A Forensic Aspect of Articulation Rate Variation in Chinese

Cao Honglin[1,2], *Wang Yingli*[3]

1 Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China;
2Dept. of Chinese Language and Literature, Peking University, Beijing, China;
3Center of Criminal Technology, Public Security Bureau of Guangdong Province, China

ABSTRACT

This study presents the statistical data for the articulation rate (AR) of 101 male Chinese speakers. 100 spontaneous telephone speech samples produced by 100 speakers and 10 samples produced by another speaker are investigated to test the inter- and intra-speaker variation of AR respectively. Two separate histograms for the global AR and the mean AR are shown to be near normal distribution. It is found that the range of AR for the one speaker is small and relatively stable when the topic and style are similar. The global AR and mean AR can be used as discriminatory features for forensic speaker identification.

Keywords: Speaker identification, articulation rate, Chinese language, spontaneous speech

1. INTRODUCTION

Speech tempo is one of the prosodic features, which can be exhibited by two methods, one is speaking rate/speech rate/syllable rate (all terms can be abbreviated to SR), and another is articulation rate (AR). Both SR and AR can be defined as "the number of output units per unit of time" [1] (e.g., syllables per second). The biggest difference between SR and AR is that the former includes pause intervals but the latter does not [1-9].

The previous studies of speech tempo showed that AR had more speaker-discriminating power than SR in English [2] and in German [6-7]. When calculating AR, one important issue is how to deal with pause. It is known that pause basically can be silent/unfilled and filled (such as um and uh in English). However, the specific methods of different investigators are not the same. All studies in [1-9] exclude silent pauses, and all except Laver [4] and Cao [8] exclude filled pauses as well. In forensic studies, Künzel [6] proposed a formula to calculate AR, which was "number of syllables/ [duration – combined duration of all pauses]". More recently, Jessen [7] refined the steps and criteria of the measurement of AR in German and made some rather persuasive conclusions.

Although the AR parameter is found to be powerful in forensic speaker identification in English and German, similar studies on Chinese are rare. The present study focuses on the AR variation of Chinese speakers and aims to provide some useful statistical data by using the method proposed in [7].

2. METHOD

2.1 Speech material

Considering that most of the forensic-phonetic casework relates to telephone recordings (TRs), a database named *FTRD 2010* was compiled at Peking University, which included a number of spontaneous TRs in the daily work. 100 different TRs of 100 male speakers (M1-100) and 10 different TRs of one male speaker (M101) were selected from the *FTRD 2010* database for evaluating the inter- and intra-speaker variation of AR respectively. The 10 TRs (all being talks with judges about legal cases) from M101 were similar in style.

The topics of the TRs were about the discussion of forensic cases in conversational style. All TRs were spoken in Mandarin Chinese with no evident regional features. The age of the 100 different speakers ranged from 22 to 55, according to a preliminary survey. And the speaker M101 was 29 years old. The speakers consisted of forensic scientists, judges, police officers, lawyers, interested parties and lab workers.

Each individual speaker's speech was selected and saved

as a single wav file through the Adobe Audition 3.0 software, i.e. the speech of irrelevant speaker (e.g. a female lab worker) was excluded. The durations of the final speech samples of speakers M1-100 and the speaker M101 were on average 51s (with standard deviation (SD) of 14s and range from 20s to 82s) and 39s (with SD of 12s and range from 26s to 57s) respectively.

2.2 Measurement

To get the AR data, three important issues have to be clarified. First, which linguistic unit should be counted? This is an easier question for the present study, because each Chinese character is concurrently one syllable [10]. So the AR will be measured in terms of monosyllabic Chinese characters per second in the present study.

The second issue is about the method and criteria for measuring of AR. We have followed Jessen [7]: The realized syllables, not canonical ones were counted. The size of speech intervals [1] were selected by the investigator's short-term memory to choose the number of syllables easily (i.e., the investigator goes through the speech signal and selects portions of fluent speech containing a certain number of syllables that can easily be retained in short-term memory.). Each memory selected stretch consisted of only fluent speech, excluding silent pauses, filled pauses, laughter sounds and any immoderate syllable lengthening. In order to minimize increasing the phrase-final lengthening effect of very short utterance on AR, the lowest number of syllables per stretch was set to be no less than four.

