Evaluation automatique du débit de la parole sur des données multilingues spontanées

Jean-Luc Rouas[1], *Jérôme Farinas*[1], *François Pellegrino*[2]

[1]Institut de Recherche en Informatique de Toulouse
UMR 5505, Université Toulouse III, 118, route de Narbonne, 31062 Toulouse Cedex 4, France
[2]Laboratoire Dynamique du Langage
UMR 5596 CNRS Université Lyon 2, ISH, 14 av. Berthelot, 69363 Lyon cedex 7, France
Mél : rouas@irit.fr, jerome.farinas@irit.fr, Francois.Pellegrino@univ-lyon2.fr

ABSTRACT

An automatic method for speaking rate estimation is developed in this paper. It is based on an unsupervised segmentation and vowel detection algorithm and thus may be costlessly applied to any language. Validation is driven on a spontaneous speech subset of the OGI Multilingual Telephone Speech Corpus. Statistics related to the speaking rate both in term of phoneme per second and syllable per second are given. The correlation between the estimated and real speaking rates are evaluated on the corpus. In term of vowel-per-second rate, it is 0.86 on average among the 6 languages for wich a phonetic transcription is available (English, German, Hindi, Japanese, Mandarin and spanish).

1. INTRODUCTION

La plupart des systèmes de traitement automatique de la parole doivent tenir compte de la variabilité du débit de la parole et de ses conséquences sur les unités segmentales et l'organisation supra-segmentale de la parole. Les applications concernées vont de l'adaptation au locuteur pour les systèmes de reconnaissance de la parole à la modélisation du rythme ou de la prosodie dans une perspective d'identification de la langue.

Cependant, à cause de la notion complexe de débit de parole, de nombreux problèmes à la fois théoriques et pratiques se posent. Le débit peut être défini de nombreuses manières (différentes unités possibles, indépendance par rapport à la langue, etc.) et sa variabilité résulte d'interactions complexes (il dépend du locuteur, évenutellement de la langue, et peut varier pendant la durée du discours). On pourra consulter les articles de Ramus [9], [8] pour un discussion plus complète sur la notion de débit de parole dans une perspective multilingue et [3] pour une méthode combinant des estimateurs liés aux phonèmes et à la syllabe.

Dans une étude précédente [6], nous avons développé un modèle statistique d'unités rythmiques dans un but d'identification automatique des langues. Cet algorithme a permis d'obtenir des résultats intéressants sur un corpus de parole lue. Cependant, il semble évident qu'une normalisation des durées des unités en fonction du débit de parole est un préalable à l'application de cet algorithme à de la parole spontanée. La suite de cet article portera sur les mesures automatiques possibles du débit de la parole sur un corpus multilingue de parole spontanée et sur l'utilisation d'un algorithme de détection automatique des voyelles pour estimer ce débit. Les différentes méthodes de mesure du débit sont discutées dans la section 2. Le corpus est présenté dans la section 3, accompagné de statistiques relatives aux débits de parole calculés à partir des transcriptions phonétiques. L'évaluation des estimateurs de débits de parole proposés est menée dans la section 4, et les résultats sont discutés dans la dernière section.

2. DÉFINITIONS ET MÉTHODES D'ESTIMATION

2.1. Définition(s) du débit de parole

La notion de débit de parole (DP) est liée à la notion de rythme et génère des problèmes de définition similaires, puisqu'ils font tous les deux intervenir le comptage d'unités par seconde. Le choix de l'unité demeure crucial : syllabes et phonèmes en constituent les meilleurs candidats. Pfitzinger a montré [7] que le débit de parole perçu est plus corrélé au débit syllabique qu'au débit phonétique (respectivement $R = 0.81$ contre $R = 0.73$). Dans une perspective de typologie rythmique ou d'identification des langues, il semble évident que les DP calculés en terme de phonèmes par seconde ou de syllabes par seconde apportent des informations complémentaires sur la structure rythmique et l'organisation phonotactique des langues. Par ailleurs, des expériences ont montré que les DP calculés en terme de syllabes ou de phonèmes sont corrélés (pour l'allemand : $R = 0.6$ [7]), tout du moins pour un débit de parole normal. Le niveau de corrélation est probablement plus élevé pour les langues ayant une structure syllabique simple en CV que pour les langues qui autorisent une plus grande complexité syllabique en termes de nombre de segments consonantiques consécutifs. À des débits de parole élevés, des stratégies de dépendance par rapport à la langue peuvent aussi intervenir (voir [1] pour une étude de l'impact du débit sur l'organisation temporelle de la parole en terme de quantité de voyelles et de variance de durées de segments de consonnes).

