Prosodic Analysis in Human-Machine Interaction

Maria Di Maro
Università degli Studi di Napoli 'Federico II'

Sara Falcone
Università degli Studi di Trento

Francesco Cutugno
Università degli Studi di Napoli 'Federico II'

In this paper we are going to present some experiments concerning the analysis of prosodic features in the spoken production of requests by human users in human-machine interactions. The main aim of this analysis is to understand if and how much a speaker adapts the spoken production to his/her virtual interlocutor, to therefore be able to choose the right register as model to improve speech recognition in spoken dialogue systems.

The conventional assumption towards human speech in conversing with computer systems is that speakers usually tend to simplify their language to avoid not being understood. The language simplification is expected to be encountered at different linguistic levels: the uttered sentences tend to be syntactically simpler, the pronunciation is more articulated, a higher intensity of voice is used, pauses and interruptions are avoided or they are posed in specific points of the utterance, the speaking rate is reduced and a reduced pitch dynamics can be observed. Even though many computer systems try, by means of natural interfaces (Valentino 2017), to reproduce a natural conversation with its basic requirements, i.e. shared knowledge and context (Bettoni 2006), the reason behind the use of a simplified register takes origin from the perception of the non-expertise in conversing naturally. This leads to the use of *Machine Talk*, a register similar to others like *Foreigner Talk* (Freed 2009), used with the aim of being understood.

On the other hand, other empirical observations show how the use of a virtual assistant on a personal smartphone can lead to a more spontaneous language production, as if the user was speaking with another human interlocutor. This difference lies in the representational perception of the other which explains how the language production can be modified according to the users' preconceptions towards a specific situation (Fischer 2011a). This means that it is not straightforward to encounter a simplified register in dialogues with machines, since this can be caused by the way the speaker understands the functionality of the system itself (Fischer 2011b) - something which can also depend on the quality of the synthesizer. Such observations prove how the importance of the Theory of Mind (Goldman 2012), in constructing mental states of the interlocutors to better produce and understand the conveyed messages, can also be applied when the interlocutor is not a human being.

In this work, we mainly focus on the analysis of the prosodic features of questions and commands posed to a domain-dependent Spoken Dialogue System and to a commercial virtual assistant implemented in a smartphone, in order to understand if the prosodic differences depend on the tool used, which causes a different representational preconception,

or if these perceptions depend on the speaker itself, with his/her cultural and sociological characteristics. To better understand the correlation between the differences and the representational constructions of the speakers, we compare the collected results with a usability test taken by the users. Moreover, the linguistic production of our users is also compared with other two different spoken uses, reading and spontaneous narration. The features we consider are the speaking rate, the pitch dynamics, pauses, and the sounds articulation.

The experiments we carried out are designed as follows: in the first session of tests, 5 users were asked to pose questions to a spoken dialogue system and were given conceptual categories names as an inquiring guide; afterwards, they had to read a text and were asked to talk about the last movie they saw; in the second session of tests, other 10 users had to play a game of task completion with a smartphone, and had to read an article, and describe a movie as well. At the end of the second test, they were asked to give a measurable opinion concerning the ease of use and satisfaction for the system. Our preliminary results show a direct correlation between class of users and speaking styles, a strategic attitude of speakers to posing pauses at specific and syntactically meaningful points of the utterances, and they confirm a general reduction of pitch dynamics. These results are going to be considered as an important means for developing spoken dialogue systems whose speech recognition module skills are better suited to the type of user.

References

Valentino, M., Di Maro M., Cutugno F. (2017). *Towards a Natural User Interface for Small Groups in Real Museum Environments*. Proceedings of SEMDIAL 2017 (SaarDial): Workshop on the Semantics and Pragmatics of Dialogue, Saarbrücken, Germany.

Bettoni, C. (2006). *Usare un'altra lingua. Guida alla pragmatica interculturale*. Roma-Bari: Laterza.

Fischer, K. (2011a). Interpersonal variation in understanding robots as social actors. In Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 53–60). ACM, New York, NY, USA.

Fischer, K. (2011b). Recipient design, alignment, interaction: The role of the addressee in so-called "simplified registers." Habilitation thesis, University of Bremen.

Freed, B. (2009). «Foreigner Talk, Baby Talk, Native Talk». *International Journal of the Sociology of Language*, 1981(28), pp. 19-40.

Goldman A.I. (2012). «Theory of Mind». *Oxford Handbook of Philosophy and Cognitive Science*, pp. 402-424.