DOI: 10.17469/O2104AISV000013

MARIA DI MARO, SARA FALCONE, FRANCESCO CUTUGNO

Prosodic analysis in human-machine interaction

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, and to which extent this could be caused by the type of user and his/her representational preconception towards the specific interaction. The collected results are, therefore, considered as an important means for developing spoken dialogue systems whose speech recognition module skills are better suited to the characteristic of the human interlocutor.

Keywords: human-machine interaction, spoken dialogue systems, pitch, speaking rate, vowel space, hyperarticulation.

1. *Introduction*

Human-machine interaction is a field of research that covers several aspects, from text to speech to pragmatic aspects of the verbal and nonverbal interaction. In this paper we focus on speech and, particularly, on the prosodic anomalies analysed in several human-machine dialogues. In this work, the analysis of the prosodic features of questions and commands posed to a domain-dependent spoken dialogue system and to Google Assistant is carried out. The collected results were compared with speech materials produced at a different point on the diaphasic continuum. In particular, spontaneous narrations were recorded, as users were asked to tell the plot of a movie they saw. The main aim of this analysis was to understand if and how much a speaker adapted the spoken production to the virtual interlocutor and if a previous acquaintance with the system had an impact on the interaction.

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 reason behind the use of a simplified register takes origin from the perception of the non-expertise of computers in conversing naturally. 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 channel of communication or context of interaction. With this analysis we would like to start a deepened pragmatic analysis of human-computer interaction, taking prosodic features as a starting point. A theory of mind of user behaviour is

the desired future goal, useful to shape a better model (Soltau, Waibel, 2000), which would improve speech recognition for spoken dialogue systems.

In this paper, we sketched out the steps of a prosodic analysis concerned with the alleged hyperarticulation of sounds in user's utterances addressed to machines. We started from what it is described in other state-of-the-art studies on hyperarticulation and simplified registers. Hyperarticulate speech is intended as a strategy to ease the communicative exchange in situations where misunderstanding or non-understanding of the speaker intention occur. Other than using simpler syntactic structures, common words, pointing gestures and other visual paralinguistic means of communication, the use of a greater articulatory effort in producing sounds, is seen as helping in increasing the success of the informative exchange. This strategy is employed in different compromising situation, where the listener could have problems to catch the message because of a noisy environment, or because he/she is a foreigner, or has hearing impairment or is a child, who is taking his/her first steps in language learning. In fact, researches on hyperarticulation are mainly concerned with infant-directed (McMurray et al., 2013; Hartman et al., 2017; Kalashnikova et al., 2018), foreigner-directed (Scarborough et al., 2007) and hearing-impaired-persons directed speech (Picheny et al., 1986). Computer-directed speech is another field of interest (Oviatt et al., 1998; Stent et al., 2008; Akira et al., 2017) where the hyperarticulation hypothesis is studied, especially as an error resolution strategy (Oviatt et al., 1998). Conversely, with this paper we start considering possible hyperarticulated realizations without paying particular attention to phonetic adaptations triggered by error resolutions, which has already been studied (Oviatt et al., 1998) and which will be however further researched in future studies, considering different classes of errors causing different types of adapting strategies. This was useful to prove if the alleged adaptations were the result of a general phenomenon occurring in human-machine interaction or if they were merely related to error resolution patterns.

The paper is structured as follows: in the first section, the description of tests used to collect data in two subcorpora is provided; the second paragraph explains the parameters taken into account in the conducted analysis; in the third section, results for both subcorpora are shown; finally, conclusions outline interpretation of the previously mentioned results.

2. *Corpus collection*

The corpus used for the analysis consists of annotated audio files recorded in two different test sessions. During the first session, users were asked to pose questions to a task-oriented dialogue system (Di Maro et al., 2017), designed to give information concerning paintings in a virtual museum. To guide users through the test, they were provided with twenty different conceptual classes, such as *Name of the artist*, *Name of the painting*, *Techniques*, *Iconography*, and similar. The combination of these classes and the paintings shown in the 3D scene – developed with the game engine Unreal Engine 4[1] – led users to ask questions

[1] https://www.unrealengine.com.

naturally. For this work, we selected one hundred questions posed by five different users, each of whom was asked to pose two questions per class (a simpler one and a more articulated one in terms of syntactic structures or vocabulary). The interaction was structured here as a question answering system. The same users were asked to recount the last movie they saw or the movie they liked the most, in order to be able to compare the previously mentioned human-machine interactions with spoken utterances collected in human-human conversations, thus in a different situational context for which the interlocutor was human and not virtual.

