1 Multimodal Analysis of Social Signals

One of the earliest books dedicated to the communication between living beings is "*The Expression of Emotion in Animals and Man*" by Charles Darwin. The text includes a large number of accurate and vivid descriptions of the way living beings express emotions: "*many kinds of monkeys, when pleased, utter a reiterated sound, clearly analogous to our laughter, often accompanied by vibratory movements of their jaws or lips, with the corners of the mouth drawn backwards and upwards, by the wrinkling of the cheeks, and even by the brightening of the eyes*" [Darwin 1872]. Darwin never uses the word "*multimodal*", but the example above - and the many similar others the book contains - makes it clear that the expression of emotions often involves the simultaneous use of multiple communication channels and, correspondingly, the simultaneous stimulation of different senses.

Around one century after the seminal insights by Darwin, research in life and human sciences started to adopt the expression "*multimodal communication*" to denote the phenomenon above and to investigate its underlying principles and laws [Partan and Marler 1999; Rowe and Guilford 1996; Scheffer et al. 1996]. In parallel, it was observed that multimodality is not peculiar of expression and perception of emotions, but it concerns any interaction between living beings (human-human, animal-animal or human-animal) [Poggi 2007]. In other words, multimodality plays a major role in the exchange of *social signals*, i.e., "*acts or structures that influence the behavior or internal state of other individuals*" [Mehu and Scherer 2012], "*communicative or informative signals which [...] provide information about social facts*" [Poggi and D'Errico 2012], or "*actions whose function is to bring about some reaction or to engage in some process*" [Brunet and Cowie 2012].

At the moment machines become powerful enough to deal with human behaviour and its subtleties, the interest for multimodal communication reaches the computing community as well [Vinciarelli et al. 2009, 2012]. Figure 1.1 shows the distribution of the number of computing oriented papers containing the word "*multimodal*" in the ACM Digital Library, one of the most important repositories of computing literature. The chart shows that the interest for the topic has been continuously increasing for the last 15 years. According to the latest technology forecasts [1], socially intelligent technologies - in particular humanoid robots - will

[1] According to Tractica, "*annual robot unit shipments will increase from 8.8 million in 2015 to 61.4 million by 2020, with more than half the volume in that year coming from consumer robots*", i.e., robots that integrate the everyday life of their users (`https://www.tractica.com/newsroom/press-releases/global-robotics-industry-to-surpass-151-billion-by-2020/`)

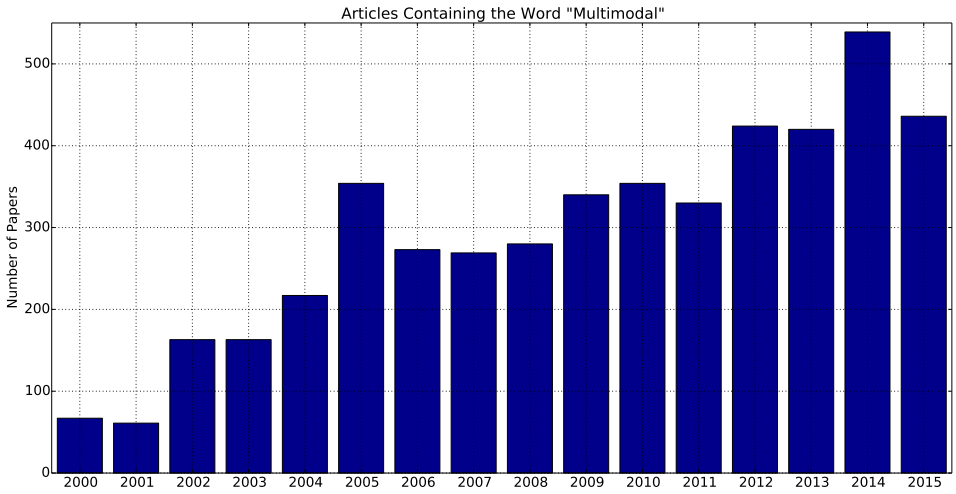

Figure 1.1 The chart shows the number of papers that the ACM Digital Library returns after submitting the query "*multimodal*". Overall, the number grows continuously since 2000.

become a ubiquitous feature of everyday life in the next 20 years. This suggests that the trend of Figure 1.1 will continue in the foreseeable future.

Overall, the brief outline above shows that multimodality and multimodal communication are concepts that attract attention in communities as diverse as life sciences, computing and human sciences. However, the exact meaning of the term "*multimodal*" is not necessarily the same in all fields. The goal of this chapter is to address, at least partially, such an issue and to show differences and commonalities (if any) behind the use of the word "*multimodal*" in the various areas. In particular, this chapter tries to show if concepts originally elaborated in life sciences can be "translated" into computational methodologies and, if yes, how.

The rest of the chapter is organised as follows: Section 1.1 provides a brief introduction to the concept of multimodality in life and human sciences; Section 1.2 describes the key-methodological issues of multimodal approaches for the analysis of social signals; Section 1.3 highlights some future perspectives and, finally, Section 1.4 draws some conclusions.

1.1 Multimodal Communication in Life and Human Sciences

According to life sciences, "*Animals communicate with their entire bodies and perceive signals with all available faculties (vision, audition, chemoreception, etc.). To best understand communication, therefore, we must consider the whole animal and all of its sensory emissions and percepts*" [Partan and Marler 2005]. However, what makes communication truly multimodal is not the joint use of multiple modalities, but their integration to achieve communicative effects that cannot be achieved individually by the various modalities involved:

