The Role of the Vocal Stream in Telepresence Communication

Banan S. Bamoallem
University of Strathclyde
75 Montrose Street
Glasgow G1 1XJ, UK
+44 7853914684
banan.bamoallem@strathclyde.ac.uk

Andrew J. Wodehouse
University of Strathclyde
75 Montrose Street
Glasgow G1 1XJ, UK
+44 1415482628
andrew.wodehouse@strath.ac.uk

ABSTRACT

Most of the work in affective computing within telepresence robot platforms adds to current research and knowledge generation as opposed to application. The main reason behind this lack of benefit is that most research does not represent reality, and the actual capabilities we have in the real world do not match the capabilities that are used in research. Taking this into consideration, this paper helps in establishing a new method to display naturalistic behaviour that can be feasibly implemented for telepresence (TP) interaction. Based on an understanding of different aspects of human-human interaction (HHI) a three phases rhythm were proven to exist between nonverbal and certain verbal behaviours of speakers and listeners. We chose the gestures related to our research, and tried to match them with the proposed TP phases by identifying which best matched the phase descriptions. Thus, this study provided step by step guidelines to govern the creation of practical user interfaces that will capture the vocal stream, and allow users to relate to natural nonverbal behaviours that spontaneously arise during speech.

CCS concepts

• Human-centered computing → Human computer interaction (HCI) • Interaction techniques → Human-centered computing → Gestural input.

Keywords

Nonverbal behaviours; head movements; face-to-face interaction; telepresence robot, body movement.

1. INTRODUCTION

Telepresence (TP) is commonly defined as a set of technologies that allow people to feel as if they are present in a location other than their true location. Various studies have investigated diverse disciplines such as psychology, communication, computer science and philosophy in relation to telepresence. These studies discuss guidelines for increasing social acceptability and also raise concerns about social issues such as the sense of social presence. However, the exploration was not based on detailed observations of human-human interaction [1]; [2]; [3].

A review of the literature on human-human interaction and communication behaviour indicates that nonverbal behaviours play an important role in message production and in involvement in a variety of different situations. This finding inspired our first study [4] to investigate the possibility of achieving similar results within the field of human-robot interaction using a telepresence robot (TP).

In the first experiment we mainly explored the head area as it supports interaction involvement and feedback functions between people. The potential for replicating the head movements which could be useful in the development of a model of human gestures to implement in a telepresence platform was investigated.

The results did not show any evidence of the possibility of identifying any improvement by incorporating head movements. However, our findings at least highlighted important implications for future research. It suggested that face-to-face interaction is complex in its own right, as it includes various behaviours that help in maintaining the connection between two people. Thus it would be difficult to find any significant result if we only focused on one of these behaviours. In general, it should be noted that real-time communication requires more than verbal communication, facial expressions and head nodding. It is important to complement them with other types of nonverbal behaviour such as posture. We needed to make robots capable of generating meaningful and recognizable gestural cues.

At present, the most accurate method of achieving a life-like experience is to use multi-modal interactions, i.e. making use of multiple signals from the different types of human interaction. Using traditional HCI multimodal methods for true whole body markerless motion interaction is not feasible; we need points of reference on the body, whether from sensors or cameras. We also require that the processing and feedback of signals is achieved in real time. Thus there is a need for further advances in machine learning techniques, or development of a new technique to deal with real time markerless interaction for a usable experience.

The challenges for Intelligent Technology and the Human Robot Interaction researchers are numerous. In our case, the challenge will be mainly related to developing methodologies for eliciting user requirements in real contexts and be acceptable in the long-term. Defining a new method to implement human behavioural traits was one of the ideas that sparked off our platform concept and design. Thus, this study will provided step by step guidelines to govern the creation of practical natural user interfaces which allow users to relate to natural behaviours that spontaneously arise during speech.

The process will be broken down into key steps:

1. Firstly, a more in-depth understanding of what the face to face interaction mechanism facilitates, in order to define the nonverbal behaviour that is of significance to our system.
2. Secondly, find a way to replicate this behaviour without losing its essence.

Finally, redefine these behaviours so they can fit into the human–robot interaction (HRI) that we are working on.

