Un modèle de durée des syllabes fondé sur les propriétés syllabiques intrinsèques et les variations locales de débit

N. Obin, X. Rodet, A. Lacheret-Dujour

IRCAM, IRCAM, Université Paris X

nobin@ircam.fr, rodet@ircam.fr anne@lacheret.com

ABSTRACT

Local speech rate is an emergent field in research on prosody. This paper introduces a syllable duration model based on intrinsic syllable duration properties and local speech rate variations. The proposed model is compared to the observed syllable durations and to a standard model in which durations are normalized according to local speech rate only. This comparison shows that our model is i) robust : significant reduction of the observed syllable dispersion, ii) consistant : this reduction comes with reduction of duration dispersion due to the syllable intrisic properties as well as prominence phenomena.

Keywords prosody, syllable duration, local speech rate, syllable intrisic properties, model

1. Introduction

L'étude du rythme est un domaine majeur dans le traitement de la parole, domaine qui a été abondamment traité tant du point de vue de la production que de la perception. La durée syllabique et/ou la durée inter centre perceptif [1] sont considérées comme les deux unités d'encodage rythmique privilégiées en traitement automatique de la parole. La compréhension de leur organisation temporelle demeure cependant un sujet d'étude ouvert dans l'analyse et la modélisation de la prosodie. A un niveau d'organisation globale, des études ont montré que le débit - défini comme l'inverse de la durée moyenne des syllabes considérées sur l'ensemble d'un corpus donné - influe significativement sur la structure de la parole, en terme d'organisation phonématique [11], syllabique (distribution de la durée des syllabes), et suprasegmentale (contraste de la durée des syllabes proéminentes et des syllabes non-proéminentes [6]). Ces études distinguent généralement 2 ou 3 classes de débit global qui influent significativement sur la distribution de la durée des syllabes [12], [6].

Dans un même temps, ces études font implicitement l'hypothèse que le débit ne varie pas significativement sur un corpus donné. Si cette hypothèse peut être partiellement soutenue pour de parole de laboratoire contrôlée [12] [6] , elle n'est en revanche plus valide dès lors que l'on aborde un corpus tout venant où les enjeux communicationnels peuvent entraîner des variations temporelles significatives d'un site discursif à un autre, ou *variations locales de débit*. Cette hypothèse résulte habituellement en une dispersion difficilement interprétable dans la distribution des durées [12]. Pour résoudre ce problème, il est nécessaire d'intégrer à l'analyse des durées la notion de débit local et d'en donner une formulation adéquate. L'intégration de cette notion dans l'étude de l'organisation prosodique des durées présente en outre l'avantage d'enrichir significativement les représentations en usage dans le sens où elle permet de quantifier la manière dont les durées des syllabes s'organisent localement.

De récentes recherches ont proposé et étudié des modèles d'estimation du débit local. Ces études se divisent en : i) l'estimation du *débit local* syllabique (respectivement phonématique) [7] [8], et ii) l'estimation du *débit relatif* [4] [5]. Nous suggérons qu'une formulation adéquate du débit local permet de mieux rendre compte de la distribution des syllabes dans l'analyse des durées et de fournir un modèle de prédiction des durée plus robuste pour la synthèse de la parole.

Dans un premier temps, nous présentons les modèles d'estimation du débit local et la manière dont cette estimation est utilisée pour réduire la dispersion des durées observées. Ensuite, nous présentons un modèle de durée syllabique fondé sur les propriétés temporelles intrinsèques de la syllabe couplées à une modulation induite par les variations locales de débit. Enfin nous comparons le modèle proposé aux durées observées des syllabes ainsi qu'aux durées des syllabes normalisées par les variations de débit local sans prise en compte des propriétés intrinsèques des syllabes.

