Patrons Rythmiques et Genres Littéraires en Synthèse de Parole

Elisabeth Delais-Roussarie[1], Damien Lolive[2], Hiyon Yoo[1] et David Guennec[2]
(1) LLF - UMR 7110 & Université Paris-Diderot, France
(2) IRISA - UMR 6074, Université Rennes 1, France
elisabeth.roussarie@wanadoo.fr, damien.lolive@irisa.fr,
yoo@linguist.univ-paris-diderot.fr, david.guennec@irisa.fr

RÉSUMÉ

Ces vingt dernières années, la qualité de la parole synthétique s'est améliorée grâce notamment à l'émergence de nouvelles techniques comme la synthèse par corpus. Mais les patrons rythmiques obtenus ne sont pas toujours perçus comme très naturels. Dans ce papier, nous comparons les patrons rythmiques observés en parole naturelle et synthétique pour trois genres littéraires. Le but de ce travail est d'étudier comment le rythme pourrait être amélioré en synthèse de parole. La comparaison des patrons rythmiques est réalisée grâce à une analyse de la durée relativement à la structure prosodique, les données audio provenant de six comptines, quatre poèmes et deux extraits de conte. Les résultats obtenus laissent penser que les différences rythmiques entre parole naturelle et synthétique sont principalement dues au marquage de la structure prosodique, particulièrement au niveau des groupes intonatifs. De fait, le taux d'allongement des syllabes accentuées en fin de groupes intonatifs est beaucoup plus important en synthèse que dans la parole naturelle.

ABSTRACT

How to improve rhythmic patterns according to literary genre in synthesized speech *.

In the last twenty years, the quality of synthesized speech has greatly improved with the emergence of new TTS techniques, including corpus-based synthesis systems. Yet the rhythmic patterns obtained do not always sound very natural. In this paper, we compare the rhythmic patterns observed in natural and synthesized speech for three literary forms. The aim of the study is to evaluate how rhythm could be improved in synthesized speech. The comparison of the rhythmic patterns is done by analyzing duration in relation to prosodic structure on a set of texts (six rhymes, four poems and two extracts from fairy tales). This approach allows showing that rhythmic differences between synthesized and natural speech are mostly due to the marking of prosodic structure, especially at the level of the intonational phrase. The lengthening rate for accented syllables located at the end of IPs is much more important in synthesized speech than in natural speech.

MOTS-CLÉS : Patrons rythmiques, phonogenre, synthèse de la parole, structure prosodique.

KEYWORDS: Rhythmic patterns, phono-genre, speech synthesis, prosodic structure..

1 Introduction

Ces dernières décennies, la qualité globale de la parole synthétisée s'est améliorée de façon notable avec l'émergence de nouvelles techniques de synthèse comme la synthèse par corpus (Sagisaka, 1988; Hunt & Black, 1996). Néanmoins, générer une prosodie naturelle qui tienne compte des genres et

∗. Cet article est tiré d'une publication par les mêmes auteurs à la conférence Speech Prosody 2016.

styles de parole reste un challenge (Schröder, 2009; Obin, 2011), en particulier pour les aspects rythmiques. De fait, la composante rythmique semble souvent peu naturelle en synthèse de parole et doit être améliorée pour permettre une meilleure utilisation de la synthèse dans de nombreuses applications (jeu vidéo, logiciel éducatif, lecture de livres audio, etc.).

Dans un projet de recherche visant à utiliser la synthèse de parole pour favoriser l'apprentissage de l'écriture à des enfants de cycle 2 (CP, CE1), il fallait améliorer le système de synthèse de parole afin qu'il puisse lire de façon claire et naturelle des contes, des poèmes et des comptines. Pour tenter d'atteindre cet objectif, nous avons comparé les patrons rythmiques obtenus en parole naturelle et en parole synthétique pour chacun des genres littéraires visés (comptines, poèmes, contes). Nous avions émis au départ l'hypothèse que les patrons rythmiques les plus précis seraient observés pour les contes, les corpus utilisés pour extraire les unités de parole lors de la synthèse contenant essentiellement des textes lus comparables à des récits. Cependant, les résultats obtenus n'ont pas confirmé cette hypothèse, les lectures de comptines étant souvent plus satisfaisantes. Aussi, avons-nous tenté de comprendre pourquoi les patrons rythmiques sont plus adéquats dans le cas des poèmes et des comptines que dans le cas des contes, alors que les corpus utilisés pour générer les stimuli ne contenaient pas ce genre littéraire.

