Understanding Public Speakers' Performance: First Contributions to Support a Computational Approach

Fábio Barros[1], Ângelo Conde[2,4], Sandra C. Soares[2,3], António J. R. Neves[1], and Samuel Silva[1]

[1] Institute of Electronics and Informatics Engineering of Aveiro,
Department of Electronics, Telecomunications and Informatics,
University of Aveiro, Aveiro, Portugal
[2] Department of Education and Psychology, University of Aveiro, Aveiro, Portugal
[3] William James Center for Research, University of Aveiro, Aveiro, Portugal
[4] CIDTFF — Research Center in Didactics and Technology in Training of Trainers,
University of Aveiro, Aveiro, Portugal
{fabiodaniel, aconde, sandra.soares, an, sss}@ua.pt

Abstract. Communication is part of our everyday life and our ability to communicate can have a significant role in a variety of contexts in our personal, academic, and professional lives. For long, the characterization of what is a good communicator has been subject to research and debate by several areas, particularly in Education, with a focus on improving the performance of teachers. In this context, the literature suggests that the ability to communicate is not only defined by the verbal component, but also by a plethora of non-verbal contributions providing redundant or complementary information, and, sometimes, being the message itself. However, even though we can recognize a good or bad communicator, objectively, little is known about what aspects – and to what extent -– define the quality of a presentation. The goal of this work is to create the grounds to support the study of the defining characteristics of a good communicator in a more systematic and objective form. To this end, we conceptualize and provide a first prototype for a computational approach to characterize the different elements that are involved in communication, from audiovisual data, illustrating the outcomes and applicability of the proposed methods on a video database of public speakers.

Keywords: Verbal and non-verbal communication · Computational methods · Posture · Facial Expressions · Voice

1 Introduction

Communication is inherent to human life, only through communication humans can interact with each other and exchange ideas and experiences. Nowadays, the capability to communicate well in public is a very important competence at the professional, academic, and personal levels.

Literature suggests that the ability to communicate is not only defined by the verbal component, but also by a set of non-verbal components since many non-verbal aspects provide redundant and/or complementary information. The literature also suggests that through the body, gestures, facial expressions and voice variations, the audience can identify several social characteristics such as competence, dominance, confidence, and others [13].

The subject of communication skills, such as in public speaking contexts, has been a topic widely studied in various areas, notably, in Education, for its relevance as a core element for the dissemination of information and knowledge. However, the advances made, through the years, about what explains good or bad communication skills has yet to reach a wider audience that could profit from it, mostly due to a arduous translation from theory to practise. One of the reasons is that researchers are often faced with a difficulty in objectively testing their hypotheses regarding the driving factors for good or bad communication due to a lack of a more quantitative setting for their research. This is where computational approaches may help. To achieve this, and as part of a long-term effort to advance the research on communication skills, the main goal of the work presented here is to bring forward a framework supporting increased objectivity in the study of communication in public, specifically by:

- Selecting, based on the literature, which aspects (channels) are most relevant in human-human communication, particularly those that have been described as having a notable impact on public communication;
- Proposing a set of computational methods describing the actions and contents present in the different communication channels identified.
- Complementing and annotating, through the proposed methods, an existing audiovisual database focusing on the study of speakers' performance.

The remainder of this paper is organized as follows: Section 2 presents a brief overview of research focusing public communication, and summarizes a wealth of notable datasets and libraries deemed relevant for the study and extraction of features from different communication channels. In Section 3 the methods proposed for the computation of different features from the selected channels are described along with a summary of the resulting data and information . Then, in Section 4, the methods are illustrated by applying them to an existing dataset of public speaker videos, a context for which communication skills assume a paramount role. In Section 5, we present a conclusion regarding the work developed and a discussion about the possible paths to advance and take advantage of the proposed approach.

2 Related Work

Human-human communication is not just about words. Although verbal communication is the main form of communication between humans, nonverbal behaviour (e.g., facial expressions, gestures, and body posture) has a very important role in communication. The information conveyed through these different channels may serve a wide variety of purposes, whether to explicitly reinforce or complement the message, as the message itself, or, inadvertently, as a barrier to its correct perception.

