LES PRISES DE SOUFFLE DANS LE DISCOURS
Quelques éléments vers une implémentation « text-to-speech »

(Bernard P.-J. & A. Auchlin)

0. Avant-propos

La recherche présentée ci-dessous a été conduite entre le 1er octobre 2002 et le 31 mai 2003 par Pierre-Jean Bernard, sous la direction d'A. Auchlin (Département de linguistique), en collaboration avec J.-P. Goldmann (LATL, Université de Genève, et AT&T). Elle a successivement bénéficié du soutien financier : du projet plurifacultaire Prosodie du Rectorat (oct. 02); du Département de linguistique (nov. 02 à mars 03); et du Fonds Charles Bally de la Société Académique (avr. – mai 03). Nous leur exprimons ici notre reconnaissance.

Cette étude préliminaire avait pour but général de mettre en évidence les rôles joués par les prises de souffle (Pds) dans la parole naturelle « spontanée », et de poser les bases d'une intégration dans un synthétiseur vocal.

Elle s'inscrit dans la continuité des arguments développés par Grobet & Auchlin (2001) qui suggèrent une première catégorisation des Pds. Quelques observations permettent de comprendre l'intérêt que peut revêtir une étude des Pds.

- Réalité du phénomène : dans l'extrait radiophonique étudié, durée 3 minutes 02, les Pds occupent 12 secondes, soit 6, 59% du temps de parole, distribution comparable à celle de phonèmes vocaliques ; pour autant une Pds est :
- Non-phonème (pas de paire minimale…), non-morphème (pas de « contenu »), c'est un « bruit »
- Non répertorié dans les bases de données exploitées par les synthétiseurs de parole (Mbrola[1])
- Les Pds font l'objet d'un « oubli » symptomatique dans la recherche linguistique
- Convergence entre les « bonnes manières oratoires » (inspirations inaudibles, chant scénique et microphone), la description scientifique, et la tradition de l'écrit…
- On peut accorder aux Pds un certain contenu instructionnel[2], et différentes valeurs :
- textuelle : annonce discursive, taille du segment amorcé; interactionnelle : orientation accord-désaccord ; stratégique ; rythmique
- Facteur d'inter-synchronisation[3], ses effets interprétatifs peuvent être minimaux voire nuls, mais ses effets expérientiels décisifs.

Nous présentons ci-après notre matériel, nos hypothèses, et les premiers résultats et mesures ; nous décrivons également l'expérience menée afin de tester nos premières données et en présentons les résultats ; les conclusions générales sont exposées à la fin du document.

1. Corpus

Tous les fichiers de notre corpus ont été enregistrés en 22'100Hz ou 44'100Hz, 16bits, mono afin de permettre des analyses sans conversions lors de l'utilisation de Praat[4].Au demeurant, la qualité médiocre de certains enregistrements et donc l'irrégularité des données ont contribué à freiner une analyse spectrale étendue.

---

[1] Logiciel de synthèse vocale : http://tcts.fpms.ac.be/synthesis/mbrola.html (voir bibliographie)

[2] Contenu « instructionnel » que l'on conçoit comme celui des marques à fonction pragmatique, connecteurs et autres marqueurs discursifs.

[3] Voir Kerbrat-Orecchioni ; Cosnier ;

[4] Logiciel d'analyse et de traitement du signal de parole ; crédit à BOERSMA P. & D. WEENINCK: www.praat.org (voir bibliographie)

1.1 Corpus radiophonique

Il s'agit de l'enregistrement de l'émission "La ligne de coeur" (Radio Suisse Romande 1, 13.06.1996), journaliste Roselyne Fayard, invitée l'écrivain(***e***) Benoîte Groult.
L'échantillon sélectionné dure 1 minute 7 secondes. Ce corpus a, par ailleurs, fait l'objet d'une partie des analyses présentées lors du colloque Prosodie 2002, organisé dans le cadre du projet plurifacultaire Prosodie de l'Université de Genève, dont les Actes sont à paraître[5].

1.2 Parole « libre »

Ces enregistrements sont extraits de séances de Formation Continue, composés de dix personnes (4 hommes et 6 femmes), et consistent en deux tâches différentes :

- Une tâche de lecture à voix haute
- Une tâche d'improvisation (se présenter devant l'auditoire).

Les durées des lectures et présentations varient de 43 secondes à 2 minutes 34 ; la dualité des tâches-situations-emplois langagiers visait à contrôler la variabilité inter-tâches par locuteur, et entre locuteurs distincts.

