View metadata, citation and similar papers at core.ac.uk brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Signal Processing in the Workplace

Daniel Gatica-Perez

According to the U.S. Bureau of Labor Statistics, during 2013 employed Americans "worked an average of 7.6 hours on the days they worked", and "83 percent did some or all of their work at their workplace" [1]. Understanding processes in the workplace has been the subject of disciplines like organizational psychology and management for decades. In particular, the study of nonverbal communication at work is fundamental as "face-to-face interaction with superiors, subordinates, and peers consume much of our time and energy" [2] and a variety of phenomena including job stress, rapport, and leadership can be revealed by and perceived from the tone of voice, gaze, facial expressions, and body cues of co-workers and managers [2].

In parallel to these developments, progress in audio-visual sensing and machine perception is making the extraction of several of these nonverbal cues feasible and scalable. This trend creates opportunities towards improving the scientific understanding of phenomena in organizations and to develop technology that supports individuals and groups at work. Furthermore, it defines a domain where signal processing researchers can find new problems while working with social scientists.

In this column, we describe a framework developed with collaborators in organizational psychology, aimed at inferring high-level constructs of interest in the workplace from nonverbal behavior. We summarize our experience tackling two tasks: identifying emergent leaders in small groups, and assessing hirability of candidates in employment interviews. The examples discussed in this column have been recorded in a standard lab setting [9], in which sensors are fixed in a specific environment that volunteer participants have to visit, but also in moderately in-the-wild settings, where a portable sensing solution has been used to bring participants to quiet indoor environments for recordings [5], which gives flexibility for recruitment of volunteers. Sensors have included webcams and commercial microphone arrays for the portable case, and high-resolution cameras and Microsoft Kinect for the lab case. As the interactions take place around a table in real workplaces, we have exploited this setting for sensor placement. One specific goal of our work with psychologists has been the deployment of the sensing lab in their institution, with the goal of promoting a wider and more frequent use of the technology in their discipline.

The material of this column is adapted from [5, 9] and the reader is referred to the original papers for details. We close with a few thoughts on what the future could bring in this domain.

A framework for social inference from nonverbal behavior

The computational framework we have developed is shown as a diagram in Figure 1 [5, 9]. It follows a supervised machine learning approach, where training and test phases are defined to automatically infer variables of interest (hirability in job interviews or emergent leadership in small groups) from dyadic or group interactions. At the onset, experiments are designed jointly by psychologists and engineers, and involve the selection and deployment of sensing technology, the design of the specific interaction to be recorded, a battery of questionnaires to be completed by study participants, and human coding tasks to be completed by external observers.

Questionnaire data completed by participants and additional coding data provided by external observers are used both for psychology research and as ground-truth data for computational analysis. Questionnaires, designed and validated by psychologists, are often adapted from previous literature and administered to participants in the experiments. Additional coding data can be produced by trained psychology students or experts. The manual annotation process in Fig. 1 involves the post-processing of the above data to define ground-truth in amenable form for machine learning tasks. Concretely, hirability scores in job interviews provided by trained coders can be used to define a regression task (e.g. estimate the actual score) or a classification task (e.g. high vs. low score levels); furthermore, questionnaire data provided by the participants in a group discussion about the perceived leadership of each team member can be aggregated to define the ground-truth in a task whose goal is to identify one person in each group.

The nonverbal feature extraction process has involved both the development of new techniques to extract cues from audio and video and the use of existing modules. Cues related to speaking activity, prosody, body and head activity, and gaze have been used in the work described here (facial expressions have been used in other instances of our work.) The bi-disciplinary approach has influenced our choices regarding the extraction of behavioral cues previously documented in psychology research with respect to their predictive value for the variables of interest (hirability or emergent leadership.) This has facilitated placing the results of our studies in the context of previous literature. At the same time, machine learning gives the possibility to extract new features, some of which might not be readily interpretable but effective for automatic inference. Moreover, the use of machine learning methods (Support Vector Machines as an example) can spur constructive dialog with psychologists, who are less familiar with these methods and in contrast are more acquainted with classical statistical methods and especially interested in interpretable approaches.