According to the literature, human beings seem to use expansive and open postures (becoming bigger and taking up more space) to project signs of power, confidence and assertiveness [12, 10]. On the other hand, counteractive and closed postures (minimizing occupied space and shrinking body) project signs of powerlessness and low confidence. The literature also states that during communication, humans use broad gestures and expansive body postures to project dominance.

Also, human beings usually produce gestures while talking. According to the literature (e.g., [15]), such gestures are actions that are directly related to the lexical and semantic content and are particularly suited to reinforce or complement the message being conveyed. The literature also states that gestures are a crucial element for speakers, as they help to expose ideas and retrieve content that is difficult to memorize [30].

On the other hand, facial expressions can transmit countless emotions without saying a single word, and unlike some forms of nonverbal communication, they are universal. Thus, facial expressions are one of the most important aspects in human communication since they can convey the speaker's emotional state, but also intentions, through facial muscle movements, such as wrinkling eyebrows or lifting lip corners. However, facial expressions are not the only ones with an important role in nonverbal facial communication. The posture of the head and the direction of the gaze are equally important indicators of communicative intention, since they influence the level of perceived naturalness and competence [24, 14].

In the contrast between verbal and non-verbal communication, literature pertaining to the verbal side states that the human being not only infers the meaning conveyed but also the way in which this is done (e.g. [19]). In this sense, prosodic clues are an integral part of human communication. Also, the literature states that there are a number of features present in audio resources that have been widely considered to understand how voices are heard and interpreted. These characteristics include, for example, volume and its variations, duration of speech, duration of pauses, consideration of a restricted lexical field (use of a group of restricted words), among others [11].

Attention to public speaking has risen, in recent years, as a valued personal communication skill and computational technology has echoed such attraction. The research community has been creating several datasets to provide the grounds to support a more systematic study of public speaking performance such as, for example, [7], [26], [10] and [5]. The present work is supported on a new dataset of 36 (18F+18M) public speaker videos. This dataset aims to establish the best social inference predictors of communicative performance. The dataset is based on thin-slice videos (30s to 50 seconds) of Portuguese TEDx orators, which were viewed by 97 participants asked to evaluate, for each video, using a visual analog scale, the emotion elicited by the presentations, first impressions (e.g., competence, warmness, confidence), and the perceived importance of several nonverbal features (e.g., gestures) on such judgments.

Considering the available datasets, one of the challenges pertains how to take the most out of the resources they provide, particularly in a way that does not loose complete sight of the theory gathered through the years to enable establishing parallels and test hypotheses. One of the most challenging aspects motivating this work is precisely the lack of a more objective knowledge about which aspects of the speaker's performance are influencing how the message is perceived and grasped. The assessments about who is a good or bad communicator are often based on the expertise of communication specialists and the subjective nature of their feedback, due to a lack of a more quantitative framework to support them, makes it difficult to understand what aspects – and their level of importance – influence the audience. Although there are several characteristics in both verbal and non-verbal communication that have been pointed out as relevant, a more systematic and quantitative approach is necessary in the improvement of studying the communication process.

In this regard, it is also important to note that the consideration of machine learning methods to build a system capable of discriminating between a good and a bad communicator is a natural goal, in the long-run, for this research. However, at this stage, we are mostly interested in adopting a framework that enables an exploratory analysis of how well we can compute features to express concepts as expansiveness, in tight collaboration with communication experts. By doing this, we hope to first create the basis for a greater multidisciplinary insight over the communication process and build a set of meaningful features that will contribute to the explicability of future machine learning approaches.

Considering the different channels that can be used to communicate, verbal and non-verbal, it is necessary to gather features allowing a computational description of their contents supporting further computational approaches to focus on the study of communication. In this regard, several technologies and software libraries can be considered.

