Conversational structures affecting auditory likeability

Benjamin Weiss[1], *Katrin Schoenenberg*[2]

[1]Technische Universität Berlin, Quality & Usability Lab, Germany
[2]Technische Universität Berlin, Assessment of IP-based Applications, Germany
Benjamin.Weiss@tu-berlin.de, Katrin.Schoenenberg@telekom.de

10.21437/Interspeech.2014-408

Abstract

Three-person telephone conferences of unacquainted people are conducted by means of pre-defined scenarios providing individual information and goals. Interlocutors rate the likeability of each other after a training session as well as after the actual conference. The recordings of the conferences are manually annotated concerning speaker's verbal contribution, pauses, and back-channels. Regression analysis reveals likeability ratings after the conference to be dominated by the ratings before the conversation, but simple parameters like number of turns also contribute significantly to a descriptive model. The regression model for ratings averaged for both interlocutors provide a similar fit as the one for pair-wise ratings. Exchanging of the manually obtained parameters by automatically estimated values still results in significant regressions, indicating facilitation for future research.

Index Terms: interpersonal perception, interaction parameters, likability, triads

1. Introduction

When strangers interact with each other, they execute not only various attribution processes to estimate or guess physical, social or even personality traits in order to properly deal with the interlocutors. They also evaluate the interaction partners concerning their social attraction, i.e. if they like or dislike the other. Such a likeability judgment is constituted quickly and usually relying on stereotypes if based on few information only [1, 2]. However, this process can be made conscious and thus reported during an evaluation process.

Much research on social attraction has focused on mid-term or long-term effects. However, studies concerned with so-called "thin slices" show that ratings of brief interactions (of minutes) or even briefer stimuli without interaction (seconds) correlate with judgments assessed after full conversation and can be used to predict even the outcome of (professional) interactions (e.g. for teachers, doctors) [1, 3]. Still, it is not clear how much the first nonverbal impression and the initial verbal interactive behavior contribute to this consistent evaluation. In this paper, we therefore study the effect of conversational structures as verbal predictors of likeability, and how these relate to the already established first impression. To separate these two assumed predictors of likeability, judgments are assessed directly before (the first impression) and after the conversation (impact of the conversational structure).

2. Related Work

Typically, quantitative analysis of human conversation is not conducted on the conversational level, but instead aspects of the interlocutors are assessed externally. For example, ideal character traits are assessed to be able to relate similarity and likability, or third-party observers are asked to rate character traits to relate similarity to likeability [4], or by letting third-party observers rate the communicative style on a questionnaire to find correlations with likeability rated by a separate group of observers [5, 6]. Others may even set up the conversation by instructing a confederate interlocutor to behave differently in each condition tested [7, 8, 9]. Although such studies reveal effects for a relationship of interaction behaviour and likeability, there is only a thin body of research on actual parameters describing a conversation.

Verbal and non-verbal (prosodic, lexical, syntactical) alignment of interlocutors is affecting likeability and smoothness, and thus success of interaction [10, 11]. For example, there is a general divergence in articulation rate found in [12], which is reversed or at least decreases for mutual liking. In this paper, however, we do not analyze such prosodic features, but concentrated on conversational structures, including similarity in parameter values.

Likeability ratings, collected from observers, correlate positively with filled pauses and contractions, and negatively with, e.g., interjections [13]. Interruptions and, surprisingly, back-channels have a negative effect though. Even those interlocutors with the role of a follower are rated better, which take turns by causing overlap instead of a pause. Concerning just the role of a follower, those interlocutors are rated better that take turns by causing overlap instead of a pause.

The concept cohesiveness is related to likeability, and was studied for the four-person "scenario" data of the AMI corpus by [14]. Within a number of extracted parameters, averaged post-meeting ratings of cohesiveness are negatively correlated with interruptions, proportion of silence in turn-taking (both in line with [13]) and with the number of the dialog act "eliciting information", but positively with turn-taking freedom and the dialog acts "providing assessments" and "comments about understanding".

