(1996, Revue Francaise de Linguistique Appliquée, n°1, Paris)

# Structures temporelles et structures prosodiques en français lu

**Laboratoire d'Analyse Informatique de la Parole**
**Lettres, UNIL, CH-1015 Lausanne, Suisse**

**Brigitte Zellner**
**Brigitte.Zellner@imm.unil.ch**

## Résumé

*Si la composante prosodique est intégrée dans les systèmes de synthèse de la parole depuis plusieurs années, une dimension temporelle a cependant été peu prise en compte. Il s'agit de la fluidité de la parole. Une parole fluide se caractérise par une gestuelle verbale produite avec aisance, avec des transitions et des attaques douces et un débit rapide et sans heurt. Il sera montré que pour le français, le manque de fluidité dans les paroles de synthèse actuelles s'explique par la génération d'une structuration temporelle trop pauvre car cette structure est supposée être congruente à la structure accentuelle.Une nouvelle approche de l'organisation temporelle de l'énoncé sera ensuite présentée.*

## Introduction

Lors de la conversion automatique d'un texte en parole artificielle, («Text-To-Speech»), il est important de disposer d'une grammaire prosodique (i.e. intonative et temporelle) appropriée. Depuis plusieurs années, la composante prosodique est intégrée dans les systèmes de synthèse de la parole en français (Bailly, 1992; Keller, 1992; Sorin et al., 1987). Cependant, une dimension temporelle a jusqu'à présent été peu prise en compte. Il s'agit de la fluidité de la parole. Une parole fluide se caractérise par une gestuelle verbale produite avec aisance, avec des transitions et des attaques douces et un débit rapide et sans heurt (Pfauwadel, 1986; Zellner, 1992, 1994). Le manque de fluidité dans les paroles de synthèse actuelles s'explique par la génération d'une structuration temporelle trop pauvre. Tout comme en anglais, l'organisation temporelle d'un énoncé en français est le plus souvent inférée à partir de sa structure accentuelle (Bailly, 1983; Barbosa, 1993; Beaugendre, 1994; Delais, 1994, sous presse; Di Cristo et Hirst, 1994; Jun et Fougeron, 1995; Martin, 1987; Mertens, 1987, 1993; Pasdeloup, 1992). La structure temporelle est supposée être congruente à la structure accentuelle. Pourtant l'accent en français ne joue pas le même rôle qu'en anglais ou en allemand puisqu'il n'a pas de valeur distinctive (Dauer, 1983). Il sera montré que la structure accentuelle est insuffisante en français pour déduire la structure temporelle complète d'un énoncé. Une nouvelle approche, qui intègre le concept de fluidité, sera ensuite proposée.

# I. Structure accentuelle et structure temporelle

La prosodie du français est généralement décrite comme une structure hiérarchique à deux niveaux: les structures intonatives qui regroupent les structures accentuelles — ou tonales— (Rossi, 1985). Traditionnellement, les structures temporelles sont directement dérivées des structures prosodiques de l'énoncé car le passage d'une unité prosodique à une autre est marquée tant sur le plan mélodique que temporel. Ces marques prosodiques se manifestent essentiellement sous la forme de proéminences du type accentuation. La durée des éléments sonores (segment, syllabe, diphone, etc.) est supposée être plus ou moins directement affectée par la structure accentuelle de l'énoncé (emplacement de l'élément dans l'unité prosodique, profondeur hiérarchique de l'unité, etc.).

En synthèse de la parole, déterminer l'organisation temporelle revenait donc pour les tenants de cette approche, à déterminer quelles étaient les structures intonative et accentuelle. Dans cette perspective, les modèles de *prédiction* des structures prosodiques se sont développés autour de deux grands axes : les modèles dérivés des structures syntaxiques et les modèles dérivés des structures phonologiques.

### *a. La prédiction des structures prosodiques à partir des structures syntaxiques*

Étant donné le nombre de travaux réalisés en syntaxe dans les années 1960-70, il semblait naturel d'assimiler les structures prosodiques aux structures syntaxiques. En effet, comme les structures prosodiques semblent organisées en termes de cohésion entre les mots, l'origine de cette cohésion entre les mots pouvait être vue comme provenant *directement* des relations syntaxiques. Ceci explique qu'il y ait eu toute une série de tentatives pour générer le phrasé à partir de l'analyse syntaxique de la phrase (p. ex., Allen, J., Hunnicutt, M.S., & Klatt, D., 1987). Les structures syntaxiques sont généralement représentées en arbre où les noeuds les plus bas (les branchements) joignent des éléments proximaux fortement reliés, tandis que les noeuds les plus hauts joignent des unités syntaxiques moins directement reliées. Dans ce type de structure, par exemple, on associe au noeud syntaxique le plus haut la pause silencieuse la plus longue. La distribution des *accents* était prédite à partir de la structure syntaxique. Ainsi, dans les énoncés neutres, la durée des syllabes accentuées est supposée servir d'indicateur de la structure syntaxique (Touati, 1987). Toutefois, les résultats obtenus ne sont pas satisfaisants car la relation entre structures syntaxiques et structures prosodiques n'est pas véritablement congruente (Grosjean et Dommergues, 1983; Sorin, 1990). Ainsi, dans la phrase suivante (Martin, 1987), la frontière prosodique majeure réalisée par le locuteur, permet d'équilibrer la longueur de constituants au lieu de respecter la structuration syntaxique en GN-GV :

Structure produite:
(J'ai vu le père) (de Pénélope)

et non pas:
(J'ai vu) (le père de Pénélope)

Lehiste (1972) a analysé les effets des frontières morphologiques et syntaxiques sur la structure temporelle de la parole (pour l'anglais). Elle a conclu que les processus de réajustement temporel tendent à ignorer les frontières de morphème et de mot. La structure des durées est selon Lehiste, conditionnée par le nombre de syllabes, plutôt que par le nombre de segments ou la présence de frontières syntaxiques. Dans le même sens, Ferreira (1993), a montré que la structure temporelle pouvait être étroitement reliée à la structure prosodique. Si une variation syntaxique (i.e. manipulation de la profondeur de frontières syntaxiques après un mot cible) est introduite dans une phrase à lire, les variables temporelles ne sont pas affectées de manière significative. Par contre, si on introduit une variation prosodique (un changement d'emphase), les variables temporelles telles que la durée des mots en position finale et les pauses sont alors affectées de manière significative.

