

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.  
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) 27-08-2024	2. REPORT TYPE Book Chapter	3. DATES COVERED (From - To) -
---	--------------------------------	-----------------------------------

4. TITLE AND SUBTITLE Speech based analysis of group interactions	5a. CONTRACT NUMBER W911NF-20-1-0214
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 611102

6. AUTHORS Evangelia Fringi, Susannah Paletz, Alessandro Vinciarelli	5d. PROJECT NUMBER 611104
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES University of Maryland - College Park The University of Maryland Office of Research Administration College Park, MD 20742 -5141	8. PERFORMING ORGANIZATION REPORT NUMBER
--	--

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	10. SPONSOR/MONITOR'S ACRONYM(S) ARO
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) 76922-HC-H.14

12. DISTRIBUTION AVAILABILITY STATEMENT
---

13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.
---

14. ABSTRACT
--------------

15. SUBJECT TERMS
-------------------

16. SECURITY CLASSIFICATION OF:	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Susannah Paletz
a. REPORT	b. ABSTRACT	c. THIS PAGE	19b. TELEPHONE NUMBER 301-405-1218

---

**REPORT DOCUMENTATION PAGE (SF298)**  
**(Continuation Sheet)**

---

**Continuation for Block 13**

Proposal/Report Number: 76922.14-HC-H

Report Title: Speech based analysis of group interactions

Report Type: Book Chapter

**Publication Type:** Book Chapter                      Peer Reviewed: Y      **Publication Status:** 4-Under Review

**Chapter Title:** Speech based analysis of group interactions

Publication Identifier Type:

Publication Identifier:

Volume:

Edition:

1st Page#:

Date Received: 27-Aug-2024

Publication Year: 2026

Publisher: Oxford University Press

Publication Location:

**Book Title:** Computational group and team dynamics: Forging an interdisciplinary science

**Authors:** Evangelia Fringi, Susannah Paletz, Alessandro Vinciarelli

**Editor:** Stephen Kozlowski, Hayley Hung, Nale Lehmann-Willenbrock, Albert Ali Salah

Acknowledged Federal Support: Y

Speech Based Analysis of Group Interactions

Evangelia Fringi<sup>1</sup>, Susannah B.F. Paletz<sup>2</sup>, Alessandro Vinciarelli<sup>1</sup>

<sup>1</sup>University of Glasgow

<sup>2</sup>University of Maryland

## Abstract

Human communication includes nonverbal elements of speech, such as intonation, pitch, loudness, turn-taking and speaking rate. While there is still a gap between analysis conducted with the methodologies of social sciences and automatic analysis approaches, these latter hold a great deal of promise in drawing connections between a physical layer (the measurable characteristics of speech signals) and an inferential layer (social and psychological information that the speech signals are expected to convey). This chapter describes the technology for speech-based analysis such as data collection, speaker diarization, paralinguistic analysis, and computational analysis of constructs. Challenges include obtaining appropriate datasets and ensuring construct validity. We cover the state-of-the-art of current analysis, touching on specific efforts such as the detection of valence, roles, personality, and group cohesion. Future work in this area should attend to limitations, such as of measurement and ecological validity, and the ethics of using artificial intelligence techniques to characterize social and psychological phenomena.

*Keywords:* speech analysis, nonverbal communication, valence, cohesion, AI

## Speech Based Analysis of Group Interactions

**Introduction**

The IEEE Signal Processing Society<sup>1</sup>, one of the most important associations of computing professionals, was founded in 1948 with some of the earliest work revolving around automatic analysis and processing of speech signals (see e.g., Davis, Biddulph, and Balashek (1952) for one of the earliest digit recognition systems). The initial focus was on Automatic Speech Recognition (Pieraccini, 2012), the task of automatically transcribing audio recordings in which one or more speakers talk. Since the early 2000s, researchers have increasingly made use of speech as a source of information about social and psychological phenomena involving individuals or groups of interacting people. Initially, the efforts focused on the recognition of emotions or inner states of individual speakers (see, e.g., Scherer (2003); Wani, Gunawan, Qadri, Kartiwi, and Ambikairajah (2021) for extensive surveys on speech-based emotion recognition). However, it did not take long before recordings of multiparty interactions were analyzed, especially in the context of meetings (McCowan et al., 2005) or broadcast material (Vinciarelli, 2007). This progress gave scientists the opportunity to analyze a broad spectrum of social phenomena taking place in interactions, including dominance (Jayagopi, Hung, Yeo, & Gatica-Perez, 2009; Worgan & Moore, 2011), roles (Zancanaro, Lepri, & Pianesi, 2006), turn-taking (Vinciarelli, 2009), etc.

From a purely technological point of view, one of the probable motivations behind the major interest on speech is that reliable technologies for acquisition and storage of speech signals became available in conjunction with the development of modern computers. Microphones, the sensors needed to capture speech and input it into machines, were available since the end of the 19th century<sup>2</sup>, and they were of sufficient quality when speech

---

<sup>1</sup> <https://signalprocessingsociety.org>

<sup>2</sup> Patent No. 372,786, dated November 8, 1887

processing research started. Speech storage has been possible for more than 150 years (the phonograph was invented in the late 1800s<sup>3</sup>) and high quality recordings (tapes first and then digital memories) were available during the time speech processing technologies were developed. Even if such technologies did not contribute directly to speech processing, they still provided the conditions for tackling it.

In addition to the above, compared to other sensors (e.g., cameras), microphones have been more robust to reality and more comfortable for the people being recorded. For instance, data collected with cameras (films and photographs first and then digital images and movies) can be very sensitive to natural changes in the environment such as lighting and differences in the relative position between a camera and its subject. However, approaches dealing with speech recordings became robust to reasonably realistic conditions at a relatively early stage. In turn, people developed familiarity with audio technologies relatively early through the use of phone and radio, two technologies available since the first half of the 20th century. These advances gave rise to the availability of large amounts of speech data and the possibility to develop technologies relying on speech input relatively early during the computer era (see, e.g., Averbuch et al. (1986); Chow et al. (1987)).

Still, technological factors cannot be the only reason behind the interest for speech processing in the computing community. Because of its role in human communication, speech cannot be considered a signal like any one else. One way to describe the importance of speech is to consider the answer of Helen Keller, a prominent disability right activist, when she was asked to explain the difference between deafness and blindness: "Blindness separates people from things; deafness separates people from people." Keller similarly highlighted speech as "the most vital stimulus - the sound of the voice that brings language, sets thoughts astir, and keeps us in the intellectual company of man." Keller's words suggest that speech can be considered as a privileged channel through which people establish social interactions. Indeed, the very physiology of human ears seems to be designed around

---

<sup>3</sup> Patent No. 227,679, dated May 18, 1880

speech. In fact, human voice is the sounds that requires the lowest energy to be heard, up to a million less times than other sounds in the environment (Fletcher, 1940).

