Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech

François Pellegrino[1], J. Farinas[2] & J.-L. Rouas[2]

[1]Laboratoire Dynamique Du Langage, UMR 5596 CNRS – Univ. Lumière Lyon 2, France
[2]Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS – Univ. Toulouse 3, France
Francois.Pellegrino@univ-lyon2.fr; {jfarinas; jean-luc.rouas}@irit.fr

Abstract

An automatic estimation of speaking rate is developed in this paper. It is based on an unsupervised 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. The correlation coefficient between the estimated and real speaking rates (evaluated in term of vowel-per-second rates) is 0.84 on average among the 6 languages for which a phonetic transcription is available (English, German, Hindi, Japanese, Mandarin and Spanish).

1. Introduction

Most of automatic speech processing systems have to cope with the variability of Speaking Rate (hereunder SR) and its consequences both on segmental units and supra-segmental organization of speech. Applications range from speaker adaptation of automatic speech recognition systems to automatic modeling of rhythm or prosody in a typological or language identification perspective.

Obviously, due to the intricate notion of speaking rate, many theoretical and practical problems arise. To sum up, let say that SR may be defined in several ways (which recurrent unit should be taken into account? Is it language independent? etc) and that its variability results from complex interactions (it depends on speaker, maybe language, and it may vary during the discourse). See Ramus [9, 10] for a more complete discussion on SR in a cross-linguistic view and Morgan & Fosler-Lussier [4] for a method combining phone level and syllable level estimators.

In a previous work [7], we developed a rhythmic unit model for language identification. This algorithm reached pretty good results on a read speech corpus. However, it seemed obvious that speaking rate normalization would have been the bottleneck to overpass before considering spontaneous speech. The following of this paper focuses on SR measurement on a multilingual spontaneous speech corpus and on using a vowel detection algorithm as a predictor of the SR. These methods are discussed in Section 2. Section 3 presents the corpus and statistics related to SR. The results are given in Section 4 while the final section summarizes the findings and discusses perspectives.

2. Methods

2.1. Defining Speaking rate

The notion of SR is linked to the notion of rhythm and generates the same kind of problems, since they both involve the counting of some pattern per second. Some argue that syllable is the right unit while others oppose that the universal relevancy of syllable is not assessed and that phonemes may be better candidates. Still, Pfitzinger showed in [8] that syllable rate is more correlated to perceptual speaking rate than phone rate (r=0.81 vs r=0.73).

Selecting which pattern is the relevant one is beyond the range of this paper and we may consider that SR calculated in terms of syllable or phoneme rates are correlated ([8] for german: r=0.6), at least in normal rate speech. The level of the correlation is probably higher for languages with simple CV syllable structure than for languages allowing more consonantal cluster complexity. At fast speaking rates, language dependent strategies may also interact (see [9] for a study of the impact of the speech rate on the temporal organization of speech in term of vowel quantity and of variance of consonantal cluster durations).

As a consequence, the observed SR results from interactions between speaker dependent and language dependent factors. Following Ramus [9], we consider that studying large corpora will lead to a better comprehension of the respective contribution of each factor. At this moment we propose to define the SR as the **number of vowels per second,** which is a good estimation of the number of syllables per second. This way, vowel detection may be done in a language independent manner (see below) and provide an estimator of it, whereas syllable detection may involve language dependent syllabation strategies.

2.2. An Algorithm for Speaking rate Evaluation

The vowel detection algorithm has been already described in [6]. It is based on a statistical segmentation combined with a spectral analysis of the signal. It is applied in a language and speaker independent way without any manual adaptation phase. Classical errors are omissions of low energy or devoiced vowels and insertions of R-like sounds.

3. Experiments

3.1. Corpus

Experiments are performed using a subset of the OGI Multilingual Telephone Speech Corpus [5] for which a hand-made phonetic transcription is provided. Table 1 gives the characteristics of the database. For each speaker, one excerpt lasting about 40 seconds is phonetically labeled and tagged as ‘spontaneous’ or ‘read’. This tagging is missing for Hindi. For the other languages, most of the excerpts are considered “spontaneous” and the size of the corpus ranges from 64 excerpts for Japanese to 144 for English.

Table 1: *Corpus Description. Number of speakers is given with the number of speakers considered as "spontaneous". Statistics about the excerpt duration (mean and standard deviation) are also given.*

Language	Number of speakers (spontaneous speech)	Mean duration per speaker (std)
English (EN)	144 (111)	47.1 (3.4)
German (GE)	98 (89)	42.7 (8.4)
Hindi (HI)	68 (n.a.)	46.5 (6.0)
Japanese (JA)	64 (55)	46.1 (5.1)
Mandarin (MA)	69 (69)	39.9 (10.7)
Spanish (SP)	108 (106)	45.6 (5.6)

3.2. Conventions and Speaking Rate Calculation

The labeling conventions developed at CSLU [3] rely on language independent rules adapted to each target language for the phoneme list. Phonemic boundaries are set with a precision of one millisecond. By convention, diphthongs are considered as one vowel in the SR calculation.