### *b. L'apport de la phonologie*

Chaque unité prosodique spécifie des caractéristiques rythmiques au niveau de la syllabe aussi bien que du mot, du syntagme et de la phrase. Par exemple, d'un point de vue prosodique, le mot est constitué d'un ensemble de *pieds*, i.e. un ensemble de syllabes forte-faible. De tels mots sont groupés en *syntagmes phonologiques* (Selkirk, 1984), syntagmes qui ne sont pas nécessairement en isomorphie avec les syntagmes syntaxiques. Finalement, les syntagmes phonologiques sont groupés en syntagmes intonatifs, qui se combinent eux-mêmes pour former une phrase. Selon ces principes, Libermann et Prince, Selkirk, Nespor et Vogel, Ladd (cf. Wightman, 1992) proposent diverses hiérarchies pour exprimer au mieux les structures prosodiques. Exemple de Bachenko et Fitzpatrick, (1990):

*Etape initiale*
**He told me last night he was coming to London for several days**

*Etape 1: formation des mots phonologiques*
**He+told+me last+night he+was+coming to+London for+several+days**

*Etape 2: formation des syntagmes phonologiques grâce à l'information syntaxique*
**(He+told+me)(last+night)(he+was+coming)(to+London)(for+several+days)**

*Etape 3: hiérarchisation*
Tout se rattache ensuite “en bouquet” pour former la phrase. Puis on assigne une force relative à chaque frontière en fonction de l'étiquetage syntaxique, de la longueur du syntagme et des frontières adjacentes.

**(He+told+me)(last+night)|(he+was+coming)(to+London)||(for+several+days)**

Selon Bachenko, l'arbre prosodique qui résulte de ce type d’analyse n'est pas toujours congruent avec l’arbre syntaxique et les structures obtenues rendent mieux compte des relations prosodiques qu’une analyse syntaxique traditionnelle. Chez Bachenko, 80% des frontières majeures ont été exactement ou approximativement prédites par comparaison avec les frontières effectivement produites et perçues comme telles. Toutes les frontières majeures prédites ont été effectivement produites, ce qui n’est pas le cas pour les frontières mineures (seulement 42%). Chez Ostendorf (1994), le taux de frontières mineures et majeures correctement détectées est de 77%. Sans aucune information syntaxique, ce score passe à 81%.

Une explication potentielle des insuffisances de ces modèles dans leur capacité à prédire un score plus élevé de frontières prosodiques, pourrait concerner le poids de ces modèles en production de la parole. En effet, Caelen-Haumont (1991) a dégagé, pour trois phrases et 12 locuteurs, des stratégies de lecture en fonction des modulations intonatives d'énoncés produites. Les groupages de mots dans un énoncé semblent fondés sur des *stratégies successives de structuration* qui sont en plus ou moins grande dépendance avec un modèle linguistique. Selon Caelen-Haumont, ces résultats montrent que 78% des groupes de mots semblent avoir été formés sur une base sémantique ou pragmatique. Selon elle, le recours à la syntaxe avait eu lieu avant tout lorsque le contexte était sémantiquement moins saillant. Or les contraintes sémantiques et pragmatiques ne sont pas prises en compte dans les modèles phonologiques.

En outre, les structures générées à partir de modèles phonologiques ou prosodiques s'appuient sur le repérage des accents. Ainsi, le processus de prédiction des durées pour le français suit généralement le mécanisme suivant. Les groupes prosodiques sont constitués en fonction de la localisation des accents. Puis, les unités rythmiques sont progressivement allongées à l'intérieur du groupe accentuel, et elles sont "ré initialisées" après l'accent, à l'image du pattern dégagé par Caelen-Haumont (1981), Pasdeloup (1990, 1992), et Keller & Zellner (1993).

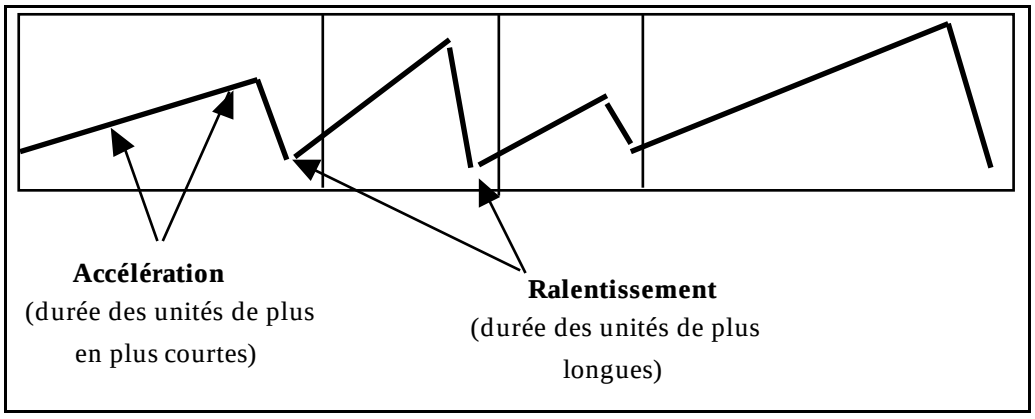



**Figure 1:** Représentation schématique de l'évolution du débit en français, au cours d'un énoncé, selon Pasdeloup (1990) et Keller & Zellner (1993): des phases successives d'accélération et de ralentissement de la vitesse de parole.

Par exemple, Delais (sous presse) a récemment proposé une nouvelle approche dans le calcul des structures prosodiques, approche qui autoriserait différentes segmentations prosodiques. Mais le calcul de la structure rythmique est encore dérivé du calcul des accents. Or le français n'est pas une langue accentuelle au même titre que l'anglais ou l'allemand. Le français n'utilise pas d'accent lexical distinctif. Ainsi, Vaissière propose la distinction "boundary language" vs "stress language" pour distinguer le français de l'anglais (Vaissière, 1991). De plus, les récentes descriptions de l'accent en français divergent (Pasdeloup, 1992; Astesano, C., Di Cristo. & Hirst., 1995). Enfin, il n'a pas été clairement établi que la relation entre l'accent et la durée en français était une relation équivalente à celle qui avait été établie pour des langues comme l'anglais. La supposition d'une telle équivalence occulte probablement certains aspects de l'organisation temporelle du français qui ont trait à la fluidité.

## II. Les structures temporelles dérivées d'un modèle psycholinguistique

Les modèles envisagés précédemment considèrent l'accent comme clé de voûte de la structure temporelle des énoncés. Il s'avèrent que ces modèles sont insuffisants pour fournir une prédiction satisfaisante (>90%) des frontières temporelles avec un taux d'erreurs faible (i.e. peu de frontières inacceptables).

D'autre part, les structures générées étant fondées sur l'organisation accentuelle de l'énoncé (syllabe accentuée - syllabe inaccentuée), il s'en suit une structure temporelle appauvrie car des phénomènes temporels comme les allongements ou les compressions ne sont pas véritablement pris en compte. Aussi, d'autres approches doivent être considérées (Wightman, 1992; Delais, 1994).