Communication via speech is also the main mechanism by which humans coordinate, collaborate, and cooperate (Tannenbaum & Salas, 2020). Verbal and nonverbal communication is key for synchronous interactions and teamwork such as remote, hybrid, or face-to-face team meetings (Allen & Lehmann-Willenbrock, 2023; Kane & van Swol, 2022). Recording and measuring these interactions may reveal patterns too brief or subtle to be detectable to the speakers, and thus not easily remembered or measured via self-report (Gottman & Notarius, 2000; Paletz, Schunn, & Kim, 2011). For instance, specific patterns in conversations between married couples can predict divorce years later (Gottman & Levenson, 2000).

Speech thus provides an ideal ground for interdisciplinary collaboration between the researchers who look at speech from a technological point of view (engineers, computer scientists, mathematicians, etc.) and those who look at it from a human sciences perspective (sociologists, psychologists, linguists, etc.). In fact, while some speech processing problems can be addressed without necessarily involving multiple disciplines (e.g., voice recognition can be performed by analyzing spectral properties of speech, irrespective of any considerations about the social identity of the speaker), others require knowledge not only about machines, but also about people (e.g., the recognition of emotions needs to be informed with knowledge about the way speakers express their inner states). It is probably because of this need for multiple fields of expertise that the computing community recognizes now inherently interdisciplinary areas such as Computational Paralinguistics (Schuller & Batliner, 2013), Affective Computing (Picard, 2000), and Social Signal Processing (Vinciarelli, Pantic, & Bourlard, 2009).

Overall, speech includes two major components, namely language (*what* people say) and paralinguage (*how* they say it). The first component is the subject of another chapter and is addressed with methodologies that are not typical of speech processing. For this

reason, this chapter will focus on paralinguistic, often referred to as an aspect of *nonverbal communication* (Knapp & Hall, 1997; Richmond & McCroskey, 2000). While nonverbal communication can include facial expression, gestures, and body language, paralinguistic includes phenomena such as intonation and pitch, loudness, speaking rate, vocalizations (laughter, sobbing, grunts, crying, etc.), pauses and everything else in speech that is not words. Far from being a secondary component of spoken communication, paralinguistic conveys social and psychological information that is not necessarily available, or at least not as clearly, in the verbal component (Wharton, 2009). Not surprisingly, previous research shows that people speaking on the phone dedicate close to 25% of the time to four of the most common nonverbal behavioural cues, namely laughter, fillers (utterances such “ehm” and “uhm” that should correspond to a word), silences, and overlapping speech (Vinciarelli, Chatziioannou, & Esposito, 2015). This finding confirms that speech is more than words; the goal of this chapter is to explore what this “more” is and the importance of its measurement.

### **Technology of Speech-Based Group Analysis**

In the context of speech-based group analysis, the goal of a speech processing approach can be thought of as bridging the gap between a *physical* layer (the measurable characteristics of speech signals) and an *inferential* layer (social and psychological information speech signals are expected to convey). The physical layer corresponds to machine detectable speech characteristics that account for what people hear, while the inferential layer corresponds to information that listeners cannot hear directly, but can infer from what they hear. Emotion recognition is a typical example. People tend to speak differently (meaning that they change the measurable properties of their speech) depending on the emotions they are trying to express, or at least failing to hide. Correspondingly, listeners tend to assume that people speaking in a different way are likely to experience different emotions. What people can hear are not the internally experienced emotions, but

the different ways of speaking corresponding to the expressions of different emotions (e.g., louder or softer, faster or slower, etc., some of which may be affected by cultural norms, see Elfenbein and Ambady (2002)). Similar considerations can be made for every social or psychological phenomenon that leaves traces in speech.

The key assumption underlying this description of the physical versus inferential layers is that the relationship between them is, for practical purposes, stable and consistent. The main evidence that such an assumption holds is that people can communicate. In fact, effective communication is only possible if most of the time, speakers express (e.g., emotions) what they mean, and in turn, if listeners interpret their speech according to the what the speakers are trying to express. Still, while being consistent, the relationship between the physical and inferential layers can be ambiguous (Larrouy-Maestri, Poeppel, & Pell, 2024). For instance, different people manifest the same emotion in different ways, the expression of emotions can be different in different cultures or settings, and different emotions can be expressed in the same way (e.g., crying can result from both sorrow and happiness; Barrett, Adolphs, Marsella, Martinez, and Pollak (2019)). Furthermore, the process of inferring information from speech is, in addition to being socially and culturally contextual, potentially subjective such that different listeners can attribute different emotions to the same speaker. These considerations made about emotions can be extended to any other social or psychological phenomenon that can be manifested through speech. For instance, tone, loudness, and emphasis, as well as the cultural norms surrounding them, can influence whether the same statement is viewed as a disagreement or not (Paletz & Schunn, 2018).

It is because of these ambiguities that the most common computational methodology for speech-based group analysis is Artificial Intelligence (AI). Computational scientists deploy AI to associate variance in speech data to variance in assessments of that data based on human annotation of the phenomenon under study. In other words, AI can associate variance in the speech signals to variance in social and psychological phenomena

that leave their traces in speech. Furthermore, AI can be designed to do it in presence of ambiguity and uncertainty, i.e., in the conditions observed in human-human communication as described previously. Figure 1 shows the main stages of a technological approach allowing AI to automatically make sense of group phenomena in speech recordings. The rest of this section describes every stage in detail and shows the main aspects from both technological and interdisciplinary points of view.

## Data Collection

There is no automatic approach without data: It is only from data that AI can make the relationship between the physical and inferential layers. The datasets for these efforts must be large because current AI models require increasingly larger amounts of data to work effectively. Furthermore the data must be *annotated*, meaning that it must include not only speech signals, but also information about the inferential layer. Such information can only be obtained by human judges listening to the speech data and providing ratings or assessments about the social or psychological information of interest. For example, in the case of emotion recognition, human experts have to listen to the speech data (e.g., conversations) and provide ratings about the specific emotions they observe. Alternatively, the speakers being recorded have to provide, while speaking or later, information about their inner states and/or their intended expressions.

Both data collection and data annotation are expensive and time consuming processes. However, they both provide major opportunities for interdisciplinary collaboration between technology and human sciences. In fact, while the former can develop appropriate sensing apparatuses (e.g., ensuring the use of proper microphones and conditions suitable for the technical success of data collection), the latter can provide theory, procedures, and measures. For instance, if the goal is to recognize emotions, the speakers must be put in condition to experience emotions. Furthermore, the social scientists can ensure that the data are ecologically valid and that social or psychological

phenomena are assessed correctly. For example, one research paradigm used by Gottman and colleagues to study married partners and predict divorce involved having the married partners discuss (in this order) daily events, a contentious issue, and a pleasant topic while the pair were video recorded and various physiological measurements were taken (Gottman & Levenson, 2000). The audiovideo data were then annotated using different schemes as the theory and understanding of these kinds of interactions developed, including a Gestalt approach of linguistic, paralinguistic, contextual, and other nonverbal information (Gottman & Levenson, 2000).

