La modélisation de l'intonation pour la synthèse de la parole

LA MODÉLISATION DE L'INTONATION POUR LA SYNTHÈSE DE LA PAROLE

Cet article présente une revue des méthodologies les plus importantes dans le domaine de la modélisation de l'intonation de la parole. De plus, il illustre l'utilisation d'un modèle particulier pour la synthèse du français. Les caractéristiques principales des quatre approches de prédiction de la mélodie (Pierrehumbert, IPO, ICP et INTSINT) sont décrits et sont comparés à l'algorithme de Fujisaki. Ce dernier est examiné plus en détail. Après une discussion des suppositions de base générales, son application au français est illustrée à l'aide d'un exemple.

1. Introduction

Une partie importante de la communication parlée n'est pas tant en relation avec les sons isolés, les syllabes, les mots et les phrases, qu'avec la manière dont tous ces éléments sont reliés entre eux, ce qui se réalise au moyen du rythme, de la mélodie et des accents. Ces phénomènes de mise en relation sont souvent désignés comme phénomènes suprasegmentaux — au sens où leur étendue n'est pas liée aux limites d'un segment sonore —, ou bien encore comme indices prosodiques de la parole.

La prosodie occupe une part majeure dans la parole de tous et chacun, ce qui pose par là un des problèmes les plus difficiles à résoudre pour les chercheurs en phonétique et en linguistique. Un certain nombre de ces complexités frappantes sont en relation avec l'*intonation* ainsi que le contrôle de la *mélodie* de parole. Tout locuteur / auditeur comprendra le rôle crucial de cette composante prosodique, si par exemple il a eu l'occasion d'écouter la mélodie «bizarre» des énoncés produits par un étranger, ou encore s'il a pu écouter l'intonation artificielle et monotone produite par une parole de synthèse de qualité moyenne.

Dans cet article, il sera présenté un exemple de ce que pourrait être un modèle intonatif de haute qualité, utilisable dans le cadre d'une synthèse de la parole.

2. Modélisation de l'intonation

Comme tout phénomène de parole, l'intonation peut être appréhendée de trois points de vue différents : l'angle articulatoire, l'angle acoustique ou l'angle de la perception. Le point de vue articulatoire comprend la génération de l'intonation dans le corps humain, via tout un système complexe d'impulsions nerveuses, de contrôles musculaires, de configurations du tractus vocal et d'autres éléments qu'il serait trop long de détailler ici. Le domaine acoustique recouvre l'étude de la transmission de la parole en tant que mise en vibration de molécules d'air. Enfin, le domaine de la phonétique perceptive comprend l'étude des processus de perception de la parole chez l'auditeur.

L'analyse de l'intonation sous l'angle articulatoire traite des différents moyens de faire varier la mélodie de la parole grâce, avant tout, à l'interaction entre la musculature du larynx et celle de la cage thoracique. L'analyse acoustique, à son tour, mesure les ondes sonores et étudie leurs structures de modulations de la fréquence fondamentale (« Fo », c'est-à-dire, principalement la vitesse d'ouverture et de fermeture des cordes vocales). Enfin, l'analyse perceptive considère les réactions que provoquent ces modulations du point de vue de l'auditeur impliqué dans la communication langagière.

Les approches adoptées pour modéliser l'intonation d'une langue spécifique sont beaucoup plus variées que ne le suggèrent ces trois démarches générales. La diversité de ces approches ne résulte pas uniquement des différences entre les niveaux phonétiques, ni seulement de différences entre les langues, mais elle provient aussi de points de vues divergents en ce qui concerne les buts, l'étendue et l'utilisation conceptuelle des modèles d'intonation. Aussi, il sera présenté dans les sections suivantes une revue de quelques modèles *prototypiques*.

a. L'école de Pierrehumbert

La thèse de Janet Pierrehumbert (Pierrehumbert, 1980) est à la base de l'une des approches les plus influentes et les plus largement utilisées dans l'analyse de l'intonation. Le système d'intonation de Pierrehumbert se représente comme une suite de tons hauts et bas. Ces tons peuvent se voir assigner différentes fonctions comme par exemple *l'accent mélodique, l'accent de syntagme ou l'accent de frontière de syntagme*. Ces tons H (haut) et B (bas) sont considérés comme des éléments de la structure phonologique profonde qui peuvent se transformer en une séquence

de valeurs de Fo grâce à une interpolation entre les cibles des tons H et B.

Un exemple de modélisation de l'intonation de l'anglais, du «type Pierrehumbert» comprend les éléments structuraux suivants :

• six types d'accent :

— deux accents, B* et H*, caractérisé par un seul ton sur la syllabe accentuée,

— quatre accents, B* + H-, B- + H*, H* + B- and H- + B*, caractérisés par un double ton, c'est-à-dire, une combinaison entre un accent à un ton (étoile) et un ton flottant qui n'est pas directement associé à une syllabe spécifique,

• deux types de tons de syntagme, B- et H-, reliant le dernier accent d'un syntagme avec la fin du syntagme,

• et deux types de tons de frontière, B% et H%, pour la fin du syntagme.

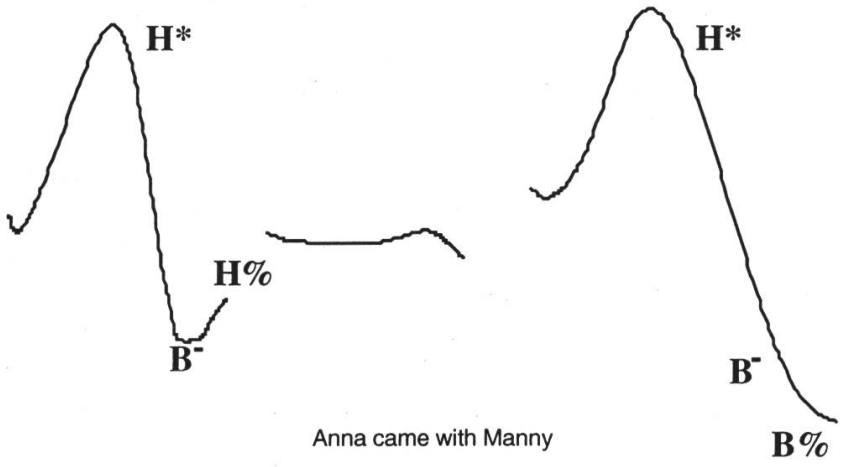


Figure 1. Exemple de contour de Fo, étiqueté selon le système de Pierrehumbert.

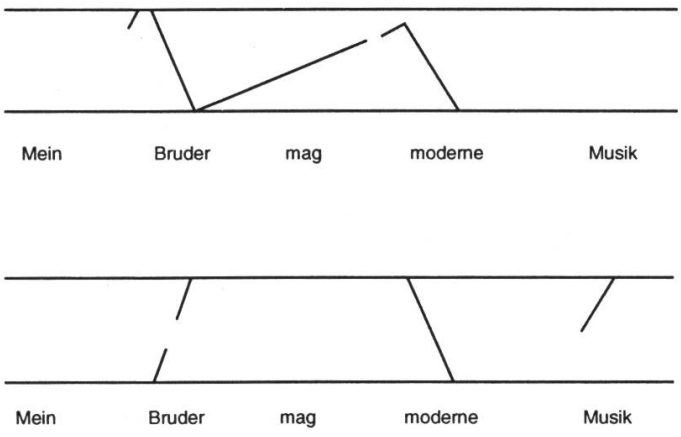


Figure 2. Deux contours de Fo différents, générés à partir des étiquettes de Pierrehumbert, pour la phrase en allemand «Mein Bruder mag moderne Musik».

La figure 1 montre l'étiquetage d'une phrase simple extraite de la thèse de Pierrehumbert. Les symboles sont attribués en fonction de plusieurs ensembles de règles qui recouvrent entre autres, l'initialisation des lignes de base et du haut, le «downstep» entre des accents consécutifs H*, ainsi que des restrictions distributionnelles.