Par conséquent, les débits observés résultent des intéractions entre les facteurs dépendants des locuteurs et/ou dépendants des langues. De même que Ramus [8], nous considérons que l'étude de grands corpus va conduire à une meilleure compréhension de la contribution respective de chaque facteur. Nous proposons dans cet article d'étudier les DP en termes de phonèmes et de syllabes par seconde dans une perspective multilingue et d'évaluer un algorithme de segmentation automatique et de détection automatique de voyelles comme estimateurs des DP.

2.2. *Estimation du débit de parole*

L'algorithme de segmentation et de détection de voyelles utilisé est décrit dans [5]. Il est basé sur une segmentation statistique combiné à une analyse spectrale du signal de parole. Il est appliqué de manière indépendante de la langue et du locuteur, sans aucune phase d'adaptation. La segmentation est intrinsèquement infra-phonémique puisqu'elle est basée sur une détection de ruptures et que les parties transitoires des phonèmes sont dissociées des parties stables. Nous testerons cependant si le nombre de segments par seconde se révèle corrélé de manière forte au nombre de phonèmes par seconde. La détection des voyelles fournit quant à elle un estimateur a priori fiable du nombre de syllabes par seconde. Les erreurs de détection les plus fréquentes sont des erreurs d'omission de voyelles de faible énergie ou dévoisées et des fausses détections de liquides.

3. ANALYSE DES DONNÉES

3.1. *Corpus*

Les expériences sont menées sur un sous ensemble du corpus "OGI Multilingual Telephone Speech Corpus" [4] pour lequel des transcriptions phonétiques manuelles sont fournies. Le tableau 1 donne les caractéristiques de cette base de données. Pour chaque locuteur, un enregistrement d'environ 40 secondes est étiqueté phonétiquement et qualifié de "spontané" ou "lu". Cette distinction n'a pas été faite pour l'hindi. Pour les autres langues, la plupart des enregistrements sont considérés "spontanés" et la taille du corpus varie de 64 fichiers pour le japonais à 144 pour l'anglais.

TAB. 1: Description du corpus, nombre total de locuteurs et nombre de locuteurs considérés comme "spontanés", durée moyenne des fichiers (et écart-type)

Langue	Nombre de locuteurs (spontanés)	Durée moyenne par locuteur (écart-type)
Anglais	144 (111)	47,1 (3,4)
Allemand	98 (89)	42,7 (8,4)
Hindi	68 (n.c.)	46,5 (5,9)
Japonais	64 (55)	46,1 (5,1)
Mandarin	69 (69)	39,9 (10,8)
Espagnol	108 (106)	45,6 (5,6)

3.2. *Conventions et calcul du débit*

Les conventions d'étiquetage développées au CSLU [2] sont basées sur des règles indépendantes des langues et adaptées à chaque langue suivant la liste des phonèmes. Les frontières phonémiques sont déterminées avec une précision de l'ordre d'une milliseconde. Par convention, les diphtongues sont considérées comme une seule voyelle dans le calcul du débit. Puisque les évenements ne correspondant pas à de la parole (pauses silencieuses, respirations, etc.) sont également étiquetés, il est possible de les écarter pour le calcul du débit.

