

Speaking Rate Effects on Discourse Prosody in Standard Chinese

Aijun LI

Institute of Linguistics,
Chinese Academy of Social Sciences
liaj@cass.org.cn

Yiqing ZU

MOTOROLA China Research Center
Yiqing.Zu@motorola.com

Abstract

What is the prosodic mechanism of faster or slower discourse speech? This paper focuses on observing the effects of speech rate on discourse prosody of Standard Chinese speech with fast, normal and slow speech rates. The investigation in discourse prosody structure demonstrates that the speaking rate effects on discourse prosody are nonlinear and need careful manipulations.

Keywords: speech rate, articulation rate, discourse prosody

1. Introduction

Speaking rate is one of the prosodic features signalling the naturalness for text-to-speech (TTS) synthesis. The measure of speaking rate is distinguished into speech rate (SR) when pauses are included, and into articulation rate (AR) when pauses are not included [6,9]. Speaking rate will affect both segmental and suprasegmental features of speech. This paper focuses only on suprasegmental aspect where there are many contributions relating to this topic. [1,4-19]

Florien and Monique analyzed the speaking rate strategy on discourse level. They found that the average syllable duration of the first run of a paragraph was longer than the overall mean value for per speaker in more than 60 % of the cases.

It has been found that French speakers use a number of strategies for consciously achieving an increase in speech rate [7]. These include a reduction in the number of phrases and a demotion of major phrases to minor phrases, achieved by deleting phrase boundaries or reducing their strength. This prosodic restructuring is reflected in a reduction in the number and mean duration of pauses. Although considerable inter-speaker variability was observed, it was shown that fast speech was largely characterized by a reduction in overall pitch range and in the amplitude of individual rising and falling pitch movements as well as a simplification of the tonal structure, achieved by the non-realization of underlying tones.

Jürgen Trouvain and Martine Grice found a considerable effect of tempo in the pausing structure either in the number of pauses, or in the mean pause duration.[10] And their results show that articulation and speaking rate cannot be used as sole indicators of an achieved rate change. And they have also observed that slowing down strategies are not always the converse of speeding up strategies, and that individual speakers differ considerably in this respect.

Zellner found that slowing down was obtained first by lengthening the duration of segments, second by producing additional syllables, and third by producing pauses [17-18].

Most studies are relevant to stress timed languages such as English.[19] The present study is trying to make a contrast analysis on the Chinese discourse among three rates, so as to disclose some strategies on articulation rate adjustment for discourse speech synthesis.

2. Chinese prosodic structure on discourse and the material used

Tseng describes the hierarchical framework of the discourse

prosody for fluent speech [2,3]. A similar prosody framework is used in this study as shown in figure 1: from small to large units are syllable (SYLL), prosodic word (PW), minor phrase (MIP) / Compound PW, Intonation Phrase (IP), Prosodic Group (PG), Turn (TN), and Discourse (DI).

The speech material used in this study is a four paragraph (PA) Chinese narrative discourse and it is read by a female no-professional speaker in Normal (N), Fast (F) and Slow (S) rates. The syllable numbers of 4 paragraphs are 134, 123, 151 and 34 respectively. 10 subjects were asked to identify the three speaking rates for the recorded discourse; all of them got 100% correct rates.

The recorded speech waveform was automatically segmented into syllables, initials and finals which were manually checked. And three-level stress of PW, MIP and IP was annotated as well.

3. Prosodic features in three speaking rates

3.1. Speaking rate on hierarchical constituents of discourse

How to discover the speaking rate (AR and SR) variation patterns for read fluent discourse speech? According to the hierarchical framework of the discourse, the speaking rates are observed within various prosodic units.

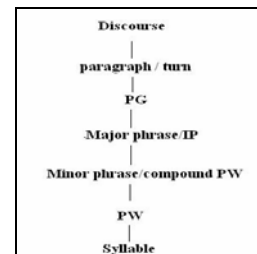


Fig 1: the hierarchical prosodic structure of Chinese discourse

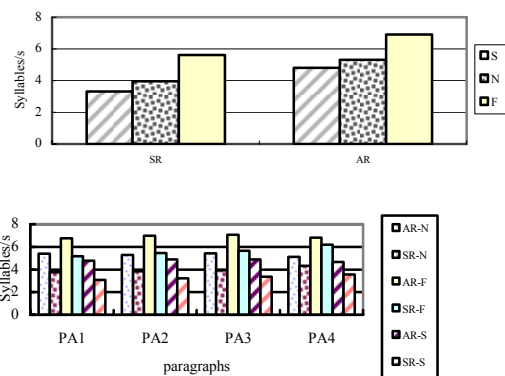


Figure 2: SR and AR in whole discourse (upper panel) and in four paragraphs (syllable/second)

