*Multimodal Technologies and Interaction*

MDPI

*Article*

# Group Leader vs. Remaining Group—Whose Data Should Be Used for Prediction of Team Performance?

**Ronald Böck** [1,2]

[1] Research Division, Genie Enterprise, Donnersbergweg 1, 67059 Ludwigshafen, Germany; rboeck@genie-enterprise.com; Tel.: +49-62-1166-39018

[2] Cognitive Systems Group, Faculty of Electrical Engineering and Information Technology, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany

**Abstract:** Humans are considered to be communicative, usually interacting in dyads or groups. In this paper, we investigate group interactions regarding performance in a rather formal gathering. In particular, a collection of ten performance indicators used in social group sciences is used to assess the outcomes of the meetings in this manuscript, in an automatic, machine learning-based way. For this, the Parking Lot Corpus, comprising 70 meetings in total, is analysed. At first, we obtain baseline results for the automatic prediction of performance results on the corpus. This is the first time the Parking Lot Corpus is tapped in this sense. Additionally, we compare baseline values to those obtained, utilising bidirectional long-short term memories. For multiple performance indicators, improvements in the baseline results are able to be achieved. Furthermore, the experiments showed a trend that the acoustic material of the remaining group should use for the prediction of team performance.

**Keywords:** group analyses; performance prediction; baseline results; machine learning-based prediction

check for updates

**Citation:** Böck, R. Group Leader vs. Remaining Group—Whose Data Should Be Used for Prediction of Team Performance? *Multimodal Technol. Interact.* **2023**, *7*, 90. https://doi.org/10.3390/mti7090090

Academic Editor: Stephan Schlögl

Received: 2 May 2023
Revised: 16 August 2023
Accepted: 12 September 2023
Published: 14 September 2023



## 1. Introduction

Humans are considered to be communicative, usually interacting in dyads or groups. Billions of interactions and hundreds of millions meetings take place every day around the world. Such gatherings can be quite short, for a brief exchange of gossip, or be rather long coordination meetings, defining the current state and future plans of entire countries. However, the evolution of such meetings and, thus, the creation and development of groups or teams is a dynamic process [1], indicating the flexibility in everyday interactions. In social sciences, those aspects are discussed already for a longer period of time (e.g., [2–5]), but also computer sciences need to address those issues, especially if it comes to multi-agent interactions (e.g., [6,7] for an overview), whereas the agents can be either human beings, technical devices, or virtual agents. Further, the millions of meetings usually should have ideally a purpose and an intended outcome [1,2,8]. Moreover, the goal as well as the gatherings context (cf. e.g., [9]) influence the way of communicating in the group (e.g., [6,7,10]). As already stated in [11]:

> "An informal family gathering is mainly related to fun aspects and the (social) feeling of closeness. In contrast, formal business meetings put the (efficient) exchange of information and solutions in focus (cf. e.g., [12])."

Regarding the meetings' outcome, two different interpretations can be seen: (1) The classical way, where the outcome relates to specific goals or results and thus, can be assessed in terms like effectiveness; (2) The interpretations discussed in [11] where a more socialising perspective is being considered.

Regarding, at first, the second aspect, being mainly related to longer-lasting meetings that are rather considered not as effective, the authors of [11] argue:

> "These meetings were called "stimulating meetings", meetings being perceived as effective in terms of outcomes and the way of interaction but are not necessarily short in the sense of measured absolute meeting time[, but] are considered as interactions where the communication partners and their interaction are propelled (by each other) such that the entire group is able to perform better."

Keeping this in mind, [13] further argues that such meeting are often used for familiarisation in the group (cf. also [9]), especially when the group members do not know each other, yet. Therefore, even a longer-lasting meeting can be perceived as effective and considered as a kind of "bet" in a perhaps more efficient interaction later on [13]. However, the term "stimulating" has also another perspective: In brainstorming-like meetings, the discussion is usually allowed to be more interactive and lively as well as open-minded to generate a (large) number of ideas. Therefore, the ongoing interaction can be fertilised by the communication partners, resulting in stimulating conversations. Ideally, such meetings are either open-ended or repetitive, avoiding restrictions and limitations.

Coming back to the first interpretation, the classical view on meetings and their outcomes is given. In each meeting, there is an inherent longing that it should be effective, being related also to the performance of the group. Ideally, the effectiveness, often put on the same level as having a good meeting performance (cf. [14,15]), should be measured already during the gathering, in an objective way. Therefore, we aim for an automatic prediction of meeting performance, in this manuscript, being represented by performance indicators (cf. Section 2.1.2), to provide an objective statement. In this sense, the current work investigates the groups' performance results of meetings in the Parking Lot Corpus (cf. Section 2.1.1). For this, particular performance indicators from social sciences were considered, explained in Section 2.1.2. Furthermore, this is the first time that baseline prediction results on the corpus are provided, following the research questions stated in Section 1.2. Although we investigate the Parking Lot Corpus in rather the classical way, we also consider the "stimulating meeting" aspect in parallel (in terms of discussions).

### *1.1. Related Work*

There is already work which refers to the Parking Lot Corpus, introduced in Section 2.1. In particular, this refers to the sound of the meeting in [11], where the prosodic–acoustic characteristics of the spoken samples were analysed. The authors show that there are indeed acoustic differences between (perceived) successful meetings and those considered to be not as efficient, mainly based on pitch-related features. These investigations are the foundation of the current manuscript, investigating acoustic samples (utilising extracted features) for the prediction of performance indicators.

Furthermore, the already discussed concept of "stimulating meetings" [11] was introduced. In a second paper, the corpus was used to investigate the influence of speech duration and number of turns on the perceived meeting effectiveness [13]. In particular, these analyses support the observations, leading to a more detailed interpretation of effectiveness as discussed in the introduction above.

In [6], an overview of communication and interaction in multi-party settings is presented, considering social and technical perspectives. This is complemented in [7], where a respective overview as well as a listing of current challenges are provided, focusing on multimodal and acoustic investigations.

In a classical view, group performance is related to a team's cohesion and the internal synchrony. In [11], this aspect is discussed as:

> "Team cohesion refers to the sense of togetherness, whereas synchrony describes the temporal alignment of activities or actions within a team. Measurement and categorisation concepts are provided by [16,17] for cohesion and, similarly, by [18,19] for synchrony. Beyond relationships within teams and team performance (cf. e.g., [20,21]), the cohesion and synchrony of teams also to some degree determine the quality of customer support (cf. e.g., [22]) and job satisfaction (cf. e.g., [5]). Investigations on estimated team outcomes are, according to [23],

influenced by "dynamics [. . .] (cf. e.g., [4,23])[, . . . meeting] satisfaction (cf. [5]), handling of errors (cf. [24]), humour (cf. [25]), and group emotions (cf. [2])".

As stated in [7], group analyses can be approached multimodally. For an overview, using rather visual cues in engagement in groups, we refer to [26]. In terms of acoustic analyses, [7] provides specific overviews, highlighting current achievements in analysing groups and their behaviour (e.g., [27–29]).

Regarding acoustic features, also utilised in the manuscript, we see that a huge variety of collections or feature sets are available in the community of speech recognition and affective computing (cf. e.g., [30]). We state that often rather larger feature sets or respective sub-sets are used, for instance, 6552 features in [31,32] or 2832 features in [33]. Rather small feature sets can be seen in, for instance, [34,35] with fewer than 300 features. One of the most prominent feature set is the *emobase* feature set [33] and respective derivatives, like in the paralinguistic challenges (cf. [36]). In our experiment, we rely on *emobase* features, being introduced in Section 2.2.3.

Finally, we mention the work in [37], where a combination of acoustic and linguistic cues is used for performance prediction in groups. The authors achieve good results using domain adaptation to overcome the drawback of a low number of samples. However, the current manuscript applies machine learning techniques to the Parking Lot Corpus, which are parameterised in such a way that they are able to handle training only on one modality (acoustics).

