HAL
open science

Social Intelligence Modeling using Wearable Devices

Alaeddine Mihoub, Grégoire Lefebvre

To cite this version:

Alaeddine Mihoub, Grégoire Lefebvre. Social Intelligence Modeling using Wearable Devices. 22nd International Conference on Intelligent User Interfaces (IUI 2017), Mar 2017, Limassol, Cyprus. pp.331 - 341, 10.1145/3025171.3025195 . hal-01493849

HAL Id: hal-01493849

https://hal.science/hal-01493849v1

Submitted on 22 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Social Intelligence Modeling using Wearable Devices

Alaeddine Mihoub
Orange Labs
Meylan, France
alaeddine.mihoub@orange.com

Grégoire Lefebvre
Orange Labs
Meylan, France
gregoire.lefebvre@orange.com

ABSTRACT
Social Signal Processing techniques have given the opportunity to analyze in-depth human behavior in social face-to-face interactions. With recent advancements, it is henceforth possible to use these techniques to augment social interactions, especially the human behavior in oral presentations. The goal of this paper is to train a computational model able to provide a relevant feedback to a public speaker concerning his coverbal communication. Hence, the role of this model is to augment the social intelligence of the orator and then the relevance of his presentation. To this end, we present an original interaction setting in which the speaker is equipped with only wearable devices. Several coverbal modalities have been extracted and automatically annotated namely speech volume, intonation, speech rate, eye gaze, hand gestures and body movements. An offline report was addressed to participants containing the performance scores on the overall modalities. In addition, a post-experiment study was conducted to collect participant's opinions on many aspects of the studied interaction and the results were rather positive. Moreover, we annotated recommended feedbacks for each presentation session, and to retrieve these annotations, a Dynamic Bayesian Network model was trained using as inputs the multimodal performance scores. We will show that our assessment behavior model presents good performances compared to other models.

Keywords
Oral presentation; Social intelligence; Wearable devices; Multimodal behavior assessment; DBN; Feedback.

INTRODUCTION

Oral communication is considered one of the most valued interpersonal skills in variety of domains such as education, business and politics [30]. This task of public speaking stresses many people and was featured in many surveys as their number one fear, even higher than death [49]. Fortunately, as stated in literature of clinical psychology [11], this skill is not only a bestowed gift to minority of charismatic individuals but rather a skill that can be learned and improved. Nevertheless, being an efficient communicator requires a lot of practice and training notably with experts and professional coaches. It requires to continually modulate both verbal and coverbal cues, and to perceive, interpret and react to other's displays and signals. The nonverbal communication in particular is acknowledged to play a significant role in social interactions [3, 28]. According to [2, 16], at least 65 percent of the information in conversations, are conveyed via nonverbal behaviors. Moreover, studies on conversational behavior have confirmed that these coverbal cues – such as body posture, hand gestures, facial expressions, eye gaze, speech energy, intonation and speech rate – are strongly involved in maintaining social attention and social glue [31]. Thus, mastering coverbal behavior in public speech is a key success for being a brilliant communicator, that is what we call the social intelligence [1].

Social intelligence as defined by Vinciarelli et al. [48] is the facet of our cognitive capabilities that guide us to interact harmoniously with others in different situations and contexts. In this paper we present a specific social face-to-face interaction carefully designed to model the social intelligence of a speaker in an oral presentation scenario. We present an original setting using only wearable devices in order to study the subtleties of human behavior in public speech. Our main challenge is to train a behavioral assessment model able to interpret the presenter nonverbal performances, estimate the cognitive state of the presenter and then predict an appropriate feedback. The feedback is aimed to enhance the awareness about coverbal attitude and thus the social intelligence of the speaker. In this work, the feedback modeling is solved in a data-driven way through a Dynamic Bayesian Network whose graphical structure describes the complex relationship between the cognitive state of the presenter, his multimodal performance scores and the adequate feedback. The long-term goal of our research is to implement our behavioral evaluation model on wearable devices and test in real-time the relevance of our modeling and the social acceptability of the setting. More generally, this research can be enlarged to many other domains besides public speaking such as job interviews, customer services and sales, communication for health-care professionals, helping people with social difficulties, etc.

The paper is organized as follows: the next section briefly reviews the state of the art of feedback systems and models. In section 3, we introduce our oral presentation interaction: the proposed scenario, study protocols, experimental setting and recorded signals are described in details. In section 4, the whole process of data annotation is described. In section 5, the post-experiment study is presented and we will show that overall opinions were positive. In section 6, the DBN model is introduced and we will show its relevance in predicting suitable feedbacks. Section 7 summarizes our contributions and concludes the paper.

RELATED WORK

This research is part of the automatic analysis of human behavior in social interactions [13]. Automatic conversations analysis tries to infer from raw signals information about social actions, social emotions, social relations and social attitudes [48]. Depending on the application domain, several computational models have been proposed to cope with these issues. For instance, Mihoub et al. [35] presented a behavioral model designed to endow social robots with social skills enabling them to engage relevant interactions with human interlocutors. The model was based on Hidden Markov Models and was able to analyze human-partner behavior and generate adequate coverbal behavior particularly eye gaze fixations. Nihei et al. [40] established classification models that estimate influential statements in group discussions using prosodic features, head motion and eye gaze information. Group discussions are used in some countries for job recruitment processes. Expert recruiters observe the social behavior of all candidates and select only the good discussion performers. Through their modeling, Nihei et al. aimed to elaborate automatic skill evaluation through automatic scene analysis. They showed that persons who make influential statements frequently present better facilitation ability in discussions. To enhance group collaboration, Kim et al. [27] introduced the Meeting Mediator (MM), a real-time system that detects social interactions in meetings and provide feedback on the interaction dynamics. The exchanged social signals - based essentially on speech features and body movements - were captured by Sociometric portable badges and visualized on mobile phones of the interacting interlocutors. They showed that the system was able - in a brainstorming scenario - to ameliorate interactivity level and to reduce behavioral differences between non-dominant and dominant people.