The sensing apparatus determines what can be accessed and what it cannot in the data. For example, it is not possible to analyze facial expressions if no camera is included in the sensing approach. In the case of speech-based group analysis, the most important choices concern the type of microphones to be used (e.g., lapel vs standard) and their positions with respect to the speakers. The most interesting aspect in the definition of the sensing approach is that microphones should be in condition to capture the variance in the physical layer that is expected to reflect variance in the inferential layer (see above). For example, if the goal of an experiment is to study turn-taking dynamics, it is probably better to use lapel microphones or microphone arrays because these can help to better identify turns, i.e., the time intervals during which only one person speaks.

Because the sensing apparatus constraints what can be analyzed, researchers should take into account both the theory of what they wish to study and practical logistics. Regarding the first, if the social constructs being studied inherently involves one type of behavior, sensor, or channel, it is necessary to try to acquire that data. For instance, a study of team participation equality in team meetings requires speech data of sufficient quality such that it is possible to detect who is speaking when. Regarding the second, some situations limit what can be detected. In capturing speech in natural settings, for instance, stakeholders and the environment itself may dictate what kinds of data can be captured. A hypothetical study of conversations during a cat adoption event at a music store may incur

ambient noise, and attendees may resist being hooked up to microphones. Similarly, in studying town hall meetings, government officials and questioners may be speaking into microphones, but side conversations by audience members may not be recorded.

## Speaker Diarization

A great deal of annotation requires the human judges, and thus any AI, to distinguish between speakers. Speaker diarization is the task of detecting *who speaks when*, i.e., to split audio recordings into non-overlapping segments during which it is expected that only one person speaks (see Figure 1). For instance, studies of meeting participation require knowing the percent of time each person speaks (Paletz & Schunn, 2011), but even studies of other team constructs derived from speech require diarization (Lehmann-Willenbrock & Chiu, 2018). Historically, this was a time-consuming manual process, where transcribers or other workers post-transcription would divide up a conversation as best they could hear according to different speaker voices (Kane & van Swol, 2022; Paletz et al., 2011). Technological advances can automate this process. The input of a diarization approach is typically a raw audio recording, while its output is a sequence of triples  $T = \{(t_k, \Delta t_k, l_k)\}$  - with  $k = 1, \dots, N$  - where  $t_k$  is the time at which a segment starts,  $\Delta t_k$  is the duration of the segment,  $l_k$  is a label that accounts for one of the speakers involved in the interaction and  $N$  is the total number of triples (see Figure 1). According to a recent survey (Park et al., 2022), there are two main aspects that characterize the latest diarization approaches, typically based on deep networks, namely whether the training is based on diarization-relevant loss functions and whether there is only one model or multiple models working jointly (e.g., one for the detection of speaking activity and the other for the actual diarization).

This step can be performed irrespective of the end goal a speech-based approach. In other words, the diarization can be performed in the same way whether the goal is to analyze group phenomena or to perform any other task (e.g., Automatic Speech

Recognition) that requires to distinguish between one speaker and the other.

### Paralanguage Analysis

Also valuable to social scientists studying different aspects of communication is the capability to process these separate turns automatically. The sequence  $T = \{(t_k, \Delta t_k, l_k)\}$  (see the section about diarization) allows one to process speech signal segments expected to include only one speaker. This technique makes it possible to obtain, for every speaker, an automatic transcription and to perform language analysis while extracting paralinguistic information that could convey social and psychological layers of meaning (Wharton, 2009). Language and paralanguage can possibly be analyzed jointly through a multimodal approach (see Baltrušaitis, Ahuja, and Morency (2018) for an extensive survey on multimodal learning). However, language and multimodal analysis are the subject of other chapters in this book and, therefore, this section will focus on paralanguage.

The extraction of paralinguistic information can take two main forms, the first is typically referred to as *Computational Paralinguistics* (Schuller & Batliner, 2013). It includes two main stages, namely the segmentation of speech signals into 20-30 ms long analysis windows that start at regular time steps of 10 – 20 ms and the extraction of *Low Level Descriptors* (LLD) from the analysis windows. The length of the windows roughly corresponds to the amount of time a person can keep stable the configuration of speech articulators (the organs allowing one to speak). This process makes it possible to assume that the part of signal enclosed within a window is *stationary*, i.e., it is the result of a stable generative process.

LLDs corresponds correspond to physical measurements expected to account for speech properties that a listener can hear. The most common measurements are Energy (also loudness or intensity Lubold and Pon-Barry (2014)), Pitch (the height of a tone), Mel Frequency Cepstral Coefficients, Speaking Rate, Jitter, Shimmer, etc. (see Schuller and Batliner (2013) for more details). The values obtained for an individual window can be

averaged over multiple windows (possibly over an entire recording). More generally, the distribution of an LLD across multiple windows can be modelled with different *statistics*, including variance, skewness, maximum, minimum, position of maximum and minimum, median, range, etc.

The key-assumption underlying Computational Paralinguistics is that LLDs capture the way a person speaks, i.e., that component of speech that gives words a psychological colour, while allowing listeners to make inferences about inner state of one or more individuals. In many cases, the choice of LLDs is informed by social science findings about, on the one hand, how people manifest social and psychological information through the way they speak and, on the other hand, how people tend to interpret the way others speak in terms of social and psychological information. From this point of view, one particularly interesting effort was the development of the Geneva Minimalistic Acoustic Parameter Set or GeMAPS (Eyben et al., 2015), a set of LLDs that were found to be particularly effective for the inference of social and psychological phenomena from speech. In fact, such a set was jointly developed by computer scientists and psychologists so that every LLD was identified as accounting for a particular aspect of speech that people can hear (e.g., energy accounts for loudness). Another line of relevant research involves studying acoustic-prosodic entrainment, or people speaking similarly to each other. This research was inspired by Communication Accommodation Theory (see Giles, Coupland, and Coupland (1991); Giles, Taylor, and Bourhis (1973), and more recently Giles, Edwards, and Walther (2023)). The research on entrainment often seeks to map the psychological antecedents and social and task correlates of dyads and teams speaking similarly on features such as pitch, loudness, and more (Levitan (2020); Paletz, Litman, Karuzis, Jones, and Rahimi (2023); Yu, Litman, and Paletz (2019)). Entrainment research is also used to design AI or robot tutors that can better match learners (e.g., Lubold and Pon-Barry (2014); Lubold, Walker, and Pon-Barry (2021)).

In more recent times, paralanguage analysis tends to be performed using speech

representations based on deep learning, i.e., using embeddings like those extracted with *wav2vec* (Baevski, Zhou, Mohamed, & Auli, 2020). Unlike LLDs, embeddings are an abstract representation that cannot be interpreted in terms of speech characteristics that people hear. However, the results that are obtained are at least as good as those that were achieved with traditional LLDs. In this respect, embeddings can be thought as a technological solution that is capable of capturing the information people actually use to manifest or perceive social and psychological information. However, it does it in a way that does not easily allow an interpretation of the results and an interdisciplinary collaboration between technology and human sciences.

