EVALITA 2011: Forced alignment task

Francesco Cutugno[1], Antonio Origlia[1], and Dino Seppi[2]

[1] LUSI-lab, Department of Physics, University of Naples “Federico II”, Naples, Italy
[2] ESAT, Katholieke Universiteit Leuven, Belgium
cutugno@unina.it antonio.origlia@unina.it Dino.Seppi@esat.kuleuven.be

Abstract. Forced alignment both for words and phones is a challenging and interesting task for automatic speech processing systems because the difficulties introduced by natural speech are many and hard to deal with. Furthermore, forced alignment approaches have been tested on Italian just in a few studies. In this task, the main goal was to evaluate the performance offered by the proposed systems on Italian and their robustness in presence of noisy data.

Keywords: Forced alignment, word alignment, phone alignment

1 Task description

For the Forced Alignment task of EVALITA 2011, the participants were asked to align spontaneous utterances to the respective manual transcriptions. Both the speech dataset and its transcription have been provided by the organizers. In other words, the participants are requested to estimate and provide the sequence of $t_s(i)..t_e(i)$ for each phoneme and word i in the utterance, where t_s and t_e define the phoneme/word first and last frame in the audio chunk.

More specifically, two *subtasks* are defined and applicants could choose to take part to one or both of them:

word segmentation: speech transliterations are provided;
phone segmentation: speech phonetic transcriptions are provided.

All participants submitted results for both subtasks. Furthermore, the participants could chose between two *modalities*:

closed: only the training data provided by the organizers could be exploited to train and tune the systems;
open: the participant could extend the provided training material with additional data.

All participants submitted results for the first of these two modalities. Additionally, two participants embraced also the second one: one system was been trained on third-party data and tuned on the distributed dataset; in the second system adaptation models were estimated on a third dataset.

2 Dataset

The training set for the EVALITA forced alignement task contains 8064 training units. Each participant should extract a portion of this data set and use it as development set. Each unit contains data regarding exactly one dialogic turn and comprises three files

- A .wav file containing the utterance (clean speech, close mic, high quality).
- A .wrd file containing the transcription of the utterance at word level
- A .phn file containing the transcription of the utterance at phone level

.wav files are encoded as PCM 16Khz mono. More details regarding each speaker (Gender, age, birthplace...) can be found in the header included in the dialogue description files (TXT) along with the full dialogue transcription. Transcription files were encoded in TIMIT format [5].

3 Tests and results

The test set for the EVALITA 2011 Forced Alignment task consisted of 89 wav files containing approximately 10 minutes of natural speech. These were extracted from a set of previously unpublished dialogues collected for the CLIPS corpus. Unaligned word level transcription for each file was provided. Regarding the phonetic transcription, we let the participants choose if they wanted to employ their own automatic transcription system or if they preferred to have a dictionary. All the participants chose to use their own transcription system. The reference phonetic transcription we used for the final evaluation did not contain phones that were not actually pronounced. Should the speakers have introduced sounds other than the expected ones, these were not included in the transcription. This was introduced to seek the participants to take into account different models of pronunciation for the same word. Differences were introduced by different dialects and by natural language reduction phenomena.

For the evaluation, we used the SCLITE and SC_STATS tools from the NIST SCTK toolset [6]. 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. Among the trascription rules, it is relevant to note that the same symbol was used for geminates and short consonants. Since different solutions were provided to the neighbouring vowels problem, to evaluate the alignment performances by taking into account all of these different solutions, we decided to include in the reference transcription a number of possible alternatives for the tool to choose from in particularly difficult cases. To do this, we used the syntax provided by the CTM file format and interpreted by the tool during the scoring process. Specifically, we chose to provide alternate transcriptions for the following cases:

- Sequence /t/tS/ or /d/dZ/ expected: using a single /tS/ or /dZ/ interval is allowed
- Neighbouring vowels separated by a glottal stop: both coupled and splitted version are accepted
- Three neighbouring vowels giving different groupings: all possible groupings involving the specified symbols are accepted
- Three or more neighbouring vowels: a single segment spanning the whole vowel interval is marked as a single substitution error instead of a sequence of deletions because this has been realized using unspecified symbols.