### *a. Un autre mode de prédiction des structures temporelles : les précurseurs*

Sorin et al (1987) ont proposé pour la synthèse de la parole un parseur prosodique indépendant de toute analyse syntaxique. La formation des groupes prosodiques est fondée sur la détection des mots grammaticaux. Diverses études psycholinguistiques ont montré en effet que d'une part, le traitement des mots grammaticaux diffère de celui des mots lexicaux et que d'autre part, les mots grammaticaux ont une durée plus brève que celle des mots lexicaux (cf. par exemple: Segui & al, 1982, 87; Zellner 1992). Cet algorithme du CNET fournissait d'après l'auteur, un parsing acceptable et il a été implanté dans un système de lecture automatique du courrier électronique. Néanmoins, ce parseur s'est avéré trop rudimentaire pour traiter par exemple des expressions lexicales composées de plusieurs mots graphiques —par exemple: «mis en place»—. D'autres concepts devaient donc être intégrés.

### *b. L'apport de la psycholinguistique*

Les études psycholinguistiques du groupement de mots sont basées sur des mesures de rappel mnésique, de prolongation de syllabes, et de pauses. Un certain consensus s'est dégagé de ces études. Les groupements de mots créés par le locuteur — ou «structures de performance» — reflètent un certain équilibre de «poids»: nombre de mots, longueur des constituants, hiérarchie des constituants, et symétrie dans l'énoncé (Grosjean et Dommergues, 1983; Monnin et Grosjean, 1993). Par exemple, si une phrase est formée d'un syntagme nominal *court* suivi d'un syntagme verbal *long*, la frontière majeure devrait, selon le principe syntaxique, se situer entre ces deux syntagmes. Mais les mesures empiriques des pauses montrent que la division majeure se situe entre le verbe et le reste du syntagme verbal (cf. ci-dessus au paragraphe I.a. l'exemple de Martin). Inversement, un syntagme nominal particulièrement long suivi par un syntagme verbal plutôt court a tendance, en structure de performance, à être subdivisé en constituants différents et plus courts, de façon à équilibrer le «poids», ou la longueur, entre les différents regroupements proximaux (Grosjean et Dommergues, 1983).

Nombre d'auteurs, à commencer par Grosjean et al (1975), ont montré que les occurrences et les longueurs des pauses étaient fortement corrélées à ce degré de cohésion inter-lexicale. Les

pauses ont par exemple tendance à être plus longues et plus fréquentes entre les mots qui montrent peu de proximité. Et les pauses sont plus rares et plus courtes entre les mots avec forte interdépendance.

Les structures de performance sont donc fortement corrélées aux structures prosodiques (Monnin & Grosjean, 1993). Ces structures dépendent peu des contraintes de respiration au niveau de la phrase. (On retrouve les mêmes structures de performance dans une phrase lue avec ou sans reprise d'air). En outre, les structures de performance varient peu d'un mode de mise en évidence à un autre, que ce soit par exemple par l'analyse des pauses ou par une épreuve de segmentation subjective (Grosjean et Dommergues, 1983).

### *c La prédiction automatique du phrasé à partir des structures psycholinguistiques*

Une fois reconnu la quasi équivalence entre la structuration psycholinguistique et la structuration prosodique (Monnin et Grosjean, 1993), il devenait possible d'élaborer des algorithmes capables de prédire cette structuration psycholinguistique de l'énoncé.

Par exemple, le modèle stochastique d'Ostendorf (1994) pour l'anglais permet de prédire l'organisation prosodique d'une phrase, en distinguant 3 niveaux hiérarchiques. Au plus bas niveau hiérarchique une représentation de facteurs grammaticaux et de longueur de constituants est saisie grâce à un arbre de décisions binaires (emplacement du mot par rapport à l'unité intonative, type du mot de droite et de gauche, etc.). D'après l'auteur, l'algorithme, qui ne comporte donc pas d'information issue d'une analyse syntaxique donne un score de 81% de frontières prosodiques correctement placées avec 4% de prédictions perçues comme étant fausses (i.e. non acceptables) sur un corpus de phrases isolées.

Pour le français, les laboratoires du CNET ont élaboré un parseur prosodique constitué d'un peu plus de 140 règles (Emerard et al, 1992). Ces règles sont fondées sur l'hypothèse que la frontière prosodique peut être déterminée à partir d'un ensemble de concepts psycholinguistiques comme par exemple le type d'un mot (lexical vs grammatical) et des mots voisins, sa longueur syllabique, la description de la syllabe et des types de segments. Puis, les frontières prosodiques étant posées, d'importantes bases de données mélodique, temporelle et pausale sont consultées afin de calculer les valeurs de durées, de fréquence fondamentale et de pauses nécessaires.

La qualité de parole émise par le système de synthèse CNETVOX/PSOLA est actuellement une des meilleures pour le français. L'une des raisons de cette bonne performance provient de la qualité du parseur puisque plus de 95% des frontières prosodiques sont correctement détectées. Ce chiffre a été mesuré sur plusieurs corpus, dont un corpus de plusieurs centaines de phrases où toutes les règles de parsing étaient mises en oeuvre. Pourtant, l'oreille accepterait une parole de synthèse encore plus fluide. De plus, l'architecture très lourde de ce système rend son implantantion très coûteuse au plan informatique.

## III. Une alternative: un modèle temporel

Si la composante prosodique est désormais intégrée dans tous les systèmes actuels de synthèse de la parole pour le français, la qualité de parole obtenue manque encore de *fluidité*. La fluence verbale se caractérise par une élocution rapide et sans heurt, avec des attaques et des transitions douces, qui donnent une impression d'aisance (Pfauwadel, 1986; Zellner, 1992, 1994). Cette dimension est donc avant tout une dimension temporelle. C'est dans cette perspective que l'élaboration d'un modèle temporel s'impose.

