A Conversational Agent that Reacts to Vocal Signals

Daniel Formolo(✉) and Tibor Bosse

Department of Computer Science, VU University Amsterdam,
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{d.formolo,t.bosse}@vu.nl

Abstract. Conversational agents are increasingly being used for training of social skills. One of their most important benefits is their ability to provide natural interaction with humans. This work proposes to extend conversational agents' benefits for social skills training by analysing the emotion conveyed by the user's speech. For that, we developed a new system that captures emotions from human voice and, combined with the context of a particular situation, uses this to influence the internal state of the agent and change its behaviour. An example of the system's use is shown and its limitations and advantages are discussed, together with the internal workflow of the system.

Keywords: Virtual agents · Social skills training · Speech analysis · Vocal signals · Emotions

1 Introduction

Embodied Conversational Agents (ECAs) can be defined as computer-generated characters 'that demonstrate many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication' [1]. As research into ECAs is becoming more mature, conversations with ECAs are increasingly being perceived as natural, or at least 'believable'. As a result, there is a growing interest in the use of ECAs for training of communicative skills, such as negotiation, conflict management or leadership skills (e.g., [2–7]). The main motivation is that a training system based on conversational agents provides a cost-effective method to replace (or at least complement) human actors, as it can be used anytime, anywhere.

Despite this promising prospect, developing effective conversational agents for communication training is far from easy. An important requirement for effective ECAs is their ability to react to behaviour of the trainee in a similar manner as a human interlocutor would do. Otherwise, there is a risk that the system reinforces the wrong behaviour. For instance, a virtual agent that only listens to you if you address it with a submissive attitude is probably not very useful for leadership training. Hence, making an ECA show the appropriate response to the appropriate behaviour of the trainee is crucial. However, this introduces another challenge, namely to define what is 'appropriate behaviour' of the trainee. Obviously, one relevant aspect of behaviour involves the *content* of what the trainee says. And indeed, most ECA-based training systems

R. Poppe et al. (Eds.): INTETAIN 2016, LNICST 178, pp. 285–291, 2017.
DOI: 10.1007/978-3-319-49616-0_29

have been designed in such a way that the ECA's responses depend on what the user says (e.g., by analysing the user's speech, or by generating appropriate responses based on selected options within a multiple choice menu).

However, although most ECAs respond to *what* the user says, they often do not respond to *how* the user says it. This is a serious limitation, as the style of a person's speech is very important during social interactions: as discussed in [8], humans heavily rely on vocal cues (such as volume, or speed of talking) to infer other people's emotions. For example, the phrase 'sorry sir, we cannot accept 100 Euro bills' can be perceived as very friendly when it is uttered calmly and gently, but it can be perceived as offensive when it is uttered with a quick and monotone voice. Especially for communication training it is important to take such differences into account, as it allows professionals to learn not only what to say during their job, but also how to say it. Hence, this paper proposes the use of ECAs for social skills training that adjust their behaviour based on vocal signals that are extracted from the user's speech[1]. The paper first presents global architecture to develop such feature in the agents, followed by a discussion on how the system can be used for specific types of communication training.

2 Emotions in Vocal Signals

Many factors influence the generation of emotion in humans. Emotions can remain stable for a long time or may come and go fast, and sometimes various emotions are mixed at same moment. In the literature, roughly three theoretical perspectives may be distinguished. First, *categorical theories* are based on the assumption that there is a limited set of basic emotions categories such as joy, sadness, fear, anger, and disgust [9]. Second, *dimensional theories* view emotions as states that can be represented as points within a continuous space defined by two (or three) dimensions, namely valence and arousal (and dominance) [10, 11]. Valence refers to the level of pleasure, while arousal refers to a general degree of intensity. Third, *componential theories* highlight the role of different components that play a role in the emotion generation process, such as the desirability and likelihood of the events that trigger the emotion, cf. appraisal theory [12]. In the current paper, we will mainly make use of the dimensional approach, using the dimensions of arousal and valence. Both valence and arousal are expressions of brain circuits involving amygdala, orbitofrontal cortex, the insula and various brain areas [13], and the emotions that arise from those areas have a direct reflection in the human voice.

