On the Sound of Successful Meetings: How Speech Prosody Predicts Meeting Performance

OLIVER NIEBUHR, University of Southern Denmark
RONALD BÖCK, Otto von Guericke University
JOSEPH A. ALLEN, University of Utah

This paper investigates the degree to which meeting success can be predicted through holistic, acoustic-prosodic measurements. The analyzed meetings are taken from the Parking Lot Corpus in which 70 groups of three to six students discuss the traffic situation at their university and come up with parking and transportation recommendations. The number, feasibility, and quality of these recommendations as well as the mean effectiveness and satisfaction ratings across group members provide the basis for correlations with three sets 15 acoustic-prosodic features that cover pitch, duration/timing, intensity, and the absolute frequencies of local events such as silent pauses. Results show that meeting success is, in fact, considerably correlated with the overall "sound" of the individual meetings, with pitch features being the most diverse and powerful predictors. In addition, we found that the "sound" of subjectively effective meetings differs from the "sound" of objectively productive meetings, i.e. meetings that generate a high output of feasible and/or high-quality recommendations. The prosodic feature patterns suggest that effective meetings are short and matter-of-fact, whereas the productive meetings are longer and have a lively speech melody that makes these meetings stimulating. We discuss the implications of our findings for future research and technological innovation.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Collaborative interaction**;

Additional Key Words and Phrases: speech prosody, interaction, pitch, intensity, duration, team performance, focus group

ACM Reference Format:
Oliver Niebuhr, Ronald Böck, and Joseph A. Allen. 2021. On the Sound of Successful Meetings: How Speech Prosody Predicts Meeting Performance. 1, 1, Article 4 (September 2021), 14 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Gathering with other people is a common habit of human beings, resulting in interactions and speech communication. Types of gatherings range from informal meetings with family members or friends to formal business meetings. There are more than 55 million workplace meetings every day in the United States alone (cf. [29]). That is, business meetings are ubiquitous with organizational life. Moreover, all types of meetings usually serve a specific goal that influences the way of communication and the interaction style (cf. e.g. [11, 50]). Further, the meeting specific goal and the interactions therein create expectations on the part of the meeting members in terms of results or outcomes (cf. [2]). 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. [30]).

Authors' addresses: Oliver Niebuhr, University of Southern Denmark, Street, Sonderborg, Denmark, olni@sdu.dk; Ronald Böck, Otto von Guericke University, Street, Magdeburg, Germany, ronald.boeck@ovgu.de; Joseph A. Allen, University of Utah, Street, Salt Lake City, USA, joseph.a.allen@utah.edu.

Unlike meetings with family or friends, the outcome of business meetings has an enormous relevance (cf. [3]) and is often measured in terms of group performance (cf. [30]). High group performance is linked to factors like synchrony and cohesion (for references we refer to Section 1.1), but also to affects (cf. [6, 8]), humour (cf. e.g. [31]), and interpersonal relationships (cf. e.g. [44]). The more important interpersonal relationships in a team, the more the dependencies between the members become crucial. Extreme examples are the NASA Mars base scenario (cf. [45]) or fire department settings (cf. e.g. [42]).

The interpersonal relationships are also influenced by the (perceived) personality of the team members (cf. e.g. [34, 43]), most prominently by that of the team leader. This subtle interplay between the team members needs to be considered in group formation processes (cf. e.g. [30, 32]). Multimodal assessments of team members can be conducted to reveal, for instance, accommodation processes (cf. e.g. [18, 22]) or the nonverbal communication patterns in teams (cf. e.g. [23]). However, maybe also the participants' voices contain indicators that show their ability to *stimulate* interaction. Based on this idea (cf. Section 1.2), we analyzed group meetings (cf. Section 2) and correlated team performance measures (to be presented in Section 2.2) with acoustic-prosodic parameters that are related to stimulation.

1.1 Related Work

In [8], pp. 40–45, the author presents an overview of communication in group interaction, reflecting both the social as well as the technical perspectives and considering multimodal and acoustic investigations. Especially for team performance in the sense of Kozlowski and colleagues [30], the team's internal dynamics is important. It is related to a team's cohesion as well as to its synchrony. Team cohesion refers to the sense of togetherness, whereas synchrony describes the temporal alignment of activities or actions within a team. Measurement and categorization concepts are provided by [47, 49] for cohesion and, similarly, by [17, 21] for synchrony. Beyond relationships within teams and team performance (cf. e.g. [7, 42]), the cohesion and synchrony of teams also to some degree determine the quality of customer support (cf. e.g. [13]) and job satisfaction (cf. e.g. [44]). Investigations on estimated team outcomes are, according to [52], influenced by "dynamics [. . .] (cf. e.g. [32, 52])[, . . . meeting] satisfaction (cf. [44]), handling of errors (cf. [39]), humour (cf. [31]), and group emotions (cf. [6])" [8].

From a technical perspective, various aspects of team dynamics and performance have been addressed, aiming at developing automatic assessment procedures. An overview of automatic processing approaches based on patterns of nonverbal behavior is provided by [23, 36], amongst others. The study of [36] provides a brief overview of relevant acoustic patterns, see also [15, 20] for supplementary features related to overlaps of speech turns bewteeen human speakers as well as between humans and technical systems. For interactions among human team members, an overview is presented in, for instance, [1, 35, 46, 48].

However, to the best of our knowledge, automatic analyses of stimulating speech in relation to team performances are not conducted as yet.

1.2 Research Objectives: Stimulating Meetings

The main objective of the current research is to determine whether or not "the sound of a meeting" already provides indicative information on the performance of the group members participating in a meeting.

