CHAPTER 15

Multimodal Fusion in Human-Agent Dialogue

Elisabeth André, Jean-Claude Martin, Florian Lingenfelser and Johannes Wagner

1. Introduction

Sophisticated fusion techniques are an essential component of any multimodal system. Historically, systems aimed at analyzing the semantics of multimodal commands and typically investigated a combination of pointing and drawing gestures and speech. The most prominent example includes the "Put-that-there" system (Bolt, 1980) that analyzes speech in combination with 3D pointing gestures referring to objects on a graphical display. Since this groundbreaking work, numerous researchers have investigated mechanisms for multimodal input interpretation mainly working on speech, gestures and gaze while the trend is moving towards intuitive interactions in everyday environments. Since interaction occurs more and more in mobile and tangible environments, modern multimodal interfaces require a greater amount of context-awareness (Johnston et al., 2011).

At the same time, we can observe a shift from pure task-based dialogue to more human-like dialogues that aim to create social experiences. Usually, such dialogue systems rely on a personification of the user interface by means of embodied conversational agents or social robots. The driving force behind this work is the insight that a user interface is more likely to be accepted by the user if the machine

is sensitive towards the user's feelings. For example, Martinovsky and Traum (2003) demonstrated by means of user dialogues with a training system and a telephone-based information system that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively. This observation shows that a system should not only analyze what the user said or gestured but also consider more subtle cues, such as psychological user states.

With the departure from pure task-based dialogue to more human-like dialogues that aim to create social experiences, the concept of multimodal fusion as originally known in the natural language community has to be extended. We do not only need fusion mechanisms that derive information on the user's intention from multiple modalities, such as speech, pointing gestures and eye gaze. In addition, fusion techniques are required that help a system assess how the user perceives the interaction with it. Accordingly, fusion mechanisms are required not only at the semantic level, but also at the level of social and emotional signals. With such systems, any interaction may indeed feature a task-based component mixed with a social interaction component. These different components may even be conveyed on different modalities and overlap in time. It is thus necessary to integrate a deeper semantic analysis in social signal processing on the one side and to consider social and emotional cues in semantic fusion mechanisms on the other side. Both streams of information need to be closely coupled during fusion since they can both include similar communication channels. For example, a system may fuse verbal and nonverbal signals to come up with a semantic interpretation, but the same means of expression may also be integrated by a fusion mechanism as an indicator of cognitive load (Chen et al., 2012).

By providing a comparative analysis of semantic fusion of multimodal utterance and fusion of social signals, this chapter aims to give a comprehensive overview of fusion techniques as components of dialogue system that aim to emulate qualities of human-like communication. In the next section, we first present taxonomies for categorizing fusion techniques focusing on the relationship between the single modalities and the level of integration. Section 3 addresses the fusion of semantic information, whereas Section 4 is devoted to the fusion of social signals. To enable a better comparison of issues handled in the two areas, both sections follow a similar structure. We first introduce techniques for fusing information at different levels of abstraction and discuss attempts to come up with standards to represent information to be exchanged in fusion engines. After that we discuss challenges that arise when moving from controlled laboratory

environments to less controlled everyday scenarios. Particular attention is given to the use of fusion mechanisms in human-agent dialogue where the mechanisms for input analysis have to be tightly coordinated with appropriate feedback to be given by the agent. Section 5 presents approaches that combine semantic interpretation with social cue analysis either to increase the robustness of the analysis components or to improve the quality of interaction. Section 6 concludes the chapter and gives an outline for future research.

2. Dimensions of Description

Most systems rely on different components for the low-level analysis of the single modalities, such as eye trackers, speech and gesture recognizers, and make use of one or several modality integrators to come up with a comprehensive interpretation of the multimodal input. In this context, two fundamental questions arise: How are the single modalities related to each other and at which level should they be integrated?

2.1 Relationships between modalities

Several combinations of modalities may cooperate in different manners. Martin et al. (1998) mention the following cases: equivalence, redundancy, complementarity, specialization and transfer.

When several modalities cooperate by *equivalence*, this means that a command or a chunk of information may be produced as an alternative, by either of them. For example, to consider the needs of a variety of users, a multimodal interface might allow them to specify a command via speech or as an alternative by pressing a button.

Modalities that cooperate by *redundancy* produce the same information. Redundancy allows a system to ignore one of the two redundant modalities. For example, when a user uniquely specifies a referent via speech and uses at the same time an unambiguous pointing gesture, only one modality needs to be considered to uniquely identify the referent.

When modalities cooperate by *complementarity*, different chunks of information are produced by each modality and have to be integrated during the interpretation process. A classical example includes a multimodal command consisting of a spoken utterance and a pointing gesture both of which contribute to the interpretation.

When modalities cooperate by *specialization*, this means that one modality provides the frame of interpretation for another. Specialization occurs, for example, when a user points to a group of

objects and specifies the intended referent by verbally providing a category that distinguishes it from alternatives.

Cooperation by *transfer* means that a chunk of information produced by one modality is used by another modality. Transfer is typically used in hypermedia interfaces when a mouse click triggers the display of an image.

Different modalities may also be used *concurrently*, i.e. produce independent chunks of information, i.e. chunks without any semantic overlap, at the same time. For example, a user may say "Hello" and at the same time point to an object. Here the chunks of information should not be merged. Earlier systems usually did not allow for a concurrent use of modalities, but required an *exclusive* use of modalities. For example, the user may utter a greeting and point to an object, but not at the same time.

While the relationship between modalities has mainly been discussed for multimodal user commands, little attempts have been made to specify the relationship between social signals. However, modalities that convey social signals may cooperate in a similar manner as modalities that convey semantic information. For example, different dimensions of emotions, such as valence and arousal, may be expressed by different channels of communication, such as the face or the voice. It is important to note, however, that it is hard to deliberately employ given channels of communication for the expression of social signals.

Different modalities do not always convey *congruent* pieces of information. In the case of semantic information, little robust input processing components typically lead to incongruent pieces of information. In the case of social signals, incongruent pieces of information often result from the fact that users are not equally expressive in all modalities. In particular, the attempt to conceal social signals may result into an inconsistent behavior.