Dans un module prosodique au sein d'un système de synthèse, le calcul des structures temporelles n'exclut évidemment pas la nécessité de générer ultérieurement la structure prosodique complète d'un énoncé. La structure prosodique est en effet cette structure de "rendez-vous" où intonation, rythme et énergie s'accordent. Par conséquent, la structure temporelle constitue un des *inputs* de la structure prosodique au lieu d'en être un *output.*

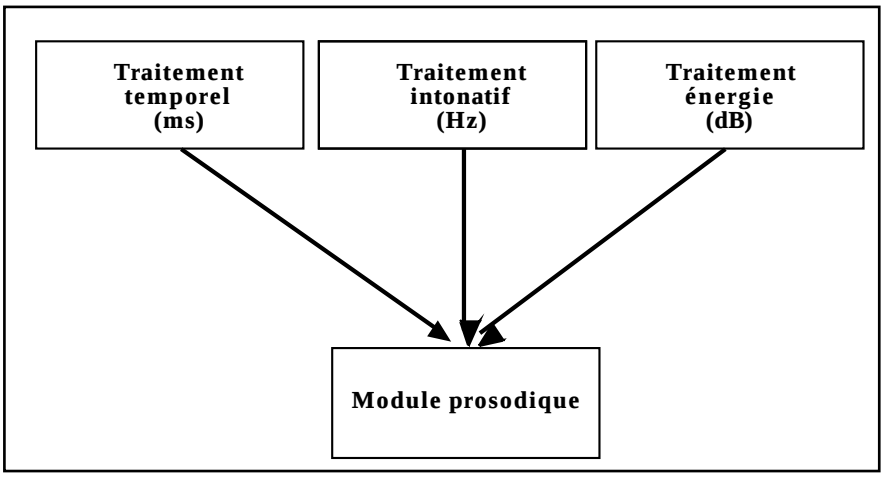



**Figure 2**: La structure prosodique est calculée *après* calcul de la structure temporelle, intonative et énergique.

### *a. Concept*

La production orale d'un énoncé étant structurée dans le temps, deux concepts sont retenus. Tout d'abord, pour désigner les éléments de cette organisation temporelle, on parlera de groupe temporel, et non pas de groupe prosodique, puisque seule la dimension temporelle est ici prise en compte. Une structuration à trois niveaux est retenue, comme chez Ostendorf et al (1994). Ainsi, une phrase peut être structurée en plusieurs groupes majeurs (GM) qui eux-mêmes sont ou non structurés en plusieurs groupes mineurs (Gm).

Le deuxième concept retenu porte sur la relation qui s'établit entre les durées. Comme toute gestuelle, la parole est un phénomène dynamique. Dans cette perspective, la durée d'une syllabe est relative car elle est aussi fonction des durées précédentes et suivantes. Ainsi, une impression de ralentissement du débit peut être créée par un *allongement* de la syllabe, tout aussi bien que par une *compression* de — voire des — la syllabe() précédente(s) et/ou suivantes.

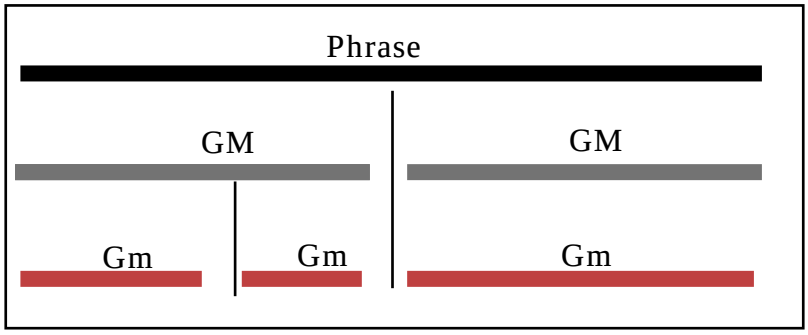



**Figure 3.** Structure temporelle hiérarchisée de la phrase. Une phrase est constituée de groupes temporels majeurs (GM) qui eux-mêmes sont constituées de groupes temporels mineurs (Gm).

### *b. Mise en évidence statistique de la structure temporelle*

Cent phrases déclaratives lues à haute voix à un débit rapide par un locuteur fluent ont été enregistrées en chambre insonorisée. Ces phrases ont été soigneusement étiquetées, selon un protocole précis défini dans le Laboratoire d'Analyse Informatique de la Parole (Thévoz et al, 1994). La structure temporelle a ensuite été dégagée à partir d'une analyse statistique de la durée des syllabes (décrite ci-après) portant sur les trente premières phrases (soit 599 syllabes), puis testée sur les 70 phrases suivantes (soit 1354 syllabes). L'unité linguistique retenue pour cette étude est la syllabe. En effet, d'une part, un consensus assez général s'est établi quant à la réalité psycho-rythmique de la syllabe en français et d'autre part la plupart des modèles prosodiques l'utilisent (pour un avis opposé, voir Barbosa, 1993).

La distinction d'au moins trois niveaux de proéminence prosodique étant largement admise dans la littérature, la dynamique temporelle de la phrase a été mise en évidence comme suit. La distribution des durées syllabiques a été normalisée par transformation logarithmique.

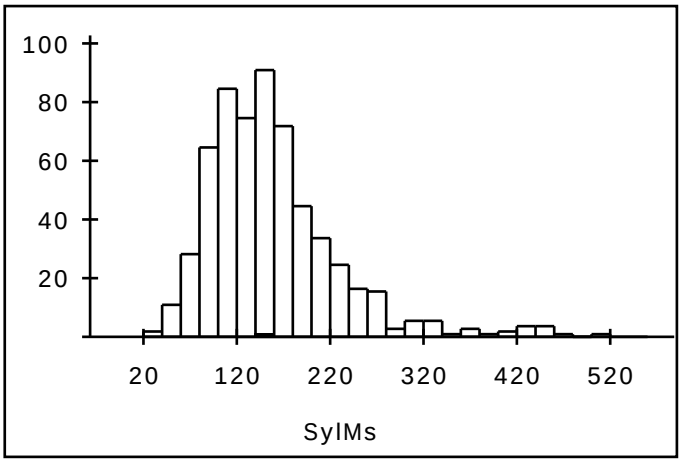



**Figure 4a:** distribution des durées syllabiques en ms (avant normalisation).

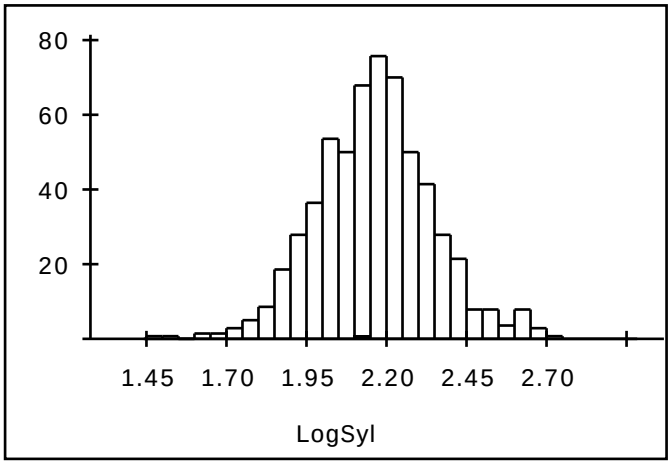



**Figure 4b:** distribution des durées syllabiques après tranformation logarithmique.

Les allongements et les compressions syllabiques qui se situent au delà d'un écart-type de la durée syllabique moyenne ont été répertoriés comme "écarts majeurs", et ceux qui s'éloignent de part et d'autre de la moyenne de plus d'un demi écart-type comme "écarts mineurs". Ceci correspond à une subdivision des durées en trois classes dans un espace normalisé (cf. figure 5).