Ce modèle est élégant par l'économie des paramètres utilisés — en effet, outre les H et B, seuls les symboles * (pour les accents), % (pour les frontières), - (pour annoter les tons) et + (pour combiner deux tons consécutifs) sont requis pour le marquage des éléments intonatifs et leurs fonctions, ainsi que pour la combinaison de ces éléments.

Par conséquent, la courbe complexe de Fo peut être exprimée comme une simple séquence de deux éléments de base. Cette réduction très notable de la complexité est probablement l'une des raisons de la popularité du modèle de Pierrehumbert, tant auprès des théoriciens phonologues que des ingénieurs travaillant dans les applications orientées vers le langage.

Un exemple d'application de cette méthode dans le cadre d'un projet actuel de synthèse de la parole[1] est montré dans la figure 2. Dans ce cas, la courbe d'intonation pour la phrase allemande «Mein Bruder mag moderne Musik» («mon frère aime la musique moderne») est variée par l'application de différentes valeurs cibles sur les noyaux syllabiques qui supportent un accent. La valeur cible de Fo est calculée directement à partir des tons de Pierrehumbert qui sont liés à la syllabe.

D'autres recherches dans ce même axe (si ce n'est avec quelques modifications et amendements mineurs) comprennent le travail de Ladd (Ladd, 1987), Gussenhoven (Gussenhoven, 1988), Beckman (Pierrehumbert & Beckman, 1988) et d'autres. Aussi, le développement du système d'étiquetage prosodique ToBI («Tones and Break Indices», voir (Silverman *et al.*, 1992)), qui est appliqué à un nombre de plus en plus important de corpora, doit beaucoup aux concepts de Pierrehumbert.

Ceci dit, l'un des inconvénients majeurs de cette approche est le manque de définition du processus de génération de la Fo. Par exemple, si l'on tente d'appliquer ce système à la génération de courbes Fo en synthèse de la parole, les règles d'interpolation sont souvent arbitraires et l'imprécision des cibles tonales par rapport aux valeurs empiriques de Fo constitue un sérieux obs-

1. Produit à l'institut IMS de l'Université de Stuttgart.

tacle. Par exemple, un H peut en fait correspondre à une valeur de Fo plus basse que celle précédant un L, et il n'existe pas de règle explicite pour calculer l'abaissement du Fo requis[2]. Ces mêmes problèmes rendent le modèle de Pierrehumbert moins adapté au travail théorique que ce qui aurait pu être attendu après la première impression de simplicité et de concision.

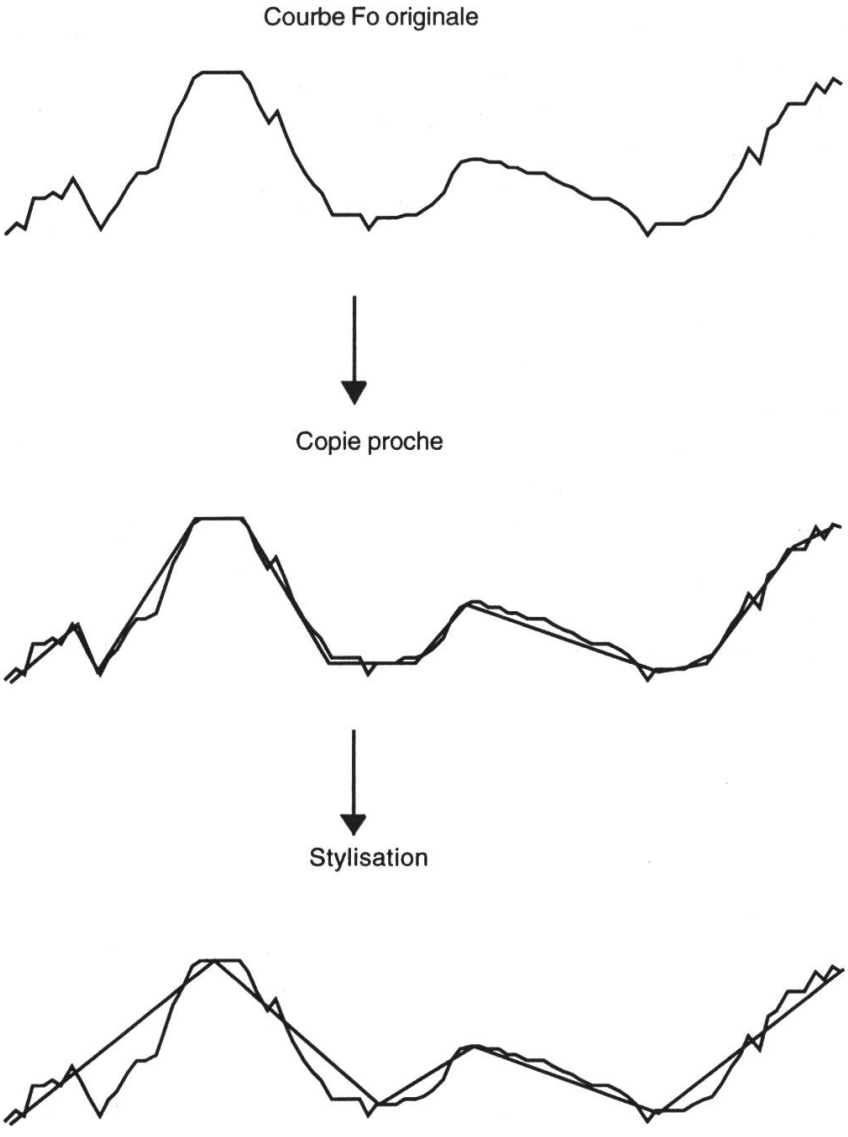


Figure 3. Un exemple de l'application du modèle IPO.

2. Voir Reinecke (1996) pour une tentative d'extraction d'information aussi quantitative que possible d'une courbe de F0 à partir d'une séquence d'étiquettes de Pierrehumbert. L'ampleur de la variation permise est très grande.

b. Le modèle IPO

L'approche développée à l'Institut pour la Recherche en Perception à Eindhoven ('t Hart *et al.*, 1990), propose une modélisation de l'intonation qui est résolument orientée vers la perception. Cette modélisation peut s'appliquer aussi bien à l'analyse qu'à la synthèse des contours intonatifs, et elle incorpore tout un ensemble de tests auditifs à tous les niveaux décisifs de la construction du modèle. La construction en particulier d'une *grammaire intonative* comprend les étapes suivantes :

• Une courbe Fo mesurée est légèrement lissée, puis convertie en une copie perceptiblement très proche.
• Cette procédure est répétée pour chaque type de courbe Fo qui est différent d'un point de vue perceptif pour une langue donnée.
• Cet inventaire des copies, aussi appelé *stylisations standard*, est classifié en fonction de trois paramètres : début, direction et vitesse du mouvement mélodique.
• Un système de règles de combinaisons est créé en utilisant des tests de perception. Ce système doit être capable de générer tous les contours types de Fo pertinents et uniquement ceux-là.
• Finalement, par tri et mise en correspondance des expériences, on obtient une plus grande formalisation des contours, les *patrons intonatifs*.