Since non speech events are also labeled on these data, it is possible to take them, and especially silent pauses, breaths, etc. into account for the computation of the actual SR.

Let *u* be the utterance for which the SR is computed. Let $N_V(u)$ be the number of vowel segments labeled along this utterance and $D(u)$, the duration of the utterance. The mean Speaking Rate along the utterance $SR(u)$ is thus defined as:

$$SR(u) = \frac{N_V(u)}{D(u)} \quad (1)$$

Considering $D_{ns}(u)$, the total *non speech* duration in *u*, the mean SR unbiased by the pauses is

$$SR_{ns}(u) = \frac{N_V(u)}{\left(D(u) - D_{ns}(u)\right)} \quad (2)$$

This global measurement of the SR is obviously limited since it underestimates the impact of local SR variation during the speech production (see Section 4.2).

The vowel detection algorithm provides an estimation of the actual number of vowel present in the waveform. It thus provides an estimate of $SR(u)$:

$$\widehat{SR}(u) = \frac{\widehat{N}_V(u)}{D(u)} \quad (3)$$

3.3. Cross-linguistic comparison

Table 2 displays the mean *SR* and SR_{ns} computed for each language of the database. The lowest mean SR is reached for Mandarin (3.0) while the fastest rate is Japanese one (4.9 but see Section 4.1).

Table 2: *Mean and standard deviation values computed in term of* ***hand-labeled*** *vowels per second.*

Language	Mean SR with pauses (± CI)	Mean SR no pauses (± CI)
EN	3.8 (± 0.11)	5.0 (± 0.09)
GE	3.6 (± 0.11)	5.0 (± 0.12)
HI	3.7 (± 0.16)	5.7 (± 0.14)
JA	4.9 (± 0.25)	7.0 (± 0.19)
MA	3.0 (± 0.19)	4.7 (± 0.16)
SP	4.2 (± 0.14)	6.0 (± 0.13)

This rating persists whether pauses are discarded or not. English and German exhibit very similar SR_{ns} rates that may be linked to their nearby rhythmic structure.

The significant differences (ANOVA performed with SPSS, $F(5)=129$, $p<.0001$) observed among languages confirm that SR is also linked to the language rhythmic structure and not only speaker characteristics. For example, the canonical Japanese syllable structure is CV, while English or German allow complex CCC clusters on attack.

4. Results

4.1. Speaking rate Estimation

Results are given in Table 3, both in terms of correlation coefficients and of linear regression (computed with SPSS) and illustrated on Figure 1 (see last page). All correlations are highly significant ($p<.0001$). The worst correlation is reached with German, but it's still pretty high. It may be explained by the slope revealed by the linear regression (0.63) that means that a significant amount of false alarms occur. For Japanese, on the opposite, the number of vowels seems to be underestimated (slope equals 1.14). This value is due to a bias introduced by the hand-labeling procedure during which phonemic long vowels are labeled as two successive segments. Merging these two segments in one unique vowel leads to mean SR of 3.9 (SR_{ns} = 5.4), R coefficient of 0.89 and a linear regression equation more conventional: y = 0.92x + 0.78.

Table 3*: Correlation and linear regression between the estimate and real SRs.*

Language	R	R^2	Linear regression
EN	0.82	0.67	y = 0.89x + 0.65
GE	0.73	0.54	y = 0.63x + 1.43
HI	0.91	0.83	y = 0.94x + 0.59
JA	0.88	0.78	y = 1.14x + 1.08
MA	0.88	0.77	y = 0.98x + 0.20
SP	0.84	0.71	y = 1.00x + 0.74
Mean	**0.84**	**0.71**	**-**

4.2. Discussion

The automatic vowel detection algorithm provides a pretty good way to estimate the mean SR. However, several parameters may influence the precision of the prediction.

First, it appears that considering a SR averaged along the excerpt may be problematic, especially when the speaking rate is widely varying along the utterance and obviously because of the presence of pauses.

Figure 2 displays an example of this effect where the curve of the number of vowels as a function of time (both in term of hand-labeled and detected vowels) is non linear. On this picture, two major pauses (from about 26s to 31s and from 32.5s to 43s) alter dramatically the estimation of the SR.

Other effects are more difficult to predict. For instance, Figure 3 displays the same kind of curves as in Figure 2 for two English speakers (call4 and call93) with very different SR (resp. 3.8 and 6.9). In each case, the red line corresponds with the detected vowels and the black line with the actual vowels.