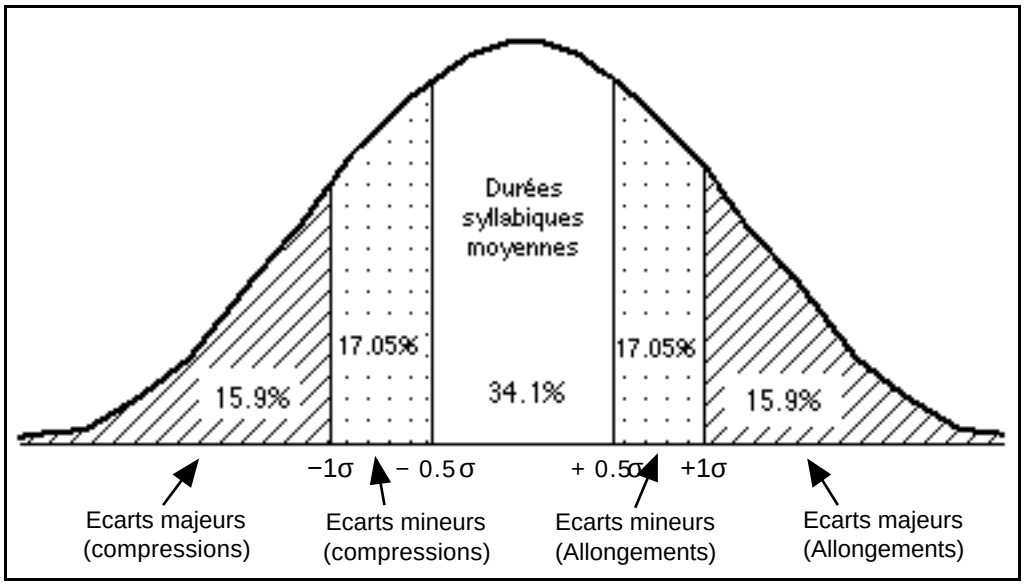



**Figure 5.** Subdvision des durées syllabiques après normalisation.

• Analyse des écarts majeurs

Il est généralement admis que plus une structure se trouve placée haut dans la hiérarchie phonosyntaxique ou prosodique, plus elle est marquée du point de vue des paramètres de F0, durée et énergie. Par conséquent, les groupes majeurs (GM) étant hiérarchiquement les plus élevés, l'hypothèse est que les frontières des GM seront marquées par les allongements les plus longs.

Après normalisation des durées syllabiques par transformation logarithmique, l'extraction des durées syllabiques majeures désigne le repérage des durées syllabiques supérieures à 1 écart-type de part et d'autre de la moyenne (pour l'ensemble des durées syllabiques de l'énoncé considéré).

• Analyse des durées éloignées de la durée moyenne (écarts mineurs)
Ces durées syllabiques se situent dans l'intervalle compris de part et d'autre de la moyenne entre 0.5 écart-type et 1 écart-type. Ces intervalles permettent de repérer les durées qui s'écartent beaucoup de la valeur moyenne — les syllabes allongées ou les syllabes compressées — sans pour autant atteindre les valeurs extrêmes (traitées ci-dessus). L'hypothèse est que les frontières de groupes mineurs (Gm) sont marquées par ce type de durées.

• Analyse des pauses
Les durées des pauses et l'environnement dans lequel elles se manifestent ont été relevés.

### *c. Résultats de l'analyse statistique*

• Frontière de groupes majeurs (GM)
*83%* des 30 phrases analysées se terminent par un allongement syllabique situé au delà de 1 écart-type par rapport à la durée syllabique moyenne. Cependant, de toutes les durées syllabiques mesurées dans l'énoncé, cet allongement final n'est l'allongement le plus long que dans 50% des cas.

La mise en évidence des frontières de GM se manifeste par un *allongement* majeur sur la syllabe finale, voire également sur la syllabe précédente. Le marquage de la frontière peut aussi être renforcé au moyen d'une *compression* de la syllabe antépénultième. L'ensemble des phrases considérées montre qu'une frontière de GM se situe proche du centre de la phrase par rapport au nombre de mots (en moyenne à 0.45 du total des mots en partant du début de la phrase).

• Frontière de groupe mineur (Gm)
Les syllabes allongées se situent toutes à la *fin des mots lexicaux* et les syllabes compressées sont plutôt placées sur les mots grammaticaux ou à l'initiale d'un mot lexical. Lorsqu'un mot lexical est suivi d'un autre mot lexical, le premier ne se finit pas toujours sur une syllabe allongée. Deux allongements successifs sont possibles si la dernière syllabe contient un schwa.

• Pauses
*80%* des pauses sont situées sur une frontière de Gm, dont 43% produites dans le contexte (C)V-V(C). Au total, 66% des pauses dans ce corpus sont produites dans le contexte (C)V-V(C) que ce soit au sein d'un Gm ou à la frontière d'un Gm, sachant qu'il s'agit d'un débit rapide.

En résumé, une syllabe voit sa durée affectée en fonction de sa place dans la structure temporelle. Quatre types de frontières sont finalement retenus pour expliquer quatre niveaux de durées syllabiques:

| | | |
|---|---|---|
| • Début de Groupe mineur | -> | syllabe brève (au moins inférieure de 0.5 écart-type par rapport à la moyenne) |
| • Milieu de Groupe mineur | -> | syllabe brève ou médium |
| • Fin de Groupe mineur | -> | syllabe longue (entre 0.5 et 1 écart-type par rapport à la moyenne) |
| • Fin de Groupe majeur | -> | syllabe très longue (au delà de 1 écart-type par rapport à la moyenne) |

***b. La prédiction de la structure temporelle***

Hormis un étiquetage morphosyntaxique des mots, le calcul de la structure temporelle ne requiert pas d'analyse syntaxique. En fonction de sa catégorie morpho-syntaxique, chaque mot est ensuite déclaré comme étant mot lexical — "l"— (i.e. noms, verbes, adjectifs ou adverbes) ou mot grammatical —"g"— (i.e. autres catégories morphosyntaxiques). Certaines unités lexicales sont insécables comme par exemple les expressions figées, les formes verbales complexes («faire faire», «faire pouvoir») et la négation («ne fait pas»). Cette propriété est signalée par le signe +.