Another classification concerns the timing of modalities. Here, we may basically distinguish between the *sequential* use of modalities and the *parallel* use of modalities which overlap in time. Semantically related modalities may overlap in time or may be used in sequence. If they are merged, the temporal distance should, however, not be too large. Algorithms for the fusion of social signals usually start from the assumption that social signals that refer to particular user state, such as frustration, emerge at exactly the same time interval. We will later see that such an assumption may be problematic.

Other researchers use similar terms to describe relationships between modalities. See Lalanne et al. (2009) for an overview.

2.2 Levels of integration

Basically, two main fusion architectures have been proposed in the literature depending on at which level sensor data are fused.

In the case of *low-level fusion*, the input from different sensors is integrated at an early stage of processing. Low-level fusion is therefore often also called *early fusion*. The fusion input may consist of either raw data or low-level features, such as pitch. The advantage of low-level fusion is that it enables a tight integration of modalities. There is, however, no declarative representation of the relationship between various sensor data which aggravates the interpretation of recognition results.

In the case of *high-level fusion*, low-level input has to pass modality-specific analyzers before it is integrated, e.g. by summing recognition probabilities to derive a final decision. High-level fusion occurs at a later stage of processing and is therefore often also called *late fusion*. The advantage of high-level fusion is that it allows for the definition of declarative rules to combine the interpreted results of various sensors. There is, however, the danger that information goes lost because of a too early abstraction process.

3. Multimodal Interfaces Featuring Semantic Fusion

In this section, we focus on semantic fusion that combines the meaning of the single modalities into a uniform representation.

3.1 Techniques for semantic fusion

Systems aiming at a semantic interpretation of multimodal input typically use a late fusion approach at a decision level and process each modality individually before fusion (see Figure 1a). Usually, they rely on mechanisms that have been originally introduced for the analysis of natural language.

Johnston (1998) proposed an approach to modality integration for the QuickSet system that was based on unification over typed feature structures. The basic idea was to build up a common semantic representation of the multimodal input by unifying feature structures which represented the semantic contributions of the single modalities. For instance, the system was able to derive a partial interpretation for a spoken natural language reference which indicated that the location of the referent was of type “point”’. In this case, only unification with gestures of type “point”’ would succeed.

Kaiser et al. (2003) applied unification over typed feature structures to analyze multimodal input consisting of speech, 3D gestures and head direction in augmented and virtual reality. Noteworthy is the fact that the system went beyond gestures referring to objects, but also considered gestures describing how actions should be performed. Among others, the system was able to interpret multimodal rotation commands, such as "Turn the table <rotation gesture> clockwise." where the gesture specified both the object to be manipulated and the direction of rotation.

Another popular approach that was inspired by work on natural language analysis used finite-state machines consisting of $n + 1$ tapes which represent the n input modalities to be analyzed and their combined meaning (Bangalore and Johnston, 2009). When analyzing a multimodal utterance, lattices that correspond to possible interpretations of the single input streams are created by writing symbols on the corresponding tapes. Multiple input streams are then aligned by transforming their lattices into a lattice that represents the combined semantic interpretation. Temporal constraints are not explicitly encoded as in the unification-based approaches described above, but implicitly given by the order of the symbols written on the single tapes. Approaches to represent temporal constraints within state chart mechanisms have been presented by Latoschik (2002) and more recently by Mehlmann and André (2012).

3.2 Semantic representation of fusion input

A fundamental problem of the very early systems was that there was no declarative formalism for the formulation of integration constraints. A noteworthy exception was the approach used in QuickSet which clearly separated the statements of the multimedia grammar from the mechanisms of parsing (Johnston, 1998). This approach enabled not only the declarative formulation of type constraints, such as "the location of a flood zone should be an area", but also the specification of spatial and temporal constraints, such as "two regions should be a limited distance apart" and "the time of speech must either overlap with or start within four seconds of the time of the gesture".

Many recent multimodal input systems, such as SmartKom (Wahlster 2003), make use of an XML language for representing messages exchanged between software modules. An attempt to standardize such a representation language has been made by the World Wide Web Consortium (W3C) with EMMA (Extensible MultiModal Annotation markup language). It enables the representation of characteristic features of the fusion process: "composite" information

(resulting from the fusion of several modalities), confidence scores, timestamps as well as incompatible interpretations ("one-of"). Johnston (2009) presents a variety of multimodal interfaces combining speech-, touch- and pen-based input that have been developed using the EMMA standard.

3.3 Choice of segments to be considered in the fusion process

Most systems start from the assumption that the complete input provided by the user can be integrated. Furthermore, they presume that the start and end points of input in each modality are given, for example, by requiring the user to explicitly mark it in the interaction. Under such conditions, the determination of processing units to be considered in the fusion process is rather straightforward. Typically, temporal constraints are considered to find the best candidates to be fused with each other. For example, a pointing gesture should occur approximately at the same time as the corresponding natural language expression while it is not necessary that the two modalities temporally overlap. However, there are cases when such an assumption is problematic and may present a system from deriving a semantic interpretation. For example, the input components may by mistake come up with an erroneous recognition result that cannot be integrated. Secondly, the user may unintentionally provide input, for example, by making a gesture that should not be taken as a gesture. In natural environments where users freely interact, the situation becomes even harder. Users permanently move their arms, but not every gesture is meant to be part of a system command. If eye gaze is employed as a means to indicate a referent, the determination of segments becomes even challenging. Users tend to fixate the objects with the eye they refer to. However, not every fixation is supposed to contribute to a referring expression. A first approach to solve this problem has been presented by Sun et al. (2009). They propose a multimodal input fusion approach to flexibly skip spare information in multimodal inputs that cannot be integrated.

3.4 Dealing with imperfect data in the fusion process

Multimodal interfaces often have to deal with uncertain data. Individual signals may be noisy and/or hard to interpret. Some modalities may be more problematic than others. A fusion mechanism should consider these uncertainties when integrating the modalities into a common semantic representation.

Usually, multimodal input systems combine several *n*-best hypotheses produced by multiple modality-specific generators. This leads to several possibilities of fusion, each with a score computed as a weighted sum of the recognition scores provided by individual modalities. In this vein, it may happen that a badly ranked hypothesis may still contribute to the overall semantic representation because it is compatible with other hypotheses. Thus, multimodality enables us to use the strength of one modality to compensate for weaknesses of others. For example, errors in speech recognition can be compensated by gesture recognition and vice versa. Oviatt (1999) reported that 12.5% of pen/voice interactions in QuickSet could be successfully analyzed due to multimodal disambiguation while Kaiser et al. (2003) even obtained a success rate of 46.4% for speech and 3D gestures that could be attributed to multimodal disambiguation.