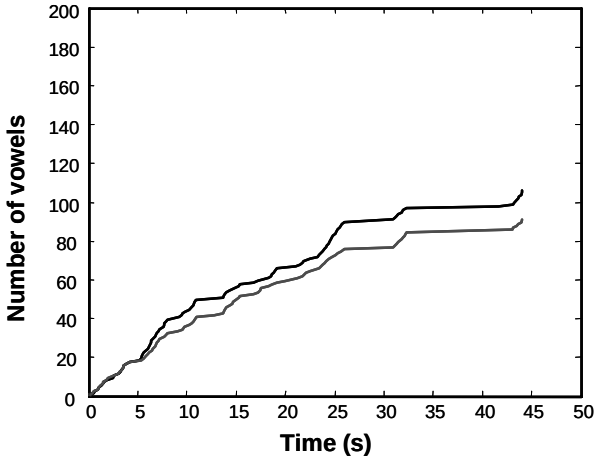

Figure 2*: Example of the evolution of the number of vowels per second (black line: hand labeled; red line: detected) for a German speaker (call145).*

Call4 is an example of correct vowel detection; even if a few vowels are omitted the two curves tend to be almost parallel, and on average, the mean SR will be correctly estimated. The result is quite different with call93: The detection algorithm regularly missed vowels and consequently the estimated curve drifted away from the theoretical one resulting in an underestimation of the SR for the utterance.

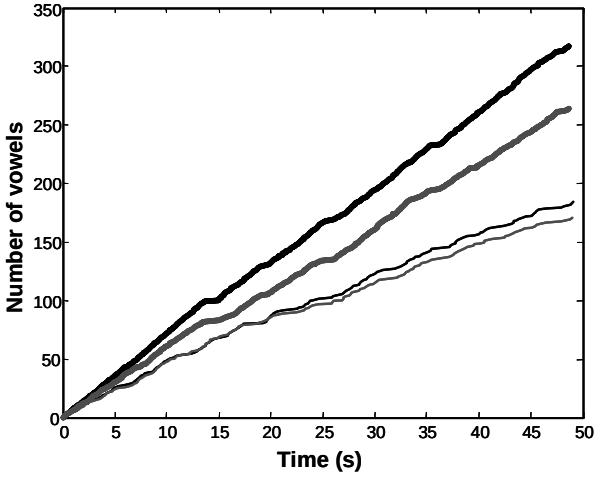

Figure 3*: Comparison of the automatic detection of vowel (red lines) with the hand-labeled phonetic transcription (black lines). Two speakers, a very fast one (bold lines, SR = 6.9) and a more standard one (thin lines, SR = 3.8) are represented.*

Despite these limitations, these first results indicate that the automatic vowel detection may be an efficient way to estimate the SR. Furthermore, combining this approach with an efficient Speech Activity Detector may lead to a correct estimation of the real duration of the speech excerpt and thus to a better SR estimation.

5. Conclusions

As a conclusion, the speaking rate detector performs well on all studied languages (on average, R = 0.84). Correlation is quite good, especially for Hindi. It means that this approach may be useful to adapt a system to a specific speaking rate and to accomplish a basic normalization for prosodic modeling purposes. However, it is still necessary to evaluate the specific impact of the SR on either vowels or consonants (e.g. see [1]).

Going further with the estimation of the *local* Speaking Rate in term of number of vowels per *effective* second of speech implies to use an efficient Speech Activity Detector and last but not least, to detect filled pauses as well. For this purpose, taking advantage from the statistical segmentation we already use is planned.

6. Acknowledgements

This research is supported by the French *Ministère de la Recherche* (program ACI "*Jeunes Chercheurs*").

7. References

[1] Crystal, T. H.; House, A.S; 1990. Articulation rate and the duration of syllables and stress groups in connected speech, *JASA*, 88(1), 101-112

[2] Dellwo, V; Wagner, P; 2003. Relations between language rhythm and speech rate. *ICPhS 03*, Barcelona, Spain

[3] Lander, T.; Hieronymus, J. L.;,1997, "The CSLU labeling guide", Technical Report, Center for Spoken Language Understanding, Oregon Graduate Institute

[4] Morgan, N; Fosler-Lussier, E., 1998. Combining Multiple Estimators of Speaking Rate. *IEEE ICASSP-98*, Seattle, 729-732.

[5] Muthusamy Y. K.; Cole, R. A.; Oshika, B. T., 1992, "The OGI multilanguage telephone speech corpus", Proc. of ICSLP, p. 895-898

[6] Pellegrino F.; André-Obrecht R., 2000, "Automatic Language identification: an alternative approach to phonetic modeling", In Signal Processing, 80, p. 1231-1244, Elsevier Science

[7] Pellegrino, F.; Chauchat, J.-H.; Rakotomalala, R.; Farinas, J.,2002, "Can automatically extracted rhythmic units discriminate among languages?" In *Proc. of International Conference on Speech Prosody*, p. 563-566.