2. Segmentation

La figure 1 donne un aperçu de la manière dont a été segmenté le signal.
Concernant le corpus radiophonique, le découpage a été effectué en : Actes (§2. 1), Syllabes (Syll.), Pauses (P), Pause avant la prise de souffle (Pav), Prise de souffle (Pds), Pause après la Pds (Pav).
La même segmentation a été appliquée pour le corpus 2 à l'exception du découpage en Actes[6].

Cette procédure permet d'obtenir une certaine mesure rythmique de la parole en fonction du nombre de P, Syll. et autres groupes Pav/P/Pap.

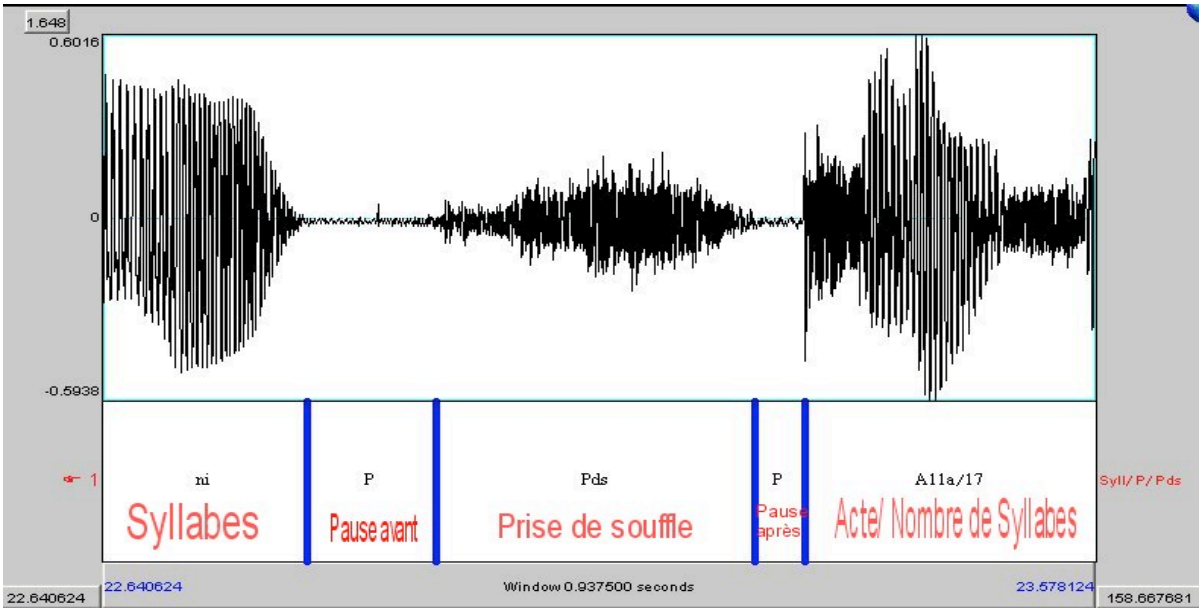

Fig. 1

2.1 Notion d' « Acte »

La segmentation discursive ou pragmatique du discours en unités dénommées « actes » suit les hypothèses de Roulet, Filliettaz, Grobet & Burger 2001[7], qui définissent l'unité minimale « textuelle » comme unité minimale d'incrémentation de la mémoire discursive ; il n'est pas assuré que cette unité, « atome » de la structure hiérarchique textuelle du discours, soit la plus pertinente

[5] *Cahiers de l'Institut de Linguistique de Louvain* 30/1-3.
[6] Unité qui a été remplacée par une unité plus empirique, intuitive mais empiriquement « réaliste », de « groupe syntagmatique majeur intonatif borné ». Voir § 2.1.
[7] Suivant une hypothèse de Berrendonner 1993.

pour saisir l'organisation des Pds dans le discours. D'autres unités, de différents niveaux ou domaines d'organisation du discours (Roulet & al), pourraient intervenir de façon plus directe (voir Simon 2004). Pour une partie du corpus, l'analyse en actes a été délaissée pour une approche plus immédiate, recourant à une unité ad hoc, que l'on peut nommer (cf. note 3) « groupe syntagmatique majeur intonatif borné » ; unité empirique qui émerge localement de la convergence d'indications morpho-syntaxiques, sémantiques, et prosodiques ; elle ne correspond pas à la notion d'acte dans sa définition ; empiriquement, elle produit un découpage globalement isomorphe à celui en actes ; la discussion théorique de cette question est menée par Simon (2004), i.a.