La fille s'est déguisée en une jolie petite fée espiègle
f l f l f f l l l l

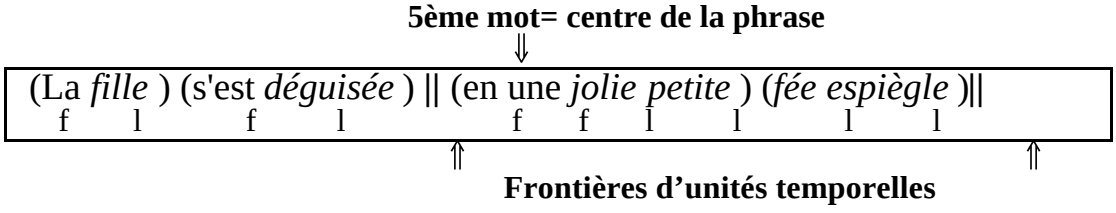
Les frontières temporelles se déterminent alors comme suit, en lisant la phrase de gauche à droite:

• Une marque de fin de groupe majeur (GM) est créée en fin de phrase et devant une virgule.

• Une marque de fin de groupe mineur (Gm) est posée dès qu'un mot grammatical apparaît après un mot lexical. Si trois mots lexicaux (ou 3 unités successives insécables) se suivent sans être séparées par une virgule, et si le dernier mot est le verbe, poser une frontière entre les premiers mots et le dernier (ou la dernière unité insécable). Sinon, la frontière de groupe sera entre le premier mot lexical (ou entre la première unité insécable) et les mots suivants. Si au moins quatre mots (ou unités insécables) lexicaux se suivent, mettre la frontière juste après le verbe s'il occupe la 1ère ou la 2ème position, avant le verbe s'il occupe la 3ème position. Dans les autres cas, la frontière de groupe sera posée après le deuxième mot.

(La *fille* ) (s'est *déguisée* ) (en une *jolie petite* ) (*fée espiègle* )
f l f l f f l l l l

• Une marque de fin de groupe majeur (GM) est créée à l'intérieur de la phrase lorsque celle-ci comprend au moins 13 syllabes (moyenne observée pour ce locuteur). Dans ce cas, en fonction du nombre de *mots*, une frontière de GM est posée à mi-phrase, *sur la frontière de Gm la plus proche.*



Puis, des pauses sont insérées, en général au moins une pause à mi-phrase si la phrase compte plus de 13 syllabes. Les pauses qui résistent le mieux quel que soit le débit de parole sont celles rencontrées à une frontière de groupe, dans le contexte suivant: (C)V#V(C).

***c. Représentation de la structure temporelle obtenue***

Soit la phrase:

> "Je me demande si vous savez vraiment parler ou si vous avez encore des choses à apprendre en prosodie, en syntaxe ou en sémantique."

La dynamique de cette phrase, telle que produite par un étudiant, est représentée ci-après d'une part sous forme de texte au moyen de la taille des caractères et d'autre part sous forme d'arbre hiérarchique (cf. figure 6). Dans la représentation en texte, les plus gros caractères correspondent aux allongements extrêmes (les deux allongements les plus longs de toute la phrase), puis viennent les allongements majeurs (supérieurs à 1 écart-type par rapport à la moyenne) et les allongements mineurs (entre 0.5 et 1 écart-type); la taille de caractère normale est utilisée pour représenter les durées moyennes. Les compressions syllabiques sont représentées par réduction de la taille des caractères selon le même principe. Le passage à la ligne suivante correspond à l'émergence d'une pause. Les pauses sont représentées par deux barres, et l'espace représenté entre ces barres est également fonction de la durée pausale par rapport à la durée syllabique moyenne. Ainsi, chaque ligne de texte correspond à un GM et les Gm sont signalisés par des barres verticales.

Je me demand(e)| si vous savez vraiment parler| ou si vous avez encor(e) | des
chos(e)s | à appren dr(e) | en proso die, |
[ ]
en syntax(e) | ou en séman tiqu(e)| .

La structure temporelle de l'énoncé est aussi représentée dans la représentation hiérarchisée en arbre (cf. figure 6). Cette représentation ne tient pas compte de la structure intonative. *Les embranchements principaux* sont créés sur la base de la structure pausale de l'énoncé. *Les embranchements mineurs* sont créés sur la base des allongements syllabiques produits aux frontières des groupes mineurs, compte tenu des compressions éventuelles adjacentes (interpolation entre la syllabe précédente et la syllabe suivante). Par exemple, un allongement mineur précédé et suivi d'une compression mineure correspond en fait à un écart équivalent à un allongement *majeur*. Sur l'arbre, ce sera donc un allongement majeur qui sera représenté. La hauteur dans l'arbre est fonction de l'allongement syllabique (majeur ou mineur) qui résulte de cette compensation entre allongement et compression; la durée syllabique moyenne constitue la base de l'arbre (valeur dans l'arbre: 0). Un Gm *domine* le précédent lorsque l'allongement produit à la frontière est supérieur à celui du précédent Gm.

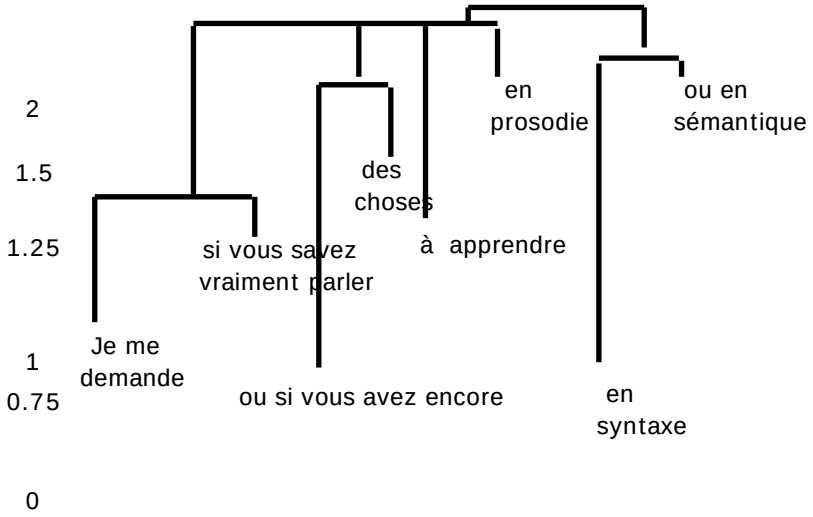



**Figure 6**: Structure temporelle hiérarchisée de la phrase lue par un étudiant. Les embranchements majeurs captent la structure pausale et les embranchements mineurs la hiérarchie des allongements syllabiques.

Les illustrations suivantes représentent la dynamique temporelle prédite avec le modèle temporel décrit ci-dessus.

Je me demand(e) | si vous savez vraiment parler | ou si vous avez encor(e) | [ ]
des chos(e)s | à apprendr(e) | en proso die, |[ ]
en syntax(e) | ou en séman tiqu(e) |.