All approaches mentioned so far focus on low-level speech processing, i.e., on extraction of information that tends to represent the physical aspects of the signal rather than the way human listeners make sense of it. Another approach to paralinguistic analysis is to detect nonverbal behavioural cues that human listeners can identify consciously and that convey social or psychological information. Common examples include the detection of laughter (Cosentino, Sessa, & Takanishi, 2016), pauses (Marzinzik & Kollmeier, 2002), fillers (Stouten & Martens, 2003), and overlapping speech (Yousefi & Hansen, 2020). Each of these cues was shown to be a good marker for information of interest in the inferential layer. For example, laughter was shown to account for changes of topic in conversation (Bonin, Campbell, & Vogel, 2014) and unshared (solo) laughter was associated with open communication and team effectiveness in dyads (Wang, Doucet, Waller, Sanders, & Phillips, 2016); fillers were shown to account for hesitation (Verkhodanova, Shapranov, & Kipyatkova, 2017); silences were shown to mark the presence of depression (Liu, Kang, Feng, & Zhang, 2017); and overlapping speech was shown to indicate conflict and aggression (Lefter & Jonker, 2017). Paralinguistic analysis based on the detection of nonverbal cues has the major advantage of allowing interdisciplinary collaboration between computational and human sciences. The main reason for this success is that behaviour-oriented literature made significant efforts towards the analysis of the

relationships between nonverbal cues and social phenomena (see, e.g., Provine (2001) for laughter, Rochester (1973) for pauses, Clark and Tree (2002) for fillers and Schegloff (2000) for overlapping speech). In this respect, it is possible to say that technologies for paralinguistic analysis are informed by social science findings and, in turns, human sciences can benefit from approaches for the detection of nonverbal cues, especially when it comes to the analysis of large amounts of data (Lehmann-Willenbrock & Hung, 2023).

### Computational Analysis of Constructs

It is at the next stage that a speech-based approach bridges the gap between physical and inferential layer. All previous stages (see Figure 1) aim at extracting the physical measurements expected to reflect the information of interest in the inferential layer. In other words, all previous stages extract the physical traces in speech anticipated to reflect a social or psychological phenomenon. These phenomena cannot be heard directly in the audio signal, but they can be inferred from it (for example, the emotional content of a group interaction can be inferred from the way people speak and interact). The philosophy of science behind much of psychology is the assumption that an observed score is comprised of the true phenomenon plus random and systematic error (Trochim & Donnelly, 2001). In fact, an established areas of psychology, *psychometrics*, revolves around the valid and reliable ways of quantifying *constructs*, i.e., abstractions of social and psychological phenomena that, while not being directly accessible to human senses, can still be represented and measured (Furr, 2021). This stage draws on computational advances to speed up the quantification of social science constructs reflected in paralinguistic.

The section about diarization shows that the first type of physical information that can be extracted from speech is turn-taking, typically represented as a sequence  $T$  of triples each corresponding to a turn. Domains such as *pragmatics* (Yule, 1996) and *Conversation Analysis* (Sacks, Schegloff, & Jefferson, 1978; Schegloff, 2007) show that, however simple,  $T$  can provide substantial information about social phenomena taking

place during an interaction, irrespective of what people say and how they say it. For example,  $T$  allows one to obtain the *adjacency matrix*  $A$  in which element  $a_{ij}$  tells how frequently speaker  $j$  talks immediately after speaker  $i$ . Such information can help one identify the presence of conflicts because during discussions people tend to talk immediately after others they disagree with (Bilmes, 1988). Furthermore,  $T$  shows when someone speaks and for how long and this was shown to account for the role being played in meetings (Garg, Favre, Salamin, Hakkani Tür, & Vinciarelli, 2008) or in broadcast material (Vinciarelli, 2007). Finally,  $T$  can be used to estimate the amount of time a person speaks, and this was shown to account, e.g., for dominance (Hung, Jayagopi, Ba, Odobez, & Gatica-Perez, 2008; Mast, 2002).

The section about paralinguistic analysis shows that individual turns (or possibly entire recordings) can be converted into individual vectors or sequences of vectors. Such a format is suitable for Machine Learning and AI approaches that, once fed with vectors representing data of interest, give as output either an element  $c$  of a predefined finite set of classes  $\mathcal{C}$  or a continuous number  $r$ . In the first case the mapping is referred to as *classification*, while in the second case, it is referred to as *regression*. In both cases, the information (element  $c$  or continuous number  $r$ ) is expected to represent the phenomenon being targeted in the inferential layer in quantitative terms.

The approach above is highly suitable for interdisciplinary collaboration between computational and social sciences in cases when those social sciences quantify social and psychological phenomena as categories or continuous numbers, already understood to be imperfect representations of constructs. Emotions are a typical example. Different camps of psychologists have historically conceptualized emotions in different ways. Basic emotion theory describes seven emotions (Ekman, 1992), namely *sadness*, *anger*, *disgust*, *fear*, *surprise* and *happiness* (or *enjoyment*), with *contempt* added later, which are universal, of short duration and quick onset, and distinguished by particular facial and other signals (Ekman & Cordaro, 2011). Another set of researchers argue for Core Affect Theory,

in which affective states are inherently socially constructed and captured within three continuous dimensions, *valence*, *arousal* and *dominance* (Barrett, Mesquita, Ochsner, & Gross, 2007; Russell & Mehrabian, 1977). In this structure, the first dimension reflects whether the emotion is positive or negative and to what extent, the second one accounts for the emotion's level of activation, and the third one for the possibility of the emotion being controlled by someone experiencing it. These two theories naturally lend themselves to classification and regression, respectively. Another set of theorists more recently have proposed semantic space theory: Arguing both theoretically and using empirical data that those two early theories are insufficient, they contend that emotional experience is best captured by a high-dimensional space of dozens of emotions and blends between them (Cowen & Keltner, 2021; Cowen, Sauter, Tracy, & Keltner, 2019). This theory includes aspects that each emotion can vary in intensity and, while recognized cross-culturally, are influenced by personal, setting, and cultural differences. That third theory would require datasets be annotated by cultural and setting experts/natives and the annotation scheme to enable the simultaneous capture of multiple emotions (Paletz, Golonka, et al., 2023). In other words, the AI approach would need to use regression but also an approach that allows for the same utterances to include multiple classification.

A large number of other constructs can be represented as categories or continuous values and, therefore, can be potentially inferred from speech in the same way as emotions. A widely addressed construct is, for example, personality (John & Srivastava, 1999; Vinciarelli & Mohammadi, 2014). The reason is that the most effective and popular personality models are based on traits (Matthews, Deary, & Whiteman, 2003), i.e., on continuous dimensions that account for the most salient characteristics of an individual and can be measured with appropriate questionnaires (e.g., Rammstedt and John (2007)). Furthermore, psychometric instruments are available for a variety of team phenomena, including, e.g., interpersonal attraction (McCroskey & McCain, 1974), conflict management style (Rahim, 1983), etc.