Turns, silence, laughter and back-channel were manually annotated for dyadic telephone calls based on a survival scenario [15]. A post-conversational questionnaire of social attractiveness shows positive correlations of the role of the caller with the number of laughter and back-channel. For the role of the receiver there is a negative correlation with laughter. The number of overlapping speech reveals a similar systematic as laughter, but is not significant.

Finally, social attractiveness of interlocutors during interviews was compared to third-party observers and related to turn duration and response latency of the interviewee (along with speaking rate) [16]. Observers ratings do not covariate with the participants' data. For the interlocutors, response latency correlates with social attraction for both roles. Additionally, interviewees are rated more negatively with increasing turn duration

and decreasing similarity in turn duration.

The aim of this study is to obtain insight into the strength of correlation between likeability and both, the first impression and interaction behavior on a surface level for the domain of unacquainted persons. To our knowledge such studies have not been reported so far. A dominant impact of the first impression will be expected [17]. As there are only few studies with participant ratings [15, 16, 14], identifying and confirming relevant interaction parameters is another goal.

3. Procedure

Three persons were recorded, verbally interacting according to prepared scenarios. The expected benefit of triads over dyads is the possibility to compare pair-wise effects to the likeability of a speaker averaged over both interlocutors.

Altogether, 39 persons took part in this experiment (9 women, 30 men, aged 36.2 years, SD=12.2). The triads did not know each other in advance. The participants were all experienced in telephone conferences to ensure familiarity with the situation and technology. On average, the participants stated to have conducted 34.65 telephone conferences in total. All participants of a group were instructed to the procedure together and then send to their individual sound proofed room (ITU-T Rec. P.800 [18]). From this personal meeting, participants gained a first impression of each other including visual information.

The three rooms were connected by a conferencing system implemented in PD [19]. It provided broadband connection with intensity attenuation of 23.1 dB SPL (test signal of 61.3 dB SPL). Closed headsets of the type Beyerdynamics DT 290 were used by the participants. Prior to the actual session a first training scenario was conducted. Following this, issues concerning the procedure could be clarified, but the subjects did not leave their individual room. All scenarios were taken from the collection described in [20]: These semi-structured tasks provide business conversations with topics like choosing a conference location or songs of a music album. For each scenario, every participant receives different information to contribute or ask for in order to stimulate the conversation. All aspects are "solved" in this manner by the three participants. Although the scenarios provide various job descriptions, no specific (conversational) roles are defined. During annotation, we gained the impression that, e.g. a leading role was taken individually, if at all, and its conversational consequences are thus reflected in the resulting conversational parameters. The training scenario was fixed, whereas the actual scenario accounting for the analysis varied. The conversations analyzed here are only the first block of a bigger experiment investiagting transmission delay in later blocks, which is why the scenario was randomized. Each conversation was initiated by a melody.

After the conversation, each participant was asked to state each partner's likeability, as well as personal attention to the call and overall quality of the transmission. The likeability ratings after the training are used as basis to control for the fist impression established so far. The scale used was continuous with the two antonyms "very likeable" and "very unlikeable" afterwards transformed to numeric values between zero and ten.

4. Annotation

The conversations have been manually segmented and annotated with ELAN[1]. As the specific task of annotating, e.g. back-channel, was actually new to us, we decided against separate annotations and κ values, but for consequent discussions and counter-insurance to ensure a single strategy. Two students of linguistics conducted the annotation by supervision of the two authors. The aim was to describe the structure of each recording to quantify for each participant the type of utterance (back-channel, turn, noticeable pause within turn) and turn management (turn change by pause, verbal overlap or a failed attempt), as found in the related work section.

From this segmentation and annotation 14 parameters have been extracted to represent structural aspects of the conversation for one interlocutor. Additional parameters can be derived from these, e.g. total time of Speaker S turns to all turns or ratio of S pause duration to S speaking duration.

Annotating all conversations was very laborious, even though relatively easy labels were used in comparison to, e.g., dialog act labels. Therefore, alternative parameters estimated automatically are used for a replication of the analysis with hand-annotated data. These parameters are extracted based on voice activity detection [21]. Subsequently, segment borders and segment durations have been calculated based on a state model [22]. However, a distinction between back-channels and actual turns is not trivial and was therefore not attempted for this data. Therefore, the automatically obtained surrogate of "a turn" includes all contributions of one interlocutor.