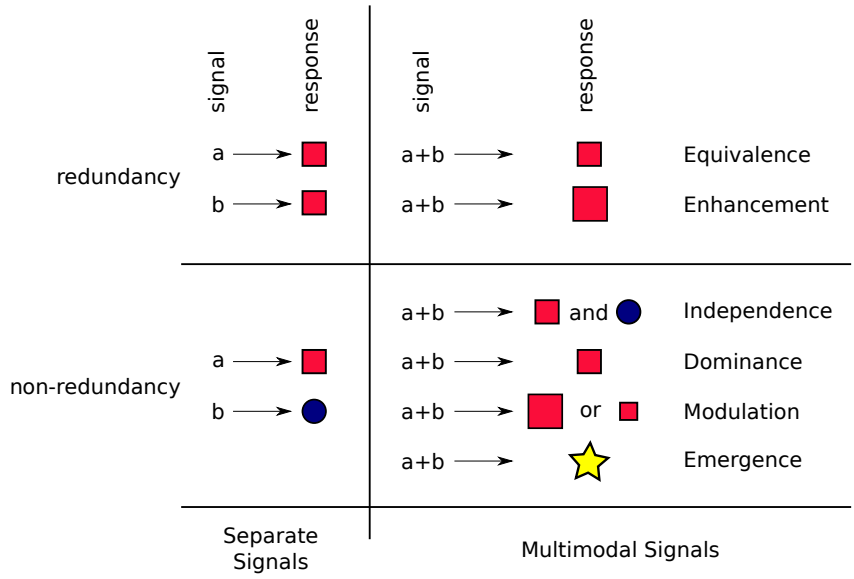

Figure 1.2 The figure reproduces the scheme proposed in [Partan and Marler 1999, 2005] and it shows the multimodal communication patterns observed in nature.

"*[the use of multimodal signals] produces unexpected psychological responses [...] which remain hidden when the components are presented alone, with the clear implication that the full significance of multicomponent animal signals cannot be understood by investigating components independently*" [Rowe and Guilford 1996]. In other words, communication is multimodal when the effect of multiple modalities is not just the sum of the individual effects achieved with individual modalities.

Figure 1.2 shows a taxonomy of multimodal communication patterns observed in nature. The first important distinction is between patterns based on *redundant* and *non-redundant* signals, corresponding to upper and lower half of the figure, respectively [Partan and Marler 1999]. In the case of redundant signals, the different modalities carry the same information and the main function of multimodality is to ensure that the message reaches the agent supposed to perceive it. A typical case in nature is the combination of acoustic signals and movements. These latter ensure that the message can be received even when there is loud noise in the environment, a condition that is frequent in natural settings. Conversely, the acoustic signals ensure that the message can be received even in absence of light or in case of visual occlusion. In the case of non-redundant signals, the main function of multimodality is to transmit a larger amount of information per unit of time. For example, the appropriate combination of pigmentation and chemical signals can discourage predators or attract sexual mates, two

functions that require the relevant information to be communicated quickly in order to survive and transmit genes, respectively [Scheffer et al. 1996].

The second important criterion that informs the taxonomy of Figure 1.2 is the response of the subject that perceives the multimodal communication pattern [Partan and Marler 2005], whether the response corresponds to a behavioural display or to a correct understanding of the message conveyed by the pattern. In the case of redundant signals, two scenarios are observed, namely *equivalence* and *enhancement*. The former corresponds to the case in which the response to the multimodal signal is the same as the responses observed when the unimodal signals are perceived individually. The latter corresponds to the case in which the response is the enhanced version of the one observed when the unimodal signals are perceived individually.

When the signals are non-redundant, the scenarios observed in nature are four, i.e., *independence*, *dominance*, *modulation* and *emergence* (see Figure 1.2). Independence means that the response to a multimodal pattern is the mere sum of the responses to the individual signals. According to the principle outlined at the beginning of this section, such a scenario does not even qualify as an example of multimodal communication. The dominance scenario covers those cases where the response is the same as the one that would be observed for one of the unimodal signals in the multimodal pattern. The modulation scenario is similar, but the magnitude of the response changes with respect to the unimodal case. Finally, the emergence scenario corresponds to those cases in which the response is different from those that can be observed when each of the unimodal signals is perceived individually.

It is not surprising to observe that the first observations on multimodal communication have been done by life scientists studying interactions between animals. The reason is that these use a much wider variety of sensory channels than how humans do, including hearing, sight, radar-like receptors of acoustic waves, olfaction, chemoreceptors, etc. [Rowe and Guilford 1996; Scheffer et al. 1996]. When it comes to human-human communication, interaction takes place mainly via speech and visual signals (facial expressions, gestures, posture, mutual distances, etc.). The other channels - touch, smell and taste - are used only rarely and only in very specific contexts (e.g., sexual intercourse). For this reason, in the case of humans, multimodality typically means *bimodality*. In particular, it is possible to distinguish between *micro-bimodality* (the combination between speech and movements like those of the lips that are necessary for the very emission of voice and articulation of phonemes) and *macro-bimodality* (the combination of speech and movements like facial expressions that are not strictly necessary for the emission of speech) [Poggi 2007].

To the best of our knowledge, no systematic attempts have been done to verify whether the taxonomy of Figure 1.2 applies to human-human communication. However, a few examples indicate that this is actually the case. A person that attracts the attention of others by shouting and waiving arms is a case of equivalence (the visual signal tries to reach distances that the voice cannot). Similarly, persons that say “No” while shaking their head enhance their

message through the use of two redundant signals. In the case of non-redundant signals (lower half of Figure 1.2), a person can manifest aggressiveness through the tone of her voice while showing fear through a defensive body posture, thus fitting the independence scenario. The dominance case applies, e.g., to someone that says to be comfortable while blushing and, hence, is perceived as someone that is actually not comfortable. For what concerns the modulation case, prosody helps to stress and emphasise certain parts of a verbal message, thus achieving a modulation effect. Finally, irony can be considered a case of emergence where, e.g., the verbal and non-verbal components of a message are opposite to one another.