3.5 Desktop vs. mobile environments

More recent work focuses on the challenge to support a speech-based multimodal interface on heterogeneous devices including not only desktop PCs, but also mobile devices, such as smart phones (Johnston, 2009).

In addition, there is a trend towards less traditional platforms, such as in-car interfaces (Gruenstein et al., 2009) or home controlling interfaces (Dimitriadis and Schroeter, 2011). Such environments raise particular challenges to multimodal analysis due to the increased noise level, the less controlled environment and multi-threaded conversations. In addition, we need to consider that users are continuously producing multimodal output and not only when interacting with a system. For example, a gesture performed by a user to greet another user should not be mixed up with a gesture to control a system. In order to relieve the users from the burden to explicitly indicate when they wish to interact, a system should be able to distinguish automatically between commands and non-commands.

Particular challenges arise in a situated environment because the information on the user's physical context is required to interpret a multimodal utterance. For example, a robot has to know its location and orientation as well as the location of objects in its physical environment, to execute commands, such as "Move to the table". In a mobile application, the GPS location of the device may be used to constrain search results for a natural language user query. When a user says "restaurants" without specifying an area on the map displayed on the phone, the system interprets this utterance as a request to provide only restaurants in the user's immediate vicinity. Such an approach

is used, for instance, by Johnston et al. (2011) in the MTalk system, a multimodal browser for location-based services.

3.6 Semantic fusion in human-agent interaction

A number of multimodal dialogue systems make use of a virtual agent in order to allow for more natural interaction. Typically, these systems employ graphical displays to which a user may refer to using touch or mouse gestures in combination with spoken or written natural language input; for example, Martin et al. (2006), Wahlster (2003) or Hofs et al. (2010). Furthermore, the use of freehand arm gestures (Sowa et al., 2001) and eye gaze (Sun et al., 2008) to refer to objects in a 3D environment has been explored in interactions with virtual agents. Techniques for multimodal semantic fusion have also attracted interest in the area of human-robot interaction. In most systems, the user's hands are tracked to determine objects or locations the user is referring to via natural language; for example, Burger et al. (2011). In addition to the recognition of hand gestures, Stiefelhagen et al. (2004) make use of head tracking based on the consideration that users typically look at the objects they refer to.

While some of the agent-based dialogue systems employ unification-based grammars (Wahlster, 2003) or chart starts (Sowa et al., 2001) as presented in Section 3.1, others use a hybrid fusion mechanism combining declarative formalisms, such as frames, with procedural elements (Martin et al., 2006). Often the fusion of semantic information is triggered by natural language components which detect a need to integrate information from another modality (Stiefelhagen et al., 2004).

In addition, attempts have been made to consider how multimodal information is analyzed and produced by humans in the semantic fusion process. Usually what is being said becomes not immediately clear, but requires multiple turns between two interlocutors. Furthermore, people typically analyze speech in an incremental manner while it is spoken and provide feedback to the speaker before the utterance is completed. For example, a listener may signal by a frown that an utterance is not fully understood. To simulate such a behavior in human-agent interaction, a tight coupling of multimodal analysis, dialogue processing and multimodal generation is required. Stiefelhagen et al. (2007) propose to allow for clarification dialogues in order to improve the accuracy of the fusion process in human-robot dialogue. Visser et al. (2012) describe an incremental model of grounding that enables the simulation of several grounding acts, such as initiate, acknowledge, request and repair, in human-agent dialogue. If the virtual agent is not able to come up with a meaning for the user's input, it generates

an appropriate feedback signal, such as a frown, to encourage more information from the user. As a consequence, the fusion process in this system may extend over a sequence of turns in a multimodal dialogue.

4. Multimodal Interfaces Featuring Fusion of Social Signals

Recently, the automatic recognition of social and emotional cues has shifted from a side issue to a major topic in human-computer interaction. The aim is to enable a very natural form of interaction by considering not only explicit instructions by human users, but also more subtle cues, such as psychological user states. A number of approaches to automated affect recognition have been developed exploiting a variety of modalities including speech (Vogt and André, 2005), facial expressions (Sandbach et al., 2012), body postures and gestures (Kleinsmith et al., 2011) as well as physiological measurements (Kim and André, 2008). Also, multimodal approaches to improve emotion recognition accuracy are reported, mostly by exploiting audiovisual combinations. Results suggest that integrated information from audio and video leads to improved classification reliability compared to a single modality—even with fairly simple fusion methods.

In this section, we will discuss applications with virtual humans and social robots that make use of mechanisms for fusing social and emotional signals. We will start off by discussing a number of design decisions that have to be made for the development of such systems.

4.1 Techniques for fusing social signals

Automatic sensing of emotional signals in real-time systems usually follows a machine learning approach and relies on an extensive set of labeled multimodal data. Typically, such data are recorded in separate sessions during which users are asked to show certain actions or interact with a system that has been manipulated to induce the desired behavior. Afterward, the collected data is manually labeled by human annotators with the assumed user emotions. Thus, a huge amount of labeled data is collected for which classifiers are trained and tested. An obvious approach to improve the robustness of the classifiers is the integration of data from multiple channels. Hence, an important decision to take concerns the level at which the single modalities should be fused.

A straightforward approach is to simply merge the features calculated from each modality into one cumulative structure, extract

the most relevant features and train a classifier with the resulting feature set. Hence, fusion is based on the integration of low-level features at the feature level (see Figure 1b) and takes place at a rather early stage of the recognition process.

An alternative would be to fuse the recognition results at the decision level based on the outputs of separate unimodal classifiers (see Figure 1c). Here, multiple unimodal classifiers are trained for each modality individually and the resulting decisions are fused by using specific weighting rules. In the case of emotion recognition, the input for the fusion algorithm may consist of either discrete emotion categories, such as anger or joy, or continuous values of a dimensional emotion model (e.g. continuous representation of the valence or the arousal of the emotions). Hence, fusion is based on the integration of high-level concepts and takes place at a later stage of the recognition process.

