ASR and TTS for Voice Controlled Child-Robot Interactions for Treating Children with Metabolic Disorders

Giacomo Sommavilla, Fabio Tesser, Giulio Paci, and Piero Cosi

Institute of Cognitive Sciences and Technologies (ISTC)
National Research Council (CNR)
Padova, via Martiri della Libertà, 2, 35137 Padova, Italia
http://www.pd.istc.cnr.it

Abstract. Artificial companion agents are becoming increasingly important in the field of healthcare, particularly when children are involved, with the aim of providing novel educational tools, supporting communication between young patients and hospital personnel and taking on the role of entertainment robots. The principal application of the European FP7 project ALIZ-E is the development of an educational robot companion for children (target age 8-11) who are affected by metabolic disorders such as diabetes and/or obesity. The purpose of this educational robot is to enhance the child's well-being and facilitate therapeutic activities in a hospital setting. Though speech comprehension, in particular acoustic analysis applied to automatic speech recognition of children's voices, has been investigated extensively by speech technology researchers over the last two decades, most of the literature is focused on the English language. Given that the primary evaluation site of the ALIZ-E project is located in an Italian hospital, ISTC-CNR researchers have focused on the application of speech technologies in the Italian language as spoken by children. This chapter also outlines the investigation of voice adaptation techniques in children ASR. It reports on an experiment of ASR in a real case child-robot interaction scenario in a hospital setting, and presents the data collection for a corpus of annotated Italian children's speech. The study shows how in the production of speech the robotic companion must be able to convey to the child the identity and the emotional state of the speaker, in addition to verbal content. In addition, the robot companion must focus on particular words that are important in the communication with the child. In this chapter, we describe the tools and the modules needed to build a Text to Speech engine implementing these features designed for the Italian language.

1 Aliz-e Project

Robots helping humans in surgery are a well established trend in healthcare technology. However, artificial companion agents are becoming increasingly important as well (Baxter et al. 2011). Companion robots can be useful in several situations (especially with regard to children), by providing health education,

supporting communication between patients and healthcare professionals and by entertaining patients in hospitals. In addition, children are eager users of new technologies, which can enrich their experience, specifically for educational and therapeutic purposes (Tartaro & Cassell 2006).

These companion agents have been proposed for, and applied to, a number of roles, such as support (Kidd 2008) and motivation (Janssen et al. 2011). Robots play these roles within defined task contexts. However, the more general aim of a robot that could act as a generic, task independent, social peer is yet to be achieved. Such a companion agent would at the very least be required to operate in the real world over extended time-scales, making use of multiple modalities (both perceptual and physical) to engage the human interactant (Tapus et al. 2007).

The challenges of Child Robot Interaction (CRI) outside of the laboratory setting are significant, requiring the confrontation of both technical and pragmatic issues. As children are not "mini-adults", and this fact is very much apparent in the context of CRI, they bring an imaginative investment to encounters with robot agents. This imaginative element is critical to the exploration of how we can develop technologies and systems for social interaction between children and companion agents.

The ALIZ-E ("Adaptive Strategies for Sustainable Long-Term Social Interaction") project[1] is a European Project funded by the European Commission's 7th Framework Programme. This project began in April 2010 and will last four and a half years.

The aim of the ALIZ-E project is to develop embodied cognitive robots and to study the theory and practice of affective interactions between children and cognitive robots for either a sporadic or an extended period of time. Specifically, the ALIZ-E project is testing robots with children (target age 8-11), affected by metabolic disorders, such as diabetes or obesity. Our ultimate goal is to employ robots for supporting the children's well-being and facilitating therapeutic activities in a hospital setting.

The project, coordinated by Dr. Tony Belpaeme (University of Plymouth), involves a consortium of seven partners: the University of Plymouth (UK), Vrije Universiteit Brussel (Belgium), Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (Germany), the Imperial College (UK), the University of Hertfordshire (UK), the National Research Council (Italy), the Netherlands Organization for Applied Scientific Research (The Netherlands), Aldebaran Robotics (France) and Fondazione San Raffaele del Monte Tabor (Italy).

The principal application of the ALIZ-E project methodologies is the development of an educational robot companion for young diabetic patients. Within the ALIZ-E project, Henkemans et al. (2012) have studied children's diabetes self-management and their experience with illness, with regard to their quality of life.

They interviewed children and caregivers, as well as parents. They discovered that usually the parents play a prominent role in diabetes self-management.

[1] http://www.aliz-e.org/

However, Henkemans et al. (2012) state that it is important for the children to become more proficient and experienced in their self-management at an earlier age, because they start developing a need for autonomy during puberty. Children seem to accept the illness as part of their life, but they experience difficulties in specific situations, especially outside their everyday life (e.g., during sports and holidays). These difficulties may have a strong negative impact on their mental and physical well-being by causing insecurity, fear, listlessness and tiredness. Henkemans et al. (2012) came to the conclusion that children could benefit from social robots offering motivation, training, and (parental) monitoring and support. In order to prevent stigmatization, a robot would need to act as a buddy and not as a support tool in managing diabetes. Experiments have been carried out in the Department of Paediatrics within the "Ospedale San Raffaele" in Milan and in summer schools for children with diabetes in Misano Adriatico. The principal testing environment for the ALIZ-E project is the Aldebaran Nao robot, a 60 cm tall humanoid already widely used in the robotics research field.

1.1 Description of the Quiz Game

The ALIZ-E robotic environment (Kruijff-Korbayová et al. 2012) implements several game-like activities that a child can undertake. In this section, the Quiz Game is described. It is the most challenging activity implemented in the research field of child-robot verbal interaction.

The Quiz Game interaction starts with the robot explaining the rules that are very similar to those of the "Who Wants to Be a Millionaire?" game show.

The child and the robot are two players who take turns asking each other questions. Whoever asks also provides multiple choice replies for the other player. Players can answer by providing the entire reply or the corresponding number (or letter). Feedback about the correctness of each answer is given: a correct answer gets positive feedback; an incorrect answer prompts another response. If the second answer is wrong, the system provides the correct answer. Feedback consists of a combination of verbal and non-verbal signs. After a round of three questions, the robot summarizes the results and the players switch their roles. The game ends when one of the two players asks to leave.

The topics of the questions and answers span the domain of diabetes and nutrition. More generic questions, taken from a Trivial Pursuit game for children, are interspersed throughout the interaction.

An LCD tablet is wirelessly connected to the ALIZ-E integrated system. It is used to display Quiz questions and answers to the players. The tablet is mounted on a support which allows one to flip the display, making it is possible to show information to both players. But while the child actually reads the sentences, the robot just pretends to take advantage of the device. Figure 1 shows the setup of the interaction.

It has been decided not to use the Nao built-in microphones for recording, since they are low quality microphones. Moreover, two of them are placed under the robot's loudspeakers, one is placed on the robot's nape, near the fan, and

Fig. 1: Screenshots taken from the video recordings of experimental interactions.

all four microphones record a lot of noise resulting from motors and electronic circuits.