1.2 Multimodal Analysis of Social Signals

The taxonomy of Figure 1.2 does not explain how multimodal communication patterns result into a given response, it simply provides criteria and terminology to describe multimodal communication in rigorous terms. Approaches for multimodal analysis of social signals do not explain the way multimodal signals produce a response either [Vinciarelli et al. 2009, 2012]. However, computer analysis requires one to express the process in operational terms. If $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_R)$ is a multimodal pattern - $\vec{x}_i$ is a vector of physical measurements extracted from modality i, meaning from the data captured with sensor i, and R is the total number of sensors adopted (every modality corresponds to a sensor) - there are two approaches to deal with it. The first is called *early fusion* and it consists in concatenating the unimodal $\vec{x}_i$ vectors to obtain a multimodal vector $\vec{x}$. This latter can then be fed to any pattern recognition approach to perform classification or regression depending on the problem. The response to $\vec{x}$ will be in this case the output of the particular approach being used (e.g., a probability, a score, a distance, etc.). The second is called *late fusion* and consists in adopting a *Multiple Classifiers System*, i.e. a combination of several classifiers that deal separately with the various modalities. The output of an MCS is a probability distribution $P(\omega_k|X)$, where $\omega_k \in \Omega = \{\omega_1, \ldots, \omega_L\}$ and Ω is the set of all possible responses. The probability distribution $P(\omega_k|X)$ allows one to make a decision about the response to X according to the following rule:

$$\hat{\omega} = \arg \max_{k=1,\ldots,L} P(\omega_k|X) = \arg \max_{k=1,\ldots,L} P(\omega_k|\vec{x}_1, \ldots, \vec{x}_R). \tag{1.1}$$

Early and late fusion are not expected to explain the way living beings respond to multimodal communication patterns. They are simply methodologies that allow a machine to map a multimodal input pattern X into a suitable response $\hat{\omega}$. In other words, early and late fusion can maybe reproduce the observations summarised in Figure 1.2, but they cannot be considered an explanation or a model of the processes that lead from stimulus to response in living beings.

The rest of this section focuses on the probabilistic framework underlying MCS and, in particular, it closely follows the approach proposed in [Kittler et al. 1998] to estimate $P(\omega_j|\vec{x}_1, \ldots, \vec{x}_R)$ (the literature does not provide similar frameworks for the early fusion).

Such a distribution is difficult to estimate because this requires to know the distribution $p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j)$ which is typically difficult to infer:

$$P(\omega_j|\vec{x}_1, \ldots, \vec{x}_R) = \frac{p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j)P(\omega_j)}{p(\vec{x}_1, \ldots, \vec{x}_R)} = \frac{p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j)P(\omega_j)}{\sum_{k=1}^{L} p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j)P(\omega_k)}. \quad (1.2)$$

For this reason, it is common to make independence assumptions that, while making the problem tractable, lead to intuitive combination rules. If the vectors $\vec{x}_k$ are assumed to be statistically independent given ω_j, then $p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j)$ boils down to the following:

$$p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j) = \prod_{k=1}^{R} p(\vec{x}_k|\omega_j) \quad (1.3)$$

The main problem with such an approximation is that the posterior probability becomes low even if just one of the terms $p(\vec{x}_k|\omega_j)$ is low. For this reason, the adoption of MCS requires the assumption that the posteriors $P(\omega_k|\vec{x}_i)$ are similar to the a-priori probabilities $P(\omega_k)$:

$$P(\omega_j|\vec{x}_i) = P(\omega_j) \cdot (1 + \delta_{ji}) \quad (1.4)$$

with $|\delta_{ji}| << 1$. This leads to the following:

$$P(\omega_j|X)\alpha P(\omega_j) \prod_{k=1}^{R} p(\vec{x}_k|\omega_j) = P(\omega_j)^{1-R} \prod_{k=1}^{R} p(\omega_j|\vec{x}_k) = P(\omega_j) \prod_{k=1}^{R} (1 + \delta_{jk}). \quad (1.5)$$

Given that $|\delta_{ji}| << 1$, it is possible to neglect the non-linear terms of the product appearing in the rightmost term:

$$P(\omega_j) \prod_{k=1}^{R} (1 + \delta_{jk}) = P(\omega_j) + P(\omega_j) \sum_{k=1}^{R} \delta_{jk}. \quad (1.6)$$

Following Equation (1.4), the expression of the δ_{jk} is as follows:

$$\delta_{jk} = \frac{p(\omega_j|\vec{x}_k)}{P(\omega_j)} - 1, \quad (1.7)$$

and by replacing the expression above in Equation 1.6, the result is:

$$P(\omega_j|\vec{x}_k)\alpha(1 - R)P(\omega_j) + \sum_{k=1}^{R} p(\omega_j|\vec{x}_k). \quad (1.8)$$

The expression above is known as *Sum Rule* and it has the major advantage that $P(\omega_j|\vec{x}_k)$ can be significantly different from zero even if one or more of the probabilities $p(\omega_j|\vec{x}_k)$ are not.

Consider the following relationship:

$$\prod_{k=1}^{R} P(\omega_j|\vec{x}_k) \leq \min_k P(\omega_j|\vec{x}_k) \leq \frac{1}{R} \sum_{k=1}^{R} P(\omega_j|\vec{x}_k) \leq \max_k P(\omega_j|\vec{x}_k). \quad (1.9)$$

The indication is that the product rule can be approximated with the minimum of the posteriors, while the sum rule can be approximated with the their maximum. This gives rise to some of the most common practices adopted in the late fusion of classifiers. The first example is the following:

$$\hat{\omega} = \arg\max_j \left[(1-R)P(\omega_j) + \sum_{k=1}^{R} p(\omega_j|\vec{x}_k) \right] =$$

$$= \arg\max_j \left[(1-R)P(\omega_j) + R \max_k p(\omega_j|\vec{x}_k) \right]. \tag{1.10}$$

When the priors are uniform (a frequent case in practical applications), the above boils down to the following:

$$\hat{\omega} = \arg\max_{jk} p(\omega_j|\vec{x}_k), \tag{1.11}$$

also known as the *Maximum Rule*. If the signals are non-redundant, this rule appears to fit the dominance case in Figure 1.2. The response is fully determined by the only signal for which the last equation is satisfied. If the signals are redundant, the rule corresponds to the equivalence scenario when the value of $p(\omega_j|\vec{x}_k)$ is the same for all k. The rule cannot reproduce the enhancement scenario because the response will never be greater than the maximum of $p(\omega_j|\vec{x}_k)$ across all possible j and k values.