Figure 2 presents the overall SRs and ARs for the whole discourse. SR-N F S: 3.963, 5.615 and 3.311 syllables/second; AR-N F S: 5.309, 6.904 and 4.801 syllables/second, and the AR and SR in 4 paragraphs. Statistics shows there is no significant difference among 4 paragraphs in 3 speaking rates (Pearson

$\chi^2=315$, $P=1.000$) and the paragraph rate is similar to that of the whole discourse as shown in the upper panel of fig.2.

With paired-samples t-test, the differences between S and N, N and F, F and S are significant. Hereafter in this paper, the t-samples are default paired t-samples, and the confidence interval is 95%.

Figures 3-4 describe the local AR and SR variation patterns in prosodic units of PG, IP (those for MIP and PW are omitted for the limit space). It is detectable that local speaking rate varying patterns are quite different among 4 paragraphs; however they have similar varying tendencies among three speaking rates even though the prosodic unit number varies.

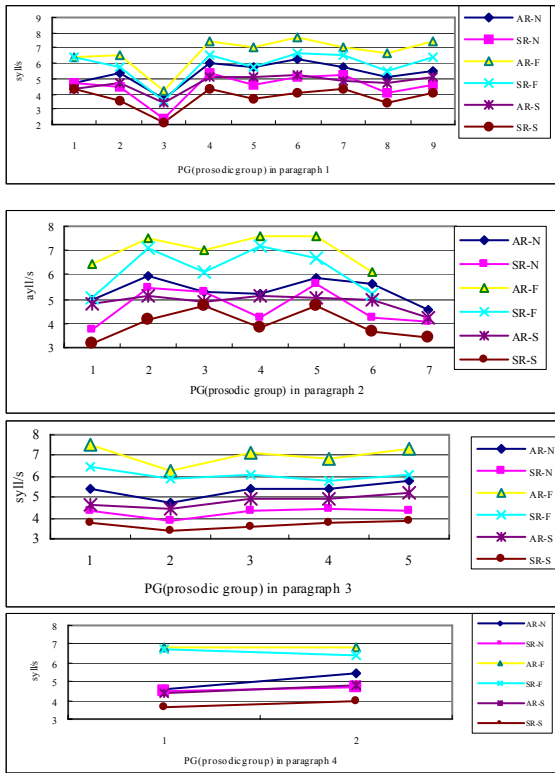


Figure 3: SR and AR in PG of four paragraphs

It seems that the local speaking rates vary randomly. But when looking for the reasons, such as the abruptly decreased local rate in the paragraph 1, we find it is caused by the information structure that the speaker expresses or transmits in that part: 5 coordinate phrases delimited by Chinese punctuation maker “、”.

One-way ANOVA (table 1) shows that significant differences exist among three ARs and three SRs for PG, IP, MIP and PW respectively. Paired-Samples T Test (table 1 right column) demonstrates that significant difference exists between the AR and SR counterparts in 3 rates for all these prosodic units.

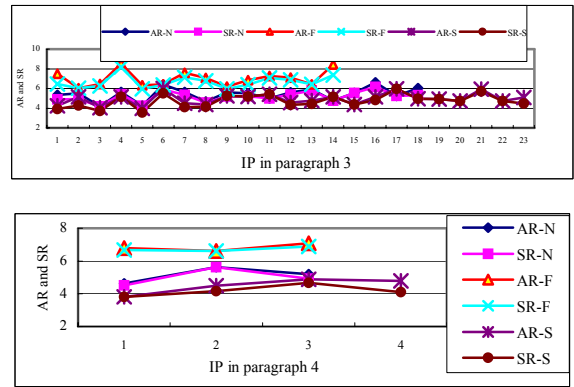
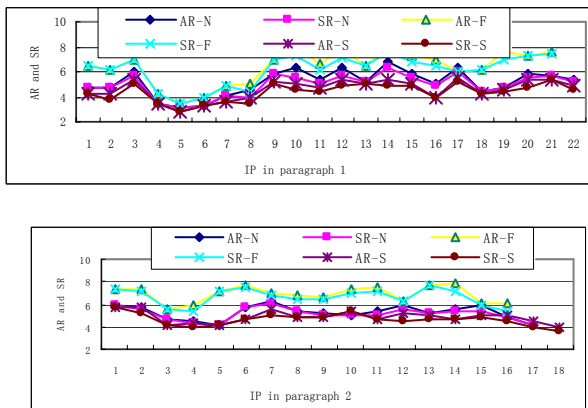


Figure 4: SR and AR in IP of four paragraphs(X-axis: IP no.)