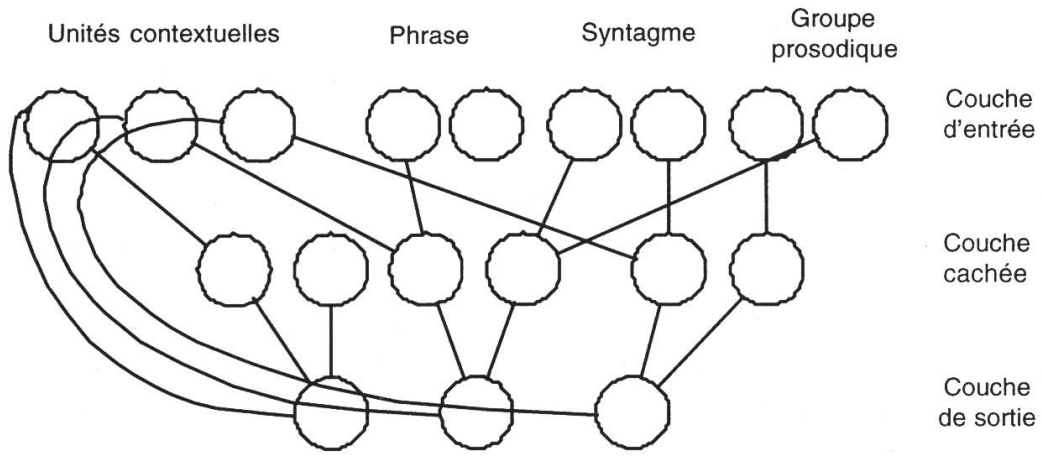
Un exemple des quatre premières étapes est montré dans la figure 3.

La beauté de cette méthode de modélisation repose sur l'application conséquente et continue de l'auto-contrôle et de l'auto-correction grâce à des tests perceptifs. Elle repose aussi sur sa clarté formelle, son état d'achèvement, ainsi que sa réversibilité. Ce système a été utilisé avec succès pour un certain nombre de langues européennes. Il reste qu'une question est toujours sans réponse et consiste à savoir dans quelle mesure le système serait suffisamment flexible pour adapter des données provenant essentiellement de différents types de langues. De plus, il est possible que des modulations mineures de Fo contribuent au naturel de la parole resynthétisée, sans pour autant être jugées différentes dans le contexte d'une expérience de pairage perceptif. Cela étant, le plus gros problème que pose cette approche est l'énorme quantité de travail requise pour effectuer les classifications répétitives des mouvements mélodiques ainsi que les vérifications, étape après étape, avec des tests auditifs sophistiqués.

c. Les approches de l'ICP

Le modèle intonatif pour la synthèse de la parole en français utilisé à l'Institut de la Communication Parlée, à Grenoble, va bien plus loin que la phonologie du type Pierrehumbert dotée des seuls points cibles. L'approche d'Aubergé, de Bailly et d'autres collaborateurs (voir par exemple Morlec *et al.*, 1995 ; Auberge & Bailly, 1995) prend en compte tout le contour de la Fo au niveau de l'énoncé, de la phrase, du syntagme et du groupe intonatif pour construire une base de données hiérarchisée de toutes les *formes intonatives* possibles.

Cette base de donnés, ce « lexique structuré » de formes Fo, contient des classes séparées de prototypes de contours pour les différents niveaux linguistiques du groupe prosodique, du syntagme, et de la phrase. Les contours sont codés comme autant de séquences à trois valeurs de Fo : une valeur pour le début, une pour le milieu et une pour la fin de chaque voyelle. Lorsque l'intonation doit être générée, les contours appropriés pour chaque niveau sont sélectionnés et superposés les uns aux autres. Les valeurs manquantes entre les points fournis sont reconstruits par interpolation. La figure 4 montre l'application de chaque prototype pour chaque niveau qui contribue au contour intonatif.



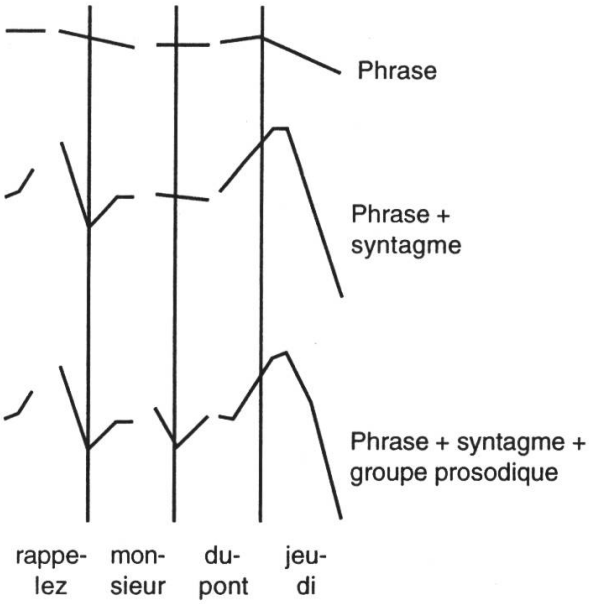


Figure 4. Page précédente: construction du contour de Fo avec un réseau de neurones séquentiel selon la méthode de l'ICP. Ci-dessus: le contour Fo par superposition de prototypes obtenus au moyen de ce réseau des neurones.

Cette méthode d'archivage d'un très grand nombre de contours intonatifs observés, puis de sélection des contours adéquats pour la synthèse de la parole, soit grâce à un jeu de règles, soit à l'aide d'un réseau de neurones, a été adoptée dans un certain nombre de systèmes (p. ex., voir aussi Traber, 1992). Probablement du fait de l'absence d'une théorie de type grammaire, cette méthode semble plus populaire auprès des ingénieurs que des linguistes. L'avantage de cette méthode est de fournir une intonation proche du naturel dans de nombreux cas. Ceci doit toutefois être pondéré par les désavantages de cette approche, à savoir un pouvoir d'explication faible et la quasi impossibilité de réduire les données archivées.

Une seconde approche développée parallèlement par l'ICP utilise les réseaux de neurone pour calculer les valeurs appropriées de Fo. Pour chacun des trois niveaux linguistiques considérés dans le modèle, deux paramètres alimentent le réseau : un paramètre prosodique, indiquant le degré de proximité entre les différentes unités en entrée, et un indice syllabique indiquant la distance syllabique entre la syllabe courante et le marqueur prosodique suivant.

Selon quelques expériences initiales de perception, l'approche par réseau de neurones semble tout aussi appropriée pour prédire

la courbe de Fo que l'a été la méthode du lexique structuré. L'inconvénient du faible pouvoir explicatif mentionné ci-avant s'applique également à la solution par réseau de neurones, du moins à première vue. Mais il faut prendre en compte les progrès accomplis dans le domaine de l'extraction automatique de règles pour de tels réseaux (Tickle *et al.*, 1996) qui pourraient potentiellement transformer les boîtes noires d'aujourd'hui en générateurs de systèmes de règles explicites pour l'avenir.

d. Le modèle d'Aix INTSINT

À l'Université d'Aix-en-Provence, un modèle bidirectionnel de l'intonation a été développé au sein d'un projet de synthèse multilingue (Hirst *et al.*, 1994; Hirst & Espesser, 1993). Durant le processus de l'extraction de la Fo, un contour d'intonation mesuré est automatiquement traduit en une séquence d'étiquettes prosodiques abstraites. Durant le processus de la resynthèse, la courbe de Fo est automatiquement générée à partir d'une séquence d'étiquettes (qui finalement peuvent être automatiquement générées à partir du texte écrit[3]).

Les étiquettes utilisées sont différentes des cibles tonales de Pierrehumbert à plusieurs égards. Non seulement, elles peuvent sans ambiguïté être transformées en contours de Fo ou être calculées à partir des contours de Fo, mais de plus, elles forment un inventaire structuré de catégories absolues et relatives. Les trois étiquettes utilisées pour la catégorisation absolue sont :

- T (pour "top", correspond à haut dans le registre du locuteur),
- B (pour "bottom", correspond à haut dans le registre du locuteur),
- M (pour "middle area", utilisé au début d'un énoncé).

Les autres catégories relatives sont :

- U (pour "step up", monter),
- D (pour "step down", descendre),
- H (pour "high"; maximum local),
- L (pour "low"; minimum local),
- S (pour "same as preceding"; identique à ce qui précède).

