Towards an Integrated Understanding of Speaking Rate in Conversation

Jiahong Yuan, Mark Liberman, Christopher Cieri

Department of Linguistics, Linguistic Data Consortium
University of Pennsylvania, USA
jiahong@ling.upenn.edu, myl@cis.upenn.edu, ccieri@ldc.upenn.edu

10.21437/Interspeech.2006-204

Abstract

We investigate factors that affect speaking rate in conversation, using large corpora of conversational telephone speech in English and Chinese. We find that speaking rate as a function of "turn" length rises rapidly for turns from one to seven words; remains level (when final words are included) or falls gradually (if final words are excluded) for turns of medium length; and rises slowly for longer turns. When talking with strangers or discussing certain topics, people tend to use longer turns but slower speech rates. In general older people have a slower speech, and males tend to speak slightly faster than females. Finally, we find that the effect of L1 (native language) on L2 (second language) speaking rate is L1 dependent.

Index Terms: speaking rate**,** speech rate, conversational speech

1. Introduction

Speaking rate has been found to be related to many factors: individual, demographic, cultural, linguistic, psychological and physiological. However, we lack a comprehensive understanding of these factors and their interactions. The problem is a difficult one because of the large number and variable definition of potentially relevant factors, the many different ways to define and analyze rate, and the great variability of the phenomena under any definition.

Among the demographic factors, the effect of age on speaking rate has been consistently reported. In general, older speakers have a slower speaking rate, perhaps due to both physiological and psychological reasons [1, 2]. This effect has also been confirmed by perception studies [3]. Studies on speaker sex and dialect region have, however, reported contradictory results. Sex and dialect region were shown to have significant effects on speaking rate in [4] but not in [5]. It was also reported that nonnative speakers have a slower speaking rate than native speakers [6].

Speaking rate is also affected by utterance length and utterance position. It has been found that there is an inverse relation between segment duration and utterance length, i.e., the longer the utterance, the shorter the average segment duration [7]. On the other hand, taking into account the effect of phrase-final lengthening [8], and more general, boundary adjacent lengthening [9, 10], a short utterance is expected to have a longer average word duration (slower speaking rate) than a long utterance. Quené (2005) found that speaking rate depends mainly on utterance length [5]. In his study, the statistical significance of the effects of age, sex, and dialect region disappeared after including utterance length as a factor.

Situational factors such as topic, speaker relationship, etc. may also affect conversational speaking rate. Such factors have been investigated in studies of disfluency [11, 12], but have, to our knowledge, not been explored in published speaking-rate studies. In this study, we investigate the effects of all these factors on speaking rate in both English and Chinese conversational telephone speech. We regard this work as a first step towards a comprehensive analysis. Because our research is based on published corpora, we look forward to future investigations that will supplement or revise our results.

2. Corpora and methods

To investigate the effects of demographic factors and conversation topics, we make use of the transcripts of Fisher English Part I (LDC2004T19 and LDC2004S13) and HKUST Mandarin Part I (LDC2005T32 and LDC2005S15). In these corpora, detailed speaker information and conversation topics are provided.

However, the conversations in these corpora are nearly all between strangers. To study whether speaker relationships affect speech rate, we further analyze corpora of conversations between friends and family members: CallHome English (LDC97T14, LDC97S42) and CallHome Chinese (LDC96T16, LDC96S34) and CallFriend Chinese (LDC98T26, LDC98S69).

During the original creation of these corpora, the recorded speech was segmented into pause groups (called "segments" hereafter) as a convenience for the human transcribers, and also as a method to anchor the transcripts for use in training speech-recognition systems. These segments are often equivalent to speaker turns, and may be regarded as a reasonable proxy for turns, though longer turns may be divided into several segments. In most cases, the boundaries of these segments are the only reliable time-stamps that are published. Unfortunately, the procedures for creating these segments are not consistent across (and sometimes even within) the cited corpora. In some procedures, each side was transcribed separately, and the segments may overlap to a considerable extent, where the speakers spoke at the same time or went back and forth rapidly. In other procedures, both speakers were transcribed in a single pass, and the time-stamped segments overlap much less often. In both methods, the initial division into convenient segments may have been done manually, or may have been done automatically with possible manual revision by the transcribers. As a result of these differences, it is necessary to be careful in interpreting or comparing segment-based calculations of speaking rate.