3.1 Word Alignment (closed mode)

The SCLITE tool was used to perform the time-mediated alignment between the reference and hypothesis files. In this mode, the weights of the word-to-word distances are calculated during the alignment on the basis of the markers distance instead of being preset. Results obtained by the systems on the word alignment task in closed mode are presented in Table 1. The SC_STATS tool was used to check the statistical differences among the proposed approaches. The standard Matched Pair Sentence Segment (MPSS) test was used to compare the different systems in the word alignment task. Results of the test in closed mode are presented in Table 2.

Table 1. Summary of the word alignment task results in closed mode

	Corr.	Sub.	Del.	Ins.	Err.	S. Err.
Bigi	97.6	1.0	1.4	1.4	3.8	17.8
Ludusan (5ms)	99.3	0.1	0.5	0.5	1.2	6.7
Ludusan (10ms)	99.2	0.2	0.6	0.6	1.4	7.8
Paci	98.4	0.4	1.2	1.2	2.8	16.7

Table 2. MPSS comparison table for the word alignment task in closed mode. A difference is marked if it can be detected at least at 95% confidence level.

statistically better than ↓	Ludusan (5ms)	Ludusan (10ms)	Bigi	Paci
Ludusan (5ms)		No	No	No
Ludusan (10ms)	No		No	No
Bigi	Yes	Yes		No
Paci	Yes	No	No	

3.2 Phone alignment (closed mode)

Overall results obtained by the systems on the phone alignment task in closed mode are presented in Table 3. For this task the Friedman two-way ANOVA by

Rank test was used. This is because of the presence of alternative transcriptions in the reference phone alignment task: the ANOVA test does not assume a single reference transcription while the MPSS test does. Results obtained in closed mode are presented in Table 4.

Table 3. Summary of the phone alignment task results in closed mode

	Corr.	Sub.	Del.	Ins.	Err.	S. Err.
Bigi	83.7	11.3	5.0	4.9	21.2	93.9
Ludusan (5ms)	93.0	5.0	2.0	8.1	15.1	80.5
Ludusan (10ms)	93.9	4.9	1.2	7.2	13.3	79.8
Paci	92.4	5.9	1.7	4.5	12.1	81.0

Table 4. Results of the ANOVA test for the phone alignment task in closed mode (confidence 95%).

statistically better than ↓	Ludusan (5ms)	Ludusan (10ms)	Bigi	Paci
Ludusan (5ms)		Yes	No	Yes
Ludusan (10ms)	No		No	No
Bigi	Yes	Yes		Yes
Paci	No	No	No	

3.3 Word alignment (open mode)

Overall results obtained on the word alignment task in open mode are shown in Table 5. The MPSS test on the word alignment task in open mode did not detect any statistically relevant difference among the proposed systems.

Table 5. Summary of the word alignment task results in open mode

	Corr.	Sub.	Del.	Ins.	Err.	S. Err.
Ludusan (5ms + VTLN)	99.0	0.2	0.8	0.8	1.8	10.0
Ludusan (10ms + VTLN)	99.3	0.2	0.5	0.5	1.2	5.6
Paci	97.4	1.2	1.5	1.5	4.1	14.4

3.4 Phone alignment (open mode)

Overall results obtained on the phone alignment task in open mode are shown in Table 6. The ANOVA test on the phone alignment task in open mode did not detect any statistically significant difference among the proposed systems.

Table 6. Summary of the phone alignment task results in open mode

	Corr.	Sub.	Del.	Ins.	Err.	S. Err.
Ludusan (5ms + VTLN)	93.0	5.2	1.8	8.2	15.1	81.6
Ludusan (10ms + VTLN)	93.6	5.1	1.3	7.2	13.6	79.1
Paci	90.6	7.3	2.1	4.6	13.9	81.3

4 Discussion

Main aim of this task was to investigate force alignment techniques of spontaneous speech. The chosen speech material was derived by the dialogue subsection of the CLIPS corpus [7], and presented some intrinsic degree of complexity for the forced alignment task. In particular: a) all glides, diphthongs and vowel clusters in general were not segmented, both across and within words, as corpus designers considered too arbitrary to put one or more segment boundary within such a continuous where no specific evidence of sound change could be assigned to a specific instant; b) every evidence of phonetic reduction was marked with some specific symbol. Elisions, insertions, non-verbal sounds, uncertain category assignments, false starts, repetitions, filled and empty pauses and all similar phenomena typically encountered in the spontaneous speech, were marked and labeled in some way; c) the dialogues were recorded in different regions of Italy and consequently presented a wide variability on the diatopic plane.

