An exploratory study on how the use of general lexical and linguistics information helps to predict the dynamic of speech rate in dyadic conversations

Simone Fuscone
Aix-Marseille Université
LPL, Aix-en-Provence, France
LIS, Marseille, France

Laurent Prévot
Aix-Marseille Université
CNRS
LPL, Aix-en-Provence, France

Benoit Favre
Aix-Marseille Université
CNRS
LIS, Marseille, France

1 Introduction

Understanding how linguistics features, produced by speakers in a dyadic conversation, evolve is a challenging task due to the not well-known relation between the conversants. As known from the literature, speakers tend to change their speech style during conversations. In particular, Giles and Coupland (1991) explains that people could accommodate their speech style with respect to their interlocutors at different levels (lexical and syntactic) according to the *Accomodation Theory*. Street and Giles (1982), Giles and Powesland (1975), Giles and Howard (1980) have proposed that conversation participants respond to one another's speech, including speech rate. Such a dynamics is potentially affected by many parameters making its study a difficult task. Previous works presented models to explain how speakers adapt and change their style, focusing mainly on the analysis of one feature. In this exploratory work we attempted to answer whether it is possible to predict the changes of speakers speech rate in the *second part* of the conversation, using features like extra linguistics variables associated to each participant and averaged features extracted from time-aligned transcripts from the *first part*. The choice of use speech rate as target variable is due to the proof of previous studies that underline a correlation between the speech rate and some properties of the speech style. As argued by Hannah and Murachver (1999), Kendall (2009)) and Babel (2012) speech rate could be influenced by the sex and age of the speakers, or by the topic of the conversation. As Goldman-Eisler (1956) outlined the number of pauses and the duration of each participant influence the production of speech rate. Newman and Smit (1989) studied the effect of turn latencies on the speech rate in Children-adult conversations, finding that the change of the time latency affects the speech rate in children. Yang (2003) in his study asses the strict correlation within the speech rate and the internal pause of the speakers. More globally there is a tension between the intuition that conversational dynamics, or adaptation of some kind, largely unconscious, to the interlocutor speaking and interacting style, as supported by various experimental studies (Babel, 2012), and the difficulty of actually finding strong effects of such phenomena in corpora as exemplified for example from recent negative results of (Weise and Levitan, 2018). Our objective is to more systematically scrutinize the variables involved in characterizing speaking and interacting style on large corpora and decipher their cross-speaker dynamics thanks to advanced machine learning techniques. As a first step, we focus on one variable, using speech rate as target variable. We tried to determine whether the use of linguistics features of both speakers led to a better performance than use features of just one speaker. No acoustic, prosodic or phonetic features were used at this stage. We present two precursory tasks in the study of the dynamic: (i) predict the speech rate evolution (`decrease / increase / no change`) of one speaker in the second half of the conversation; and (ii) the difference of speech rate of the two speakers respectively increase, decrease or remains stable also in the second half of the conversation. For this purpose, we used transcripts data from the Switchboard corpus, splitting each conversation in two halves. To predict the behavior of speech rate, after experimenting with different learning methods and parameters we settled in using a Random Forest algorithm testing different subsets of input variables. Finally a bootstrap approximation method was applied to asses if there is a difference within the different subsets of variables we used in prediction task.

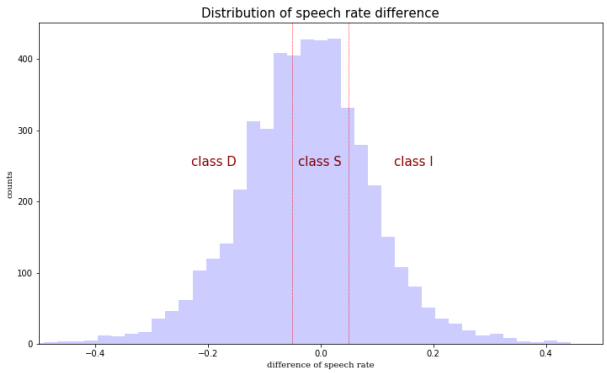

Figure 1: The picture shows the distribution of the difference speech rate δSP_1 for all the speakers of the corpus. The red lines delimit the three zones that correspond to the three target classes D (speech rate Decreases), I (speech rate Increases) and S (speech rate remains Stable).