One indication of the nature of this problem can be seen by exploring various definitions of speaking rate applied to a modest-sized corpus that has been carefully aligned at the word level. This is the version of English Switchboard [13] corrected and aligned at ICSI, comprising 2,438 conversations. If we

calculate the overall speaking rate for this corpus, by simply adding up the number of words spoken by both participants, and dividing by the total elapsed time of the conversations, we find an overall average rate of 196 words per minute (WPM). For individual conversations, the rate measured in this way ranges from 111 WPM to 291 WPM. If we use the word alignments to exclude silences, non-speech noises and so on, we find an average "net" speaking rate of 236 WPM, with a minimum of 158 WPM and a maximum of 312 WPM. However, if we calculate the rate by adding up all of the turns for each speaker in order to get the total time, we find that there is so much overlap in the turn boundaries that the average turn-wise rate (total word count divided by the sum of segment times) is only 164 WPM, or 14% less than the rate calculated using the total conversational time.

Therefore, we begin our analysis with two corpora that have been carefully word-aligned: the just-mentioned version of the English Switchboard corpus, and similarly word-aligned Chinese CallHome and CallFriend [14]. In the case of the English "Fisher" corpus, we are using only the portion that was transcribed by the WordWave company and segmented at BBN.

Table 1 is a summary of the corpora used in this study:

Table 1. *Sizes of the corpora used in the study.*

Corpora	Sides	Segments
English Fisher I (WordWave part)	10,150	943,044
English Switchboard	4,876	248,479
English CallHome	200	27,542
Chinese Fisher I (HKUST)	1,746	189,568
Chinese Callhome/CallFriend	284	47,004

3. Results and discussion

Effect of segment length and position

Figure 1 shows the effect of segment length, i.e. the number of words in a segment, on speaking rate.

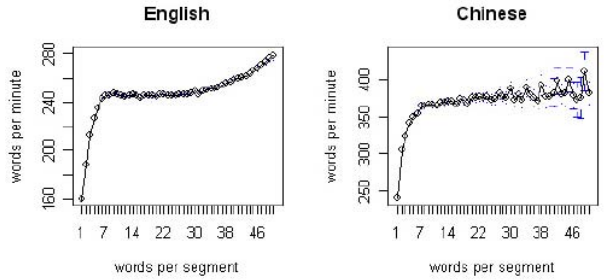

Figure 1. *The average speaking rate for each segment length (only the segments containing less than 50 words are shown).*

In doing this calculation, we used the word-aligned version of the English Switchboard corpus, and the word-aligned version of the Chinese CallHome and CallFriend corpora. The segments are those used in the (cited revision of the) published versions of these corpora.

We have excluded the time associated with any within-segment silences or non-speech noises. In the case of Chinese, since the published transcriptions have been carefully segmented into words as opposed to characters, we have shown the rate in terms of words per minute rather than characters per minute. We can see that in both English and Chinese, there is an abrupt rise of speaking rate for the segments containing from one to seven words. For the segments having eight to about 30 words, however, the speaking rate stays level. And then, especially in English, the speaking rate rises again, but with a more gradual slope.

To explore the contribution of boundary-adjacent lengthening to the initial abrupt rise in speaking rate, we calculated the duration of words at different positions of shorter segments (1 to 12 words) in the English Switchboard corpus. The results are shown in Figure 2.

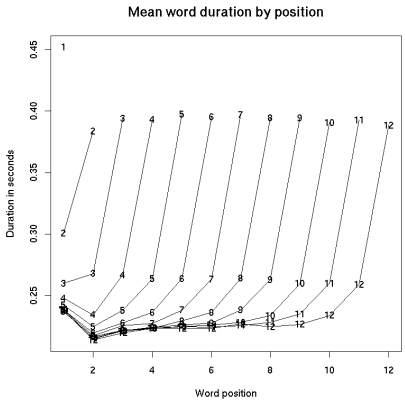

Figure 2. *The average word duration at each segment position (only the segments containing less than 12 words are shown).*

Word duration is highly dependent on segment position: the second word is a little shorter than the first; then the words have similar durations until the third to the last word; the second to the last word is longer than the previous ones, and the last word is the longest. Figure 3 separates word categories defined by POS when calculating word duration by position. Short words, for example, refer to the words that belong to ['DT', 'IN', 'PDT', 'PRP', 'TO', 'WP', 'WRB'], each of which has an average duration of less than 0.2 seconds across all positions.

