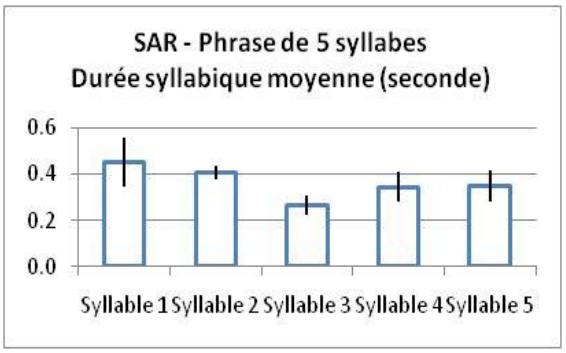
Ironie sarcastique (SAR)

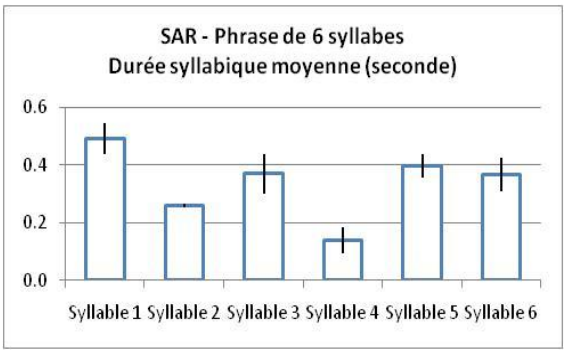


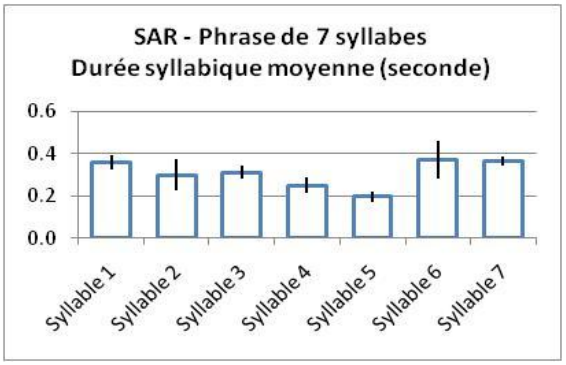


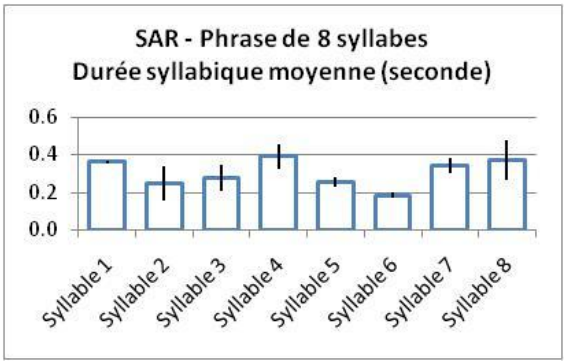








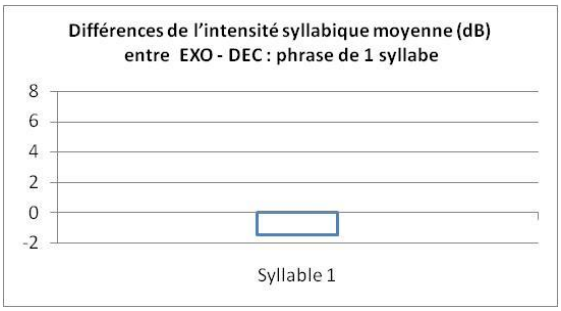


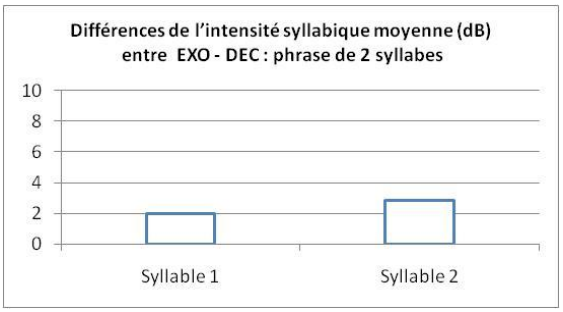