These measures have the potential to be at different levels of analysis, be they instantiated at the utterance level, turn level, interactions, or whole team meeting. For instance, the brief expression of an emotion from one person might be at the turn level or go over a few turns. Even more intriguing are group-level phenomena that occur across turns. For instance, disagreement might be one turn by one person, or might extend over several turns as during an ongoing conflict between two or more teammates (Paletz & Schunn, 2018). Phenomena such as team viability could be detected over a minute or more (e.g., Cao et al. (2021)), and the aforementioned phenomenon of entrainment might be detected as change over the course of a team meeting (Levitan (2020); Rahimi and Litman (2018)).

The considerations above show that psychometrics provides an ideal ground for interdisciplinary collaboration between computing and human sciences, but, at the same time, it represents a bottleneck. Given the way Artificial Intelligence works, it is not possible to work on phenomena that cannot be represented as categories or continuous numbers and, correspondingly, it is not possible to apply any theory that does not represent social and psychological phenomena in quantitative terms. A variety of excellent social science research brings us knowledge about the social world through qualitative analyses, and at this time AI cannot represent that information until it is quantified.

### **State-of-the-Art**

The previous section shows the main technological components underlying a computational approach for the analysis of groups. Overall, what technology can do today is to infer social or psychological constructs (typically resulting from the judgment of a human observer) from behavioural data captured through multiple sensors. Such a level of analysis is far from the fine grained investigations conducted in social sciences, but machines are a bottleneck and do not allow one to preserve complexity and specificity accessible to human observers. On the other hand, while maybe being simplistic from a

social sciences point of view, the inference of constructs from data allowed the investigation of a wide spectrum of phenomena that were never considered before in computing science (Beyan, Vinciarelli, & Del Bue, 2023). This section focuses mostly on those works in which audio-based features play an important role.

One of the first problems addressed was the recognition of *roles*, defined as functions that people play in a group interaction. After a few works focusing on roles that are highly specific of a particular setting (e.g., the *anchorman* in a radio or television program Salamin, Favre, and Vinciarelli (2009)), the attention shifted towards roles that are as independent as possible from the situation and rather reflect social dynamics such as, e.g., *protagonist*, a person who manages to manage the floor and assert her authority, or *supporter*, an individual who tends to cooperate with others (see Vinciarelli, Valente, Yella, and Sapru (2011) for a description of a commonly used set of these roles). The possibility to generalize these roles over multiple settings was at the core of their success and resulted into several approaches aimed at their detection (Dong, Lepri, Pianesi, & Pentland, 2012; Fotedar, Gaonkar, Chatterjee, & Ghosh, 2016; Sapru & Bourlard, 2015; L. Zhang & Radke, 2020).

The attempt to recognize constructs independent of the particular setting attracted attention on *personality traits* because these, by definition, are expected to be stable over the life of an individual and, therefore, people are expected to manifest them, at least to a certain extent, in all situations. Not surprisingly, the experiments on personality recognition were performed over multiple types of data, from meetings, the most common case (Aran & Gatica-Perez, 2013; Dotti, Popa, & Asteriadis, 2020; Staiano, Lepri, Kalimeri, Sebe, & Pianesi, 2011; Valente, Kim, & Motlicek, 2012) to movies (Srivastava, Feng, Roy, Yan, & Sim, 2012) and surveillance videos (Favaretto, Knob, Musse, Vilanova, & Costa, 2019). Traits are typically measured in terms of continuous scores, but most experiments try to classify people as *high* or *low* along the various traits, meaning that they are above or below a statistical threshold (e.g., the median or the average). Such an

approximation is necessary because the trait distributions tend to be peaked around values at the centre of the scales and, correspondingly, the variance tends to be low (Vinciarelli & Mohammadi, 2014).

Other research efforts targeted *dominance*, the tendency to impose one's will over the one of others. The interesting aspect of such a construct is that the literature provides clear indications about the cues that someone must look at and automatic approaches were highly effective at capturing them. For example, it was shown that dominant people tend to attract more visual attention than others and, correspondingly, approaches based on automatic analysis of Visual Focus of Attention (Hung et al., 2008) were actually shown to be effective. Similarly, the tendency to speak and grab the floor (detectable through speaker diarization) was shown to be an effective way to identify the most dominant person in an interaction (Hung, Huang, Friedland, & Gatica-Perez, 2010). In a similar vein, several works addressed the problem of detecting leadership, whether this means to identify who is the leader in a group (Jayagopi & Gatica-Perez, 2010) or what are the leadership styles of different individuals such as, e.g., *autocratic vs legalitarian, considered vs free-rein*, etc. (Beyan, Capozzi, Becchio, & Murino, 2017; Feese et al., 2011).

All problems considered so far can be considered as an attempt to identify individuals' traits and characteristics manifested while participating in group interactions. In parallel, several efforts were made to analyze a group as whole. For example, the focus in Jayagopi, Sanchez-Cortes, Otsuka, Yamato, and Gatica-Perez (2012) was on understanding what a group is doing and how (e.g., *cooperative versus competitive, brainstorming versus decision-making or formal vs. informal*). A similar problem was addressed in Matic, Osmani, and Mayora-Ibarra (2014) with the help of modalities different from audio, but the focus was on activities specific of a workplace. More limited efforts were made to analyze phenomena such as performance, how effectively groups address a task (Kubasova, Murray, & Braley, 2019), satisfaction, how happy are people to be part of a group (Lai & Murray, 2018), and conflict, how people manifest the presence of

incompatible interests (Kim, Valente, Filippone, & Vinciarelli, 2014). Overall, the group-level problem that attracted most interest is cohesion, i.e., the tendency of group members to like each other (referred to as *social* cohesion) and to achieve tasks in a shared way (referred to as *task* cohesion). The problem was addressed mostly in the case of meetings (Sharma, Ghosh, & Dhall, 2019; Y. Zhang, Olenick, Chang, Kozlowski, & Hung, 2018) and audio-related features were accompanied by other measurements that account for different modalities.

## Conclusions

When social scientists are flooded with data, big data analyses may be a solution to feasibly analyzing and distilling that data into construct measurement. From the computational side, data availability remains a crucial issue because AI methodologies require increasingly larger amounts of material to be trained. Given that data collection is typically an expensive and time consuming process, such an issue risks to become a bottleneck that prevents research from progressing. One possible solution is to collect material “in the wild”, i.e., without setting up an experiment and inviting participants, but simply recording material wherever groups of individuals gather and interact (e.g., public spaces, restaurants, cafés, etc.). This approach can allow one to accumulate large amounts of data in a short time, but there is still the problem of the annotation process and, even more importantly, there are significant ethical issues. In fact, it would be quite difficult to obtain the informed consent of all people recorded in a naturalistic environment, not to mention that collecting data in the wild is likely to violate privacy regulations in most countries. Similar considerations apply to data collected in activities that involve audio recording as a natural component like, e.g., the interaction with call centre operators or the therapeutic interviews between clinicians and mental health patients. In many countries or states, recording requires the consent of all parties (Wiles, Crow, Charles, & Heath, 2007). Furthermore, the data generated in organizations such as companies and hospitals are

unlikely to become public because of privacy concerns and because they are too much an important asset to be shared.