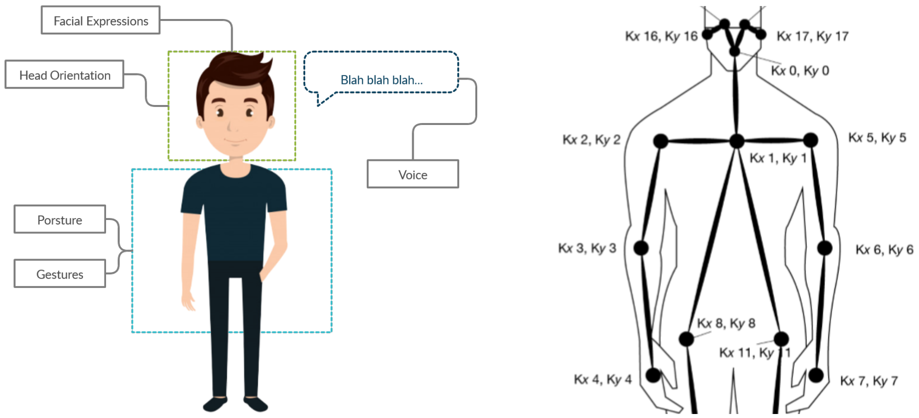
Observations performed during the communication process allow the extraction of the necessary data for the analysis of body posture, movements and gestures, via some alternatives such as Kinect Skeletal Tracking[29], ArtTrack[16] and DeeperCut[17]. However, the OpenPose[3] library was presented in April 2017, which revolutionized the field of visual computing. Taking video frames as *input* it allows the detection and collection of values at two dimensions of the main parts of the human body in a total of 130 *keypoints*, 15 or 18 for the body, 21 for each hand, and 70 for the face.

For the extraction of facial data a few libraries have been considered, such as Menpo [21], LEAR [20], and OpenFace 2.0 [1]. From these, OpenFace 2.0, an open source library, provides the wider range of features. It detects face landmarks, head position, gaze direction, and enables recognizing the activation and intensity of several key elements of the face (Action Units) enabling a more detailed study of facial activity.

The tools OpenEAR [8], SPAC [22], and Praat [2] are some examples of what can be considered for processing and analysis of the audible component of communication. As an alternative to these, OpenSMILE (The Munich open-Source Media Interpretation by Large feature-space Extraction) [9] is strongly used by the community of researchers in the areas of voice and emotion recognition, and MIR (Music Information Retrieval). It is a flexible and modular library for signal processing and machine learning applications. Regarding voice related resources, OpenSMILE, allows the extraction of Mel Frequency Cepstral Coefficients MFCCs, Pitch, Jitter, Energy, Intensity, Zero crossing rate and others.

3 Computation Approach to Study Communication Skills

In order to better understand the phenomenon of communication at the different levels of the communication channels, a set of methods was developed that allow a computationally-based description of what happens in each of these channels (see Figure 2. In this regard, the literature on the assessment and discussion of comunnication skills provided clues regarding which aspects could be considered. Two processing stages were considered: the first, focused on obtaining low-level characteristics e.g., wrist position, over time, describing what happens in each of the channels; the second carried out the transformation of these characteristics into high-level annotations (e.g., from hand movement coordinates into "rising hand") with the aim of allowing greater readability and better identification of relevant activities, so that they may encompass the different levels of study and the multidisciplinary nature of the research team.



(a) (b)

Fig. 2: On the left, identification of the different sources of communicative content (channels) considered for this work. On the right, identification of the skeleton key points considered to infer different measures describing posture, gestures, and head orientation.

3.1 Features per communication channel

According to Koppensteiner et al. [18], horizontal and vertical body movements affect the formation of impressions in a different way. Similarly, and according to Carney et al. [4], expansive and open postures project signs of power, dominance, confidence and assertiveness, and on the other hand, counteractive and closed postures project signs of impotence and low self-esteem. In this way and, using OpenPose, the key points of the skeleton were extracted. These were then used to compute several features characterizing communicator performance.

Human's face is another indicator of communicative intention. Facial muscle movements (Action Units), such as wrinkling eyebrows or lifting lip corners and head postures, are very important aspects in communication, since they influence the level of perceived naturalness and competence. In order to that, we extracted values related to head posture (pitch and yaw), , and the activation and intensity of Action Units using Openface.

Last but not least, variation of intensity, duration of speech, duration of pauses and pitch have been investigated in order to assess the voice. However, prosodic characteristics also become one of the pillars of the recognition of paralinguistic traits [6]. In this sense, with the use of Opensmile a set of features deemed relevant, based on the literature, e.g., [25], such as speech intensity, energy, pitch, jitter, loudness, and MFCCs and voice activity detection (VAD) [23], were extracted.