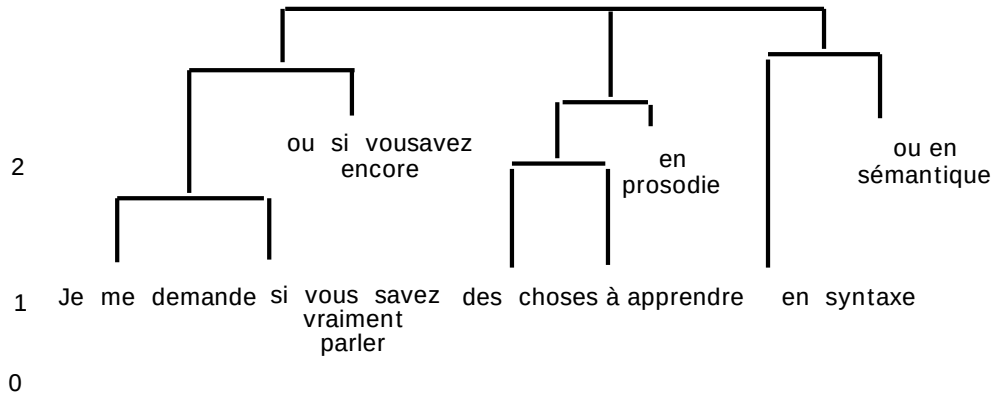



**Figure 7**: Structure temporelle hiérarchisée de la phrase telle que calculée par le modèle temporel. Les embranchements majeurs captent la structure pausale et les embranchements mineurs la hiérarchie des allongements syllabiques.

Cette structure ressemble à la structure produite naturellement par l'étudiant, bien que comprenant une pause supplémentaire. Elle respecte la tendance à faire converger tous les allongements d'une branche vers le haut : plus les allongements sont proches de l'embranchement (GM), plus ils sont longs.

### *c. Évaluation de la prédiction*

Afin d'évaluer les performances de ce modèle temporel, les frontières prédites ont été comparées à un premier modèle (KZ) de prédiction des durées en français (Keller et Zellner, 1995). Ce modèle KZ avait été conçu sur un autre corpus, pour le même locuteur. Il s'agit d'un modèle statistique qui permet de calculer dans une première étape les durées segmentales "intrinsèques" (cf. figure 8), c'est-à-dire dégagées de toutes influences suprasegmentales (syllabiques, phrastiques, etc.).

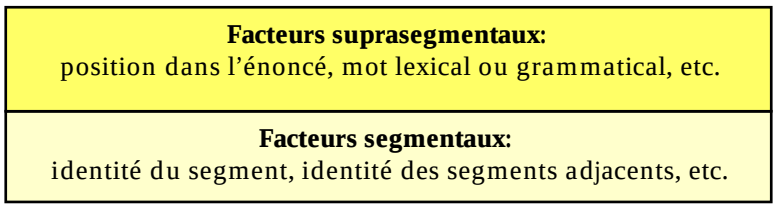



**Figure 8**. Le modèle temporel statistique KZ (1995) calcule la durée syllabique en deux étapes: la durée segmentale, puis la durée suprasegmentale.

La prédiction des frontières temporelles générées par le modèle temporel a été évaluée à l'aide de cet outil statistique. Un delta a été calculé par soustraction de la durée syllabique totale, mesurée dans les données, avec la durée syllabique "intrinsèque" calculée— i.e. la somme des durées segmentales calculées par le modèle KZ lors de la première étape. L'avantage de cette méthode est qu'elle permet de calculer la corrélation entre les facteurs *supra*segmentaux qui influent sur ce delta et les types de frontières temporelles. Dans cette évaluation, les influences qui appartiennent au domaine segmental ont donc été exclues.

Une Anova a été calculée à l'aide du package DataDesk. La corrélation obtenue pour 70 phrases (1354 syllabes) entre le delta et les frontières temporelles est de **0.609.** Ceci correspond à une explication de variance de 37% pour les facteurs suprasegmentaux (cf. figure 9). Cela représente une amélioration du modèle Keller-Zellner (1995) qui expliquait environ 30% de la variance.

Les coefficients obtenus lors de cette Anova montre que les syllabes allongées sont avant tout les syllabes situées en fin de GM devant une pause ou en finale de phrase. Les fins de Gm ne sont pas tant marquées par un allongement de la syllabe finale que par une compression de la syllabe initiale du Gm suivant.

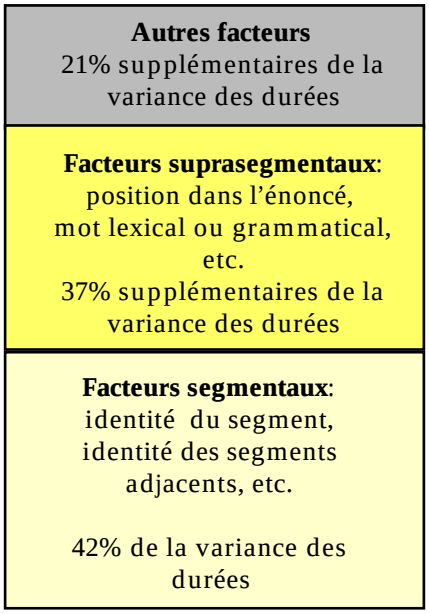



**Figure 9.** Explication de la variance des durées syllabiques en français. Les facteurs suprasegmentaux permettent d'expliquer 37% supplémentaires de la variance.

## Conclusion

Diverses contraintes linguistiques (phonologique, syntaxique, sémantique), psycholinguistique (contraintes liées à la production de la parole), émotives et pragmatiques s'exercent dans la dimension temporelle de la parole. En français, les algorithmes de prédiction des durées basés essentiellement sur des modèles phono-syntaxiques s'avèrent insuffisants. Dans ces modèles, la structure temporelle d'un énoncé est prédite à partir de la structure intonative et de la localisation des accents. Or le français n'est pas une langue accentuelle au même titre que l'anglais ou l'allemand. C'est sans doute une des raisons pour lesquelles les structures calculées au moyen de ces algorithmes diffèrent des structures pouvant réellement être produites. De plus, elles ne permettent pas de générer une parole fluide.

Un modèle temporel est proposé, dans lequel aucune information accentuelle n'est requise. En effet, l'accent peut être intégré plus tard dans le modèle prosodique car il ne constitue pas, pour le français, la clé de voûte de la structure temporelle même si sa présence peut localement introduire des variations de durée.

Les premières évaluations de ce modèle sont prometteuses globalement — corrélation élevée — et localement — dans la mesure où le modèle permet de générer des structures temporelles proches de celles que pourraient produire un locuteur, sans "erreur de performance". Ce modèle devra par la suite intégrer le traitement d'autres types de phrases et de textes. D'autres évaluations sur d'autres corpus devront également permettre de tester la robustesse du modèle, en termes de groupements de mots acceptables et inacceptables.