La figure 5 montre l'application de ces étiquettes au contour de Fo de la phrase test. La courbe lissée, qui est prétendue être difficilement discernable de l'original (Véronis *et al.*, 1997), est une

3. Voir Véronis *et al.* (1997), pour davantage d'informations concernant cet aspect du traitement linguistique.

stylisation automatique du contour mesuré, grâce à une technique mathématique appelée régression quadratique modale asymétrique. Cette technique permet de détecter les valeurs cibles de Fo et l'interpolation entre ces valeurs et une fonction spline. Les positions des valeurs cibles sont automatiquement codées en étiquettes INTSINT.

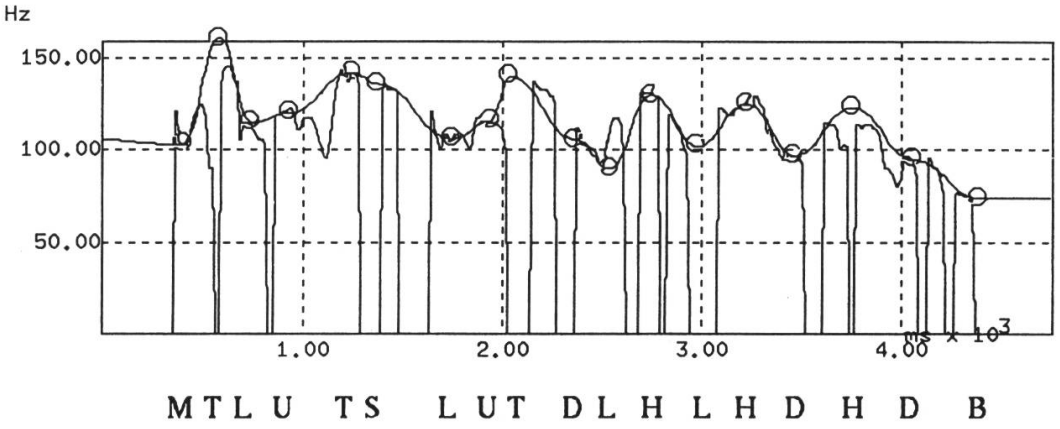


Figure 5. Fo de la phrase "Son étude ethnologique porte sur la relation entre les acupuncteurs et les cent'naires afghans" avec les étiquettes INTSINT et l'interpolation spline (produite par le programme MESSIGNAIX).

L'instauration d'une relation univoque entre les séquences d'étiquettes et la contrepartie en courbes lissées de Fo constitue certainement un des attraits majeurs de cette approche, tant pour la recherche fondamentale que pour la recherche appliquée. Ajouté au fait de l'équivalence perceptuelle entre la courbe brute et la courbe lissée, ce modèle semble offrir une alternative prometteuse aux autres modèles, même si les méthodes de génération de la parole pourraient encore être améliorées (Véronis *et al.*, 1997).

e. Résumé

Plusieurs remarques critiques peuvent être faites en ce qui concerne les modèles présentés jusqu'ici.

1. Ils ne prennent pas explicitement en compte le processus de production de la parole, i.e. l'aspect articulatoire.
2. Ils ne permettent pas la modélisation des caractéristiques spécifiques à un locuteur individuel.
3. Ils ne permettent pas la génération de valeurs à une résolution arbitraire, mais proposent soit un certain nombre de points cibles, soit des approximations de courbes lissées.

La section suivante décrira un modèle intonatif qui ne présente aucune de ces déficiences.

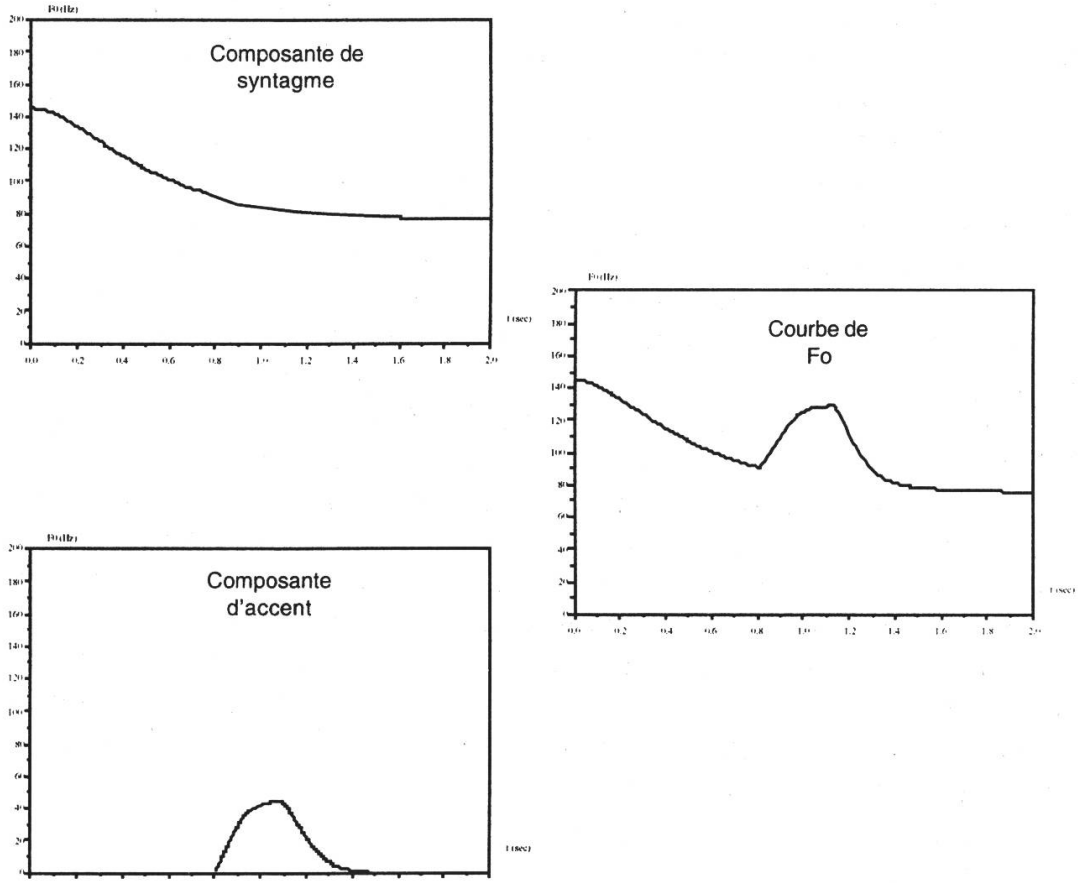


Figure 6. Les composantes d'une courbe de Fo selon le modèle de Fujisaki.

3. Le modèle intonatif de Fujisaki

a. Présentation générale

Le modèle phonétique d'intonation de Fujisaki utilise une approche qui diffère des modèles précédemment discutés en ce sens qu'il cherche à obtenir une description précise du contour de Fo. Pour cela, il se base sur une simulation des mécanismes articulatoires. Conçu comme une représentation fonctionnelle de la production de la mélodie de la parole, ce modèle simule l'activité des deux groupes musculaires fixés aux cordes vocales, les plus importants d'un point de vue phonatoire. Cette approche permet d'approcher avec beaucoup de précision les contours de Fo produits naturellement, en n'utilisant qu'un nombre réduit de paramètres de contrôle.

Fujisaki (Fujisaki, 1997) a montré que son modèle quantitatif est un outil très utile pour l'analyse et la synthèse des contours complexes de Fo pour différentes langues. Il est basé sur un modèle filtre à structure hiérarchique conçu à l'origine pour le sué-

dois par Öhman (Öhman, 1967). Pour chaque point dans le temps, le modèle calcule la somme d'une valeur de base de Fo (Fmin), d'une composante de syntagme et d'une composante d'accent, comme montré dans la figure 6.

Ainsi, le modèle interprète les contours d'intonation comme une combinaison de deux types de contours différents, la composante de syntagme et la composante d'accent. La composante de syntagme représente la pente générale de Fo avec ses variations lentes dans l'énoncé, tandis que les changements locaux et plus rapides de Fo sont représentés par la composante d'accent. Ces modifications de Fo peuvent être reliées à la réalisation de syllabes accentuées qui se surimposent au contour global.