[8] Pfitzinger, H., 1998. Local speaking rate as a combination of syllable and phone rate. In *Proceeding of ICSLP 1998*.

[9] Ramus, F., 2002, "Acoustic correlates of linguistic rhythm: Perspectives", In *Proc. of International Conference on Speech Prosody*

[10] Ramus, F.; Nespor, M.; Mehler, J., 1999, "Correlates of linguistic rhythm in the speech signal", Cognition, 73(3), p. 265-292

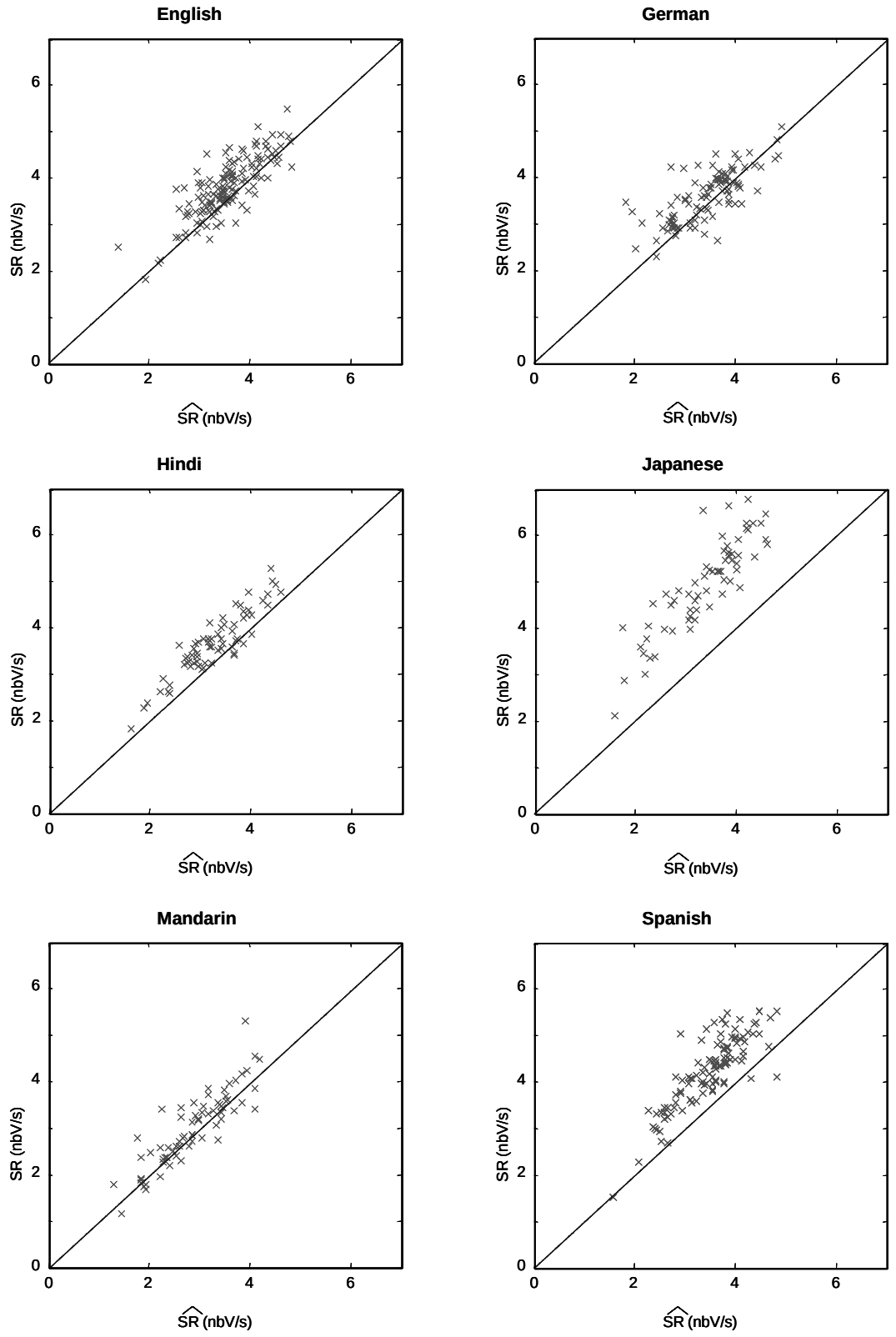

Figure 1: *Correlation between the estimated SR (X-axis) and actual SR (Y-axis) for the six languages. Each cross corresponds with a different speaker and the plain line is the* $SR = \widehat{SR}$ *line.*