2 Methods

We used the definition provided in the work of Cohen Priva et al. (2017) to compute speech rate, as the ratio between utterance duration and utterance expected duration. We split each conversation in two halves and take the average speech rate in the first and second half, denoted as $SR_1^{'}$, $SR_1^{''}$ for one speaker (that we call *speaker 1*) and respectively $SR_2^{'}$, $SR_2^{''}$ for the other speaker, (called *speaker 2*). We analyzed two types of target variables based on the previous definition of speech rate:

- difference speech rate of one speaker in the second and first half of the conversation, called $\delta SR_1 = SR_1^{''} - SR_1^{'}$
- difference of the two speakers speech rate in the second half of the conversation, $\Delta SR_{12}^{''} = SR_1^{''} - SR_2^{''}$

We labeled the two target variables into three classes called I (`increases`), D (`decreases`), S (`remains stable`) as stated by the following:

$$\begin{aligned} \text{Class D if} \quad & \delta SR_1 < -\epsilon \\ \text{Class I if} \quad & \delta SR_1 > +\epsilon \\ \text{Class S if} \quad & -\epsilon \leq \delta SR_1 \leq \epsilon \end{aligned}$$

where ϵ is a threshold, that is chosen in order to obtain a comparable number of example for the 3 classes. We repeated the same process for the variable $\Delta SR_{12}^{''}$.

The distribution of δSR_1 is shown in Figure 1.

For the input variables, as described by Bell (1984) we could divide the factors of the speech variation in two groups: *Linguistics features* and *Extra-Linguistics features*. In our study, linguistics features are all the variables linked to the speech style of the speakers or that describe the reciprocal interaction, extracted just from time-aligned transcripts since acoustic features haven't been taken into account at this stage. For **extra-Linguistics features (ELF)** we refer about information of the speakers like *Age, Sex, Level of study, Geographical place* and in addition the *topic of the conversation*. **Cross-speaker Linguistic Features (CLF)** take into account the relation between the speakers: *Cosine Similarity of discourse markers* (It is the distribution of discourse markers produced by the two speakers in the first half of the conversation. We use a short list of items for this purpose, selected for their frequency : $['hm','oh','right','uh','um','yeah']$. We use *Laplace Smoothing* to avoid zero count for an item.), *Total Overlap time of the conversation* (it is the duration during which the speech of the two speakers overlaps.), *Difference of Speech Rate* (Difference of speech rate within the speakers, in the first part of the conversation, $\Delta SR_{12}^{'}$). **Linguistics Features (LF)** refer to the linguistics style of one speaker: *Percentage of token* (it is computed as the ratio between the number of token produced by the speaker divided by the total number of token in the first half of conversation. It captures floor dominance); *Lexical Density* (As Johansson (2009) described, lexical density is the proportion of content words to

the total number of tokens); *% Stop Words* (the ratio between the Stop Words produced by the speaker and the total words produced in the first part); *Average Time duration of utterance* (it is the averaged time of the utterance of the speaker, dropping out silence, noise and laughter); *Number of significant turns* (is the total count of turns for each speaker that have a minimum duration of $2s$ and contains at least 3 *content words*); *Turn Latency* (it is the total time of the latency response. Street (1984) defined it as the pause between two consecutive turns belonged to different speakers. Basically is the time that a speaker occurs for answering to the turn of the other speaker); *Overlap time* (it is the duration during which a speaker overlaps his turn to the turn of the other speaker divided by the speaking duration of that speaker); *Discourse markers* (For each discourse markers like 'hm','oh', 'right', 'uh', 'um', 'yeah' we computed the total count produced by the speakers and use their relative frequencies as a singular variables).

We used transcripts from the Switchboard corpus (Godfrey et al., 1992), formed by participants that took part in multiple telephonic conversations. There are 543 speakers in the corpus, with about 2400 conversations. The averaged duration is 6 minutes. We dropped out conversation that were too short to compute the input and output variables. So after pre-processing we obtain 4864 "conversation sides". To predict the class (D, I or S) of the variables in the two tasks, we used a *Random Forest classifier* implemented in the Scikit Learn package (Buitinck et al., 2013). We tested different sets of features grouping the input variables described [1] as listed in Table 1.

Table 1: The table represents the Accuracy scores for the different sets of variables in the case of the target variable δSR_1.