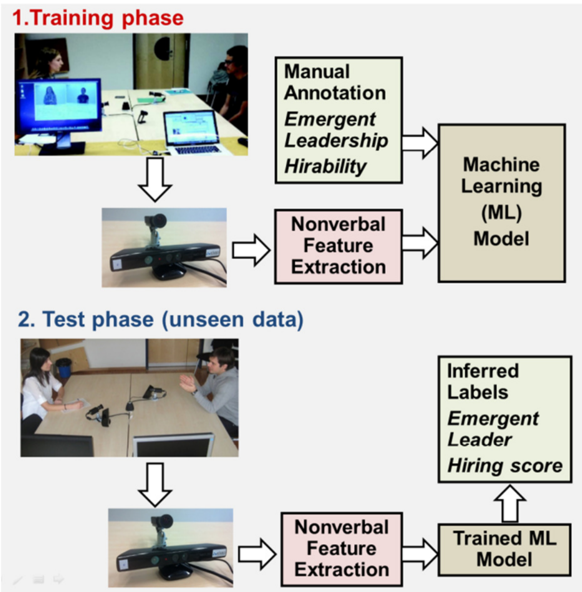

Figure 1. Computational framework to study work-related tasks

Emergent Leadership in Small Groups

In the context of groups, the so-called vertical dimension of social relations includes constructs like dominance, status, and leadership, all referring to the position that members occupy in a group [3]. In particular, research on leadership in organizational psychology and management has characterized leadership styles used to direct groups as well as emerging phenomena. Emergent leaders are individuals who rise among the members of a group and gain power from the group members themselves, instead of doing so from external entities (e.g. upper management) [4]. As much work nowadays gets done in groups, identifying emergent leaders is relevant in practice for recruitment, training, and development in organizations.

Connections between nonverbal behavior and emergent leadership have been studied for several decades [4]. While an extensive discussion cannot be provided here, different studies have found connections between ratings of perceived emergent leadership and manually coded cues like speaking time, arm movements and gaze (including given and received gaze and joint patterns of looking/speaking.) Some of these cues have also been linked to dominance, a related but not identical concept related to a tendency to control others via observable acts [3].

In [5], we followed the approach described in Figure 1 to identify the emergent leader in three- to four-person groups. We used two webcams and a Dev-Audio Microcone microphone array as sensors. Each camera covers two people, and the Microcone provides audio for prosody feature extraction while generating a segmentation of the speech of each person (Figure 2.) Groups of unacquainted people were asked to play the Winter Survival Task, a commonly used exercise to study group decision making and performance. In the task, participants need to rank a list of items according to their relevance for survival in a hypothetical plane crash in winter. Individuals first generate their own rankings, and then discuss and collectively agree on a final list, the interaction eliciting the possible emergence of a leader. After concluding the list, participants were asked to fill out questionnaires to characterize the other group members, including variables like perceived leadership, perceived dominance, and perceived competence. The resulting Emergent LEAdership (ELEA) corpus includes audio, video, and questionnaire data for 40 groups (148 individuals), and is publicly available for academic research.

Standard speech processing and computer vision methods were used to extract a variety of nonverbal cues. From the audio track for each participant, this included the amount of speaking time, number and average length of speaking turns, number of interruptions, speech spectral flatness, energy variation, and pitch variation. From video, features included a head activity measure obtained from a head tracker and optical flow estimates, and a body activity measure based on an improvement of classic motion templates (Motion Energy Images). Details can be found in [5]. In subsequent work [6], head pose (as a proxy for gaze) and joint looking/speaking patterns were also extracted using visual trackers based on particle filtering.