La structure du modèle de Fujisaki ne permet pas seulement des simulations extrêmement détaillées des contours de Fo, mais elle offre aussi le grand attrait — en particulier pour le linguiste — de manipuler individuellement les accents (pour les mots accentués, l'emphase, le contraste, etc.) indépendamment de la forme de la courbe générale de Fo, qui est plutôt reliée au mode de la phrase (déclaratif, interrogatif, etc.). Ceci sera expliqué plus en détail dans la section suivante.

b. Une description plus détaillée

Le processus de production de la parole dans le modèle de Fujisaki obéit à des mécanismes (consistant en filtres amortis) qui ont besoin en entrée de l'information de syntagme, d'accent et des données caractéristiques d'un locuteur. Puis un contour de Fo continu est calculé en sortie. Pour chaque syntagme, il doit être fourni une commande de syntagme, et pour chaque accent, une commande d'accent. Ces commandes discrètes indiquent au modèle quand un syntagme commence et où doivent être placés les accents dans ce syntagme, et à quelle hauteur et avec quelle raideur les courbes de syntagme et d'accent doivent être générées. L'information d'entrée doit être fournie sous la forme des paramètres[4] suivants :

Fmin	Valeur minimum dans l'étendue fréquentielle d'un locuteur X
a_i	Facteur d'amortissement pour la i.ème commande de syntagme
b_j	Facteur d'amortissement pour la j.ème commande d'accent
Ap_i	Amplitude de la i.ème commande de syntagme
Aa_j	Amplitude de la j.ème commande d'accent
TO_i	Alignement temporel de la énième commande de syntagme

4. Le paramètre constant γ a été omis dans cette liste. C'est une valeur non pertinente pour cette discussion.

$T1_j$ Début de la j.ème commande d'accent
$T2_j$ Fin de la j.ème commande d'accent

La manière dont ces différents paramètres interagissent et influencent la forme finale de la courbe Fo peut être mieux visualisée à partir de la figure 7.

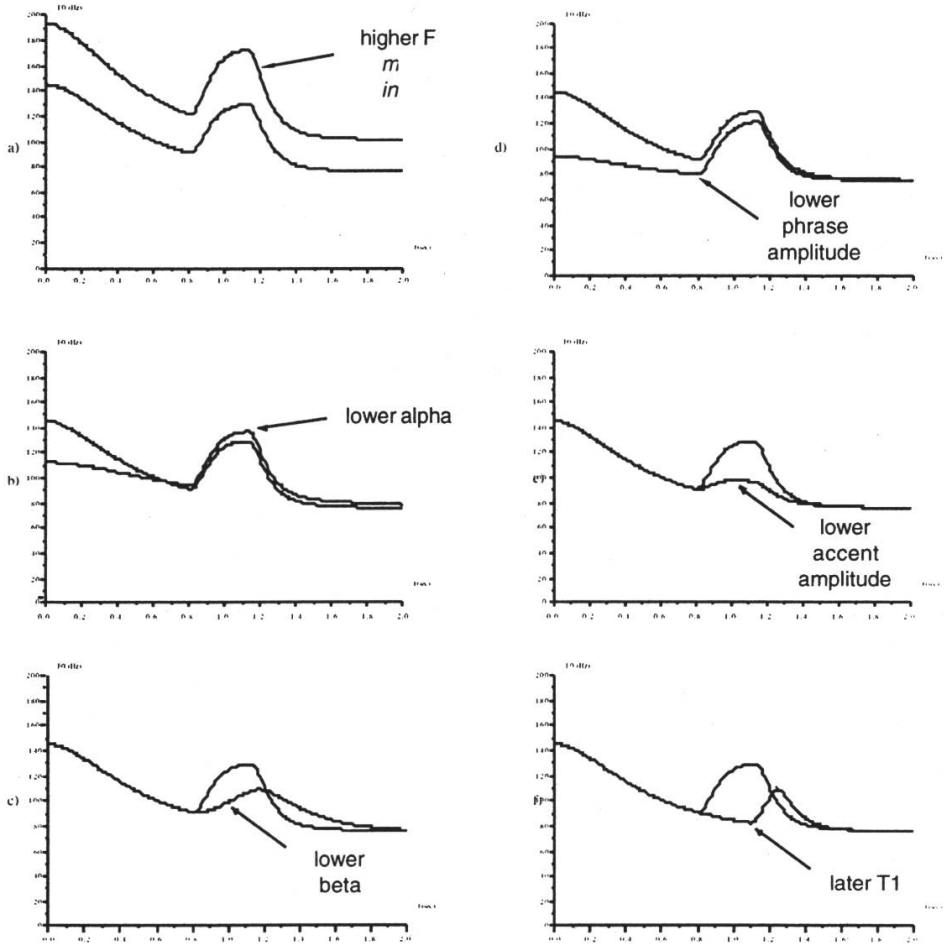


Figure 7. Fonctions des différents paramètres du modèle de Fujisaki pour un contour de Fo avec une commande de syntagme et une commande d'accent. a) Fmin, b) a, c) b, d) Ap, e) Aa, f) T1.

Un des avantages inhérents à ce modèle fondé sur des principes articulatoires est de simuler une observation universelle en matière d'intonation : la *pente de déclinaison de Fo*. Cette pente de déclinaison consiste en l'abaissement progressif du Fo qui se produit fréquemment chez un locuteur au cours d'un énoncé. Beaucoup des autres modèles doivent ajouter séparément cet effet *a posteriori* à leurs calculs, comme par exemple le « downstep » dans le système de Pierrehumbert. Dans le modèle de Fujisaki, la composante de syntagme génère cette déclinaison naturelle du Fo à partir de son mécanisme d'amortissement.

L'utilisation de fonctions discrètes comme entrée pour le mécanisme de l'accent est un autre avantage. Ce type de fonction a un début et une fin définis par les frontières de syllabe accentuée avec laquelle l'accent est associé. La durée de l'accent peut varier en modifiant les paramètres T1 et T2, et le niveau de proéminence de l'accent peut varier au moyen du paramètre d'amplitude Aa.

Le modèle de Fujisaki est le seul modèle capable de fournir une synthèse exacte d'un contour de Fo. En effet, il a été montré qu'il était possible, avec ce modèle, de s'approcher avec une très grande précision de contours réels de Fo pour toute une variété de langues (cf. par exemple Möbius, 1993 ; Taylor, 1994). Toutefois, il n'a pas toujours été précisé la quantité d'efforts nécessaires pour spécifier les paramètres d'entrée de façon à ce que la sortie soit optimale, ni selon quelle méthode cet objectif était atteint.

C'est pourquoi la prochaine section traitera en détail des questions de détermination et de standardisation des valeurs des paramètres du modèle.

c. Application du modèle de Fujisaki à une langue spécifique

Le problème essentiel de l'application du modèle de Fujisaki (ou de tout autre modèle intonatif) à une langue particulière vient de la complexité des corrélations entre les structures linguistiques et les structures intonatives. Malheureusement, les préjugés des chercheurs concernant la détermination des éléments d'énoncés pertinents au niveau intonatif — et inversement la détermination des contours intonatifs qui présentent un intérêt linguistique — constituent moins une aide qu'un obstacle supplémentaire sur le chemin de la solution simple, la perception de la parole n'étant pas directement équivalente aux mesures des courbes de Fo[5]. Par conséquent, une approche empirico-statistique systématique est requise. Même si ensuite les résultats de cette analyse mettent à mal certaines de nos idées préconçues (comme le marquage intonatif des mots accentués d'un point de vue sémantique ou pragmatique), nous ne pouvons pas nous permettre d'éviter cette phase de départ qui fait « table rase » de tout préjugé.