For this, we reviewed the literature (cf. Section 1.1 for a brief overview) and found that most studies focus on other aspects of language than speech acoustics, and prosody in particular [23]. These other aspects are, for example, turn-taking behavior or patterns of gesture and mimic, or the degree to which group members use the same wording or morphosyntactically defined styles [24, 27]. And those studies that do look at speech acoustics and/or speech prosody

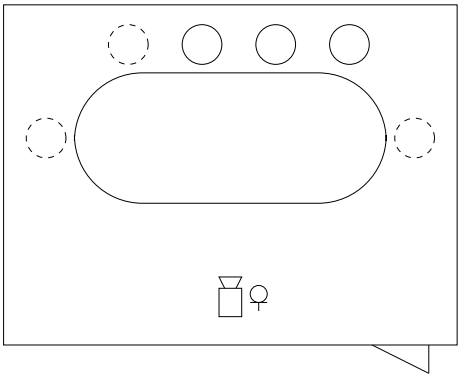

Fig. 1. Schematic visualization of the meeting's setting indicating the relative position of the participants as well as the camera and microphone. Dashed seats represent optional participants.

have so far focused on either the assessment or status/cohesion (rather than the performance) of a group or was concerned with particular group members such as the group leaders and their charisma or nonverbal communication signals [14, 26, 37, 40, 41]. However, the particular expressiveness or persuasiveness of individual group members is not of interest here. Rather, what is of interest here is whether the group as a whole creates an acoustic pattern that correlates with external group factors, which are, moreover, not related to group assessment or status, but to group performance in terms of creating valuable and feasible outcomes. The question we address here is: Is the group discussion *stimulating*? By stimulating, we mean that the communication partners and their interaction are propelled (by each other) such that the entire group is able to perform better.

Furthermore, with a focus on stimulation, we evaluate meeting performances beyond efficiency and its obvious physical counterpart duration (cf. e.g. [13]). However, looking at stimulation also includes the option to incorporate aspects like satisfaction, humor, synchrony, relationships, etc. (cf. e.g. [31, 34, 42, 44]), in this way complementing the time-centered notion of an efficient meeting by the concept of a "worthwhile" meeting. In other words, even longer-lasting meetings can be both efficient – in the sense of its outcomes – (cf. Section 4 and Fig. 2(a)) and worthwhile – in the sense of the communication "performance" of the meeting participants (in line with e.g. [43]). These aspects are investigated, analyzing the acoustic speech signals of meeting participants in the Parking Lot Corpus.

2 CORPUS DESCRIPTION AND PERFORMANCE MEASURES

This section introduces the corpus and utilized material, showing the details of the data collection. Further, the indicative measures, assessing either the overall experience or the group's performance, are presented. The meaning of each measure is discussed, allowing a better interpretation of our results.

2.1 The Parking Lot Corpus

The Parking Lot Corpus was set up based upon a meeting science experiment. In total, 245 undergraduate students from a public Midwestern United States university were recruited through the psychology department's online participant subject pool and were given class credits as compensation. Participants convened in 70 meeting groups ranging in size from three to six participants each. Given the current regulations, the corpus is (currently) not publicly distributed.

Each group met separately in a conference room (cf. Fig. 1). Each meeting occurred independently, and the different sized groups resulted from intentional non-specification of the number of participants per group. Participants were told that the purpose of the meeting, which was video and audio recorded using a (simple) video camera, was to generate campus parking and transportation recommendations that would eventually be shared with the appropriate university administrative office. Therefore, the experimenter provided a meeting agenda that included discussion questions and randomly selected a group leader by rolling a die. Further, the experimenter gave a brief overview of the tasks that group leaders typically perform (e.g., guiding the discussion, keeping attendees on track). Then meeting groups were given 20 minutes to discuss and record their recommendations as hardcopy. Upon completion of the meeting, leaders and meeting attendees completed different versions of a questionnaire that included the measures allowing an assessment of the meeting performance.

Given this data collection, we currently analyzed a subset of groups that provide – from a psychological perspective – a reasonable representative cross-section in both, the groups' constellations as well as the related performances. Therefore, a set of groups ($N = 14$) was processed and evaluated according to the performance measures presented in Section 2.2 and the speech parameters introduced in Section 3. However, the presented trends are first insights to the material.

2.2 Performance Measures

From the questionnaires and collected data, we derived eight performance measures that will highlight correlations of speech prosody analyses and group outcomes. Supporting the discussion of results (cf. Section 5), we ordered the measures into two categories:

Subjective performance indicators: measuring subjective ratings and subjective meeting results like efficiency or ideas.

- **ME_mean.** Four items measuring satisfaction with meeting processes were taken from Briggs et al. [10]. Participants rated phrases such as "satisfied with how the meeting was conducted" on a scale from 1 (strongly disagree) to 5 (strongly agree).
- **MSO_mean.** Four items measuring satisfaction with meeting outcomes were taken from Briggs et al. [10]. Participants rated phrases such as "satisfied with meeting results overall" on a scale from 1 (strongly disagree) to 5 (strongly agree).
- **MSP_mean.** Meeting satisfaction with process was assessed using a five item measure created by Briggs et al. [10]. Instructions read "Please indicate your agreement with the following statements" and a sample item is "I feel good about today's meeting process". Items were rated on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).
- **Mean_Q_Rec.** For quality, two independent raters had an agreement that was satisfactory (Cohen's $\kappa = 0.83$). The score was computed by summing the ratings for each recommendation and dividing by the total number of recommendations. Higher scores indicate greater quality of the ideas generated, on average.

Objective performance indicators: measuring objective and countable deliveries like number of ideas or feasible recommendations.

- **TS_Rec.** Total recommendations generated by the team were assessed by counting the written recommendations provided by the team at the conclusion of the meeting. Higher numbers indicate more ideas/recommendations were generated by the team.

Table 1. Overview on the utilized features.

pitch	minimum
	maximum
	mean
	final level
	range
	variability
	maximum velocity
duration	total duration
	IPU duration
	speaking rate
	silent pause duration
	location pitch maximum
IPU's mean intensity level	
total number of silent pauses	
total number of syllables	

- **Mean_F_Rec.** Two independent raters rated the recommendations for both, feasibility and quality, on a five point scale for each (extremely low to extremely high). For feasibility, the raters had an agreement that was satisfactory (Cohen's $\kappa = 0.86$). The score was computed by summing the ratings for each recommendation and dividing by the total number of recommendations. Higher scores indicate greater feasibility of the ideas generated, on average.
- **High_Rec.** As another indicator of team performance, we computed the total number of highly feasible and quality recommendations for each group. We took the sum of the recommendations that score either a 4 or 5 on either feasibility or quality ratings. Higher numbers indicate larger number of feasible and quality recommendations by a given team.