Besides promoting social dynamics in-group meetings, one major application of social augmenting is to enhance social behavior in public speak. In this context, three configurations are identified in the literature: (1) Behavior assessment in training sessions, i.e. the feedback is provided before a real presentation situation. (2) Feedback provided online, i.e. during a real presentation situation. (3) Feedback provided offline, i.e. after a real presentation situation. The presentation trainer [42] is a recent example of systems within the first configuration. It is a multimodal tool designed to track the user's body and voice (body posture, gestures, volume, pauses, etc.) using a microphone and a Kinect®. The feedback is transmitted through a visual graphical interface and a connected wristband (haptic information). Two types of feedback (corrective and interruptive) were provided using a rule-based model. For the latter type, the system considers the user's mistake as severe and interrupts the speaking. To reduce the cognitive load of the presenter only one feedback is instructed at a time and the system was demonstrated to have significant effect on users' learning. MACH (My Automated Conversation coacH) [21, 22] is another well-known system in the literature. Through a virtual agent the system simulates a job interview with human participant. At end of session, a summary feedback is provided concerning many aspects of the nonverbal communication such as total pause duration, speaking rate, weak language (filler words such as "basically" and "umm"), pitch variation, smiles, head gestures, loudness, etc. In the same context, Chollet et al. [5] explored an interactive virtual audience paradigm to create a forgiving and engaging environment for public speak training. The virtual audience was projected on a life-size screen and was animated by a wizard of oz strategy. The system was able to give direct visual feedback on the screen but also implicit feedback via the nonverbal behavior of the animated characters. For instance a character could exhibit a forward or a backward posture depending on speaker's performances. The system was shown to have positive effect on learners' audiovisual behavior.

The second type of systems proposed in literature, are those delivering online feedback during real presentations. Knowing that human attention is a limited source [43] and that human brain is not adept for multitasking [41, 45], the challenge is to convey a relevant feedback in a convenient format and with an optimal frequency, so that speaker is not distracted from his task. For instance, Rhema [46] is an intelligent user interface for wearable Google® glasses, developed to help presenters modulate speech volume and speech rate during their talk. The paper shows that optimal feedback should be one instruction at a time and in single words format (e.g. louder, faster, etc.). One limitation of this work is the lack of multimodality in the behavior assessment. Logue [7] is another system designed to deliver in-situ real-time feedbacks about the nonverbal behavior of a speaker, especially speech rate, body energy and openness. The social sensing is performed by a microphone, a Kinect® placed in front of the speaker and wearable Vuzix® glasses which were used also for feedback rendering. The feedback generation was provided by a rule-based model and a set of manually fixed thresholds. With two user studies, the authors showed that Logue system was perceived to be helpful, and that it has got significant impact on users' performances. The multi-sensor self-quantification framework [12] is a feedback system belonging to the third aforementioned configuration. It

generates an offline report about many aspects of the oral presentation such as speaking rate, liveliness, body movement, gestures, speaker attention, audience engagement, etc. It utilizes a whole motion capture system composed of two static cameras, one Kinect®, and three wearable Google® glasses (one for the user and two for public). The feedback model was based on Support Vector Machine models trained with annotated multimodal data. The ground truth was collected using manual annotation of videos segmented in multiple 10-second clips. The user study showed that analytics report generated after each presentation received positive opinions from the experiment participants.

Compared to literature, our research paradigm is intended to offer the final user the advantages of the three configurations. In fact, in this paper we claim the following contributions:

- An original social sensing setting not based on classic motion capture systems but only wearable devices that are easily portable and susceptible to be shortly available for general public.
- At this state, our system offers training and practice possibilities as well as detailed offline reports on behavioral performances. It is intended to integrate real-time feedbacks in the near future.
- Behavior evaluation is performed to be as exhaustive as possible in order to deal with almost coverbal modalities (speech features, hand/body motions and eye gaze).
- Contrary to existing feedback models (rule-based, classic classifier, wizard of oz, etc.), we propose an evaluation model based on a probabilistic graphical model (DBN) that is more appropriate in modeling temporal complexity of human behavior dynamics.

BEHAVIOR ASSESSMENT EXPERIMENT

Proposed scenario

The objective of the proposed interaction is to collect multimodal behaviors of many speakers in an oral presentation scenario (cf. Figure 1). The presenters were asked to give a "pitch" presentation for five minutes. Each presentation was composed of five slides in which each slide corresponds to one of the following items: presenter education, professional experiences, technical/interpersonal skills, current projects, and future projects. In our experiment, this pitch scenario was particularly chosen since it highly motivates participants to be at the top of their performances compared to other presentation topics.

Study protocols

11 presenters (7 male and 4 female) were recruited to participate in our experiment. Our study was limited to French native speakers in order to have homogeneous behavioral patterns and a standard annotation protocol. Participants' ages ranged from 22 to 48 years old with an average of 34. They had different functions: interns, developers, PhD students, researchers, ergonomists and project managers. Two weeks before recording, participants were asked to fill an online form about (1) their demographic information (2) self-assessment of skills in oral presentations and (3) their personal Curriculum Vitae. On a 10-point scale question (1 = never, 10 = very frequent), our participants rated themselves on average 5.36 for frequency of giving presentations, 6.27 (1 = not at all, 10 = very comfortable) for their comfort in public speak and 6.45 (1 = very bad, 10 = very good) on how good they think their nonverbal communication is. Experimenters used the third information (i.e. personal CV) to build presentation slides that were sent to participants one week before their talk. This protocol helped us to have standard presentations for all speakers, which neutralizes preparation effects on behavior performances. Each presenter was asked to give two presentations on two different days to minimize learning effect, which leads to a total of 22 presentations. These presentations were recorded in 7 sessions with groups of three or four participants. The mean length of a session was about 1 hour and 15 minutes. The groups were designed to mix the 11 participants in different groups. Since we have two presentations per participant, this protocol avoided presenters' lassitude caused by the repetition of a talk in front of the same person. For each group (for instance a 4-person group), we have one presenter and three observers who constitute the audience.