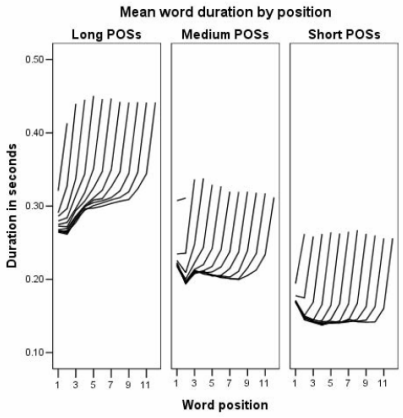

Figure 3 *The average word duration at each segment position for different POS categories.*

The different POS categories show the same pattern, suggesting that the correlation between word duration and position shown in Figure 2 is not an artifact of unbalanced

distribution of word classes, but an inherent effect of word position in a segment on its duration.

Effect of speaker relationship

In the Switchboard and Fisher corpora, the speakers discuss assigned topics, nearly always with strangers (though in a very small fraction of cases, the speakers may have been acquaintances). In the CallHome and CallFriend corpora, the speakers are friends or family members, and the topics were freely chosen by the speakers. The effect of this difference on speaking rate is shown in Table 2. In order to minimize any possible effects of differences in segmentation practices, we are comparing overall speaking rate, in which the total number of words used by all participants in a conversation is divided by the total amount of elapsed time. We have also shown the average segment durations in the two cases, since this is a reasonable proxy measure for turn length.

We can see that in both English and Chinese, people tend to use longer segments (in terms of word or character count) but slower speaking rates when talking with strangers than when talking with friends or family members (intimates). This is opposite to the general pattern that longer turns have faster speaking rate. It may be that people are more polite or more formal when talking with strangers, and use longer turns and slower speaking rate in consequence. It is also plausible that shared knowledge among intimates permits faster speaking rates without loss of intelligibility (but note that the rate differences are only about 10%).

Table 2. *The effects of speaker relationship on speaking rate and segment length, with .95 confidence intervals.*

	Rate	Segment length
English CH	214±6.73 wpm	7.84±0.95 words
English Fisher	193±0.71 wpm	10.00±0.02 words
Chinese CH/CF	247±10.2 cpm	9.76±0.83 chars
Chinese Fisher	228±2.85 cpm	10.42±0.04 chars

Effect of conversation topics

Both conversation topics and detailed speaker information are provided in the English and Chinese Fisher corpora, which contain telephone conversations between strangers. The following analyses use these corpora.

Because of differences in segmentation practices, we are using only the portion of part 1 of the English Fisher corpus that was transcribed by WordWave. Also, the Chinese Fisher corpus (collected and transcribed at HKUST) was not segmented into words, so we are using characters rather than words as the unit.

Figure 4 shows the average speaking rate as well as turn length for each conversation topic. The topics were assigned to the speakers at the time that each conversation was recorded.

From Figure 4 we can see that conversation topics significantly affect both speaking rate and segment length. In English, for example, the average speaking rate ranges from 152 words per minute to nearly 170 words per minute, and the averaged segment length ranges from 9 words per turn to 11 words per segment. More interestingly, some conversation topics tend to have both longer speaker turns and slower speaking rates. One possibility is that when talking about these topics people tend to produce more "important" or "unexpected" turns, and allow more time for the listeners to process their words. It has been shown in the literature that important or unpredictable portions are spoken at a relatively slower rate [15].