Soit u la phrase sur laquelle on calcule le débit. Soit $N_v(u)$ le nombre de segments étiquetés "voyelle" dans cette phrase et $D(u)$ la durée de la phrase. Le débit moyen en terme de syllabes par seconde mesuré sur la phrase ($DP(u)$) est alors défini par :

$$DP(u) = \frac{N_v(u)}{D(u)} \quad (1)$$

En considérant la durée totale des évenements ne correspondant pas à de la parole $D_{np}(u)$, le débit moyen non biaisé est alors défini par :

$$DP_{np}(u) = \frac{N_v(u)}{(D(u) - D_{np}(u))} \quad (2)$$

Cette mesure globale du débit est évidemment limitée puisqu'elle sous-estime l'impact des variations locales de débit qui ont lieu pendant la production de parole. Elle doit cependant permettre d'évaluer l'impact de la variabilité inter-locuteur et inter-langue sur le DP.

L'algorithme de détection automatique des voyelles fournit une estimation du nombre de voyelles présentes dans le signal. Il permet ainsi d'estimer le débit $\widehat{DP}(u)$:

$$\widehat{DP}(u) = \frac{\widehat{N_v}(u)}{D(u)} \quad (3)$$

Le débit de parole peut également être calculé en terme de nombre de phonèmes par seconde. On remplacera alors N_v dans les formules précédentes par N_{phon}.

3.3. *Comparaisons inter-langues*

Débit syllabique Le tableau 2 donne les débits moyens (DP et DP_{np}) calculés pour chaque langue du corpus en fonction du nombre de voyelles par seconde. La première constatation est que, même en écartant les pauses, les différences inter-langues sont significatives (ANOVA F(5) = 15 ; p<.0001). Le débit d'information (en terme de syllabes par seconde) est donc dépendant de la langue ce qui confirme indirectement que le débit d'information global résulte non seulement du niveau phonético-phonologique mais également des niveaux morpho-syntaxiques. Si l'on écarte les pauses, le plus faible débit est obtenu pour le mandarin (4,61) tandis que le plus important est obtenu pour l'espagnol (5,71). L'ordre est quasi-identique que l'on considère ou non les pauses. L'anglais et l'allemand montrent des débits DP_{np} très proches qui peuvent être liés au fait que ces deux langues sont très proches rythmiquement.

TAB. 2: Moyenne et écart-type du DP syllabique (étiquetage manuel)

Langue	DP (NbVoy/s) (avec pauses)	DP_{np} (NbVoy/s) (sans pauses)
Anglais	3,80 ±0,11	4,71 ±0,09
Allemand	3,60 ±0,11	4,68 ±0,11
Hindi	3,67 ±0,16	5,40 ±0,14
Japonais	3,89 ±0,20	5,21 ±0,15
Mandarin	3,04 ±0,18	4,61 ±0,16
Espagnol	4,24 ±0,15	5,71 ±0,13

Débit phonémique Le tableau 3 donne les débits moyens (DP et DP_{np}) calculés pour chaque langue du corpus en fonction du nombre de phonèmes par seconde. La mise en correspondance avec les résultats du tableau 2 donne des indices intéressants sur la structure syllabique des langues présentées. Par exemple, l'allemand présente le débit syllabique le plus faible et le débit phonémique le plus important, révélant ainsi une structure syllabique complexe.

TAB. 3: Moyenne et écart-type du DP phonémique (étiquetage manuel)

Langue	DP (NbPhon/s) (avec pauses)	DP_{np} (NbPhon/s) (sans pauses)
Anglais	11,61 ±0,30	13,73 ±0,25
Allemand	11,44 ±0,33	14,20 ±0,31
Hindi	9,90 ±0,42	13,54 ±0,37
Japonais	11,47 ±0,51	14,63 ±0,38
Mandarin	8,82 ±0,52	12,45 ±0,48
Espagnol	10,96 ±0,34	13,95 ±0,30

Corrélation entre les deux types de débits proposés Le tableau 4 montre les corrélations entre les deux estimateurs de débit proposés. Ces résultats montrent que ces deux mesures de débit sont très corrélées, de manière plus importante même que la corrélation indiquée pour l'allemand dans [7]. La pente des régressions linéaires permet d'estimer que la longueur moyenne des syllabes est de 2,8 phonèmes (maximum 3,1 pour l'allemand et minimum 2,4 pour le japonais).