2.2 Premières données

Après écoute, une première classification quantitative des Pds semble se dégager :

- les Pds courtes[8] (< ou = à 300ms)[9],
- les Pds longues (<500ms>)[10].

Même en ce premier état (le paradigme sera élargi par la suite), il s'agit de « Pds-type », c'est-à-dire dont la réalisation phonétique, comme celle des phonèmes (les « voyelles ou consonnes-type »), est sujette à variation selon l'environnement, local et global.

Puis, ce que nous appelons la « couleur » de ces Pds. A savoir :

- Pds vocalisées, c'est-à-dire dotées d'un timbre de voyelle, ou identifiables comme voyelle (par exemple : /a/, /e/, /yi/)[11],
- Pds « bruit », déterminées par des paramètres de co-articulation, ne produisant pas un signal périodique.

De nombreux autres paramètres devront être pris en compte afin d'avoir une classification plus fine, plus précise (§ 5.3).

3. Mesures et Analyses

Les comptages et mesures suivants ont été effectués :

a) nombre total de Syll., P, Pav, Pds, Pap pour chaque séquence, pour chaque locuteur,
b) durée totale du signal avec et sans P, Pav, Pds, Pap,
c) durées moyennes des P, Pav, Pds, Pap,
d) durée moyenne des Syll. par locuteur,
e) nombre de Pds par nombre de Syll.

Les résultats obtenus (essentiellement des rapports de durées) permettront d'extraire les informations nécessaires en vue de notre implémentation Text-To-Speech (TTS).

3.1 Premiers résultats

S'il était question, dans les premières hypothèses, de dégager une seule Pds « standard », les résultats montrent, conformément aux premières observations (§ 2.2), qu'il faut au moins distinguer deux types de Pds : les Pds courtes (PdsC) et les Pds longues (PdsL).

Les résultats indiquent également que les Pav ont une durée supérieure aux Pap, et ce pour l'ensemble de nos corpus (graphique page suivante).

[8] Ou *brèves* – terminologie non encore fixée.

[9] Le temps est donné en millisecondes pour la seule Pds (c'est-à-dire sans les Pav et Pap attenantes). Cela sera modifié par la suite.

[10] Il faudra, par la suite, tenir compte de Pds dites « moyennes » (cf. § 5.3).

[11] Il s'agit d'un timbre non voisé ; F0 n'est pas donnée par les cordes vocales, mais un effet de sifflet, et les F1, F2, F3 sont le produit ordinaire des positions articulatoires ; en partie déterminées par co-articulation, mais ce dernier facteur ne peut rendre compte des variations de timbre des Pds.