Sets of Features	Acc. Test	Acc. Validation
$\mathbf{CLF + LF_1 + ELF_1^\star}$	$\mathbf{0.5068}$	0.4922 ± 0.0085
$\mathbf{LF_{12}}$	$\mathbf{0.5041}$	0.4913 ± 0.0078
$LF_{12} + ELF_1$	0.4959	0.4937 ± 0.0100
$LF_1 + ELF_2^\star$	0.4931	0.4937 ± 0.0092
$LF_1 + ELF_1^\star$	0.4931	0.4932 ± 0.0110
LF_1	0.4904	0.4905 ± 0.0043
ELF_{12}	0.4904	0.4884 ± 0.0112
$CLF + LF_{12}$	0.4890	0.4869 ± 0.0067
$LF_2 + ELF_2^\star$	0.4863	0.4913 ± 0.0170
$LF_2 + ELF_1^\star$	0.4863	0.4922 ± 0.0171
$CLF + ELF_1^\star$	0.4835	0.4843 ± 0.0116
LF_2	0.4822	0.4925 ± 0.0152
CLF	0.4808	0.4736 ± 0.0099
$LF_{12} + ELF_{12}$	0.4794	0.4940 ± 0.0096
ELF_2	0.4781	0.4763 ± 0.0083
$ELF_1^\star$	0.4740	0.4697 ± 0.0123
ELF_1	0.4712	0.4804 ± 0.016
$SR_1^{'}$	0.4452	0.4591 ± 0.0018

Table 2: The table represents the Accuracy scores for the different sets of variables in the case of the target variable $\Delta SR_{12}^{''}$

Set of Features	Acc. Test	Acc. Validation
$\mathbf{LF_2 + ELF_1^\star}$	$\mathbf{0.5425}$	0.4961 ± 0.0068
$LF_2 + ELF_2^\star$	0.5342	0.4930 ± 0.0074
ELF_2	0.5301	0.4883 ± 0.0165
$CLF + LF_{12}$	0.5287	0.4988 ± 0.0142
ELF_{12}	0.5246	0.4912 ± 0.0123
$LF_1 + ELF_2^\star$	0.5232	0.4963 ± 0.0134
LF_{12}	0.5205	0.4980 ± 0.0069
$LF_{12} + ELF_1$	0.5205	0.5060 ± 0.0101
$LF_1 + ELF_1^\star$	0.5205	0.4949 ± 0.0162
LF_2	0.5192	0.4932 ± 0.0056
$CLF + LF_1 + ELF_1^\star$	0.5178	0.4997 ± 0.0109
$LF_{12} + ELF_{12}$	0.5109	0.4833 ± 0.0099
CLF	0.5096	0.4627 ± 0.0079
LF_1	0.5082	0.4976 ± 0.0075
$CLF + ELF_1^\star$	0.5081	0.4840 ± 0.0044
$ELF_1^\star$	0.5000	0.4862 ± 0.0064
ELF_1	0.4972	0.4934 ± 0.0119
$\Delta SR_{12}^{'}$	0.4698	0.4681 ± 0.0078

For the first task, each subset contains the variable $SP_1^{'}$. The Extra Linguistics subsets marked by *, contain just the age and sex information. We divided the data in a Training, Validation and Test set. We performed a K-Fold approach (K = 3) testing different parameters choosing that ones which best performed on the Validation set.

As the parameters were fixed, we computed the accuracy score on the Test set. The Accuracy value of the Baseline is 0.41 (corresponding to the majority class). As noted regarding Table 1 the score for each group of subset is greater than the Accuracy Baseline. The lower score corresponds to the set in which we just use the speech rate of one speaker. It suggests that the use of more information helps to slightly increase the score. In particular, the ELF (Extra Linguistics Features) don't change significantly if we use the information of speaker 1 and speaker 2 separately. Using just Linguistics Feature (LF)

[1] We use the subscript (1), (2), (12) to indicate respectively if the variable refers to speaker 1, speaker 2 or both speaker 1 and speaker 2 (e.g.; $LF_{12} = LF_1 + LF_2$)

we can note that the best score is obtained using the linguistics features of both the speakers, and that $LF_1 > LF_2$. In general, the highest Accuracy score is obtained using the LF of speaker 1 in addition of ELF and CLF. In order to compare the results, we performed a Bootstrap significance Test using the Random Approximation method as described by Yeh (2000). We computed the Significance Test on the Accuracy using couple-match approach. All the variable (except for ELF_1, $ELF_1^\star$) sets result to be significantly different (with a p value $p < 0.05$) from the set SR_1. This result suggests that the use of LF and both LF and ELF improve the performance of the classification. Moreover, the set of features $CLF + LF_1 + ELF_1^\star$, LF_{12} are significantly different compared to the sets formed by just ELF. This indicates that the speech style influences the speech rate of the speaker than the age and sex of the singular speakers.

