Prosody of Dialogues: Influence of Recognition Failure on Local Speech Rate

Olga I. Dioubina

Department of Phonetics and Speech Communication
University of Munich, Germany
olga.dioubina@phonetik.uni-muenchen.de

Abstract

The variation in speech rate influences the performance of the automatic information-providing devices and leads to the recognition failures in the process of man-machine communication. While this problem has been generally recognized, there are few studies that provide detailed information on the variation in the speech rate in a specific context. Our research focuses on the variation that occurs in a clearly defined context of recognition failure, which evokes an abrupt change in verbal (and non-verbal) behavior on part of the subjects. In addition, the method we applied in measuring the speech rate improves the reliability of the results.

1. Introduction

Synthetic speech of the automatic information-providing devices should master temporal organization of the human speech. For example, the simulation of speech rhythm requires a realistic imitation of the manner in which speakers organize their information flow in time, e.g. the word grouping strategy [11]. As for the speech rate, it would be good to know in which communication situations during man-machine interaction the speakers use to talk slower or faster, and what segmental and prosodic phenomena take place when the speech rate is being changed. Another approach is to ask human speakers to produce some speech material at different speech rates and then to investigate, for example, the duration of speech units at both rates. The initial durational model for French in [11] was conceived on the basis of the analysis of read sentences produced at two speech rates by the same speaker. In [10] the effect of speech rate on the temporal organization of syllable production of Hong Kong Cantonese was investigated on the base of meaningful text syllables of different syllable structures recorded at a normal conversational rate and at a fast rate. Our research focuses on the variation of speech rate in German spontaneous speech that occurs in a clearly defined context of recognition failure, which evokes an abrupt change in verbal (and non-verbal) behavior on part of the subjects.

Constant speech rate in an interaction course between two or more humans is an idealization. The speakers vary their speech rate with respect to the changes in their own paralinguistic state, the paralinguistic state of the listener(s) and with respect to any modification which might be caused by environmental factors such as time, place and circumstances of interaction. While assuming that there is a continuous speech rate variation for one speaker, it does not seem to upset the capability of a listener to assess the overall (or global) speech rate of the speaker, e. g. to name it slow, moderate or fast (overall speech rate). However, the speech rate variation for one speaker can be shown by comparing two or more short speech strings with each other (local speech rate).

For the sake of the spontaneous character of the interaction, we should not manipulate the changes in the paralinguistic state of the speaker. However, we can predefine the communication partner and the setting: (i) Putting an automatic system into the place of the listener will limit the paralinguistic state of the listener to the set of predefined verbal reactions; (ii) Defining place, time and circumstances of interaction will restrict environmental factors to the predefined setting which would be similar for all speakers.

In such predefined setting of man-machine interaction we observed a general tendency among 52 human subjects to speak slower as soon as a particular predefined verbal reaction of the machine, namely recognition failure, occured. This observation of rather general character was thereafter proved by comparing short speech strings articulated immediately before the recognition failure with speech strings of equal duration articulated immediately after the recognition failure. The results of the acoustical analysis and perception experiment supported our preliminary observation of speech rate slowing-down-effect after recognition failure.

We believe that the method we applied for the phonetical analysis of the speech rate contributes to the research on this topic by submitting more reliable information on the grade of speech rate variation. We have applied three different procedures of measuring speech rate in our data. On the one hand, we applied two methods used in numerous studies dedicated to the variation in "speaking style," namely, computing the speech rate in the number of phones and syllables per time unit ([1], [9]). On the other hand, we applied another method of measuring the mean perceptual speech rate as linear combination of the syllable and the phone rate, which has been developed recently in an attempt to find a more adequate way of describing this phenomenon.

2. Recognition failure in man-machine interaction

Spoken dialogue systems often indicate their failure to recognize the verbal input on part of the user. The users seek to resume the interaction. To achieve it, the users try a variety of communication strategies. Most frequently, they switch to a prosodically "marked" speaking style, which is characterized by hyper-articulation of the sounds, and by the decrease in the speech rate ([6], [4], [5]). While these strategies are usually very effective in the context of the human-human interaction, in the context of the man-machine dialogues they lead to some additional problems in automatic speech recognition ([9]).