For the second session, we selected a different dialogue system, which was the well known general-purpose virtual assistant Google Assistant. This choice is motivated by the necessity to understand if different tools, their designed purposes and the quality of ASRs could be important in affecting the language production by human users. This test was divided in four phases. The first one was concerned with collecting users' personal data, useful to later map the extracted language features to their specific characteristics. Those who never used a virtual assistant on their smartphone were introduced to it through a brief explanation and examples. Afterwards, they were asked to complete three different tasks. Specifically, they had to interact with the assistant to find a place (street, route or a specific place) with Google Maps, to memorize an appointment on the calendar, specifying time, place and everything that was important to them, and to send a message to someone using Whatsapp. In this third phase, we collected 67 different recordings, corresponding to conversational turns by the users. The resulting corpus comprises clauses of different type: 65,67% consists of imperative clauses, 16,41% declarative clauses, 7,48% infinitive clauses, 4,47% noun clauses, and 5,97% interrogative clauses. The dialogue act (requesting information, saving an appointment, etc) were fulfilled sometimes in one turn and some other times in more than one turn, according to the expertise or preference of the user. Not-understood or misunderstood user inputs, for which error resolution reformulations were needed, were only the 15,67% of the total turns collected. Finally, they were asked to complete a questionnaire, to evaluate usability and satisfaction of the interaction. The collected evaluations were used to better interpret the results. The ten selected participants, who did not overlap with users tested in the previous experiment, were different for age, gender and attitude towards technology. As for the first test, we also recorded spontaneous narrations given by the same participants.

The recordings were paired with TextGrid files containing graphic and phonetic transcriptions. The phone alignment was automatically processed using WebMAUS (Kisler et al., 2017), whose outputs have been manually corrected. These files were therefore used to extract suprasegmental features, precisely pitch mean, pitch range, speaking rate and stressed vowels' formants, as we are going to illustrate in the next section.

3. *Corpus analysis*

Starting from an empirical observation of our subcorpora, it was noticeable that the collected human-machine interactions were characterized by specific hyperarticulation-related acoustic properties, such as loudness, reduced speaking rate, and the absence of

hypo-clear sounds. Hyperarticulate speech is used with "at-risk" listeners, such as children, hearing-impaired interlocutors, and non-native speakers. In each of these communicative risky situations, different parameters have been noticed: in infant-directed speech elevated pitch, higher pitch rate and stress on new words are used (Ferguson, 1977); with hearing-impaired speakers, amplitude and frequency values are higher and speaking rate is lower (Picheny et al., 1986). Since there are no distinct characteristics in determining hyperarticulation and stated that hyperarticulation in human-computer interaction is still not well defined (Oviatt et al., 1998), our long-term goal here is to find specific traits characterizing this peculiar context of spoken production. As a starting point, prosodic parameters, which are going to be presented in the course of this section, were selected.

As user commands appear to be hyperarticulated, their speaking rate is expected to be lower. As a matter of fact, in hyperarticulate speech, all syllables are pronounced, and many short pauses are used, resulting in an increased speech time. Pauses are here mainly used to pragmatically segment units of meaning, such as phrases, as if avoidance of system information overloading was intended. To get accurate results, we manually counted the perceived syllables (i.e. the ones really produced by the speaker, and not the ones expected to be produced) divided by the seconds of speech production.

Many studies on infant-directed speech (Fernald et al., 1989; Song et al., 2010; Gauthier, Shi, 2011), pointed out that higher pitch values, exaggerated pitch contours, and wider pitch ranges occur when asking for attention, communicating intentions, or even for lexical teaching purposes. Since being clear is fundamental in those situations, we may consider pitch trends to be crucial to also differentiate human-machine interaction utterances compared to spontaneous language productions. In fact, one of the pragmatic difference arising in human-computer interaction is concerned with the necessity of being understood, by means of a clear language, in a situation where the linguistic expertise of the interlocutor is perceived to be lower. For this research, we therefore decided to start from computing pitch mean and pitch range values, where the former is important to collect single utterances pitch trends, and where the latter is useful to compute differences in trends for each utterance. These values are manually calculated for every single file recording in our corpus, using the pitch frequency waves computed in Praat (Boersma, Weenink, 2018). The average values obtained for each user are also noted.