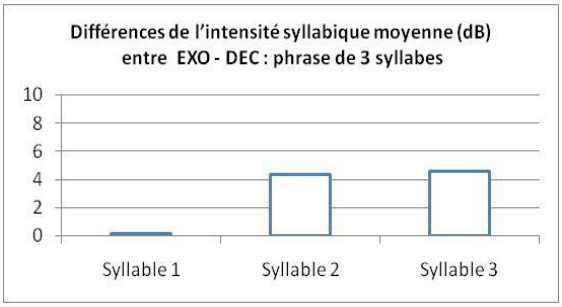
Annexe 4: Différences d'intensité syllabique moyenne

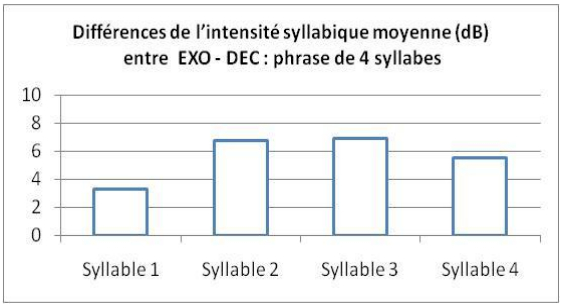
Différences d'intensité syllabique moyenne entre les attitudes sélectionnées et la déclaration

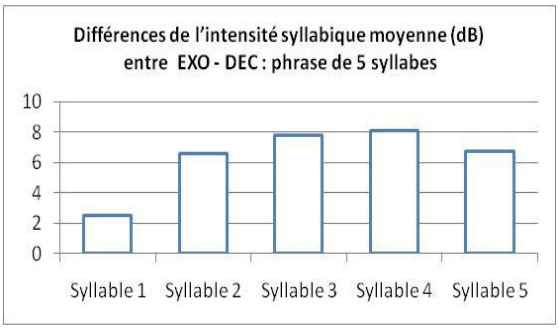
Surprise neutre (EXo)