Consider the following:

$$\hat{\omega} = \arg\max_j P(\omega_j|\vec{x}_1, \ldots, \vec{x}_R) \alpha p(\vec{x}_1, \ldots, \vec{x}_R|\omega_j) P(\omega_j) = \prod_{k=1}^{R} p(\vec{x}_k|\omega_j), \tag{1.12}$$

where the last step is possible when the priors $P(\omega_j)$ are uniform and the $\vec{x}_k$ are assumed to be statistically independent given ω_j. According to Equation (1.9) the product is bound by the minimum and the decision rule becomes as follows:

$$\hat{\omega} = \max_j \left[\min_k P(\omega_j|\vec{x}_k) \right] \tag{1.13}$$

meaning that the response corresponds to the class fo which the minimum of $P(\omega_j|\vec{x}_k)$ across the various modalities is the highest. In this case as well, the combination rule, known as the *Minimum Rule*, corresponds to the dominance and equivalence scenarii of Figure 1.2. In both cases, it is one of the modalities that the determines the response.

When the priors are uniform, the sum rule:

$$\hat{\omega} = \arg\max_j \left[(1-R)P(\omega_j) + \sum_{k=1}^{R} p(\omega_j|\vec{x}_k) \right] \tag{1.14}$$

boils down to:

$$\hat{\omega} = \arg\max_j \sum_{k=1}^{R} p(\omega_j|\vec{x}_k) = \arg\max_j \frac{1}{R} \sum_{k=1}^{R} p(\omega_j|\vec{x}_k) \tag{1.15}$$

meaning that the response corresponds to the class $\hat{\omega}$ for which the average $P(\hat{\omega}|\vec{x}_k)$ over the modalities $\vec{x}_k$ is the highest. However, given that the number of modalities is typically limited, it is better to use the median to avoid the effect of outliers:

$$\hat{\omega} = \arg\max_j M_k \left[p(\omega_j|\vec{x}_k)\right] \tag{1.16}$$

where $M_k[.]$ is the median over k. The above is called *Median Rule* and the response is determined by one of the modalities, but it is not possible to know a priori which one. In this respect the median rule appears to fit the dominance scenario in the case of non-redundant signals.

One last rule, the *Majority Vote Rule*, can be obtained by hardening the expression of the $P(\omega_j|\vec{x}_i)$ as follows:

$$P(\omega_j|\vec{x}_i) = \Delta_{ji} = \begin{cases} 1 & \text{if } P(\omega_j|\vec{x}_i) = \max_k P(\omega_j|\vec{x}_k) \\ 0 & \text{otherwise} \end{cases} \tag{1.17}$$

In this case, starting from the Sum Rule with the assumption that the priors are uniform, it is possible to write that:

$$\hat{\omega} = \arg\max_j \sum_{k=1}^{R} \Delta_{jk}. \tag{1.18}$$

This means that the response is the class that maximizes $P(\hat{\omega}|\vec{x}_k)$ for the largest number of modalities. In other words, the response corresponds to the class with respect to which the input modalities are most redundant.

1.2.1 Classifier Diversity

Section 1.2 shows the principles underlying early and late fusion. However, it is not sufficient to combine multiple modalities to improve the effectiveness of a multimodal approach. Unlike in human and life sciences, where all observed multimodal communication patterns appear to have some advantages (see Section 1.1), in multimodal Social Signal Processing not all multimodal approaches are effective. In the case of the early fusion, many works show that not all concatenations of features extracted from different modalities lead to performance improvements, e.g., [Ramanarayanan et al. 2015; Subramanian et al. 2013]. Furthermore, feature selection approaches applied to concatenations of unimodal feature vectors do not always retain features coming from all modalities involved. In the case of the late fusion, the combination of classifiers works only when the inputs are *diverse* [Kuncheva and Whitaker

2003]. However, the literature does not provide clear and consensual definitions of the diversity and no approach for its measurement appears to clearly outperform the others: "*despite the popularity of the term diversity, there is no single definition and measure of it [...] none of [its] measures is proven superior to the others and why these diversity measures are useful is still unclear*" [Tang et al. 2006]. Still, the discussions about the topic echo the considerations of life and human sciences about redundancy across signals (see Section 1.1). In particular, the diversity is proposed as a criterion to identify the effective classifiers combinations, i.e., those in which the performance of the combination is higher than the performance of the best individual classifier to a statistically significant extent [Kuncheva and Whitaker 2003; Tang et al. 2006].

The only point that attracts consensus is that the diversity is a trade-off between two conflicting needs. On the one hand, it is intuitive that the performance of an MCS is higher when the performance of the individual classifiers tends to be higher. On the other hand, the performance of the individual classifiers should not be too high because diverse classifiers are expected to give different responses for the same input (otherwise they are not diverse) and this means that a fraction of the classifiers involved in the same MCS should make mistakes. In other words, MCS and, hence, multimodal approaches are useful only when unimodal approaches are in a certain performance range.

A few diversity measures are *pairwise*, i.e., they can be applied to pairs of classifiers (one of the most common cases in the literature). In case an MCS includes more than two classifiers, then the overall diversity is simply the average of all pairwise diversities. When two classifiers are combined and tested over a dataset, they are both right in a fraction $r_1 r_2$ of the cases, both wrong in a fraction $w_1 w_2$ of cases, the first is right while the second is wrong in a fraction $r_1 w_2$ of the cases and viceversa in a fraction $w_1 r_2$ of the cases. The four fractions sum up to 1: $r_1 r_2 + w_1 w_2 + r_1 w_2 + w_1 r_2 = 1$.

The Q statistic is as follows:

$$Q = \frac{r_1 r_2 \cdot w_1 w_2 - r_1 w_2 \cdot w_1 r_2}{r_1 r_2 \cdot w_1 w_2 + r_1 w_2 \cdot w_1 r_2}, \tag{1.19}$$

it ranges between -1 (when at least one between $r_1 r_2$ and $w_1 w_2$ are null) and 1 (when at least one between $r_1 w_2$ and $w_1 r_2$ is null). The diversity is maximum when Q is close to zero, i.e. when the probability of both classifiers giving the same answer or a different answer is roughly the same. In other words, Q tends to be close to zero when the probability of the two classifiers giving a different response is roughly the same as the probability giving the same response. This means that there is equilibrium between the classifiers being redundant (potentially both right) and non-redundant (at least one of them must be wrong).