One assumption about hyperarticulate speech is that exaggerating speech sounds production leads to an extension of the vowel space (Story, Bunton, 2017; Wedel et al., 2018). This means that when vowels are hyperarticulated, they tend to be further from the centroid in the vowel space triangle. Conversely, when speakers do not articulate sounds carefully, they tend to produce vowels closer to the centroid, meaning that their articulation differences are less detectable. To compute the vowel position in the triangle, we extracted the formant values (F1 and F2) for each manually annotated stressed vowel in the corpus, using a Praat script. Moreover, since speakers can show their own articulation differences, we calculated the vowel space dispersion using the centroid of

each speaker's vowel triangle. Specifically, centroids were computed by grand means of vowels' average formant values (Koopmans-Van Beinum, 1983).

4. *Results*

In this session, the experimental results are presented. Firstly, we will have a look to the selected parameters (speaking rate, pitch contours and vowel space) within the first test, then the same will be displayed for the second one. Discussions and interpretations follow the outcomes.

4.1 Dialogue system-directed speech vs. spontaneous speech

As far as the speaking rate is concerned, a tendency can be outlined: speakers tend to talk slowly in interacting with a goal-oriented dialogue system (Figure 1). The system was indeed new to them; therefore, they thought that speaking fast could have jeopardized the understanding by the virtual interlocutor. Only for the fourth user the tendency is not applicable. Nevertheless, it must be noticed that this user started the task with a very high speed (7,3) and ended it with a lower value (3,49). In Figure 2, we can observe how the aforementioned speaker continuously changes the speed value, as a gradual attempt to adjust his speech to the interlocutor. The general tendency of a lower speaking rate is a first value in favour of the hypothesis of hyperarticulation in this particular speech use.

Figure 1 - *Speaking rate results – human-human vs. human-machine interaction (Users 1, 2= female; users 3, 5= male)*

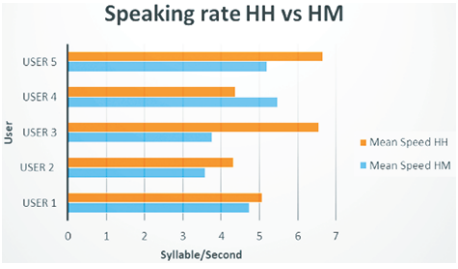


Figure 2 - *Speaking rate values – user 4 (male) – on the x-axis is the id-number of the utterance, whereas on the y-axis is the speed (syllables/second)*

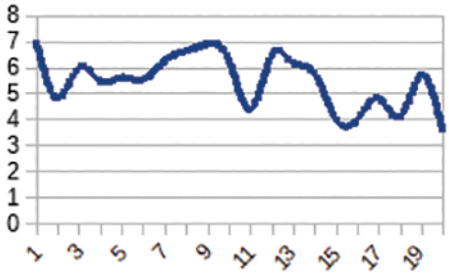


Figure 3 - *Mean pitch results – human-human vs. human-machine interaction (Users 1, 2= female; users 3, 5= male)*

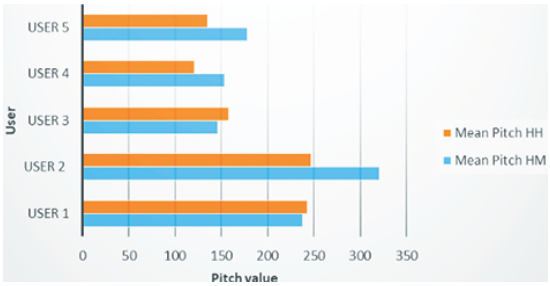
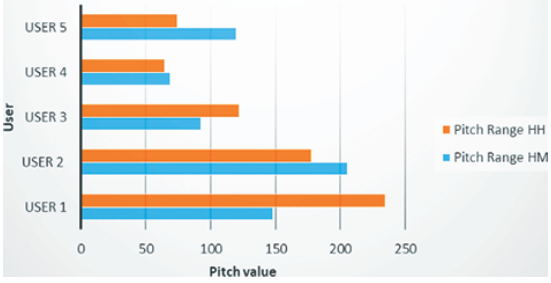


Figure 4 - *Pitch Range Results – Human-Human vs. Human-Machine Interaction [Users 1-2= female; Users 3-5= male]*



Conversely, the hypothesis of an evidence in pitch contours cannot be confirmed. As a matter of fact, pitch mean is higher in human-machine interaction in three out of five users, which is a difference not that significant in defining a tendency. But, although this increasing dynamics, pitch mean values are not so low compared to the ones occurring in narration. In fact, HM and HH values for user 1 and 3 are very closed to each other (Figure 3). Therefore, also pitch range values are not significant to underline the hypothesized tendency (Figure 4). The tendency of increasing pitch contours was empirically selected especially because of the speech performance of some users, such as users 2 and 5, who tended to stress sounds in an unnatural way. The non-confirmation of this prosodic adaptation can be actually motivated by the fact that in human-machine interaction there is no need to focus the computer's attention to specific lexical items via a pitch variation, as it could be useful in infant-directed speech (Oviatt et al., 1998). The pragmatic needs in this diaphasic situation are in fact of different kinds.