Pour appliquer ce modèle à une langue particulière, un certain nombre d'étapes doivent être suivies. Les paramètres du modèle de Fujisaki peuvent être extraits, après l'analyse empirique des contours de Fo dans la langue en question, ou plus exactement ils

5. Se reporter aux résultats des expériences en psycho-acoustique ('t Hart, 1976).

peuvent être inférés à partir des résultats de cette analyse. Le processus d'inférence est loin d'être immédiat. Généralement, il se déroule en trois temps au moins : une première approximation qui établit quelles sont les caractéristiques individuelles du locuteur les plus évidentes, une période de mesures supplémentaires, d'analyses et de calculs statistiques[6], et une phase de vérification finale avec l'appui de tests perceptifs. Les deux premières étapes seront décrites plus en détail dans les sections suivantes.

Réglage des paramètres : méthodes empiriques. Le premier pas est la mesure empirique de la Fo dans des données de langue naturelle. Malheureusement, la prise de telles mesures n'est pas aussi simple que cela peut paraître au premier abord, car il s'agit de capter les parties essentielles de la mélodie de la parole, et cela, à une grande précision. En dépit des nombreux efforts apportés au cours des dernières décennies, une méthode exacte ou à tout le moins globalement satisfaisante, pour déterminer le Fo à partir d'un signal acoustique n'existe toujours pas. Au lieu de cela, beaucoup de méthodes avec ajustement manuel des paramètres et avec beaucoup de résultats approximatifs différents sont en utilisation. Comme disait un chercheur bien connu dans le domaine de la fréquence fondamentale, la mesure du Fo « est un art, pas une science » (Hess, 1983).

De fait, l'inspection visuelle et la correction manuelle de la mesure de Fo sont inévitables. De plus, un lissage de la courbe peut être requis de façon à éliminer les fluctuations microprosodiques non pertinentes ou de consistance mineure.

Lorsque de telles précautions sont prises, l'extraction subséquente du paramètre Fmin — (cf. section 3.3.2), qui correspond à la fréquence fondamentale de base du locuteur, c'est-à-dire la plus basse valeur de Fo atteinte en situation de parole régulière — ainsi que les valeurs de durée et d'amplitude des composantes de syntagme et d'accent peuvent être calculées automatiquement sur un ordinateur en utilisant une application algorithmique ou un réseau de neurones.

Réglage des paramètres : valeurs liées au locuteur et autres valeurs de base. Du fait que les composantes de syntagme et d'accent sont simplement superposées par addition arithmétique, le calcul des paramètres de la commande de syntagme et de la valeur

6. Ou la confiance en la stratégie «essai - erreur» avec ajustement manuel des paramètres. Il y a de bonnes raisons pour ne pas choisir cette méthode (voir ci-dessus).

de base Fmin peut être séparé du calcul des paramètres de la commande d'accent. Le contour résultant du Fmin (qui peut être obtenu par comparaison directe) et des paramètres de syntagme est d'abord approximé à la courbe mesurée de Fo. Puis, lorsque les paramètres de la composante de syntagme ont été optimisés, la différence résiduelle sur la courbe peut être attribuée à la composante d'accent.

Outre le paramètre dépendant du locuteur Fmin et les deux paramètres de syntagme dépendant à la fois du locuteur et du mode de phrase, a et b peuvent aussi être calculés à ce stade. Il a été montré (Möbius, 1993) que ces deux paramètres peuvent essentiellement être considérés comme constants pour tout locuteur et mode de phrase donnés.

Réglage des paramètres : durée de l'accent et amplitudes. Pour chaque accent observé sur la courbe originale du Fo, une impulsion de commande d'accent appropriée doit être trouvée. Voici une façon fiable de procéder :

1. Trouver les endroits où le contour de Fo mesuré s'écarte clairement de la courbe de syntagme (qui est en construction, comme spécifié dans la section précédente).
2. Utiliser ces endroits pour les marques de durées comme paramètres T1 qui déclencheront le mécanisme de l'accent.
3. Trouver les endroits où la courbe de Fo mesurée se stabilise après ces montées.
4. Utiliser ces endroits pour les marques de durées comme paramètres T2 qui mettent fin au mécanisme de l'accent.
5. Calculer les angles de déclinaison reliant T1 à T2 (mathématiquement parlant, cela signifie prendre leur première dérivée ; sur un plan plus informel, cela signifie observer les angles de montée des courbes d'accents).
6. Utiliser les angles de déclinaison pour régler les paramètres d'amplitude Aa.

Réglage des paramètres : les règles linguistiques sous-jacentes. Les groupes intonatifs (séquences de mots comprenant une syllabe accentuée et dont l'accentuation repose sur une modification de l'intonation), les syllabes accentuées, les voyelles et autres sons voisés des syllabes accentuées sont quelques-unes de ces entités linguistiques et phonétiques évidentes qui doivent être identifiées. La durée de ces éléments phrastiques doit être systématiquement comparée avec la durée des commandes d'accent du modèle de Fujisaki afin d'obtenir les indices des structures intonatives pertinentes. Une comparaison systématique à ce

point consiste davantage qu'en une simple recherche des correspondances qui peut être minutieuse mais demeure toutefois informelle. Cela signifie l'exploitation conséquente des outils statistiques, comme par exemple l'analyse par corrélation et par régression.

Lorsque cette comparaison de la position temporelle et de la durée des portions voisées des événements linguistiques avec les données d'impulsion d'accent a donné un ou plusieurs systèmes de règles probables, ces systèmes sont utilisés pour générer artificiellement l'intonation.

La même démarche est appliquée pour l'amplitude d'accent : les connaissances sur la raideur des différents accents doivent être comparées avec les catégories des événements linguistiques. S'il y a une variation significative parmi les amplitudes d'accent, ceci doit être mis en relation avec un trait linguistique et c'est cette relation qui exprime une règle.

Problèmes du réglage des paramètres : le particulier vs. le général. Le modèle de Fujisaki est extrêmement flexible, et il est capable d'approximer pratiquement n'importe quel contour avec une grande précision (pour quelques exceptions, se reporter à Taylor, 1994). Mais cette capacité à imiter très fidèlement l'évolution d'une courbe Fo spécifique pose le problème de la généralisation et donc du pouvoir explicatif des valeurs paramétriques qui ont été extraites. Or, il peut s'avérer difficile de relier les configurations des paramètres acquis par approximation d'un grand nombre de contours de Fo produits naturellement par différents locuteurs dans différentes conditions. Ceci signifie que le chercheur doit être très prudent au moment de distinguer ce qui appartient à la modélisation d'un locuteur particulier de ce qui appartient à la construction d'un modèle intonatif pour une langue.

Mais ceci ne doit pas être vu comme un désavantage. Au contraire, une fois qu'un nombre suffisant de locuteurs ont été analysés, une recherche ultérieure peut être engagée avec des règles générales d'intonation plus précises que celles obtenues par l'approche traditionnelle qui présente cet inconvénient d'ignorer les différences individuelles pour viser d'emblée la généralisation.

d. Les problèmes de l'approche Fujisaki

Le modèle de Fujisaki présente toutefois certains désavantages. Sans doute le plus sérieux de tous est son incapacité de base à modeler certains contours de Fo qui peuvent être observés dans la parole naturelle. Un exemple fréquemment relevé dans la littérature

spécialisée (par exemple, Taylor, 1994; ou Ladd, 1995) est celui de la réalisation de l'accent mélodique («pitch accent») anglais avec une descente graduelle. Un tel mouvement de Fo ne peut en aucune façon être imité dans le cadre de Fujisaki. Sauf en admettant la possibilité de modifier continuellement dans le modèle les amplitudes d'accent ou quelque chose de similaire — ce qui serait en contradiction avec un certain nombre des règles fondamentales motivées par des considérations articulatoires — il n'y a pas moyen d'approximer de façon satisfaisante des contours de ce type.