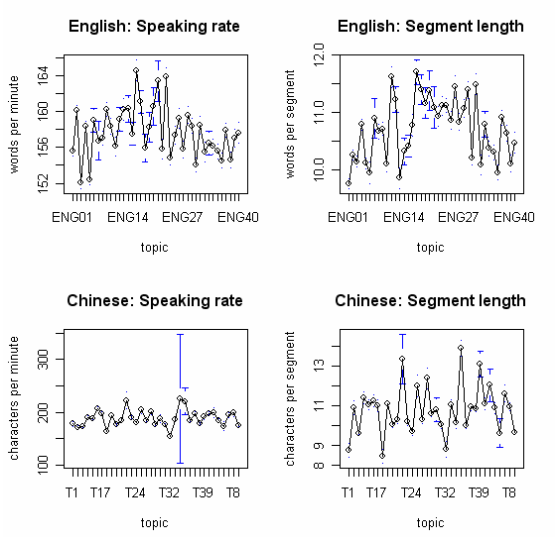

Figure 4. *The effects of conversation topics on speaking rate and segment length.*

Effect of age and sex

Figure 5 and 6 show the effects of speaker age and sex respectively.

From Figure 5 we can see that older people tend to have a slower speaking rate, and they produce significantly more variation in the length of their turns than younger speakers do. Similar results have been reported in [5].

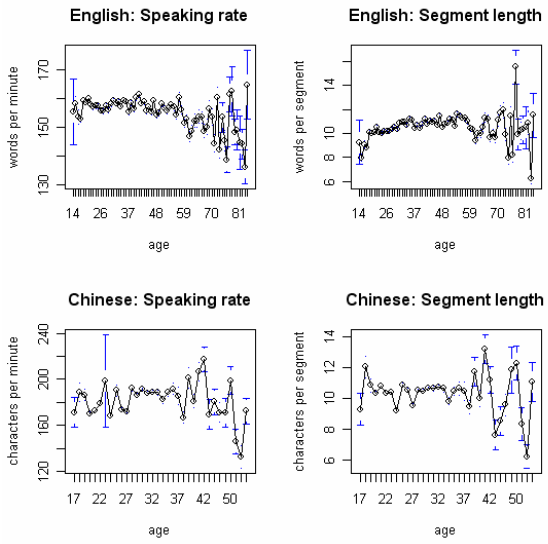

Figure 5. *The effects of speaker age on speaking rate and segment length.*

Males tend to speak faster than females, as shown in Figure 6. The difference between them is, however, very small, only about 4 to 5 words or characters per minute (2%), though it is statistically significant. It might be due to things that we would

not normally think of as speech-rate parameters, such as differences in word-frequency distributions. The opposite patterns of segment-length difference between male and female in Chinese and English are interesting, and need more study.

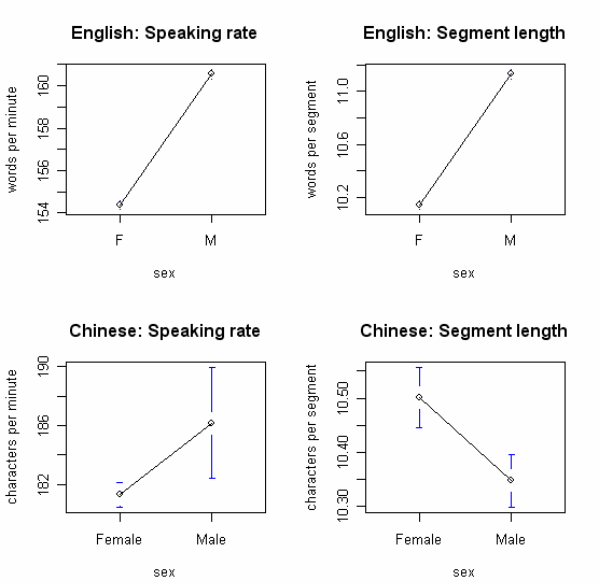

Figure 6. *The effects of speaker sex on speaking rate and segment length.*

Effect of native languages

Figure 7 displays the average speaking rate and segment length of speakers with different native languages (L1) when speaking in English as their second language (L2). All these speakers are fluent in English.