A similar principle can be observed in another pairwise measure of the diversity, namely the correlation:

$$\rho = \frac{r_1 r_2 \cdot w_1 w_2 - r_1 w_2 \cdot w_1 r_2}{\sqrt{(r_1 r_2 + r_1 w_2)(r_1 r_2 + w_1 r_2)(w_1 w_2 + r_1 w_2)(w_1 w_2 + w_1 r_2)}}. \tag{1.20}$$

It is possible to prove that ρ and Q have always the same sign and that $|\rho| \leq |Q|$. In this case as well, the diversity is maximum when there is equilibrium between the probability of both classifiers giving the same answer and the two classifiers giving a different answer.

When it comes to diversity measures that apply to a whole set of classifiers, the *Kohavi-Wolpert variance* [Kohavi and Wolpert 1996] is one of the most commonly applied. If a MCS includes L classifiers, and $l(\vec{X}) \leq L$ is the number of these that classify correctly an input $\vec{x}$, then the probabilities of the output being correct or incorrect are as follows:

$$\begin{aligned} P(\omega_r|\vec{x}) &= \tfrac{l(\vec{x})}{L} \\ P(\omega_w|\vec{x}) &= \tfrac{L-l(\vec{x})}{L} \end{aligned} \tag{1.21}$$

where ω_r and ω_w correspond to the right and wrong outcome for $\vec{x}$, respectively, and $l(\vec{x})$ is the number of classifiers that give ω_r as response. If a training set contains N patterns $\vec{x}_i$, then the KW variance of the MCS is as follows:

$$kw = \frac{1}{NL^2} \sum_{i=1}^{N} l(\vec{x}_i) \cdot [L - l(\vec{x}_i)]\,. \tag{1.22}$$

If the expression above is maximised with respect a generic $l(\vec{x}_i)$, the result is that the maximum is achieved when $l(\vec{x}_i) = L/2$, i.e., when only half of the classifiers map $\vec{x}_i$ into the right response. In this case as well, the maximum diversity is achieved when there is equilibrium between the probability of a classifier in the MCS being right and the probability of it being wrong.

The literature provides mixed evidence about the relationship between diversity and performance of an MCS [Tang et al. 2006]. The possible explanations are that the current measurements available in the literature are not fully suitable and that the diversity does not depend only on the MCS, but also on the data at disposition. However, there is consensus that the diversity is desirable for an MCS and that there is a difference between diverse and non-diverse ensembles of classifiers. In this respect, the technical literature appears to echo the considerations of human and life sciences about redundancy and non-redundancy in multimodal communication. However, while these latter consider that redundant signals can still be useful in certain cases (see Section 1.1), the technical literature indicates that the lack of diversity tends to make the combination of multiple classifiers unuseful.

Reference	Construct	A	V	K	W	L	FA	M
[Chatterjee et al. 2015]	Passion, Credibility	Y	Y			Y	L	Y
[Ramanarayanan et al. 2015]	Oral Delivery	Y	Y	Y			E	Y
[Curtis et al. 2015]	Oral Delivery	Y	Y				L	Y
[Wörtwein et al. 2015]	Oral Delivery	Y	Y				E	Y
[Nguyen and Gatica-Perez 2015]	Hirability	Y	Y				E	Y
[Siddiquie et al. 2015]	Persuasiveness	Y	Y			Y	EL	Y
[Strohkorb et al. 2015]	Dominance	Y	Y				E	Y
[Salim et al. 2015]	Engagement	Y	Y				E	Y
[Dibeklioğlu et al. 2015]	Depression	Y	Y				E	Y
[Demyanov et al. 2015]	Deception		Y				N	N
[Grafsgaard et al. 2014]	Engagement, Frustration	Y	Y			L	E	Y
[Park et al. 2014]	Persuasiveness	Y	Y				E	Y
[Nihei et al. 2014]	Influence	Y	Y				E	N
[Chen et al. 2014]	Oral Delivery	Y	Y	Y		Y	E	Y
[Ghosh et al. 2014]	Distress	Y	Y			Y	EL	Y
[Vail et al. 2014]	Engagement	Y	Y			Y	E	Y
[Subramanian et al. 2013]	Personality		Y				E	Y
[Aran and Gatica-Perez 2013]	Personality	Y	Y				E	Y
[Mohammadi et al. 2013]	Personality	Y	Y			Y	N	N
[Kalimeri et al. 2013]	Personality	Y			Y		E	Y
[Biel et al. 2013]	Personality	Y	Y			Y	E	Y
[Scherer et al. 2013]	Depression	Y	Y				E	Y
[Batrinca et al. 2012]	Personality	Y	Y				E	Y
[Nakano and Fukuhara 2012]	Dominance	Y	Y				N	Y
[Raiman et al. 2011]	Deception	A	V			Y	E	Y
[Hung and Kröse 2011]	Proxemics		Y				E	Y
[Batrinca et al. 2011]	Personality	Y	Y				E	Y
[Delaherche and Chetouani 2011]	Coordination	Y	Y				N	N

Table 1.1 The table includes the main works on multimodal analysis of social signals between 2011 and 2015. For every article, the table shows whether it uses microphones (A), cameras (V), depth cameras (K), or wearable sensors (W). Furthermore, the table shows whether the fusion is late or early (FA) and whether the use of a multimodal approach improves the performance of the best individual modality (M). Finally, N stands for no and Y stands for yes.

1.2.2 State-of-the-Art in Multimodal Analysis of Social Signals

This section surveys the works on multimodal analysis of social signals that have been presented at the *ACM International Conference on Multimodal Interaction* (ICMI) between 2011 and 2015 included. In particular, the section takes into account works that have been published in the main proceedings of the conference (works presented in workshops, challenges, demo sessions and doctoral consortium have not been taken into account). The *rationale* behind this choice is that the ICMI articles are likely to be representative of the current *state-of-the-art* as well as of the most recent and emerging trends.