3.2 High level features and annotations

Once the low-level descriptors associated with each of the identified communication channels were extracted, we transform these data into high-level annotations in order to give them better readability and an easier identification of relevant activities. In this sense, the high level annotations that we provide concern: posture, head movements, emotion (as expressed by the face), horizontal and vertical gestures of the harms and hands, and, also, related to the voice, moments of silence/voicing and variation of the audio intensity.

Hand gestures and head position — we use the same approach based on assessing the variation of the relevant keypoints in each sequence of ten frames. Table 1 summarize all the annotations considered and indicates which values are used to set each annotation.

Horizontal and vertical expansiveness — The amplitude of the horizontal movements is obtained by adding the distance between the neck (k_x1) and the left wrist (k_x7) and right k_x4. $H(x) = |k_x1 - k_x4| + |k_x1 - k_x7|$ (1).

The amplitude of the vertical movements is obtained by adding the distance between the neck (k_y1) and the left wrist (k_y7) and right k_y4. $H(y) = (k_x4| - k_y1) + (k_x7 - k_x1)$ (2).

Occupied area — To calculate the area occupied in each frame the values of K_xmax, K_xmin, K_ymax and K_ymin from the human body are obtained. For these values, we check which key points represent the minimum and maximum

Table 1: On the left, criteria used to perform high level annotation of gestures and head position; on the right, the action units considered to infer the emotion expressed by the speaker.

Annotation	**Criteria** (10 frames)
Head Moving Right	ΔYaw > 0
Head Moving Left	ΔYaw < 0
Head Moving Up	ΔPitch < 0
Head Moving Down	ΔPitch > 0
Approach Hands	ΔWrists < 0
Separate Hands	ΔWrists > 0
Arm Going Down	$\Delta Y_{wrist} > 0$
Arm Going Up	$\Delta Y_{wrist} < 0$

Emotion	**Action Units**		
	Weight		
	50%	25%	12.5%
Fear	AU 1	AU 5	AU 25
Happy	AU 12	AU 6	AU 25
Sad	AU 1	AU 4	AU 17
Angry	AU 25	AU 4	AU 9
Surprised	AU 26	AU 17	AU 2
Disgust	AU 9	AU 7	AU 4

values for axes (X) and (Y) in each frame. $A = |K_x max - K_x min| \times |K_y max - K_y min|$ (3)

Figure 3 illustrates the annotation of "Head Moving Right" in a sequence of frames, since the head's yaw increases over the interval.

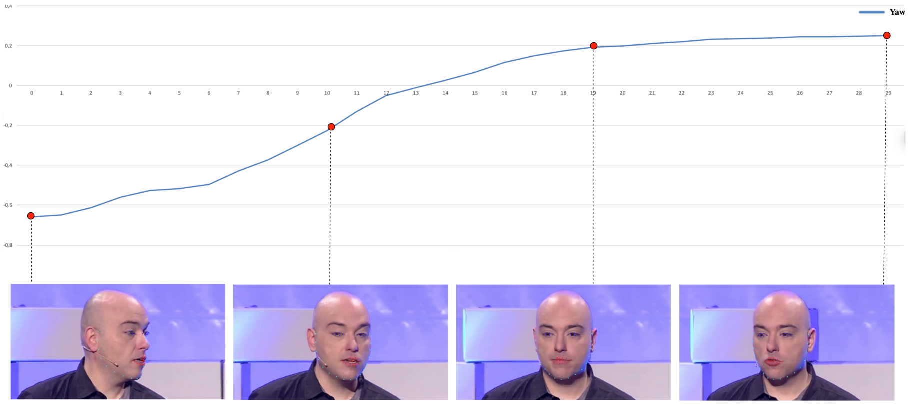


Fig. 3: Characterization of head movement: illustration of yaw variation and key video frames associated with head movement from left to right.

Emotion from facial expression — in each moment of presentation we used a simple approach based on action units intensities. Firstly, based on [28, 27], we selected the three most frequent action units present in each emotion: fear, happy, sad, angry, surprise and disgust, as depicted to the right of Table 1, ordered from left to right according to their predominance.

Then, for each emotion is attributed a numerical value based on the intensity of the action unit and weight of predominance. Finally, we select the emotion with the highest value, but, if the value is lower than a threshold devised empirically, during development, we attribute the neutral emotion. Figure 4 illustrates how the emotional state of “Happy” is attributed to a speaker.