### *1.2. Research Questions*

Following the results presented in [11,13], we extend the investigations focusing on the central question: What are the particular influences of acoustic contributions by the leader of a group compared to the remaining group members?

For this, the overall goal of this manuscript is the automatic prediction of the group's performance results regarding a variety of performance indicators (cf. Section 2.1.2), i.e., the prediction of objective and subjective performance indicators, especially in relation to the acoustic samples provided by the group's leader vs. the remaining group. Taking this into consideration, we focus on two research questions in the manuscript:

RQ1: What are the particular baselines for the prediction of performance indicators, based on the acoustic material in the Parking Lot Corpus?

RQ2: What is the prediction capability of neural networks that are tailored to handle temporal dependencies?

Given these research questions, we state the following hypothesis:

**Hypothesis 1.** *The use of neural networks capable of processing temporal dependencies improves the prediction performance for the indicators.*

## 2. Materials and Methods

### *2.1. The Parking Lot Corpus and Performance Indicators*

#### 2.1.1. Corpus Description

The investigations are based on the Parking Lot Corpus [11] comprising, in total, 70 meetings recorded at a public Midwestern United States university. The general setting of the recordings is visualised in Figure 1, wherein the particular occurrence slightly varies depending on the utilised location within the university. In the experiments, the group size ranges from three to six participants (mean group size: 3.6). For recruitment, in total, 245 undergraduate students from the psychology department were invited, obtaining class credits compensating for participation; no further selections of participants were made. As stated in [11]: "Each meeting occurred independently, and the different sized groups resulted from intentional non-specification of the number of participants per group."

The corpus comprises audio–visual recordings of discussions, aiming for recommendation to improve the university's parking situation. Each group was instructed accordingly,

providing also a list of questions to be considered during the discussion. However, the group's interaction can still be considered non-scripted and spontaneous in the sense of [38]. For each group, a leader was determined by rolling a die. For the group leader, the experimenter introduced typical tasks to perform during the interaction, for instance, guiding the discussion, keeping attendees on track, etc. The total time of discussion was controlled by the experimenter to allow comparable conditions, important for some of the performance indicators (e.g., number of recommendations). Therefore, each group was given 20 min for interaction and filing the discussion results. After the meeting, each participant filled several questionnaires (cf. [11]), providing self-annotations of the meeting, using different performance indicators (cf. Section 2.1.2 for details on indicators and references to questionnaires). Furthermore, the meetings were assessed by trained annotators, given annotations on additional performance indicators.

The corpus' data were recorded using a (simple) video camera and an internal microphone (cf. Figure 1). Therefore, in the analyses, we have to deal with naturalistic material (cf. [38]) in a quality that is, however, still fine for investigations. Data contributors did not ensure occlusion-free or masked-free data, which limits video-based analyses. Regarding the acoustic quality, it is proper and no noise is included, resulting from the conference room-like environment. In our experiments, we rely only on the acoustic samples, being extracted from the video material. Details on the limitations of the corpus are discussed in Section 2.1.3, and restrictions in the experiments are stated in Section 2.2.1.

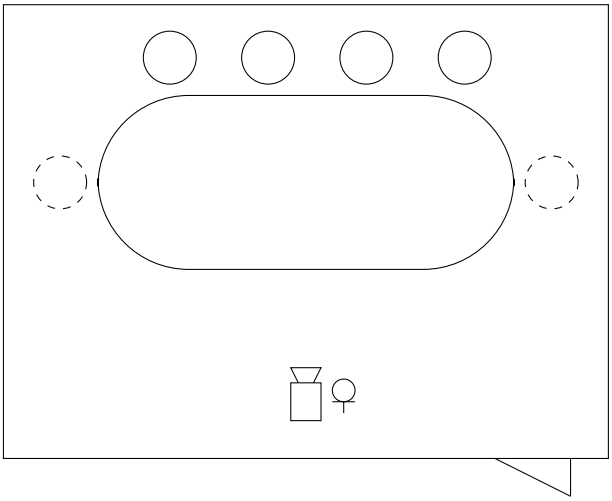

**Figure 1.** Sketch of the recording setting. The recordings take place in a conference room-like environment, providing a conference table and multiple seats (dashed seats indicate variations in group sizes; cf. Section 2.1.1). Orientation of the camera and microphone are also visualised. Figure is taken from [11].

#### 2.1.2. Performance Indicators

As already mentioned in the corpus description (cf. Section 2.1.1), performance indicators are based on either individual assessments of the participants or on the external validation by qualified raters. For the individual ratings, questionnaires (we refer to the particular performance indicator for references) are used, utilising a varying number of items, indicated on a Likert scale, to request and guide the assessment. These ratings are provided along with the recordings and were pre-processed by the team of Joseph A. Allen, University of Utha. In particular, the self-assessments allow an internal view of the perceived performance of the group, which enabled discussions in relation to aspects of "stimulating meetings" [11]. This issue might be of interest in the discussion of our experiments in Section 4 , as well.

Following the suggestion in [11], we distinguish two categories of performance indicators, namely, subjective and objective indicators. These are briefly introduced in the following.

*Subjective performance indicators* provide an individually perceived impression of the meeting situation and are further used to obtain an internal view of the group itself. Mean indicates that the indicator is finally averaged across all group members to achieve an evaluation of the entire group.

- **ME_mean.** Six items measured meeting effectiveness following Leach et al. [39]. Participants rated the meeting's achievements on a 5-point scale from "extremely ineffective" to "extremely effective".
- **MSA_mean.** Four items measured satisfaction with meeting processes following Briggs et al. [40]. Participants rated mainly how the meeting was conducted on a 5-point scale from "strongly disagree" to "strongly agree".
- **MSO_mean.** Four items measured satisfaction with meeting outcomes following Briggs et al. [40]. Participants rated their satisfaction with the overall meeting results on a 7-point scale from "strongly disagree" to "strongly agree".
- **MSP_mean.** Five items measured satisfaction with process following Briggs et al. [40]. Participants rated the perceived satisfaction with the current process on a 7-point scale from "strongly disagree" to "strongly agree".
- **MSBS_mean.** Twenty-seven items measured boredom in the meeting following Fahlman et al. [41]. Participants rated the level of boredom in the current meeting on a 7-point scale from "strongly disagree" to "strongly agree".
- **ANX_mean.** Five items measured anxiety in the meeting. The scale was generated in the group of Joseph A. Allen based on general and social anxiety scales (e.g., [42,43]). Participants rated the anxiety in the current meeting on a 5-point scale from "strongly disagree" to "strongly agree".

*Objective performance indicators* measure the objective and countable outcomes of the meeting, being based mainly on the provided written recommendations. Thus, an external view on the meeting and its "productiveness" is already (somehow) given.

- **TS_Rec.** Total recommendations is the counting indicator stating the written recommendations provided by the group. The higher the numbers, the more ideas or recommendations were generated.
- **Mean_F_Rec.** Each recommendation was assessed by two independent raters regarding both feasibility and quality on a scale from "extremely low" to "extremely high". As stated in [11] "[f]or feasibility, the raters had an agreement [. . . of] Cohen's $\kappa = 0.86$". The individual scores per recommendation were accumulated and finally averaged across all recommendations per group.
- **Mean_Q_Rec.** Each recommendation was assessed by two independent raters regarding both feasibility and quality on a scale from "extremely low" to "extremely high". As stated in [11] "[f]or quality, [the] two independent raters had an agreement [. . . of] Cohen's $\kappa = 0.83$". The individual scores per recommendation were accumulated and finally averaged across all recommendations per group.
- **High_Rec.** Based on the scores in F_Rec. and Q_Rec. the recommendations were considered highly feasible and high quality for each group. The current indicator sums the number of recommendations, achieving a score of either a four or five on either feasibility or quality ratings.