Instead of the Nao's microphones, a hand-free close-talk radio microphone[2] has been used. This microphone has been selected in order to ensure good sound quality while interfering as little as possible with the child-robot interaction and allowing freedom of movement to the user. The microphone has been connected to a Zoom H4n sound card that can record by either using a computer (the device is used as an USB audio input interface) or saving audio data on an SD memory card.

1.2 The Aliz-e Integrated System

From a technical point of view, one of the main problems in the ALIZ-E integrated system is related to the need of having components that should (a) allow access to low-level hardware devices of the Robot, (b) perform heavy computations, (c) typically run on separate, more powerful machines, (d) be coordinated concurrently and/or (e) react to (typically asynchronous) events.

Programming languages such as C/C++ can well suit low-level, heavy computational tasks, but it can be difficult and time consuming to manage concurrency, network communication and event handling with those languages.

The Urbi environment, developed by Aldebaran ALIZ-E project's partner, provides the high level `urbiscript` scripting language which can orchestrate

[2] Proel RM300H, Radio frequency range: UHF High Band 750-865 MHz, Microphone: headset HCM-2.

complex organisations of low level components called "UObjects" in highly concurrent settings.

A C/C++/Java program can be made accessible as an UObject by "wrapping" the upstream program into a C++ (or Java) class inheriting from Urbi::UObject, then binding in `urbiscript` the methods that should be accessible from there. Also, events can be defined in order to orchestrate the components concurrently. An `urbiscript` event can eventually carry a "payload", i.e., transmitted data attached to the event itself.

The most interesting issues about the integration of the ASR and the TTS components into the ALIZ-E system are explained Sections 2.4 and 3.4, respectively.

1.3 Speech Technology in Aliz-e

Speech is the principal mode of communication for the child-robot interactions of the ALIZ-E project. For this reason, a significant amount of research and development has been dedicated to investigate and develop specific Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems for the Italian language, which is used with children involved in the experimental part of the project.

The Padova Institute of Cognitive Sciences and Technologies (ISTC) of the National Research Council (CNR) is the partner in charge of carrying out these studies within the ALIZ-E project. Through its contribution to the ALIZ-E project, advances in the research field of speech technology (with the focus on child-robot interaction) have been accomplished by:

1. Studying voice adaptation techniques for ASR (see Section 2.7);
2. Experimenting with ASR in a real case child-robot interaction scenario (see Section 2.8);
3. Collecting and making available three new Italian child speech annotated corpora, made up by read sentences and spontaneous utterances, along with recordings from a listen and repeat experiment (see Section 2.2);
4. Providing TTS technology for the ALIZ-E integrated system (see Section 3.4);
5. Experimenting with TTS technologies suitable for children-robot interaction (see Section 3.5);
6. Investigating expressive TTS techniques (see Section 3.5).

2 Automatic Speech Recognition

This section describes the ASR system for Italian children's voices that has been built for the ALIZ-E project. While the task of recognizing the children's language is challenging, it has been largely studied by the speech research community. Unfortunately, the scientific literature related to the Italian language has only begun during the last decade. Moreover, available children's speech corpora for Italian are not plentiful.

ISTC-CNR researchers committed themselves to collect audio data (and their transcriptions) of young speakers. Section 2.1 gives an overview of the literature of ASR regarding children's voices, with specific attention to the Italian language. The collection of Italian children's speech corpora within the ALIZ-E project is described in Section 2.2.

Since the ALIZ-E integrated system handles several components in real-time, a high-speed, small memory footprint ASR decoder has been chosen for this task. Details about the software used to develop the ASR system are reported in 2.3, and the integration issues in 2.4. Sections 2.5 and 2.6 respectively describe the acoustic and the linguistic models that have been built for the system. One of the key strategies for coping with the high variability among children's voices in order to obtain better results in ASR experiments is the use of voice adaptation techniques. In 2.7, adaptation experiments with children's voices are explained. Finally, Section 2.8 gives the ASR results that have been achieved by means of the ALIZ-E children's voice data.

2.1 Children's Speech Recognition

ASR and acoustic analysis of children's voices have been studied extensively by speech technology researchers. Although most of the literature focuses on native English speakers, in recent years the Italian language has been studied as well. As a result, speech corpora in the Italian language have been collected and made available to the scientific community.

Lee et al. (1999) have investigated spectral acoustic parameters of children's speech as a function of age and gender and compared them to those of adults. This study has shown that some parameters converge to adult levels around age 12, while most of the acoustic speech characteristics becomes fully established around age 15.

Another important work is that of Potamianos and Narayanan (2003) who have been one of the first to apply algorithms related to automatic recognition of children's speech. What they have shown is that in addition to anatomical and morphological differences in the vocal-tract geometry with respect to adults, children have proven to introduce more disfluencies than adults, not only in spontaneous speech but also in read speech. These disfluencies are due to a not yet mature control of articulators and suprasegmental aspects of speech such as tone, stress, and prosody. The authors studied the age-dependent variability of speech characteristics, which were due to developmental changes. They found that such factors worsen ASR results when applied to children's voices. An analysis of age-dependent scaling in formant frequencies, in particular the first two, F1 and F2, showed that they change almost linearly with increasing age. Also, it proved that for children it is more difficult to classify patterns based on spectral feature as there is a high dynamic range for acoustic parameters values. The study ends with speech recognition tests with adult acoustic models on children's voice inputs. Extensive experiments, conducted on Connected Digit and Command Phrase recognition tasks, showed that results become worse as age decreases. On average, ASR experiments with adult Acoustic Model (AM)

tested on children lead to a WER that is 2-5 times worse than recognition of adult AM tested on adults. Finally, the authors applied Speaker Normalization and Frequency Warping, showing an improvement on ASR performance.

Also Gerosa et al. (2007) have studied children's read speech, both for Italian and English languages, with the purpose of analyzing acoustic characteristics related to ASR. In particular, they focused on the analysis of phone duration, intra-speaker variability and acoustic space. This work also described experiments carried out on speech recognition of children's voices in matched (i.e., training and testing on voices of Italian children aged 7-13) and unmatched conditions (i.e., testing on children's speech with models trained on adult speech). It is important to note, for the benefit of the goals of the ALIZ-E project, that these tests showed that an acoustic model trained on children's recordings lead to better ASR results for children's inputs than a model trained on adult voices.

Gerosa et al. (2007) have also investigated and analyzed the correspondence among vocal tract morphology, speech acoustics and formant patterns in children and in adults. Their work confirmed literature results also about the Italian language: formant frequency values of recordings of the "ChildIt" corpus and "APASCI" corpus, the former comprising children's voices, the latter consisting of adult speech.