[1]EUDICO Linguistic Annotator: `http://tla.mpi.nl/tools/tla-tools/elan/`.

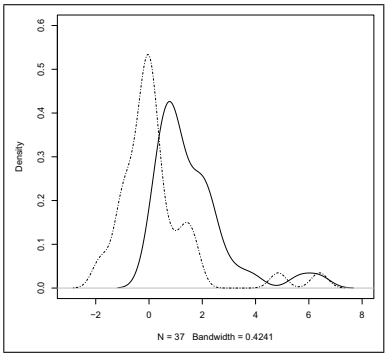

Figure 1: *Density distribution of the averaged ratings.*

5. First Inspection

Unfortunately, some participants failed to rate the likeability in some cases: From the 39 participants, 2 interlocutors did not rate at all, and for 8 participants, there is only one rating available for averaging. A first inspection also reveals two potential outliers for the averaged data as well as for the relative ratings, i.e. the difference between the basic rating and the final likeability (see Figure 1), which are actually identical. Both outliers differ from the mean more than two standard deviations, and are excluded for further analysis as they might distort the regression models. Interestingly, there is no relevant agreement between two raters on the third interlocutor (Intra-Class-Correlation, ICC=.23, p=.11). This holds also for reciprocal ratings (ICC=−0.25, p=.91), which is unexpected, as it is not in line with results for reciprocal interpersonal attraction comprised in the Attraction Theory. As a result, interdependence does not have to be considered in the linear models [23]. According to our expectations, however, there is a moderate correlation for the basic (pre-conversational) and post conversational likeability (r=.54, t(50)=4.58, p<.0001).

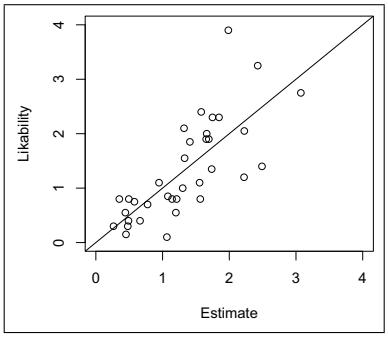

Figure 2: *Averaged likeability and the estimates from the linear model with basic rating and 2 additional parameters.*

6. Results

The results are presented in two sections: first, the relationship of likeability averaged for both conversational partners with surface parameters of the rated person will be examined; second, pair-wise relative parameters for each of the three pairs of interlocutors will be calculated.

6.1. Describing averaged likeability

Likeability ratings of an interlocutor averaged for the two others constitute the dependent variable, while the parameters from annotation as well as the basic likeability rating form the independent variables. Data from 35 participants is used, with 7 single ratings instead of averaged ones due to missing ratings.

In a first step, simple correlations are used to include only related variables ($p<.10$) in the multiple linear regression to avoid over-fitting during the parameter selection. Thus, from originally 36 variables extracted from annotation, only four items were used in a step-wise multiple linear regression in addition to the basic likeability rating (number of S turns ending with overlap, ratio of S turns ending with overlap to S speaking duration, ratio of S turn to all turns, ratio of S turns ending by pause to all turns ending by pause). The final regression model includes likeability and two of the four parameters based on AI-criterion ($F_{(3,31)}$=13.66***, R^2=.57, $R^2_{adj.}$=.53, RMSE=.59), confer Table 1 and Figure 2.

Table 1: *Results of the multiple linear model for averaged likeability ratings – from annotation.*

Parameter	beta	t-value	p-value
basic likeability rating	0.56	5.131	<.001***
number of S turns ending with overlap	0.29	2.671	<.05*
percentage of S turns to all turns	-0.37	-3.375	<.01**

A related model for both interactive parameters only (without the basic likeability rating) is still significant, but covers considerably less variability ($F_{(2,32)}$=4.09*, R^2=.20, $R^2_{adj.}$=.15, RMSE=.80). It has to be taken into account that building a model without basic ratings might have led to a model with more parameters and a better fit, if the task had been to estimate a model based on interaction parameters only.