Regarding the introduced indicators, it is to be noticed "that preceding data analyses showed that the subjective and objective performance indicators are correlated. However, correlations are weak (below $r = 0.3$) [, but nevertheless statistically significant,] and indicate relationships, but no clear directional effects" [11].

In the experiments (at least in the beginning), all indicators were used for the development of prediction models (cf. Section 2.2). However, regarding the indicators MSBS_mean and ANX_mean for various groups, no, or fragmentary, assessments were provided, resulting in a large number of outliers. Therefore, the results achieved on these indicators have limited significance and should be considered with caution. Further details on this aspect are given in Sections 2.1.3 and 3.1.

#### 2.1.3. Limitation of the Data

Given the perspective of [44], in our experiments, the Parking Lot Corpus data can be considered secondary data since they are not self-collected, and inherent variables of the data collection were not influenced by the manuscript's authors. From discussions with the data provider, we can summarise that participants were not pre-selected according to any characteristics; the only restriction was that they were enrolled in the Midwestern United States university as already explained in Section 2.1.1. There were only two interventions by the experimenter: (1) the group leader was not elected by the group itself but by rolling a die, and (2) the discussion was limited in time for comparison reasons (cf. Section 2.1.1). For our experiments, the Parking Lot Corpus comprises appropriate conditions from an experimental design perspective. The corpus provides acoustic samples from group interactions that have characteristics of naturalistic interactions (as in [38]): no scripted interaction, no pre-defined wording, but rather spontaneous interaction of the group members. Further, we have direct access to a wide-spread collection of performance indicators (cf. Section 2.1.2), being used for the assessment of groups. Finally, given the recorded 70 group interactions, a suitable number of acoustic samples are available (for details, we refer to Section 2.2.1), allowing a training of (rather shallow) machine learning-based prediction models (cf. Sections 2.2.4 and 2.2.5). It is to be noticed that we did not contribute to neither the experimental design nor the data collection. Therefore, we used the provided material as such, except the exclusion of "outliers" as explained in Section 2.2.1.

Regarding the original intention of the data collection, studying performance indicators in group interactions from a work psychology or social science perspective, the experiments presented in the manuscript are established in a post hoc manner. We focus on an acoustic-based perspective on the groups' performance results, posing the research questions in Section 1.2 or as stated in the manuscript's title, "Whose Data Should be Used for Prediction of Team Performance?". Given the definition of [45], our experiments can be seen as a kind of *ex post facto* experiment since we interpret given material in a novel sense, especially in an acoustic way. We use the data to tap them for automatic analyses and (maybe) for future use in the human–machine interaction domain, as the "value of co-relational [ex post facto . . .] studies lies chiefly in their exploratory [. . .] character" [45].

However, the Parking Lot Corpus has also limitations; we focus on those affecting the current experiments. The manuscript's authors were not being able to influence neither the recording conditions nor the parameters of the data collection. Given the underlying task, namely coming up with ideas on the current parking situation on the campus, the used terms and phrases are mainly focused on this issue. This might restrict variation in terms of the used vocabulary and possible acoustic variations, being seen in spontaneous interactions. However, the situation reflects a common, task-oriented group interaction. Finally, only the pre-defined performance indicators can be considered in the experiments and, thus, the prediction models are constructed specifically for those indicators.

Generally, we were not involved in the data collection and design process, which results in rather a retrospective and external view of the collection setting, which can be seen as *ex post facto* [45]. We nevertheless state that the experimental design as well as the provided data fit our needs for the training and testing of prediction models. Also regarding the number of available acoustic samples, which is usually a question during the training of machine learning approaches, we argue that a suitable amount of data is provided. In particular, we use rather shallow networks and methods (cf. Sections 2.2.4 and 2.2.5) and thus, the need for data is reduced. For the investigations on group leaders, the Parking Lot Corpus contributes in total 5225 acoustic samples and 5222 acoustic samples with respect to self-restrictions (cf. Section 2.2.1). Regarding acoustic samples for the remaining groups, 9076 samples can be used, already taking into account self-restrictions.

### *2.2. Experimental Setting*

Based on the research questions stated in Section 1.2, we select an approach to handle the corpus' material and further choose prediction methods to assess the performance

indicators (cf. Section 2.1.2). For this, we distinguish two perspectives: (1) The data flow and how the data are generally processed to achieve a prediction—this is visualised in Figure 2 and discussed in Section 2.2.1. (2) The in-depth handling of the data to obtain a prediction—an overview of the experimental workflow is presented in Figure 3, whereas the approaches are discussed especially in Sections 2.2.4 and 2.2.5.

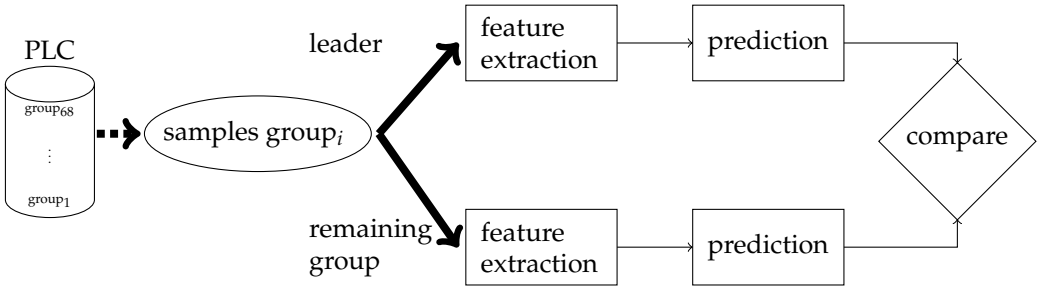


**Figure 2.** Flow of data in the prediction experiments. Data (taken from the Parking Lot Corpus (PLC); indicated by dotted arrow) per group are split into acoustic samples of the group's leader, and the remaining group is processed separately. Details on the workflow are visualised in Figure 3. Finally, a comparison of the performance indicators is conducted, currently on a manual base.

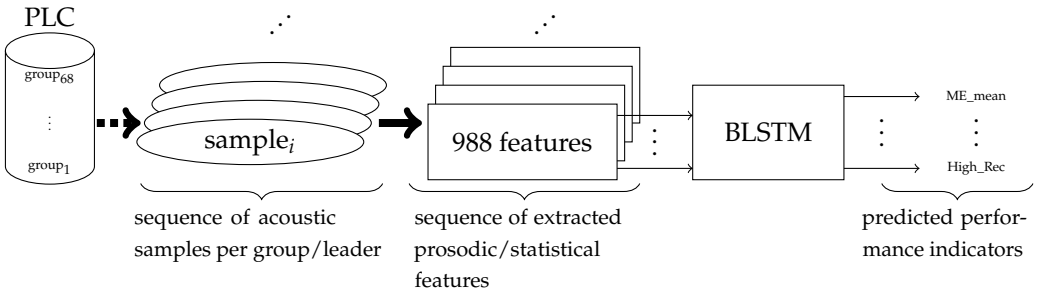


**Figure 3.** Workflow of the prediction. For each remaining group or leader from the Parking Lot Corpus (PLC; indicated by dotted arrow), a respective sequence of acoustic samples is used for prediction. In total, 988 features (*emobase* feature set) are extracted per sample *i* and fed to the bidirectional LSTM (BLSTM). The performance indicators are predicted simultaneously.

#### 2.2.1. Separation of Data

At first, we consider the material, which is available in the Parking Lot Corpus. The corpus comprises audio–visual recordings of 70 group sessions, where for each group, a group leader is chosen randomly. Relying in the experiments on the acoustic material, each group contributes a different number of speech acts/samples (i.e., the number of spoken statements). We neglect those groups with fewer than ten speech acts (self-restriction) since this can be assumed to be a minimum amount of material for making any adaptation in the prediction models. Therefore, the total number of available groups reduces to 68, comprising a total number of 14,298 acoustic samples. As suggested and used in [13], the entire material can be split into data from the leading person and those from the remaining participants (usually called remaining group in the manuscript). The flow of the data within the experiments is visualised in Figure 2.