Their most important conclusion regarding the goals of the ALIZ-E project is that Gerosa et al. (2007) confirmed that children of 7-13 years are not a homogeneous group of speakers. Specifically, at about 12 male children's fundamental frequency and their first three formants abruptly change. In order to cope with the variability of spectral parameters among different age groups, VTLN experiments have been tested and proven to be effectively useful to reduce errors in ASR.

2.2 Data Collection

This section describes the corpora of children's speech as regards the Italian language that have been built within the ALIZ-E project.

Read Speech Data collection of read speech is useful to increase audio plus transcription children's data. These data are meant to be used to train the Acoustic Model.

The major advantage of collecting read speech is the straightforward acquisition of the transcriptions corresponding to the audio. Thus, it is a relatively short time consuming task (compared to that of transcribing spontaneous recordings).

With regard to the text of the recordings, it has been decided to use the FBK ChildIt's prompts which are phonetically balanced sentences, selected from children's literature. During each session the input coming from the four Nao microphones, a close-talk microphone and a panoramic one[3] has been recorded. The close-talk[4] and the panoramic microphones were connected to a digital audio

[3] AKG Perception 200, -10 dB, flat equalization.

[4] for the Limena session, a Shure WH20QTR Dynamic Headset; for the other sessions, the same Proel radio microphone used in 1.1.

Recording date	Location	Number of children	Age
July 2011	Summer school at Limena	31	6-10 years old
August 2012	Summer school for children with diabetes at Misano Adriatico	5	9-14 years old
March-April 2013	Istituto Comprensivo "Gianni Rodari", Rossano Veneto	52	11-14 years old
August 2013	Summer school for children with diabetes at Misano Adriatico	8	11-13 years old

Table 1: The four recording sessions of read speech data collection.

USB interface[5]. Children read text prompts from an LCD monitor. Synchronisation of the sources has been accomplished by using a chirp-like sound, played by an external loudspeaker, at the beginning of every utterance.

The four main sessions that have been recorded can be seen in Table 1. 96 Italian young speakers have been recorded, for a total amount of 4875 utterances, resulting in more than 10 hours of children's speech. All data will be made available to the research community.

Spontaneous Speech Read Speech is very useful for increasing the size of speech training data to be used in the AM training procedure, but it is not well suited for building a reliable test set for ASR in the ALIZ-E project. A proper test set should consist of audio collected in a scenario as close as possible to the real one.

For this reason it has been decided to collect and manually transcribe and annotate speech data recorded during the Quiz Game experiments that took place at "Ospedale San Raffaele" and at Misano Adriatico Summer Schools. Moreover, it allows us to collect and classify those non-verbal sounds occurring specifically in the child-robot Quiz Game interaction, such as Nao's speech and motor noise.

The collected audio data consist of spontaneous speech recordings of children's utterances produced during real interactions with Nao in a non-autonomous modality (i.e., recognition and understanding of user input are not automatic, but performed by a human operator). An experimenter welcomes the child, introduces him/her to the interaction and takes care of submitting questionnaires at the end of the game; another experimenter, the "Wizard of Oz", remotely controls the system, by entering user input data.

The experiments consist of the robot posing questions to the child. Then, after more or less four answers, they exchange roles and the child reads the questions (and the answer options) to the robot. In the latter case, the child speech cannot be considered entirely "spontaneous". However, since it is part of the real system interaction, those data can be considered a reliable test set.

[5] Zoom H4n Handy Recorder.

The procedure and the setup of the Quiz Game interaction has been described in more detail in Section 1.1.

The partners of the ALIZ-E project have agreed to manually transcribe speech recordings of the above mentioned interactions, and, in particular, to annotate a specific category called "domain objects", which are speech events relevant to Natural Language Understanding (NLU) components. Domain objects are sentences that carry a special meaning, important to specific ALIZ-E scenarios. For the Quiz Game, two main labels have been used to mark the domain objects: (a) a *Question*, used when the child poses a Quiz question (e.g. "Chi allattò Romolo e Remo?") and (b) an *Answer*, used to mark a Quiz answer option (e.g. "Una lupa", "La loro mamma", "La prima"). A domain object can be tagged as "inappropriate" (for example when the provided answer was not on the answers' list) and/or "incomplete" (if the user did not utter the sentence completely).

The tool used for annotating the audio files is Transcriber[6]. Transcription has been carried out at different levels: speech, interruptions, domain objects, noise and fillers.

Globally, the 76 interactions have been annotated, totalling about 20 hours of total audio, containing more or less 3.6 hours of actual children's speech.

All the annotated data (audio and transcriptions) from ALIZ-E Quiz Game experiments will be available to the research community.

Listen and Repeat Experiment Another data collection session has been set up at "Istituto Comprensivo Berna" (Mestre, Italy). Children aged 7-11 were asked to listen to about 40 sentences generated by the Italian ALIZ-E TTS system (see Section 3) and to repeat them aloud. Children listened to the sentences in two steps: (1) half of the sentences were uttered by the Nao robot and the young users were recorded by means of a close-talk radio microphone and the four Nao microphones; (2) half of the sentences were played through headphones worn by children who were recorded by menas of a close-talk wired microphone.

The prompts were randomly generated by an automatic program that provides grammatically correct but semantically anomalous sentences. A lexicon with the most common words used by Italian children has been applied. Words with CV (Consonant-Vowel) patterns have been favoured, and only a few very common words containing CC (Consonant-Consonant) patterns, like "scuola", have been allowed. These kind of sentences has been used because this recording session is part of a wider experiment that involves testing the intelligibility of the Italian ALIZ-E TTS system.

While the recording session was running, the annotations of the recorded audio sentences were automatically generated, assuming that the child repeated exactly what the TTS system has uttered. However, if the child mispronunced some words, right after the end of the utterance the experimenter could modify the transcription accordingly (or tag the sentence to be corrected afterwards).

Recordings were performed in April 2014. 95 children contributed so that we were able to collect almost 5 hours of children's speech.

[6] Transcriber: http://trans.sourceforge.net/en/presentation.php

2.3 Julius

Julius[7] has been chosen as the ASR decoder for the ALIZ-E project, mainly because it is designed for real-time decoding and modularity (Lee et al. 2001). Its well-designed decoder API made it very easy to implement and incorporate speech recognition into the ALIZ-E integrated system. The core engine is a C library with very low system requirements. High-speed decoding is achieved by means of a small memory footprint. It is also possible to swap language models at run-time. Finally, Julius supports several AM normalisation algorithms: Cepstral Mean Normalisation (CMN), Cepstral Variance Normalisation (CVN) and Vocal Tract Length Normalisation (VTLN). For these reasons, Julius is particularly suited for the ALIZ-E integrated system, which needs to handle several components in real-time. Its configuration is modular (i.e., each configuration file can embed another one covering only one specific aspect of the configuration).