Eyben et al. (2011) propose a mechanism that fuses audiovisual social behaviors at an intermediate level based on the consideration that behavioral events, such as smiles, head shakes and laughter, convey important information on a person's emotional state that might go lost if information is fused at the level of low-level features or at the level of emotional states.

Which level of modality integration yields the best results is usually hard to predict. Busso et al. (2004) report on an emotion-specific comparison of feature-level and decision-level fusion that was conducted for an audiovisual database containing four emotions, sadness, anger, happiness, and neutral state, deliberately posed by an actress. They observed for their corpus that feature-level fusion was most suitable for differentiating anger and neutral while decision-level fusion performed better for happiness and sadness. Caridakis et al. (2007) presented a multimodal approach for the recognition of eight emotions that integrated information from facial expressions, body gestures, and speech. They observed a recognition improvement of more than 10% compared to the most successful unimodal system and the superiority of feature-level fusion to decision-level fusion. Wagner et al. (2011a) tested a comprehensive repertoire of state-of-the-art fusion techniques including their own emotion-specific fusion scheme on the acted DaFEx corpus and the more natural CALLAS corpus. Results were either considerably improved (DaFEx) or at least in line with the dominating modality (CALLAS). Unlike Caridakis and colleagues, Wagner and colleagues found that decision-level fusion yielded more promising results than feature-level fusion.

W3C EmotionML (Emotion Markup Language) has been proposed as a technology to represent and process emotion-related data and to enable the interoperability of components dedicated to emotion-oriented computing. An attempt towards a language that is not limited to the representation of emotion-related data, but directed to the representation of nonverbal behavior in general has been recently made by Scherer et al. (2012) with PML (Perception Markup Language). As in the case of semantic fusion, the authors identified a specific need to represent uncertainties in the interpretation of data. For example, a gaze away from the interlocutor may signal a moment of high concentration, but also be an indicator of disengagement.

4.2 Acted versus spontaneous signals

Most emotion recognition systems still rely exclusively on acted data for which very promising results have been obtained. The way emotions are expressed by actors may be called prototypical, and independent observers would largely agree on the emotional state of these speakers. A common example includes voice data from actors for which developers of emotion recognition systems reported accuracy rates of over 80% for seven emotion classes. In realistic applications, there is, however, no guarantee that emotions are expressed in a prototypical manner. As a consequence, these applications still represent a great challenge for current emotion recognition systems, and it is obvious to investigate whether the recognition rates obtained for unimodal non-acted data can be improved by considering multiple modalities.

Unfortunately, the gain obtained by multimodal fusions seems to be lower for non-acted than for acted data. Based on a comprehensive analysis of state-of-the-art approaches to affect recognition, D'Mello and Kory (2012) report on an average improvement of 8.12% for multimodal affect recognition compared to unimodal affect recognition while the improvement was significantly higher for acted data (12.1%) than for spontaneous data (4.39%).

One explanation might be that experienced actors are usually able to express emotions consistently across various channels while natural speakers do not have this capacity. For example, Wagner et al. (2011a) found that natural speakers they recorded for the CALLAS corpus were more expressive in their speech than in their face or gestures—probably due to the fact that the method they used to elicit emotions in people mainly affected vocal emotions. As a consequence, they did not obtain a high gain for multimodal fusion compared to the unimodal speech-based emotion classifier. At least, they were able

to handle disagreeing modalities in a way so that competitive results to the best channel could be achieved.

4.3 Offline versus online fusion

Characteristic of current research on the multimodal analysis of social and emotional signals is the strong concentration on posteriori analyses. Out of the many methods discussed in the recent analysis by D'Mello and Kory (2012), hardly any one of them was tested in an online scenario where a system responds to users' social and emotional signals while they are interacting with it. The move from offline to online analysis of social and affective cues raises a number of challenges for the multimodal recognition task. While in offline analysis the whole signal is available and analysis can fall back on global statistics, such a treatment is no longer possible for online analysis. In addition, offline analysis usually focuses on a small set of pre-defined emotion classes and neglects, for example, data that could not be uniquely assigned to a particular emotion class. Online analysis has, however, to take into account all emotion data. Finally, while there are usually no temporal restrictions for offline analysis, online analysis has to be very fast to enable a fluent human-robot dialogue. A fusion mechanism specifically adapted to the needs of online fusion has been used in the Callas Emotional Tree, an artistic Augmented Reality installation of a tree which responds to the spectators' spontaneous emotions reactions to it; see Gilroy et al. (2008). The basic idea of this approach is to derive emotional information from different modality-specific sensors and map it onto the 3D of the Pleasure-Arousal-Dominance model (PAD model) by Mehrabian (1980). In addition, to the input provided by a modality-specific sensor at a particular instance of time, the approach considers the temporal dynamics of modality-specific emotions by integrating the current value provided by a sensor with the previous value. The fusion vector then results from a combination of the vectors representing the single modality-specific contributions. Unlike traditional approaches to sensor fusion, PAD-based fusion integrates contributions from the single modalities in a frame-wise fashion and is thus able to respond immediately to a user's emotional state.

4.4 Choice of segments to be considered in the fusion process

Even though it is obvious that each modality has a different timing, most fusion mechanisms either use processing units of a fixed duration

or linguistically motivated time intervals, such as sentences. Kim et al. (2005) suggested for a corpus consisting of speech and biosignals choosing the borders of the single segments in such a way that it lies in the middle between two spoken utterances. Lingenfelser et al. (2011) used the time interval covered by a spoken utterance for all considered modalities, i.e. audio and video. These strategies suffer from two major problems. First, significant hints for emotion recognition from different modalities are not guaranteed to emerge at exactly the same time interval. Second, they might occur in a shorter time period than a sentence only. Classification accuracy could be expected to improve, if modalities were segmented individually and the succession and corresponding delays between occurrences of emotional hints in different signals could be investigated more closely. A promising step into this direction is the event-based fusion mechanism developed for the Callas Emotional Tree (Gilroy et al., 2011). Rather than computing global statistics in a segmentation-based manner, the approach aims to identify changes in the modality-specific expression of an emotion and is thus able to continuously respond to emotions of users while they are interacting with the system.