Third, both AR for the entire recording, which was called "global AR (GAR)" [7] and AR for each selected speech stretch, which was called "local AR (LAR)" [7] was to be calculated. To get GAR of one speech sample, the total duration of its all selected stretches was divided by the total number of syllables of all selected stretches. And LAR was calculated by the number of syllables dividing by the duration for each stretch.

Both the number of syllables and the duration of each stretch were extracted from the "TextGrid" file generated by the Praat (version 5143) [11]. The procedure is illustrated in Figure1.

The numbers of memory stretches counted were on average 30 for M1-100 (with SD of 8.5 and range from 12 to 54), and 26 for M101 (with SD of 7.7 and range from of 18 to 41). The number of syllables per stretch for M1-100 was range from of 4 to 22 and on average 7.8.

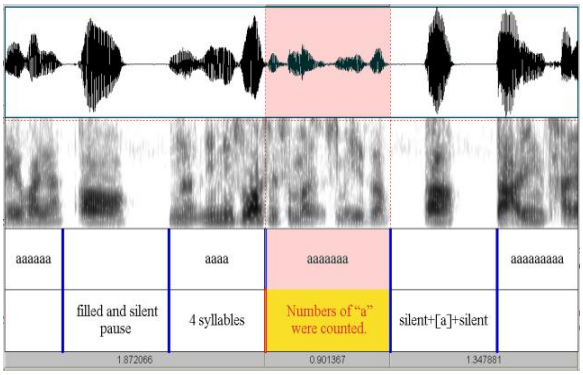

Figure 1: The annotation procedure for each memory stretch. One letter "a" stands for one syllable.

3. RESULTS AND DISCUSSION

The global articulation rate (GAR) value for the 101 speakers (M1-100 and M101) and the mean articulation rate value across memory stretches for each speech sample (LARmean) were calculated. In Figure 2-3, results are shown in form of histograms that stand for how many speakers lie within a particular interval of GAR and LARmean values.

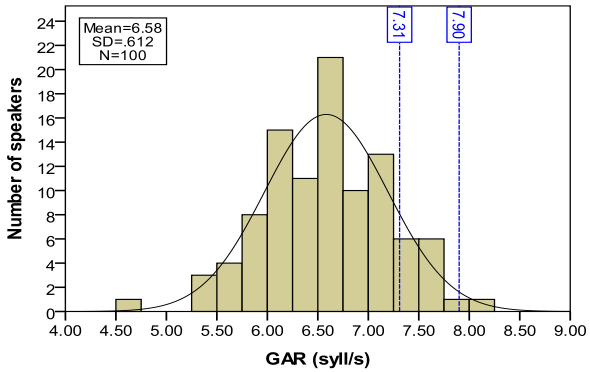

Figure 2: Histogram for GAR parameter. The blue dashed lines stand for the range of GAR values of M101 (see below).

Illustrated in Figure 2, the statistical results of the parameter GAR values of speakers M1-100 show an approximate normal distribution. For example, the values from 6.50 to 6.75 syll/s are found in many more speakers (21, 21% of

[1] As recommended by one reviewer, two possible types of speech intervals can be chosen for calculating AR, which are the "inter-pause intervals" and the "intonation phrase". However, they are "not without empirical or methodological problems"; the present method is simpler and more pragmatic (for more details see [7]).

100 speakers) than values at the lowest and highest margins of the distribution (both found in only one speaker, 1% of 100 speakers). This result provides valuable reference data for Chinese population statistics in forensic casework. Based on this statistical result, the GAR parameter does not successfully discriminate some of the speakers with GAR values in the central area. However, for those speakers who strongly deviate from the central trend, the GAR becomes a salient discriminatory parameter.