TAB. 4: Corrélation entre les deux types de débits proposés (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,94	0,89	$DP_{syl} = 0{,}35 DP_{pho}$ - 0,23
Allemand	0,93	0,86	$DP_{syl} = 0{,}32 DP_{pho}$ - 0,04
Hindi	0,96	0,92	$DP_{syl} = 0{,}37 DP_{pho}$ + 0,03
Japonais	0,98	0,95	$DP_{syl} = 0{,}38 DP_{pho}$ - 0,46
Mandarin	0,96	0,91	$DP_{syl} = 0{,}34 DP_{pho}$ + 0,07
Espagnol	0,96	0,91	$DP_{syl} = 0{,}41 DP_{pho}$ - 0,21

4. EVALUATION DES ALGORITHMES COMME ESTIMATEURS DU DÉBIT

4.1. Estimation du débit syllabique par le nombre de voyelles détectées par seconde

Les résultats sont donnés dans le tableau 5 à la fois en termes de coefficients de corrélation (R) et de régression linéaire, et sont illustrés par la figure 1. Toutes les corrélations sont très significatives ($p < .001$). La plus mauvaise corrélation est obtenue pour l'espagnol, mais elle demeure élevée ($R = 0,79$).En moyenne, la corrélation obtenue est de $R = 0,86$ ce qui indique que le détecteur de voyelles est un bon à très bon estimateur du débit syllabique. La qualité du détecteur de voyelles est également confirmée par les valeurs des pentes, proches de l'unité (en moyenne 0,89).

TAB. 5: Corrélation entre débits syllabiques réels et estimés par la détection des voyelles (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,84	0,70	$DP_{syl} = 0{,}90 \widehat{DP_{syl}} + 0{,}41$
Allemand	0,81	0,65	$DP_{syl} = 0{,}75 \widehat{DP_{syl}} + 0{,}85$
Hindi	0,89	0,80	$DP_{syl} = 0{,}91 \widehat{DP_{syl}} + 0{,}58$
Japonais	0,92	0,85	$DP_{syl} = 0{,}97 \widehat{DP_{syl}} + 0{,}44$
Mandarin	0,90	0,81	$DP_{syl} = 0{,}94 \widehat{DP_{syl}} + 0{,}11$
Espagnol	0,79	0,62	$DP_{syl} = 0{,}88 \widehat{DP_{syl}} + 1{,}05$

4.2. Estimation du débit phonémique par le nombre de segments détectés par seconde

L'utilisation de l'algorithme de segmentation comme estimateur du débit phonémique montre des résultats plus contrastés (tableau 6 et figure 2). La pente moyenne (0,55) confirme le caractère infra-phonémique de la segmentation. Les coefficients de corrélation s'étendent de 0,51 pour l'allemand à 0,86 pour l'hindi. Il est difficile d'estimer la part de cette variation iée à la structure syllabique des langues et celle liée à un éventuel biais de la segmentation en fonction des langues.

TAB. 6: Corrélation entre débits phonétiques réels et estimés par la segmentation (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,74	0,55	$DP_{pho} = 0{,}52 \widehat{DP_{pho}} + 3{,}91$
Allemand	0,51	0,27	$DP_{pho} = 0{,}37 \widehat{DP_{pho}} + 6{,}16$
Hindi	0,86	0,74	$DP_{pho} = 0{,}62 \widehat{DP_{pho}} + 1{,}71$
Japonais	0,74	0,55	$DP_{pho} = 0{,}51 \widehat{DP_{pho}} + 4{,}80$
Mandarin	0,72	0,51	$DP_{pho} = 0{,}64 \widehat{DP_{pho}} + 1{,}47$
Espagnol	0,74	0,55	$DP_{pho} = 0{,}56 \widehat{DP_{pho}} + 3{,}48$