At the beginning of each session, the experimenter presents the study and its progress phases. He also takes few minutes to expose the importance of nonverbal communication in a public speak and gives some recommendations about it. This step guarantees that participants have similar baseline knowledge about coverbality in public speak. After each presentation, one form is filled by the speaker for self-assessment and a second form is filled by an audience member that we call the evaluator. Two tablets were used to this end; collected information from these questionnaires will be discussed in the post-study experiment section. After each presentation, a role changing strategy is applied by the experimenter so that each person participates as a presenter but also as observer for the rest. Particularly, he/she will be an evaluator for a specific speaker by rating his/her behavior performances through the second questionnaire. Each evaluator was discreetly informed from the beginning about the person to evaluate. Note that, our changing role strategy ensured that two persons couldn't evaluate theirself mutually. The evaluators were instructed to pay attention to the coverbal behavior of the speakers especially voice, gaze, and body language.

Experimental setting

The aim of the setting (cf. Figure 1 and Figure 2) is to sense the audiovisual social behavior of a presenter particularly voice attributes (volume, intonation, speech rate), eye gaze fixations, hand gestures and body movements. In order to

make our evaluation system portable and usable in different contexts, only the presenter was equipped with wearable devices (cf. Figure 2). For same reasons, we omitted the use of heavy motion capture systems by utilizing only wearables in part already democratized for public. Therefore, the multimodal behavior of a presenter was captured by:

- A smartphone (Samsung®) put in the pocket of the presenter, used to track body movements.
- A smartwatch (Sony®) worn on the directional hand, used to track hand gestures.
- Smart glasses (Epson Moverio BT-200®), a head-mounted device used to capture eye gaze fixations.
- A microphone connected to the glasses, used to monitor the presenter's speech signal.

For the purpose of annotation we also equipped the scene with a static camera in order to film the whole interaction.

Figure 1: Experimental setting of a presentation

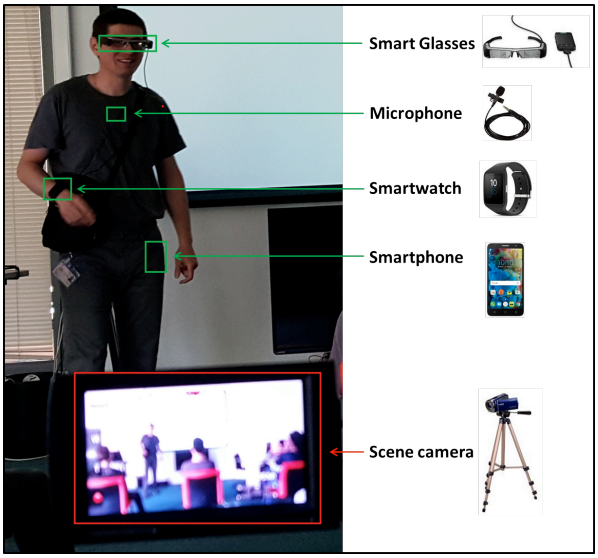

Figure 2: List of devices used to capture presenter's behavior

Recorded signals

For data acquisition a recording platform was developed in order to collect multimodal and synchronized data from all devices. The recorded signals are as follow:

- Accelerometer and gyrometer data from the pocket smartphone
- Accelerometer and gyrometer data from the smartwatch
- Accelerometer and gyrometer data from the glasses
- Egocentric video scene from the front camera of the glasses
- The speech signal from the microphone
- Video environment from the static camera

Human behavior is paced by subtle temporal coordination between different modalities. Thus, the challenge for our recording platform was to collect those signals in highly synchronized way; otherwise all social behavior modeling will be invalid. To this end, two Android® applications were developed to handle recordings. The first application runs on the experimenter smartphone (Samsung®). It allows triggering recording for both presenter smartphone and smartwatch. The second Android® application monitors the glasses and the connected microphone. The two applications were controlled by the experimenter to help orators focus on their presentation task. Note that manual claps are also performed before and after each presentation in order to synchronize precisely the different modalities especially audio and video.

DATA ANNOTATION

Similar to many previous works [12], we chose to segment each presentation on 10-second clips. This frame granularity was used for all modalities in our annotation process. Mean duration of the presentations was about 37 frames (i.e. 6 minutes and 10 seconds long) and the overall time of presentations was about 2 hours and 15 minutes. The experiment design was thought to handle a list of selected modalities known to largely contribute on public talk pertinence. Next paragraphs describe the set of signal processing based rules utilized to infer these modalities from recorded signals.

Volume

A principal characteristic of human voice is volume. Known also as intensity or speech energy, the volume represents the loudness of an audio signal and plays fundamental role for transmitting clear messages for audience [10]. To analyze volume, we used an application based on the intensity extraction method of the Praat software [4]. We extracted also silent intervals for all segmented clips with a silent threshold of –30 dB, a minimum silent interval duration of 0.1 sec and a minimum sounding interval duration of 0.05 sec. For each 10-second clip, we first delete silent intervals and we calculate the mean from remaining intensities. Then, using two intensity thresholds [17] (50 dB and 60 dB), the volume of the interval is classified into three categories: low, good and high. Thus, volume levels under 50 dB were considered as

low while volume levels above 60 dB were considered as high. Note that these thresholds may slightly change depending on room noise and distance from audience.

Intonation

One major attribute of human prosody is intonation. It represents the way speakers modulate their voices and plays significant role for retaining audience's attention [10]. Pitch is known to be the main acoustic feature that correlates with intonation. Hence, for intonation analysis, we automatically extract pitches using an application based on Praat software [4]. For the extraction, we used a minimum pitch of 75 Hz and defaults parameters of the Praat autocorrelation method. To quantify pitch variation, it is possible to use the raw pitch standard deviation as a metric. However, because of differences between speakers especially males and females, this metric will give invalid and unfair comparison results [20]. Therefore, as proposed by R. Hincks [19, 20] the standard deviation should be expressed as a percentage of the pitch mean. This normalization is performed by dividing the pitch standard deviation by the mean; the resulting quotient is known by the acronym PVQ (Pitch Variation Quotient). Note that in Hincks works, the PVQ is also calculated for 10-second clips. In our work, we first calculate for each 10-second segment the corresponding PVQ. Then, using two thresholds inspired from [19] (0.10 and 0.23), the intonation is classified into three categories: monotone, medium and lively. In this way, PVQ values under 0.10 characterize monotonous speech while values above 0.23 reflect lively voice.