2. Les modèles de débit local

2.1. Le débit local

Le *débit local syllabique* (respectivement phonématique) [7] [8] repose sur l'estimation de l'inverse de la durée moyenne des syllabes considérées dans une fenêtre d'observation à court terme. Les auteurs ont proposé à partir de cette estimation un modèle de durées des phonèmes dans lequel la durée de chaque phonème est normalisée par rapport aux variations de débit local observées. Ce modèle présente l'intérêt particulier de réduire significativement la dispersion de la durée des phonèmes [9]. Cependant une analyse de la distribution des durées résultant de cette normalisation montre que les facteurs types de syllabes et accentuation conservent une influence significative dans la dispersion des durées observées. Ceci suggère une certaine inconsistance du modèle proposé dans le sens où nous attendrions que la réduction de la dispersion des durées s'accompagne d'une réduction

de l'influence du phénomène de proéminence dont le modèle est justement supposé rendre compte. Nous faisons l'hypothèse que cette propriété est principalement due au fait que ce modèle ne prend pas en compte les propriétés intrinsèque des phonèmes, ce qui a pour conséquence de ne pas distinguer la part respective de la nature des phonèmes et des variations locales de débit dans l'observation des durées. Ce problème s'accroit dès lors que l'on se place dans le cadre de l'étude de la durée des syllabes : les propriétés intrinsèques des syllabes diffèrent largement entre elles, étant elles-mêmes composées non seulement de phonèmes présentant des propriétés intrinsèques différentes mais encore dans des proportions (nombre de phonèmes) et des séquences (agencement des ces phonèmes dans la syllabe) variables.

2.2. Le débit relatif

L'estimation du *débit relatif* [4] [5] repose sur un alignement temporel des coefficients cepstraux d'une phrase observée et d'une phrase considérée comme référence. S'il permet de s'affranchir de l'estimation des frontières des syllabes, ce modèle présente par contre des inconvénients majeurs : i) il introduit une notion de référence arbitraire et difficilement interprétable, ii) la référence proposée par les auteurs nécessite un nombre suffisant de répétitions pour chaque phrase considérée, ce qui ne semble pas réaliste dans le cadre de systèmes de synthèse de la parole à base de corpus, et iii) le modèle n'a pas été étudié en terme de propriétés sur l'observation des durées.

3. Modèle de durée syllabique

Sur la base des observations présentées, nous présentons un modèle de durée qui se place dans le cadre théorique de [9] tout en y intégrant une formulation de la notion de référence présentée dans le modèle de *débit relatif*. Nous faisons l'hypothèse que la durée d'une syllabe observée dépend i) des propriétés temporelles intrinsèques de la syllabe considérées comme référence. Une telle définition d'une référence devient consistante puisqu'elle explicite le processus de représentation phonologique de la durée des syllabes. ii) des variations locales de débit occurant sur cette syllabe étant données ses propriétés intrinsèques. Ce modèle devrait rendre consistant l'analyse de la distribution des durées en permettant de réduire la dispersion des durées observées et d'accompagner cette réduction d'une réduction de l'influence de la nature des syllabes et du phénomène de proéminence. Dans notre modèle, la durée d'une syllabe observée D_{obs} résulte de : i) une durée résultant des propriétés de durée intrinsèques D_{int} de la syllabe considérée (composante micro-prosodique), ii) une modulation $\alpha_{macro|int}$ résultant des variations locales de débit occurant sur la syllabe considérée étant donné ses propriétés intrinsèques (composante macro-prosodique).

$$D_{model} = \alpha_{macro|int} D_{int}$$

L'estimation des paramètres de notre modèle nécessite en conséquence l'estimation des propriétés intrinsèques des durées des syllabes ainsi que l'estimation du débit local conditionnellement aux propriétés intrinsèques des syllabes.