Cet article est organisé de la manière suivante. La section 2 fournit une description des données et de la méthode utilisée. Dans la section 3, les résultats obtenus grâce à la comparaison des patrons de durée et de débit de parole dans les deux types de parole (parole naturelle vs. parole synthétisée) sont présentés pour chaque genre. Ces résultats sont ensuite discutés dans la section 4, le but de cette discussion étant essentiellement de voir comment améliorer les systèmes de synthèse.

2 Corpus et méthodologie

2.1 Corpus

Le corpus utilisé pour étudier les patrons rythmiques en parole naturelle et synthétique est constitué de trois types distincts de textes adressés à des enfants : six comptines, quatre poèmes et deux extraits de contes. Le tableau 1 présente la composition quantitative du corpus par genre littéraire. Les différences de réalisation entre les locuteurs (qui sont indirectement notées par l'écart entre le nombre de syllabes par genre à multiplier par cinq, c'est-à-dire le nombre de locuteurs, et le nombre effectif de syllabes) résultent principalement de l'insertion ou de l'élision de schwas, ou de l'omission d'un mot. Le nombre effectif de syllabes obtenu pour la parole synthétique est donné entre parenthèses.

Genre littéraire	Nombre de mots	Nombre de syll.	Nombre de syll. effectif
Comptines	158	228	1137 (454 for synth.)
Poèmes	290	422	2155 (808 for synth.)
Contes	522	777	3861 (1538 for synth.)
Total	970	1427	7153 (2800)

TABLE 1 – Composition du corpus

L'ensemble des textes a été produit par cinq voix différentes (deux voix synthétiques et trois voix naturelles). Pour les voix naturelles, le corpus a été enregistré par trois locuteurs (deux hommes et une femme) dans un studio d'enregistrement. Les participants ont eu le temps de lire les textes et de

les répéter avant l'enregistrement. Parmi ces locuteurs, deux ont lu les textes de la même manière que des parents liraient une histoire à leurs enfants, tandis que le troisième est un acteur confirmé et les a lus avec beaucoup plus d'expressivité.

Les stimuli synthétisés ont été produits grâce au système de synthèse par corpus présenté dans (Guennec & Lolive, 2014), pour lequel des filtres de pré-sélection sont utilisés en lieu et place d'un coût cible. Pour cette étude, les filtres ordonnés utilisés au niveau du phonème sont les suivants :

1. Identité dans les étiquettes associées aux segments (obligatoire).
2. Nature de l'unité : phonème ou autre ? (obligatoire)
3. Est-ce que le segment/phonème est dans la dernière syllabe de la phrase ?
4. Le segment se situe-t-il dans la dernière syllabe d'un groupe prosodique majeur (IP) ?
5. Le segment est-il dans la dernière syllabe d'un mot ?
6. Le segment est-il dans une syllabe réalisée avec une intonation montante ?

Lors de la recherche, si le nombre d'unités correspondant à un ensemble de filtres est insuffisant, le dernier filtre de l'ensemble est relâché. Cela permet d'élargir le champ de recherche en réduisant le nombre de contraintes appliquées. Dans tous les cas, les deux premiers filtres sont toujours utilisés. De manière complémentaire, une pénalité est appliquée pour les classes de phonèmes pour lesquelles la concaténation semble risquée (Alain *et al.*, 2015). Par exemple, une concaténation sur une voyelle est plus sujette à l'apparition d'un artéfact audible qu'une concaténation réalisée sur la partie silencieuse d'une plosive ou même sur une fricative.

Concernant la prosodie, aucun traitement spécifique n'est réalisé, et les seules contraintes susceptibles d'améliorer le rythme de la parole générée sont les filtres de pré-sélection. De fait, certains d'entre eux permettent d'appliquer aux unités sélectionnées des contraintes positionnelles comme par exemple la fin d'énoncés ou la fin de mot. En ce qui concerne les pauses, elles sont placées de manière arbitraire, et une pause de durée fixe est insérée après chaque marque de ponctuation.