Une autre objection concerne les définitions un peu floues de commande d'accent et de syntagme, et la frontière entre les deux. En effet, certains phénomènes ne peuvent être modélisés de façon acceptable qu'en voilant la distinction linguistique entre accent et syntagme. Pour d'autres phénomènes, c'est le concept articulatoire du mécanisme d'accent ou de syntagme qui doit être interprété avec liberté.

Un autre problème est celui du fondement articulatoire lui-même. Aucune évidence suffisante n'a été apportée jusqu'à présent qui accréditerait les thèses de Fujisaki. S'il est indéniable que son modèle capte certains des processus articulatoires qui sont impliqués dans le contrôle de l'intonation, il ne les capte pas tous. Et pour l'heure, aucun chercheur n'est en mesure de dire quels sont les plus importants. Par conséquent, l'avantage à intégrer la connaissance sur les processus articulatoires dans ce modèle le prédispose en même temps à des conclusions prématurées.

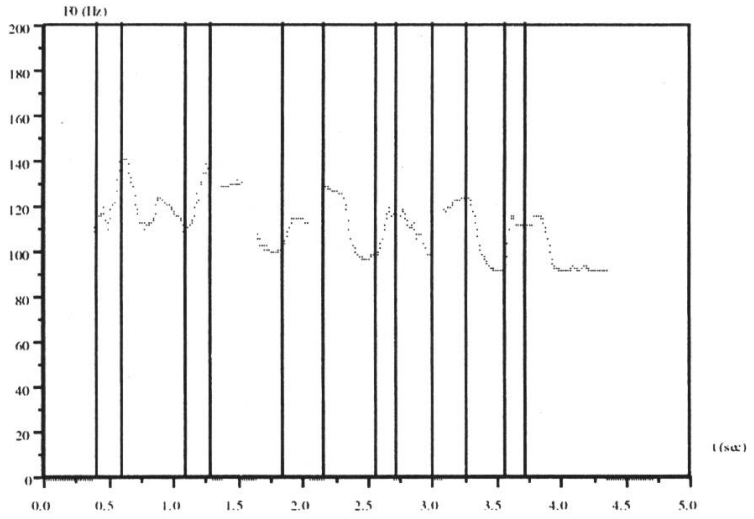


Figure 8. Fo de l'énoncé "Son étude ethnologique porte sur la relation entre les acupuncteurs et les cent'naires afghans" avec des marques des débuts et fins de montées.

4. Application du modèle de Fujisaki au français

Il y n'a eu auparavant qu'une seule tentative pour appliquer l'approche de Fujisaki à la modélisation de l'intonation du français. Dans sa thèse, Bailly (Bailly, 1983) a utilisé une implantation *ad hoc* du modèle, où les valeurs paramétriques ne sont pas exclusivement fixées sur la base des résultats d'une analyse empirique explicite mais plutôt sur la base d'heuristiques et de mise au point manuelle.

Dans ce qui suit, j'aimerais au contraire mettre l'accent sur l'application explicite des mesures et de leur évaluation, afin de parvenir à un algorithme de contrôle de génération de l'intonation pour le français.

a. Réglage des paramètres pour le français

L'application des étapes décrites dans la section 3.3 s'établit essentiellement selon les mêmes lignes, quelle que soit la langue choisie. Seules les interprétations linguistiques qui entrent plus tard dans le modèle varient.

Sur la figure 8, les débuts et les fins des montées de Fo sont marqués. Ces points correspondent aux limites temporelles des impulsions d'accent dans le modèle de Fujisaki. Lorsque ces événements cruciaux sont marqués sur la courbe intonative, il est tenu compte des imprécisions dans les mesures, des interférences microprosodiques et autres phénomènes qui compliquent l'interprétation de la courbe de Fo (cf. section 3.3.1).

La figure 9 montre la courbe Fo artificielle, générée par le modèle de Fujisaki, lorsque lui sont fournies les valeurs temporelles obtenues comme dans la figure 8, en tant que valeurs de début et de fin de commandes d'accents — la composante de syntagme ayant déjà été construite (cf. section 3.3.2). Remarquer que les amplitudes d'accent devront être calculées séparément pour les montées de Fo, sans quoi l'imitation de la courbe naturelle est loin d'être parfaite.

Après l'étiquetage d'un nombre suffisamment grand de courbes, on effectue l'analyse statistique : il s'agit de déterminer en quoi les positions des moments des accents sont corrélées aux indices phonétiques et linguistiques[7].

Dans le cas du français, la caractéristique principale des montées de Fo pour des phrases déclaratives semble être la position de

---

7. Les amplitudes d'accent, qui contrôlent la pente de montée de F0, ne sont pas mentionnées ici parce que les opérations nécessaires à leur calcul ne diffèrent guère de la détermination des paramètres temporels de l'accent.

la syllabe accentuée d'un *mot de contenu*[8]; on obtient alors la corrélation maximale. La syllabe accentuée correspond à la première syllabe dans des mots proéminents par emphase, sans quoi elle correspond plutôt à la dernière syllabe du mot dans les parties neutres de l'énoncé, la solution par défaut.

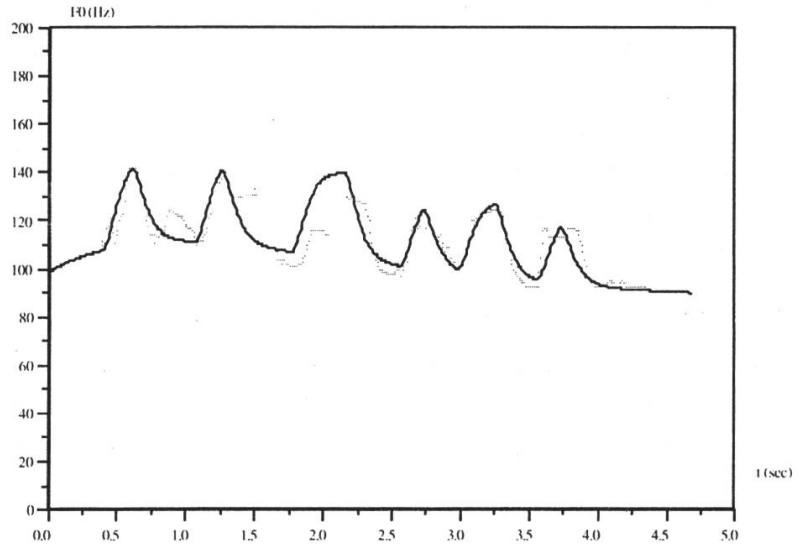


Figure 9. Courbe Fo de l'énoncé "Son étude ethnologique porte sur la relation entre les acupuncteurs et les cent'naires afghans" re-généré avec le modèle de Fujisaki et des amplitudes d'accent constantes.

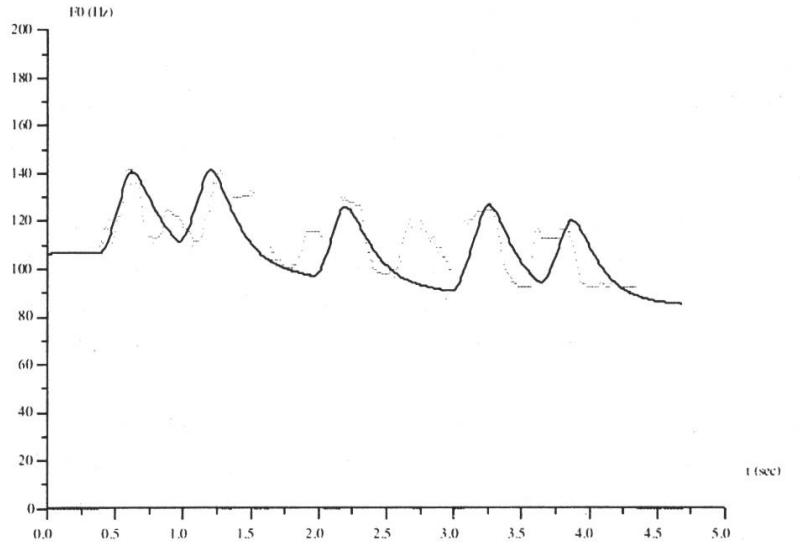


Figure 10. Courbe Fo de l'énoncé "Son étude ethnologique porte sur la relation entre les acupuncteurs et les cent'naires afghans" généré par règles avec le modèle de Fujisaki.