**Remerciements**
Mes chaleureux remerciements à Eric Keller, Geneviève Caelen-Haumont, Philip Keller, Jacqueline Vaissière et Helène Huot sans lesquels cet article n'aurait pas vu le jour.

# Références

Astesano, C., Di Cristo, A. & Hirst, D.J. (1995). Discourse based empirical evidence for a multi-class accent system in French.

Allen, J., Hunnicutt, M.S., & Klatt, D. (1987). *From text to speech. The MITalk system.* Cambridge, England: Cambridge University Press.

Bachenko J., Fitzpatrick E. (1990). A computationnal grammar of discourse- neutral prosodic phrasing in English, *Computationnal Linguistics, 16.* 155 -170.

Bailly, G. (1983). *Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Etablissement d'un modèle de génération.* Thèse d'ingénieur, Institut National Polytechnique de Grenoble.

Bailly, G. & Benoit, C. (1992),*Talking Machines. Theories, Models, and Designs* . Elsevier Science Publishers.

Barbosa, P. A. (1994). *Caractérisation et génération automatique de la structuration rythmique du français.* Thèse de Doctorat. U.R.A. CNRS n°368 - INPG/ENSERG, Université Stendhal, Grenoble.

Beaugendre, F. (1994). Une étude perceptive de l'intonation du français. Thèse de Doctorat en Sciences de l'Université Paris XI. LIMSI n°94 - 25.

Caelen-Haumont, G. (1991). *Stratégies des locuteurs et consignes de lecture d'un texte: Analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques,* Thèse d'Etat, Aix-en-Provence.

Dauer, R.M. (1983). Stress-timing and syllable timing reanalyzed. *Journal of Phonetics, 11*. 51-62.

Delais, E. (1994). Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en français. *20èmes Journées d'Etude sur la Parole* (pp. 379-384). Trégastel.

Delais, E. (1995). Rythme et structure prosodique en Français. *Proceedings of Congrès Annuel de l'Association pour l'Etude de la Langue française,* Aix-Marseille.

Delais, E. (to appear). Towards a dynamic model of the prosodic structure. *Proceedings of the HILP 2.* Amsterdam, Hollande (25-27 janvier 95)

Di Cristo, A. & Hirst, D. (1994). Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français. *Travaux de l'Institut de Phonétique d'Aix, 15.* 13-24.

Ferreira, F., (199**3**). Creation of Prosody During Sentence Production *Psychological Review, 2*.. 233-253.

Grosjean, F. , & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica, 31.* 144-184

Grosjean, F. , & Dommergues, J.Y. (1983). Les structures de performance en psycholinguistique.*L'Année psychologique, 83*. 513-536.

Jun, S-A. & Fougeron, C. (1995). The accentual phrase and the prosodic structure of French. *XIIIème Congrès International des Sciences Phonétiques, 2* (pp. 722-725). Stockholm.

Keller, E. (1992). *Le choix d'un modèle informatique pour la prosodie en synthèse de la parole.* Inaugural Lesson, Faculté des Lettres, Université de Lausanne, Lausanne, October. (d)

Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The prediction of prosodic timing: Rules for final syllable lenthening in French. *Proceedings ESCA Workshop on Prosody* (pp. 212-215). *September 27-29. Lund, Sweden.*

Keller, E., & Zellner, B. (1995). A statistical timing model for French. *XIIème Congrès International des Sciences Phonétiques, 3* (pp. 302-305). Stockholm.

Lehiste, I.(1972). The timing of Utterances and Linguistic Boundaries.*Journal.of the Acoustical Society of America,69.* 2018–2024.

Martin, Ph. (1987). Structure rythmique de la phrase française. Statut théorique et données expérimentales. *Proceedings des 16e Journées d'Etude sur la Parole* (pp. 255-257). *Hammamet.*

Mertens, Piet. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique.* Thèse doctorale, Katholieke Universiteit Leuven.

Mertens, P. (1993). Intonational grouping, boudaries and syntactic structure in French. *Proceedings ESCA Workshop on Prosody, September 27-29. Lund, Sweden.* 156-159.

Monnin, P , & Grosjean, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique, 93.* 9-30.

Ostendorf, M. & Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics, 20.* 1-27.

Pasdeloup, V. (1990). Organisation de l'énoncé en phases temporelles: Analyse d'un corpus de phrases réitérées, *18émes Journées d'Etudes sur la Parole.*(pp. 254 - 258). Montréal.

Pasdeloup, V. (1992). Durée intersyllabique dans le groupe accentuel en Français. *Actes des 19émes Journées d'Etudes sur la Parole.* (pp. 531-536). Bruxelles.

Pfauwadel, M.C. (1986), *Etre bègue,* Paris: Retz.

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access. *Neuropsychologia, 6.* 615-627.

Segui, J., Frauenfelder, U., Lainé, C., & Mehler, J. (1987). The word frequency effect for open- and closed-class items. *Cognitive Neuropsychology, 4.* 33-44

Selkirk, E.O. (1984). *Phonology and syntax: The relation between sound and structure.* MIT Press, Cambridge, MA.

Sorin, C., Larreur, D. & Llorca, R. (1987). A rythm-based prosodic parser for text-to-speech systems in French. *XIème Congrès International des Sciences Phonétiques* (pp. 125-128). Talinn, Estonie. Proceedings.

Sorin, C. (1990). Synthèse de la parole à partir du texte: "Etat des recherches et des applications". *Publications and communications of the "RCP" Department, CNET*(pp. 143-158). Lannion.

Thévoz, N., & Enkerli, A. (1994).*Critères de segmentation.* Laboratoire d'Analyse informatique de la Parole (LAIP), Université de Lausanne, August, (rapport interne).

Touati, P. (1987). *Structures prosodiques du suédois et du français.* Lund: University Press.

Wightman, C. W. (1992). Automatic detection of prosodic constituents for parsing. Doctoral dissertation. Boston University Graduate School.

Vaissière, J. (1991). Rhythm, accentuation and final lengthening, in J. Sundberg L. Nord, R. Carlson (Eds).*French in Music, Language, Speech and Brain* (pp.108-120). Wenner-Gren International Symposium Series Macmillan Press, Vol. 59

Zellner, B. (1992). Le bé bégayage et euh... l'hésitation en français spontané. *Actes des 19ème Journées d'Etudes sur la Parole (J.E.P*), Bruxelles. 481-487.

Zellner, B. (1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) *Fundamentals of speech synthesis and speech recognition.* (pp. 41-62). Chichester: John Wiley.