Human-Robotic Agent Speech Interaction

INSIDE Technical Report 002-16

Rubén Solera-Ureña and Helena Moniz

February 23, 2016

1. Introduction

Project INSIDE investigates how robust symbiotic interactions can be established between children with Autism Spectrum Disorder (ASD) and robotic (virtual) agents during the joint execution of collaborative tasks conducted in a hospital environment. This report summarizes the work conducted by the Spoken Language Systems Laboratory (L^2F/INESC-ID Lisboa) on Project INSIDE Taks 2: Multimodal interactions in networked systems, during the period August 2014 – January 2016.

2. Robust child-robotic (virtual) agent interactions

2.1. Automatic speech recognition (ASR)

Regarding the ASR line of work, our goal here is to endow the robot with the ability to automatically recognize single keywords or more complex utterances spoken by their teammates at therapeutic sessions (children with ASD but also therapists).

This scenario is far from those more typical in the automatic speech recognition field, where well known, stablished ASR solutions provide competitive results nowadays. This particular context presents a number of severe difficulties that must be addressed in this project: multiple, moving and different type of speakers (children and adults), distant speech, room reverberations, different noise sources (background conversations, robots' engines and fans), microphone distortions, and mechanical vibrations, among others. Specially challenging regarding the objectives of this project are the spontaneous, limited and very short speech interactions typical of children with ASD, and the severe restrictions on experimental work imposed by the hospital scenario and the involvement of children.

Work on this line has focused on an in-depth initial analysis of the characteristics of the project's scenario and an evaluation of the practical significance of the aforementioned identified difficulties. A preliminary experimental setup for signal acquisition and automatic speech recognition, upon which the final ASR system will be built, was then proposed; this baseline system is currently under continuous development. Both laboratory-work and field-tests at the hospital scenario (pilots held on May 14 and July 23, 2015) are being conducted to identify potential problems, refine the speech acquisition and ASR modules and evaluate their performance. A brief description of the work made in this line is given below.

2.1.1. Speech acquisition

The design of the preliminary speech recording system has been made in accordance to the following three main criteria:

- Subjective quality of the recorded speech.
- Influence of the recording front-end on the objective performance of the automatic speech recognition system.

- Appropriateness of the setup for a hospital scenario with autistic children attending their therapeutic sessions, i.e., the recording system must not be invasive, scare, distract or draw the attention of the children.

The speech acquisition setup must address the difficulties mentioned above. For this purpose, two main alternatives were considered: on-board and off-board configurations. Microphones mounted on the robot seem a more natural approach towards our final objective of endowing the robot with some degree of autonomy. Furthermore, they would be closer to the speech sources but also more exposed to noise and mechanical vibrations from the robot. A network of microphones distributed around the walls/ceiling of the therapy room would be less affected by noise from the robot but could suffer from distant speech, which we consider potentially more harmful in this case.

Several on-board microphone options have been considered and a first proposal based on the simultaneous use of a directional RODE VideoMic Pro microphone and an omnidirectional circular array of 8 Micro Electrical Mechanical System (MEMS) microphones, both mounted in the robot, was adopted. The former is placed in the frontal part of the robot looking forwards, whereas the latter is placed on the robot's head. The directional RODE microphone provides the best speech quality when the speaker is facing the robot. On the other hand, the MEMS array presents some interesting advantages such as very low cost, being small and unobtrusive, geometrical flexibility, an omnidirectional pattern more suitable when children/therapists are moving around the room, and the possibility to implement array processing techniques, at the expense of slightly lower overall quality. The 4-microphone linear array of a Kinect system inside the robot's shell (used for video acquisition) was discarded due to an unsuitable arrangement of the microphones, lower speech quality with respect to the other options, and suffering from a greater extent from the robot's noise.

Further and deeper evaluation at the hospital scenario of the proposed setup will be carried out in the future. In particular, a decision on the convenience of keeping both microphones in our setup (or just one of them) must be made. The MEMS array currently used will soon be replaced with a new, updated version provided by STMicroelectronics[1]. This new, definitive setup will allow us to accelerate the work on the automatic speech recognition module. The possibility of using the MEMS array to track the position of the children in the therapy room will be also studied. Finally, the intention by one of the project partners to use several Kinect sensors spread around the therapy room for other purposes will also allow us to make a more accurate comparison of the proposed setup against an off-board configuration.

2.1.2. Automatic speech recognition module

Our work on this subsystem starts from a preliminary, general-purpose speech recognizer implemented using the well-known *Kaldi* toolkit[2]. A noise-free, large-vocabulary, continuous speech Portuguese database (adult speakers) [1] has been used to train a classical ASR system with Gaussian Mixture Models (GMM). Preliminary, non-exhaustive laboratory tests using the selected

[1] http://www.st.com
[2] http://kaldi-asr.org

microphone setup show lower ASR performance with respect to the database baseline results due to the expected acoustic environment mismatch (sound proof room speech vs. office environment speech). The arrival of a new version of the MEMS array will also allow us to evaluate with a higher precision the effect on the speech recognition module of the recording setup used on the robot.