Julius also integrates a GMM-based and Energy-based Voice Activity Detector (VAD). It is very useful since it helps to avoid unnecessary coding and transmission from the ASR component. It also prevents the ASR system from spotting false positive results (such as words that may come from speech recognition of noise events).

2.4 ASR Component

In the ALIZ-E integrated system, ASR is provided as a `urbiscript` component API, whose functions can be accessed by other components (e.g., the DM module). When an ASR output result is available, an event is launched and the result is provided as a payload, so that every component needing this information can access it.

The ASR component for the ALIZ-E project is basically made up of two modules. The first is a configuration structure (that holds data for AM and LM) and a main recognition loop function (also called "Julius stream"). The second contains an internal VAD and outputs the recognition result. The principal methods of this component are: load/free/switch configuration and start/stop main recognition loop.

A diagram describing the function call and the data exchange among spoken interaction components in the ALIZ-E system can be seen in Figure 2. In the Figure, Dialogue Manager (DM) and Natural Language Understanding (NLU) are two components that are connected to the ASR module. The former can specify an ASR configuration and decides when to start/stop the recognition loop. The latter takes as input the words which are recognized and it is responsible for interpreting them.

Julius can express its output as nbest lists or word lattices. The former is the list of n most likely sentences recognized by the ASR system. The latter, also called Word Graph in Julius terminology, is an acyclic ordered graph in

[7] Open-Source Large Vocabulary Continuous Speech Recognition Engine Julius: http://julius.sourceforge.jp/en_index.php

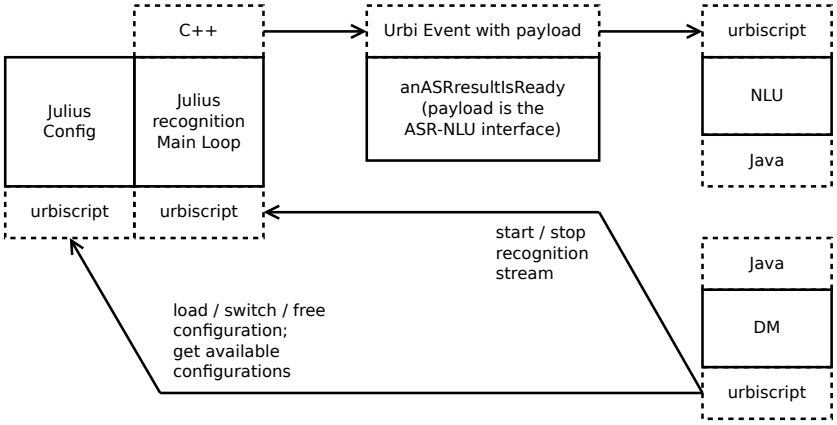

Fig. 2: *ASR communications through* `urbiscript`.

which nodes are words and edges are transition probabilities, weighted by the acoustic and language model probabilities. A Word Graph is a more powerful representation than nbest lists for Spoken Language Understanding (SLU) since lattices provide a larger set of hypotheses and a more accurate representation of the hypothesis space. Both the nbest list and the lattice structures have been implemented in the Julius C++/Urbi module, exposing most of the original Julius data structure to `urbiscript`.

Julius can perform decoding by means of parallel, multiple language and acoustic models specified in one single configuration instance called "search instance" in Julius terminology. This means that from the same audio input, a different result is given for every acoustic and/or language model. This feature can help in improving the accuracy of the ASR module, by keeping the result from the search instance that has the greatest acoustic likelihood. Moreover, thanks to this feature it will also be possible to create models for input rejection so that unwanted speech events can be detected and discharged as needed. The ASR component for the ALIZ-E project is capable of switching search models at run time.

2.5 Acoustic Model

Since the Julius distribution does not include specific tools for Acoustic Model (AM) training, the HTK tools (Young et al. 2006) have been used for this task. Also a procedure to build multi-gaussian Acoustic Model (AM) has been implemented. More details on the AM training procedure can be found in the work of Paci et al. (2013).

The AM for the work of Paci et al. (2013) was built only with data taken from the Italian FBK ChildIt Corpus. This corpus consists of Italian children's voices, amounting to almost 10 hours of speech from 171 different children. Audio

prompts consist of adults interviewing children about their favourite books, TV shows, hobbies, sports, etc. Audio recordings were performed by means of a Shure SM10A head-worn microphone at 48 kHz and down-sampled at 16 kHz, 16 bit linear.

The Acoustic Model for Italian children's ASR experiments described in this work has been created by utilizing the FBK ChildIt Corpus and also data from the "read speech" recording sessions collected within the Aliz-e project, as described in Section 2.2. This corpus counts more than 10 hours of children's speech from a total of 96 young speakers, aged 6-14. While the utterances of this collection comprise read speech, they contain also non verbal sounds (such as laughing, breathing, coughing, etc.). The ASR tool Julius has been configured to perform a specific Forced Alignment procedure designed to spot optional non verbal sounds between words. The assumption is that these fillers are not present in the orthographic transcription reference, but may occur in the actual utterance. In order to achieve this task, two different approaches have been implemented: (1) at phoneme level, for each filler, the creation of an alternative pronunciation for every word by appending the filler phoneme at the end of the phoneme sequence; (2) at word level, the optional inclusion of fillers among words: that means using a specific grammar allowing an optional transition to a "filler-word" before processing the next word.

Once this FA task has been run on all files, the non verbal sounds have been incorporated accordingly in the text reference of the corpus and used as additional information for the AM routine.

2.6 Language Model

Julius supports N-gram, grammar and isolated word Language Models (LM), although its distribution does not include any tool to create them. An external program must be used to create an LM.

The SRILM toolkit (Stolcke 2002) has been used to train a 4-gram model for question recognition of the Quiz Game Aliz-e scenario. The Quiz questions and answers database has been used as training material for a "question recognition" model. Also a simple grammar model for Quiz answers has been built, automatically including the answers in the Quiz database and then adding rules to handle common answers and filler words. Details on how language models have been trained can be found in the work of Paci et al. (2013).

2.7 Adaptation

As discussed in 2.1, voices of children aged 7-13 form an extremely heterogeneous set of spectral features. This means that the standard HMM approach to ASR leads to poor results. The literature in this field suggest the use of adaptation techniques to cope with this problem.

Through the VTLN technique (Young et al. 2006, Zhan & Waibel 1997), it is possible to divide the group of young speakers into more "homogeneous" groups. Also, since each child will interact several times with the robot, data

from previous interactions can be used to adapt the models. For these reasons, the VTLN technique has been proven to be extremely suitable for the goals set through the ALIZ-E project.

VTLN experiments have been carried out by Paci et al. (2013). The procedure for applying VTLN to ASR has been the following: N recognitions, with N different configurations for VTLN parameters have been run simultaneously by using an AM trained by means of the ChildIt corpus. The configuration with the best confidence score is chosen.