Overall, an approach for the multimodal analysis of social signals can be described using the following three dimensions:

- *Extraction of behavioural cues*. This dimension includes both the sensors adopted to record people and the cues actually targeted in the work. The two aspects are considered jointly because the choice of the cues determines the choice of the sensors and, conversely, the choice of the sensors makes certain cues detectable and others not.
- *Inference of social / psychological phenomena or constructs*. This dimension includes the methodology adopted to map the cues into the phenomenon or construct of interest (typically based on machine learning or pattern recognition) and the phenomenon or construct itself. The two aspects are considered jointly because the construct determines the choice of the methodology. In particular, dimensional constructs (e.g., personality traits) require the use of regression techniques while cathegorical constructs (e.g., the Bales social roles) require the use of classifiers.
- *Fusion of multiple modalities*. This dimension corresponds to the methodology adopted to combine the multiple modalities and, in particular, whether the methodology is an *early* or *late* fusion approach.

Table 1.1 provides a synopsis of the articles based on the three dimensions above. Given the wide variety of cues targeted in the various works, the table adopts the sensors as a proxy for the type of behavioural cues that are actually extracted in the works. These are available in Table 1.2 that shows that the cues most commonly detected are prosody (the way people speak) and face or head behaviour (gaze contact, facial expressions and head moevements). Such a distribution confirms the primacy of speech and face in social interactions [Vinciarelli et al. 2009, 2012]. As a consequence, cameras and microphones are the two sensors most commonly adopted. It is important to observe that speech processing and computer vision have developed a large number of approaches that, while having been designed for other tasks, are suitable for detecting verbal and nonverbal behavioural cues [Vinciarelli et al. 2009, 2012]. Thus, the use of microphones and cameras allows one to rely on a solid and extensive body of knowledge.

Reference	Turn-Taking	Prosody	Vocalisations	Verbal	Face and Head	Bodily Movements	Gestures	Proxemics
[Chatterjee et al. 2015]		Y		Y	Y			
[Ramanarayanan et al. 2015]		Y			Y	Y		
[Curtis et al. 2015]		Y			Y	Y		
[Wörtwein et al. 2015]		Y			Y		Y	
[Nguyen and Gatica-Perez 2015]	Y	Y			Y	Y		
[Siddiquie et al. 2015]				Y				
[Strohkorb et al. 2015]	Y				Y		Y	
[Salim et al. 2015]		Y	Y			Y		
[Dibeklioğlu et al. 2015]		Y			Y			
[Demyanov et al. 2015]					Y			
[Grafsgaard et al. 2014]				Y	Y	Y	Y	
[Park et al. 2014]		Y		Y	Y	Y		
[Nihei et al. 2014]		Y			Y			
[Chen et al. 2014]		Y		Y	Y	Y		
[Ghosh et al. 2014]		Y		Y	Y			
[Vail et al. 2014]				Y	Y			
[Subramanian et al. 2013]					Y			Y
[Aran and Gatica-Perez 2013]		Y			Y	Y		
[Mohammadi et al. 2013]								
[Kalimeri et al. 2013]		Y						Y
[Biel et al. 2013]	Y	Y		Y	Y			
[Scherer et al. 2013]		Y			Y	Y		
[Batrinca et al. 2012]	Y	Y				Y		
[Nakano and Fukuhara 2012]	Y				Y			
[Raiman et al. 2011]	Y					Y		
[Hung and Kröse 2011]								
[Batrinca et al. 2011]		Y			Y	Y		
[Delaherche and Chetouani 2011]	Y						Y	

Table 1.2 The table shows the cues most commonly detected in the articles considered in this section.

The construct most frequently addressed in ICMI works is the personality, at least when it comes to multimodal analysis of social signals (7 works out of the 28 considered in this section). The main probable reason is that personality is predictive of a large number of life aspects, including the ability to establish statisfactory relationships, professional success, tendency to criminal or antisocial behaviour, etc. [Ozer and Benet-Martinez 2006]. In this respect, personality appears to be an important target for any technology expected to interact with people. Furthermore, the most important models available in the Psychology literature represent personality as a vector in a low-dimensional space, a representation particularly suitable for computer processing.

When it comes to the combination approach, early fusion appears to be by far the most common and effective approach (22 works out of the 28 considered). This might be surprising because, unlike the case of late fusion (see Section 1.2), the literature does not provide a theoretic framework explaining why such a methodology works and how. Furthermore, the concatenation of multiple $\vec{x}_i$ often leads to high-dimensional feature vectors that require the application of dimensionality reduction techniques. A possible explanation is that late fusion requires classifiers that estimate $p(\omega_k|\vec{x}_i)$, but some of the most effective classification approaches give as output scores that cannot be interpreted as probabilities (e.g., the Support Vector Machines). This can make it difficult to apply late fusion approaches.

1.3 Next Steps

Early and late fusion are the established state-of-the-art in the analysis of multimodal social signals, i.e., the two methodologies that are applied in the large majority of the works presented in the literature (see above). Both approaches are known to work well, but more sophisticated approaches are emerging that promise to better account for the properties of multimodal data and, in particular, to better exploit the relationships between multiple modalities, whether these correspond to redundancy or complementarity in conveying a given social signal. The extensive survey proposed by [Baltrusaitis et al. 2017] identifies five major issues when it comes to the analysis of multimodal data, namely *representation* ("*how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities*", according to the definition provided in the survey), *translation* ("*how to translate (map) data from one modality the other*" in the words of the survey), *alignment* ("*to identify the direct relations between between (sub)elements from two or more different modalities*" following the description of the survey), *fusion* ("*to join information from two or more modalities to perform a prediction*" in the survey's terms) and *co-learning* ("*transfer knowledge between modalities*" in the definition available in the survey).