More promising solutions to the data problem come from two other sources. The first is the recent availability of foundational models that, while being trained to perform a certain task, are effective in many other contexts too. A typical case in speech processing are feature extractors such as *wav2vec* (Baevski et al., 2020) and *whisper* (Radford et al., 2023). They were trained over large corpora to perform Automatic Speech Recognition, but they are effective for many other tasks for which there is less data. Similarly, methodologies like transfer learning (Niu, Liu, Wang, & Song, 2020) allow one to “adapt” large models trained over major databases to tasks for which only limited material is available. In this respect, the Deep Learning community appears to provide approaches for benefiting from large models trained with big, but typically private datasets, while having at disposition only limited material for a task at hand. For these methods to be applicable, however, researchers need to explicitly test and be assured that the base datasets do not include bias or significantly missing data, such as on minority groups (e.g., Ntoutsis et al. (2020)), or be applied cross-linguistically or cross-culturally without care. Last, but not least, there is discussion on whether *good* data might be a valuable alternative to *big* data (Ghasemaghaei & Calic, 2020). This possibility is of particular interest in the case of group analysis because the interdisciplinary collaboration between computing and human sciences might help to ensure that even a limited amount of data contains sufficient information for a model to learn how to bridge the gap between physical and inferential layer.

Another important issue is that the most effective and widely applied AI methodologies are based on associative statistics and not on causal explanations. In other words, most approaches learn the association between variance in the physical layer and variance in the inferential layer (e.g., variance in loudness is associated to variance in emotion), but this association does not allow one to say that one is the explanation of the other or one influences the other. In other words, while current speech-based group

analysis approaches can lead to good results, they do not necessarily provide an explanation of the phenomena. On the one hand, this provides an opportunity for interdisciplinary collaboration because human sciences can help to make sense of the results achieved with associative statistics. On the other hand, this leads to the risk that the results simply depend on spurious associations in the data at disposition that do not necessarily generalize to other data.

As new AI technologies raise a host of issues, the United States and the European Union are rolling out standards and guidelines for AI (see, e.g., the NIST AI Risk Framework<sup>4</sup> or the European Union AI Act<sup>5</sup>). In addition to issues of privacy, additional ethical concerns include the explainability and transparency of the AI methodologies. These issues revolve around the ability of AI approaches to simultaneously include a number of variables that would be impossible for humans to parse but also the lack of transparency in revealing exactly how results are generated, including the degree of undesirable biases that may be lurking. The possibility of modelling jointly large numbers of variables is certainly a positive aspect because it allows one to detect and benefit from relationships impossible to detect by naked eye, but it is also a problem because it means that it is not possible to make sense of the process leading from input (physical aspects of speech in the case of this chapter) to output (inferential aspects of group phenomena in this chapter). Major efforts are underway to develop both approaches that can provide causal explanations (see, e.g., Prospero et al. (2020)) or intelligible explanations of their outcomes (see, e.g., Confalonieri, Coba, Wagner, and Besold (2021)). The social science of technologies can also help inform future directions in AI and generate suggestions of how to overcome these issues.

In conclusion, the collaboration between social and computational scientists in

---

<sup>4</sup> <https://www.nist.gov/itl/ai-risk-management-framework>, last accessed on March 30, 2024.

<sup>5</sup> <https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf>, last accessed on March 30th, 2024.

leveraging AI to study teams has great promise. The best theory, findings, methods, and knowledge of constructs from social science, combined with the most exciting and rigorous computational approaches, have the potential to provide us with a deeper understanding of how teams work—in this chapter’s case, regarding paralanguage features. At its best, such an approach would be sensitive to cross-cultural differences in speech, not just in words but in both the expression and interpretation of the social and psychological phenomena under study.

## References

- Allen, J. A., & Lehmann-Willenbrock, N. (2023). The science of workplace meetings: Integrating findings, building new theoretical angles, and embracing cross-disciplinary research. *Organizational Psychology Review*, *13*(4), 351–354.
- Aran, O., & Gatica-Perez, D. (2013). One of a kind: Inferring personality impressions in meetings. In *Proceedings of the ACM International Conference on Multimodal Interaction* (pp. 11–18).
- Averbuch, A., Bahl, L., Bakis, R., Brown, P., Cole, A., Daggett, G., . . . Spinelli, P. (1986). An IBM PC based large-vocabulary isolated-utterance speech recognizer. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 11, p. 53-56).
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 12449–12460).
- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68.
- Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Reviews of Psychology*, *58*, 373–403.
- Beyan, C., Capozzi, F., Becchio, C., & Murino, V. (2017). Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Transactions on Multimedia*, *20*(2), 441–456.
- Beyan, C., Vinciarelli, A., & Del Bue, A. (2023). Co-located human–human interaction analysis using nonverbal cues: A survey. *ACM Computing Surveys*, *56*(5), 1–41.

- Bilmes, J. (1988). The concept of preference in conversation analysis. *Language in Society*, *17*, 161-181.
- Bonin, F., Campbell, N., & Vogel, C. (2014). Time for laughter. *Knowledge-Based Systems*, *71*, 15–24.
- Cao, H., Yang, V., Chen, V., Lee, Y. J., Stone, L., Diarrassouba, N. J., . . . Bernstein, M. S. (2021). My team will go on: Differentiating high and low viability teams through team interaction. In *Proceedings of CHI* (Vol. 4, pp. 1–27).
- Chow, Y., Dunham, M., Kimball, O., Krasner, M., Kubala, G., Makhoul, J., . . . Schwartz, R. (1987). BYBLOS: The BBN continuous speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 12, pp. 89–92).
- Clark, H., & Tree, J. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.
- Confalonieri, R., Coba, L., Wagner, B., & Besold, T. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(1), e1391.
- Cosentino, S., Sessa, S., & Takanishi, A. (2016). Quantitative laughter detection, measurement, and classification—a critical survey. *IEEE Reviews in Biomedical Engineering*, *9*, 148-162.
- Cowen, A., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, *25*(2), 124–136.
- Cowen, A., Sauter, D., Tracy, J. L., & Keltner, D. (2019). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, *20*(1), 69–90.
- Davis, K., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, *24*(6), 637–642.
- Dong, W., Lepri, B., Pianesi, F., & Pentland, A. (2012). Modeling functional roles