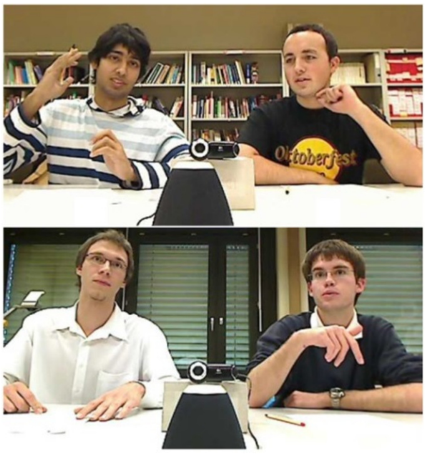

Figure 2. A snapshot from the Emergent LEAdership (ELEA) corpus (taken from [5]).

A correlation analysis of the perceived variables from the questionnaires first showed that the emergent leader was significantly perceived as a dominant person, with a second, less strong correlation effect between perceived leadership and competence. This is an interesting finding that relates different organizational constructs with one another. Furthermore, a correlation analysis between the perceived questionnaire variables and the nonverbal features showed that emergent leadership is linked to participants who talk more, take more turns, interrupt more, and move their body more. This motivated the automatic recognition approach from these cues. Using standard classification techniques (SVMs or ranked feature fusion), the method identified the emergent leader in a group with accuracy between 70-85% depending on the modalities and classifiers used. Two results relevant for signal processing are that the cues derived from the audio track were more discriminant than the visual cues, and when combined, visual cues can bring a slight performance improvement.

Hirability in Job Interviews

Interviews are an integral part of the recruitment process, and as such they have been extensively studied in organizational psychology and management [7, 8]. From the social computing perspective, employment interviews are an important subject of study because of their impact in people's life, their expressiveness, and the volume with which they are generated. Automatic analysis could be used to provide feedback to candidates, to support training programs, or to summarize large volumes of data in big organizations.

Previous literature on nonverbal communication has studied links between a number of features and job interview perceptions and outcomes. Interviewers most often do not meet the applicants in person before the interview; they interact on the basis of previous information provided by the applicant (CV, reference letters, LinkedIn profiles) and the behavior during the interview itself. Interviewers make impressions about a number of attributes of the candidate, hirability being one of them, and use these impressions and other available information to make decisions. Studies based on manually coded cues have found that candidates who are perceived as more hirable and competent (or who are actually hired) display an array of cues including smiling, eye contact, nodding, reduced interpersonal distance, body posture (oriented towards the interviewer), and specific speaking patterns [7,8]. Taken together, this so-called immediacy behavior might convey a sense of larger availability or closeness, which as some literature suggests can lead to positive impressions on interviewers and, as a consequence, more positive assessments of candidates.

In [9], we analyzed job interviews following the approach in Figure 1. We first collected a corpus of 62 interviews where candidates applied for a real (albeit short) paid job, related to recruiting volunteers on the street for future psychology experiments. The job itself had therefore connections to a sales position. We used Microsoft Kinect and high-resolution cameras to collect video, and the Microcone to collect audio (Figure 3). The interviews were structured (i.e., they consisted of a fixed number of questions, asked in the same order to each candidate) and behavioral (i.e., the questions were designed to elicit behavioral responses from the candidates). Interviews lasted 11 minutes on average. Among a variety of questionnaire and manual coding data that were collected, a hirability measure was provided by a psychology student who watched the interview using audio and video from both the candidate and the interviewer and who was trained at the task.

Figure 3. Interviewer and candidate in a job interview (taken from [9]).

Regarding nonverbal cues, in addition to audio features similar to the ones described in the previous section, we developed methods to extract head nods [10] and body cues [11]. In [10], we demonstrated the advantages in terms of performance of a multimodal approach for nodding recognition, in which the observation of the (self) speaking state of a person (speaking or silent) is used to learn two separate nodding/non-nodding classifiers, one for each speaking state. In [11], we developed a method to extract body cues from RGB video, by first detecting a person's face and hands, then inferring an approximation of the 3-D pose of the upper body, and finally using this representation to do recognition of basic conversational cues like self-touch and gestures (see Figure 4). These approaches have been later extended to use the depth information from Kinect.