Oncoming work on this line will pursue incremental improvements of the preliminary speech recognizer, with the goal of achieving a robust, specific speech recognition module for our particular scenario (therapeutic sessions with autistic children in a hospital environment). Some specific difficulties will be successively addressed in this context: distant speech, room reverberations, microphone distortions, and different types of noise (fan, background conversations, robot's engine), among others. For this purpose, work on different sections of the baseline recognizer will be carried out including voice activity detection, speech enhancement, robust speech parameterization, and acoustic model adaptation (to noise, distortions, and children's speech). The latter case is especially important in this project, since the acoustic characteristics of children's speech are markedly different from those of adult's speech used to train the baseline recognizer. For this purpose, a specific children speech database will be used [2]. In addition, the difficulty of dealing with the scarce and very short speech interactions typical of children with ASD is under assessment and could force us to reorient work on this line towards other speech interaction scenarios (keyword spotting, guided activities using a virtual agent on a tablet device…). Expected work on this line would end with the integration of the final speech recognizer on the robot's architecture and the execution of human-robot speech interaction field-tests at the hospital scenario.

2.2. Text-to-speech synthesis (TTS)

Project INSIDE features a robot that interacts with children with Autism Spectrum Disorders by using speech, movement and a representation of facial expressions with speech-lip synchronization. Regarding the TTS synthesis line of work, our goal here is to endow the robot with a natural vocal interface to interact with the children. Synthesized speech and data required for lip synchronization on the robot (phoneme chain and their durations) are provided by L^2F/INESC-ID and VoiceInteraction own speech technologies (ASR and TTS).

The DIXI TTS synthesizer engine [3] was employed in the first pilot (May 14, 2015) performing an European Portuguese male voice (named "Vicente"). Lip synchronization was not developed on the robot at that moment. Thus, although the required phoneme and duration data needed for lip synchronization was available, only synthesized speech files were supplied and played on the robot speakers.

Due to the difficulty of achieving natural emotions in synthesized speech, which are essential for an improved engagement experience of the children with the robot, pre-recorded human speech was employed for the robot's vocal output in the second pilot held at the hospital (July 23, 2015). Since lip synchronization was then implemented in the robot setup, phonemes and their durations were also provided. To obtain the phoneme chain and their durations from the pre-recorded speech audio files,

the ASR module of the AUDIMUS system [4] was employed, using monophone acoustic models to perform a forced alignment of the audio files with their corresponding transcriptions.

All this processing is performed on a dedicated computer which supplies a network service accessed by the robot to get the audio files, phonemes and their durations. In the particular scenario of the experiments no internet connection was available; therefore, a laptop providing the service through a local network has been used.

3. Computational paralinguistics (social signal processing)

The main goal of this line of work is to extract any non-linguistic information from children's speech that can be used to create more natural and engaging interactions between children and their robotic teammates in joint cooperative activities. Such information will be used to plan and drive the robot's course of action in order to perform a proactive, anticipatory and collaborative behavior towards the children subject to the therapeutic sessions. In our work, we primarily focus on researching acoustic features of autistic children's speech that can be used to:

- Determine the presence and degree of severity of Autistic Spectrum Disorder in children.
- Determine long-term personality traits of children.
- Recognize short-term emotional states of children during the course of therapeutic sessions.

Work performed in this line includes a deep study of the state-of-the-art in the field and the implementation of a baseline system for personality attribution on adults, based on the works presented in the Interspeech 2012 Speaker Trait Challenge - Personality sub-challenge [5]. We are now exploring and evaluating novel speech features computed on specific temporal/lexical contexts that are expected to be more appropriate than the rather generic features commonly used in the literature. Also, two psychologists annotated a subset of non-autistic children interactions regarding personality traits. The annotation procedure followed three main steps:

- A subset of Interspeech 2012 Personality sub-challenge.
- A subset of children interactions with only audio information.
- A subset of children interactions with audio and video information.

The goals of these tasks are to establish a baseline model for adult and non-autistic children speech tackling the comparison of personality traits predictions amongst the two age dependent groups, and then compare with data from autistic children, either from the small samples of the pilots conducted already at the hospital scenario or from other corpora obtained during the project. Comparisons of personality traits across age groups and autistic children are quite scarce in the literature and may enable us to better understand the correlations between personality traits, emotions, and Autism Spectrum Disorders.

Oncoming work seeks to further develop our baseline system and build solid personality and emotion models for child interactions that can be finally exported to the project scenario with autistic children. One of the biggest difficulties in this field is the scarcity and small size of the speech datasets

available for researching personality and emotion recognition, especially with children. Thus, a continuous search for new datasets will be performed.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, under Post-doc grant SFRH/PBD/95849/2013, and by project CMUP-ERI/HCI/0051/2013 (INSIDE).

References

[1] J. Neto, C. Martins, H. Meinedo and L. Almeida: "The Design of a Large Vocabulary Speech Corpus for Portuguese". In Proceedings of EUROSPEECH 97, Rhodes, Greece, 1997.

[2] A. Hämäläinen, F. Miguel Pinto, S. Rodrigues, A. Júdice, S. Morgado Silva, A. Calado, M. Sales Dias: "A Multimodal Educational Game for 3-10-year-old Children: Collecting and Automatically Recognising European Portuguese Children's Speech". In Workshop on Speech and Language Technology in Education, Grenoble, France, 2013.

[3] S. Paulo, L.C. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana and H. Moniz: "DIXI – A Generic Text-to-Speech System for European Portuguese". In 8th International Conference on Computational Processing of the Portuguese Language (PROPOR 2008), LNAI 5190, pp. 91-100, Springer-Verlag, Heidelberg, Germany.

[4] H. Meinedo, D. Caseiro, J. Neto and I. Trancoso: "AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language". In 6th International Workshop on Computational Processing of the Portuguese Language (PROPOR 2003), LNAI 2721, pp. 9-17, Springer-Verlag, Heidelberg, Germany.

[5] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi and B. Weiss: "A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge". Computer Speech & Language, vol. 29, n. 1, pp. 100-131, 2015.