Note that preceding data analyses showed that the subjective and objective performance measures are correlated. However, correlations are weak (below $r = 0.3$) and indicate relationships, but no clear directional effects. Two graduate research assistants conducted the ratings of the group's recommendations in terms of quality and feasibility. The assistants were trained in coding the key behaviors of interest. However, they were blind to the research project hypotheses and did not code together, which makes them independent raters.

3 SPEECH PARAMETERS

The current section presents and introduces the investigated speech-based acoustic-prosodic parameters. This compiled set represents a selection of well-known parameters also be used in speech-based assessments of human affects (for an overview cf. e.g. [4, 5, 8, 36]).

The acoustic-prosodic features were measured automatically using adapted PRAAT scripts written by De Jong & Wempe [16], De Looze et al. [33], and Xu [51]. Measurements were conducted based on inter-pausal units (IPUs), meaning portions of speech that were separated by silent pauses longer than 300 ms. The recorded meetings included between 170 and 379 IPUs. Note that the IPU number also represents the number of measurements and the sample size per parameter.

A total of 15 acoustic-prosodic features (cf. Tab. 1) were analyzed for the present study. They cover four different phenomenological areas: *pitch, duration/timing, intensity/energy,* and *local events*. While the first three areas deal with holistic levels or patterns, the fourth area is about absolute frequencies, such as pauses and syllables. Measurements of the first three areas were taken based on an analysis window of 40 ms and a window shift of 10 ms (default settings in PRAAT). Pilot tests showed that these analysis settings were suitable for the analyzed speech material and resulted in a manageable number of measurement errors – which were manually corrected whenever detected and possible. If manual correction was not possible, then obviously implausible measurements were removed from the dataset.

As regards the features themselves, we measured for pitch (i.e. fundamental frequency, f0): the minimum, maximum, mean, and final pitch level inside each IPU (in Hz), the IPU's pitch range (in semitones), and pitch variability (in terms of the standard deviation, in Hz) as well the maximum velocity (st/s) of pitch change within these rising and falling pitch patterns. The latter measure could take both negative and positive values, indicating that the fastest movement was embedded in either a falling or a rising slope.

Duration and timing measurements included: the meetings' total duration (s), the IPU duration (s), the speaking rate (syllables/s, excluding silent pauses), silent pause duration (> 300 ms, in s), and the location (in s) of the pitch maximum inside the IPU.

Further measured parameters were: The IPU's mean intensity level (RMS, dB) and the total number of silent pauses (> 300 ms) and syllables.

Moreover, we additionally calculated the skewness and variance of measurements[1]. Skewness relates to how the measurements of a sample (here the meeting IPUs) are internally distributed. A skewness of 0 means that all values are symmetrically distributed around a central mean value. A skewness of < 0, which is a negative skewness, means that the distribution is biased to the right, for instance, in the sense that there are more smaller and larger values in the sample. The opposite is indicated by a skewness > 0, which is a positive skewness. Variance is an indicator of how widely the measured values are distributed around the central mean. A small variance thus means that all measured values were more or less similar, whereas a large variance means that the individual measurements differed considerably from each other and around the central mean.

A number of studies have shown in the past decades that acoustic-prosodic features like those measured here are consistently linked to affects or related concepts like passion and charismatic, stimulating speech (cf. e.g. [4, 5, 25, 28, 38, 41, 43]). In a nutshell, compared with factual, neutral reference conditions or conditions with less intense affects/expressions, an increase in arousal leads to an increase in the measured values of pitch, intensity, and speaking rate. This is especially true for affects with positive valence, in particular joy, surprise, happiness, excitement, and the like. It does not apply in the same way to intense negative affects such as anger, rage, or disgust. However, we can exclude this spectrum of affects for our meetings anyway and assume that the relevant spectrum in the analyzed meetings varied somewhere between neutral and positive - or at most reached into weak negative expressions such as boredom (which differs from positive affects in a similar way as a neutral, factual way of speaking).

Higher pitch levels, larger pitch ranges, faster pitch changes and, generally, higher levels of the variability and variance of pitch features as well as low or negative skewness levels are therefore positively correlated with the expression of joy or related positive affects in the meeting. The same applies to higher and more variable intensity and speaking rate measurements. Correlations in the opposite direction stand for meetings that were characterized by a more factual and neutral or possibly bored tone of voice. Alternatively and/or additionally, higher levels of variance

[1]Cf. for instance https://www.che.utah.edu/ tony/course/material/Statistics/12_descriptive.php (last accessed: June 01, 2021)