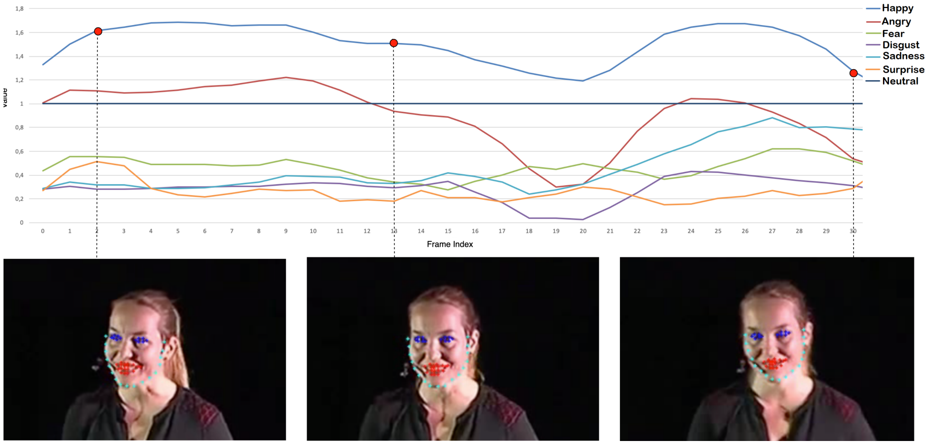


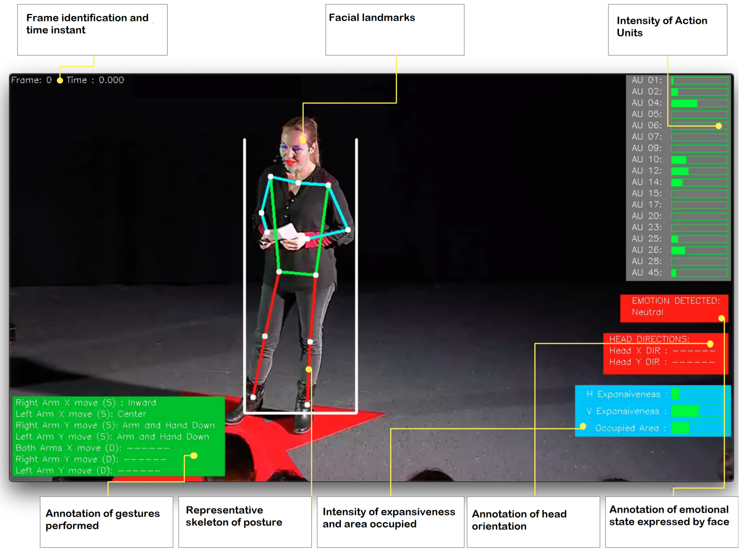
Fig. 4: Determination of emotion from facial expressions: the intensity for each emotion is determined, over time, based on AU activation, and the most intense is chosen.

Voice activity and intensity — Based on the VAD probability values, we annotate moments of silence and speech (VAD > 0). Additionally, we also annotate voice intensity variations by using both the VAD and Voice intensity values extracted. For that purpose, we calculate the average of the intensity excluding the intervals marked as silent. Then, we consider that there is an increase in voice intensity at all moments for which the intensity is above average.

All the computed low-level descriptors along with the determined annotations and activities are stored in JSON file to facilitate, e.g., exporting the data into other tools implementing the computation of different high-level features.

4 Results

The methods presented in this paper include a large set of data and information considered relevant for the characterisation of the different communication channels. In this sense, it was considered important to propose a visualization tool that would also support the analysis of the contributions of this work by researchers from other areas, e.g., Education and Psychology, to favour exploratory analysis of the dataset and harness expert insights in finding synergies between the computed measures and observed speaker behaviour. So, we propose a tool that performs the overlay of the computed data and information sets, on video,



Frame identification and time instant

Facial landmarks

Intensity of Action Units

Annotation of gestures performed

Representative skeleton of posture

Intensity of expansiveness and area occupied

Annotation of head orientation

Annotation of emotional state expressed by face

Fig. 5: Illustration of the software tool created to enable a visualization and analysis of the extracted information along the videos.