2.1. Speech corpus of SmartKom

The speech data we used for your research, were recorded in the so called Wizard-of-Oz experiments. Wizard-of-Oz (WOZ) is a technical term for simulated man-machine interaction. "Simulated" means that in fact interaction takes place between two human subjects though one of them thinks to be talking to the fully automated system. The "wizards" who play the role of the automated system talk in a computer-like manner, use predefined vocabulary, their voice is manipulated, and the subjects can not see them. Each dialogue consists of ca. 10-40 turns in the alternative succession of the user and the system turns. As much as 450 WOZ dialogues were recorded at the University of Munich in 2000–2003 as part of the SmartKom project [3].

2.2. Speech data

2.2.1. Context of recognition failure

One of the predefined utterances of the "wizard" indicated that a recognition failure had been taken place. The predefined utterance of the "wizard" can be translated from German as follows: "I did not understand you. Try to repeat or say it in other words. I can understand gestures as well."

If the recognition failure happened for the first time in the otherwise fluent dialogue, then two turns of the users, one turn before and one turn after predefined utterance of the "wizard", were extracted from the audio recordings for the purpose of our research.

The reason why we selected only the first recognition failure per dialogue was in conformity with the focus of our research to examine most abrupt changes in verbal behavior on part of the users in a clearly defined specific context.

2.2.2. Selection of data

In total 52 dialogues contained recognition failure which was introduced by the above mentioned predefined utterance of the "wizard" and took place for the first time in the dialogue. The subjects in these dialogues were 29 female and 23 male Germans aged from 12 to 64 with the mean age of 27.5. Most of them were students from Bavaria, the place where Wizard-of-Oz experiments of SmartKom project took place.

Our speech data consisted of 104 one-second speech strings without pauses which were extracted from the very beginning of the turn before and after the recognition failure.

3. Acoustical analysis of speech rate

3.1. Definition of speech rate

As we have already mentioned in the Introduction, we have applied three different procedures of measuring speech rate in our data. On the one hand, we applied two methods used in numerous studies dedicated to the variation in "speaking style," namely, computing the speech rate in the number of phones and syllables per time unit ([1], [9]). On the other hand, we applied another method of measuring the mean perceptual speech rate as linear combination of the syllable and the phone rate, which has been developed recently in an attempt to find a more adequate way of describing this phenomenon.

Campbell [2] believed that simple measures of speech rate, such as "syllables per second" or even "words per second" are inadequate to describe local changes in rate, due to the effects of differences in the structure of words and syllables. For example measuring speech in "syllables per second" will be affected by the structure of syllables: a string of simple units will yield a higher rate than the same number of more complex one.

Pfitzinger [7], [8] suggests that neither the syllable nor the phone rate on its own represents the speech rate adequately. Pfitzinger carried out a perception experiment in which the subjects were instructed to compare and assess the speech rate of short speech signals. Based on the results of the perception experiment he suggested a model for determining speech rate from a linear combination of syllable and phone rate. He [8] demonstrates that linear correlation $r \approx 0.6$ between local (window width of analysis 625 ms) syllable rate and local phone rate indicates that the information which phone rate and syllable rate carries about the speech rate is different. Therefore a combination of both factors should be considered during the measurement of the speech rate. In [7] a local speech rate measurement method based on linear combination of the local syllable rate and the local phone rate was suggested. This method applies a "momentary" or local approach to acoustic speech rate measurements and correlates well with perceived local speech rate ($r = 0.91$). The formula for estimating local perceptual speech rate for a 625 ms window width is as follows:

$$8.6 * syllable\ rate + 3.6 * phone\ rate - 0.2(\%) \quad (1)$$

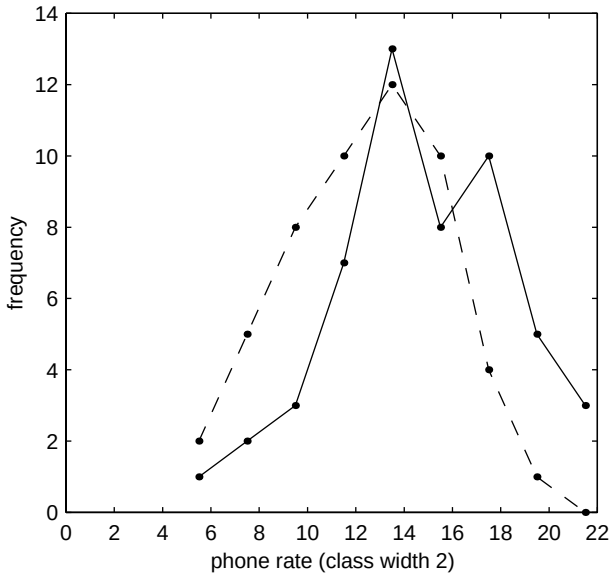

Figure 1: *Phone rate on the frequency plot: solid line illustrates the phone rate before the recognition failure, dashed line – after the recognition failure.*

3.2. Annotation procedure

Our speech data were 104 one-second speech signals without pauses which were extracted from the very beginning of the turn before and after recognition failure. They were manually segmented into phones and syllables by a trained phonetician.