4.3. Corrélation entre les estimateurs de débits syllabiques et phonémiques

TAB. 7: Corrélation entre les estimateurs de débits syllabiques et phonémiques (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,60	0,35	$\widehat{DP_{syl}} = 0{,}14 \widehat{DP_{pho}} + 1{,}66$
Allemand	0,55	0,31	$\widehat{DP_{syl}} = 0{,}15 \widehat{DP_{pho}} + 1{,}57$
Hindi	0,82	0,66	$\widehat{DP_{syl}} = 0{,}22 \widehat{DP_{pho}} + 0{,}47$
Japonais	0,79	0,62	$\widehat{DP_{syl}} = 0{,}20 \widehat{DP_{pho}} + 0{,}94$
Mandarin	0,72	0,52	$\widehat{DP_{syl}} = 0{,}22 \widehat{DP_{pho}} + 0{,}62$
Espagnol	0,71	0,51	$\widehat{DP_{syl}} = 0{,}20 \widehat{DP_{pho}} + 0{,}89$

De faibles valeurs de corrélation sont observées pour certaines langues. L'algorithme de segmentation est basé sur la détection de ruptures dans le signal. Les phonèmes complexes ont tendance à être sursegmentés, ce qui augmente le débit phonémique et détériore la mesure de corrélation.

5. CONCLUSION ET PERSPECTIVES

Les statistiques présentées dans la section 3 montrent que le débit de parole est dépendant non seulement du locuteur mais également de la langue. elles montrent également que, même pour des langues accentuelles comme l'anglais, pour lesquelles une grande variabilité de la complexité syllabique est avérée ([9]), les débits de parole phonémique et syllabique sont extrêmement corrélés ($R = 0,94$) tout comme dans des langues ayant des structures syllabiques plus simples (espagnol ou mandarin par exemple).

Les résultats présentés dans la Section 4 montrent quant à eux que le détecteur de voyelles fournit un bon estimateur du débit syllabique (en moyenne $R = 0,86$) et que par contre, le comportement du nombre de segments automatiques comme estimateur du débit phonémique est dépendant de la langue, la corrélation étant plutôt bonne pour

l'hindi et faible pour l'allemand. On peut penser que ces différences inter-langues sont liées aux inventaires phonémiques des langues étudiées.

Cependant plusieurs autres paramètres peuvent influencer l'estimation du débit. On peut citer par exemple la variation du débit au cours du temps, qui peut être liée à plusieurs aspects linguistiques (allongement final, etc.), ou encore la qualité de la prise en compte des pauses silencieuses et remplies dans le signal.

Le couplage des algorithmes d'estimation de débits proposés ici avec un détecteur d'activité vocale et un détecteur de pauses remplies est l'extension la plus évidente du travail proposé.

6. REMERCIEMENTS

Cette recherche est soutenue par le programme "Société de l'Information" du CNRS (projet RAIVES), le programme EMERGENCE de la Région Rhône-Alpes et le Ministère de la Recherche (programme *ACI Jeunes Chercheurs*).

RÉFÉRENCES

[1] Volker Dellwo and Petra Wagner. Relations between language rhythm and speech rate. In *XVth International Congress of Phonetic Sciences*, Barcelone, Espagne, August 2003.

[2] Terry Lander and James L. Hieronymus. The cslu labeling guide. Technical report, Center for Spoken Language Understanding - Oregon Graduate Institute, 1997.

[3] Nelson Morgan and Eric Fosler-Lussier. Combining multiple estimators of speaking rate. In *IEEE 23rd International Conference on Acoustics Speech and Signal Processing*, volume 2, pages 729–732, Seattle, WA, USA, May 1998.

[4] Yeshwant Kumar Muthusamy, Ronald A. Cole, and B. T. Oshika. The ogi multilanguage telephone speech corpus. In *International Conference on Speech and Language Processing*, volume 2, pages 895–898, October 1992.