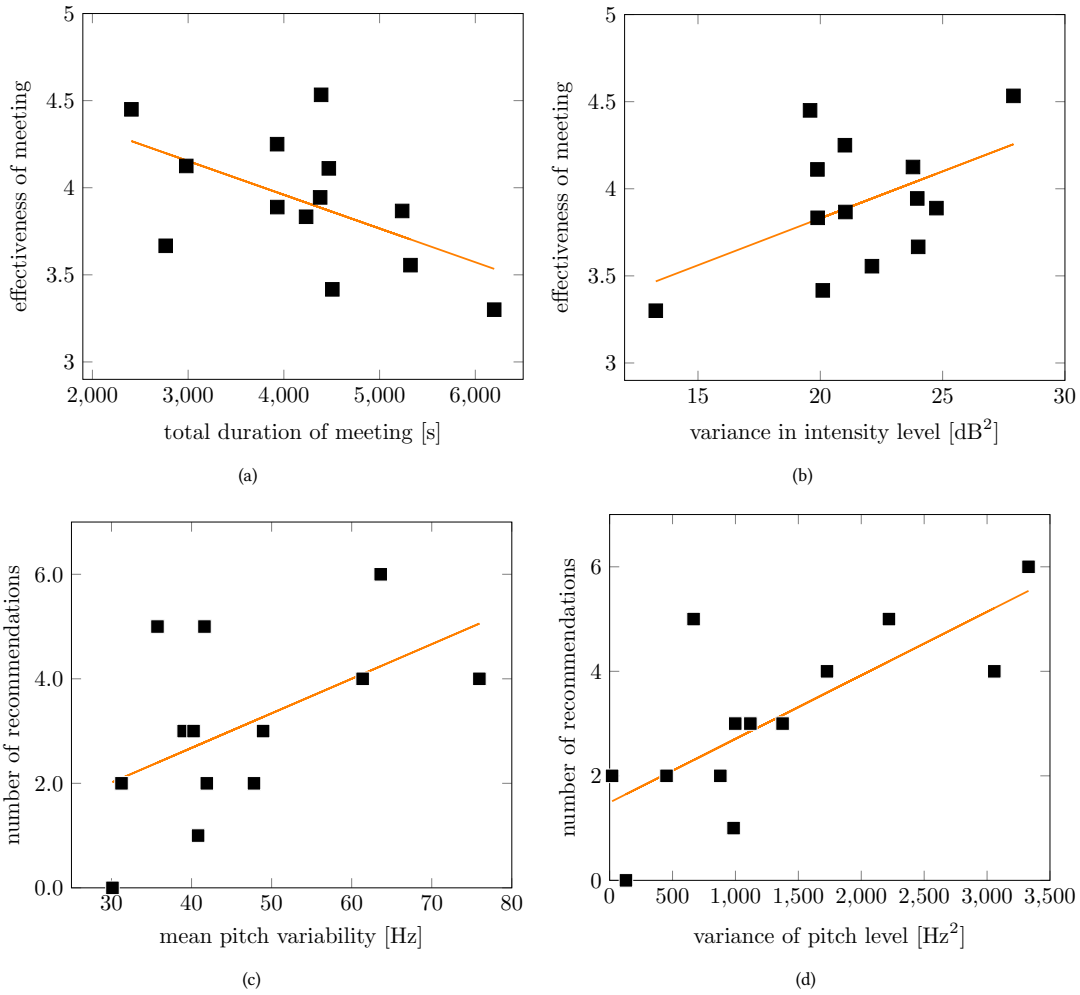

Fig. 2. Illustrations of selected significant correlations; ME_mean ratings with (a) total meeting duration and (b) variance in intensity level, as well as High_Rec with (c) mean pitch variability and (d) variance in pitch level.

and variability can also mean that more meeting participants with individually different prosodies participated actively or more often in the discussion, which, like with more intensive positive affects, can also be interpreted in the sense of a more lively and stimulating meeting.

4 PHONETIC ANALYSES AND RESULTS

The acoustic-prosodic features were correlated with the meeting ratings or scores using Pearson Product-Moment correlations (Pearson correlation). Note that, despite the relatively small number of analyzed dialogues, resulting in a sample size of $N = 14$, each dialogue included between 170-379 measured values per acoustic-prosodic feature. Given this solid empirical basis and considering our study's pilot nature, the below summary includes both significant

correlations and significant trends. For $N = 14$ (or df = 12), significant correlations (with $p \leq 0.05$) require a correlation coefficient $r \leq |0.49|$. Significant trends (with $p \leq 0.1$) require a correlation coefficient $r \leq |0.42|$. Significant correlations are marked with a double asterisk (**) and significant trends with a single asterisk (*).

Meeting effectiveness (ME_mean): With a total of 12 significant correlations, the retrospective ratings on meeting effectiveness are most closely linked to the acoustic-prosodic characteristics of the meeting. In specific, we found negative correlations between the mean rating of a meeting's effectiveness and the total duration of that meeting ($r = -0.56^{**}$, cf. Fig. 2(a)), the mean pitch range used by the participants in that meeting ($r = -0.56^{**}$), the maximum pitch velocity within that pitch range ($r = -0.54^{**}$), and the location of the highest pitch peak inside prosodic phrases ($r = -0.49^{**}$). The distribution measures were largely positively correlated with meeting effectiveness. For skewness, this concerns the duration of prosodic phrases ($r = 0.48^{*}$), the maximum velocity of pitch change ($r = 0.52^{**}$), and the location of the highest pitch peak within the participants' prosodic phrases ($r = 0.55^{**}$). For variance, positive correlations emerged for pitch range ($r = 0.50^{**}$), mean intensity ($r = 0.50^{**}$, cf. Fig. 2(b)), and maximum pitch velocity ($r = 0.44^{*}$). In addition, there were two negative correlations that concerned the duration of the meeting's in prosodic phrases ($r = -0.50^{**}$) as well as the location of the highest pitch peak within these phrases ($r = -0.50^{**}$).

Total score on recommendations (TS_Rec): This output indicator was only linked to two acoustic-prosodic features. Both are related to the changeability (i.e. vividness) of the participants' speech melody. A higher degree of pitch variability was positively correlated with the total score on recommendations ($r = 0.51^{**}$). Similarly, a larger variance of the mean pitch levels between prosodic phrases was positively correlated with the total score on recommendations ($r = 0.62^{**}$)

Feasibility of recommendations (Mean_F_Rec): In terms of how feasible participants rated their recommendations, we found negative correlations with the mean speaking rate in the meetings' prosodic phrases ($r = -0.55^{**}$) and the number of silent pauses in between these phrases ($r = -0.51^{**}$). Furthermore, in terms of distribution measures, we found that feasibility ratings were higher when the skewness of the phrases' mean pitch levels ($r = -0.57^{**}$) and final pitch levels ($r = -0.47^{*}$) was lower. For variance, there were two negative correlations, one with the mean pitch level ($r = -0.43^{*}$) and one with the maximum pitch velocity ($r = -0.43^{*}$). In other words, the less melodic variability a meeting showed the higher were the feasibility ratings for the recommendations made in that meeting.

Quality of recommendations (Mean_Q_Rec): Not a single acoustic-prosodic features correlated significantly with this performance indicator. The quality of a recommendation is usually based on or related to the particular content of the recommendation. Therefore, such an assessment needs to be driven by contextual and semantical investigations which are out of the paper's scope. In future analyses, we link our work with investigations of [9] or [19] modeling contextual information in interactions.

Number of highly feasible/HQ recommendations (High_Rec): In terms of the actual goal of having a meeting, this can be considered a key performance indicator. It was correlated primarily with different aspects of changes in speech melody. For example, we found positive correlations with mean pitch range ($r = 0.53^{**}$) and mean pitch variability ($r = 0.53^{**}$, cf. Fig. 2(c)) as well as with the variance in pitch maximum level ($r = 0.61^{**}$), pitch minimum level ($r = 0.59^{**}$), and mean pitch level ($r = 0.74^{**}$, cf. Fig. 2(d)).