*Leader Data*: Since the leading person has a distinct position in the group ([1] pp. 91–111 or [46] e.g., pp. 5–7), as she/he has the option for influencing and accentuating the behaviour and also the acoustics of this specific group, this member is of special interest. Therefore, we decided to analyse the acoustic statements of the group leader separately, applying, however, the same methods as for the remaining group (cf. Sections 2.2.4 and 2.2.5). To conduct the investigations, we split the entire Parking Lot Corpus' material, selecting those acoustic samples which were assigned to the respective group leader. This annotation was provided by the corpus' distributor, cross-checked for an arbitrary selection of samples

of the subset. Since we already cleaned the data set as mentioned above, the number of leaders remained at 68, providing a range of acoustic samples from 17 to 156 statements per speaker.

*Remaining Group*: The remaining group part contains the acoustic material of those participants who are collaborating within the current group but were not randomly selected as the group's leader. As discussed in, for instance, [46], the forming of a group is a dynamic process which is usually ongoing during the discussion. Although we are not focused on this interesting issue, we see that the forming process should not be neglected as already discussed in Section 1 and in [11,13]. However, in the current analyses, we focus rather on the influence of the acoustic statements to the group performance indicators, using prediction experiments (cf. Section 2.2.4 and 2.2.5) to obtain (ideally) objective assessments of group performance results in the future. Again, we cross-checked the split material on a random base and applied the same self-imposed restrictions as already described. This led to the same 68 remaining groups, contributing a range of acoustic samples from 55 to 274 statements. The reader should keep in mind that the absolute number is not directly related to any details of the group since the group's size varies between three and six participants, resulting in a remaining group size's range of two to five. Further, some groups are more interactive than others, which also influences the number of statements, additionally reflecting the discussion on "stimulating meetings" [11].

#### 2.2.2. Validation Paradigm and Measures

Based on the grouping on meta-level (cf. Section 2.2.1) and given the group information comprised by the corpus, we applied two validation strategies in our experiments, aiming for conclusions on generalisation or generalised perspectives.

*Leader Data*: For the group's leader, the validation strategy is based on an individual level since only the acoustic samples of one participant are used. As we split the corpus' material accordingly, we were able to apply a Leave-One-Speaker-Out (LOSO) paradigm. This means that all acoustic samples of one group leader are only used in the test set; the remaining material of the other group leaders is employed for the training. To ensure a smooth training of the models, the training set was further split into a real training set and a validation set (10% of training material), which allows an assessment of the training process. The whole process was repeated 68 times, utilising each group leader once as a test person.

*Remaining Group Data*: Regarding the validation of the groups' investigations, we decided on a similar approach as for the leader. Again, we opted for a general perspective (generalisation) in the prediction experiments. Given the material, the test and training sets were constructed as follows: for testing, the acoustic statements of a particular group were applied; the samples of the rest of the groups were merged to form the training set. To assess and control the training process, the training set was split in 10% validation and 90% training material. The whole process was repeated 68 times, utilising each remaining group once for testing. This resulted in a Leave-One-Group-Out (LOGO) validation paradigm.

*Validation Measure*: Although we predicted the performance indicator values (either point scale or counting values— being used for objective indicators like number of recommendations), we applied the Root Mean Square Error (RMSE) according to Equation (1) for the validation of the prediction experiments:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2}, \tag{1}$$

where $\hat{y}$ is the predicted output, $y$ is the expected output, and $n$ is the number of observations.

The RMSE was calculated for each leader and group, respectively. This allows an assessment of the differences between the predicted performance indicator values and the respective human assessment (cf. Section 3.1), being considered as a kind of ground truth. For an overall assessment of the current experiments, the individual RMSE values were

further averaged across all leaders or groups, which allows a generalised discussion on the results achieved with a particular parameter setting for the predictors and further enables a ranking of the achievements (the best performing parameter settings are highlighted in tables visualising experimental results).

*Statistical Significance*: The achieved results were also analysed regarding statistical significance. For this, we used the Kruskal–Wallis test (cf. [47]) with internal Bonferroni correction. To obtain a high significance, we selected $p < 0.05$ as the significance level. Analyses showed that $p < 0.01$ was reached when statistical significance was obtained (cf. Section 3.3.3).

#### 2.2.3. Extracted Features

Nowadays, two approaches for the development of features are distinguished, namely, hand-crafted and learnt features (e.g., [7,48]). In our experiment, we focused on the hand-crafted features since they provide an option to interpret the respective input used for prediction. In future research, these investigations can be extended to learnt features, and both approaches might be compared.

Features derived from the raw signal by the application of functions and operations are usually called hand-crafted since "manual" effort and additional (expert) knowledge are used to obtain those features. There is a long tradition of using such features, which is currently gaining attention again in the sense of explainability. Since the raw data are processed "manually", the obtained values are (usually) more easily interpreted as in the case of automatically learnt features. For hand-crafted features, we decided on the *emobase* feature set (cf. [33]), which is a well-established set in the community of affective computing from speech. In our experiment, this feature set is applied to acoustic samples of either the group leader or remaining group. Given the wide range of covered speech characteristics, comprising a balanced set of spectral and prosodic features, as well as use in multiple speech-related investigations (cf. e.g., [30,36,49]), we assume that *emobase* is also applicable in the current task (supported by the review in [7]).

The feature extraction is handled as follows: The features are extracted on the utterance level using a common windowing approach (usually a Hamming window), applying the openSMILE toolkit [36]. In total, *emobase* contains 988 features constructed from 52 Low Level Descriptors (LLDs) as well as 19 functionals applied to each LLD. The set of LLDs contains, for instance, Mel-Frequency Cepstral Coefficients, intensity, loudness, etc., as well as their respective delta values. The list of functionals includes amongst others mean, minimum, maximum, various quantiles, ranges, etc.

#### 2.2.4. Baseline Setting

The Parking Lot Corpus is a data set which is novel to both communities, social group investigations as well as computational group analysis and thus, is not yet frequently used. Recent analyses (cf. [11,13]) focused on the qualitative and quantitative acoustic-based assessments of group characteristics in relation to performance indicators. An experimental prediction of such indicators, based on acoustic material, has not yet been performed on the data. Therefore, we first conducted experiments which allowed to define a baseline for further comparison. The overall workflow of baseline experiments was adapted from those visualised in Figure 3, using an alternative machine learning approach.

For baseline experiments, we applied a rather simple approach using Support Vector Regression (SVR) using the implementation of the Python library `sklearn`. Besides the default parameters, the kernel function was set to the Radial Base Function. For the evaluation, the respective LOSO and LOGO paradigms (cf. Section 2.2.2) were applied, using the hand-crafted features introduced in Section 2.2.3. To fit the requirements of SVR, we "compressed" the features as follows: for each feature of the *emobase* features set [33], we calculated the respective mean values across all samples per leader or remaining group. This results in a representation of the leader or the remaining group in an $n$-dimensional

space, which can be used for SVR. The baseline results are presented and discussed in Section 3.2.

#### 2.2.5. Prediction Models

In the following, we introduce the methods used in the paper's main experiments as well as the respective parameters. An overview on the entire workflow is given in Figure 3. We try our very best to communicate the most relevant aspects to also achieve an option of reproducibility, already highlighting that the parameters not mentioned in the description are kept as the default.