Speech rate

Balanced speech rate is another important attribute featuring voice quality. In fact, a slow rate may cause boredom while an accelerated rate may cause incomprehension and ambiguity. To measure speaking rate one possibility is to use the word per minute (WPM) metric. Nevertheless, because of major difference in word length, this measure can be imprecise in many contexts [15]. Expressing speech rate in syllables per second (SPS) solves this problem and presents many advantages. It allows a local analysis of voice and more efficient track of its rate variations [20]. In our work, the SPS measurement is used to characterize speech rate. In particular, we relied on Jong and Wempe algorithm [25] to segment our audio clips into syllables. The speech rate is then computed by dividing the number of detected syllables by the length of clips, i.e. 10 seconds. Afterwards, using two thresholds inspired from [8] (5.8 sps and 10 sps), the speech rate is classified into three categories: low, good and high. Accordingly, SPS values fewer than 5.8 feature low speaking rates while values above 10 distinguish highly speech rates.

Gaze

Besides retaining their attention, sustained eye contacts with listeners make the orator look engaged, more believable and more convincing. Further, it helps the presenter to better receive listeners' nonverbal signals (e.g. facial expressions) and then, responds and enhances his message deliver. In our work, eye gaze performances were annotated semi-automatically. First, using the videos of the glasses, we annotated manually fixations over four regions of interest: audience, personal computer, projection screen and elsewhere. From these annotations, we computed for each 10-second segment a gaze distribution that contains the percentage of each region of interest. Then we established a rule-based model that uses a list of thresholds in order to automatically classify the gaze segment into three categories: improvable, medium and good. At this state, fixations are annotated manually but we intend in the future to use computer vision techniques to automatically detect regions of interest. In particular, we may use accelerometer and gyrometer data of the glasses - coupled with visual features - in order to get better estimations.

Hand gestures & Body movements

Hand and body movements are known to be powerful tools to enhance social influence. Hand gestures for instance help to add emphasis and clarity to spoken words. Body movements help also to reinforce verbal messages and further ameliorate audience attraction [47]. In our setting, hand and body movements were sensed by wearables devices. In particular, hand gestures were inferred from accelerometer data from the smartwatch, while body movements were extracted from accelerometer data from the pocket smartphone. For both modalities, a high-pass filter was applied to eliminate gravity force. Then, using the 10-second segments, we computed an energy measurement - based on effective power of signal [34] - in order to characterize gesture energy and body energy of the speaker. Based on two thresholds, the gesture energy was classified into three categories: low, good and excessive. Similarly, two other thresholds were used to assign body energy to one of those classes. Notice that in our annotation, we used only accelerometer data to describe overall energy of the presenter. In the future, we may also exploit gyrometer data in order to get more precisions on types and categories of gestures.

Feedback

After analyzing all videos, nine specific feedbacks about nonverbal behavior of the presenter were proposed with an audience point of view:

- + volume : louder speech volume
- + liveliness : more voice liveliness, used when intonation is perceived as monotone
- + speech rate : faster speech rate
- - speech rate : lower speech rate
- + public : more eye contacts with the public
- + gestures : more hand gestures
- + body : more body movements

- none : no feedback, used when it is not appropriate to deliver a feedback at that moment
- good : used when nonverbal behavior is evaluated as good

As mentioned before, human brain is not adept for multitasking [41, 45], for this reason we chose to deliver only one feedback per 10-second segment. Efficiency of this paradigm has been already demonstrated in the literature [42, 46]. Each 10-second clip was affected to one of those feedbacks by two expert annotators that have excellent knowledge about nonverbal communication in public speech. ELAN software [44] was used to visualize and to annotate each presentation. Annotators were instructed to give priority firstly to voice attributes and gaze features, and secondly to hand gestures and body movements. These instructions were based on results from the post-experiment study. In fact, participants were asked about importance of each modality in public presentation and the final statistics (cf. Figure 3) showed that voice features, as well as gaze, were declared more important than hand and body energies. Annotators were also instructed to consider annotated feedback history during the annotation process in order to get diversified feedbacks and avoid sustained emphasis on a particular modality. Note that in case of mismatch between annotators, they were invited to debate on video segment and choose only one final feedback. In the Feedback modeling section we will propose a Dynamic Bayesian Network model able of retrieving these annotated feedbacks from automatic multimodal scores.

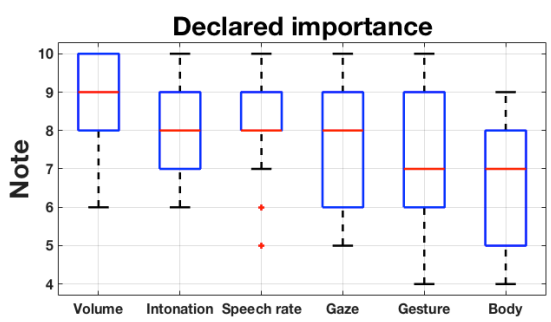

Figure 3: Declared importance (from audience) of coverbal modalities for presentation relevance (1= not important at all, 10=very important). Voice is underlined the most important modality especially volume. Gaze fixations are also emphasized but with more diverging opinions. Finally, gestures and body movements are perceived as significant but less crucial.

POST-EXPERIMENT STUDY

The aim behind our post-experiment study is to (1) survey both presenters and audience about many aspect of the experiment and (2) collect participants' opinions about the quality of our automatic assessment system. Three questionnaires were established to this end: the first one is filled by the speaker just after his/her presentation, the second one is filled by an audience evaluator and the third one is filled by the speaker after getting the automatic evaluation. In this section we first propose qualitative results about the studied interaction and then qualitative results about the automatic offline scores generated by our rule-based annotation models.