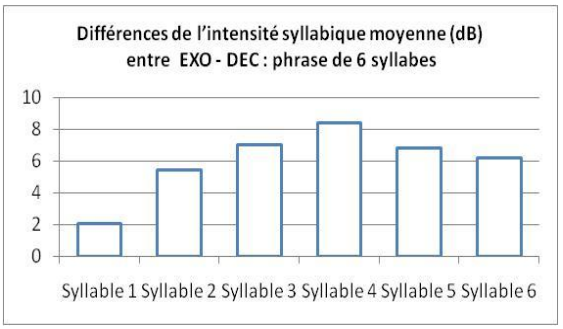


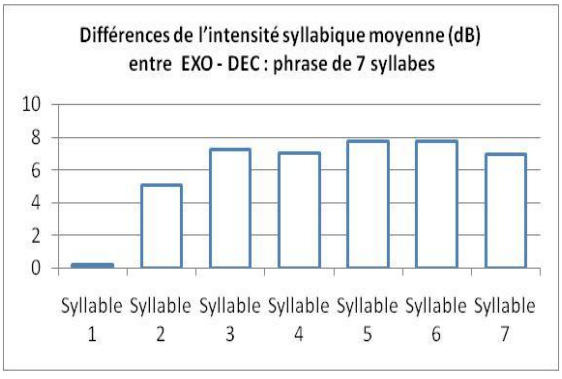


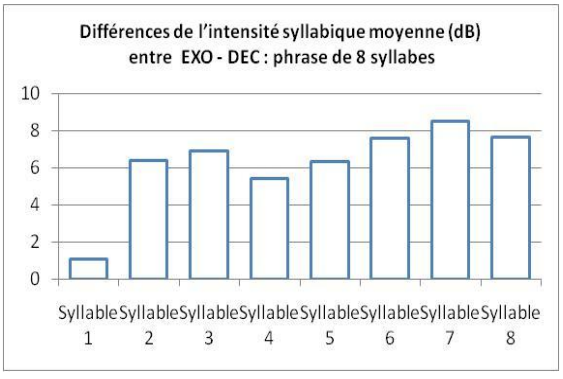




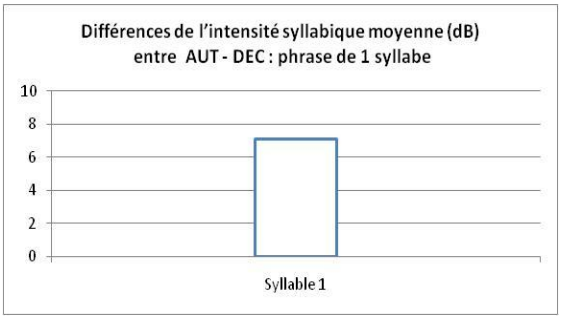


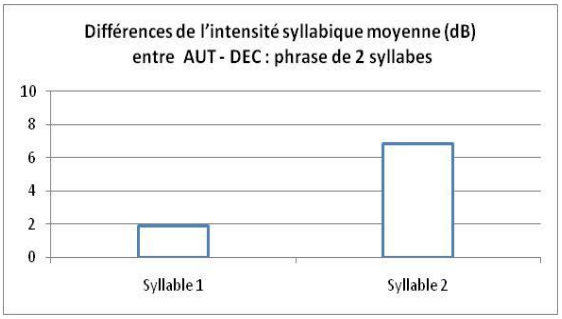


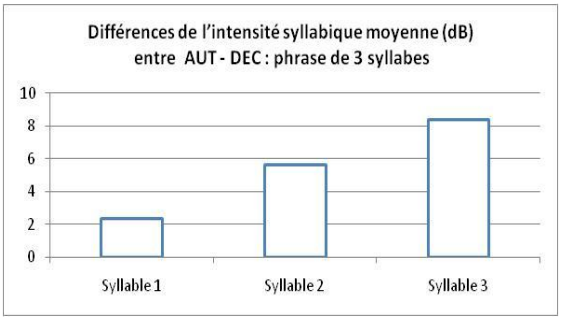


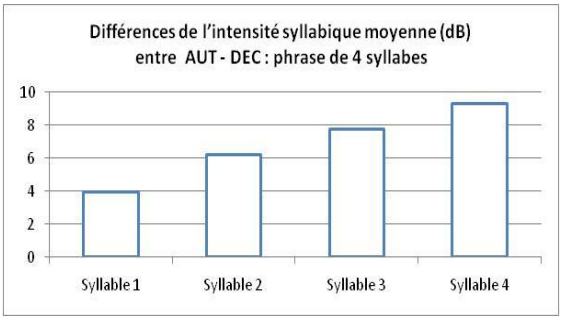


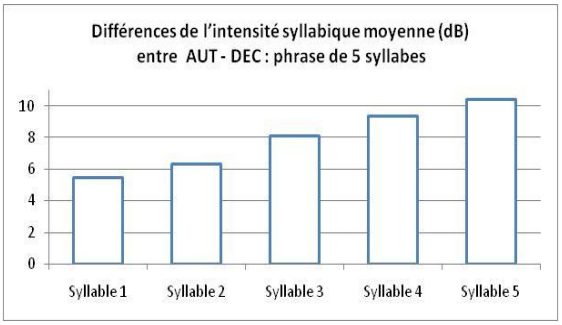
Autorité (AUT)