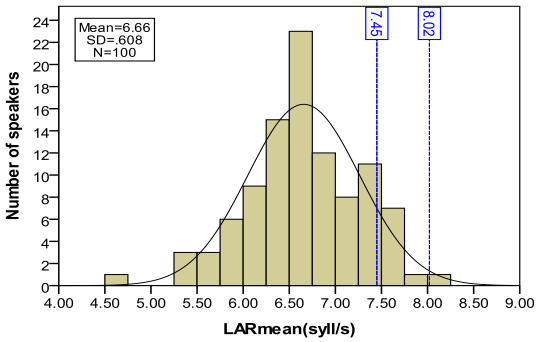

Figure 3: Histogram for LARmean parameter. The blue dashed lines stand for the range of LARmean values of M101 (see below).

Figure 3 shows that the result of the LARmean values of M1-100 also form a good approximation of a normal distribution. Not surprisingly, the distributions of GAR values and LARmean values are similar. Comparatively, 23 speakers are found in the center of the LARmean distribution (values from 6.50 to 6.75 syll/s). One speaker appears at the margin of each distribution. However, the two distributions are not exactly the same. Across the 100 speakers, mean values of GAR and LARmean are 6.58 syll/s and 6.66 syll/s respectively. The former is a little lower. After examining the two groups of data, an interesting result is found.

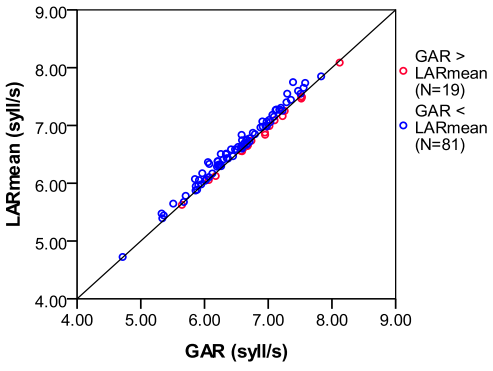

Figure 4: Scatterplot for GAR and LARmean values.

As shown in Figure 4, 19 speakers' GAR values are higher than their LARmean values (all difference between them are less than 0.07 syll/s), whereas the other 81 speakers' GAR values are lower than their LARmean values (the difference range from 0.00 to 0.36 syll/s with an average of 0.10 syll/s). This explains why the two distributions are different, e.g. in the range from 6.00 to 6.25 syll/s, the numbers of speakers are 15 and 9 in figure 2 and 3 respectively. Pearson correlations are run in order to determine whether and to what extent the GAR values and the LARmean values correlate with each other. A significant positive correlation was found (r=0.990, p=0.01) between the two AR values. However, given the difference between the two different calculating methods, it is unwise to mix them in one case at the same time, and the GAR values and LARmean values should not be compared.

Table 1: Literature on AR (syll/s) of male speakers in three languages (L). G – German, E – English, C – Chinese, Spont – spontaneous.

Study	Men	AR mean	L	Speech
[5]	27	5.74	G	Reading
[6]	5	5.89	G	Spont talk
[7]	100	5.19	G	Spont phone
[9]	47	5.2	E	Informal talk
[8]	2	5.65 [2]	C	Reading
present	100	6.58/6.66	C	Spont phone

The data in the previous studies of AR for German, English and Chinese are presented in Table 1. The number of male speakers in the present study is more than [5,6,8,9], excepted for [7]. The major difference among these studies in the literature lies in the average value of AR (GAR or LARmean). The present results are the highest in all studies. The difference may be caused by factors such as number of speakers, age of speakers, calculating method, language and speech style. Interestingly, as we follow the method in [7], the differences between [7] and the present results are still significant (both in the average values and the whole distribution). Compared with German and English, Chinese syllable structure is simple. The maximal Chinese syllable construction is #CVVC/V# and there are no consonant clusters [10]. Comparatively, a syllable in English/German can contain up to three consonants at the beginning, as in *stray/strick*, and

[2] The value 5.65 syll/s is the average value of the two speakers' AR values (5.3 syll/s for male 1 and 6.0 syll/s for male 2) in Cao [8].

up to four consonants at the end, as in *glimpsed/herbst* [10,12]. As pointed out in [7], "speakers of a language with simple syllable structure are expected to produce more syllables … than speakers of a language with more complex syllable structures, hence show higher AR", we can infer that except for Cao [8], language may be the most critical factor. As it is known that the syllable structures of Japanese and many African languages, like Yoruba, are less complex than Chinese, to find whether or not AR data of these languages speakers will be lower than the present result, more researches are needed. The low AR value of Cao [8] is easily explained since only silent pauses were excluded when the AR was measured and the subjects were confined to two male speakers.