Meeting satisfaction with processes (MSP_mean): This performance indicator was higher, when the mean duration of the prosodic phrases of that meeting and the speaking rate inside these phrases were lower ($r = -0.45^{*}$; $r = -0.57^{**}$).

Moreover, meeting satisfaction increased when the skewness of the pitch-range distribution ($r = 0.44^{*}$) and the variance in pitch level were higher ($r = 0.50^{**}$).

Meeting satisfaction with outcomes (MSO_mean): Compared to the process-satisfaction ratings (e.g. MSP_mean), the output-satisfaction ratings were - with six instead of only four correlations - more tightly linked to the meetings' prosodic characteristics. We found correlations of output satisfaction ratings with mean pitch level ($r = 0.49^{*}$), mean intensity level ($r = -0.44^{*}$), and the number and/or duration of silent pauses ($r = -0.48^{*}$) as well as with the speaking rate ($r = -0.48^{*}$). Moreover, with respect to distribution measures, there were positive correlations with the skewness of the maximum velocity of pitch change ($r = 0.53^{**}$) and with the variance of the pitch level between the prosodic phrases of a meeting ($r = 0.47^{*}$).

5 DISCUSSION

The aim of this study was to use the Parking Lot Corpus (cf. Sec. 2.1) to find out whether there are correlations between the objective and subjective performance indicators of meetings on the one hand and their acoustic-prosodic characteristics on the other. That is, the question was whether or not "the sound of a meeting" already provides information about how productive it was and how satisfied the meeting participants were, especially with regard to output, process, and effectiveness.

The level of our acoustic analysis was fairly basic from a socio-phonetic point of view. For example, neither did we analyze the prosody of the meeting leader separately, nor did we break up the meeting into the individual participants' turns and examine their duration, number, and interaction/interplay separately. These limitations will be considered in our extended future investigations that will also take into account the full number of groups of the speech corpus ($N = 70$). However, this basic approach was pursued for the reason that the basic bottom-up method we have currently chosen, is appealing from a pragmatic, application-oriented point of view. If, for example, a first rough automatic assessment of the performance of a meeting were possible without running any software for automatic speech recognition and automatic speaker recognition in the background, then the technical hurdles of such a system (and, thus, also its price) would be relatively low and simultaneously, due to the system's simplicity, its robustness (error resistance) would be relatively high.

Against this background, we were indeed able to find several significant correlations (at $p<0.1$ or $p<0.05$) of the acoustic-prosodic features with the performance indicators that were obtained from participants after their meetings (cf. Section 2.2). In other words, our study provides initial empirical evidence that objective and subjective success criteria can be derived from "the sound of a meeting" to a certain extent. A total of 35 correlations were found between these success criteria and the meetings' sound characteristics. Despite the relatively large amount of analyzed prosodic features and performance indicators, this is a considerable number. Moreover, the fact that some specific performance indicators yielded no correlations at all (quality of recommendations, cf. explanation in Section 4), while others were associated with almost a third of all correlations (effectiveness rating), additionally speaks for the non-random nature and validity of the findings.

The following overarching results patterns have emerged.

First, the feature set of intonation played the biggest role in the correlations with the performance indicators. Almost 71% of all correlations (25 out of 35) concerned mean values or distribution measures (skewness or variance) of pitch features. In comparison to this, the dimensions of the acoustic energy, which means, the RMS/intensity level and its distribution parameters as well as the frequency and extension of silent pauses, played only a marginal role. Only about

Fig. 3. Word cloud illustrating the frequency of occurrence of the acoustic-prosodic labels and parameters in the significant correlations with meeting performance scores.

8% of all correlations (3 out of 35) resulted from this feature set. Note that this means a further advantage from an application-oriented point of view, because in practice the dimensions of the acoustic energy would strongly depend on the room acoustics and the speakers' mouth-to-microphone distances and could, therefore, hardly be measured reliably. With 21% of all correlations (7 out of 35) duration and tempo measures took a middle position in the relevance ranking of the three prosodic feature sets.

The dominance of pitch for estimating both subjective and, particularly, objective meeting success becomes visually tangible in the word cloud. The font size reflects the absolute frequency of occurrence of the acoustic-prosodic label elements that yielded significant correlations with meeting performance scores (cf. Fig. 3). The analyzed acoustic parameter and features labels of all significant correlations (at $p<0.05$ and $p<0.1$) served as input for the word cloud. Accordingly, we furthermore see in Fig. 3 that pitch level and maximum characteristics played a bigger role than pitch minimum and final pitch, and that mean values played a bigger role than values of skewness and variance.

Second, a remarkable difference emerged between the objective and subjective performance indicators. Both objective indicators, which means, the total score on recommendations and the number of highly feasible or high-quality recommendations, correlated exclusively with pitch features. Moreover, an increase in the variation or variability of these pitch features, both within and between prosodic phrases, was linked to an increase in objective performance. That is, the livelier and more changeable the intonational sound of a meeting was, the more ideas and, in particular, the more feasible and valuable ideas did the meeting generate. Defining the nature of this intonational sound more specifically must be postponed to future studies, though, as our basic and holistic measurements leave too much room for interpretation. For example, the fact that a higher variance in the mean pitch levels of prosodic phrases is positively correlated with the total score on recommendations allows at least two interpretations: either affectively more varied debates characterized those meetings that ended with a higher total score on recommendations; or the total speaking time of participants with different voice-pitch levels was more balanced in such meetings. Similarly, that the number of highly feasible or high-quality recommendations was positively correlated with the mean location of the highest pitch peak inside in the participants' prosodic phrases can also mean two things: 1) either the utterances with a later peak-maximum location were more often comments on given than introductions of new information (in terms of a pitch-accent marking of the theme-rheme structure, cf. e.g. [12]); 2) or there were more questions in these meetings that ended in high rising intonations, expressing activation and listener orientation. More in-depth phonological and turn-based data analyzes are required to resolve such different interpretations.