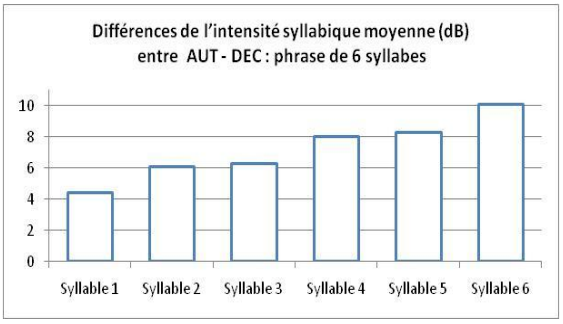


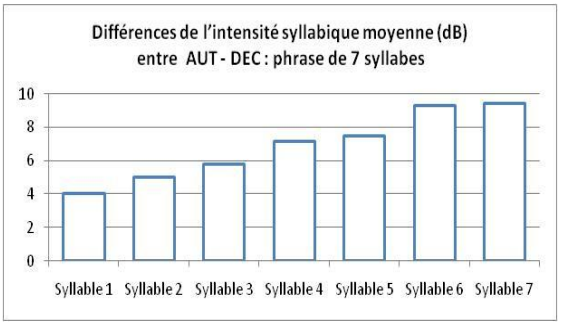


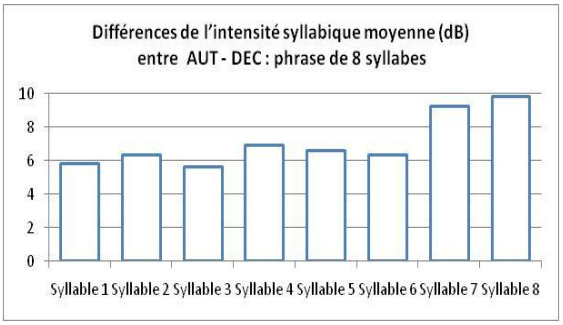




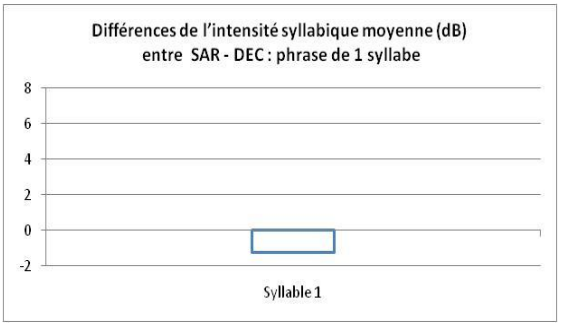


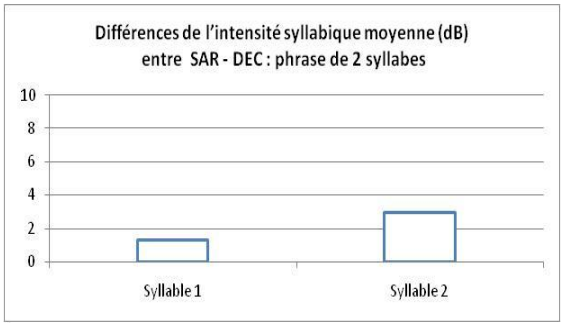


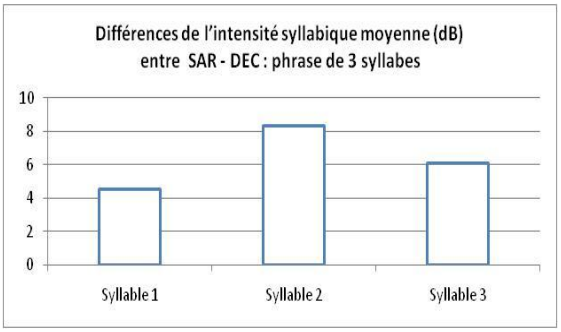


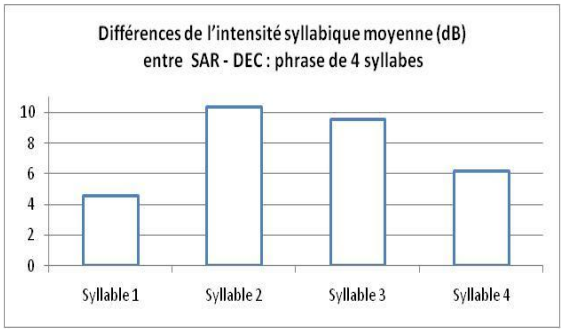


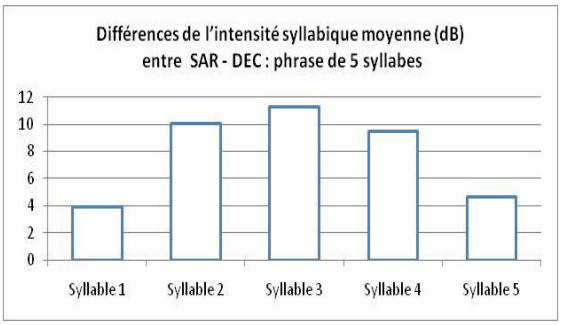
Ironie sarcastique (SAR)