3.1. Modélisation des propriétés temporelles intrinsèques de la syllabe

Les phonèmes ainsi que les syllabes ont des propriétés de durée intrinsèques dépendant de leur nature (nature du phonème et structure de la syllabe). D'un point de vue perceptif, ces propriétés renvoient à la notion de durées phonologiques qui repose essentiellement sur une représentation mentale de la durée attendue d'une syllabe. Le modèle que nous proposons pour l'estimation des propriétés syllabiques intrinsèques repose sur trois étapes : i) une description symbolique des attributs de la structure syllabique, ii) une sélection des attributs qui permettent de rendre compte de manière optimale de la durée des syllabes et une estimation des durée intrisèques par arbre de décision.

Description symbolique de la syllabe L'estimation des propriétés intrinsèques des syllabes suppose que ces dernières présentent des caractéristiques attendues qui sont fonction uniquement de leurs structures propres, c'est-à-dire indépendantes de toute manifestiation prosodique particulière. Suivant cette hypothèse, nous proposons d'estimer ces durées exclusivement à partir d'une description interne de la syllabe. Nous avons choisi de décrire la structure interne de la syllabe sur la base de l'ensemble des attributs symboliques suivant : S_{symb} = {nombre de phonèmes dans la syllabe, agencement des phonèmes dans la syllabe (ex. CVC), label du nucleus, nombre de phonèmes de l'onset, agencement des phonèmes dans l'onset, nombre de phonèmes de la coda, agencement des phonèmes dans la coda}. Notre représentation exclue de fait tout attribut qui donnerait lieu à une information prosodique (position de la syllabe dans la phrase,...), de manière à neutraliser l'effet des paramètres prosodiques lors de l'estimation des propriétés intrinsèques de la syllabe.

Modèle de durée intrinsèque par arbre de décision L'estimation des durées phonologiques à partir de l'ensemble complet des attributs symboliques S_{symb} pose de fait le problème sa robustesse. Nous n'avons en effet pas la garantie d'avoir un nombre suffisant d'observations pour chaque description symbolique observée dans notre corpus. Dans un même temps, nous désirons réduire la dimension de notre espace symbolique à un sous-ensemble optimal en regard de l'estimation des durées correspondantes. Ce faisant, nous avons choisi d'estimer ce sous-ensemble à partir d'une sélection de descripteur par arbre de décision. Cette étape nous permet de réduire l'ensemble symbolique intitial S_{symb} au sous-ensemble S_{opt} = {type des phonèmes dans la syllabe, label du nucleus} optimal suivant le critère de maximisation du gain d'information sur la durée des syllabes. Le choix de ce sous-ensemble nous permet d'accéder à une estimation robuste des propriétés des durées syllabiques à partir d'une représentation compacte de leurs attributs. Les propriétés intrinsèques des syllabes sont alors estimées par arbre de décision sur la base de la description symbolique S_{opt}. Les propriétés intrin-

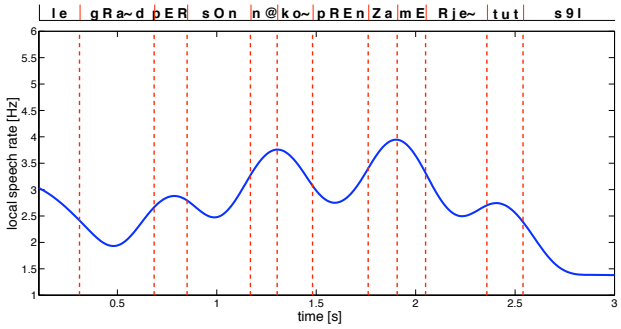

Fig. 1: Estimation du débit local sur la phrase : "*Les grandes personnes ne comprennent jamais rien toutes seules*". Les lignes verticales pointillées représentent les frontières des syllabes.

sèques d'une syllabe sont alors définies par la moyenne et la déviation standard ($\mu_{\mathcal{C}}$,$\sigma_{\mathcal{C}}$) de la classe $\mathcal{C} \in S_{opt}$ correspondante dans l'arbre de décision.