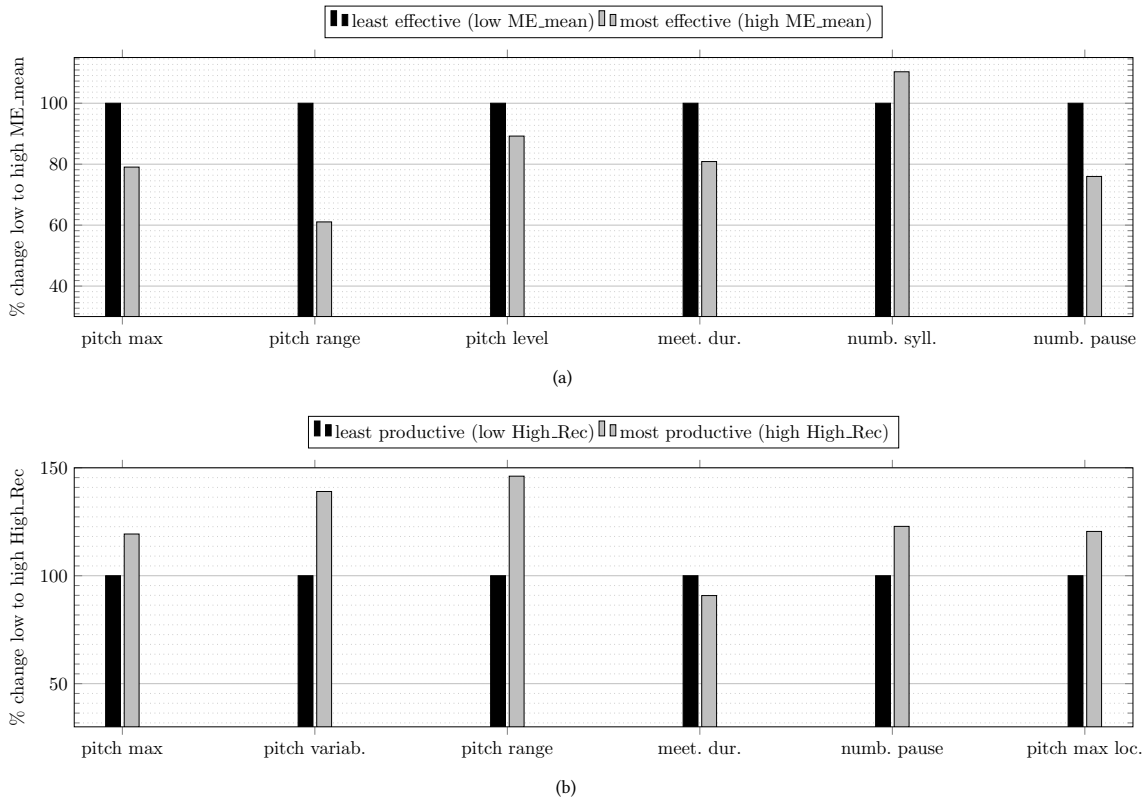

Fig. 4. Prosodic profiles (parameter mean values) of meetings that (a) scored the lowest and the highest in terms of subjective effectiveness (ME_mean) and that (b) yielded the lowest and highest output in terms of the number of feasible, high-quality ideas (High_REC).

However, one crucial point is clear already: Meetings that were a success in subjective terms sounded differently than those that were a success in objective (output) terms. Fig. 4(a)-(b) illustrate these conclusions with references to selected mean values. Fig. 4(a) shows the percentage changes that distinguish the four meetings with the subjectively lowest effectiveness rating from the four meetings with the subjectively highest effectiveness rating in terms of ME_mean. We see that the sound of subjectively more effective meetings is characterized by an approximately 40% narrower pitch range and 20% lower pitch maxima and a 10% lower pitch level. Furthermore, being subjectively more effective means decreases of about 20% in meeting duration and in the absolute frequency of pauses, whereas the number of syllables produced in a meeting increases by about 10% on average.

In contrast, when it comes to the actual productive output of a meeting in terms of High_Rec, which is the number of highly feasible or high-quality recommendations, then virtually inverse sound characteristics of meetings emerge (cf. Fig. 4(b)). The most productive meetings were those whose pitch variability and pitch range are 40-45% higher and whose pitch maxima reach about 20% higher and occur about 20% later within the IPUs. The absolute frequency of pauses increases rather than decreases by about 20%, whereas the meeting duration remains more or less constant.

Overall, the results indicate that meetings that were affectively calmer and had a simpler and shorter prosodic structure were perceived to be better (more effective). However, the better meetings in objective terms, which means, those that generated many good ideas, actually required the exact opposite: a lively, interactive, stimulating prosody. This conclusion is also in line with recent findings on the role of expressive, prosodic-based stimulating characteristics in creativity challenges (cf. e.g. [40, 41]).

The two subjective performance indicators, perceived effectiveness and satisfaction with output, differ similarly but not quite as strongly from the objective indicators. The other subjective indicators are feasibility of recommendations and satisfaction with the meeting process. Unlike the latter two, perceived effectiveness and satisfaction with output included duration and intensity as further relevant features. For example, meetings with a higher subjective output satisfaction were quieter; and subjectively more effective meetings were one thing above all else: shorter. This applies to both the overall duration of the meeting and the duration of the individual prosodic phrases inside the meeting (compare the correlation of effectiveness ratings with a positively skewed phrase duration distribution). In contrast, the ratings on perceived effectiveness and satisfaction with output also correlated positively, at least in parts, with higher intonational variability. For example, ratings on average increased for greater variances in pitch range. Combined with the positive correlation found for the mean pitch level, this suggests that the subjective satisfaction with meeting outputs in particular also depends, at least to some extent, on a variable, more expressive/affective prosody (cf. also e.g. [34, 40]). In this respect these ratings were more similar than all others to the objective performance indicators.

6 CONCLUSION

We investigated the option to assess performance indicators, evaluating the overall outcomes of and satisfaction with meetings. For this, we analyzed a set of 35 prosodic-acoustic features extracted from the acoustic material of group interactions, in particular, a subset of 14 groups of the Parking Lot Corpus (cf. Section 2). The features – generated using PRAAT scripts (cf. e.g. [16]) – were correlated with subjective and objective meeting scores by applying Pearsons Product-Moment correlations.

Regarding the word cloud in Fig. 3, the most influential parameters are mean pitch level and variance as well as their maximum and skewness. But, we also found a significant influence of the intensity velocity or silent pauses on the group's performance. Interestingly, there are correlations of ratings with the meeting's duration. However, the strength of this effect is not as prominent as it could be expected (cf. Fig. 3 and analytical results in Section 4). Therefore, from an acoustic perspective, we conclude that lively and variable voices contribute a lot to meeting satisfaction and outcome for performing groups. This is also reflected and visualized in Fig. 4, which compares the least and most effective meetings' prosodies. These findings also suggest what a stimulating meeting or interaction actually is – something that is not based on (total) duration.