Concerning the vowel space based on formants values, we can infer that users tend to articulate sounds differently. For the female users (Figures 5, 6), the dialogue system's vowel triangle appears to be less extended that the one resulting from the spontaneous speech, especially what front vowels' F2 is concerned. Despite this reduced extension, for user 2 (Figure 6) the back closed vowel is further from the centroid compared to its equivalent in narration. On the other hand, for the other three users, vowel spaces are much more extended in interacting with the virtual agent, confirming the hypothesis of hyperarticulation.

Figure 5 - *Vowel chart - human-human vs. human-machine interaction (female)*

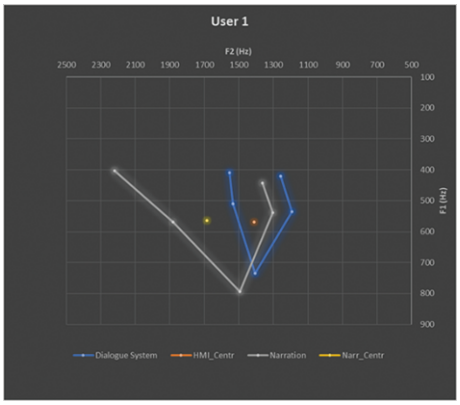


Figure 6 - *Vowel chart - human-human vs. human-machine interaction (female)*

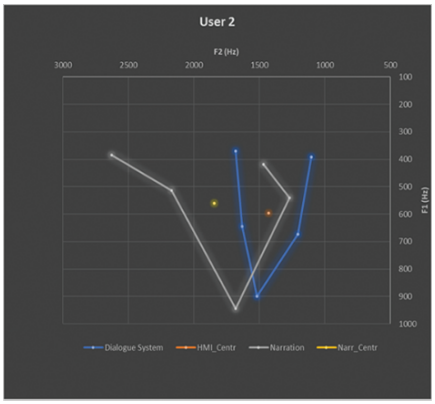


Figure 7 - *Vowel chart - human-human vs. human-machine interaction (male)*

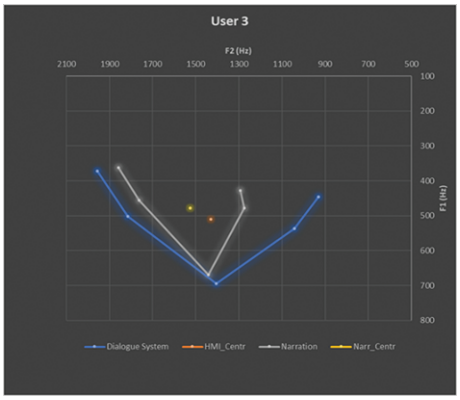


Figure 8 - *Vowel chart - human-human vs. human-machine interaction (male)*

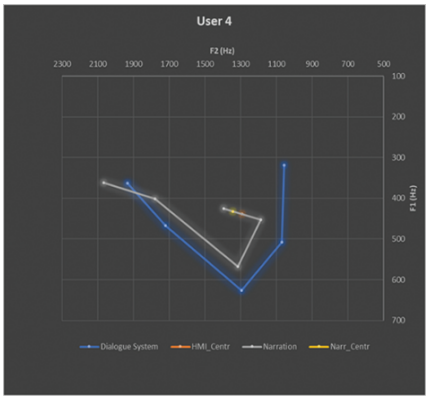


Figure 9 - *Vowel chart - human-human vs. human-machine interaction (male)*

4.2 Google assistant-directed speech vs. spontaneous speech

Even when interacting with Google Assistant, speakers tend to decrease their speaking rate, although the difference registered was not always significantly remarkable (Figure 10). Only for three (user 3, 7, 8) out of ten users this tendency was not observed. In Figure 11, we can notice how both users 3 and 7 start with a medium speed (on average 4-5 syllable/second), increase it up to 6, and eventually drop it to a slower value (2,94 for user 3 and 4,42 for user 7). User 8's behaviour is less stable, since his speaking rate increases and decreases at every utterance. Interestingly, the first utterance of each task (we refer to the points 1, 3, 4 on the x-axis in Figure 11 – user 8) is always slower than the others. The reason may lie in the tendency to be hesitant about the correct speaking rate to use to avoid misunderstanding; therefore, once the first turn is understood, the others uttered to complete the pragmatic act and to get the desired information is faster produced.

