10.12871/clicit2014223

Forced Alignment on Children Speech

Piero Cosi
Vincenzo Galatà
ISTC-CNR, UOS Padova
Via Martiri della libertà, 2
35137 Padova, Italy
piero.cosi@pd.istc.cnr.it
vincenzo.galata@pd.istc.cnr.it

Francesco Cutugno
Antonio Origlia
Dip.Sc.Fisiche Sez.Informatica,
Università di Napoli "Federico II",
Via Cinthia, I-80126 Napoli, Italy
cutugno@unina.it
antonio.origlia@unina.it

Abstract

English. In this Forced Alignment on Children Speech (FACS) task, systems are required to align audio sequences of children read spoken sentences to the provided relative transcriptions, and the task has to be considered speaker independent.

Italiano. *In questo task di EVALITA 2014 dal nome "Forced Alignment on Children Speech" (FACS), tradotto in "Allineamento Forzato su Parlato Infantile", ai partecipanti è stato richiesto di allineare alcune sequenze audio di parlato letto infantile alle corrispondenti trascrizioni fonetiche. I sistemi in esame sono da considerarsi indipendenti dal parlatore.*

1 Introduction

As with other international evaluation campaigns, guidelines describing the FACS task were distributed among the participants, who were also provided with training data and had the chance to test their systems with the evaluation metrics and procedures used in the formal evaluation. As for FACS, two subtasks were defined, and applicants could choose to participate in any of them:

- phone segmentation
- word segmentation

Two modalities were allowed:

- **closed**: only distributed data are allowed for training and tuning the system
- **open**: the participant can use any type of data for system training, declaring and describing the proposed setup in the final report.

The final formal evaluation is based on Unit Boundary Positioning Accuracy. The evaluation methodology follows the standard described in the documentation of the NIST SCLite evaluation tool (NIST, 2015). The SCLite tool itself was used as scorer.
Finally, there was only one participant for the FACS task and this was the SPPAS system by Brigitte Bigi (Bigi, 2012).

2 Data

Training and development data were available quite in advance of test data and participant had only one week to submit their system results to organizers.

2.1 Training data (adult speech)

About 15 map task dialogues recorded by couples of speakers exhibiting a wide variety of Italian variants from the CLIPS corpus (Savi, Cutugno, 2009). Dialogues length ranges from 7/8 minutes to 15/20 minutes. It is up to participants to split these data in train and development subsets. For each dialogue, the following files are provided:

- full dialogue manually performed transcriptions;
- single turn audio files: PCM-encoded mono WAV files (16KHz). Each file is referenced to turns into the full transcription by means of its name;
- single turn phonetic labeling;
- single turn word labeling.

2.2 Training data (children speech)

About 40 sentences read by 20 female and 20 male children speakers taken from the new CHILDIT-2 corpus (Cosi et al., 2015a) collected by ISTC CNR within the ALIZ-E Project (Cosi

et al., 2015b). Sentences length ranges from 2/3 seconds to 5/6 seconds. It is up to participants to split these data in train and development subsets. For each sentence, the following files are provided:

- full sentences automatic performed transcriptions;
- audio files: PCM-encoded mono WAV files (16KHz). Each file is referenced to turns into the full transcription by means of its name;
- phonetic labeling;
- word labeling,

2.3 Test data (children speech)

About 20 sentences read by 5 unseen new female and 5 unseen new male children speakers from the same CHILDIT-2 training corpus cited above. Sentences length ranges from 2/3 seconds to 5/6 seconds. For each sentence, the following files are provided:

- full sentences automatic performed transcriptions;
- audio files: PCM-encoded mono WAV files (16KHz). Each file is referenced to turns into the full transcription by means of its name.

2.4 Reference data (children speech)

Reference transcriptions were automatically created by a recent KALDI ASR system trained on the FBK CHILDIT corpus. The performances of this system are up to now the best obtained so far on this type of material (Cosi et al., 2015b).

3 Test and Results

As previously stated, unaligned phonetic transcription for each file was provided together with the corresponding wav waveform. The reference phonetic transcription we used for the final evaluation did not contain phones that were not actually pronounced. For the evaluation, we used the SCLite tool from the NIST SCTK toolset (NIST, 2015). Participants were requested to send back to the organizers the results of the alignment process in the same format that was used in the training set. Transcriptions were then converted in the CTM format used to perform evaluation by the SCLITE tool. This was to ensure that the conversion from samples to time instants for the boundary markers would have been performed on the same machine for all the participants and for the reference transcription.

