PIERO COSI, GIULIO PACI, GIACOMO SOMMAVILLA, FABIO TESSER [1]

Building Resources for Verbal Interaction
Production and Comprehension within the ALIZ-E Project

Abstract: The European FP7 project ALIZ-E ("Adaptive Strategies for Sustainable Long-Term Social Inter- action") has recently concluded and its main goal was to study children-robot interactions in the field of healthcare. ISTC CNR, UOS Padova has been the partner of the ALIZ-E project responsible of carrying out studies in the field of speech technologies. One of its main achievements has been the collection of three new Italian children's speech annotated corpora, made up by read sentences, spontaneous utterances and recordings from a listen and repeat experiment.

1 Introduction

The European FP7 project ALIZ-E ("Adaptive Strategies for Sustainable Long-Term Social Interaction") (Belpaeme et al., 2013) has recently concluded. Its main goal was to study children-robot interactions in the field of healthcare. The Padova Institute of Cognitive Sciences and Technologies (ISTC) of the National Research Council (CNR) has been the partner of the ALIZ-E project responsible of carrying out studies in the field of speech technologies, as described in (Tesser et al., 2013) and (Paci et al., 2013).

One of its main achievements has been the collection of three new Italian children's speech annotated corpora, made up by read sentences, spontaneous utterances and recordings from a listen and repeat experiment.

2 Data Collection

Three new Italian children's speech annotated corpora, made up by read sentences, spontaneous utterances and recordings from a listen and repeat experiment were collected during the European FP7 project ALIZ-E.

[1] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche - Unità Organizzativa di Supporto di Padova - Italy
[piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser]@pd.istc.cnr.it.

2.1 Read Speech

With respect to read speech data collection, prompts from the FBK ChildIt corpus (Gerosa et al., 2007) have been used in order to extend the training material of children speech ASR systems. They are simple phonetically balanced sentences, selected from children's literature[2].

As illustrated in Figure 1, during each session the input coming from the four microphones of Nao (a robot used in the ALIZ-E project), a close-talk microphone and a panoramic one has been recorded.

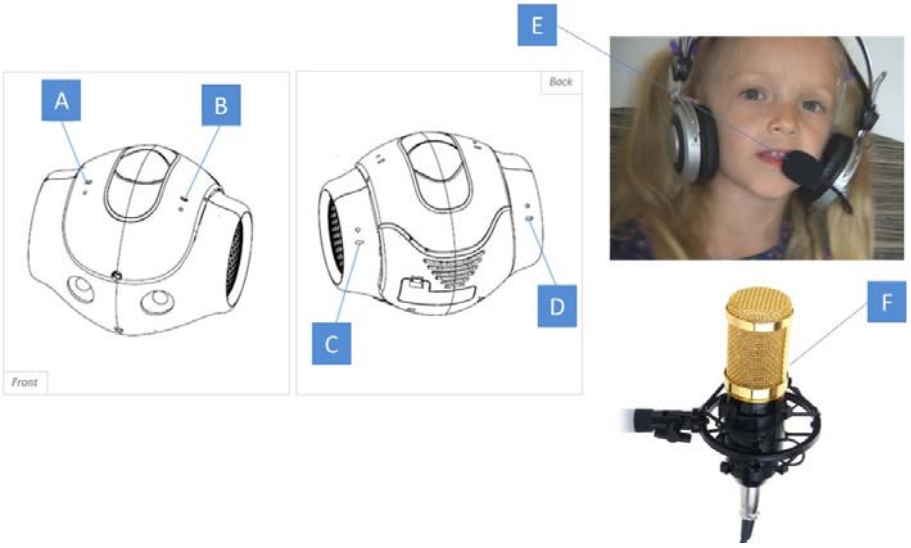

Figure 1 - *Data Collection framework: A,B,C,D - 4 microphones of Nao (the robot used in the ALIZ-E project); E - 1 close-talk microphone; F - 1 panoramic microphone*.

Four main recording sessions have been performed during the ALIZ-E project. In July 2011, 31 children (age 6-10) have been recorded at a Summer school at Limena (PD, Italy); in August 2012, at a Summer school for children with diabetes, recordings from 5 children (age 9-14) have been collected. In 2013 two final sessions have been carried out: the first one (March-April 2013, at Istituto Comprensivo "Gianni Rodari", Rossano Veneto) involved 52 young users aged between 11 years to 14 years; in the second one (August 2013), eight children aged between 11 and 13 years have been recorded at the Summer school for children with diabetes at Misano Adriatico. All recording sessions consist of data from 96 Italian young speakers, for a total amount of

[2] Typical sentences: "il fabbro usa la tenaglia con forza", "nella zona avvenne uno scoppio", ecc..