In contrast to the baseline experiments, we applied neural networks for the prediction experiments on hand-crafted features. In general, since, on the one hand, the creation of a team is a dynamic process ([1] and discussion in Section 1), and, on the other hand, the interaction within a group is based on a sequence of interactions between group members (cf. Figure 3), we need an approach that is able to handle such dynamics. In the current experiments, we focused on the sequential characteristics of an interaction and thus, the interplay of communication partners needs to be tackled. Therefore, a recurrent neural network approach appears appropriate for this task. Currently, multiple state-of-the-art approaches are available, which might solve the issue, albeit often requiring large/big datasets. Given the Parking Lot Corpus with 14,298 acoustic samples, we decided to use a rather shallow realisation of recurrent neural networks. For this, in particular, we relied on Long-Short Term Memory (LSTM)-based networks (e.g., [50,51]), specifically, BLSTMs. Preliminary experiments show that especially the bidirectional characteristic of BLSTMs provides advantages for the prediction task since the context in the interaction (BLSTMs preserves past and future information) is beneficial for an assessment of the entire discussion. From our point of view, this allows a better ranking of the entire group's performance.

BLSTMs were implemented and trained using the `keras` framework [52] in `Python`. For the fitting process, the Adam optimiser was utilised based on the following parameters: learning rate $\eta = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-7}$. As the neuron's activation function, we selected the sigmoid function across all neurons. Regarding the number of layers, we used a shallow model, keeping it simple, and fixed the setting to two hidden layers (a BLSTM and a dense layer). Those network parameters, neither currently mentioned nor being varied in the experiments, were kept on the default setting provided by the `keras` framework and thus, not being reported in the manuscript. The parameters which are varied in the experiments are (1) the number of units (in LSTMs and BLSTMs, a unit is the fundamental processing (memory) cell, being a composition of basic (artificial) neural structures and trainable gates controlling the cell; for details, we refer to, for instance, [51]), where the range is indicated in Tables 3 and 4; and (2) the number of training epochs (in the current experiments, either 50,000 or 100,000 epochs). Furthermore, we implemented early stopping on the validation loss. Although, in general, a LOSO or LOGO validation paradigm is applied for the training of the models, for internal evaluation during the individual training process, an internal validation set (size: six leaders or remaining groups, respectively) is randomly selected. All input values are normalised using the L2-norm.

Given the validation paradigm, we trained individual models for each leader or remaining group as well as for each performance indicator to be predicted. As discussed finally, the average performance predictions were calculated for a general discussions of the results.

## 3. Results

In this section, we present the results and give a first discussion of the achievements. A discussion in a broader sense is given in Section 4, especially with respect to external results and stimulating meetings. We divide the presentation into baseline (cf. Section 3.2) and prediction results (cf. Section 3.3). This allows approaching the Parking Lot Corpus with respect to an automatic prediction of performance indicators as well as a more detailed investigation with state-of-the-art neural predictors.

### *3.1. Human Assessments of Performance Indicators*

Before we dive into the particular results, we briefly introduce the human annotations (can also be called human assessment) for the performance indicators. This refers to the aspect that the group meetings were assessed by at least two qualified raters [11], providing mainly an external view of the groups and their performance. The average performance indicators of humans across all groups are given in Table 1. These are used as the ground truth for both the training of the prediction models as well as the comparison in our experiments, where no distinction is made whether the leader or remaining group is investigated since the respective performance is evaluated for the entire group (cf. also [11]).

Hint: Regarding the individual values in Table 1, the results for ANX (mean) and MSBS (mean) are considered outliers. During the human assessment, at least one human expert did not rate either ANX (seven groups are affected) or MSBS (four groups are affected). To solve and highlight this issue, the corpus distributor marked these events by a value of −99. Computing the mere mean across all groups per performance indicator, this results in negative average values for ANX and MSBS (cf. Table 1). Neglecting those particular groups generally for all performance indicators across all prediction experiments would further reduce the number of samples for those indicators not affected by annotation issues. For this, we decided to discard only the two performance indicators in further experiments, although for the majority of groups, human assessments would be available.

**Table 1.** Human rater assessments for each performance indicator based either on a point scale or on counting values (cf. Section 2.1.2). For this table, the respective values are averaged across all groups. For details on the values of ANX and MSBS we refer to explanations in Section 3.1.

| | **MSA** | **ME** | **TS_Rec** | **F_Rec** | **Q_Rec** | **High_Rec** | **ANX** | **MSBS** | **MSP** | **MSO** |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 3.8661 | 3.9676 | 7.3529 | 3.1503 | 2.3676 | 3.0588 | −8.7538 | −3.6194 | 4.6346 | 6.037 |

### *3.2. Baseline Results*

As already mentioned in the introduction (cf. Section 1) and the description of the baseline architecture (cf. Section 2.2.4), the Parking Lot Corpus is a rather novel data set. Therefore, we still lack baseline experiments to compare the current achievements. This is being resolved now, giving first indications for predictions using a rather simple statistical approach and further providing results utilising state-of-the-art models (cf. Section 3.3).

Table 2 presents the baseline results, relying on SVR experiments estimating respective predictions values, providing the predicted mean performance per indicator (either point scale or counting value). Therefore, the results can be directly compared to the human assessments in Table 1.

**Table 2.** Baseline results per performance indicator using SVR, distinguishing the performance estimated on the acoustic material of either the leader or the remaining group (Rem_Group). The results are averaged across all groups, representing the automatically obtained prediction values in terms of either point scale or counting value (cf. Section 2.1.2).

| | **MSA** | **ME** | **TS_Rec** | **F_Rec** | **Q_Rec** | **High_Rec** | **MSP** | **MSO** |
|---|---|---|---|---|---|---|---|---|
| Leader | 3.8967 | 3.9603 | 6.9698 | 3.4507 | 2.5192 | 2.9021 | 6.2361 | 6.1471 |
| Rem_Group | 3.8707 | 3.9859 | 7.0005 | 3.4298 | 2.5459 | 2.9069 | 6.2678 | 6.1572 |

Regarding the achieved results and comparing them to the human gold standard (cf. Table 1), we see already good results for both experimental settings. This is also supported by the RMSE values, measuring the differences between the human assessed performance indicators and predicted ones. It is to be noticed that we decided to avoid stating individual RMSE values for clarity of presentation; however, if needed, the values can be calculated according to Equation (1).

The indicators can be grouped by low, medium, and high RMSEs. Comparing Tables 1 and 2, we see that low RMSE values appear for MSA and ME; medium values are given for TS_Rec, F_Rec, Q_Rec, High_Rec, and MSO; and high values are seen for MSP. Given these baseline results, we see options for improvement, especially in the medium and high RMSE indicators. Therefore, we ran additional experiments, utilising higher-level prediction methods as introduced in Section 2.2.5.

### *3.3. Results of Prediction Models*

For more advanced prediction models, we relied on BLSTM networks as introduced in Section 2.2.5. This approach is usually known as a good option to predict sequences, especially using long-term dependencies. From the experimental results, we learnt that investigations on leader and remaining group samples are highly different and thus, separated observations are recommended.

#### 3.3.1. Leader

In Table 3, the prediction results, providing predicted mean performance per indicator (either point scale or counting value; cf. Section 2.1.2), using BLSTM-based networks, are visualised, varying the number of units being used in the networks and the number of training epochs. The highlighted cells indicate those settings which show the best results, relying on the RMSE, throughout the prediction experiment per performance indicator. Given the results, we see a spread across the networks' parameters, obtaining the best performance per indicator (respective column in Table 3). However, there are two settings, namely 750 and 1000 units, that provided the best performance results for half of the indicators. Nevertheless, no clear effect of the network setting could be identified in relation to subjective or objective performance indicators.

