Speech Rate Effects on Prosodic Features

Yinqing ZU
MOTOROLA Research Center China
Yiqing.Zu@motorola.com

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

Yang LI
Tianjin Univ.
Visiting student of Chinese Academy of Social Sciences
liyang8334@gmail.com

Abstract

Speech corpus for data-driven TTS system is usually recorded in a medium and even speech rate. How to produce faster or slower speech from the normal rate speech? In spite of articulation rate, prosodic features needs carefully manipulated as well. In this paper, speech rate effects on prosodic features are investigated for fast, normal and slow speech, including prosodic structure variation, prosodic duration variation, F0 distribution and variation, accent placement variation.

1. Introduction

In TTS, speaking rate is one of the prosodic features to signal the naturalness of the synthesized speech. Speaking rate is distinguished into speech rate when pauses are included and articulation rate when pauses are not included (Florien and Monique 1996; Cao, 2003). There are many contributions relating to this topic. Some of them are summarized as followings.

Florien and Monique analysis the speaking rate strategy on discourse level, they found that the average syllable duration of the first run of a paragraph is longer than the overall mean value for per speaker in more than 60 % of the cases. Inspection of the quartiles of runs with highest ASD-values and those with lowest ASD-values for each of the speakers shows quite different structures, which can be explained on the basis of partly local and partly global discourse characteristics.

It has been found that French speakers use a number of strategies for consciously achieving an increase in speech rate (Fougeron, C. & Jun S.-A., 1998). These include a reduction in the number of phrases and the demotion of major 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 (1999) found a considerable effect of tempo in the pausing structure either in the number of pauses, or in the mean pause duration. 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 is obtained first by lengthening the duration of segments, second by producing additional syllables, and third by producing pauses (Zellner, B. 1998).

The present study is trying to make a contrast analysis on the prosodic features among three rate speech, so as to disclose some strategies on speech rate (SR) adjustment on prosodic features.

2. Speech corpus and annotation

50 declarative sentences were selected from MOTO multilingual speech corpus (Zu et al. 2006), in three rates: Slow, Normal, and Fast (S, N, and F hereafter) , read by a female British English speaker. Totally 150 utterances are selected. The average sentence lengthen is 17.9 words (deviated from 8 to 25 words). Utterances are automatically segmented into words and phones and manually checked,. Prosodic annotations include tags of pitch accent, intonational phrase accent, boundary tones, and pitch accent implementation domain (ID) and intonational phrase boundaries (IP). (JIA and LI, 2005)

3. The analysis of prosodic features

3.1. Speech rate analysis

The duration of each utterance contains silent pause duration excluding the beginning and the end silent pause. Table 1 shows the average syllable and word numbers per second and table 2 presents the average duration of an utterance. Along with the rate increases, the number of syllables and words also increase, but the duration of sentence decrease.

SR	Syll./s	Dev.	Words/s	Dev.	utterances
S	3.94	0.10	2.40	0.14	50
N	4.10	0.09	2.49	0.09	50
F	4.41	0.15	2.68	0.14	50

Table 1: Speech rates (SR) in syllable or word

With t-paired test, the differences between S and N, N and F, F and S are significant. In the present

paper, if we don't make supplementary explanation, the t-samples are t-paired samples, and the confidence interval is 95%.

	Ave.	Dev.	utterances
S	7.60	5.23	50
N	7.29	4.64	50
F	6.78	3.97	50

Table 2: Average duration for an utterance

3.2. Prosodic structures analysis

This part we analysis the prosodic boundary differences of ID and IP among three rates. Because of the diversity and variability of articulation, the same sentence uttered at different times may have different prosodic structures to transmit the same or different meanings (Chu,2001). Our speech materials are context free and focus uncontrolled utterances, so we can't exclude this diversity or variability even though the utterances were recorded in neutral emotionless state.

Figure 1 shows the average numbers of ID and IP boundaries in an utterance. For ID: N>F>S and for IP: S>N>F. With speech rate increases, the number of IP decreases, while that of ID doesn't show this tendency.

Next, the boundary consistency ratio (CR) is observed. Here we discarded one invalid utterance.

Figure 2 are the consistency ratio (CR) of ID and IP. We can see that either ID or IP, the consistency ratio between S and F is the lowest and is the highest between N and F. And the consistency ratio of IP is higher than ID.