The suite of nonverbal cues was used in [9] both for correlation analysis and a regression task, where the hirability measure provided by the trained student was the variable to be predicted. Regarding correlation, the results showed that candidates who spoke longer and faster, and who took longer speaking turns received higher hirability scores. Visual features related to the amount of head motion also showed positive effects with hirability. For the regression task, using the coefficient of determination (R^2) as performance measure, the approach achieved a best result of R^2= 0.36 using ridge regression and all features extracted from a candidate. This initial result shows promise, but overall the problem is challenging. As in the case of emergent leadership, cues from the audio track were more discriminative compared to video cues. Finally, some of the cues of the interviewer turned out to be predictive of hirability, which suggests that the behavior of the interacting partner can also be informative about the self, and highlights the importance to think about this problem in contextual terms.

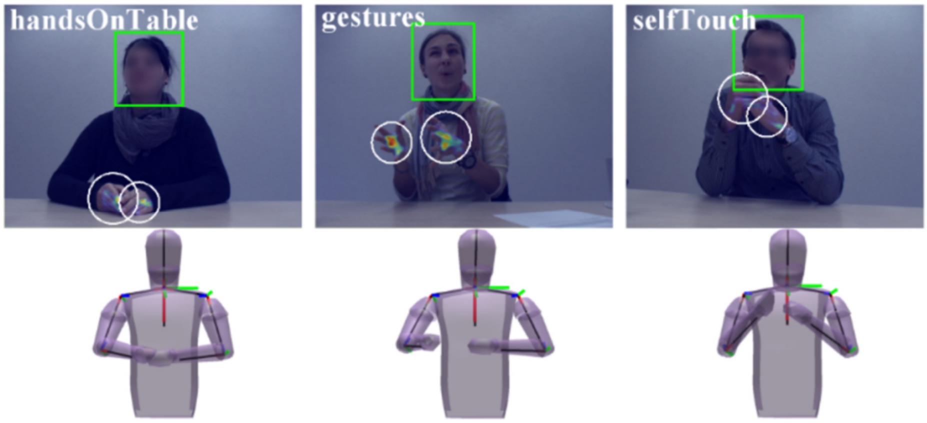

Figure 4. Interview frames with face and hands detection outputs, recognized activity, and estimated 3-D upper body pose (taken from [11]).

Perspectives

This column summarized our experience studying two research problems in organizational psychology using automatically measured nonverbal behavior and machine learning. More generally, how can research at the boundaries between signal processing and organizational psychology be expanded? Three possible directions are the following.

First, we need to communicate the possibilities of multimodal signal processing and machine learning methods within the social and organizational psychology communities, creating further partnerships where common goals can be defined and pursued. In their discipline, our collaborators have advocated for the benefits of this approach in their specific research, and have shared experiences on how similar work could be incorporated into other research lines [12]. As with other examples of multidisciplinary work, there are important issues of language, methodology, expectations, and practices that need to be sorted out. Should engineers only be service providers for psychology labs? What is the level at which automation should stop? What is the value (and the place) of computational approaches for recognition that are high performing but less interpretable? What is the level of experimental control that a discipline is willing to lose in order to conduct experiments in the wild? These are a few questions that we have encountered in our own work.

Second, from the perspective of ubiquitous applications, interactivity is key. Some aspects of the methodology presented here could be embedded in real-time awareness tools to support sectors in industry where privacy-sensitive feedback at work would be positive. This includes hospitality, sales, and public communication. Another relevant dimension is training [13]. In addition to smartphones, the current surge of wearable devices including wristbands, smart watches, and glasses are opening new ways to sense and interact. Ethics and privacy need to be a fundamental part of future designs.

Finally, as new studies from the lab towards real workplaces become possible, computational models to handle longitudinal and relational data are needed. While lab studies are intrinsically localized in time, future work that aims at understanding teams in the workplace over days, weeks, or months require of thinking about time and relations in a different way (for example, dynamic graphs with multidimensional attributes at multiple time scales.) This is a direction where signal processing methods could be especially useful, both via adaptation of existing techniques and through the development of new frameworks.