Concerning pitch values, six out of ten users employed this prosodic strategy regarding speech adaptation to computers, despite a major difference was evident only in users 1, 5, 7 and 9 (Figure 12). Pitch range values confirm the unreliability of this parameter in this context of use (Figure 13), as explained for the first test.

Contrary to what being observed for the previous test, the hyperarticulation hypothesis cannot be confirmed when analysing vowel spaces as a matter of fact, its extension is wider only in three users (1, 5 and 6), although this difference is not statistically relevant. For users 4 and 7, only specific vowels are over-articulated: the closed back rounded vowel for user 4 and user 7 (Figures 16, 20) and the open unrounded vowel for user 7 (Figure 20). Interestingly, in user 9 (Figure 22), the reduction phenomenon occurring in the interaction with the conversational agent is so strong that the back closed rounded vowel overlaps with the centroid.

The observable reductions can be explained with reference to the perceived experience during the interaction. Only users 2 and 4 stated that they had never conversed with a dialogue system. Consequently, not being an unprecedented experience, it makes them believe that they have a better expertise and can interact more naturally. The users who perceived themselves as experts are also the ones with whom the assistant had understanding problems. As a consequence, they evaluated the system use negatively.

Figure 10 - *Speaking rate results - human-human vs. human-machine interaction (users 1, 5 = females; users 6-10 = males)*

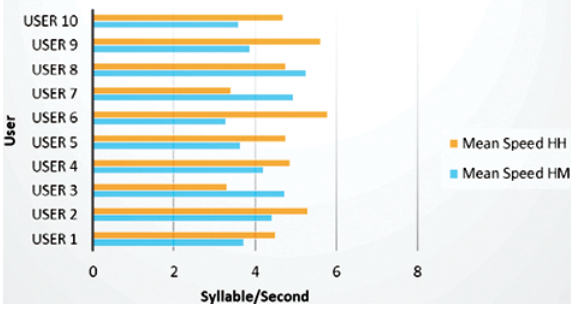


Figure 11 - *Speaking rate values - users 3, 7, 8 On the x-axis is the id-number of the utterance, whereas on the y-axis is the speed (syllables/second)*

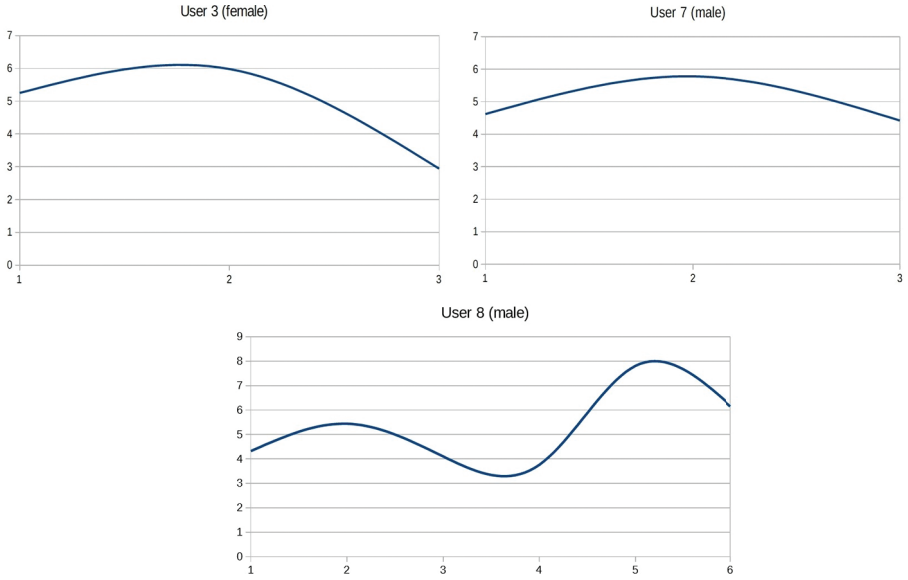


Figure 12 - *Mean pitch results - human-human vs. human-machine interaction (users 1, 5 = females; users 6, 10 = males)*