Comme nous l'avons déjà mentionné, deux voix de synthèse ont été utilisées pour cette étude :

— la voix d'homme SY-P, construite à partir de 10 heures de parole extraites d'un livre audio (un roman lu par un acteur) ;
— la voix de femme SY-A, construite à partir de 7 heures de parole lue, les éléments lus ayant été sélectionnés spécifiquement pour la construction d'un système de synthèse. Les différences de contenu et de taille des corpus amènent à considérer SY-P comme une voix plus expressive que SY-A, qui est plus neutre.

Pour générer les stimuli de synthèse, la structure des strophes et des vers dans les poèmes et comptines a été représentée par des marques de ponctuation. Ainsi, pour obtenir la version synthétisée, les trois strophes sous (1), extraites du poème "La fourmi" de R. Desnos ont été mises en forme comme indiqué sous (2).

(1) *Une fourmi traînant un char*
plein de pingouins et de canards
ça n'existe pas, ça n'existe pas

Une fourmi parlant français
parlant latin et javanais
ça n'existe pas, ça n'existe pas

eh ! et pourquoi pas !

(2) Une fourmi traînant un char, plein de pingouins et de canards, ça n'existe pas, ça n'existe pas. Une fourmi parlant français, parlant latin et javanais, ça n'existe pas, ça n'existe pas. Eh ! Et pourquoi pas !

Comme on peut le voir, la fin des strophes est systématiquement indiquée par un point même s'il n'y avait pas de ponctuation dans le texte original. La fin des vers est retranscrite par une virgule, sauf dans le cas où une ponctuation existait déjà.

2.2 Méthodologie

Les enregistrements audio ont d'abord été transcrits et segmentés en phrases sous PRAAT (Boersma & Weenink, 2001). La transcription orthographique a ensuite été phonétisée, et le signal acoustique automatiquement segmenté en phones, syllabes, et mots avec EASYALIGN (Goldman, 2011). Les transcriptions phonétiques et les segmentations acoustiques ont été vérifiées et corrigées, si nécessaire. L'ensemble des données a été utilisé pour l'analyse rythmique et prosodique.

La voyelle plutôt que la syllabe a été choisie comme unité de base pour générer les patrons de durée et pour analyser et comparer les durées des pauses et les débits en fonction des genres et des locuteurs. Ce choix résulte du fait que les structures syllabiques varient beaucoup en français (entre, par exemple, des syllabes de forme CCVC et d'autres de forme CV). La durée des syllabes ne constituent donc pas un indicateur robuste pour évaluer les taux d'allongement. Comme le nombre de voyelles par contextes prosodiques était limité en raison de la taille du corpus, aucune normalisation des durées n'était possible. Aussi avons-nous décidé de faire une distinction entre voyelles courtes et voyelles longues, même si cette distinction n'existe pas dans le système phonologique du français. Les voyelles nasales ([ɔ̃],[ɑ̃], [ɛ̃] et [œ̃]) et les séquences composées d'une semi-voyelle et d'une voyelle en position de noyau (par exemple, [jɛ̃] dans *tiens* [tjɛ̃], [wa] dans *noir* [nwaʁ]) ont été codées comme des voyelles longues, tandis que les autres voyelles ont été considérées comme courtes.

De nombreux travaux consacrés à la prosodie du français ont montré que l'intonation et l'accentuation sont très liées dans cette langue (cf. (Post, 2011)) ; aussi avons-nous décidé de partir des découpages prosodiques pour étudier les schémas rythmiques. Les différents textes ont donc été segmentés en groupes prosodiques, une distinction étant faite entre trois niveaux de structuration : le mot prosodique (MP) qui correspond à un mot lexical précédé des mots grammaticaux qui en dépendent, le syntagme phonologique (SP) qui est borné à droite par une tête lexicale de projection syntagmatique maximale et le groupe intonatif (IP). Pour pouvoir comparer les données malgré les différences possibles de réalisation et pour éviter une certaine circularité, nous avons décidé de dériver les unités prosodiques à partir du texte, et plus précisément des informations morpho-syntaxiques, voir (Delais-Roussarie, 1996; Martin, 1987; Pasdeloup, 1992). De plus, comme la dernière syllabe des groupes prosodiques est considérée en français comme accentuée et est habituellement allongée (Fletcher, 1991), trois catégories de syllabes accentuées ont été retenues pour comparer les taux d'allongement par rapport à la position prosodique :