The number of configurations N and the VTLN-specific parameters have been tuned by means of a grid search experiment carried on children's speech training data. All training sentences (more than 10000 audio files) with VTLN-alpha values spanning from 0.7 to 1.3 (with a 0.1 step) have been processed. For each training file the alpha value that maximizes the acoustic score has been selected. Results showed that most optimal alpha values fell around the value of 1 and that alpha values of 0.7 (and below) and 1.3 (and above) were very unlikely. Moreover, very little fluctuation of alpha values is observed in each speaker's recordings. Finally, a further experiment with a 0.05 step for alpha values proved no significant acoustic score increase. For these reasons, a *N=5* parallel recognition (with alpha values spanning from 0.8 - 1.2, 0.1 step) has been chosen as the optimal configuration for the ChildIt corpus.

2.8 ASR Results

Preliminary results of the ALIZ-E ASR system for Italian children's voices have been discussed in the work of Paci et al. (2013). Those results were obtained with an AM smaller than the one presented here (see Section 2.5 for details). This section describes ASR experiment results that take into account almost all ALIZ-E children's speech corpora (the data from the listen and repeat recording session have not been incorporated).

Also, the test set was much shorter: while the one used by Paci et al. (2013) included 64 sentences and 559 words, the one described here is almost ten times larger. Data from spontaneous speech data collection (see 2.2) have been used and a test set including 540 sentences from 46 speakers, totalling 5423 words, has been created.

ID	#Snt	#Wrd	WCR%	Sub%	Del%	Ins%	WER%
C	540	5423	62.9	33.5	3.6	20	57.2
C+R	540	5423	63.2	32.7	4.1	16	52.8

Table 2: ASR Results obtained with AM trained only with FBK ChildIt corpus ("C" ID) compared with those obtained with AM trained with both ChildIt corpus and read speech data ("C+R" ID).

Table 2 shows the results of ASR applied to real-case scenario audio files, where a child poses Quiz Game questions to the NAO robot. The first line of the table is referring to the ASR results using the AM trained only with FBK ChildIt, while the second one shows the results from the AM described in Section 2.5. No speaker adaptation procedure has been applied. The 4-gram language model described in Section 2.6 has been used.

It is worth noting that the "C+R" AM provides little improvement (+0.3%) in terms of Word Correct Rate (WCR), while giving a sensible improvement (+4.4%) in terms of Word Error Rate (WER). This is a due to the fact that the system trained with ChildIt and read speech data gives better results in term of inserted words. It is likely that the "C+R" system is more robust against noise; this can be considered a consequence of the FA with the filler sounds spotting procedure described in Section 2.5.

3 Text To Speech Synthesis

The term "Speech synthesis" means the artificial production of human speech. A Text-To-Speech (TTS) system is able to artificially produce speech starting from a textual input. The use of text input allows the creation of intelligent systems able to speak automatically when the appropriate text is provided to the TTS. It is manifest how the TTS technology is important for the purpose of building talking robots.

In the ALIZ-E project, a robot must be able to speak to the child in order to communicate information about the various activities or to ask questions. The voice is one of the human-preferred means of communication allowing the transmission of verbal messages; moreover speech communication allows one to relay or carry much more information.

In fact, it is known that several messages are contained in the speech signal as it is summarized in Table 3. The table shows the main speech messages and their main speech correlates.

Message	**Acoustic correlates**
verbal content of the speech	spectral envelope
speaker's identity	spectral envelope, voice quality
emotional state of the speaker	voice quality, prosody
focus	prosody

Table 3: Messages contained in the speech signal and their main acoustic correlates.

For example, it is well known that the verbal content of a speech signal is mainly included in its spectral envelope, whilst the identity of the speaker is mainly linked to the spectral envelope and the voice quality and to a lesser

extent to its prosody (Lavner et al. 2000). With regard to emotional speech, many scientist (Scherer 2003) have identified clear correlates between emotional categories and acoustic features such as intonation, loudness, rhythm, and voice quality. Finally, many researchers have confirmed that the focus[8] has a prosodic reflection in various languages, affecting prosodic phrasing, prominence and/or intonation (Frota 2002).

Within the ALIZ-E project, the robot has to convey these messages, encoded by means of particular speech patterns to the child. For this reason, a speech synthesizer has necessarily been developed in order to generate these speech patterns for communicating those different messages to the child. Moreover, the Nao robot can also integrate the audio/speech channel with other media: movements, gestures and blinking lights.

MaryTTS (Modular Architecture for Research on speech sYnthesis (Schröder & Trouvain 2003)) satisfies this requisite and it has been

chosen for the ALIZ-E project. Another good reason to use MaryTTS is that it is released as an open-source project[9].

A TTS system is usually made up of two components, a front-end and a back-end. The front-end takes care of performing NLP (Natural Language Processing) tasks. Its three main purposes are the following:

1. to normalize the input text, an operation also called "tokenization", (which means, for example, converting numbers and abbreviations into their written-out words equivalent);
2. to perform text-to-phoneme (or grapheme-to-phoneme) conversion (which means the process of assigning phonetic transcriptions to words); and
3. to divide and mark the input text into prosodic units, such as phrases, clauses, and sentences.

The output of the front-end is the so-called "symbolic linguistic representation", that consists of phonetic transcriptions and prosody information. The NLP modules developed for Italian MaryTTS are described in Section 3.1.

The back-end – often referred to as the synthesizer or the vocoder – is designed to convert the symbolic linguistic representation into sound.

The HMM Speech Synthesis approach, a declination of Statistical Parametric Synthesis (Zen et al. 2009), has been chosen for the task of modelling the voice of the robot, as it allows a more extensive modification of the produced acoustic patterns and to provide greater flexibility than other TTS approaches, such as the Unit Selection technology (Black & Campbell 1995). In HMM systems, the back-end module needs also to compute the target prosody (pitch contour and phoneme durations), which is then imposed on the output speech.

In HMM-based speech synthesis systems, a symbolic representation of the speech segments, together with their phonetic and prosodic context, is extracted from the input text. Such representation is defined by the so-called "full context

[8] Here the term *focus* refers to the part of a sentence which expresses the centre of attention.

[9] https://github.com/marytts/marytts

labels". In order to generate the speech signal, a machine learning algorithm uses these labels to generate the appropriate control parameters (usually excitation and spectral parameters) and then employ them as input for a vocoder. Figure 3 shows a functional diagram of an HMM-speech synthesizer.