2. NONVERBAL BEHAVIOUR

As individuals we subconsciously seek information whenever we enter the presence of others, as regardless of whether or not we are talking, we still communicate through embodied expression. We seek or process the information we already have to define a situation, or signal in advance to others what an individual may expect from them, and vice-versa. Individual embodied expressions have two radically different kinds of sign activity; verbal signals or substitute nonverbal signals that aid the transfer of verbal information [5].

Nonverbal behaviours refer to actions as distinct from speech, and they include facial expressions, hand and arm gestures, posture, positions and various movements of the body or the legs and feet [6]. These behaviours have a direct link to the verbal part, as they can function to qualify whatever an individual means by a statement. Furthermore, such behaviours may convey information about the speaker's social attributes, about their own conceptions, about others present, and about the setting.

In respect to our research both types are of interest, because as an individual we expect to see confirmation and consistency between both types in order to emulate natural human interaction as we understand it, which is not the case with human computer interaction. Different attempts have been made to create a system that can develop a sense of presence for the user. However, they fail to produce an accurate personal presence as they mainly focus on one type and ignore the other. The consistency between verbal signals and its substitute signals is missing from most of the current research, as it is only focusing on maximizing the ease and speed of recognition instead of the naturalness of the interactions.

In the next section, both types of interactions will be covered in conjunction with a description of implementation with respect to current technological limitations.

2.1 The coordination of verbal and nonverbal

This link goes beyond not just conveying information to covering similarities between nonverbal and certain verbal behaviours of individuals, as supported by different studies. An early microscopic analysis of the coordination between movement and speech was carried out by Kendon in the early 1960s. He was able to match body movement with a speech transcript in a close study of sound and film recordings of interactions. A rhythm was proven to exist at even at the most microscopic levels (e.g., spoken syllables) where the points of change in the flow of sound are coincident with the points of change in body movement. Nevertheless, Argyle [7] highlighted that body movement has a hierarchy which corresponds to different verbal unit sizes, e.g. emphasising through change in loudness or pitch for speech can be emulated through hand or head movements. This also can be found between body movements, vocal hesitations and pauses in speech [8], [9].

Condon and Ogston [10], [11] suggested that listener's actions were modelled on a speaker's speech stream and vice-versa, e.g. a phoneme change can be seen in a speaker's talk which resulted from the small movements produced by the listener's head, eyes, wrist, mouth and fingers. They further explain that as a rule, speakers and listeners are in synchrony up to the word level, as any variation in the configuration of movement of the listener will match the variation in the speaker's configuration at word, syllable and phonic levels…

"if speaker and listener are in synchrony and the listener lifts a cigarette to his lips, draws on it, and lowers the cigarette again, the boundaries of the major components of this action will coincide with boundaries in the behaviour flow of the speaker, but these boundaries will not necessarily also be boundaries of the larger waves of behaviour in the speaker, for instance the boundaries of his phrases"

This emphasises the importance of synchronisation between vocal stream and nonverbal behaviour in regard to regulating and organising dialogue itself in group interaction situations, through sharing attentions and expectancies. In other words, this coordination provides one of the ways in which two people signal that they are open to one another, and not to others

2.2 Human interaction phases

As previously highlighted, one of the early works providing detailed analysis of human interaction is Kendon works [12]. In specific, we looked at the detailed analysis to use it as a source for examples of different behaviours between listeners and speakers. This analysis is of particular interest as it provides movement phases during interaction, with a full description for each phase which helps to frame the outline of our system.

Three phases of movement and speech rhythm between speakers and listeners were generated as resulted of this analysis which are

1. *First Phase (opening position)*

At the beginning of the interaction there is an associated movement called the opening position. This phase serves to visually validate that the speaker is speaking to the right person, and for onlookers it clarifies to whom the speech is being directed. Shared-movement rhythmicity can be seen here; a mirrored movement which only happens between the speaker and the person he directly addresses to grab their attention.

2. *Second Phase*

As the speakers become more confident that they have the attention of the listeners, the movement more or less ceases, apart from mouth movements, eye shifts and blinking of the eyes.

3. *Third Phase*

Finally, as a result of the familiarisation between both sides, in this last phase the listener's behaviour is followed by the speaker's, and related to the variation of the pitch level of speaker's voice.