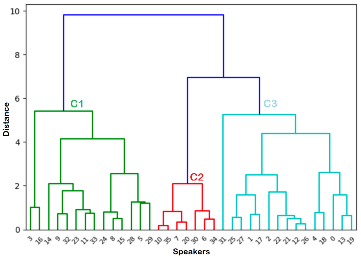
in real time. Figure 5 illustrates the visualization tool developed, where the computed data and annotations (e.g., intensity of the *Action Units* and gestures) are overlayed on the video stream.

This paper proposes an approach to obtain a set of data describing the activity in each communication channel, from digital audio and video, providing a first level of quantitative support for an exploratory study about understanding the communication with the adoption of computational methods.

4.1 Illustrative Application Example

To illustrate how the outcomes of the work presented in this article may support shedding some light over the study and impact of communication skills, we present an example of using the outcomes of the proposed methods to explore how the "Occupied Area" annotation might relate with the perception of good or bad communicators..

According to the literature, and as previously mentioned, more expansive and open people project signs of confidence and assertiveness. Thus, using the data computed for the occupied area, described by its (mean, standard deviation, maximum and minimum), unsupervised machine learning methods, concretely Agglomerative Clustering, were applied in order to distribute the speakers of the video dataset considered in this work. Thus, we obtained the distribution of the speakers, relative to the occupied area, during their presentation, in three groups, as illustrated in Figure 6.



Clusters	Area	Conf.	Assert.
C1	27%	61	57
C2	30%	51	50
C3	37%	49	48

Fig. 6: Illustrative example considering the computed occupied area for clustering speakers and its relation with first impressions annotated by participants: on the left, resulting dendrogram, considering Ward's distance; on the right, average scores for computed occupied area, and corresponding annotated confidence and assertiveness.

Then, the average human provided annotations for confidence and assertiveness (part of the dataset) were computed, for each of the groups, as shown in Fig. 6.

With this basic approach and by relating the clustering with the human annotations available on data set, we have some evidence that audiences do not seem to interpret confidence and assertiveness solely based on expansive/constrained (larger/smaller occupied area) postures, hinting on the paramount importance of an understanding of the synergies among the multiple verbal and nonverbal components of communication. The work presented here is, in our opinion, a relevant first step towards that goal.

5 Conclusions

Given that public communication is an area that is still very little explored, this work has managed to make a positive contribution to its progress. Based on the literature, and since it is noticeable that communication is multimodal, i.e., a mix of the contents of several channels, it was possible to select a set of characteristics present in verbal and non-verbal communication that are deemed to have a relevant impact on human communication, in public.

Through the use of different computational tools, we successfully extract a set of elements for the characterization of multiple nonverbal features playing a role in human-human communication involving the body, face, and voice. Through these, it was also possible to implement some methods to annotate activities considered relevant that facilitate the description and interpretation of a set of actions/contents occurring during communication (e.g., raising the harms). While the proposed methods have been illustrated over a particular audiovisual dataset, they are applicable to any other videos of communicative tasks, although limited to a single speaker. In the future, our goal is to generalize our approach to encompass multiple speakers at the same time, such as in a conversation.

Interestingly, the simple application example presented, albeit preliminary, hints that some of the theoretical aspects for communication skills assessment need to be further explored, since a direct relationship between expansiveness, as represented by the occupied area, and the perception of assertive/confident speakers, as annotated by a human audience, does not stand out.

6 Acknowledgements

This work is partially funded by IEETA Research Unit funding (UIDB/00127-/2020), by Portugal 2020 under the Competitiveness and Internationalization Operational Program, and the European Regional Development Fund through project SOCA – Smart Open Campus (CENTRO-01-0145-FEDER-000010) and project MEMNON (POCI-01-0145-FEDER-028976).