[5] François Pellegrino and Régine André-Obrecht. Automatic Language Identification : an alternative approach to phonetic modeling. In *Signal Processing*, volume 80, pages 1231–1244. Elsevier Science, jul 2000.

[6] François Pellegrino, Jean-Hugues Chauchat, Ricco Rakotomalala, and Jérôme Farinas. Can automatically extracted rhythmic units discriminate among languages ? In *International Conference on Speech Prosody*, pages 563–566, Aix-en-provence, France, April 2002.

[7] Hartmut R. Pfitzinger. Local speaking rate as a combinaison of syllable and phone rate. In *5th International Conference on Spoken Language Processing*, volume 3, pages 1087–1090, December 1998.

[8] Franck Ramus. Acoustic correlates of linguistic rhythm : Perspectives. In *International Conference on Speech Prosody*, pages 115–120, Aix-en-Provence, France, April 2002.

[9] Franck Ramus, Marina Nespor, and Jacques Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3) :265–292, 1999.

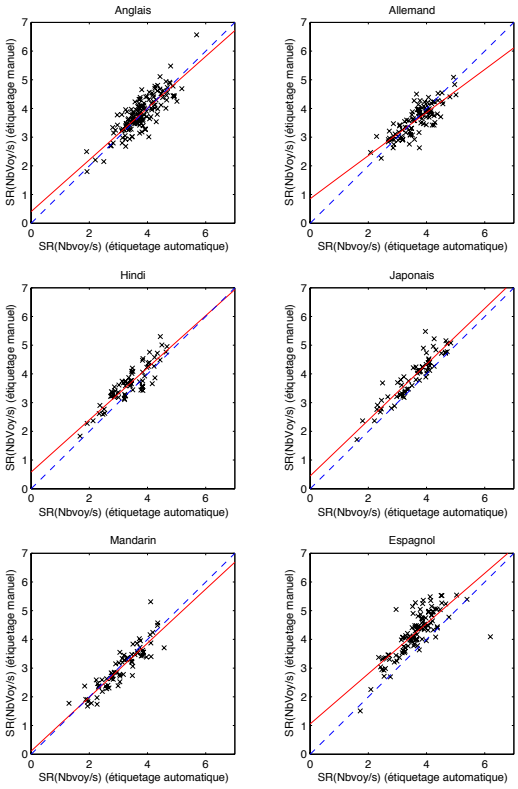

FIG. 1: DP syllabique estimé par le nombre de voyelles par seconde

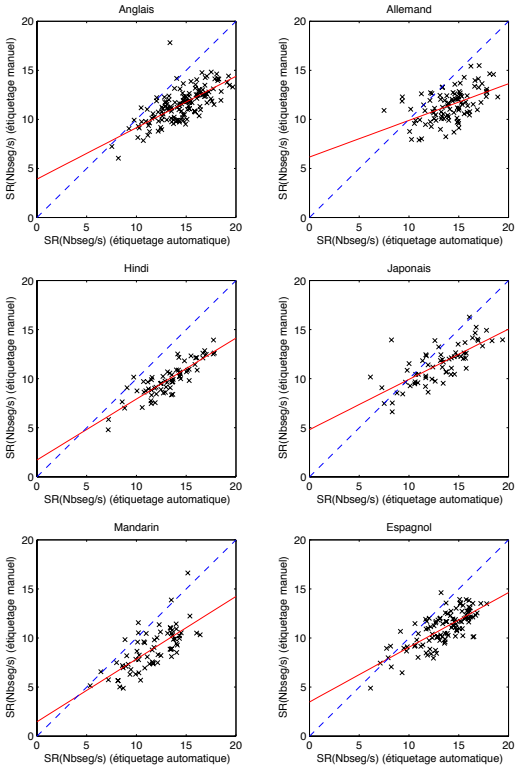

FIG. 2: DP phonémique estimé par le nombre de segments par seconde