4.5 Dealing with imperfect data in the fusion process

Most algorithms for social signal fusion start from the assumption that all data from the different modalities are available at all time. As long as a system is used offline, only this condition can be easily met by analyzing the data beforehand and omitting parts where input from one modality is corrupted or completely missing. However, in online mode, a manual pre-selection of data is not possible and we have to find adequate ways of handling missing information. Generally, various reasons for missing information can be identified. First of all, it is unrealistic to assume that a person continuously provides meaningful data for each modality. Second, there may be technical issues, such as noisy data due to unfortunate environmental conditions or missing data due to the failure of a sensor. As a consequence, a system needs to be able to dynamically decide which channels to exploit in the fusion process and to what extent the present signals can be trusted. For the case that data is partially missing a couple of treatments have been suggested in literature, such as the removal of noise or the interpolation of missing data from available data. Wagner et al. (2011a) present a comprehensive study that successfully applies adaptations of state-of-the-art fusion techniques to the missing data problem in multimodal emotion recognition.

While semantic fusion is driven by the need to exploit the complementarity of modalities, fusion techniques in social signal processing make less explicit use of modality-specific benefits. Nevertheless, such an approach might help improve the gain obtained by current fusion techniques. For example, there is evidence that arousal is recognized more reliably using acoustic information while facial expressions yield higher accuracy for valence. In addition, context information may be exploited to adapt the weights to be assigned to the single modalities. For example, in a noisy environment less weight might be given to the audio signal. A first attempt to make use of the complementarity of modalities has been by Wagner et al. (2011a). Based on evaluation of training data, experts for every class of the classification problem are chosen. Then the classes are rank ordered, beginning with the worst classified class across all classifiers and ending with the best one.

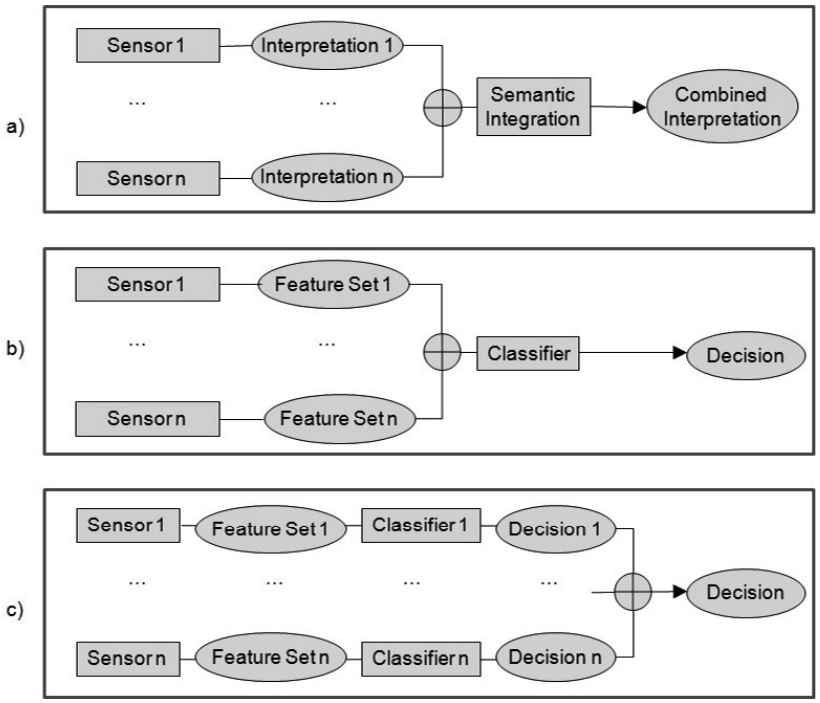

Figure 1. Different fusion mechanisms: (a) Semantic fusion, (b) feature-level fusion and (c) decision-level fusion.

4.6 Evaluation of schemes for social signal fusion

Since it is usually obvious which intention a user aims to convey to a system, the evaluation of schemes for semantic fusion is rather straightforward. On the opposite, the evaluation of fusion schemes for social and emotional signals raises a number of challenges. First of all, it is not obvious how to acquire ground truth data against which to evaluate the performance of automated emotion-recognition components. Users show a great deal of individuality in their emotional responses, and there is no clear mapping between behavioral cues and emotional states. To avoid a too-high degree of subjective interpretation, ground truth data are typically obtained by requesting multiple annotators label corpora. A voting scheme can be used if annotators disagree. One question to consider in this context is the amount of information to be made available to the annotators. Should the annotator just have access to one modality, such as speech or video, or to all available modalities? To provide a fair comparison between human and machine performance, it seems reasonable to make all the modalities that will be considered in the fusion process available to the human annotator. However, to acquire realistic ground truth values, one might consider giving the human annotator as much information as possible even if the fusion process will only employ part of it.

Another question concerns the processing units which should be taken into account for an evaluation. Since online fusion processes data frame by frame, an evaluation should consider time serious over a whole experimental session as opposed to computing global statistics over a longer period of time. One option might be to make use of an annotation tool, such as Feeltrace, that allows for the continuous annotation of data.

Instead of experimentally testing the robustness of the fusion process, a number of approaches have rather tested the effect of it. Gratch et al. (2007) presented an artificial listener that was able to recognize a large variety of verbal and nonverbal behaviours from a human user including acoustic features, such as hesitations or loudness, as well as body movements, such as head nods and posture shifts, and responded to it by providing nonverbal listening feedback. The system does not seem to employ a fusion engine, but rather responds to cues conveyed in a particular modality, such as a head nod, directly. An evaluation of the system revealed that the responsive agent was more successful in creating rapport with the human user than the non-responsive agent.

Typically, most experimental studies investigating the potential of social signal processing in human-agent dialogues have been

performed offline, i.e. after the interaction between the human and the agent. Such an approach may, however, be problematic because the participants of an experiment might have forgotten what they experienced at a particular instance of time. As an alternative, Gilroy et al. (2011) present an evaluation approach which compares the results of the fusion process with the users' physiological response during the interaction with the system.

4.7 Social signal fusion in human-agent interaction

Starting the recent years, various attempts have been made to explore the potential of social signal processing in human interaction with embodied conversational agents and social robots.

Sanghvi et al. (2011) analyzed body postures and gestures as an indicator of the emotional engagement of children playing chess with the iCat robot. They came to the conclusion that the accuracy of the detection methods was high enough to integrate the approach into an affect recognition system for a game companion. Even though the approach above addressed an attractive scenario for multimodal social signal fusion, it was only tested in offline mode. An integration of the approach into an interactive human-robot system scenario did not take place.

Increasing effort has been made on the multimodal analysis of verbal and nonverbal backchannel behaviors during the interaction with a virtual agent. An example includes the previously mentioned artificial listener by Gratch et al. (2007) that aims to create rapport with a human interlocutor through simple contingent nonverbal behaviors. A more recent example is the virtual health care provider recently presented by Scherer et al. (2012). This agent is able to detect and respond multimodal behaviors related to stress and post-traumatic stress disorder. For example, when the patient pauses a lot in the conversation, the agent tries to encourage her to continue speaking. Even though both systems are able to analyze multiple modalities, they do not seem to employ a fusion engine, but rather directly respond to cues conveyed in a particular modality, such as a head nod.

An exemplary application that is based on a fusion algorithm adapted to the specific needs of online processing is the Affective Listener "Alfred" developed by Wagner et al. (2011b). Alfred is a butler-like virtual character that is aware of the user and reacts to his or her affective expressions. The user interacts with Alfred via acoustics of speech and facial expressions (see Figure 2). As a response, Alfred simply mirrors the user's emotional state by appropriate facial expressions. This behavior can be interpreted as a simple form of showing empathy.

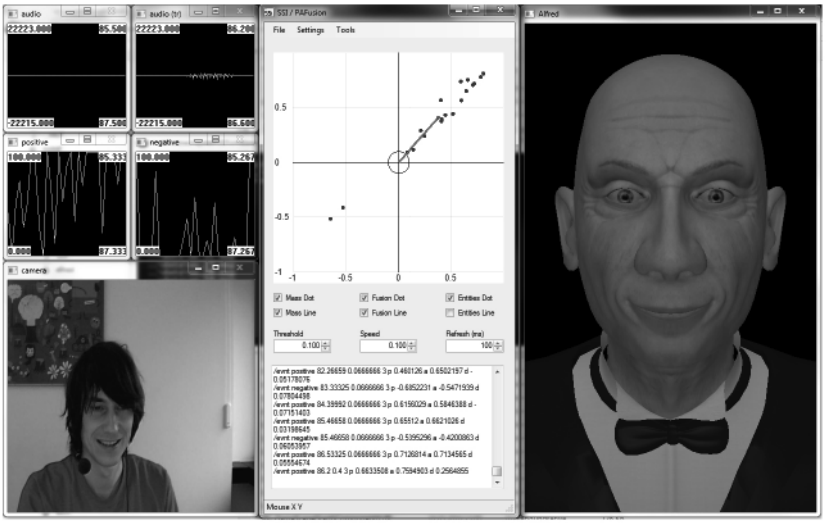

Figure 2. Affective Listener Alfred: the current user state is perceived using SSI, a framework for social signal interpretation (Wagner et al., 2011b) framework (upper left window); observed cues are mapped onto the valence and arousal dimensions of a 2D emotion model (upper middle window); values for arousal and valence are combined to a final decision and transformed to a set of Facial Animation Coding System (FACS) parameters, which are visualized by the virtual character Alfred (right window).

(Color image of this figure appears in the color plate section at the end of the book.)

The fusion approach is inspired by that developed for the Augmented Reality Tree. However, while Gilroy et al. (2011) generate one vector per modality, Wagner et al. (2011b) generate one vector for each detected event. This way they prevent sudden leaps in case of a false detection. Since the strength of a vector decreases with time, the influence of older events is lessened until the value falls under a certain threshold and is completely removed.

5. Exploiting Social Signals for Semantic Interpretation

Few systems combine semantic multimodal fusion for task-based command interpretation and multimodal fusion of social signals. A few studies nevertheless mention some interaction between the two communication streams. Such combinations occur in users' behaviors. For example, a user may say "Thanks" to a virtual agent and at the same time start a new command using gesture (Martin et al., 2006). In another study about multimodal behaviors of users when interacting with a virtual character embedded in a 3D graphical environment, such concurrent behaviors were also observed. In such cases, speech input was preferred for social communication with the virtual character ("how old are you?"), whereas 2D gesture input was used in parallel

for task command (e.g. to get some information about one of the graphical object displayed in the environment) (Martin et al., 2006).

An example of a system managing these two streams includes the SmartKom system, which features adaptive confidence measures. While the user is speaking (possibly for task commands), the confidence value of the mouth area recognizer is decreased for the module that detected emotions expressed in user's facial expression (Wahlster, 2003). The SmartKom system thus uses a mixture of early fusion for analyzing emotions from facial expressions and speech and late fusion for analyzing the semantics of utterances.

Rich et al. (2010) presented a model of engagement for human-robot interaction that took into account direct gaze, mutual gaze, relevant next contribution and back channel behaviors as an indicator of engagement in a dialogue. Interestingly, the approach was used for modeling the behavior of both the robot and the human. As a consequence, it was able to explain failures in communication from the perspective of both interlocutors. Their model demonstrates the close interaction between the communication streams required for semantic processing and social signal processing because it integrates multimodal grounding with techniques for measuring experiential qualities of a dialogue. If communication partners fail to establish a common understanding of what a dialogue is about, it is very likely that they will lose interest in continuing the interaction.

Bosma and André (2004) presented an approach to the joint interpretation of emotional input and natural language utterances. Especially short utterances tend to be highly ambiguous when solely the linguistic data is considered. An utterance like "right" may be interpreted as a confirmation as well as a rejection, if intended cynically, and so may the absence of an utterance. To integrate the meanings of the users' spoken input and their emotional state, Bosma and André combined a Bayesian network to recognize the user's emotional state from physiological data, such as heart rate, with weighted finite-state machines to recognize dialogue acts from the user's speech. The finite-state machine approach was similar to that presented by Bangalore and Johnson (2009). However, while Bangalore and Johnston used finite-state machines to analyze the propositional content of dialogue acts, Bosma and André focused on the speaker's intentions. Their objective was to discriminate a proposal from a directive, an acceptance from a rejection, etc., as opposed to Bangalore and Johnston who aimed at parsing user commands that are distributed over multiple modalities, each of the modalities conveying partial information. That is, Bosma and André did not expect the physiological modality to contribute to the propositional interpretation of an utterance. Instead, the emotional

input was used to estimate the probabilities of dialogue acts, which were represented by weights in the finite-state machines.

Another approach that fuses emotional states with natural language dialogue acts has been presented by Crook et al. (2012) who integrated a system to recognize emotions from speech developed by Vogt et al. (2008) into a natural language dialogue system order to improve the robustness of a speech recognizer. Their system fuses emotional states recognized from the acoustics of speech with sentiments extracted from the transcript of speech. For example, when the users employ words to express their emotional state that are not included in the dictionary, the system would still be able to recognize their emotions from the acoustics of speech.

6. Conclusion and Future Work

In this chapter, we discussed approaches to fuse semantic information in dialogue systems as well as approaches to fuse social and emotional cues. While the fusion of semantic information has been strongly influenced by research done in the natural language community, the fusion of social signals has heavily relied on techniques from the multimedia community. Recently, the use of virtual agents and robots in dialogue systems has led to stronger interactions between the two areas of research. The use of social signal processing in dialogue systems may not only improve the quality of interaction, but also increase their robustness. Vice versa research in social signal processing may profit from techniques developed for semantic fusion. Most systems that integrate mechanisms for the multimodal fusion of social signals in human-agent dialogue only consider a supplementary use of multiple signals. That is, the system responds to each cue individually, but does not attempt to resolve ambiguities by considering additional modalities. One difficulty lies in the fact that data have to be integrated in an incremental fashion while mechanisms for social signal fusion usually start from global statistics over longer segments. A promising avenue for future research might be to research to what extent techniques from semantic fusion might be included to exploit the complementary use of social signals. Among other things, this implies a departure from the fusion of low-level features in favor of higher level social cues, such as head nods or laughters.

Acknowledgement

The work described in this chapter is partially funded by the EU under research grants CEEDS (FP7-ICT2009-5), TARDIS (FP7-ICT-2011-7) and ILHAIRE (FP7-ICT-2009-C).

REFERENCES

Bangalore, S. and M. Johnston. 2009. Robust Understanding in Multimodal Interfaces. *Computational Linguistics*, **35(3):**345–397.

Bolt, Richard A. 1980. Put-that-there: Voice and Gesture at the Graphics Interface. Proceedings of the 7'th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '80, pp. 262–270. ACM, New York, NY.

Bosma, W. and E. André. 2004. Exploiting Emotions to Disambiguate Dialogue Acts. Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI '04, pp. 85–92. ACM, New York, NY.

Burger, B., I. Ferrané, F. Lerasle and G. Infantes. 2011. Two-handed gesture recognition and fusion with speech to command a robot. *Autonomous Robots*, **32(2):**129–147.

Busso, C., Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann and S. Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. International Conference on Multimodal Interfaces (ICMI 2004), pp. 205–211.

Caridakis, G., G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis and K. Karpouzis. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In Artificial Intelligence and Innovations (AIAI 2007), pp. 375–388.

Chen, F., N. Ruiz, E. Choi, J. Epps, A. Khawaja, R. Taib and Y. Wang. 2012. Multimodal Behaviour and Interaction as Indicators of Cognitive Load. ACM Transactions on Interactive Intelligent Systems, Volume 2, Issue 4, Article No. 22, pp. 1-36.

Crook, N., D. Field, C. Smith, S. Harding, S. Pulman, M. Cavazza, D. Charlton, R. Moore and J. Boye. 2012. Generating context-sensitive ECA responses to user barge-in interruptions. *Journal on Multimodal User Interfaces,* **6:**13–25.

Dimitriadis, D.B. and J. Schroeter. 2011. Living rooms getting smarter with multimodal and multichannel signal processing. IEEE SLTC newsletter. Summer 2011 edition, http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2011-07/living-room-of-the-future/

D'Mello, S.K. and J. Kory. 2012. Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. International Conference on Multimodal Interaction (ICMI 2012), pp. 31–38.

Eyben, F., M. Wöllmer, M.F. Valstar, H. Gunes, B. Schuller and M. Pantic. 2011. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. Automatic Face and Gesture Recognition (FG 2011), pp. 322–329.

Gilroy, S.W., M. Cavazza and V. Vervondel. 2011. Evaluating multimodal affective fusion using physiological signals. Intelligent User Interfaces (IUI 2011), pp. 53–62.

Gilroy, S.W., M. Cavazza, R. Chaignon, S.-M. Mäkelä, M. Niranen, E. André, T. Vogt, J. Urbain, M. Billinghurst, H. Seichter and M. Benayoun. 2008. E-tree: Emotionally driven augmented reality art. *ACM Multimedia* pp. 945–948.

Gratch, G., N. Wang, J.Gerten, E. Fast and R. Duffy. 2007. Creating Rapport with Virtual Agents. Intelligent Virtual Agents (IVA 2007), pp. 125–138.

Gruenstein, A., J. Orszulak, S. Liu, S. Roberts, J. Zabel, B. Reimer, B. Mehler, S. Seneff, J.R. Glass and J.F. Coughlin. 2009. City browser: Developing a conversational automotive HMI. In Jr., Dan R. Olsen, Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson and Saul Greenberg (eds.), Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Extended Abstracts Volume, Boston, MA, April 4–9, pp. 4291–4296. ACM.

Hofs, D., M. Theune and R. den Akker. 2010. Natural interaction with a virtual guide in a virtual environment: A multimodal dialogue system. *Multimodal User Interfaces*, **3:**141–153.

Johnston, M. 1998. Unification-based Multimodal Parsing. In Proceedings of the International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (Coling-ACL), Montreal, Canada, pp. 624–630.

Johnston, M. 2009. Building multimodal applications with EMMA. In Crowley, James L., Yuri A. Ivanov, Christopher Richard Wren, Daniel Gatica-Perez, Michael Johnston, and Rainer Stiefelhagen (eds.), Proceedings of the 11th International Conference on Multimodal Interfaces, ICMI 2009, Cambridge, Massachusetts, USA, November 2–4, 2009, pp. 47–54. ACM.

Johnston, M., G. Di Fabbrizio and S. Urbanek. 2011. mTalk: A multimodal browser for mobile services. In INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, pp. 3261–3264. ISCA.