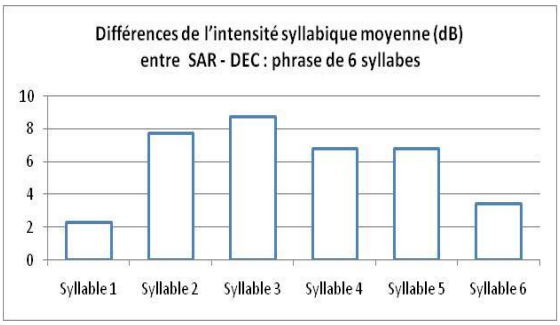


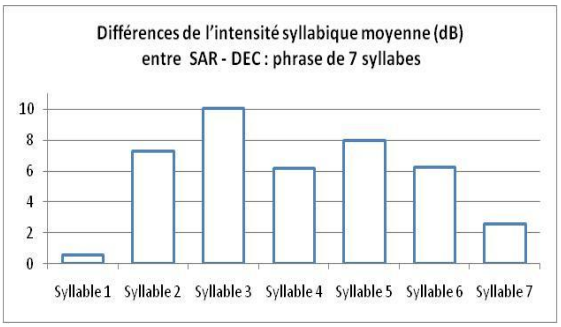


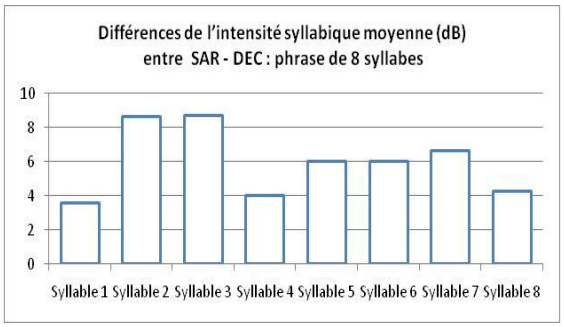












Annexe 5: Paramètres du modèle pour générer la prosodie des attitudes

Comme nous l'avons mentionné, notre corpus contient 16 affects sociaux. Les évaluations perceptives ont permis d'en dégager 4 parmi les 16 qui sont les plus performantes dans la modalité acoustique. Nous avons présenté dans le Chapitre 7 l'extraction des paramètres prosodiques et la modélisation du contour prosodique des 4 attitudes sélectionnées. En fait, notre modèle peut s'appliquer pour toutes les attitudes. Nous présentons dans cette annexe les paramètres de notre modèle pour générer la prosodie pour toutes les 16 attitudes. Ces paramètres pourront être utilisés pour notre travail au futur.