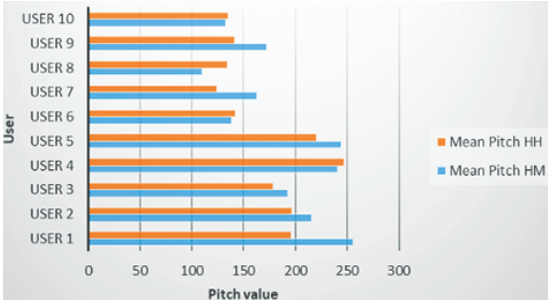


Figure 13 - *Pitch Range Results - human-human vs. human-machine interaction (users 1, 5= females; users 6, 10= males)*

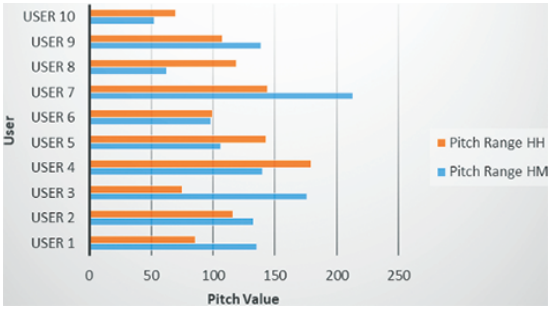


Figure 14 - *Vowel chart - human-human vs. human-machine interaction (female)*

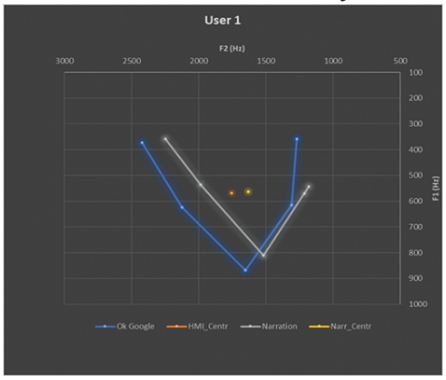


Figure 15 - *Vowel chart - human-human vs. human-machine interaction (female)*

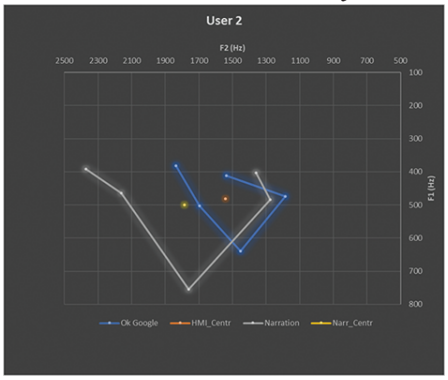


Figure 16 - *Vowel chart - human-human vs. human-machine interaction (female)*

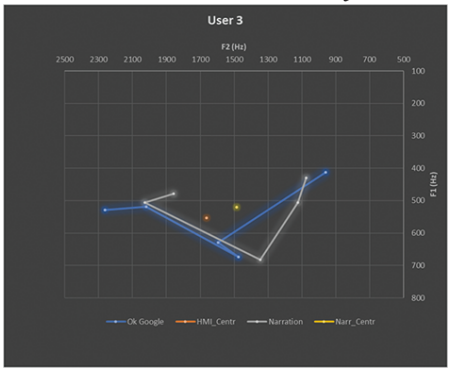


Figure 17 - *Vowel chart - human-human vs. human-machine interaction (female)*

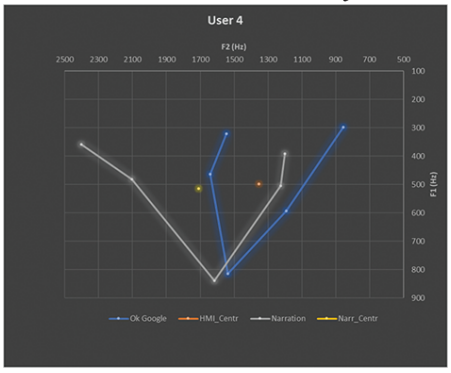


Figure 18 - *Vowel chart - human-human vs. human-machine interaction (female)*

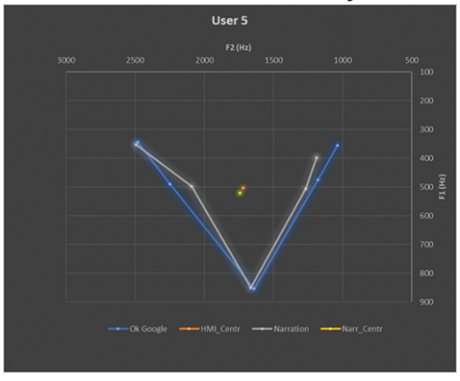


Figure 19 - *Vowel chart - human-human vs. human-machine interaction (male)*

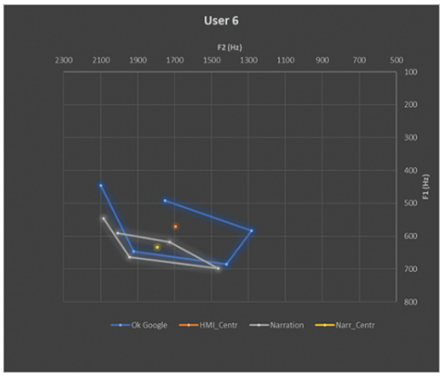


Figure 20 - *Vowel chart - human-human vs. human-machine interaction (male)*

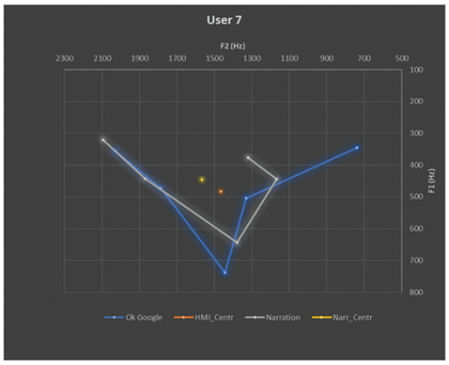


Figure 21 - *Vowel chart - human-human vs. human-machine interaction (male)*

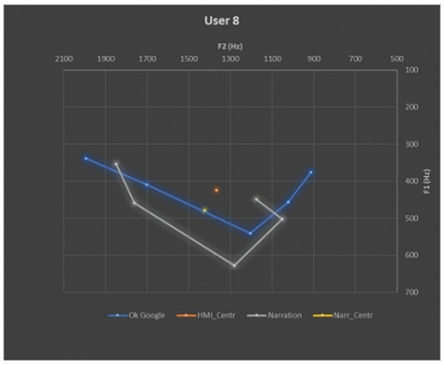
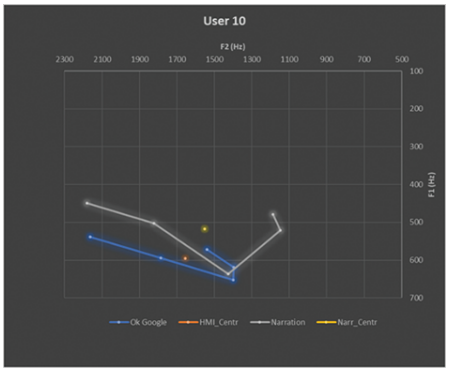


Figure 22 - *Vowel chart - human-human vs. human-machine interaction (male)*

Figure 23 - *Vowel chart - human-human vs. human-machine interaction (male)*



5. *Conclusions*

Although human-computer dialogue could be considered as an "at risk" context of interaction because of the present limited expertise of machines in the field of encyclopaedic knowledge and social signal processing, our results show that, as far as the selected parameters are concerned, no measurable acoustic difference were observed. As a matter of fact, only the speaking rate was slower as expected. Pitch and vowel space expansion did not change significantly, confuting the hypothesis of hyperarticulation. Precisely, pitch values are considered not important in defining hyperarticulation in human-machine