4. Modèle de débit local

4.1. Définition du débit local

Nous introduisons ici une formulation du débit local fondée sur l'intégration de la durée des syllabes dans une fenêtre d'observation à court-terme. Soit $\{d_n\}_{1\leqslant n\leqslant N}$ une séquences de durées de syllabes correspondant à la durée des N syllabes d'une phrase donnée. Soit $\{t_n\}_{1\leqslant n\leqslant N}$ la séquence de temps correspondante. Nous définissons au préalable la fonction T comme suit :

$$T(t) = \frac{1}{d_i}, t \in [t_i t_{i+1}]_{1\leqslant i\leqslant N}$$

Cette fonction transforme une séquence de durées discrètes en une fonction échelon par palliers définie à tout instant t. Soit w une fenêtre d'observation de durée w_T, et $t_{p-pause}$ et $t_{n-pause}$ respectivement les temps de fin de la précédente pause et le temps de début la pause suivante au temps t. Soit $t'_0 = max(t-\frac{w_T}{2}, t_{p-pause})$ and $t'_1 = min(t+\frac{w_T}{2}, t_{n-pause})$. Nous définissons alors la fonction d'estimation du débit local à tout instant t de la manière suivante (un exemple est présenté sur la figure 1) :

$$lsr(t) = \alpha \int_{t'_0}^{t'_1} w(\tau)T(\tau)d\tau$$

où α est un coefficient de normalisation

$$\alpha = \frac{1}{w_T} \frac{\int_{t-\frac{w_T}{2}}^{t+\frac{w_T}{2}} T(\tau)d\tau}{\int_{t-\frac{w_T}{2}}^{t+\frac{w_T}{2}} w(\tau)d\tau} \frac{\int_{t-\frac{w_T}{2}}^{t+\frac{w_T}{2}} w(\tau)d\tau}{\int_{t'_0}^{t'_1} w(\tau)d\tau}$$

Les hypothèses qui sous-tendent cette normalisation sont : i) le débit local est réinitialisé à chaque pause, et ii) lors d'une observation partielle, l'observation totale est inférée proportionnellement par rapport à l'observation partielle.

4.2. Débit local normalisé

Dans l'expression de l'estimation du débit local telle que définie dans la section précédente, les durées des syllabes sont toutes également considérées sans égards aux différences observées entre les propriétés intrinsèques des syllabes considérées. Pour remédier à ce point et intégrer l'estimation de ces propriétés intrinsèques, nous proposons de définir un débit local normalisé dans lequel chaque syllabe observée est normalisée préalablement par rapport à ses propriétés intrinsèques. Nous définissons ainsi le durée normalisée pour une syllabe appartenant à la classe symbolique $\mathcal{C}$ de la manière suivante :

$$d_{norm} = (d_{obs} - \mu_{\mathcal{C}})\frac{\sigma_{ref}}{\sigma_{\mathcal{C}}} + \mu_{ref}$$

Cette transformation normalise chaque durée de syllabe en fonction de leur propriétés intrinsèques ($\mu_{\mathcal{C}}$,$\sigma_{\mathcal{C}}$) sur une base normée (μ_{ref},σ_{ref}). Nous considérons ici (μ_{ref},σ_{ref}) respectivement la moyenne et la déviation standard de l'ensemble de la distribution des durées considérées. Nous nommons dès lors *débit local normalisé* le débit local estimé sur ces durées normalisées.

5. Evaluation du modèle

Notre étude a été menée dans le cadre du système IrcamCorpusTools [10]. Le corpus considéré représente environ 6 heures de parole lue par un locuteur mâle. Les frontières des phonèmes sont estimées par un modèle HMM d'étiquetage et alignement automatique de phonèmes [2]. Ce corpus représente un total de 78803 syllabes. Les syllabes proéminentes ont été prédite à partir d'un un modèle GMM [3]. Les paramètres de notre modèle de durée ont été estimé ensuite sur l'ensemble du corpus. Le débit local a été calculé avec une fenêtre de hanning de 650 ms [7]. Le corpus étudié a un débit moyen de 4.69 syllabes par seconde et une déviation relative de 26%. Pour tester la robustesse et la consistance de notre modèle, nous avons analysé la dispersion des durées résultant de notre modèle ainsi que l'influence des causes type de syllabe et phénomène de proéminence sur cette dispersion.