	Longueur de phrase	Valeurs moyennes des points stylisés pour les contours fonctionnels de F0										Valeurs moyennes du rapport de la durée syllabique (écarts types ≤ 0.1)			Les valeurs de la différence de l'intensité syllabique moyenne (dB) par rapport à la déclaration (écarts types ≤ 0.1)		
		F0 en demi-tons (écart type ≤ 0.2)						Distances relatifs (écart type ≤ 0.05				première syllabe	syllabe du milieu	dernière syllabe	première syllabe	syllabe du milieu	dernière syllabe
		$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	$\overline{P}_4$	$\overline{P}_5$	$\overline{P}_6$	$\overline{T}_{1-2}$	$\overline{T}_{1-3}$	$\overline{T}_{4-6}$	$\overline{T}_{5-6}$						
QUE	1 syllabe	1.2	1.5	1.5	1.8	2.8	1.5	0.3	0.5	0.2	0.1	0.9			0.5		
	2 syllabes	1.5	2.2	1.8	2.5	3.5	2.8	0.7	1.0	0.9	0.2	0.8		0.9	0.6		0.5
	≥ 3 syllabes	2.0	2.5	1.8	3.5	3.9	3.0	0.3	0.9	0.9	0.3	0.7	0.8	0.8	0.2	1.2	1.5
EXo	1 syllabe	5.6	3.9	4.2	4.8	6.3	6.3	0.2	0.4	0.4	0.2	1.1			-1.4		
	2 syllabes	6.8	6.6	7.9	9.1	10.1	9.8	0.3	0.9	0.7	0.1	0.7		0.8	2.0		2.8
	≥ 3 syllabes	7.6	8.8	9.2	11.0	11.7	11.0	0.3	0.9	0.9	0.2	0.6	0.7	0.9	1.6	6.9	6.3
EXp	1 syllabe	3.2	5.2	8.3	8.6	5.3	4.1	0.2	0.4	0.5	0.2	1.1			4.5		
	2 syllabes	5.3	6.8	10.9	12.1	91	8.8	0.4	0.9	0.9	0.5	0.9		1.0	2.5		3.0
	≥ 3 syllabes	5.6	8.5	9.2	10.0	12.7	11.3	0.5	0.9	0.9	0.4	07	0.7	1.0	1.8	2.2	3.3
EXn	1 syllabe	-0.5	0.5	0.4	0.4	1.2	1.5	0.3	0.4	0.3	0.1	0.9			-1.8		
	2 syllabes	1.0	0.8	1.1	1.2	1.8	1.5	0.5	0.9	0.9	0.5	0.9		0.8	-0.8		-0.7
	≥ 3 syllabes	-1.0	0.2	1.1	1.0	1.5	1.4	0.3	0.9	0.9	0.5	1.0	0.9	0.8	-1.0	0.1	0.2
EVI	1 syllabe	1.5	2.5	4.6	3.2	2.5	-2.2	0.3	0.5	0.3	0.1	0.8			1.9		
	2 syllabes	0.1	5.2	4.5	4.7	6.8	-1.2	0.6	0.9	0.9	0.5	0.7		0.8	1.3		2.2
	≥ 3 syllabes	-0.5	4.3	3.2	5.2	7.2	2.3	0.3	0.9	0.9	0.3	0.8	0.8	0.7	1.2	1.8	2.3
DOU	1 syllabe	-1.5	-1.2	0.3	1.2	2.8	4.5	0.3	0.5	0.3	0.1	1.6			-1.5		
	2 syllabes	1.6	2.2	3.5	4.5	3.2	6.5	0.5	0.9	0.9	0.7	1.1		1.9	0.5		1.2
	≥ 3 syllabes	5.6	7.2	7.6	8.3	8.6	9.6	0.3	0.9	0.9	0.6	0.8	0.8	1.1	1.6	2.1	2.3
AUT	1 syllabe	2.2	3.0	3.7	3.8	1.8	-3.6	0.2	0.2	0.4	0.2	1.1			7.1		
	2 syllabes	-0.3	-0.4	3.0	5.0	6.0	0.0	0.2	0.7	0.7	0.4	0.7		0.8	1.9		6.8
	≥ 3 syllabes	1.2	2.7	4.2	4.6	5.5	0.9	0.4	0.9	0.8	0.4	0.9	0.7	0.7	4.3	7.0	9.6
IRR	1 syllabe	0.6	0.8	1.2	1.1	3.2	-0.5	0.3	0.6	0.2	0.1	0.6			3.5		
	2 syllabes	-1.2	4.5	3.2	4.2	7.8	6.3	0.5	0.9	0.9	0.5	0.9		0.8	1.5		5.2
	≥ 3 syllabes	1.8	5.6	7.1	8.2	9.1	8.8	0.7	0.9	0.9	0.6	0.6	0.7	0.6	5.6	6.4	6.2