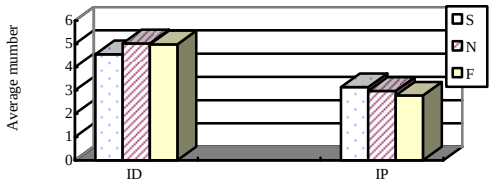

Figure1: Average boundary numbers in 3 rates

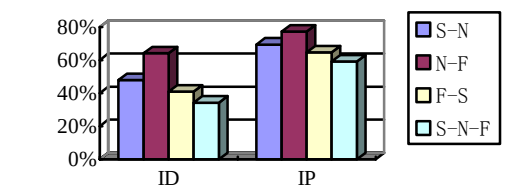

Figure2:Consistency Ratios of ID and IP

3.3. Length of Prosodic units

Figure 3 presents the average number of words and syllables in an ID and an IP. For ID, the numbers of words and syllables are S>F>N. With t-paired test, there is no significant difference between each other (syllable test: N-S: P=0.068, N-F P>0.1, F-S P>0.1, word test: all P>0.1). While for IP, the numbers of words and syllables increase when rate increases. With t-paired test, F-N has significant difference both in words (P=0.01) and syllables (P=0.012), S-F: word test: P=0.049; syllable test: P=0.076, S-N: word test: P=0.681;syllable test: P=0.855.

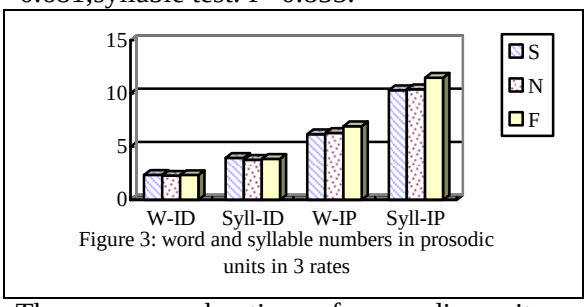

Figure 3: word and syllable numbers in prosodic units in 3 rates

The average duration of prosodic units are statistically calculated and shown in figure 4. We can see that the average duration of ID increases when speech rate decreases. Duration of ID in slow rate is the longest. From the t-paired test results, there is significant difference between any two rates (P<0.001).

While for the average duration of IP, S>F>N. T-test demonstrates there is no significant difference between each other (F-N:P=0.652,N-S:P=0.374,F-S:P=0.640). Combining previous result, we found that from normal rate to fast rate or to slow rate the duration of IP keeps consistent, while a number of significant changes exist for IP boundary. It implies that the shorting of ID and IP boundary variation makes the speech rate increase, in another word, articulation rate increases.

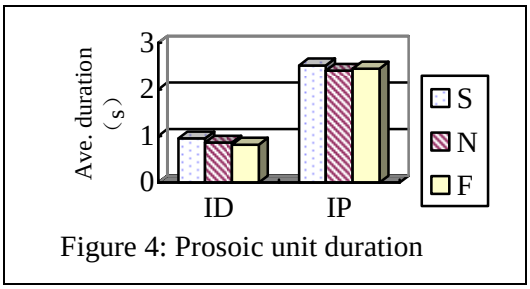

Figure 4: Prosoic unit duration

3.4. Accent placement and its duration

Here we calculate the consistency ratios of pitch accent and intonation phrase accent in three rates. Here a consistent accent placement is broadly defined as two accents are placed at the same words between two utterances with same text.

Figure 5 shows the consistency ratios of pitch accent and intonation phrase accent. Although the consistency ratio of intonation phrase boundary is higher, the consistency ratio of intonation phrase accent is lower. The consistency ratio of pitch accent is not high either. The consistency ratio between N and F is higher than that of between N and S, and the consistency ratio between S and F is the lowest.

The low consistency ratio for pitch accent demonstrates that the pitch accent changes greatly . But you can also see the ratios are different for F-N and N-S with the former higher than the later, the ratios between F and L is the lowest. Whether these pitch accent deviations are mainly caused by speech rate variation or other factors needs further investigation.

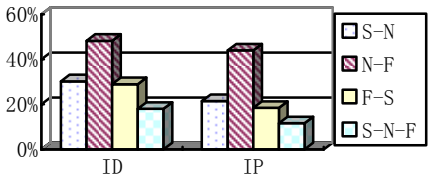

Figure5: Consistency ratios of pitch accent and IP accent

As shown in figure 6, we got the duration distributions for accented vowels bearing pitch accent or phrase accent respectively. The average duration of phrase accent is longer than that of pitch accented: 20ms longer under normal and fast rate, while 6 ms longer under slow rate. With t-paired test, pitch accent between F and N (P=0.036) has significant differences, while for other two situations, N-S (P=0.719) and F-S (P=0.093), have no significant differences. For intonation phrase accent: F-N (P=0.039) and N-S (P=0.002) have significant differences, and F-S (P=0.204) has no significant differences.