Apart from these phases, it has been found that some facial expressions or head movements appear at specific junctures in the speech of our partners; for example, head nods and movements of hands and feet tend to occur at the end of rhythmical units of the talker's speech i.e. at pauses within phonemic clauses but mainly at junctures between these clauses. Vocally stressed words also tend to be accompanied by movements

3. PROPOSED SYSTEM

Since we aim to improve the interaction between both sites in mediated interaction, we focused on identifying the cluster of signals that distinguished a positive evaluation of an interaction partner from a negative one as identified by [13]. We used the signals related to our research and tried to match them with the phases that best matched the phase description. Three gestures were chosen from Mehrabian's [13] list based on our design specification, these were

- More forward lean
- Closer proximity
- More direct body orientation

In the next section we will give an outline of movement specification in respect to Kendon [14] interaction phases

3.1 First phase (opening position)

As it is called an opening position, we thought a forward movement would help in maintaining the exchange of talk between two people, and of course ensure there were no physical obstructions to block them from addressing each other in an encounter. This movement will be a translation of the starting conversation from the person in front of the device.

While investigating the optimal range of viewing distance for desktop monitors, a study by [15] recommended a minimum distance of 635 mm. However, this was not the case in every situation; another study relating to viewing distances for LCD monitors found that screen reflections affected the viewing distance which resulted in a shorter distance compared to a normal desktop monitor [16]. Although this was also supported [17], they argue that this effect is minimal. Therefore, the system will try to keep approximately 400mm and 600mm between one side and other when moving forward. This forward movement will range between 500mm/s and 700mm/s.

3.2 Second phase

The second phase, described as quiescent by [12] - the analysis did not show big movements between both sides, only slight movements. Therefore, we decided to add visibly small movements into this phase. A Slight backward lean movement will accommodate the initiation of the talk (from the person in front of the device) and a slight forward movement when the person on the device is talking, as suggested by [18]. These slight movements will be used as a way to enhance communicative attention between both sides. This slight lean will adjust the viewing angle to the optimum viewing angle which is between 15° and 20° beneath the horizontal sight line

3.3 Third phase

As we cannot indicate when the conversation will reach an end, we decided to add some movement to fit the description of the phase. This phase is described as interchange phase, where speaker and listener mirror one another's posture, and such posture shifts often occur synchronously. This synchrony of movements can be seen as a signal of understanding, agreement or support from listener. In our case we will replicate this with a slight upward shrug movement in relation the rise of the pitch of voice, when the primary stress points in the speech are accrued as reported by [17]. In addition, forward movement of the head, will happen during silent parts of the conversation where movements seem to peak. This will be within range of 5°-20°.

If we examine these three phases, we find that there are two scenarios in respect to the vocal stream translation. The first two phases aim to translate the vocal stream into movement with respect to both sides (both the person in the device and in the front of the device). Whereas the last phase, concerns the volume only, without any differentiation between each side (see Figure 1). However, these movements were generated to cover all the possibilities of the technological capabilities that can feasibly be implemented.

DOI: http://dx.doi.org/10.1145/12345.67890

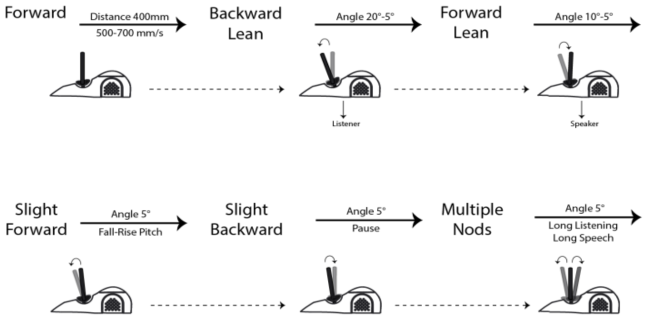

Figure 1. Proposed Movements Specification

4. CONCLUSION

In order to design telepresence robots that are able to emulate, and therefore enhance, natural interaction it is vital to make robots capable of generating meaningful and recognisable gestural cues. Currently the most accurate method of achieving this it is the use of multi-modal interactions, i.e. utilising multiple signals from different types of human interactions; for example, physical motion capture, physiological inputs, the normal five etc. However, even given such an approach, generating a useable experience is still problematic as it requires the processing and feedback of signals to be in real time, and feedback alongside the gesture recognition thresholds is still not high enough in all domains.