5.1. Distribution de la durée des syllabes

Nous avons estimé la distribution des durées pour chaque type de syllabe en fonction de i) les durées observées, ii) les durées normalisées par le débit local, iii) les durées normalisées par notre modèle (Table 1).

D'une manière générale, ces résultats mettent en lumière la réduction significative de la dispersion des durées, par comparaison aux durées observées, à laquelle conduit la méthode employée (durées normalisées par le débit local et par notre modèle) En menant une analyse plus fine des résultats obtenus, on notera que le modèle de normalisation par le débit local tend à une surestimation de la durée syllabique lorsque celles-ci sont intrinsèquement courtes (de V à VC) et, à l'inverse, à une sous-estimation de la durée des syllabes lorsque celles-ci sont intrinsèquement longues (de CVC à CVCCC). Un tel comportement peut s'expliquer par le fait que ce modèle ne prend pas en compte les propriétés intrinsèques des durées des syllabes : une syllabe intrinsèquement courte a tendance à augmenter localement le débit local et inversement. Les durées moyennes des syllabes estimées

	moyenne (ms.)			dév std (ms.)			%dsr		
syllabe	obs.	norm.	mod.	obs.	norm.	mod.	obs.	norm.	mod.
V	125	167	115	50	57	27	39.8	34.4	23.8
CV	203	224	188	63	48	38	30.8	21.3	20.5
VC	212	228	194	73	51	32	34.4	22.5	16.7
CCV	279	265	259	74	49	49	26.5	18.5	19.0
VCC	322	322	287	93	63	87	28.9	23.0	30.4
CVC	337	278	295	121	54	61	35.8	19.3	20.7
CCCV	346	294	324	86	51	50	24.8	17.4	15.7
CCVC	429	428	375	139	54	67	32.5	18.1	18.0
CVCC	432	309	399	99	57	77	22.9	18.6	19.5
VCCC	492	314	492	10	28	36	2.2	9.1	7.5
CCCVC	544	350	506	128	75	76	23.4	21.4	15.1
CVCCC	551	332	544	56	49	73	10.1	14.9	13.4
total	390	290	368	79	50	53	**23.8**	**18.5**	**16.6**

Tab. 1: Distribution de la durée des syllabes (moyenne, déviation standard et déviation standard relative) dans le cas des durées observées, des durées normalisées par le débit local et les durées normalisées par notre modèle en fonction du type de syllabe

par notre modèle sont en revanche systématiquement plus basses que la moyenne des durées observées : cette propriété s'explique par le fait que notre modèle élimine conjointement les durées intrinsèquement longues ainsi que les allongements important résultants du phénomène de proéminence.

5.2. Effets syllabiques et accentuels sur la distribution des durées

Pour étudier la consistance de notre modèle (la réduction de la dispersion des durées doit s'accompagner d'une réduction de l'influence du type de syllabe et du phénomène d'accentuation), nous avons testé les effets des facteurs *proéminences* et *type de syllabe* sur i) les durées observées, ii) le résiduel des durées normalisées par le débit local, et iii) le résiduel obtenu par notre modèle (Table 2). Tandis que les durées résiduelles obtenues par normalisation par le débit local ne montre pas de réduction significative de la variance expliquée par rapport aux durées observées, notre modèle conduit à une forte réduction de ces effets. Ce résultat met en exergue la consistance du modèle puisqu'il permet de prendre en compte de manière efficace ces effets et de réduire significativement leur influence sur les durées observées.