**Table 3.** Results per performance indicator of the group's leader using BLSTMs varying the number (#) of units utilised in the respective network. The results are presented as either point scale values or counting values (cf. Section 2.1.2). The grey cells highlight the best result per performance indicator, using the RMSE as the decision's foundation. † represents the network being trained for 50,000 epochs instead of 100,000 (default).

| # Units | MSA | ME | TS_Rec | F_Rec | Q_Rec | High_Rec | MSP | MSO |
|---|---|---|---|---|---|---|---|---|
| 500 † | 1.1282 | 1.1405 | 4.7781 | 1.0700 | 0.8042 | 1.6120 | 4.7134 | 3.1978 |
| 500 | 1.0725 | 1.2364 | 4.8481 | 0.9946 | 0.9004 | 1.5873 | 4.8736 | 3.2662 |
| 750 | 1.0925 | 1.1022 | 3.2831 | 1.9346 | 2.5412 | 2.4506 | 2.6557 | 1.1749 |
| 1000 | 3.7664 | 3.7353 | 2.7967 | 4.5159 | 5.2854 | 4.8647 | 2.9846 | 1.7951 |
| 1250 | 6.9774 | 6.7886 | 4.1614 | 7.6412 | 8.3177 | 7.8256 | 6.1693 | 4.6949 |
| 1500 | 10.4070 | 10.0766 | 7.0673 | 11.0224 | 11.7454 | 11.0756 | 9.4374 | 8.1379 |
| 2000 | 17.2842 | 17.2160 | 13.8987 | 18.1328 | 18.8412 | 18.2975 | 16.6489 | 15.2709 |

There are two interesting observations.

At first, regarding Table 3, we saw that the prediction performance varies with the number of BLSTM units, which is somehow expected. In the first instance, we assumed that rather small networks might be able to solve the task already. But we saw that this depends on the number of epochs (cf. rows 1 and 2 in Table 3), which indicates that the systems need quite a while to learn the necessary dependencies. To provide the network more flexibility to learn the characteristics of group interactions, we increased the number of BLSTM units, showing, by contrast, that already a limited range is appropriate to handle the prediction task per performance indicator. All indicators could be tackled using network settings in the range of 500 to 1250 units. Already with more than 1500 units, we observed a drastic decline in prediction power, increasing with increased unit numbers.

Given these achievements, we additionally ran similar experiments, also varying the number of epochs and even the architecture to LSTMs. No evidence for improvement or

difference in the prediction could be seen. Therefore, we decided to neglect the results in this manuscript.

Second, comparing the prediction performance in Tables 2 and 3 to the human annotations, no improvement could be achieved across all indicators, except MSP. For this specific performance indicator, a significant improvement could be achieved. Taking the argumentation of [11] into account, that especially MSP is related to short-term speech characteristics, we saw the benefit of BLSTMs handling short-term dependencies in relation to the statistical SVR approach in the baseline. For the remaining indicators, the variation in spoken acoustic/prosodic characteristics is on a rather long-term indication, which can be already fetched by the statistical model using fewer parameters. Additionally, it seems that especially for MSA and ME, BLSTMs is also able to handle the larger variations in prosody (cf. respective analyses in [11]), although they currently do not beat the baseline achievements. However, this is a matter of further research, investigating the particularly affected network parameters in the sense of explainable AI.

#### 3.3.2. Remaining Group

Regarding the results predicting the performance using the material of the remaining group only (cf. Table 4, also presented as predicted mean performances per indicator (either point scale or counting value); cf. Section 2.1.2), we saw a more diverse spread across the parameters (in this case, number of units and number of training epochs). The results vary more than those for the leader-only experiments (cf. Table 3) since a broader spectrum of speaker characteristics needs to be covered. Interestingly, the predictions for F_Rec, Q_Rec, and High_Rec are condensed in a particular setting, namely 750 units. In general, the BLSTM models show performance peaks, being different from those seen in the leader setting in Table 3. This is related to the diversity in the samples since the models need to learn a higher variations in acoustics since the variation across the members per remaining group has to be modelled. Given these results and also the analyses in [11], the complexity in the specific performance indicators can be assumed.

However, comparing the human annotations (cf. Table 1) and baseline results (cf. Table 2) to the achievements of the current models, we see an improvement in the performance for the indicators MSA, ME, TS_Rec, and MSP. Therefore, using material of the remaining group members could be a benefit for the prediction.

**Table 4.** Results per performance indicator of the remaining group using BLSTMs varying the number (#) of units utilised in the respective network. The results are presented as either point scale values or counting values (cf. Section 2.1.2). The grey cells highlight the best result per performance indicator, using the RMSE as the decision's foundation. † represents the network being trained for 50,000 epochs instead of 100,000 (default).

| # Units | MSA | ME | TS_Rec | F_Rec | Q_Rec | High_Rec | MSP | MSO |
|---|---|---|---|---|---|---|---|---|
| 500 † | 1.0896 | 1.2002 | 4.6993 | 1.1273 | 0.8819 | 1.6044 | 4.9064 | 3.1810 |
| 500 | 1.1085 | 1.2401 | 4.6161 | 1.0401 | 1.0161 | 1.5704 | 4.9151 | 3.2914 |
| 750 | 1.1886 | 1.0012 | 3.3591 | 1.8171 | 2.6134 | 2.6469 | 2.8615 | 1.1118 |
| 1000 | 3.8181 | 3.7263 | 2.8752 | 4.5782 | 5.3608 | 5.0242 | 3.0933 | 1.6040 |
| 1250 | 6.9038 | 6.7865 | 4.2911 | 7.4682 | 8.2923 | 7.8843 | 6.1861 | 4.7702 |
| 1500 | 10.1821 | 10.2862 | 7.0843 | 10.9336 | 11.7347 | 10.9835 | 9.5881 | 8.0243 |
| 2000 | 17.3228 | 17.1411 | 13.9539 | 18.1687 | 19.1816 | 18.1517 | 16.7636 | 15.2154 |

#### 3.3.3. Statistical Comparison Leader vs. Remaining Group

Driven by the improvements being achieved by the models on the remaining groups' material, we also calculated the statistical significance for the results. For this, we compared the predictions of leader and remaining groups across performance indicators. Using the Kruskal–Wallis test [47] (further details in Section 2.2.2), applying a significance level of $p < 0.05$, we saw two results: on the one hand, if statistical significance was achieved, this was obtained at the $p < 0.001$ level; in particular, this is the case for TS_Rec, F_Rec,

Q_Rec, and High_Rec. On the other hand, for the other performance indicators, no statistical significance can be seen. The calculated *p*-values are far from the pre-defined significance level.

## 4. Discussion

The following discussion is tailored to the research questions and the hypothesis stated in Section 1.2. Furthermore, we embed our findings in a broader discussion already started in [7,11].

*Research Question RQ1:* Regarding the first research question, we establish the baseline results on the Parking Lot Corpus. Since this is a rather novel data collection, these results (cf. Table 2) are the first automatic prediction values using acoustic material. So far, the data were analysed rather in psychological and social backgrounds, providing human-based assessments. The human annotations in Table 1 are average values across at least two certified annotators (cf. Sections 2.1.2 and 2.2.4, and [11]), which can be seen as ground truth for the upcoming investigations.

Given a rather simple SVR approach, good predictions were already able to be achieved. These results indicate that predictions of performance per indicator, based on acoustics, can be set up. Further, if compared to the results of higher-level neural approaches, reasonable results were obtained. This shows that already "simple" models are able to retrieve the necessary details, providing first-level predictions. These are also being related to human assessments.

*Research Question RQ2:* To answer the second research question, we selected specific networks, namely BLSTM, which fit the aspect of temporal handling (further details are given in Section 2.2.5). Similar to the baseline experiments, human annotations are being used as ground truth during the evaluation of models. As shown in Section 3.3, no clear preferences for particular network settings could be shown. In both scenarios of leader or remaining group, a variation across the network settings is given for the performance indicators. Some benefits can be seen, especially in the recommendation-related indicators. However, this is not ultimately fixed and is rather a pointer towards further investigations.

