PIERO COSI [1], GIULIO PACI [2], GIACOMO SOMMAVILLA [2], FABIO TESSER [1]

CHILDIT2 – A New Children Read Speech Corpus

Abstract: 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.

1 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 et al., 2013) 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 various new Italian children's speech annotated corpora (Cosi et al., 2015) and in this paper the design, building and development.of CHILDIT2, a new read children's speech corpus, is described in detail.

[1] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche - Unità Organizzativa di Supporto di Padova – Italy
http://www.pd.istc.cnr.it
[piero.cosi, fabio.tesser]@pd.istc.cnr.it.

[2] MIVOQ S.R.L, Padova – Italy
http://www.mivoq.it

[giulio.paci, giacomo.sommavilla]@mivoq.it

2 Data Collection

CHILDIT2 is made up by sentences read by young children, and prompts from the FBK CHILDIT corpus (Gerosa et al., 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:

- Channels: 1
- Sample Rate: 16000 (originally 48000)
- Precision: 16-bit
- Sample Encoding: 16-bit Signed Integer PCM

3 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 et al., 2015), (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).

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 %

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

In a set of recent and still not published experiments, KALDI 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.

CHILDIT + CHILDIT2	KALDI	KALDI (DNN)
	12.5 %	7.9 %
	7.9 %	7.3 %

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

CHILDIT2 is freely available to the research community and it is licensed by FBK and ISTC CNR, UOS Padova, under a Creative Commons Attribution-NonCommercial-ShareAlike 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).

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.

Benoıt, 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. doi: 10.1016/0167-6393(96)00026-X

Cosi, P., Nicolao, M., G., Paci, G., Sommavilla, G., Tesser, F. (2014), Comparing Open Source ASR Toolkits on Italian Children Speech, in onLine Proceedings of WOCCI 2014, 4th Workshop on Child Computer Interaction, Satellite Event of INTERSPEECH 2014, Singapore, September 19, 2014,

Cosi, P., G., 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, 12-17 July 2015, Killarney, Ireland, CD-paper 15079.

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