There were several reasons why we measured just one second of speech from the beginning of the turns before and after recognition failure: (i) speech strings of 1s among 52 users: manual segmentation is a reliable but time consuming process, at the same time it was important to measure the variation in the speech rate of as many different users as possible, so that we could statistically prove the general tendency to slow down the speech rate immediately after the recognition failure; (ii) first second of speech in the turns: we were interested in providing the information on the extreme variation of the speech rate, e.g. the abrupt change from the normal speech rate of users to the slower speech rate immediately after an unexpected communication problem; (iii) 1s window of analysis instead of 625 ms (see more in 3.1): one second is a convenient unit to handle and it stays in agreement with our attitude to the assessment of speech rate, e. g. combination of local and overall speech rate in a speech string.

3.3. Estimation of speech rate

3.3.1. Phone and syllable rate

The results of computing phone and syllable rate in the 52 pairs of one-second speech strings are shown in the Table 1. The difference between the mean number of phones before and after recognition failure is 2.36 phones per second which supports the hypothesis that there is a general tendency to decrease the phone rate after the recognition failure. The difference between the mean number of syllables before and after recognition failure is 0.9 syllables per second which also supports the decrease in the syllable rate after recognition failure.

We plotted the results on phone and syllable rate on the frequency plots (Figure 1 and Figure 2, class width – 2) to illustrate the distribution of the number of phones and syllables, e.g. how many phones and syllables are more or less frequently articulated during one second before and after recognition fail-

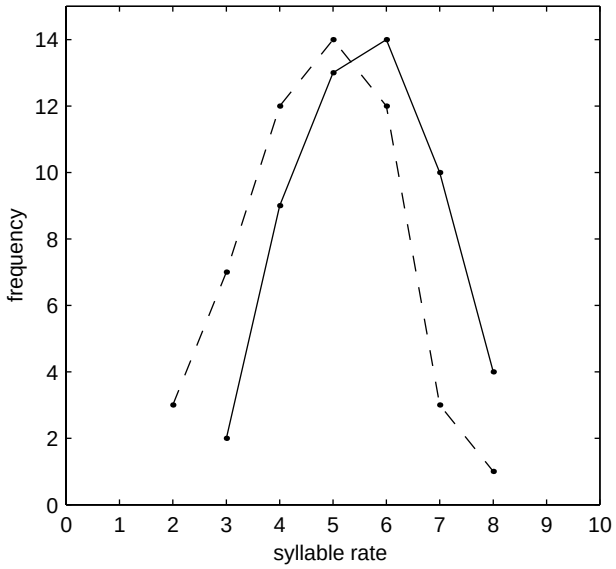

Figure 2: *Syllable rate on the frequency plot: solid line illustrates syllable rate before, dashed line after recognition failure.*

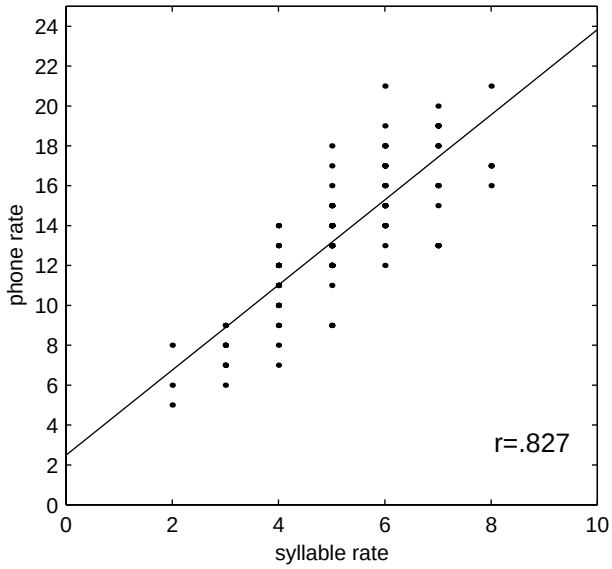

Figure 3: *Linear correlation phone vs. syllable rate. The resulting coefficient is r = 0.827.*

ure. The frequency plot of the phone rate illustrates that the most frequent number of phones per second lies in both cases between 12 and 14 phones per second. The main difference between two curves could be seen before and after the common peak on the plot. The dashed line which illustrates the distribution of phones after recognition failure rises and falls earlier than the solid line illustrating the distribution of phones before the recognition failure.

