# Analyzing Explicit(Speech) Modalities from Human-Human Interactions for building Context about a Robot-Assisted Dressing Task

Antonella Camilleri[1]

***Abstract*— Robots that assist in the Activities of Daily Living (ADL), such as dressing, can support an aging population and lack of caregivers. These interactions are treacherous due to the implication of having end users like older adults who require great care and control on the overall interaction between the human and robot. The goal of collaborative interactions, such as ADLs, incorporates the success of the ADL task while taking into consideration the direct or indirect effect of the surrounding environment on the task and the user itself. Therefore a need to measure distractions or lack of collaboration between user and robot is vital. Collaboration in tasks like these evolve around the task itself and any measure of information from the interaction used to acknowledge progress needs to be carefully evaluated. Progress in task and state of interaction is context; and not being able to identify this, can be a sound indicator of distractions or lack of motivation to collaborate. Looking at human-human interactions for an assisted dressing task, speech utterances modality, together with other modalities, are used as a method of classification of the progress in the assisted task; by using LSTMs. A better classification of sequence state of task, due to speech utterances, would indicate that this explicit modality can be important to measure the level of collaboration and progress in such task.**

## I. INTRODUCTION

Human-Robot Interaction (HRI) revolves around the robot's capability in practical tasks to allow an effective and successful human-robot collaboration. Assisting humans with daily tasks requires a robotic system which is capable of assessing the current state (*context*) of all entities in the interaction environment [1]. This current state needs to be based on the latest past, present or/and abrupt forthcoming interactions. Modeling or simulating interactions with humans is challenging. The complex and dynamic nature of the different states of interaction requires knowledge on states such as spatial, temporal and resources of the robotic system, the states of the human and objects interacting with the robot. All of these states form part of the environment surrounding the interaction which often holds information that humans use to infer about the required action. This ability enables the execution of tasks in a very practical and collaborative way. Furthermore, the collaboration between humans comes with the ability to distinguish between shared common grounds, ones own knowledge and collaborators' knowledge [2]. This knowledge shapes the required action making the interaction realistic and to a safer extent because the action is based on current states and not only on past modeled states.

Possibly, the obvious modality used to infer information about a collaborative task between humans is the symbolic representation of the explicit modality of speech utterances. Nonetheless, once the knowledge of how to carry out a task is known, speech utterances are observed to be limited in interactions between humans [3]. The hypothesis is, that due to the lack of speech utterance, such explicit modality is only related to extreme interaction states such as acknowledgment of progress in task or to correct actions. If this hypothesis is proved to be true than the lack or erroneous speech utterances can indicate different unknown states (*situations* that arise provided a specific *context*) in the interactions. These errors can be deduced as distractions from the collaborative tasks. Hence the objective of this paper is to examine if speech utterance in time are directly linked to affirming progress in the sequence of the task, meaning sequence prediction of the task is improved. The collaborative task examined is the assistive task of dressing an outer layer of jacket between humans.

[1]Antonella Camilleri is a PhD student at the University of the West of England, Bristol Robotics Laboratory, Coldharbour Lane, BS16 1QY, Bristol, United Kingdom. antonella.camilleri@uwe.ac.uk

## II. MOTIVATION

A robot that provides support with dressing has to detect when the user is distracted and what the current states are in order to complete the task safely. These distractions can be instilled from noise or commotion in the environment or simply by a lack of attention from the user. Hence, knowing what lacks in explicit modalities when a user is distracted or not can allow the robot to adapt and perform the right action. The assisted task of dressing is complicated and depending on which part/sequence of task the robot is in, the current state and selection of actions can vary along progress of task. Consequently, establishing if an explicit speech utterance in time is related to the sequence of task is or not is imperative for a safe HRI.

## III. EARLIER WORK

In the past, belief models for situation awareness have been implemented using Markov Logic provided that model spatio-temporal frames include epistemic information. Analyzing HRI requires methods that can handle multivariate time series inputs. One machine learning method used to train such data is the variation of recurrent networks called Long Short-Term Memory Units (LSTMs). This has been used to extract contextual features from multi-modal inputs in order to classify emotion or sentiment or action selection. Furthermore, in [4] skeleton data trained in a three-layer LSTM has been implemented to infer users interactive intent. Prediction of sequential tasks can be seen in [5] in which goal location of reaching motion is implemented and combined with LSTM to predict the next steps in sequence. The benefit

of using LSTM is the ability of creating long-term dependencies first by extracting features from each modality and then by looking at the relationship between the modalities. This method particularly holds the state of the neurons which is ideal for predicting sequences.