The BNF of the CTM format is defined as follows:

CTM :==< F >< C >< BT >< DUR > phoneme

where :

< F >: the waveform filename;
< C >: the waveform channel;
< BT >: the begin time (seconds) of the phoneme, measured from the start of the file;
< DUR >: the duration (seconds) of the phoneme.

Among the transcription rules, it is relevant to note that the same symbol was used for geminates and short consonants. Only 5 vowels were considered, thus eliminating the difference of open and closed feature. A single allophone was considered bot for nasal phoneme m and n.
The SCLite tool was used to perform the time-mediated alignment (TMA) between the reference and hypothesis files and the phoneme-to-phoneme distance was replaced by the following formulas:

D(correct) = |T1(ref)-T1(hyp)| + |T2(ref)-T2(hyp)|
D(insertion) = T2(hyp)-T1(hyp)
D(deletion) = T2(ref)-T1(ref)
D(substit.) = |T1(ref)-T1(hyp)| + |T2(ref)-T2(hyp)| + 0.001

In this mode, the weights of the phoneme-to-phoneme distances are calculated during the alignment based on the markers distance instead of being preset. Results obtained by the only system participating to FACS on the phone alignment task are presented in Table 1 for three different conditions. The ”Closed A” model was trained using CHILDIT-2 and CLIPS corpora, the ”Closed B” model using only CHILDIT-2 and the ”Open” model using both CHILDIT-2 and CLIPS corpora plus a free corpus available on the web named ”read-Torino”, available at http://sldr.org/ortolang-000894.

	Corr	Sub	Del	Ins	Err	S Err
open	96.7	1.2	2.1	1.1	4.4	48.6
closedA	96.8	1.1	2.1	1.1	4.3	49.8
closedB	96.9	1.2	2.0	1.0	4.1	48.6

Table 1. SCLite Time Mediated Alignment results for the open, closedA, amd closedB case.

Results in Table 2 refer instead to the % of markers correctly assigned within 5, 10, 15, 20, 25 ms.

	5ms	10ms	15ms	20ms	25ms
open	43.5	58.7	75.7	85.5	90.3
closedA	45.2	60.6	77.1	86.7	91.1
closedB	43.7	59.2	76.3	85.9	90.6

Table 2. Percentage of markers correctly assigned within 5,10,15,20,25 ms for the open, closedA, amd closedB case.

4 Conclusion

The main aim of this task was to investigate force alignment techniques on read children speech. We explicitly avoid using spontaneous speech in order to evaluate the force alignment of only children speech quality, without considering the difficulties of having to tackle the problem of elisions, insertions, non-verbal sounds, uncertain category assignments, false starts, repetitions, filled and empty pauses and all similar phenomena typically encountered in spontaneous speech. The SPPAAS systems obtained reasonable high performances in all three presented conditions, and results are quite comparable to the state of the art in other languages. Due to the read speech material, reducing the phone inventory to the target one resulted in no difficulties in the alignment task and, even if it is not statistically significant, a dedicated system (closedB case) resulted the best in term of TMA SCLITE alignment errors.

Unfortunately, the SPPAAS system was the only one participating to the FACS task, thus an incomplete analysis of FACS on children speech had been possible because of the lack of comparison of different systems and techniques.

References

NIST (2015), NIST Scoring Toolkit Version 0.1, ftp://jagar.ncsl.nist.gov/current_docs/sctk/doc/sctk.htm

Brigitte Bigi. 2012. SPPAS: a tool for the phonetic segmentations of Speech. In: *Proceedings of LREC 2012, the eight international conference on Language Resources and Evaluation*, Istanbul (Turkey), 1748-1755, ISBN 978-2-9517408-7-7.

Renata Savy, Francesco Cutugno. 2009. CLIPS: Diatopic, Diamesic and Diaphasic Variations of Spoken Italian. In: *Proceedings of Corpus Linguistics Conference 2009*, http://ucrel.lancs.ac.uk/publications/cl2009/213_FullPaper.doc

Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser. 2015a. Building Resources for Verbal Interaction Production and Comprehension within the Project ALIZ-E. In: *Proceedings of AISV 2015* (to be published - 2015).

Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser. 2015. KALDI: Yet Another AST Toolkit? Experiments on Adult and Children Italian Speech. In: *Proceedings of AISV 2015* (to be published - 2015).