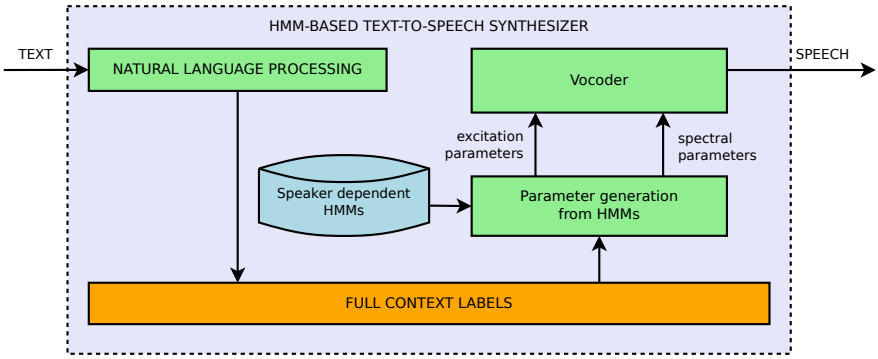

Fig. 3: Functional diagram of a HMM-based TTS system.

3.1 Italian MaryTTS NLP Modules

MaryTTS allows one to easily add support for new languages. As a matter of fact, when MaryTTS was born, it was originally developed for the German language; nowadays it provides voices and support for the following languages: American English, British English, German, Turkish, Russian, Italian, French and Telugu.

ISTC-CNR researchers have developed the Italian NLP modules (Tesser et al. 2013a), and added Italian to the list of the languages supported by MaryTTS.

MaryTTS supports the creation of HMM voices for new languages by using the *Multilingual Voice Creation* tool (Pammi et al. 2005).

The standard procedure, based on the use of the freely available Wikipedia dump of the new language, allows one to create basic language-specific NLP modules, such as the LTS (Letter To Sound) rules for out-of-vocabulary words and a minimal POS (Part Of Speech) tagger. For the Italian language, more sophisticated NLP modules have been developed. Hereafter, brief descriptions of each of them are listed.

Lexicon and LTS Rules Pronunciation of Italian words can be obtained from the pronunciation lexicon or from the Letter To Sound (LTS) rules module. The Italian lexicon for MaryTTS has been adapted from an existing one and improved. Specifically, an automatic algorithm capable of generating inflected forms of verbs with clitics has been implemented and used to build a lexicon with transcriptions, generating about 2.6 millions words.

Finally, the Letter To Sound rules have been inferred by using an automatic MaryTTS procedure from lexicon examples.

Numbers Expansion Number expansion and pronunciation has been implemented for cardinal and ordinal numbers, allowing the pronunciation of numbers written in digit form. Expansion is a prerequisite for several expansion modules such as percentages, charts, currencies, and dates. Cardinal numbers expansion has been completed. It permits the reading of huge numbers, with no other limit but the size of the maximum long integer. Floating point numbers are also expanded.

Part of Speech Tagger A context dependent Part Of Speech (POS) tagger has been developed to predict whether words are nouns, verbs, or other grammatical categories. A manually annotated corpus (Attardi et al. 2008, Zanchetta & Baroni 2005), containing 4000 sentences for a total of 113k words, with 36 POS categories, has been used to train an Italian OpenNLP POS tagger[10].

Homograph Pronunciation Disambiguation Homograph words with different pronunciations are common in the Italian language[11].

Luckily, most of these pronunciation ambiguities occur between two words with different parts of speech. Thus the correct pronunciation can be found by identifying the part of speech for each word. A new lexicon look-up method that make use of POS tags has been implemented.

ToBI Rules The ToBI standard (Silverman et al. 1992) defines a symbolic representation of the prosody of a sentence. *Break indices* are used to describe the degree of disjuncture between consecutive words and the tones associated with *phrase boundaries* and *pitch accents*. Principal rules to predict ToBI labels from punctuation and POS (part Of Speech) of words have been implemented for the Italian language.

3.2 Italian Corpus Based HMM Voice

The HMM Speech Synthesis approach belongs to the class of the so called "corpus-based" TTS systems. This means that a database of speech audio data and their transcriptions must be designed and recorded.

In order to get high quality synthetic voices, the task of sentence selection is crucial, because the set of sentences (also called the "script corpus") must be phonetically and prosodically balanced. Given a large portion of text, MaryTTS provides a method for optimal text selection capable of ensuring good phonetic and prosodic coverage. For the Italian language, phonetic and prosodic information has been extracted from the entire Italian wikipedia dump by using the Lexicon modules and the Symbolic Prosody predictor described in Section 3.1.

1400 sentences have been selected and uttered by the 20 years old Italian native female speaker *Lucia* and recorded in a quasi soundproof chamber, by

[10] OpenNLP library: http://opennlp.apache.org/

[11] For example, the word *ancora*, can be used as a noun (English translation: *anchor*) or as an adverb (English translation: *again*). These two Italian words have different pronunciations.

means of a Shure WH20QTR Dynamic Headset. The finale corpus totals more than 2 hours of speech.

Finally, the resulting speech corpus is provided to the training TTS procedure, that consists of the HMM voice models estimation.

The context-dependent speech units used in this work symbolize phonetic context (triphone/quinphone models) but also prosodic and linguistic contexts such as stress, syllable accent, boundary tones, part of speech, and sentence information. The context depended HMM models have been trained on speech audio data whith full-context labels. The latter are automatically generated by the MaryTTS *Voice Import Tools* from the text sentences. Phonetic Forced Alignment (FA) has been done by using HTK 3.4.1 (Young et al. 2006) and the HTS HMM speech synthesis toolkit version 2.2 (Zen et al. 2007) has been employed to estimate the models.

3.3 *Signal-driven* TTS Training

The main concept under the "*Signal-driven* TTS training" refers to the fact that the speaker sometimes completes the utterance in a different way (in term of prosody or pronunciation) from what is estimated by the TTS front-end. A method to obtain more coherent data for the training of the TTS system is to:

- automatically detect these differences between the speaker and the TTS front-end,
- impose the detected speaker's prosody and pronunciation in the training data.

This concept has lead to the development of three techniques:

- *signal-driven* symbolic prosody;
- break pauses and punctuation correction; and
- multiple pronunciation disambiguation.

Signal-driven Symbolic Prosody Within the ALIZ-E project, experiments on symbolic prosody (Tesser et al. 2013b) have been conducted. The *Signal-driven* term indicates that the symbolic prosody is evaluated from the actual speech signal, as opposed to the *text-driven* symbolic prosody.

Modern TTS systems usually evaluate symbolic prosody parameters solely from text (*text-driven* symbolic prosody), by using both handwritten rules or statistical methods. While most of the parametric speech synthesis features (like phonetic features, syllable features, parts of speech, ...) are determined by linguistic rules applied only to textual information, the symbolic prosody is also strongly related to the way in which a speaker has acoustically uttered a particular sentence.

This preliminary work aims at advancing the state of the art in the field of prosody prediction for corpus-based Text-To-Speech synthesis systems, by taking advantage of acoustic information to improve naturalness of synthetic voices through a better prosody prediction, thus generating parameters that

model the symbolic prosody of a specific speaker, or his/her particular speaking style used in the corpus.