Qualitative results about the experiment

Figure 4 proposes evaluations from the presenters about eight criteria of the studied interaction. These criteria correspond to the following sub-questions:

On a scale of 1 to 10, do you agree or disagree with the following statements (1 if you totally disagree, 10 if you totally agree):

1. **Awareness**: *"I was aware of my coverbal behavior during my presentation."*
2. **Concentration**: *"I was very focused during my presentation."*
3. **Comfort**: *"I was comfortable during my presentation."*
4. **Attention**: *"I was attentive to audience reactions."*
5. **Encumbrance:** *"The experimental devices took up too much space during the presentation."*
6. **Perturbation**: *"The experimental devices disrupted my social interaction with the audience."*
7. **Online Feedback**: *"To improve my presentations, it will be good to use the experimental devices to get feedbacks about my coverbal behavior **during** my presentation."*
8. **Offline Feedback**: *"To improve my presentations, it will be good to use the experimental devices to get feedbacks about my coverbal behavior **after** my presentation."*

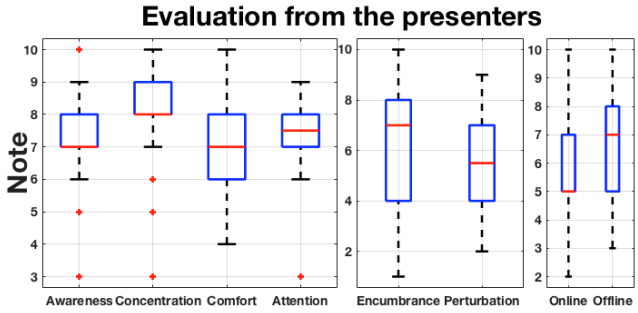

Figure 4: Evaluations from the presenters. Notation scale was from 1 to 10 (1=total disagreement, 10=total agreement).

About the first four criteria, participants have given quite good evaluations, showing in part that the experimenter presentation about coverbal behavior was well perceived. One participant gave the following comment: "*My attention*

was more focused on my body movements during the second session because I was more comfortable". About experimental device perception, statistics shows divided opinions on the perturbation criteria. However, the box plots show relatively high median for the encumbrance criteria, which is due principally to the glasses type. One user explains "*I find it extremely heavy*". This particular issue would be probably solved in the future thanks to more recent and light devices. Concerning online versus offline feedbacks, the evaluations from the presenters give a preliminarily preference to offline feedbacks. One participant judges online feedbacks as follows: "*I imagine it would be too disturbing*". Consequently, the challenge for an online assessment system will be to optimize timing and manner of the feedback render. Evaluations from audience evaluators are presented in Figure 5. They concern three criteria that correspond to the following sub-questions:

1. ***Attention**: "I was careful to the coverbal behavior of the presenter."*
2. ***Comfort**: "I was comfortable during the presentation even if the presenter wears connected devices."*
3. ***Perturbation**: "The experimental setting disrupted social interactions between presenter and audience."*

The figure plots expose relevant values about the audience attention, which respected the experiment instructions. For the other criteria, a mixed feeling is observed. We noticed positive feelings about comfort criterion but also relatively negative ones about perturbation. Indeed, some people complain about the glasses opacity, limiting thus eye contacts. This result joins presenter evaluation about used glasses.

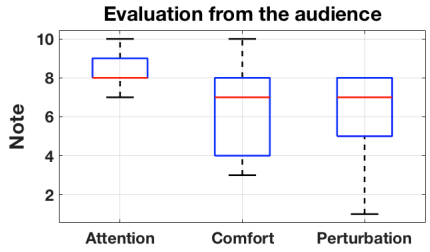

Figure 5: Evaluations from the audience. Notation scale was from 1 to 10 (1=total disagreement, 10=total agreement).

Qualitative results about the proposed offline feedbacks

The multimodal signal analysis gave us the opportunity to provide offline feedbacks to the participants about their coverbal behaviors. Figure 6 presents one of them. For the two sessions (S1 and S2) of the concerned participant, an automatic classification is computed for the six coverbal modalities. First, a learning effect is proved between the two sessions (i.e. for all participants better scores were overall recorded for their second session). In case of testing online feedback system, it will be necessary to exactly quantify learning effect in order to prove that amelioration over time is due to the intelligent system and not simply to a learning effect. For this particular presenter, the speech analysis presents good performances concerning volume, intonation - even if this dimension could be improvable - and speech rate. In addition, charts show good performance on gaze which indicates that it was oriented to the audience most of the time. Gesture energy was positively evaluated except few excessive segments during the second session. Looking on videos, excessive intervals corresponded to sudden and saccadic gestures performed by the speaker. An improvement on body movements is also notably observed between the two sessions.

Moreover, we collected some verbatim and evaluations from our participants concerning these offline feedbacks. Some verbatim are positive: "*I find very interesting having a numerical target, because it helps to realize and quantify our performance, and this is perhaps more credible than the comments of a third party*", "*Interesting in order to progress, particularly the energy levels*". Some verbatim are less positive: "*To be honest, I expected such results because I am aware of my weaknesses (it does not mean that I know how to resolve them)*". Some verbatim give some perspectives: "*For intonation, I would have to "perceive" what I can change (e.g. a difference with another person who does better)*", "*For the eye contacts, it could be interesting to visualize in which direction I look the most*".

Finally, on 10-point scale questions, our participants rated on average 7.45 for the relevance of provided statistics in general, 7.36 for the relevance of those statistics to better understand coverbal behaviors, 6.55 for their relevance to improve presentations, and 6.82 as a motivation level to participate in future experiments with online feedbacks.