To recognise such affective features in human speech, the presented approach builds upon a vast body of previous work. For instance, in [14] an approach was put forward to detect emotions in speech in terms of arousal and valence. Similarly, [15] has shown that more specific emotions (e.g., aggression) can be identified as well. Moreover, Rodriguez et al. analyse changes of vocal patterns in humans when they interact with ECAs [16].

[1] Obviously, vocal signals are not the only aspect of behaviour that is relevant for communication training. Other aspects include facial expression, gestures, and posture, among others. However, these aspects are beyond the scope of this paper.

Inspired by these developments, a number of recent systems use vocal cues to trigger the behaviour of virtual agents. For example, in [17] vocal cues are used to generate *backchannels* (i.e., non-intrusive signals provided during the speaker's turn). Acosta and Ward proposed a system that uses speech and prosody variation to build rapport between human and agent [18], and Cavazza et al. used vocal signals for character-based interactive storytelling [19]. Furthermore, the virtual human SimSensei Kiosk uses voice, speech and other features to analyse user emotions in the context of healthcare decision support [20]. As can be observed, these works are closely related to the proposed system, although they focus on different applications than social skills training. In contrast, one recent system that does focus on communication training (in the context of job interviews) is put forward in [21]. This paper presents an ECA that adapts its behaviour to vocal cues according to social constructs such as attitude and relationship. One way in which the current system extends this work is by considering the context of the conversation more explicitly.

3 The System

The proposed system is expected to be easily integrated within different serious games or other specialised systems. The final module is a library that is available in the Windows platform as a DLL and that may be extended to Linux-like operating systems. Figure 1 shows an overview of the system (i.e., the ECA). It contains various modules, including an interface to capture the user's speech, the off-the-shelf openEar tool to process this speech [22], and a module to generate a response to the user.

The openEar tool performs the task of identifying which emotion is currently experienced by user; however, it can also be replaced by any other tool, because the sub-components are completely independent. One only needs to adjust the connection between them. Some voice features used by openEar and consequently by the system to analyse emotions are Pitch, Formants and Bandwidth, and Temporal characteristics. The output of openEar is a set of emotions and their values. That information is processed by the Context Awareness Module, which deals with ambiguous outputs received by the previous module through a decision tree algorithm combined with

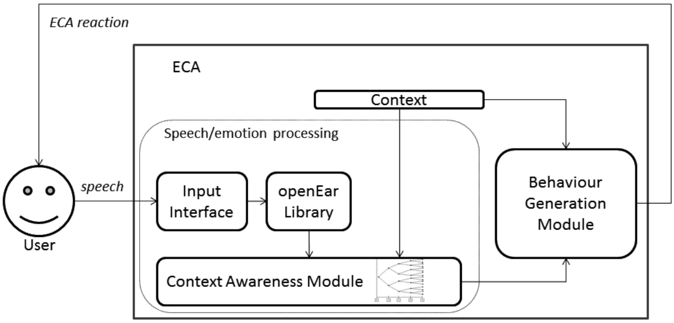

Fig. 1. Flow diagram of the proposed system.

context information provided by the ECA's beliefs. As mentioned before, it is difficult to distinguish between emotions with similar arousal, even if their valence levels are completely opposite (like anger and joy), [13, 23]. However, by using the context information, this module can minimise problems like that. The output is a set of emotions and their levels, varying from 0 to 100.

Next, this output is provided to the Behaviour Generation Module, which generates an appropriate response to the user. Obviously, this module can be very complex in itself (e.g., including modules for dialogue management and speech generation), but this is outside the scope of this paper. As a simple proof of concept, the Behaviour Generation Module currently just makes the ECA show a facial expression that is similar to the human emotion that is perceived. However, in other situations it might be more effective to respond in a different way to the perceived emotion (see the more extensive discussion below).

Currently, the system is still in development, and need improvements mainly in the Context Awareness and Behaviour Generation modules.

To illustrate its working, one prototype application composed of an ECA that responds to the inferred emotion captured from voice was developed. Figure 2 shows 4 different emotions expressed by the ECA, which reflect the voice of the user. However, the system could also be applied in many different situations in which the ECA not only mirrors the emotions of the user but also shows variations of those, like in negotiations, where a happy emotion from the user could produce an angry reaction from the ECA.