interaction, as stated in other studies (Oviatt et al., 1998), whereas F1 and F2 extension is an intriguing approach which did not lead to the expected results, even in studies on infant-directed speech (Miyazawa et al., 2017). Nevertheless, a slightly difference in prosody, due to length and type of pauses or to amplitude and tone, can still be perceivable. For this reason, further phonetic and pragmatic analysis are needed to define how hyperarticulation or clearer speech production is generated in human-machine interaction.

All in all, users, even the ones who were not used to talk with a virtual agent as stated in the questionnaires, appeared to perceive the interaction as customary, especially as far as the interaction with the digital personal assistant was concerned, since speakers tend to identify these technological tools as something which is no longer far from our experiential horizon. In fact, in drafting a theory of mind (Goldman, 2012) applied to this special context of use, we can assert that users tend to interact accordingly to their preconception of the affordances of the tool: since nowadays technology is perceived as having unlimited capacities, speakers start adopting a communicative code which is closer to the one used in interacting with other humans.

The counter-adaptation resulting from the non-hyperarticulation of sounds in human-machine interactions can be explained with the increased error rate occurring when speakers try to speak clearer, as shown in other studies (Oviatt et al., 1998b; Soltau and Waibel, 1998). For this particular reason, the prosodic characteristics triggered in error resolution scenarios or in noisy environments (virtual agents for call-routing or for driver assistance) represent an interesting future investigation. Finally yet importantly, increasing the awareness of the specific pragmatic traits arising from this context of interaction will be advantageous in the development of better-performing acoustic and linguistic models.