— AC-MP correspond à la dernière syllabe accentuée d'un mot prosodique, c'est-à-dire d'un mot d'une catégorie lexicale tels que le verbe V, le nom N, l'adjectif A ou l'adverbe Adv (voir, entre autres, (Mertens *et al.*, 2001; Nespor & Vogel, 1986)) ;
— AC-SP coïncide avec la dernière syllabe accentuée d'un syntagme phonologique, c'est-à-dire la dernière syllabe accentuée d'une tête lexicale de projection syntaxique (voir, par exemple, (Delais-Roussarie, 1996; Post, 2000; Selkirk, 1986)) ;

— AC-IP correspond à la dernière syllabe accentuée d'un IP, les frontières d'IP étant localisées à la fin des clauses, des constituants syntaxiques détachés, ou, dans les poèmes et les comptines, des vers (voir, entre autres, (Nespor & Vogel, 1986; Delais-Roussarie *et al.*, 2015; Portes & Bertrand, 2011)).

3 Résultats

L'étude des durées observées pour les voix naturelles et synthétiques a permis de comparer les débits de parole, la durée et la distribution des pauses, et le marquage de la structure prosodique.

3.1 Débit de parole et pauses

La durée totale des textes lus a été utilisée pour calculer, pour chaque locuteur et pour chaque genre, le débit de parole, le taux d'articulation et les durées des pauses. Les différences entre débit de parole et taux d'articulation reposent sur le fait que les pauses ne sont pas prises en compte dans le second cas (Simon *et al.*, 2010). Le tableau 2 présente les résultats obtenus pour chaque locuteur dans les trois genres. Pour chacun d'entre eux, les deux premières lignes concernent le débit de parole et le taux d'articulation, tandis que les deux dernières lignes portent sur la durée.

Comptines	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	9.9	7.35	7.08	7.63	9.09
Taux d'articulation moyen (ph./sec)	12.09	7.83	8.53	9.79	12.61
Durée totale des pauses (ms)	2178.92	1449.22	2573.76	3025	3000
% de pause moyen	25.27	13.60	24.15	29.06	33.80
Poèmes	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	10.6	8.16	6.28	8.26	9.45
Taux d'articulation moyen (ph./sec)	13.60	9.32	8.72	10.70	12.85
Durée totale des pauses	1534	1373.36	2590.52	2000	2000
% de pause moyen	27.38	18.10	33.85	28.17	31.29
Contes	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	10.58	8.74	8.18	9.31	10.79
Taux d'articulation moyen (ph./sec)	14.99	10.08	11.09	11.36	13.68
Durée totale des pauses	1331.33	763.40	1482.06	992	992.14
% de pause moyen	32.82	18.06	29.96	21.96	24.79

TABLE 2 – Débit de parole et taux d'articulation en phones/sec, durée et pourcentage de pauses (relativement à la durée totale de lecture).

Les taux d'articulation et les débits de parole observés pour chaque genre varient de manière importante, mais on ne peut pas dire que les voix de synthèse diffèrent des voix naturelles : LOD et SY-P possèdent pour les trois genres considérés un débit plus rapide que les locuteurs SY-A, GOR et DRE (qui ont des débits plus lents, mais relativement semblables). Si on compare pour un genre donné les débits des différentes voix, on s'aperçoit que les taux d'articulation obtenus par la synthèse sont dans les limites de ceux observés pour les voix naturelles.

Une comparaison inter-genres montre que les locuteurs adaptent leur débit de parole et leur taux d'articulation en fonction du genre, des débits plus lents étant mis en œuvre pour la lecture des comptines et des poèmes. Cette adaptation est, comme on s'y attendait, moins claire pour la parole synthétique. De fait, pour une voix donnée et pour tous les genres, le même corpus et la même procédure de sélection d'unités sont utilisés. Néanmoins, les différences entre voix naturelles et voix synthétiques demeurent mineures, ce qui signifie que l'adaptation découle également de la composition interne des textes.

Concernant les pauses, une différence importante existe entre parole naturelle et parole synthétique dans tous les genres. La proportion de pauses est moins importante dans les comptines que dans les contes pour les trois locuteurs ; en revanche, il y a plus de pauses dans les comptines que dans les contes pour les deux voix de synthèse. Vus les mécanismes de placement des pauses utilisés par le synthétiseur, ces résultats sont tout à fait logiques. De plus, en parole naturelle, la durée des pauses semble dépendre du taux d'articulation (la proportion de pauses est en effet moins importante lorsque le débit est lent, comme par exemple dans les comptines et les poèmes) ; mais une telle corrélation n'apparaît pas en synthèse, une durée fixe étant assignée aux pauses en fonction de la force de la frontière prosodique.