Another application that we are currently focusing on is aggression de-escalation training. In this domain, there is an interesting difference between so-called *emotional aggression* and *instrumental aggression*. The main difference is that emotional (hot-blooded) aggression is caused by an agent's goals being frustrated, whereas instrumental (cold-blooded) aggression is caused by an agent using intimidation as a means to achieve its goals [2]. This distinction is interesting for our system because an *emotionally aggressive agent* will calm down if the user approaches it empathically. Concretely, this means that the ECA first identifies the emotion conveyed in the user's

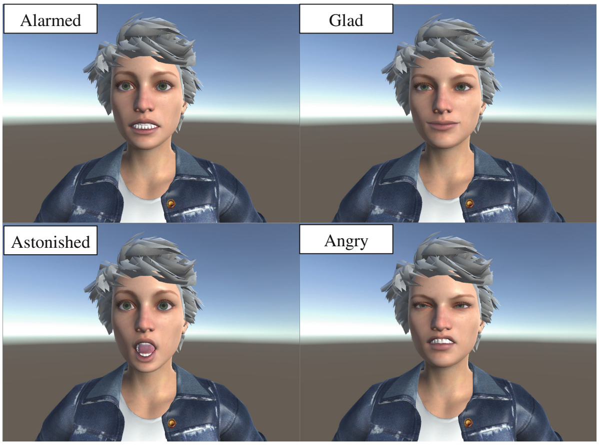

Fig. 2. Example of recognised emotions transferred to avatar.

voice, and if it recognises this as an empathic reaction it will become less aggressive. Similarly, if it interprets the user's utterance as non-cooperative, it will become more aggressive. Instead, for *instrumentally aggressive agents* this will be the other way around: if such an agent identifies the user's behaviour as empathic, it will become more aggressive, and if it interprets it as non-cooperative, it will calm down. Based on such an application, users could train to take the more suitable conversation style in the appropriate situation. This is very relevant, e.g., for employees in domains such as law enforcement and public transport [2].

Note that we presented here two possible applications of the system, where an ECA generates mirroring and opposite emotions, respectively. However, the application is not limited to these two situations. Generally, the system provides ECA developers the possibility to develop variance in emotional representation for a variety of aspects, for example, developing a sentiment of trust during a conversation. As another example, during an interaction, an supportive ECA could perceive irony in a user's voice and as a consequence become less empathic for the rest of the conversation. In principle, the possibilities cover the entire spectrum of human interactions; nevertheless, its success depends on the capacity of capturing the real emotion that the user is transmitting.

4 Discussion

This paper proposes the use of vocal signals that are extracted from the user's speech as one additional component to adjust ECAs' behaviour. To achieve this goal, we developed an adaptable system that processes human voice and returns a set of emotions and their intensity levels. The system can be easily plugged in into ECAs or other specialised systems that can enrich user experience. Especially for ECAs, the emotional information of a person's voice provides a new element to model their internal behaviour, which may make the interaction between ECAs and humans more natural and effective for training applications. A second innovation is the use of context information to extract emotions from human speech more accurately. Often, context conveys crucial information that is neglected by systems and serious games.

Nevertheless, there are circumstances that might limit the use of the proposed system; for example, when the user's environment is noisy or has more than one person speaking at the same time, the system cannot provide precise information. In other cases, the user might not interact much with the system, which could also limit the emotional information extracted by the system. Besides this, it is important to combine the emotional information provided by the user's voice with other sources like facial expressions, gestures and text. Despite these limitations, the system is an important addition to the state-of-the-art of the development of ECAs. For future work, it is necessary to refine the system, analyse its accuracy in different contexts, and test it in real world applications.

Acknowledgments. This research was supported by the Brazilian scholarship program Science without Borders - CNPq {scholarship reference: 233883/2014-2}.