The two aspects more relevant to this chapter are representation and fusion (not to be confused with the meaning that the latter word has in the expressions "early" and "late fusion"), the two main issues that have to be addressed in analyzing automatically social

signals. For what concerns the representation, the survey by [Baltrusaitis et al. 2017] proposes a distinction between *joint* and *coordinated* representations. In the first case, the multimodal representation $\vec{y}$ can be expressed as a function of unimodal representations $\vec{x}_i$, where $i = 1, \ldots, R$ and R is total number of modalities involved: $\vec{y} = f(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_R)$. The early fusion described earlier in the chapter - the simple concatenation of the unimodal feature vectors - is a simple form of joint representation. In the second case - the coordinated representation - the approach is to use an individual mapping function $f_m(\vec{x}_m)$ for every individual unimodal vector with the constraint that $f_i(\vec{x}_i) \sim f_j(\vec{x}_j)$, where the symbol $\sim$ accounts for a relationship like, e.g., the maximization of the correlation. For what concerns the fusion, the survey describes a large number of approaches that range from graphical models - joint probability distributions defined over sets of random variables - to deep neural networks (the interested reader can refer to [Baltrusaitis et al. 2017] for more details). The adoption of these most recent approaches promises to change the way multimodal analysis of social signals has been performed so far and, hopefully, to improve the performances achieved so far.

Another important issue that, to the best of our knowledge, has been addressed only to a limited extent is whether it is possible to detect constellations of nonverbal behavioral cues through unsupervised approaches, i.e., whether it is possible to understand the way people convey social and psychological information by combining multiple cues - typically belonging to different modalities - into one individual social signal. One possible approach to address such a task is the application of topic models or co-clustering approaches capable to detect and model frequent associations of multiple cues. Topic models (see a recent survey by [Blei 2012]) have been developed for texts, but they can be applied to any type of data that can be represented with vectors where every component accounts for how frequently a given *word* (a cue in the case of behavior) has been observed. Co-clustering "is the simultaneous clustering of both points and their attributes" [Berkhin 2006] and it can help in addressing those application domains where people with different characteristics are expected to display different behavioral cues. A typical example is psychiatry, where pathologies are often represented in terms of *Homeostatic Property Clusters* [Zachar 2014], i.e., groups of individuals that tend to manifest the same behavioral symptoms (e.g., the tendency to avoid direct gaze in the case of autistic individuals). The main advantage of unsupervised approaches in these cases is that they might be capable to detect combinations of cues that are not easy to grasp for human observers (in the case of psychiatry, clinical observation is still the main approach towards the identification of the behavioral cues that characterize a pathology).

A last important issue to take into account is the explanatory power of the approaches adopted in the multimodal analysis of social signals. As it is a multidisciplinary field, Social Signal Processing does not simply target the performance of an approach - e.g., in terms of correct percentage of times a given social signal is interpreted correctly - it also tries to

provide an insight about the way a given social signal is used and why. In particular, Social Signal Processing tries to provide human sciences researchers with alternative approaches to formulate their findings about the way people behave. One common approach is to use feature selection approaches to identify the features that perform better in inferring a given psychological phenomenon from behavioral observations. Compared with the methodologies that the psychologists adopt, such an approach considers subsets of features rather than individual features and, hence, it better accounts for interactions between multiple features. Another approach is to use the parameters of the models to obtain insights about what are the cues other low-level behavioral observations that better explain the outcome of a classification and, indirectly, better explain the phenomenon that is the target of the classification.

1.4 Conclusions

This chapter revolves around multimodal communication between living beings and shows how such a phenomenon is addressed in life sciences and computing technology. Section 1.1 shows how life sciences have organised the multimodal communication patterns observed in nature into a coherent taxonomy. Section 1.2 introduces early and late fusion, the two main approaches aimed at dealing with multimodal stimuli. In addition, the section includes a brief meta-analysis of the works on multimodal analysis of social signals that have been presented at the ACM International Conference on Multimodal Interaction between 2011 and 2015. Finally, Section 1.3 outlines some future perspectives.

Studies on multimodal communication in life sciences and computing have been conducted, presumably, with limited or no mutual influence at all. However, there are interesting parallelisms between the two areas. The first is that both areas focus on the response that a system gives to multiple stimuli known to account for the same input information. The second is that both areas recognize the redundancy across multiple modalities as a crucial issue and as a criterion to distinguish between different cases. The main difference is that nature, the subject of life sciences, appears to take advantage of all possible patterns of multimodal communication - whether they involve redundancy across modalities or not and whether the response is different from the one obtained with individual modalities or not - while technology appears to benefit only of those cases where the redundancy is limited and the most desirable response is enhanced with respect to the use of individual modalities. However, it is important to observe that parallelisms and similarities do not mean that the methodologies dealing with multimodal data are a model or an explanation of the way living beings deal with multisensory inputs.

An interesting aspect of computing technologies is the need to express every process in operational terms, i.e., in terms suitable for the implementation of a computer program. This has led to the development of the computational frameworks described in Section 1.2 and, more extensively, in [Kittler et al. 1998]. Furthermore, it has led to the definition of the *diversity*, a property that a multimodal approach is expected to have in order to work

properly. Intuitively, the diversity is a measure of the lack of redundancy across modalities and this leads to an interesting difference between multimodality in nature and multimodality in machines. In the former case, redundancy can still be useful when one of the modalities can face obstacles in conveying its message (e.g., acoustic signals in a noisy environment). In the latter case, redundant modalities are never useful unless it is possible to know which one should be trusted in case the others fail due to noise or other technical problems. Last, but not least, a multimodal approach is considered to be successful when its performance is higher, to a statistically significant extent, than the performance achieved with the best individual modality. This means that, unlike in nature, if there is a modality that performs above a certain threshold, multimodality becomes unnecessary. In other words, multimodality appears to be an advantage only when individual modalities do not work well enough.

The meta-analysis of the works presented at ICMI between 2011 and 2015 shows that the large majority of the proposed approaches adotps an early fusion approach. In most cases, multimodal approaches based on different modalities lead to an improvement with respect to the best individual modality. In other words, multimodal approaches appear to be particularly suitable for the analysis of social signals that, both for their inherent ambiguity and the technical difficulties involved in their detection, cannot be detected and interpreted individually with high performance.