Most of the explained variance is covered by the basic rating, but parameters describing the surface structure of the conversation also significantly contribute to the final likeability. Only two parameters are included in the model, while an increasing proportion of the turn number has a negative effect, the increase of number of turns ending with overlap have a positive one. Interestingly, there is no observed effect for back-channels.

Automatically obtained estimates of the two parameters used do correlate significantly with the respective variable extracted from the manual annotations (Pearson's correlation: number of S turns ending with overlap, r=.60, p<.001; percentage of S turns to all turns: r=.75, p<.0001). Despite providing only moderate correlations, a repetition of the model estimation with automatically obtained parameters results only in a slightly lower fit than with manual data ($F_{(3,31)}$=11.15***, R^2=.52, $R^2_{adj.}$=.47, RMSE=.87).

Table 2: *Results of the multiple linear model for averaged likeability ratings – automatic estimates.*

Parameter	beta	t-value	p-value
basic likeability rating	0.63	5.278	<.001***
number of S turns ending with overlap	0.33	2.754	<.01**
percentage of S turns to all turns	-0.33	-2.812	<.01**

6.2. Describing individual likeability

In order to test, whether the averaging of the likeability ratings removes information systematically related to the interaction parameters, an analysis similar to the previous one was conducted. However, this time individual ratings were chosen as the dependent variable, as well as pair-wise relative parameter values, to cover each of the three pairs within one dyad. To represent each pair of interlocutors within one triad, the ratio of absolute values has been calculated as independent variables, e.g. ratio of average turn duration of S to the one of the rater Sx, i.e. S/Sx. For counts, the number of pair-wise turn events (changes and attempts) are considered for each individual pair, e.g. number of turn changes from S to Sx and vice versa, as well as the ratio of both variables (each value shifted by one, as the data did contain zeros).

Data from 27 participants is available, resulting in 54 pairs, from which the above mentioned outliers are removed already. This results in about double the data points compared to averaged ratings, as every participant is mostly rated twice, but this of course also induces more variability.

Following the same procedure as for the averaged ratings, simple regressions reveal four parameters out of 14 derived ($p<.10$) in addition to the basic likeability ratings: 1. ratio of total turn duration of S to the one of Sx, 2. ratio of turn number of S to the one of Sx, 3. ratio of duration of doubletalk excluding back-channel ($S\underset{\rightarrow}{^{Sx}}S$ / $Sx\underset{\rightarrow}{^{S}}Sx$) and 4. ratio of number of turn changes with overlap excluding back-channel ($S{\rightarrow}Sx/Sx{\rightarrow}S$).

The subsequent multiple linear regression including the basic likeability rating and three additional parameters is significant. It provides a fit comparable to the one for the averaged ratings ($F_{(4,47)}$=18.79***, R^2=.62, $R^2_{adj.}$=.58, RMSE= .83), cf. Table 3 and Figure 3. The respective model without the basic rating is also significant, but covers much less variability ($F_{(3,48)}$=5.33**, R^2=.25, $R^2_{adj.}$=.20, RMSE=1.16).

Automatically obtained replacements of the parameters do again correlate with the manually extracted ones (ratio of turn number of S to the one of Sx, r=.83, p<.0001; ratio of total turn duration of S to the one of Sx, r=.97, p<.0001; ratio of number of turn changes with overlap for both, r=.31, p<.05).

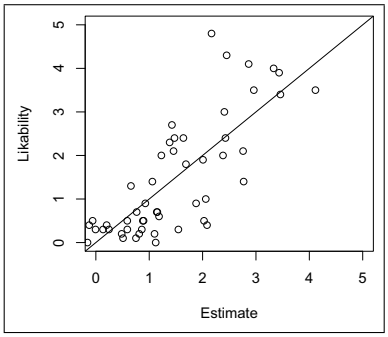

Figure 3: *Individual likeability and the estimates from the linear model with basic rating and 3 additional pair-wise parameters.*

Table 3: *Results of the multiple linear model for individual likeability ratings – from annotation*

Parameter	beta	t-value	p-value
basic likeability rating	0.82	6.346	<.001***
turn number of S / the one of Sx	-0.43	-3.205	<.01**
total turn duration of S / the one of Sx	-0.33	-2.498	<.05*
ratio of no. turn changes with overlap	0.33	2.535	<.05*

The subsequent fit is considerable lower, but exhibits a similar structure (see Table 4; $F_{(4,47)}$=9.63***, R^2=.45, $R^2_{adj.}$=.40, RMSE=1.35).