	Longueur de phrase	Valeurs moyennes des points stylisés pour les contours fonctionnels de F0										Valeurs moyennes du rapport de la durée syllabique (écarts types ≤ 0.1)			Les valeurs de la différence de l'intensité syllabique moyenne (dB) par rapport à la déclaration (écarts types ≤ 0.1)		
		F0 en demi-tons (écart type ≤ 0.2)						Distances relatifs (écart type ≤ 0.05)				première syllabe	syllabe du milieu	dernière syllabe	première syllabe	syllabe du milieu	dernière syllabe
		$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	$\overline{P}_4$	$\overline{P}_5$	$\overline{P}_6$	$\overline{T}_{1-2}$	$\overline{T}_{1-3}$	$\overline{T}_{4-6}$	$\overline{T}_{5-6}$						
SAR	1 syllabe	-0.8	1.5	5.1	5.0	3.7	1.5	0.3	0.6	0.2	0.1	3.7			-1.2		
	2 syllabes	1.5	3.2	10.3	12.7	10.6	4.9	0.3	0.9	0.9	0.5	3.2		1.9	1.3		3.0
	≥ 3 syllabes	7.7	9.1	14.1	9.1	9.1	5.5	0.3	1.2	0.9	0.6	2.0	1.5	1.5	3.2	8.0	4.6
MEP	1 syllabe	4.5	3.8	2.9	1.8	0.2	-1.2	0.3	0.5	0.3	0.1	1.8			1.2		
	2 syllabes	-1.1	0.8	1.2	2.2	1.2	0.1	0.5	0.9	0.9	0.5	1.2		0.5	-0.5		-1.2
	≥ 3 syllabes	1.1	5.6	10.2	4.2	2.1	0.2	0.3	0.9	0.9	0.6	1.8	1.2	1.0	1.8	1.5	-0.3
POL	1 syllabe	-1.1	-0.8	0.2	1.0	0.7	0.6	0.3	0.5	0.3	0.1	1.1			-0.8		
	2 syllabes	-1.5	0.5	-0.6	0.1	0.9	-1.1	0.5	1.0	0.9	0.4	0.9		1.1	-1.2		-1.9
	≥ 3 syllabes	-1.2	0.2	1.0	0.9	0.7	1.1	0.4	0.9	0.9	0.3	0.9	1.1	0.9	-1.5	-0.5	-0.5
ADM	1 syllabe	2.8	3.2	2.5	3.1	3.8	3.5	0.3	0.5	0.2	0.1	1.1			0.5		
	2 syllabes	3.5	2.1	3.8	5.8	4.5	2.5	0.3	0.9	0.9	0.5	0.9		1.2	1.5		1.5
	≥ 3 syllabes	3.2	5.2	7.8	8.3	7.5	7.1	0.3	1.0	0.9	0.6	0..9	0.8	1.0	1.2	1.5	1.5
MAT	1 syllabe	0.2	2.5	4.2	7.5	9.2	-2.5	0.2	0.4	0.3	0.1	3.5			2.5		
	2 syllabes	1.2	2.8	5.2	6.4	10.1	1.2	0.3	0.9	0.9	0.3	2.8		3.6	2.2		2.8
	≥ 3 syllabes	8.9	9.2	8.8	8.5	12.6	4.3	0.5	0.9	0.9	0.3	2.6	1.5	3.2	2.0	1.8	2.3
SED	1 syllabe	1.8	0.8	1.5	1.8	2.5	3.1	0.2	0.4	0.4	0.2	1.1			0.8		
	2 syllabes	-1.8	-2.2	-1.5	-0.5	1.2	0.8	0.5	0.9	0.9	0.4	0.9		1.0	-0.5		-1.0
	≥ 3 syllabes	1.2	2.2	0.5	1.0	2.5	0.8	0.4	1.2	0.8	0.4	0.8	0.7	0.7	0.5	0.8	-0.5
FAM	1 syllabe	-3.8	-3.3	-2.5	2.2	3.1	4.1	0.3	0.6	0.2	0.1	1.0			-2.0		
	2 syllabes	-5.2	-2.2	2.1	0.1	-0.5	0.5	0.3	0.9	0.9	0.5	1.1		1.2	-2.3		-3.1
	≥ 3 syllabes	-4.1	-3.2	-2.5	-1.5	1.2	0.5	0.5	0.9	0.9	0.6	1.5	1.2	1.3	-1.1	-1.0	0.2