## IV. PROPSED METHOD

The proposed method of implementation is LSTM. LSTMs are a variation of RNN with the ability to perceive previously in time and classify sequential input [6]. However, LSTMs are better because of the no vanishing gradient problem, but mostly because they have a forget gate with a purpose of linking distant occurrences to a final output [7] . Furthermore, LSTM preserve the error and can be back propagated through time and layers allowing recurrent nets to continue to learn over many time steps. This opens a channel to associate causes and effects remotely. This property in LSTM addresses the challenges of having delayed reward signal in realist environment interactions. Also, having a stacked LSTM architecture allows the hidden state of each level to operate at different timescale which is very likely to happen in this kind of interaction. This implies that user distractions or successfully completion of task can be predicting by observing the state of the memory cell representing this data input.

### *A. Experiments Procedure*

The dataset used in this work was gathered during a dressing task [3] where 12 users were given assistance from another human posing as a robot. The users were wearing a motion tracking suit (Xsens) to record the spatio-temporal, position and orientation, of 23 points on the body. In the task, the users had to collaborate to put on a jacket several times. Each user put the garment on three times. Each dressing task took approximately 40s. For eye gaze tracking, the users were wearing a Tobi Pro Glasses which recorded eye gaze during the task. Video recordings were used to extract speech utterance in time. Speech was the modality used by the human getting dressed to provided instructions the other human performing the dressing task. Data processing of gaze with respect to shoulder and torso from the 23 point on the body were extracted and used for the prediction model. Additionally, the three main sequence steps (hand-elbow-shoulder) of the dressing task where encoded and presented as part of the output of the model to be predicted. Being able to obtain a higher categorical classification with utterances in the interaction would indicate that speech utterances frequency can be linked to progress in a collaborative task. Due to having discontinuities in speech utterances, a dual pipeline approach to the LSTM network was considered. A LSTM network was used to process word embeddings (explicit utterances) and another LSTM networks for extracting features from the implicit modalities. Provided that speech utterances are not continuous the two LSTMs models will be trained separately and combined on a concatenation layer. These combined features are fed through fully connected LSTM layers leading to an output layer with 3 outputs (hand-elbow-shoulder). These outputs would respectively represent the dressing up to the hand, elbow or shoulder (completed task).

## V. EXPECTED RESULTS

The preliminary analysis of time difference between speech utterances and the progression between the three stages of the dressing task suggest that sequence classification of task state is likely to be improved. This indicates that explicit speech utterances can be related to the main state changes of this collaborative task and such modality can be used to extract context of progression to achieve the final goal in a collaborative task.

A better prediction of task sequence indicates that speech utterances are an explicit modality of interaction used to dictate progression or need of change in the interaction approach between human-human interactions.

## VI. CONCLUSIONS

The current paper introduces an approach of examining the importance of explicit speech utterances in relation to progress in a collaborative task between humans. The preliminary results indicate that explicit speech utterances are important to measure collaboration and progress in the final task objectives.

As future work, we plan to finalize results and further evaluate the incorporation of similar explicit speech utterances as one of the modalities to assess the level of collaboration between a robot and a human in a real scenario. Specifically, the evaluation of collaboration will be based on alterations in the modality of speech when distraction are introduced in the environment of the interaction.

## VII. ACKNOWLEDGEMENT

This work has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 721619 for the SOCRATES project.

## REFERENCES

[1] P. Menezes, J. Quintas, and J. Dias, "The Role of Context Information in Human-Robot Interaction," *RoMan 2014 Workshop on Interactive Robots for aging and/or impaired people*, pp. 4–7, 2014.

[2] J. Quintas, G. S. Martins, L. Santos, P. Menezes, and J. Dias, "Toward a Context-Aware Human-Robot Interaction Framework Based on Cognitive Development," *Ieee Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–11, 2018.

[3] G. Chance, P. Caleb-Solly, A. Jevtic, and S. Dogramadzi, "What's up? Resolving interaction ambiguity through non-visual cues for a robotic dressing assistant," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 284–291, IEEE, 8 2017.

[4] K. Li, S. Sun, J. Wu, X. Zhao, and M. Tan, "Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-like Decision Mechanism," 2018.

[5] H. C. Ravichandar, A. Kumar, A. P. Dani, and K. R. Pattipati, "Learning and Predicting Sequential Tasks Using Recurrent Neural Networks and Multiple Model Filtering," pp. 331–337, 2016.

[6] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," Tech. Rep. September, 2014.

[7] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks," *PLoS ONE*, vol. 11, no. 1, 2016.