Dans l'ensemble, on n'observe pas de grosses différences entre parole naturelle et synthétique pour le débit de parole et le taux d'articulation. En effet, les voix de synthèse et les voix naturelles varient dans les mêmes proportions. En revanche, pour la durée et la proportion des pauses, il existe des différences entre la synthèse et les voix naturelles.

3.2 Structure prosodique et durée

En règle générale, les allongements indiquent en français le phrasé et l'accentuation. Les syllabes accentuées, qui correspondent à la dernière syllabe pleine à chaque niveau de structuration prosodique, sont allongées, leur taux d'allongement étant proportionnel à la force de la frontière prosodique, voir, entre autres, (Post, 2000; Portes & Bertrand, 2011). Aussi avons-nous voulu vérifier si cela se retrouve dans les voix de synthèse. Pour ce faire, les taux d'allongement ont été calculés en comparant les durées des voyelles dans les syllabes non accentuées aux durées des segments vocaliques dans les syllabes accentuées, et cela à tous les niveaux de hiérarchie prosodique (mot prosodique, syntagme phonologique et groupe intonatif). Le tableau 3 présente les résultats obtenus par genre.

Il existe une variation relativement importante de la durée des voyelles non accentuées dans les différents genres, et cela pour les trois voix naturelles. De manière générale, les voyelles en position non accentuée sont plus longues dans les comptines et les poèmes que dans les contes. Par comparaison, aucune variation claire n'est observée entre genres pour les voix synthétisées. Ce résultat confirme le fait que les locuteurs adaptent leur débit de parole en fonction du genre, ce que ne fait pas la synthèse de parole.

En ce qui concerne le marquage de la structuration prosodique, des allongements se produisent toujours à la fin des groupes prosodiques à tous les niveaux (mot prosodique, syntagme phonologique et groupe intonatif), en synthèse comme en parole naturelle. Pour tous les genres et tous les locuteurs, les taux d'allongement varient

- de 10 à 40%, au niveau du mot prosodique, avec une moyenne aux alentours de 20% ,
- de 20 à 100% au niveau du syntagme phonologique, avec une moyenne à 40%,
- de 60 à 190% au niveau du groupe intonatif, avec une moyenne de 98% (avec 77% en parole naturelle et 128% en synthèse).

Les taux moyens (à l'exception des IP en parole synthétique) correspondent à ceux souvent donnés dans les travaux sur les patrons de durée en français (Delais-Roussarie, 1996; Pasdeloup, 1992). Dans les comptines, les taux d'allongement ne permettent pas toujours de clairement distinguer les trois niveaux de structuration, en particulier les mots prosodiques des syntagmes phonologiques. Dans les poèmes, la distinction entre SP et IP n'est pas clairement marquée dans les taux d'allongement chez LOD et GOR. On peut aussi noter que les taux d'allongement qui indiquent les frontières des IP sont plus nettement marqués en synthèse qu'en parole naturelle, dans tous les genres, et plus particulièrement chez SY-A.

Comptines	LOD	DRE	GOR	SY-A	SY-P
Durée moyenne voy. non accentuée	66 ms	130 ms	93 ms	81 ms	68 ms
Taux d'allongement AC-MP	30%	20%	20%	20%	30%
Taux d'allongement AC-SP	20%	20%	50%	30%	10%
Taux d'allongement AC-IP	90%	70%	70%	150%	60%
Poèmes	**LOD**	**DRE**	**GOR**	**SY-A**	**SY-P**
Durée moyenne voy. non accentuée	67 ms	110 ms	95 ms	78 ms	69 ms
Taux d'allongement AC-MP	10%	20%	40%	20%	20%
Taux d'allongement AC-SP	60%	40%	100%	50%	40%
Taux d'allongement AC-IP	60%	70%	80%	190%	80%
Contes	**LOD**	**DRE**	**GOR**	**SY-A**	**SY-P**
Durée moyenne voy. non accentuée	59 ms	99 ms	78 ms	77 ms	65 ms
Taux d'allongement AC-MP	10%	20%	10%	20%	20%
Taux d'allongement AC-SP	20%	40%	50%	40%	30%
Taux d'allongement AC-IP	80%	80%	100%	190%	100%

TABLE 3 – Durées moyennes des voyelles dans les syllabes non accentuées (en ms.) et taux d'allongement (en %) pour les trois niveaux de structuration (MP, SP et IP).