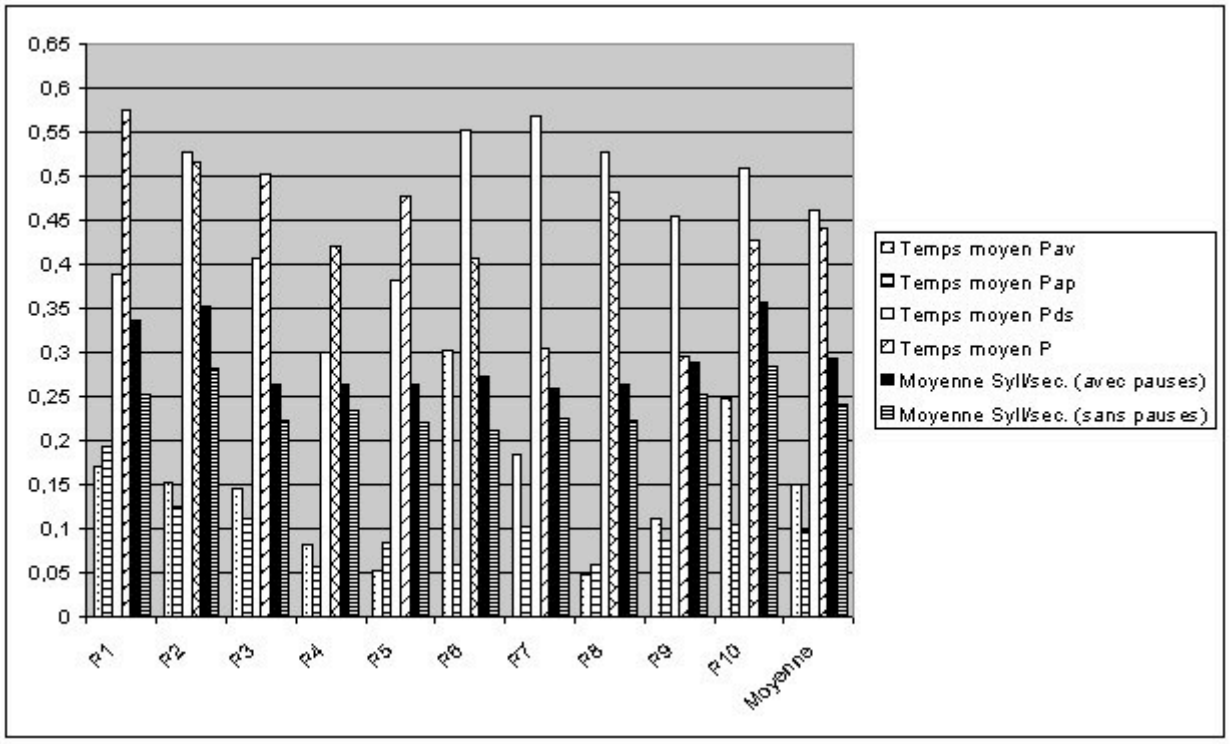

Fig. 2

Ces données ne prennent pas en compte les paramètres qualitatifs tels le timbre, la hauteur (F0), l'intensité, le contexte phonétique. Le travail, fondé principalement sur des moyennes temporelles, rythmiques, se justifie par le souci d'une mise en place prototypique de Pds aisément implémentables dans un contexte de TTS.

Les résultats obtenus quant au nombre de syllabes produites avant l'apparition d'une Pds ne permettent pas, en l'état, d'extraire une régularité pour l'ensemble des locuteurs. Autrement dit, il n'est pas possible de faire une prédiction fiable sur le moment où une Pds va apparaître uniquement à partir de moyennes.

Néanmoins, ces données sont suffisamment complètes pour autoriser une première création de Pds insérables dans un système TTS.

Les valeurs de durée retenues dans ce but pour les Pds sont les suivantes :

PdsC = 330ms
PdsL = 550ms

3.2 Prototypisation

Les premiers fichiers-tests synthétisés (5 au total) ont été créés manuellement. La procédure a été la suivante :

a) suppression de tous les groupes Pav/Pds/Pap du fichier original non-synthétisé, en prenant bien soin de noter l'emplacement de chaque groupe dans le signal sonore;
b) synthétisation complète du signal ;
c) synthétisation des Pds « standard » (cf. ***3.1***);
d) réglage du niveau sonore pour c) ;
e) insertion des Pds synthétisées dans le signal obtenu en b) aux emplacements marqués ;
f) réglage du volume pour l'ensemble du nouveau fichier synthétisé.

Les premiers résultats obtenus ont été globalement convaincants : l'apport au « naturel » de la parole de synthèse des Pds ajoutées est indéniable – en tout cas lorsqu'il est testé informellement, et si le matériau testé a été « manufacturé ».

L'étape suivante a consisté à insérer automatiquement ces Pds à l'aide de scripts Praat via MBRola[12], et à soumettre les résultats à une évaluation par juges.

4. Observations théoriques

D'après Hird K. & K. Kirsner (2002) il n'y a pas d'ajustement systématique entre la fin d'une ligne de déclinaison et les Pds (en lecture à voix haute) ; ces paramètres sont plus indépendants lors de discours libre. Les Pds apparaissent régulièrement aux frontières grammaticales lors d'une tâche de lecture ; dans le discours spontané, Winkworth, Davis , Adams & Ellis (1995) remarquent que 68% des Pds ont lieu à des frontières grammaticales, mais avec une grande variabilité d'un locuteur à un autre. Ils signalent également que des Pds peuvent apparaître sur un simple mot ou à l'intérieur de syllabes; cela revient à dire que certaines Pds ont lieu à des moments sans rapport avec un simple besoin physiologique. Par exemple dans le cas d'hésitations, où la même syllabe est entrecoupée d'une ou de plusieurs Pds.

Les auteurs mettent la longueur des groupes respiratoires en relation avec les besoins sémantiques liés au message. Leur conclusion est que le groupe respiratoire reflète davantage une « unité de sens » qu'une unité de nature syntaxique. Pour Garret (1982), la durée et la profondeur de l'inspiration sont anticipées, et coordonnés avec la planification « d'unités de sens ».

Ces remarques posent le problème de la définition et de la délimitation stricte de ce qu'est un « groupe respiratoire » en discours libre, de la définition ou caractérisation des « unités de sens », et des relations qu'elles entretiennent avec les unités linguistiques et discursives à différents niveaux.

La notion d' « unité de sens » n'est pas une réalité tangible pour la synthèse de la parole. On peut cependant supposer que la « phrase » possède cette capacité de délimiter des unités de sens ; cela expliquerait assez simplement pourquoi, en lecture à voix haute, les Pds sont relativement synchronisées avec les phrases syntaxiques (dans certaines conditions). Nous nous sommes ainsi limités aux paramètres précédemment énoncés pour mettre en place une étude de jugement de « naturalité » de la Pds dans un discours synthétisé.

5. Une première expérience

Nous supposons que l'insertion de prises de souffle dans un signal de parole synthétisée peut concourir à améliorer son « naturel » ; en soumettant à des juges des échantillons de parole de synthèse, certains pourvus de Pds ajoutées, nous prédisons (espérons…) que ces derniers seront mieux notés en terme de naturel. Conformément aux conclusions de nos analyses, deux types de Pds ont été utilisées : les courtes et les longues. Les Pds sélectionnées sont d'intensité et de timbre identiques. Ce seront les mêmes utilisées tout au long de l'expérience.

Notre choix s'est porté sur les Pds dont la qualité sonore était la meilleure : peu de bruits parasites, audibilité, frontières parfaitement délimitées.

5.1 Procédure

Les phrases sélectionnées (26) appartiennent aux deux corpus et ont été synthétisées à l'aide de MBRola. Les Pds ont été introduites à la main, aux endroits exacts de leur apparition dans les fichiers originaux. Les durées pour les Pav et les Pap ont été calculées automatiquement, en fonction des paramètres précédemment établis, à l'aide d'un script Praat.

Trois types de phrases composent cette expérience :

- phrases sans Pds,
- phrases avec PdsC,
- phrases avec PdsC et PdsL.

Les vingt-six phrases ont ensuite été présentées dans un ordre aléatoire (utilisation de Winamp[13]) et séparées les unes des autres par une pause de 5 secondes.

---

[12] La base de diphones la plus utilisée en TTS – qui ne contient pas de Pds insérable dans la synthèse.

[13] NullsoftWinamp : http://www.winamp.com (voir bibliographie)

5.2 Déroulement

Les phrases ont été présentées à un auditoire de 30 personnes qui devaient juger de leur degré de naturalité selon une échelle allant de 1 (pas naturel du tout) à 5 (tout à fait naturel). Les auditeurs ont eu trois phrases d'entraînement (répétées deux fois), leur permettant de se familiariser avec la voix synthétisée.

L'audition s'est faite à l'aide d'un ordinateur portable et de deux haut-parleurs de puissance suffisante placés au centre de la salle.

5.3 Résultats et commentaires

Les résultats obtenus démentent nos prédictions.

Une telle issue était à certains égards prévisible : en ne prenant en compte que les paramètres temporels (et encore, selon une quantification qui pourrait être affinée), à l'exclusion des traits de timbre, nous pouvions nous attendre à ce que le degré de naturalité des phrases avec Pds soit moins bien jugé que celui des phrases sans Pds.

Deux autres paramètres expliquent cette invalidation :

- niveau sonore des Pds trop élevé par rapport au signal général,
- Pav et Pap trop longues.

L'effet audible était une mise en évidence des Pds, rendues bien plus manifestes qu'elles ne devraient l'être.

Il ressort donc que :

a) il n'est pas possible de ne prendre en compte que les seuls paramètres temporels pour obtenir une Pds « naturelle » ;
b) le contexte phonétique affecte de manière significative la durée de la Pav ;
c) le contexte sémantico-pragmatique influe directement sur le type de Pds produite ; les restreindre à deux types standardisés (PdsC et PdsL) ne permet pas de recouvrir l'intégralité des cas de leur apparition dans le discours libre spontané ;
d) il est impératif d'analyser, de qualifier les paramètres acoustiques (intensité, Fo, timbre) en vue de produire des Pds qui font réellement sens.

