Towards Autonomous Child-Robot Interaction


I. INTRODUCTION

The ALIZ-E project [1] aims at designing and developing long-term, adaptive social interaction between robots and child users (8-11 years old) in real-world settings, for which a conversational human-robot interaction system has been developed [2]. In this context we present the auditory and visual perception components that have been specifically build for the purpose of supporting verbal and non-verbal human-robot interaction.

II. AUTONOMOUS PROCESSING

A. Operator’s Voice Modification

For studying the child-robot interaction a Wizard-of-Oz set-up is being used that allows children to interact with the robot being partially operated by an unseen human operator. This is an essential step towards building an autonomous system as it enables the collection of data and experimenting without relying on specific system components. To this purpose, we have developed a voice modification that allows the operator to speak in real-time through the robot with a voice resembling the robot’s voice. The incoming operator’s speech signal is time-scaled using the robust WSOLA algorithm [3]. The modified signal is then resampled to its original length and played though the robot’s speakers. This operation results in shifting the original signal’s spectrum which when played through the robot’s speakers creates a “robotized” voice effect. Currently, additional techniques are being explored for individually modifying parameters of the speech, e.g. formants and voice quality.

B. Sound Source Localization

An aspect which is important in child-robot interaction is the ability of the robot to localize the direction of sounds. Robot audition systems often estimate the time delay on arrival (TDOA) utilizing microphone arrays which consist of many sensors located in free space. These assumptions are not fulfilled in real-world robots such as Nao, where a small number of microphones is mounted on the robot. Scattering of the sound wave along the shape of the robot has a significant influence on the conventional TDOA estimation. In order to address this, we have implemented a Generalized Cross-Correlation (GCC) based method which utilizes a set of pre-measured TDOAs, followed by parabolic interpolation [4]. In addition, our experiments have highlighted the importance of speech denoising for performing localization under noisy conditions. The effect and performance of different denoising techniques have been assessed under internal and ambient noise types. Finally, we have demonstrated that taking into account the microphones’ frequency responses in the GCC improves the localization’s accuracy [5].

C. Non Verbal-Behavior Recognition

As means of improving the ability of the robot to respond in a way that facilitates productive and enjoyable interaction experience, we proposed several components for the recognition of the non-verbal behavior of the child. More precisely, we developed automated recognition and dynamical analysis of body gestures and facial expression. Namely, tracking and recognition of human arms movements [6], 3D sensor based body gesture recognition [7], facial expressions recognition [8], [9], audio speech emotion recognition [10] and audio visual emotion recognition [11].

REFERENCES