References

[1] American Time Use Survey Summary, June 18, 2014, http://www.bls.gov/news.release/atus.nr0.htm

[2] M. Remland, “Uses and consequences of nonverbal communication in the context of organizational life,” In V. Manusov, & M. Patterson (Eds.), *The SAGE Handbook of Nonverbal Communication,* SAGE Publications, 2006.

[3] J. A. Hall, E. J. Coats, and L. Smith, “Nonverbal behavior and the vertical dimension of social relations: A meta-analysis,” *Psychological Bulletin*, 131(6):898–924, 2005.

[4] R. T. Stein, “Identifying emergent leaders from verbal and nonverbal communications,” *Personality and Social Psychology*, 32(1):125–135, 1975.

[5] D. Sanchez Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez, ``A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups,'' *IEEE Trans. on Multimedia*, Vol. 14, No. 3-2, pp. 816-832, Jun. 2012.

[6] D. Sanchez Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, and D. Gatica-Perez, "Emergent Leaders through Looking and Speaking: from Audio-Visual Data to Multimodal Recognition,''

Journal of Multimodal User Interfaces, Special Issue on Multimodal Corpora, Vol. 7, Issue 1-2, pp. 39-53, Mar. 2013.
[7] A. S. Imada and M. D. Hakel, "Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews," *Applied Psychology*, 62(3):295–300, 1977.
[8] R. J. Forbes and P. R. Jackson, "Non-verbal behaviour and the outcome of selection interviews" *Occupational Psychology*, 53(1):65–72, 1980.
[9] L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez, "Hire Me: Computational Inference of Hirability in Employment Interviews based on Nonverbal Behavior," *IEEE Trans. on Multimedia*, Vol. 16, No. 4, Jun. 2014.
[10] L. S. Nguyen, J.-M. Odobez, and D. Gatica-Perez,``Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interaction,'' in *Proc. ACM Int. Conf. on Multimodal Interaction (ICMI)*, Santa Monica, Oct. 2012.
[11] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. S. Nguyen, and D. Gatica-Perez, ``Body Communicative Cue Extraction for Conversational Analysis,'' in *Proc. IEEE Int. Conf. on Face and Gesture Recognition (FG)*, Shanghai, Apr. 2013.
[12] D. Frauendorfer, M. Schmid Mast, L. S. Nguyen, and D. Gatica-Perez, ``Nonverbal Social Sensing in Action: Unobtrusive Recording and Extracting of Nonverbal Behavior in Social Interactions Illustrated with a Research Example,'' *Journal of Nonverbal Behavior*, Special Issue on Contemporary Perspectives in Nonverbal Research, Vol. 38, Issue 2, pp. 231-245, Jun. 2014.
[13] M. E. Hoque, R. W. Picard, "Rich Nonverbal Sensing To Enable New Possibilities in Social Skills Training," *IEEE Computer*, Vol. 47, no. 4, April, 2014.

Acknowledgments

The research discussed here is joint work with colleagues at the University of Lausanne, Switzerland (Marianne Schmid Mast and Denise Frauendorfer), Idiap (Dairazalia Sanchez-Cortes, Oya Aran, Laurent Nguyen, Alvaro Marcos, Dinesh Babu Jayagopi, and Jean-Marc Odobez), and other institutions (Tanzeem Choudhury, Cornell University; Marta Marron, University of Alcala, Spain, and Daniel Pizarro, University of Auvergne, France.) The author thanks all of them, and acknowledges the support by the Swiss National Science Foundation (SONVB and UBImpressed projects) and the European Commission (NOVICOM project).

Author

Daniel Gatica-Perez (gatica@idiap.ch) is Head of the Social Computing Group at Idiap Research Institute and Maître d'Enseignement et de Recherche at the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland.