Workshop on Emotion and Computing

Current Research and Future Impact

Jahrestagung Künstliche Intelligenz 2011

October 4, 2011, TU Berlin

- HANDOUT –

Session 1 : 9:00 - 10:30 am

Welcome and Introduction
D. Reichardt

A computational model of emotional alignment
O.Damm, K. Malchus, F. Hegel, P. Jaecks, P. Stenneken, B. Wrede, and M. Hielscher-Fastabend

CoVE: Coping in Virtual Emergencies
C. Becker-Asano, D. Sun, B. Kleim, C. Scheel, B. Tuschen-Caffie

Session 2 : 11:00 - 12:30 pm

Topic and Emotion Classification of Customer Surveys
D. Suendermann, J. Liscombe, J. Bloom, R. Pieraccini

Acoustic analysis of politeness and efficiency in a cooperative time-sensitive task
M. Charfuelan, P. M. Brunet

Multidimensional meaning annotation of listener vocalizations for synthesis
S. Pammi, M. Schroeder, and M. Charfuelan

Session 3 : 14:00 - 15:30 pm
Discussion on "Applications of Emotional Computing" and Demo

An experimental triangulative research design for analyzing consumer behavior (Demo)
Y.Zajontz, V.Kollmann, M.Kuhn, D.Reichardt

A computational model of emotional alignment

Oliver Damm, Karoline Malchus, Frank Hegel, Petra Jaecks, Prisca Stenneken, Britta Wrede, and Martina Hielscher-Fastabend

University of Bielefeld,
33106 Bielefeld, Germany
{odamm,fhegel,bwrede}@techfak.uni-bielefeld.de
{karoline.malchus,petra.jaecks,prisca.stenneken}@uni-bielefeld.de
{hielscherfastabend}@ph-ludwigsburg.de

Abstract. In order to make human-robot interaction more smooth and intuitive it is necessary to enable robots to interact emotionally. Having a robot which can align to interlocutors emotion expressions will enhance its emotional and social competence. Therefore we propose a computational model of emotional alignment. This model regards emotions from a communicative and interpersonal view and is based on three layers: The first layer comprises the automatic emotional alignment, the second layer the schematic emotional alignment and the third one the conceptual emotional alignment. In a next step we have to implement our model on a robotic platform and to evaluate it.

1 Introduction

In social interaction, expressing and understanding emotions is essential [11]. Furthermore, personal mood, attitudes and evaluations implicitly influence communication processes ([12]; [?]). Therefore, interaction is always emotionally colored [10]. This statement is in harmony with Schultz von Thun [14], who postulates that in addition to the objective meaning, emotional information is also conveyed (e.g. the revealing of the self or the relation to the interaction partner). A study by Eyssl et al. [7] exemplifies the relevance of emotions in human-robot interaction. They found that people sympathize more strongly with a robot if it communicates emotions.

One reason for this might be that people expect a behavior, which they often express themselves in their real-life interactions [23]. In other words: There are many emotion expressions in human- human interaction (HHI), and therefore people presume the same for human-robot interaction (HRI).

With regard to the account of alignment, postulated by Pickering & Garrod [16], there are communicative mechanisms, which lead to an adaptation between the interlocutors. This adaption is an essential part of human-human interaction ([8]; [13]), according to this the contextual aspects of emotional processing have to be taken into account for building social robots.

With regard to the account of alignment, postulated by Pickering & Garrod [16], there are communicative mechanisms, which lead to an adaptation between the interlocutors. In contrast to other communicative theories, alignment is based on automatic and resource-saving processes [18]. In this context alignment is an essential part of human-human interaction [8]. That alignment is also an important part of human-computer interaction was illustrated for example by Suzuki and Katagiri [20] or Branigan et al. [4]. Concerning emotions, we understand a communication as emotionally aligned if both interaction partners show adequate reactions to expressed emotions. This can be a simple mirroring or copying of the emotion expression, an emotional reaction based on emotional contagion or an empathic reaction (see Part 3). Linking these different levels of affective adaptation processes, we propose a layer model of emotional alignment between humans and robots as the basis for a computational model that will produce emotion expressions. These emotion expressions are influenced by the emotional adaptation process in communication on the one hand and contextual and situational aspects on the other hand.

2 Related Work

Most computational models of emotions are influenced by anatomic approaches (e.g. [21]) or appraisal and dimensional theories of emotions. As an example, Marsella and Gratch presented EMA, a computational model of appraisal dynamics. They assume the dynamics arises from perceptual and inferential processes operating on a persons interpretation of their relationship to the environment. A model based on the dimensional approach were proposed by Gebhard [9]. The ALMA integrates three major affective characteristics emotions, moods and personality and covers short, medium, and long term affect. They implemented their model of mood with the three traits pleasure (P), arousal (A), and dominance (D) as described by Mehrabian.
The WASABI Affect Simulation Architecture by Becker-Asano [2] puts appraisal and dimensional theories together. Becker-Asano models emotions by representing aspects of each secondary emotions connotative meaning in PAD space, he also combines them with facial expressions, that are concurrently driven by primary emotions.
In communicative approaches the expression of emotions fulfils two functions. On the one hand the interactant is informed of one's mental state, on the other hand the expression is used to request changes in others behavior. A computational model of these approaches enable the social robot to decide on it's own when an emotional display will fulfill the expectations of the user.
A model for multimodal mimicy of human users were developed and implemented by Caradakis et. al [5]. In this case the mimicry is realized in a loop of perception, interpretation, planning and animation of the expressions.The result is not in an exact duplicate of the human but an expressive model of the users original behavior.

Paiva [15] describes an empathy-based model for agents, which involves two stages. The first one is the empathic appraisal, the second one the empathic response. Boukricha et al. [3] propose an emotion model for a virtual agent, too. Thereby the authors focus on alignment processes based on empathy.
In this paper we want to demonstrate a computational model, which isn't only limited to one level of emotional alignment. Hence, we believe that our three-layered computational model of emotional alignment is a promising extension to established approaches. In the following sections the model and especially the layers will be described in detail.

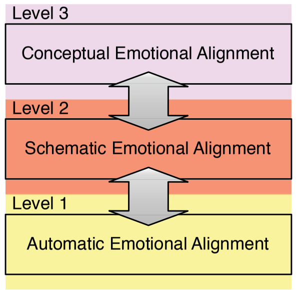

Fig. 1. Theoretical layer model of emotional alignment

3 Layer model of emotional alignment

Each communication signal is part of a bidirectional process [10]. Therefore we propose a layer-model of communicating emotions which regards emotion expressions from a more social and interpersonal view [6]. This model, called layer model of emotional alignment (see fig. 1), has three layers: The first layer comprises the automatic emotional alignment, the second layer the more schematic emotional alignment and the third layer the conceptual emotional alignment. Based on these levels we are able to describe the functions of emotion expressions in human-robot interaction and their underlying processes. It is important to note that the different layers do not represent different categories of emotions (e.g. primary emotions vs. secondary emotions). Our model presents a distinction between automatic, schematic and conceptual emotional adaptive reactions (=alignment) to the interaction partner. While it is still under debate if these mechanisms are distinct alternatives in human-human interaction, we suppose our layer model of emotional alignment to be highly relevant and helpful in designing human-robot communication [6]. In the following we introduce the computational model based on this layer model and describe the different layers in detail.

4 Computational Model of emotional Alignment

Developing a computational model of emotional alignment requires building a system which is able to produce the similar phenomena that can be observed in human-human interaction. Such phenomena might be mimicking an emotional expression, emotional contagion or empathy.
In the following section we describe a computational approach to implement the proposed layer model of emotional alignment on a robotic platform. According to the theoretical model, the computational model (fig. 2) can be split into three levels of computational complexity. In the following sections, the main components of the proposed model will be described in detail. Thereafter the levels of processing will be specified.

Perception and Expression of emotional Stimuli
In human-computer interaction it is useful to get visual as well as auditory input to analyze the given situation and react in an appropriate manner. The input component (fig. 2, box 1) of the system takes different input-sources into account. The model uses a multi-modal approach to compute the emotion. It is not restricted to the inference of only one channel (e.g. only facial expressions) but rather uses a broader spectrum of information and applies different techniques, such as speech processing and pattern recognition, to make an inference from this data. Because any given sensor will have various problems with signal noise and reliability, and a single signal will contain limited information about emotion, the use of multiple sensors should also improve robustness and accuracy of inference. The promising approach seems to be the combination of recognition of emotional features from voice (e.g. [22]) and the analysis of facial expressions (e.g. [17]).

Recognition of Context
According to our model of interpersonal emotions it is indispensable to take the whole situation or even parts of it into account and extract the relevant features for the current interaction. In a natural interaction factors like the expected reaction of the interlocutor (congruent/incongruent with the expectation), the human relation (private/occupational) or the sympathy for each other determines the situational context.
This context (fig. 2, box 2) is divided into an external part and an internal part, which is a situational memory of the robot. The internal situational knowledge is necessary for several reasons, e.g. for the formation of expectations during an interaction or to model the essential background knowledge about the current

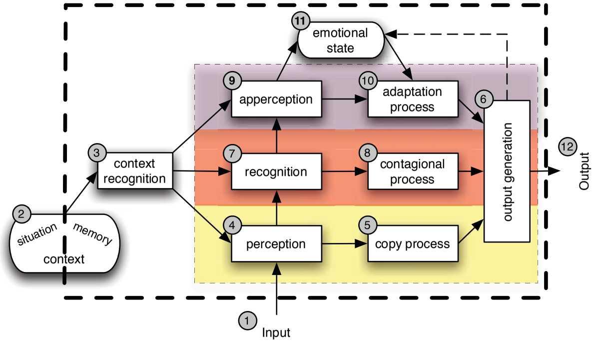

Fig. 2. The proposed computational model for emotional alignment.

application.
The external part of the context models the surroundings relevant to the robot. That is, the current interlocutor and all visual and auditory stimuli are part of the situational context. All of these objects and events may influence the robot, the kind of reaction as well as the level of processing.
The recognition and evaluation of the context is mainly dependent on the current task. The relevant factors in a storytelling-situation may differ from those in a child-parent situation. In the storytelling-situation, a smile of the interlocutor can be related to a funny part of the story, but it can also be the reaction to the robots expression. So the reason of the smile may differ: It can convey an emotion of the teller or mirror the observed smile.

In the parent-child interaction, a parent may smile after a child has succeeded at a difficult task. Even so, there can be a parents smile after the child has failed at that task. The message of these smiles differ. In one case, the smile can be an expression of happiness and pride; in the other case the smile can be seen as an encouraging signal ([1]).
With regard to the importance and the complexity of the context, the recognition of the context influences emotional alignment on every level. In this way the situational context is also involved in the decision on which level the emotional alignment occurs.

Internal Model of Emotions
Artificial emotions in a robotic system can fulfill several conditions, beginning with the computation of facial expressions up to influencing the whole behavior.

In our interpersonal model the emotional state (fig. 2, box 11) is first and foremost important for the conceptual level of emotional alignment. According to the intended purpose, the emotional state is mainly influenced by the apperception process of the conceptual layer. In addition a feedback from output generation will enable a synthesized utterance to influence the internal emotional state.
According to several findings, facial feedback influences the own experience of an emotion [19]. The link between *output generation* (fig. 2, box 6) and *emotional state* (fig. 2, box 11) realizes a kind of facial feedback. By linking the process to the emotional state, a synthesized emotion can influence the internal state of the robot.

Layers of Processing
As aforementioned the processing of the emotional feedback may occur on several layers of complexity. The choice of the level depends on the level of understanding and the necessity, i.e. in case of non-understanding only the level of automatically emotional alignment can be reached. On the lowest level the processing is limited to *perception* (fig. 2, box 4) of an emotion and the *copy process* (fig. 2, box 5). The middle level, named *recognition* (fig. 2, box 7) and *contagional process* (fig. 2, box 8), uses the features previously extracted by the underlying level to compute a hypothesis with respect to the observed expression. The third level, the *apperception* (fig. 2, box 9) and the *adaptation process* (fig. 2, box 10), is the top-level process.
In the following paragraphs we describe how the three levels process a given stimulus and produce an emotional reaction.

Level 1: Automatic Emotional Alignment
On the lowest level the processing is limited to *perception* of an emotion and the *copy process* without a classification of the emotion. This means that the visual and auditory information will be captured and analyzed on the signal processing level.
According to our model a given stimulus will take a route starting from *perception* (fig. 2, box 4). In this component, the presented stimulus will be analyzed on a level of signal processing. The gained features are provided to the following component (fig. 2, box 5). Depending on the modality of the stimulus, this process maps the received features into motor-commands or prosodic features of the emotional display. With this mapping the next component (fig. 2, box 6) will be able to synthesize an emotional utterance with similar or even perhaps the same emotional feature as the perceived. On this level the module of context recognition (fig. 2, box 2) may influence the way and the frequency of automatic adaptation of emotional expressions.

Level 2: Schematic Emotional Alignment
The second level of emotional alignment processing builds on the automatic level. But, schematic emotional alignment uses the perceived motor movements to recognize the observed emotion by analyzing its distinct features (e.g. vi-

sual or prosodic cues)(fig. 2, box 7). In the following *contagional processing* the relevant emotional expression is chosen (fig. 2, box 8) and information for output generation is transferred to (fig. 2, box 6), where a motor program produces an emotionally aligned output on all relevant channels. With respect to a storytelling-situation, the process can be described as follows: The narrator reads a passage to the robot. At the same time he expresses a specific emotion, e.g. sadness by a sad facial expression and tears. The whole expression is perceived by the robot, which combines the different features to recognize the correct emotion. Based on emotional schema, the social robot will then align with the narrator. For example, it will show sadness by a sad facial expression and an altered prosody, although the human interaction partner did not speak with a sad voice but expressed his sadness by tears. Nevertheless, the robot recognizes the emotion and expresses it itself exceeding mimicry and automatic emotional alignment.

Level 3: Conceptual Emotional Alignment

The third layer of emotional alignment is the most complex level. Similar to the underlying, this layer receives contextual informations as well as the preprocessed sensory input. On this level the emotional input has to be classified and analyzed with regard to it's influence on the internal emotional state (fig. 2, box 11). The third layer consists of the components *apperception* (fig. 2, box 9) and *adaptation process* (fig. 2, box 10). The process of apperception can be described as a conscious recognition of an perceived emotion whereas the input of the *context recognition* (fig. 2, box 3) is taken into account.

In the process of adaptation (fig. 2, box 10) the robots takes the own emotional state (fig. 2, box 11) as well as the result of the apperception process into account. This generates an emotional response to the given stimuli.

With respect to the storytelling-situation, the process can be described as follows: As on the schematic level the narrator reads a passage to the robot and expresses a specific emotion, e.g. through his face and voice (fig. 2, box 1). The whole expression is perceived by the robot (fig. 2, box 4), recognized (fig. 2, box 7) and consciously percepted (fig. 2, box 9). Influenced by the situational context and the internal emotional state, the social robot will then align with the narrator. For example, if the robot percepts a sad facial expression and the evaluation of the situational context implies that the narrator read a sad part of the story it will try to cheer him up. In summary, this model is not limited to describe only one alignment process, e.g. empathy or mimicry. It regards emotional interaction processes from a more communicative perspective and integrates alignment processes, which can be allocated to the three layers (automatic, schematic, conceptual). In addition, the model is influenced on all 3 layers of processing by internal and external context factors. Communication with an (emotionally) aligning robot is supposed to be much easier than with less adaptive partners.

5 Conclusions and Outlook

In this paper we argue that the current state-of-the-art models of artificial emotions should include communicative adaptation processes to reliably model human-robot interaction. A robot, which aligns in communication to the emotions expressed by the human partners, will not only be perceived more natural and emotionally more competent but will also enhance successful communication. As an extension to Pickering and Garrods model of alignment in communication, we presented a computational model of emotional alignment.

Even though the alignment approach is still a relatively new theory in human-human communication research, we think that our model is a useful addition to human-robot interaction studies. The next steps are twofold. The first one is to implement the here presented model into our robotic platform "Flobi". We want the robot to react emotionally to its communication partner alternatively on the three described layers. In the second step, we are going to evaluate our model. To validate the difference between the single layers we plan a set of empirical interaction studies including factors such as context, situation or communicative goal. The results of the experiments allow us to refine our model in order to support an emotionally aligned communication with social robots.

6 Acknowledgements

This research is partially supported by the German Research Foundation (DFG) in the Collaborative Research Center 673 "Alignment in Communication".

References

1. K. Barrett and G. C. Nelson-Goens. Emotion Communication and the Development of the Social Emotions. *New directions for child development*, 1997.
2. C. Becker-Asano and I. Wachsmuth. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49, 2010.
3. H. Boukricha and I. Wachsmuth. Empathy-Based Emotional Alignment for a Virtual Human: A Three-Step Approach. *KI - Künstliche Intelligenz*, 25(3):195–204, May 2011.
4. H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean. Linguistic alignment between people and computers. pages 1–14, 2010.
5. G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. Virtual agent multimodal mimicry of humans. *Computers and the Humanities*, pages 1–36.
6. O. Damm, K. Dreier, F. Hegel, P. Jaecks, P. Stenneken, B. Wrede, and M. Hielscher-Fastabend. Communicating emotions in robotics: Towards a model of emotional alignment. *Proceedings of the workshop "Expectations in intuitive interaction" on the 6th HRI International conference on Human-Robot Interaction*, Jan. 2011.
7. F. Eyssel, F. Hegel, G. Horstmann, and C. Wagner. Anthropomorphic inferences from emotional nonverbal cues: A case study. In *Proceedings of the 19th IEEE International Symposium in Robot and Human Interactive Communication (RO-MAN 2010)*, pages 681–686, 2010.

8. a. H. Fischer and G. a. van Kleef. Where Have All the People Gone? A Plea for Including Social Interaction in Emotion Research. *Emotion Review*, 2(3):208–211, Apr. 2010.
9. P. Gebhard. ALMA A Layered Model of Affect. *Artificial Intelligence*, pages 0–7, 2005.
10. K. H Delhees. *Soziale Kommunikation: psychologische Grundlagen für das Miteinander in der modernen Gesellschaft*. Opladen: Westdt. Verlag, 1994.
11. R. Harré. The discursive mind. *books.google.com*, 1994.
12. M. Hielscher. Emotion und Sprachproduktion. *Gert Rickheit/Theo Herrmann/Werner Deutsch (Hg.), Psycholinguistics/Psycholinguistik. Ein internationales Handbuch, Berlin/New York*, pages 468–490, 2003.
13. R. E. Kraut and R. E. Johnston. Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, 37(9):1539–1553, 1979.
14. I. Langer and F. von Thun. Sich verstandlich ausdrucken. *wwwzb.fz-juelich.de*, 1981.
15. A. Paiva. Empathy in Social Agents. *International Journal*, 10(1):65–68, 2011.
16. M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, pages 1–58, 2004.
17. A. Rabie, C. Lang, M. Hanheide, M. Castrillon-Santana, and G. Sagerer. Automatic Initialization for Facial Analysis in Interactive Robotics. *Computer Vision Systems*, pages 517–526, 2008.
18. G. Rickheit. Alignment und Aushandlung im Dialog. *Zeitschrift für Psychologie*, 213(3):159–166, July 2005.
19. F. Strack and L. Martin. Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social . . .*, 1988.
20. N. Suzuki and Y. Katagiri. Prosodic alignment in humancomputer interaction. *Connection Science*, 19(2):131–141, June 2007.
21. J. D. Velasquez. A computational framework for emotion-based control. In *Conference*, pages 62–67, 1998.
22. T. Vogt, E. Andre, and N. Bee. EmoVoice A framework for online recognition of emotions from voice. *Perception in Multimodal Dialogue Systems*, pages 188–199, 2008.
23. A. Weiss, N. Mirnig, and F. Forster. What users expect of a Proactive Navigation Robot. In *Proceedings of the workshop "Expectations in intuitive interaction" on the 6th HRI International conference on Human-Robot Interaction*, 2011.

CoVE: Coping in Virtual Emergencies

C. Becker-Asano, D. Sun, B. Kleim, C. N. Scheel, B. Tuschen-Caffier and B. Nebel

Freiburg Institute for Advanced Studies, Albert-Ludwigs-Universität Freiburg,
Starkenstraße 44, 79104 Freiburg, Germany
{basano, sun, nebel}@informatik.uni-freiburg.de, b.kleim@psychologie.uzh.ch,
{corinna.scheel, brunna.tuschen-caffier}@psychologie.uni-freiburg.de

Fig. 1. Outline of the overall design of a first empirical study

Abstract. The applicability of appropriate coping strategies is important in emergencies or traumatic experiences such as car accidents or human violence. However, research on human reactions to traumatic experiences is very challenging and most existing research uses retrospective assessments of these variables of interest. Thus, we are currently developing and evaluating novel methods to investigate human behavior in cases of emergency. Virtual Reality (VR) scenarios of emergencies are employed to enable an immersive interactive engagement (e.g., dealing with fire inside a building) based on the modification of Valve's popular Source™ 2007 game engine.

Preliminary results of a first empirical study (cp. Figure 1) suggest that our VR scenario has a similar fear-inducing effect as a short movie clip (Becker-Asano, Sun, Kleim, Scheel, Tuschen-Caffier, & Nebel, 2011), which previously has been evaluated to induce fear. In addition, the neutral VR experiences during the training sessions did never elicit fear in our participants, letting us conclude that the interactively presented emergency itself was indeed the fear eliciting factor in the experimental sessions. In the long run, we aim at a more detailed analysis that includes the personality questionnaire and physiological data, which will be analyzed in correlation with the trajectories of the participants in the VR emergency.

References

Becker-Asano, C., Sun, D., Kleim, B., Scheel, C. N., Tuschen-Caffier, B., & Nebel, B. (2011). Outline of an empirical study on the effects of emotions on strategic behavior in virtual emergencies. *Emotion in Games workshop in conj. with ACII2011*. Memphis, USA: Springer. (accepted)

Topic and Emotion Classification of Customer Surveys

D. Suendermann[1,2], J. Liscombe[2], J. Bloom[2], and R. Pieraccini[2]
{david,jackson,jonathanb,roberto}@speechcycle.com

[1] Baden-Wuerttemberg Cooperative State University, Stuttgart, Germany
[2] SpeechCycle Labs, New York, USA

Abstract. A method to assess the quality of customer service phone interactions is to point callers to an online survey where they can express their opinions, wishes, complaints, commendations, etc. by way of free-form text input. This paper investigates to which extend semantic classification can be applied to large amounts of surveys (thousands) in order to answer questions such as those in the following examples:

– Which callers are calling about their bill, technical issues, product pricing, etc.?
– Has the percentage of callers complaining about long hold time on the phone increased from month to month?
– Who is asking for a call-back or is threatening to cancel service with the company being called?
– Is the caller conveying positive, negative, or neutral emotion referring to a certain topic?

Three statistical classifiers (Ripper, SVM, naïve Bayes) were evaluated on a manually annotated set of 5589 surveys using ten-fold cross-validation. In doing so, 15 different topics (classes) were investigated. In an additional set of experiments, each class was associated with an emotion flag (positive/negative/neutral) to add valence to the picture. In order to cope with the occurence of multiple classes and emotion flags in a single survey, we introduced a novel annotation language encoding semantics, emotion flags, and temporal sequence of topics. A demo system can be accessed at `http://suendermann.com/verbatim.php5`.

1 Introduction

In a world where product and service features barely differ among competitors of certain businesses, the quality of customer service is an important differentiator. E.g., in the telecommunication industry, bundle services nowadays include cable TV, high-speed Internet, landline and wireless service whose features are largely identical among different providers. In addition to lower pricing, providers try to differentiate their services by means of superior customer service and support. Consequently, one of the main focuses of the customer service departments of large companies is to constantly monitor the quality of services rendered [8].

A frequently used method to assess customer support is to survey customers [16, 7] . This can be done in a number of ways including

1) out-bound calling customers and asking a number of questions,
2) asking customers who are calling into a service hotline a number of questions right after their service interaction,
3) sending customers a personal e-mail after a completed service interaction with a link to a survey web portal.

Survey questions are generally of these types:

A) yes/no (e.g., *Were you satisfied with this customer service interaction?*),
B) multiple choice (e.g., *Which was the reason for your call: billing, payment, technical support, general inquiry, or something else?*), or
C) free-form (e.g., *What was the reason for your call?*).

Responses to questions of Type A or B can be evaluated in a rather straightforward fashion by calculating frequency distributions over the number of possible choices (e.g., 85% of the callers were satisfied [Type A], or 21% of the people called about billing, 18% wanted to make a payment, etc. [Type B]). Type C allows customers to express their opinions and desires in an unconstrained way, which has the potential of conveying lots of useful and detailed information. E.g., by matching customers to the call center representative serving them, it can provide very specific feedback. An example of a Type-3 survey response collected via a web interface of a large cable service provider is

> *Cynthia's assistance went above and beyond. However, even though Cynthia offered me a new contractual option with your company, (Which I will give it a 1 year trial) I feel that my rates for cable & internet are extremely high and if they continue to rise, I will discontinue my service with your company.*

It is certainly worthwhile for customer service managers to read such survey responses every now and then to hear the direct voice of the customers. However, in companies processing millions of customer interactions every week [13], the manual processing of free-form customer feedback becomes unfeasible. Instead, in this paper, we propose the application of semantic classifiers to textual features in order to identify surveys belonging to predefined topics (classes). This method can be useful to answer a variety of questions of primary interest to stakeholders in customer service departments. Examples include the ones listed in the abstract:

– Which callers are calling about their bill, technical issues, product pricing, etc.?
– Has the percentage of callers complaining about long hold time on the phone increased from month to month?
– Who is asking for a call-back or is threatening to cancel service with the company being called?
– Is the caller conveying positive, negativ, or neutral emotion referring to a certain topic?

Section 2 will focus on the derivation of topics and emotion flags; the annotation scheme will be discussed in Section 3. Then, in Section 4 we will provide details on the experimental setup around this work and present results.

2 Topics and Emotion Flags

Topics of particular interest to customer service departments, e.g. in the cable provider market vertical, include surveys about

- an **A**utomated system,
- the **B**illing department,
- the **C**osts of services,
- a billing **D**ispute,
- an **E**mergency situation (e.g., callers threatening to cancel service),
- a request to **F**ollow up with the caller (call-back request),
- a **H**uman representative,
- the automated **I**nternet troubleshooting system [1],
- **O**ther topics,
- a **P**roduct,
- the general-purpose call **R**outer [5],
- a vague mentioning of an automated trouble-**S**hooting system [1],
- a **T**ruck roll or a **T**echnician on site,
- the automated cable **TV** troubleshooting system [1],
- **W**ait time in line.

The bolded letters are unique to each topic and will be used to refer to topics in the scope of the annotation scheme introduced in Section 3.

A fixed number of unique classes to distiguish in written documents generally suggests the application of a semantic classifier similar to what is being used for the task of call routing [4]. There, callers are asked to briefly describe the reason for their call in response to a system prompt such as

Briefly tell me what you are calling about today.

After applying large-vocabulary speech recognition to the caller response, a semantic classifier is applied to the recognition hypthesis returning one of a number of possible call reasons (classes). High-resolution call routers sometimes distinguish hundreds of classes [14].

However, it turns out that responses to call routing system prompts and survey responses of unlimited input length differ considerably in their nature. The example given above is prototypical for free-form responses in that they are not limited to a unique topic but contain a time sequence of topics. The topic sequence of this particular example is decoded in Table 1.

Reviewing this example, we observe that the mentioning of a topic can be associated with a certain emotion. The emotional flavor of a customer comment is clearly of special interest to the customer service department. It is crucial to know whether people like or hate their services, whether they had a positive or negative experience with the call center agent or spoken dialog system, or whether product costs are considered cheap or expensive. For this purpose, we introduce a three-point emotion scale (positive/neutral/negative).

In Table 1, each topic is also associated with an emotion flag, so, we see for instance that a human agent is mentioned twice, once in a positive way (*went above and beyond*) and once neutral (*Cynthia offered me*).

Table 1. Example for a time sequence of topics

text	annotation	emotion flag
Cynthia's assistance went above and beyond.	H	+
However, even though Cynthia offered me...	H	
a new contractual option with your company, (Which I will give it a 1 year trial)...	O	
I feel that my rates for cable & internet are extremely high...	C	-
and if they continue to rise, I will discontinue my service with your company.	E	-

3 Annotation

As motivated in Section 1, we want to apply statistical classifiers in order to automatically analyze the membership of a given survey to the classes and emotion flags introduced in Section 2. In order to train the classification models, we need to establish respective training data. In our case, we need to map the survey text to the canonical classes and emotion flags it represents. This process is often done in a supervised manner (i.e., manually) and is referred to as *annotation*.

Semantic annotation as required for a standard call routing task (see Section 2) maps exactly one class to a given utterance/text [5]. Figure 1 shows annotation software which lists a number of caller responses to the aforementioned example prompt *Briefly tell me what you are calling about today.* On the left, possible classes are shown in a hierarchical fashion (similar to a folder structure). The annotation task consists now of dragging and dropping utterances into one of the classes on the left. For example, the utterance *I am having a problem ordering a movie* refers to cable TV service (aka *Video*), it is about an *Order* and describes a *Problem*. The correct class would hence be

```
Video_Order_Problem
```

Sometimes, callers refer to multiple reasons at once (e.g., *I'd like to order a show and pay last month's bill*). Since the above described annotation method is not designed to accomodate multiple classes for a single utterance, this utterance would be mapped to a generic multiple-symptom class. Since, usually, these cases are negligible (0.4% for our example call router), no special handling for mappings to multiple classes is required.

As shown in Section 2, the situation is completely different for the case of unrestricted surveys. In fact, the corpus used in our experiments (see Section 4), contained 64% surveys with multiple classes.

To cover all possible scenarios of classes and emotion flags which can be associated with a given survey, we came up with a simple language describing

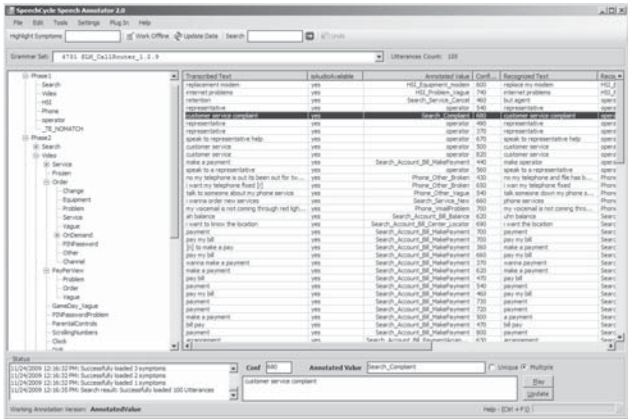

Fig. 1. Annotation software processing data of a call routing task

the time sequence of topics and emotion flags encountered in the survey. Here, the coding scheme of Table 1 is used, so, for the table's example, the semantic annotation string is

`H+HOC-E-`

Generally, our annotation language l can be expressed as

$$l := c[l]$$
$$c := t[e]$$
$$t \in \{\mathtt{P}, \mathtt{H}, \mathtt{W}, \mathtt{T}, \mathtt{B}, \mathtt{D}, \mathtt{A}, \mathtt{R}, \mathtt{I}, \mathtt{V}, \mathtt{C}, \mathtt{F}, \mathtt{E}, \mathtt{O}\}$$
$$e \in \{+, -\}$$

Figure 2 shows how the same annotation software we have applied to the call routing scenario can be used to produce the annotation string. While reading the survey, the annotating person writes the string into the *Annotated Value* field.

4 Experiments

4.1 The Classification Framework

A practical way to answer the questions raised in the introduction of this paper is to train separate classifiers for each topic (class). These classifiers would be

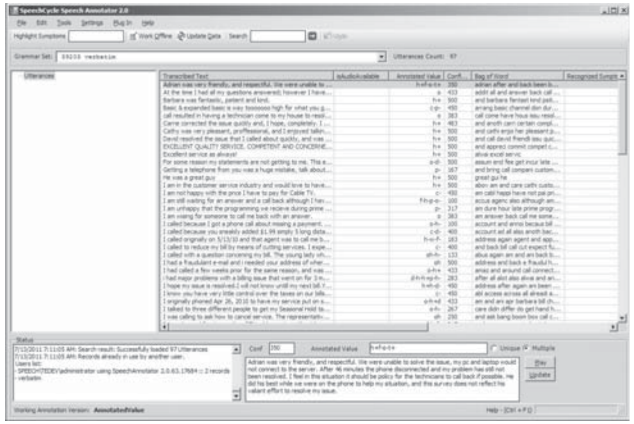

Fig. 2. Annotation software processing data of the survey task

binary when discarding emotion flags at the first place, i.e., the classifier would return `1` in the case it is confident that the survey is about a certain topic, otherwise `0`. This means that as many classifiers have to be trained as there are distinct classes, i.e., in our case 15.

When adding emotion flags to the picture, one has to be aware of the fact that a single survey can possibly contain multiple mentionings of the same topic with different emotion flags each. Principly, every single combination of positive, negative, and neutral are possible in a single survey for a single class (in our example in Section 2, we had positive and neutral for the class `H`. Consequently, when we would intend to use a single classifier per topic, it would have to be able to return every possible combination of emotion flags: `+`, `-`, `0`, `+-`, `+0`, `-0`, `+-0`, so, seven distinct return values. Here, `0` stands for *neutral*.

Another possibility to cope with emotion flags in this framework is to train separate binary classifiers for each topic/emotion flag combination. I.e., we would have an `H+` classifier, an `H-` classifier, and an `H0` classifier for the topic `H`.

4.2 Measuring Performance

In addition to the substantial difference between the annotation scheme of a call router and that of the free-form surveys we introduced in Section 3, there is also a major difference in the way classifier performance should be measured. In spoken-language understanding tasks as for instance in call routing, the classification hypothesis is simply compared with the canonical class (which a human annotator produced for the utterance in question). Here, the hypothesis is either correct of wrong. The metric True Total is the number of correct matches divided by the total number of samples in a test corpus, i.e., it is the percentage of correct responses of the classifier on a given test corpus [15].

Theoretically, one can calculate the True Total also for the binary classification scenario of the current work. However, as it turns out, the result can be misleading. This is because some of the topics have a very low likelihood of occurrence. For instance, only 0.2% of the surveys analyzed in this work mentioned `I` (see Table 2). That means, if we build a trivial classifier that exclusively returns the majority vote (in this case `0`), it would be correct in 99.8% of the cases, a True Total that seems extraordinarily good. However, it missed all the cases that *did* mention `I` rendering it completely useless.

Table 2. Distribution of topics in the corpus.
Note: Percentages describe the fraction of surveys in which the topic/the topic with a certain emotion flag was found. Due to multiple occurrences of topics/emotion flags in some surveys, `total` does not add up to 100%, and `+` and `-` do not necessarily add up to `total`.

topic	description	`total`	`+`	`-`
A	automation	3.8%	0.2%	3.4%
B	billing	0.6%	0.0%	0.5%
C	cost	10.5%	0.3%	9.9%
D	dispute	4.4%	0.1%	3.6%
E	emergency	8.4%	0.0%	6.2%
F	follow-up	3.8%	0.5%	3.0%
H	human	66.6%	50.3%	17.6%
I	Internet	0.2%	0.0%	0.1%
O	other	34.0%	6.7%	20.2%
P	product	23.1%	2.5%	20.3%
R	call router	1.0%	0.0%	0.9%
S	troubleshooter	1.4%	0.2%	1.2%
T	truck	10.9%	6.3%	3.7%
V	TV	0.2%	0.0%	0.2%
W	wait	3.5%	0.3%	3.2%

In cases like these, the machine learning community usually considers the standard metrics Precision, Recall, and F-Measure [11]. Precision is the percentage of correctly accepted tokens in the set of accepted tokens. So, Precision describes the *quality* of accepted tokens. Recall, on the other hand, is the percentage of the correctly accepted tokens in the set of all tokens which *should have been* accepted. That is, Recall describes the *completeness* of accepted tokens. Finally, F-Measure is a harmonic mean of Precision and Recall.

Depending on the specifics of the classification task, Precision and Recall may not be of equal importance, a fact that is accounted for by different flavors of F-Measures. F_1, the most commonly used metric, treats Precision and Recall identically, whereas F_2 weights Recall twice as strong as Precision. In the current work, F_2 turned out to be a more appropriate metric than F_1 because missing tokens of some of the topics (such as emergency callers, requests for follow-up, or billing disputes) are considered critical, i.e., missing instances of such topics are more expensive than false alarms. At any rate, since the above mentioned trivial majority vote classifier would not accept any tokens, its Recall would consequently be zero, so would be *any* F-Measure, including F_2.

4.3 Corpus and Experimental Results

For a large cable service provider [1] with a call volume of several million calls every month to its service hotline, we collected free-form online surveys as described in the introduction of this paper. The collected surveys amounted to about ten thousand every month. For a first proof of concept, we focused on a single month (May 2010) for which a number of 5589 randlomly selected surveys were annotated according to the scheme described in Section 3. We did not separate fixed training and test sets but instead used ten-fold cross-validation [3] in our experiments.

In a first round of experiments, we compared the performance of several state-of-the-art classifiers on this task. We selected the following classifiers from the WEKA toolbox [6] for this work:

- Ripper (a decision tree learner) [2],
- Sequential Minimal Optimization (SMO), a fast support vector machine implementation [9],
- naïve Bayes [4].

All these classifiers rely on sets of feature vectors and their associated class labels as training data, so, the survey text had to be converted into a feature representation. There are multiple techniques to represent utterances or texts in vector form, out of which we have been using the following ones:

- **wpres1.** Each vector element represents one word type in the vocabulary. For a specific text, all those elements representing words present in the respective text are `1`, all the others are `0`.
- **wpres5.** The same as **wpres1**, but only types whose total count in the training data is five or more are considered in the vector.
- **wcount1.** The same as **wpres1**, but instead of `1` to indicate the presence of a word in the text, the *count* of the word is used as element value.

- **wcount5.** The same as **wcount1** but discarding types with a total count of four or less.
- **bowpres1.** The same as **wpres1**, but before establishing vocabulary and vector elements, texts are converted into a bag-of-word representation, a compressed but semantically almost identical form of the text [10, 4].
- **bowpres5.** The same as **bowpres1** but discarding types with a total count of four or less.
- **tfidf1.** The same as **wpres1**, but the element values represent the text's words' TF-IDF scores [12].
- **tfidf5.** The same as **tfidf1**, but discarding types with a total count of four or less.
- **tfidfbow1.** The same as **tfidf1**, but after conversion into a bag-of-word representation.
- **tfidfbow5.** The same as **tfidfbow1**, but discarding types with a total count of four or less.

For the first experiment (to compare classifiers), we limited analysis to **tfidf1** features which are very common in information retrieval and data mining. We performed topic classification as well as joint classification of topics and emotion flags as discussed in Section 4.1.

At a first glance, the results seem to be slightly disappointing, with many results below 0.5 and even some 0. At this point, we have to remind the reader of the motivation behind using F_2 which was that a classifier can only be deemed useful when there is a Recall greater than 0 which means, at least one test sample has to be correctly identified. Given the extremely sparse and, at the same time, linguistically diverse set of examples for certain classes, it is almost impossible for a classifier to produce reasonable output. Nonetheless, this first experiment clearly indicates that the classification tree algorithm Ripper outperforms its competitors SMO and naïve Bayes and will therefore be used in the continuation of this project. Furthermore, we will use a consolidated score across classes (the weighted average as shown in the last row of Table 3) in order to help drawing conclusions more easily.

Looking at the joint classification of topics and emotion flags (in parenthesis in Table 3, classifier is Ripper), it is interesting how similar the results are to pure topic classification. For some of the topics, subdivision into more classes by adding emotion flags even results in a performance gain.

Results of our experiments to compare different feature vectors are shown in Table 4. Here, we used the Cost Sensitive Meta Classifier offered by WEKA which allowed us to optimize results towards our target metric F_2. This is why, this time, **tfidf1** achieved a higher score than in Table 3.

5 Conclusion

According to these results, the well-established TF-IDF metric performed lower than bag-of-word vectors. The absolute values of $F_2 = 0.71$ indicate that the

Table 3. Comparing classifiers for topics and emotion tags. In bold, results (F_2) greater than 0.5.

topic	Ripper (±)		SMO	naïve Bayes
A	**0.53**	(0.36)	0.04	0.02
B	0.23	(0)	0	0
C	**0.61**	**(0.58)**	0.08	0.32
D	0.20	(0.32)	0.01	0.05
E	0.25	(0.22)	0.05	0.19
F	0.19	(0.24)	0	0.02
H	**0.83**	**(0.85)**	**0.89**	**0.81**
I	0	(0)	0	0
O	0.31	(0.33)	0.24	0.49
P	0.26	(0.32)	0.11	0.50
R	0.15	(0.15)	0	0
S	0.08	(0.17)	0	0.02
T	**0.58**	**(0.61)**	0.07	0.22
V	0.10	(0)	0	0
W	**0.57**	**(0.59)**	0.07	0.47
avg	**0.63**	**(0.61)**	0.22	0.42

Table 4. Comparing features. Winners in bold.

feature	Precision	Recall	F_2
wpres1	0.56	0.75	0.70
wpres5	0.54	0.72	0.67
wcount1	0.50	0.74	0.67
wcount5	0.58	0.72	0.68
bowpres1	0.71	0.71	**0.71**
bowpres5	0.60	0.74	**0.71**
tfidf1	0.65	0.65	0.65
tfidf5	0.50	0.69	0.64
tfidfbow1	0.72	0.7	0.70
tfidfbow5	0.80	0.66	0.69

technique can indeed be useful when trying to detect infrequent surveys of specific topics in large amounts of data. Taking **bowpres5** as example: A Recall of 0.74 means that the classifier is missing only 26% of the topic's surveys. In contrast, a Precision of 0.6 means that 60% of the surveys returned by the classifier actually referred to the topic. Without classification, this percentage would be much much lower, e.g. $< 10\%$ for most of the topics shown in Table 2. Hence, topic classification as preprocessing step can significantly reduce the manual workload associated with the screening of tens of thousands of surveys every month specifically for rare topics and emotion flags.

References

1. Acomb, K., Bloom, J., Dayanidhi, K., Hunter, P., Krogh, P., Levin, E., Pieraccini, R.: Technical Support Dialog Systems: Issues, Problems, and Solutions. In: Proc. of the HLT-NAACL. Rochester, USA (2007)
2. Cohen, W.: Fast Effective Rule Induction. In: Proc. of the International Conference on Machine Learning. Lake Tahoe, USA (1995)
3. Devijver, P., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, USA (1982)
4. Evanini, K., Suendermann, D., Pieraccini, R.: Call Classification for Automated Troubleshooting on Large Corpora. In: Proc. of the ASRU. Kyoto, Japan (2007)
5. Gorin, A., Riccardi, G., Wright, J.: How May I Help You? Speech Communication 23(1/2) (1997)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
7. Hone, K., Graham, R.: Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI). Natural Language Engineering 6(3-4) (2000)
8. Neustein, A.: Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics. Springer, New York, USA (2010)
9. Platt, J.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Tech. rep., Microsoft Research, Seattle, USA (1998)
10. Porter, M.: An Algorithm for Suffix Stripping. Program 14(3) (1980)
11. van Rijsbergen, C.: Information Retrieval. Butterworths, London, UK (1979)
12. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw Hill, New York, USA (1983)
13. Suendermann, D.: Advances in Commercial Deployment of Spoken Dialog Systems. Springer, New York, USA (2011)
14. Suendermann, D., Hunter, P., Pieraccini, R.: Call Classification with Hundreds of Classes and Hundred Thousands of Training Utterances ... and No Target Domain Data. In: Proc. of the PIT. Kloster Irsee, Germany (2008)
15. Suendermann, D., Liscombe, J., Dayanidhi, K., Pieraccini, R.: A Handsome Set of Metrics to Measure Utterance Classification Performance in Spoken Dialog Systems. In: Proc. of the SIGdial Workshop on Discourse and Dialogue. London, UK (2009)
16. Walker, M., Litman, D., Kamm, C., Abella, A.: PARADISE: A General Framework For Evaluating Spoken Dialogue Agents. In: Proc. of the ACL. Madrid, Spain (1997)

Acoustic analysis of politeness and efficiency in a cooperative time-sensitive task

Marcela Charfuelan[1], Paul M. Brunet[2]

[1] DFKI GmbH, Language Technology Lab
Alt-Moabit 91c, D-10559, Berlin, Germany
{marcela.charfuelan}@dfki.de
[2] School of Psychology, Queen's University Belfast, United Kingdom
{p.brunet}@qub.ac.uk

Abstract. We present an acoustic analysis of politeness and efficiency in a cooperative time-sensitive task experiment. In the experiment sixteen dyads completed 20 trials of the "Maze Task", where one participant (the *navigator*) gave oral instructions for the other (the *pilot*) to follow. For half of the trials, navigators were instructed to be polite, and for the other half to be efficient. We investigate what are the main acoustic factors that are associated with greater politeness in the polite condition and lesser politeness in the efficient condition.

Keywords: Prosody, Voice quality, Vocal social signals, Politeness, Acoustic measures, Acoustic correlates

1 Introduction

Detection, analysis and synthesis of social signals are topics increasingly applied in computing technologies. Sensitive Artificial Listeners (SAL), which are machines that possess some social and emotional intelligence capabilities [7], pedagogical agents that exhibits social intelligence [10] or predictors of behavioural outcomes in social situations [8] are just some examples where social signals play and important role.

Social signals like politeness, empathy, hostility, (dis)-agreement and any other stances towards others, can be expressed through verbal and non-verbal means in different modalities [9]. One of these modalities is *vocal nonverbal behaviour* – not *what* is said, but *how* it is said. This includes prosodic features such as pitch, energy and rhythm, as well as voice qualities such as harsh, creaky, tense, etc. Brown and Levinson [1] predicted that sustained high pitch (maintained over a number of utterances) will be a feature of negative-politeness usage, and creaky voice a feature of positive-politeness usage, and that a reversal of these associations will not occur in any culture.

Social signals, like politeness, typically occur in interactions among people; this makes it natural to study them in corpora of spontaneous interactions rather than in material produced by an actor out of context [4]. In this study we analyse the recordings of a cooperative time-sensitive task experiment designed to study vocal expression of politeness and efficiency [2]. In the experiment sixteen dyads completed 20 trials of the "Maze Task", where one participant (the *navigator*) gave oral instructions (mainly "up", "down", "left", "right") for the other (the *pilot*) to follow. For half of the trials, navigators were instructed to be polite, and for the other half to be efficient. In this experiment, task accuracy is an objective measure calculated by the distance from the cursor position at the end of the trial and the end point.

In a preliminary analysis of the experiment, it was found that although the task was very simple and users had few ways to express politeness, it significantly affected task accuracy and pilots's subjective ratings indicate that it was perceived [2]. So in this paper we investigate what are the main acoustic factors that are associated with greater politeness in the polite condition and lesser politeness in the efficient condition. We use Principal Component Analysis (PCA) to analyse possible clusters on the data and multiple linear regression to find the acoustic features that better predict task accuracy. If the task accuracy is systematically affected by the politeness/efficiency condition we would like to know whether there are predominant acoustic features in each condition.

The paper is organised as follows. In Section 2 we start describing the experiment, data and methodology used in this study. Then in Section 3 we briefly describe the acoustic measures extracted from the data. Results are presented in Section 4 and conclusions in Section 5.

2 Data and method

The study consisted of participants engaging in a cooperative task with a partner. The participant was positioned in front of a computer monitor in one room, while the partner was in a second room. The assigned task was a computerized maze task requiring the dyad to guide the cursor from the starting point of the maze to the endpoint. The participant could see the maze on the computer monitor but did not have the means to directly move the cursor. The other dyad member could not see the maze (instead they saw the participants face via a webcam) but with the arrow keys of the keyboard could move the cursor. The dyad could communicate via microphones and speakers. Consequently, the participant had the role of navigator and was responsible for verbally guiding the partner's cursor movements. In total the dyad completed 20 trials. The experimental trials were broken into 4 blocks of 5 trials. In each block, the trials became increasingly more difficult by increasing the black squares by 5%, also for the second and fourth blocks the vertical and horizontal cursor controls were flipped (participants were informed of this change). For the first 10 trials, the participant was instructed to be polite, the second 10 trials to be efficient. Half of the participants were instructed to be efficient first, then polite for the second part. The trials were time sensitive (less than a minute allotted) and errors (i.e. hitting the walls) decreased the allotted time limit.

The blocks and trials of every session and the words or command words used by the navigators were manually segmented. Acoustic features were extracted from these small segments and averaged if the extracted measure is frame based. The distribution of data is presented in Table 1, due to technical problems with the recordings we have analysed 14 of the 16 dyads, corresponding to 4 male and 10 female navigators. In this table the data has been split according to the difference score (Diff score) between the average accuracy scores of the polite and efficient sessions.

Table 1: Distribution of data. Diff score is the difference between the task accuracy score obtained on the polite condition and the score obtained on the efficient condition.

	Diff score > 10		Diff score ≤ 10		
Condition	female	male	female	male	Total
efficient	1127	379	958	562	3026
polite	1452	517	959	382	3310

For the analysis of the data, first we use Principal Component Analysis (PCA) to analyse possible clusters on the data and the two conditions. Then we perform multiple linear regression using the task accuracy score of each trial as objective measure and several acoustic features as explanatory variables. We search for the acoustic features that better predict the accuracy score of each trial using ten repetitions of ten-fold sequential floating forward selection - multiple linear regression (SFFS-LM).

3 Acoustic measures

The acoustic measures used in this study are described in detail in [3], here we mention them briefly:

1. Low level acoustic measures
   - Voicing strengths: full-band and multi-band: str, str1, str2, str3, str4, str5
   - Pitch harmonics magnitude: first ten magnitudes: mag1...mag10.
   - Spectral features: Melcepstrum coefficients (mcep0...mcep24), Spectral entropy (full-band and multi-band: spec_entropy, spec_entropy1,..., spec_entropy5)
   - Articulatory-based features: Formants, Formant bandwidths, Formant dispersion
2. Prosody acoustic measures
   - Fundamental frequency or pitch
   - Pitch entropy (calculated as the spectral entropy)
   - maximum, minimum, and range of f0
   - Duration of the utterance in seconds
   - Voicing rate calculated as the number of voiced frames per time unit
   - Energy, calculated as the short term energy $\sum x^2$
3. Voice quality acoustic measures
   - Hamm_effort = LTAS_{2-5k}
   - Hamm_breathy = (LTAS_{0-2k} - LTAS_{2-5k}) - (LTAS_{2-5k} - LTAS_{5-8k})
   - Hamm_head = (LTAS_{0-2k} - LTAS_{5-8k})
   - Hamm_coarse = (LTAS_{0-2k} - LTAS_{2-5k})
   - Hamm_unstable = (LTAS_{2-5k} - LTAS_{5-8k})
   - slope_ltas: least squared line fit of LTAS in the log-frequency domain (dB/oct)
   - slope_ltas1kz: least squared line fit of LTAS above 1 kHz in the log-frequency domain (dB/oct)
   - slope_spectrum1kz: least squared line fit of spectrum above 1 kHz (dB/oct)

Low level acoustic measures are extracted at frame level, with a frame length of 25 ms. and a frame shift of 5 ms. The frame based measures are averaged per word. Prosody features are classical features related to pitch, energy, duration, etc. And voice quality measures are measures mostly used in emotion research. Prosody and voice quality measures are extracted at word level.

4 Results

4.1 PCA analysis

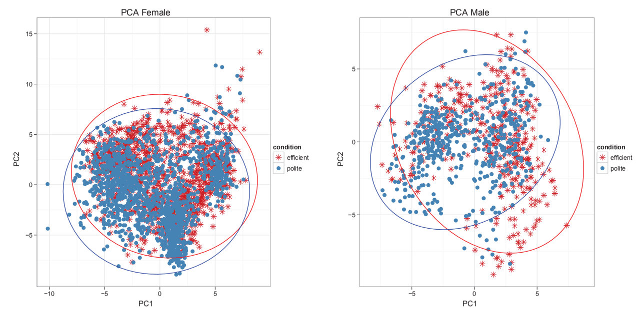

Fig. 1: PCA analysis of male and female data with Diff score > 10. The first two PCs in female data explain 32% of the variance and in male data the first two PCs explain 24% of the variance.

Since we do not have perceptual annotations of how polite the users were when they were asked to be polite, just their subjective impressions collected through a questionnaire, for the first experiment with PCA we selected the sessions where the difference score is high. That is, the sessions where the task accuracy score obtained on the polite condition was higher than the score on the efficient condition (in this experiment a high score means low accuracy). As we mentioned in the introduction, in a preliminary study it was already detected a consistent acoustic separation in individual sessions where the polite and efficient scores were very different.

In Figure 1 a scatter plot of the first two principal components of the PCA analysis is presented. In this analysis we have used all the acoustic features and the data where the Diff score is > 10 (see Table 1). We expected that the clusters were more apparent when there is a big score difference between the polite and efficient condition. An ellipse in these figures indicate clusters of words used used during the polite and efficient sessions. The clusters for male data seem to be more separated than for female data, but there is also less male speakers in this data. PC1 in both cases separate better the clusters.

Table 2: Main loadings for acoustic features for the male and female PCA analysis presented in Figure 1.

Female PCA				Male PCA			
PC1		PC2		PC1		PC2	
Feature	score	feature	score	Feature	score	feature	score
spec_entropy1	-0.22	spec_entropy4	-0.23	mcep5	-0.22	spec_entropy1	-0.24
spec_entropy	-0.20	mcep2	-0.21	mcep18	-0.21	mcep7	-0.21
mcep6	-0.20	Hamm_breathy	-0.21	str4	-0.21	mcep6	-0.20
mcep11	-0.20	spec_entropy5	-0.20	mcep21	-0.19	mcep2	-0.19
...	...	...	...	...	...	...	...
mcep0	0.17	str4	0.21	voicing_rate	0.18	Hamm_effort	0.20
formant_disp	0.18	logpow	0.22	mcep1	0.20	pitch_entropy	0.21
pitch_entropy	0.19	Hamm_effort	0.23	B4	0.22	str1	0.22
voicing_rate	0.21	str3	0.24	spec_entropy4	0.27	mcep0	0.27

The higher positive and negative loadings of the PCA analysis are presented in Table 2. For PC1 mostly spectral features are the more loaded and also voicing rate. For PC2 spectral features, voicing strengths and voice quality features are highly loaded. Is interesting to notice that prosody features did not appear as good discriminators of the two conditions. An analysis of variance of these measures (one way ANOVA) indicates that almost all the measures are significantly different between polite and efficient condition with p-value < 0.001 except for str4, mcep6 and Hamm_effort on the male data.

4.2 SFFS-LM analysis

In Table 3 the features that best predict task accuracy for male and female data are presented. In this case all the data was used irrespective of the difference score. If task accuracy is systematically affected by the politeness/efficiency condition we would like to know whether there are predominant acoustic features in each condition. In this case task accuracy in the polite condition seem to be better predicted by prosody features like max_f0, min_f0, std_f0, energy, and also some spectral features. Task accuracy in the efficiency condition seems to be less dependent on prosody features. An analysis of variance of these measures showed that most of these measures are not significantly different between the two classes polite and efficient. Here again the spectral features are more significantly different among the two conditions.

Table 3: Main acoustic predictors of accuracy for all the data. In parentheses is indicated the prediction error for each case. p-value after ANOVA of measures between the two classes polite and efficient is indicated by the significance codes: ***<0.001, **<0.01, *<0.05, . <0.1, ∘ <1.

Predicted accuracy Female				Predicted accuracy Male			
Polite (14.3%)		Efficient (13.95%)		Polite (4.85%)		Efficient (5.15%)	
mcep23	***	str2	∘	std_f0	**	mcep13	∘
max_f0	.	spec_entropy1	***	min_f0	*	std_f0	∘
spec_entropy2	**	spec_entropy2	**	energy	***	energy	***
min_f0	∘	mag2	∘	mcep10	∘	mcep4	***
mag1	*	mcep23	***	mcep18	***	spec_entropy4	***
std_f0	∘	min_f0	∘	mcep0	∘	str	∘
pitch_entropy	.	pitch_entropy	.	mcep6	***	str5	***
spec_entropy1	***	str1	***	pitch_entropy	∘	min_f0	*
mcep1	∘	max_f0	∘	str	∘	mcep7	***
str1	***	mcep5	***	mcep16	***	mcep10	∘

5 Conclusions

In this paper we have presented an acoustic analysis of politeness and efficiency in a cooperative time-sensitive task experiment.

In the PCA experiment we have found not so clear clusters or tendencies on the data analysed, although some individual sessions present clear clusters. One explanation could be that actually for some speakers there is no acoustic difference between the two conditions. In that case it would be necessary to perceptually annotate the words in the sessions so we can be sure that at perception level some words actually sound polite. This is in fact an important issue when analysing social signals, affect or emotions in spontaneous interactions, since it is not easy to find relevant speech material that includes corresponding perceptual annotations.

In the SFFS-LM experiment we have found that task accuracy in the polite condition is better predicted by prosody features and task accuracy in the efficient condition seems to be less dependent on prosody features. This result seems to be more in line with the general tendency described on the literature that pitch is a good predictor of politeness [1, 5]. However, the analysis of variance of the features that better predict task accuracy showed that these features do not discriminate well among the two conditions polite and efficient. So we can not conclude that the politeness condition was the only (or main) factor that affected task accuracy. One hypothesis, that will be analysed in future work, is that in the experiment task accuracy would have been also affected by task or cognitive load.

During the maze task, the trials in a block became increasingly more difficult, and in the second and fourth blocks the cursor controls were flipped. The participants were informed about this change so they have to concentrate more on these blocks. In the literature it has been reported that speech rate, energy contour, F0 and spectral parameters are correlated with task load and stress [6], so we will analyse whether these features discriminate different levels of task load among the four blocks of the experiment.

Acknowledgements. The research leading to these results has received funding from the EU Programme FP7/2007-2013, under grant agreement no. 231287 (SSPNet).

References

1. Brown, P., Levinson, S.C.: Politeness some universals in language usage. Cambridge University Press (1987)
2. Brunet, P., Charfuelan, M., Cowie, R., Schröder, M., Donnan, H., Douglas-Cowie, E.: Detecting politeness and efficiency in a cooperative social interaction. In: Proc. Interspeech. Makuhari, Japan (2010)
3. Charfuelan, M., Schröder, M.: The vocal effort of dominance in scenario meetings. In: Proc. Interspeech. Florence, Italy (2011)
4. Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J., Devillers, L., Abrilian, S., Batliner, A., Amir, N., Karpouzis, K.: The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In: Affective Computing and Intelligent Interaction, pp. 488–500 (2007)
5. Grawunder, S., Winter, B.: Acoustic correlates of politeness: prosodic and voice quality measures in polite and informal speech of korean and german speakers. In: Proc. Speech Prosody 2010. Chicago, Illinois, USA (2010)
6. Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, T.: Acoustic correlates of task load and stress. In: Proc. Interspeech. ISCA, Denver, Colorado, USA (2002)
7. Schöder, M., McKeown, G.: Considering social and emotional artificial intelligence. In: Proc. AISB 2010 Symposium "Towards a Comprehensive Intelligence Test". Leicester, UK. (2010)
8. Soman, V., Madan, A.: Social signaling: Predicting the outcome of job interviews from vocal tone and prosody. In: Proc. IEEE ICASSP. Dallas, Texas, USA (2010)
9. Vinciarelli, A., Salamin, H., Pantic, M.: Social signal processing: Understanding social interactions through nonverbal behavior analysis. In: IEEE Computer Vision and Pattern Recognition Workshops. pp. 42–49 (2009)
10. Wang, N., Johnson, W.L., Mayer, R.E., Rizzo, P., Shaw, E., Collins, H.: The politeness effect: Pedagogical agents and learning gains. Frontiers in Artificial Intelligence and Applications 125, 686–693 (2005)

Multidimensional meaning annotation of listener vocalizations for synthesis

Sathish Pammi, Marc Schröder, and Marcela Charfuelan

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany and
Alt-Moabit 91c, D-10559, Berlin, Germany
{firstname.lastname}@dfki.de

Abstract. Listener vocalizations convey affective and epistemic states behind the listener's intentions while the interlocutor is talking. The meaning annotation of such vocalizations is a crucial step in synthesis of listener vocalizations. This paper presents a perception study to annotate meaning of vocalizations. In this study, subjects annotate (characterize) a set of listener vocalizations using a multi-dimensional set of meaning descriptors. The set of stimulus vocalizations is selected based on intonation clustering. We investigate the typical impressions and the appropriateness of meanings conveyed by vocalizations, based on high agreement ratings provided by the participants. We also discuss the suitability of the annotation procedure to generate expressive listener vocalizations.

Keywords: listener vocalizations, perception study, meaning, speech synthesis

1 Introduction

Nowadays spoken and multimodal dialogue systems attempt to model the computer's part of the dialogue in both the speaker and the listener role [12, 16]. That means the machine must emit signs of listening while the user is speaking: backchannels [19] or expressive feedback signals [1]. In multimodal dialogue systems, some of these signals can be visual, such as head nods, smiles, or raised eyebrows [5]; in the vocal channel, backchannel and feedback signals can be realized as listener vocalizations. Listener vocalizations like *mhm, right, yeah, uh-huh* are not only produced to make the interaction more natural but also to signal affective meanings such as *anger, amusement* and epistemic meanings such as *interested, agreeing*.

Yngve [19] investigated responses such as *uh-huh*, *yes*, *okay*; he called them as "behavior in the back channel". Duncan [6] attempted to correlate meaning with segmental forms like *yeah*, *right* and *I see*; whereas Schegloff [17] and McCarthy [10] noted the multifunctioning of vocalizations. Later studies [8, 18] indicates that several behavior properties like segmental form, intonation, voice-quality have influence on the meaning conveyed by vocalizations.

Although several studies attempted to understand meanings of vocalizations, there has been not much focus on how these vocalizations can be used for synthesis. An integrative account of all these studies must be considered in a bigger picture. It requires the following sequence of steps: (i) identification of suitable meaning descriptors; (ii) annotation of appropriateness for each meaning descriptor; (iii) identifying a typical impression of meanings for each vocalization; (iv) analyzing the impact of behavioral properties like segmental form and intonation on perceived meaning. We attempt the above steps in this paper.

In order to synthesize an appropriate listener vocalization, we require two kinds of information about each of the available vocalizations [13]: a typical impression of the meaning that the vocalization could convey; and how appropriate is the vocalization for a given meaning. In this paper, we experiment a methodology to find meanings of vocalizations that are usable for synthesis. We conduct a listening test where subjects annotate (characterize) a set of listener vocalizations using a multi-dimensional set of meaning descriptors.

Considering the possibility to improve acoustic variability using imposed intonation contours [14], we also investigate the relevance of intonation and segmental form on the perceived meaning. This motivates the procedure of stimuli selection for the experiment. The paper is organized as follows. In Section 2 the vocalizations database used in this study is described. Section 3 describes our meaning descriptors used in this study. In Section 4 our approach to select representative vocalizations is explained. In this section the perception experiment is also explained. In Section 5 main results are discussed and in Section 6 findings are summarized.

2 Vocalizations database

To collect natural listener vocalizations from dialogue speech, we recorded about half an hour of free dialogue with professional British actors. Four British actors were selected for four Sensitive Artificial Listener (SAL) voices: cheerful (Poppy), neutral (Prudence), gloomy (Obadiah), and aggressive (Spike) voices. The British actors were originally chosen for the recordings required for building new TTS voices. In addition to speech synthesis recordings, free dialogue of around 30 minutes was recorded with each of the British speakers. The recording setup and instructions given to the actors are described in [15].

	Prudence	Poppy	Spike	Obadiah
Corpus duration (in minutes)	25	30	32	26
number of vocalizations	128	174	94	45

Table 1: British English listener vocalizations recorded for the four SAL characters

Once the dialogue was recorded for all four characters, listener vocalizations were marked on the time axis and transcribed as a single (pseudo-)word, such as *myeah* or *(laughter)*. With respect to the number of listener vocalizations they produced the speakers varied enormously. Whereas Obadiah produced only 45 vocalizations, Poppy produced 174 (see Table 1).

3 Meaning descriptors

We started by establishing a list of meaning dimensions, based on three sources: the most frequent categories in an exploratory annotation study on German listener vocalizations [15]; the most frequently used annotations of the SEMAINE corpus [11] – a large and annotated collection of dialogue of the SAL domain; and a set of affective-epistemic descriptors used to describe visual listener behavior [4].

Descriptors	**Scale type**	**Source**
anger	unipolar	Emotional categories
sadness	unipolar	
amusement	unipolar	
happiness	unipolar	
contempt	unipolar	
solidarity	unipolar	IPA categories
antagonism	unipolar	
(un)certain	bipolar	Baron-Cohen's categories
(dis)agreeing	bipolar	
(un)interested	bipolar	
(high/low)anticipation	bipolar	

Table 2: Consolidated list of meaning descriptors used in this study

The three sources were consolidated into a list of 11 descriptors as shown in Table 2. The table shows the scale type (unipolar/bipolar) of meaning descriptors. We made sure that these categories are derived from three different backgrounds, emotional categories [7], Baron-Cohen's epistemic mental states [3] and Bales Interaction Process Analysis (IPA) [2]. Whereas epistemic states can be used to transmit attitudinal mental states of listener, IPA labels can be used to convey social meanings in dialogue.

4 Approach

This section describes our approach to annotate meanings of listener vocalizations. Annotation of meaning for all listener vocalizations is a tedious and time consuming process. Instead, annotation of selective vocalizations would be more cost effective. As literature [8, 18] suggests that the meaning of vocalization highly correlates with segmental form and intonation, we propose a semi-automatic procedure to select representative vocalizations of segmental forms and intonation contours in the corpus. This also facilitates us to investigate the relevance of segmental form and intonation on the perceived meaning.

4.1 Stimuli selection

The stimuli are selected based on a semi-automatic clustering of intonation contours. For clustering vocalizations according to intonation, a contour was automatically computed for each vocalization by fitting a 3rd-order polynomial to f0 values extracted using the Snack pitch tracker [9]. Polynomials can approximate intonation contours of speech signal in unvoiced regions. Separately for each speaker, we used K-means clustering of intonation contours to identify the vocalizations with a similar intonation.

Two sets of stimuli were manually extracted from the clustered data for the purpose of selecting representative vocalizations that cover the maximum number of possible segmental forms and intonation contours. We aimed for two sets that contain, on one hand, stimuli with the same segmental form (as determined from the single-word description) varying in intonation (identified in the following as *fixed segmental form*); and on the other hand, stimuli with the same intonation (flat intonation contour) and varying in segmental form (henceforth, *fixed intonation contour*). Thus we manually selected samples from clusters as follows: (i) in order to get wide range of contour shapes, we selected one or two representative samples from each cluster with same segmental form (i.e. *yeah*); (ii) we selected samples with different segmental forms from a single cluster where contour shape is constant. Table 3 shows the number of selected stimuli for the experiment.

Character	*Fixed segmental form*	*Fixed intonation contour*
Poppy	15	8
Spike	10	9
Obadiah	5	8
Prudence	8	9
Total	38	34

Table 3: Character wise number of vocalizations selected for meaning annotation

4.2 Perception experiment

Scale-based ratings capture inherent ambiguity more than forced-choice test. We designed a web-based perception study for participants. The first page provided instructions, the second page collected demographic information and the following pages present the audio and rating scales one at a time, as shown in Figure 1. The stimuli were presented to the participants in a random order for eliminating order and fatigue effects. Participants could play the audio as many times as they liked before providing meaning ratings. A 5-points Likert scale for each meaning was used: from 1 (absolutely no attribution) to 5 (extremely high attribution) for unipolar meaning categories; from -2 (extremely negative attribution) to +2 (extremely positive attribution) for bipolar meaning categories. "No Real Impression" option was provided for each meaning scale in case the participant is unsure.

44 participants (20 women, 24 men) took part in the annotation study. 22 participants provided ratings for the vocalizations in test set *fixed segmental form* (9 women, 13 men) and 22 participants rated vocalizations in test set *fixed intonation contour* (11 women, 11 men).

5 Results and discussion

In order to study each of the vocalizations per meaning, we first introduce the term *meaning-vocalization* combination that is used in the rest of this paper. Each vocalization can convey maximally 11 meanings used in the corpus annotation. One stimulus

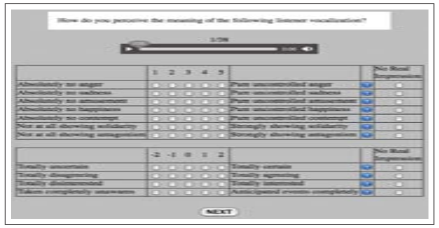

Fig. 1: A screenshot of the web page for the perception study

indicates 11 *meaning-vocalization* combinations. For example, in the case of Prudence (see Table 4), 187 *meaning-vocalization* combinations (17 stimuli * 11 meaning categories) were available for analysis.

5.1 High versus Low agreement

Table 4 shows the high variability on agreement of *meaning-vocalization* combinations for Prudence. In this table high agreement is identified with circles or arrows and low agreement is identified with a dot (·). In order to identify high agreement versus low agreement of *meaning-vocalization* combinations, we computed the interquartile range (IQR) of ratings provided for each combination. We considered that a combination has high agreement if the IQR of the combination is less than one third of the meaning scale range. In other words, a combination has high agreement if more than 50% of the raters agree within one third of the meaning scale range. The high agreement combinations indicates typical impression of the meaning on the vocalization.

Table 4 shows that the number of low agreement annotations (identified as ·) are higher in the *fixed intonation contour* set when compared to the *fixed segmental form* set for Prudence. The same tendency was observed when taking into account all the vocalizations in our corpus, that is 792 (72 stimuli * 11 categories) *meaning-vocalization* combinations, from which 418 combinations belong to the *fixed segmental form* set and 374 belong to the *fixed intonation contour* set. Figure 2 shows a global picture of high agreement versus low agreement combinations for all the corpus. While around 60% of the *fixed segmental form* combinations show high-agreement, only 40% of the *fixed intonation contour* combinations show high-agreement. This seems to indicate that the participants perceived more distinguishable information from intonation when compared to segmental form. In other words, this evidence indicates that the intonation contour is highly relevant for signaling meaning when compared to phonetic segmental form.

5.2 Appropriateness of high agreement annotations

Not all vocalizations with high agreement may be suitable to convey a specific meaning for synthesis. In this work the suitability of a *meaning-vocalization* combination is calculated by computing the median of ratings provided for that combination. However, we can not conclude about suitability of low agreement ratings.

Fixed segmental form													Fixed intonation contour												
segmental form	intonation-contour	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation	segmental form	intonation-contour	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah		○	·	○	○	·	↑	○	↑	↑	○	○	tsyes		·	·	·	·	·	·	·	↑	○	·	·
yeah		○	⇑	○	○	○	·	○	○	○	·	○	tsyeah		·	·	·	○	·	·	·	·	·	·	○
yeah		○	·	○	○	·	·	○	↑	↑	○	↑	mhm		·	·	○	○	·	↑	·	·	○	·	·
yeah		○	○	·	·	○	⇑	○	↑	↑	↑	·	yeah		·	·	○	○	·	·	·	↑	·	·	·
yeah		○	·	○	○	○	·	·	○	↑	·	·	yes		·	·	○	○	·	·	·	○	○	○	↑
yeah		○	○	⇑	⇑	○	⇑	○	↑	↑	↑	·	right		·	○	○	○	·	·	·	·	·	·	·
yeah		○	·	○	○	○	·	○	○	○	↓	○	tsright		·	·	○	○	·	·	·	↑	·	·	○
yeah		○	·	○	○	·	↑	·	·	·	↓	○	aha		○	○	·	·	·	⇑	○	↑	↑	↑	↑
													tsgosh		○	○	·	·	·	·	○	○	○	○	○

Table 4: Segmental form, intonation contour and meaning of Prudence's stimuli. *Meaning-vocalization* combination is represented using the following symbols.
○: vocalization is not appropriate for the meaning;
↑ or ↓ : vocalization is somewhat appropriate;
⇑ or ⇓ : vocalization is very appropriate for the meaning;
·: the annotation has low agreement (we can not conclude on appropriateness);
↓ and ⇓ : negative sides of bipolar scales

We distinguish three levels of appropriateness based on where the participants tend to agree on the meaning scale. A *meaning-vocalization* combination is very appropriate if the participants tend to agree on positive (in case of unipolar and bipolar scales) or negative (in case of bipolar scale) end of meaning scale. The combination is not appropriate to convey the meaning if they tend to agree on '0'. In other words, we can say that the combinations are "very appropriate", "somewhat appropriate", and "not appropriate" when the median is greater than two third of meaning scale, between one third and two third, and less than one third respectively. Among high-agreement *meaning-vocalization* combinations available in our corpus, it was found that, 7.2% (30) are very appropriate, 22.4% (93) are somewhat appropriate, and 70.4% (293) are not appropriate combinations. This result is highly relevant in speech synthesis, that is, one vocalization can be "not appropriate", "somewhat appropriate" or "very appropriate" for several different meanings at the same time. These three categories can be used, for example, in an algorithm for unit-selection synthesis (i.e. vocalization selection) that considers

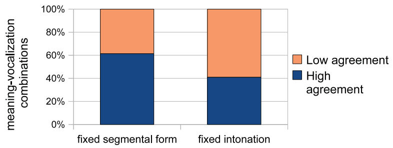

Fig. 2: Percentage of high and low agreement *meaning-vocalization* combinations

appropriateness to realize a particular intended (target) meaning. The evaluation of such unit-selection algorithm has been presented in [13].

5.3 Inherent ambiguity of listener vocalizations

According to Table 4, the vocalization *aha* can convey 5 meanings (*solidarity, certain, agreeing, interested, anticipation*), whereas the vocalization *right* does not convey any meaning available in our descriptors. Figure 3 shows the histogram of possible meanings for the listener vocalizations in our corpus. Among 72 stimuli, 14 vocalizations (19.5%) convey no meaning, 27 (37.5%) convey single meaning, and the remaining 31 (43%) convey multiple meanings. On average, a single vocalization in this corpus can convey 1.68 meanings, this confirms the argumentations already made in the literature [10,17]. Indeed the inherent ambiguity of listener vocalizations is a very interesting feature to exploit in speech synthesis, because a single vocalization can be used in multiple instances.

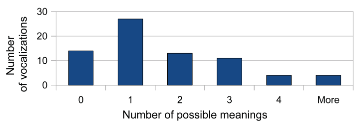

Fig. 3: Histogram of multiple meanings

6 Conclusion

In this paper, we explored a multi-dimensional annotation methodology to annotate listener vocalizations in view of conversational speech synthesis. We conclude the following issues from this study: (i) this methodology can provide a typical impression of meanings from high agreement annotations; (ii) unit-selection algorithms can benefit from the annotation of meaning on scales: it captures appropriateness of listener vocalizations for a given meaning; (iii) one vocalization can convey several meanings, which is useful for the usage of the same vocalization in several instances; (iv) the evidence indicates that the intonation contour is highly relevant for signaling meaning when compared to the phonetic segmental form - in support for improving acoustic variability using imposed-intonation contours.

7 Acknowledgements

This research has received funding from the European Community's Seventh Framework Programme (FP7-ICT) under grant agreement no. 211486 (SEMAINE), 248116 (ALIZ-E) and 231287 (SSPNet). We would like to thank Professor Roddy Cowie and Dr. Gary McKeown for useful discussions.

References

1. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9(1), 1–26 (1992)
2. Bales, R.: Interaction process analysis. Cambridge, Mass (1950)
3. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.: Mind Reading: The Interactive Guide to Emotions. Jessica Kingsley Publishers, London (2004)
4. Bevacqua, E., Heylen, D., Pelachaud, C., Tellier, M.: Facial feedback signals for ECAs. In: AISB 2007 Annual convention, workshop "Mindful Environments". Newcastle, UK (2007)
5. Bevacqua, E., Pammi, S., Hyniewska, S., Schröder, M., Pelachaud, C.: Multimodal backchannels for embodied conversational agents. In: IVA 2010. Philadelphia, USA (2010)
6. Duncan, S.: On the structure of speaker–auditor interaction during speaking turns. Language in society 3(02), 161–180 (1974)
7. Ekman, P., Dalgleish, T., Power, M.: Handbook of cognition and emotion. Chihester, UK: Wiley (1999)
8. Kowtko, J.: The function of intonation task-oriented dialogue (1996)
9. KTH: The snack sound toolkit. http://www.speech.kth.se/snack (2006)
10. McCarthy, M.: Talking back:" small" interactional response tokens in everyday conversation. Research on Language & Social Interaction 36(1), 33–63 (2003)
11. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: Proc. IEEE ICME 2010. Singapore (2010)
12. Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive ECA system. In: Proc. AAMAS. p. 1399–1400 (2009)
13. Pammi, S., Schröder, M.: Evaluating the meaning of synthesized listener vocalizations. In: INTERSPEECH 2011 (2011)
14. Pammi, S., Schröder, M., Charfuelan, M., Türk, O., Steiner, I.: Synthesis of listener vocalisations with imposed intonation contours. In: SSW7 Workshop. Kyoto, Japan (2010)
15. Pammi, S., Schröder, M.: Annotating meaning of listener vocalizations for speech synthesis. In: Proc. Affective Computing & Intelligent Interaction. Amsterdam, The Netherlands (2009)
16. Pfleger, N., Alexandersson, J.: Modeling non-verbal behavior in multimodal conversational systems. Information Technology 46(6), 341–345 (2004)
17. Schegloff, E.: Discourse as an interactional achievement: Some uses of "uh-huh" and other things that come between sentences. Analyzing discourse: Text and talk 71(93) (1982)
18. Ward, N.: Non-lexical conversational sounds in american english. Pragmatics &# 38; Cognition 14(1), 129–182 (2006)
19. Yngve, V.H.: On getting a word in edgewise. In: Chicago Linguistic Society. Papers from the 6th regional meeting. vol. 6, pp. 567–577 (1970)

An experimental triangulative research design for analyzing consumer behavior

Y.Zajontz, V.Kollmann, M.Kuhn, D.Reichardt

Duale Hochschule Baden-Württemberg Stuttgart
Baden-Wuerttemberg Cooperative State University Stuttgart
70178 Stuttgart, Germany
{zajontz, kollmann, kuhn, reichardt}@dhbw-stuttgart.de

Abstract. The first couture house Yves Saint Laurent or the leading fashion designer Tom Ford are only a few that create print advertisements which particularly go back to nudity and erotism as a mean of design. Despite controversial discussions sensual elements, romantic themes or sexual illustrations have become almost commonplace in advertising campaigns. The use of erotic stimuli in advertisements seems especially interesting for products which must compete strongly for consumer's attention.

The literature proposes different theoretical models through which the effects of these stimuli may be understood. The Stimulus-Organism-Response (SOR) framework is used as the basis of the research project to explore and to identify the emotional states (intervened variables) which influence various dimension of purchase behavior (see Kroeber-Riel/Weinberg, 2008, S.30 et seqq). This outlines the following main research question of the related project: "What kind of emotional reactions can be derived by erotic and sexual stimuli in advertisements?" The first attempt is to identify how persons respond to different kind of stimuli by classifying the intensity of brain waves.

Only little research has been done to measure the impact of sexual-oriented advertisements on consumer's attitude (emotions) or behavior (purchase). Against this theoretical background the purpose of the research is to close this gap in creating an experimental triangulative research design that enables different kind of implicit methods of emotion measurement that prevents bias which usually arises (only) from questioned surveys. The combination of the implicit research methods EEG (electroencephalographic), eye tracking, biometric measurement (heart rate, galvanic skin response etc.) (see Gröppel-Klein/Braun 2001) and facial emotion measurement equipment is applied to enhance the measurement quality with regard to understand consumers' deep subconscious responses to sexual-oriented stimuli. We assume that the triangulative approach (combining different instruments to measure the same research object [emotions]) presents the best way to achieve reliable results. Whilst the EEG measure the means of brainwave activity (see Rothschild et al. 1988), the eye tracking enables the measurement of the actually perceived information that comes from the human visual process (see Kroeber-Riel/Weinberg, 2008, p.264). To get valuable information the eye tracker is used to verify via infrared, where and how long test persons are looking at different parts of the displayed frames (see Duchowski, 2007, p.263). Biometric and the facial emotion measurement provide further data to accompany brainwave monitoring and eye tracking. The last stage in the research project contains an explicit standardised recall survey among the participants. The test persons are asked to give an evaluation of every advertisement. The combination of different implicit methods provides a picture of how persons respond to stimuli material and which emotional engagement they have.

References

Duchowski, Andrew (2005). Eye Tracking Methodology – Theory and Practice. 2th ed. London: Springer.

Gröppel-Klein, Andrea / Braun, Dorothea (2001); The Role of Customers' Arousal for Retail Stores - Results from An Experimental Pilot Study Using Electrodermal Activity as Indicator; Advances in Consumer Research; Jan; 2001(1).

Kroeber-Riel, Werner / Weinberg, Peter (2008). Kosumentenverhalten. 8th ed. München: Vahlen.

Michael L. Rothschild / Yong J. Hyun / Byron Reeves /Esther Thorson/ Robert Goldstein; (1988); Hemispherically Lateralized EEG as a Response to Television Commercials; Journal of Consumer Research; Sep88, Vol. 15