From Fig. 4, we also see that the overall (objective) output of the meetings is influenced by acoustic-prosodic indicators and the subjective assessment of the meetings' effectiveness. In combination with Fig. 2(a), also the expected effect of meeting duration can be seen, although its influence on ratings is relatively small.

In general, we conclude that multiple acoustic-prosodic features or indicators are significantly relevant for the assessment of successful meetings and group performance. From an application-based perspective, we further showed that 1) the number of those indicators is rather small (compared to common feature sets in, for instance, affect recognition from speech) and that 2) these indicators (already) work at the holistic level of the entire group. Especially the latter allows an technologically easily implementable analysis or prediction of group performance without detailed examinations of particular group members. Nevertheless, we hypothesize that fine-graded analyses will further improve

the prediction of meeting performance. Additionally, such more fine-grained analyses could also allow for automatic assessments of group constellations and inner-group performance, thus helping anticipate and treat group "conflicts" or "mismatches" objectively and effectively. In this sense, we recommend both more fine-graded analyses of group members as well as automatic, neural-based holistic prediction methods to assess meeting performance.

ACKNOWLEDGEMENTS

We acknowledge support by the project “Adaptive Strategies in Assistive Technologies in Multi-Person Interaction” (ASAMI) funded by the Federal State of Sachsen-Anhalt, Germany, under the grant number: I 138. Note that the first author, ON, is co-founder of the speech-technology company AllGoodSpeakers ApS. For a corresponding note on conflict of interest, please see https://www.allgoodspeakers.com/coi

REFERENCES

[1] Oleg Akhtiamov, Dmitrii Ubskii, Evgeniia Feldina, Aleksei Pugachev, Alexey Karpov, and Wolfgang Minker. 2017. Are You Addressing Me? Multimodal Addressee Detection in Human-Human-Computer Conversations. In *Speech and Computer*, Alexey Karpov, Rodmonga Potapova, and Iosif Mporas (Eds.). Springer, Cham, 152–161.

[2] J.A. Allen, N. Lehmann-Willenbrock, and S. Rogelberg. 2015. *The Cambridge Handbook of Meeting Science.* Cambridge University Press.

[3] J.A. Allen and S.G. Rogelberg. 2013. Manager-Led Group Meetings: A Context for Promoting Employee Engagement. *Group & Organization Management* 38, 5 (2013), 543–569.

[4] J. A. Bachorowski and M. J. Owren. 2008. Vocal expressions of emotion. *Handbook of emotions* 3 (2008), 196–210.

[5] R. Banse and K. R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology* 70, 3 (1996), 614–636.

[6] S. G. Barsade and D. E. Gibson. 2012. Group Affect: Its Influence on Individual and Group Outcomes. *Current Directions in Psychological Science* 21, 2 (2012), 119–123.

[7] Kylie Bartolo and Brett Furlonger. 2000. Leadership and job satisfaction among aviation fire fighters in Australia. *Journal of Managerial Psychology* 15, 1 (2000), 87–93.

[8] Ronald Böck. 2020. *Anticipate the User: Multimodal Analyses in Human-Machine Interaction towards Group Interactions.* TUDpress, Dresden, Germany.

[9] Ronald Böck and Britta Wrede. 2019. Modelling Contexts for Interactions in Dynamic Open-World Scenarios. In *Proceedings of the 2019 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1475–1480.

[10] R.O. Briggs, G.-J. de Vreede, and B.A. Reinig. 2003. A theory and measurement of meeting satisfaction. In *Proc. of the 36th Annual Hawaii International Conference on System Sciences*. 8 pp.

[11] R. Brown. 2000. *Group Processes – Dynamics within and between Groups.* Blackwell Publishers, Oxford, UK.

[12] Sasha Calhoun. 2012. The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics* 40, 2 (2012), 329–349.

[13] Z. Chen, J. Zhu, and M. Zhou. 2015. 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. *Journal of Applied Psychology* 100 (2015), 511–521. Issue 2.

[14] H. W. Chou, Y. H. Lin, H. H. Chang, and W. W. Chuang. 2013. Transformational Leadership and Team Performance: The Mediating Roles of Cognitive Trust and Collective Efficacy. *SAGE Open* 3, 3 (2013), s.p. https://doi.org/10.1177/2158244013497027

[15] S. A. Chowdhury, M. Danieli, and G. Riccardi. 2015. Annotating and categorizing competition in overlap speech. In *Proc. of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5316–5320.

[16] N.H. de Jong and T. Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behav Res Methods* 41, 2 (2009), 385–390.

[17] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal Synchrony : A Survey Of Evaluation Methods Across Disciplines. *IEEE Transactions on Affective Computing* 3, 3 (2012), 349–365.

[18] Joseph Dippong. 2020. Status and Vocal Accommodation in Small Groups. *Sociological Science* 7, 12 (2020), 291–313.

[19] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk K.J. Heylen, Hayley Hung, Mark Neerincx, and Khiet Phuong Truong. 2019. Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction*. IEEE, United States, 206–212.

[20] Olga Egorow and Andreas Wendemuth. 2017. Emotional features for speech overlaps classification. In *INTERSPEECH-2017*. ISCA, 2356–2360.

[21] S.D. Farley. 2014. Nonverbal reactions to an attractive stranger: The role of mimicry in communicating preferred social distance. *Journal of Nonverbal Behavior* 38 (2014), 195–208. Issue 2.

[22] Cindy Gallois and Howard Giles. 2015. *Communication Accommodation Theory.* American Cancer Society, 1–18.

[23] Daniel Gatica-Perez. 2009. Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review. *Image Vision Computation* 27, 12 (2009), 1775–1787.

[24] A.L. Gonzales, J.T. Hancock, and J.W. Pennebaker. 2011. Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research* 37, 1 (2011), 3–19.

[25] K. Hammerschmidt and U. Jürgens. 2007. Acoustical correlates of affective prosody. *Journal of voice* 21, 5 (2007), 531–540.