Table 1: One-way ANOVA in three ARs and SRs and paired-Samples T Test between AR and SR for different prosodic units

Units	F (AR/ SR) (P=0.0)	T(F N S) (P=0.0)
PG	F(2,65)=73.42/64.404	T=9.734,9.086,11.113
IP	F(2,178)=87.87/99.48	T=6.880,6.760,7.773
MIP	F(2,394)=110.77/ 107.61	T=6.379,6.07, 4.042
PW	F(2,592)= 96.18/ 92.32	T=4.425,4.65, 3.473

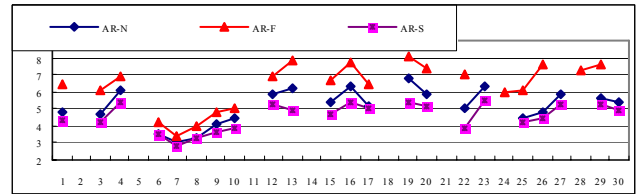


Figure 5: Three ARs in IPs grouped by PG in paragraph 1 (X-axis: prosodic unit, Y-axis: AR in Syllable/S)

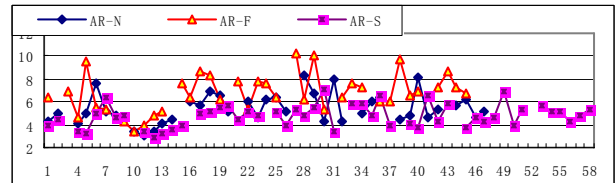


Figure 6: 3 ARs in MIPs grouped by PG in paragraph 1

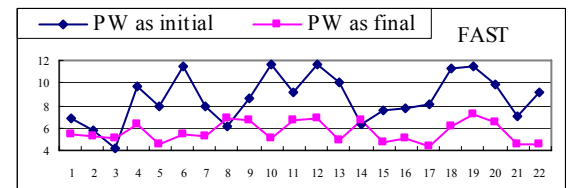
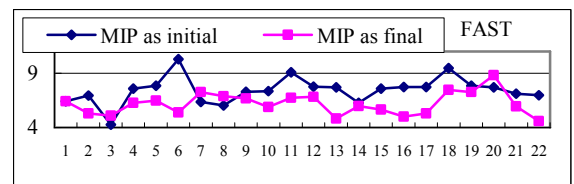
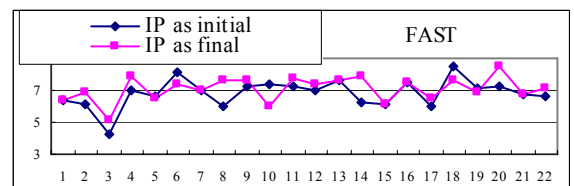


Figure 7: Initial and final ARs in IP, MIP and PW for all PGs (X-axis: prosodic unit, Y-axis: AR in Syllable/S)

To observe the initial and final local ARs for prosodic units of PG, IP and MIP, the samples of all PGs are plotted in figure 7, where the dark lines stand for the initial ARs and the light lines stand for the final ARs. The results demonstrate that the initial and final ARs try to keep similar relations at each kind of boundary. However, when boundary rate (the rate of the boundary units) in different units is observed, it is shown that the initial rates of PG are faster than the final rates for MIP (77%) and PW (86%). Likewise, IP initial is faster for PW (F: 100%, N: 94%, S: 86%) than IP final in all 3 ARs.

These kinds of 'initial faster' and 'final slower' for PG and IP are final PW lengthening. By contrast with the initial lengthening as we know, the initial PW is shortening.