Bibliography

Akira, H., Vogel, C., Luz, S. & Campbell, N. (2017). Speech Rate Comparison when Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. *Proceedings of Interspeech 2017*, 3286-3290.

Boersma, P., Weenink, D. (2018). *Praat: Doing Phonetics by Computer*. Version 6.0.37. Accessed 14.03.2018 from http://www.praat.org/

Di Maro, M., Valentino, M., Riccio, A. & Origlia, A. (2017). Graph Databases for Designing High-Performance Speech Recognition Grammars. IWCS 2017 – 12th International Conference on Computational Semantics – Short papers.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., De Boysson-Bardies, B. & Fukui, I. (1989). A crosslanguage study of prosodic modifications in mothers' and fathers' speech to preverbal infants, In *Journal of Child Language*, 16, 477-501.

Ferguson, C.A. (1977). Baby Talk as a Simplified Register. In Snow C.E., Ferguson C.A. (Eds.), *Talking to Children*. Cambridge, Cambridge University Press, 209-235.

Gauthier, B., Shi, R. (2011). A Connectionist Study on the Role of Pitch in Infant-directed Speech, In *The Journal of the Acoustical Society of America*, 130(6), EL380-EL386.

Goldman, A.I. (2012). Theory of Mind. In Margolis, E., Samuels, R. & Stich, S.P. (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science*. New York: Oxford University Press, 402-424.

Hartman, K.M., Ratner, N.B. & Newman, R.S. (2017). Infant-Directed Speech (IDS) vowel clarity and child language outcomes. In *Journal of child language*, 44(5), 1140-1162.

Kalashnikova, M., Goswami, U. & Burnham, D. (2018). Mothers speak differently to infants at-risk for dyslexia." In *Developmental science*, 21(1), 10-15.

Kisler, T., Reichel U.D. & Schiel, F. (2017). Multilingual Processing of Speech via Web Services. In *Computer Speech & Language*, 45, 326-347.

Koopmans-Van Beinum F.J. (1983). Systematics in Vowel System. In van den Broecke M., van Heuven V. & Zonneveld W. (Eds.), *Sound Structures*, FORIS Publications, Dordrecht, 159-171.

McMurray, B., Kovack-Lesh, K.A., Goodwin, D. & McEchron, W. (2013). Infant Directed Speech and the Development of Speech Perception: Enhancing Development or an Unintended Consequence? In *Cognition*, 129(2), 362-378.

Oviatt, S.L., MacEachern, M. & Levow, G. (1998). Predicting Hyperarticulate Speech During Human-Computer Error Resolution. In *Speech Communication*, 24(2), 87-110.

Oviatt, S.L. (1998). The CHAM Model of Hyperarticulate Adaptation During Human-Computer Error Resolution. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia paper 49.

Picheny, M.A., Durlach, N.I. & Braida L.D. (1986). Speaking Clearly for the hard of hearing II: Acoustic Characteristics of Clear and Conversational Speech. In *Journal of Speech, Language, and Hearing Research*, 29(4), 434-446.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y. & Brenier, J. (2007). An Acoustic Study of Real and Imagined Foreigner-Directed Speech. In *Journal of the Acoustical Society of America*, 121(5), 3044.

Soltau, H., Waibel, A. (1998). On the Influence of Hyperarticulated Speech on Recognition Performance". *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia paper.

Soltau, H., Waibel, A. (2000). Specialized Acoustic Models for Hyperarticulated Speech. In *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3, 1779-1782.

Song, J.Y., Demuth, K. & Morgan, J. (2010), Effects of the Acoustic Properties of Infant-Directed Speech on Infant Word Recognition. In *The Journal of the Acoustical Society of America,* 128, 389-400.

Stent, A.J., Huffman, M.K. & Brennan, S.E. (2008). Adapting Speaking after Evidence of Misrecognition: Local and Global Hyperarticulation. In *Speech Communication*, 50(3), 163-178.

Story, B.H., Bunton, K. (2017), Vowel Space Density as an Indicator of Speech Performance. In *The Journal of the Acoustical Society of America*, 141(5), EL458-EL464.

Wedel, A., Nelson, N. & Sharp, R. (2018), The phonetic specificity of contrastive hyperarticulation in natural speech. In *Journal of Memory and Language*, 100, 61-88.