DOI: 10.17469/O2102AISV000016

PIERO COSI, GIULIO PACI, GIACOMO SOMMAVILLA, FABIO TESSER

CHILDIT2 – A New Children Read Speech Corpus

One of the main achievement of the recently concluded European FP7 project ALIZ-E ("Adaptive Strategies for Sustainable Long-Term Social Inter-action") has been the collection of various new Italian children's speech annotated corpora. From some of this speech material the CHILDIT2 corpus has been created and this paper describes in detail its design, building and development.

Key words: children, speech, corpus.

Introduction

The Padova Institute of Cognitive Sciences and Technologies (ISTC) of the National Research Council (CNR) has been the partner of the ALIZ-E ("Adaptive Strategies for Sustainable Long-Term Social Interaction") project (Belpaeme, Baxter, Read, Wood, Cuayahuitl, Kiefer, Racioppa, Kruijff-Korbayová, Athanasopoulos, Enescu, Looije, Neerincx, Demiris, Ros-Espinoza, Beck, Cañamero, Hiolle, Lewis, Baroni, Nalin, Cosi, Paci, Tesser, Sommavilla & Humbert, 2013) responsible of carrying out studies in the field of speech technologies, as described in (Tesser, Paci, Sommavilla & Cosi, 2013) and (Paci, Sommavilla, Tesser & Cosi, 2013).

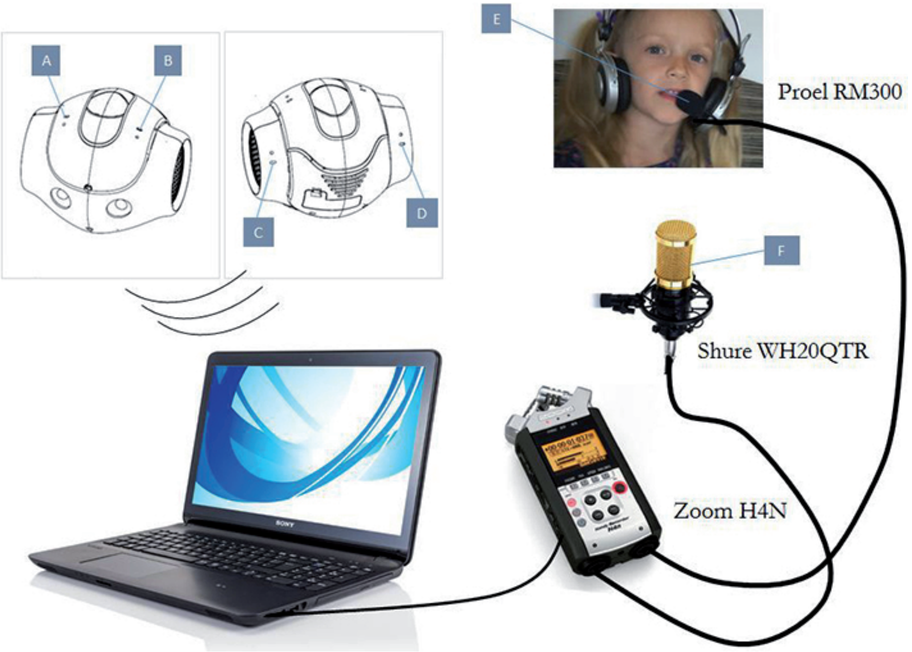
One of its main achievements has been the collection of various new Italian children's speech annotated corpora (Cosi, Paci, Sommavilla & Tesser, 2015) and in this paper the design, building and development of CHILDIT2, a new read children's speech corpus, is described in detail.

1. *Data Collection*

CHILDIT2 is made up by sentences read by young children, and prompts from the FBK CHILDIT corpus (Gerosa, Giuliani & Brugnara, 2007) have been used. They are phonetically balanced sentences, selected from children's literature.

In the original recording set-up, 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, and for CHILDIT2, only the close talk microphone has been taken into consideration.

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

For all recording sessions, an external Zoom H4N device connected to a laptop computer's USB port has been used (see Fig. 1). A Shure WH20QTR Dynamic Headset or a Proel RM300 close talk microphone, plugged into the Zoom's input, has been indifferently chosen for recording, depending on the different sessions and the audio format is characterized by the following set: Channels: 1, Sample Rate: 16000 (originally 48000), Precision: 16-bit and Sample Encoding: 16-bit Signed Integer PCM.