Globalement, les patrons de durée obtenus pour la parole synthétique sont relativement comparables à ceux observés en parole naturelle : les différents niveaux de phrasé sont toujours indiqués par un allongement dont le taux varie, très souvent, proportionnellement à la force de la frontière (Post, 2000; Delais-Roussarie *et al.*, 2015; Delais-Roussarie & Feldhausen, 2014).

4 Discussion

Les comparaisons effectuées ne révèlent pas de différences notables entre parole synthétique et parole naturelle. De fait, les variations qui apparaissent pour le débit de parole et le taux d'articulation ne permettent pas de distinguer la parole naturelle de la parole synthétique. En ce qui concerne les allongements et le marquage des frontières prosodiques, l'analyse montre clairement que des allongements sont réalisés en fin de groupement prosodique dans les deux types de parole (naturelle et synthétique), même si des différences apparaissent dans l'importance des taux d'allongement observés au niveau des IP (plus important en synthèse qu'en parole naturelle). Néanmoins, on peut douter que ces différences expliquent à elles seules le manque de naturel de la parole synthétique. De plus, en écoutant les stimuli synthétisés, nous avons été surpris par la qualité des patrons rythmiques observés dans les comptines, en particulier pour SY-A : ils semblent très naturels en comparaison de ceux obtenus pour les contes. En conséquence, les problèmes de rythme rencontrés en synthèse ne

peuvent pas être attribués à des "sur-allongements" au niveau des IPs.

Les taux d'articulation et le débit d'une part, et le marquage par la durée de la structuration prosodique d'autre part, ne peuvent expliquer le manque de naturel dans les motifs rythmiques. Il faut donc trouver d'autres explications. Deux pistes de recherche méritent selon nous d'être explorées : (i) aucune corrélation entre le débit de parole, la force des frontières et la durée des pauses n'a été observée en parole synthétique, alors que cette corrélation existe en parole naturelle (en français, de nombreuses études ont montré que les groupes prosodiques comme le syntagme phonologique ou le groupe intonatif visent soit à posséder le même nombre de syllabes soit la même durée (Delais-Roussarie, 1996; Martin, 1987; Pasdeloup, 1992; Wioland, 1991), les durées des pauses pouvant alors jouer un rôle important dans la recherche de l'isochronie) ; et (ii) les patrons intonatifs jouent probablement un rôle dans la réalisation des motifs rythmiques : en insérant une virgule à la fin de chaque vers dans les poèmes et les comptines, on a forcé la réalisation d'un contour mélodique non final montant (contour de continuation majeure), ce qui a conduit à la répétition régulière d'une forme mélodique et a renforcé l'impression de rythme, ces résultats laissant penser que la récurrence de motifs mélodiques est cruciale pour le rythme.

5 Conclusion et perspectives

L'analyse des patrons de durée observés en parole naturelle et synthétique dans trois genres littéraires montre clairement que la durée ne peut pas, à elle seule, expliquer le manque de naturel de la synthèse de parole, notamment pour les aspects rythmiques. Les valeurs obtenues pour les durées segmentales et pour le marquage de la structure prosodique sont comparables dans bien des cas. Des travaux complémentaires sur des corpus plus grands sont donc nécessaires. De plus, trois points peuvent être avantageusement intégrés pour améliorer la procédure de sélection d'unités dans le système de synthèse de parole, et par là-même, les patrons rythmiques et prosodiques :

— Distinguer clairement les différents niveaux de structuration prosodique et les prendre en compte dans la sélection des unités, notamment comme contrainte positionnelle ;
— Prendre en compte la forme des mouvements mélodiques réalisés sur les syllabes accentuées : la procédure qui a été utilisée dans les comptines et les poèmes en forçant à l'insertion de contours intonatifs similaires à intervalles réguliers a donné des résultats très satisfaisants ;
— Adapter les taux d'allongements, les taux d'articulation et la durée des pauses en fonction des genres, mais aussi pour rendre compte des corrélations qui existent entre ces éléments.