Bibliography

O. Aran and D. Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 11–18, 2013.

T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. Technical report, arXiv, 2017. URL http://arxiv.org/abs/1705.09406.

L. Batrinca, B. Lepri, N. Mana, and F. Pianesi. Multimodal recognition of personality traits in human-computer collaborative tasks. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 39–46, 2012.

L.M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: Automatic personality assessment using short self-presentations. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 255–262, 2011.

P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping multidimensional data*, pages 25–72. Springer Verlag, 2006.

J.-I.-Isaac Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez. Hi youtube!: Personality impressions and verbal content in social video. In *Proceedings of ACM International Conference on Multimodal Interaction*, pages 119–126, 2013.

D.M. Blei. Probabilistic topic models. *Communications of ACM*, 55(4):77–84, 2012.

P. Brunet and R. Cowie. Towards a conceptual framework of research on Social Signal Processing. *Journal of Multimodal User Interfaces*, 6(3-4):101–115, 2012.

M. Chatterjee, S. Park, L.-P. Morency, and S. Scherer. Combining two perspectives on classifying multimodal data for recognizing speaker traits. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 7–14, 2015.

L. Chen, G. Feng, J. Joe, C.W. Wee Leong, C. Kitchen, and C.M. Lee. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 200–203, 2014.

K. Curtis, G.J.F. Jones, and N. Campbell. Effects of good speaking techniques on audience engagement. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 35–42, 2015.

C. Darwin. *The expression of emotion in animals and man*. John Murray, 1872.

E. Delaherche and M. Chetouani. Characterization of coordination in an imitation task: Human evaluation and automatically computable cues. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 343–350, 2011.

S. Demyanov, J. Bailey, K. Ramamohanarao, and C. Leckie. Detection of deception in the mafia party game. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 335–342, 2015.

H. Dibeklioğlu, Z. Hammal, Y. Yang, and J.F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages

307–310, 2015.

S. Ghosh, M. Chatterjee, and L.-P. Morency. A multimodal context-based approach for distress assessment. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 240–246, 2014.

J.F. Grafsgaard, J.B. Wiggins, A.K. Vail, K.E. Boyer, E.N. Wiebe, and J.C. Lester. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 42–49, 2014.

H. Hung and B. Kröse. Detecting F-formations as dominant sets. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 231–238, 2011.

K. Kalimeri, B. Lepri, and F. Pianesi. Going beyond traits: Multimodal classification of personality states in the wild. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 27–34, 2013.

J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

R. Kohavi and D. Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the International Conference on Machine Learning*, pages 275–83, 1996.

L.I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

M. Mehu and K. Scherer. A psycho-ethological approach to Social Signal Processing. *Cognitive Processing*, 13(2):397–414, 2012.

G. Mohammadi, S. Park, K. Sagae, A. Vinciarelli, and L.-P. Morency. Who is persuasive?: The role of perceived personality and communication modality in social multimedia. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 19–26, 2013.

Y. Nakano and Y. Fukuhara. Estimating conversational dominance in multiparty interaction. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 77–84, 2012.

L.S. Nguyen and D. Gatica-Perez. I would hire you in a minute: Thin slices of nonverbal behavior in job interviews. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 51–58, 2015.

F. Nihei, Y.I. Nakano, Y. Hayashi, H.-H. Hung, and S. Okada. Predicting influential statements in group discussions using speech and head motion information. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 136–143, 2014.

D.J. Ozer and V. Benet-Martinez. Personality and the prediction of consequential outcomes. *Annual Reviews of Psychology*, 57:401–421, 2006.

S. Park, H.S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 50–57, 2014.

S.R. Partan and P. Marler. Communication goes multimodal. *Science*, 283(5406):1272–1273, 1999.

S.R. Partan and P. Marler. Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2):231–245, 2005.

I. Poggi. *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, 2007.

I. Poggi and F. D'Errico. Social Signals: a framework in terms of goals and beliefs. *Cognitive Processing*, 13(2):427–445, 2012.

N. Raiman, H. Hung, and G. Englebienne. Move, and i will tell you who you are: Detecting deceptive roles in low-quality data. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 201–204, 2011.

V. Ramanarayanan, C.W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 23–30, 2015.

C. Rowe and T. Guilford. Hidden colour aversions in domestic chicks triggered by pyrazine odours of insect warning displays. *Nature*, 383(6600):520–522, 1996.

F.A. Salim, F. Haider, O. Conlan, S. Luz, and N. Campbell. Analyzing multimodality of video for user engagement assessment. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 287–290, 2015.

S.J. Scheffer, G.W. Uetz, and G.E. Stratton. Sexual selection, male morphology, and the efficacy of courtship sgnalling in two wolf spiders (araneae: Lycosidae). *Behavioral Ecology and Sociobiology*, 38(1):17–23, 1996.

S. Scherer, G. Stratou, and L.-P. Morency. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 135–140, 2013.

B. Siddiquie, D. Chisholm, and A. Divakaran. Exploiting multimodal affect and semantics to identify politically persuasive web videos. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 203–210, 2015.

S. Strohkorb, I. Leite, N. Warren, and B. Scassellati. Classification of children's social dominance in group interactions with robots. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 227–234, 2015.

R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 3–10, 2013.

E.K. Tang, P.N. Sugnathan, and X. Yao. An analysis of diversity measures. *Machine Learning*, 65(5): 247–271, 2006.

A.K. Vail, J.F. Grafsgaard, J.B. Wiggins, J.C. Lester, and K.E. Boyer. Predicting learning and engagement in tutorial dialogue: A personality-based model. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 255–262, 2014.

A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of Social Signal Processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer. Multimodal public speaking performance assessment. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 43–50, 2015.

P. Zachar. Beyond natural kinds: Toward a “relevant” “scientific” taxonomy in psychiatry. In H. Kincaid and J.A. Sullivan, editors, *Classifying Psychopathology*, pages 75–104. MIT Press, 2014.