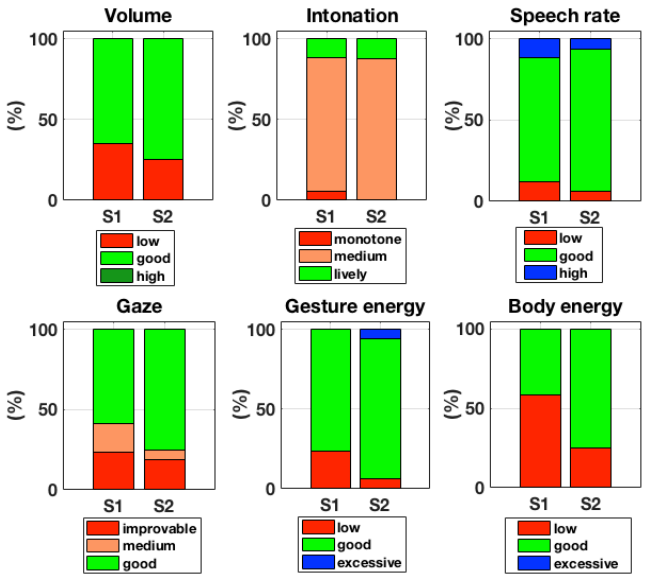

Figure 6: Offline multimodal scores for one presenter. S1 corresponds to session 1 and S2 to session 2.

FEEDBACK MODELING

The main challenge of our feedback modeling is to coach social intelligence to public speakers. To this end, we propose new models – based on Dynamic Bayesian Networks (DBNs) – for behavior assessment and feedback generation. In the next paragraph, we start with a brief review on DBNs.

Dynamic Bayesian Networks

Bayesian Networks (BNs) [39] belong to the family of Probabilistic Graphical Models (PGMs) [29]. PGMs have married in a same formalism graph theory and probability theory in order to provide intuitive and effective tools to represent a joint probability distribution over a set of random variables [26]. Formally, a Bayesian Network (BN) is a directed acyclic graph whose nodes are random variables and whose edges represent conditional dependencies between these variables. An edge linking a parent node X to a child node Y semantically means that node X exerts direct (or causal) influence over node Y. Dynamic Bayesian Networks (DBNs) [38] has extended static Bayesian Networks (BNs) by modeling the temporal evolution of the variables. Beyond intuitive graphical representation and uncertainty dealing, they efficiently model complex temporal relationships between variables. Since we consider processes that are only Markovian (i.e. variables at slice t are only dependent on t-1 slice variables) and stationary (i.e. conditional probability distributions do not change over time), a DBN can be sufficiently depicted by a 2-timeslice structure called 2-timeslice Bayesian Network (2-TBN) (see example in Figure 7) [14]. Due to their multiple characteristics, DBNs have been used successfully in many domains; in particular, they represent an attractive and a powerful formalism for modeling dynamics of human behavior in social interactions [23, 36].

DBN-based models

Cognitive state

In this work, we propose to integrate the cognitive state (CS) of the presenter to our modeling in order to get better prediction of feedbacks. In fact, additional to annotated variables already described in section 4 (Data annotation), we introduced a novel variable (CS) that tries to model the cognitive state of the speaker. Cognitive states condition perception-action loops of the human behavior: they reflect cognitive processes that coordinate modalities of human behavior but also contextualize the interaction by considering its circumstances and its evolution. Examples of CS could be “informing”, “nodding”, “listening”, “turn taking”, etc. Notice that one single CS could also sequence many sensory-motor states together. Similar concepts have been proposed in the literature of multimodal behavior modeling [23, 32, 35]. In our application, we used data mining techniques to explore the cognitive states of the presenter. Actually, we used the K-means clustering algorithm [24] to infer these states from both multimodal scores and feedbacks (volume, intonation, speech rate, gaze, gesture energy, body energy, feedback). Indeed, associating scores and feedbacks this way, allows the estimated states to better contextualize interaction phases. For information, Weka Java API [18] has been used to realize the clustering task. After many empirical tests (9, 18, 27, 36), we chose the optimal number of 27 clusters, which corresponds, to 27 specific CS.

Proposed Model

Based on Dynamic Bayesian Networks (DBNs), our model uses machine learning techniques to intrinsically associate CS and multimodal performance scores to annotated feedbacks. The main objective is to retrieve annotated feedbacks and to prove usefulness of this model for future online coaching system. List of used variables, as well as their respective cardinalities were as follows:

- CS : cognitive state of the presenter, 27 states
- V : Volume, 3 levels (low, good, high)
- I : Intonation, 3 classes (monotone, medium, lively)
- S : Speech rate, 3 levels (low, good, high)
- G : Gaze fixations, 3 classes (improvable, medium, good)
- H : Hand gesture energy, 3 levels (low, good, excessive)
- B : Body energy, 3 levels (low, good, excessive)
- F : Feedback, 9 types (+ volume, + liveliness, + speech rate, - speech rate, + public (gaze), + gestures, + body, none, good)

Our proposed model is illustrated in Figure 7. The internal structure of our DBN was designed to efficiently model causal relationships between different nodes. This causality graph presents interesting properties:

- The cognitive state of the speaker influences his multimodal behavior (blue arrows).
- The appropriate feedback is a direct consequence of multimodal scores of nonverbal behavior (green arrows).
- The appropriate feedback is also directly influenced by the mental state of the speaker (red arrows).
- The network reflects the temporal correlation between cognitive states since a CS at a certain time is influenced by the previous state of the speaker (grey arrows).
- The network reproduces an important instruction that was given to annotators concerning feedback; this instruction has incited annotators to diversify feedbacks by considering feedback history. Looking to the structure, we can easily identify temporal dependence between feedbacks (black arrows).

To demonstrate the efficiency of the proposed causality network, we compare our principal model with a second baseline DBN whose structure does not consider the CS of the presenter (cf. Figure 8).