3.2. Size and duration of prosodic units

The size and duration of prosodic units are usually considered as rhythmic relatives. Figure 8 presents the average number of syllables in a PG, IP, MIP and PW. The numbers of syllables in PG and PW are irrelative to speaking rate [PG: $F(2,65)=.054$, $p=.947$; PW: $F(2,591)=.397$, $p=.942$], while those of IP and MIP are relative to speaking rate [IP: $F(2,178)=3.428$, $p=.035$, into two classes: N&F, N&S; MIP: $F(2,394)=22.955$, $p=.000$, into three classes].

Therefore, the increase or decrease of speaking rate will mainly affect the size of IP and MIP, while the size of PG and PW have little changes.

The average durations of prosodic units (excluding internal silent pause) are calculated and shown in figure 9. The average duration of PW is significantly different in 3 rates [$F(2,592)=41.605$, $P=0.0$], while for MIP, IP, PG and PA, there are no significant differences [$F(2,401)=.618$, $P=.540$; $F(2,171)=2.089$, $P=.127$; $F(2,65)=2.226$, $P=.116$; $F(2,9)=.575$, $P=.582$].

The average durations of prosodic units (including internal silent pause) are also calculated. This time, durations of PW and PG are significantly different [$F(2,592)=38.905$, $P=0.0$; $F(2,65)=3.402$, $P=.039$], while others are not, which seems consistent with the varied unit size.

So, increase and decrease of speaking rate will reconstruct the prosodic structure. The duration of PW and PG is remarkably affected.

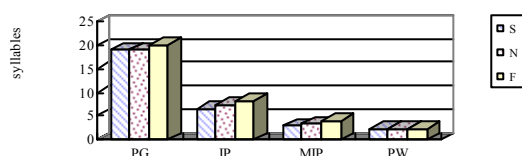


Figure 8: Prosodic unit size in 3 rates (in syllable No.)

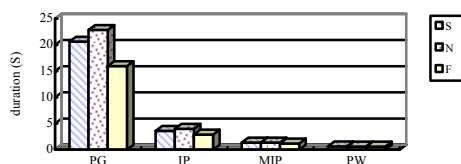


Figure 9: Prosodic unit duration in 3 rates (Second)

3.3. Silent pause

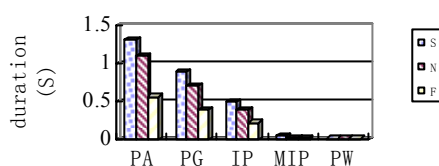


Figure 10: Silent pause at boundaries in 3 rates (in Second)

The silent pauses at the boundaries will be involved into the adjustment per se. Figure 10 shows that along with the rate increases, the silent pause decreases except that after PW which is $F>S>N$; the varying ratio to Normal rate is almost the same for IP, PG and PA, about 1.2 times for rate increase and 0.5 times for decrease. When rate increases, the silent pause after MIP and PW is shortened about 0.8 times, while it is shortened a half after PW and lengthened 2.6 times when rate decreases.

So, speaking rate has nonlinear effect on the silent pauses after different prosodic boundaries.

3.4. Prosodic boundary

With rate increases and decreases, the number of prosodic boundaries except PW decrease and increase about 0.8-0.95 and 1-1.2 times to normal rate respectively. Varying amplitude is bigger for smaller units than larger units. Interestingly, PW boundary number decreases (0.8 times) when slowing down and increases (1.3 times) when speeding up.

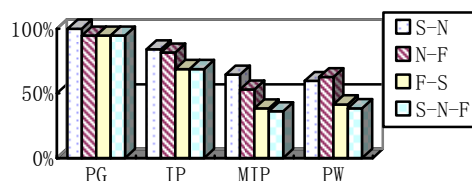


Figure 11: Consistency ratios of prosodic boundaries

With boundary level decreases, the boundary consistent ratio decreases (consistent boundaries refer to those boundaries with the same places and the same levels in different rates), figure 11 shows that S-N get the highest consistency, then F-N and F-S, the lowest are S-F-N and S-F. The rate effect on prosodic structure is nonlinear as well.

3.5. Stress distribution and its length

The stress consistency ratios of IP, MIP and PW are calculated in three rates. A consistent placement is broadly defined as the stress is placed at the same syllable between two rates.

IP gets higher ratio than that of MIP and PW, and N-S ratio is higher than those of other conditions for IP and MIP, while N-F is highest for PW (shown in figure 12).