- dynamics in small group interactions. *IEEE transactions on Multimedia*, 15(1), 83–95.
- Dotti, D., Popa, M., & Asteriadis, S. (2020). Being the center of attention: A person-context cnn framework for personality recognition. *ACM Transactions on Interactive Intelligent Systems*, 10(3), 1–20.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4), 169–200.
- Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4), 364–370.
- Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, 128(2), 203.
- Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., . . . Truong, K. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.
- Favaretto, R. M., Knob, P., Musse, S. R., Vilanova, F., & Costa, Â. B. (2019). Detecting personality and emotion traits in crowds from video sequences. *Machine Vision and Applications*, 30, 999–1012.
- Feese, S., Muaremi, A., Arnrich, B., Troster, G., Meyer, B., & Jonas, K. (2011). Discriminating individually considerate and authoritarian leaders by speech activity cues. In *Proceedings of the IEEE International Conference on Social Computing* (pp. 1460–1465).
- Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, 12(1), 47.
- Fotedar, G., Gaonkar, A., Chatterjee, S., & Ghosh, P. K. (2016). Automatic recognition of social roles using long term role transitions in small group interactions. In *Proceedings of Interspeech* (pp. 2065–2069).
- Furr, R. M. (2021). *Psychometrics: An introduction*. SAGE Publications.
- Garg, N., Favre, S., Salamin, H., Hakkani Tür, D., & Vinciarelli, A. (2008). Role

- recognition for meeting participants: an approach based on lexical information and social network analysis. In *Proceedings of the ACM International Conference on Multimedia* (pp. 693–696).
- Ghasemaghaei, M., & Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, *108*, 147–162.
- Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.
- Giles, H., Edwards, A. L., & Walther, J. B. (2023). Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, *99*.
- Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society*, *2*(2), 177–192.
- Gottman, J. M., & Levenson, R. W. (2000). The timing of divorce: Predicting when a couple will divorce over a 14-year period. *Journal of Marriage and Family*, *62*(3), 737–745.
- Gottman, J. M., & Notarius, C. I. (2000). Decade review: Observing marital interaction. *Journal of Marriage and Family*, *62*(4), 927–947.
- Hung, H., Huang, Y., Friedland, G., & Gatica-Perez, D. (2010). Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 847–860.
- Hung, H., Jayagopi, D. B., Ba, S., Odobez, J.-M., & Gatica-Perez, D. (2008). Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 233–236).

- Jayagopi, D., & Gatica-Perez, D. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia*, 12(8), 790–802.
- Jayagopi, D., Hung, H., Yeo, C., & Gatica-Perez, D. (2009). Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3), 501–513.
- Jayagopi, D., Sanchez-Cortes, D., Otsuka, K., Yamato, J., & Gatica-Perez, D. (2012). Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the ACM International Conference on Multimodal Interaction* (pp. 433–440).
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*. Guilford Press.
- Kane, A., & van Swol, L. (2022). Harnessing a language analysis perspective to uncover emergent group processes. In *Handbook of language analysis in psychology*. The Guilford Press.
- Kim, S., Valente, F., Filippone, M., & Vinciarelli, A. (2014). Predicting continuous conflict perception with bayesian gaussian processes. *IEEE Transactions on Affective Computing*, 5(2), 187–200.
- Knapp, M., & Hall, J. (1997). *Nonverbal communication in human interaction*. Harcourt Brace College Publishers.
- Kubasova, U., Murray, G., & Braley, M. (2019). Analyzing verbal and nonverbal features for predicting group performance. In *Proceedings of Interspeech* (p. 1896-1900).
- Lai, C., & Murray, G. (2018). Predicting group satisfaction in meeting discussions. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data* (pp. 1–8).
- Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological*

*Science*.

- Lefter, I., & Jonker, C. (2017). Aggression recognition using overlapping speech. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (p. 299-304).
- Lehmann-Willenbrock, N., & Chiu, M. M. (2018). Igniting and resolving content disagreements during team interactions: A statistical discourse analysis of team dynamics at work. *Journal of Organizational Behavior*, *39*(9), 1142–1162.
- Lehmann-Willenbrock, N., & Hung, H. (2023). A multimodal social signal processing approach to team interactions. *Organizational Research Methods*.
- Levitan, R. (2020). Developing an integrated model of speech entrainment. In *Proceedings of the International Joint Conference on Artificial Intelligence*.
- Liu, Z., Kang, H., Feng, L., & Zhang, L. (2017). Speech pause time: A potential biomarker for depression detection. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine* (pp. 2020–2025).
- Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the ACM Workshop on Multimodal Learning Analytics* (pp. 5–12).
- Lubold, N., Walker, E., & Pon-Barry, H. (2021). Effects of adapting to user pitch on rapport perception, behavior, and state with a social robotic learning companion. *User Modeling and User-Adapted Interaction*, *31*, 35–73.
- Marzinzik, M., & Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, *10*(2), 109–118.
- Mast, M. (2002). Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research*, *28*(3), 420–450.
- Matic, A., Osmani, V., & Mayora-Ibarra, O. (2014). Mobile monitoring of formal and informal social interactions at workplace. In *Proceedings of the ACM International*

- Joint Conference on Pervasive and Ubiquitous Computing* (pp. 1035–1044).
- Matthews, G., Deary, I., & Whiteman, M. (2003). *Personality traits*. Cambridge University Press.
- McCowan, L., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., & Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3), 305-317.
- McCroskey, J., & McCain, T. (1974). *The measurement of interpersonal attraction*. Taylor & Francis.
- Niu, S., Liu, Y., Wang, J., & Song, H. (2020). A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, *1*(2), 151-166.
- Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.-E., . . . Staab, S. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.
- Paletz, S. B., Golonka, E. M., Pandža, N. B., Stanton, G., Ryan, D., Adams, N., . . . others (2023). Social media emotions annotation guide (SMemo): Development and initial validity. *Behavior Research Methods*, 1–51.
- Paletz, S. B., Litman, D., Karuzis, V., Jones, K. M., & Rahimi, Z. (2023). Speaking similarly: team personality composition and acoustic-prosodic entrainment. *Small Group Research*, *54*(6), 860–898.
- Paletz, S. B., Schunn, C., & Kim, K. (2011). Intragroup conflict under the microscope: Micro-conflicts in naturalistic team discussions. *Negotiation and Conflict Management Research*, *4*(4), 314-351.
- Paletz, S. B., & Schunn, C. D. (2011). Assessing group-level participation in fluid teams: Testing a new metric. *Behavior Research Methods*, *43*, 522–536.
- Paletz, S. B., & Schunn, C. D. (2018). Micro-conflict coding scheme. In *The cambridge handbook of group interaction analysis* (p. 472-482). Cambridge University Press.