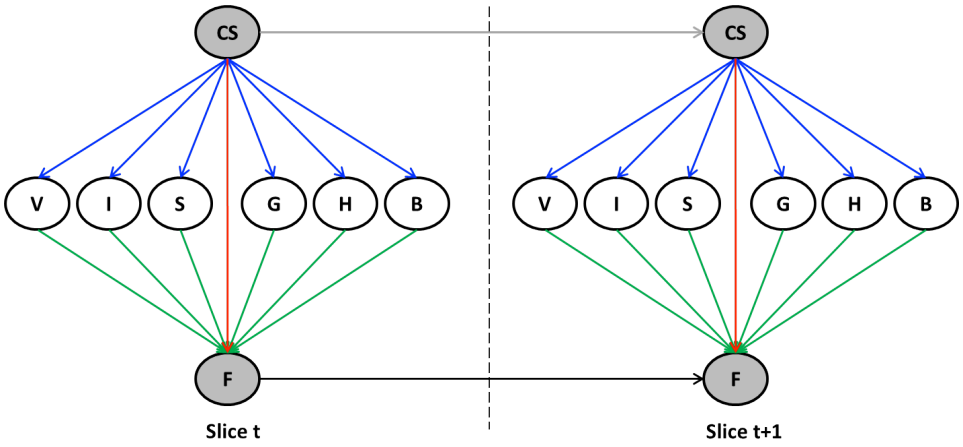

Figure 7: Our proposed DBN model (2-TBN model). Variables with white background are observable while grey ones represent variables to predict.

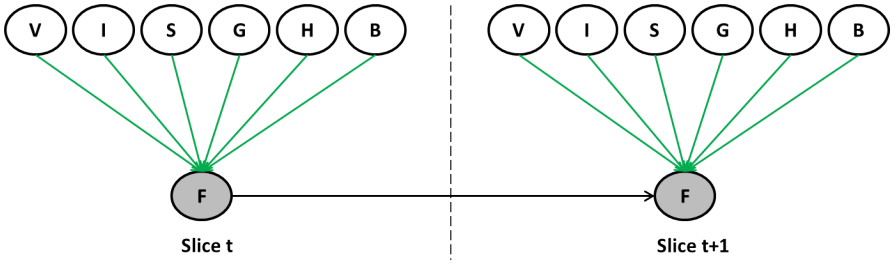

Figure 8: Baseline model: DBN structure without cognitive state (CS) modeling.

Evaluation and results

Bayes Net toolbox [37] was used for DBNs learning and feedback estimation. EM algorithm [9] is applied for training while Junction tree algorithm [6] is applied for inference. In particular, online inference is used for prediction i.e. feedback (F) at a slice t is estimated only with observable data (V,I,S,G,H,B) till that instant t. Offline inference would rather use the total sequence which is not appropriate for real-time systems. We remind that we have 22 sequences (11 participants and 2 presentations per participant). After many empirical tests (1, 2, 5, 10 seconds), we chose the initial temporal granularity of our multimodal data, i.e. 10 sec per time slice. With this granularity, mean duration of sequences filled to our DBNs was 37 frames. Using a leave-2-out-cross validation, we end up with 11 models; each model is trained on sessions of 10 participants and then tested on the two sessions of the remaining participant. Two metrics were used to quantify models performances. First, we used the precision of prediction, which corresponds to the percentage of retrieved feedbacks from original sequences. Because we are comparing similarity between two sequences, we propose a second approach based on Levenshtein distance [33]. In fact, Levenshtein distance computes a minimum number of operations to transform one sequence to another. From this optimal alignment the F-measure metric is directly computed. This method is more adequate to our problem since it tolerates small miss-alignments between original and predicted values. The final performance results are presented in Table 1. Table values correspond to calculated means on the 11 trained models. Furthermore, for the sake of comparison, we computed also random performances from the empirical distributions of annotations. The empirical distributions reflect the frequency of each feedback type in the annotation database.

	Precision (%)	F-measure (%)
Our DBN model (with CS)	**62.47**	**67.17**
Baseline DBN (without CS)	59.14	64.57
Random model	17.53	36.91

Table 1: Mean performances for feedback estimation

As we can see, performances of DBN models are largely higher than random levels. Thanks to CS modeling, our proposed model outperforms the baseline DBN (without CS) on both metrics: 62.47 vs. 59.14 for precision and 67.17 vs. 64.57 for F-measure. This performance gap

proves the relevance of our causality network in modeling the studied interaction and confirms the significant contribution of cognitive states in enhancing feedback estimation.

CONCLUSION AND PERSPECTIVES

In this paper we studied human behavior analysis in social interactions. Especially, we were interested in modeling nonverbal communication of a public speaker in an oral presentation scenario. The challenge was to develop an automatic assessment system able to assist presenters in becoming brilliant and charismatic speakers. For this purpose and contrary to many works in literature, we designed a social experiment in which we used only wearables devices for multimodal signals sensing. Using social signal processing techniques we developed rule-based models that computed performance scores concerning multimodal behavior of the speaker. In our analysis, we proceeded with a multimodal approach that allowed us to interpret not only voice attributes (volume, intonation, speech rate) but also eye gaze, hand gestures and body movements. Performance scores were communicated to participants in an offline reports (after their presentations) and received positive evaluations about their relevance and their usefulness. Similarly, overall opinions were quite positive about the experimental setting. In fact, some issues were raised up especially about the heaviness of the smart glasses. This problem will be overcome in future research by replacing that device with more appropriate one. Moreover, one step toward online feedback delivery is ensured thanks to our DBN model. Indeed, we showed that our proposed causality network (with explicit CS modeling) has led to good performance rates compared to baseline models. Starting from actual models, we will enrich in future our setting with real-time capacities in order to deliver online feedbacks, which enables the speakers to improve in real-time and in-situ their nonverbal communication.