Most of the research conducted on improving interactions within telepresence robot platforms adds to current research and knowledge generation as opposed to application. Consequently, users do not get clear benefits from them in the real world. The main reason behind this lack of benefit is that most research does not represent reality, and the actual capabilities we have in the real world do not match the capabilities that are used in research. Therefore, contemporary telepresence robot design should be based on an understanding of different aspects of human-computer interaction (HCI) in regard to that which can be feasibly implemented. Thus, this study provided step by step guidelines to govern the creation of practical natural user interfaces which allow users to relate to nonverbal behaviours that spontaneously arise during speech.

Based on this we have changed the concept of the current research by developing a platform which is able to deliver nonverbal signals for video-mediated conversations as representations of real world gestures, with respect to the actual capabilities we have in the real world. We provided a way to use a single model approach by relying on the detailed analysis of Kendon's work of the coordination between movement and speech within HHI as a framework to govern the production of non-verbal signals within HRI. Our initial plan is to implement these human behavioural traits using the vocal stream via audio sensors.

To our knowledge, this approach is unique as we have looked at the HHI from different angles which revealed a new, valuable way to implement the nonverbal in respect to the technology capabilities we have nowadays. Our future work will involve demonstrating this approach to enhance the users' presence.

5. REFERENCES

1. C. Liu, et al., "Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction," Proc. Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on, IEEE, 2012, pp. 285-292.

2. D. Heylen, “Head gestures, gaze and the principles of conversational structure,” International Journal of Humanoid Robotics, vol. 3, no. 03, 2006, pp. 241-267.

3. A. Duquette, et al., “Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism,” Autonomous Robots, vol. 24, no. 2, 2008, pp. 147-157.

4. B.S. Bamoallem, et al., “The impact of head movements on user involvement in mediated interaction,” Computers in Human Behavior, vol. 55, 2016, pp. 424-431.

5. E. Goffman, The Presentation of Self in Everyday Life, Peter Smith Publisher, Incorporated, 1999.

6. A. Mehrabian, Nonverbal communication, Transaction Publishers, 1977.

7. M. Argyle, Bodily communication, Routledge, 2013.

8. D.S. Boomer, “Hesitation and grammatical encoding,” Language and speech, vol. 8, no. 3, 1965, pp. 148-158.

9. G.L. Trager and H.L. Smith, An outline of English structure, Рипол Классик, 2009.

10. W.S. Condon and W.D. Ogston, “Sound film analysis of normal and pathological behavior patterns,” The Journal of Nervous and Mental Disease, vol. 143, no. 4, 1966, pp. 338-347.

11. W.S. Condon and W.D. Ogston, “A segmentation of behavior,” Journal of psychiatric research, vol. 5, no. 3, 1967, pp. 221-235.

12. A. Kendon, “Some relationships between body motion and speech,” Studies in dyadic communication, vol. 7, 1972, pp. 177.

13. A. Mehrabian, “Nonverbal communication,” 1972.

14. A. Kendon, “Movement coordination in social interaction: Some examples described,” Acta psychologica, vol. 32, 1970, pp. 101-125.

15. D.R. Ankrum, “Viewing distance at computer workstations,” Workplace Ergonomics, vol. 2, no. 5, 1996, pp. 10-13.

16. K.-K. Shieh, “Effects of reflection and polarity on LCD viewing distance,” International Journal of Industrial Ergonomics, vol. 25, no. 3, 2000, pp. 275-282.

17. K.-K. Shieh and D.-S. Lee, “Preferred viewing distance and screen angle of electronic paper displays,” Applied Ergonomics, vol. 38, no. 5, 2007, pp. 601-608.

18. M. Knapp, et al., Nonverbal communication in human interaction, Cengage Learning, 2013.

6. ACKNOWLEDGMENTS

The first author would like to thank Umm Al-Qura University, Makkah, Saudi Arabia for funding this work as part of her PhD scholarship program.