[26] Hayley Hung, Dinesh Jayagopi, Chuohao Yeo, Gerald Friedland, Sileye Ba, Jean-Marc Odobez, Kannan Ramchandran, Nikki Mirghafori, and Daniel Gatica-Perez. 2007. Using Audio and Video Features to Classify the Most Dominant Person in a Group Meeting. In *Proc. of the 15th ACM International Conference on Multimedia*. ACM, New York, NY, USA, 835—-838.

[27] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J.W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science* 22, 1 (2011), 39–44.

[28] T. Johnstone. 2017. *The effect of emotion on voice production and speech acoustics*. Ph.D. Dissertation. The University of Western Australia.

[29] Elise Keith. 2015. 55 Million: A Fresh Look at the Number, Effectiveness, and Cost of Meetings in the U.S. (Dec 2015). https://blog.lucidmeetings.com/blog/fresh-look-number-effectiveness-cost-meetings-in-us

[30] Steve W.J. Kozlowski, Stanley M. Gully, Earl R. Nason, and Eleanor M. Smith. 1999. *Developing adaptive teams: A theory of compilation and performance across levels and time*. Wiley, Hoboken, USA, 240–292.

[31] Nale Lehmann-Willenbrock and Joseph A. Allen. 2014. How fun are your meetings? Investigating the relationship between humor patterns in team interactions and team performance. *Journal of Applied Psychology* 99, 6 (2014), 1278–1287.

[32] Daniel Levi. 2015. *Group dynamics for teams*. SAGE, Los Angeles, USA.

[33] Céline De Looze and Daniel Hirst. 2008. Detecting changes in key and range for the automatic modelling and coding of intonation. In *Proc. of the Speech Prosody 2008*. 135–138.

[34] Paulo N. Lopes, Peter Salovey, and Rebecca Straus. 2003. Emotional intelligence, personality, and the perceived quality of social relationships. *Personality and Individual Differences* 35, 3 (2003), 641–658.

[35] Usman Malik, Mukesh Barange, Julien Saunier, and Alexandre Pauchet. 2019. Using Multimodal Information to Enhance Addressee Detection in Multiparty Interaction. In *Proc. of the International Conference on Agents and Artificial Intelligence*. scitepress, s.p.

[36] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22–35.

[37] S. J. Miles and G. Mangold. 2002. The impact of team leader performance on team member satisfaction: the subordinate's perspective. *Team Performance Management: An International Journal* 8, 5/6 (2002), 113–121.

[38] S. J. L. Mozziconacci. 1998. *Speech variability and emotion: Production and perception*. Ph.D. Dissertation. Technical University Eindhoven.

[39] Joseph E. Mroz, Joseph A. Allen, Dana C. Verhoeven, and Marissa L. Shuffler. 2018. Do We Really Need Another Meeting? The Science of Workplace Meetings. *Current Directions in Psychological Science* 27, 6 (2018), 484–491.

[40] Oliver Niebuhr. 2021. Advancing higher-education practice by analyzing and training students' vocal charisma: Evidence from a Danish field study. In *Proc. 7th International Conference on Higher Education Advances*. Editorial Universitat Politecnica de Valencia, 743–751.

[41] Oliver Niebuhr and Radek Skarnitzl. 2019. Measuring a speaker's acoustic correlates of pitch - but which? A contrastive analysis for perceived speaker charisma. In *Proc. of the 19th Int. Congress of Phonetic Sciences (Proc. of the International Congress of Phonetic Sciences)*, Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren (Eds.). Australasian Speech Science and Technology Association Inc, 1774–1778.

[42] Mark D. Peterson, Daniel J. Dodd, Brent A. Alvar, Matthew R. Rhea, and Mike Favre. 2008. Undulation Training for Development of Hierarchical Fitness and Improved Firefighter Job Performance. *The Journal of Strength & Conditioning Research* 22 (2008), 1683—-1695. Issue 5.

[43] Amalia Petrovici and Tatiana Dobrescu. 2014. The Role of Emotional Intelligence in Building Interpersonal Communication Skills. *Procedia – Social and Behavioral Sciences* 116 (2014), 1405–1410.

[44] Steven G. Rogelberg, Joseph A. Allen, Linda Shanock, Cliff Scott, and Marissa Shuffler. 2010. Employee satisfaction with meetings: A contemporary facet of job satisfaction. *Human Resource Management* 49, 2 (2010), 149–172.

[45] Eduardo Salas, Scott I. Tannenbaum, Steve W. J. Kozlowski, Christopher A. Miller, John E. Mathieu, and William B. Vessey. 2015. Teams in Space Exploration: A New Frontier for the Science of Team Effectiveness. *Current Directions in Psychological Science* 24, 3 (2015), 200–207.

[46] Ingo Siegert, Shuran Tang, and Alicia Flores Lotz. 2018. Acoustic addressee-detection - analysing the impact of age, gender and technical knowledge. In *Proc. of the 29. Konferenz Elektronische Sprachsignalverarbeitung*. TUDpress, 113–120.

[47] T. Treadwell, N. Lavertue, V. K. Kumar, and V. Veeraraghavan. 2001. The Group Cohesion Scale-Revised: Reliability and validity. *International Journal of Action Methods: Psychodrama, Skill Training, and Role Playing* 54, 1 (2001), 3–12.

[48] T. J. Tsai, A. Stolcke, and M. Slaney. 2015. A Study of Multimodal Addressee Detection in Human-Human-Computer Interaction. *IEEE Transactions on Multimedia* 17, 9 (2015), 1550–1561.

[49] Konstantin O. Tskhay and Nicholas O. Rule. 2013. Accuracy in Categorizing Perceptually Ambiguous Groups: A Review and Meta-Analysis. *Personality and Social Psychology Review* 17, 1 (2013), 72–86.

[50] Alessandro Vinciarelli, Maja Pantic, and Herve Boulard. 2009. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing* 12, 27 (2009), 1743–1759.

[51] Y. Xu. 2023. ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis.. In *Proc. of Tools and Resources for the Analysis of Speech Prosody*. 7–10.

[52] Michael Yoerger, Joseph A. Allen, and John Crowe. 2018. The Impact of Premeeting Talk on Group Performance. *Small Group Research* 49, 2 (2018), 226–258.