References

1. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: Proc. 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018). pp. 59–66. IEEE (2018)
2. Boersma, P.: Praat, a system for doing phonetics by computer. Glot. Int. **5**(9), 341–345 (2001)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
4. Carney, D.R., Cuddy, A.J., Yap, A.J.: Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. Psychological science **21**(10), 1363–1368 (2010)
5. Chen, L., Feng, G., Leong, C.W., Joe, J., Kitchen, C., Lee, C.M.: Designing an automated assessment of public speaking skills using multimodal cues. Journal of Learning Analytics **3**(2), 261–281 (2016)
6. Cullen, A., Hines, A., Harte, N.: Perception and prediction of speaker appeal–a single speaker study. Computer Speech & Language **52**, 23–40 (2018)
7. Echeverría, V., Avendaño, A., Chiluiza, K., Vásquez, A., Ochoa, X.: Presentation skills estimation based on video and kinect data analysis. In: Proc. of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge. pp. 53–60 (2014)
8. Eyben, F., Wöllmer, M., Schuller, B.: OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In: Proc. 3rd Int. Conf. on affective computing and intelligent interaction and workshops. pp. 1–6. IEEE (2009)
9. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proc. 18th ACM Int. Conf. on Multimedia. pp. 1459–1462 (2010)
10. Gan, T., Wong, Y., Mandal, B., Chandrasekhar, V., Kankanhalli, M.S.: Multisensor self-quantification of presentations. In: Proc. 23rd ACM Int. Conf on Multimedia. pp. 601–610 (2015)
11. Giannakopoulos, T.: pyaudioanalysis: An open-source python library for audio signal analysis. PloS one **10**(12) (2015)
12. Gronau, Q.F., Van Erp, S., Heck, D.W., Cesario, J., Jonas, K.J., Wagenmakers, E.J.: A bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. Comprehensive Results in Social Psychology **2**(1), 123–138 (2017)

13. Hall, J.A., Knapp, M.L.: Welcome to the handbook of nonverbal communication. Nonverbal Communication. Berlin: De Gruyter Mouton pp. 3–10 (2013)
14. Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., Özyürek, A.: Social eye gaze modulates processing of speech and co-speech gesture. Cognition **133**(3), 692–697 (2014)
15. Iani, F., Bucciarelli, M.: Mechanisms underlying the beneficial effect of a speaker's gestures on the listener. Journal of Memory and Language **96**, 110–121 (2017)
16. Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B.: Arttrack: Articulated multi-person tracking in the wild. In: Proc. of the IEEE conf. on computer vision and pattern recognition. pp. 6457–6465 (2017)
17. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. pp. 34–50. Springer (2016)
18. Koppensteiner, M., Stephan, P., Jäschke, J.P.M.: Moving speeches: Dominance, trustworthiness and competence in body motion. Personality and Individual Differences **94**, 101–106 (2016)
19. Kreitewolf, J., Friederici, A.D., von Kriegstein, K.: Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. Neuroimage **102**, 332–344 (2014)
20. Martinez, B., Valstar, M.F., Binefa, X., Pantic, M.: Local evidence aggregation for regression-based facial point detection. IEEE transactions on pattern analysis and machine intelligence **35**(5), 1149–1163 (2012)
21. Alabort-i Medina, J., Antonakos, E., Booth, J., Snape, P., Zafeiriou, S.: Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In: Proc. 22nd ACM Int. Conf. on Multimedia. pp. 679–682 (2014)
22. Özseven, T., Düğenci, M.: Speech acoustic (spac): A novel tool for speech feature extraction and classification. Applied Acoustics **136**, 1–8 (2018)
23. Park, T.J., Chang, J.H.: Dempster-shafer theory for enhanced statistical model-based voice activity detection. Computer Speech & Language **47**, 47–58 (2018)
24. Sadoughi, N., Liu, Y., Busso, C.: Meaningful head movements driven by emotional synthetic speech. Speech Communication **95**, 87–99 (2017)
25. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.S.: The interspeech 2010 paralinguistic challenge. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
26. Tanveer, M.I., Zhao, R., Chen, K., Tiet, Z., Hoque, M.E.: Automanner: An automated interface for making public speakers aware of their mannerisms. In: Proc. 21st Int. Conf. on Intelligent User Interfaces. pp. 385–396 (2016)
27. Velusamy, S., Kannan, H., Anand, B., Sharma, A., Navathe, B.: A method to infer emotions from facial action units. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 2028–2031. IEEE (2011)
28. Vick, S.J., Waller, B.M., Parr, L.A., Pasqualini, M.C.S., Bard, K.A.: A cross-species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (FACS). J. of nonverbal behavior **31**(1), 1–20 (2007)
29. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia **19**(2), 4–10 (2012)
30. Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., Tian, X.: Manual directional gestures facilitate cross-modal perceptual learning. Cognition **187**, 178–187 (2019)