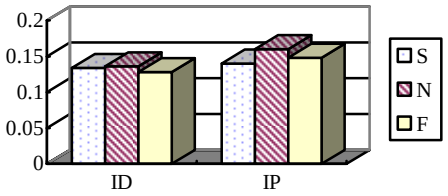

Fig 6: Duaration of pitch accent and phrase accent（s）

3.5. 3.5 F0 characteristics

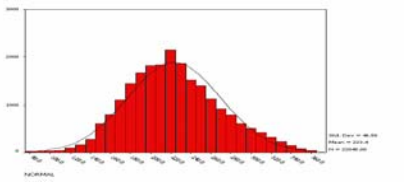
Figure 7: F0 Histogram at Normal rate

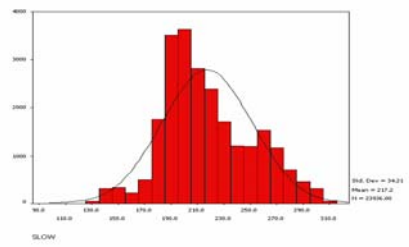
Figure 8: F0 Histogram at Slo w rate

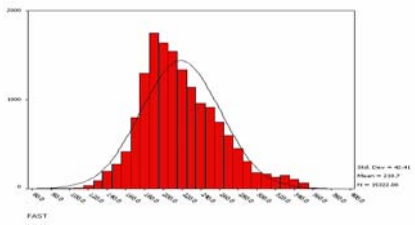
Figure 9: F0 Histogram at Fast rate

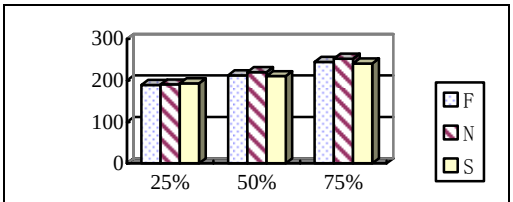

Figure 10: F0 at points of accumulative frequency 25%,50% and 75%;

Figure 7-9 are histograms of all f0 values for normal、slow and fast rate respectively. Figure 10 shows the quartiles of all f0 values under three rates respectively, and also the mean, pitch range between 25% and 75%. F0 mean values of fast, normal and slow rate are 218.7Hz、223.4Hz、217.2Hz respectively and significantly differences exist among three rates (P<0.001, t-test, not paired t-test). All F0 distributions approximate Normal Distribution; normal rate speech has better distribution than the other two. It implies that amplitude of individual rising and falling pitch movements for fast rate is the biggest among three rates.

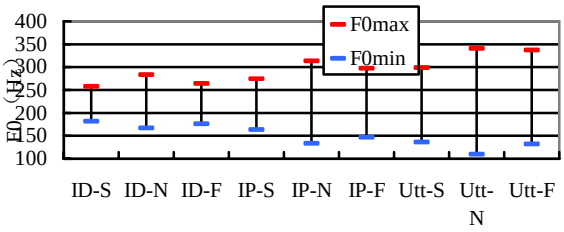

Figure 11: Max and min f0 in different prosodic units

In an utterance, there maybe contain several IDs and IPs. Figure 11 shows the average values of maxima and minima f0 in different prosodic units, and the corresponding pitch range in semitone calculated by $12*[\lg(F_{F0max}/F_{F0min})/\lg 2]$. The order of the maxima of f0 is N>F>S, while the minima order is: S>F>N, the f0 range is N>F>S. With t-paired samples, for the maxima f0 of the utterance, F-N (P=0.207) has no significant difference, S-F and S-N (both P<0.001) have significant differences. And the results of the minima f0 are: S-F (P=0.171), S-N and F-N (both P<0.001). For ID、IP, either the maxima or the minima values ,the results have significant differences.

Statistic analysis was also made to get the maxima and minima f0 values of pitch accent and intonational phrase accent, shown in table 3. For pitch accent, the order of the maxima of f0 is: N>F>S, and for intonation phrase accent is N>S>F. While the order of the minima, for pitch accent is F>S>N; for intonation phrase accent is S>F>N. The t-paired test results of the maxima of pitch accent: F-S (P=0.100), F-N and N-S (both P<0.001), while the minima: S-N (P=0.271),N-F (P=0.162) and F-S (P=0.797) all show no significant differences. The t-paired test results of maxima of intonation phrase accent are S-N (P=0.849), N-F (P=0.267) and F-S (P=0.501), all show no significant differences, while the minima: S-N (P<0.001), N-F (P<0.005)

and F-S (P<0.001) all show significant differences.

accent	SR	F0max	F0min	Pitch range (ST)
Pitch accent	S	234.73	215.01	1.52
	N	249.44	211.68	2.84
	F	239.13	215.73	1.78
Intonation phrase accent	S	245.78	226.36	1.43
	N	246.69	194.37	4.07
	F	242.44	207.21	2.72

Table 3: F0 values of pitch accent and international phrase accent (Hz)

Finally, the average pause duration after the ID and IP (except the last one) was calculated. We can see from figure 12, for the average pause duration after ID, normal rate is the longest. For the average pause duration after IP, fast rate is the longest. The result of T-paired test for ID: N-F (P=0.019) has significant difference, while F-S (P=0.434) and S-N (P=0.236) have no significant difference and for IP: N-F (P=0.442) has no significant difference, F-S (P=0.004) and S-N (P=0.009) have significant difference.