Different TTS systems have been trained by using the *Signal-driven* prosody prediction and have been compared to the baseline MaryTTS system based on *text-driven* symbolic prosody. Experiments have been carried out by using data from a male and a female speaker. A test set has been extracted from each corpora, and its content has not been used in the voice training process.

For each text sentence in the test set, the symbolic prosody parameters have been calculated by the *Signal-driven* and the baseline system respectively. The parameters trajectories extracted from natural speech have been compared to the generated speech trajectories. Subsequently an objective evaluation has been performed on these measures which has shown how the proposed systems improve the prediction accuracy of phonemes duration and F0 trajectories.

Break Pauses and Punctuation Correction The forced alignment phase detects when the speaker has made a pause in every utterance. This technique is capable of checking whether the speaker:

- has carried out a speech pause that is not consistent with the punctuation in the text;
- has not paused when the punctuation marks demanded it.

The text is automatically modified by creating a version of the text which contains a punctuation more consistent with the prosody (the break pauses) that the speaker has actually carried out during the recordings. Table 4 shows the improvements (objective tests) with regard to the training of Italian female and male voices. The proposed method increases the correlation between the predicted and the original pitch and decreases the prediction error.

Gender	**Type**	**Correlation**	**RMSE** (Hz)
female	original	0.645	32
female	proposed	0.666	31
male	original	0.537	28
male	proposed	0.545	21

Table 4: Objective results (10% of the entire corpus) of pitch prediction.

Multiple Pronunciation Disambiguation The issue of homograph words with different pronunciations affects also the training procedure of a TTS system: a speaker could have uttered a word with phonemes that are different from those expected by the TTS training system. Ignoring this difference leads to a bad estimation of the models. Being able to automatically recognize which phonemes have been actually uttered by the speaker permits the building of a system which estimates HMM models with more appropriate phonetic labels.

Such a forced alignment procedure, one that takes into account multiple pronunciations for a single word, can be used to select the most accurate pronunciation according to the speaker's utterance.

3.4 Integration in the Robotic Environment

A TTS component has been developed for the integrated system of the ALIZ-E project. This component is a client for the MaryTTS server.

In order to make Nao a believable robot, it is crucial for the robot to emit the speech signal from its loudspeakers located in its head. However, this is not a trivial achievement. In fact, due to the Nao computational limitations, it has been decided to run the MaryTTS server on a remote machine connected to the robot through a Wi-Fi network.

MaryTTS has been integrated in the `urbiscript`-based robotic system in a straightforward manner by means of an URBI-Aldebaran UObject, called `robot.proxy.ALAudioPlayer.playWebStream`, capable of directly calling the MaryTTS server and playing the speech signal using a specific http request. Moreover the low-latency of this process has been granted by the MaryTTS stream modality, which permits one to stream the synthetic speech signal as soon as the first audio data are produced by the remote TTS server.

3.5 Flexible TTS for Aliz-e

As explained at the beginning of this section, the HMM Speech Synthesis approach has been chosen because it allows one to widely manipulate the acoustic parameters. For example, by using the HMM speech synthesis technology it is possible to:

- change the speaker's identity of the synthetic voice by tuning vocoder parameters (Imai 1983, Fukada et al. 1992) or, alternatively, by using speaker adaptation techniques (Yamagishi et al. 2009),
- stress the focus of a sentence by applying specific prosodic patterns,
- change the emotional content of the synthetic speech by applying different prosodic settings and patterns.

The prosodic settings and patterns mentioned beforehand are drawn either from previously acquired knowledge and experience (e.g., it is known that happy voices usually adopt a higher pitch with respect to the standard voice) or from modules capable of learning these from real data.

To make child-robot interactions more realistic, interesting and expressive, TTS parameters manipulations have been subject to rigorous experimentation and three relevant features have been introduced: (a) the robot voice timbre has been modified to make it sound more "childish", (b) prosodic modifications have been implemented according to the focus of the sentence, and (c) a robot can express a particular emotional state through speech. These three features are explained in more details in the following paragraphs.

Robot Voice Identity A vocal tract scaler, which can simulate a longer or shorter vocal tract, has been used in order to obtain a child-like voice, starting from a female adult voice. In this implementation, the frequency axis warping method has been used.

Focus Prominence by Prosodic Modification The Natural Language Generation component of the ALIZ-E integrated system is the module responsible for generating the text sentences that will be uttered by the robot. This component can mark the words to be stressed during the verbal output generation process. Subsequently, the speech synthesis process takes care of emphasizing the focus by using adequate speech parameters. Moreover, prosody modifications have been realized by forcing the corresponding prosody changes in the words that bear the focus.

After some informal listening test, the prosody on the focus words is implemented in the following way:

- the speech rate is decreased by 10% with respect to the speech normal production;
- the pitch is raised by 25% with respect to the normal speech production.

An example of the SSML[12] control parameters which have been used is shown in Listing 1.1.

```
<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/
   synthesis" xml:lang="it">
  <p>
    <prosody rate="+0%" pitch="+0%">
Ora la <prosody rate="-10%" pitch="+25%"> seconda </
   prosody> domanda. Come si chiama il Ministero che
   cura i servizi automobilistici e ferroviari?
    </prosody>
  </p>
</speak>
```

Listing 1.1: *Prosodic prominence SSML implementation of the input text: "Ora la seconda domanda. Come si chiama il Ministero che cura i servizi automobilistici e ferroviari?"*

Emotional Prosodic Modifications The ALIZ-E system is capable of deciding when the verbal output should be rendered with (non-neutral) emotional coloring, either "sad" or "happy".

According to that, the speech paralinguistic feedback is carried out by increasing the speech rate (+5%) and the pitch contour (+25%) in the "happy"

[12] http://www.w3.org/TR/speech-synthesis11/

case, while in the "sad" case, the speech rate and pitch contour are both decreased (-20%). Listings 1.2 and 1.3 show some realizations using the SSML prosody control command.

```
<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/
   synthesis" xml:lang="it">
  <p>
    <prosody rate="+5%" pitch="+25%">
Sei proprio forte.
    </prosody>
  </p>
</speak>
```

Listing 1.2: *Emotional (Happy) SSML implementation of the input text: "Sei proprio forte."*

```
<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/
   synthesis" xml:lang="it">
  <p>
    <prosody rate="-20%" pitch="-20%">
No, non è corretto.
    </prosody>
  </p>
</speak>
```

Listing 1.3: *Emotional (Sad) SSML implementation of the input text: "No, non è corretto."*

4 Conclusion

Investigation in speech technologies is an essential part in the ALIZ-E project, since verbal interaction plays a central role in child-robot interactions. Voice controlled robots supporting hospitalized children need to incorporate adequate speech comprehension and production tools, which have to be set up for the Italian language.

A robot is needed to understand the children's language, which itself is a challenging scientific task. Moreover, the scientific literature with regard to the Italian language in this matter is inadequate if compared to that of the English language.