Kaiser, E., A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen and S. Feiner. 2003. Mutual Disambiguation of 3D multimodal interaction in augmented and virtual reality. In Proceedings of the 5th International Conference on Multimodal Interfaces, ICMI '03, pp. 12–19. ACM, New York, NY, USA.

Kim, J., E. André, M. Rehm, T. Vogt and J. Wagner. 2005. Integrating information from speech and physiological signals to achieve emotional sensitivity. INTERSPEECH 2005, pp. 809–812.

Kim, J. and E. André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30(12):** 2067–2083.

Kleinsmith, A., N. Bianchi-Berthouze and A. Anthony Steed. 2011. Automatic Recognition of Non-Acted Affective Postures. *IEEE Transactions on Systems, Man and Cybernetics, Part B* **41(4):**1027–1038.

Lalanne, D., L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt and J.-F. Ladry. 2008. Fusion engines for multimodal input: A survey. In Proceedings of the 10th International Conference on Multimodal Interfaces ICMI 2008, pp. 153–160.

Latoschik, M.E. 2002. Designing transition networks for multimodal vr-interactions using a markup language. In Proceedings of ICMI' 02, pp. 411–416.

Lingenfelser, F., J. Wagner and E. André. 2011. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In Proceedings of the 13th International Conference on Multimodal Interfaces ICMI 2011, pp. 19–26.

Martin, J.C., R. Veldman and D. Béroule. 1998. Developing multimodal interfaces: A theoretical framework and guided propagation networks. In H. Burt, R.J. Beun and T. Borghuis (eds.), Multimodal Human-Computer Communication (Vol. 1374, pp. 158–187). Berlin: Springer Verlag.

Martin, J.-C., S. Buisine, G. Pitel and N.O. Bernsen. 2006. Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. *Journal of Signal Processing. Special issue on Multimodal Human-computer Interfaces,* **86(12):**3596–3624.

Martinovsky, B. and D. Traum. 2003. Breakdown in Human-Machine Interaction: The Error is the Clue. Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems, pp. 11–16.

Mehlmann, G. and E. André. 2012. Modeling multimodal integration with event logic charts. In Proceedings of the 14th ACM International Conference on Multimodal Interfaces, ICMI 2012, Santa Monica, USA, October 22–26, pp. 125–132.

Mehrabian, A. 1980. Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies. Cambridge, MA: Oelgeschlager, Gunn & Hain.

Oviatt, S.L. 1999. Mutual disambiguation of recognition errors in a multimodel architecture. In Williams, Marian G. and Mark W. Altom (eds.), Proceeding of the CHI '99 Conference on Human Factors in Computing Systems: The CHI is the Limit, Pittsburgh, PA, USA, May 15–20, pp. 576–583. ACM.

Rich, C., B. Ponsleur, A. Holroyd and C.L. Sidner. 2010. Recognizing engagement in human-robot interaction. Human-Robot Interaction, (HRI 2010), pp. 375–382.

Sandbach, G., S. Zafeiriou, M. Pantic and L. Yin. 2012. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vision Comput.*, **30(10):**683–697.

Sanghvi, J., G. Castellano, I. Leite, A. Pereira, P.W. McOwan and A. Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. Human Robot Interaction (HRI 2011), pp. 305–312. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A.A. and Morency, L.P. (2012). Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. Intelligent Virtual Agents (IVA 2012), pp. 455–463.

Scherer, S., S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo and L.-P. Morency. 2012. Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. Intelligent Virtual Agents. Y. Nakano, M. Neff, A. Paiva and M. Walker, Springer-Verlag Berlin Heidelberg, **7502:**455–463

Sowa, T., M. Latoschik and S. Kopp. 2001. A communicative mediator in a virtual environment: Processing of multimodal input and output. Proc.

of the International Workshop on Multimodal Presentation and Natural Multimodal Dialogue—IPNMD 2001. Verona, Italy, ITC/IRST, pp. 71–74.

Stiefelhagen, R., H. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, and A. Waibel. 2007. Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot. *IEEE Transactions on Robotics, Special Issue on Human-Robot Interaction*, **23(5):**840–851.

Sun, Y., Y. Shi, F. Chen and V. Chung. 2009. Skipping spare information in multimodal inputs during multimodal input fusion. Proceedings of the 14th International Conference on Intelligent User Interfaces, pp. 451–456, Sanibel Island, USA.

Sun, Y., H. Prendinger, Y. Shi, F. Chen, V. Chung and M. Ishizuka. 2008. THE HINGE between Input and Output: Understanding the Multimodal Input Fusion Results in an Agent-Based Multimodal Presentation System. CHI '08 Extended Abstracts on Human Factors in Computing Systems, pp. 3483–3488, Florence, Italy.

Visser, T., D. Traum, D. DeVault and R. op den Akker. 2012. Toward a model for incremental grounding in spoken dialogue systems. In the 12th International Conference on Intelligent Virtual Agents.

Vogt, T. and E. André. 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. Proceedings of the 2005 *IEEE International Conference on Multimedia and Expo*, ICME 2005, July 6–9, 2005, pp. 474–477. Amsterdam, The Netherlands.

Vogt, T., E. André and N. Bee. 2008. Emovoice—A framework for online recognition of emotions from voice. In André, Elisabeth, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, Roberto Pieraccini, and MichaelWeber (eds.), Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany, June 16–18, 2008, Proceedings, volume 5078 of Lecture Notes in Computer Science, pp. 188–199. Springer.

Wagner, J., E. André, F. Lingenfelser and J. Kim. 2011a. Exploring fusion methods for multimodal emotion recognition with missing data. *T. Affective Computing*, **2(4):**206–218.

Wagner, J., F. Lingenfelser, N. Bee and E. André. 2011b. Social signal interpretation (SSI)—A framework for real-time sensing of affective and social signals. *KI*, **25(3):**251–256.

Wahlster, W. 2003. Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. In Günter, Andreas, Rudolf Kruse, and Bernd Neumann (eds.), KI 2003: Advances in Artificial Intelligence, 26th Annual German Conference on AI, KI 2003, Hamburg, Germany, September 15–18, 2003, Proceedings, volume 2821 of Lecture Notes in Computer Science, pp. 1–18. Springer.