Table 4: Results of the multiple linear model for individual likeability ratings – automatic estimates.

Parameter	beta	t-value	p-value
basic likeability rating	0.80	5.288	<.001***
turn number of S / the one of Sx	-0.28	-1.659	=.104
total turn duration of S / the one of Sx	-0.26	-1.511	=.137
ratio of no. turn changes with overlap	0.20	1.355	=.182

7. Discussion

We cannot replicate the effects of back-channels and interruptions from [13, 14, 15]. However, the reported positive effect of a turn-transition by overlapping speech [13] is found, albeit for the turn holder, not the taker. It can be interpret as the participant's degree to invite fluent turn taking or collaborative behavior. The negative effect of longer turn duration [16] is only reflected in the pair-wise data, not in the averaged data. Instead, an increasing number of turns has a negative impact on averaged and pair-wise ratings.

Of course, these studies, including our own, are not directly comparable. We argue to not neglect the first impression in favor of interaction parameters. This might result in artificial parameter effects. Despite the dominating first impression, aggregated parameters describing the conversational structure are significantly complementing the description of the persons' likeability. Using conversational parameters only, would certainly increase the relevance of those descriptors, but also lead to a lower fit, as we know for example from Paradise models using subjectively described task success versus instrumental measures [24]. Still, the aim of this analysis was not automatic prediction. Instead, the aim was to test if parameters of conversational structures add to initial subjective ratings, thus motivating to study such a source of information increasingly in the field of Computational Paralinguistics. For this, we find evidence from the results.

There are two, respectively three parameters significantly contributing to the description models. Although a less conservative approach (i.e. choosing from all 35/14 parameters, not defining "outliers") might lead to more parameters and thus more explained variance, the method used ensures a reasonable number of parameters for interpretation. The negative effect of the proportion of turn numbers might refer to a lacking of politeness or attention towards the rater, as less proportion is rated more positively. Despite the rich body of results for similarity, the two parameter from the pair-wise analysis (relative turn number and turn duration) do not exhibit an effect of similarity. Furthermore, there seems to be an asymmetry for turn changes with overlap, as participants seemed to apply the pattern "the more I can interrupt S to take the turn, the more I like S".

Both models (averaged and pair-wise ratings) perform similar, with the pair-wise model having one parameter in addition.

The latter model contains a measure of "turn number" and one of "turn changes with overlap" similar to the model for averaged ratings. The third parameter is the relation of the speaker's turn length to the participant's turn length and might reflect an impression related to the turn number.

Indeed, the mere parameters describing surface features of the conversations alone do not provide sufficient information to properly interpret the resulting parameters, especially because replicating studies are lacking so far. Reproducing the results in experiments, e.g. with confederates might be a fruitful approach for future investigations.

8. Conclusion

The results obtained here emphasize the first impression on interpersonal attraction. However, the results also clearly indicate a relationship of conversational structure and changes in likability ratings. This relationship is independent from the effect of initial (e.g. facial, vocal) likeability on the examined parameters. One important finding is the impact of easily assessable parameters like turn number and turn duration, which can in principle be estimated automatically with high validity. Just one parameter was not estimated satisfyingly.

Although the interactive situation is somewhat artificial, goal-oriented conversation could be enabled. Even though, the corpus can certainly not be called representative, it can still be regarded as authentic; which is a prerequisite for studying the relationship of conversational surface structure and likeability.

Several important aspects are not considered in this initial analysis, i.e. aspects of speech acts and politeness, but also group processes and roles. For this first analysis, group processes were not considered. It may be interesting to examine them in future works.

The method itself, namely using triads to assess two instead of one rating can be improved further by collecting acoustic only initial judgments and by omitting the training session for these professionals.

9. Acknowledgements

We want to thank Charlotte, Anne and Pierre for their tremendous support. This work was financially supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant WE 5050/1-1.