The results achieved by using acoustic material from the remaining groups are more promising. We see that for half of the indicators, an improvement against the baseline is achieved. Comparing those achievements also to the leaders' results, statistical significance is reached for some indicators. However, the current results are the first findings unlocking the treasure of the Parking Lot Corpus. Detailed analyses need to be conducted to show why particular performance indicators gained improvements in the neural approach. A first interpretation might be that the recommendation part directly corresponds to specific acoustic characteristics. This is in line with the findings of [11], showing that this particular characteristic is spread across the group rather than being established in the leader's acoustics. In contrast, the decrease in the remaining performance indicators shows that neural approaches need well-selected tasks. The interplay of statistical features (given in the *emobase* features set [33]) and a quite robust prediction method already cover the characteristics. More detailed investigation on the specific acoustic variation responsible in the neural approaches and the analyses in [11] is necessary in the future research.

*Hypothesis 1*: In general, we expected an improvement in the prediction performance using networks, with the capability to model temporal characteristics and incorporate context. It should be possible to model a development in a contextual sense, especially for the remaining group; for the leader, without any group context, this is much harder. However, given the Parking Lot Corpus, this was not the case for BLSTM models and all performance indicators. We obtained improvements for some indicators, especially those related to the recommendations. Given these results, further investigations are necessary to clarify reasons for the achievements. This can be related to either the data itself, the underlying course of interaction, the models' fine tuning, or the utilised features, directly asking for interdisciplinary collaborations to assess the respective characteristics and issues.

*General Discussion*: In general, we provide the first prediction results for the performance of groups in the Parking Lot Corpus based on performance indicators (cf. Section 2.1.2). In this sense, the paper taps the corpus and also provides additional insights into an automatic assessment of performance predictions, aiming for an objective evaluation in the future. What we see from the data and, thus, from the results shows two different indications.

On the one hand, usually the leader has an important role in the meeting (e.g., [1,46]), but given the acoustic analyses, this is currently rather not being seen. In contrast, the prediction approaches are in fact confused by the particular material, especially in the neural-based approach. On the other hand, in contrast, the remaining group provides reasonable acoustic evidence for a prediction of performance, in both cases, baseline and BLSTMs. In this sense, the "holistic" view on the group enables the system to establish an understanding of the communication, leading to an option of performance assessment. The combination of understanding the ongoing communication and a link to contextual ("holistic") information enables not only the evaluation of the current performance but is also related to aspects like familiarisation. The authors in [9] see familiarisation as an essential part of communication (amongst humans only or in human–machine interaction), especially in "open-world" scenarios (cf. naturalistic interaction in [38]), helping to improve task solving in further interactions. This is also in line with the arguments presented in [11] and Section 1, where already the benefit of a "stimulating meeting" was introduced. Such meetings use the option to familiarise and thus bet on upcoming meetings. However, this enforces more fine-graded analyses in both social as well as computer sciences, longing for respective interdisciplinary research. The current work provides some steps toward such interpretation and combined/interdisciplinary investigations.

## 5. Conclusions

The current manuscript analysed the Parking Lot Corpus and the performance indicators assigned to the group interactions. We presented first acoustic-based prediction results on the corpus, being divided into baseline results and achievements using neural approaches. In particular, the baseline setting used statistical features derived from the *emobase* feature set [33] and SVR for prediction. In addition to this, in further experiments, BLSTMs were applied, utilising the *emobase* feature set [33], and were compared to the baseline results. Improvements in the prediction performance could be achieved only for parts of the indicators (cf. Section 3.3), showing some benefit of higher-level neural networks. The achievements are discussed in Section 4, also highlighting the relation of the investigations of [11].

Particularly, we summarise the findings in term of takeaway messages:

- Baseline results of performance indicators for the Parking Lot Corpus based on acoustic samples and statistical features were provided.
- Performance prediction per indicator was contributed based on acoustic samples, utilising the *emobase* feature set [33], and BLSTMs.
- Comparison of baseline and BLSTM-based results.
- BLSTM prediction on remaining group level is beneficial for particular objective performance indicators.
- Regarding whose acoustic samples (leader or remaining group) might be used for prediction tasks, currently a (slight) trend towards the remaining group's samples is given.
- Contextual information (as in the remaining group) is beneficial for an improvement in the performance predictions (related to (theoretical) discussions in [9,11,13]).

In future research, additional features might be tested in relation to neural approaches but also linking those results to meta-discussions as already started in Section 4 (General Discussion). In particular, the familiarisation process, being part of task-solving interactions (cf. [9]), is an important aspect of establishing a better understanding of contextual relationships during an interaction and how this is being linked to the assessment of

group performance results. This establishes further collaborations with and relations to the social sciences.

**Funding:** This research was partly funded by the Federal State of Saxony-Anhalt, Germany under the grant number: I 138 (project "ASAMI").

**Institutional Review Board Statement:** Not applicable. Data is secondary data as stated in the respective Section.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Data is available upon request by the data distributor.

**Acknowledgments:** I thank Joseph A. Allen for providing the recordings and ratings as well as for the valuable discussions regarding the performance indicators. Further, I acknowledge current support by Genie Enterprise as well as previous support by the project "ASAMI".

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Brown, R. *Group Processes—Dynamics within and between Groups*; Blackwell Publishers: Oxford, UK, 2000.
2. Barsade, S.G.; Gibson, D.E. Group Affect: Its Influence on Individual and Group Outcomes. *Curr. Dir. Psychol. Sci.* **2012**, *21*, 119–123. [CrossRef]
3. Cohen, M.A.; Rogelberg, S.G.; Allen, J.A.; Luong, A. Meeting design characteristics and attendee perceptions of staff/team meeting quality. *Group Dyn. Theory Res. Pract.* **2011**, *15*, 90–104. [CrossRef]
4. Levi, D. *Group Dynamics for Teams*; SAGE: Los Angeles, CA, USA, 2015.
5. Rogelberg, S.G.; Allen, J.A.; Shanock, L.; Scott, C.; Shuffler, M. Employee satisfaction with meetings: A contemporary facet of job satisfaction. *Hum. Resour. Manag.* **2010**, *49*, 149–172. [CrossRef]
6. Böck, R. *Anticipate the User: Multimodal Analyses in Human-Machine Interaction towards Group Interactions*; TUDpress: Dresden, Germany, 2020.
7. Böck, R. Affects in Groups: A review on automated affect processing and estimation in groups. *IEEE Signal Process. Mag.* **2021**, *38*, 74–83. [CrossRef]
8. Allen, J.; Lehmann-Willenbrock, N.; Rogelberg, S. *The Cambridge Handbook of Meeting Science*; Cambridge Handbooks in Psychology; Cambridge University Press: Cambridge, UK, 2015.
9. Böck, R.; Wrede, B. Modelling Contexts for Interactions in Dynamic Open-World Scenarios. In Proceedings of the 2019 IEEE International Conference Systems, Man, and Cybernetics, Bari, Italy, 6–9 October 2019; IEEE: New York, NY, USA, 2019; pp. 1475–1480.
10. Vinciarelli, A.; Pantic, M.; Boulard, H. Social Signal Processing: Survey of an Emerging Domain. *Image Vis. Comput.* **2009**, *12*, 1743–1759. [CrossRef]
11. Niebuhr, O.; Böck, R.; Allen, J.A. On the Sound of Successful Meetings: How Speech Prosody Predicts Meeting Performance. In Proceedings of the 23rd ACM International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 240–248
12. Kozlowski, S.W.J.; Gully, S.M.; Nason, E.R.; Smith, E.M. Developing adaptive teams: A theory of compilation and performance across levels and time. In *The Changing Nature of Performance: Implications for Staffing, Motivation, and Development*; Ilgen, D.R., Pulakos, E.D., Eds.; Wiley: Hoboken, NJ, USA, 1999; pp. 240–292.
13. Böck, R. Times and Turns in Stimulating Meetings. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*; Niebuhr, O., Lundmark, M.S., Weston, H., Eds.; TUDpress: Dresden, Germany, 2022; pp. 89–96.
14. Radnor, Z.; Barnes, D. Historical analysis of performance measurement and management in operations management. *Int. J. Product. Perform. Manag.* **2007**, *56*, 384–396. [CrossRef]
15. Wiltshire, T.J.; van Eijndhoven, K.; Halgas, E.; Gevers, J.M.P. Prospects for Augmenting Team Interactions with Real-Time Coordination-Based Measures in Human-Autonomy Teams. *Top. Cogn. Sci.* **2022**, *0*, 1–39. [CrossRef]
16. Treadwell, T.; Lavertue, N.; Kumar, V.K.; Veeraraghavan, V. The Group Cohesion Scale-Revised: Reliability and validity. *Int. J. Action Methods Psychodrama Ski. Train. Role Play.* **2001**, *54*, 3–12.
17. Tskhay, K.O.; Rule, N.O. Accuracy in Categorizing Perceptually Ambiguous Groups: A Review and Meta-Analysis. *Personal. Soc. Psychol. Rev.* **2013**, *17*, 72–86. [CrossRef]
18. Delaherche, E.; Chetouani, M.; Mahdhaoui, A.; Saint-Georges, C.; Viaux, S.; Cohen, D. Interpersonal Synchrony: A Survey of Evaluation Methods Across Disciplines. *IEEE Trans. Affect. Comput.* **2012**, *3*, 349–365. [CrossRef]
19. Farley, S. Nonverbal reactions to an attractive stranger: The role of mimicry in communicating preferred social distance. *J. Nonverbal Behav.* **2014**, *38*, 195–208. [CrossRef]
20. Bartolo, K.; Furlonger, B. Leadership and job satisfaction among aviation fire fighters in Australia. *J. Manag. Psychol.* **2000**, *15*, 87–93. [CrossRef]