For the second task, predicting how the variable ΔSR_{12} changes, we applied the same procedure as we described previously, assigning one of the three classes (D, I or S) to the target variable. In this case, the set CLF contains just the *Similarity Score of the Discourse Markers* and *Total Overlap* because we have already taken into account the variable $\Delta SR'_{12}$ for all the sets of features. The baseline Accuracy value on the Test set is 0.38 (by selecting the majority class).

In Table 2 the accuracy scores for the different sets are reported. As the previous case, all the sets of features have a better accuracy compared to the use of just $\Delta SR'_{12}$. In particular, the use of features of the speaker 2 improves the performance on the Accuracy.

Applying the bootstrap significance test, described in the previous task, it comes out that all the variables (except for ELF_1, $ELF_1^\star$) set result to be significantly different from the set $\Delta SR'_{12}$. This result suggests that the use of LF and in addition of ELF improve the performance of the classification but the only ELF of the singular speakers don't help to predict the speech rate difference of the conversation. Moreover, the set of features $LF_2 + ELF_1^\star$, result to be significantly different from the others variables except from LF_2, ELF_2, CLF_{12}, ELF_{12} and $LF_1 + ELF_2^*$.

3 Discussion

In this exploratory work, we presented two experiments as an approach to test whether is possible predict the changes of the speech rate using various linguistic and extra-linguistic parameters, as averaged values produced by the speakers in the first part of the conversation. We tested different sets of features showing that the use of these sets of variables gives an improvement on the Accuracy score compared with the baseline (majority class) in both the experiments. We are especially interested in assess the difference among the different sets of variables in order to investigate a possible relation between the speech style and the speech rate production during the conversation. The accuracy scores among these sets of features are compared through a robust Significance test, a Bootstrap approach, due to the use of the Switchboard corpus that contains a large number of conversations. In the first experiment (predicting speech rate of one speaker) the use of Linguistics features, and both Linguistics features and extra linguistics information of the profile of the speakers perform significantly better than just use sex, age information. Instead, extra linguistics features of the speaker are not significantly better compared to the simple speech rate. As regards the second experiment, also in this case all the sets of features perform better than use just $\Delta SR'_{12}$. Moreover, the significance test shows that the use of linguistics features of speaker 2 and extra linguistics information of the speaker 1 performs significantly better compared with the the linguistics features of both speakers, cross-speaker features and the extra linguistics information of the speaker 1. This could indicate that the speech rate and his dynamic depend by the whole speech style of both speakers, and the dynamic cant be explained just knowing extra linguistics information of the speaker. Anyway, these results should be interpreted as a starting point to study how the speech style of the speakers could influence the dynamic of speech style using this approach. Indeed, it is necessary to deepen some aspects. At first the accuracy scores reached at best 54%. This value seems quite low and should be justified considering the complexity of the phenomenon and the rough nature of the method. A better result maybe derive including acoustic features that are an important cues of the speech production. Additional experiments are also necessary to better specify what are the features that mainly influence the speech rate and the relation between the speakers.

References

Molly Babel. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1):177 – 189.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Uriel Cohen Priva, Lee Edelist, and Emily Gleason. 2017. Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, 141(5):2989–2996.

Howard Giles and Nikolas Coupland. 1991. *Language: Contexts and consequences.* Thomson Brooks/Cole Publishing Co.

Giles and Howard. 1980. Accommodation theory: Some new directions. *York Papers in Linguistics*, 9(105):30.

Howard Giles and Peter F Powesland. 1975. *Speech style and social evaluation.* Academic Press.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Frieda Goldman-Eisler. 1956. The determinants of the rate of speech output and their mutual relations. *Journal of Psychosomatic Research*, 1(2):137–143.

Annette Hannah and Tamar Murachver. 1999. Gender and conversational style as predictors of conversational behavior. *Journal of Language and Social Psychology*, 18(2):153–174.

Victoria Johansson. 2009. Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, 53:61–79.

T.S. Kendall. 2009. Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project. *Phd Thesis, Duke University*.

Linda L Newman and Ann B Smit. 1989. Some effects of variations in response time latency on speech rate, interruptions, and fluency in children's speech. *Journal of Speech, Language, and Hearing Research*, 32(3):635–644.

Richard L Street and Howard Giles. 1982. Speech accommodation theory: A social cognitive approach to language and speech behavior. *Social cognition and communication*, 193226.

Richard L. Street. 1984. Speech convergence and speech evaluation in fact-finding iinterviews. *Human Communication Research*, 11(2):139–169.

Andreas Weise and Rivka Levitan. 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 297–302.

Li-Chiung Yang. 2003. Duration and pauses as phrasal and boundary marking indicators in speech. In *Proceedings of 15th ICPhS*, pages 1791–1794.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.