Effet	obs.	norm.	modèle
Proéminence	39.0	43.8	9.0
Type Syl.	46.1	33.9	5.9

Tab. 2: Pourcentage de variance expliquée en fonction des causes *proéminences* et *type de syllabe* pour les durées observées, les durées résiduelles obtenues après normalisation par le débit local, et les durées résiduelles obtenues après normalisation par notre modèle

6. Conclusion

Nous avons présenté un modèle de durée fondé sur une estimation des propriétés intrinsèques des syllabes et d'une modulation due aux variations locales de débit. Nous avons montré que notre modèle était robuste en tant qu'il réduit la dispersion des durées observées et consistant dans la mesure où cette réduction s'accompagne d'une réduction significative de l'influence des facteurs type de syllabe et accentuation sur la dispersion. Le modèle soulève de nouvelles questions quant aux propriétés intrinsèques des syllabes et à leur perception : ces propriétés devraient être étudiées sous l'angle de la perception et de la représentation phonologique de la durée des syllabes pour lesquelles nous ne connaissons pas d'étude à ce jour. L'influence de ces propriétés sur la perception de débit local doit être validée par des expériences psycho-acoustiques : ainsi, la perception de la proéminence dépend-elle uniquement d'un allogement par rapport à une référence attendue, ou bien alors certaines syllabes lourde (naturellement longues) n'attirent-elles pas de manière absolue la perception d'une proéminence ? Enfin, nous souhaitons intégrer notre modèle à un modèle prosodique global comme facteur de contrôle de l'allongement de la durée des syllabes.

Références

[1] P. Barbosa. *Caractérisation et génération automatique de la structuration rythmique du français*. PhD thesis, Institut de la Communication Parlée, Institut National Polytechnique, 1994.

[2] P. Lanchantin, A.C. Morris, X. Rodet, and C. Veaux. Automatic phoneme segmentation with relaxed textual constraints. In *accepted to The 6th edition of the Language Resources and Evaluation Conference*, Marrakech, 2008.

[3] N. Obin, X. Rodet, and A. Lacheret-Dujour. Prominence model : a probabilistic framework. In *submitted to The 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, U.S.A, 2008.

[4] S. Ohno and H. Fujisaki. A method for quantitative analysis of the local speech rate. In *Proc. of EUROSPEECH'95*, volume 1, pages 421–424, Madrid, Spain, 1995.

[5] S. Ohno, M. Fukumiya, and H. Fujisaki. Quantitative analysis of the local speech rate and its application to speech synthesis. In *Proc. of International Conference on Speech Language Processing (ICSLP '96)*, pages 2254–2257, Philadelphia, PA, 1996. 4.

[6] V. Pasdeloup. Figures et fond dans la scène prosodique : leur résistance face aux variations du débit de parole. In *Symposium Interface Discours-Prosodie - IDP05*, Aix-en-Provence, France, 2005.

[7] H. Pfitzinger. Two approaches to speech rate estimation. In *proc. of the 6th Australian Int. Conf. on Speech Science and Technology (SST'96)*, pages 421–426, Adelaide, Australia, 1996.

[8] H. Pfitzinger. Local speech rate as a combination of syllable and phone rate. In *Proc. of International Conference on Speech Language Processing (ICSLP'1998)*, volume 3, pages 1087–1090, Sydney, Australia, 1998.

[9] H. Pfitzinger. Reducing segmental duration variation by local speech rate normalization of large spoken language resources. In *proc. of the Third Int. Conf. on Language Resources and Evaluation*, volume 1, pages 313–320, Gran Canaria, Canary Islands, Spain, 2002.

[10] C. Veaux, B. Beller, D. Schwarz, and X. Rodet. Ircamcorpustools : an extensible plateform for speech corpora exploitation. In *accepted to The 6th edition of the Language Resources and Evaluation Conference*, Marrakech, 2008.

[11] B. Zellner. *Caractérisation et prédiction du débit de parole en français - une étude de cas*. PhD thesis, Université de Lausanne, 1998.

[12] B. Zellner. Fast and slow speech rate : a characterisation for french. In *proc. of the Vth International Conference on Spoken Language Processing (ICSLP'95)*, volume 7, pages 3159–3163, Syndney, Australia, 1998.