A slightly different development of the curves could be seen on the frequency plot 2 illustrating distribution of syllables before and after recognition failure. There are two clear peaks on the plot demonstrating that the most frequent number of syllables before recognition failure was 6 syllables per second (solid line) and after recognition failure – 5 syllables per second. The both curves rise and fall in a similar way.

The results of the paired t-test for phone rate (t = 3.493, α = 0.001) and syllable rate (t=3.432, α = 0.001) both demonstrated a statistically significant decrease in phone and syllable rate from the turn before to the turn after the recognition failure.

We measured linear correlation coefficients of phone and syllable rate before recognition failure vs. phone and syllable rate after recognition failure: the coefficient of $r = 0.778$ of phone rate vs. syllable rate before recognition failure was lower than the linear correlation coefficient $r = 0.837$ of phone rate vs. the syllable rate after the recognition failure. However, the difference between linear correlation coefficients was proved to be statistically not significant.

Since we found no significant difference between the linear correlation coefficients before and after recognition failure, we then calculated linear correlation for all 104 speech strings. The scatter plot (Figure 3) illustrates high positive correlation of phone vs. syllable rate ($r = 0.827$): high scores on the X-axis are associated with high scores on the Y-axis.

The resulting coefficient ($r = 0.827$) is higher than the one indicated by Pfitzinger [7] ($r \approx 0.6$). We believe that the difference between our results and those of Pfitzinger could be explained by larger amount of experimental data he used.

Table 1: *Phone and syllable rate before and after recognition failure. The mean for phone (14.73) and syllable (5.63) rate are higher than those after recognition failure (phone rate – 12.37, syllable rate – 4.73).*

	mean	max.	min.	stand.dev.
phone rate (before)	14.73	21	6	3.55
phone rate (after)	12.37	19	5	3.34
syllable rate (before)	5.63	8	3	1.29
syllable rate (after)	4.73	8	2	1.37

3.3.2. Mean perceptual speech rate

In section 3.1 we introduced the method of combining phone and syllable rate for measuring the speech rate. With respect to the formula (1) we measured the mean perceptual phone and syllable rate for one-second speech strings before and after recognition failure. The results of our measurement are shown in the Table 2.

Table 2: *Data on mean perceptual speech rate. The decrease from 101.29 to 85.00 suggests that the speech rate after the recognition failure was slower.*

	before recognition failure	after recognition failure
mean	101.29	85.00
variance	511.26	521.31
standard deviation	22.61	22.83

The difference between the mean speech rate before the recognition failure (101.29) and the mean speech rate after the recognition failure (85.00) was 16.29 units per second.

The results of the paired t-test (t = 3.655, α = 0.001) supported statistically high significance of the decrease in the speech rate after the recognition failure.

4. Perception experiment

A forced-choice perception test was carried out in order to verify the results of the acoustical analysis such as that immediately after the first recognition failure in the dialogue the speech rate is slower than immediately before recognition failure.

4.1. Experimental procedure

4.1.1. Subjects, stimuli and procedure

A total of twenty subjects, 10 male and 10 female, aged 25-45, took part in the perception experiment. All of them were either students of phonetics, phoneticians or technicians at the Department of Phonetics, University of Munich, Germany.

The stimuli were presented in 52 pairs. Each pair included a speech string of 1s duration before and after recognition failure. To assess the difference in the speech rate, the subjects had to press two buttons on the computer-aided user interface, and the stimuli of one pair were reproduced. The subjects had to make a forced-choice decision on whether they could hear the difference between two stimuli or not. In case they could perceive the difference between two speech strings of the pairs then the pairs were split into two groups, the "faster" and the "slower" group. If the subjects were not able to state any difference between the signals of one pair, the both buttons representing one pair were pulled into the group "no difference".

4.1.2. Preliminary results

Out of 1040 possible judgments (20 subjects x 52 pairs) 66,54% (692) of all judgments split the pairs into faster and slower speech strings. There was a significant difference between the judgments of phonetically trained subjects (73% of pairs were judged as being articulated with different speech rate) and of technicians (only 60% of pairs were judged as being articulated with different speech rate).