References

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.: Embodied Conversational Agents. MIT Press, Cambridge (2000)
2. Bosse, T., Provoost, S.: Towards aggression de-escalation training with virtual agents: a computational model. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2014. LNCS, vol. 8524, pp. 375–387. Springer, Heidelberg (2014). doi:10.1007/978-3-319-07485-6_37
3. Bruijnes, M., Linssen, J.M., op den Akker, H.J.A., Theune, M., Wapperom, S., Broekema, C., Heylen, D.K.J.: Social behaviour in police interviews: relating data to theories. In: D'Errico, F., Poggi, I., Vinciarelli, A., Vincze, L. (eds.) Conflict and Multimodal Communication. Computational Social Sciences, pp. 317–347. Springer, Heidelberg (2015)
4. Hays, M., Campbell, J., Trimmer, M., Poore, J., Webb, A., Stark, C., King, T.: Can role-play with virtual humans teach interpersonal skills? In: Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) (2012)
5. Jeuring, J., Grosfeld, F., Heeren, B., Hulsbergen, M., IJntema, R., Jonker, V., Mastenbroek, N., Smagt, M., Wijmans, F., Wolters, M., Zeijts, H.: Communicate! — a serious game for communication skills —. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 513–517. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24258-3_49
6. Kim, J., Hill, R.W., Durlach, P., Lane, H.C., Forbell, E., Core, C., Marsella, S., Pynadath, D., Hart, J.: BiLAT: a game-based environment for practicing negotiation in a cultural context. Int. J. Artif. Intell. Educ. **19**(3), 289–308 (2009)
7. Vaassen, F., Wauters, J.: deLearyous: Training interpersonal communication skills using unconstrained text input. In: Proceedings of ECGBL, pp. 505–513 (2012)
8. Juslin, P.N., Scherer, K.R.: Vocal expression of affect. In: Harrigan, J.A., et al. (eds.) The New Handbook of Methods in Nonverbal Behavior Research. Oxford Press, Oxford (2005)
9. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
10. Russel, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
11. Yik, M., Russel, J., Steiger, J.: A 12-point circumplex structure of core affect. Emotion **11** (4), 705–731 (2011)
12. Scherer, K.R., Shorr, A., Johnstone, T.: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Canary (2001)
13. ElAyadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**, 572–587 (2011)
14. Truong, K.P., van Leeuwen, D.A., de Jong, F.M.G.: Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Commun. **54**, 1049–1063 (2012)
15. Lefter, I., Rothkrantz, L.J.M., Burghouts, G.: Aggression detection in speech using sensor and semantic information. Text, Speech and Dialogue **7499**, 665–672 (2012)
16. Rodriguez, H., Beck, D., Lind, D., Lok, B.: Audio analysis of human/virtual-human interaction. In: Prendinger, H., Lester, J., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 154–161. Springer, Heidelberg (2008). doi:10.1007/978-3-540-85483-8_16
17. Bevacqua, E., Pammi, S., Hyniewska, S.J., Schröder, M., Pelachaud, C.: Multimodal backchannels for embodied conversational agents. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS (LNAI), vol. 6356, pp. 194–200. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15892-6_21
18. Acosta, J.C., Ward, N.G.: Achieving rapport with turn-by-turn, user-responsive emotional coloring. Speech Commun. **53**, 1137–1148 (2011)

19. Cavazza, M., Pizzi, D., Charles, F., Vogt, T., Andre, E.: Emotional input for character-based interactive storytelling. In: Proceedings of the 8th International Conference on Autonomous Agents and multi-agent systems, AAMAS 2009, pp. 313–320 (2009)
20. DeVault, D., et al.: SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 13th International Conference on Autonomous Agents and Multi-agent Systems, AAMAS 2014, pp. 1061–1068 (2014)
21. Ben Youssef, A., Chollet, M., Jones, H., Sabouret, N., Pelachaud, C., Ochs, M.: Towards a socially adaptive virtual agent. In: Brinkman, W.-P., Broekens, J., Heylen, D. (eds.) IVA 2015. LNCS (LNAI), vol. 9238, pp. 3–16. Springer, Heidelberg (2015). doi:10.1007/978-3-319-21996-7_1
22. Eyben, F., Wöllmer, M., Schuller, B.: openEAR - introducing the munich open-source emotion and affect recognition toolkit. In Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009). IEEE, Amsterdam (2009)
23. Nwe, T., Foo, S., De Silva, L.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**, 603–623 (2003)