21. Peterson, M.D.; Dodd, D.J.; Alvar, B.A.; Rhea, M.R.; Favre, M. Undulation Training for Development of Hierarchical Fitness and Improved Firefighter Job Performance. *J. Strength Cond. Res.* **2008**, *22*, 1683–1695. [CrossRef] [PubMed]
22. Chen, Z.; Zhu, J.; Zhou, M. How does a servant leader fuel the service fire? A multilevel model of servant leadership, individual self identity, group competition climate, and customer service performance. *J. Appl. Psychol.* **2015**, *100*, 511–521. [CrossRef]
23. Yoerger, M.; Allen, J.A.; Crowe, J. The Impact of Premeeting Talk on Group Performance. *Small Group Res.* **2018**, *49*, 226–258. [CrossRef]
24. Mroz, J.E.; Allen, J.A.; Verhoeven, D.C.; Shuffler, M.L. Do We Really Need Another Meeting? The Science of Workplace Meetings. *Curr. Dir. Psychol. Sci.* **2018**, *27*, 484–491. [CrossRef]
25. Lehmann-Willenbrock, N.; Allen, J.A. How fun are your meetings? Investigating the relationship between humor patterns in team interactions and team performance. *J. Appl. Psychol.* **2014**, *99*, 1278–1287. [CrossRef]
26. Oertel, C.; Castellano, G.; Chetouani, M.; Nasir, J.; Obaid, M.; Pelachaud, C.; Peters, C. Engagement in Human-Agent Interaction: An Overview. *Front. Robot. AI* **2020**, *7*, 92. [CrossRef]
27. Ottl, S.; Amiriparian, S.; Gerczuk, M.; Karas, V.; Schuller, B. Group-Level Speech Emotion Recognition Utilising Deep Spectrum Features. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 25–29 October 2020; ACM: New York, NY, USA, 2020; pp. 821–826.
28. Franzoni, V.; Biondi, G.; Milani, A. Emotional sounds of crowds: Spectrogram-based analysis using deep learning. *Multimed. Tools Appl.* **2020**, *79*, 36063–36075. [CrossRef]
29. Sreenivas, V.; Namdeo, V.; Kumar, E.V. Group based emotion recognition from video sequence with hybrid optimization based recurrent fuzzy neural network. *J. Big Data* **2020**, *7*, 56. [CrossRef]
30. Schuller, B. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
31. Schuller, B.; Vlasenko, B.; Eyben, F.; Wöllmer, M.; Stuhlsatz, A.; Wendemuth, A.; Rigoll, G. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Trans. Affect. Comput.* **2010**, *1*, 119–131. [CrossRef]
32. Zhang, Z.; Weninger, F.; Wöllmer, M.; Schuller, B.W. Unsupervised learning in cross-corpus acoustic emotion recognition. In Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding Workshop, Waikoloa, HI, USA, 11–15 December 2011; pp. 523–528.
33. Eyben, F.; Wöllmer, M.; Schuller, B. openSMILE—The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 2010 ACM Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
34. Lefter, I.; Rothkrantz, L.J.M.; Wiggers, P.; van Leeuwen, D.A. Emotion Recognition from Speech by Combining Databases and Fusion of Classifiers. In Proceedings of the Text, Speech & Dialogue, Brno, Czech Republic, 6–10 September 2010; Sojka, P., Horák, A., Kopecek, I., Pala, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6231, pp. 353–360.
35. Lefter, I.; Nefs, H.T.; Jonker, C.M.; Rothkrantz, L.J.M. Cross-corpus analysis for acoustic recognition of negative interactions. In Proceedings of the 2015 IEEE International Conference on Affective Computing and Intelligent Interaction, Xi'an, China, 21–24 September 2015; pp. 132–138.
36. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 835–838.
37. Murray, G.; Oertel, C. Predicting Group Performance in Task-Based Interaction. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 14–20.
38. Valli, A. The design of natural interaction. *Multimed. Tools Appl.* **2008**, *38*, 295–305. [CrossRef]
39. Leach, D.J.; Rogelberg, S.G.; Warr, P.A.; Burnfield, J.L. Perceived meeting effectiveness: The role of design characteristics. *J. Bus. Psychol.* **2009**, *24*, 65–76. [CrossRef]
40. Briggs, R.; de Vreede, G.J.; Reinig, B. A theory and measurement of meeting satisfaction. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003; p. 8.
41. Fahlman, S.A.; Mercer-Lynn, K.B.; Flora, D.B.; Eastwood, J.D. Development and validation of the multidimensional state boredom scale. *Assessment* **2013**, *20*, 68–85. [CrossRef]
42. Heimberg, R.G.; Becker, R.E. *Cognitive-Behavioral Group Therapy for Social Phobia: Basic Mechanisms and Clinical Strategies*; Guilford Publications: New York, NY, USA, 2002.
43. Spitzer, R.L.; Kroenke, K.; Williams, J.B.; Löwe, B. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch. Intern. Med.* **2006**, *166*, 1092–1097. [CrossRef] [PubMed]
44. Leedy, P.D.; Ormrod, J.E. *Practical Research—Planning and Design*, 11th ed.; Pearson: Essex, UK, 2016.
45. Cohen, L.; Manion, L.; Morrison, K. *Research Methods in Education*, 5th ed.; Routledge Falmer: London, UK, 2005.
46. Baron, R.; Kerr, N.; Miller, N. *Group Process, Group Decision, Group Action*; Brooks/Cole: Pacific Grove, CA, USA, 1992.
47. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
48. Schuller, D.M.; Schuller, B.W. A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice. *Emot. Rev.* **2021**, *13*, 44–50. [CrossRef]
49. Böck, R.; Egorow, O.; Höbel-Müller, J.; Requardt, A.F.; Siegert, I.; Wendemuth, A. Anticipating the User: Acoustic Disposition Recognition in Intelligent Interactions. In *Innovations in Big Data Mining and Embedded Knowledge*; Esposito, A., Esposito, A.M., Jain, L.C., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 203–233.

50. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
51. Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*; Kolen, J.F., Kremer, S.C., Eds.; IEEE Press: New York, NY, USA, 2001.
52. Chollet, F. (Original Author). Keras. 2015. Available online: https://keras.io (accessed on 25 April 2023).