The table 3 illustrates the relationship between the decrease in phone and syllable rate and the decrease in the number of votes for the decision that before recognition failure the speech rate was faster than after recognition failure.

The preliminary results show that the difference in the phone (5.54) and syllable (2.46) rate is required to achieve an agreement among 18–20 subjects who had to assess the difference in the speech rate between two speech signals.

Table 3: *Data on perception test. The column "Votes" indicates the number of votes given to those pairs of speech strings where the string before recognition failure was perceived as articulated faster than the string after it. The columns "phone rate" and "syllable rate" indicate the increase/decrease in phone and syllable rate required to undertake the decision on the speech rate.*

Votes	No. of pairs	phone rate	syllable rate
18–20	13	+5.54	+2.46
15–17	9	+3.67	+1.33
12–14	3	+2.67	+1.33
9–11	5	+0.20	+0.60
6–8	6	+2.17	+0.67
3–5	5	+0.20	-0.40
0–2	11	-0.82	-0.91

5. Discussion

The main finding of this paper is that there is a general tendency to decrease the speech rate during the first second of the speech string that follows the first recognition failure in a man-machine dialogue. This finding is supported by both our acoustical analysis and preliminary results of the perception experiment.

By applying the hypothesis to our data we can see that a mean decrease of 2.36 phones and 0.9 syllables per second characterizes the extent of the speech rate variation in a man-machine interaction. As for the mean perceptual speech rate which we measured following the method of Pfitzinger there is a speech rate variation of 16.29 units.

The preliminary results of our perception experiment show that in fact a larger difference in the phone (5.54) and syllable (2.46) rate is required to achieve the agreement among 18–20 subjects.

The variation in speech rate influences the performance of the automatic information-providing devices and leads to the recognition failures in the process of man-machine communication.

To avoid or to limit the number of recognition failures in speech recognition, the automatic information-providing devices should know how to capture temporal organization of the human speech, e.g. variations in speech rate. In order to be able to describe the variations in speech rate we would need to consider (i) the extent of speech rate variation which could be observed for a prototypical speaker, (ii) situations in which extreme speech rate variation could be observed, e.g. situation of the recognition failure, and at last (iii) perceptional capability of a human listener to hear the variation in speech rate in a way which could be different from the information in the acoustics. Out research is an attempt to receive some preliminary answers to these considerations.

6. Acknowledgments

This research was partly done within the framework of the SmartKom project (grant no. 01 IL 905). I am very grateful to Hartmut Pfitzinger, Parham Mokhtari, Christoph Draxler, Uwe Reichel, Michiko Inoue, Silke Steininger and Ulrich Reubold for discussions and remarks on this paper.

7. References

[1] Bell, L.; Gustafson, J., 1999, Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer directed speech, *The XIVth International Congress of Phonetic Sciences (ICPhS)*, San Francisco, 2:1221–1224.

[2] Campbell, W. N., 1988, Speech-rate variation and the prediction of duration, *The XIIth International Conference on Computational Linguistics (COLING)*, Budapest, 93–95.

[3] Dioubina, O. I., 2003, Annotation of expressive speech, *ISCA Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, Geneva, 173–177.

[4] Krahmer, E.; Swerts, M.; Theune, M.; Weegels, M, 2002, The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates, *Speech Communication* 36, 133–145.

[5] Levow, G.-A., 2002, Adaptations in spoken corrections: Implications for models of conversational speech, *Speech Communication* 36, 147–163.

[6] Oviatt, S. L.; MacEachern, M; Levow, G.-A., 1998, Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication* 24, 87–110.

[7] Pfitzinger, H. R., 1998, Local speech rate as a combination of syllable and phone rate, *The Vth International Conference on Spoken Language Processing (ICSLP)*, Sydney, 3:1087–1090.

[8] Pfitzinger, H. R., 1999, Local speech rate perception in german speech, *The XIVth International Congress of Phonetic Sciences (ICPhS)*, San Francisco, 2:893–896.

[9] Swerts, M.; Litman, D.; Hirschberg, J., 2000, Corrections in spoken dialogue systems, *The VIth International Conference on Spoken Language Processing (ICSLP)*, Beijing, 2:615–618.

[10] Zee, E, 2002, The Effect of Speech Rate on the Temporal Organization of Syllable Production in Cantonese, *Speech Prosody 2002*, Aix-en-Provence.

[11] Zellner Keller, B., 2002, Revisiting the Status of Speech Rhythm, *Speech Prosody 2002*, Aix-en-Provence, 727–730.