Even if none of the speaker used her/his specific local dialect, Italian produced in the different dialogues in the various places of recording introduced a wide variety of pronunciation for each speech sound in the various words. A further step into complexity was then generated by the processing that us, as task organizers, introduced during the preparation of the train and test material. We decided to suppress most of the indications related to phonetic reduction originarily available changing the speech sound label of any 'degenerated' speech sound into a garbage symbol #. This means that participants to the task could train their acoustic models only with those sounds that CLIPS labelers identified as more close to the prototypical form. A consequence of this two participants on the three involved in the task decided to manually intervene on the material to add information on the audio portion labeled as garbage. This was the main corrective procedure that was introduced by the contestants, only one of them successively used these new data to introduce alternative pronunciation/transcriptions models in the vocabulary. Original CLIPS dialogues were recorded on two separated channels and divided in turns based on the orthographic transcription. In principle this should guarantee that speakers' voices could be listened only in the respective channels. One participant reported the presence of a phantom audio due to the second speaker in the regions in which the first was silent. The participant was afraid that this could introduce alteration in the Markov Chain train process, for this reason he made an attempt to automatically hide these audio portions during training. In this case we can empirically evaluate differences between this condition and the one in which this correction was not performed and results appear as not easily differentiable. One participant, as suggested by specific literature about ASR robustness on spontaneous speech, tried to reduce the advancing step of the analysis window during features extraction: he compared a typical 10 ms advancement with a more fine 5 ms step. Differences, however, do not appear to be relevant with the exception of the closed mode phone alignment task in which the version employing 10ms performed better than the version using a time step of 5ms in a statistically significant way. This result, however, has to be taken carefully because of the confidence value employed during the test. This said, all three contestants obtained results that can be considered very close to the state of art for other languages.

Concerning the state of the art in Italian, there is almost nothing in literature, especially for what word alignment concerns. References for this task can be only found in [4,2,1,3]. Difficulties introduced by our processing at phonetic level made, in general, the phones alignment task less performing. Bigi states that results of this task suggest that the classic approach based on vocabularies and expected pronunciations is to be at least refined, if not deeply reviewed, when phonetic alignment approaches spontaneous speech data, and we agree on this. However we are faced with the very poor availability of speech data manually labeled that are an unrenounceable requirement for this task. At the same time the scientific community still lacks in finding an agreement, especially for Italian, on what a phonetic reduction is, and, even more important, which standard annotation system must be used to describe all the phenomena that fall under the name of reduction.

References

1. Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., Omologo, M.: Automatic segmentation and labeling of english and italian speech databases. In: Proc. of Eurospeech. pp. 653–656 (1993)
2. Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., Omologo, M.: A baseline of a speaker independent continuous speech recognizer of italian. In: Proc. of Eurospeech. pp. 847–850 (1993)
3. Cangemi, F., Cutugno, F., Ludusan, B., Seppi, D., Van Compernolle, D.: Automatic speech segmentation for italian (ASSI): tools, models, evaluation and applications. In: Proc. of AISV (2011)
4. Cosi, P., Falavigna, D., Omologo, M.: A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In: Proc. of Eurospeech. pp. 693–696 (1991)
5. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993)
6. NIST: NIST SCTK Toolkit, ftp://jaguar.ncsl.nist.gov/current_docs/sctk/doc/sctk.htm
7. Savy, R., Cutugno, F.: CLIPS: diatopic, diamesic and diaphasic variations of spoken italian. In: Proc. of Corpus Linguistics Conference (2009), http://ucrel.lancs.ac.uk/publications/cl2009/213_FullPaper.doc