2. *Final Considerations*

Free available speech data are essential for small labs to build and develop new ASR systems and to improve their knowledge on speech of specific group of people, such as the children one.

As illustrated in previous papers (Cosi, Nicolao, Paci, Sommavilla & Tesser, 2014; Cosi, 2015) the original CHILDIT corpus was quite useful in the past to build children speech ASR systems, and it was extensively tested with various open-source ASR systems producing very good PER (phoneme-error-recognition) results (see Table 1).

Table 1 - *PER (phoneme-error-recognition) for various open-source systems tested on CHILDIT*[1]

CHILDIT	SPHINX	BAVIECA	SONIC	KALDI	KALDI (DNN)
Applied Adaptation Methods	VTLN+MLLR (5 Loops)	MLLR (5 Loops)	VTLN + SMAPLR (5 Loops)	LDA+MLLT SGMM+MMI (4 Loops)	DNN+ SMBR
Baseline	18.7 %	16.9 %	15.03 %	13.8 %	8.5 %
Best Score	17.3 %	14.7 %	12.4 %	8.6 %	8.1 %

In a set of recent and still not published experiments, KALDI (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlicek, Qian, Schwarz, Silovsky, Stemmer & Vesely, 2011; Kaldi ASR-web) was tested on CHILDIT+CHILDIT2. Results, shown in Table 2, are quite better than those obtained with the previous experiments where only CHILDIT was used, showing both the importance of using more data to improve recognition performance and also that the quality of the data in the newly created CHILDIT2 corpus is the same as that of CHILDIT.

Table 2 - *PER (phoneme-error-recognition) for KALDI ASR system tested on CHILDIT+CHILDIT2*

CHILDIT + CHILDIT2	KALDI	KALDI (DNN)
	LDA+MLLT SGMM+MMI (4 Loops)	DNN+ SMBR
	12.5 %	7.9 %
	7.9 %	7.3 %

[1] VTLN,: Vocal Tract Length Normalization; MLLR: Maximum Likelihood Linear Regression SMAPLR: Structural Maximum A Posteriori Linear Regression; LDA: Linear Discriminant Analysis; MLLT: Maximum Likelihood Linear Transform; SGMM: Subspace Gaussian Mixture Models; MMI: Maximum Mutual Information; DNN: Deep Neural Network; SMBR: State-level Minimum Bayes Risk.

CHILDIT2 is freely available to the research community[2] and it is licensed by FBK and ISTC CNR, UOS Padova, under a Creative Commons Attribution-Non-Commercial-Share-Alike 4.0 International License.

Acknowledgements

We acknowledge FBK and in particular, Diego Giuliani, for inspiring and guiding the development of the whole CHILDIT2 project. This work was partially supported by the EU FP7 "ALIZ-E" project (grant number 248116).

Bibliography

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayahuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G. & Humbert, R. (2013). Multimodal Child-Robot Interaction: Building Social Bonds. In *Journal of Human-Robot Interacion*, vol. 1, 2, 33-53.

Cosi, P., Nicolao, M., Paci, G., Sommavilla, G. & Tesser, F. (2014). Comparing Open Source ASR Toolkits on Italian Children Speech. In online proceedings of *4th Workshop on Child Computer Interaction (WOCCI 2014)*, Satellite Event of Interspeech 2014, Singapore, 19 September 2014.

Cosi, P., Paci, G., Sommavilla, G. & Tesser, F. (2015). Building Resources for Verbal Interaction – Production and Comprehension within the ALIZ-E Project. In *Atti AISV 2015, XI Convegno Nazionale dell'Associazione Italiana di Scienze della Voce. "Il farsi e il disfarsi del linguaggio. L'emergere, il mutamento e la patologia della struttura sonora del linguaggio"*, Alma Mater Studiorum, Università di Bologna, 28-30 gennaio 2015.

Cosi, P. (2015). A KALDI-DNN-Based ASR System for Italian Experiments on Children Speech. In CD-Rom *Proceedings of IJCNN 2015*, Killarney, Ireland, 12-17 July 2015, CD-paper 15079.

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

KALDI ASR. http://kaldi-asr.org.

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.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii, US, December 2011.

[2] For further info mail-to: piero.cosi@pd.istc.cnr.it.

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, 2013.