8. Par opposition aux mots fonction comme les articles, les conjonctions ou les prépositions.

b. Exemples de génération d'intonation du français avec le modèle de Fujisaki

La courbe de la figure 10 a été obtenue avec les paramètres temporels d'accents suivants :

- T1 Début de la syllabe accentuée — 80 ms — log10 (vowel duration)
- T2 Fin de la voyelle accentuée — 100 ms

Ces paramètres sont le résultat d'une étude pilote portant sur 50 phrases comprenant un seul syntagme et lues par un même locuteur. Il faut noter que plus le nombre de phrases analysées dans ce cadre augmente, plus la corrélation et même certaines valeurs des paramètres sont susceptibles de fluctuer.

Les courbes de Fo originales et celles générées par règles ne sont pas seulement visuellement très proches, elles sont aussi indiscernables au plan perceptif: des auditeurs testés sont incapables d'identifier correctement et avec régularité les versions naturelles et artificielles.

5. Conclusion

La modélisation de l'intonation est une tâche complexe et malgré les nombreux progrès effectués ces dernières années en linguistique, en phonétique et dans les technologies de la parole, la recherche actuelle sur l'intonation n'est toujours pas en mesure d'offrir des solutions complètes. Toutefois, il existe quelques approches prometteuses comme il a été montré dans les paragraphes précédents. Que ce soit le travail effectué dans le domaine de l'analyse phonologique, inspiré par Pierrehumbert, ou les systèmes de catégorisation basés sur d'importants tests perceptifs comme ceux développés à l'IPO, ou les méthodes de l'ICP pour la génération de la Fo via des prototypes extraits d'un lexique structuré, ou via un système connexionniste, ou la conversion bidirectionnelle entre des étiquettes bien définies et des courbes de Fo dans le projet INTSINT, ou encore la tentative de Fujisaki de construire un modèle quantitatif continu sur des fondements articulatoires, tous fournissent d'importants éclairages sur la communication humaine, et des outils valables pour la simulation partielle.

L'approche Fujisaki a révélé plusieurs avantages par rapport aux autres approches, en particulier dans le domaine de l'applica-

tion à la synthèse de la parole. Il a été montré comment cette méthode pouvait être aisément appliquée avec efficacité à une langue donnée et générer une intonation proche d'une intonation naturelle.

D'un autre côté, il a été aussi montré qu'en dépit de ses mérites évidents, l'approche de Fujisaki est loin de fournir une solution idéale pour tous les problèmes inhérents à la modélisation de l'intonation. Cette approche a des défauts, en partie ouverts à l'amélioration au sein du cadre, et en partie inhérents à la conception structurelle.

Ainsi, il se pourrait que les développements et progrès futurs dans le domaine de l'intonation émergent de l'interaction entre les différentes approches présentées (et quelques autres qui n'ont pas été mentionnées ici). De nouveaux modèles hybrides peuvent être construits à partir des composantes des modèles existants, tirant parti de leurs forces et évitant leurs faiblesses*.

Stefan WERNER
Département d'allemand
Université de Joensuu, Finlande
stefan.werner@joensuu.fi

* Cet article a été traduit en français par Brigitte Zellner. L'auteur exprime ses remerciements à la traductrice pour sa traduction précise et claire.

Références

AUBERGÉ, V., & BAILLY, G. (1995). Generation of intonation: A global approach. In *Eurospeech '95, Proceedings of he 4th European Conference on Speech Communication and Technology*. Madrid, 2065-2068.

BAILLY, G. (1983). *Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique.* Institut National Polytechnique de Grenoble.

BECKMAN, M. (1986). *Stress and Non-Stress Accent*, Dordrecht: Foris.

FUJISAKI, H. (1997). Modeling the process of fundamental frequency control of speech for synthesis of tonal features of various languages. In *Proceedings of the 1997 China-Japan Symposium on Advanced Information Technology.*

GUSSENHOVEN, C. (1988). Adequacy in intonation analysis: The case of Dutch. In H. van der Hulst & N. Smith (eds), *Autosegmental Studies on Pitch Accent.* Dordrecht: Foris, 95-121.

HESS, W. (1983). *Pitch Determination of Speech Signals*, Berlin: Springer.

HIRST, D., & ESPESSER, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. In *Travaux de l'Institut de Phonétique d'Aix.* Université d'Aix-en-Provence, 71-85.

HIRST, D., IDE, N., & VÉRONIS, J. (1994). Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project. In *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis.* New Paltz, NY, 77-81.

LADD, D. R. (1987). A model of intonational phonology for use with speech synthesis by rule. In J. Laver & M. Jack (eds), *Proceedings of the European Conference on Speech Technology.* 21-24.

LADD, D. R. (1995). "Linear" and "overlay" descriptions: An autosegmental-metrical middle way. In Kjell Elenius & Peter Branderud (eds), *Proceedings of the XIIIth International congress of Phonetic Sciences, vol. 2.* Stockholm: KTH, 116-123.

MÖBIUS, B. (1993). *Ein quantitatives Modell der deutschen Intonation. Analyse und Synthese von Grundfrequenzverläufen.* Tübingen: Niemeyer.

MORLEC, Y., AUBERGÉ, V., & BAILLY, G. (1995). Evaluation of automatic generation of prosody with a superposition model. In Kjell Elenius & Peter Branderud (éds.) *Proceedings of the XIIIth International congress of Phonetic Sciences, vol. 4.* Stockholm: KTH, 224-227.

ÖHMAN, S. (1997). Word and sentence intonation: A quantitative model (réimprimé). In J. Swedenmark & J. Anward (eds), *9 x Öhman, Reports from Uppsala University Linguistics, 30.* Uppsala: Uppsala University, 39-94.

PIERREHUMBERT, J. (1980). *The Phonology and Phonetics of English Intonation.* Cambridge, MA: MIT Press.

PIERREHUMBERT, J., & BECKMAN, M. (1988). *Japanese Tone Structure.* Cambridge, MA: MIT Press.

REINECKE, J. (1996). *Resynthese als Hilfsmittel bei der prosodischen Etikettierung*. Verbmobil-Report 162. Braunschweig: Universität Braunschweig.

SILVERMAN, K. E. A., BLAAUW, E., SPITZ, J., & PITRELLI, J. F. (1992). Towards using prosody in speech recognition/understanding systems: Differences between read and spontaneous speech. In *Proceedings of the Fifth DARPA Workshop on Speech and Natural Language*. New York: Harriman.

'T HART, J. (1976). Psychoacoustic backgrounds of pitch contour stylization. In *IPO Annual Progress Report 11*. Eindhoven: IPO, 11-19.

'T HART, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation.* Cambridge, UK: Cambridge University Press.

TAYLOR, P. (1994). *A Phonetic Model of Intonation in English.* Bloomington, IN: Indiana University Linguistics Club.

TICKLE, A. B., Orlowski, M., & Diederich, J. (1996). DEDEC: A methodology for extracting rules from trained artificial neural networks. In *Proceedings of on From Trained Artificial Neural Networks Workshop.* Brighton: University of Sussex, 90-102.

TRABER, C. (1992). F0 generation with a database of natural F0 patterns and with a neural network. In G. Bailly & C. Benoît (eds), *Talking Machines: Theories, Models and Designs*. Amsterdam: Elsevier, 287-304.

VÉRONIS, J., DI CRISTO, Ph., COURTOIS, F., & LAGRUE, B. (1997). A stochastic model of intonation for French text-to-speech synthesis. In *Eurospeech '97, Proceedings of the 5th European Conference on Speech Communication and Technology*, CD-ROM, Rhodes, Greece.