Remerciements

Le travail présenté ici a été soutenu par l'opération PPC 7 du Labex "Empirical Foundations in Linguistics" (ANR-10-LABX-0083). Il a également bénéficié du soutien financier de l'Agence Nationale de la Recherche dans le cadre du projet ANR SynPaFlex (ANR-15-CE23-0015).

Références

ALAIN P., CHEVELU J., GUENNEC D., LECORVÉ G. & LOLIVE D. (2015). The IRISA Text-To-Speech system for the blizzard challenge 2015. In *Proc. of the Blizzard Challenge 2015 Workshop*.

BOERSMA P. & WEENINK D. (2001). Praat, a system for doing phonetics by computer.

DELAIS-ROUSSARIE E. (1996). *Phonological phrasing and accentuation in French*, In M. NESPOR & N. SMITH, Eds., *Dam Phonology : HIL phonology papers II, den Haag : Holland Academic Graphics*, p. 1–38.

DELAIS-ROUSSARIE E. & FELDHAUSEN I. (2014). "variation in prosodic boundary strength : a study on dislocated XPs in french". In *Proc. of Speech Prosody*, p. 1052—1056.

DELAIS-ROUSSARIE E., POST B., AVANZI M., BUTHKE C., DI CRISTO A., FELDHAUSEN I., JUN S., MARTIN P., MEISENBURG T., RIALLAND A. *et al.* (2015). *Intonational Phonology of French : Developing a ToBI system for French*, In S. FROTA & P. PRIETO, Eds., *Intonation in Romance*, p. 63–100. Oxford University Press.

FLETCHER J. (1991). Rhythm and final lengthening in french. *Journal of phonetics*, **19**(2), 193–212.

GOLDMAN J.-P. (2011). Easyalign : an automatic phonetic alignment tool under praat. In *Proc. of Interspeech*, p. 3233–3236.

GUENNEC D. & LOLIVE D. (2014). Unit selection cost function exploration using an A* based text-to-speech system. In *Proc. of TSD*, p. 432–440 : LNCS, Springer, Heidelberg.

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP*, volume 1, p. 373–376.

MARTIN P. (1987). Prosodic and rhythmic structures in french. *Linguistics*, **25**(5), 925–950.

MERTENS P., GOLDMAN J.-P., WEHRLI E. & GAUDINAT A. (2001). La synthèse de l'intonation à partir de structures syntaxiques riches. *Traitement Automatique des Langues*, **42**(1), 145–192.

NESPOR M. & VOGEL I. (1986). Prosodic phonology, vol. 28. *Dordrecht : Foris*.

OBIN N. (2011). *MeLos : Analysis and modelling of speech prosody and speaking style*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

PASDELOUP V. (1992). *A prosodic model for French text-to-speech synthesis : A psycholinguistic approach*, In C. B. G. BAILLY & T. SAWALLIS, Eds., *Talking Machines. Theories, Models and Designs*, p. 335–348. Elsevier Science Publishers.

PORTES C. & BERTRAND R. (2011). Permanence et variation des unités prosodiques dans le discours et l'interaction. *Journal of French Language Studies*, **21**(01), 97–110.

POST B. (2000). Tonal and phrasal structures in french intonation. *The Hague : Holland Academic Graphics*.

POST B. (2011). *The multi-faceted relation between phrasing and intonation contours in French*, In C. GABRIEL & C. LLEÒ, Eds., *Intonational Phrasing in Romance and Germanic : Crosslinguistic and bilingual studies*, p. 44–74. Amsterdam : Benjamins.

SAGISAKA Y. (1988). Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. of ICASSP*, volume 1, p. 679–682.

SCHRÖDER M. (2009). Expressive speech synthesis : Past, present, and possible futures. In *Affective information processing*, p. 111–126. Springer.

SELKIRK E. (1986). On derived domains in sentence phonology. *Phonology*, **3**(01), 371–405.

SIMON A.-C., AUCHLIN A., AVANZI M., GOLDMAN J.-P. *et al.* (2010). *Les phonostyles : une description prosodique des styles de parole en français*, In M. ABÉCASSIS & G. LEDEGEN, Eds., *Les voix des français : en parlant, en écrivant*. Bern : Lang.

WIOLAND F. (1991). *Prononcer les mots du français. Des sons et des rythmes*. Paris : Hachette.