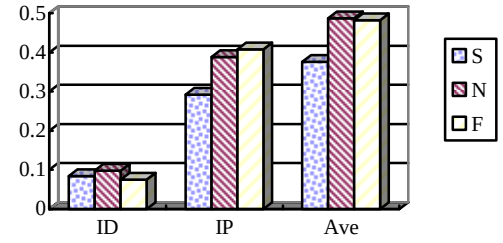

Figure 12: silent pause after prosodic units(S)

4. Concluding remarks

In this paper, speech rate effects on prosodic features are contrastively investigated for fast, normal and slow rate speech. The major effects are concluded as follows:

(1)With rate increases, intonation phrase (IP) number decreases significantly, while no statistic change found for pitch accent implementation domains (ID). IP boundary agreement ratio is higher than that of ID boundary in 3 rates. IP contains more words or syllables with rate increase. This indicates that articulation rate increases indeed.

(2) Placement of Intonation Phrase accent varies greatly in 3 rates and has low consistency ratio. However whether this variation is caused by speech rate or other factors needs to be carefully studied in the future work.

(3) F0 ranges for IP and ID are ordered as normal > fast > slow; F0 ranges of vowels bearing IP accent or pitch accent are ordered as normal > fast > slow as well. The max and min F0 values have no consistent relation among 3 rates.

(4) F0 mean values of fast, normal and slow rate are 218.7Hz、223.4Hz、217.2Hz respectively and significantly difference exist among three rates (P<0.001). F0 distribution of approximates Normal Distribution, normal rate speech has better distribution than the other two.

(5) Vowel duration of IP accent is significantly longer than that of pitch accent, but there is no noticeable relation among 3 rates.

(6) Average duration of silent pause in each utterance is longer at normal rate than at fast rate, shortest at low rate. Silent pause between IP is ordered as fast > normal > slow, while for ID is normal > slow > slow.

The results consist with some previous research on French (Fougeron and Jun,1998), while have some discrepancy with that on Chinese (Cao,2003).

From research on Germany, Jorgen and Martine found the speakers used different strategy to change their speech rate to slow and to fast (1999), similarly, we found our speaker also used asymmetry strategy from normal to fast and to slow.

However, currently some results can't be explained without further investigation, such as the great deviation of accent placements among three rates. The ongoing wording is to look for the reducing or lengthening strategy for accented and unaccented component, i.e., the rhythm variation.

We are also focusing also on the situation for Mandarin Chinese. Are there any differences with English? How is the speech rate variation on discourse level? How effective t to use speech rate variation strategy to improve the naturalness in TTS?

References

Cao, J. (2003), Speech rate and its varitions, Phonetic research report, Phonetics Lab., Institute of Linguistics, Chinese Academy of Social sciences.

Chu, M. (2001), Prosodic research and naturalness of the synthesized speech. In proceeding of the 5th national modern phonetics conference- New Century Phonetics, Beijing, Ts inghua Publisher.

Fougeron, C. & Jun, S.-A. (1998): Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics* 26, 45-69.

Florien J. Koopmans-van Beinum and Monique E. van Donzel (1996), Relationship between discourse structure and dynamic speech rate. In: HT Bunnell & W. Idsardi (Eds), *Proceedings ICSLP96*.

Jia, Y., Li, A. (2005)，An introduction the English intonation labeling system IVi, *Acoustic Technology*, Vol24. 3,2005.

Jürgen Trouvain & Martine Grice (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, 1999.

Zellner, B. (1998). Fast and Slow Speech Rate: a Characterisation for French. *ICSLP, 5th International Conference on Spoken Language Processing*. (Volume 7, pp. 3159 - 3163), December 1998, Sydney (Australia).

Tseng, C. (2006), Fluent Speech Prosody and Discourse Organization: Evidence of Top-down Governing and Implications to Speech Technology, Keynote speech of Speech Prosody 2006, Dresden, Germany.

Zu Y., Cao, Z., et.al.(2006), Multi-lingual TTS Speech Corpus Development, *ISCSLP2006*, Singapore,2006.