For these reasons, particular attention has been paid to speech recognition adaptation techniques and a huge effort has been made regarding the collection and annotation of children's speech data. Both read sentences (which allows one to easily obtain transcriptions) and spontaneous utterances (taken from real case interactions) have been collected.

In addition, speech data from a listen and repeat experiment have been recorded. These data will be published as speech corpora and they will be freely available to the scientific community.

On the other hand, the robotic verbal output, applied to interactions with hospitalized children, should convey expressivity and emotions in order to involve and engage young users as much as possible. Finally, stress on particular words or phrases is crucial for children to comprehend the most important health-related educational topics of the interactions. Specific Text to Speech modules and tools for the Italian language have been studied and developed.

Acknowledgements

This research was partly funded by EU-FP7 project ALIZ-E (ICT-248116).

Bibliography

Attardi, G., Montemagni, S., Simi, M. & Lenci, A. (2008), Tanl - Text Analytics and Natural Language processing: Analisi di Testi per il Semantic Web e il Question Answering, Technical report.

Baxter, P., Belpaeme, T., Canamero, L., Cosi, P., Demiris, Y. & Enescu, V. (2011), Long-Term Human-Robot Interaction with Young Users, *in* 'IEEE/ACM Human-Robot Interaction 2011 Conference (Robots with Children Workshop)'.

Black, A. & Campbell, N. (1995), Optimising selection of units from speech databases for concatenative synthesis, *in* 'Eurospeech 1995', number September, Madrid, Spain, pp. 581–584.

Frota, S. (2002), The prosody of focus: a case-study with cross-linguistic implications, *in* 'Speech Prosody 2002, International Conference', Aix-en-Provence, France, pp. 315–318.

Fukada, T., Tokuda, K., Kobayashi, T. & Imai, S. (1992), An adaptive algorithm for mel-cepstral analysis of speech, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 92, pp. 137–140.

Gerosa, M., Giuliani, D. & Brugnara, F. (2007), 'Acoustic variability and automatic recognition of children's speech', *Speech Communication* **49**, 847–860.

Henkemans, O. A. B., Hoondert, V., Schrama-Groot, F., Looije, R., Alpay, L. L. & Neerincx, M. A. (2012), "'I just have diabetes": children's need for diabetes self-management support and how a social robot can accommodate their needs', *Patient Intelligence* .

Imai, S. (1983), Cepstral analysis synthesis on the mel frequency scale, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 8, pp. 93–96.

Janssen, J. B., van der Wal, C. C., Neerincx, M. A. & Looije, R. (2011), Motivating children to learn arithmetic with an adaptive robot game, *in* 'ICSR Conference', pp. 153–162.

Kidd, C. D. (2008), Designing for Long-term Human-robot Interaction and Application to Weight Loss, PhD thesis, Cambridge, MA, USA. AAI0819995.

Kruijff-Korbayová, I., Cuayáhuitl, H., Kiefer, B., Schröder, M., Cosi, P., Paci, G., Sommavilla, G., Tesser, F., Sahli, H., Athanasopoulos, G., Wang, W., Enescu, V. & Verhelst, W. (2012), Spoken Language Processing in a Conversational System for Child-Robot Interaction, *in* 'Proceedings of Workshop on Child, Computer and Interaction (WOCCI)'.

Lavner, Y., Gath, I. & Rosenhouse, J. (2000), 'The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels', *Speech Communication* **30**(1), 9–26.

Lee, A., Kawahara, T. & Shikano, K. (2001), Julius - an open source real-time large vocabulary recognition engine, *in* 'Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)', pp. 1691–1694.

Lee, S., Potamianos, A. & Narayanan, S. S. (1999), 'Acoustics of children's speech: Developmental changes of temporal and spectral parameters', *Journal of the Acoustical Society of America* **105**(3), 1455–1468. Selected Research Article.

Paci, G., Sommavilla, G., Tesser, F. & Cosi, P. (2013), Julius ASR for Italian children speech, *in* '9th national congress, AISV (Associazione Italiana di Scienze della Voce)', Venice, Italy.

Pammi, S., Charfuelan, M. & Schröder, M. (2005), Multilingual voice creation toolkit for the MARY TTS platform, *in* 'Proc. Int. Conf. Language Resources and Evaluation', Valleta, Malta.

Potamianos, A. & Narayanan, S. S. (2003), Robust recognition of children's speech., Vol. 11, pp. 603–616.

Scherer, K. R. (2003), 'Vocal communication of emotion: A review of research paradigms', *Speech communication* **40**(1-2), 227–256.

Schröder, M. & Trouvain, J. (2003), 'The German text-to-speech synthesis system MARY: A tool for research, development and teaching', *International Journal of Speech Technology* **6**(4), 365–377.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), ToBI: A standard for labeling English prosody, *in* 'Second International Conference on Spoken Language Processing', Vol. 2, Banff, Canada, pp. 867–870.

Stolcke, A. (2002), SRILM - an extensible language modeling toolkit, *in* 'Proceedings of eventh International Conference on Spoken Language Processing (ICSLP)', ISCA, pp. 901–904.

Tapus, A., Matarić, M. J. & Scassellati, B. (2007), 'The grand challenges in socially assistive robotics', *IEEE Robotics and Automation Magazine* **14**(1), 35–42.

Tartaro, A. & Cassell, J. (2006), Using virtual peer technology as an intervention for children with autism, *in* J. Lazar, ed., 'Universal Usability: Designing Computer Interfaces for Diverse User Populations', John Wiley & Sons, New York, pp. 231–262.

Tesser, F., Paci, G., Sommavilla, G. & Cosi, P. (2013a), A new language and a new voice for MARY-TTS, *in* '9th national congress, AISV (Associazione Italiana di Scienze della Voce)', Venice, Italy.

Tesser, F., Sommavilla, G., Paci, G. & Cosi, P. (2013b), Experiments with Signal-Driven Symbolic Prosody for Statistical Parametric Speech Synthesis, *in* '8th Speech Synthesis Workshop (SSW)', Barcelona, Spain, pp. 183–187.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S. & Renals, S. (2009), 'Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis', *IEEE Transactions on Audio, Speech, and Language Processing* **17**(6), 1208–1230.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. C. (2006), *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK.

Zanchetta, E. & Baroni, M. (2005), Morph-it! A free corpus-based morphological resource for the Italian language, *in* 'Proceedings of Corpus Linguistics 2005', Birmingham.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. & Tokuda, K. (2007), The HMM-based speech synthesis system (HTS) version 2.0, *in* 'The 6th International Workshop on Speech Synthesis', Bonn, Germany, pp. 294–299.

Zen, H., Tokuda, K. & Black, A. W. (2009), 'Statistical parametric speech synthesis', *Speech Communication* **51**(11), 1039–1064.

Zhan, P. & Waibel, A. (1997), Vocal tract length normalization for large vocabulary continuous speech recognition, Technical report, CMU Computer Science Technical Reports.