4875 utterances, resulting in more than eight and a half hours of children's speech.

2.2 Semi-Spontaneous Speech

While read Speech is very useful for increasing the size of speech training data to be used in the AM training procedure, on the other hand it is not well suited for building a reliable test set for ASR, since it does not represent well a real case of recorded audio from an interaction of the ALIZ-E project.

In order to build a proper test set, speech data recorded during a project's experiments in a Wizard of Oz modality have been collected using the same recording framework adopted for read speech (see Figure 1) and manually transcribed and annotated. The interactions took place at the "San Raffaele" hospital and at *Summer Schools* for diabetic children, The experiments consisted in a "Quiz Game", with the robot and the young user posing questions to each other.

Transcription and annotation of semi-spontaneous speech recordings of the above mentioned experiments have been manually performed by means of the "Transcriber" tool. The ALIZ-E project partners agreed to perform a special level of transcription: speech events relevant to Natural Language Understanding (NLU) components (such as questions and answers) have been annotated.

Also non-verbal sounds occurring specifically in such interactions (for example robot's speech and motor noise) have been collected and classified. Globally, the 76 ALIZ-E verbal interactions have been annotated at the word level, totaling about 20 hours of total audio, containing more or less 3.6 hours of actual children's speech. This semi-spontaneous speech annotated corpus has allowed to build a test set for ASR that is almost ten times larger than the one used initially in the ALIZ-E project. It comprises 540 sentences from 46 speakers, totaling 5423 words.

2.3 Listen and Repeat Speech

Another data collection session has been set up at "*Istituto Comprensivo Berna*" (Mestre, Italy) as part of "ITACIE" (Italian TTS Adult Children Intelligibility Experiment).

Children aged 7-11 were asked to listen to about 40 sentences generated by the Italian ALIZ-E TTS system played by either a robot's loudspeakers or by earphones and to repeat them aloud. These recordings served in a TTS in-

telligibility experiment. Usually, in standard intelligibility tests the participants are asked to write down or type what they understand. This kind of tasks is quite tedious and tiring for children. For this reason a listen and repeat experiment has been set up. It has proven to be a convenient and fast technique to collect annotated speech data: it is more reliable than making the children read a text, and it is much faster to transcribe than recording spontaneous speech. The text prompts have been built using Semantically Unpredictable Sentences (SUS) [3] (Benoît et al., 1996). The aim is to prevent the child to be able to "guess" the words in sentences thanks to the semantic context. For this reason, the prompts were randomly generated by an automatic program that provides grammatically correct but semantically anomalous sentences. In order to avoid the problem that the sentences were too difficult to pronounce for the children, a lexicon with the most common words used by Italian children has been applied. Words with CV (Consonant-Vowel) patterns have been favored, and only a few common words containing CC (Consonant-Consonant) patterns, like "scuola", have been allowed. 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 pronounced. If the child mispronounced some words, however, right after the end of the utterance, the experimenter could modify the transcription accordingly (or tag the sentence to be corrected afterwards). Ninety-five children aged 7-11 have been recorded and almost three and a half hours of children's speech have been collected.

3 Final Remarks

ISTC-CNR intends to make all data collected available to the research community. ISTC-ChildIT (read speech) or CHILDIT-2 will be released soon, probably in 2015. ISTC-CNR is currently discussing with ALIZ-E partner "Fondazione San Raffaele del Monte Tabor" about the possibility of publishing the Spontaneous Speech audio data.

Acknowledgements

This work was entirely supported by the EU FP7 "ALIZ-E" project (grant number 248116).

[3] Typical SUS (grammatically correct but semantically anomalous) sentences: "la parola piena beve la scuola", "la pace marrone odia la figlia", ecc..

References

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuay´ahuitl, H., Kiefer, B., et al. (2013). Multimodal Child-Robot Interaction: Building Social Bonds, *Journal of Human-Robot Interacion*, Vol. 1, no. 2, 33-53.

Benoît, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences, *Speech Communication*, Vol. 18, no. 4, 381–392.

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

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

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

Author name, affiliation and email

Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser

Istituto di Scienze e Tecnologie della Cognizione
Consiglio Nazionale delle Ricerche -
Unità Organizzativa di Supporto di Padova
Via Martiri della libertà, 2
35137 Padova

[piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser]@pd.istc.cnr.it