L'invalidation de nos prédictions par le mauvais score de « naturel » des échantillons pourvus de Pds est à mettre au compte de la réalisation pratique du matériau testé, et du paramétrage par défaut de dimensions phonétiques que nous n'avons pu prendre en compte (timbre). Cela tendrait à confirmer l'importance du timbre des Pds, associé à leur durée, et, plus globalement, la pertinence langagière des Pds.

Cette expérience avait été montée dans le but de présenter quelques résultats chiffrés à l'appui de nos observations au congrès Eurospeech ; les résultats obtenus ne justifiant plus une présentation, nous avons renoncé à cette participation.

6. Problèmes d'insertion automatique

Outre les problèmes de réalisation technique à partir de données statistiques, un autre phénomène s'est fait jour. Le calcul des Pav et Pap via Praat et MBRola ne s'est pas avéré des plus précis. Des décalages compris entre 40 et 60ms, de surcroît irréguliers, ont été générés par l'un ou l'autre des programmes.

Il faut donc envisager une autre interface de programmation qui permettrait de supprimer ces décalages indésirables. Une fois encore, le temps à disposition n'a pas permis de mettre en place une structure plus solide et fiable.

7. Conclusion

La recherche n'aura pas épuisé la description, tant phonétique que fonctionnelle, des prises de souffle dans le discours, pas plus qu'elle n'est parvenue à un algorithme permettant d'insérer de façon automatique des Pds dans la parole de synthèse lui faisant gagner en naturel. De tels objectifs étaient d'ailleurs bien au delà de ses ambitions.

La recherche a confirmé l'importance du phénomène des Pds dans la parole spontanée ; elle a mis en évidence différentes régularités temporelles (durées des Pav/Pap ; proportionnalité des durées ; etc), et quelques éléments de leur morphologie ; elle confirme l'importance du timbre des Pds, notamment dans leur apport communicatif ou fonctionnel.

En ce qui concerne l'implémentation TTS, la pertinence de la démarche d'insertion de Pds peut être vérifiée en recourant à des échantillons « manufacturés » - mais les résultats obtenus en matière d'automatisation font apparaître différentes difficultés ; celle-ci sont liées aux limitations techniques de l'interfaçage de plusieurs systèmes, mais aussi, surtout, à l'état de nos connaissances sur les Pds.

La synthèse de la parole joue ici parfaitement le rôle de vérification d'hypothèses linguistiques que lui reconnaît E. Keller (2004) ; elle confirme l'importance du rôle des Pds dans la parole, leur diversification formelle et fonctionnelle, et la finesse de leurs réglages. Cette recherche préliminaire pose les fondations d'une étude de plus grande envergure, et démontre son intérêt tant au plan de la connaissance de l'usage communicatif du langage, qu'à celui de l'ergonomie et du confort des machines de TTS.

8. Références

Berrendonner A. (1993) « Périodes », in Parret H. (1993) (éd.) *Temps et discours*, Louvain, Presses Universitaires de Louvain, 47-61.

Boersma P. & D. Weeninck, PRAAT, logiciel d'analyse du signal : http://www.praat.org/

Garret M. (1982), Production of speech : observation from normal and pathological language use. In A. Ellis (Ed.), *Normality and pathology in cognitive functions.* UK : Academic Press.

Grobet A. & A. Auchlin (2001), « A l'attaque! Vers une typologie des différentes prises d'élan du discours », Cahiers de linguistique française 23, 165-187.

Hird, K. & Kirsner, K. (2002) « The Relationship between Prosody and Breathing in Spontaneous Discourse », *Brain and Language 80,* 536-555.

Keller E. & al. (2004), "La vérification d'hypothèses linguistiques au moyen de la synthèse de parole", *Cahiers de l'Institut de Linguistique de Louvain* 30/1-3.

MBRola, The MBROLA Project, Towards a Freely Available Multilingual Speech Synthesizer: http://tcts.fpms.ac.be/synthesis/mbrola.html

Nullsoft Winamp : http://www.winamp.com

Roulet E., Filliettaz L., Grobet A., Burger, M. (2001), *Un modèle et un instrument d'analyse de l'organisation du discours*, Berne, Lang.

Simon A. C. (2004), *Segmentation et structuration prosodiques du discours*. Bern : Peter Lang.

Winkworth, A., Davis,P. , Adams, R. & Ellis,E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research*, **38**(1), 124-144.