Table 2 lists GAR and LARmean values of M101's 10 different speech samples. The minimal and maximal values of GAR and LARmean are illustrated in figure 2 and 3 (see the blue dashed lines) respectively. The ranges of the GAR and LARmean values are both relatively centralized and both occupy the position near the top margin of the two distributions. Because the topics and styles of the 10 speech samples are similar, the intra-speaker variations of GAR and LARmean are relatively stable and (in this case) smaller than the inter-speaker variations shown in figure 2 and 3. The intra-speaker variation may be larger if the styles of the speech samples differ. Forensically, when AR parameters are used, it is critical to get the most stylistically similar speech samples (including other possible factors, such as emotional factors), compared with the unknown samples.

Table 2: A list of GAR and LARmean values of M101's 10 different speech samples.

Number	GAR	LARmean
1	7.52	7.52
2	**7.31**	**7.45**
3	7.87	7.96
4	7.55	7.73
5	7.43	7.76
6	**7.90**	**8.02**
7	7.55	7.66
8	7.86	7.80
9	7.85	7.69
10	7.100	7.80
Mean	7.74	7.64
SD	0.17	0.21

CONCLUSIONS

This study provides valuable population statistics on Chinese speakers' articulation rates. Two histograms are shown for the articulation rates (GAR and LARmean) of 100 male Chinese speakers, which show approximate normal distributions. Our findings are not very similar with previous data in German or English, presumably because the syllable structure in Chinese is simpler than that in German and English. Both GAR and LARmean parameters, which are significantly correlated, can discriminate individual speakers. However in the present study, it is hard to estimate which one is more powerful in discriminating individuals. 10 different spontaneous speech samples of one speaker, of which the topics and styles are similar, were investigated. The results show that the intra-speaker variation of AR is relatively stable and lower than the inter-speaker variation. In forensic casework the investigators should pay attention to the possible mismatch in stylistic factors, which may cause high intra-speaker variation. Since more variables have to be included as shown in [9], this study is also a platform for further investigation.

4. ACKNOWLEDGEMENTS

This research was funded by the China National Natural Science Funds (Grant 61073085). We thank two anonymous reviewers for the comments on the earlier version of the paper. And we also thank Michael Jessen for providing us the article [6].

5. REFERENCES

[1] Tsao, Y.C., Weismer, G., Iqbal, K. 2006. Interspeaker variation in habitual speaking rate: Additional evidence. *JSLHR*, 49, 1156–1164.

[2] Goldman-Eisler, F. 1968. *Psycholinguistics. Experiments in Spontaneous Speech*, Academic Press, London

[3] Miller, J.L., Grosjean, F., Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41, 215–225.

[4] Laver, J. 1994. *Principles of Phonetics*, Cambridge University Press: Cambridge.

[5] Künzel, H.J., Masthof, H.R., Köster, J.P. 1995. The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition, *Science and Justice*, 35, 291–295.

[6] Künzel, H.J. 1997. Some general phonetic and forensic aspects of speaking tempo, *Forensic Linguistics*. 4, 48–83.

[7] Jessen, M. 2007. Forensic reference data on articulation rate in German. *Science and Justice*, 47, 50-67.

[8] Cao, J.F. 2003. Articulation rate and its variations (in Chinese), *Proceedings of the 6th National Conference on Modern Phonetics*, Tianjin, 143-148.

[9] Jacewicz, E., Fox, R.A., O'Neill. C., Salmons, J. 2009. Articulation rate across dialect, age, and gender, *Language Variation and Change*, 21, 233–256.

[10] Wang, H.J. 2008. *Chinese Non-Linear Phonology* (in Chinese), Peking University Press, Beijing.

[11] http://www.fon.hum.uva.nl/praat/ download_win.html

[12] Fox, A. 2005. *The structure of German*. 2nd edition. Oxford University Press, Oxford.