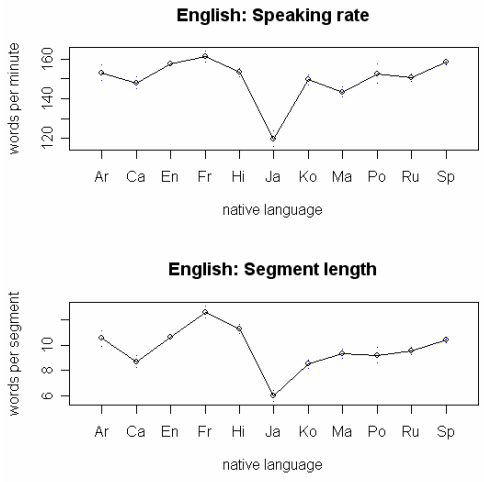

Figure 7. *The effects of native languages on English speaking rate and segment length.*

We can see that when speaking in English, the Japanese speakers have a slower speaking rate than the others. This might result from some characteristic of the Japanese language substrate. For example, a previous study [16] showed that less advanced Japanese learners of English make less durational contrast between lexically stressed and unstressed vowels than native English speakers, apparently due to the fact that there is no durational contrast between pitch-accented and unaccented moras in Japanese. It might also be due to a Japanese culture of "politeness" [17], or differences in L2 teaching methods.

It has been claimed that nonnative speakers have a slower speaking rate than the native speakers [6]. We can see from Figure 7 that this holds for most (but not all) of the languages. However, our results also show that the speaking rate in L2 is L1 dependent. Further research is needed to find out whether this effect mainly depends on culture differences or language differences, or both.

4. References

[1] Amerman, J. D. and Parnell, M. M., "Speech timing strategies in elderly adults", *Journal of Phonetics*, 20, 65-76, 1992.

[2] Verhoeven J. et al., "Speech rate in a pluricentric language", *Language and Speech*, 47, 297-308, 2004.

[3] Stölten, K. and Engstrand, O., "Effects of perceived age on perceived dialect strength: A listening test using manipulations of speaking rate and F0", *PHONUM*, 9, 29-32, 2003.

[4] Quené, H., "On the just noticeable difference for tempo in speech", Manuscript.

[5] Quené, H., "Modeling of between-speaker and within-speaker variation in spontaneous speech tempo", *INTERSPEECH-2005*, 2457-2460, 2005.

[6] Riggenbach, H., "Toward an understanding of fluency: a microanalysis of nonnative speaker conversations," *Discourse Processes*, 14, 423-441, 1991.

[7] Nakatani, L. H., O'Connor, J. D., and Aston, C. H., "Prosodic aspects of American English speech rhythm", *Phonetica*, 38, 84-106, 1981.

[8] Oller, D. K., "The effect of position in utterance on speech segment duration in English", *JASA*, 54, 1235-1247, 1973.

[9] Byrd, D. and Saltzman, E., "Intragestural dynamics of multiple phrasal boundaries", *Journal of Phonetics, 26, 173-199*, 1998

[10] White, L. S., *English speech timing: a domain and locus approach*, University of Edinburgh PhD dissertation, 2002.

[11] Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E., "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender", *Language and Speech*, 44, 123-147, 2001.

[12] Liberman, M., "Young men talk like old women," http://itre.cis.upenn.edu/~myl/languagelog/archives/002629.html, 2005.

[13] ICSI transcripts of English Switchboard (01/29/03), http://www.cavs.msstate.edu/hse/ies/projects/switchboard/.

[14] Yuan, J. and Jurafsky, D., "Detection of questions in Chinese conversational speech", *ASRU-2005*, 2005.

[15] Nooteboom, S. G. and Eefting, W., "Evidence for the adaptive nature of speech on the phrase level and below", *Phonetica*, 51, 92–98, 1994.

[16] Ueyama, M., "Phrase-Final Lengthening and Stress-Timed Shortening in the Speech of Native Speakers and Japanese Learners of English", *ICSLP-96*, 610-613, 1996.

[17] Ofuka, E., McKeown, J. D., Waterman, M. G., and Roach, P. J., "Prosodic cue for rated politeness in Japanese Speech", *Speech Communication*, 32, 199-217, 2000.