- Park, T., Kanda, N., Dimitriadis, D., Han, K., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language, 72*, 101317.
- Picard, R. (2000). *Affective computing*. MIT Press.
- Pieraccini, R. (2012). *The voice in the machine: building computers that understand speech*. MIT Press.
- Prosperi, M., Guo, Y., Sperrin, M., Koopman, J., Min, J., He, X., . . . Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence, 2*(7), 369–375.
- Provine, R. (2001). *Laughter: A scientific investigation*. Penguin.
- Radford, A., Kim, J., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning* (pp. 28492–28518).
- Rahim, M. (1983). A measure of styles of handling interpersonal conflict. *Academy of Management Journal, 26*(2), 368–376.
- Rahimi, Z., & Litman, D. (2018). Weighting model based on group dynamics to measure convergence in multi-party dialogue. In K. Komatani, D. Litman, K. Yu, A. Papangelis, L. Cavedon, & M. Nakano (Eds.), *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue* (p. 385-390). Association for Computational Linguistics.
- Rammstedt, B., & John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality, 41*(1), 203–212.
- Richmond, P., & McCroskey, J. (2000). *Nonverbal behavior in interpersonal relations*. Allyn & Bacon.
- Rochester, S. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research, 2*, 51–81.

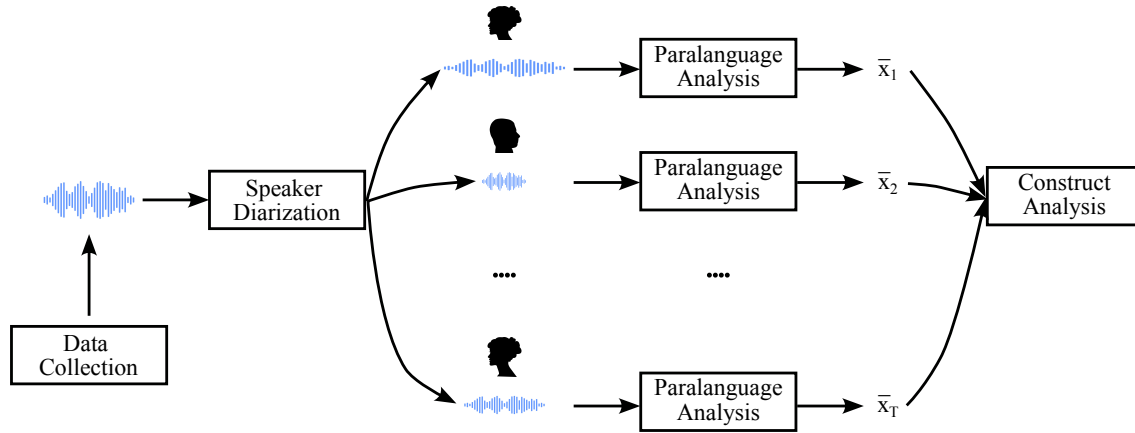
- Russell, J., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3), 273–294.
- Sacks, A., Schegloff, E., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (p. 7-55). Academic Press.
- Salamin, H., Favre, S., & Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, 11(7), 1373–1380.
- Sapru, A., & Bourlard, H. (2015). Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, 17(5), 746–760.
- Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1), 1–63.
- Schegloff, E. (2007). *A primer in conversation analysis*. Cambridge University Press.
- Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2), 227–256.
- Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. John Wiley & Sons.
- Sharma, G., Ghosh, S., & Dhall, A. (2019). Automatic group level affect and cohesion prediction in videos. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 161–167).
- Srivastava, R., Feng, J., Roy, S., Yan, S., & Sim, T. (2012). Don't ask me what i'm like, just watch and listen. In *Proceedings of the ACM International Conference on Multimedia* (pp. 329–338).
- Staiano, J., Lepri, B., Kalimeri, K., Sebe, N., & Pianesi, F. (2011). Contextual modeling of personality states' dynamics in face-to-face interactions. In *Proceedings of the IEEE International Conference on Social Computing* (pp. 896–899).
- Stouten, F., & Martens, J.-P. (2003). A feature-based filled pause detection system for

- dutch. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 309–314).
- Tannenbaum, S., & Salas, E. (2020). *Teams that work: the seven drivers of team effectiveness*. Oxford University Press.
- Trochim, W. M., & Donnelly, J. P. (2001). *Research methods knowledge base*. Atomic Dog Publishing Cincinnati.
- Valente, F., Kim, S., & Motlicek, P. (2012). Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *Proceedings of Interspeech* (pp. 1183–1186).
- Verkhodanova, V., Shapranov, V., & Kipyatkova, I. (2017). Hesitations in spontaneous speech: acoustic analysis and detection. In *Proceedings of Speech and Computer* (pp. 398–406).
- Vinciarelli, A. (2007). Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(6), 1215–1226.
- Vinciarelli, A. (2009). Capturing order in social interactions [social sciences]. *IEEE Signal Processing Magazine*, 26(5), 133–152.
- Vinciarelli, A., Chatziioannou, P., & Esposito, A. (2015). When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, 2, 4.
- Vinciarelli, A., & Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3), 273–291.
- Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12), 1743–1759.
- Vinciarelli, A., Valente, F., Yella, S. H., & Sapru, A. (2011). Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the AMI meeting corpus. In *Proceedings of the IEEE International Conference on*

- Systems, Man, and Cybernetics* (pp. 374–379).
- Wang, L., Doucet, L., Waller, M., Sanders, K., & Phillips, S. (2016). A laughing matter: Patterns of laughter and the effectiveness of working dyads. *Organization Science*, *27*(5), 1142–1160.
- Wani, T., Gunawan, T., Qadri, S., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, *9*, 47795–47814.
- Wharton, T. (2009). *Pragmatics and non-verbal communication*. Cambridge University Press.
- Wiles, R., Crow, G., Charles, V., & Heath, S. (2007). Informed consent and the research process: following rules or striking balances? *Sociological Research Online*, *12*(2), 99–110.
- Worgan, S., & Moore, R. (2011). Towards the detection of social dominance in dialogue. *Speech Communication*, *53*(9-10), 1104–1114.
- Yousefi, M., & Hansen, J. (2020). Block-based high performance CNN architectures for frame-level overlapping speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 28–40.
- Yu, M., Litman, D., & Paletz, S. (2019). Investigating the relationship between multiparty linguistic entrainment, team characteristics and the perception of team social outcomes. In R. Bart & K. Brawner (Eds.), *Proceedings of the 32nd International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (p. 227-232). Association for the Advancement of Artificial Intelligence.
- Yule, G. (1996). *Pragmatics*. Oxford University Press.
- Zancanaro, M., Lepri, B., & Pianesi, F. (2006). Automatic detection of group functional roles in face to face interactions. In *Proceedings of the ACM International Conference on Multimodal Interfaces* (pp. 28–34).
- Zhang, L., & Radke, R. J. (2020). A multi-stream recurrent neural network for social role

detection in multiparty interactions. *IEEE Journal of Selected Topics in Signal Processing*, 14(3), 554–567.

Zhang, Y., Olenick, J., Chang, C.-H., Kozlowski, S. W., & Hung, H. (2018). TeamSense: assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), 1–22.



*Figure 1.* The scheme shows the most common and standard processing steps in an approach for audio-based group analysis. Vectors  $\vec{x}_k$  are paralinguistic representations extracted from every turn and  $T$  is the total number of turns.