REFERENCES

1. Karl Albrecht. 2006. *Social Intelligence: The New Science of Success*. John Wiley & Sons.
2. Michael Argyle. 1988. *Bodily Communication*. Routledge, London; New York.
3. Ray L. Birdwhistell. 1970. *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press, Philadelphia.
4. P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot International* 5, 9/10: 341–345.
5. Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. 2015. Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp '15), 1143–1154.
6. Robert G. Cowell, Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. 2003. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer New York.
7. Ionut Damian, Chiew Seng (Sean) Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting Social Interactions: Realtime Behavioural Feedback Using Social Signal Processing Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 565–574.
8. Volker Dellwo, Emmanuel Ferragne, and François Pellegrino. 2006. The perception of intended speech rate in English, French, and German by French speakers. In *Proc. 3rd International Conference of Speech Prosody, Dresden, Deutschland*, 101–104.
9. Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1976. Maximum Likelihood from Incomplete Data via the EM Algorithm. Retrieved April 9, 2015 from http://dash.harvard.edu/handle/1/3426318
10. Joseph A. DeVito. 2014. *The Essential Elements of Public Speaking*. Pearson.
11. Stephen B. Fawcett and L. Keith Miller. 1975. Training public-speaking behavior: an experimental analysis and social validation. *Journal of Applied Behavior Analysis* 8, 2: 125–135.
12. Tian Gan, Yongkang Wong, Bappaditya Mandal, Vijay Chandrasekhar, and Mohan S. Kankanhalli. 2015. Multi-sensor Self-Quantification of Presentations. In *Proceedings of the 23rd ACM International Conference on Multimedia* (MM '15), 601–610.
13. Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing* 27, 12: 1775–1787.
14. Kevin Gimpel, Daniel Rudoy, Lincoln Laboratory, United States Missile Defense Agency, and United States Air Force. 2008. *Statistical Inference in Graphical Models*. MIT, Lincoln Laboratory.
15. Roger Griffiths. 1990. Speech Rate and NNS Comprehension: A Preliminary Study in Time-Benefit Analysis. *Language Learning* 40, 3: 311–336.
16. Laura K. Guerrero and Kory Floyd. 2005. *Nonverbal Communication in Close Relationships*. Routledge, Mahwah, N.J.
17. T. Hacki. 1996. Comparative speaking, shouting and singing voice range profile measurement: physiological and pathological aspects. *Logopedics, Phoniatrics, Vocology* 21, 3–4: 123–129.
18. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 11, 1: 10–18.
19. Rebecca Hincks. 2004. Processing the prosody of oral presentations. In *InSTIL/ICALL Symposium 2004*.
20. Rebecca Hincks. 2005. Measures and perceptions of liveliness in student oral presentation speech: A

proposal for an automatic feedback mechanism. *System* 33, 4: 575–591.
21. M.E. Hoque and R.W. Picard. 2014. Rich Nonverbal Sensing Technology for Automated Social Skills Training. *Computer* 47, 4: 28–35.
22. Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My Automated Conversation Coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp '13), 697–706.
23. Chien-Ming Huang and Bilge Mutlu. 2014. Learning-based Modeling of Multimodal Behaviors for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction* (HRI '14), 57–64.
24. Anil K. Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31, 8: 651–666.
25. Nivja H. de Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41, 2: 385–390.
26. Michael I. Jordan (ed.). 1998. *Learning in Graphical Models*. A Bradford Book, Cambridge, Mass.
27. Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting Mediator: Enhancing Group Collaborationusing Sociometric Feedback. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (CSCW '08), 457–466.
28. Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. 2013. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing, Boston, MA.
29. Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.
30. Patrick C. Kyllonen. 2012. Measurement of 21st century skills within the common core state standards. In *Invitational Research Symposium on Technology Enhanced Assessments. May*, 7–8.
31. Jessica L. Lakin, Valerie E. Jefferis, Clara Michelle Cheng, and Tanya L. Chartrand. 2003. The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior* 27, 3: 145–162.
32. Jina Lee, Stacy Marsella, David Traum, Jonathan Gratch, and Brent Lance. 2007. The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (IVA '07), 296–303.
33. Vladimir Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10, 8: 707–710.
34. Richard G. Lyons. 2010. *Understanding Digital Signal Processing*. Prentice Hall, Upper Saddle River, NJ.
35. Alaeddine Mihoub, Gérard Bailly, Christian Wolf, and Frédéric Elisei. 2015. Learning multimodal behavioral models for face-to-face social interaction. *Journal on Multimodal User Interfaces*: 1–16.
36. Alaeddine Mihoub, Gérard Bailly, Christian Wolf, and Frédéric Elisei. 2016. Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognition Letters* 74: 82–89.
37. Kevin P. Murphy. 2001. The Bayes Net Toolbox for MATLAB. *Computing Science and Statistics* 33: 2001.
38. Kevin Patrick Murphy. 2002. Dynamic bayesian networks: representation, inference and learning.
39. Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker. 2007. *Réseaux bayésiens*. Eyrolles, Paris.
40. Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions Using Speech and Head Motion Information. In *Proceedings of the 16th International Conference on Multimodal Interaction* (ICMI '14), 136–143.
41. H. Pashler. 1994. Dual-task interference in simple tasks: data and theory. *Psychological Bulletin* 116, 2: 220–244.
42. Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. 2015. Presentation Trainer, Your Public Speaking Multimodal Coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (ICMI '15), 539–546.
43. Richard M. Shiffrin and Gerald T. Gardner. 1972. Visual processing capacity and attentional control. *Journal of Experimental Psychology* 93, 1: 72–82.
44. Han Sloetjes and Peter Wittenburg. 2008. Annotation by Category: ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.
45. David L. Strayer, Frank A. Drews, and Dennis J. Crouch. 2006. A comparison of the cell phone driver and the drunk driver. *Human Factors* 48, 2: 381–391.
46. M. Iftekhar Tanveer, Emy Lin, and Mohammed (Ehsan) Hoque. 2015. Rhema: A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 286–295.
47. Toastmasters International. 2011. Gestures: your body speaks. Retrieved from https://www.toastmasters.org/~/media/E202D7AA84E24A758D1BAAE8A77FD496.ashx
48. A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. 2012. Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing* 3, 1: 69–87.
49. Lilyan Wilder. 1999. *7 Steps to Fearless Speaking*. John Wiley & Sons, New York.