Figure 13 shows that for PW, MIP and IP stresses, the durations of the stressed finals have significant difference among three rates [PW: $F(2,195)=10.34$, $P=0.0$; MIP: $F(2,213)=29.76$, $P=0.0$; IP: $F(2,178)=12.81$, $P=0.0$]. However, there is no difference between N and S. For PW, MIP and IP stresses, the durations of the stressed syllables have significant difference among three rates [PW: $F(2,195)=11.41$, $P=0.0$; MIP: $F(2,213)=34.38$, $P=0.0$; IP: $F(2,178)=16.10$, $P=0.0$]. However, there is no difference between N and S for PW and MIP.

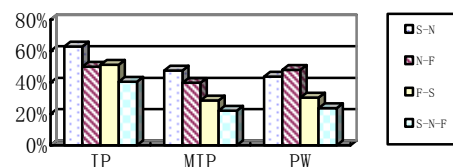


Figure 12: Consistency ratios of stresses

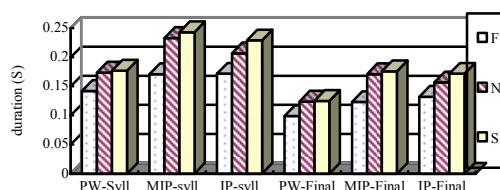


Figure 13: Duration of stressed syllable and final in 3 rates

Therefore, speeding up shortens the stressed finals and syllables, while slowing down makes no effect on stressed finals and syllables except lengthening the stressed syllable of IP.

3.6. F0 characteristics

Mean F0 values of fast, normal and slow rates are 240.35 Hz, 236.95Hz, and 231.79Hz respectively, and significant differences are found between them [F - N :t=3.803, P=0.0<0.01; S - N:t=4.717, P=0.0<0.01; F - S :t=9.354, P=0.0<0.01]. The F0 range between 75% and 25% is: N(6.10st)> F(6.02st)> S(5.97st).

F0 variations in different prosodic units are: (1) F0max and F0rang: PA>PG>IP>MIP>PW; F0min: PA<PG<IP<MIP<PW. (2) F0max of IP, PG and PA in 3 rates are F>S & N. One way ANOVA shows no difference among 3 rates for all units [PW: F(2,592)=.88, P=.412; MIP: F(2,394)=.81, P=.441; IP: F(2,178)=.82, P=.441; PG: F(2,65)=.72, P=.486; PA: F(2,9)=.36, P=.703]. (3) F0min: One way ANOVA shows no difference among 3 rates for all units [MIP: F(2,394)= .085, P=.919; IP: F(2,178)= .357, P=.700; PG: F(2,65)=1.48, P=.234; Paragraph: F(2,9)= .458, P=.646] except PW whose S & N has no difference with F [PW: F(2,592)= 4.95, P=0.07, F > S & N]. (4) F0 range: One way ANOVA shows no difference among 3 rates for all units [MIP: F(2,394)= .930, P=.395; IP: F(2,178)= .440, P=.645; PG: F(2,65)= 0.258, P=.774; Paragraph: F(2,9)= .689, P=.257] except PW whose S & N has no difference with F [PW:F(2,592)= 4.95, P=0.07, F<S& N].

Speaking rate variation affects the minimal F0 value of PW. Increasing makes the minimal F0 value raise and compresses the range, while no significant effect can be found on the overall F0 features of other units.

Additionally, statistics on F0 overall values of stresses shows that F0 of PW stress has difference among 3 rates[F(2,195)=3.26, p=0.04], N and F are at the same level, there are no effects on stressed F0 of other units (high, low and range) among 3 rates (P>.05 for all).

The effect of speaking rate on stressed F0 is limited to prosodic words.

4. Concluding remarks

This pilot study displays a preliminary panorama on the speaking rate effects on discourse prosody with only one speaker's material. It leaves more aspects for further investigation: (1) The segmental variation caused by the articulation reduction referred to as the fifth dimension of prosody by Pfitzinger (keynote speech of Speech Prosody 2006). (2) Units contributing to articulation rate from perceptual point of view, run, IP, PW, PG or discourse? (3) Unit to calculate the articulation rate: syllable, phone or phoneme per second? [8] (4) Language specific issues, such as Chinese tone movement and alignment as in [5].

The major results of this study demonstrate that: (1) It is the expressive content or information structure of discourse that causes the local speaking rate varying. For this reason, the overall rate varying tendencies are similar among three rates of N, F and S. (2) Initial faster and final slower exist for PG and IP if PW as initial and final unit. (3). Speaking rate has nonlinear effect on the silent pauses after different prosodic boundaries. Increase and decrease of speaking rate will reconstruct the prosodic structure, and the speaking rate effecting on prosodic structure is nonlinear as well. The duration of PW and PG is remarkably affected. (4) Speeding up shortens the stressed finals and syllables; slowing down makes no effect on stressed finals and syllables except lengthening the stressed syllable of IP. This result is quite different from English in [19]. (5) Speaking rate variation affects the overall mean F0 values. Speeding up will raise F0 mean and slowing down will lower F0 mean. However, the F0 range between 75% and 25% is: N>F>S.(6) Speaking rate variation affects the F0 minimal value of PW, increasing makes

it raise and compresses the range, while no significant effect can be found on F0 overall features of other units. The effect of speaking rate on stressed F0 is limited to prosodic word, different from that in English [19].

5. References

- [1] Zellner, B., 1998. Temporal structures for fast and slow speech rate, *ESCA/COCOSDA. Third International Workshop on Speech Synthesis*, Australia.
- [2] Tseng, C.Y., 2006. Fluent Speech Prosody and Discourse Organization: Evidence of Top-down Governing and Implications to Speech Technology, Keynote speech of *Speech Prosody*, Dresden, Germany.
- [3] Tseng, C. Y., 2006. Prosody analysis, in Chin-Hui Lee, Hanzhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, *Advances in Chinese Spoken Language Processing*, World Scientific Publishing.
- [4] Dankovičová, J. 1999. "Articulation rate variation within the intonation phrase in Czech and English," *Proc. 14th ICPhS*, San Francisco, Vol. 1, pp. 269-272.
- [5] Ladd, D.R., aulknerm. D.F., et.al. 1999. Constant "segmental anchoring" of F0 movements under changes in speech rate, *JASA* 106 (3), Pt. 1, September 1999.
- [6] Koopmans-van Beinum, F. J. and van Donzel, M. E. 1996. Relationship between discourse structure and dynamic speech rate. In: HT Bunnell & W. Idsardi (Eds), *Proceedings ICSLP96*.
- [7] Fougeron, C. & Jun, S.-A. 1998. Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics* 26, 45-69.
- [8] Tillmann, H. G. & Hartmut, R. P 2006. Local Speech Rate: Relationships between Articulation and Speech Acoustics. *Third. Int. Conf. on Speech Prosody*, Dresden.
- [9] Cao, J-f (2003), Speech rate and its variations, Phonetic research report, Phonetics Lab., Institute of Linguistics, CASS. (in Chinese)
- [10] Trouvain, J. & Grice, M., 1999. The Effect of Tempo on Prosodic Structure, In *14th International Congress of Phonetic Sciences (ICPhS)*, August 1-7, pages 1067-1070, San Francisco, USA.
- [11] Koreman, J. 2003. "The perception of articulation rate," *Proc. 15th Int. Congress of Phonetic Sciences (ICPhS)*, Barcelona.
- [12] Koreman, J., 2006. Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech, *J. Acoust. Soc. Am.* 119(1), 582-596.
- [13] Koreman, J., 2006. The role of articulation rate in distinguishing fast and slow speakers, *Third. Int. Conf. on Speech Prosody*, Dresden.
- [14] Rietveld, T., and Gussenhoven, C. 1987. "Perceived speech rate and intonation," *J. Phonetics* 15, 273-285.
- [15] Trouvain, J., and Grice, M., 1999. "The effect of tempo on prosodic structure," *Proc. 14th ICPhS*, San Francisco.
- [16] Pfitzinger, H. R. 1999. "Local speech rate perception in German speech," *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS)*, San Francisco, Vol. 2, pp. 893-896.
- [17] Zellner, B., 1998. Fast and Slow Speech Rate: a Characterisation for French. *5th ICSLP*, Volume 7, pp. 3159 - 3163, Sydney, Australia.
- [18] Zellner Keller, B., 2002. Revisiting the Status of Speech Rhythm, *Proceedings of Speech Prosody 2002*. Aix-en-Provence, France.
- [19] ZU, Y.Q., LI, A. J.; LI, Y., 2006. Speech Rate Effects on Prosodic Features, *O-COCOSDA 2007*, Malaysia.