10. References

[1] N. Ambady and J. J. Skowronski, Eds., *First Impressions*. New York: Guilford Press, 2008.

[2] J. Kreiman and D. Van Lancker Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Chichester: Wiley-Blackwell, 2011.

[3] J. Curhan and A. Pentland, "Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes," *Journal of Applied Psychology*, vol. 92, pp. 802–811, 2007.

[4] J. LaPrelle and R. Hoyle, "Interpersonal attraction and descriptions of the traits of others: Ideal similarity, self similarity, and liking," *Journal of Research in Personality*, vol. 24, pp. 216–240, 1990.

[5] D. Brandt, "On liking social performance with social competence: Some relations between communicative and attributions of interpersonal attractiveness and effectiveness," *Human Communication Research*, vol. 5, pp. 223–226, 1979.

[6] R. W. Norton and L. S. Pettegrew, "Communicator style as an effect determinant of attraction," *Communication Research*, vol. 4, pp. 257–282, 1977.

[7] S. Goldbrand, "Imposed latencies, interruptions and dyadic interaction: Physiological response and interpersonal attraction," *Journal of Reseach in Personality*, vol. 15, pp. 221–232, 1981.

[8] A. Baker and J. Ayres, "The effect of apprehensive behavior on communication apprehension and interpersonal attraction," *Communication Research Reports*, vol. 11, pp. 45–51, 1994.

[9] L. Kohn and R. Dipboye, "The effect on interview structure on recruiting outcomes," *Journal of Applied Social Psychology*, vol. 28, pp. 821–843, 1998.

[10] J. Bradac, A. Mulac, and A. House, "Lexical diversityand magnitudeof convergent versus divergent style shifting perceptual and evaluative consequences," *Language and Communication*, vol. 8, pp. 213–228, 1988.

[11] J. Lakin, V. Jefferis, C. Cheng, and T. Chartrand, "The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 145–162, 2003.

[12] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proc. Interspeech*, 2013.

[13] A. Gravano, R. Levitan, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," in *Proc. Interspeech*, 2011, pp. 97–100.

[14] C. Lai, J. Carletta, and S. Renals, "Modelling participant affect in meetings with turn-taking features," in *Proc. Workshop of Affective Social Speech Signals*, 2013.

[15] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, "From nonverbal cues to perception: personality and social attractiveness," in *COST'11 Proceedings of the 2011 international conference on Cognitive Behavioural Systems*, 2011, p. 60–72.

[16] J. Richard L. Street, "Participant-observer differences in speech evaluation," *Journal of Language and Social Psychology*, vol. 4, pp. 125–130, 1979.

[17] M. J. Harris and C. P. Garris, "You never get a second chance to make a first impression: behavioral consequences of first impressions," in *First Impressions*, N. Ambady and J. J. Skowronski, Eds. New York: Guilford Press, 2008, pp. 147–170.

[18] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," International Telecommunication Union, Geneva, 1996.

[19] M. Puckette, "The theory and technique of electronic music," http://puredata.info/, 2007.

[20] A. Raake, C. Schlegel, K. Hoeldtke, M. Geier, and J. Ahrens, "Listening and conversational quality of spatial audio conferencing," in *Proc. 40th Conference of Audio Engineering Society (AES)*, vol. 3, 2010, pp. Paper Number:4–7.

[21] I. Luengo, E. Navas, I. Odriozola, I. Saratxaga, I. Hernaez, I. Sainz, and D. Erro, "Modified LTSE-VAD algorithm for applications requiring reduced silence frame misclassification," in *Proc. of the LREC*, 2010, pp. 1539–1544.

[22] K. Hoeldtke and A. Raake, "Conversation analysis of multi-party conferencing and its relation to perceived quality," in *Proc of the. Int. Conf. on Communications (ICC), IEEE*. Kyoto, Japan, 2011, pp. 1–5.

[23] D. A. Kenny, "Models of non-independence in dyadic research," *Journal of Social and Personal Relationships*, vol. 13, pp. 279–294, 1996.

[24] S. Möller, K.-P. Engelbrecht, and R. Schleicher, "Predicting the quality and usability of spoken dialogue services," *Speech Communication*, vol. 50, pp. 730–744, 2008.